Stochastic process-based modelling for hydrological systems

Georgia Papacharalampous

PhD thesis

National Technical University of Athens
School of Civil Engineering
Department of Water Resources and Environmental Engineering

Stochastic process-based modelling for hydrological systems

Διατριβή

για την απόκτηση του τίτλου της
Διδάκτορος Μηχανικού
από το Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ)
μετά από δημόσια υπεράσπιση

την Παρασκευή, 23 Οκτωβρίου 2020 ώρα 15:00

από

την Γεωργία Παπαχαραλάμπους

διπλωματούχο Πολιτικό Μηχανικό, ΕΜΠ (2014) και
απόφοιτο του Διατμηματικού Προγράμματος
Μεταπτυχιακών Σπουδών (ΔΠΜΣ)
«Επιστήμη και Τεχνολογία Υδατικών Πόρων», ΕΜΠ (2016)

Advisory committee

Professor Demetris Koutsoyiannis (Supervisor)	National Technical University of Athens
Professor Alberto Montanari	University of Bologna
Associate Professor Nikos Mamassis	National Technical University of Athens

Examination committee

Professor Demetris Koutsoyiannis (Supervisor)	National Technical University of Athens
Professor Alberto Montanari	University of Bologna
Associate Professor Nikos Mamassis	National Technical University of Athens
Professor Bellie Sivakumar	Indian Institute of Technology Bombay
Associate Professor Nikos Lagaros	National Technical University of Athens
Associate Professor Andreas Langousis	University of Patras
Associate Professor Vissarion Papadopoulos	National Technical University of Athens

This thesis has been conducted at the Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens.

Keywords: Autoregressive fractionally integrated moving average; benchmarking time series forecasts; case studies; cross-case synthesis; generalized random forests; gradient boosting machine; ensemble learning; hydrological model; hyperparameter optimization; lagged variable selection; large-scale hydrology; machine learning; multi-step ahead forecasting; neural networks; no free lunch theorem; one-step ahead forecasting; precipitation forecasting; probabilistic prediction; Prophet; quantile averaging; quantile regression; quantile regression forests; quantile regression neural networks; random forests; river discharge; simple exponential smoothing; stochastic hydrology; support vector machines; temperature forecasting; time series; time series forecasting; uncertainty quantification

The research work was supported by the Hellenic Foundation for Research and Innovation (HFRI) under the HFRI PhD Fellowship grant (Fellowship Number: 1388).

This thesis has been written using Microsoft Office and printed in Athens.

To my beloved people

Whatever you do, work at it with all your heart.

Acknowledgements

I would like to thank several people, institutions and organizations for supporting me, directly or indirectly, as a PhD student.

I wish to greatly thank Professor Demetris Koutsoyiannis for his sincere support, advice and collaboration during my PhD studies. I wish to warmly thank him for offering me the opportunity for this PhD, for always being available and happy to advise me, for extending my original research interests with the "MK blueprint" for probabilistic hydrological modelling, and for progressively accepting (or continuously trying to accept) my interest in machine learning algorithms and ARFIMA models. I also wish to thank him for sharing with me his strong opinions and scholarly publishing experiences, along with his input to our research works during our meetings. All that he has done for me during this PhD are greatly appreciated.

I also wish to extend my sincere gratitude to Professor Alberto Montanari and Associate Professor Nikos Mamassis. Their kind support, advice and collaboration are greatly appreciated and valued. I would like to warmly thank Professor Montanari for patiently answering my very initial questions on the MK blueprint, for carefully checking several paper drafts, for his continued promptness to discuss and brainstorm on our research works through email, and for always being kind and friendly. I wish to thank Professor Mamassis for his valuable practical advice and support during my PhD studies, for some extensive discussions on how good theses are written, for his feedback to our research works, and finally for his bright positivity.

I am sincerely thankful and appreciative to the members of the examination committee, Professor Bellie Sivakumar, Associate Professor Nikos Lagaros, Associate Professor Andreas Langousis and Associate Professor Vissarion Papadopoulos, for their thoughtful acceptance to examine this thesis, and their good example and kindness.

I am deeply grateful to my main collaborator and boyfriend, Dr Hristos Tyralis. I wish to wholeheartedly thank Hristos for his invaluable partnership and advice, for being my mentor and my friend, for all the creative moments that we share each and every day of our lives, and (of course) for his love, trust and tremendous support. Once upon a time, I was an MSc student and Hristos turned me into a "happy researcher". He passed me on his programming experiences, he patiently guided me in writing our first common paper, and very soon, he considered me as his close research collaborator. This PhD is largely a product of our close collaboration and idea sharing, and I am really happy to know that he likes its outcomes. Hristos was always there for me during my best and worst PhD moments, and never lost his trust on our research works. He instead made me feel that we are looking together at the same direction. His guidance gave me clarity and trust on our research works. His trust gave me confidence and strength. His interest gave me energy and motivation. There are many other reasons for me to thank Hristos, but this is just the preface of a PhD thesis.

During my PhD, I had also the chance to interact and collaborate with Associate Professor Andreas Langousis, Professor Amithirigala Jayawardena and Professor Bellie Sivakumar, whom I wish to greatly thank (additionally) for their big support and fruitful discussions on our co-authored paper. Constructive comments by the editors and reviewers of the seven papers included in this PhD thesis, as well as the thoughtful and detailed reviews on drafts of the thesis by the advisory and examination committees, have considerably contributed to the thesis' final version, and are therefore gratefully acknowledged in this preface.

Moreover, I wish to warmly thank Lia Theofanidou, Olga Kitsou, Nefeli Lagopati and Maria Dromazou for their valuable help with my PhD applications and paperwork, and for their kindness and positive attitude.

PhD is about research, but research can be made without holding a PhD. The greatest truth-seeker and self-taught researcher of my life is definitely my mom. She has made huge sacrifices for supporting my previous studies and this PhD, and gave me this big chance with her whole heart. She has greatly contributed to forming my passions and interests, and I think that this PhD

degree should make her more proud of her big efforts. I wish to also greatly thank my dad, who is deeply loved and missed, and my brother Andreas for his sincere love. A big "thank you" should, in fact, go to all the people making my everyday life beautiful and interesting. Among them are also my friends Martha, Antonia, Magda, Anna and Dimitris, my cousin Callie, my aunt Jan, my uncle Manolakis, my uncle Thalassinos, my godmother Kiki, and Hristos' family.

I am also very grateful for having received my PhD financial support. My PhD studies have been primarily supported by the Hellenic Foundation for Research and Innovation (November 2019 – October 2020). As a PhD student and early-career scientist, I have also received the «Thomaidis award» for scientific papers from National Technical University of Athens (2017), conference participation funding from the same institution (2018), the «Early Career Scientist's Travel Support» from the European Geoscience Union (2018), and conference registration funding from the Hydrology Section of the American Geoscience Union and the Consortium of Universities for the Advancement of Hydrologic Sciences (2020). As regards the received «Early Career Scientist's Travel Support», I am very grateful to the Conveners Serena Ceola, Demetris Koutsoyiannis, Alberto Montanari, Christophe Cudennec and Harry Lins for their positive evaluation of my application. Their kind gesture is very much appreciated.

Other forms of support that should be acknowledged in this preface are the invitation by the organizing committee of the «10th World Congress on Water Resources and Environment. "Panta Rhei"» to submit our work presented in the congress (and included in this thesis) to the Special Issue «Water Resources and Environment» of the «Water Resources Management» Journal (2017), and the open-access publication funding received from the Asia Oceania Geosciences Society (2018) and the «Water» Journal, MDPI (2019) for two papers included in this thesis.

Finally, I wish to express my sincere gratitude for having been awarded with the International Scientific Prize of the Dimitris N. Chorafas Foundation (2020) for the best graduating doctorate students in 21 selected universities in Europe, North America and Asia (including the National Technical University of Athens). I am deeply grateful to the Nominating Committee of the National Technical University of Athens and the Foundation's Committee for respectively nominating and awarding me (and this thesis) with this honourable prize.

Publications

The below list contains the publications of the author during her time as a PhD student.

Publications in scientific journals

Included in the PhD thesis

[1] **Papacharalampous GA**, Koutsoyiannis D, Montanari A (2020) Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models. Advances in Water Resources 136:103471. https://doi.org/10.1016/j.advwatres.2019.103471

[2] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D, Montanari A (2020) Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. Advances in Water Resources 136:103470. https://doi.org/10.1016/j.advwatres.2019.103470

[3] **Papacharalampous GA**, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019) Probabilistic hydrological post-processing at scale: Why and how to apply machine learning quantile regression algorithms. Water 11(10):2126. https://doi.org/10.3390/w11102126

[4] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2019) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Stochastic Environmental Research and Risk Assessment 33(2):481–514. https://doi.org/10.1007/s00477-018-1638-6

[5] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2018) Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. Water Resources Management 32(15):5207–5239. https://doi.org/10.1007/s11269-018-2155-6

[6] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2018) Predictability of monthly temperature and precipitation using automatic time series forecasting methods. Acta Geophysica 66(4):807–831. https://doi.org/10.1007/s11600-018-0120-7

[7] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2018) One-step ahead forecasting of geophysical processes within a purely statistical framework. Geoscience Letters 5(1):12. https://doi.org/10.1186/s40562-018-0111-1

Not included in the PhD thesis

[8] Tyralis H, **Papacharalampous GA**, Langousis A (2020) Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05172-3

[9] **Papacharalampous GA**, Tyralis H (2020) Hydrological time series forecasting using simple combinations: Big data testing and investigations on one-year ahead river flow predictability. Journal of Hydrology 590:125205. https://doi.org/10.1016/j.jhydrol.2020.125205

[10] Tyralis H, **Papacharalampous GA**, Burnetas A, Langousis A (2019) Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. Journal of Hydrology 577:123957. https://doi.org/10.1016/j.jhydrol.2019.123957

[11] Blöschl G, **et al.** (2019) Twenty-three Unsolved Problems in Hydrology (UPH) – A community perspective. Hydrological Sciences Journal 64(1):1141–1158. https://doi.org/10.1080/02626667.2019.1620507

[12] Tyralis H, **Papacharalampous GA**, Langousis A (2019) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910

[13] Tyralis H, **Papacharalampous GA**, Tantanee S (2019) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. Journal of Hydrology 574:628–645. https://doi.org/10.1016/j.jhydrol.2019.04.070

[14] **Papacharalampous GA**, Tyralis H (2018) Evaluation of random forests and Prophet for daily streamflow forecasting. Advances in Geosciences 45:201–208. https://doi.org/10.5194/adgeo-45-201-2018

[15] Tyralis H, **Papacharalampous GA** (2018) Large-scale assessment of Prophet for multi-step ahead forecasting of monthly streamflow. Advances in Geosciences 45:147–153. https://doi.org/10.5194/adgeo-45-147-2018

[16] Tyralis H, **Papacharalampous GA** (2017) Variable selection in time series forecasting using random forests. Algorithms 10(4):114. https://doi.org/10.3390/a10040114

[17] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) Forecasting of geophysical processes using stochastic and machine learning algorithms. European Water 59:161–168

Working publications in scientific journals

Not included in the PhD thesis

[18] **Papacharalampous GA**, Tyralis H, Papalexiou SM, Langousis A, Khatami S, Volpi E, Grimaldi S (2020) Global-scale massive feature extraction from monthly hydroclimatic time series: Statistical characterizations, spatial patterns and hydrological similarity. https://arxiv.org/abs/2010.12833

[19] Tyralis H, **Papacharalampous GA** (2020) Boosting algorithms in energy research: A systematic review. https://arxiv.org/abs/2004.07049

Book chapters

Not included in the PhD thesis

[20] Tyralis H, **Papacharalampous GA**, Langousis A (2020) Streamflow forecasting at large time scales using statistical models. In: Sharma P, Machiwal D (Eds) Advances in Streamflow Forecasting, Elsevier. In press

Fully evaluated conference publications

Not included in the PhD thesis

[21] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2018) Error evolution patterns in multi-step ahead streamflow forecasting. 13th International Hydroinformatics Conference, Palermo, Italy:1598–1607. https://doi.org/10.29007/84k6

Popular science publications

Not included in the PhD thesis

[22] **Papacharalampous GA**, Tyralis H (2020) Machine learning for probabilistic hydrological forecasting. HEPEX blog post. https://hepex.inrae.fr/machine-learning-for-probabilistic-hydrological-forecasting

Conference publications and presentations with evaluation of abstract

Included in the PhD thesis

[23] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D, Montanari A (2021) Large-scale calibration of conceptual rainfall-runoff models for two-stage probabilistic hydrological post-processing. To be submitted

[24] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D, Montanari A (2020) The wisdom of the crowd in probabilistic predictive modelling: Large-scale application to monthly rainfall-runoff problems. American Geoscience Union Fall Meeting 2020, San Francisco, USA. Accepted for presentation

[25] **Papacharalampous GA**, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019) Large-scale comparison of machine learning regression algorithms for probabilistic hydrological modelling via post-processing of point predictions. European Geosciences Union General Assembly 2019, Vienna, Austria: EGU2019-3576. https://doi.org/10.6084/m9.figshare.8018342.v1

[26] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2018) A step further from model-fitting for the assessment of the predictability of monthly temperature and precipitation. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-864. https://doi.org/10.6084/m9.figshare.7325783.v1

[27] **Papacharalampous GA**, Koutsoyiannis D, Montanari A (2018) Toy models for increasing the understanding on stochastic process-based modelling. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-1900-1. https://www.itia.ntua.gr/1813

[28] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) Large scale simulation experiments for the assessment of one-step ahead forecasting properties of stochastic and machine learning point estimation methods. Asia Oceania Geosciences Society 14th Annual Meeting, Singapore: HS06-A002. https://doi.org/10.13140/RG.2.2.33273.77923

[29] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) Comparison between stochastic and machine learning methods for hydrological multi-step ahead forecasting: All forecasts are wrong!. European Geosciences Union General Assembly 2017, Vienna, Austria: EGU2017-3068-2. https://doi.org/10.13140/RG.2.2.17205.47848

Not included in the PhD thesis

[30] Tyralis H, **Papacharalampous GA**, Langousis A, Burnetas A (2019) Stacking of probabilistic predictions for improving hydrological forecasts. European Geosciences Union General Assembly 2019, Vienna, Austria: EGU2019-2827. https://doi.org/10.6084/m9.figshare.8018510.v1

[31] **Papacharalampous GA**, Tyralis H, Mamassis N (2018) Conceptual hydrological modelling at daily scale: Aggregating results for 340 MOPEX catchments. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-3759. https://doi.org/10.6084/m9.figshare.7325771.v1

[32] Tyralis H, **Papacharalampous GA** (2018) Multi-step ahead forecasting of monthly streamflow discharge time series using a variety of algorithms. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-3571. https://www.itia.ntua.gr/1803

[33] **Papacharalampous GA**, Tyralis H (2018) Illustrating important facts about multi-step ahead forecasting of univariate hydrological time series. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-3570. https://www.itia.ntua.gr/1804

[34] **Papacharalampous GA**, Tyralis H (2018) One-step ahead forecasting of annual precipitation and temperature using univariate time series methods. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-2298. https://doi.org/10.6084/m9.figshare.7325777.v1

[35] **Papacharalampous GA**, Tyralis H (2018) Large-scale assessment of random forests for data-driven hydrological modelling at monthly scale. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-1902. https://doi.org/10.6084/m9.figshare.7325759.v1

[36] Tyralis H, **Papacharalampous GA** (2018) Univariate time series forecasting properties of random forests. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-1901. https://doi.org/10.6084/m9.figshare.4696312.v2

[37] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) A set of metrics for the effective evaluation of point forecasting methods used for hydrological tasks. Asia Oceania Geosciences Society 14th Annual Meeting, Singapore: HS01-A001. https://doi.org/10.13140/RG.2.2.19852.00641

[38] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) Investigation of the effect of the hyperparameter optimization and the time lag selection in time series forecasting using machine learning algorithms. European Geosciences Union General Assembly 2017, Vienna, Austria: EGU2017-3072-1. https://doi.org/10.13140/RG.2.2.20560.92165/1

[39] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) Multi-step ahead streamflow forecasting for the operation of hydropower reservoirs. European Geosciences Union General Assembly 2017, Vienna, Austria: EGU2017-3069. https://www.itia.ntua.gr/1692

Supplementary material to articles (containing codes and data)

Included in the PhD thesis

[40] **Papacharalampous GA**, Tyralis H (2018) Supplementary material for the paper "Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes". Figshare. https://doi.org/10.6084/m9.figshare.7092824

[41] **Papacharalampous GA**, Tyralis H (2018) One-step ahead forecasting of geophysical processes within a purely statistical framework: Supplementary material. Figshare. https://doi.org/10.6084/m9.figshare.5357359.v1

Not included in the PhD thesis

[42] **Papacharalampous GA**, Tyralis H (2018) Error evolution patterns in multi-step ahead streamflow forecasting: Supplementary material. Mendeley Data, v1. https://doi.org/10.17632/dxkm8n3g99.1

[43] Tyralis H, **Papacharalampous GA** (2017) Supplementary material for the paper "Variable selection in time series forecasting using random forests". Mendeley Data, v1. https://doi.org/10.17632/nr3z96jmbm.1

Supplementary material to articles (not containing codes and data)

Included in the PhD thesis

[44] **Papacharalampous GA**, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019) Supplementary material for the paper "Probabilistic hydrological post-processing at scale: Why and how to apply machine learning quantile regression algorithms". Figshare. https://doi.org/10.6084/m9.figshare.9496262.v2

[45] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D, Montanari A (2019) Supplementary material for the paper "Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale". Figshare. https://doi.org/10.6084/m9.figshare.7959473.v2

Not included in the PhD thesis

[46] **Papacharalampous GA**, Tyralis H (2020) Additional material to the paper entitled "Hydrological time series forecasting using simple combinations: Big data testing and investigations on one-year ahead river flow predictability". Mendeley Data, v1. https://doi.org/10.17632/z2mdnnxghg.1

[47] **Papacharalampous GA**, Tyralis H, Koutsoyiannis D (2017) Forecasting of geophysical processes using stochastic and machine learning algorithms: Supplementary material. Mendeley Data, v3. https://doi.org/10.17632/p8sw8pzkcd.3

Abstract

This thesis falls into the scientific areas of stochastic hydrology, hydrological modelling and hydroinformatics. It contributes with new practical solutions, new methodologies and large-scale results to predictive modelling of hydrological processes, specifically to solving two interrelated technical problems with emphasis on the latter. These problems are:

(A) hydrological time series forecasting by exclusively using endogenous predictor variables (hereafter, referred to simply as "hydrological time series forecasting"); and

(B) stochastic process-based modelling of hydrological systems via probabilistic post-processing (hereafter, referred to simply as "probabilistic hydrological post-processing").

For the investigation of these technical problems, the thesis forms and exploits a novel predictive modelling and benchmarking toolbox. This toolbox is consisted of:

(i) approximately 6 000 hydrological time series (sourced from larger freely available datasets),

(ii) over 45 ready-made automatic models and algorithms mostly originating from the four major families of stochastic, (machine learning) regression, (machine learning) quantile regression, and conceptual process-based models,

(iii) seven flexible methodologies (which together with the ready-made automatic models and algorithms consist the basis of our modelling solutions), and

(iv) approximately 30 predictive performance evaluation metrics.

Novel model combinations coupled with different algorithmic argument choices result in numerous model variants, many of which could be perceived as new methods. All the utilized models (i.e., the ones already available in open software, as well as those automated and proposed in the context of the thesis) are flexible, computationally convenient and fast; thus, they are appropriate for large-sample (even global-scale) hydrological investigations. Such investigations are implied by the (mainly) algorithmic nature of the methodologies of the thesis. In spite of this nature, the thesis also provides innovative theoretical supplements to its practical and methodological contribution.

Technical problem (A) is examined in four stages. During the first stage, a detailed framework for assessing forecasting techniques in hydrology is introduced. Complying with the principles of forecasting and contrary to the existing hydrological (and, more generally, geophysical) time series forecasting literature (in which forecasting performance is usually assessed within case studies), the introduced framework incorporates large-scale benchmarking. The latter relies on big hydrological datasets, large-scale time series simulation by using classical stationary stochastic models, many automatic forecasting models and algorithms (including benchmarks), and many forecast quality metrics. The new framework is exploited (by utilizing part of the predictive modelling and benchmarking toolbox of the thesis) to provide large-scale results and useful insights on the comparison of stochastic and machine learning forecasting methods for the case of hydrological time series forecasting at large temporal scales (e.g., the annual and monthly ones), with emphasis on annual river discharge processes. The related investigations focus on multi-step ahead forecasting.

During the second stage of the investigation of technical problem (A), the work conducted during the previous stage is expanded by exploring the one-step ahead forecasting properties of its methods, when the latter are applied to non-seasonal geophysical time series. Emphasis is put on the examination of two real-world datasets, an annual temperature dataset and an annual precipitation dataset. These datasets are examined in both their original and standardized forms to reveal the most and least accurate methods for long-run one-step ahead forecasting applications, and to provide rough benchmarks for the one-year ahead predictability of temperature and precipitation.

The third stage of the investigation of technical problem (A) includes both the examination-quantification of predictability of monthly temperature and monthly precipitation at global scale, and the comparison of a large number of (mostly stochastic) automatic time series forecasting

methods for monthly geophysical time series. The related investigations focus on multi-step ahead forecasting by using the largest real-world data sample ever used so far in hydrology for assessing the performance of time series forecasting methods.

With the fourth (and last) stage of the investigation of technical problem (A), the multiple-case study research strategy is introduced –in its large-scale version– as an innovative alternative to conducting single- or few-case studies in the field of geophysical time series forecasting. To explore three sub-problems associated with hydrological time series forecasting using machine learning algorithms, an extensive multiple-case study is conducted. This multiple-case study is composed by a sufficient number of single-case studies, which exploit monthly temperature and monthly precipitation time series observed in Greece. The explored sub-problems are lagged variable selection, hyperparameter handling, and comparison of machine learning and stochastic algorithms.

Technical problem (B) is examined in three stages. During the first stage, a novel two-stage probabilistic hydrological post-processing methodology is developed by using a theoretically consistent probabilistic hydrological modelling blueprint as a starting point. The usefulness of this methodology is demonstrated by conducting toy model investigations. The same investigations also demonstrate how our understanding of the system to be modelled can guide us to achieve better predictive modelling when using the proposed methodology.

During the second stage of the investigation of technical problem (B), the probabilistic hydrological modelling methodology proposed during the previous stage is validated. The validation is made by conducting a large-scale real-world experiment at monthly timescale. In this experiment, the increased robustness of the investigated methodology with respect to the combined (by this methodology) individual predictors and, by extension, to basic two-stage post-processing methodologies is demonstrated. The ability to "harness the wisdom of the crowd" is also empirically proven.

Finally, during the third stage of the investigation of technical problem (B), the thesis introduces the largest range of probabilistic hydrological post-processing methods ever introduced in a single work, and additionally conducts at daily timescale the largest benchmark experiment ever conducted in the field. Additionally, it assesses several theoretical and qualitative aspects of the examined problem and the application of the proposed algorithms to answer the following research question: *Why and how to combine process-based models and machine learning quantile regression algorithms for probabilistic hydrological modelling?*

Περίληψη

Η παρούσα διδακτορική διατριβή εμπίμπτει στους επιστημονικούς κλάδους της στοχαστικής υδρολογίας, της υδρολογικής μοντελοποίησης και της υδροπληροφορικής. Συνεισφέρει με νέες πρακτικές λύσεις, νέες μεθοδολογίες και αποτελέσματα μεγάλης κλίμακας στην μοντελοποιήση υδρολογικών διεργασιών, συγκεκριμένα στην επίλυση δύο στενά συνυφασμένων τεχνικών προβλημάτων με έμφαση στο δεύτερο. Τα προβλήματα αυτά είναι:

(A) η πρόβλεψη της μελλοντικής συμπεριφοράς υδρολογικών διεργασιών χρησιμοποιώντας αποκλειστικά ενδογενείς μεταβλητές πρόβλεψης (στο εξής αναφερόμενη ως «πρόβλεψη υδρολογικών χρονοσειρών»), και

(B) η στοχαστική μοντελοποίηση υδρολογικών συστημάτων μέσω πιθανοτικής μετεπεξεργασίας αποτελεσμάτων διεργασιακής υδρολογικής μοντελοποιήσης (στο εξής αναφερόμενη ως «πιθανοτική μετεπεξεργασία αποτελεσμάτων υδρολογικής μοντελοποίησης»).

Για τη διερεύνηση των εν λόγω τεχνικών προβλημάτων, αναπτύσσεται και αξιοποιείται εργαλειοθήκη πρότυπης μοντελοποίησης και συγκριτικής αξιολόγησης αποτελούμενη από:

(i) περίπου 6 000 υδρολογικές χρονοσειρές προερχόμενες από μεγαλύτερες ελεύθερα διατιθέμενες βάσεις δεδομένων,

(ii) περισσότερα από 45 αυτοματοποιημένα μοντέλα και αλγορίθμους (διαθέσιμα σε ανοιχτό λογισμικό), τα οποία κατά κύριο λόγο προέρχονται από τις τέσσερις μεγάλες οικογένειες των στοχαστικών μοντέλων, των μοντέλων παλινδρόμησης (συμπεριλαμβανομένων μοντέλων μηχανικής μάθησης), των μοντέλων παλινδρόμησης ποσοστημορίου (συμπεριλαμβανομένων μοντέλων μηχανικής μάθησης) και των διεργασιακών υδρολογικών μοντέλων,

(iii) επτά ευέλικτες μεθοδολογίες, οι οποίες μαζί με τα διαθέσιμα σε ανοιχτό λογισμικό αυτοματοποιημένα μοντέλα και αλγορίθμους (βλ. σημείο (ii) παραπάνω) συνιστούν τη βάση των διενεργούμενων μοντελοποιήσεων, και

(iv) περίπου 30 μέτρα για την αξιολόγηση της ποιότητας των διενεργούμενων μοντελοποιήσεων.

Νέοι συνδυασμοί μοντέλων και αλγορίθμων, συνοδευόμενοι από διαφορετικές αλγοριθμικές επιλογές παραμέτρων, οδηγούν σε πολυάριθμες παραλλαγές μοντέλων, πολλές από τις οποίες μπορούν να θεωρηθούν ως νέες μέθοδοι. Όλα τα χρησιμοποιούμενα μοντέλα (τόσο τα ήδη διαθέσιμα σε ανοιχτό λογισμικό όσο και τα αυτοματοποιημένα στο πλαίσιο της διατριβής) είναι ευέλικτα, υπολογιστικά εύχρηστα και γρήγορα στην εφαρμογή. Κατά συνέπεια, είναι κατάλληλα για διερευνήσεις μεγάλης κλίμακας, ακόμη και για διερευνήσεις παγκόσμιας κλίμακας. Τέτοιες διερευνήσεις επιβάλλονται από τον (κυρίως) αλγοριθμικό χαρακτήρα των μεθοδολογιών της διατριβής. Παρά τον συγκεκριμένο χαρακτήρα, η διατριβή παρέχει επίσης καινοτόμα θεωρητικά συμπληρώματα στην πρακτική και μεθοδολογική της συμβολή.

Η διερεύνηση του τεχνικού προβλήματος (A) γίνεται σε τέσσερα στάδια. Κατά το πρώτο στάδιο εισάγεται ένα νέο μεθοδολογικό πλαίσιο για την αξιολόγηση τεχνικών πρόγνωσης στην υδρολογία. Όντας σύμφωνο με τις αρχές που θα πρέπει να διέπουν την πρόβλεψη χρονοσειρών και σε αντίθεση με την υπάρχουσα βιβλιογραφία της πρόβλεψης υδρολογικών (και γενικότερα γεωφυσικών) χρονοσειρών (στην οποία η αξιολόγηση μεθόδων συνήθως βασίζεται στη διενέργεια μελετών περίπτωσης), το προτεινόμενο πλαίσιο ενσωματώνει συγκριτική αξιολόγηση μεθοδολογιών μεγάλης κλίμακας. Η τελευταία βασίζεται σε μεγάλα σύνολα υδρολογικών δεδομένων, στην πρακτική της στοχαστικής προσομοίωσης χρονοσειρών μεγάλης κλίμακας χρησιμοποιώντας στάσιμα κλασσικά στοχαστικά μοντέλα, σε έναν μεγάλο αριθμό πλήρως αυτοματοποιημένων μοντέλων και αλγόριθμων πρόβλεψης (συμπεριλαμβανομένων μοντέλων αναφοράς) και σε έναν ικανό αριθμό μέτρων για την ποσοτικοποίηση της ποιότητας των προβλέψεων. Το νέο μεθοδολογικό πλαίσιο αξιοποιείται (χρησιμοποιώντας τμήμα της εργαλειοθήκης της διατριβής) για την παροχή αποτελεσμάτων μεγάλης κλίμακας, καθώς και χρήσιμης κατανόησης σχετικά με τη σύγκριση των στοχαστικών μεθόδων και των μεθόδων

μηχανικής μάθησης στην πρόβλεψη υδρολογικών διεργασιών σε μεγάλες χρονικές κλίμακες (π.χ., την ετήσια και την μηνιαία), με έμφαση στις ετήσιες διεργασίες απορροής ποταμών. Οι σχετικές διερευνήσεις γίνονται για προβλέψεις πολλαπλών βημάτων.

Κατά το δεύτερο στάδιο της διερεύνησης του τεχνικού προβλήματος (Α) επεκτείνεται το μεθοδολογικό πλαίσιο του πρώτου σταδίου για διερευνήσεις σχετικές με την πρόβλεψη ενός βήματος μπροστά των ετήσιων γεωφυσικών χρονοσειρών. Έμφαση δίνεται στην μελέτη δύο συνόλων δεδομένων πραγματικού κόσμου, ενός συνόλου δεδομένων ετήσιας κατακρήμνισης και ενός συνόλου δεδομένων ετήσιας θερμοκρασίας. Τα συγκεκριμένα σύνολα δεδομένων εξετάζονται τόσο στην αρχική όσο και στην τυποποιημένη μορφή τους με κύριο στόχο την ανάδειξη των ακριβέστερων μεθόδων για πρακτικές εφαρμογές πρόβλεψης ενός βήματος μπροστά, και δευτερεύοντα στόχο την παροχή αρχικών σημείων αναφοράς για την προβλεψιμότητα της ετήσιας κατακρήμνισης και της ετήσιας θερμοκρασίας.

Το τρίτο στάδιο της διερεύνησης του τεχνικού προβλήματος (Α) περιλαμβάνει τόσο την μελέτη-ποσοτικοποίηση της προβλεψιμότητας της μηνιαίας θερμοκρασίας και της μηνιαίας κατακρήμνισης σε παγκόσμια κλίμακα, όσο και τη σύγκριση ενός μεγάλου αριθμού πλήρως αυτοματοποιημένων (κυρίως στοχαστικών) μεθόδων πρόβλεψης κατάλληλων για εποχιακές γεωφυσικές διεργασίες. Οι διερευνήσεις πραγματοποιούνται για προβλέψεις πολλαπλών βημάτων χρησιμοποιώντας το μεγαλύτερο συνόλο δεδομένων πραγματικού κόσμου που έχει χρησιμοποιηθεί μέχρι σήμερα στον χώρο της πρόβλεψης υδρολογικών χρονοσειρών.

Με το τέταρτο (και τελευταίο) στάδιο της διερεύνησης του τεχνικού προβλήματος (Α) εισάγεται η διεξαγωγή εκτεταμένων μελετών πολλαπλών περιπτώσεων ως μία καινοτόμος στρατηγική στον χώρο της πρόβλεψης γεωφυσικών χρονοσειρών. Με κύριο στόχο τη διερεύνηση τριών επιμέρους προβλημάτων που αφορούν την πρόβλεψη των συγκεκριμένων χρονοσειρών χρήσει αλγορίθμων μηχανικής μάθησης, πραγματοποιείται μια μελέτη πολλαπλών περιπτώσεων, αποτελούμενη από έναν ικανό αριθμό μελετών περιπτώσεων. Οι τελευταίες αφορούν μηνιαίες χρονοσειρές θερμοκρασίας και κατακρήμνισης παρατηρημένες στην Ελλάδα. Τα υπό μελέτη επιμέρους προβλήματα είναι η επιλογή μεταβλητών πρόβλεψης, η επιλογή των υπερπαραμέτρων, και η σύγκριση μεθόδων μηχανικής μάθησης και στοχαστικών μεθόδων.

Η διερεύνηση του τεχνικού προβλήματος (Β) γίνεται σε τρία στάδια. Κατά το πρώτο στάδιο αναπτύσσεται μια νέα μεθοδολογία πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης, χρησιμοποιώντας ως σημείο εκκίνησης ένα θεωρητικά συνεπές γενικό σχήμα πιθανοτικής υδρολογικής μοντελοποίησης δύο σταδίων. Επίσης, διεξάγονται διερευνήσεις πρότυπης μοντελοποίησης, οι οποίες καταδεικνύουν τη χρησιμότητα της προτεινόμενης μεθοδολογίας και δείχνουν πώς η κατανόηση μας για το μοντελοποιούμενο σύστημα μπορεί να μας οδηγήσει στην επίτευξη βελτιωμένης προγνωστικής μοντελοποίησης.

Κατά το δεύτερο στάδιο της διερεύνησης του τεχνικού προβλήματος (Β), μελετάται σε ένα μεγάλο σύνολο πραγματικών προβλημάτων και σε μηνιαία χρονική κλίμακα η μεθοδολογία πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης που αναπτύσσεται στο προηγούμενο στάδιο. Με τις πραγματοποιούμενες διερευνήσεις αποδεικνύεται εμπειρικά η μεγαλύτερη ευρωστία της εν λόγω μεθοδολογίας σε σχέση με τις επιμέρους προβλέψεις που συνδυάζονται από αυτήν και, κατ 'επέκταση, σε σχέση με βασικές μεθοδολογίες πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης δύο σταδίων. Επίσης, αποδεικνύεται η ικανότητα της μεθοδολογίας να αξιοποιεί τη σοφία του πλήθους.

Τέλος, κατά το τρίτο στάδιο της διερεύνησης του τεχνικού προβλήματος (Β) εισάγεται ο μεγαλύτερος αριθμός πιθανοτικών μεθόδων υδρολογικής μοντελοποίησης που έχουν μέχρι στιγμής εισαχθεί σε μια εργασία, και επιπρόσθετα διεξάγεται σε ημερήσια χρονική κλίμακα το μεγαλύτερο πείραμα συγκριτικής αξιολόγησης που έχει διεξαχθεί μέχρι στιγμής στον χώρο της πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης. Επιπρόσθετα, αξιολογούνται θεωρητικές και ποιοτικές πτυχές του επιλυόμενου προβλήματος και της χρήσης των επιλεγμένων αλγορίθμων υπό το πρίσμα της ακόλουθης ερευνητικής ερώτησης: *Γιατί και πώς να συνδυάσει κανείς διεργασιακά μοντέλα και αλγορίθμους μηχανικής μάθησης για πιθανοτική υδρολογική μοντελοποίηση;*

Εκτενής περίληψη

Συνολική σύνοψη και κύριοι στόχοι

Η παρούσα διδακτορική διατριβή εμπίμπτει στους επιστημονικούς κλάδους της στοχαστικής υδρολογίας, της υδρολογικής μοντελοποίησης και της υδροπληροφορικής. Συνεισφέρει με νέες πρακτικές λύσεις, νέες μεθοδολογίες και αποτελέσματα μεγάλης κλίμακας στην μοντελοποιήση υδρολογικών διεργασιών, συγκεκριμένα στην επίλυση δύο στενά συνυφασμένων τεχνικών προβλημάτων. Τα προβλήματα αυτά είναι:

(A) η πρόβλεψη της μελλοντικής συμπεριφοράς υδρολογικών διεργασιών χρησιμοποιώντας αποκλειστικά ενδογενείς μεταβλητές πρόβλεψης (στο εξής αναφερόμενη ως «πρόβλεψη υδρολογικών χρονοσειρών»), και

(B) η στοχαστική μοντελοποίηση υδρολογικών συστημάτων μέσω πιθανοτικής μετεπεξεργασίας αποτελεσμάτων διεργασιακής υδρολογικής μοντελοποιήσης (στο εξής αναφερόμενη ως «πιθανοτική μετεπεξεργασία αποτελεσμάτων υδρολογικής μοντελοποίησης»).

Τα εν λόγω τεχνικά προβλήματα διερευνώνται εκτενώς στα Κεφάλαια 3–6 και στα Κεφάλαια 7–9, αντίστοιχα. Επιπρόσθετα, στο Κεφάλαιο 2 γίνεται μια σύντομη επισκόπηση του θεωρητικού, μεθοδολογικού και τεχνικού υποβάθρου της διατριβής. Στο ίδιο Κεφάλαιο περιγράφεται η εργαλειοθήκη πρότυπης μοντελοποίησης και συγκριτικής αξιολόγησης, όπως αυτή έχει αναπτυχθεί και αξιοποιείται στο πλαίσιο της διατριβής. Η συγκεκριμένη εργαλειοθήκη αποτελείται από:

(i) περίπου 6 000 υδρολογικές χρονοσειρές προερχόμενες από μεγαλύτερες ελεύθερα διατιθέμενες βάσεις δεδομένων,

(ii) περισσότερα από 45 αυτοματοποιημένα μοντέλα και αλγορίθμους (διαθέσιμα σε ανοιχτό λογισμικό), τα οποία κατά κύριο λόγο προέρχονται από τις τέσσερις μεγάλες οικογένειες των στοχαστικών μοντέλων, των μοντέλων παλινδρόμησης (συμπεριλαμβανομένων μοντέλων μηχανικής μάθησης), των μοντέλων παλινδρόμησης ποσοστημορίου (συμπεριλαμβανομένων μοντέλων μηχανικής μάθησης) και των διεργασιακών υδρολογικών μοντέλων,

(iii) επτά ευέλικτες μεθοδολογίες, οι οποίες μαζί με τα διαθέσιμα σε ανοιχτό λογισμικό αυτοματοποιημένα μοντέλα και αλγορίθμους (βλ. σημείο (ii) παραπάνω) συνιστούν τη βάση των διενεργούμενων μοντελοποιήσεων, και

(iv) περίπου 30 μέτρα για την αξιολόγηση της ποιότητας των διενεργούμενων μοντελοποιήσεων.

Νέοι συνδυασμοί μοντέλων και αλγορίθμων, συνοδευόμενοι από διαφορετικές αλγοριθμικές επιλογές παραμέτρων, οδηγούν σε πολυάριθμες παραλλαγές μοντέλων, πολλές από τις οποίες μπορούν να θεωρηθούν ως νέες μέθοδοι. Ιδιαιτέρως σημαντικό –από πρακτική άποψη– είναι το γεγονός ότι όλα τα χρησιμοποιούμενα μοντέλα είναι ευέλικτα, υπολογιστικά εύχρηστα και γρήγορα στην εφαρμογή. Κατά συνέπεια, είναι κατάλληλα για διερευνήσεις μεγάλης κλίμακας, ακόμη και για διερευνήσεις παγκόσμιας κλίμακας. Η διεξαγωγή τέτοιων διερευνήσεων υπήρξε σημαντική προτεραιότητα για τη συγκεκριμένη διατριβή, όπως και η ανάπτυξη νέων μεθοδολογιών και νέων πρακτικών λύσεων. Η προτεραιότητα αυτή επιβάλλεται από τον προσανατολισμό της διατριβής και τον (κυρίως) αλγοριθμικό χαρακτήρα των μεθοδολογιών της.

Επιπρόσθετα, είναι σημαντικό να σημειωθεί ότι τα περισσότερα από τα αυτοματοποιημένα μοντέλα του σημείου (ii) βασίζονται σε πολλά εμπιμέρους, καθιστώντας έτσι δύσκολο για την παρούσα διατριβή να περιγράψει ξεχωριστά καθένα από τα μοντέλα που χρησιμοποιεί (ή ακόμα και να τα μετρήσει). Παραταύτα, η θεωρητική κατανόηση των περισσότερων (αλλά όχι όλων) των χρησιμοποιούμενων μοντέλων δεν θα μπορούσε να βοηθήσει στην ερμηνεία και κατανόηση των αλγοριθμικά αποκτηθέντων αποτελεσμάτων της διατριβής. Υπό το πρίσμα αυτό, ένα πλεονέκτημα (και παράλληλα περιορισμός) που χαρακτηρίζει τη διατριβή και απορρέει από τους στόχους της είναι ο αλγοριθμικός της χαρακτήρας. Παρά τη φύση και τον κύριο προσανατολισμό

των μεθοδολογικών μας πλαισίων, η παρούσα διατριβή παρέχει καινοτόμα θεωρητικά συμπληρώματα στην πρακτική και μεθοδολογική της συμβολή.

Στη συνέχεια, συνοψίζουμε το περιεχόμενο των Κεφαλαίων 3–9, δίνοντας έμφαση τόσο στις κύριες καινοτομίες που τα χαρακτηρίζουν (όπως αυτές προκύπτουν υπό το πρίσμα της βιβλιογραφίας) όσο και στην τεχνογνωσία που αυτά παρέχουν. Συζητάμε ακόμη τον τρόπο με τον οποίο τα Κεφάλαια αυτά χτίζουν το ένα πάνω στο άλλο για (α) την παροχή νέων τεχνικών λύσεων και νέων μεθοδολογιών, (β) την απάντηση πρακτικών και θεωρητικών ερευνητικών ερωτημάτων, και (γ) την βελτίωση της κατανόησης των διερευνούμενων τεχνικών προβλημάτων μέσα από συγκρίσεις και αξιολογήσεις μοντέλων σε μεγάλη κλίμακα.

Πρόβλεψη υδρολογικών διεργασιών

Στοχαστικές μέθοδοι έναντι μεθόδων μηχανικής μάθησης στην πρόβλεψη πολλαπλών βημάτων

Το Κεφάλαιο 3 έχει ως γενικό του στόχο την προώθηση συγκρίσεων μεγάλης κλίμακας στον χώρο της πρόβλεψης των υδρολογικών διεργασιών. Το Κεφάλαιο ξεκινά με μια σύντομη επισκόπηση και κριτική θεώρηση της σχετικής βιβλιογραφίας. Η συγκεκριμένη βιβλιογραφία επικεντρώνεται συχνά στη σύγκριση στοχαστικών μεθόδων και μεθόδων μηχανικής μάθησης, καθώς και στη διερεύνηση νέων «υβριδικών» μεθοδολογιών, αποκλειστικά διεξάγοντας μελέτες περιπτώσεων. Οι συγκεκριμένες μελέτες αδυνατούν να υποστηρίξουν οποιαδήποτε γενίκευση περί της χρησιμότητας μεθόδων πρόβλεψης, παρότι χρησιμοποιούνται συχνά για τον συγκεκριμένο σκοπό. Εντούτοις, επιτρέπουν την ανάδειξη σημαντικών σημείων, παρέχοντας αμεσότητα και παραστατικότητα. Είναι, επομένως, εξαιρετικά χρήσιμες όταν συνοδεύουν αναλυτικές διερευνήσεις ή εμπειρικές διερευνήσεις μεγάλης κλίμακας. Μόνο τέτοιες διερευνήσεις παρέχουν (εν δυνάμει) γενικεύσιμα αποτελέσματα. Στη βιβλιογραφία, έχουν διεξαχθεί αναλυτικές διερευνήσεις για διάφορες μεθόδους πρόβλεψης (κυρίως για τις λιγότερο ευέλικτες από αυτές). Ωστόσο, τέτοιου είδους διερευνήσεις είναι πολύ απαιτητικές (έως σχεδόν αδύνατες) για πολλές άλλες μεθόδους (κυρίως για τις πιο ευέλικτες μεθόδους μηχανικής μάθησης). Ως εκ τούτου, θεωρητικά συνεπείς αξιολογήσεις και συγκρίσεις μεθόδων πρόβλεψης υδρολογικών διεργασιών απαιτούν αναγκαστικά την εξέταση ενός επαρκώς μεγάλου και αντιπροσωπευτικού δείγματος περιπτώσεων. Αναδεικνύουμε το συγκεκριμένο γεγονός για πρώτη φορά στη σχετική βιβλιογραφία.

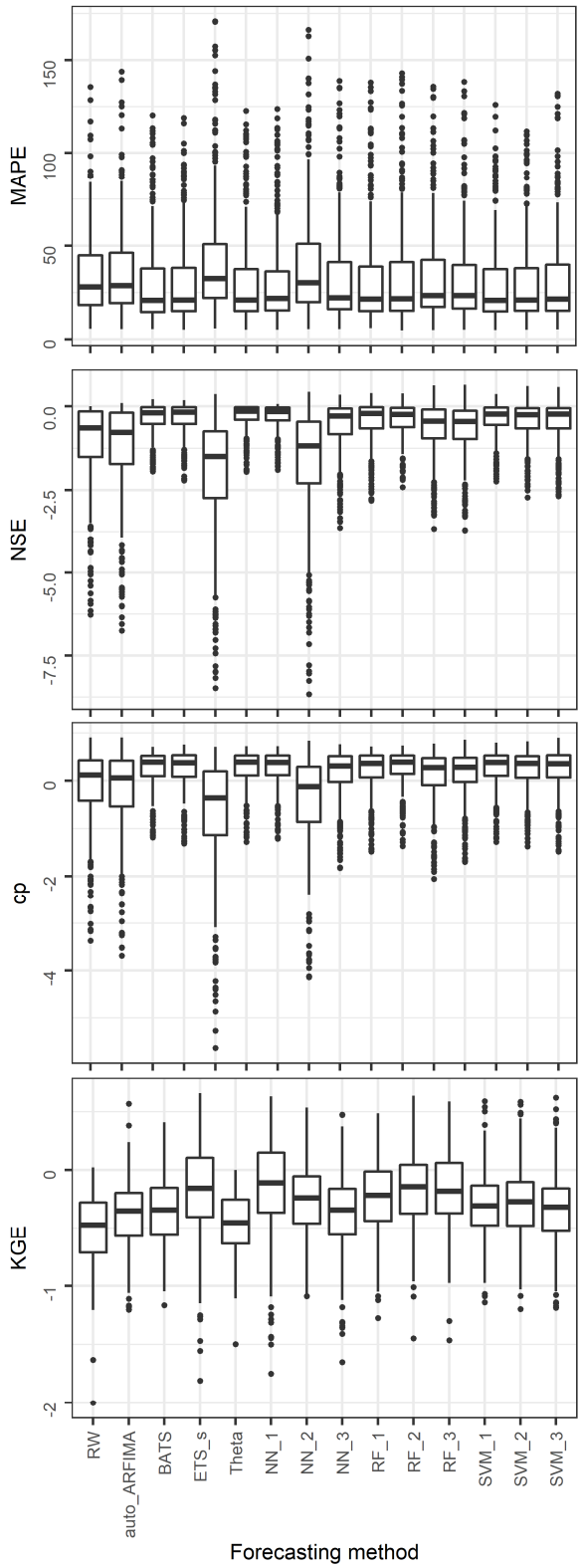
Διατυπώνουμε και διερευνούμε διεξοδικά το εξής ερευνητικό ερώτημα: *Αντιστοιχεί το δίπολο στοχαστικές μέθοδοι – μέθοδοι μηχανικής μάθησης σε κάποια σαφή διαφορά στην προγνωστική επίδοση των μεθόδων*; Για να δοθεί απάντηση στο συγκεκριμένο ερώτημα, αναπτύσσουμε και υιοθετούμε ένα νέο μεθοδολογικό πλαίσιο για την αξιολόγηση τεχνικών πρόγνωσης στην υδρολογία. Όντας σύμφωνο με τις αρχές που θα πρέπει να διέπουν την πρόβλεψη χρονοσειρών, το προτεινόμενο πλαίσιο ενσωματώνει συγκριτική αξιολόγηση μεθοδολογιών μεγάλης κλίμακας. Η τελευταία βασίζεται σε μεγάλα σύνολα υδρολογικών δεδομένων, σε στοχαστική προσομοίωση χρονοσειρών μεγάλης κλίμακας χρησιμοποιώντας στάσιμα κλασσικά στοχαστικά μοντέλα, σε έναν μεγάλο αριθμό πλήρως αυτοματοποιημένων μοντέλων και αλγόριθμων πρόβλεψης (συμπεριλαμβανομένων μοντέλων αναφοράς) και σε έναν ικανό αριθμό μέτρων για την ποσοτικοποίηση της ποιότητας των προβλέψεων. Συγκεκριμένα, στόχος μας είναι να παρέχουμε αποτελέσματα μεγάλης κλίμακας και χρήσιμη κατανόηση σχετικά με τη σύγκριση των στοχαστικών μεθόδων και των μεθόδων μηχανικής μάθησης στην πρόβλεψη υδρολογικών διεργασιών σε μεγάλες χρονικές κλίμακες (π.χ., ετήσιες και μηνιαίες), με έμφαση στις ετήσιες διεργασίες απορροής ποταμών.

Συγκρίνουμε 11 στοχαστικές μεθόδους και εννέα μεθόδους μηχανικής μάθησης στην πρόβλεψη πολλαπλών βημάτων. Οι στοχαστικές μέθοδοι περιλαμβάνουν απλά μοντέλα, μοντέλα από τις συχνά χρησιμοποιούμενες οικογένειες autoregressive moving average (ARMA) και autoregressive fractionally integrated moving average (ARFIMA), innovations state space μοντέλα και μοντέλα εκθετικής εξομάλυνσης, ενώ οι μέθοδοι μηχανικής μάθησης είναι νευρωνικά δίκτυα (neural networks), τυχαία δάση (random forests) και support vector machines. Από τα παραπάνω μοντέλα μόνο τα AR(FI)MA, νευρωνικά δίκτυα (neural networks) και support vector

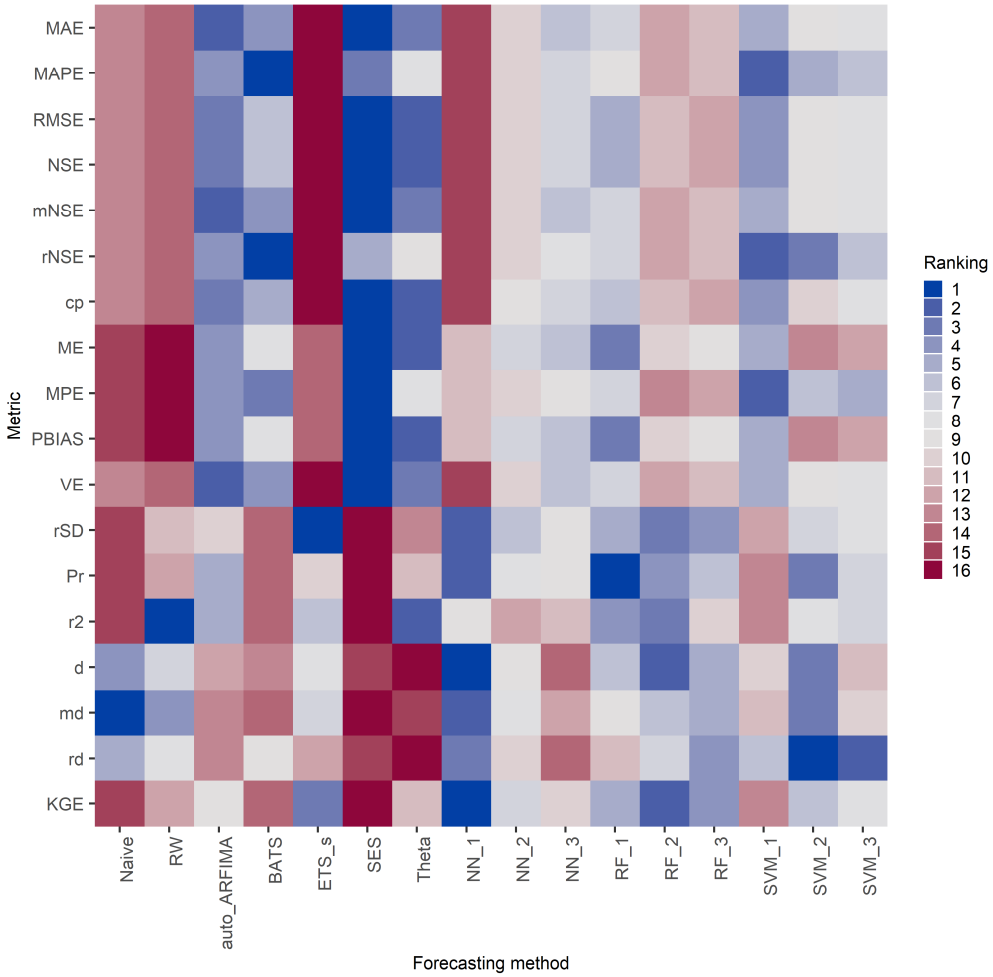
machines έχουν χρησιμοποιηθεί ευρέως για υδρο-μετεωρολογικές προβλέψεις, συνήθως όμως σε μη αυτοποιημένη μορφή. Οι περισσότερες από τις υπόλοιπες μεθόδους έχουν χρησιμοποιηθεί σε λιγοστές εργασίες (π.χ., στα Κεφάλαια 4, 5 και 6). Χρησιμοποιούμε έτοιμα αυτοματοποιημένα μοντέλα και αλγόριθμους πρόβλεψης, και παράλληλα συνδυάζουμε διαφορετικούς αλγορίθμους για την αυτοματοποίηση νέων. Ειδικά για τις μεθόδους μηχανικής μάθησης, προτείνουμε τρεις αντικειμενικές μεθόδους για την επιλογή των μεταβλητών πρόβλεψης (μεταξύ των οποίων μία εμπνευσμένη από έναν έτοιμο πλήρως αυτοματοποιημένο αλγόριθμο) και τρία σύνολα τιμών πλέγματος για τη βελτιστοποίηση υπερπαραμέτρων μέσω αυτοματοποιημένης αναζήτησης. Το προτεινόμενο σύνολο μεθόδων θα μπορούσε να χρησιμοποιηθεί για τη συγκριτική αξιολόγηση της επίδοσης οποιασδήποτε νέας μεθόδου πρόβλεψης υδρολογικών διεργασιών. Επίσης, δίνεται σε μορφή κώδικα.

Διενεργούμε 12 υπολογιστικά πειράματα μεγάλης κλίμακας βασιζόμενα σε προσομοιώσεις. Καθένα από τα διενεργούμενα πειράματα χρησιμοποιεί διαφορετικό μοντέλο στοχαστικής προσομοίωσης. Τα επιλεγμένα μοντέλα στοχαστικής προσομοίωσης αντιστοιχούν σε διαφορετικούς τύπους αυτοσυσχέτισης. Πραγματοποιούμε κάθε πείραμα προσομοίωσης δύο φορές, την πρώτη φορά χρησιμοποιώντας προσομοιωμένες χρονοσειρές 100 τιμών και τη δεύτερη φορά χρησιμοποιώντας προσομοιωμένες χρονοσειρές 300 τιμών. Επιπλέον, πραγματοποιούμε ένα πείραμα πραγματικού κόσμου χρησιμοποιώντας 405 μέσες ετήσιες χρονοσειρές απορροής ποταμών, καθεμία από τις οποίες αποτελείται από 100 τιμές. Ο συνολικός αριθμός των προβλέψεων είναι 858 480, εκ των οποίων 6 480 παράγονται εντός του πειράματος πραγματικού κόσμου. Ποσοτικοποιούμε την επίδοση των μεθόδων πρόβλεψης χρησιμοποιώντας 18 μέτρα. Αυτά τα μέτρα δεν έχουν ένα-προς-ένα σχέση μεταξύ τους, δίνοντας έμφαση σε −περισσότερο ή λιγότερο− διαφορετικές πτυχές της ίδιας πληροφορίας. Έχουν επιλεγεί για να παρέχουν μια πολύπλευρη αξιολόγηση της ποιότητας προβλέψεων πολλαπλών βημάτων των υδρολογικών διεργασιών.

Τα αποτελέσματα μεγάλης κλίμακας (βλ. π.χ., Σχήματα 1 και 2) καταδεικνύουν ότι οι στοχαστικές μέθοδοι και οι μέθοδοι μηχανικής μάθησης δεν διαφέρουν δραματικά, όπως συνήθως υποστηρίζεται στη βιβλιογραφία. Στην πραγματικότητα, μέθοδοι και από τις δύο αυτές κατηγορίες είναι εξίσου χρήσιμες στην πρόβλεψη υδρολογικών διεργασιών σε μεγάλες χρονικές κλίμακες. Αυτό το αποτέλεσμα είναι ιδιαίτερα ενδιαφέρον, δεδομένων των ισχυρισμών ότι οι μέθοδοι μηχανικής μάθησης είναι πιθανότερο να υπερέχουν σε "μη γραμμικές καταστάσεις". Συχνά υποστηρίζεται ότι οι διεργασίες απορροής ποταμών προσιδιάζουν σε τέτοιες καταστάσεις. Γενικά, δεν μπορούμε να αποφασίσουμε για κάποια καθολικά καλύτερη ή χειρότερη μέθοδο πρόβλεψης, ούτε μπορούμε να κατατάξουμε τις μεθόδους πρόβλεψης με βάση τα αποτελέσματα μεγάλης κλίμακας. Οποιαδήποτε κατάταξη των μεθόδων πρόβλεψης θα απαιτούσε την εκ των προτέρων επιλογή ενός πειράματος και ενός κριτηρίου ενδιαφέροντος, καθώς και την εφαρμογή μιας απλούστευτικής διαδικασίας, και συνεπώς δεν θα ήταν γενική. Ωστόσο, η συσταδοποίηση-ομαδοποίηση των μεθόδων πρόβλεψης με βάση ομοιότητες ή διαφορές στην επίδοση σε σχέση με διάφορα κριτήρια είναι δυνατή, αν και μόνο σε κάποιο βαθμό.



Σχήμα 1. Ενδεικτικά θηκογράμματα (boxplots) για την συγκριτική αξιολόγηση των μεθόδων πρόβλεψης του Κεφαλαίου 3 όσον αφορά την επίδοση τους στο πείραμα πραγματικού κόσμου. Τα μακρυνά εξωκείμενα σημεία (outliers) έχουν αφαιρεθεί.



Σχήμα 2. Θερμογράφημα (heatmap) για την συγκριτική αξιολόγηση των μεθόδων πρόβλεψης του Κεφαλαίου 3 όσον αφορά την μέση κατάταξη τους από την 1η (καλύτερη) έως στην 16η (χειρότερη) για το πείραμα πραγματικού κόσμου.

Μια άλλη σημαντική συνεισφορά του Κεφαλαίου 3 σχετίζεται με το θεώρημα «no free lunch». Σύμφωνα με το συγκεκριμένο θεώρημα, στο χώρο όλων των πιθανών περιπτώσεων ενός προβλήματος, δεν υπάρχει κάποιο μοντέλο που να λειτουργεί πάντα καλύτερα από άλλα, ελλείψει σημαντικών επιπρόσθετων πληροφοριών για το συγκεκριμένο πρόβλημα. Τα αποτελέσματα μεγάλης κλίμακας συμβαδίζουν με αυτό το θεώρημα, αν και το θεώρημα αναφέρεται σε άπειρο χώρο προβλημάτων, ενώ εμείς εξετάζουμε τον πεπερασμένο χώρο προβλημάτων που ορίζεται από τις υπό διερεύνηση προσομοιωμένες χρονοσειρές και ετήσιες χρονοσειρές απορροής ποταμών. Στην πραγματικότητα, η εύρεση του καταλληλότερου αλγορίθμου εξαρτάται κυρίως από την κατανόηση του συστήματος, η οποία προφανώς θα πρέπει να είναι βαθύτερη από τη γνώση των στατιστικών ιδιοτήτων του (π.χ., από την γνώση της μέσης τιμής, της διακύμανσης και της συνάρτησης αυτοσυσχέτισης). Όσον αφορά τον βαθμό στον οποίο τα συμπεράσματά μας θα μπορούσαν να είναι γενικεύσιμα για την πρόβλεψη υδρολογικών διεργασιών σε μεγάλες χρονικές κλίμακες, τονίζουμε ότι η παραδοχή της στασιμότητας και η λογική περί καταλληλότητας αυτής για τη μοντελοποίηση των γεωφυσικών διεργασιών είναι σύμφωνες με το θεώρημα «no free lunch». Συγκεκριμένα, σε περιπτώσεις που δεν μπορούμε να εξηγήσουμε τη συμπεριφορά μιας γεωφυσικής διεργασίας βασιζόμενοι σε κάποιον προσδιοριστικό μηχανισμό, τότε τα καταλληλότερα μοντέλα για να την περιγράψουμε είναι τα στάσιμα.

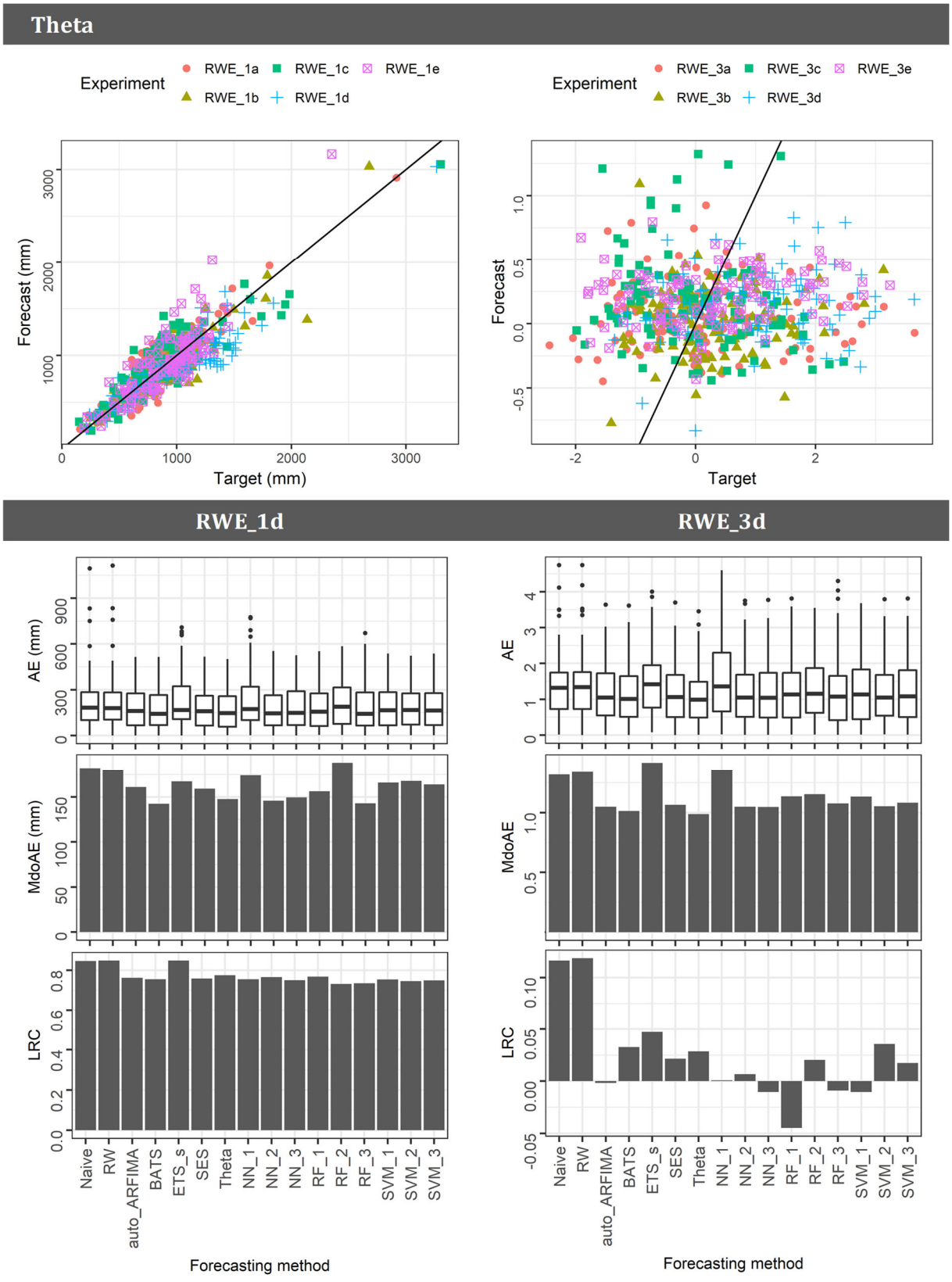
Πρόβλεψη ενός βήματος μπροστά της ετήσιας θερμοκρασίας και κατακρήμνισης

Το Κεφάλαιο 4 έχει ως γενικό του στόχο την προώθηση των παραδοσιακών μεθόδων και αρχών πρόβλεψης χρονοσειρών στις γεωεπιστήμες. Το Κεφάλαιο ξεκινά με την παροχή λεπτομερών

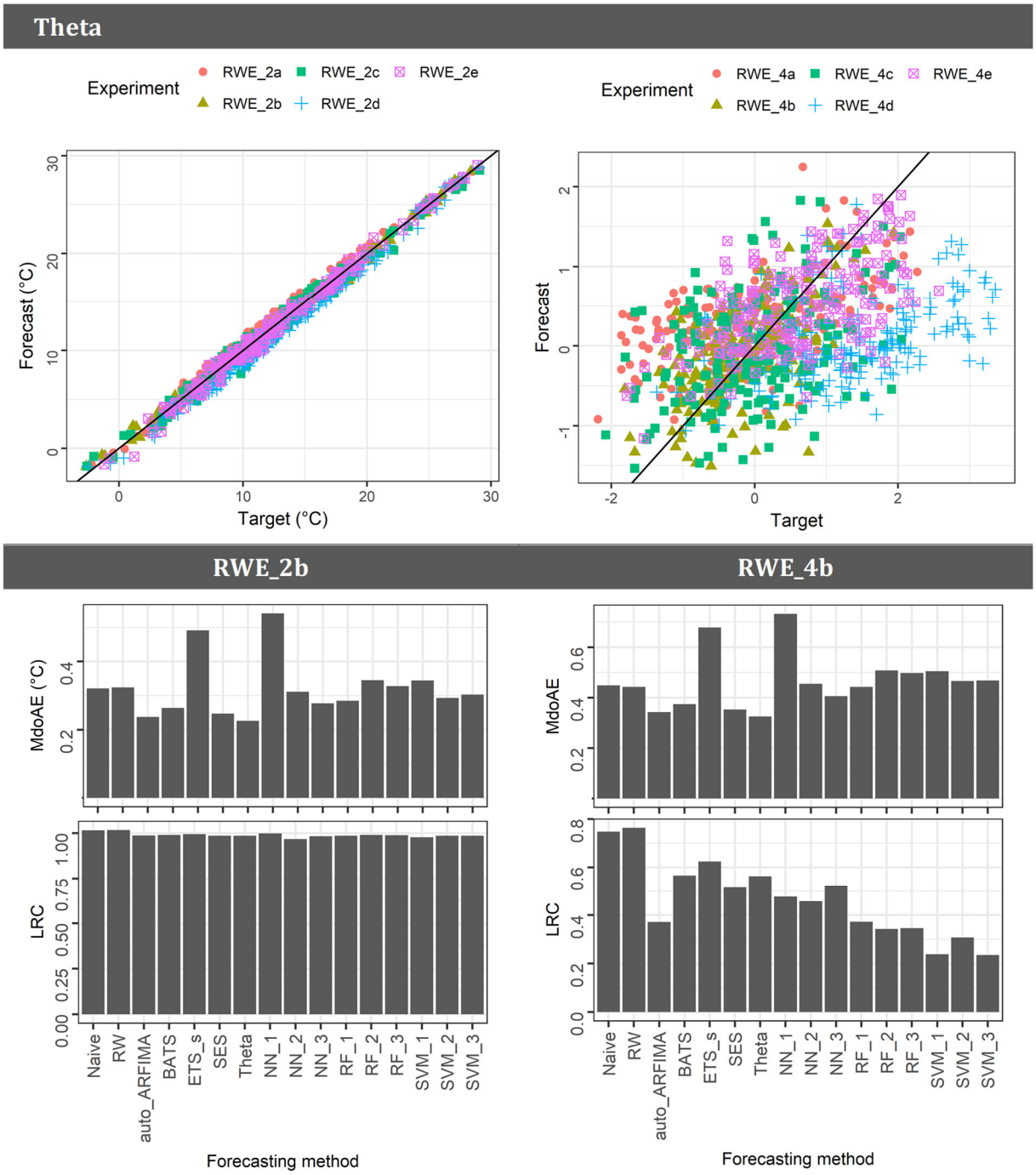
πληροφοριών σχετικά με τη χρήση στατιστικών μεθόδων (στοχαστικών μεθόδων και μεθόδων μηχανικής μάθησης) στον χώρο της πρόβλεψης υδρο-μετεωρολογικών διεργασιών, συμπληρώνοντας έτσι το εισαγωγικό υποκεφάλαιο του Κεφαλαίου 3. Συγκεκριμένα, το Κεφάλαιο στοχεύει (α) στη διερεύνηση του θεμελιώδους προβλήματος της πρόβλεψης ενός βήματος μπροστά εντός ενός αμιγώς στατιστικού πλαισίου (το οποίο τεκμηριώνεται από ειδικούς στον κλάδο της πρόβλεψης χρονοσειρών) στις γεωεπιστήμες, και (β) στην θέσπιση των αποτελεσμάτων που προκύπτουν από την εξέταση των τυποποιημένων χρονοσειρών πραγματικού κόσμου ως αρχικών σημείων αναφοράς για την προβλεψιμότητα ενός βήματος μπροστά των γεωφυσικών διεργασιών. Η θέσπιση τέτοιων σημείων αναφοράς έχει ιδιαίτερο νόημα για τις εφαρμογές πρόβλεψης ενός βήματος μπροστά, καθώς οι τελευταίες αποτελούν τις απλούστερες εφαρμογές πρόβλεψης και η ακρίβεια με την οποία διενεργούνται μπορεί να ποσοτικοποιηθεί χρησιμοποιώντας ένα μόνο μέτρο, συγκεκριμένα την απόλυτη τιμή του σφάλματος πρόβλεψης.

Για να επιτύχουμε τους παραπάνω στόχους, επεκτείνουμε το μεθοδολογικό πλαίσιο του Κεφαλαίου 3, διερευνώντας τις ιδιότητες των μεθόδων του, όταν αυτές εφαρμόζονται για την πρόβλεψη ενός βήματος μπροστά σε ετήσιες γεωφυσικές χρονοσειρές. Έμφαση δίνεται στην διερεύνηση δύο συνόλων δεδομένων πραγματικού κόσμου, ενός συνόλου δεδομένων κατακρήμνισης και ενός συνόλου δεδομένων θερμοκρασίας, που μαζί περιέχουν 297 ετήσιες χρονοσειρές 91 τιμών. Τα συγκεκριμένα δεδομένα εξετάστηκαν τόσο στην αρχική όσο και στην τυποποιημένη μορφή τους. Συμπληρωματικά, πραγματοποιούμε πειράματα μεγάλης κλίμακας βασιζόμενοι σε 12 προσομοιωμένα σύνολα δεδομένων. Αυτά τα σύνολα δεδομένων αποτελούνται από συνολικά 24 000 χρονοσειρές των 91 τιμών. Τα διεξαγόμενα πειράματα προσομοίωσης συμπληρώνουν επιτυχώς τα πειράματα πραγματικού κόσμου, επιτρέποντας την εξέταση μιας μεγάλης ποικιλίας διεργασιακών συμπεριφορών. Παράλληλα, είναι σε κάποιο βαθμό ελεγχόμενα, διευκολύνοντας έτσι πιθανές γενικεύσεις, ενώ επίσης αυξάνουν την κατανόηση του εξεταζόμενου προβλήματος. Χρησιμοποιούμε τις πρώτες 50, 60, 70, 80 και 90 τιμές της εκάστοτε χρονοσειράς για την προσαρμογή (και επιλογή) των μοντέλων, και διενεργούμε προβλέψεις που αντιστοιχούν στην 51η, 61η, 71η, 81η και 91η τιμή της χρονοσειράς, αντίστοιχα. Ο συνολικός αριθμός των προβλέψεων που παράγονται είναι 2 177 520, μεταξύ των οποίων 47 520 παράγονται στο πλαίσιο των πειραμάτων πραγματικού κόσμου. Η αξιολόγηση βασίζεται σε οκτώ μέτρα και στατιστικά σφαλμάτων.

Τα πειράματα προσομοίωσης καταδεικνύουν τις περισσότερο και λιγότερο ακριβείς μεθόδους για πρακτικές εφαρμογές πρόβλεψης ενός βήματος μπροστά, αποδεικνύοντας επίσης ότι οι απλές μέθοδοι είναι ιδιαιτέρως ανταγωνιστικές σε συγκεκριμένες περιπτώσεις. Ακόμη προκύπτει πως η σχετική επίδοση των μεθόδων πρόβλεψης εξαρτάται ελάχιστα από το μήκος της χρονοσειράς (σημειωτέον ότι εστιάζουμε σε χρονοσειρές 51, 61, 71, 81 και 91 τιμών), ενώ εξαρτάται έντονα από την υπό διερεύνηση διεργασία. Όσον αφορά τα αποτελέσματα των πειραμάτων πραγματικού κόσμου που χρησιμοποιούν τις πρωτότυπες (τυποποιημένες) χρονοσειρές (βλ. π.χ., Σχήματα 3 κα 4), στο πλαίσιο αυτών προκύπτουν ελάχιστες και μέγιστες διάμεσοι απόλυτων σφαλμάτων ίσες με 68 mm (0.55) και 189 mm (1.42), αντίστοιχα για την κατακρήμνιση, και 0.23°C (0.33) και 1.10°C (1.46), αντίστοιχα για τη θερμοκρασία. Τα αποτελέσματα που προκύπτουν χρησιμοποιώντας τις τυποποιημένες χρονοσειρές πραγματικού κόσμου θα μπορούσαν να χρησιμοποιηθούν ως αρχικά σημεία αναφοράς για την προβλεψιμότητα της ετήσιας κατακρήμνισης και της ετήσιας θερμοκρασίας, καθώς δεν υπάρχει σχετική πληροφορία στη βιβλιογραφία.



Σχήμα 3. Συνοπτική σύγκριση ανάμεσα στα αποτελέσματα των πειραμάτων του Κεφαλαίου 4 που χρησιμοποιούν χρονοσειρές ετήσιας κατακρήμνισης (αριστερά) και στα αποτελέσματα των πειραμάτων του ίδιου Κεφαλαίου που χρησιμοποιούν χρονοσειρές τυποποιημένης ετήσιας κατακρήμνισης (δεξιά).



Σχήμα 4. Συνοπτική σύγκριση ανάμεσα στα αποτελέσματα των πειραμάτων του Κεφαλαίου 4 που χρησιμοποιούν χρονοσειρές ετήσιας θερμοκρασίας (αριστερά) και στα αποτελέσματα των πειραμάτων του ίδιου Κεφαλαίου που χρησιμοποιούν χρονοσειρές τυποποιημένης ετήσιας θερμοκρασίας (δεξιά).

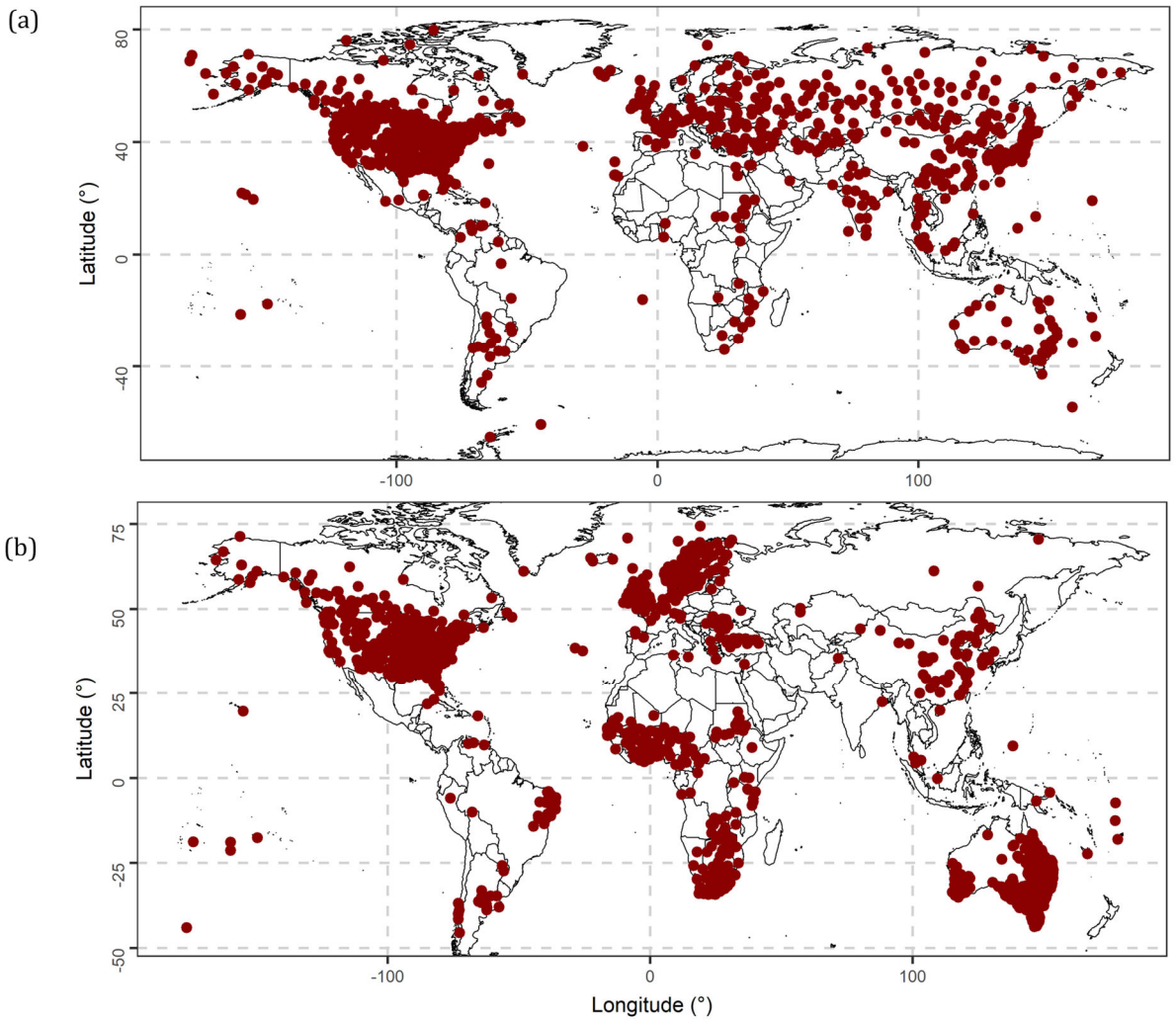
Πρόβλεψη πολλαπλών βημάτων της μηνιαίας θερμοκρασίας και κατακρήμνισης

Το Κεφάλαιο 5 έχει ως γενικό του στόχο την προώθηση της χρήσης πλήρως αυτοματοποιημένων μεθόδων πρόβλεψης στις γεωεπιστήμες. Η μη αυτόματοποιημένη ή υποκειμενική προσέγγιση του προβλήματος της πρόβλεψης χρονοσειρών υιοθετείται συχνά στη γεωεπιστημονική βιβλιογραφία (συμπεριλαμβανομένης της υδρολογικής βιβλιογραφίας), ενώ απαιτεί την εκ των προτέρων διεξαγωγή διερευνητικής ανάλυσης δεδομένων για κάθε συγκεκριμένη περίπτωση που πρέπει να προβλεφθεί και, συνεπώς, ανθρώπινη παρέμβαση κατά τη διάρκεια της πρόβλεψης. Ως εκ τούτου, η εφαρμογή της μπορεί να περιοριστεί σημαντικά από παράγοντες

κλίμακας. Η πλήρως αυτοματοποιημένη πρόβλεψη χρονοσειρών είναι απαραίτητη, για παράδειγμα, σε περιπτώσεις που μας έχει ζητηθεί ένας μεγάλος αριθμός προβλέψεων.

Διεξάγουμε δύο διερευνήσεις σε παγκόσμια κλίμακα. Ποσοτικοποιούμε την προβλεψιμότητα της μηνιαίας θερμοκρασίας και της μηνιαίας κατακρήμνισης εφαρμόζοντας 24 πλήρως αυτοματικοποιημένες μεθόδους πρόβλεψης σε 985 και 1 552 μηνιαίες χρονοσειρές θερμοκρασίας και κατακρήμνισης, αντίστοιχα. Το δείγμα αυτό είναι το μεγαλύτερο που έχει χρησιμοποιηθεί στην υδρολογία για την αξιολόγηση της επίδοσης μεθόδων πρόβλεψης χρονοσειρών. Χρησιμοποιούμε πλήρως αυτοματοποιημένα μοντέλα (διατιθέμενα σε ανοιχτό λογισμικό) με διαφορετικές αλγοριθμικές επιλογές (στο βαθμό του μας το επιτρέπουν τα αυτοματοποιημένα μοντέλα) και, παράλληλα, προβαίνουμε σε διάφορους συνδυασμούς για την αυτοματοποίηση νέων μοντέλων. Οι πλήρως αυτοματοποιημένες μέθοδοι του Κεφαλαίου περιλαμβάνουν: (α) την εποχιακή μέθοδο αναφοράς (η οποία βασίζεται στις μηνιαίες τιμές του τελευταίου έτους), (β) τέσσερις μεθόδους βασιζόμενες στο μοντέλο τυχαίος περίπατος (random walk), (γ) τέσσερις μεθόδους βασιζόμενες σε ένα αυτόματο μοντέλο ARFIMA, (δ) έξι μεθόδους βασιζόμενες στο μοντέλο BATS (μοντέλο εκθετικής εξομάλυνσης που ενσωματώνει μετασχηματισμό Box-Cox, διόρθωση σφαλμάτων μέσω μοντέλων ARMA, όρους τάσεων και εποχιακούς όρους), (ε) τέσσερις μεθόδους βασιζόμενες στο μοντέλο απλής εκθετικής εξομάλυνσης, (στ) δύο μεθόδους βασιζόμενες στο μοντέλο Theta, και (ζ) τρεις μεθόδους βασιζόμενες στο μοντέλο Prophet.

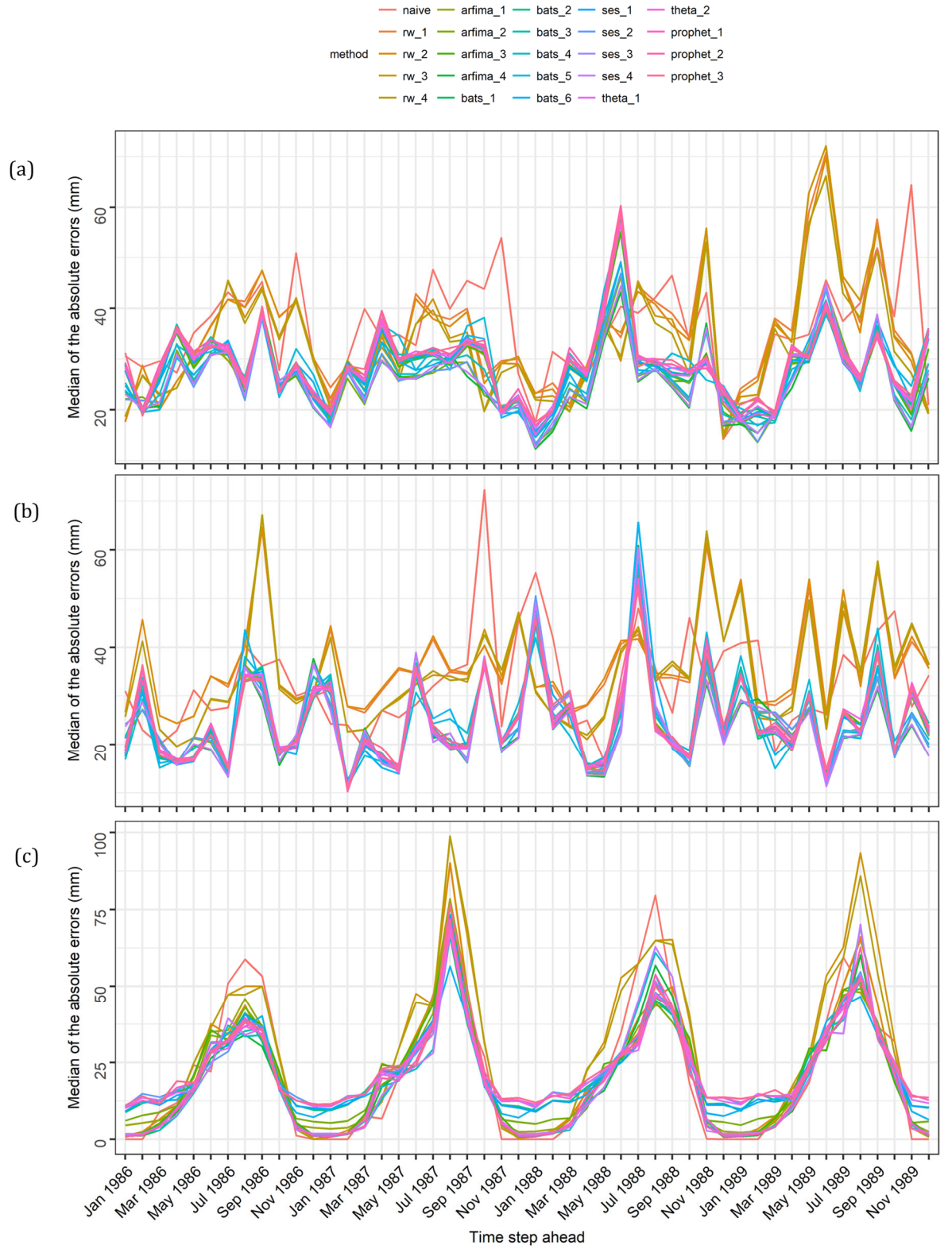
Το μοντέλο Prophet είναι ένα πρόσφατο μοντέλο, εμπνευσμένο από τη φύση των χρονοσειρών που μελετώνται για τη λειτουργία του Facebook. Στο Κεφάλαιο 5, το εν λόγω μοντέλο εφαρμόζεται για πρώτη φορά σε υδρο-μετεωρολογικές χρονοσειρές. Αντιθέτως, τα μοντέλα ARFIMA χρησιμοποιούνται ευρέως −όμως με μη αυτοματοποιημένο τρόπο− στην υδρολογική βιβλιογραφία, ενώ τα υπόλοιπα μοντέλα χρησιμοποιούνται πολύ σπάνια (π.χ., στα Κεφάλαια 3 και 4), παρότι θεωρούνται θεμελιώδη στο επιστημονικό πεδίο της πρόβλεψης χρονοσειρών. Στα Κεφάλαια 3 και 4, δεν γίνονται διερευνήσεις σχετικά με το πώς οι διαφορετικές επιλογές μοντελοποίησης της εποχιακότητας και χειρισμού της μη κανονικότητας επηρεάζουν την επίδοση των μοντέλων. Τέτοιες διερευνήσεις αποτελούν έναν από τους κύριους στόχους του Κεφαλαίου 5 (επομένως, εξετάζονται κατάλληλες παραλλαγές των μεθόδων), μαζί με την αξιολόγηση της επίδοσης των επιλεγμένων μοντέλων σε μηνιαίες υδρο-μετεωρολογικές χρονοσειρές και τη σύγκριση του μοντέλου Prophet με τα υπόλοιπα. Οι υπό διερεύνηση χρονοσειρές έχουν μήκος 480 μήνες και είναι πλήρεις (χωρίς ελλείπουσες τιμές). Έχουν παρατηρηθεί από τον Ιανουάριο του 1950 έως το Δεκέμβριο του 1989 σε σταθμούς που καλύπτουν ένα σημαντικό μέρος της επιφάνειας της Γης (βλ. Σχήμα 5) και συνεπώς περιλαμβάνουν διάφορες διεργασιακές συμπεριφορές πραγματικού κόσμου. Τα μοντέλα προσαρμόζονται στα πρώτα 36 έτη (432 μήνες) και στη συνέχεια δοκιμάζονται στη διεξαγωγή προβλέψεων πολλαπλών βημάτων για τα τελευταία τέσσερα χρόνια (48 μήνες). Τα αποτελέσματα συνοψίζονται σε παγκόσμια στατιστικά, ενώ ομαδοποίηση των σταθμών έχει οδηγήσει σε πέντε επιμέρους κατηγορίες στατιστικών για τη θερμοκρασία και έξι για την κατακρήμνιση. Η ομαδοποίηση των σταθμών γίνεται σύμφωνα με τη γεωγραφική γειτνίαση τους.



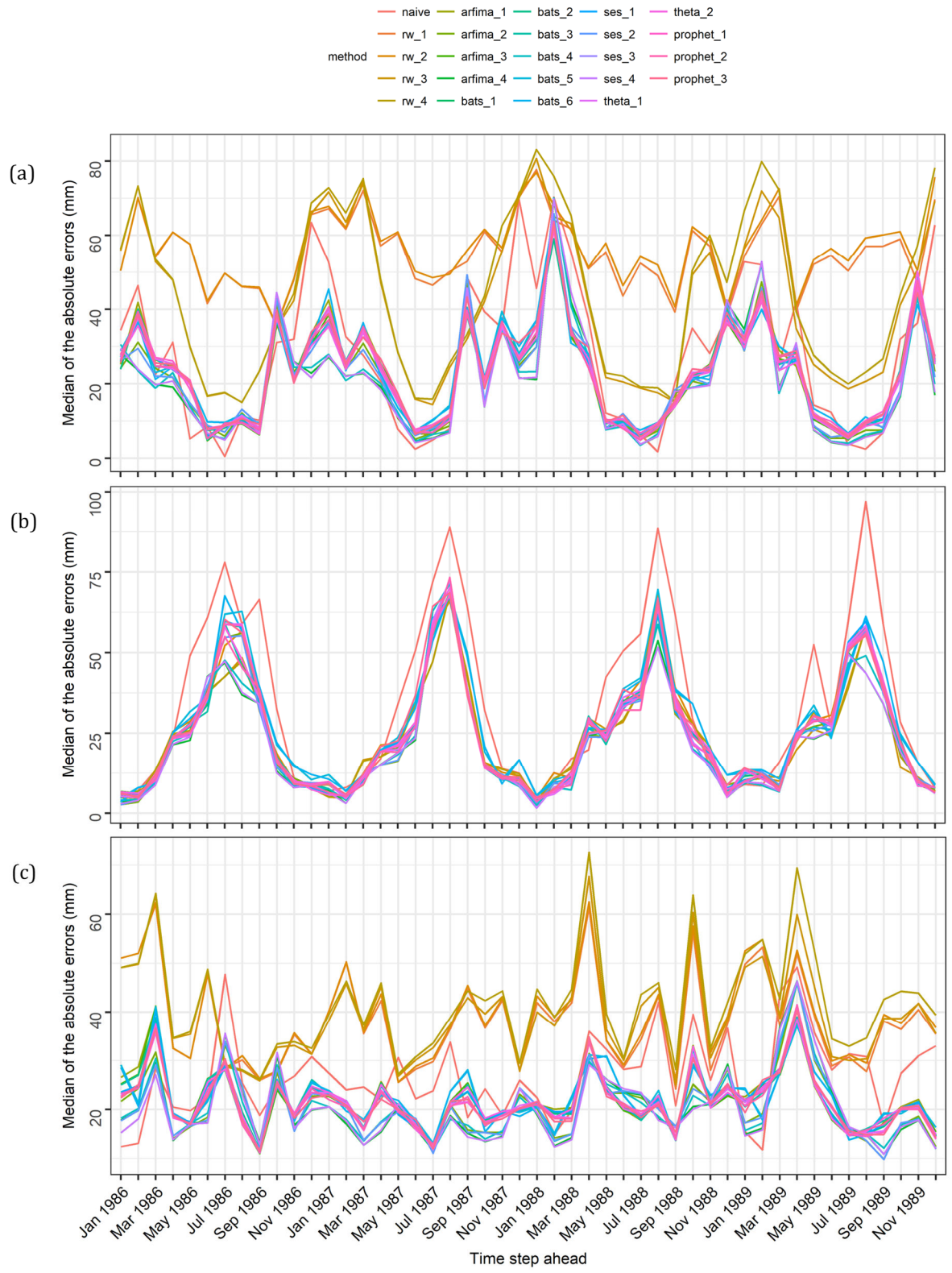
Σχήμα 5. Γεωγραφικές θέσεις των σταθμών μέτρησης (a) θερμοκρασίας και (b) κατακρήμνισης. Μηνιαία δεδομένα από τους συγκεκριμένους σταθμούς χρησιμοποιούνται για τα πειράματα του Κεφαλαίου 5.

Τα αποτελέσματα (βλ. π.χ., Σχήματα 6–8) καταδεικνύουν ότι όλες οι εξεταζόμενες μέθοδοι, εκτός από την εποχιακή μέθοδο αναφοράς και τις μεθόδους τυχαίου περιπάτου (random walk), είναι αρκετά ακριβείς ώστε να χρησιμοποιούνται σε πρακτικές εφαρμογές. Ακόμα και οι μέθοδοι απλής εκθετικής εξομάλυνσης και τα μοντέλα Theta που παρουσιάζουν μάλλον μέτρια επίδοση σε όρους ρίζας μέσου τετραγωνικού σφάλματος (root mean square error – RMSE) και Nash-Sutcliffe στα πειράματα προσομοίωσης του Κεφαλαίου 3, στο Κεφάλαιο 5 προκύπτουν εξίσου ανταγωνιστικές με τις μεθόδους ARFIMA και BATS. Οι τελευταίες δύο μέθοδοι προκύπτουν ως οι πιο ακριβείς σε όρους RMSE και Nash-Sutcliffe στο Κεφαλαίου 3. Αυτό θα μπορούσε να εξηγηθεί ως εξής: Τα πειράματα προσομοίωσης του Κεφαλαίου 3 εξετάζουν μη εποχιακές προσομοιωμένες διεργασίες με διαφορετική προβλεψιμότητα από τις μηνιαίες διεργασίες θερμοκρασίας και κατακρήμνισης. Η εποχιακότητα μπορεί να θεωρηθεί ως το προσδιοριστικό κομμάτι μιας διεργασίας, ενώ κατάλληλη μοντελοποίηση της μπορεί να οδηγήσει σε σημαντική βελτίωση των προβλέψεων. Το παραπάνω ποιοτικό αποτέλεσμα συμφωνεί με τα αποτελέσματα των 50 μελετών περιπτώσεων του Κεφαλαίου 6. Αυτές οι μελέτες περιπτώσεων χρησιμοποιούν επίσης μηνιαία δεδομένα θερμοκρασίας και κατακρήμνισης. Στο ίδιο Κεφάλαιο, η εποχιακότητα μοντελοποιείται χρησιμοποιώντας το πολλαπλασιαστικό μοντέλο και το προσθετικό μοντέλο για τις χρονοσειρές θερμοκρασίας και κατακρήμνισης, αντίστοιχα. Όσον αφορά τα αποτελέσματα σχετικά με την καταλληλότητα των διαφορετικών επιλογών μοντελοποίησης της εποχιακότητας και χειρισμού της μη κανονικότητας, δεν προκύπτει κάποιος από τους διερευνώμενους συνδυασμούς εξωτερικών επιλογών περισσότερο αποτελεσματικός από τους υπόλοιπους.

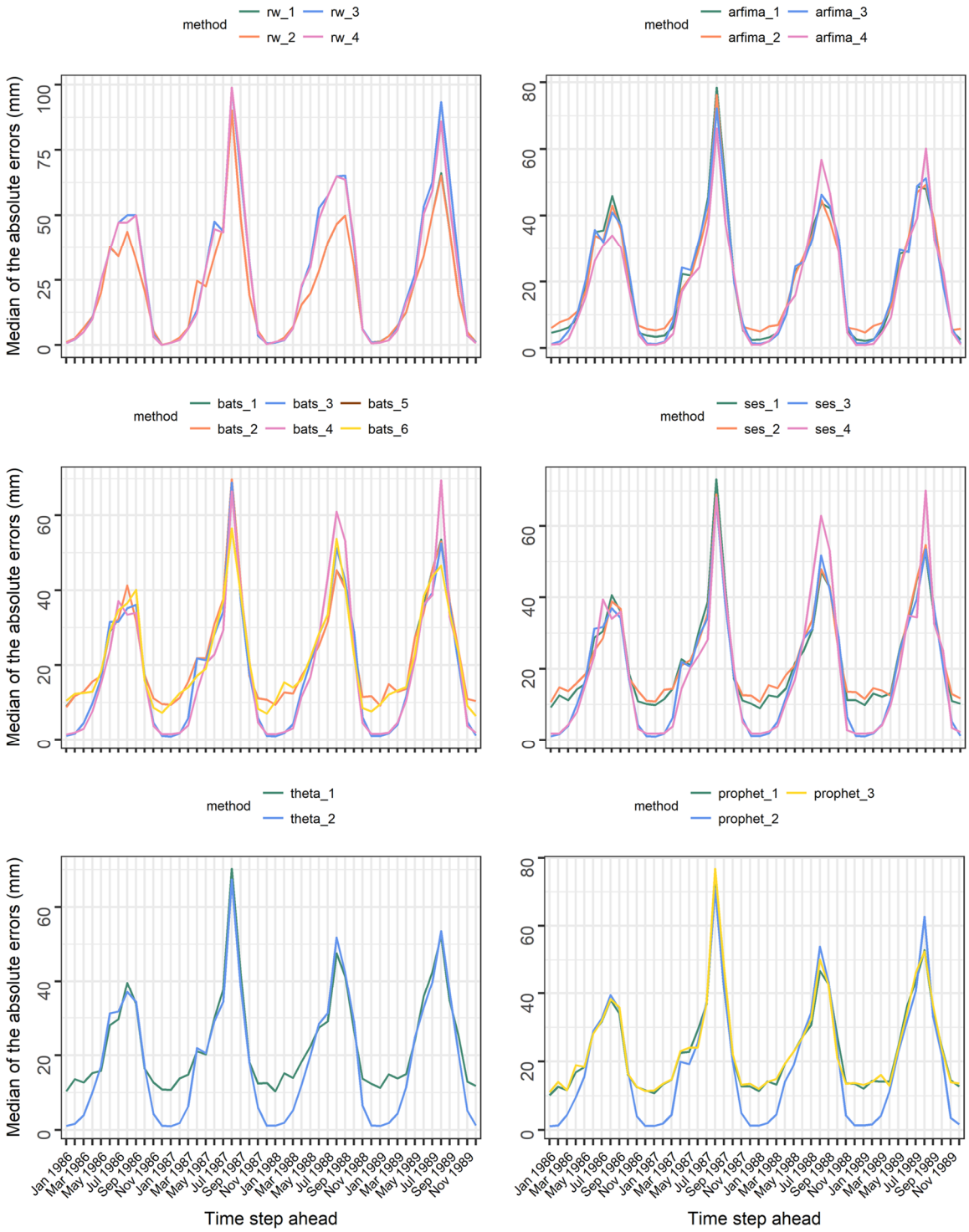
Ωστόσο, προκύπτει ότι η μοντελοποίηση της εποχιακότητας μέσω των μοντέλων BATS και Prophet (δηλαδή των μόνων μοντέλων που προσφέρουν αυτή τη δυνατότητα ανάμεσα στα χρησιμοποιούμενα) δίνει λιγότερο ακριβείς προβλέψεις από την εξωτερική μοντελοποίηση, ειδικά για το πρώτο μοντέλο.



Σχήμα 6. Διάμεσοι των απόλυτων σφαλμάτων πρόβλεψης σε κάθε βήμα του ορίζοντα πρόβλεψης για τις χρονοσειρές κατακρήμνισης που έχουν παρατηρηθεί (a) στην Βόρεια Αμερική, (b) στην Βόρεια Ευρώπη και (c) στην Βόρεια Αφρική στα αντίστοιχα πειράματα του Κεφαλάιου 5.



Σχήμα 7. Διάμεσοι των απόλυτων σφαλμάτων πρόβλεψης σε κάθε βήμα του ορίζοντα πρόβλεψης για τις χρονοσειρές κατακρήμνισης που έχουν παρατηρηθεί (a) στην Νότια Αφρική, (b) Ανατολική Ασία και (c) Αυστραλία στα αντίστοιχα πειράματα του Κεφαλάιου 5.



Σχήμα 8. Διάμεσοι των απόλυτων σφαλμάτων πρόβλεψης σε κάθε βήμα του ορίζοντα πρόβλεψης για τις χρονοσειρές κατακρήμνισης που έχουν παρατηρηθεί στην Βόρεια Αφρική στα αντίστοιχα πειράματα του Κεφαλαίου 5: Σύγκριση μεθόδων που βασίζονται στο ίδιο μοντέλο πρόβλεψης.

Τα ποσοτικά αποτελέσματα του Κεφαλαίου 5 είναι επίσης σημαντικά, καθώς εκφράζουν άμεσα την προβλεψιμότητα της μηνιαίας θερμοκρασίας και της μηνιαίας κατακρήμνισης. Η ελάχιστη και μέγιστη διάμεσος των απόλυτων σφαλμάτων των προβλέψεων θερμοκρασίας προκύπτουν περίπου ίσες με 0.25 K και 8.20 K, αντίστοιχα. Επιπλέον, υπολογίζεται μηδενικός μέσος όρος απόλυτων σφαλμάτων για τις προβλέψεις κατακρήμνισης τους ξηρούς μήνες σε γεωγραφικές περιοχές με σχετικά κανονική μεταβλητότητα κατακρημνίσεων, ενώ ο μέγιστος μέσος όρος είναι περίπου ίσος με 100 mm. Αυτές οι τιμές θα μπορούσαν να μελετηθούν σε

σύγκριση με την ελάχιστη και την μέγιστη διάμεσο των απόλυτων σφαλμάτων πρόβλεψης της ετήσιας θερμοκρασίας και κατακρήμνισης, όπως αυτές προκύπτουν χρησιμοποιώντας δύο σύνολα δεδομένων πραγματικού κόσμου με συνολικά 297 χρονοσειρές στο Κεφάλαιο 4. Αυτές οι διάμεσοι είναι περίπου ίσες με 0.23 K και 1.10 K, και 68 mm και 189 mm, αντίστοιχα. Επιπλέον, οι προκύπτουσες τιμές RMSE κυμαίνονται μεταξύ 1.01 K και 3.65 K για τη θερμοκρασία, και 36.16 mm και 70.17 mm για την κατακρήμνιση. Οι αντίστοιχες τιμές Nash-Sutcliffe είναι 0.79 και 0.98 για τη θερμοκρασία, και −0.55 και 0.71 για την κατακρήμνιση.

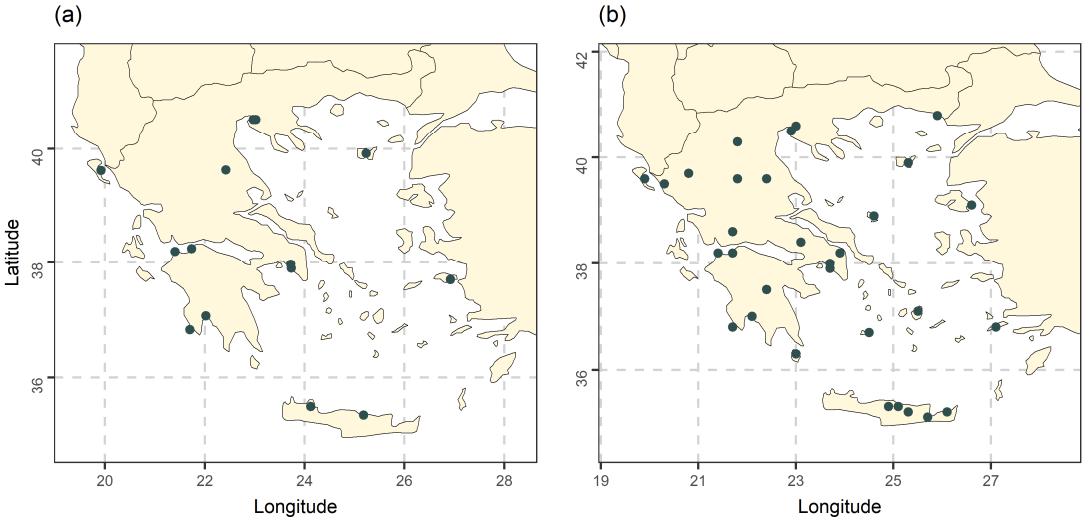
Εξαιρώντας την απλή μέθοδο αναφοράς και τις παραλλαγές του μοντέλου τυχαίου περιπάτου (random walk), οι αντίστοιχες τιμές RMSE κυμαίνονται μεταξύ 1.01 K και 2.84 K για τη θερμοκρασία, και 36.16 mm και 51.71 mm για την κατακρήμνιση. Λεπτομερέστερα, για το σύνολο των χρονοσειρών θερμοκρασίας, η χρήση ενός μοντέλου ARFIMA, BATS, απλής εκθετικής εξομάλυνσης, Theta ή Prophet, αντί της απλής μεθόδου αναφοράς, οδηγεί σε περίπου 19–29% πιο ακριβείς προβλέψεις όσον αφορά το RMSE, ή ακόμα και περίπου 30–32% ακριβέστερες προβλέψεις ειδικά για τις χρονοσειρές θερμοκρασίας που έχουν παρατηρηθεί στη Βόρεια Ευρώπη. Για το σύνολο των χρονοσειρών κατακρήμνισης, η χρήση όλων αυτών των μεθόδων οδηγεί σε 21–22% καλύτερες προβλέψεις από τη χρήση της απλής μεθόδου αναφοράς, ενώ για τις γεωγραφικές περιοχές της Βόρειας Αμερικής, της Βόρειας Ευρώπης και της Ανατολικής Ασίας τα ποσοστά αυτά είναι 26–29%, 22–24% και 32–38%, αντίστοιχα. Αυτός ο υψηλότερος βαθμός ακρίβειας είναι αξιοσημείωτος και ιδιαίτερα σημαντικός για εφαρμογές μακροχρόνιου ορίζοντα. Επίσης σημαντικό είναι το γεγονός ότι το μοντέλο Prophet προσφέρει από 13% έως και 32%, και από 16% έως και 38% καλύτερα αποτελέσματα από την απλή μέθοδο αναφοράς για τις χρονοσειρές θερμοκρασίας και κατακρήμνισης, αντίστοιχα. Επιπλέον, οι ελάχιστες και μέγιστες διάμεσοι Nash-Sutcliffe για τα μοντέλα ARFIMA, BATS, απλής εκθετικής εξομάλυνσης, Theta και Prophet είναι 0.89 και 0.98 για τη θερμοκρασία, και −0.04 και 0.71 για την κατακρήμνιση. Οι πρώτες τιμές Nash-Sutcliffe υποδηλώνουν καλές προβλέψεις, και οι τελευταίες είναι αποδεκτές έως μέτριες. Η μεγαλύτερη προβλεψιμότητα της μηνιαίας θερμοκρασίας σε σύγκριση με τη μηνιαία κατακρήμνιση αναμένεται ήδη από τη σύγκριση των αντίστοιχων τιμών τυπικής απόκλισης των εποχιακά αποσυντεθημένων χρονοσειρών. Οι συγκεκριμένες έχουν διάμεσους περίπου 1.70 K και 42 mm, αντίστοιχα. Θεωρούμε ότι το επίπεδο της ακρίβειας των προβλέψεων θα μπορούσε να βελτιωθεί ελάχιστα χρησιμοποιώντας άλλες μεθόδους, όπως καταδεικνύουν τα πειράματα του Κεφαλαίου 3.

Μια μελέτη πολλαπλών περιπτώσεων με έμφαση στους αλγόριθμους μηχανικής μάθησης

Το Κεφάλαιο 6 έχει ως γενικό του στόχο την προώθηση της διεξαγωγής μελετών πολλαπλών περιπτώσεων −στην εκτεταμένη τους κλίμακα− ως μίας καινοτόμου στρατηγικής και εναλλακτικής λύσης σε σχέση με τη διεξαγωγή μελετών μεμονωμένης περίπτωσης στον τομέα των προβλέψεων υδρολογικών χρονοσειρών. Η στρατηγική αυτή περιλαμβάνει την εξέταση περισσοτέρων της μίας μελετών περίπτωσης, διευκολύνοντας έτσι την παρατήρηση συγκεκριμένων φαινομένων από πολλαπλές οπτικές γωνίες ή εντός διαφορετικών πλαισίων. Για την ανίχνευση συστηματικών προτύπων σε κάθε μεμονωμένη περίπτωση, μπορεί να πραγματοποιηθεί δια-περιπτωσιακή σύνθεση. Δεδομένου του γεγονότος ότι τα όρια μεταξύ των φαινομένων και του πλαισίου δεν είναι ξεκάθαρα, είναι σημαντικό κάθε μεμονωμένη περίπτωση να διατηρεί την ταυτότητα της σε μια μελέτη πολλαπλών περιπτώσεων, έτσι ώστε ο ενδιαφερόμενος να μπορεί να επικεντρωθεί ειδικά σε αυτήν, εφόσον το επιθυμεί. Η διερεύνηση του συνόλου των περιπτώσεων (αλλά και μεμονωμένων περιπτώσεων από το σύνολο) μπορεί να προσφέρει ενδιαφέρουσες κατανοήσεις σχετικά με τα εξεταζόμενα φαινόμενα, καθώς και μια μορφή γενίκευσης που ονομάζεται "πιθανή εμπειρική γενίκευση", διατηρώντας παράλληλα την αμεσότητα της μεθόδου της μελέτης περίπτωσης.

Διεξάγουμε μια εκτεταμένη μελέτη πολλαπλών περιπτώσεων, αποτελούμενη από 50 επιμέρους μελέτες περιπτώσεων. Οι τελευταίες χρησιμοποιούν μηνιαίες χρονοσειρές θερμοκρασίας και κατακρήμνισης παρατηρημένες στην Ελλάδα (βλ. τις γεωγραφικές θέσεις των σταθμών μέτρησης στο Σχήμα 9). Εξετάζουμε αυτές τις δύο γεωφυσικές διεργασίες, επειδή παρουσιάζουν διαφορετικές ιδιότητες, οι οποίες μπορεί να επηρεάσουν διαφορετικά τα

αποτελέσματα των διερευνήσεων. Ο κύριος στόχος της διενεργούμενης μελέτης πολλαπλών περιπτώσεων είναι η διερεύνηση τριών προβλημάτων που σχετίζονται με την πρόβλεψη των υδρολογικών χρονοσειρών χρησιμοποιώντας αλγόριθμους μηχανικής μάθησης. Τα υπό διερεύνηση προβλήματα είναι: (α) η επιλογή μεταβλητών πρόβλεψης, (β) η επιλογή των υπερπαραμέτρων, και (γ) η σύγκριση των μεθόδων μηχανικής μάθησης και των στοχαστικών μεθόδων. Παρουσιάζουμε επίσης ποσοτικές πληροφορίες σχετικά με την ποιότητα των προβλέψεων (ιδιαίτερα σημαντικές για την περίπτωση της Ελλάδας) και αναζητάμε στοιχεία σχετικά με την ύπαρξη πιθανών σχέσεων μεταξύ της ποιότητας της πρόβλεψης και των εκτιμήσεων μέγιστης πιθανοφάνειας της τυπικής απόκλισης, του συντελεστή μεταβλητότητας και της παραμέτρου Hurst της ανέλιξης fractional Gaussian noise για τις εποχιακά αποσυντεθειμένες χρονοσειρές (που χρησιμοποιούνται για την προσαρμογή των μοντέλων).

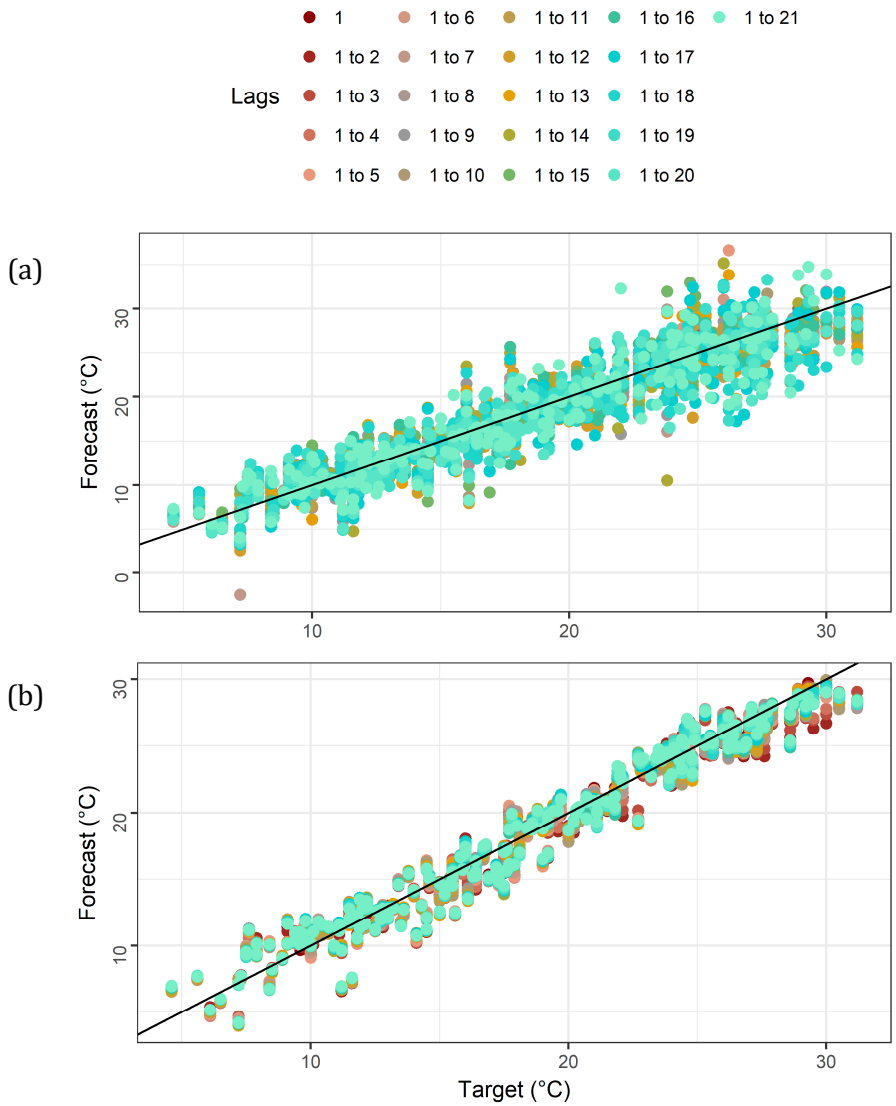


Σχήμα 9. Γεωγραφικές θέσεις σταθμών μέτρησης (a) θερμοκρασίας και (b) κατακρήμνισης. Μηνιαία δεδομένα από τους συγκεκριμένους σταθμούς χρησιμοποιούνται για τα πειράματα του Κεφαλαίου 6.

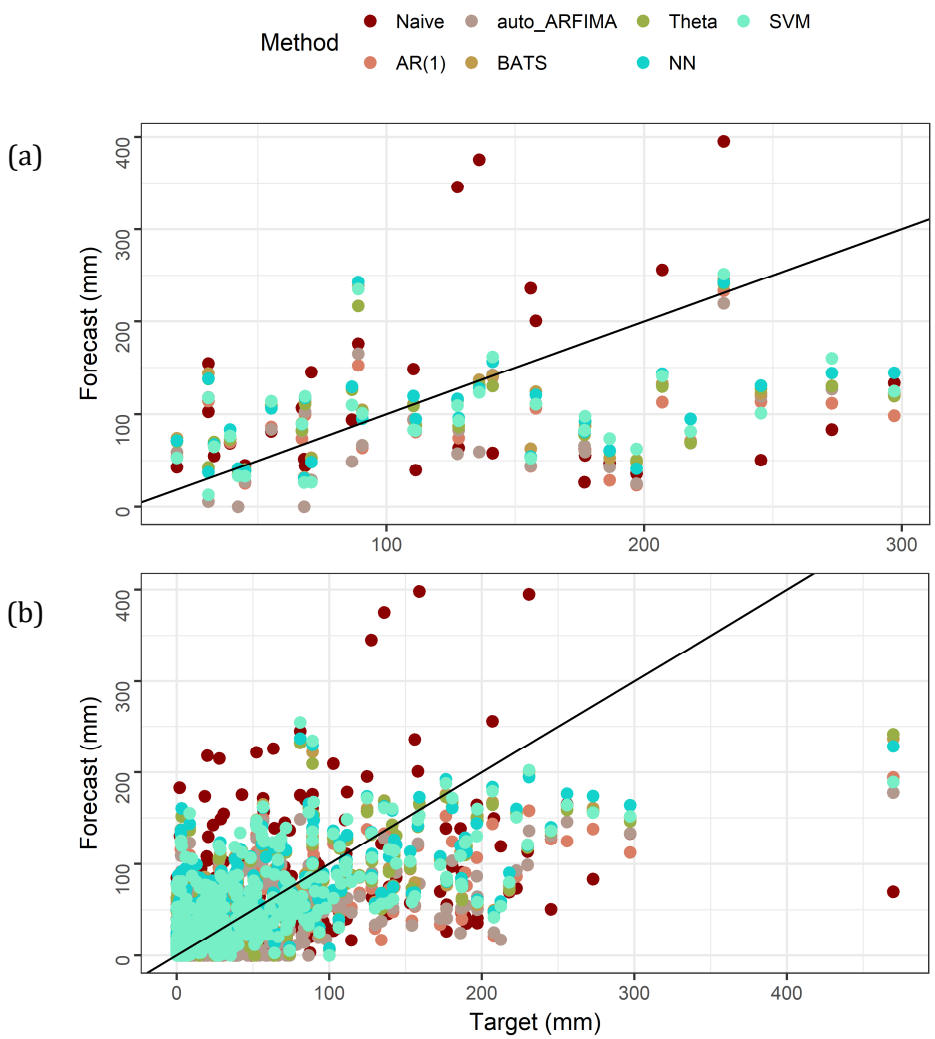
Επικεντρωνόμαστε σε δύο αλγορίθμους μηχανικής μάθησης, συγκεκριμένα στα νευρωνικά δίκτυα (neural networks) και στα support vector machines, ενώ επίσης συμπεριλαμβάνουμε τέσσερις στοχαστικές μεθόδους και την εποχιακή μέθοδο αναφοράς στις συγκρίσεις μας. Οι στοχαστικές μέθοδοι είναι (i) ένα πλήρως αυτοματοποιημένο μοντέλο AR(1), (ii) ένα πλήρως αυτοματοποιημένο μοντέλο ARFIMA, (iii) το μοντέλο BATS, και (iv) το μοντέλο Theta. Εφαρμόζουμε κοινή μεθοδολογία σε κάθε μεμονωμένη περίπτωση και, στη συνέχεια, πραγματοποιούμε δια-περιπτωσιακή σύνθεση για να διευκολύνουμε την ανίχνευση επαναλλαμβανόμενων μοτίβων. Προσαρμόζουμε τα μοντέλα σε εποχιακά αποσυντεθειμένες χρονοσειρές και ανακτούμε την εποχιακότητα στις προβλέψεις. Συγκρίνουμε την επίδοση των αλγόριθμων στην πρόβλεψη ενός και δώδεκα βημάτων μπροστά. Η αξιολόγηση της επίδοσης των μεθόδων στην πρόβλεψη ενός βήματος μπροστά βασίζεται στο απόλυτο σφάλμα της πρόβλεψης της τελευταίας μηνιαίας παρατήρησης. Η αξιολόγηση της επίδοσης των μεθόδων στην πρόβλεψη πολλαπλών βημάτων γίνεται για τις μηνιαίες παρατηρήσεις του τελευταίου έτους και βασίζεται σε πέντε μέτρα. Τα τελευταία είναι το RMSE, το Nash-Sutcliffe, ο λόγος τυπικών αποκλίσεων, ο συντελεστής συσχέτισης και ο δείκτης συμφωνίας.

Τα αποτελέσματα (βλ. π.χ., Σχήματα 10 και 11) καταδεικνύουν ότι μέθοδοι πρόβλεψης που βασίζονται στον ίδιο αλγόριθμο μηχανικής μάθησης μπορεί να παρουσιάζουν πολύ διαφορετικές επιδόσεις, σε βαθμό που εξαρτάται κυρίως από τον αλγόριθμο και την υπό διερεύνηση περίπτωση. Πράγματι, ο αλγόριθμος νευρωνικών δικτύων (neural networks) μπορεί να δώσει προβλέψεις αρκετά διαφορετικής ποιότητας για μια συγκεκριμένη περίπτωση, σε αντίθεση με τα support vector machines. Η επίδοση του πρώτου αλγορίθμου φαίνεται να επηρεάζεται περισσότερο από την επιλογή των μεταβλητών πρόβλεψης παρά από τη διαδικασία επιλογής των υπερπαραμέτρων (χρήση προκαθορισμένων υπερπαραμέτρων ή επιλογή μετά από

βελτιστοποίηση). Παρόλο που κανένα από τα συγκρινόμενα σετ μεταβλητών πρόβλεψης δεν οδηγεί σε συστηματικά καλύτερες προβλέψεις από τα υπόλοιπα, τόσο για τα νευρωνικά δίκτυα (neural networks) όσο και για τα support vector machines, τα αποτελέσματα ευνοούν περισσότερο τη χρήση λιγότερων και πρόσφατων μεταβλητών πρόβλεψης. Επιπλέον, η βελτιστοποίηση υπερπαραμέτρων φαίνεται να μην οδηγεί απαραιτήτως σε καλύτερες προβλέψεις από τη χρήση των προεπιλεγμένων τιμών υπερπαραμέτρων για τους υπό διερεύνηση αλγορίθμους. Όσον αφορά τις συγκρίσεις που πραγματοποιούνται μεταξύ των αλγορίθμων μηχανικής μάθησης και των κλασσικών αλγορίθμων, τα αποτελέσματα δείχνουν ότι μέθοδοι και από τις δύο κατηγορίες μπορούν να φανούν εξίσου χρήσιμες. Η καλύτερη μέθοδος εξαρτάται από την περίπτωση που εξετάζεται και το κριτήριο ενδιαφέροντος, ενώ μπορεί να είναι είτε μηχανικής μάθησης είτε κλασσική. Ακολουθούν ορισμένες πληροφορίες δευτερεύουσας σημασίας, όπως αυτές προκύπτουν από τα πειράματα του Κεφαλαίου: Η μέση επίδοση των αλγορίθμων που χρησιμοποιούνται για τις προβλέψεις θερμοκρασίας ενός και δώδεκα βημάτων μπροστά κυμαίνεται μεταξύ 0.66°C και 1.00°C (σε όρους απόλυτου σφάλματος πρόβλεψης), και 1.14°C και 1.70°C (σε όρους RMSE πρόβλεψης), αντίστοιχα. Για τις μηνιαίες προβλέψεις κατακρήμνισης οι αντίστοιχες τιμές είναι 39 mm και 72 mm, και 41 mm και 52 mm. Τέλος, από την μελέτη πολλαπλών περιπτώσεων δεν προκύπτει κανένα στοιχείο που να καταδεικνύει την ύπαρξη σχέσης ανάμεσα στην ποιότητα των προβλέψεων και στις εκτιμήσεις μέγιστης πιθανοφάνειας της τυπικής απόκλισης, του συντελεστή μεταβλητότητας και της παραμέτρου Hurst της ανέλιξης fractional Gaussian noise για τις εποχιακά αποσυντεθειμένες χρονοσειρές.



Σχήμα 10. Προβλέψεις μηνιαίας θερμοκρασίας δώδεκα βημάτων μπροστά, παρηγμένες στο Κεφάλαιο 6 για τη διερεύνηση του Προβλήματος 1 και τους αλγορίθμους (a) NN and (b) SVM, σε σύγκριση με τις αντίστοιχες παρατηρημένες τιμές μηνιαίας θερμοκρασίας.



Σχήμα 11. Προβλέψεις μηνιαίας κατακρήμνισης (a) ενός και (b) δώδεκα βημάτων μπροστά, παρηγμένες στο Κεφάλαιο 6 για τη διερεύνηση του Προβλήματος 3, σε σύγκριση με τις αντίστοιχες παρατηρημένες τιμές μηνιαίας κατακρήμνισης.

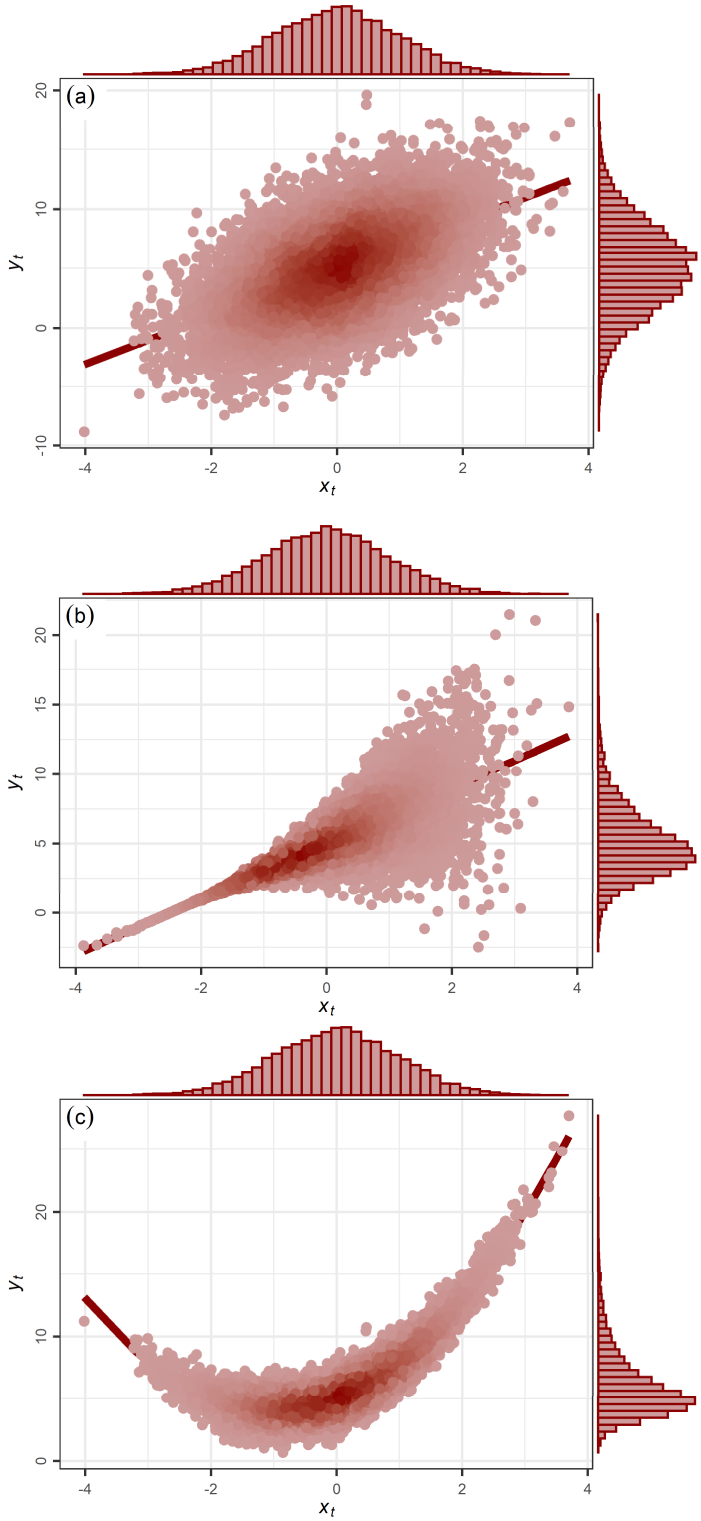
Πιθανοτική μετεπεξεργασία αποτελεσμάτων υδρολογικής μοντελοποίησης

Μια νέα μεθοδολογία και η διερεύνησή της μέσω πειραμάτων πρότυπης μοντελοποίησης

Στο Κεφάλαιο 7 αναπτύσσουμε μια νέα μεθοδολογία πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης, χρησιμοποιώντας ως σημείο εκκίνησης ένα θεωρητικά συνεπές γενικό σχήμα πιθανοτικής υδρολογικής μοντελοποίησης. Η προτεινόμενη μεθοδολογία υποδιαιρείται σε τρεις εναλλακτικές παραλλαγές. Εν συντομία, παράγει ένα μεγάλο αριθμό σημειακών (συνήθως προσδιοριστικών) προβλέψεων χρησιμοποιώντας ένα μόνο διεργασιακό υδρολογικό μοντέλο, αλλά με διαφορετικές τιμές παραμέτρων. Αυτές οι "αδελφές προβλέψεις" μετατρέπονται στη συνέχεια σε βοηθητικές πιθανοτικές προβλέψεις (καθεμία από τις οποίες αποτελείται από έναν αριθμό προβλέψεων ποσοστημορίων) μέσω της επίλυσης ενός προβλήματος παλινδρόμησης ποσοστημορίου. Η επίλυση αυτή βασίζεται σε έναν κατάλληλο αλγόριθμο (στο εξής αναφερόμενο ως το "μοντέλο σφάλματος" της μεθοδολογίας). Οι βοηθητικές πιθανοτικές προβλέψεις τελικώς συνδυάζονται υπολογίζοντας τον μέσο όρο των προβλέψεων των ποσοστημορίων με την ίδια πιθανότητα. Από όσο γνωρίζουμε, η νέα μεθοδολογία είναι η πρώτη μεθοδολογία πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης που παράγει και αξιοποιεί διαφορετικά σύνολα πληροφορίας χρησιμοποιώντας ένα μόνο (διεργασιακό) μοντέλο με διαφορετικές τιμές παραμέτρων και ένα μοντέλο παλινδρόμησης ποσοστημορίου.

Μέσω της χρήσης μοντέλων παλινδρόμησης ποσοστημορίου επιτυγχάνεται κάποια πρόοδος (σε σχέση με την αρχική υλοποίηση του γενικού σχήματος και τις παραλλαγές του τελευταίου που είναι προγενέστερες των Κεφαλαίων 7 και 8) όσον αφορά την ευελιξία στη μοντελοποίηση. Τα μοντέλα αυτά προβλέπουν ποσοστημόρια με δεδομένη πιθανότητα και όχι ολόκληρη την κατανομή της μεταβλητής ενδιαφέροντος, ενώ παράλληλα είναι κατάλληλα για τη μοντελοποίηση της ετεροσκεδαστικότητας. Τέτοια μοντέλα (βλ. επίσης το Κεφάλαιο 9) είναι το μοντέλο παλινδρόμησης ποσοστημορίου (quantile regression), το μοντέλο γενικευμένα τυχαία δάση (generalized random forests) για παλινδρόμηση ποσοστημορίου, το μοντέλο γενικευμένα τυχαία δάση (generalized random forests) για παλινδρόμηση ποσοστημορίου μιμούμενο το μοντέλο τυχαία δάση για παλινδρόμηση ποσοστιμορίου (quantile regression forests), το μοντέλο gradient boosting machine, το μοντέλο model-based boosting με γραμμικά μοντέλα βάσης και το μοντέλο νευρωνικά δίκτυα παλινδρόμησης ποσοστιμορίου (quantile regression neural networks). Η δυνατότητα για εκμετάλλευση της ευελιξίας που παρέχεται από τα μοντέλα παλινδρόμησης ποσοστημορίου θα πρέπει να θεωρηθεί σημαντικό πλεονέκτημα της προτεινόμενης μεθοδολογίας από πρακτική άποψη.

Δείχνουμε τη χρησιμότητα της προτεινόμενης μεθοδολογίας και πώς η κατανόηση μας για το μοντελοποιούμενο σύστημα μπορεί να μας οδηγήσει στην επίτευξη βελτιωμένης προγνωστικής μοντελοποίησης διεξάγοντας διερευνήσεις πρότυπης μοντελοποίησης (βασιζόμενες στα σύνολα δεδομένων του Σχήματος 12). Στο πλαίσιο των συγκεκριμένων διερευνήσεων, εστιάζουμε στην ακαταλληλότητα της παραδοχής της ομοσκεδαστικότητας, όταν αυτή γίνεται κατά την μοντελοποίηση του σφάλματος πρόβλεψης του υδρολογικού μοντέλου, και στο πώς η επιλογή ενός κατάλληλου μοντέλου παλινδρόμησης οδηγεί σε βελτιωμένες πιθανοτικές προβλέψεις. Δείχνουμε, επίσης, τη σημασία της χρήσης ενός πιο ακριβούς υδρολογικού μοντέλου για την παροχή πιθανοτικών προβλέψεων που να είναι ταυτόχρονα αξιόπιστες και όσο το δυνατόν πιο στενές. Τέλος, χρησιμοποιούμε τα αποτελέσματα της πρότυπης μοντελοποίησης για να δείξουμε πώς η προτεινόμενη μεθοδολογία χαρακτηρίζεται από μεγαλύτερη ευρωστία. Η τελευταία επιτυγχάνεται υπολογίζοντας τον μέσο όρο πολλών ποσοστημοριακών προβλέψεων.



Σχήμα 12. Προσομοιωμένα σύνολα δεδομένων (a–c) 1–3. Τα συγκεκριμένα σύνολα δεδομένων χρησιμοποιούνται για τα πειράματα του Κεφαλαίου 7.

Παρά το γεγονός ότι επικεντρωνόμαστε στην προτεινόμενη μεθοδολογία, ορισμένα από τα αποτελέσματα του Κεφαλαίου μπορούν να χρησιμοποιηθούν για να την απόκτηση γενικής εικόνας σχετικά με τον τρόπο λειτουργίας των μεθοδολογιών πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης δύο βημάτων και τις συνθήκες κάτω από τις οποίες μεγιστοποιείται η προγνωστική επίδοση τους. Τα παρουσιαζόμενα παραδείγματα πρότυπης μοντελοποίησης καταδεικνύουν την μεγάλη σημασία τόσο του μοντέλου παλινδρόμησης ποσοστημορίου όσο και του υδρολογικού μοντέλου για μια μεθοδολογία

πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης δύο βημάτων, ενώ προχωρούν ορισμένα βήματα μπροστά σε σχέση με ορισμένες παραδειγματικές (αλλά ταυτόχρονα και βασικές) δοκιμές πρότυπης μοντελοποίησης που έχουν πραγματοποιηθεί μέχρι στιγμής για την ερμηνεία διάφορων μεθοδολογιών για την ποσοτικοποίηση της προγνωστικής υδρολογικής αβεβαιότητας. Τέτοιες δοκιμές πρότυπης μοντελοποίησης υιοθετούν, ως επί το πλείστον, την παραδοχή της ομοσκεδαστικότητας και ένα τέλειο "υδρολογικό μοντέλο", ενώ οι διερευνήσεις πρότυπης μοντελοποίησης του Κεφαλαίου 7 είναι εμπνευσμένες και από πρόσφατα πειράματα προσομοίωσης που δεν βασίζονται στις συγκεκριμένες παραδοχές.

Δύο ελκυστικές και ταυτόχρονα χρήσιμες ιδιότητες της προτεινόμενης μεθοδολογίας (εκτενώς διερευνώμενες στο Κεφάλαιο 8) είναι: (α) η μεγαλύτερη ευρωστία της σε σύγκριση με τις επιμέρους προβλέψεις που συνδυάζονται στο πλαίσιό της και, κατά συνέπεια, σε σύγκριση με βασικές μεθοδολογίες πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης δύο βημάτων (οι οποίες παράγουν μία μόνο πιθανοτική πρόβλεψη και δεν συνδυάζουν προβλέψεις), και (β) η ικανότητα της να "αξιοποιεί τη σοφία του πλήθους". Η τελευταία ορίζεται στη βιβλιογραφία της πρόβλεψης χρονοσειρών ως η ιδιότητα ορισμένων συνδυασμών προβλέψεων να είναι τουλάχιστον το ίδιο καλές –συνήθως καλύτερες– ως προς ένα ορισμένο μέτρο από τον μέσο όρο των τιμών που λαμβάνει το ίδιο μέτρο για καθεμία από τις επιμέρους προβλέψεις που συνδυάζονται. Στην πραγματικότητα, όσο μεγαλύτερος είναι ο αριθμός των συνδυαζόμενων προβλέψεων (ίσος με τον αριθμό των παραγόμενων αδελφών προβλέψεων), τόσο πιο μεγάλη είναι η ευρωστία της προτεινόμενης μεθοδολογίας και τόσο περισσότερο αξιοποιείται η σοφία του πλήθους.

Η προτεινόμενη μεθοδολογία χαρακτηρίζεται από ορισμένα πρόσθετα πλεονεκτήματα, τα οποία είναι ιδιαίτερα σημαντικά υπό το πρίσμα της προγνωστικής μοντελοποίησης. Πρώτον, είναι υπολογιστικά βολική υπό την έννοια ότι μπορεί εύκολα να εκφραστεί σε αλγοριθμική μορφή και να προγραμματιστεί χρησιμοποιώντας ανοιχτό λογισμικό. Δεύτερον, προσφέρει ορισμένες επιλογές μοντελοποίησης που θα μπορούσαν να αξιοποιηθούν για τη μεγιστοποίηση της προγνωστικής της επίδοσης. Παραδείγματος χάριν, οι δύο από τις τρεις παραλλαγές επιτρέπουν την αξιοποίηση από το μοντέλο σφάλματος ενός μεγάλου αριθμού διαφορετικών συνόλων πληροφορίας, αντί του ενός συνόλου πληροφορίας (που αξιοποιεί η τρίτη παραλλαγή), διευκολύνοντας έτσι τη διεύρυνση του χώρου δειγματοληψίας των παρατηρούμενων σφαλμάτων πρόβλεψης του υδρολογικού μοντέλου. Αυτή η διεύρυνση θα μπορούσε να είναι ιδιαίτερα σημαντική στην περίπτωση της μοντελοποίησης αυτών των σφαλμάτων χρησιμοποιώντας μεθόδους που δεν κάνουν προέκταση (extrapolation), όπως το μοντέλο δάση παλινδρόμησης ποσοστημορίου (quantile regression forests). Τέλος, επιτρέπει την πλήρη αξιοποίηση της παραγόμενης πληροφορίας, με την έννοια ότι κάθε αδελφή πρόβλεψη μετατρέπεται σε πιθανοτική πρόβλεψη και όχι σε μια πιθανή πραγματοποίηση της διεργασίας που μας ενδιαφέρει (την αρχική υλοποίηση του γενικού σχήματος και τις παραλλαγές που είναι προγενέστερες των Κεφαλαίων 7 και 8).

Θα πρέπει επίσης να συζητηθούν ορισμένοι περιορισμοί που συνοδεύουν την προτεινόμενη μεθοδολογία. Αυτοί περιλαμβάνουν περιορισμούς που απορρέουν από την φύση της ίδιας της μεθοδολογίας (που επιτάσσει την αξιοποίηση της ιστορικής πληροφορίας σε δύο διαδοχικά βήματα). Τέτοιοι είναι ο βαθμός στον οποίο τα αποτελέσματα της μοντελοποίησης μπορούν να ερμηνευτούν (ειδικά όσον αφορά την παραγωγή και αξιοποίηση ερμηνεύσιμων τιμών παραμέτρων) και οι σημαντικές απαιτήσεις για μεγάλο μέγεθος ιστορικών χρονοσειρών. Αν και αυτός ο τελευταίος περιορισμός πρέπει να σημειωθεί και ίσως να ληφθεί υπόψιν σε πρακτικές εφαρμογές, οι ημερήσιες χρονοσειρές βροχής-απορροής είναι συνήθως ικανοποιητικού μήκους. Επιπλέον, στο Κεφάλαιο 8 αποδεικνύεται εμπειρικά ότι, ακόμη και όταν η διαθέσιμη ιστορική πληροφορία είναι λίγη, η προτεινόμενη μεθοδολογία έχει καλή επίδοση όταν η υλοποίηση της βασίζεται στο μοντέλο παλινδρόμησης ποσοστημορίου (quantile regression) ως μοντέλο σφάλματος.

Επιπλέον, οι υπολογιστικές απαιτήσεις της προτεινόμενης μεθοδολογίας είναι (επί του παρόντος) μεγάλες όταν (α) επιλέγονται υπολογιστικά δαπανηροί αλγόριθμοι (π.χ., αλγόριθμοι Markov Chain Monte Carlo) για τη βαθμονόμηση του υδρολογικού μοντέλου, και/ή (β) το μοντέλο

σφάλματος αξιοποιεί την ιστορική πληροφορία όπως επιτάσσουν δύο από τις τρεις παραλλαγές της, εκτός από την περίπτωση που η υλοποίηση της περιορίζεται στην παραγωγή και αξιοποίηση ενός μικρού αριθμού αδελφών προβλέψεων. Θα πρέπει να σημειωθεί, στο σημείο αυτό, ότι ένας υπολογιστικά βολικός και απλός αλγόριθμος δεν είναι απαραίτητα και υπολογιστικά γρήγορος. Είναι επίσης σημαντικό να διευκρινιστεί ότι ο παραπάνω περιορισμός ισχύει μόνο για εφαρμογές σε εκατοντάδες λεκάνες απορροής και χρονικές κλίμακες μικρότερες από την μηνιαία, καθώς και για εφαρμογές μέσω συνήθων προσωπικών υπολογιστών. Δεν ισχύει για εφαρμογές σε μικρό αριθμό λεκανών απορροής ούτε για εφαρμογές σε μηνιαία και ετήσια χρονική κλίμακα. Ακόμα, εφαρμογές μεγάλης κλίμακας σε ημερήσια σύνολα δεδομένων υποστηρίζονται επαρκώς από την τρίτη παραλλαγή της μεθοδολογίας, όταν αυτή η παραλλαγή υλοποιείται με τη χρήση υπολογιστικά γρήγορων αλγορίθμων για την βαθμονόμηση του υδρολογικού μοντέλου.

Εκτός από τα παραπάνω ζητήματα και σε αντίθεση με αρκετές στατιστικές μεθοδολογίες πιθανοτικής πρόβλεψης, ένα ευρέως παραδεκτό μειονέκτημα των ευέλικτων μοντέλων μηχανικής μάθησης για πρόβλεψη ποσοστημορίων (που αποτελούν την βάση της προτεινόμενης μεθοδολογίας) είναι η ακαταλληλότητα τους για την μοντελοποίηση της μακροπρόθεσμης εμμονής. Η συγκεκριμένη μοντελοποίηση κατά την επίλυση προβλημάτων πρόβλεψης είναι σημαντική προτεραιότητα στην βιβλιογραφία της εφαρμοσμένης στοχαστικής υδρολογίας (βλ. π.χ., τις διερευνήσεις μεγάλης κλίμακας των Κεφαλαίων 3–5 και τη συγκριτική μελέτη περίπτωσεων του Κεφαλαίου 6). Παραταύτα, εμπειρικά αποτελέσματα καταδεικνύουν ότι η Μαρκοβιανή παραδοχή (που κατά κάποιον τρόπο επιτρέπεται από την προτεινόμενη μεθοδολογία χρησιμοποιώντας ως μεταβλητή πρόβλεψης στην παλινδρόμηση την πρόβλεψη του υδρολογικού μοντέλου για την χρονική στιγμή $t-1$) είναι εύλογη για την μοντελοποίηση των σφαλμάτων πρόβλεψης των υδρολογικών μοντέλων. Γενικά, συμπεριλαμβάνοντας περισσότερες (από μία) μεταβλητές πρόβλεψης (π.χ., τις προβλέψεις του υδρολογικού μοντέλου για τις χρονικές στιγμές t, $t-1$, $t-2$, κ.λπ.) στο πρόβλημα παλινδρόμησης, μπορούμε να αυξήσουμε την ποσότητα πληροφορίας που αξιοποιείται και να βελτιώσουμε την πρόβλεψη, όπως αποδεικνύεται εμπειρικά για προβλήματα βροχής-απορροής του Κεφαλαίου 9 της παρούσας διατριβής.

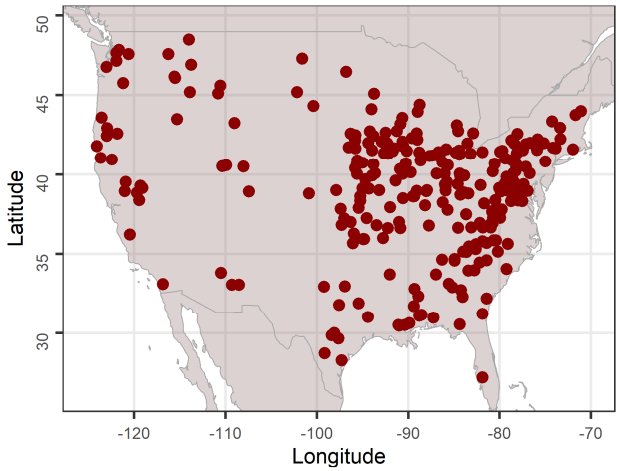
Εν κατακλείδι, το βασικό δίλημμα που καλείται να αντιμετωπίσει κανείς κατά την επιλογή μεταξύ της προτεινόμενης μεθοδολογίας και των βασικών μεθοδολογιών πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης δύο σταδίων (όταν γίνεται χρήση του ίδιου μοντέλου σφάλματος) είναι αυτό ανάμεσα (α) στην μεγαλύτερη ευρωστία που χαρακτηρίζει την πρώτη και στην ικανότητα της να αξιοποιεί τη σοφία του πλήθους, και (β) στις πολύ λιγότερες υπολογιστικές απαιτήσεις των τελευταίων μεθοδολογιών. Πιστεύουμε ότι από την σκοπιά της διαχείρισης κινδύνου η διάθεση των επιπρόσθετων υπολιστικών πόρων είναι συμφέρουσα, όπως καταδεικνείει το μεγάλης κλίμακας πείραμα πραγματικού κόσμου του Κεφαλαίου 8.

Διερευνήσεις μεγάλου υδρολογικού δείγματος με έμφαση στην αξιολόγηση της ευρωστίας

Το Κεφαλαίο 8 έχει ως γενικό του στόχο τη διερεύνηση σε πραγματικά προβλήματα της μεθοδολογίας πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης που αναπτύσσεται στο Κεφάλαιο 7. Η συγκεκριμένη μεθοδολογία υιοθετεί βασικά στοιχεία από μια θεωρητικά συνεπή και ευέλικτη μεθοδολογία πιθανοτικής υδρολογικής μοντελοποίησης, ενώ επίσης στηρίζεται σε απλές μεθόδους συνδυασμού προβλέψεων από το πεδίο της πρόβλεψης χρονοσειρών. Χρησιμοποιεί ένα οποιοδήποτε διεργασιακό υδρολογικό μοντέλο για να δημιουργήσει έναν μεγάλο αριθμό «αδελφών προβλέψεων» υιοθετώντας ισάριθμα σετ παραμέτρων. Οι παράμετροι του διεργασιακού υδρολογικού μοντέλου προκύπτουν χρησιμοποιώντας Μπεϋζιανά ή άλλα σχήματα βαθμονόμησης. Επομένως, αυτή η μεθοδολογία δεν έχει κάποια ιδιαίτερη σχέση εκ κατασκευής με Μπεϋζιανές μεθόδους, όπως ισχύει και για την μητρική μεθοδολογία. Ένα μοντέλο παλινδρόμησης ποσοστημορίου (βλ. π.χ., τα μοντέλα που διερευνώνται στο Κεφάλαιο 9 της παρούσας διατριβής) χρησιμοποιείται στη συνέχεια για την μοντελοποίηση του σφάλματος πρόβλεψης του υδρολογικού μοντέλου. Μέσω αυτής της μοντελοποίησης οι αδελφές προβλέψεις μετατρέπονται σε πιθανοτικές προβλέψεις. Οι

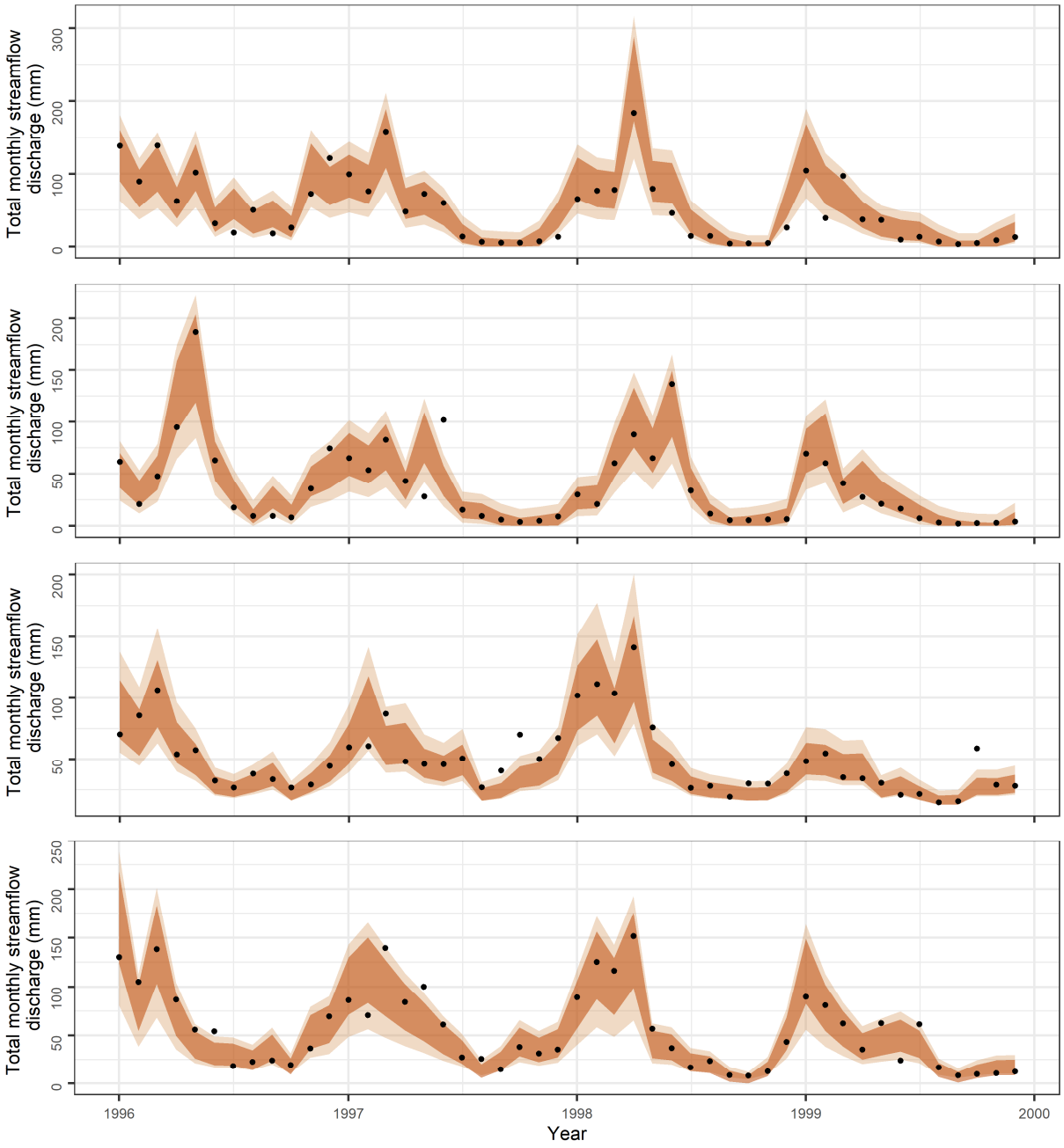
τελευταίες τελικά συνδυάζονται με απλό τρόπο για να δώσουν τις παραδοτέες προβλέψεις ποσοστιμορίων. Η μεθοδολογία υπό αξιολόγηση υποδιαιρείται σε τρεις εναλλακτικές παραλλαγές, οι οποίες διαφέρουν μόνο ως προς την εκπαίδευση του μοντέλου παλινδρόμησης ποσοστημορίου.

Πραγματοποιούμε ένα πείραμα πραγματικού κόσμου σε μηνιαία κλίμακα. Στο πλαίσιο του συγκεκριμένου πειράματος χρησιμοποιούμε πλήρεις (χωρίς ελλείπουσες τιμές) ημερήσιες χρονοσειρές 50 ετών για 270 λεκάνες απορροής στις Ηνωμένες Πολιτείες (βλ. π.χ., τις γεωγραφικές θέσεις των σταθμών μέτρησης της απορροής στο Σχήμα 13). Προκειμένου να βελτιώσουμε την κατανόηση γύρω από την πιθανοτική υδρολογική μοντελοποίηση, επιμένουμε στην χρήση ερμηνεύσιμων μοντέλων και στη συγκριτική αξιολόγηση εντός όλων των διεξαχθεισών δοκιμών. Χρησιμοποιούμε το φειδωλό διεργασιακό υδρολογικό μοντέλο GR2M και δύο (σε μεγάλο βαθμό) ερμηνεύσιμα μοντέλα παλινδρόμησης, συγκεκριμένα το γραμμικό μοντέλο παλινδρόμησης και μοντέλο παλινδρόμησης ποσοστημορίου (quantile regression). Εφαρμόζουμε έξι σχήματα πιθανοτικής προγνωστικής μοντελοποίησης, όλα βασιζόμενα στην προτεινόμενη μεθοδολογία. Εκείνα τα σχήματα που βασίζονται στο γραμμικό μοντέλο (τρία σε αριθμό) χρησιμοποιούνται ως σημεία αναφοράς για τα υπόλοιπα σχήματα (επίσης τρία σε αριθμό). Εκείνα τα σχήματα που βασίζονται στο ίδιο μοντέλο παλινδρόμησης χρησιμοποιούν διαφορετικές παραλλαγές της υπό αξιολόγηση μεθοδολογίας. Η επίδοση των έξι σχημάτων πιθανοτικής προγνωστικής μοντελοποίησης ποσοτικοποιείται υπολογίζοντας τις πιθανότητες κάλυψης, τα μέσα πλάτη και τις μέσες τιμές του μέτρου διαστήματος πρόβλεψης των διαστημάτων πρόβλεψης, καθώς επίσης και μέσω συγκριτικής αξιολόγησης των αποτελεσμάτων που παρέχουν σε σχέση με τα αποτελέσματα αφελών πιθανοτικών μοντέλων.

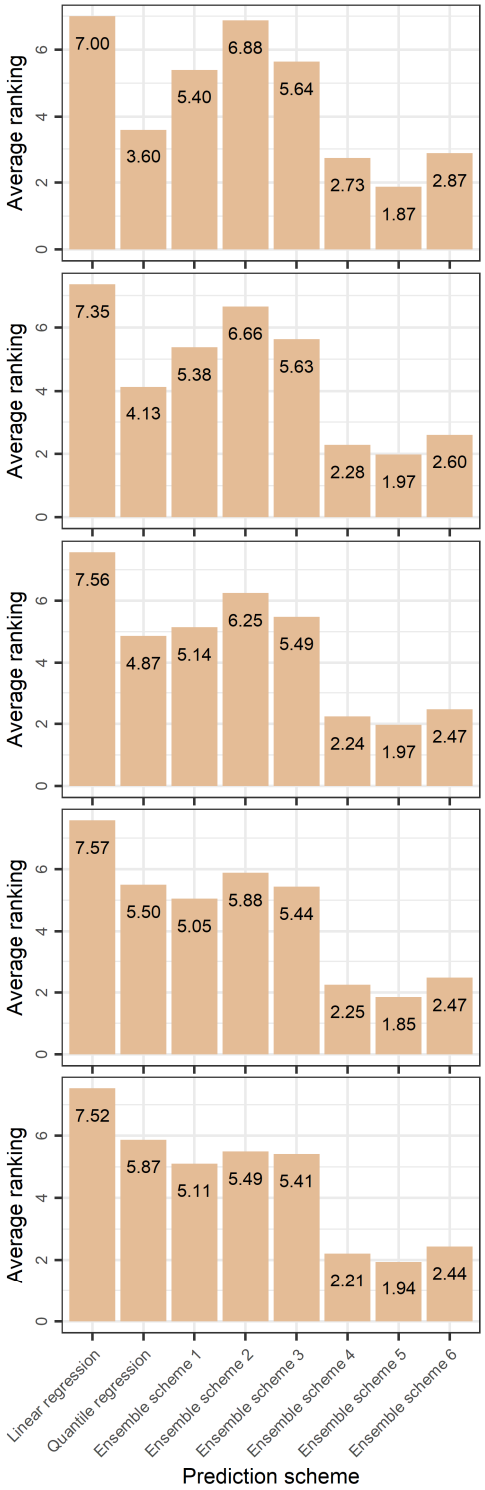


Σχήμα 13. Γεωγραφικές θέσεις 270 σταθμών απορροής ποταμών. Μηνιαία δεδομένα από τους συγκεκριμένους σταθμούς χρησιμοποιούνται για τα πειράματα του Κεφαλαίου 8.

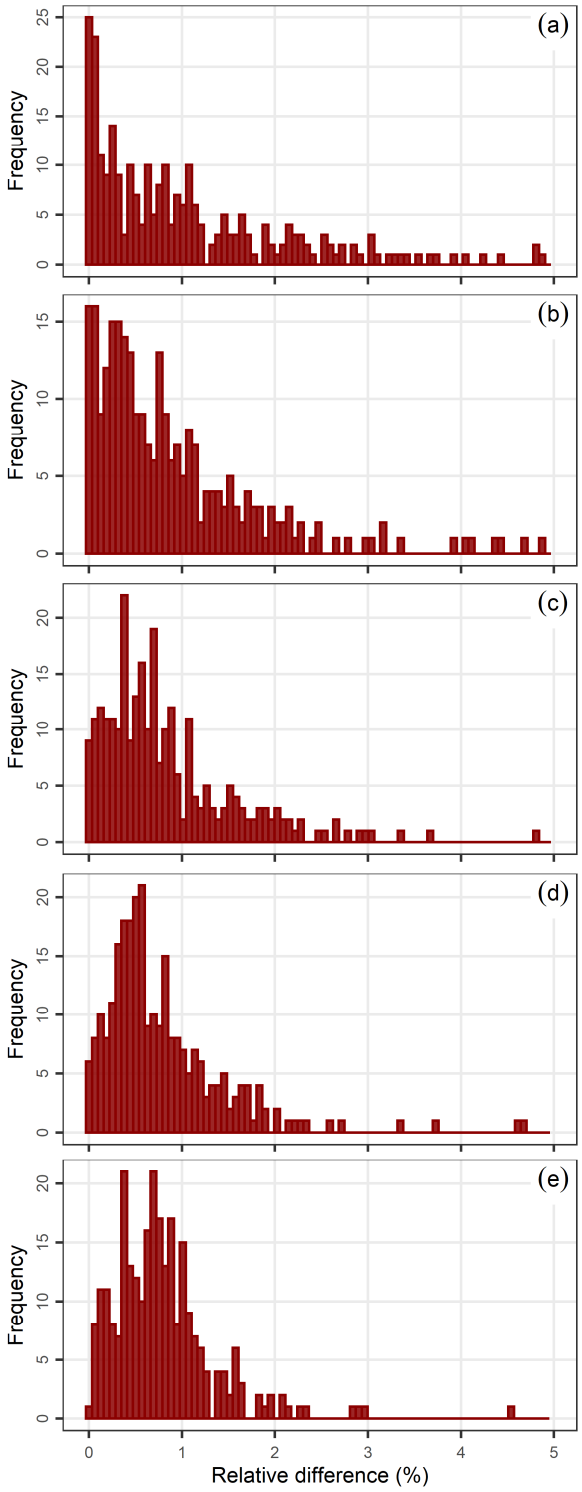
Τα πειραμαματικά αποτελέσματα (τιμές μέτρων για 4 870 800 διαστήματα πρόβλεψης) υποδεικνύουν τη χρησιμότητα της υπό αξιολόγηση μεθοδολογίας για την απόκτηση πιθανοτικών προβλέψεων υδρολογικών μεταβλητών (βλ. π.χ., τα Σχήματα 14–16). Η παραλλαγή με την καλύτερη επίδοση προσφέρει μια μέση σχετική βελτίωση έως και 5.46% σε σχέση με τις εναλλακτικές παραλλαγές, όταν υλοποιείται με τη χρήση του μοντέλου παλινδρόμησης ποσοστημορίου. Η συγκεκριμένη παραλλαγή εκπαιδεύει το μοντέλο παλινδρόμησης σε ένα μεγάλο σύνολο δεδομένων. Το τελευταίο χρησιμοποιεί πληροφορία από το σύνολο των αδελφών προβλέψεων. Οι μέσες σχετικές βελτιώσεις όταν χρησιμοποιείται το μοντέλο παλινδρόμησης ποσοστημορίου (quantile regression) έναντι του γραμμικού μοντέλου παλινδρόμησης, φτάνουν μέχρι περίπου 37% ως προς τη μέση τιμή του μέτρου διαστήματος πρόβλεψης. Το τελευταίο αριθμητικό αποτέλεσμα θα πρέπει να αξιολογηθεί με βάση το γεγονός ότι μόνο το πρώτο από αυτά τα μοντέλα μπορεί να μοντελοποιήσει την ετεροσκεδαστικότητα. Η παραδοχή της ομοσκεδαστικότητας γίνεται συχνά στη βιβλιογραφία κατά την μοντελοποίηση του σφάλματος πρόβλεψης του υδρολογικού μοντέλου.



Σχήμα 14. Ενδεικτικά διαστήματα πρόβλεψης, παρηγμένα στο Κεφάλαιο 8 κάνοντας χρήση της υπό διερεύνηση μεθοδολογίας, για τέσσερις τυχαίες λεκάνες απορροής και μία κοινή χρονική υποπερίοδο της περιόδου δοκιμών (έτη 1996–1999). Τα μαύρα σημεία υποδηλώνουν τις πραγματικές τιμές της απορροής, ενώ οι ανοιχτές και σκούρες πορτοκαλί περιοχές υποδηλώνουν τις διαστήματα πρόβλεψης 95% και 80%, αντίστοιχα.



Σχήμα 15. Μέσες τιμές κατάταξης των σχημάτων πιθανοτικής πρόβλεψης του Κεφαλαίου 8 σύμφωνα με την μέση τιμή του μέτρου διαστήματος πρόβλεψης για την χρονική περιόδο δοκιμών (έτη 1975–1999). Οι μέσες τιμές κατάταξης έχουν υπολογιστεί για τα διαστήματα πρόβλεψης 99%, 97.5%, 95%, 90% και 80% (από πάνω προς τα κάτω). Τα σχήματα πιθανοτικής πρόβλεψης κατατάσσονται από το 1ο (καλύτερο) στο 8ο (χειρότερο). Κάθε ράβδος συνοψίζει 270 τιμές.



Σχήμα 16. Ενδεικτικές σχετικές διαφορές υπολογισμένες για την απόδειξη της ιδιότητας της υπό διερεύνηση μεθοδολογίας να “αξιοποιεί τη σοφία του πλήθους”. Οι συγκεκριμένες σχετικές διαφορές έχουν υπολογιστεί στο Κεφάλαιο 8 για το σύνολο των λεκανών απορροής που διερευνώνται στο πλαίσιο του, και για τα διαστήματα πρόβλεψης (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80%. Τα συγκεκριμένα διαστήματα πρόβλεψης έχουν παραχθεί για την χρονική περιόδο δοκιμών (έτη 1975–1999). Ο οριζόντιος άξονας έχει συντμηθεί στην τιμή 5%. Κάθε ιστόγραμμα συνοψίζει 270 τιμές.

Τέλος, αποδεικνύουμε την μεγαλύτερη ευρωστία της υπό διερεύνηση μεθοδολογίας σε σχέση με τις επιμέρους προβλέψεις που συνδυάζονται από αυτήν και, κατ 'επέκταση, σε σχέση με βασικές μεθοδολογίες πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής

μοντελοποίησης δύο σταδίων. Η ικανότητα της μεθοδολογίας να αξιοποιεί τη σοφία του πλήθους αποδεκνύεται εμπειρικά (βλ. π.χ., Σχήμα 16). Διαπιστώνεται ότι οι προβλέψεις ποσοστημορίων καθενός εκ των έξι σχημάτων πιθανοτικής προγνωστικής μοντελοποίησης είναι τουλάχιστον το ίδιο καλές –συνήθως καλύτερες– ως προς την μέση τιμή του μέτρου διαστήματος πρόβλεψης από τον μέσο όρο των τιμών που λαμβάνει το ίδιο μέτρο για καθεμία από τις επιμέρους προβλέψεις που συνδυάζονται από το εκάστοτε σχήμα. Οι αντίστοιχες μέσες σχετικές διαφορές ευνοούν την πρώτη ποσότητα έναντι της δεύτερης έως περίπου 37%, ενώ οι μέσες τιμές τους κυμαίνονται μεταξύ 0.19% και 1.83%. Εξαρτώνται τόσο από το διάστημα πρόβλεψης όσο και από την παραλλαγή της υπό αξιολόγηση μεθοδολογίας. Για το σχήμα με τις καλύτερες επιδόσεις, οι αντίστοιχες μέσες σχετικές διαφορές είναι περίπου 1%. Εν κατακλείδι, η ευρωστία και η ικανότητα αξιοποίησης της σοφίας του πλήθους αναγνωρίζονται ως δύο βασικές ιδιότητες της υπό διερεύνηση μεθοδολογίας.

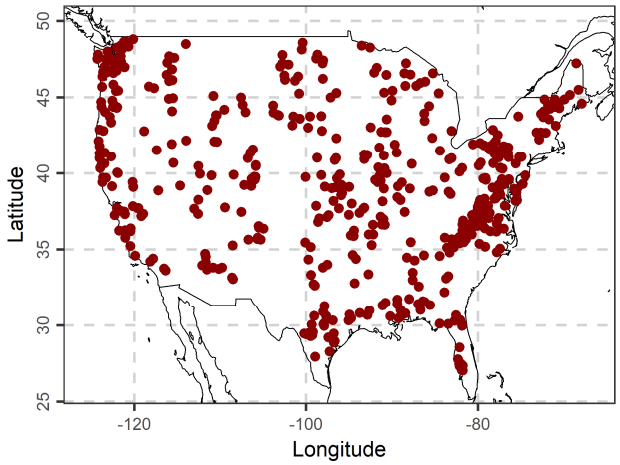
Γιατί και πώς να συνδυάσει κανείς διεργασιακά μοντέλα και αλγορίθμους μηχανικής μάθησης

Το Κεφάλαιο 9 έχει ως γενικούς του στόχους: (α) την προώθηση της χρήσης αλγορίθμων μηχανικής μάθησης στους τομείς της πιθανοτικής υδρολογικής μοντελοποίησης και της υδρο-μετεωρολογικής πρόβλεψης, (β) τη διάδοση στον χώρο της υδρολογίας της ιδέας ότι οι μέθοδοι μηχανικής μάθησης μπορούν να χρησιμοποιηθούν για πιθανοτικές προβλέψεις, (γ) την προώθηση μεθοδολογιών που βασίζονται στο συνδυασμό μοντέλων μηχανικής μάθησης, και (δ) την προώθηση της χρήσης μεγάλων συνόλων δεδομένων και πρότυπης συγκριτικής αξιολόγησης όταν χρησιμοποιούνται μέθοδοι μηχανικής μάθησης στην υδρολογία. Το Κεφάλαιο εισάγει τον μεγαλύτερο αριθμό πιθανοτικών μεθόδων υδρολογικής μοντελοποίησης που έχουν μέχρι στιγμής εισαχθεί σε μια εργασία (βασιζόμενων σε ένα ευέλικτο μεθοδολογικό σχήμα) και επιπρόσθετα διεξάγει το μεγαλύτερο πείραμα συγκριτικής αξιολόγησης που έχει διεξαχθεί μέχρι στιγμής σχετικά με τη χρήση αλγορίθμων παλινδρόμησης ποσοστιμορίου για την πιθανοτική μετεπεξεργασία αποτελεσμάτων υδρολογικής μοντελοποίησης δύο σταδίων. Επικεντρωνόμαστε στην ακόλουθη ερευνητική ερώτηση: *Γιατί και πώς να συνδυάσει κανείς διεργασιακά μοντέλα και αλγορίθμους μηχανικής μάθησης για πιθανοτική υδρολογική μοντελοποίηση;* Ως εκ τούτου, η συμβολή του Κεφαλαίου περιλαμβάνει την επιθεώρηση και την αξιολόγηση τόσο ποσοτικών όσο και ποιοτικών πτυχών σχετικών με την χρήση των αλγορίθμων.

Συζητάμε μερικά βασικά οφέλη που προκύπτουν από τον συνδυασμό διεργασιακών μοντέλων και μοντέλων μηχανικής μάθησης, όπως αυτά γίνονται αντιληπτά από την οπτική της μείωσης της αβεβαιότητας. Συζητάμε επίσης ορισμένα πρακτικά πλεονεκτήματα που απορρέουν από τον συγκεκριμένο συνδυασμό. Εν ολίγοις, με την ενσωμάτωση διεργασιακών υδρολογικών μοντέλων σε μεθοδολογίες πιθανοτικής μετεπεξεργασίας αποτελεσμάτων υδρολογικής μοντελοποίησης δύο σταδίων, επωφελούμαστε από την εμπειρία που βρίσκεται ενσωματωμένη στα διεργασιακά υδρολογικά μοντέλα (και, ως εκ τούτου, η αβεβαιότητα μειώνεται σε κάποιο βαθμό), και ταυτόχρονα ποσοτικοποιούμε την προγνωστική υδρολογική αβεβαιότητα. Επιπλέον, οι αλγόριθμοι παλινδρόμησης ποσοστημορίου μπορούν να χρησιμεύσουν αποτελεσματικά ως στατιστικά μοντέλα μετεπεξεργασίας, δεδομένου ότι μοντελοποιούν από κατασκευής την ετεροσκεδαστικότητα, συμβάλλοντας έτσι περαιτέρω στον στόχο μείωσης της αβεβαιότητας. Είναι ακόμη απλά στην εφαρμογή, πλήρως αυτοματοποιημένα (δηλαδή η υλοποίηση τους δεν απαιτεί καμία ανθρώπινη παρέμβαση), είναι διαθέσιμα σε ανοιχτό λογισμικό, υπολογιστικά βολικά και γρήγορα. Έτσι, είναι ιδιαιτέρως κατάλληλα για υδρολογικές μελέτες μεγάλου δείγματος.

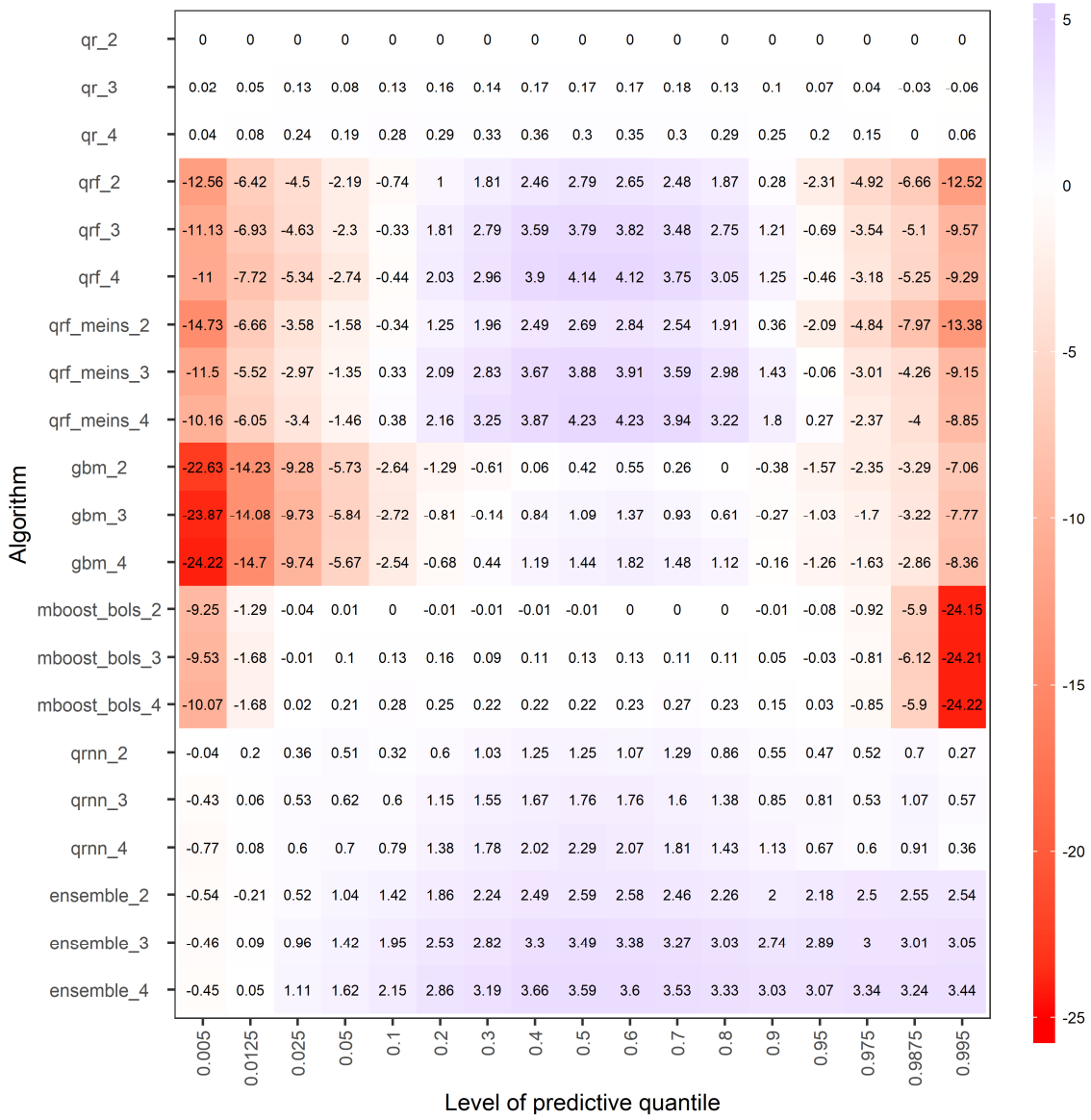
Το διενεργούμενο πείραμα μεγάλης κλίμακας χρησιμοποιεί ημερήσιες χρονοσειρές βροχόπτωσης, θερμοκρασίας, εξατμοδιαπνοής και απορροής ποταμών με μήκος 34 ετών από 511 λεκάνες απορροής στις Ηνωμένες Πολιτείες (βλ. τις γεωγραφικές θέσεις των σταθμών μέτρησης απορροής ποταμών στο Σχήμα 17). Οι σημειακές υδρολογικές προβλέψεις παράγονται χρησιμοποιώντας το διεργασιακό υδρολογικό μοντέλο GR4J και αξιοποιούνται ως μεταβλητές πρόβλεψης κατά την επίλυση των προβλημάτων παλινδρόμησης. Έξι αλγόριθμοι παλινδρόμησης ποσοστημορίου και ένας απλός συνδυασμός αυτών των αλγορίθμων χρησιμοποιούνται για την πρόβλεψη ποσοστημορίων των σφαλμάτων πρόβλεψης του υδρολογικού μοντέλου. Οι έξι

επιλεγμένοι αλγόριθμοι είναι το μοντέλο παλινδρόμησης ποσοστημορίου (quantile regression), το μοντέλο γενικευμένα τυχαία δάση (generalized random forests) για παλινδρόμηση ποσοστημορίου, το μοντέλο γενικευμένα τυχαία δάση (generalized random forests) για παλινδρόμηση ποσοστημορίου μιμούμενο το μοντέλο δάση παλινδρόμησης ποσοστιμορίου (quantile regression forests), το μοντέλο gradient boosting machine, το μοντέλο model-based boosting με γραμμικά μοντέλα βάσης και το μοντέλο νευρωνικά δίκτυα παλινδρόμησης ποσοστημορίου (quantile regression neural networks). Οι προβλέψεις ποσοστημορίων για τα σφάλματα πρόβλεψης του υδρολογικού μοντέλου μετατρέπονται σε προβλέψεις ποσοστημορίων για την ημερήσια απορροή ποταμού. Οι τελευταίες προβλέψεις αξιολογούνται χρησιμοποιώντας κατάλληλα μέτρα επίδοσης και τεχνικές πρότυπης συγκριτικής αξιολόγησης. Η αξιολόγηση αφορά προβλέψεις ποσοστημορίων με διάφορες πιθανότητες, ενώ γίνεται ανεξάρτητα από το μέγεθος της απορροής και σε συνάρτηση με αυτό.



Σχήμα 17. Γεωγραφικές θέσεις 511 σταθμών απορροής ποταμών. Ημερήσια δεδομένα από τους συγκεκριμένους σταθμούς χρησιμοποιούνται για τα πειράματα του Κεφαλαίου 9.

Τα αποτελέσματα (βλ. π.χ., Σχήμα 18) μπορεί να φανούν χρήσιμα σε τεχνικές εφαρμογές. Εν συντομία, οι αλγόριθμοι πρέπει να χρησιμοποιούνται κατά τρόπο που θα μεγιστοποιεί τα οφέλη και θα μειώνει τους κινδύνους από τη χρήση τους. Αυτό μπορεί να επιτευχθεί μέσω του συνδυασμού αλγορίθμων (π.χ., μέσω της αξιοποίησης της μεθοδολογίας των Κεφαλαίων 7 και 8) και μέσω της ενσωμάτωσης αλγορίθμων εντός συστηματικών πλαισίων (π.χ. μέσω της χρήσης διαφορετικών αλγορίθμων για προβλέψεις ποσοστημορίων με διάφορες πιθανότητες ή με την επιλογή αλγορίθμων σύμφωνα με την ικανότητά τους στην πρόβλεψη χαμηλών, μέσων ή μεγάλων ροών, ξεχωριστά για τα διάφορα ποσοστημόρια). Εάν ενδιαφερόμαστε πρωτίστως να δώσουμε αποτελέσματα γρήγορα, τότε πιθανότατα θα πρέπει να επιλέξουμε το μοντέλο παλινδρόμησης ποσοστημορίου (quantile regression). Αυτή η επιλογή θα πρέπει να γίνει έχοντας κατά νου ότι το μοντέλο αυτό είναι έως και 3.5% χειρότερο ως προς την μέση τιμή του μέτρου ποιότητας ποσοστημορίου από ότι ο απλός συνδυασμός των έξι αλγορίθμων του Κεφαλαίου. Δείχνουμε ότι ο συγκεκριμένος απλός συνδυασμός έχει την καλύτερη επιδόση συνολικά, επιβεβαιώνοντας την αξία της μάθησης του συνόλου γενικά και της μάθησης του συνόλου μέσω της απλής μέτρησης του μέσου όρου. Η αξία αυτή είναι ευρέως αναγνωρισμένη στο πεδίο της πρόβλεψης χρονοσειρών, αλλά δεν έχει λάβει ακόμη την απαραίτητη προσοχή τόσο στην βιβλιογραφία της υδρολογικής μοντελοποίησης όσο και στην βιβλιογραφία της υδρο-μετεωρολογικής πρόβλεψης. Παρά την εξαιρετική του επίδοση, ο απλός συνδυασμός των έξι αλγορίθμων αυτού του Κεφαλαίου αναμένεται, με τη σειρά του, να έχει χειρότερη επίδοση από ορισμένους από τους μεμονωμένους αλγορίθμους σε πολλές περιπτώσεις μοντελοποίησης. Γενικά, κανένας αλγόριθμος δεν αναμένεται να είναι (ούτε πρέπει να παρουσιάζεται ως) ο καλύτερος ως προς όλα τα κριτήρια.



Σχήμα 18. Διάμεσοι των σχετικών βελτιώσεων που προσφέρει καθένα από τα υπό διερεύνηση σχήματα πιθανοτικής πρόβλεψης (%) σε σχέση με το σχήμα πιθανοτικής πρόβλεψης qr_2 σε όρους μέσης τιμής του μέτρου ποιότητας ποσοστημορίου.

Contents

List of Figures

List of Tables

1. Introduction

1.1 Motivation, main objectives and principles

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk", ~ John von Neumann. In fact, model fitting can be made quite satisfactorily when using an adequate number of model parameters (see e.g., the cartoon-like implementation by Mayer et al. 2010 of the above quote using five complex parameters, adapted in Figure 1.1), and can also be very interesting, inspiring and creative. In particular, by computing and analysing descriptive features of real-world processes, one may achieve significant advancements in terms of real-world process understanding, and may also be able to facilitate comparisons within and across real-world processes, thereby strengthening this understanding further. Probably due to this indisputable value, a significant part of the hydrological literature is devoted to assessing the descriptive power of various methodologies (and to increasing this power by adding parameters, trend or other type, to them) in the context of geophysical time series analysis (Papacharalampous et al. 2018b). Moreover and since models are usually required to also have practical implications along with their mathematical and theoretical value, many of the conducted works extend their conclusions by claiming that their models are also appropriate for predictive modelling (distinguished from descriptive modelling in Shmueli 2010), without however having tested their predictive ability. Although the descriptive and predictive perspectives may indeed be connected (to a larger or smaller extent) with each other (see e.g., the investigations for North America and Europe on the relationships between selected predictive and descriptive annual river flow features in Papacharalampous and Tyralis 2020), this behaviour is rather "cheating" and often appears due to the ignorance of one simple rule, which nonetheless constitutes the *"most powerful idea in data science"* (towardsdatascience.com/the-most-powerful-idea-in-data-science-78b9cd451e72): The same data point cannot be used both (i) for forming an opinion (e.g., on how useful a model is) and (ii) for generalizing this opinion, i.e., generalizations cannot be supported by a single dataset (formed by data points). By considering this simple rule, we understand that all possibly expected connections between descriptive power and predictive power should first be proven valid, before trusted in practice. This is what actually makes prediction *"very difficult, especially about the future"*, while one can casually fit elephants or cats to data.

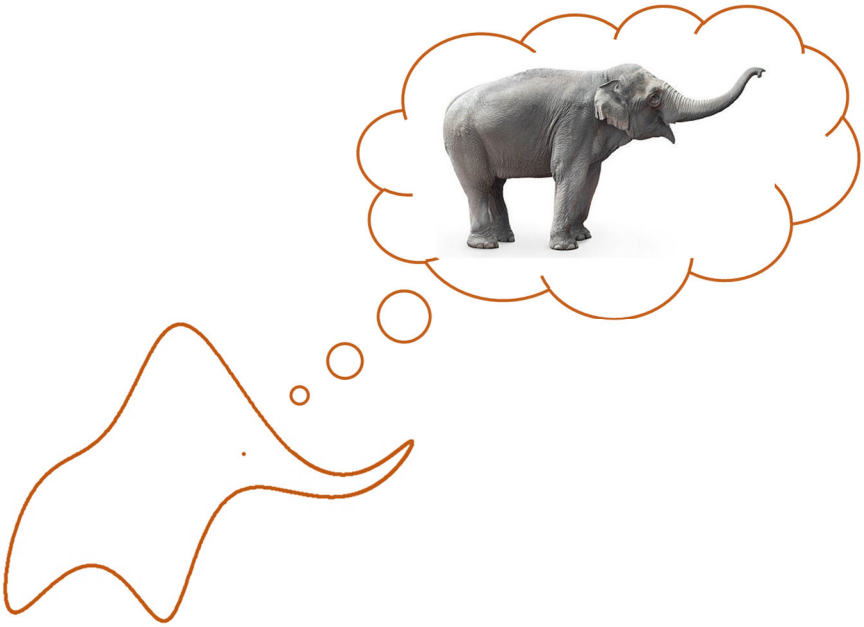


Figure 1.1. *"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk"*, ~ John von Neumann. A relevant discussion and a Python code for the implementation of the method of Mayer et al. (2010), i.e., for drawing the orange elephant on the left, can be found at: johndcook.com/blog/2011/06/21/how-to-fit-an-elephant.

Moving a step further from model fitting, this thesis is devoted to the challenging task of predictive modelling of hydrological processes. The thesis falls into the scientific areas of stochastic hydrology, hydrological modelling and hydroinformatics, and aspires to contribute with new practical solutions, new methodologies and large-scale results to the following two interrelated technical predictive modelling problems, with emphasis on the latter one:

(A) point forecasting of hydrological processes by exclusively considering endogenous predictor variables within purely statistical frameworks (hereafter referred to simply as "hydrological time series forecasting", unless specified differently); and

(B) stochastic process-based modelling for hydrological systems via probabilistic post-processing (hereafter referred to simply as "probabilistic hydrological post-processing", unless specified differently)

Within the context of the thesis, hydrological time series forecasting is performed by using either stochastic forecasting models or machine learning regression algorithms, while probabilistic hydrological post-processing is performed by using conceptual process-based hydrological models in combination with (machine learning) quantile regression algorithms. There exists a widespread misconception in the minds of hydrologists that machine learning algorithms are by nature deterministic. Nonetheless, machine learning methods are all statistical, while the quantile regression ones are also ideal for predictive uncertainty quantification. It is also relevant to highlight that hydrological time series forecasts obtained by exclusively using endogenous predictor variables (see e.g., Koutsoyiannis et al. 2008) are, in general, accurate enough when delivered at large time scales (i.e., the annual, seasonal and monthly ones); therefore, problem (A) is herein solved at such time scales. On the contrary, at fine time scales (i.e., the daily and sub-daily ones) exogenous predictor variables (e.g., observed or forecasted values of various hydrometeorological variables) can be very informative and, thus, their consideration (e.g., by utilizing autoregressive moving average models with exogenous predictor variables − ARMAX, autoregressive fractionally integrated moving average models with exogenous predictor variables − ARFIMAX or machine learning algorithms) can result in large improvements in forecasting performance (see e.g., the investigations by Papacharalampous and Tyralis 2018, and Tyralis et al. 2020b). In contrast to problem (A), problem (B) involves the consideration of exogenous predictor variables by definition, and is herein solved both at the monthly time scale and at the daily time scale.

Importantly, all the models and algorithms exploited in the thesis are flexible, computationally convenient and fast; thus, they are appropriate for large-sample (even global-scale) hydrological investigations. Conducting such investigations and large-scale simulation tests has been of major priority herein, together with the introduction of new methodologies and new practical solutions. This priority is implied by the fact that analytical investigations of many of its methods (especially, the most flexible machine learning ones) can be highly demanding (to nearly impossible). Therefore, within the context of the thesis generalizations are empirically achieved by using large datasets under the data splitting approach, i.e., as implied by the *"most powerful idea in data science"* (see above). Millions of out-of-sample predictions are used for generalizing our opinion about the predictive performance of numerous modelling approaches. Finally and in spite of its main orientation, this thesis also provides innovative theoretical supplements and justifications to many of its algorithmically obtained outcomes.

1.2 Original research works and roadmap

The remainder of this thesis is structured as follows: Chapter 2 presents the theoretical, methodological and technical background of the thesis. It also presents its predictive modelling and benchmarking toolbox, formed and exploited for achieving its aims. The latter are explicitly stated in the introductory sections of Chapters 3–9. These seven Chapters present original research works in the areas of hydrological time series forecasting (Chapters 3–6) and probabilistic hydrological post-processing (Chapters 7–9). Finally, Chapter 10 summarizes the content and main contributions of the thesis by emphasizing its innovative points.

Seven original research works have been conducted within the context of the thesis, each being the basis of a different Chapter from Chapters 3–9, as detailed in the following:

- Chapter 3 has been based on the work conducted under the title "Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes" (Papacharalampous et al. 2019a). The original work is reproduced with adaptations. The Chapter is fully reproducible; all codes and data, as well as their outcome results, are available in Papacharalampous and Tyralis (2018b). Closely related preliminary investigations can be found in Papacharalampous et al. (2017a).
- Chapter 4 has been based on the work conducted under the title "One-step ahead forecasting of geophysical processes within a purely statistical framework" (Papacharalampous et al. 2018d). The original work is reproduced with adaptations. The Chapter is fully reproducible; all codes and data, as well as their outcome results, are available in Papacharalampous and Tyralis (2018b). Closely related preliminary investigations can be found in Papacharalampous et al. (2017c).
- Chapter 5 has been based on the work conducted under the title "Predictability of monthly temperature and precipitation using automatic time series forecasting methods" (Papacharalampous et al. 2018e). The original work is reproduced with adaptations. Closely related preliminary investigations can be found in Papacharalampous et al. (2018b).
- Chapter 6 has been based on the work conducted under the title "Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece" (Papacharalampous et al. 2018f). The original work is reproduced with adaptations. Closely related preliminary investigations can be found in Papacharalampous et al. (2017b).
- Chapter 7 has been based on the work conducted under the title "Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models" (Papacharalampous et al. 2020a). The original work is reproduced with adaptations. Closely related preliminary investigations can be found in Papacharalampous et al. (2018a).
- Chapter 8 has been based on the work conducted under the title "Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale" (Papacharalampous et al. 2020b). The Chapter is supplemented by Papacharalampous et al. (2019d). The original work is reproduced with adaptations.
- Chapter 9 has been based on the work conducted under the title "Probabilistic hydrological post-processing at scale: Why and how to apply machine learning quantile regression algorithms" (Papacharalampous et al. 2019d). The Chapter is supplemented by Papacharalampous et al. (2019e). The original work is reproduced with adaptations. Closely related preliminary investigations can be found in Papacharalampous et al. (2019b).

Segments of the above-listed works have been compiled for writing Chapter 2, and have, therefore, been omitted from Chapters 3–9. The latter Chapters have been formulated according to the guidelines for improved research in the field of practical hydroinformatics by Abrahart et al. (2008). These guidelines emphasize more on reproducibility and less on exhaustive descriptions of models and algorithms, unless these models and algorithms are entirely new. The adopted writing strategy offers an important benefit: The thesis can be read either by following the "bottom-up" approach or by following the "top-down" approach. In the former case, the reading of Chapter 2 should precede the reading of Chapters 3–9. Although the "bottom-up" approach is often preferred, from a practitioner's point of view the content of Chapter 2 is only auxiliary for solving the technical problems of Chapters 3–9. Under this latter view, Chapters 3–9 can be read directly (and in any order) by drilling down to their theoretical and methodological foundations, when necessary, and by inspecting their framework inputs through Chapter 2. Chapter 10 can be read independently of the remaining Chapters.

2. Theoretical, methodological and technical background and toolbox

This thesis applies a variety of models and algorithmic procedures to a variety of datasets and modelling contexts. In this Chapter, we present its theoretical, methodological and technical background by reviewing the literature, when necessary. The interested reader is also referred to several specialized and detailed books, textbooks, technical works and journal articles for the complete documentation of the existing algorithms, models and methodologies exploited in the context of this thesis. To ease the reading of Chapters 3–9, we also provide an overview of the basic methodological elements combined in the thesis. In what follows, random variables are underscored, following the Dutch convention.

2.1 Stochastic time series modelling

For this thesis, we exploit several stochastic models (also referred to as "time series models") and related procedures. The exploitation is made, either directly or indirectly (i.e., through wider modelling approaches), and concerns time series simulation, time series processing (e.g., standardizations, decompositions, gap-filling), time series characterization and time series forecasting, as summarized in Section 2.9. In this Section, we briefly present the mathematical background of the exploited stochastic models.

2.1.1 Basic definitions and concepts

In this Section, we provide some basic definitions and concepts underlying time series modelling (see also Wei 2006, pp. 6–16). A time series in discrete time is defined as a sequence of observations x_1, x_2, … of a certain phenomenon, while the time t is stated as a subscript to each value x_t. A time series can be modelled by a stochastic process. The latter is a family of random variables $\underline{x}_1$, $\underline{x}_2$, …. A random variable is a function that maps events from the sample space to the real numbers.

Let us consider a stochastic process of normally distributed random variables. The mean function (μ_t) of the stochastic process is defined with the following Equation:

$$\mu_t := \mathrm{E}[\underline{x}_t] \tag{2.1}$$

The standard deviation function (σ_t) of the stochastic process is defined with the following Equation:

$$\sigma_t := \sqrt{\mathrm{Var}[\underline{x}_t]} \tag{2.2}$$

The covariance function between $\underline{x}_{t_1}$ and $\underline{x}_{t_2}$ of the stochastic process, denoted with $\gamma(t_1,t_2)$, is defined with the following Equation:

$$\gamma(t_1, t_2) := \mathrm{E}[(\underline{x}_{t_1} - \mu_{t_1})(\underline{x}_{t_2} - \mu_{t_2})] \tag{2.3}$$

The correlation function between $\underline{x}_{t_1}$ and $\underline{x}_{t_2}$ of the stochastic process, denoted with $\rho(t_1,t_2)$, is defined with the following Equation:

$$\rho(t_1, t_2) := \gamma(t_1,t_2)/(\sigma_{t_1}\, \sigma_{t_2}) \tag{2.4}$$

For a strictly stationary stochastic process, Equations (2.5)–(2.8) must be satisfied:

$$\mu_t = \mu \;\forall\; t \in \{1, 2, \ldots\} \tag{2.5}$$

$$\sigma_t = \sigma \;\forall\; t \in \{1, 2, \ldots\} \tag{2.6}$$

$$\gamma(t_1, t_2) = \gamma(t_1 + k, t_2 + k) \;\forall\; t_1, t_2, k \text{ integers} \tag{2.7}$$

$$\rho(t_1, t_2) = \rho(t_1 + k, t_2 + k) \;\forall\; t_1, t_2, k \text{ integers} \tag{2.8}$$

In this case, let us consider that:

$$t_1 := t - k,\; t_2 := t \tag{2.9}$$

Then we have:

$$\gamma(t_1, t_2) = \gamma(t - k, t) = \gamma(t, t + k) = \gamma_k \quad (2.10)$$

$$\rho(t_1, t_2) = \rho(t - k, t) = \rho(t, t + k) = \rho_k \quad (2.11)$$

Using the Equations (2.5)–(2.11), the autocovariance function (γ_k) and the autocorrelation function (ρ_k) of a stationary stochastic process are defined with Equations (2.12) and (2.13), respectively.

$$\gamma_k := \mathrm{E}[(\underline{x}_t - \mu)(\underline{x}_{t+k} - \mu)] \quad (2.12)$$

$$\rho_k := \gamma_k / \sigma^2 \quad (2.13)$$

For a stationary stochastic process, the partial autocorrelation function P_k is defined by

$$P_k := \mathrm{Corr}[(\underline{x}_t, \underline{x}_{t+k} \mid \underline{x}_{t+1}, \ldots, \underline{x}_{t+k-1})] \quad (2.14)$$

The partial autocorrelation function is the correlation between two random variables $\underline{x}_t$ and $\underline{x}_{t+k}$, with the linear dependency between the intervening variables $\underline{x}_{t+1}$, …, $\underline{x}_{t+k-1}$ removed.

A strictly stationary stochastic process $\{\underline{a}_t\}$ is called a white noise process, if it is a sequence of uncorrelated random variables. Let us consider, hereinafter, that the white noise is a normal variable with zero mean, unless mentioned otherwise, and standard deviation σ_a.

2.1.2 Sample autocorrelation and partial autocorrelation functions

For a given time series in discrete time $x_1, x_2, \ldots$, the sample autocovariance function (denoted with $\hat{\gamma}_k$) and the sample autocorrelation function (denoted with $\hat{\rho}_k$) can be computed through Equations (2.15) and (2.16), respectively (Wei 2006, pp. 18–23).

$$\hat{\gamma}_k := (1/n) \sum_{i=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) \quad (2.15)$$

$$\hat{\rho}_k := \hat{\gamma}_k / \hat{\gamma}_0 := \sum_{i=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}) / \sum_{i=1}^{n} (x_t - \bar{x})^2 \quad (2.16)$$

In these Equations, $\bar{x}$ denotes the sample mean of the time series, defined with the following Equation:

$$\bar{x} := (1/n) \sum_{i=1}^{n} x_i \quad (2.17)$$

A recursive method for calculating the sample partial autocorrelation function (denoted with $\hat{\varphi}_{k+1,k+1}$) is given by Equations (2.18) and (2.19).

$$\hat{\varphi}_{k+1,k+1} := (\hat{\rho}_{k+1} - \sum_{j=1}^{k} \hat{\varphi}_{kj} \hat{\rho}_{k+1-j}) / (1 - \sum_{j=1}^{k} \hat{\varphi}_{kj} \hat{\rho}_j) \quad (2.18)$$

$$\hat{\varphi}_{k+1,j} := \hat{\varphi}_{kj} - \hat{\varphi}_{k+1,k+1} \hat{\varphi}_{k,k+1-j}, \; j = 1, \ldots, k \quad (2.19)$$

2.1.3 Autoregressive moving average processes

In this Section, we provide some basic definitions and concepts underlying time series modelling (see also Wei 2006, pp. 23–87). The stochastic process $\{\underline{y}_t\}$ is defined with the following Equation:

$$\underline{y}_t := \underline{x}_t - \mu \quad (2.20)$$

Let us consider the operator B, which is defined with the following Equation:

$$\mathrm{B}^j \underline{x}_t := \underline{x}_{t-j} \quad (2.21)$$

Then the operator $\varphi_p(\mathrm{B})$ is defined with the following Equation:

$$\varphi_p(\mathrm{B}) := (1 - \varphi_1 \mathrm{B} - \ldots - \varphi_p \mathrm{B}^p) \quad (2.22)$$

The stochastic process $\{\underline{x}_t\}$ is an AR(p), if the following equation holds:

$$\varphi_p(\mathrm{B}) \underline{y}_t = \underline{a}_t \quad (2.23)$$

that can be written in the following form:

$$\underline{y}_t = \varphi_1 \underline{y}_{t-1} + \ldots + \varphi_p \underline{y}_{t-p} + \underline{a}_t \quad (2.24)$$

Let us also consider the operator $\theta_q(\mathrm{B})$, which is defined with the following Equation:

$$\theta_q(\mathrm{B}) := 1 + \theta_1 \mathrm{B} + \ldots + \theta_q \mathrm{B}^q \quad (2.25)$$

The stochastic process $\{\underline{x}_t\}$ is a MA(q), if the following equation holds:

$$\underline{y}_t = \theta_q(\mathrm{B})\underline{a}_t \quad (2.26)$$

that can be written in the following form:

$$\underline{y}_t = \underline{a}_t + \theta_1 \underline{a}_{t-1} + \ldots + \theta_q \underline{a}_{t-q} \quad (2.27)$$

The stochastic process $\{\underline{x}_t\}$ is an ARMA(p,q), if the following equation holds:

$$\varphi_p(\mathrm{B})\underline{y}_t = \theta_q(\mathrm{B})\underline{a}_t \quad (2.28)$$

that can be written in the following form:

$$\underline{y}_t = \varphi_1 \underline{y}_{t-1} + \ldots + \varphi_p \underline{y}_{t-p} + \underline{a}_t + \theta_1 \underline{a}_{t-1} + \ldots + \theta_q \underline{a}_{t-q} \quad (2.29)$$

2.1.4 Autoregressive integrated moving average processes

Let d be a natural number. Then the stochastic process $\{\underline{x}_t\}$ is an ARIMA(p,d,q), if the following equation holds:

$$\varphi_p(\mathrm{B})(1-\mathrm{B})^d \underline{x}_t = \theta_0 + \theta_q(\mathrm{B})\underline{a}_t \quad (2.30)$$

If $d = 0$, then we have an ARMA(p,q) and for θ_0 we obtain:

$$\theta_0 = (1 - \varphi_1 - \ldots - \varphi_p)\mu \quad (2.31)$$

If $d \geq 1$, then θ_0 is called deterministic trend term and is usually omitted from the model, unless it is truly required. This specific stochastic process is non-stationary (Wei 2006, p. 69).

2.1.5 Autoregressive fractionally integrated moving average processes

Let $d \in (-0.5, 0.5)$. The stochastic process $\{\underline{x}_t\}$ is an ARFIMA(p,d,q), if the following equation holds:

$$\varphi_p(\mathrm{B})(1-\mathrm{B})^d \underline{x}_t = \theta_q(\mathrm{B})\underline{a}_t \quad (2.32)$$

In contrast to ARIMA(p,d,q), ARFIMA(p,d,q) is stationary (Wei 2006, p. 489). This specific stochastic process is widely applied in hydrology (see e.g., Montanari et al. 1997, 1999, 2000). In general, it can be used to model processes that are characterized with long-range dependence, with its parameter d being indicative of the magnitude of this dependence and, therefore, fitted to serve as its measure. The long-range dependence is an inherent property of some geophysical processes (see, for example, Tyralis and Koutsoyiannis 2011 and the references therein).

2.1.6 Fractional Gaussian noise process

Let $\{\underline{x}_t\}$, $t = 1, 2, \ldots$ be a fractional Gaussian noise process, a stationary stochastic process of normally distributed random variables in discrete time. Then, its parameters μ, σ, H are defined with Equations (2.1), (2.2) and (2.33), respectively (Tyralis and Koutsoyiannis 2011).

$$\rho_k := \mathrm{Corr}[\underline{x}_t, \mathrm{x}_{t+k}] = |k+1|^{2H} / 2 + |k-1|^{2H} / 2 - |k|^{2H},\ k = 0, 1, \ldots,\ H \in (0, 1) \quad (2.33)$$

The parameters μ and σ are the mean and the standard deviation of the stochastic process, respectively, while the parameter $H \in (0, 1)$, known as its "Hurst parameter", is assumed to be informative about the magnitude of long-range dependence in geophysical time series. The long-range dependence is strong when H is high, while $H = 0.5$ corresponds to uncorrelated random variables.

We fit this stochastic process to non-seasonal or seasonally decomposed time series by using the maximum likelihood method (Tyralis and Koutsoyiannis 2011). Furthermore, we estimate the coefficient of variation of the fractional Gaussian noise process according to the following Equation:

$$\text{cv} := \sigma/\mu \tag{2.34}$$

We also perform non-seasonal time series standardization by using the Equation:

$$z_t := (y_t - \mu)/\sigma \tag{2.35}$$

In Equation (2.35), y_t and z_t denote the original and standardized data, respectively, at time t.

2.1.7 Time series decomposition

We perform classical time series decomposition by applying the additive and multiplicative models. These models are defined with Equations (2.36) and (2.37), respectively (Hyndman and Athanasopoulos 2018, Chapter 6.1).

$$y_t = S_t + T_t + R_t \tag{2.36}$$

$$y_t = S_t\, T_t\, R_t \tag{2.37}$$

In these Equations, y_t denotes the data at time t, while S_t, T_t and R_t denote the seasonal, trend-cycle and remainder components, respectively, at time t. The additive model is suitable when the seasonal fluctuations do not depend on the level of the time series, while the multiplicative model is suitable for modelling seasonal fluctuations which are proportional to the level of the time series (Hyndman and Athanasopoulos 2018, Chapter 6.1).

Classical time series decomposition has its routes in the 1920s (Hyndman and Athanasopoulos 2018, Chapter 6.3). It uses moving averages (Hyndman and Athanasopoulos 2018, Chapter 6.2) in a relatively simple procedure, and has been used as a starting point for building most of the other available time series decomposition methods (Hyndman and Athanasopoulos 2018, Chapter 6.3). Such methods are the X11, the SEATS (acronym for "Seasonal Extraction in ARIMA Time Series") and STL (acronym for "Seasonal and Trend decomposition using Loess") methods (see e.g., Hyndman and Athanasopoulos 2018, Chapters 6.4–6.6).

2.1.8 Mathematical transformations

Mathematical transformations are often used to improve time series modelling. To this respect, logarithmic and power transformations can be applied, among others. A popular transformation that is also adopted herein is the Box-Cox transformation, introduced by Box and Cox (1964). This transformation considers both logarithms and power transformations (Hyndman and Athanasopoulos 2018, Chapter 3.2), and is given by the following equation for $x > 0$:

$$f_\lambda(x) = \begin{cases} (x^\lambda - 1)/\lambda \text{ if } \lambda \neq 0 \\ \ln(x) \text{ if } \lambda = 0 \end{cases} \tag{2.38}$$

In Equation (2.38), x denotes the variable to be transformed and λ is a parameter that is estimated from data. This estimation is herein made by using the method of Guerrero (1993), as implemented in Hyndman et al. (2018).

2.1.9 State space models and Kalman filtering

The state space representation of a system (also referred to as its "Markovian representation"; Wei 2006, p. 463) is described by the measurement equation (also referred to as "observation equation") and some state equations. The former equation describes the observed data, while the latter equations describe how unobserved states change over time (Hyndman and Athanasopoulos 2018, Chapter 7.5; see also Wei 2006, Chapter 18.1). The state of a system is defined as the minimum set of information from present and past required to "completely" describe the future behaviour of the same system given any present state and future input (Wei 2006, p. 463). The state space representation of a system is closely related to the Kalman filter (see Wei 2006, Chapter 18.5). This model was originally proposed by Kalman (1960) and meets a wide range of applications. In this thesis, we use it for time series gap-filling.

2.1.10 Parameter estimation, model selection and automatic methods

Several parameter estimation methods are available for time series models. Among the most popular ones are the method of moments (see e.g., Wei 2006, Chapter 7.1) and the maximum likelihood method (see e.g., Wei 2006, Chapter 7.2). Because of its nice properties, the latter method is used as part of many of the automated schemes exploited in this thesis. Ordinary least squares estimation (see e.g., Wei 2006, Chapter 7.4), originally developed for linear regression algorithmic approaches (see Section 2.3.1), can also be used in time series analysis contexts; however, it is less efficient than the afore-mentioned methods (in time series analysis contexts). For an explanation, see Wei (2006, pp. 151, 152). Automatic time series (forecasting) methods (e.g., those available in software packages) usually allow the user to select among different parameter estimation methods, which may vary in terms of computational requirements and/or efficiency.

Since there might exist numerous models that adequately fit the data, another task to be completed when building automatic time series (forecasting) models is to objectively select a single one. This is possible by applying information (or model discrimination) criteria, such as the original Akaike information criterion (AIC) by Akaike (1974), the Akaike information criterion with a correction for finite sample sizes (AICc) by Hurvich and Tsai (1993), and two Bayesian information criteria, i.e., BIC by Schwarz (1978) and KIC by Kashyap (1982). Within our modelling approaches, we minimize either AIC or AICc. These criteria are defined with Equations (2.39) and (2.40), respectively (see e.g., Hyndman and Athanasopoulos 2018, Chapter 5.5). In these Equations, L is the likelihood of the candidate model, k is the total number of parameters (and initial states) and n is the sample size.

$$\mathrm{AIC} = -2\log(L) + 2k \tag{2.39}$$

$$\mathrm{AICc} = \mathrm{AIC} + 2k(k+1)/(n-k-1) \tag{2.40}$$

We note that AICc reduces asymptotically to AIC as the sample increases (Ye et al. 2008), while for small fitting samples the minimization of AIC tends to lead to larger number of model parameters compared to the minimization of AICc (Hyndman and Athanasopoulos 2018, Chapter 5.5). Ye et al. (2008) report on a debate in hydrology on the selection between commonly used information criteria, while task-oriented comparisons of information criteria can be found, for instance, in Ye et al. (2004), Billah et al. (2005), Ye et al. (2008) and Emiliano et al. (2014).

2.2 Time series forecasting

While the available time series forecasting models are numerous, the basic ones are quite few (Hong and Fan 2016). In this thesis, we exploit fully automatic methods originating from the forecasting literature and, in many cases, combine different models to automate new ones. The primary forecasting algorithms are well documented in the literature. Therefore, in Chapters 3–6 we place emphasis on their software implementation (see also Section 2.9.4). Even the information here compiled from books, textbooks and journal articles is limited to their key concepts and their basic theoretical background. Further theoretical details, available in the provided references, are here omitted for reasons of brevity. We note that the understanding from a theoretical point of view of most methods could hardly help in interpreting the algorithmically obtained outcomes of this thesis.

2.2.1 Time series forecasting using simple models

We implement two simple forecasting methods, i.e., the naïve and random walk ones. For the non-seasonal (e.g., the annual) time series, the naïve method simply sets all forecasts equal to the last value of the training period. For the monthly time series, the forecast of the naïve method for each month of the testing period is equal to the observed value for the same month of the last year of the training period. The random walk method, a variant of the naïve forecasting method, fits a random walk model with drift to the training segment and then uses the fitted model for forecasting. This method is equivalent to drawing a line between the first and the last values, and

extrapolating it into the future (Hyndman and Athanasopoulos 2018, Chapter 3.1). Both simple methods are based on modelling the data using discrete-time martingales (see e.g., Palma 2007, Chapter 1.1.9). The difference is that, in the case of random walk, a drift is added to the model. Sometimes, simple methods perform surprisingly well; therefore, it is important to use them for benchmarking purposes.

2.2.2 Time series forecasting using ARIMA and ARFIMA models

ARIMA and ARFIMA methods are also included in the comparisons. We apply both fixed- and optimum-order ARIMA methods. For the fixed-order ARIMA methods, the numbers of the AR and MA parameters (p and q respectively) are set to be the same to those used in the time series simulation process (see Section 3.2.1), while the number of differencing (d) is set to zero. On the contrary, the optimum-order ARIMA methods automatically estimate the order of the ARIMA models (and, therefore, the utilized lagged predictor variables) as summarized in the following. First, the d values are estimated via repeated Kwiatkowski–Phillips–Schmidt–Shin tests (Kwiatkowski et al. 1992). Once the d value has been obtained, the p and q values are estimated using a stepwise algorithm aiming at the minimization of AICc (see Section 2.1.10). AICc is preferred in the herein adopted implementation by Hyndman et al. (2018), while other available options are AIC and BIC. The exact procedure adopted by the optimum-order ARIMA methods for order estimation is available in Hyndman and Khandakar (2008), and Hyndman and Athanasopoulos (2018, Chapter 8.6). As explained in the latter-mentioned textbook's chapter, the d value is not estimated simultaneously with the p and q values using AICc, because in this case the estimation would be suboptimal. Once the p, d and q values have been estimated, the all ARIMA methods apply the maximum likelihood method to estimate the AR and MA model parameters (Hyndman and Athanasopoulos 2018, Chapter 8.6).

We apply optimum-order ARFIMA methods. Similarly to the optimum-order ARIMA methods, these methods estimates d first, and thereupon follows a stepwise procedure to select p and q. Subsequently, it implements the algorithm of Haslett and Raftery (1989) to estimate the ARFIMA parameters. A final value of d is estimated as well in this last step. The latter information is sourced from Hyndman et al. (2018) and Fraley et al. (2012), where related detailed descriptions can be found. The definitions of the ARMA, ARIMA and ARFIMA models are given in Sections 2.1.3, 2.1.4 and 2.1.5, respectively (see also Wei 2006, pp. 6–87, 489–494).

2.2.3 Time series forecasting using exponential smoothing and state space models

Another family of time series (or stochastic) methods considered herein (that is also broader than the family of ARIMA models; Gardner 2006) includes the exponential smoothing models and their underlying methods, i.e., the (Innovations) state space methods (see Section 2.1.9) for exponential smoothing. Their forecasts are weighted averages of past values, with the weights decaying exponentially as these values get distant in time (Hyndman and Athanasopoulos 2018, Chapter 7). Informative reviews by Gardner (1985, 2006) discuss older and latest advances in forecasting with exponential smoothing, from the introducing works by Brown and Holt that are available in Brown (1959) and Holt (2004) respectively (the latter paper is a reprinted version of Holt's report of 1957) up to more recent studies (e.g., Assimakopoulos and Nikolopoulos 2000; Hyndman et al. 2002; Hyndman and Billah 2003). The reader is also referred to Hyndman et al. (2008), and Hyndman and Athanasopoulos (2018, Chapters 7, 8.3) for further details on the theoretical background of the exponential smoothing and state space models.

We implement the simple exponential smoothing (SES) and Theta methods. The former method is introduced by Brown (1959) and described, for example, in Hyndman et al. (2008, p. 13). It computes the forecast of the next period (f_{t+1}) based on the forecast of the previous period (f_t), the latter adjusted using its error ($x_t - f_t$), according to the following Equation (2.41). In this Equation, α is a parameter to be estimated from the data. Similarly to the naïve and average methods (the forecasts of the latter are simply the average of all training values), SES produces flat forecasts; therefore, it is considered the most simple of its class. An interpretation of the concept behind SES is provided by Hyndman and Athanasopoulos (2018, Chapter 7.1). According

to this interpretation, SES is a more general version of both the naïve and average methods. The parameters of SES are herein estimated by using procedures by Hyndman et al. (2018).

$$f_{t+1} = f_t + a(x_t - f_t),\ a \in (0, 1) \quad (2.41)$$

The Theta method by Assimakopoulos and Nikolopoulos (2000) is equivalent to SES with a drift parameter (Hyndman and Billah 2003). As shown in Hyndman and Billah (2003), the drift parameter is half the slope of the linear trend fitted to the data. There are several variants of Theta, each defined by the so-called "Theta lines", i.e., the auxiliary time series (modified versions of the original time series provided as input to the method) used for model fitting and forecasting. A Theta line is characterized by its local curvature, which is determined by the Theta coefficient θ (different for each Theta line). Extrapolations of all Theta lines are averaged to produce the forecast. We implement the version of Theta that performed well in the M3 competition (Makridakis and Hibon 2000), i.e., the one defined by two Theta lines, specifically for $\theta = 0$ and $\theta = 2$ (see Assimakopoulos and Nikolopoulos 2000).

Moreover, we implement two state space methods for exponential smoothing. Models from this category produce expected value forecasts and, additionally, provide information about the forecast error variances (Hyndman et al. 2005; see also Hyndman and Athanasopoulos 2018, Chapter 7.5). This information can be used either for constructing prediction intervals or for running an exponential smoothing model in simulation mode. The first implemented state space model for exponential smoothing is ETS. This model comprises automatic selection of the Error, Trend and Seasonal components (ETS) using the AICc (Hyndman and Athanasopoulos 2018, Chapter 7.6). The expected value forecasts of this model on the M competition and M3 competition data are found to be comparable with the best obtained in these competitions (Hyndman et al. 2002). Another state space method implemented herein is BATS. This method uses the point forecasts from an exponential smoothing state space model with several key features, i.e., capability of performing Box-Cox transformation and/or including ARMA errors correction, Trend and Seasonal components (BATS), also allowing an optimal model selection using the Akaike Information Criterion (AIC). The original model is introduced and fully documented in De Livera et al. (2011).

2.2.4 Time series forecasting using the Prophet model

The Prophet method, introduced by Taylor and Letham (2018), considers time series forecasting as a curve-fitting exercise, while it does not explicitly consider the temporal dependence of the time series. It uses the additive decomposable time series model by Harvey and Peters (1990), which is similar to the generalized additive model by Hastie and Tibshirani (1987). The Prophet method is inspired by the nature of the time series forecasted at Facebook, which are characterized by trend, multiple seasonality and holidays (an example of similar time series in the water science is the water demand). Furthermore, this method is designed to *"forecast at scale"* and to fit to the data very fast. Details on Prophet are available in Taylor and Letham (2018, Section 3).

2.2.5 Time series forecasting using (machine learning) regression algorithms

To forecast time series, we also use the machine learning (ML) regression algorithms outlined in Sections 2.3.2–2.3.4. Time series forecasting using regression algorithms is traditionally based on different strategies than those discussed so far (e.g., in Section 2.1.10) for the time series models (also known as "stochastic models"). The input to a regression model is the data matrix used in the regression process (hereafter referred to as "input data matrix"). In time series forecasting using regression algorithms, the input data matrix is built using a single time series holding the total information provided to the regression algorithm. One column of the input data matrix holds information about the predictand variable and the remaining columns information about lagged (predictor) variables that are assumed to be informative about the predictand. Variable selection (or feature selection) is known as a factor that might affect the performance of regression algorithms in both typical regression and forecasting applications (see e.g., Anctil et al. 2009;

Chapter 6 herein). Thus, many studies specifically focus on the examination of this problem (e.g., Kohavi and John 1997; Tyralis and Papacharalampous 2017). A usual practice in the literature, also adopted herein, is to use a priori determined lagged variables and place emphasis on hyperparameter optimization during the training process (see e.g., the implementations by Khan and Coulibaly 2006; Lin et al. 2006; Wang et al. 2009).

Hyperparameters are parameters that can be optimized (or tuned) to limit overfitting (known to deteriorate the forecasting performance of an algorithm), thereby improving the performance of a ML algorithm (Witten et al. 2017, pp. 171–172). This specific utility of hyperparameters justifies their artificial distinguishment from the parameters of the stochastic models and the basic parameters of the ML models. Several examples of hyperparameters can be found in Luo (2016). A common approach to hyperparameter optimization is the herein implemented automatic grid search (Hutter et al. 2015). In optimization via grid search, a complicated optimization problem is solved as the simplified problem of selecting between several candidate model configurations during the training process. The candidate configurations are defined by different predetermined hyperparameter values (Witten et al. 2017, pp. 171–172). In this thesis, hyperparameter optimization is performed using a single validation set extracted from the fitting set.

2.2.6 Time series forecasting using decompositions

Time series forecasting using decompositions (see Section 2.1.7) is a usual practice in the forecasting literature (Hyndman and Athanasopoulos 2018, Chapter 6.8). This practice offers the flexibility of using any forecasting model and algorithm, independently of whether it can automatically consider seasonality (and/or trends). It is herein considered as a flexible methodology for forecasting seasonal time series. We use it, as detailed in the following: First, we estimate the seasonal component of the fitting segment by fitting to it a time series decomposition model. Second, we forecast the time series values in the testing period by training the models on the seasonally decomposed fitting set. Finally, we recover the seasonality to the produced forecasts by assuming that the seasonal component is unchanging.

2.3 Regression algorithmic modelling

2.3.1 Linear and quadratic regression

We apply the linear regression model (see e.g., James et al. 2013; Hastie et al. 2009), whose errors are zero-mean Gaussian i.i.d. (James et al. 2013). We also apply the quadratic regression model. The multiple linear regression model can accommodate quadratic (and polynomial) relationships, as described in James et al. (2013, Chapter 3.3.2). The linear regression model focuses on describing how the mean of the response variable changes with the changes of the predictor variables. For instance, let us assume the simple linear regression model, expressed by Equations (2.42) and (2.43). In these equations, $\underline{y}$ and $\underline{x}$ are the predictand and predictor variables respectively, θ_0 and θ_1 are the regression coefficients, and $\underline{\varepsilon}_o$ is the fixed-variance error term, assumed i.i.d. and normal.

$$\underline{y} = \theta_0 + \theta_1 \underline{x} + \underline{\varepsilon}_o \tag{2.42}$$

$$\underline{\varepsilon}_o \sim N(\mu_o = 0, \sigma_o^2) \tag{2.43}$$

This model is trained on the given sample by:

- Assuming a linear relationship for the mean μ and fixed variance σ^2 for the residuals, as expressed by Equations (2.42) and (2.43). In Equation (2.42), θ_0 and θ_1 are the regression coefficients to be estimated during training.
- Optimizing the objective expressed by Equation (2.44) to estimate θ_0 and θ_1.

$$\min \sum_{j=1}^{\gamma} \varepsilon_{o,j}^2 \tag{2.44}$$

With the estimation of θ_0 and θ_1 two degrees of freedom are lost; therefore, the mean square error MSE that is defined by Equation (2.45) could serve as unbiased estimator of σ^2 (Neter et al. 1983, p. 47).

$$\text{MSE} := (\sum_{j=1}^{\gamma} \varepsilon_{0,j}^2)/(\gamma - 2) \tag{2.45}$$

When γ is large (in practice larger than 30), any new central prediction interval $(1 - \alpha)$, where $\alpha \in (0, 1)$, can be approximated conditional on the new x_j and the training sample exploited in a preceding step by using Equation (2.46), where Φ^{-1} is the inverse standard normal cumulative distribution function (Neter et al. 1983, p. 81).

$$q_p = (\theta_0 + \theta_1 x_j) \pm \Phi^{-1}(1 - \alpha/2)\ (\text{MSE})^{1/2} \tag{2.46}$$

In Equation (2.46), q_p denotes the quantile of level $p \in \{\alpha/2, 1 - \alpha/2\}$.

2.3.2 Regression using neural networks

Artificial neural networks (or neural networks) are an ensemble approach to regression (Hastie et al. 2009, p. 623) and, by extension, to forecasting (see Section 2.2.5), often perceived to mimic the human brain's function. They are perhaps the most widespread machine learning algorithm in hydrology (see e.g., the review by Maier et al. 2010). The main concept of neural networks is to extract linear combinations of the predictor variables as derived features. The dependent variable is then modelled as a nonlinear function of these features (Hastie et al. 2009, p. 389). The main reasons for using neural networks are their high predictive performance and their ability to extract linear combinations of features (Hastie et al. 2009, p. 351). Some of their drawbacks are that: (a) they are prone to overfitting; (b) the inclusion of too many predictor variables can decrease the predictive performance (unlike, for example, random forests); (c) they perform sub-optimally when needed to extrapolate beyond the range of the training set; (d) there are many model structures and architectures to choose from (albeit this can be viewed as an advantage due to offering higher flexibility); (e) appropriate optimization of the model hyperparameters can be important for improving their predictive performance (Maier et al. 2010); and (f) they are computationally slow (Hastie et al. 2009, p. 351).

Detailed information about neural networlds is available, for instance, in Lippmann (1987), Murtagh (1991), Lanc (1992, pp. 7–28), Zhang et al. (1998), Hastie et al. (2009, pp. 389–416), Marsland (2011, pp. 71–110), and Hyndman and Athanasopoulos (2018, Chapter 11.3), while the below synopsis of this information is largely adapted to our computations. We utilize a single-hidden-layer multilayer perceptron (MLP), which consists of interconnected computational units known as nodes or neurons grouped into three layers, namely the input, hidden and output layers. The employed MLP is feed-forward, i.e., the information moves in one direction, specifically from the input nodes to the output nodes through the hidden nodes. This information transit is achieved via (weighted) connections, while all computations are performed in the nodes. The input nodes are inactive, i.e., they do not apply any transfer function (e.g., a sigmoid function) to their inputs before passing them forward, while each of the hidden and output nodes computes the (weighted) sum of its inputs and subsequently applies a transfer function (usually different for the two layers) to this sum. In fact, each group of nodes has its own characteristics that are related to its utility. The number of input nodes is simply the number of lagged variables or the number of time lags. Moreover, the number of output nodes is set to be one, even for multi-step ahead forecasts, since the latter are produced iteratively using one-step ahead predictions as inputs (Cortez 2016; Hyndman and Athanasopoulos 2018, Chapter 11.3).

2.3.3 Regression using random forests

Random forests can also be considered as ensemble methods (Hastie et al. 2009, p. 605; Scornet et al. 2015). Herein we use the original random forests algorithm by Breiman (2001a), i.e., an evolution of the bagging algorithm by Breiman (1996) applied to regression trees (Liaw and Wiener 2002). The term BAGGING is an acronym for Bootstrap AGGregatING (Breiman 1996). Bagging or bootstrap aggregation is an iterative scheme for building a large number of individual

predictors by sampling from the input dataset to finally aggregate the results obtained by them to get the prediction of interest (Biau 2012; Scornet et al. 2015; Biau and Scornet 2016). For continuous variables, the aggregation is made by computing the average of all values obtained by bagged predictors (Sutton 2005; Moisen 2008). This averaging reduces the variance of an estimated prediction function leading to more accurate predictions (Sutton 2005; Hastie et al. 2009, pp. 282–288). Nonetheless, the reduction in variance is limited by large correlation values between pairs of bagged predictors. Random forests are designed to dominate their precursor by offering a further improvement in terms of variance reduction. This improvement is achieved by reducing the correlation between the tree-structured predictors through random selection of the input variables in the tree-growing process (Hastie et al. 2009, pp. 587–588).

The Breiman's random forests algorithm is described in detail in Tyralis et al. (2019b). Some important properties of random forests and their variants, as summarized in the latter study, are that: (a) they have high predictive performance; (b) they are non-linear and non-parametric; (c) they are fast compared to other machine learning algorithms; (d) they are straightforward, easy-to-use and require little tuning of the parameters (default values of the parameters are of high predictive performance); and (e) they are stable and robust to the inclusion of noisy predictor variables. An important drawback of random forests is that they do not extrapolate beyond the range of the training dataset. A systematic review on the use of random forests in water science and technology is also provided in Tyralis et al. (2019b).

2.3.4 Regression using support vector machines

A to-the-point summary of SVM is available in Solomatine and Ostfeld (2008), while Hastie et al. (2009, pp. 417–438) review the theoretical background of these models, and Smola and Schölkopf (2004) provide an overview of their underlying idea with an emphasis on regression and forecasting problems. In contrast to NN and RF that can be conceptualized as structured models with fixed and random architecture respectively (see the above paragraphs), SVM are usually perceived as models utilizing a hyperplane for the separation in a two-dimensional space of two different classes in classification (see e.g., Solomatine and Ostfeld 2008). They are introduced in Cortes and Vapnik (1995) as an extension of the Vapnik's method of optimal hyperplanes. This method is applicable to separable training data, i.e., training data that can be separated without errors, while SVM can be implemented on non-separable training data as summarized subsequently. The input vectors are non-linearly mapped into a high-dimensional feature space, where the hyperplanes are linearly constructed in a way pursuing generalizable (to unobserved situations) solutions. The optimal separating hyperplane is defined as the one that maximizes the margin between the classes in the separable case, and as the one that simultaneously minimizes the number of errors and separates with maximal margin the correctly classified elements in the non-separable case (Smola and Schölkopf 2004). The optimization problem to be solved in regression is a convex optimization problem defined as follows. The objective is to find a function f that simultaneously is as flat as possible and deviates less or equal to ε from all input data values. In cases where this problem is not solvable or we want to allow some errors, the formulation changes so that there is a predefined trade-off between the flatness of f and deviations larger than ε. This trade-off is determined by a constant $C > 0$ (Smola and Schölkopf 2004). Sigma inverse kernel width is a hyperparameter to be specified when using the radial basis and the Laplacian kernel functions for the computations in the feature space (Karatzoglou et al. 2004).

2.4 Process-based hydrological modelling and related procedures

2.4.1 Process-based hydrological modelling at monthly timescale

We perform process-based hydrological modelling at monthly timescale by implementing the Génie Rural à 2 paramètres Mensuel (GR2M) model by Mouelhi et al. (2006b), a parsimonious lumped conceptual model comprising only two parameters, that has been widely applied in the literature (see e.g., Paturel et al. 1995; Niel et al. 2003; Huard and Mailhot 2008; Louvet et al. 2016). This model was developed by adopting a stepwise procedure aiming to identify the most

useful components of a five-parameter model. The latter was inspired from the structures of the monthly model by Makhlouf and Michel (1994), and the daily GR4J model by Perrin et al. (2003; see also Edijatno et al. 1999, Perrin et al. 2001). The first parameter (θ_1) is the maximum capacity of the soil moisture reservoir expressed in mm, while the second one (θ_2) represents water exchange between the studied and adjacent catchments. Values of the second parameter larger (smaller) than 1 indicate water supply from (to) adjacent catchment(s).

2.4.2 Process-based hydrological modelling at daily timescale

We perform process-based hydrological modelling at daily timescale by implementing the Génie Rural à 4 paramètres Journalier (GR4J) model by Perrin et al. (2003), a four-parameter conceptual hydrological model. This model is widely applied in the literature (see e.g., Anctil et al. 2004; Oudin et al. 2005, 2006; Andréassian et al. 2007; Oudin et al. 2010; Wang et al. 2012; Tian et al. 2013; Evin et al. 2014; Lebecherel et al. 2016; Hernández-López and Francés 2017; Tyralis et al. 2019a), while its reliability is well-supported by large-sample empirical results (see Perrin et al. 2003). It was developed by using as starting point the GR3J model by Edijatno et al. (1999), i.e., a three-parameter conceptual hydrological model. A large-sample investigation of the latter can be found in Perrin et al. (2001). GR4J was proposed as an improved (but still parsimonious) version of its precursor model, selected through extensive computational tests among 235 (preliminary) modifications of the latter. Its four parameters are the maximum capacity of the production store (expressed in mm), the groundwater exchange coefficient (expressed in mm), the one-day ahead maximum capacity of the routing store (expressed in mm) and the time base of the unit hydrograph (expressed in days). Its inputs are daily precipitation and potential evapotranspiration, while the output is daily streamflow. For its mathematical formulation, the reader is referred to Perrin et al. (2003).

2.4.3 Procedures supporting process-based hydrological modelling

Process-based hydrological modelling is supported by few additional modelling procedures and choices. These are the following:

- We estimate daily potential evapotranspiration (input to the GR4J model; see Section 2.4.2) by using the formula by Oudin et al. (2005). This formula is the following:

$$\mathrm{PE} = \begin{cases} (0.408\, R\, (T + 5))/100 \text{ if } (T + 5) > 0 \\ 0 \text{ if } (T + 5) \leq 0 \end{cases} \quad (2.47)$$

  In Equation (2.47), PE denotes the daily potential evapotranspiration, R denotes the extraterrestrial solar radiation (MJ m^{-2} d^{-1}) given by the Julian day and the latitude, and T denotes the mean air temperature (°C).

- We apply both the GR2M and GR4J models by using one-year warming-up periods. One-year warming-up periods are often assumed adequate for achieving an optimal state initialisation, while also allowing the full exploitation of the available historical information (see e.g., Edijatno et al. 1999; Perrin et al. 2003; Kim et al. 2018; see also the implementations in Xu 2001; Perrin et al. 2001; Mouelhi et al. 2006b; Vrugt et al. 2008).

- We use the optimization algorithm by Michel (1991; see also the summary available at https://rdrr.io/cran/airGR/man/Calibration_Michel.html) for hydrological model calibration. This algorithm combines a global and a local optimization methodology to optimize a selected objective function. The algorithm begins by performing a screening based on a predefined grid or a list of initial parameter sets for deciding on a single set of parameters. The latter is used as a starting point for a local search procedure, after simple mathematical transformations are applied to them. At each local search iteration, the calibration algorithm determines and tests new parameter set candidates to select a single one to be used as a starting point for the next local search iteration. When the search step becomes smaller than a predefined value, the calibration algorithm stops.

- We simulate the posterior distribution of hydrological model parameters, as detailed in Section 2.5.2.

2.5 Simulation of posterior distributions of model parameters

2.5.1 Simulation of posterior distributions of linear regression model parameters

Some technical remarks on the simulation of the posterior distributions of the parameters of the linear regression models should be made. These remarks are a summary of the information provided by Savel'ev et al. (2015). They are made for the case of the simple linear regression model, while the generalization to the multiple linear regression model is straightforward.

Let us assume the simple linear regression model, expressed by Equations (2.42) and (2.43). Let us also assume that we are given a historical sample $\{(x_i, y_i),\ i = 1, \ldots, \beta\}$, which could be also expressed by Equations (2.48), (2.49) and (2.50).

$$\boldsymbol{x}_{\{1,\ldots,\beta\}} := (x_1, \ldots, x_\beta)^{\mathrm{T}}: \beta \times 1 \tag{2.48}$$

$$\boldsymbol{x}_B := [(1, \ldots, 1)^{\mathrm{T}}, \boldsymbol{x}_{\{1,\ldots,\beta\}}] = [(1, \ldots, 1)^{\mathrm{T}}, (x_1, \ldots, x_\beta)^{\mathrm{T}}]: \beta \times 2 \tag{2.49}$$

$$\boldsymbol{y}_B := \boldsymbol{y}_{\{1,\ldots,\beta\}} := (y_1, \ldots, y_\beta)^{\mathrm{T}}: \beta \times 1 \tag{2.50}$$

This sample can be exploited for simulating the posterior joint distribution of $\underline{\theta}_0$, $\underline{\theta}_1$ and $\underline{\sigma}^2$ by using the herein adopted Gibbs sampler. The latter is described by Equations (2.51) and (2.52), where N_2 denotes the bivariate normal distribution, $\boldsymbol{x}_B'$ the transpose of $\boldsymbol{x}_M$, Inv-Gamma the inverse gamma distribution and $(\theta_0, \theta_1)'$ the transpose of (θ_0, θ_1).

$$\underline{\theta}_0, \underline{\theta}_1 \mid \sigma^2, \boldsymbol{x}_B, \boldsymbol{y}_B \sim N_2((\boldsymbol{x}_B' \boldsymbol{x}_B)^{-1} (\boldsymbol{x}_B' \boldsymbol{y}_B), \sigma^2 (\boldsymbol{x}_B' \boldsymbol{x}_B)^{-1}) \tag{2.51}$$

$$\underline{\sigma}^2 \mid \theta_0, \theta_1, \boldsymbol{x}_B, \boldsymbol{y}_B \sim \text{Inv-Gamma}(M/2, (\boldsymbol{y}_B' \boldsymbol{y}_B - (\theta_0, \theta_1)' \boldsymbol{x}_B' \boldsymbol{y}_B - \boldsymbol{y}_B' \boldsymbol{x}_B (\theta_0, \theta_1) + (\theta_0, \theta_1)' \boldsymbol{x}_B' \boldsymbol{x}_B (\theta_0, \theta_1))^{-1}/2) \tag{2.52}$$

2.5.2 Simulation of posterior distributions of hydrological model parameters

We simulate the posterior distribution of process-based hydrological model parameters within a Bayesian Markov chain Monte Carlo (MCMC) framework. We run parallel Markov chains with different initial values. The iterative simulation is performed by using the Delayed rejection adaptive Metropolis (DRAM) algorithm by Haario et al. (2006). This algorithm combines the concept of adaptive Metropolis sampler and the concept of delayed rejection. We assess the approximate convergence of the simulated chains by implementing the algorithm of Brooks and Gelman (1998), i.e., a multivariate version of the algorithm of Gelman and Rubin (1992). Amongst the outputs of this algorithm is a point estimate that is assumed to be informative about the approximate convergence, while it is based on a comparison of within-chain and between-chain variances. Point estimates substantially larger than 1 indicate lack of convergence. The simulation process is repeated until a point estimate smaller than a predefined value is delivered.

2.6 Quantile regression algorithmic modelling

2.6.1 Basic definitions and concepts

From an applied perspective, quantile regression algorithms are explained in the tutorial article of Waldmann (2018), while a review with up-to-date progress in the field is available in Koenker (2017). Quantile regression algorithms quantify the relationship (within a regression setting) between the predictor variables $\boldsymbol{x}$ (input to the algorithm) and a conditional quantile of the dependent variable y. The quantile $q_\tau(y)$ of random variable y at level (or with probability) $\tau \in (0, 1)$ is defined by the following Equation:

$$q_\tau(y) := F_y^{-1}(\tau) \tag{2.53}$$

In Equation (2.53), where F_y denotes the CDF of y. Moreover, the respective conditional quantile $q_\tau(y|\boldsymbol{x})$ is defined with the following Equation:

$$q_\tau(y|\boldsymbol{x}) := F^{-1}_{y|\boldsymbol{x}}(\tau|\boldsymbol{x}) = y_\tau(\boldsymbol{x}) \quad (2.54)$$

In Equation (2.54), $F_{y|\boldsymbol{x}}$ denotes the CDF of y conditional on $\boldsymbol{x}$. Quantile regression is equivalent to standard regression, with the difference that the former focuses on modelling conditional quantiles instead of modelling conditional means. Most quantile regression algorithms are based on minimization of the average quantile score over all observations. The quantile score (QS_τ; see e.g., Koenker and Machado 1999; Gneiting and Raftery 2007) is defined by Equations (2.55) and (2.56).

$$QS_\tau(u) := (\tau - I\{u < 0\})\, u \quad (2.55)$$

$$u := y_\tau(\boldsymbol{x}) - y \quad (2.56)$$

In Equation (2.55), $I\{\cdot\}$ denotes the indicator function. When the average quantile score is minimized, observations of the dependent variable are divided approximately to two groups including $100\,\tau$ % and $100\,(1-\tau)$ % of the data. This observation has been theoretically confirmed (see Koenker 2017).

According to Waldmann (2018), quantile regression is appropriate when: (a) the interest is in events at the limit of probability; (b) the conditional distribution of the dependent variable is not known or is hard to deduce; (c) there are numerous outliers among the observations of the dependent variable; and (d) heteroscedasticity needs to be modelled. Drawbacks of quantile regression algorithms are also enumerated by Waldmann (2018). A main drawback, shared by most algorithms from this category due to estimating separately different quantiles, is quantile crossing. Furthermore, parameter estimation is harder in quantile regression than in standard regression.

2.6.2 Linear-in-parameters quantile regression

Quantile regression and its variants are extensively analysed in Koenker (2005). The linear-in-parameters quantile regression (or simply "quantile regression") algorithm was introduced by Koenker and Bassett (1978; see also Koenker 2005), following the exploration of quantile estimation problems by Koenker and colleagues in the 70s (see Koenker 2017). Being the linear variant of all quantile regression algorithms, its role is similar to that of standard linear regression in regular regression problems. The method estimates the quantiles of a dependent variable conditional upon selected predictor variables by using similar techniques to linear regression. Intuitively, quantile regression is performed by fitting a linear model and bisecting the data so that $100\,\tau$ % lie below the predicted values of the fitted model. In practice, this is performed by fitting a linear model to the data and minimizing the average quantile score. Two advantages of the quantile regression algorithm, as emphasized by López López et al. (2014) are the robustness of the model with respect to outliers and the fact that no assumption is required for the PDF of the predictand variable.

Some technical remarks on the application of the quantile regression algorithm should also be made. These remarks focus, among others, on the appropriateness of this algorithm for modelling heteroscedasticity (Koenker 2005, p. 25). They are made by compiling information that mostly originates from Neter et al. (1983), Koenker and Hallock (2001), Koenker (2017) and Waldmann (2018).

Let us assume that we are interested in modelling the relationship between the random variables y and x given a training sample $\{(x_j, y_j), j = 1, \ldots, \gamma\}$, so that we can probabilistically predict y conditional on x in general later on. Let also $y_\tau(x)$ denote a quantile at level $\tau \in (0, \ldots, 1)$ of y conditional on x.

In summary, the quantile regression model is trained on the given sample separately for each probability p by:

- Assuming that all quantiles at level τ (or with probability τ) share a common linear relationship with x expressed by Equation (2.57), where $\theta_{0,p}$ and $\theta_{1,p}$ are the regression coefficients to be estimated.

$$y_p(x) = \theta_{0,p} + \theta_{1,p}\, x \tag{2.57}$$

- Optimizing the objective expressed by Equations (2.58) and (2.59) to estimate $\theta_{0,p}$ and $\theta_{1,p}$. Note that the right side of Equation (2.58) has been obtained by also exploiting Equation (2.57) above.

$$u_j := y_{p,j}(x_j) - y_j = \theta_{0,p} + \theta_{1,p}\, x_j - y_j \tag{2.58}$$

$$\min \sum_{j=1}^{\gamma} \left(\tau - \mathrm{I}(u_j < 0)\right) u_j \tag{2.59}$$

Therefore, by using the quantile regression model we are able to model quantiles of random variables "independently of distributional assumptions yet conditional on the data" (Waldmann 2018), with the focus being on describing how selected quantiles of the response variable change with changes of the predictor variable(s). As a result, quantile regression is appropriate for modelling heteroscedasticity.

An illustrative example of modelling heteroscedasticity by using the quantile regression model and a comparison with the solution provided by the linear regression model for the same problem are given in Figure 2.1. We train the quantile regression algorithm by implementing the training algorithm by Koenker and d'Orey (1987, 1994).

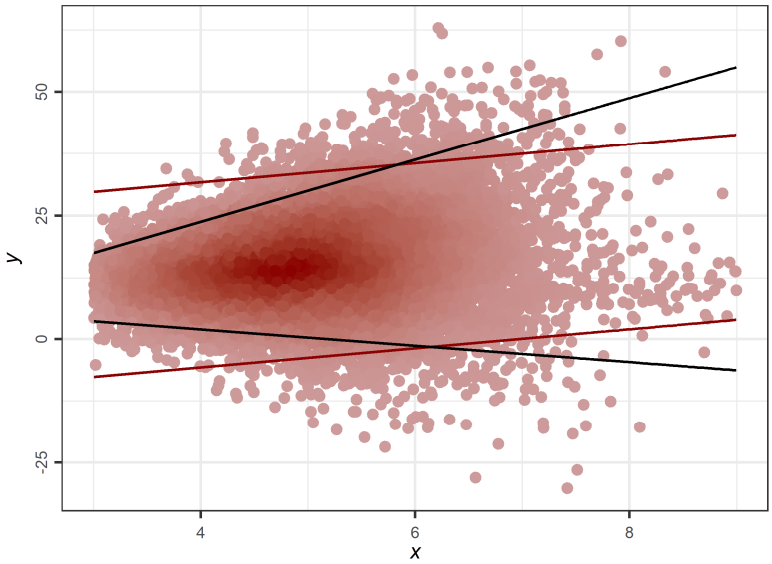


Figure 2.1. Technical illustration of modelling heteroscedasticity using the quantile regression model and comparison with the linear regression model. The training data points are depicted with coloured bubbles (pink for low density and red for high density). The 90% central prediction intervals obtained for this training dataset using the linear regression and quantile regression models are depicted with red and black lines respectively.

2.6.3 Quantile regression forests and generalized random forests

Quantile regression forests were introduced by Meinshausen (2006). They are based on Breiman's (2001a) random forests, which have been extensively used in hydrology (see the review by Tyralis et al. 2009b). Practically, random forests are regression algorithms that average an ensemble of decision trees (see a review on ensemble learning by Sagi and Rokach 2018). The ensemble is created by bagging (abbreviation for bootstrap aggregation; Breiman 1996) regression trees using an additional randomization process. With this additional randomization, the splitting in the nodes of the regression tree is conducted by randomly selecting a fixed number of predictor variables. An extensive description of the procedure for training decision trees and random forests can be found in Hastie et al. (2009, pp. 587–604).

While regression forests can approximate the conditional mean of the dependent variable, quantile regression forests approximate its conditional quantiles. Diverging from other quantile

regression algorithms (see Sections 2.6.2, 2.6.4, 2.6.5), quantile regression forests are not based on the minimization of the quantile score. While random forests estimate the conditional mean by averaging the outcomes of the individual decision trees, quantile regression forests average the indicator functions of the event that the outcome of the decision tree in the test set is lower than q_τ.

Generalized random forests (Athey et al. 2019) and their related quantile prediction algorithms differ from random forests in the implemented partitioning mechanism in the nodes of the decision trees. Due to this procedure, they are theoretically more suitable to model heterogeneities in the observed data compared to quantile regression forests.

2.6.4 Gradient boosting machine and model-based boosting

A general view of boosting methods can be found in Mayr et al. (2014), and Tyralis and Papacharalampous (2020). The concept behind boosting is to iteratively improve (boost) weak learners (i.e., algorithms of low predictive ability) to form a strong learner. A particular type of boosting algorithms, introduced by Friedman (2001), is gradient boosting machine. It is described as an "off-the-shelf" method by Hastie et al. (2009, p. 352). Gradient boosting algorithms minimize a loss function via steepest gradient descent in function space. The main idea is to fit the weak learner to the negative gradient vector of the loss function evaluated at the previous iteration (Mayr et al. 2014). In plain language, boosting is an ensemble learning method in which new models are added to the ensemble sequentially. In particular, at each iteration the new model is trained to minimize the error of the ensemble learnt up until now (Natekin and Knoll 2013). The weak learners used in our case are decision trees. The loss function (i.e., error that has to be minimized) used is the quantile score.

While many of the random forests' properties are shared by gradient boosting machine since both use decision trees as base learners, a major difference is that gradient boosting machine is theoretically expected to perform better due to being highly parameterized (Efron and Hastie 2016, p. 324). However, in practice, random forests often perform better, because optimization required for boosting algorithms is not trivial, while also depends on how accustomed the user is to using the particular algorithm. Instead, random forests are easy to use and perform very well with little tuning.

The most critical parameter in gradient boosting is the number of iterations performed to fit the algorithm. Too few iterations may result in sub-optimal fitting and too many may result in overfitting. While there are different approaches to optimize the parameters of the algorithm (Natekin and Knoll 2013), these approaches are computationally costly in such big datasets. Other drawbacks of gradient boosting machine are: (a) that they are memory-consuming due to a large number of iterations; (b) their evaluation speed; and (c) they are slower to learn compared to random forests.

In addition to decision trees, we also boost linear base learners using the quantile loss function. The relevant theory and implementation are presented by Bühlmann and Hothorn (2007), Hothorn et al. (2018), and Hofner et al. (2014).

2.6.5 Quantile regression using quantile regression neural networks

Artificial neural networks (see Section 2.3.2) can predict conditional quantiles, if they are fitted by minimizing the quantile score. This approach, termed as quantile regression neural networks, was proposed by Taylor (2000). An improved version of quantile regression neural networks by Cannon (2011) is implemented in the present thesis. This version uses the standard multilayer perceptron (MLP) artificial neural networks.

2.7 Probabilistic hydrological modelling and post-processing

2.7.1 Data-driven probabilistic hydrological modelling

Purely statistical (or data-driven) probabilistic prediction models (e.g., those outlined in Section 2.6) could be exploited directly for probabilistic hydrological modelling. Such approaches are herein adopted as benchmarks to probabilistic hydrological post-processing methodologies. The latter consider information provided by process-based models (e.g., those outlined in Section 2.4). Considering this type of information is important in the field of (probabilistic) hydrological modelling, mainly because of the hydrological experience encompassed.

2.7.2 Basic two-stage probabilistic hydrological post-processing

Two-stage post-processing methodologies are implemented by dividing the historical dataset into two independent segments. To outline the main steps and concepts adopted within a basic two-stage hydrological post-processing framework, we first define the time period $T = \{1, \ldots, (n_1+n_2+n_3)\}$, and its three distinct sub-periods $T_1 = \{1, \ldots, n_1\}$, $T_2 = \{(n_1+1), \ldots, (n_1+n_2)\}$ and $T_3 = \{(n_1+n_2+1), \ldots, (n_1+n_2+n_3)\}$. Let us now assume a historical rainfall-runoff dataset extending in the period $\{T_1, T_2\}$. Let us also assume that a probabilistic hydrological prediction is needed for the period T_3. Then the first segment of the historical dataset, extending in the period T_1, is used for calibrating the hydrological model, while information from the period T_2 is used to (a) apply the calibrated hydrological model, and (b) model the hydrological model's error conditional on selected variables (e.g., the hydrological model predictions at times t–1 and t) by using the predicted time series resulted from step (a) alongside with its target values. Under the stationarity and ergodicity assumptions (see e.g., Koutsoyiannis and Montanari 2015 for the implications of these assumptions in hydrological contexts), the trained "error model" can then be applied in the period T_3 for converting a point hydrological prediction obtained using the same hydrological model with the same parameters into a probabilistic hydrological prediction. The error model could fall into the category of conditional distribution models (see e.g., Montanari and Brath 2004; Montanari and Grossi 2008) or the category of (machine learning) quantile regression models (which can directly provide predictive quantiles instead of predictive PDFs; see e.g., Dogulu et al. 2015; López López et al. 2014; Tyralis et al. 2019a; see also Section 2.6), amongst other model categories.

2.7.3 Probabilistic hydrological modelling blueprint

A considerable part of this thesis is devoted to the probabilistic hydrological modelling blueprint by Montanari and Koutsoyiannis (2012). This flexible methodology (referred to as "MK blueprint methodology" in this thesis) is a theoretically consistent two-stage post-processing methodology that exploits information from a large number m of point predictions. Each point prediction is obtained by utilizing the same hydrological model yet with different parameter values and input data. The hydrological model typically falls into the category of process-based hydrological models (see Section 2.4). The hydrological model's parameters are obtained by using data from the period T_1 (defined in Section 2.7.2), while modelling and explicitly considering input data uncertainty imply the availability of input data error information. Information about the hydrological model's error, obtained from the period T_2 (defined in Section 2.7.2), is then used to convert the sister predictions for the period T_3 (defined in Section 2.7.2) to ensemble simulations of the process of interest. The m ensemble simulations are retained as potential realizations of the process of interest, thus collectively composing a probabilistic prediction. For instance, if we are interested in delivering the 90% prediction interval and m = 1 000, then we simply have to pick at each time $t \in T_3$ the 50th and 950th highest values (resulted via ranking) from the spaghetti plot of the 1 000 retained simulations. In absence of relevant information, the MK blueprint methodology can also be applied without explicitly considering input data uncertainty, i.e., by not running ensemble simulations for the hydrological model's input, without any loss of its generality (see e.g., the implementations in Quilty et al. 2019).

2.8 Predictive model output combination and assessment

2.8.1 Predictive model output combination

Prediction combination methodologies are increasingly adopted in various scientific fields for improving predictive modelling (see e.g., the review on ensemble learning methods by Sagi and Rokach 2018). In this thesis, we have combined the outputs of quantile regression algorithms (see Section 2.6) by using the equal-weight combiner (see e.g., Lichtendahl et al. 2013). This combiner simply assigns equal weights to the outputs of the individual algorithms considered. Simple quantile averaging has been made within new probabilistic hydrological post-processing methodologies. For the related background of this thesis, the reader is referred to Section 2.7. For other hydrological predictive model output combination methodologies (besides the equal-weight combiner), the reader is referred to Tyralis et al. (2019a, 2020b), and Papacharalampous and Tyralis (2020).

2.8.2 Point prediction model testing and evaluation

In this Section, we define the metrics exploited for assessing the quality of point predictions (e.g., point forecasts). For the definitions of these metrics, let us consider the point predictions or forecasts $\{f_i, i = 1, \ldots, n\}$ and their corresponding target values $\{x_i, i = 1, \ldots, n\}$.

The E_i metric is defined with the following Equation:

$$E_i := f_i - x_i \tag{2.60}$$

The AE_i metric is defined with the following Equation:

$$\mathrm{AE}_i := |f_i - x_i| \tag{2.61}$$

The PE_i metric is defined with the following Equation:

$$\mathrm{PE}_i := 100(f_i - x_i)/x_i \tag{2.62}$$

The APE_i metric is defined with the following Equation:

$$\mathrm{APE}_i := |100(f_i - x_i)/x_i| \tag{2.63}$$

The ME metric is defined with the following Equation:

$$\mathrm{ME} := (1/n) \sum_{i=1}^{n} (f_i - x_i) \tag{2.64}$$

The MPE metric is defined with the following Equation:

$$\mathrm{MPE} := (-1/n) \sum_{i=1}^{n} (100(f_i - x_i)/x_i) \tag{2.65}$$

The MAE metric is defined with the following Equation:

$$\mathrm{MAE} := (1/n) \sum_{i=1}^{n} |f_i - x_i| \tag{2.66}$$

The MdAE metric is defined with the following Equation:

$$\mathrm{MdAE} := \mathrm{median}_n\{|f_i - x_i|\} \tag{2.67}$$

The MAPE metric is defined with the following Equation:

$$\mathrm{MAPE} := (1/n) \sum_{i=1}^{n} |100(f_i - x_i)/x_i| \tag{2.68}$$

The MdAPE metric is defined with the following Equation:

$$\mathrm{MdAPE} := \mathrm{median}_n\{|100(f_i - x_i)/x_i|\} \tag{2.69}$$

The RMSE metric is defined with the following Equation:

$$\text{RMSE} := \sqrt{(1/n)\sum_{i=1}^{n}(f_i - x_i)^2} \tag{2.70}$$

The PBIAS metric is defined with the following Equation (Yapo et al. 1996):

$$\text{PBIAS} := 100\sum_{i=1}^{n}(f_i - x_i)/\sum_{i=1}^{n}(x_i) \tag{2.71}$$

Let $\bar{x}$ be the mean of the observations, which is defined by Equation (2.17). Let also s_x be the standard deviation of the observations, which is defined by with the following Equation:

$$s_x := \sqrt{(1/(n\text{-}1))\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.72}$$

Let $\bar{f}$ be the mean of the forecasts and s_f be the standard deviation of the point predictions, which are defined with Equations (2.73) and (2.74), respectively.

$$\bar{f} := (1/n)\sum_{i=1}^{n} f_i \tag{2.73}$$

$$s_f := \sqrt{(1/(n\text{-}1))\sum_{i=1}^{n}(f_i - \bar{f})^2} \tag{2.74}$$

The ratio of standard deviations (rSD) metric is defined with the following Equation (Zambrano-Bigiarini 2017a):

$$\text{rSD} := s_f/s_x \tag{2.75}$$

The Nash-Sutcliffe Efficiency (NSE) metric is defined with the following Equation (Nash and Sutcliffe 1970):

$$\text{NSE} := 1 - (\sum_{i=1}^{n}(f_i - x_i)^2/\sum_{i=1}^{n}(x_i - \bar{x})^2) \tag{2.76}$$

The modified Nash-Sutcliffe Efficiency (mNSE) metric is defined with the following Equation (Krause et al. 2005):

$$\text{mNSE} := 1 - (\sum_{i=1}^{n}|f_i - x_i|/\sum_{i=1}^{n}|x_i - \bar{x}|) \tag{2.77}$$

The relative Nash-Sutcliffe Efficiency (rNSE) metric is defined with the following Equation (Krause et al. 2005):

$$\text{rNSE} := 1 - (\sum_{i=1}^{n}((f_i - x_i)/x_i)^2/\sum_{i=1}^{n}((x_i - \bar{x})/\bar{x})^2) \tag{2.78}$$

The index of agreement (d) metric is defined with the following Equation (Krause et al. 2005):

$$d := 1 - (\sum_{i=1}^{n}(f_i - x_i)^2/\sum_{i=1}^{n}(|f_i - \bar{x}|+|x_i - \bar{x}|)^2) \tag{2.79}$$

The modified index of agreement (md) metric is defined with the following Equation (Krause et al. 2005):

$$\text{md} := 1 - (\sum_{i=1}^{n}|f_i - x_i|/\sum_{i=1}^{n}(|f_i - \bar{x}|+|x_i - \bar{x}|)) \tag{2.80}$$

The relative index of agreement (rd) metric is defined with the following Equation (Krause et al. 2005):

$$\text{rd} := 1 - (\sum_{i=1}^{n}((f_i - x_i)/x_i)^2/\sum_{i=1}^{n}((|f_i - \bar{x}|+|x_i - \bar{x}|)/\bar{x})^2) \tag{2.81}$$

The persistence index (cp) metric is defined with the following Equation (Kitanidis and Bras 1980):

$$\text{cp} := 1 - (\sum_{i=2}^{n}(f_i - x_i)^2/\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2) \tag{2.82}$$

The Pearson's correlation coefficient (Pr) metric is defined with the following Equation (Krause et al. 2005):

$$\mathrm{Pr} := (\sum_{i=1}^{n}(x_i - \bar{x})(f_i - \bar{f}))/(\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(f_i - \bar{f})^2)^{0.5} \quad (2.83)$$

The coefficient of determination (r2) metric is defined with the following Equation (Krause et al. 2005):

$$\mathrm{r2} := (\mathrm{Pr})^2 \quad (2.84)$$

The linear regression coefficient (LRC) metric measures the dependence of the point predictions f_i on their corresponding target values x_i, when this dependence is expressed by the following linear regression model:

$$f_i = \mathrm{LRC}\, x_i + b \quad (2.85)$$

The Kling-Gupta efficiency (KGE) metric is defined with the following Equation (Gupta et al. 2009):

$$\mathrm{KGE} := 1 - \sqrt{(\mathrm{Pr} - 1)^2 + ((s_f/s_x) - 1)^2 + ((\bar{f}/\bar{x}) - 1)^2} \quad (2.86)$$

The Volumetric Efficiency (VE) metric is defined with the following Equation (Criss and Winston 2008):

$$\mathrm{VE} := 1 - (\sum_{i=1}^{n}|f_i - x_i|/\sum_{i=1}^{n} x_i) \quad (2.87)$$

2.8.3 Probabilistic prediction model testing and evaluation

In this Section, we define the metrics exploited for assessing the quality of probabilistic predictions. For these definitions, let us assume that probabilistic predictions for a period T_3 have been obtained, e.g., by using one of the flexible methodologies of Sections 2.7.2 and 2.7.3. For a specific prediction interval of level $(1 - \alpha)$, $0 < \alpha < 1$, formed by the predictive quantiles $\{w_t, t \in T_3\}$ and $\{l_t, t \in T_3\}$, where w_t and l_t are the upper and lower quantiles, respectively, at time t, the coverage probability (CP_α), average width (AW_α) and average interval score (AIS_α) are defined with Equations (2.88), (2.89) and (2.90), respectively. In these equations, y_t is the targeted observation at time $t \in T_3$ and $|T_3|$ is the number of the target data points included in period T_3.

$$\mathrm{CP}_\alpha := \sum_t (\mathrm{I}\{l_t < y_t < w_t\})/|T_3| \quad (2.88)$$

$$\mathrm{AW}_\alpha := \sum_t (w_t - l_t)/|T_3| \quad (2.89)$$

$$\mathrm{AIS}_\alpha(l_t, w_t; y_t) := \sum_t ((w_t - l_t) + (2/\alpha)(l_t - y_t)\,\mathrm{I}\{y_t < l_t\} + (2/\alpha)(y_t - w_t)\,\mathrm{I}\{y_t > w_t\})/|T_3| \quad (2.90)$$

For a predictive quantile of level τ, $0 < \tau < 1$, the average quantile score (AQS_τ) is defined with the following Equation (see also Equations (2.55) and (2.56)):

$$\mathrm{AQS}_\tau(y_\tau(\boldsymbol{x}_t); y_t) := \sum_t ((\tau - \mathrm{I}\{(y_t - y_\tau(\boldsymbol{x}_t)) < 0\})(y_t - y_\tau(\boldsymbol{x}_t)))/|T_3| \quad (2.91)$$

Some remarks should be made on the above scores. In probabilistic modelling, the aim is to maximize the sharpness of the predictive PDFs, subject to reliability (Gneiting and Katzfuss 2014). Reliability (or calibration) is the statistical correspondence between the probabilistic forecasts and the observations, while sharpness is the concentration of the predictive PDFs in absolute terms (Gneiting and Katzfuss 2014; see also Gneiting and Raftery 2007; Gneiting et al. 2007). Reliability and sharpness are both important criteria for assessing the usefulness of probabilistic predictions. Reliability can be assessed by measuring the coverage of the delivered prediction intervals, i.e., the percentage of data points included in these intervals; see Equation (2.88) above. Sharpness can be assessed by computing the average widths of the obtained interval predictions. For engineering applications, narrower prediction intervals are preferred for avoiding excessively precautionary design or decisions.

AIS_α and AQS_τ provide an objective co-assessment of reliability and sharpness, and are, therefore, suggested as "proper scores" in Gneiting and Raftery (2007). In this view, smaller values

of these scores indicate more useful probabilistic predictions. Some remarks on the (average) interval score are made in the following: This score is appropriate for assessing probabilistic predictions in the form of prediction intervals (Gneiting and Raftery 2007, Section 6.2). It has three components (see Equation (8.4) above). The first component is the width of the prediction interval. As smaller values of the (average) interval score indicate better predictions than larger values (for a specific prediction problem), this component penalizes more the wider prediction intervals than the narrower ones (thereby rewarding narrow prediction intervals). The two remaining components quantify the distance between each of the two predictive quantiles forming the prediction interval and the observed value, in case that the latter falls outside of the prediction interval, and penalize larger distances more than smaller distances. In general, the (average) interval score should become smaller as we move from the outer to the inner prediction intervals. The reader is referred to Gneiting and Raftery (2007, Section 6.2) for detailed information on how to interpret this score. The origins of the interval score (see e.g., Gneiting and Raftery 2007) trace back to Dunsmore (1968) and Winkler (1972). This score, also known as Winkler score, rewards narrow prediction intervals, while penalizing prediction intervals missed by observations. The size of the penalty depends on the prediction interval (Gneiting and Raftery 2007).

For benchmarking purposes we also compute the relative improvements (RI_{α,P_1,P_2}), obtained when using a prediction interval P_1 of level $(1 - \alpha)$ (provided by a predictor of interest) with respect to another prediction interval P_2 of the same level (provided by a benchmark predictor) in terms of average width. This computation is made according to Equation (2.92). In this equation, AW_{α,P_1} and AW_{α,P_2} denote the average widths of the former and latter prediction intervals, respectively. We also compute the relative improvements (provided by each predictor of interest with respect to a benchmark predictor) in terms of average interval score and average quantile score. These latter computations are made by using Equations analogous to Equation (2.92).

$$RI_{\alpha,P_1,P_2} := (AW_{\alpha,P_2} - AW_{\alpha,P_1})/AW_{\alpha,P_2} \tag{2.92}$$

2.8.4 Predictive model hierarchical clustering

We perform hierarchical clustering of time series forecasting methods by conducting hierarchically clustered heatmaps. All hierarchical cluster analyses are performed by using a set of dissimilarities for the n forecasting methods being clustered at each time. These dissimilarities are derived from computed performance metric values (see Section 2.8.2). The hierarchical clustering algorithm begins by assigning each forecasting method is to its own cluster and progresses by gradually joining the two most similar clusters, until a single cluster is obtained. At each stage, it also re-computes the distances between the clusters. Hierarchical clustering is explained in detail in Hastie et al. (2009, Chapter 14.3.12).

2.9 Predictive modelling and benchmarking toolbox

2.9.1 Original and processed hydrological datasets

Our frameworks rely on large hydrological datasets, which are part of the technical background of this thesis. The exploited original datasets are summarized in Table 2.1. Moreover, a summary of the processed hydrological datasets is presented in Table 2.2.

Table 2.1. Original real-world datasets. These datasets are used for forming the ones of Table 2.2.

S/n	Code name	Main references	Chapter						
			3	4	5	6	7	8	9
1	GRDC	GRDC (2017)	✓	×	×	×	×	×	×
2	GHCN-temp	Lawrimore et al. (2011)	×	✓	✓	✓	×	×	×
3	GHCN-prec	Peterson and Vose (1997)	×	✓	✓	✓	×	×	×
4	MOPEX	Schaake et al. (2006)	×	×	×	×	×	✓	×
5	CAMELS	Newman et al. (2014); Addor et al. (2017a)	×	×	×	×	×	×	✓

Table 2.2. Processed real-world datasets. The total number of the exploited real-world time series is 5 929.

S/n	Chapter	Dataset type	Original dataset (see Table 2.1)	Hydrometeorological process	Data level	Number of time series	Time series length (years)
1	3	Typical	GRDC	River discharge	Annual	405	100
2	4		GHCN-temp	Temperature	Annual	185	91
3				Standardized temperature		185	
4			GHCN-prec	Precipitation		112	
5				Standardized precipitation		112	
6	5		GHCN-temp	Temperature	Monthly	985	40
7			GHCN-prec	Precipitation		1 552	
8	6		GHCN-temp	Temperature	Monthly	17	10–125
9			GHCN-prec	Precipitation		33	10–119
10	8	Rainfall-runoff	MOPEX	Precipitation	Monthly	270	50
				Potential evaporation		270	
				River discharge		270	
11	9		CAMELS	Precipitation	Daily	511	34
				Temperature		511	
				River discharge		511	

2.9.2 Automatic models and flexible methodologies

A summary of the ready-made automatic models and algorithms exploited in this thesis is presented in Table 2.3, while their utilities are reported independently in Table 2.4. These models and algorithms are exploited individually or by combination with various algorithmic argument choices. Most of them incorporate several others, which are here omitted for reasons of brevity. Flexible methodologies incorporating these models are outlined in Table 2.5.

Table 2.3. Ready-made automatic models and algorithms implemented and combined within the context of the thesis. Most of these models and algorithms incorporate several others, which are here omitted for reasons of brevity. Flexible methodologies incorporating these models are outlined in Table 2.5.

S/n	Model or algorithm	Description	Chapter						
			3	4	5	6	7	8	9
1	Sample autocorrelation function (ACF)	Section 2.1.2	✓	✓	×	×	×	×	×
2	Sample partial autocorrelation (PACF)		✓	×	×	×	×	×	×
3	White noise	Section 2.1.1	×	×	×	×	✓	×	×
4	Autoregressive moving average (ARMA)	Section 2.1.3	✓	✓	×	×	×	×	×
5	Autoregressive fractionally integrated moving average (ARFIMA)	Section 2.1.5	✓	✓	×	×	×	×	×
6	Fractional Gaussian noise (fGn)	Section 2.1.6	✓	✓	✓	✓	×	×	×
7	Kalman filter	Section 2.1.9	×	×	×	✓	×	×	×
8	Additive model	Section 2.1.7	×	×	✓	✓	×	×	×
9	Multiplicative model		×	×	✓	✓	×	×	×
10	Box-Cox transformation	Section 2.1.8	×	×	✓	✓	×	×	×
11	Square-root transformation		×	×	×	×	×	✓	×
12	Yeo-Johnson transformation		×	×	×	×	×	✓	×
13	Ordered quantile transformation		×	×	×	×	×	✓	×
14	Non-seasonal naïve	Section 2.2.1	✓	✓	×	×	×	×	×
15	Seasonal naïve		×	×	✓	✓	×	×	×
16	Random walk (RW)		✓	✓	✓	×	×	×	×
17	Fixed-order autoregressive moving average (ARMA)	Section 2.2.2	✓	✓	×	✓	×	×	×
18	Optimum-order autoregressive integrated moving average (ARIMA)		✓	✓	✓	✓	×	×	×
19	Optimum-order autoregressive fractionally integrated moving average (ARFIMA)		✓	✓	✓	✓	×	×	×
20	Exponential smoothing state space with Box-Cox transformation, ARMA errors correction, trend and seasonal components (BATS)	Section 2.2.3	✓	✓	✓	✓	×	×	×
21	Exponential smoothing with error, trend and seasonal components (ETS)		✓	✓	×	×	×	×	×
22	Simple exponential smoothing (SES)		✓	✓	✓	×	×	×	×
23	Theta		✓	✓	✓	✓	×	×	×
24	Prophet	Section 2.2.4	×	×	✓	×	×	×	×
25	Sample autocorrelation function (ACF)	Section 2.1.2	✓	✓	×	×	×	×	×
26	Autoregressive (AR) model	Section 2.1.3	✓	✓	×	×	×	×	×
27	Sliding window model	Section 2.2.5	✓	✓	×	✓	×	×	✓
28	Grid search		✓	✓	×	✓	×	×	×
29	Linear regression	Section 2.3.1	×	×	×	×	✓	✓	×
30	Quadratic regression		×	×	×	×	✓	×	×
31	Neural networks (NN)	Section 2.3.2	✓	✓	×	✓	×	×	×
32	Random forests (RF)	Section 2.3.3	✓	✓	×	×	×	×	×
33	Support vector machines (SVM)	Section 2.3.4	✓	✓	×	✓	×	×	×
34	Oudin's formula	Section 2.4.3	×	×	×	×	×	×	✓
35	Michel's algorithm		×	×	×	×	×	✓	✓
36	Génie Rural à 2 paramètres Mensuel (GR2M)	Section 2.4.1	×	×	×	×	×	✓	×
37	Génie Rural à 4 paramètres Journalier (GR4J)	Section 2.4.2	×	×	×	×	×	×	✓
38	Gibbs sampler	Section 2.5.1	×	×	×	×	✓	×	×
39	Delayed rejection adaptive Metropolis (DRAM) sampler	Section 2.5.2	×	×	×	×	×	✓	×
40	Brooks and Gelman's algorithm		×	×	×	×	×	✓	×
41	Quantile regression (qr)	Section 2.6.2	×	×	×	×	✓	✓	✓
42	Generalized random forests for quantile regression (qrf)	Section 2.6.3	×	×	×	×	×	×	✓
43	Generalized random forests for quantile regression emulating quantile regression forests (qrf_meins)		×	×	×	×	×	×	✓
44	Gradient boosting machine with trees as base learners (gbm)	Section 2.6.4	×	×	×	×	×	×	✓
45	Model-based boosting with linear models as base learners (mboost_bols)		×	×	×	×	×	×	✓
46	Quantile regression neural networks (qrnn)	Section 2.6.5	×	×	×	×	×	×	✓
47	Hierarchical clustering	Section 2.8.4	✓	✓	×	×	×	×	×

Table 2.4. Utilities of the individual ready-made automatic models and algorithms implemented and combined within the context of the thesis. These models and algorithms are defined in Table 2.3. Flexible methodologies incorporating these models are outlined in Table 2.5 together with their utilities.

Model or algorithm	Utility																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
	Process-based hydrological model calibration	Hyperparameter optimization	Input data matrix building	Objective variable selection	Potential evapotranspiration modelling	Predictive model clustering	Process-based hydrological modelling	Quantile regression algorithmic modelling	Regression algorithmic modelling	Simulation of posterior distributions of model parameters	Time series characterization	Time series decomposition	Time series forecasting	Time series gap-filling	Time series simulation	Time series standardization	Time series transformation
Sample ACF	×	×	×	✓	×	×	×	×	×	×	✓	×	×	×	×	×	×
Sample PACF	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×
White noise	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×
Fixed-parameter ARMA	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×
Fixed-parameter ARFIMA	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×
fGn	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	✓	×
Kalman filter	×	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×
Additive model	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×
Multiplicative model	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×
Box-Cox transformation	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓
Square-root transformation	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓
Yeo-Johnson transformation	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓
Ordered quantile transformation	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	✓
Non-seasonal naïve	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Seasonal naïve	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
RW	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Fixed-order ARMA	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	✓	×	×
Optimum-order ARIMA	×	×	×	✓	×	×	×	×	×	×	×	×	✓	×	✓	×	×
Optimum-order ARFIMA	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
BATS	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
ETS	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
SES	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Theta	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Prophet	×	×	×	×	×	×	×	×	×	×	×	×	✓	×	×	×	×
Sliding window model	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Grid search	×	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Linear regression	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×
Quadratic regression	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×
NN	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×
RF	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×
SVM	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×
Oudin's formula	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×	×
Michel's algorithm	✓	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
GR2M	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×
GR4J	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×
Gibbs sampler	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×
DRAM sampler	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×
Brooks and Gelman's algorithm	×	×	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×
qr	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×
qrf	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×
qrf_meins	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×
gbm	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×
mboost_bols	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×
qrnn	×	×	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×
Hierarchical clustering	×	×	×	×	×	✓	×	×	×	×	×	×	×	×	×	×	×

Table 2.5. Flexible methodologies considered for predictive modelling. The serial numbers continue from Table 2.3. Their utilities are also reported. The serial numbers of these utilities continue from Table 2.4.

S/n	Strategy or methodology	Utility		Description	Chapter						
		Outline	s/n		3	4	5	6	7	8	9
48	Transformation-based time series modelling	Improved time series modelling	18	Section 2.1.8	×	×	✓	×	×	✓	×
49	Regression-based time series forecasting	Time series forecasting	13	Section 2.2.5	✓	✓	×	✓	×	×	×
50	Decomposition-based time series forecasting			Section 2.2.6	×	×	✓	✓	×	×	×
51	Approximate convergence of parallel Markov chains with different initial values	Simulation of posterior distributions of model parameters	10	Section 2.5.2	×	×	×	×	×	✓	×
52	Simple quantile averaging	Prediction combination	19	Section 2.8.1	×	×	×	×	✓	✓	✓
53	Basic two-stage probabilistic hydrological post-processing	Probabilistic hydrological post-processing	20	Section 2.7.2	×	×	×	×	✓	✓	✓
54	Probabilistic hydrological modelling blueprint			Section 2.7.3	×	×	×	×	✓	✓	×

2.9.3 Predictive model evaluation metrics

A summary of the metrics exploited in this thesis for predictive model testing and evaluation is presented in Table 2.6.

Table 2.6. Metrics exploited for predictive model testing and evaluation. The metrics are defined in Section 2.8.2, while $(1 - \alpha)$, $0 < \alpha < 1$, denotes the level (or probability) of a prediction interval.

S/n	Full name	Notation	Definition	Values	Optimum value	Chapter						
						3	4	5	6	7	8	9
1	Error	E_i	Equation (2.60)	(-∞,+∞)	0	×	✓	✓	×	×	×	×
2	Absolute error	AE_i	Equation (2.61)	[0, +∞)	0	×	✓	✓	×	×	×	×
3	Percentage error	PE_i	Equation (2.62)	(-∞,+∞)	0	×	✓	×	×	×	×	×
4	Absolute percentage error	APE_i	Equation (2.63)	[0, +∞)	0	×	✓	×	×	×	×	×
5	Mean error	ME	Equation (2.64)	(-∞,+∞)	0	✓	×	×	×	×	×	×
6	Mean percentage error	MPE	Equation (2.65)	(-∞,+∞)	0	✓	×	×	×	×	×	×
7	Mean absolute error	MAE	Equation (2.66)	[0, +∞)	0	✓	×	×	×	×	×	×
8	Median absolute error	MdAE	Equation (2.67)	[0, +∞)	0	×	✓	×	×	×	×	×
9	Mean absolute percentage error	MAPE	Equation (2.68)	[0, +∞)	0	✓	×	×	×	×	×	×
10	Median absolute percentage error	MdAPE	Equation (2.69)	[0, +∞)	0	×	✓	×	×	×	×	×
11	Root mean square error	RMSE	Equation (2.70)	[0, +∞)	0	✓	×	✓	✓	×	×	×
12	Percent bias	PBIAS	Equation (2.71)	(-∞,+∞)	0	✓	×	×	×	×	×	×
13	Ratio of standard deviations	rSD	Equation (2.75)	(-∞, 1]	1	✓	×	×	✓	×	×	×
14	Nash-Sutcliffe efficiency	NSE	Equation (2.76)	(-∞, 1]	1	✓	×	✓	✓	×	×	×
15	Modified Nash-Sutcliffe efficiency	mNSE	Equation (2.77)	(-∞, 1]	1	✓	×	×	×	×	×	×
16	Relative Nash-Sutcliffe Efficiency	rNSE	Equation (2.78)	(-∞, 1]	1	✓	×	×	×	×	×	×
17	Index of agreement	d	Equation (2.79)	[0, 1]	1	✓	×	×	✓	×	×	×
18	Modified index of agreement	md	Equation (2.80)	[0, 1]	1	✓	×	×	×	×	×	×
19	Relative index of agreement	rd	Equation (2.81)	(-∞, 1]	1	✓	×	×	×	×	×	×
20	Persistence index	cp	Equation (2.82)	(-∞, 1]	1	✓	×	×	×	×	×	×
21	Pearson's correlation coefficient	Pr	Equation (2.83)	[-1, 1]	1	✓	×	×	✓	×	×	×
22	Coefficient of determination	r2	Equation (2.84)	[0, 1]	1	✓	✓	×	×	×	×	×
23	Linear regression coefficient	LRC	Equation (2.85)	(-∞,+∞)	1	×	✓	×	✓	×	×	×
24	Kling-Gupta efficiency	KGE	Equation (2.86)	(-∞, 1]	1	✓	×	×	×	×	×	×
25	Volumetric efficiency	VE	Equation (2.87)	(-∞,+∞)	1	✓	×	×	×	×	×	×
26	Reliability score	RS_α	Equation (2.88)	[0, 1]	$(1 - \alpha)$	×	×	×	×	✓	✓	✓
27	Average width	AW_α	Equation (2.89)	[0, +∞)	0	×	×	×	×	✓	✓	✓
28	Average interval score	AIS_α	Equation (2.90)	[0, +∞)	0	×	×	×	×	✓	✓	✓
29	Average quantile score	AQS_τ	Equation (2.91)	[0, +∞)	0	×	×	×	×	×	×	✓

2.9.4 Statistical software information

The analyses and visualizations are performed in R Programming Language (R Core Team 2019). We use the contributed R packages summarized in Table 2.7. Many of these R packages rely on

others that are here omitted for reasons of brevity. These additional R packages can be found in the provided references under the category "latest version exploited".

Table 2.7. R packages directly exploited in the thesis. Most of these R packages rely on others that are here omitted for reasons of brevity.

S/n	R package	Latest version exploited	Other references	Chapter						
				3	4	5	6	7	8	9
1	airGR	Coron et al. (2019)	Coron et al. (2017)	×	×	×	×	×	✓	✓
2	BayesSummaryStatLM	Savel'ev et al. (2015)	–	×	×	×	×	✓	×	×
3	bestNormalize	Peterson (2019)	Peterson (2017)	×	×	×	×	×	✓	×
4	cgwtools	Witthoft (2015)	–	✓	×	×	×	×	×	×
5	coda	Plummer et al. (2019)	Plummer et al. (2006)	×	×	×	×	×	✓	×
6	data.table	Dowle and Srinivasan (2019)	–	×	×	×	×	✓	✓	✓
7	devtools	Wickham et al. (2019c)	–	✓	✓	✓	✓	✓	✓	✓
8	dplyr	Wickham et al. (2019b)	–	×	×	×	×	×	✓	✓
9	EnvStats	Millard (2018)	Millard (2013)	✓	×	×	×	×	×	×
10	FME	Soetaert and Petzoldt (2016)	Soetaert and Petzoldt (2010)	×	×	×	×	×	✓	×
11	forecast	Hyndman et al. (2018)	Hyndman and Khandakar (2008)	✓	✓	✓	✓	×	×	×
12	fracdiff	Fraley et al. (2012)	–	✓	✓	✓	✓	×	×	×
13	gbm	Greenwell et al. (2019)	Friedman (2001)	×	×	×	×	×	×	✓
14	gdata	Warnes et al. (2017)	–	✓	✓	✓	✓	✓	✓	✓
15	ggExtra	Attali (2018)	–	×	×	×	×	✓	×	×
16	ggplot2	Wickham et al. (2019a)	Wickham (2016a)	✓	✓	✓	✓	✓	✓	✓
17	ggridges	Wilke (2018)	–	×	×	×	×	×	✓	×
18	ggpubr	Kassambara (2019)	–	×	×	×	×	×	×	✓
19	grf	Tibshirani and Athey (2019)	Meinshausen (2006)	×	×	×	×	×	×	✓
20	hddtools	Vitolo (2017)	Vitolo (2018)	×	×	×	×	×	✓	×
21	HKprocess	Tyralis (2016)	Tyralis and Koutsoyiannis (2011)	✓	✓	✓	✓	×	×	×
22	hydroTSM	Zambrano-Bigiarini (2017b)	–	×	×	×	✓	×	×	×
23	kernlab	Karatzoglou et al. (2018)	Karatzoglou et al. (2004)	✓	✓	×	✓	×	×	×
24	knitr	Xie (2019)	Xie (2014, 2015)	✓	✓	✓	✓	✓	✓	✓
25	maps	Brownrigg et al. (2018)	–	×	✓	✓	✓	×	✓	✓
26	MASS	Ripley (2019)	Venables and Ripley (2002)	×	×	×	×	✓	×	×
27	matrixStats	Bengtsson (2018)	–	×	×	×	×	✓	✓	×
28	mboost	Hothorn et al. (2018)	Hofner et al. (2014)	×	×	×	×	×	×	✓
29	nnet	Ripley (2016)	Venables and Ripley (2002)	✓	✓	×	✓	×	×	×
30	plyr	Wickham (2001)	Wickham (2016b)	✓	×	×	×	✓	✓	✓
31	prophet	Taylor and Letham (2017)	Taylor and Letham (2018)	×	×	✓	×	×	×	×
32	qrnn	Cannon (2019)	Cannon (2011)	×	×	×	×	×	×	✓
33	quantreg	Koenker (2019)	Koenker and Bassett (1978)	×	×	×	×	✓	✓	✓
34	randomForest	Liaw (2018)	Liaw and Wiener (2002)	✓	✓	×	×	×	×	×
35	readr	Wickham et al. (2018)	–	✓	✓	✓	✓	×	✓	✓
36	reshape	Wickham (2018)	Wickham (2007)	×	×	×	×	✓	✓	×
37	reshape2	Wickham (2017)	–	×	×	×	×	×	×	✓
38	rmarkdown	Allaire et al. (2019)	–	✓	✓	✓	✓	✓	✓	✓
39	rminer	Cortez (2016)	Cortez (2010)	✓	✓	×	✓	×	×	×
40	stringi	Gagolewski (2019)	–	×	×	×	×	×	×	✓
41	stringr	Wickham (2019)	–	×	×	×	×	×	×	✓
42	tidyr	Wickham and Henry (2019)	–	✓	×	×	✓	✓	✓	×
43	zoo	Zeileis et al. (2019)	Zeileis and Grothendieck (2005)	×	×	✓	✓	×	✓	×

3. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes

Research within the field of hydrology often focuses on the comparison between stochastic and machine learning (ML) forecasting methods. The performed comparisons are based on case studies, while a study providing large-scale results on the subject is missing. In this Chapter, we compare 11 stochastic and nine ML methods regarding their multi-step ahead forecasting properties by conducting 12 extensive computational experiments based on simulations. Each of these experiments uses 2 000 time series generated by linear stationary stochastic processes. We conduct each simulation experiment twice; the first time using time series of 100 values and the second time using time series of 300 values. Additionally, we conduct a real-world experiment using 405 mean annual river discharge time series of 100 values. We quantify the forecasting performance of the methods using 18 metrics. The results indicate that stochastic and ML methods may produce equally useful forecasts.

3.1 Introduction

The fundamental problem of statistically producing point forecasts of univariate time series by exploiting information from their past values only (hereafter *"forecasting"*, unless specified differently) is of traditional interest to hydrological scientists (Yevjevich 1987). Right after the introduction of the currently classical autoregressive integrated moving average (ARIMA) models by Box and Jenkins (1968), Carlson et al. (1970) used several stationary models of this specific family, i.e., autoregressive moving average (ARMA) models, to forecast the evolution of four annual time series of streamflow processes. Today the available models for time series forecasting are numerous and can be classified according to De Gooijer and Hyndman (2006) into eight categories, i.e., (a) exponential smoothing, (b) ARIMA, (c) seasonal models, (d) state space and structural models and the Kalman filter, (e) nonlinear models, (f) long-range dependence models, e.g., the family of autoregressive fractionally integrated moving average (ARFIMA) models, (g) autoregressive conditional heteroscedastic/generalized autoregressive conditional heteroscedastic (ARCH/GARCH) models, and (h) count data forecasting. The models from the categories (a)−(g) are of potential interest in hydrology, while they can be implemented for both one- and multi-step ahead forecasting.

The theoretical properties of the models of categories (a)−(d), (f), (g) (hereafter referred to as "stochastic") have been more or less investigated, in contrast to those of the nonlinear models and in particular the machine learning (ML) algorithms, also referred to in the literature as "black-box models". These two main categories of models are known to represent two different cultures in statistical modelling, i.e., the "data modelling culture" and the "algorithmic modelling culture" (Breiman 2001b). The former assumes that an analytically formulated stochastic model is behind the generation of the data, while the latter that behind this process is something complex and unknown, which does not have to be analytically formulated, as long as a purely algorithmic model can offer high forecast accuracy. In other words, profoundly understanding and properly modelling the (future) behaviour of a process are strongly connected within the data modelling culture, but completely irrelevant within the algorithmic modelling culture. The distinction between causal explanation, prediction and description is acknowledged and clarified in terms of modelling in Shmueli (2010). Still, one could question whether the (rather artificial) separation of models with respect to the "stochastic-ML dipole" actually corresponds to a striking difference in their forecasting performance.

What cannot be questioned, on the other hand, is the popularity that the various ML forecasting methods have gained in many scientific fields, including hydrology. Amongst the most popular ML algorithms are the neural networks (NN), random forests (RF) and support vector machines (SVM). The latter two algorithms are presented in their current forms in Breiman (2001a), and Cortes and Vapnik (1995; see also Vapnik 1995, 1999), respectively. For the implementation of NN for time series forecasting the reader is referred to Zhang et al. (1998) and Zhang (2001), while a review of SVM forecasting applications can be found in Sapankevych and

Sankar (2009). The large number of hydrological studies implementing NN and SVM forecasting methods is imprinted in Maier and Dandy (2000), and Raghavendra and Deka (2014), respectively. Moreover, Abrahart et al. (2012) collectively review the NN streamflow forecasting and rainfall-runoff applications (see e.g., De Vos 2013). A major difference between these two families of applications is the use of exogenous variables in the latter. In contrast to NN and SVM, RF are barely utilized for hydrological process forecasting.

To explore the related background and facilitate the following discussion, in Table 3.1 we present some literature information on hydrometeorological time series forecasting emphasizing a few key aspects and concepts. As it is apparent, hydrological research often focuses on ML or hybrid (e.g., combinations of ARMA and ML) forecasting methods and, in particular, on the comparison between stochastic (mainly ARMA and ARFIMA) and ML methods. However, the culture of assessing the performance of forecasting methods on large datasets is not customary in hydrology. Therefore, the assessment is made within case studies. Concerning the testing procedure, while the available forecast quality metrics are a lot, most of the studies use only a few (Krause et al. 2005), understating the importance of the testing process despite relevant suggestions (see e.g., Armstrong 2001; Abrahart et al. 2008; Humphrey et al. 2017). Likewise, the number of the compared forecasting methods is usually small and simple benchmarks are rarely included in the comparisons, although their use is highly recommended in the (hydrological) forecasting literature (see e.g., Harvey 1984; Pappenberger et al. 2015; Hyndman and Athanasopoulos 2018, Chapter 3.1).

Table 3.1. Methodological information on case studies focusing on hydrometeorological time series forecasting within a purely statistical framework (see also Table 4.1).

S/n	Study	Primary focus	Hydrometeorological process				Data level				Horizon		
			Temperature	Precipitation	Streamflow or river discharge	Other	Hourly	Daily	Monthly	Annual	One-step ahead	Multi-step ahead	Not clear
1	Atiya et al. (1999)	NN methods	×	×	✓	×	×	×	✓	×	×	✓	×
2	Lambrakis et al. (2000)		×	×	✓	×	✓	×	×	×	✓	×	×
3	Kişi (2007)		×	×	✓	×	×	✓	×	×	✓	✓	×
4	Cheng et al. (2008)		×	×	✓	✓	×	×	✓	✓	✓	✓	×
5	Yaseen et al. (2016)		×	×	✓	×	×	✓	✓	×	✓	×	×
6	Sivapragasam et al. (2001)	SVM methods	×	✓	✓	×	×	✓	×	×	✓	×	×
7	Shi and Han (2007)		×	×	✓	✓	×	×	✓	✓	✓	✓	×
8	Lu and Wang (2011)		×	✓	×	×	×	✓	×	×	✓	×	×
9	Hu et al. (2001)	Hybrid methods	×	✓	×	×	×	×	×	✓	✓	×	×
10	Kim and Valdés (2003)		×	×	×	✓	×	×	✓	×	✓	✓	×
11	Pai and Hong (2007)		×	✓	×	×	✓	×	×	×	✓	×	×
12	Hong (2008)		×	✓	×	×	✓	×	×	×	✓	×	×
13	Kişi and Cimen (2011)		×	×	✓	×	×	×	✓	×	✓	×	×
14	Liong and Sivapragasam (2002)	SVM vs NN methods	×	×	×	✓	×	✓	×	×	✓	✓	×
15	Guo et al. (2011)		×	×	✓	×	×	×	✓	×	×	×	✓
16	Kişi and Cimen (2012)		×	✓	×	×	×	✓	×	×	✓	×	×
17	He et al. (2014)		×	×	✓	×	×	✓	×	×	✓	×	×
18	Jain et al. (1999)	Stochastic vs ML methods	×	×	✓	×	×	×	✓	×	✓	×	×
19	Ballini et al. (2001)		×	×	✓	×	×	×	✓	×	✓	✓	×
20	Kişi (2004)		×	×	✓	×	×	×	✓	×	✓	✓	×
21	Khan and Coulibaly (2006)		×	×	×	✓	×	×	✓	×	✓	✓	×
22	Lin et al. (2006)		×	×	✓	×	×	×	✓	×	×	×	✓
23	Mishra et al. (2007)		×	×	×	✓	×	×	✓	×	✓	✓	×
24	Yu and Liong (2007)		×	×	✓	×	×	✓	×	×	×	✓	×
25	Koutsoyiannis et al. (2008)		×	×	✓	×	×	×	✓	×	✓	×	×
26	Wang et al. (2009)		×	×	✓	×	×	×	✓	×	×	✓	×
26	Abudu et al. (2010)		×	×	✓	×	×	×	✓	×	✓	×	×
28	Kişi et al. (2012)		×	×	×	✓	×	✓	×	×	✓	✓	×
29	Shabri and Suhartono (2012)		×	×	✓	×	×	×	✓	×	✓	×	×
30	Valipour et al. (2013)		×	×	✓	×	×	×	✓	×	×	×	✓
31	Patel and Ramachandran (2015)		×	×	✓	×	×	×	✓	×	×	✓	×

Researchers have long been chasing the most accurate forecast for their data, a "universally best technique". On the other hand, there is an argument that it is the data and the application of interest that determine the proper methodology for each case, rather than vice versa (Hong and Fan 2016). Another argument is that perhaps research should invest more on probabilistic forecasting (e.g., using Bayesian statistics, as made in Tyralis and Koutsoyiannis 2014) and less on point forecasting (Krzysztofowicz 2001b). In fact, the opinions on forecast evaluation are often diverging, as they tend to depend on the perspective from which the forecasts are examined. An interesting study on this subject can be found in Murphy (1993). The latter identifies three criteria for this specific evaluation, which are adopted as a foundation for further discussion in later studies (e.g., Ramos et al. 2010; Weijs et al. 2010). These criteria are (1) the consistency during the forecasting process, (2) the quality or the correspondence between the forecasts and the target values, and (3) the value or the profit that the forecast provide to the decision makers. Weijs et al. (2010) note that criterion (2) concerns more the pure science, while criterion (3) is closer related to the decisions made within the engineering applications (of science), rather than science itself. Thus, only a few studies are dedicated to criterion (3), such as Ramos et al. (2010) and Ramos et al. (2013), while the greatest part of the literature focuses on criterion (2). The latter likewise largely applies to the present Chapter and to all of its references aiming to deal with the

modelling issue (*which model should I use?*) within specific hydrological contexts. Another criterion of practical importance is the computational (running) time required for obtaining the forecasts. This information might be significant depending on the forecasting task, while it could also be decisive, especially when one has to select between methods producing equally useful forecasts. The computational requirements are known to depend on the primary algorithm and its complexity, as well as on its software implementation, while the computational time also depends on the computer.

Regarding the so far conducted comparisons between forecasting methods, their majority in all scientific fields is based on case studies. Nevertheless, in some few cases beyond hydrology the number of the examined real-world time series is quite large. These time series are realizations of several phenomena, which however are fundamentally different from being hydrological, and their examination includes concepts that are rather inappropriate in hydrological terms (e.g., paying attention to small quantitative differences in the forecasting performance of the methods). Examples of such studies can be found in Makridakis et al. (1987), Makridakis and Hibon (2000), and Ahmed et al. (2010), which examine 1 001, 3 003 and 1 045 time series respectively. Within these studies a statistical analysis is performed and the results are presented accordingly. Furthermore, the literature includes two studies, specifically Zhang (2001) and Thissen et al. (2003), in which the performance of the methods is assessed on simulated time series from linear stochastic processes. The scale of the simulation experiment is small in both cases. Thissen et al. (2003) examine one long time series from the ARMA family, and Zhang (2001) examine eight stochastic processes from the ARMA family and 30 simulated time series for each stochastic process. The forecasting methods are ARMA models, NN and SVM in the former study, and ARMA models and NN in the latter study, while Makridakis et al. (1987), Makridakis and Hibon (2000), and Ahmed et al. (2010) do not focus their comparisons on the stochastic-ML dipole.

Admittedly, the studies mentioned in the previous paragraph pursue generalized results to greater extent than most of the available studies. However, the gap still remains. What specifically needs to be addressed is whether the stochastic-ML dipole actually corresponds to a clear difference in the forecasting performance of the methods, especially in the light of published studies, which claim that they found a technique better than others. Given the fact that each forecasting case is indisputably unique, this task would necessarily require the examination of a sufficiently large and representative sample of forecasting cases within the same (properly designed) methodological framework. Extensive simulations combined with statistical analysis and benchmarking (i.e., evaluation in comparison to standard approaches and/or theoretically expected outcomes) can constitute, nevertheless, a highly effective approach to solving the problem under discussion. In more detail, for the generalized comparison of stochastic and ML forecasting methods, a sufficient number of different and representative of the underlying phenomena time series could be used for the estimation of the expected performance of forecasting methods regarding several criteria of interest. The need of using simulated time series to assess the performance of forecasting methods is emphasized by forecasting experts (Bontempi 2013). The analytical approach in assessing the performance of ML algorithms is not possible; therefore, the only alternative approach is using simulations. Apparently, the larger the scale of the simulation experiments, the more general would be the results. Real-world experiments of large scale could be used to complement the results of the simulation experiments in alignment with specific applications. Some suggestions for the design of large-scale comparisons and the incorporation of benchmarking into methodological frameworks are available in Alpaydin (2010) and Hothorn et al. (2005), respectively.

In the context described so far, we perform an extensive comparison between several stochastic and ML methods for the forecasting of hydrological processes by conducting large-scale computational experiments based on simulations. The comparison refers to the multi-step ahead forecasting properties of the methods. The simulated time series are 48 000 in total, while they are generated by linear stationary stochastic processes. The latter are commonly used for modelling hydrological processes. In fact, the linearity assumption starts to become reasonable when modelling hydrological variables at large time scales (e.g., annual or monthly), while at fine

time scales (e.g., daily or hourly) non-linear modelling approaches start to prevail (e.g., due to intermittency). Moreover, stationary models, in contrast to the non-stationary ones, are established as the appropriate modelling choice when dealing with natural processes, unless tangible and quantitative information that can fully support a deterministic description (not based on data but on physical laws) of change in time is available (Koutsoyiannis 2011; Koutsoyiannis and Montanari 2015). Additionally to the simulation experiments, we examine 405 real-world time series. Our aim is to fill the gap detected in the literature by providing large-scale results and useful insights on the comparison of stochastic and ML forecasting methods for the case of hydrological time series forecasting at large time scales, with an emphasis on annual river discharge processes. A strength (and limitation) of the present Chapter (implied by its aim) is the adopted approach to the problem, i.e., the algorithmic or data-driven approach.

3.2 Methodology

In this Section, we present the basic methodological elements of this Chapter and the way that these elements are combined into a framework for evaluating forecasting methods in hydrology. Basic information on the methods' implementation is also provided, while the total of the exploited `R` packages is independently listed in Section 2.9.4. Hereafter, to specify an implemented `R` function, we state its name accompanied by the name of the `R` package. The latter name is given between curly brackets (`{}`). To imply that we implement a built-in-`R` function, we accompany its name with "`{stats}`". All `R` functions are used as specified in this methodology overview. If no specification is made, then the default values are adopted. We note that the use of default values is acknowledged in the literature as a "reasonable and justified choice" in most cases (see e.g., Arcuri and Fraser 2013). To ensure reproducibility, the `R` codes and data are available in Papacharalampous and Tyralis (2018b; hereafter referred to as "Chapter's supplement").

3.2.1 Simulated processes

We simulate time series according to several stochastic models from the frequently used families of ARMA and ARFIMA. This modelling approach is considered appropriate for the achievement of our aim and has been widely applied in hydrology (see e.g., Montanari et al. 1997, 1999, 2000; Ballini et al. 2001; Wang et al. 2009; Valipour et al. 2013). The simulated stochastic processes are presented in Table 3.2, while for the related definitions the reader is referred to Section 2.1. These 12 stochastic models correspond to different types of autocorrelation.

Table 3.2. Simulated stochastic processes. Their definitions are given in the Section 2.1. The parameters μ and σ of the simulated stochastic processes are set to 0 and 1 respectively.

S/n	Stochastic process	Autoregressive and/or moving average parameters	`R` function
1	AR(1)	$\varphi_1 = 0.7$	`arima.sim` `{stats}`
2	AR(1)	$\varphi_1 = -0.7$	
3	AR(2)	$\varphi_1 = 0.7$, $\varphi_2 = 0.2$	
4	MA(1)	$\theta_1 = 0.7$	
5	MA(1)	$\theta_1 = -0.7$	
6	ARMA(1,1)	$\varphi_1 = 0.7$, $\theta_1 = 0.7$	
7	ARMA(1,1)	$\varphi_1 = -0.7$, $\theta_1 = -0.7$	
8	ARFIMA(0,0.45,0)		`fracdiff.sim` `{fracdiff}`
9	ARFIMA(1,0.45,0)	$\varphi_1 = 0.7$	
10	ARFIMA(0,0.45,1)	$\theta_1 = -0.7$	
11	ARFIMA(1,0.45,1)	$\varphi_1 = 0.7$, $\theta_1 = -0.7$	
12	ARFIMA(2,0.45,2)	$\varphi_1 = 0.7$, $\varphi_2 = 0.2$, $\theta_1 = -0.7$, $\theta_2 = -0.2$	

3.2.2 Real-world time series

We examine 405 mean annual river discharge time series of 100 values, sourced from GRDC (2017). For the exploration of these time series, we compute the sample autocorrelation function (ACF; see e.g., Section 2.1.2) and the sample partial autocorrelation function (PACF; see e.g.,

Section 2.1.2). Herein, the computation is made by using the `R` function `acf {stats}`. The side-by-side boxplots of the ACF and PACF estimates are presented in Figure 3.1. To describe the long-range dependence in river discharge processes, we estimate the Hurst parameter (H) of the fractional Gaussian noise process (see Section 2.1.6) for each time series. The fitting is made by using the `R` function `mleHK {HKprocess}`. The latter implements the maximum likelihood method. A histogram of the H estimates is presented in Figure 3.1. By its examination, we observe that the magnitude of the long-range dependence is mostly significant in the real-world time series.

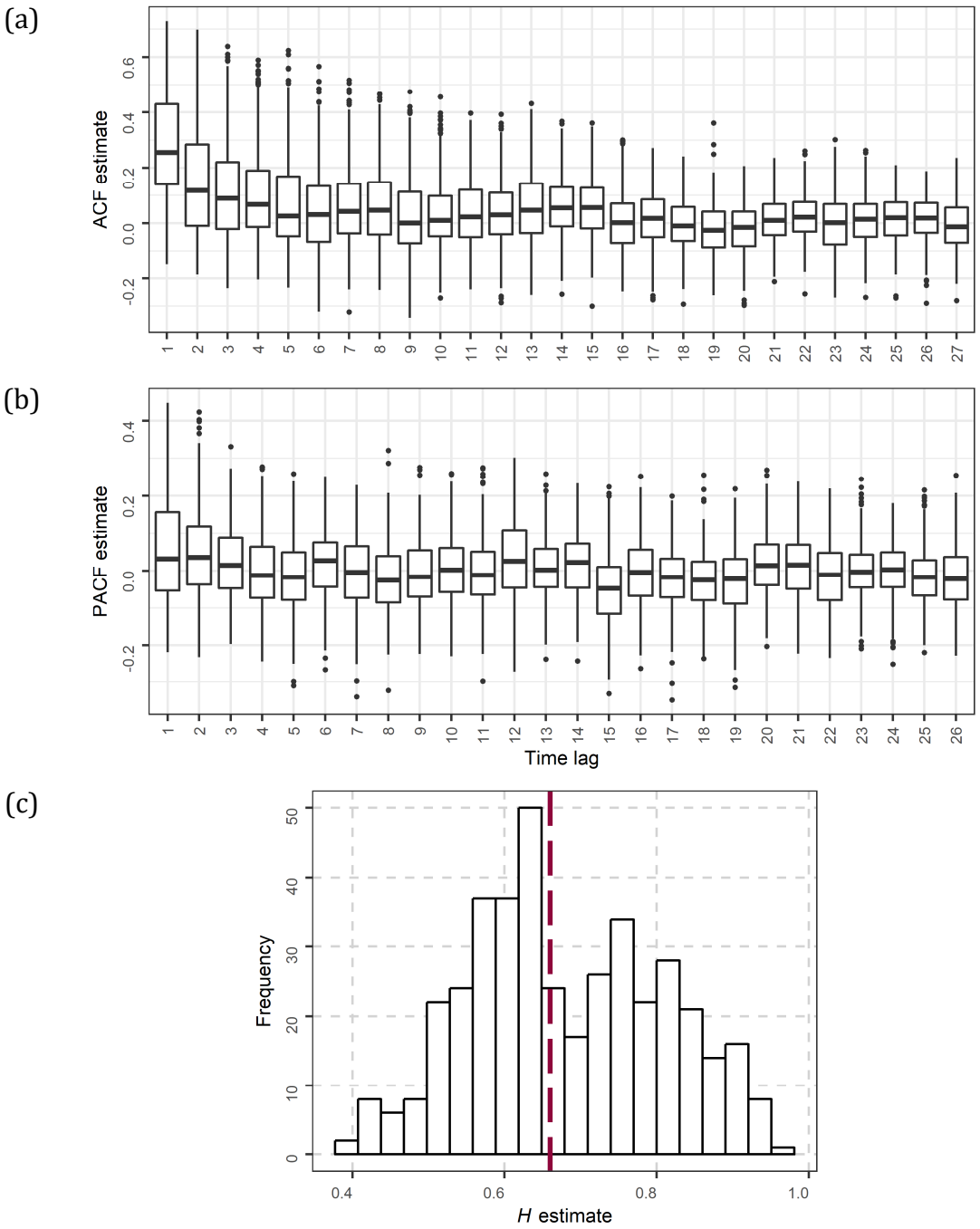


Figure 3.1. Estimates of the (a) autocorrelation function, (b) partial autocorrelation function, and (c) Hurst parameter (H) of the fractional Gaussian noise process for the mean annual river discharge time series sourced from GRDC (2017). The red dashed line in (c) denotes the median of the H estimates.

3.2.3 Forecasting methods

We compare 11 stochastic and nine ML forecasting methods. The primary forecasting models and algorithms are detailed in Section 2.2 (see also the references therein), while here we place emphasis on their reproducibility. We note that the understanding from a theoretical point of view of most methods could hardly help in interpreting the algorithmically obtained outcome of the comparison.

The stochastic methods are presented in Table 3.3, together with their classification into five general categories. In the same Table, we provide the `R` functions used for their implementation, and refer the reader to their detailed methodological descriptions and theoretical explanations in Section 2.2. ARIMA_f and ARIMA_s are implemented with their numbers of the autoregressive (AR) and moving average (MA) parameters (p and q, respectively) set to be the same to those used in the time series simulation process (see Section 3.2.1), while the number of differencing (d) is set to zero. On the contrary, auto_ARIMA_f, auto_ARIMA_s and auto_ARFIMA automatically estimate the order of the AR(F)IMA models, as summarized in Section 2.2.2 (see the descriptions for the optimum-order ARIMA and ARFIMA models therein). It is essential to also note that ARIMA_s and auto_ARIMA_s are simulation models, while the innovations are set to zero by the ARIMA_f, auto_ARIMA_f and auto_ARFIMA methods, i.e., the forecasts produced by the latter three methods are the expected future values from the AR(F)IMA model selected during the training process. For the definitions of the ARMA, ARIMA and ARFIMA models, the reader is referred to Sections 2.1.3, 2.1.4 and 2.1.5, respectively.

Table 3.3. Stochastic methods and their implementation. The forecasting methods are available in code form in Chapter's supplement. All R functions are used with predefined values, unless specified differently.

S/n	Abbreviated name	Corresponding model from Table 2.3	General category	Description	R functions	Implementation notes
1	Naïve	Non-seasonal naïve	Simple	Section 2.2.1	–	–
2	RW	Random walk			rwf {forecast}	(drift = TRUE)
3	ARIMA_f	Fixed-order autoregressive integrated moving average (ARIMA)	ARIMA	Section 2.2.2	Arima {forecast}, forecast {forecast}	Arima {forecast} (include.mean = TRUE, include.drift = FALSE, method = "ML")
4	ARIMA_s				Arima {forecast}, simulate {stats}	
5	auto_ARIMA_f	Optimum-order ARIMA			auto.arima {forecast}, forecast {forecast}	–
6	auto_ARIMA_s				auto.arima {forecast}, fracdiff {fracdiff}, simulate {stats}	
7	auto_ARFIMA	Optimum-order autoregressive fractionally integrated moving average (ARFIMA)	ARFIMA		arfima {forecast}, forecast {forecast}	arfima {forecast} (estim = "mle")
8	BATS	Exponential smoothing state space with Box-Cox transformation, ARMA errors correction, trend and seasonal components (BATS)	Innovations State Space	Section 2.2.3	bats {forecast}, forecast {forecast}	–
9	ETS_s	Exponential smoothing with error, trend and seasonal components (ETS)			ets {forecast}, simulate {stats}	–
10	SES	Simple exponential smoothing (SES)	Exponential Smoothing		ses {forecast}	–
11	Theta	Theta			thetaf {forecast}	–

The ML methods are presented in a compact form in Tables 3.4 and 3.5. In the same Tables, we list the R functions used for their implementation and refer the reader to specific Sections of Chapter 2, in which their documentation is provided. The training of the ML forecasting methods involves lagged variable selection and hyperparameter optimization procedures, discussed in detail in Section 2.2.5. The considered hyperparameter values and the adopted procedures for selecting the time lag(s) (one at minimum) are reported in Tables 3.4 and 3.5, respectively, while some supporting information to the former table are provided subsequently.

Table 3.4. Machine learning methods. The serial numbers continue from Table 3.3. The time lag selection procedures adopted are defined in Table 3.5. The forecasting methods are available in code form in Chapter's supplement. All `R` functions are used with predefined values, unless specified differently.

S/n	Abbreviated name	Corresponding model from Table 2.3	Description	Key model information	`R` functions	Implementation notes	
						Hyperparameter optimized (grid values)	Time lag selection procedure
12	NN_1	Neural networks	Section 2.3.2	Single-hidden-layer multilayer perceptron	`CasesSeries {rminer}`, `fit {rminer}`, `lforecast {rminer}`, `nnet {nnet}`	Number of hidden nodes (0, 1, ..., 15)	1
13	NN_2						2
14	NN_3				`nnetar {forecast}`, `nnet {nnet}`		3
15	RF_1	Random forests	Section 2.3.3	Breiman's random forests algorithm with 500 grown trees	`CasesSeries {rminer}`, `fit {rminer}`, `lforecast {rminer}`, `randomForest {randomForest}`	Number of variables randomly sampled as candidates at each split (1, ..., 5)	1
16	RF_2						2
17	RF_3						3
18	SVM_1	Support vector machines	Section 2.3.4	Radial basis kernel "Gaussian" function, $C = 1$, $\varepsilon = 0.1$	`CasesSeries {rminer}`, `fit {rminer}`, `lforecast {rminer}`, `ksvm {kernlab}`	Sigma inverse kernel width (2^n, $n = -8, -7, ..., 6$)	1
19	SVM_2						2
20	SVM_3						3

Table 3.5. Lagged variable selection procedures adopted for the machine learning methods of Table 3.4. The forecasting methods are available in code form in Chapter's supplement. All `R` functions are used with predefined values, unless specified differently.

S/n	Selected time lags	Corresponding model from Table 2.3	`R` function
1	The corresponding to an estimated value for the autocorrelation function (ACF; see Section 2.1.2), i.e., the time lags 1, ..., 19 for a time series of 90 values and the time lags 1, ..., 24 for a time series of 290 values	Sample ACF	`acf {stats}`
2	The corresponding to a statistically significant estimated value for the ACF. If there is no statistically significant estimated value for the ACF, the corresponding to the largest estimated value		
3	According to the `R` function `nnetar {forecast}`, i.e., the time lags 1, ..., k, where k is equal to the maximum between 1 and the number of parameters of an autoregressive (AR) model fitted to the time series data. The optimal number of AR parameters is decided using the Akaike information criterion (AIC; see Section 2.1.10)	AR model	`ar {stats}`

We here use three objective methods to select the lagged variables to be used in regression (see Table 3.5). The first of these methods is inspired by the `R` function `nnetar {forecast}`, while the remaining two are new. Given the selected lagged variables, we then perform hyperparameter optimization via automatic grid search (see Section 2.2.5). In time series forecasting using neural networks, the number of hidden nodes is an (integer-valued) hyperparameter to be optimized during the training process. The candidate architecture configurations are defined by fixed numbers of layers, input and output nodes according to the above-outlined information, and different possibilities for the number of hidden nodes according to Table 3.4. Zero number of hidden nodes and, consequently, no hidden layer is a feasible option within our experiments. In time series forecasting using random forests, the optimized hyperparameter is the number of variables randomly sampled as candidates at each split (integer-valued hyperparameter) during the tree-growing process, while the candidate configurations to choose from during the training process are five. Finally, in time series forecasting using support vector machines, we adopt the default kernel function and the default C and ε values, and optimize sigma inverse kernel width (continuous hyperparameter) during the training process according to Table 3.4.

3.2.4 Forecast quality metrics

We utilize the forecast quality metrics briefly presented in Table 3.5. These metrics do not share one-to-one relationships with each other, emphasizing -more or less- different aspects of the same information. Their classification into six main categories according to the criterion/criteria that is/are (co-)assessed through their use is also presented in Table 3.6. These criteria are two types of accuracy, the capture of the variance and the correlation. By type 1 accuracy we mean the closeness of the forecasted time series to the target time series, while by type 2 accuracy we mean the closeness of the mean of the forecasts to the mean of the target values. The definitions of the forecast quality metrics are listed in Section 2.8.2, while in the below paragraphs we justify their combined use in this Chapter.

Table 3.6. Forecast quality metrics. Their definitions are given in Section 2.8.2. Their possible and optimum values are given in Table 2.6.

S/n	Abbreviated name	Full name	Criterion/criteria	Condition (preferred values)
1	MAE	Mean absolute error	Type 1 accuracy	smaller MAE
2	MAPE	Mean absolute percentage error		smaller MAPE
3	RMSE	Root mean square error		smaller RMSE
4	NSE	Nash-Sutcliffe efficiency		larger NSE
5	mNSE	Modified Nash-Sutcliffe efficiency		larger mNSE
6	rNSE	Relative Nash-Sutcliffe efficiency		larger rNSE
7	cp	Persistence index		larger cp
8	ME	Mean error	Type 2 accuracy	smaller |ME|
9	MPE	Mean percentage error		smaller |MPE|
10	PBIAS	Percent bias		smaller |PBIAS|
11	VE	Volumetric efficiency		smaller |VE - 1|
12	rSD	Ratio of standard deviations	Capture of the variance	larger min{rSD, 1/rSD}
13	Pr	Pearson's correlation coefficient	Correlation	larger Pr
14	r2	Coefficient of determination		larger r2
15	*d*	Index of agreement	Type 1 accuracy, capture of the variance	larger *d*
16	md	Modified index of agreement		larger md
17	rd	Relative index of agreement		larger rd
18	KGE	Kling-Gupta efficiency	Type 2 accuracy, capture of the variance, correlation	larger KGE

MAE provides an easily interpretable assessment with respect to the type 1 accuracy criterion, while it is also amongst the most frequently used forecast quality metrics (Hyndman and Koehler 2006). Similarly, the computation of MAPE and RMSE is implied by their traditional use in the forecasting field (Armstrong and Collopy 1992; Hyndman and Koehler 2006). Although RMSE is more sensitive to outliers than MAE (Fildes 1992; Hyndman and Koehler 2006), the former is usually preferred to the latter by forecasting scientists mainly because of its "theoretical relevance in statistical modelling" (Hyndman and Koehler 2006). Furthermore, MAPE is a scale-independent metric, offering an advantage in comparing forecasting methods across different datasets. Nonetheless, this metric is particularly affected by target values close to zero (Fildes 1992; Hyndman and Koehler 2006). The ME and MPE metrics are also utilized herein as they constitute analogues (with similar advantages and disadvantages) to MAE and MAPE respectively for the assessment according to the type 2 accuracy criterion.

Some limitations of the correlation metrics, i.e., the Pr and r2 ones, mainly related to an over-sensitivity to outliers and to the fact that their optimum value does not indicate by itself a perfect forecast, are well understood in hydrology and beyond (see e.g., Legates and McCabe 1999; Armstrong 2001). However, their use is of traditional significance (Legates and McCabe 1999; Krause et al. 2005) and could not harm the interpretation of the results, when these metrics are used attentively and collectively with others (Krause et al. 2005). Perhaps the most widely used metric in the field of hydrology is the introduced by Nash and Sutcliffe (1970) NSE, while another traditional metric is *d* (Legates and McCabe 1999; Krause et al. 2005; Schaefli and Gupta 2007). Consequently, these two metrics are also considered helpful in communicating the results of the present Chapter. The use of their original versions, which are known to be over-sensitive and under-sensitive to high and low outliers respectively (Krause et al. 2005), is herein complemented

by the use of their modified and relative versions, i.e., the mNSE, rNSE, md and rd metrics. These four metrics can provide improved forecast evaluation depending on the data (Krause et al. 2005). Moreover, Zambrano-Bigiarini (2017a) places cp, VE, PBIAS, rSD and KGE amongst the metrics of potential interest to hydrological scientists and provides references about their use in the hydrological field (see e.g., Kitanidis and Bras 1980; Yapo et al. 1996). VE and KGE are introduced by Criss and Winston (2008) and Gupta et al. (2009) to overcome some drawbacks of NSE (and mNSE).

The rationale of using this large set of forecast quality metrics is also supported by suggestions made by experts in the field of hydrology and beyond; see Abrahart et al. (2008) and Armstrong (2001) respectively. According to the latter study, when feasible, multiple metrics should be used collectively with an emphasis on the most relevant ones. Herein, we place some emphasis on type 1 accuracy, since a good performance with respect to this criterion is a major pursuance in most of the forecasting applications. Finally, we note that amongst the utilized forecast quality metrics the MAPE, NSE, mNSE, rNSE, cp, MPE, PBIAS, VE, rSD, Pr, r2, *d*, md, rd and KGE ones are dimensionless, while MAE, RMSE and ME are expressed in the same units as the data (and the forecasts).

3.2.5 Methodology outline

To compare the forecasting methods of Section 3.2.3, we conduct 12 large-scale computational experiments based on simulations. Within each of these experiments we simulate 2 000 time series according to a stochastic process (see Section 3.2.1). We conduct each simulation experiment twice; the first time using time series of 100 values and the second time using time series of 300 values. The simulation experiments are hereafter referred to under their code names. The latter are composed by two parts separated by an underscore. The first part is "SE" (acronym for Simulation Experiment), while the second part is the serial number of the simulated process, as reported in Table 3.2, followed by the letter "a" or "b" to denote the length of the simulated time series, i.e., 100 or 300 values respectively. Additionally, we conduct a real-world experiment using the time series presented in Section 3.2.2. Within the experiments using ARMA simulated processes we test all the forecasting methods except for auto_ARFIMA. The latter method is tested within the experiments using ARFIMA simulated processes or real-world time series instead of the ARIMA_f, ARIMA_s, auto_ARIMA_f and auto_ARIMA_s ones. The total number of forecasts is 858 480, among which 6 480 are produced within the real-world experiment.

For the application of the stochastic methods, we divide each time series into two segments, i.e., the training segment and the test segment, which contain n_1 and n_2 values respectively, as indicated in Figure 3.2(a). We fit the stochastic models to the former and produce forecasts for the latter using the recursive multi-step ahead forecasting method. For the total of the conducted experiments n_2 equals 10, while n_1 equals 90 or 290 depending on the length of the used time series. For the application of the ML forecasting methods, we additionally divide the segment of n_1 values into two parts, as presented in Figure 3.2(b). The tail of the training segment is hereafter referred to as "validation segment" and serves hyperparameter selection, as delineated subsequently. We fit the ML model several times to the first $[2n_1/3]$ values of the training segment, each time using different hyperparameter values according to Table 3.4. The fitted configurations of the ML model are then utilized to produce forecasts for the validation segment. We compute the RMSE values of these forecasts using the actual values of the validation segment as reference information and decide on an optimum hyperparameter value (i.e., the corresponding to the smallest RMSE). Finally, we fit the ML model with the selected hyperparameter value to the whole training segment and produce forecasts for the test segment. The rationale of adopting this procedure is explained in Witten et al. (2017, pp. 171–172; see also Section 2.2.5). In summary, both the validation and test segments are used for testing and comparing models that have been previously fitted to independent (with respect to these segments) information. The former testing facilitates the decision on a ML method variant, so that the ML method is afterwards considered fully trained, while the latter enables the comparison between all the (fully trained) forecasting methods.

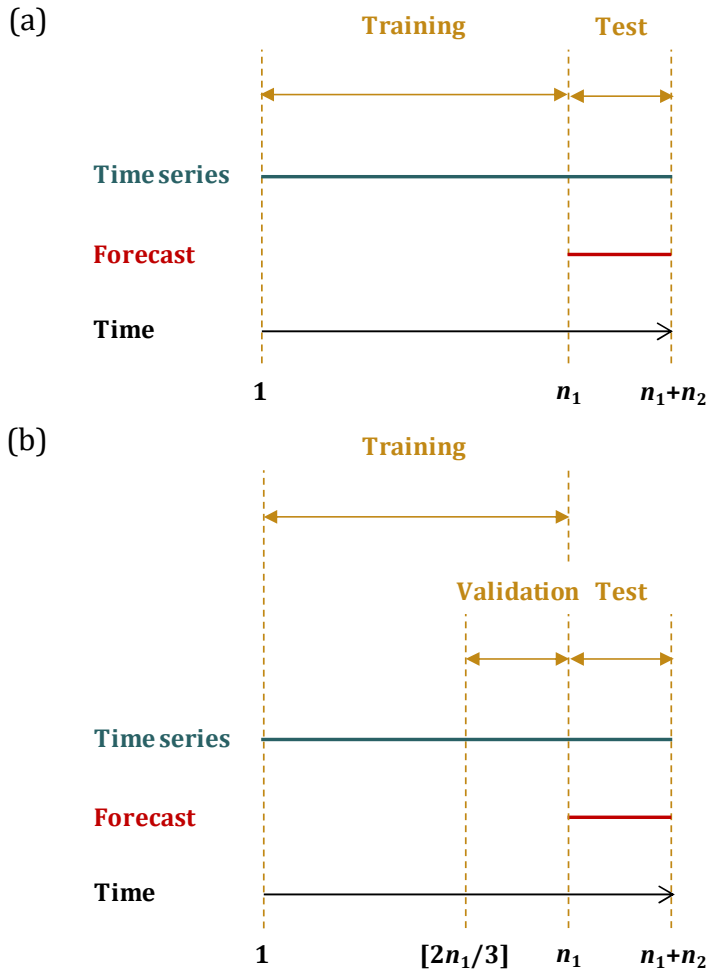


Figure 3.2. Time series segment division for the application of the (a) stochastic and (b) machine learning methods. For the latter category the validation segment serves the hyperparameter optimization procedure.

We provide a multi-faced assessment and comparison of the forecasting methods by utilizing the forecast quality metrics briefly presented in Section 3.2.4. The values of these metrics are computed for each forecasting test (conducted for a specific forecasting method and a specific time series) on the test segment. We mainly compare the medians and interquartile ranges (iqr) of the metric values, as computed for each forecasting method per experiment. We compare the medians, as described in Table 3.6, while the smallest the iqr the better the forecasts. We also apply a clustering analysis on the forecasting methods based on the median values of the forecast quality metrics. This analysis can ease the extraction of information from the experiments. It can also facilitate the identification of possible repeating patterns in the clustering of the forecasting methods. The presence or absence of such repeating patterns could be strongly connected to algorithmic aspects and elements that we aim to reveal with the conducted experiments. In particular for the real-world experiment, we rank the forecasting methods for each individual test and further compute an average-case ranking for each metric. We place our emphasis on the 18 average-case rankings and not directly in the mean or median values of the metrics, because the latter might be more affected by the results of specific time series. This practice was first adopted by Tyralis and Papacharalampous (2017).

Finally, we measure the total computational time consumed by each forecasting method within the various experiments using the `R` function `system.time {stats}`. We present these measurements to allow a simplified and easily interpretable comparison of the implemented methods in terms of computational requirements. The computations are performed in our regular home PC, while the computational times could differ significantly for other PCs.

3.2.6 Benchmarking information

Although our computational experiments are designed to produce new knowledge in the field of hydrological time series forecasting, there are several outcomes rather well known at the forefront of our methodological framework. In more detail, ARIMA_f is expected to produce optimal forecasts with respect to the type 1 accuracy criterion, mainly in terms of RMSE, on the time series resulting from the simulation of ARMA processes because of its theoretical background, specifically for two reasons. Firstly, it uses by design the p, d, q numbers that are used in the simulation procedure; therefore, in its case the forecasting procedure is in essence the inverse of the simulation procedure. Furthermore, it produces minimum mean square error forecasts by setting the innovations to zero (see Wei 2006, pp. 88–93 for the related theoretical proof). Moreover, auto_ARIMA_f should be slightly worse, since it exploits information about the type of the simulated processes, although to a lesser extent, since the values of p, d, q are not known a priori (but they are estimated during the training process). Similarly to the ARIMA_f and the auto_ARIMA_f methods, auto_ARFIMA is expected to exhibit the best performance in terms of RMSE when applied to the time series resulting from the simulation of ARFIMA processes. Finally, ARIMA_s and auto_ARIMA_s are expected to be best performing in capturing the variance exhibited by the simulated time series, while together with ETS_s are expected to not be amongst the most accurate. The six forecasting methods mentioned in the above lines play the role of benchmarks within our methodological approach, since they serve as a reference for the assessment of the remaining methods within the simulation experiments. Other benchmarks used herein are the simple methods. These two methods are amongst the most commonly used benchmarks in the forecasting field (Hyndman and Athanasopoulos 2018, Chapter 3.1). The above-outlined information is used in interpreting and discussing our results.

3.3 Results

3.3.1 Simulation experiments

This Section aims at providing a synopsis of the results of the simulation experiments. To support our key findings, here we present a small representative sample of the entire information. For the about 13 000 figures, conducted in the context of an exploratory visualization, as well as for the numerical summaries of the results in table form, the reader is referred to the fully reproducible reports, which are available together with their codes in Chapter's supplement. In the latter, we also enclose the report entitled "Selected figures for the qualitative comparison of the forecasting methods", which includes Figures S.1–S.24. These figures can support the main conclusions of this paper in a satisfactory manner.

In Figures 3.3–3.9, we present the side-by-side boxplots of the values of the forecast quality metrics computed within the SE_1a simulation experiment. These figures can provide a rough outline of the forecasting methods and the utility of the forecast quality metrics within this Chapter. By their examination, we observe that the ARIMA_f and auto_ARIMA_f benchmarks are the best performing with respect to type 1 accuracy, as assumed in Section 3.2.6, while BATS exhibits a very close to these methods performance, perhaps because it uses information from an ARMA model. We also note that the total of the ML methods except for NN_1 are competitive with BATS and with each other, while they are also better than the stochastic SES and Theta. The latter forecasting methods share a quite similar performance, a fact also applying to Naïve and RW. These simple benchmarks are better than NN_1 and the simulation models (ARIMA_s, auto_ARIMA_s, ETS_s), amongst which ETS_s produces forecasts with the most varying metric values and the worst median. Regarding the type 2 accuracy, all the methods seem to have rather equally good average-case performance, since the differences in the latter are small and not perceivable from these figures. However, the metric values computed for ETS_s are the most scattered with respect to each other, while the opposite applies to the metric values computed for ARIMA_f, auto_ARIMA_f, BATS and all the ML methods apart from NN_1. The metric values

computed for the remaining forecasting methods are scattered with respect to each other to an extent in between.

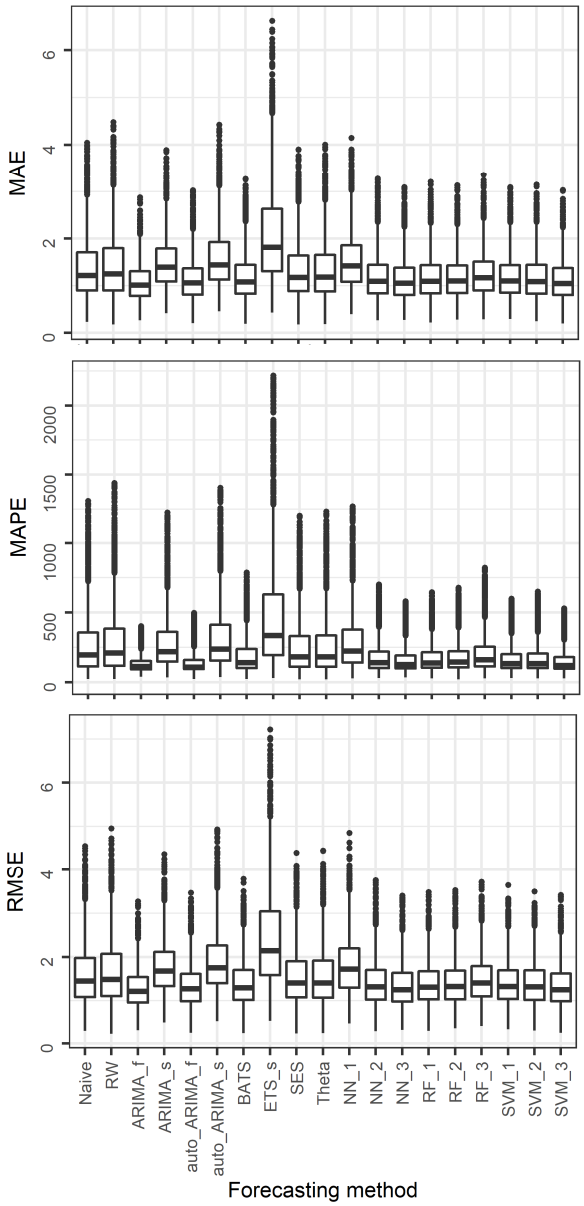


Figure 3.3. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the type 1 accuracy criterion within the SE_1a simulation experiment (part 1). The far outliers have been removed.

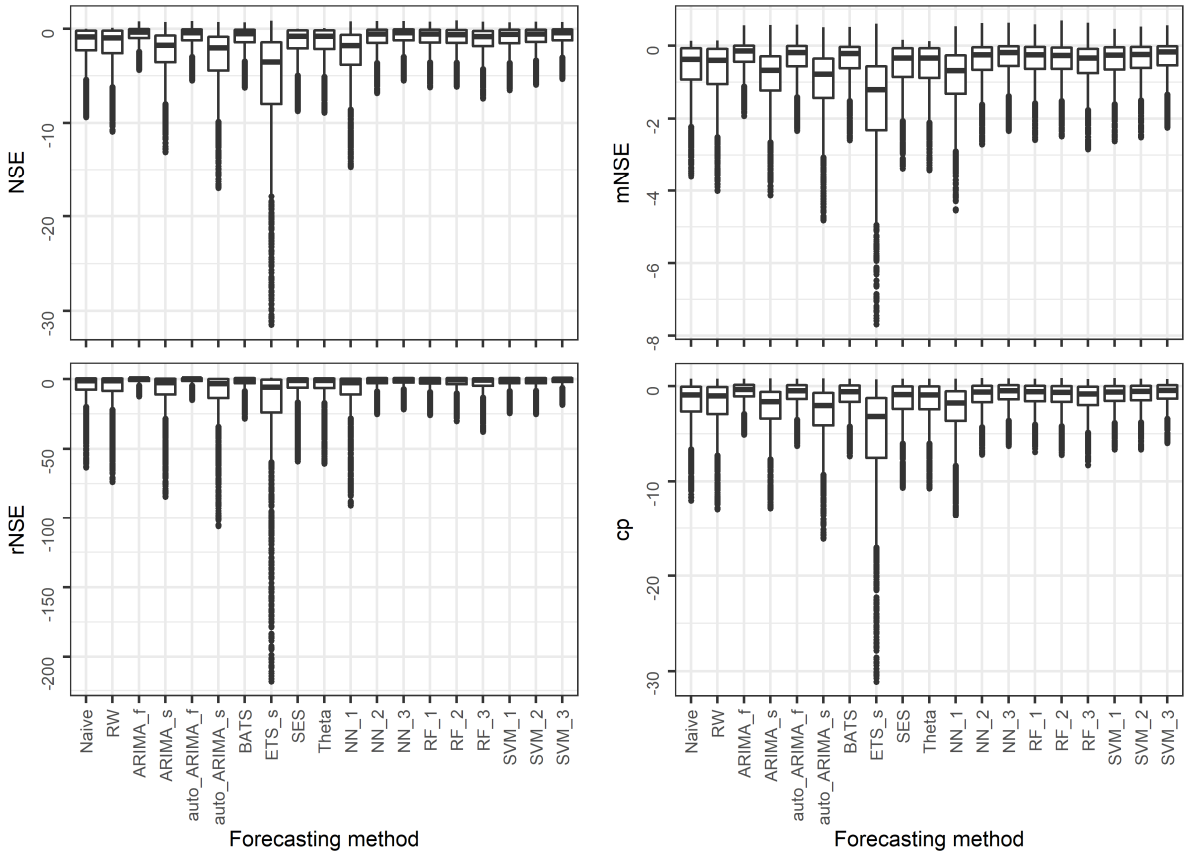


Figure 3.4. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the type 1 accuracy criterion within the SE_1a simulation experiment (part 2). The far outliers have been removed.

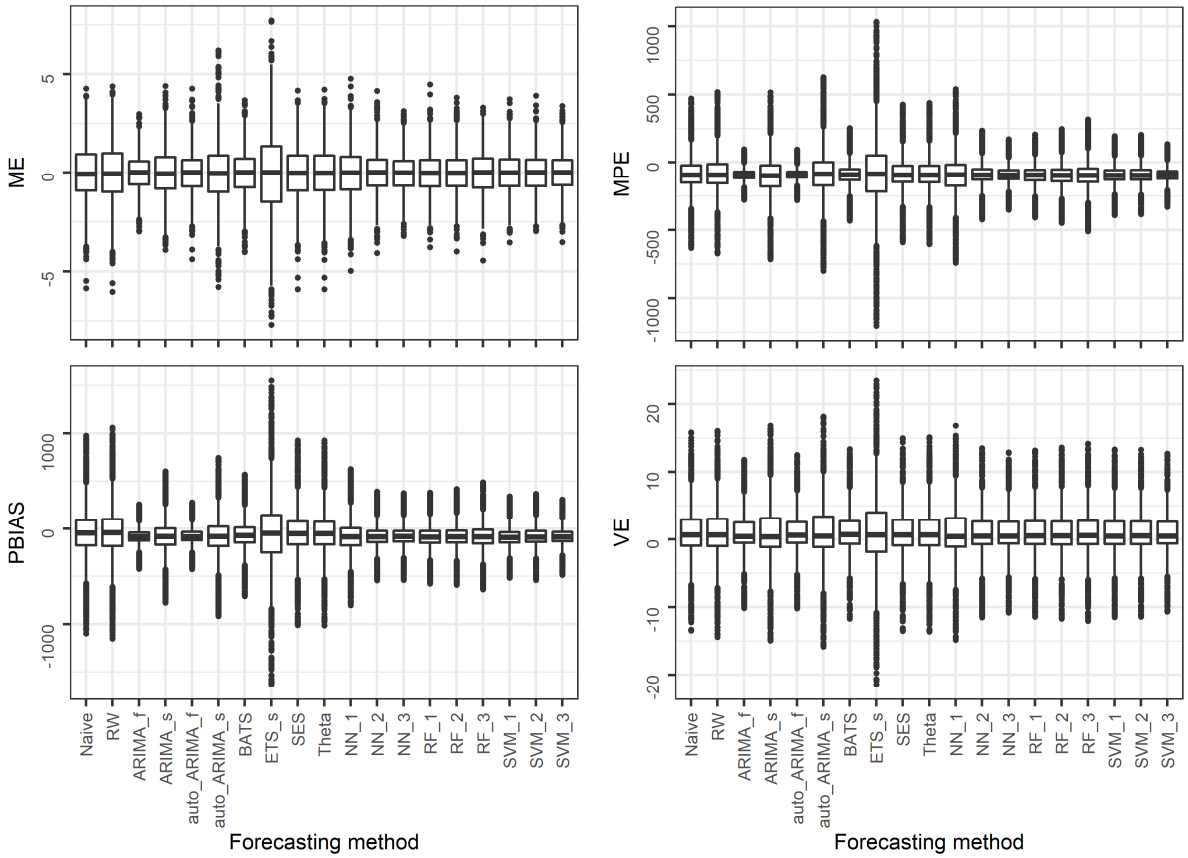


Figure 3.5. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the of type 2 accuracy criterion within the SE_1a simulation experiment. The far outliers have been removed.

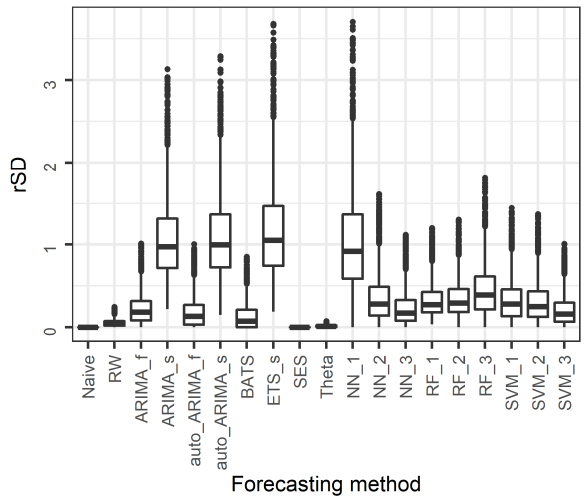


Figure 3.6. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the capture of the variance criterion within the SE_1a simulation experiment. The far outliers have been removed.

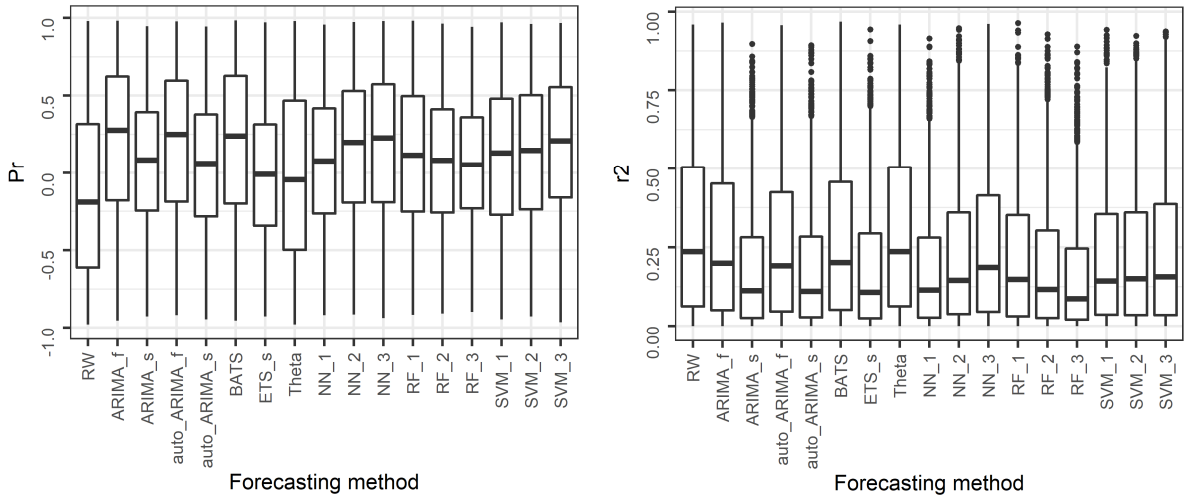


Figure 3.7. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the correlation criterion within the SE_1a simulation experiment. The Pr and r2 metrics are not defined for the forecasts produced by the Naïve and SES forecasting methods and, thus, the corresponding boxplots are not presented.

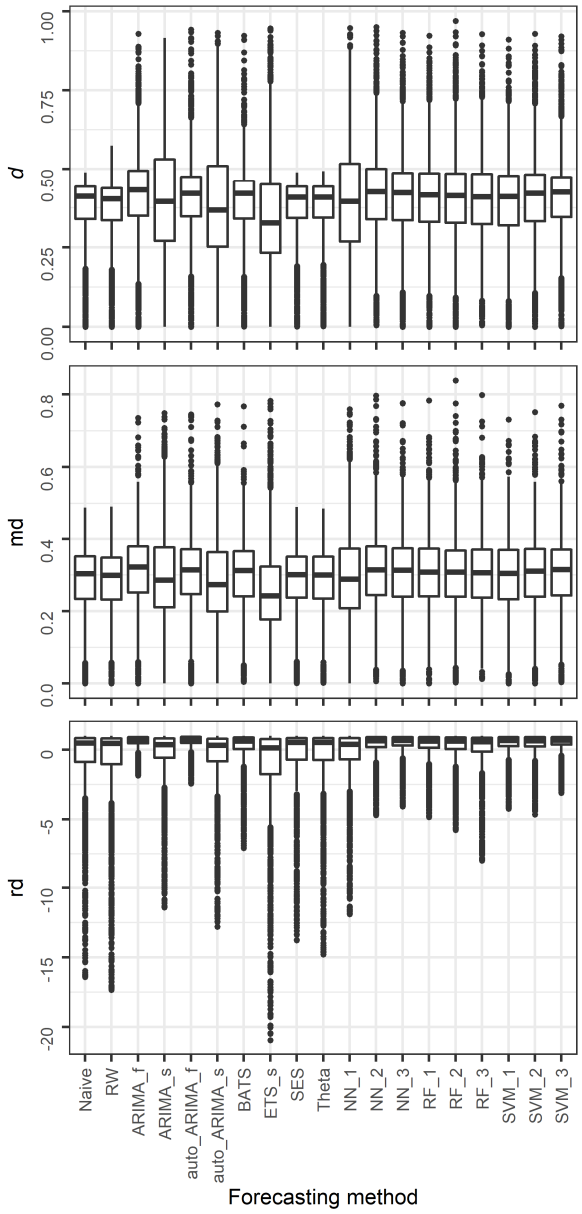


Figure 3.8. Side-by-side boxplots for the comparative co-assessment of the forecasting methods regarding their performance according to the type 1 accuracy and capture of the variance criteria within the SE_1a simulation experiment. The far outliers have been removed from the side-by-side boxplots of the rd values.

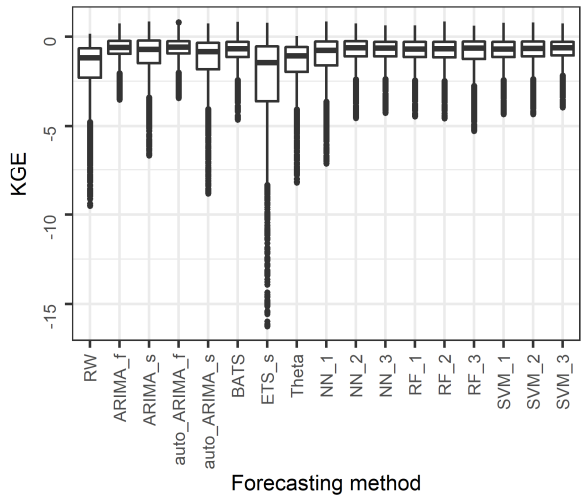


Figure 3.9. Side-by-side boxplots for the comparative co-assessment of the forecasting methods regarding their performance according to the type 2 accuracy, capture of the variance and correlation criteria within the SE_1a simulation experiment. The far outliers have been removed. The KGE metric is not defined for the forecasts produced by the Naïve and SES forecasting methods and, thus, the corresponding boxplots are not presented.

In terms of rSD, the image is mostly reversed compared to the one produced by the type 1 accuracy metrics. Naïve, RW, SES and Theta are clearly the worst, while the ML methods are more segregated. The average-case performance of NN_1, ARIMA_s, auto_ARIMA_s and ETS_s is good. Nevertheless, the rSD values for these four forecasting methods can vary significantly from the one forecasting attempt to the other, more than the rSD values computed for the remaining forecasting methods, a fact also applying to the rest of the forecast quality metrics. Regarding the average-case performance with respect to correlation, ARIMA_f, auto_ARIMA_f and BATS are the best, followed by NN_3. With respect to both type 1 accuracy and capture of the variance, ARIMA_f, auto_ARIMA_f, BATS and all the ML methods except for NN_1 are clearly better than the simple benchmarks and competitive with each other. SES and Theta, on the other hand, exhibit a very close performance to the one of Naïve and RW. Finally, in terms of KGE, the best performing methods are the same three stochastic and eight ML ones. NN_1, ARIMA_s and auto_ARIMA_s are better than Theta, which is competitive with RW. Overall, we observe that for the SE_1a simulation experiment the forecast quality metrics (even the corresponding to the same criterion) provide different aspects of the same information to an extent larger or smaller (as it is expected; see Section 3.2.4), while these 18 different aspects may also be conflicting to each other.

Subsequently, we state the main observations obtained from the total of the simulation experiments. To base these observations, in Figure 3.10 we present the heatmaps of the average-case performance of the forecasting methods within the SE_1a, SE_1b, SE_2a and SE_2b simulation experiments, while in Figures 3.11–3.13 we present the heatmaps formed using the medians of the total of the RMSE, rSD and *d* values respectively. In these figures the scaling is performed in the row direction and the darker the colour the better the forecasts. The conducted clustering analysis on the forecasting methods based on their performance is also presented. Some observations obtained from SE_1a apply to the rest of the simulation experiments as well. These are the following (see e.g., Figures 3.10–3.13): (a) forecasting methods from both the stochastic and ML categories are amongst the best and worst performing ones, (b) the metrics can provide significantly different, even conflicting, image regarding the performance of the forecasting methods, (c) the ARIMA_f, auto_ARIMA_f and auto_ARFIMA benchmarks are the best performing in terms of type 1 accuracy, while ETS_s, ARIMA_s and auto_ARIMA_s exhibit a good average-case performance in terms of rSD, (d) the image produced by rSD is mostly reversed with respect to the one produced by the type 1 accuracy metrics, i.e., methods that are well performing according to the latter criterion are bad performing with respect to the capture of the variance of the time series, (e) BATS is very close to ARIMA_f, auto_ARIMA_f and auto_ARFIMA, and (f) Naïve and RW,

as well as SES and Theta, exhibit similar performance to each other. Nevertheless, the Pr, r2 and KGE metrics are not defined for the forecasts produced by Naïve and SES. Finally, by the examination of the side-by-side boxplots produced for each and every of the simulation experiments we note that (g) ARIMA_s, auto_ARIMA_s, ETS_s and NN_1 seem to share a form of instability, i.e., their metric values vary more than the metric values of other forecasting methods. The latter concerns the results obtained from all the forecast quality metrics except for Pr and r2.

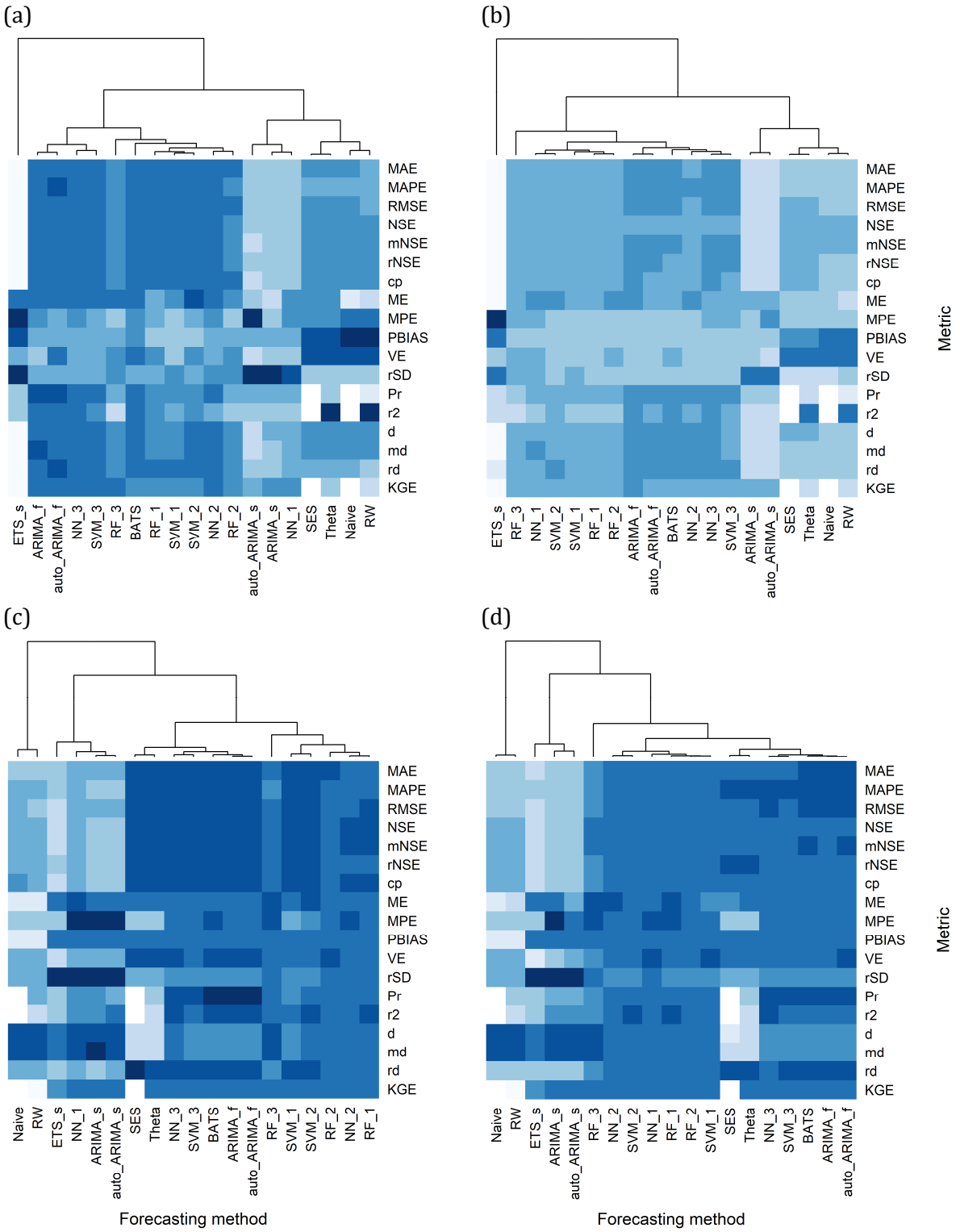


Figure 3.10. Heatmaps for the comparative assessment of the forecasting methods within the (a) SE_1a, (b) SE_1b, (c) SE_2a, (d) SE_2b simulation experiments according to the median values of the forecast quality metrics and the conditions listed on Table 3.6. The Pr, r2 and KGE metrics are not defined for the forecasts produced by the Naïve and SES forecasting methods. Their missing values are not taken into consideration during the comparative assessment and are imprinted with white colour.

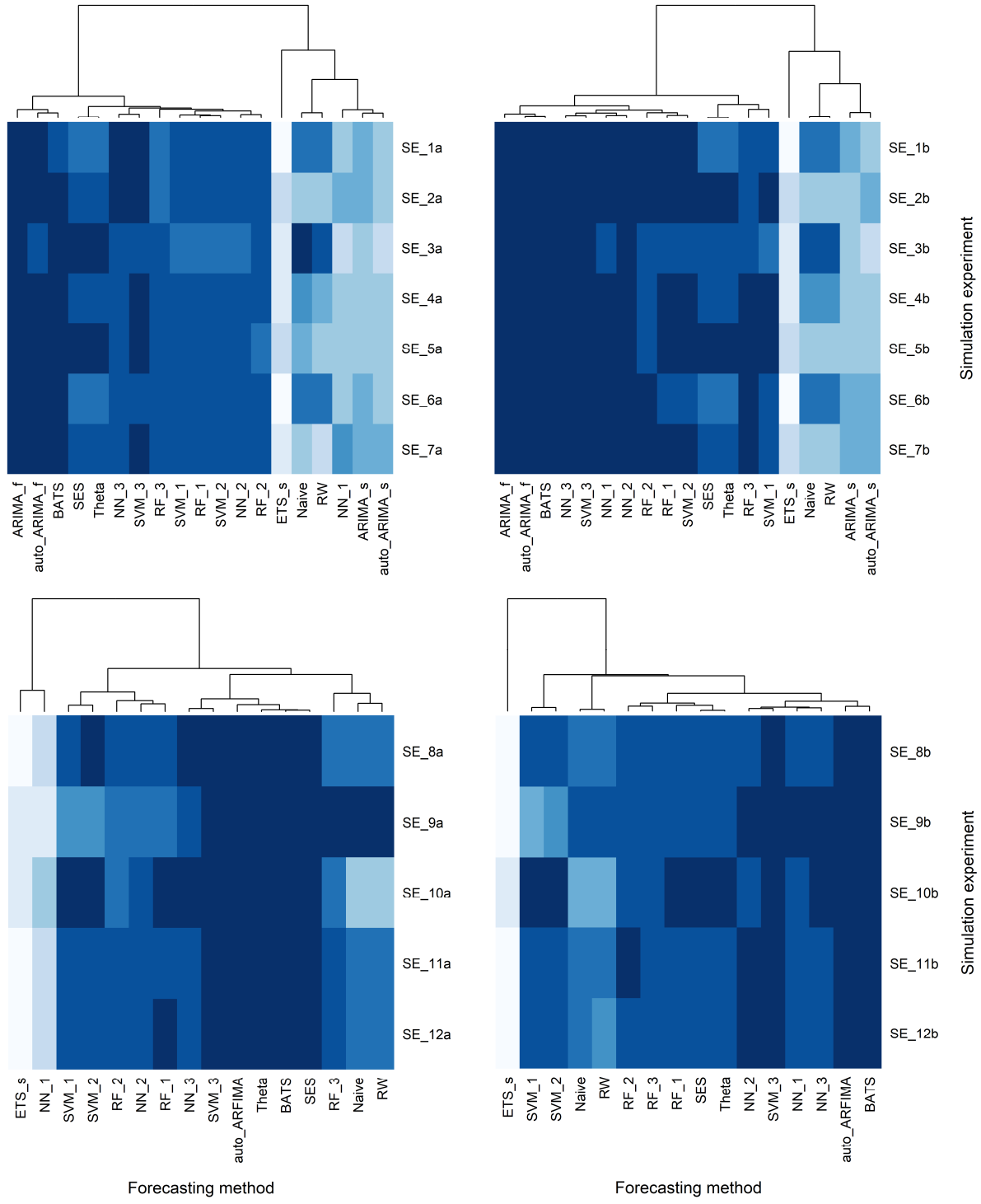


Figure 3.11. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the RMSE metric and the condition stated on Table 3.6.

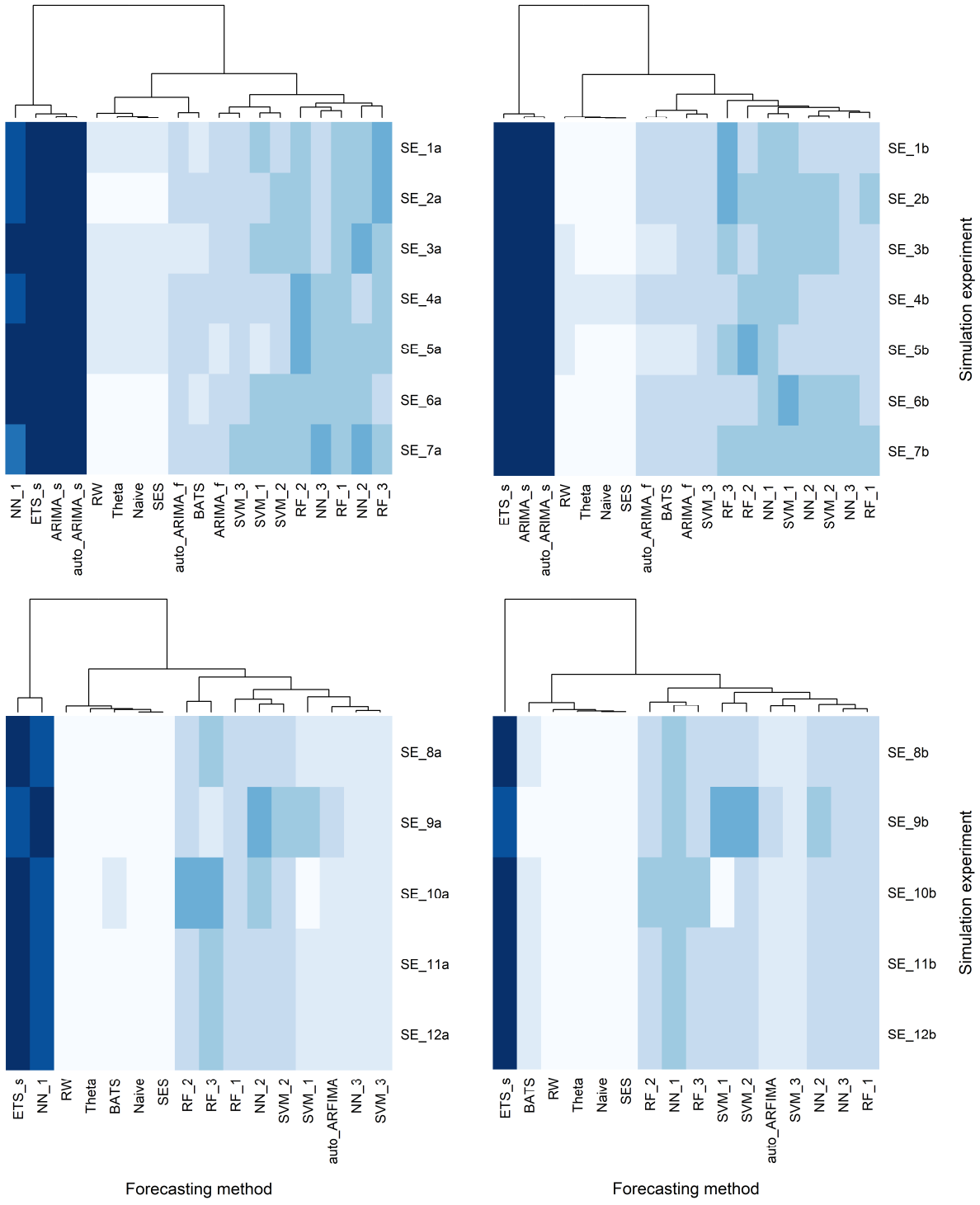


Figure 3.12. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the rSD metric and the condition stated on Table 3.6.

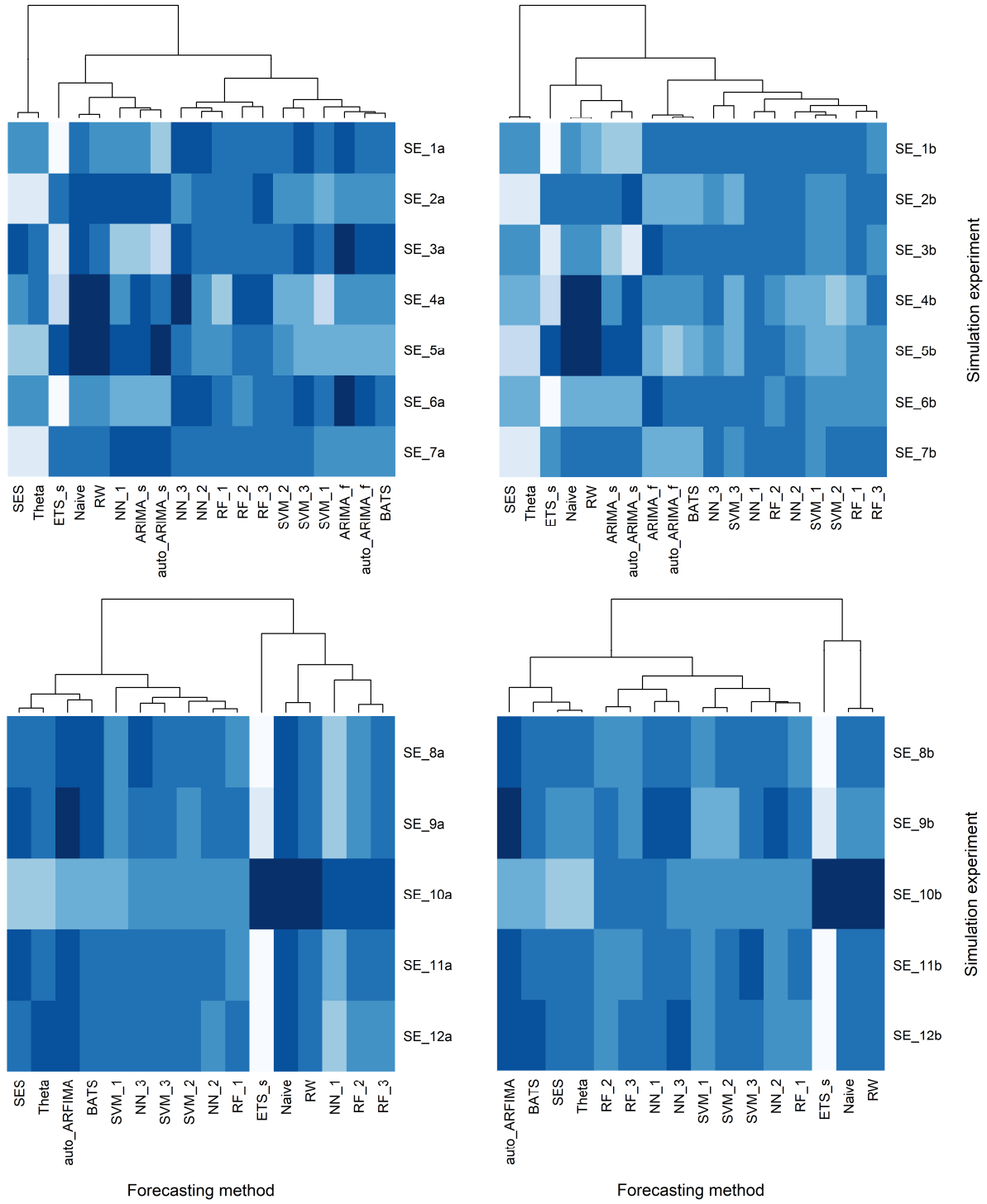


Figure 3.13. Heatmaps for the comparative assessment of the forecasting methods according to the median values of the *d* metric and the condition stated on Table 3.6.

By the examination of Figures 3.10–3.13 (or Figures S.1–S.24 of Chapter's supplement), we observe that the image provided by the metrics and the resulted grouping of the forecasting methods can also vary from the one simulation experiment to the other. Especially Figures 3.11–3.13 (or Figures S.7–S.24 of Chapter's supplement) allow us to easily perceive that the differences in the results of the various simulation experiments, also depicted in the grouping of the forecasting methods, are more related with the information provided by specific metrics and mostly concern specific forecasting methods. In fact, the heatmaps formed for the MAE, MAPE, RMSE, NSE, mNSE, rNSE, cp, rSD and KGE metrics are smoother than those formed for the

remaining forecast quality metrics. In particular, the pictures obtained from ME, MPE, VE, r2, *d* and md are the most dispersed. On the other hand, the Naïve, RW, ARIMA_s, auto_ARIMA_s, ETS_s, SES, Theta and NN_1 forecasting methods are more likely to have a varying performance (which results in varying grouping of forecasting methods). For example, we observe that Naïve and RW exhibit rather the best average-case performance in terms of *d* (see Figure 3.13) and md (see Figure S.22 in Chapter's supplement), while they have either bad, moderate or good average-case performance in terms of MAE, MAPE, PBIAS and VE depending on the simulation experiment (see Figures S.7, S.8, S.16 and S.17, respectively, in Chapter's supplement). The same applies to SES and Theta in terms of *d*, etc. We also note that forecasting methods resulting from the implementation of the same algorithm can exhibit a far distant or always close performance depending on the algorithm, as it is also perceivable by the examination of the resulted grouping of the forecasting methods. For instance, NN_1 and NN_2 (or NN_3) may differ with each other to a great extent, a fact also applying to ARIMA_s and ARIMA_f, but not to the RF and SVM forecasting methods. Interestingly, we observe that the training length largely affects the performance of NN_1 in a systematic way, while the performance of the remaining forecasting methods is less or even slightly affected. The latter effect depends on the forecasting method, as well as on the simulated process. In detail, the NN_1 forecasting method exhibits a bad performance with respect to type 1 accuracy (and a good one in terms of rSD; see Figure 3.12) within the simulation experiments using time series of 100 values, i.e., for 90-value training segments. On the contrary, its performance is good with respect to type 1 accuracy (and bad in terms of rSD) within the simulation experiments using time series of 300 values, i.e., for 290-value training segments. The latter observations concerning NN_1 might apply to a small extent to some of the remaining ML methods.

Next, we summarize some important information about the best performing forecasting methods in terms of type 1 accuracy, which has been identified as the criterion of focus herein. In terms of MAE (see Figure S.7 in Chapter's supplement) BATS is very close to the ARIMA_f, auto_ARIMA_f and auto_ARFIMA benchmarks, while SES, Theta and all the ML methods except for NN_1 have always a good or moderate performance. With respect to the MAPE metric (see Figure S.8 in Chapter's supplement) SVM_3 and BATS are mostly close to ARIMA_f, auto_ARIMA_f and auto_ARFIMA, and NN_2, NN_3, RF_1, RF_2, RF_3, SVM_1, SVM_2, SVM_3, SES and Theta are well performing for the greatest part of the simulation experiments. The same observations apply with respect to RMSE (see Figure 3.11). Nevertheless, for this metric NN_2 and NN_3 are rather very close to the good benchmarks as well. Regarding the NSE, mNSE, rNSE and cp values (see Figures S.10, S.11, S.12 and S.13, respectively, in Chapter's supplement), most of the stochastic and ML methods are competitive to each other and to the good benchmarks. The only ones that are not competitive are the simulation models, the simple benchmarks and NN_1 (the latter for 90-value training segments).

Finally, in Tables 3.7 and 3.8 we present the total computational time consumed by the forecasting methods within the simulation experiments. In summary, the following related observations are important. Naïve, SES, Theta, ARIMA_s, ARIMA_f, ETS_s and RW consume considerably less time than the remaining methods. Moreover, NN_3 is faster than auto_ARIMA_f, auto_ARIMA_s and auto_ARFIMA for the 90-value training segments, and faster than BATS for both lengths of training segments. The computational time consumed by RF_2 and RF_3 is mostly comparable with the computational time consumed by auto_ARIMA_f, auto_ARIMA_s and auto_ARFIMA for the 90-value training segments, while it is much higher for 290-value training segments. This computational time is also lower (higher) than the computational time reported for BATS for the former (latter) category of experiments. The three SVM methods are mostly faster than BATS, which in turn consumes less time than RF_1 for 290-value training segments. NN_1 and NN_2 are found to be the most computationally intensive. Overall, the ML methods collectively consume disproportionately more computational time than the stochastic ones.

Table 3.7. Total computational time (s) consumed by the forecasting methods within the simulation experiments (part 1). The numbers have been rounded up to the nearest integer. The computations have been performed in a regular home PC.

	Naïve	RW	ARIMA_f	ARIMA_s	auto_ARIMA_f	auto_ARIMA_s	BATS	ETS_s	SES	Theta	NN_1	NN_2	NN_3	RF_1	RF_2	RF_3	SVM_1	SVM_2	SVM_3
SE_1a	0	18	11	7	127	124	331	15	3	4	1301	827	90	343	178	141	312	215	187
SE_2a	0	19	13	10	173	171	1003	24	5	6	1679	1099	129	447	242	184	449	328	278
SE_3a	0	22	23	17	196	192	410	23	5	6	1797	1312	140	440	316	184	448	448	287
SE_4a	0	21	15	12	168	163	926	22	4	5	1597	946	189	466	186	223	445	266	309
SE_5a	0	24	17	12	186	180	885	24	5	6	1693	965	198	467	178	222	452	268	302
SE_6a	0	23	19	15	255	251	562	23	5	6	1748	1073	195	405	225	194	393	259	265
SE_7a	0	21	21	17	223	217	1381	21	5	6	1614	1127	213	433	249	209	397	323	297
SE_1b	0	18	15	12	148	146	1083	51	7	8	6364	3645	391	3061	1421	808	890	643	539
SE_2b	0	22	16	13	109	105	1222	56	8	9	6353	3726	421	3038	1466	802	892	650	531
SE_3b	0	25	37	30	161	155	579	51	8	8	6401	5349	543	2995	2414	808	894	801	529
SE_4b	0	24	21	16	129	124	1218	49	7	8	6282	2986	823	3148	766	1020	786	482	542
SE_5b	0	26	20	17	114	109	1159	53	8	9	6032	2829	817	3069	811	1098	895	547	620
SE_6b	0	26	30	24	184	180	1517	51	7	9	6352	4012	940	2952	1561	1124	882	674	625
SE_7b	0	25	28	22	126	123	1782	49	7	8	6555	4212	954	3062	1591	1089	834	630	583

Table 3.8. Total computational time (s) consumed by the forecasting methods within the simulation experiments (part 2). The numbers have been rounded up to the nearest integer. The computations have been performed in a regular home PC.

	Naïve	RW	auto_ARFIMA	BATS	ETS_s	SES	Theta	NN_1	NN_2	NN_3	RF_1	RF_2	RF_3	SVM_1	SVM_2	SVM_3
SE_8a	0	23	207	457	21	4	5	1614	1050	127	436	234	183	417	295	262
SE_9a	0	23	277	458	25	5	5	1908	1445	172	457	312	201	461	369	284
SE_10a	0	25	217	689	27	5	6	1681	964	127	479	176	199	432	255	265
SE_11a	0	18	178	402	19	4	5	1488	966	119	404	216	170	381	271	240
SE_12a	0	20	184	406	18	4	5	1496	970	117	406	218	169	383	272	227
SE_8b	0	24	199	743	44	6	7	6426	5111	654	2882	1999	872	752	667	524
SE_9b	0	26	242	902	56	6	9	6558	5395	525	2480	2083	665	716	625	417
SE_10b	0	23	196	860	61	9	10	6189	2600	564	2796	696	897	722	462	464
SE_11b	0	20	168	641	38	5	6	5602	4142	533	2480	1839	773	683	593	453
SE_12b	0	23	175	653	38	5	6	5614	4107	543	2483	1820	780	683	590	449

3.3.2 Real-world experiment

In full correspondence to the results of the simulation experiments, the results of the real-word experiment are presented in both quantitative and qualitative forms. In Figures 3.14–3.17, we present the side-by-side boxplots of the MAPE, NSE, cp, MPE, *d* and KGE values. Additionally, in Table 3.9 we present the median values of the dimensionless metrics, while in Figure 3.18 the average-case rankings of the forecasting methods. Here as well, we observe small differences between most of the methods, especially with respect to specific forecast quality metrics (e.g., MAPE, cp, MPE, *d*). For example, the median values of MAPE computed for auto_ARFIMA, BATS, SES, Theta, NN_3, RF_1, SVM_1, SVM_2 and SVM_3 are very close to each other. The same applies to the median values of NSE computed for the same methods, although the differences in the respective side-by-side boxplots seem to be larger in the latter case than in the former. Because of the small differences in the performance of the forecasting methods, the median metric values of Table 3.9 (e.g., the median MAPE values) may result to a different ranking of the forecasting methods than the average-case ranking presented in Figure 3.18.

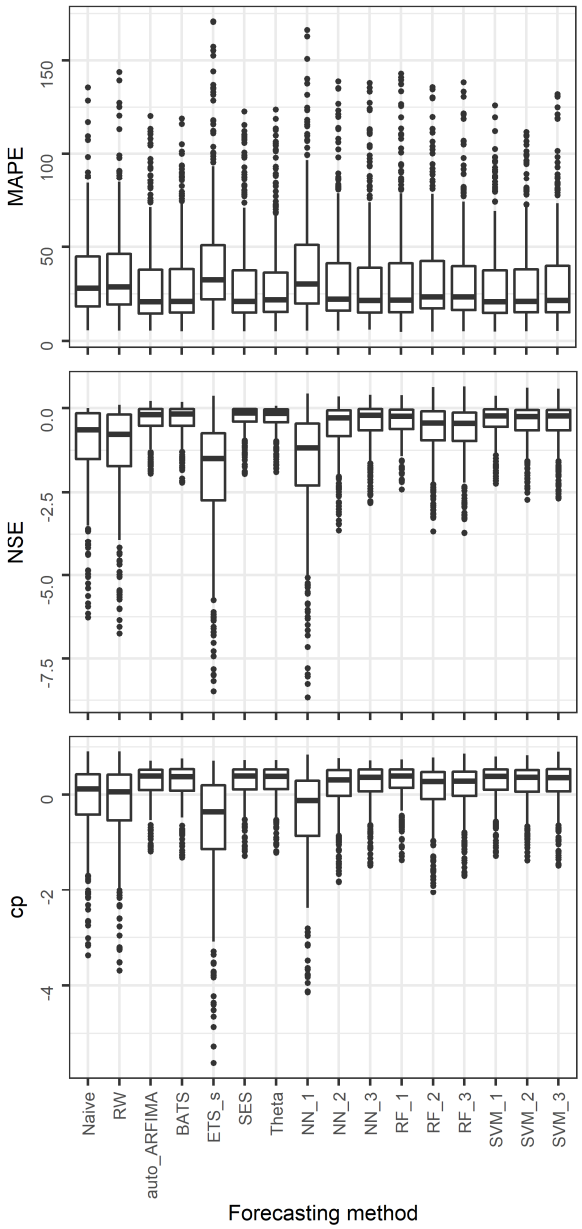


Figure 3.14. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the type 1 accuracy criterion within the real-word experiment. The far outliers have been removed.

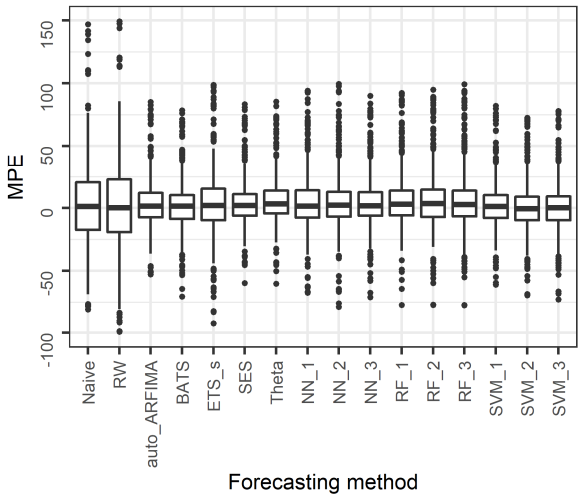


Figure 3.15. Side-by-side boxplots for the comparative assessment of the forecasting methods regarding their performance according to the type 2 accuracy criterion within the real-word experiment. The far outliers have been removed.

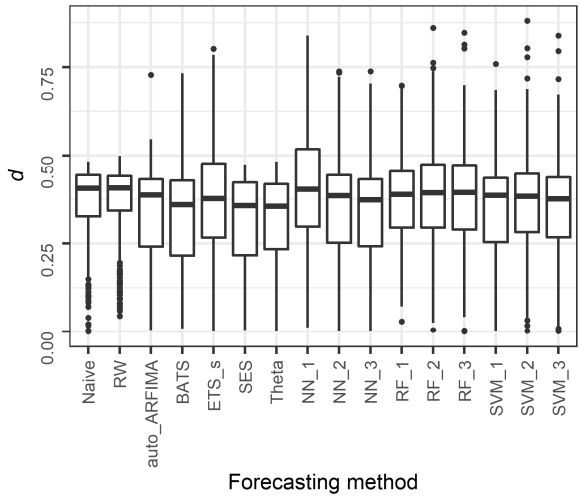


Figure 3.16. Side-by-side boxplots for the comparative co-assessment of the forecasting methods regarding their performance according to the type 1 accuracy and capture of the variance criteria within the real-word experiment.

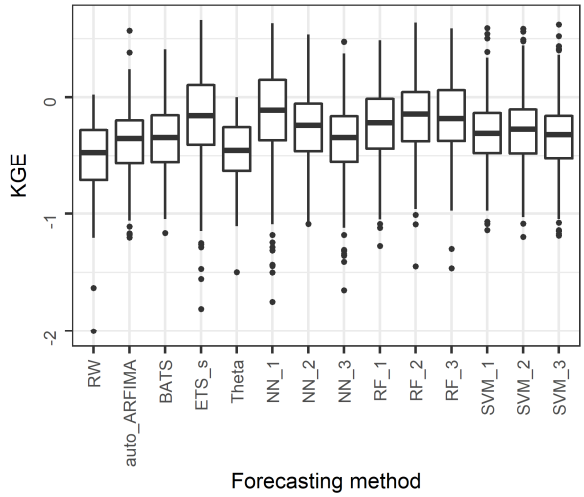


Figure 3.17. Side-by-side boxplots for the comparative co-assessment of the forecasting methods regarding their performance according to the type 2 accuracy, capture of the variance and correlation criteria within the real-word experiment. The far outliers have been removed. The KGE metric is not defined for the forecasts produced by the Naïve and SES forecasting methods and, thus, the corresponding boxplots are not presented.

Table 3.9. Median values of the dimensionless metrics computed within the real-word experiment.

	Naïve	RW	auto_ARFIMA	BATS	ETS_s	SES	Theta	NN_1	NN_2	NN_3	RF_1	RF_2	RF_3	SVM_1	SVM_2	SVM_3
MAPE	29.21	29.83	22.04	22.04	33.81	22.02	22.86	32.30	24.05	22.95	23.06	25.19	24.81	22.03	22.24	22.29
NSE	−0.72	−0.84	−0.20	−0.19	−1.57	−0.17	−0.18	−1.26	−0.13	−0.22	−0.25	−0.47	−0.46	−0.24	−0.26	−0.23
mNSE	−0.27	−0.31	−0.07	−0.07	−0.61	−0.06	−0.07	−0.51	−0.14	−0.09	−0.11	−0.20	−0.19	−0.09	-0.10	−0.10
rNSE	−0.81	−0.90	−0.35	−0.39	−2.24	−0.35	−0.45	−1.83	−0.59	−0.45	−0.46	−0.86	−0.78	−0.36	-0.40	−0.42
cp	0.09	0.03	0.39	0.38	-0.37	0.39	0.38	−0.16	0.30	0.36	0.37	0.27	0.25	0.38	0.35	0.34
MPE	2.83	1.47	2.99	2.20	3.29	3.32	5.07	2.94	4.61	3.36	4.31	4.62	3.96	3.00	1.17	1.49
PBIAS	−6.34	−6.34	−3.14	−4.25	−2.72	−2.90	−1.56	−3.05	−2.09	−2.41	−1.19	−1.80	−2.59	−4.50	−5.84	−4.60
VE	0.71	0.71	0.78	0.78	0.67	0.78	0.78	0.69	0.76	0.78	0.78	0.75	0.76	0.78	0.78	0.78
rSD	0.00	0.03	0.05	0.00	1.02	0.00	0.01	0.94	0.21	0.05	0.24	0.42	0.40	0.00	0.12	0.07
Pr	–	−0.05	0.06	0.04	0.00	–	−0.04	0.08	0.08	0.02	0.08	0.08	0.04	0.08	0.07	0.05
r2	–	0.07	0.06	0.05	0.06	–	0.07	0.05	0.06	0.06	0.06	0.07	0.06	0.06	0.06	0.06
d	0.41	0.41	0.39	0.36	0.38	0.36	0.36	0.40	0.39	0.37	0.39	0.39	0.39	0.39	0.38	0.38
md	0.31	0.31	0.28	0.28	0.28	0.27	0.18	0.30	0.29	0.28	0.28	0.29	0.30	0.29	0.30	0.29
rd	0.29	0.30	0.25	0.26	0.30	0.22	0.18	0.33	0.28	0.22	0.30	0.30	0.31	0.29	0.34	0.30
KGE	–	−0.47	−0.35	-0.34	-0.17	–	−0.46	−0.12	−0.24	−0.35	−0.22	−0.15	−0.19	−0.31	−0.27	−0.32

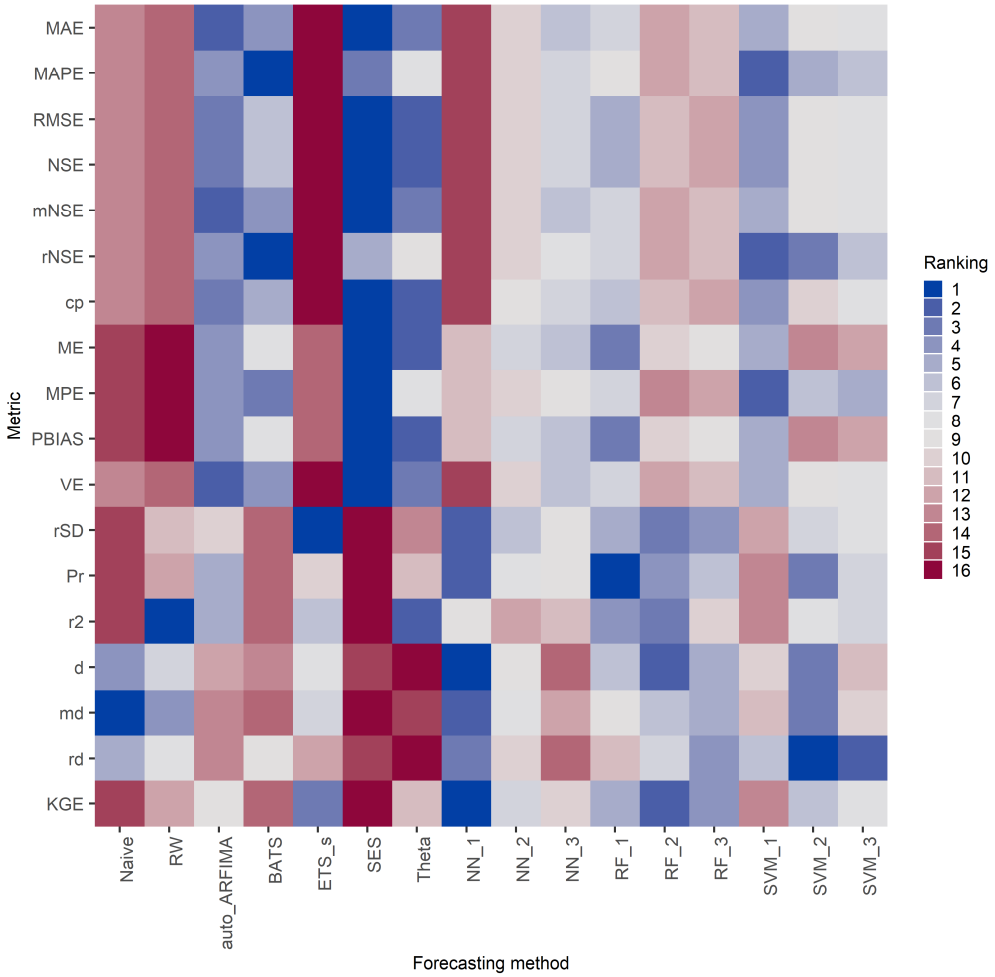


Figure 3.18. Heatmap for the comparative assessment of the forecasting methods within the real-world experiment according to their average-case rankings. The latter are based on the values of the forecast quality metrics and the conditions listed on Table 3.6. The Naïve and SES forecasting methods are ranked 15th and 16th according to rSD, Pr, r2 and KGE. Their rSD values are 0, while the Pr, r2 and KGE metrics are not defined for their forecasts.

Furthermore, while the average-case rankings with respect to accuracy mostly favour stochastic methods (SES, Theta, auto_ARFIMA and BATS), SVM_1 is also ranked amongst the best performing methods. In more detail, SES is ranked first according to MAE, RMSE, NSE, mNSE, cp, ME, MPE, PBIAS and VE, but it is worse than SVM_1, and SVM_1 and SVM_2 according to MAPE and rNSE respectively. According to the latter metrics, the best performing method is BATS. This method has a rather moderate overall performance in terms of accuracy. The less accurate methods, on the other hand, are Naïve, RW, ETS_s and NN_1, as it is expected from the simulation experiments. With respect to the remaining criteria, SES is clearly the worst performing method, while Theta, Naïve, BATS, SVM_1, NN_3 and auto_ARFIMA are also ranked behind the remaining ML methods, amongst which NN_1 is mostly ranked first. Finally, in terms of computational requirements within this real-world experiment the methods could be ranked from best (1st) to worst (16th) as follows: Naïve, SES, Theta, RW, ETS_s, NN_3, auto_ARFIMA, RF_3, RF_2, SVM_3, SVM_2, SVM_1, RF_1, NN_2, BATS and NN_1 (see also Table 3.10).

Table 3.10. Total computational time (s) consumed by the forecasting methods within the real-world experiment. The numbers have been rounded up to the nearest integer. The computations have been performed in a regular home PC.

Naïve	RW	auto_ARFIMA	BATS	ETS_s	SES	Theta	NN_1	NN_2	NN_3	RF_1	RF_2	RF_3	SVM_1	SVM_2	SVM_3
0	3	26	200	6	1	1	283	174	20	70	32	31	62	42	40

3.4 Discussion

3.4.1 Contribution in hydrology and beyond

The present Chapter contributes by developing a detailed framework for assessing forecasting techniques in hydrology. Furthermore, its findings can provide new insights into the nature of short hydrological time series forecasting at large time scales, while they concern all natural processes that could be modelled by linear stationary processes. A first view of the results suggests that the differences in the forecasting performance of the methods are mostly small (insignificant for hydrometeorological applications; see also the experiments of Chapter 5 herein), while the stochastic and ML methods can share a quite similar forecasting performance when implemented to hydrological time series of small length and small temporal resolution (e.g., annual or monthly). In fact, methods from both these categories are found to perform better or worse mainly depending on the forecast quality metric, but on the experiment as well. Regarding the type 1 accuracy, in the simulation experiments BATS is always close to the ARIMA_f, auto_ARIMA_f and auto_ARFIMA benchmarks, probably because it uses information from an ARMA model, while most of the ML methods (e.g., NN_3 and SVM_3) are amongst the best performing and often better than SES and Theta. Nevertheless, in the real-world experiment SES is mostly ranked first, followed by auto_ARFIMA, BATS, SVM_1 and Theta, while NN_3, RF_1, SVM_2, and SVM_3 are also close to the latter methods. A possible interpretation of this outcome is that for a different sample of river discharge time series, the average-case rankings would differ as well, and that there might be no particular reason to choose some methods over others for this specific process. Given the claims that in linear situations (e.g., the simulation experiments of this Chapter) the ML methods are more likely to be inferior to the stochastic ones, while in non-linear situations, as it is usually asserted to apply to river discharge processes, the ML methods are more likely to outperform, the algorithmically obtained results of the present Chapter are even more interesting. Noteworthy is also the fact that our results differ from the results of Makridakis et al. (2018), which favour the stochastic methods, probably due to the different experimental setting adopted therein (determined by the required degree of forecast accuracy, the lengths of the examined time series, the selected algorithms for performing multi-step ahead forecasting, the forecast quality metrics used for evaluating the methods and the optimization procedures of the ML methods, among others).

In this view, we would like to emphasize that the ML algorithms are accurate enough. Yet, they have the worth-mentioning particularity that their forecasting performance might be largely affected by the number of the utilized lagged variables. This number is directly related to the length of the segment used for model fitting (Tyralis and Papacharalampous 2017). Specifically, a significant decrease of this length may deteriorate the forecasting performance of a ML algorithm, as largely perceivable through the examination of the results obtained for the NN methods of the present Chapter. In detail, for the simulation experiments using 90-value training segments, NN_1 exhibits a bad performance in terms of type 1 accuracy, a fact not applying to NN_2 and NN_3 that use less and very few lagged variables respectively. On the contrary, for the simulation experiments using 290-value training segments, NN_1 is amongst the most accurate methods. The same number of lagged variables is used by RF_1 and SVM_1. Nevertheless, the performance of the herein implemented RF and SVM algorithms seems to be less affected by the number of lagged

variables than the NN algorithm. These large-scale results on time lag (or lagged variable) selection could be encountered as contributed information to the subject.

Another particularity of the ML methods is related to their computational requirements, which seem to considerably increase with increasing length of the training segment. In fact, for our regular home PC the computational time consumed by the NN and SVM methods is found to be approximately four to eight times higher for 290-value training segments than for 90-value training segments. The respective difference in computational time is smaller for the SVM methods. The number of lagged variables seems to also affect the computational requirements. Specifically, the computational time increases when moving from the third to the first time lag selection procedures of Table 3.5, i.e., from less to more lagged variables, indicating increasing computational requirements (although the length of the lagged time series decreases), with this increase to be higher for the NN methods. Overall, the computational time collectively consumed by the herein implemented ML methods is considerably higher than the respective time measured for the stochastic methods. Nonetheless, it is also shown that there are computational intensive stochastic methods (mainly BATS), as well as ML methods with lower or comparable computational requirements with stochastic methods (e.g., NN_3, RF_3).

While there are forecasting methods regularly better or worse than others with respect to specific criteria, this does not apply to all the forecasting methods neither to all the criteria. For example, we observe that Theta can exhibit good, moderate or bad average-case performance in terms of specific forecast quality metrics depending on the simulation experiment. Furthermore, sophisticated forecasting methods (such us the above mentioned ones) do not necessarily (but mostly) provide better forecasts than the simple Naïve and RW, as also shown in previous studies (e.g., Makridakis and Hibon 2000; Cheng et al. 2017). These two methods perform almost identically in the experiments of the present Chapter, but not for longer forecast horizons (see Papacharalampous et al. 2018a; Chapter 5 herein). Another pair of similarly performing forecasting methods is SES and Theta. This latter outcome is consistent with Hyndman and Billah (2003).

In general, we cannot decide on a universally best or worst forecasting method (stochastic or ML), neither we can rank the forecasting methods based on the results of the simulation experiments. Even the relative metrics, i.e., the corresponding to the same criterion (see Table 3.6), provide measurements which lead us to different aspects of the same information to an extent larger or smaller depending on the pair of forecast quality metrics considered. Some of these 18 different aspects are also conflicting to each other. Any ranking of the forecasting methods would require the a priori selection of an experiment and a criterion of interest, as well as the application of a simplification procedure (e.g., use of the median values of the selected metric) and, thus, would not be general. However, the grouping of the forecasting methods is possible, though only to some extent. This grouping could be based on the similar or contrasting performance of the forecasting methods with respect to the various metrics. For example, the simulation models (ARIMA_s, auto_ARIMA_s and ETS_s) exhibit the best average-case performance with respect to the capture of the variance, while they are clearly the worst performing in terms of type 1 accuracy. This happens, since these two criteria are contradictory. For instance, the optimum forecast for an ARFIMA model is obtained when the innovations are set to zero.

Our contribution in the field of hydrology also includes the implementation of several forecasting models barely used in hydrometeorological concepts, but commonly used in the forecasting field (RW, BATS, ETS, SES and Theta) or for regression purposes (RF). This innovation holds, especially if we could exclude from the hydrological literature Chapters 2, 5 and 6 of this thesis, as well as their large-scale companion works by Tyralis and Papacharalampous (2017), and Papacharalampous et al. (2018c), while its practical value is indisputable. One could claim that there may be an undiscovered forecasting method (stochastic or ML), which will be better than the existing ones. As regards the *"myth of the best method"* the reader is referred to Hong and Fan (2016), who mention that the original techniques are countable and have been exhausted,

while the hybrid techniques, i.e., combinations of original techniques, cannot further improve the forecasting performance.

Another important contribution of the present work is related to the so-called *"no free lunch theorem"* by Wolpert (1996). According to this theorem, in the space of all possible problem instances, there is not a model that will always perform better than the other models in the absence of significant information for the problem at hand. The empirical work presented in this Chapter shows that even in the finite space of simple (simulated) and real-world time series examined herein there is not an optimal forecasting solution. Finding the best algorithm mostly depends on our knowledge of the system. For example, using ARFIMA models for forecasting the ARFIMA simulated time series is obviously the best choice, due to the prior known information about the system. The other methods are equivalent in performance since they cannot incorporate this knowledge. In the specific class of hydrological process forecasting finding information about the examined system could be possible, for example, with the application of principles of physics, such as the maximum entropy principle, incorporation of information from deterministic models (see e.g., Tyralis and Koutsoyiannis 2017), understanding the processes from a chaotic perspective (see e.g., Sivakumar 2004) and other approaches. Obviously, the knowledge of the system is not simply equivalent to the knowledge of its statistical properties (e.g., the mean, variance, ACF), but should be deeper. Therefore, the frequently met in the hydrological literature blind use of forecasting methods is not suggested.

Additionally, it seems that major advancements in the time series forecasting performance of all methods can be achieved by incorporating appropriate exogenous variables in the model, while the potential for improving their performance in univariate time series forecasting seems limited. The latter in our opinion is also due to the nature of the problem, which is simple. Therefore, methods that are more complicated will not necessarily yield better results. A relevant example is, for instance, the difference in the games of tic-tac-toe (see Figure 3.19) and Go (see Figure 3.20). The former game is simple and can be solved by simple algorithms; therefore, the choice of a complex method is not necessary. On the other hand, the best performance on the more complex game of Go was achieved by the use of complicated machine learning algorithms (see Silver et al. 2016).

Figure 3.19. The "tic-tac-toe" game. Source: https://en.wikipedia.org/wiki/Tic-tac-toe.

Figure 3.20. The "Go" game. Source: https://www.latimes.com/entertainment/movies/la-et-mn-capsule-alpha-go-review-20171026-story.html.

Regarding the extent to which the conclusions could be generalizable for the forecasting of short hydrological time series at large time scales, we note that the stationarity assumption and the reasoning of its appropriateness for the modelling of geophysical properties, documented in Koutsoyiannis and Montanari (2015), is consistent with the no free lunch theorem. In particular, if we cannot explain the behaviour of a geophysical process based on a deterministic mechanism, then the most appropriate models are stationary. Even in cases of deterministic systems, stochastic approaches are appropriate (Koutsoyiannis 2010). This is a frequently met case in the modelling of geophysical processes (i.e., there is not an adequate explanation for the behaviour of the geophysical process), proving that our conclusions could be generalizable.

3.4.2 On the methodological approach

The above section highlights the efficiency of our methodological approach in producing large-scale and representative for the field of hydrology results. Moreover, the real-world experiment particularly accounts for the case of river discharge forecasting. Someone who examines both the results of the simulation experiments and the real-world experiment has a more complete picture of the underlying phenomena than whom considering only the results of the simulation experiments. On the other hand, the use of simulated processes combined with benchmarking has proved pivotal in achieving our aim under the linearity and stationarity assumptions. Additionally, the use of an adequate number of forecasting methods and forecast quality metrics in the present Chapter is also of crucial importance. Using fewer forecasting methods and fewer forecast quality metrics would have led to a very different overall picture, particularly if those fewer metrics corresponded to fewer criteria. Besides, the comparison is rather the only available research method for any evaluation and, consequently, the larger its scale the more generalized the derived results. For this specific reason, the novel (mainly for hydrology) methodological approach of the present Chapter is considered appropriate for the assessment of forecasting methods in hydrology. Furthermore, the qualitative form of the results facilitates their handy examination and, thus, eases the delivery of the large-scale findings. In fact, our methodology enables the assessment of the failure risk or, alternatively worded, the available opportunities for success that accompany the use of a specific forecasting method to a significant extent, while it also leads to the recognition of several advantages/disadvantages characterizing the latter. This knowledge is fundamental to the forecasters and the users of the forecasts, since a specific forecasting method can be both useful and useless, depending on the forecasting task.

3.5 Conclusions

We conduct an extensive comparison between several stochastic and machine learning methods for the multi-step ahead forecasting of hydrological processes by performing large-scale computational experiments based on simulations under the linearity and stationarity

assumptions. The implemented stochastic methods include simple models, models from the frequently used families of autoregressive moving average and autoregressive fractionally integrated moving average, as well as innovations state space and exponential smoothing models, while the machine learning ones are neural networks, random forests and support vector machines. The aim is to provide large-scale results, while the respective comparisons in the literature are usually based on case studies. We also run a real-world experiment on the largest river discharge dataset ever used for forecasting purposes within a framework that is purely statistical. Despite this specific focus, the results concern all natural processes in large time scales (e.g., annual or monthly) that could be modelled by stationary processes. The findings suggest that stochastic and machine learning methods do not differ dramatically. In fact, methods from both these categories are found to be equally useful in univariate short time series forecasting at large time scales. This is particularly important, because it reveals that the forecast quality is subjected to limitations. The latter are imposed by the nature of the examined problem and manifest themselves in the computed forecast quality metric values. We have empirically proved that these values do not favour any specific forecasting method, stochastic or machine learning, in a long run. In fact, the results are consistent with the no free lunch theorem, albeit the theorem refers to an infinite space of problem instances, while here we examine a finite space of problems. The empirical investigation shows that in the given finite space, formed by simulated and annual river discharge time series, the no free lunch theorem is still satisfied.

4. One-step ahead forecasting of geophysical processes within a purely statistical framework

The simplest way to forecast geophysical processes is to implement stochastic or machine learning models within a purely statistical framework. These models are in general fast-implemented, in contrast to the computationally intensive global circulation models, which constitute the most frequently used alternative for precipitation and temperature forecasting. For their simplicity and easy applicability, the former have been proposed as benchmarks for the latter by forecasting scientists. In this Chapter, we assess the one-step ahead forecasting performance of 20 univariate time series forecasting methods, when applied to a large number of geophysical and simulated time series of 91 values. We use two real-world annual datasets, a dataset composed by 112 time series of precipitation and another composed by 185 time series of temperature, as well as their respective standardized datasets, to conduct several real-world experiments. We further conduct large-scale experiments using 12 simulated datasets. These datasets contain 24 000 time series in total, which are simulated using stochastic models from the families of AutoRegressive Moving Average and AutoRegressive Fractionally Integrated Moving Average. We use the first 50, 60, 70, 80 and 90 data points for model-fitting and model-validation and make predictions corresponding to the 51st, 61st, 71st, 81st and 91st respectively. The total number of forecasts produced herein is 2 177 520, among which 47 520 are obtained using the real-world datasets. The assessment is based on eight error metrics and accuracy statistics. The simulation experiments reveal the most and least accurate methods for long-term forecasting applications, also suggesting that the simple methods may be competitive in specific cases. Regarding the results of the real-world experiments using the original (standardized) time series, the minimum and maximum medians of the absolute errors are found to be 68 mm (0.55) and 189 mm (1.42) respectively for precipitation, and 0.23 °C (0.33) and 1.10 °C (1.46) respectively for temperature. Since there is an absence of relevant information in the literature, the numerical results obtained using the standardized real-world datasets could be used as rough benchmarks for the one-step ahead predictability of annual precipitation and temperature.

4.1 Introduction

Forecasting geophysical processes in various time scales and horizons is useful in technological applications (e.g., Giunta et al. 2015), but a difficult task as well. Precipitation and temperature forecasting is mostly based on deterministic models as the Global Circulation Models (GCMs), which simulate the Earth's atmosphere using numerical equations; therefore, deviating from traditional time series forecasting. This particular deviation has been questioned by forecasting scientists (Green and Armstrong 2007; Green et al. 2009; Fildes and Kourentzes 2011; see also the comments in Keenlyside 2011; McSharry 2011). Traditional time series forecasting can be performed using several classes of models, as reviewed in De Gooijer and Hyndman (2006), while the two major classes are stochastic and machine learning. These models are in general fast-implemented in contrast to their computationally intensive alternative in precipitation and temperature forecasting, i.e., the GCMs. For their simplicity and easy applicability, the former have been proposed as benchmarks for the latter by Green et al. (2009).

Recognizing the necessity of introducing traditional forecasting methods in temperature and precipitation forecasting, Armstrong and Fildes (2006) have recommended a relevant issue in one of the Journals specialized in forecasting. Since then and despite the fact that considerable parts of books in hydrology are devoted to such methods (Sivakumar 2017, pp. 63–145; Remesan and Mathew 2015, pp. 71–110), there has not been a systematic approach to the subject. However, studies adopting statistical forecasting approaches in geoscience are sporadically published in a variety of Journals. Within a statistical framework, Tyralis and Koutsoyiannis (2014, 2017) use Bayesian techniques for probabilistic climate forecasts, under the established assumption of long-range dependence of the observed time-series. In the latter study, information from GCMs is used to improve the performance of the time series forecasting methods. Moreover, Table 4.1 presents some examples of studies using univariate time series forecasting approaches that do not utilize

exogenous predictor variables to forecast precipitation or temperature variables, and streamflow or river discharge variables. The former can be considered as climatic or meteorological variables depending on the time scale of interest, while the latter can be considered as the results of precipitation (and other) variables and are more frequently modelled by describing this dependence using either deterministic or statistical methods. Such statistical approaches to modelling hydrological variables can be found in Chen et al. (2015), Gholami et al. (2015), and Taormina and Chau (2015).

Table 4.1. Methodological information on case studies focusing on hydrometeorological time series forecasting within a purely statistical framework (see also Table 3.1).

S/n	Study	Geophysical process	Number of time series	Forecast time scale	Forecast horizon (steps ahead)	Univariate time series forecasting methods
1	Hong (2008)	Precipitation	9	Hourly	1	○ Support vector machines ○ Hybrid model exploiting recurrent neural networks and support vector machines
2	Chau and Wu (2010)		2	Daily	1, 2, 3	○ Neural networks ○ Hybrid model exploited neural networks and support vector machines
3	Htike and Khalifa (2010)		1	Monthly, biannually, quarterly, yearly	1	○ Neural networks
4	Wu et al. (2010)		4	Monthly, daily	1, 2, 3	○ Linear regression ○ k-nearest neighbours ○ Neural networks ○ Hybrid model exploititng neural networks
5	Narayanan et al. (2013)		6	Yearly	21 × 3 (months)	○ AutoRegressive Integrated Moving Average (ARIMA)
6	Wang et al. (2013)		1	Monthly	12	○ Seasonal AutoRegressive Integrated Moving Average (SARIMA)
7	Babu and Reddy (2012)	Temperature	1	Yearly	10	○ ARIMA ○ Wavelet based ARIMA
8	Chawsheen and Broom (2017)		1	Monthly	121	○ SARIMA
9	Lambrakis et al. (2000)	Streamflow or river discharge	1	Daily	1	○ Farmer's model ○ Neural networks
10	Ballini et al. (2001)		1	Monthly	1, 3, 6, 12	○ AutoRegressive Moving Average (ARMA) ○ Neural networks ○ Neurofuzzy networks
11	Yu et al. (2004)		2	Daily	1	○ Support vector machine coupled with an evolutionary algorithm ○ Standard chaos technique ○ Naïve ○ Inverse approach ○ ARIMA
12	Komorník et al. (2006)		7	Monthly	1, 3, 6, 12	○ Threshold AutoRegression (AR) with aggregation operators ○ Logistic smooth transition AR ○ Self-exciting threshold AR ○ Naïve
13	Yu and Liong (2007)		2	Daily	1	○ Support vector machine coupled with decomposition ○ Standard chaos technique ○ Naïve ○ Inverse approach ○ ARIMA
14	Koutsoyiannis et al. (2008)		1 × 12 (months)	Yearly	1	○ Stochastic ○ Analogue method (k-nearest neighbours) ○ Neural networks
15	Wang et al. (2015)		3	Yearly	12	○ SARIMA

In this Chapter, we examine the fundamental problem of one-step ahead forecasting, also complementing the results of the four above-mentioned studies. In more detail, we expand the work presented in Chapter 3 of this thesis by exploring the one-step ahead forecasting properties of its methods, when applied to geophysical time series. Emphasis is put on the examination of two real-world datasets, a precipitation dataset and a temperature dataset, together containing 297 annual time series of 91 values. These datasets are examined in both their original and standardized forms. We further perform experiments using 24 000 simulated time series of 91 values. These experiments complement the real-world ones by allowing the examination of a large variety of process behaviours, while they are also controlled to some extent, facilitating generalizations and increasing the understanding on the examined problem. The number of forecasts produced using these real-world and simulated datasets are 47 520 and 2 130 000 respectively, i.e., the largest among its companion studies. Our aim is twofold; to provide generalized results regarding one-step ahead forecasting within a purely statistical framework (justified, for example, in Hyndman and Athanasopoulos 2018) in geoscience and hopefully to establish the results obtained by the examination of the standardized real-world datasets as rough benchmarks for the one-step ahead predictability of annual precipitation and temperature. The establishment of forecasting benchmarks is meaningful, especially for the one-step ahead attempts, as the latter constitute the most simple ones and their accuracy can be quantified using a single metric, i.e., the absolute error.

4.2 Data and methods

In this Section, we present the data and methods of the Chapter. Basic information on the methods' implementation is also provided, while the total of the exploited `R` packages is independently listed in Section 2.9.4. All `R` functions are used with their predefined values, unless specified differently. To ensure reproducibility, the `R` codes and data are available in Chapter's supplement. Hereafter, to specify an implemented `R` function, we state its name accompanied by the name of the `R` package. The latter name is given between curly brackets (`{}`). To imply that we implement a built-in-`R` function, we accompany its name with "`{stats}`".

We use the datasets briefly described in Tables 4.2 and 4.3. The PrecDat and TempDat datasets are annual and originate from two larger monthly datasets, available in Peterson and Vose (1997), and Lawrimore et al. (2011), respectively. The sample period is from 1910 to 2000, so that the following two conditions are simultaneously met: 1) there are no missing values; and 2) the number of stations around the globe is the largest possible. We note that for sample periods extending after 2000 the number of retained stations would decrease rapidly. Figure 4.1 presents the maps of the retained stations. The precipitation ones create a sufficiently dense network in the United States of America and in Scandinavia, while the retained temperature stations in the United States of America, in Japan and in a part of South Korea. As it is apparent from Table 4.2, the StandPrecDat and StandTempDat datasets simply contain the standardized time series of PrecDat and TempDat respectively. The standardization is made by using the mean and standard deviation maximum likelihood estimates of the fractional Gaussian noise process (see Section 2.1.6), obtained through the `R` function `mleHK {HKprocess}`. This latter function implements the maximum likelihood method. The standard deviation estimates would be considerably different if we modelled the time series using independent normal variables (Tyralis and Koutsoyiannis 2011).

Table 4.2. Summary of the real-world datasets. The exploited stations are presented in Figure 4.1.

S/n	Abbreviated name	Process	Type	Primal dataset	Number of time series
1	PrecDat	Precipitation	Original	Peterson and Vose (1997)	112
2	TempDat	Temperature		Lawrimore et al. (2011)	185
3	StandPrecDat	Precipitation	Standardized	PrecDat	112
4	StandTempDat	Temperature		TempDat	185

Table 4.3. Summary of the simulated datasets. The definitions of the stochastic processes are given in the Section 2.1. In the simulation, the parameters μ and σ of the simulated stochastic processes are set to 0 and 1 respectively.

S/n	Abbreviated name	Stochastic process	Autoregressive parameters	Moving average parameters	Number of time series
5	SimDat_1	AR(1)	$\varphi_1 = 0.7$		2 000
6	SimDat_2	AR(1)	$\varphi_1 = -0.7$		
7	SimDat_3	AR(2)	$\varphi_1 = 0.7, \varphi_2 = 0.2$		
8	SimDat_4	MA(1)		$\theta_1 = 0.7$	
9	SimDat_5	MA(1)		$\theta_1 = -0.7$	
10	SimDat_6	ARMA(1,1)	$\varphi_1 = 0.7$	$\theta_1 = 0.7$	
11	SimDat_7	ARMA(1,1)	$\varphi_1 = -0.7$	$\theta_1 = -0.7$	
12	SimDat_8	ARFIMA(0,0.30,0)			
13	SimDat_9	ARFIMA(1,0.30,0)	$\varphi_1 = 0.7$		
14	SimDat_10	ARFIMA(0,0.30,1)		$\theta_1 = -0.7$	
15	SimDat_11	ARFIMA(1,0.30,1)	$\varphi_1 = 0.7,$	$\theta_1 = -0.7$	
16	SimDat_12	ARFIMA(2,0.30,2)	$\varphi_1 = 0.7, \varphi_2 = 0.2$	$\theta_1 = -0.7, \theta_2 = -0.2$	

Figure 4.1 also presents the histograms of the Hurst parameter (H) maximum likelihood estimates (Tyralis and Koutsoyiannis 2011) of the formed real-world time series. These estimates are of importance within this Chapter for two reasons: 1) we implement a univariate time series forecasting method (see later on in this section) that takes advantage of this information under the established assumption of long-range dependence, 2) we standardize the original real-world time series using the mean and standard deviation maximum likelihood estimates (estimated simultaneously with the Hurst parameter) of the fractional Gaussian noise process (see above). The magnitude of the long-range dependence is mostly significant in the real-world time series.

For consistency purposes with respect to the real-world datasets (but also to approximate the typical length of annual geophysical time series), we simulate time series of 91 values (see Table 4.3). The generating models originate from the families of autoregressive moving average (ARMA) and autoregressive fractionally integrated moving average (ARFIMA). The definitions of these processes are given in Section 2.1.3 and 2.1.5, respectively (see also Wei 2006, pp. 6–65, 489–494). We simulate the ARMA processes by using the `R` function `arima.sim {stats}` and the ARFIMA processes by using the `R` function `fracdiff.sim {fracdiff}`. The simulations are performed with mean 0 and standard deviation of 1.

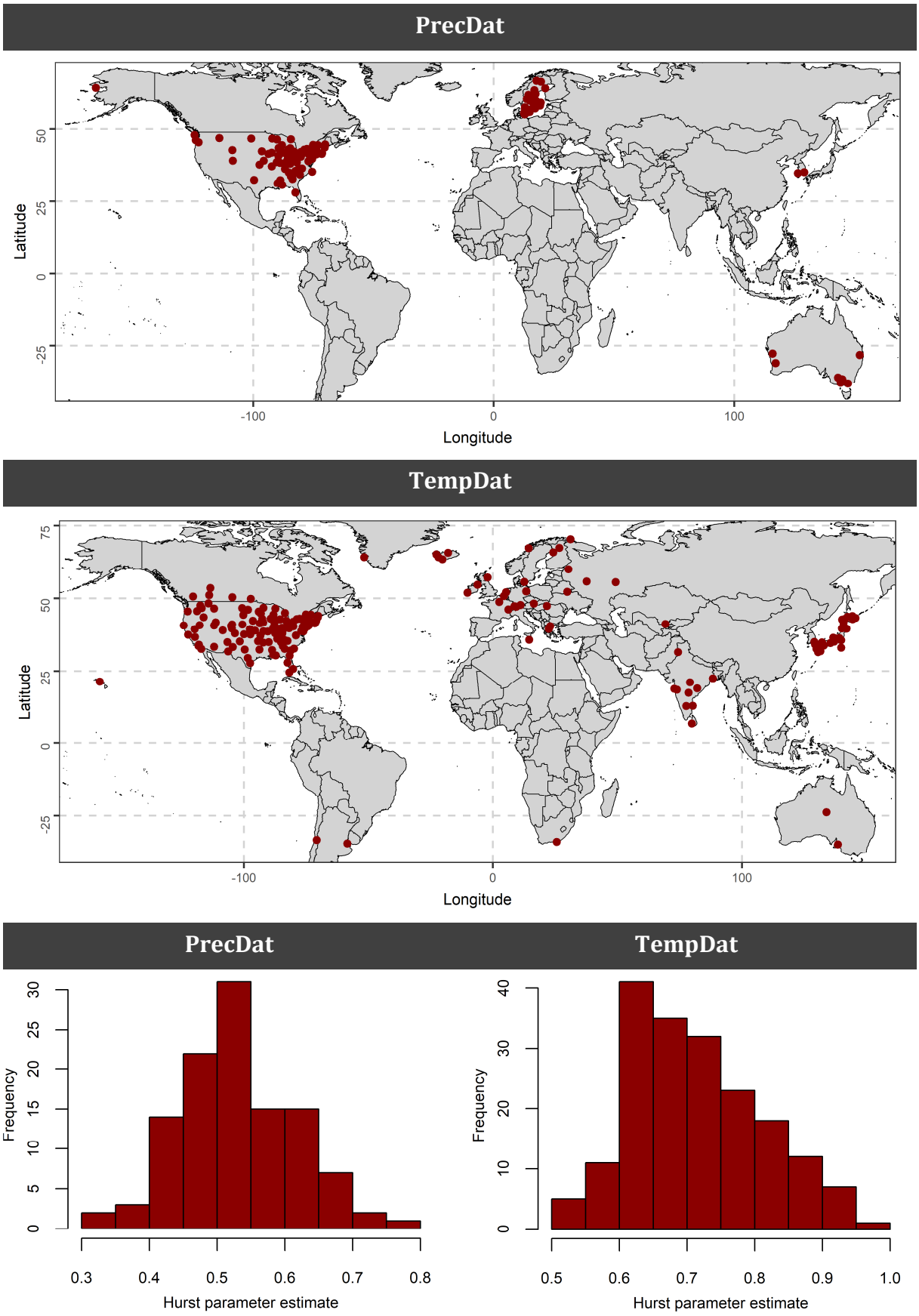


Figure 4.1. Maps of the exploited stations, and histograms of the estimated Hurst parameter (H) of the fractional Gaussian noise process for the original precipitation and temperature data. The data are sourced from Peterson and Vose (1997), and Lawrimore et al. (2011), respectively.

We implement the forecasting methods of Chapter 3. For consistency in the presentation, these methods are summarized in Table 4.4. In the same Table, the reader is referred to specific Sections of Chapter 2, in which the forecasting models and algorithms are documented. Since the theoretical examination is not possible for all the implemented methods, understanding most of them from a theoretical point of view could hardly help in interpreting the algorithmically obtained outcome of the Chapter. We should also note that ARIMA_s and auto_ARIMA_s are simulation methods. The models underlying these methods are the same that underly the ARIMA_f and auto_ARIMA_f ones, respectively; nonetheless, the innovations are set to zero in these latter ones for achieving accurate forecasts (for the related theoretical proof, see Wei 2006, pp. 88–93). This also applies to the auto_ARFIMA method, which can be used for modelling processes that are assumed to exhibit long-range dependence (see Section 2.1.5).

Table 4.4. Forecasting methods. Benchmarking information is provided in Section 3.2.6. Software implementation information is provided in Tables 3.3–3.5. The forecasting methods are available in code form in Chapter's supplement.

S/n	Abbreviated name	Corresponding model from Table 2.3	General category	Description
1	Naïve	Non-seasonal naïve	Simple	Section 2.2.1
2	RW	Random walk (RW)		
3	ARIMA_f	Fixed-order autoregressive integrated moving average (ARIMA)	ARIMA	Section 2.2.2
4	ARIMA_s			
5	auto_ARIMA_f	Optimum-order ARIMA		
6	auto_ARIMA_s			
7	auto_ARFIMA	Optimum-order autoregressive fractionally integrated moving average (ARFIMA)	ARFIMA	
8	BATS	Exponential smoothing state space with Box-Cox transformation, ARMA errors correction, trend and seasonal components (BATS)	Innovations State Space	Section 2.2.3
9	ETS_s	Exponential smoothing with error, trend and seasonal components (ETS)		
10	SES	Simple exponential smoothing (SES)	Exponential Smoothing	
11	Theta	Theta		
12	NN_1	Neural networks (NN)	Machine learning regression	Section 2.3.2
13	NN_2			
14	NN_3			
15	RF_1	Random forests (RF)		Section 2.3.3
16	RF_2			
17	RF_3			
18	SVM_1	Support vector machines (SVM)		Section 2.3.4
19	SVM_2			
20	SVM_3			

The assessment of the one-step ahead forecasting performance is based on the error metrics and accuracy statistics of Table 4.5.

Table 4.5. Error metrics and accuracy statistics. Their definitions are given in Section 2.9.3.

S/n	Abbreviated name	Full name	Category	Values	Optimum value
1	E	Error	Error metrics	$(-\infty,+\infty)$	0
2	AE	Absolute error		$[0,+\infty)$	0
3	PE	Percentage error		$(-\infty,+\infty)$	0
4	APE	Absolute percentage error		$[0,+\infty)$	0
5	MdoAE	Median of the absolute errors	Accuracy statistics	$[0,+\infty)$	0
6	MdoAPE	Median of the absolute percentage errors		$[0,+\infty)$	0
7	LRC	Linear regression coefficient		$(-\infty,+\infty)$	1
8	R2	Coefficient of determination		$[0,1]$	1

We conduct the experiments described in Tables 4.6 and 4.7. We use each dataset in five experiments; every time examining different part of the time series according to Table 4.7. While the application of the stochastic methods does not require a validation set (since all the model parameters are estimated using other procedures, such as the maximum likelihood estimation), the same does not apply to the application of the machine learning methods (except NN_3). For each of the latter, we fit the candidate models defined by all the considered hyperparameter values (see Table 3.4) to the fitting set, i.e., the first 33, 40, 47, 53 or 60 values, and subsequently use them to make predictions corresponding to the validation set, i.e. the next 17, 20, 23, 27 or 30 values respectively. Finally, we decide on the "optimal" model, i.e. the one exhibiting the smallest root mean square error on the validation set. We fit this model to the first 50, 60, 70, 80 or 90 values and make predictions corresponding to the 51st, 61st, 71st, 81st or 91st value respectively.

Table 4.6. Conducted experiments. The symbol *i* can take the values stated in Table 4.7.

S/n	Abbreviated name	Category	Dataset (see Tables 4.2, 4.3)	Forecasting methods (see Table 4.4)	Metrics used (see Table 4.5)
1	RWE_1*i*	Real-world	PrecDat	1, 2, 7–20	1–8
2	RWE_2*i*		TempDat		
3	RWE_3*i*		StandPrecDat	1, 2, 7–20	1, 2, 5, 7, 8
4	RWE_4*i*		StandTempDat		
5	SE_1*i*	Simulation	SimDat_1	1–6, 8–20	1, 2, 5, 7, 8
6	SE_2*i*		SimDat_2		
7	SE_3*i*		SimDat_3		
8	SE_4*i*		SimDat_4		
9	SE_5*i*		SimDat_5		
10	SE_6*i*		SimDat_6		
11	SE_7*i*		SimDat_7		
12	SE_8*i*		SimDat_8	1, 2, 7–20	
13	SE_9*i*		SimDat_9		
14	SE_10*i*		SimDat_10		
15	SE_11*i*		SimDat_11		
16	SE_12*i*		SimDat_12		

Table 4.7. Part of the time series used within each experiment according to the *i* value.

S/n	*i*	Data points of each time series used for the model-fitting (required for all models) and model-validation (required for the machine learning models)	Data points of each time series used for model testing
1	a	1, 2, 3, …, 50	51
2	b	1, 2, 3, …, 60	61
3	c	1, 2, 3, …, 70	71
4	d	1, 2, 3, …, 80	81
5	e	1, 2, 3, …, 90	91

The only assumption of our methodological approach concerns the application of the auto_ARFIMA method within the real-world experiments and is that the annual precipitation and temperature variables can be sufficiently modelled by the normal distribution. This assumption is rather reasonable (implied by the Central Limit Theorem; Koutsoyiannis 2008, Chapter 2.5.6) and could hardly harm the results. In general, such fundamental assumptions are preferable to the introduction of extra parameters, e.g., to using the Box-Cox transformation (see Section 2.1.8) to normalize the data. The rest of the methods are non-parametric and, thus, not affected by the possible non-normality.

To take advantage of some well-known theoretical properties, in the SE_1*i*–SE_7*i* simulation experiments the ARIMA_f and ARIMA_s methods are given the same AutoRegressive (AR) and Moving Average (MA) orders used in the respective simulation process, while the number of differencing (*d*) is set 0. These two methods, as well as the simple, auto_ARIMA_f, auto_ARIMA_s and auto_ARFIMA methods serve as reference points within our approach. In particular, ARIMA_f, auto_ARIMA_f and auto_ARFIMA are theoretically expected to be the most accurate within our

simulation experiments (for an explanation, see Section 3.2.6), while BATS is also expected to perform well in these experiments, since it comprises an ARMA model. In summary, the experiments are controlled to some extent, while their components (datasets, methods and metrics) are selected to provide a multifaceted approach to the problem of one-step ahead forecasting in geoscience.

4.3 Results and discussion

In this section, we summarize the basic quantitative and qualitative information gained from the experiments of the present Chapter, while the total amount is available in Chapter's supplement. We further discuss the findings and explicate their contribution in light of the literature.

4.3.1 Experiments using the precipitation datasets

For the experiments using the PrecDat dataset the minimum AE value is 0 (practically) and the maximum around 1 750 mm (for forecasts produced by the simple forecasting methods, i.e. Naïve and RW), while the respective values for the APE error metric are 0 (practically) and 1.64 (for a forecast produced by NN_1). The MdoAE and MdoAPE values are summarized in Tables 4.8 and 4.9 respectively. The minimum MdoAE is 68 mm, while the maximum is 189 mm. These two values are in the same order of magnitude as the smallest and average standard deviation estimates of the time series respectively. The minimum MdoAPE value is 0.09 and the maximum 0.22, while the respective LRC values are 0.73 and 1.18. The best LRC value (1.00) is measured within RWE_1c for the simple forecasting methods, while the best R2 value (0.84) is measured within RWE_1d for BATS. The worst LRC and R2 values are 0.73 for RF_2 within RWE_1d and 0.54 for NN_1 within RWE_1a respectively.

Table 4.8. Minimum, maximum and mean values of the median of absolute errors within the experiments using the precipitation dataset.

Experiment	Minimum (mm)	Maximum (mm)	Mean (mm)
RWE_1a	111 (RF_1)	172 (NN_1)	135
RWE_1b	**68** (SVM_1)	146 (ETS_s)	91
RWE_1c	91 (SVM_3)	171 (ETS_s)	119
RWE_1d	143 (BATS)	**189** (RF_2)	162
RWE_1e	98 (Theta)	150 (NN_1)	122

Table 4.9. Minimum, maximum and mean values of the median of absolute percentage errors within the experiments using the precipitation dataset.

Experiment	Minimum	Maximum	Mean
RWE_1a	0.12 (RF_1)	0.21 (RW)	0.16
RWE_1b	**0.09** (SVM_1)	0.18 (ETS_s)	0.12
RWE_1c	0.12 (SVM_3)	0.21 (NN_1)	0.15
RWE_1d	0.15 (BATS)	**0.22** (NN_1)	0.17
RWE_1e	0.12 (Theta)	0.18 (NN_1)	0.15

In Figure 4.2, we present a graphical summary of the experiments using the PrecDat dataset. The values in the three upper heatmaps are scaled in the row direction and the darker the colour within a specific row the better the forecasts. In fact, heatmaps are used in this Chapter instead of conventional tables, since they allow the easy extract of qualitative information. The relative performance of the forecasting methods differs to some degree across the various RWE_1*i* experiments, with ETS_s and NN_1 being the worst performing in terms of MdoAE and MdoAPE, followed by the simple methods. On the other hand, in terms of LRC Naïve and RW exhibit rather the best overall performance. In the downer heatmap of Figure 4.2 we zoom into the RWE_1b experiment. By its examination we observe that all the implemented forecasting methods can perform well or bad, depending on the individual case. This fact is also apparent in the side-by-side boxplots of Figure 4.2. Furthermore, we observe that for one specific time series the AE values measured are very high for all the forecasts apart from those produced by the simple forecasting methods.

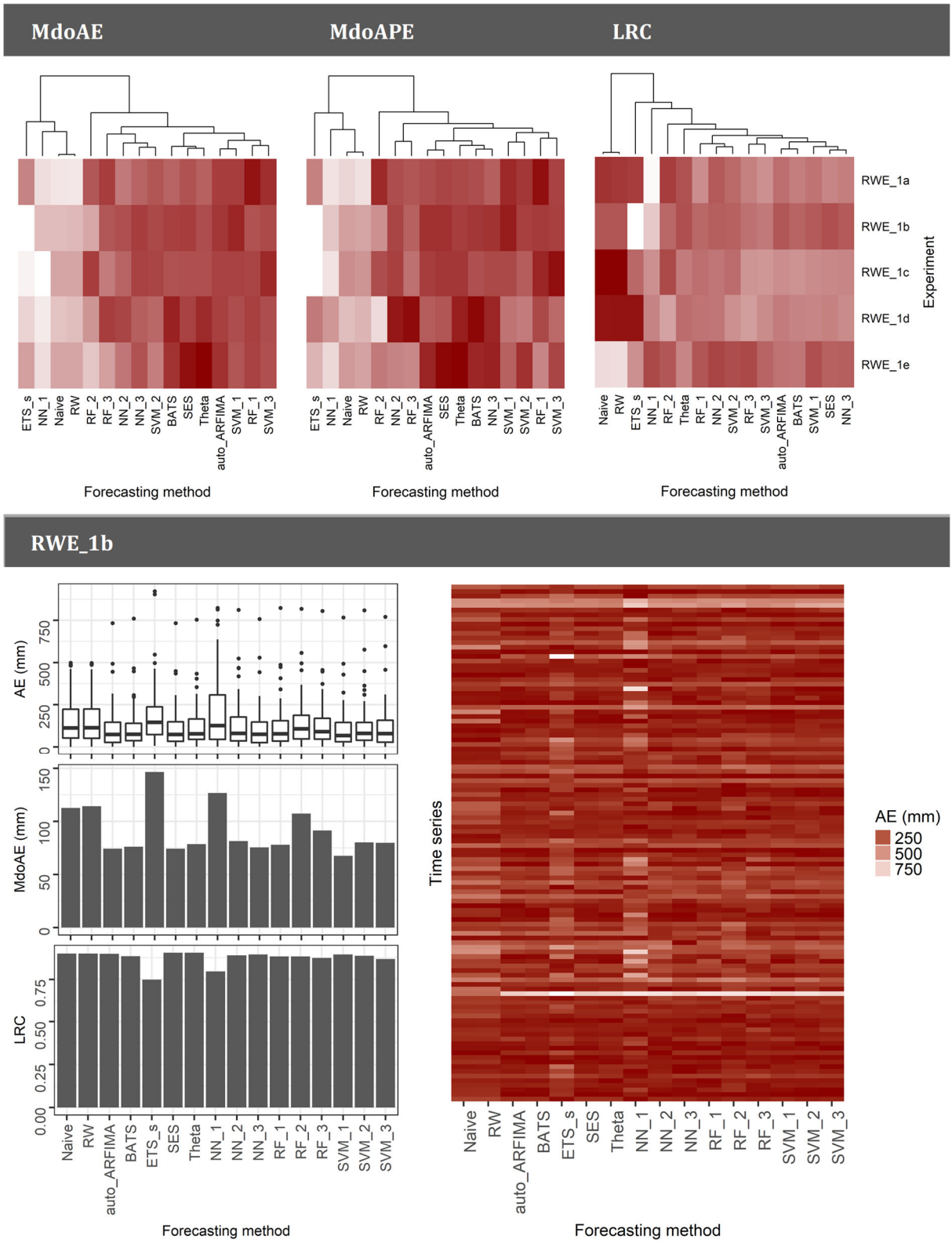


Figure 4.2. Results in brief of the experiments using the precipitation dataset.

Regarding the experiments using the StandPrecDat dataset, the minimum AE value is 0 and the maximum around 10. The MdoAE values are summarized in Table 4.10. The minimum MdoAE is 0.55, while the maximum is 1.42. These two values are 45% smaller and 42% larger than 1 (standard deviation of the standardized time series) respectively. Since there is an absence of relevant information in the literature, these values could be used as rough benchmarks for the predictability of annual precipitation. Most preferably, a representative sample set of univariate

time series forecasting methods could be implemented at least for benchmarking purposes alongside with any other forecasting attempt. Moreover, the minimum and maximum LRC values are −0.25 and 0.25 respectively, the former measured for ETS_s and the latter for RW. Finally, the minimum R2 value is 0 (practically), while the maximum is 0.09, measured within SE_3a for ETS_s. In addition to this numerical information, Figure 4.3 presents a brief comparison between the experiments using the PrecDat and StandPrecDat datasets. As illustrated in this figure, the relative performance of the forecasting methods with respect to AE and MdoAE in the experiments using the latter dataset is mostly similar to the one in the experiments using the former dataset. Nevertheless, the LRC (and R2) values are far worse when using the standardized datasets. In fact, standardization results to processes with different predictability with respect to the original.

Table 4.10. Minimum, maximum and mean values of the median of absolute errors within the experiments using the standardized precipitation dataset.

Experiment	Minimum	Maximum	Mean
RWE_3a	0.70 (RF_1)	1.22 (NN_1)	0.92
RWE_3b	**0.55** (SVM_2)	0.95 (ETS_s)	0.69
RWE_3c	0.72 (BATS)	**1.42** (NN_1)	0.86
RWE_3d	0.99 (Theta)	**1.42** (ETS_s)	1.14
RWE_3e	0.69 (Theta)	1.07 (ETS_s)	0.89

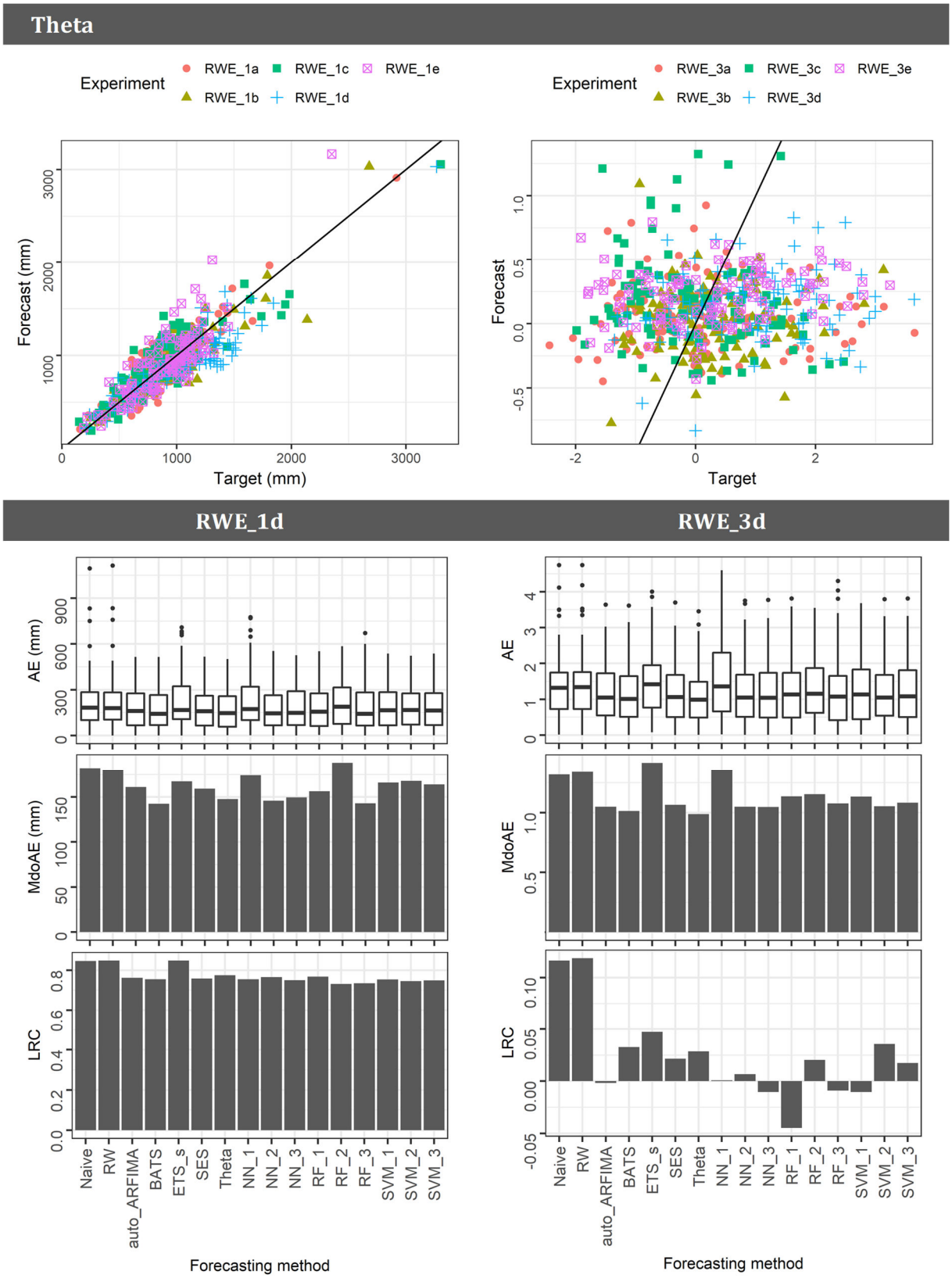


Figure 4.3. Comparison in brief between the experiments using the precipitation and the standardized precipitation datasets.

4.3.2 Experiments using the temperature datasets

In Figure 4.4, we present a graphical summary of the experiments using the TempDat dataset. For these experiments the minimum AE value is 0 and the maximum around 43 °C (for a forecast produced by NN_2), while the respective APE values are 0 and 9.64 (for a forecast produced by ETS_s). The MdoAE and MdoAPE values are summarized in Tables 4.11 and 4.12 respectively. The minimum MdoAE is 0.23 °C, while the maximum is 1.10 °C. These two values are in the same order of magnitude as the smallest and largest standard deviation estimates of the temperature time series respectively. The respective values for the MdoAPE are 0.02 and 0.08. The minimum LRC value is 0.95 and the maximum is 1.02; all the LRC values are close to the optimum. Finally, the minimum R2 value is 0.78, measured for NN_2 within RWE_2b, while all the rest R2 values are higher than 0.97 with maximum 1 (practically), measured for the auto_ARFIMA method within RWE_2b. In summary, the relative performance of the forecasting methods vary across the different experiments conducted using the TempDat dataset. The auto_ARFIMA, BATS, SES, Theta and NN_3 seem to be well performing in terms of MdoAE and MdoAPE when applied to these temperature time series compared to the overall picture, while the simple methods are far the best in terms of MdoAE within the RWE_2d experiment. ETS_s and NN_1 are the worst performing within all the experiments apart from RWE_2c, in which the simple methods exhibit the worst performance. Finally, by comparing the numerical results of the experiments using the PrecDat and TempDat dataset, we observe the fact that temperature is more predictable than precipitation.

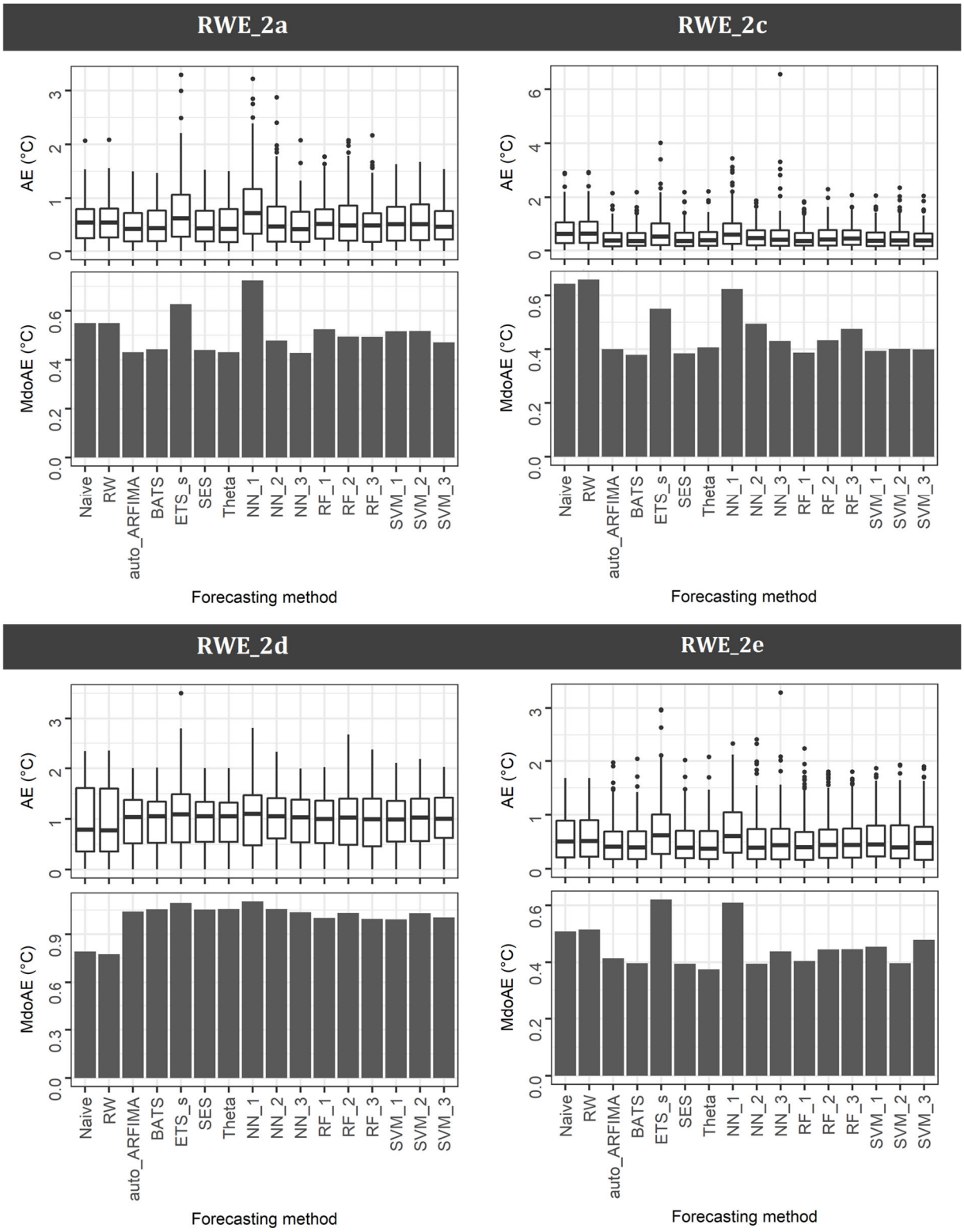


Figure 4.4. Results in brief of the experiments using the temperature dataset.

Table 4.11. Minimum, maximum and mean values of the median of absolute errors within the experiments using the temperature dataset.

Experiment	Minimum (°C)	Maximum (°C)	Mean (°C)
RWE_2a	0.42 (NN_3)	0.72 (NN_1)	0.51
RWE_2b	**0.23** (Theta)	0.54 (NN_1)	0.32
RWE_2c	0.38 (BATS)	0.66 (RW)	0.47
RWE_2d	0.78 (RW)	**1.10** (NN_3)	1.01
RWE_2e	0.38 (Theta)	0.62 (ETS_s)	0.46

Table 4.12. Minimum, maximum and mean values of the median of absolute percentage errors within the experiments using the temperature dataset.

Experiment	Minimum	Maximum	Mean
RWE_2a	0.04 (Theta)	0.06 (NN_1)	0.04
RWE_2b	**0.02** (auto_ARFIMA)	0.05 (ETS_s)	0.03
RWE_2c	0.03 (SVM_1)	0.06 (RW)	0.04
RWE_2d	0.07 (Naïve)	**0.08** (NN_1)	0.08
RWE_2e	0.03 (RF_1)	0.05 (NN_1)	0.04

Regarding the experiments using the StandTempDat, the minimum AE value is 0 and the maximum around 18.91. The MdoAE values are summarized in Table 4.13. The minimum MdoAE value is 0.33, while the maximum is 1.46. These two values are 67% smaller and 46% larger than 1 (standard deviation of the standardized time series) respectively and could be used as rough benchmarks for the predictability of annual temperature (for an explanation, see the subsection entitled "Experiments using the precipitation datasets"). The minimum LRC value is 0.04 and the maximum is 0.76, the former measured for SVM_1 and the latter for RW. Finally, the minimum R2 value is 0.03, while the maximum is 0.48. The latter value is measured for Naïve in RWE_4a. Figure 4.5 facilitates a comparison between the experiments using the TempDat and StandTempDat datasets. Here as well, we observe that the relative performance of the forecasting methods with respect to AE and MdoAE in the experiments using the standardized precipitation time series mostly does not vary from the respective relative performance when using the original temperature time series. We further note that the LRC (and R2) values are worse when using the standardized temperature dataset, while they are better for the latter than for the standardized precipitation dataset.

Table 4.13. Minimum, maximum and mean values of the median of absolute errors within the experiments using the standardized temperature dataset.

Experiment	Minimum	Maximum	Mean
RWE_4a	0.61 (BATS)	0.93 (ETS_s)	0.71
RWE_4b	**0.33** (Theta)	0.73 (NN_1)	0.47
RWE_4c	0.56 (SES)	0.96 (ETS_s)	0.69
RWE_4d	1.20 (NN_1)	**1.46** (Theta)	1.36
RWE_4e	0.48 (Theta)	0.82 (ETS_s)	0.61

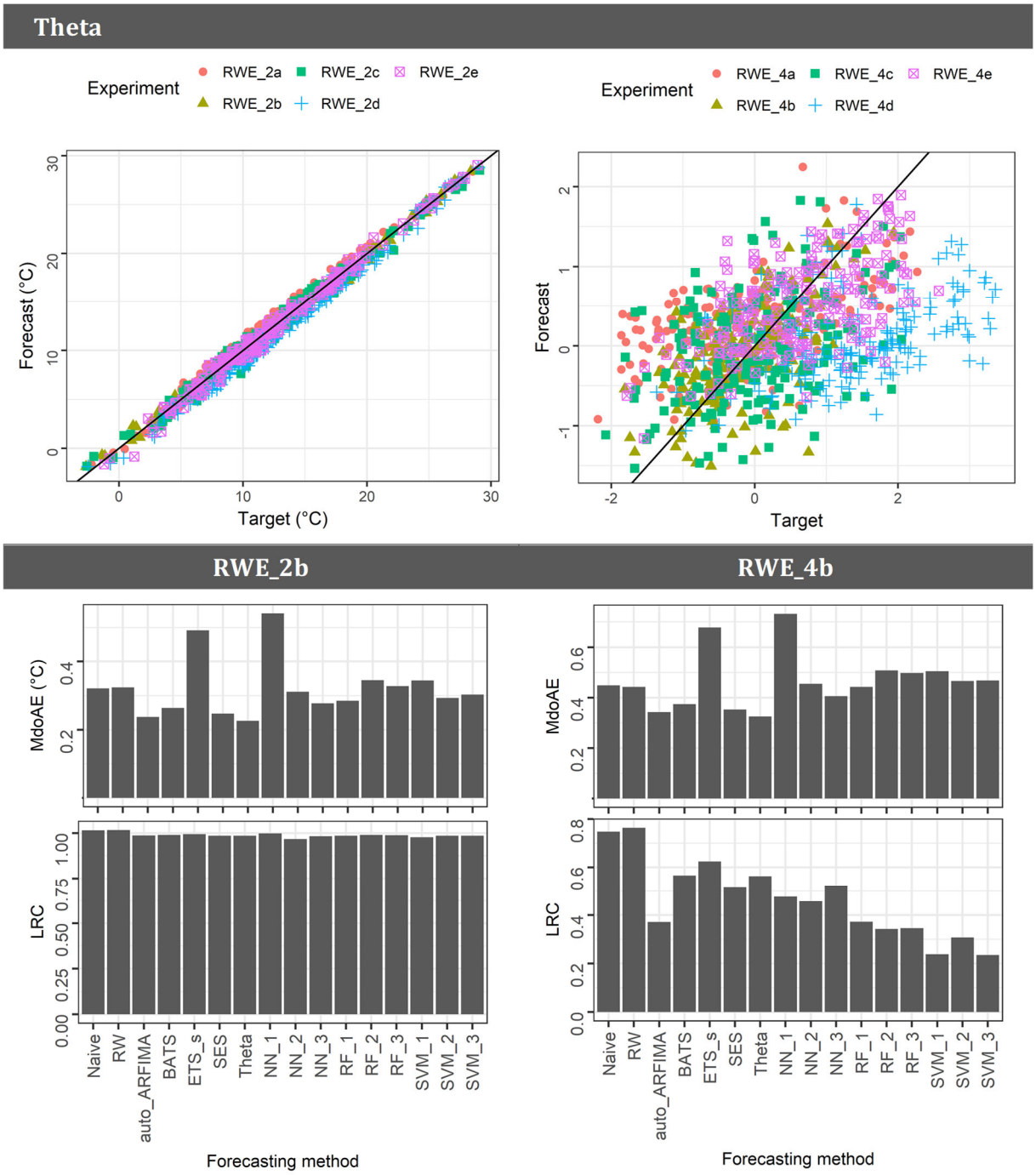


Figure 4.5. Comparison in brief between the experiments using the temperature and standardized temperature datasets.

4.3.3 Experiments using the simulated datasets

The subsequently reported information constitutes the provided empirical solution to the problem of one-step ahead forecasting in geoscience. Nonetheless, this solution is rather qualitative than quantitative (although the results are also stated quantitatively), since the respective experiments use unscaled data, that could be assumed as real-world data in a standardized form (such as StandPrecDat and StandTempDat) with different predictability than the original (for example, see the subsections entitled "Experiments using the precipitation datasets" and "Experiments using the temperature datasets"). In fact, the experiments using standardized precipitation and temperature can facilitate a connection between the experiments

using the same data in their original form and the experiments using the simulated datasets. A graphical summary of the latter experiments is available in Table 4.7.

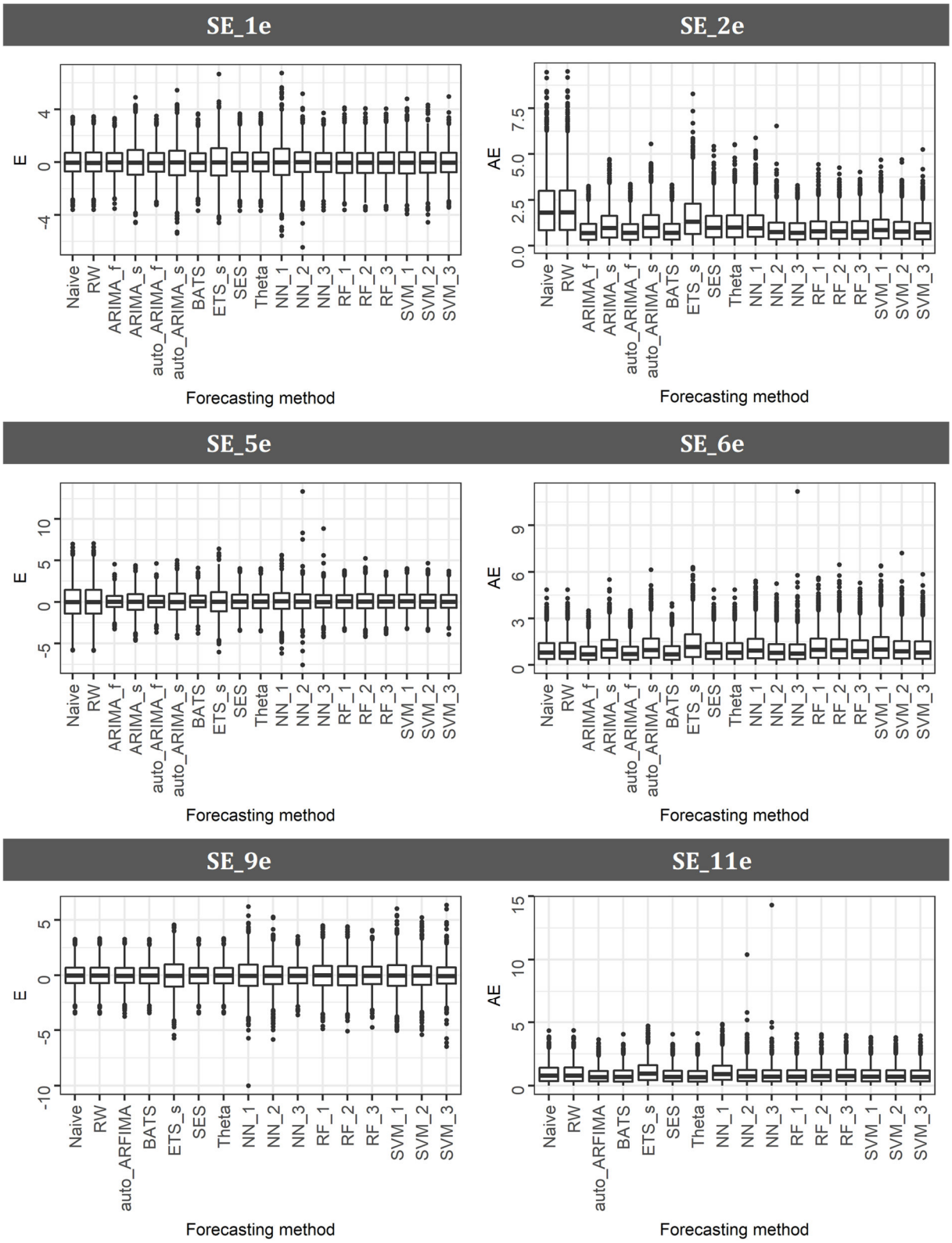


Figure 4.6. Results in brief of the experiments using the simulated datasets.

The generalized findings of the present Chapter are the following:

(1) The *E* values are approximately symmetric around 0 (mean value of the simulations).

(2) The results may vary significantly across the simulation experiments using different simulated datasets and across the different time series within a specific experiment depending on the forecasting method.

(3) Consequently, the relative performance of the forecasting methods may also vary significantly across the simulation experiments using different simulated datasets.

(4) On the contrary, the relative performance of the forecasting methods is slightly affected by the length of the time series for the experiments of the present Chapter. The same has been found to mostly apply to the multi-step ahead forecasting performance of the same methods in Chapter 3 of this thesis for two different time series lengths.

(5) Some forecasting methods are more accurate than others. The best-performing methods are ARIMA_f, auto_ARIMA_f, auto_ARFIMA, BATS, SES and Theta. This good performance of the former four methods when applied to ARMA and ARFIMA processes is expected from theory, while the Theta forecasting method has also performed well in the M3 Competition (Makridakis and Hibon 2000) and is expected to have a similar performance with SES (Hyndman and Billah 2003). The five above-mentioned forecasting methods are all stochastic.

(6) All the machine learning methods except for NN_1 (mostly NN_3 and SVM_3) are comparable to the best-performing methods, as it has also been found to apply in the experiments of Chapters 3 and 6 of this thesis. Likewise, in Tyralis and Papacharalampous (2017) random forests are competitive with the ARFIMA and Theta benchmarks.

(7) The simple methods are competitive in specific simulation experiments, as also suggested for specific cases in Cheng et al. (2017), Makridakis and Hibon (2000) and Chapter 3 herein. Nevertheless, they stand out because of their bad performance in other simulation experiments.

(8) Most of the far outliers are produced by neural networks.

The minimum AE value for the forecasts is 0 (practically) and the maximum around 155 (produced by NN_2). The MdoAE values are summarized in Tables 4.14 and 4.15. Especially the latter is useful in supporting observations (5), (6) and (7). The minimum MdoAE is 0.65, while the maximum is 2.91. These two values are 35% smaller and 191% larger than 1 (standard deviation of the simulations) respectively. Furthermore, in spite of observation (4), the MdoAE values may decrease on the level of the second or even the first decimal, when moving from the simulation experiments using time series of 51 values to those of 91 values, with the NN_1 forecasting method exhibiting the largest improvement. The minimum LRC value is −0.88 and the maximum is 0.94, both measured for RW, while the minimum and maximum values produced by Naïve differ in the second and third decimal respectively. This range holds a complete interpretation of the observed within the real-world experiments variants in the performance of the simple methods in terms of LRC from extremely good to extremely bad (with respect to the overall picture). Finally, the minimum R2 value is 0 (practically), measured for ETS_s within several experiments, while the maximum is 0.84 within SE_9b for Naïve.

Table 4.14. Minimum, maximum and mean values of the median of absolute errors within the simulation experiments.

Experiment	Minimum	Maximum	Mean
SE_1*i*	0.68 (ARIMA_f | SE_1a)	1.05 (NN_1 | SE_1a)	0.80
SE_2*i*	0.67 (ARIMA_f | SE_2c)	1.82 (RW | SE_2e)	0.95
SE_3*i*	**0.65** (ARIMA_f | SE_3c)	1.04 (NN_1 | SE_3a)	0.81
SE_4*i*	0.67 (ARIMA_f | SE_4c)	1.21 (ETS_s | SE_4a)	0.84
SE_5*i*	0.66 (ARIMA_f | SE_5e)	1.48 (RW | SE_5c)	0.90
SE_6*i*	0.68 (ARIMA_f | SE_6b)	1.20 (ETS_s | SE_6d)	0.89
SE_7*i*	0.66 (auto_ARIMA_f | SE_7d)	**2.91** (RW | SE_7e)	1.22
SE_8*i*	0.67 (auto_ARFIMA | SE_8c)	1.02 (NN_1 | SE_8a)	0.77
SE_9*i*	0.67 (auto_ARFIMA | SE_9d)	1.05 (NN_1 | SE_9b)	0.80
SE_10*i*	0.67 (auto_ARFIMA | SE_10e)	1.22 (RW | SE_10e)	0.83
SE_11*i*	0.68 (Theta | SE_11e)	1.10 (NN_1 | SE_11a)	0.77
SE_12*i*	0.69 (auto_ARFIMA | SE_12b)	1.06 (NN_1 | SE_12a)	0.78

Table 4.15. Minimum, maximum and mean values of the median of absolute errors for each forecasting method.

Method	Minimum	Maximum	Mean
Naïve	0.68 (SE_3c)	2.88 (SE_7a)	1.12
RW	0.69 (SE_9c)	**2.91** (SE_7e)	1.13
ARIMA_f	**0.65** (SE_3c)	0.72 (SE_7a)	0.69
ARIMA_s	0.91 (SE_2a)	1.04 (SE_3a)	0.96
auto_ARIMA_f	0.66 (SE_7d)	0.75 (SE_6c)	0.70
auto_ARIMA_s	0.91 (SE_4c)	1.02 (SE_3d)	0.97
auto_ARFIMA	0.67 (SE_10e)	0.73 (SE_10d)	0.69
BATS	0.67 (SE_3c)	0.76 (SE_6c)	0.71
ETS_s	0.93 (SE_3d)	2.11 (SE_7e)	1.14
SES	0.66 (SE_3c)	1.52 (SE_7e)	0.83
Theta	0.66 (SE_3c)	1.57 (SE_7a)	0.84
NN_1	0.90 (SE_7e)	1.16 (SE_7a)	1.01
NN_2	0.72 (SE_8c)	0.89 (SE_5b)	0.79
NN_3	0.69 (SE_8c)	0.84 (SE_6c)	0.74
RF_1	0.71 (SE_8c)	1.08 (SE_6a)	0.82
RF_2	0.72 (SE_8c)	1.04 (SE_6c)	0.83
RF_3	0.72 (SE_3c)	0.98 (SE_6c)	0.80
SVM_1	0.71 (SE_8e)	1.23 (SE_7a)	0.86
SVM_2	0.68 (SE_8c)	1.01 (SE_7a)	0.81
SVM_3	0.68 (SE_8c)	0.92 (SE_6c)	0.76

4.4 Conclusions

The simulation experiments reveal the most and least accurate methods for long-term one-step ahead forecasting applications, also suggesting that the simple methods may be competitive in specific cases. Furthermore, the relative performance of the forecasting methods is slightly affected by the time series length for the simulation experiments of this Chapter (using time series of 51, 61, 71, 81, 91 values), while it strongly depends on the process. Also importantly, the experiments using the original real-world time series result to minimum and maximum medians of the absolute errors of 68 mm and 189 mm for precipitation, and 0.23 °C and 1.10 °C for temperature respectively. Additionally, the experiments using the standardized real-world time series suggest that the minimum and maximum medians of the absolute errors are 0.55 and 1.42 for precipitation, and 0.33 and 1.46 for temperature respectively. These latter numerical results could be used as a rough upper boundary for the one-step ahead predictability of annual precipitation and temperature.

We subsequently state the limitations of this Chapter and some future directions. The provided empirical solution to the problem of one-step ahead forecasting in geoscience is rather

qualitative than quantitative, while the experiments using standardized precipitation and temperature data have offered rough benchmarks only. In the future more real-world data could be used to develop improved benchmarks for assessing the respective predictabilities. It would be of interest to further investigate how these predictabilities depend on the location from which the data originate. In this case, more stations spanning around the globe would be required. Moreover, a direct and large-scale comparison, set on a common base (if this is feasible), between deterministic and statistical approaches to forecasting geophysical processes would be useful and interesting. Another limitation of this Chapter is related to the adopted modelling approach, i.e. the data-driven one, according to which the selection of the model does not depend on the properties of the examined process and, therefore, the latter are mostly not investigated. Furthermore, the improvement of the performance of the machine learning models requires extensive comparisons between different procedures of hyperparameter optimization and lagged variable selection. Finally, future research could focus on the examination of the respective predictabilities, when using exogenous predictor variables as well (e.g., ARMAX and ARFIMAX models), while a definitely worth-stated future direction is related to the adoption of probabilistic forecasting methods, instead of the point forecasting ones.

5. Predictability of monthly temperature and precipitation using automatic time series forecasting methods

In this Chapter, we investigate the predictability of monthly temperature and precipitation by applying automatic univariate time series forecasting methods to a sample of 985 40-year long monthly temperature and 1 552 40-year long monthly precipitation time series. The methods include a naïve one based on the monthly values of the last year, as well as the random walk (with drift), autoregressive fractionally integrated moving average (ARFIMA), exponential smoothing state space model with Box-Cox transformation, ARMA errors, trend and seasonal components (BATS), simple exponential smoothing, Theta and Prophet methods. Prophet is a recently introduced model, inspired by the nature of time series forecasted at Facebook. In this Chapter, it is applied for the first time in the literature to hydrometeorological time series. Moreover, the use of random walk, BATS, simple exponential smoothing and Theta is rare in hydrology. The methods are tested in performing multi-step ahead forecasts for the last 48 months of the data. We further investigate how different choices of handling the seasonality and non-normality affect the performance of the models. The results indicate that (a) all the examined methods apart from the naïve and random walk ones are accurate enough to be used in long-run applications, (b) monthly temperature and precipitation can be forecasted to a level of accuracy which can barely be improved using other methods, (c) the externally applied classical seasonal decomposition results mostly in better forecasts compared to the automatic seasonal decomposition used by the BATS and Prophet methods and (d) Prophet is competitive, especially when it is combined with externally applied classical seasonal decomposition.

5.1 Introduction

The role of univariate time series forecasting methods for hydrometeorological and climate forecasting has been emphasized by forecasting experts (Armstrong and Fildes 2006; Green and Armstrong 2007; Green et al. 2009). Relevant reviews linking geosciences with the forecasting scientific field are available in the literature (e.g., in Bărbulescu 2016; Sivakumar 2017), while critical reviews of studies applying time series point forecasting methods in hydrology are available in the introductory Sections of Chapters 3 and 4 herein. Moreover, time series analysis is an essential tool for better forecasts; consequently, analysis and forecasting are usually presented simultaneously in specialized textbooks (e.g., in Hyndman and Athanasopoulos 2018; Wei 2006).

While time series forecasting is of interest in hydrology, the principles of forecasting are not always conscientiously applied by hydrological scientists. This fact is revealed by the experiments of Chapter 3. Furthermore and despite the growing literature specialized in time series prediction, and in the examination of subtleties related to the development, application and assessment of methods (see e.g., the review by De Gooijer and Hyndman 2006), the gained technical know-how has not been fully exploited by geoscientists. This is also suggested by the low number of geoscientific papers citing literature from relevant leading journals (e.g., the International Journal of Forecasting). Armstrong and Fildes (2006) have recognized the need for promoting forecasting in geosciences and proposed the publication of a special issue on applications of traditional forecasting methodologies to climate sciences.

Admittedly, the recent trend in geosciences is focusing on the development of soft computing methods for time series forecasting. These methods can be equally accurate, yet more computationally intensive, compared to the classical forecasting approaches, as presented in Chapter 3. In addition to the right above distinction, the various forecasting alternatives can be classified as automatic and non-automatic. The non-automatic or subjective approach to the problem of time series forecasting requires the prior conduct of an exploratory data analysis for each specific individual case to be predicted and human intervention during the forecasting process (Chatfield 1988). Therefore, its implementation can be significantly limited by scale-dependent factors. Taylor and Letham (2018) identify three types of scale in forecasting related to the number of people making forecasts (and their varying backgrounds), the diversity of the

characteristics of each forecasting problem and the number of forecasts needed. Automatic time series forecasting is essential, for example, when a large number of time series forecasts is required (Hyndman and Khandakar 2008). Consequently, specialized software for automatic time series forecasting has been developed (Hyndman and Khandakar 2008; Hyndman et al. 2018; Taylor and Letham 2017, 2018).

The `R` package `forecast` mostly includes methods based on exponential smoothing (Hyndman et al. 2008) or autoregressive integrated moving average (ARIMA) and related stochastic processes. The ARIMA processes and their relevant applications introduced by Box and Jenkins (1968) have been consistently used in hydrology from earlier years (e.g., in Carlson et al. 1970) until more recent times (e.g., in Montanari et al. 1997, 2000), while exponential smoothing has been used less frequently (e.g., in Chapters 3 and 4 herein). Recent methods for automatic time series forecasting (that have been used rarely in hydrology) include the Theta method (Assimakopoulos and Nikolopoulos 2000; see also Hyndman and Billah 2003) and the BATS (acronym for Box-Cox transformation, ARMA errors, Trend, and Seasonal components) method (De Livera et al. 2011). A more recently introduced forecasting model is Prophet (Taylor and Letham 2018). This latter model is inspired by the nature of time series forecasted at Facebook, and is available in the `R` package `prophet`.

Time series forecasting methodologies can also be classified into two groups according to the forecast horizon, i.e., one- and multi- step ahead forecasting. The latter is more difficult compared to the former. Still, it has been frequently used in hydrology (e.g., in Singh et al. 2011; Valipour et al. 2013; Tyralis and Koutsoyiannis 2014; Papacharalampous and Tyralis 2018a; Papacharalampous et al. 2018c; Tyralis and Papacharalampous 2018; Chapters 3 and 6 herein). Relevant reviews on multi-step ahead forecasting methodologies can be found in Chevillon (2007) and Taieb et al. (2012). Multi-step ahead forecasting has been examined theoretically (e.g., in Pemberton 1987; Stoica and Nehorai 1989; De Gooijer and Klein 1992; De Gooijer and Kumar 1992; Wei 2006, pp. 88–107; Franses and Legerstee 2010; Taieb and Atiya 2016) and empirically (e.g., in Papacharalampous et al. 2018a; Chapters 3 and 6 herein). The theoretical examination is not possible for all methods, while the performance of the latter can vary considerably in real-world case studies. An alternative framework for assessing the performance of forecasting methods is through their application to large datasets. Thus, competitions are organized, in which methods are improved and compared with each other (e.g., Makridakis et al. 1987; Makridakis and Hibon 2000), and later published (e.g., Andrawis et al. 2011).

The culture of evaluating the forecasting performance of methods by using large datasets is not a usual practice in hydrology, as also noted in Chapter 3 herein. The latter Chapter compares stochastic and machine learning methods in multi-step ahead forecasting by using a large dataset consisted of both simulated and streamflow data. A similar approach has been used in Chapter 4, in which annual values of mean temperature and precipitation, as well as simulated processes, are used for performing one-step ahead forecasting experiments. These works are among the first, in which forecasting methods popular in the time series forecasting field (e.g., BATS, Theta and simple exponential smoothing) are used in hydrology. This innovation has been justified by the detailed literature review conducted by Tyralis et al. (2020a).

In this Chapter, we apply automatic univariate time series forecasting methods to a large sample of 985 40-year long monthly temperature and 1 552 40-year long monthly precipitation time series. This sample is the largest used in hydrology for assessing the performance of forecasting methods. We implement a naïve forecasting method based on the monthly values of the last year, as well as the random walk (with drift), AutoRegressive Fractionally Integrated Moving Average (ARFIMA), BATS, simple exponential smoothing, Theta and Prophet forecasting methods. The methods' multi-step ahead forecasting performance is assessed using the last 48 months of the data. The aims of the present Chapter are to:

1) Assess the performance of the BATS, simple exponential smoothing and Theta methods when forecasting monthly geophysical time series.

2) Compute the minimum forecasting error, which directly expresses the predictability of monthly temperature and precipitation.
3) Assess whether using Box-Cox transformations and/or classical seasonal decomposition (additive or multiplicative) results in a better performance of the forecasting methods. For the latter see the relevant discussion in Mills (2011, pp. 375–395).
4) Assess the performance of the Prophet method.

5.2 Methodological framework

In this Section, we present the data and methods of the present Chapter. Basic information on the methods' implementation is also provided, while the total of the exploited `R` packages is independently listed in Section 2.9.4. Hereafter, to specify an implemented `R` function, we state its name accompanied by the name of the `R` package. The latter name is given between curly brackets (`{}`). To imply that we implement a built-in-`R` function, we accompany its name with "`{stats}`". All `R` functions are used with their predefined values, unless specified differently.

5.2.1 Global temperature and precipitation datasets

We use monthly temperature and precipitation instrumental data from the stations presented in Figure 5.1. The primary datasets are documented in Lawrimore et al. (2011), and Peterson and Vose (1997), respectively. The data span from 1950 to 1989 (i.e., 480 months). Since we need a large dataset without missing values, we do not use more recent data. Indeed, the number of stations without missing data decreases rapidly after 1990. Furthermore, 480 months are sufficient for a reliable inference regarding the performance of the forecasting methods. The stations are located in regions with different climates; therefore, our testing could be affected by the dispersion of the stations. To mitigate this effect, we group the stations according to their locations, as presented in Table 5.1 for the temperature stations and Table 5.2 for the precipitation ones. We note that 130 temperature and 149 precipitation stations are left out of the formed groups, due to their remote locations compared to the rest of the stations.

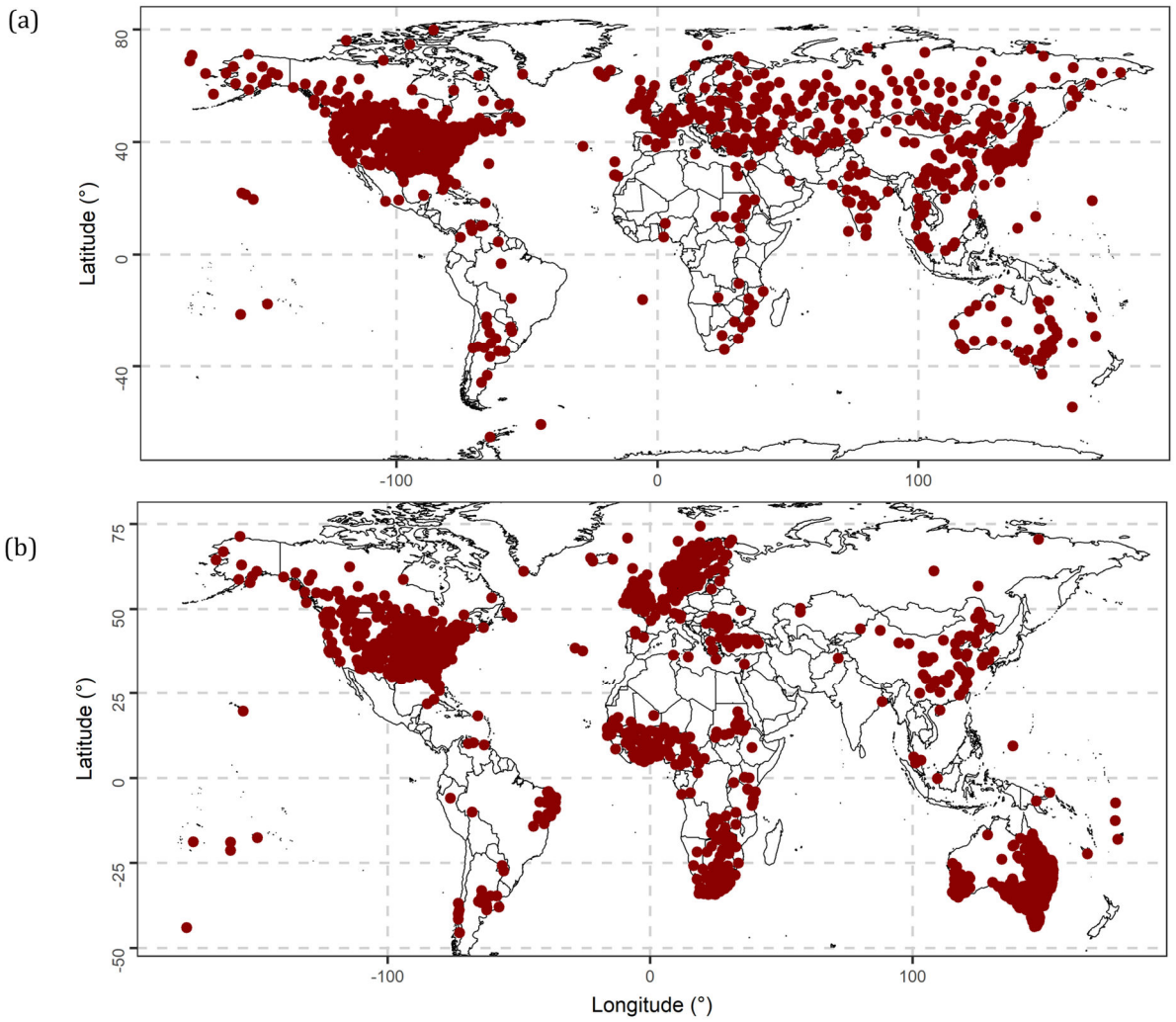


Figure 5.1. Maps of the (a) temperature and (b) precipitation stations; their sources are Lawrimore et al. (2011), and Peterson and Vose (1997), respectively.

Table 5.1. Groups of temperature stations with the respective number of stations per group and regions' geographical boundaries.

Geographical region	Number of stations	Longitude (°)	Latitude (°)
North America	410	[-140, -50]	[20, 65]
North Europe	80	[-15, 40]	[45, 75]
Siberia	70	[40, 175]	[50, 75]
Asia (except Siberia)	259	[40, 150]	[5, 50]
Oceania	36	[105, 170]	[-50, -10]

Table 5.2. Groups of precipitation stations with the respective number of stations per group and regions' geographical boundaries.

Geographical region	Number of stations	Longitude (°)	Latitude (°)
North America	388	[-135, -60]	[20, 55]
North Europe	182	[-15, 35]	[50, 75]
North Africa	100	[-20, 40]	[0, 20]
South Africa	120	[-20, 45]	[-35, 0]
East Asia	50	[95, 135]	[15, 50]
Australia	563	[110, 155]	[-45, - 15]

Moreover, we decompose the time series by using the classical additive model (see Section 2.1.7). This model is implemented through the `R` function `decompose {stats}`. We subsequently fit the fractional Gaussian noise process (see Section 2.1.6) to each seasonally

decomposed time series. This fitting is made by using the maximum likelihood method, implemented through the `R` function `mleHK {HKprocess}`. In Figure 5.2, we present the histograms of the H parameter estimates of the fractional Gaussian noise process for the total of the seasonally decomposed real-world time series, along with the estimated means (denoted with μ) and standard deviations (denoted with σ) of the same process. The magnitude of the long-range dependence is significant in the seasonally decomposed temperature time series, while long-range dependence is also observed in the seasonally decomposed precipitation time series. This is important in the forecasting procedure, since we implement a method that can model long-range dependence and take advantage of this prior knowledge (see Section 5.2.3).

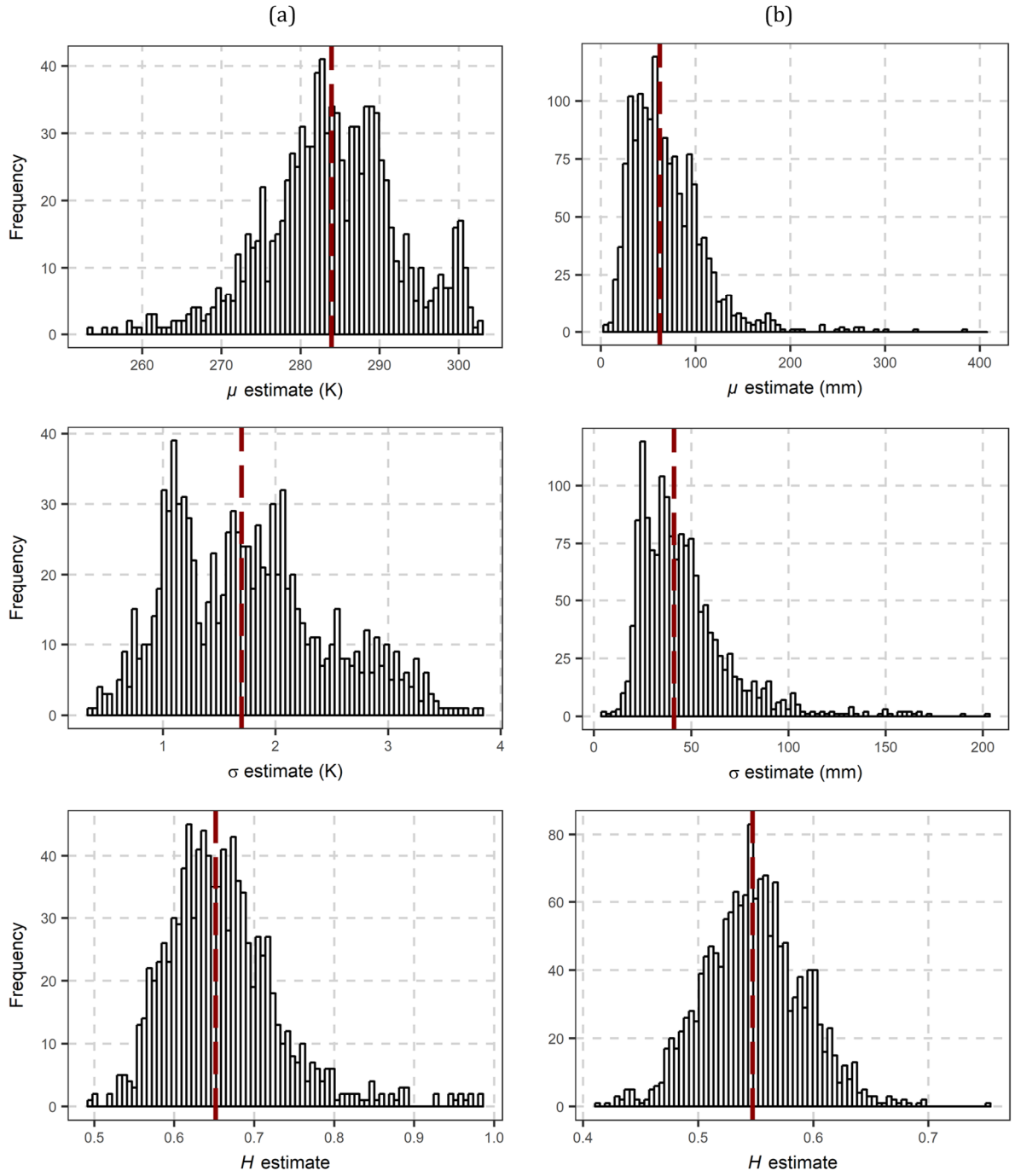


Figure 5.2. Estimates of mean (μ), standard deviation (σ) and Hurst parameter (H) of the fractional Gaussian noise process for the total of the deseasonalized (a) temperature and (b) precipitation time series. The vertical red dashed line denotes the median value of the estimates.

5.2.2 Definition of the forecasting problem

Each time series is forecasted based on its past values. Specifically, we forecast the monthly time series values in the period 1986–1989 based on its values in the period 1950–1985. The observed values in the period 1986–1989 are used for testing the performance of the forecasting methods (and are not used for the fitting of the models). Let also $x_1, x_2, \ldots, x_n$ represent the observations (in the period 1986–1989) and $f_1, f_2, \ldots, f_n$ represent their forecasts.

In Figure 5.3, we present the medians of the observed temperature values to be forecasted for all groups of Table 5.1. The seasonal patterns are obvious in each region, while the minima and maxima clearly depend on the hemisphere. In Figure 5.4, we present a similar illustration for the precipitation time series. The seasonal patterns are again apparent, while zero precipitation appears in Africa regions.

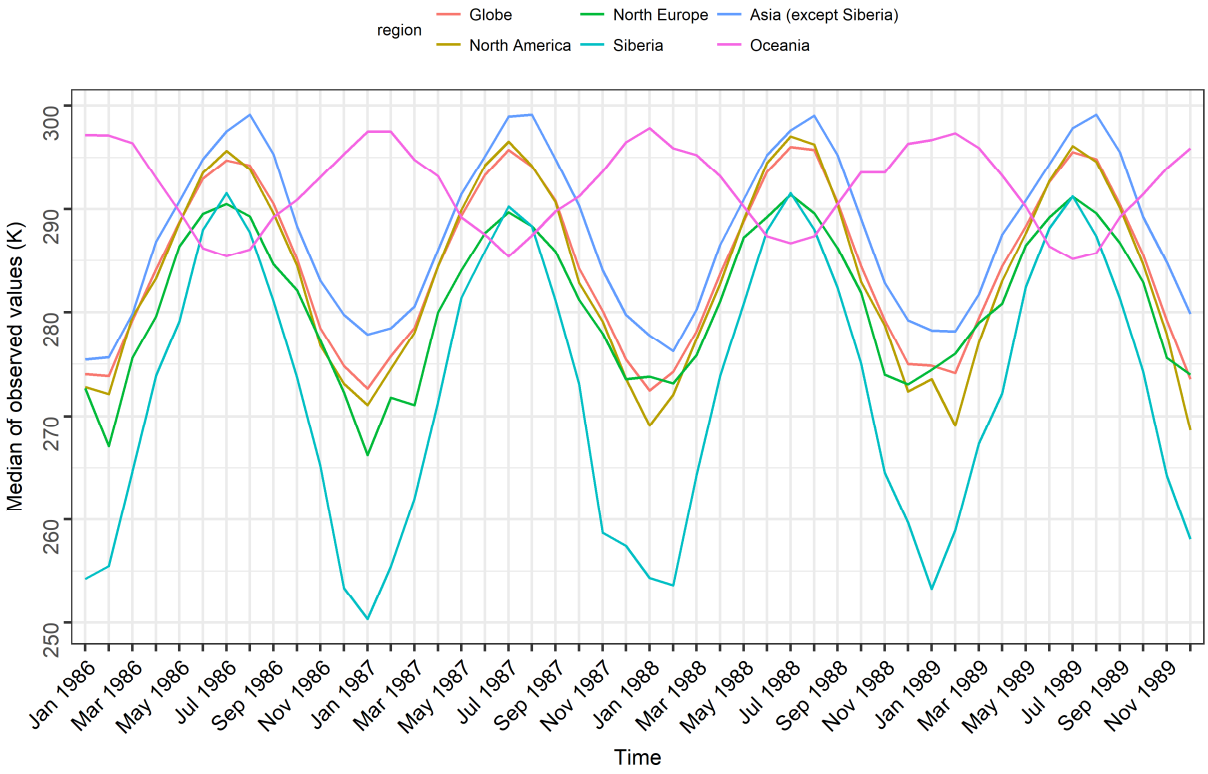


Figure 5.3. Medians of the observed temperature values to be forecasted per group presented in Table 5.1.

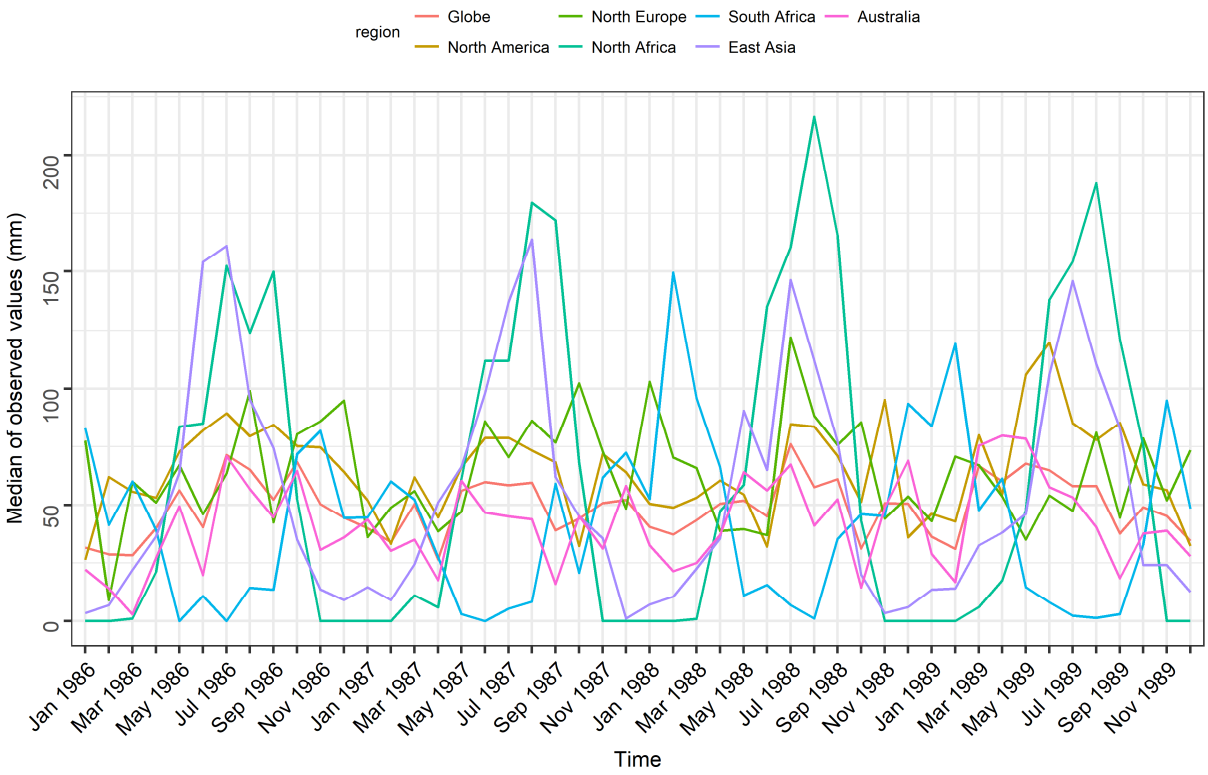


Figure 5.4. Medians of the observed precipitation values to be forecasted per group presented in Table 5.2.

5.2.3 Forecasting methods

We use the seven forecasting methods presented in Table 5.3. In the same Table, we present information on their software implementation, important from a practical point of view, and refer the reader to their documentation. The latter is independently provided in Chapter 2.

Table 5.3. Forecasting methods and the `R` functions used for their implementation. All `R` functions are used with predefined values, unless specified differently. For implementation notes on the `R` functions `rwf {forecast}` and `arfima {forecast}`, see Table 3.3.

S/n	Code name	Corresponding model from Table 2.3	General category	Description	`R` functions
1	naive	Seasonal naïve	Simple	Section 2.2.1	–
2	rw	Random walk			`rwf {forecast}`
3	arfima	Optimum-order autoregressive fractionally integrated moving average (ARFIMA)	ARFIMA	Section 2.2.2	`arfima {forecast}`, `fracdiff {fracdiff}`, `forecast {forecast}`
4	bats	Exponential smoothing state space with Box-Cox transformation, ARMA errors correction, trend and seasonal components (BATS)	Innovations state space	Section 2.2.3	`bats {forecast}`, `forecast {forecast}`
5	ses	Simple exponential smoothing	Exponential smoothing		`ses {forecast}`
6	theta	Theta			`thetaf {forecast}`
7	prophet	Prophet	Curve fitting	Section 2.2.4	`as.Date {zoo}`, `prophet {prophet}`, `make_future_dataframe {prophet}`, `predict.prophet {prophet}`

5.2.4 Seasonality and non-normality

Due to the seasonality and non-normality assumed to characterize the processes underlying monthly temperature and precipitation data, some forecasting methods would be less efficient, if directly applied to the data of this Chapter. Some other methods can automatically transform the data, without external handling.

When the examined processes are characterized by seasonality, two possible transformations of the exploited data before applying the forecasting methods (see Section 5.2.3) are the classical additive and multiplicative seasonal decomposition approaches (see Section 2.1.7). These transformations are adopted herein under the flexible methodology for time series forecasting through time series decomposition (see Section 2.2.6), which is hereafter referred to as "external handling of seasonality". In the particular case of the multiplicative seasonal decomposition of the precipitation time series, we add 10 mm to each value, since zero observed values would result in zeros during the seasonal decomposition. Obviously, the inverse transform involves the subtraction of 10 mm for the forecasted values. This approach may affect the decomposition pattern; however, it is the most practical in this case.

Some methods are applied under the normality assumption. Since non-normality is often assumed for the processes underlying the monthly geophysical data, an appropriate transformation of such data could be used to obtain possible predictive performance improvements. The most popular relevant transformation, also adopted in this Chapter, is the Box-Cox one, defined in Section 2.1.8.

In Table 5.4, we present the variants of the forecasting methods (see Section 5.2.3). The variants depend on whether a transformation is used and on which this specific transformation is. In Table 5.5, we present the transformations accounting for seasonality, while in Table 5.6 we present the transformations accounting for non-normality. Each variant of the method is assigned to a particular combination of the transformations presented in Table 5.5 and Table 5.6,

respectively. When a transformation is applied, then the forecasts are obtained through the inverse transform. The BATS and the Prophet methods can also handle the seasonal patterns automatically. Therefore, their variants include cases in which the time series are seasonally decomposed manually, as well as cases in which seasonality is considered through an automatic scheme.

Table 5.4. Variants of the methods of Table 5.3.

S/n	Abbreviated name	Primal method (see Table 5.3)	Handling of seasonality (see Table 5.5)	Handling of non-normality (see Table 5.6)
1	naïve	1	1	1
2	rw_1	2	2	1
3	rw_2	2	2	2
4	rw_3	2	3	1
5	rw_4	2	3	2
6	arfima_1	3	2	1
7	arfima_2	3	2	2
8	arfima_3	3	3	1
9	arfima_4	3	3	2
10	bats_1	4	2	1
11	bats_2	4	2	2
12	bats_3	4	3	1
13	bats_4	4	3	2
14	bats_5	4	4	1
15	bats_6	4	4	2
16	ses_1	5	2	1
17	ses_2	5	2	2
18	ses_3	5	3	1
19	ses_4	5	3	2
20	theta_1	6	2	3
21	theta_2	6	3	3
22	prophet_1	7	2	3
23	prophet_2	7	3	3
24	prophet_3	7	4	3

Table 5.5. Choices for the handling of seasonality.

S/n	Handling	Corresponding additional model from Table 2.3	Additional R functions
1	Time series offset	–	–
2	Classical seasonal decomposition using the additive model and subsequent addition of the seasonal component to the forecasts	Additive model	`ts {stats}`, `decompose {stats}`
3	Classical seasonal decomposition using the multiplicative model and subsequent multiplication of the forecasts by the seasonal component	Multiplicative model	
4	Through the forecasting algorithm	–	–

Table 5.6. Choices for the handling of non-normality.

S/n	Handling	Corresponding additional model from Table 2.3	Additional R function
1	–	–	–
2	Box-Cox transformation through the forecasting algorithm	Box-Cox transformation	`BoxCox.lambda {forecast}`
3	Default	–	–

5.2.5 Forecast quality assessment

At each time step i of the forecast horizon, we compute the error and absolute error for each forecasting attempt, denoted with E_i and AE_i, respectively. The computation is made, as prescribed by the definitions provided in Section 2.8.2. We also compute, separately for each time step i of the forecast horizon, the median of the AE_i values. Finally, we compute the root mean square error

(RMSE) and the Nash-Sutcliffe efficiency (NSE) of each multi-step ahead forecast. The RMSE and NSE metrics are defined in Section 2.8.2.

5.3 Results

5.3.1 Experiments using the temperature time series

Section 5.3.1 is devoted to the results of the analysis using the temperature time series. In Figure 5.5, we present the side-by-side boxplots of the errors at each time step of the forecast horizon, as formed for all the temperature forecasts produced by the naïve, rw_1 and prophet_1 methods. The random walk variants create a similar image to each other (see the one of rw_1), while the same applies to the set of ARFIMA, BATS, simple exponential smoothing, Theta and Prophet variants (see the one of prophet_1). To illustrate this closeness in the forecasting performance of the methods, in Figure 5.6 we present the median values of the absolute errors computed for the total of the temperature time series. The random walk variants are the least accurate at almost every step of the examined horizon with a minimum median of absolute errors approximately equal to 1.3 K and a maximum approximately equal to 4 K, while naïve is also worse than the rest of the automatic methods with minimum and maximum medians approximately equal to 0.8 K and 4.2 K, respectively. The best median performance is approximately equal to 0.5 K, which is about 70% smaller than the median of the estimated standard deviations of the deseasonalized time series (see Figure 5.2). We further observe that the presented time series tend rather to run in parallel than to intersect each other and, therefore, the good/bad forecasts of the various methods are rather grouped in the horizontal direction. This behaviour may be explained by the fact that the magnitude of the error of the forecast produced by a specific method largely depends on the value to be forecasted, i.e., some forecasting attempts are by nature more difficult than others. As a result, the worst median performance of each and every of the methods is observed for January 1989, a month exhibiting higher temperature than the one expected from seasonality (see Figure 5.3), while the second worst for February 1989 for the opposite reason.

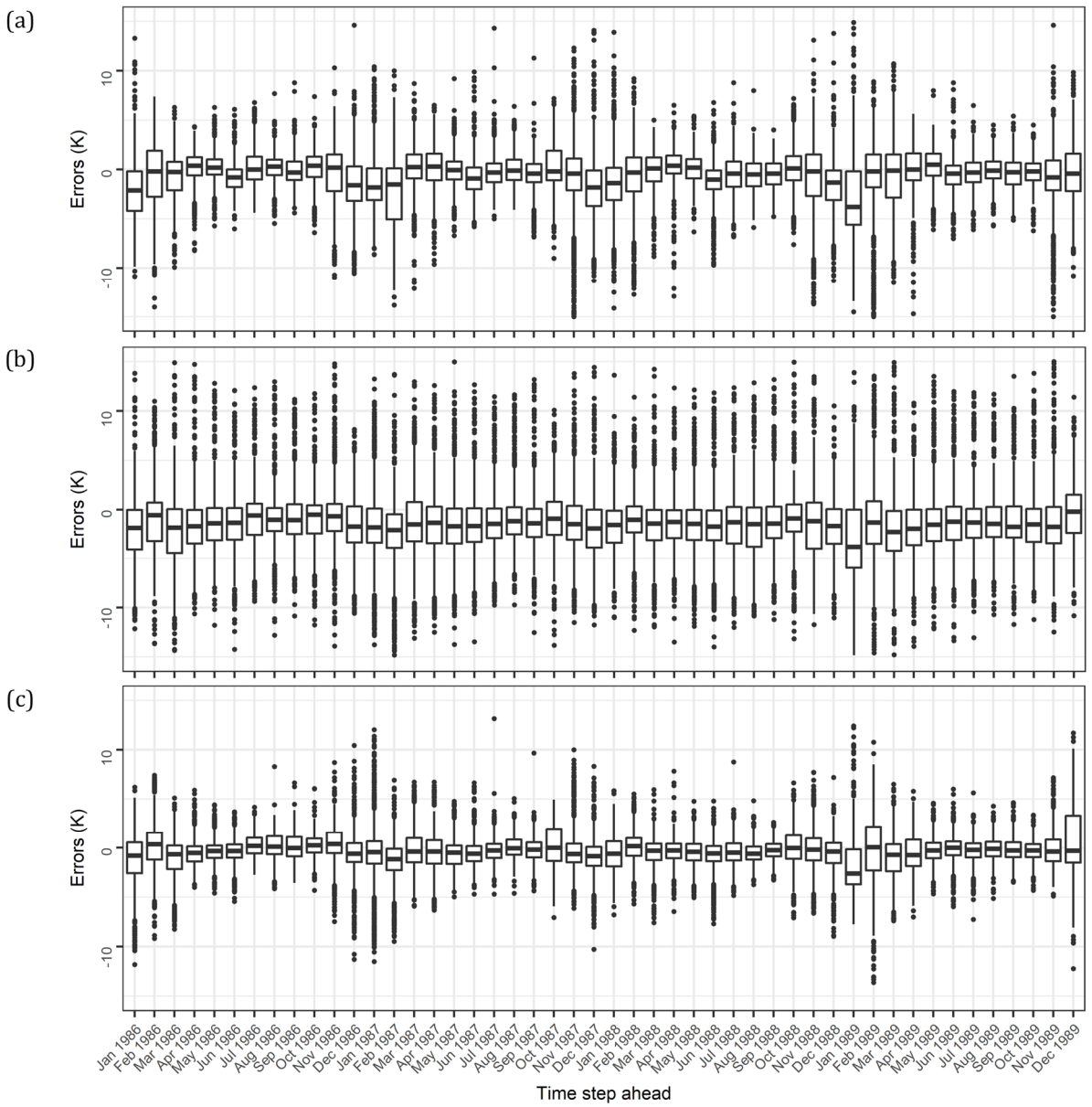


Figure 5.5. Errors at each time step of the forecast horizon for the total of the temperature time series, and the (a) naïve, (b) rw_1 and (c) prophet_1 methods. The outliers with absolute value larger than 15 K are omitted.

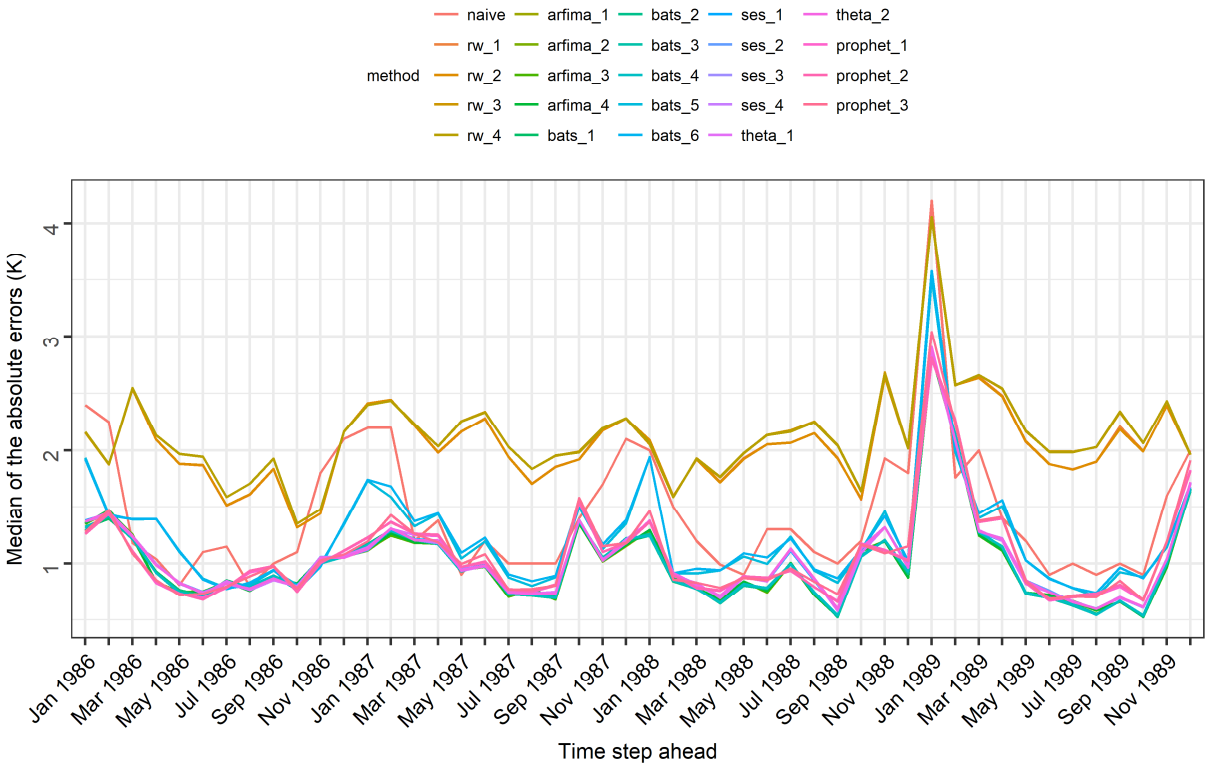


Figure 5.6. Medians of the absolute errors at each time step of the forecast horizon for the total of the temperature time series.

In Figures 5.7 and 5.8, we present the median values of the absolute errors computed for each of the groups of stations of Table 5.1. Since most of the temperature time series are observed in North America, the results corresponding to this geographical region affect the total (presented in Figure 5.6) to a significant extend. In fact, Figure 5.6 is more similar to Figure 5.7(a) than to Figure 5.7(b,c) or Figure 5.8. However, the medians of the absolute errors are larger in North America than worldwide. In the remaining geographical regions, the performance of the random walk methods are closer to the performance of the rest automatic methods, while for the specific cases of North Europe and Siberia the random walk variants are better than naïve as well. The best average performances are measured for Oceania with a minimum median of absolute errors approximately equal to 0.25 K and a maximum around 2.1 K, while the respective values for Asia (except Siberia) (approximately equal to 0.30 K and 4 K) are also better than the overall.

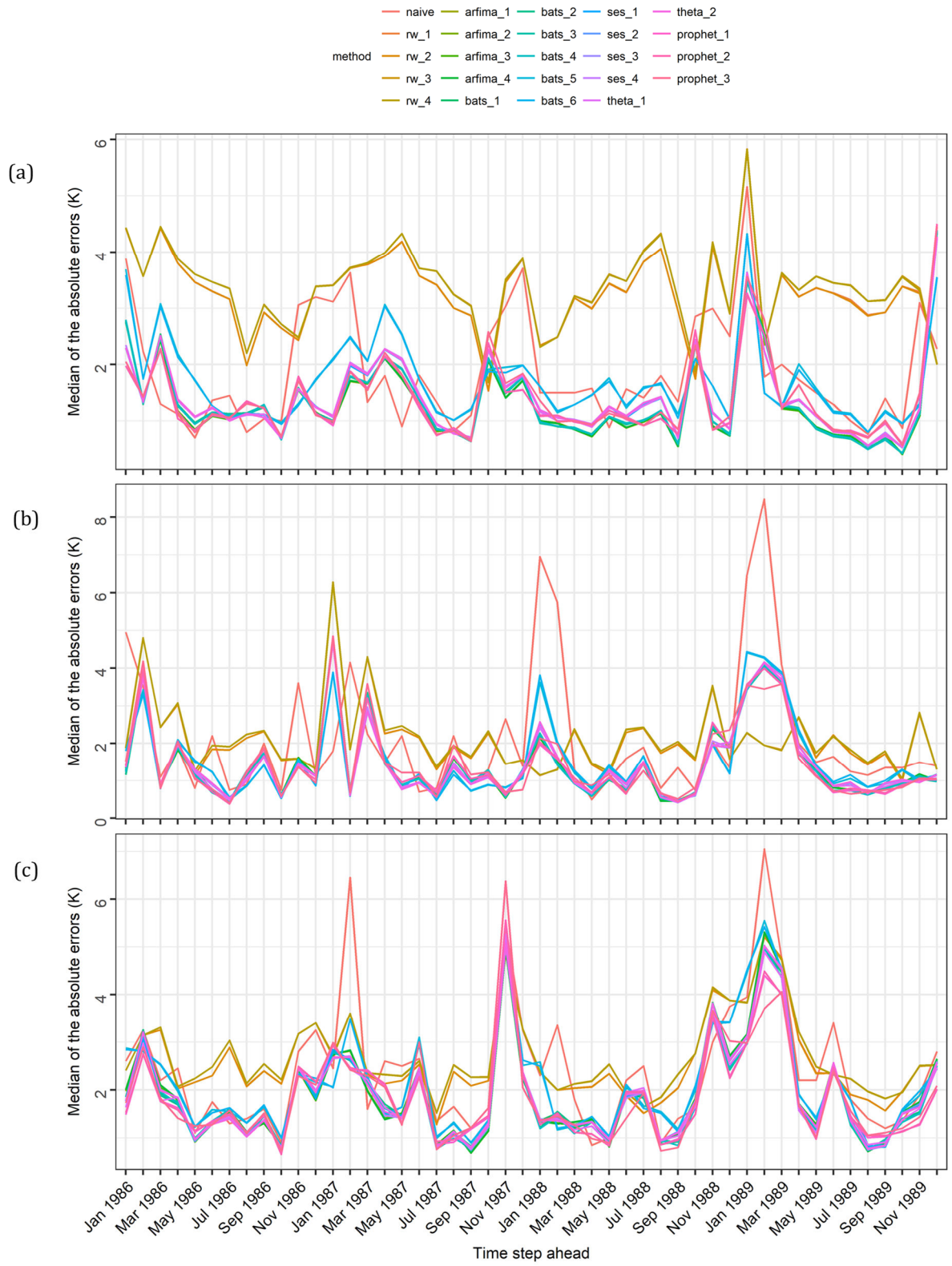


Figure 5.7. Medians of the absolute errors at each time step of the forecast horizon for the temperature time series observed in: (a) North America, (b) North Europe and (c) Siberia.

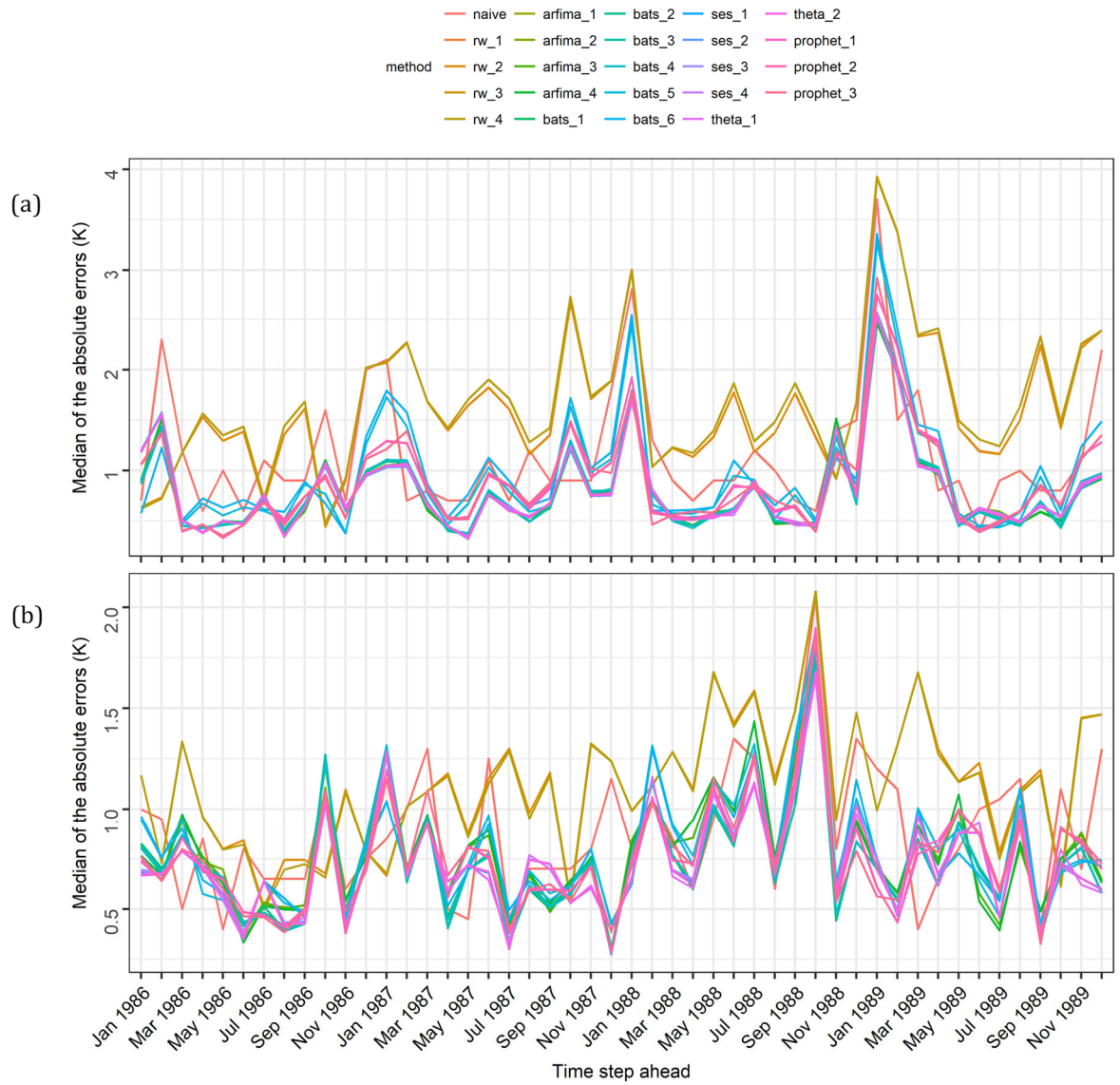


Figure 5.8. Medians of the absolute errors at each time step of the forecast horizon for the temperature time series observed in: (a) Asia (except Siberia) and (b) Oceania.

In Tables 5.7 and 5.8, we present the medians of the measured RMSE and NSE values of the global dataset and for each of the groups of stations of Table 5.1, while the side-by-side boxplots of the RMSE values are presented in Figure 5.9. These numerical results can facilitate a comparison on a common basis of the methods regarding their performance in the experiments using the temperature time series. In terms of RMSE, for the total of the temperature time series the use of a random walk variant (offering a median of 2.60 K or 2.66 K) instead of naïve (offering a median of 2.29 K) leads to about 14–16% less accurate forecasts, while the use of the remaining automatic methods (offering median values between 1.62 K and 1.86 K) to about 19–29% more accurate forecasts. For the time series observed in North America, North Europe, Siberia, Asia (except Siberia) and Oceania these latter percentages are 18–30%, 30–32%, 19–25%, 17–31% and 10–15% respectively. The median values of the latter methods are close to the median of the estimated standard deviations of the deseasonalized time series (see Figure 5.2). On the other hand, all the variants of the ARFIMA, BATS, simple exponential smoothing, Theta and Prophet methods except for the bats_5, bats_6 and prophet_3 are rather competitive to each other, while each of these categories of methods exhibits better or worse performance in comparison to the rest depending on the examined sample of time series. For example, prophet_1 exhibits the smallest median RMSE for the temperature forecasts for North Europe and Siberia, while offering

13–32% (depending on the examined set of time series) more accurate results than naïve. The variants belonging to each of the {arfima_1, arfima_2, arfima_3, arfima_4}, {bats_1, bats_2, bats_3, bats_4}, {ses_1, ses_2, ses_3, ses_4}, {theta_1, theta_2}, {prophet_1, prophet_2} sets differ in their performance at maximum about 1%, while the results do not suggest any specific combination of choices for the external handling of seasonality and non-normality as best for each algorithm. Nevertheless, the handling of the seasonality through the BATS and Prophet forecasting algorithms leads to less accurate forecasts than the external one, especially for the former algorithm. These facts are illustrated in Figure 5.9 as well. Finally, all NSE values of Table 5.8 indicate a good forecasting performance for the total of the methods.

Table 5.7. Medians of the RMSE values (K) of the forecasts for each group of temperature stations. The best performance (when rounding to more than four digits) for each model is in bold.

Method	Globe	North America	North Europe	Siberia	Asia (except Siberia)	Oceania
naïve	**2.29**	**2.70**	**3.14**	**3.49**	**1.61**	**1.19**
rw_1	**2.60**	3.59	2.84	3.22	**1.99**	1.43
rw_2	2.60	**3.57**	**2.84**	**3.22**	1.99	1.43
rw_3	2.66	3.65	2.84	3.26	2.02	**1.42**
rw_4	2.66	3.64	2.84	3.25	2.02	1.42
arfima_1	**1.63**	1.90	2.13	2.66	1.11	**1.03**
arfima_2	1.63	**1.90**	2.13	2.66	**1.11**	1.04
arfima_3	1.63	1.90	2.13	2.66	1.11	1.03
arfima_4	1.63	1.90	**2.13**	**2.66**	1.11	1.06
bats_1	1.62	**1.92**	2.13	**2.66**	**1.11**	1.01
bats_2	1.62	1.93	**2.13**	2.67	1.12	1.01
bats_3	**1.62**	1.92	2.13	2.68	1.12	1.01
bats_4	1.64	1.94	2.13	2.67	1.12	**1.01**
bats_5	1.84	2.21	2.16	2.84	1.26	1.07
bats_6	1.86	2.20	2.18	2.84	1.33	1.04
ses_1	1.68	1.99	2.15	**2.66**	1.12	1.02
ses_2	1.68	**1.99**	**2.15**	2.67	**1.12**	1.02
ses_3	**1.68**	2.00	2.15	2.67	1.12	**1.02**
ses_4	1.68	2.00	**2.15**	2.68	1.12	1.02
theta_1	1.68	**1.99**	2.17	**2.64**	1.12	**1.01**
theta_2	**1.68**	2.00	**2.15**	2.67	**1.12**	1.02
prophet_1	**1.70**	**1.99**	**2.12**	**2.62**	**1.25**	**1.03**
prophet_2	1.71	1.99	2.13	2.63	1.25	1.03
prophet_3	1.75	2.04	2.19	2.70	1.26	1.03

Table 5.8. Medians of the NSE values of the forecasts for each group of temperature stations. The best performance (when rounding to more than four digits) for each model is in bold.

Method	Globe	North America	North Europe	Siberia	Asia (except Siberia)	Oceania
naïve	**0.91**	**0.90**	**0.79**	**0.93**	**0.95**	**0.90**
rw_1	0.89	0.81	0.83	0.94	**0.93**	0.89
rw_2	**0.89**	**0.81**	**0.83**	**0.94**	0.93	0.89
rw_3	0.89	0.80	0.82	0.94	0.93	0.89
rw_4	0.89	0.80	0.82	0.94	0.93	**0.89**
arfima_1	0.95	0.95	0.91	0.96	0.98	**0.93**
arfima_2	0.95	**0.95**	**0.91**	0.96	**0.98**	0.93
arfima_3	**0.95**	0.95	0.91	0.96	0.98	0.93
arfima_4	0.95	0.95	0.91	**0.96**	0.98	0.92
bats_1	**0.95**	0.95	0.91	0.96	0.98	**0.93**
bats_2	0.95	0.95	**0.91**	0.96	0.98	0.93
bats_3	0.95	**0.95**	0.91	**0.96**	**0.98**	0.93
bats_4	0.95	0.95	0.91	0.96	0.98	0.93
bats_5	0.94	0.93	0.89	0.95	0.97	0.93
bats_6	0.94	0.93	0.89	0.95	0.97	0.93
ses_1	**0.95**	**0.95**	**0.90**	0.96	0.98	0.93
ses_2	0.95	0.95	0.90	0.96	0.98	**0.93**
ses_3	0.95	0.95	0.90	0.96	0.98	0.93
ses_4	0.95	0.95	0.90	**0.96**	**0.98**	0.93
theta_1	**0.95**	**0.95**	**0.90**	**0.96**	0.98	**0.93**
theta_2	0.95	0.95	0.90	0.96	**0.98**	0.93
prophet_1	**0.95**	**0.95**	0.91	0.96	**0.97**	0.93
prophet_2	0.95	0.95	**0.91**	**0.96**	0.97	0.93
prophet_3	0.95	0.95	0.90	0.96	0.97	**0.93**

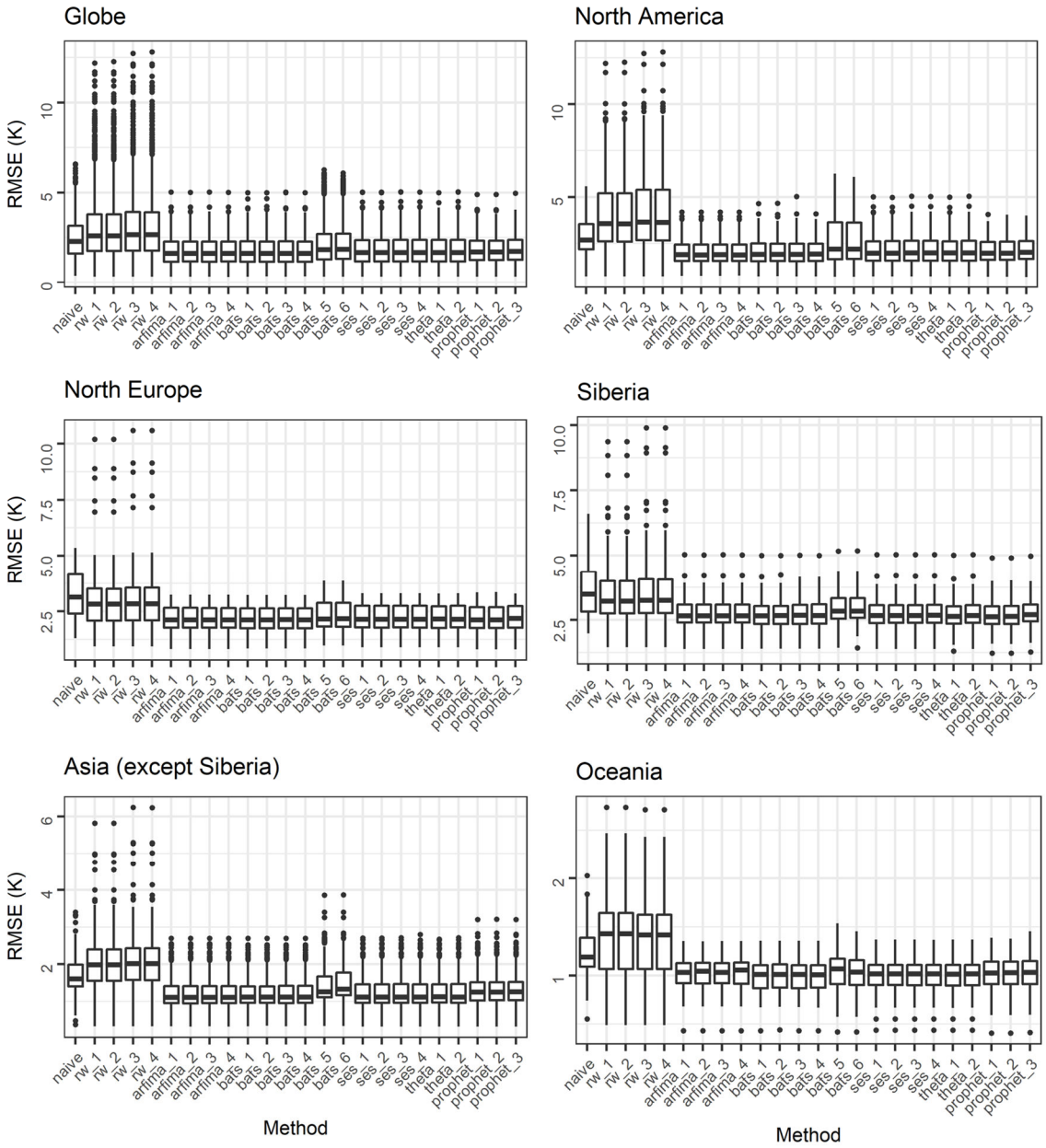


Figure 5.9. RMSE for the temperature time series.

5.3.2 Experiments using the precipitation time series

This Section is devoted to the results of the analysis using the precipitation time series. In Figure 5.10, we present the median values of the absolute errors computed for the total of the precipitation time series. Here as well, the random walk variants are the least accurate at almost every step of the examined horizon with a minimum median around 25 mm and a maximum around 48 mm. The naïve method is also worse than the rest with minimum and maximum medians around 20 mm and 39 mm respectively. The best average performance is around 15 mm, which is about 70% smaller than the median of the estimated standard deviations of the deseasonalized time series (see Figure 5.2), the same as applying to the temperature forecasts. Moreover, in Figures 5.11 and 5.7 we present the median values of the absolute errors computed for each of the groups of stations of Table 5.2. First, we observe that Figure 5.10 approximates more to Figure 5.12(c) than to Figure 5.11 or Figure 5.12(a,b). This is a rather expected outcome, since most of the totally examined precipitation time series originate from Australia. Nevertheless, the medians of the absolute errors are larger in this geographical region than

worldwide. Furthermore, in Australia and North Europe there is not a clear pattern of seasonality in the presented medians, while for the case of North America there is one, but only for the years 1988 and 1989. In North Africa and East Asia, on the contrary, the medians between April and October are clearly higher than for the rest of the months with a peak in August (or July), indicating a larger difficulty in their corresponding forecasting attempts. This is due to a more regular precipitation variability in these geographical regions. The same applies, to a smaller extend, to the case of South Africa, for which the precipitation variables between October and April are found to be the least predictable. Finally, in North Africa, South Africa and East Asia some medians of the absolute errors are equal or very close to zero. All the above-stated facts may be largely explained in Figure 5.4.

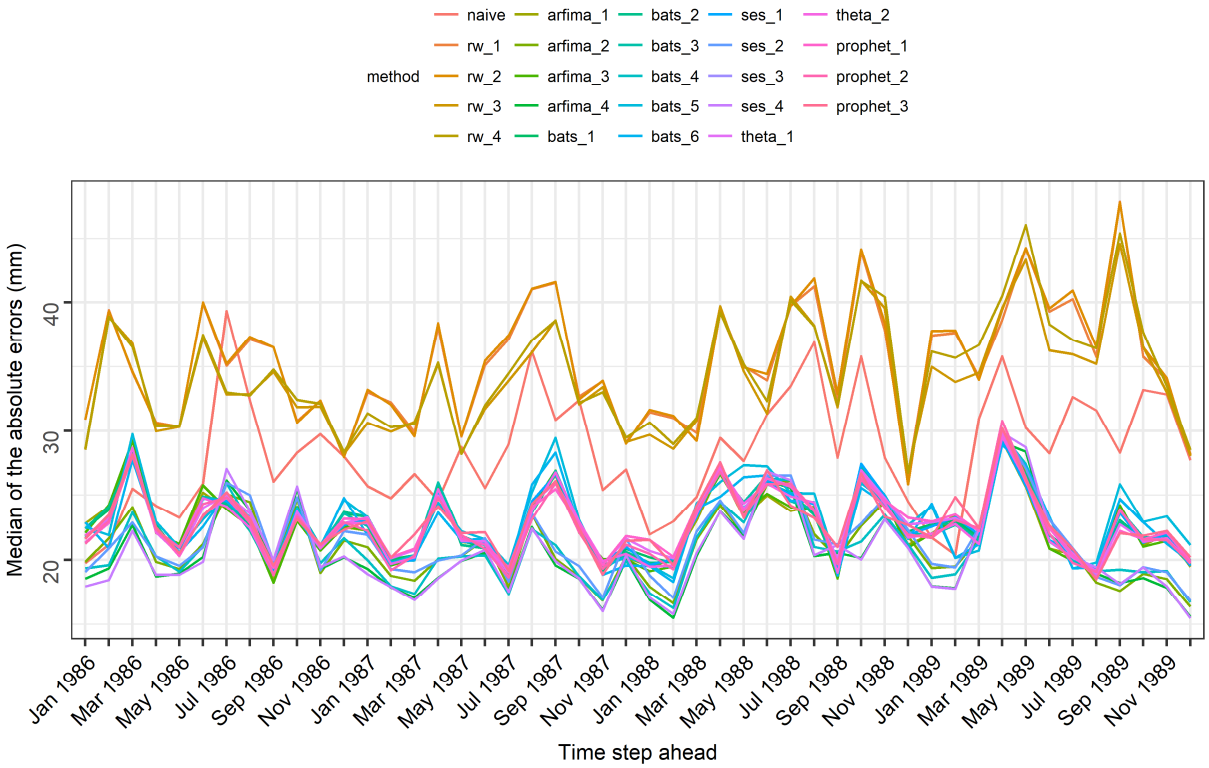


Figure 5.10. Medians of the absolute errors at each time step of the forecast horizon for the total of the precipitation time series.

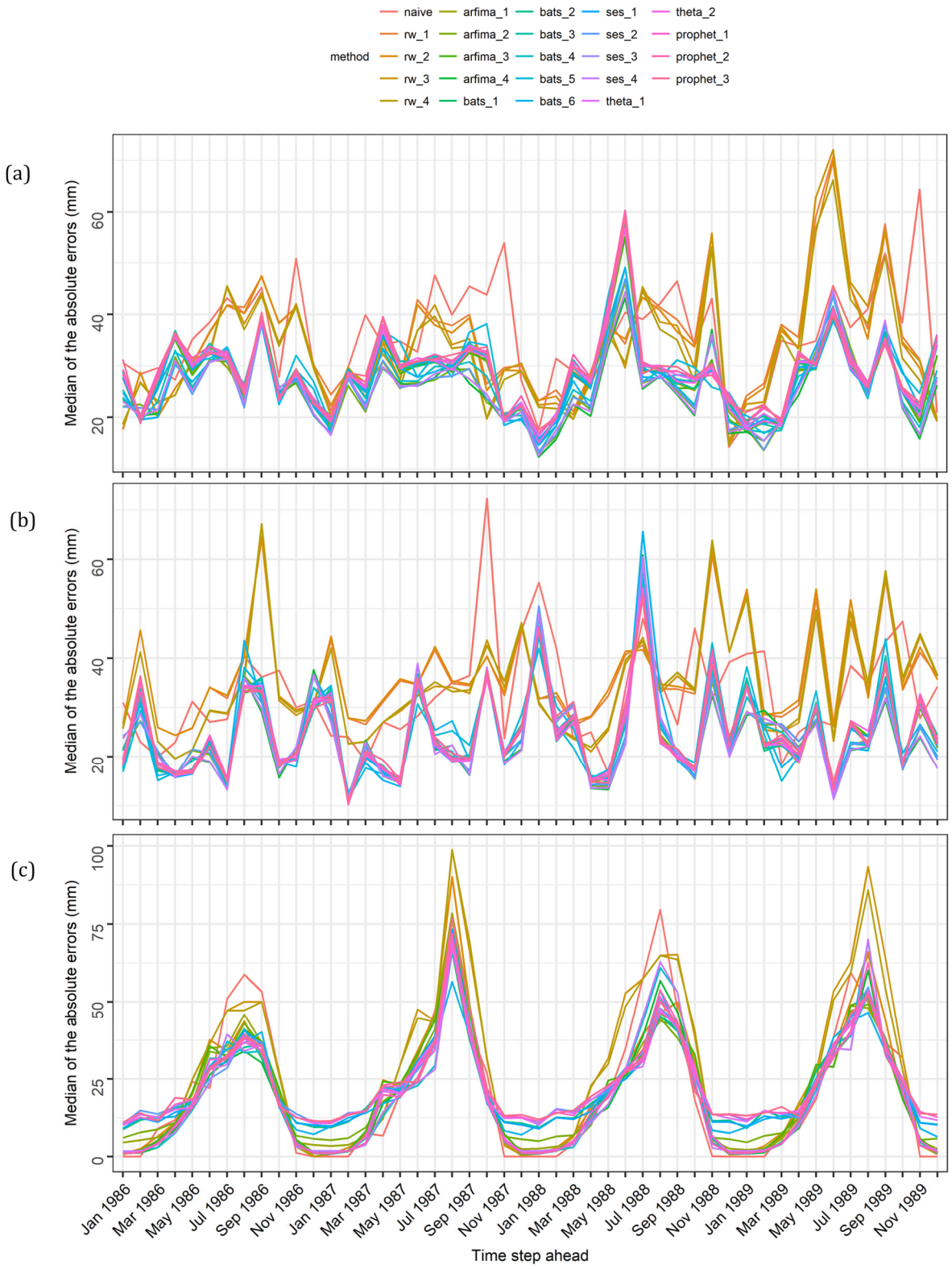


Figure 5.11. Medians of the absolute errors at each time step of the forecast horizon for the precipitation time series observed in: (a) North America, (b) North Europe and (c) North Africa.

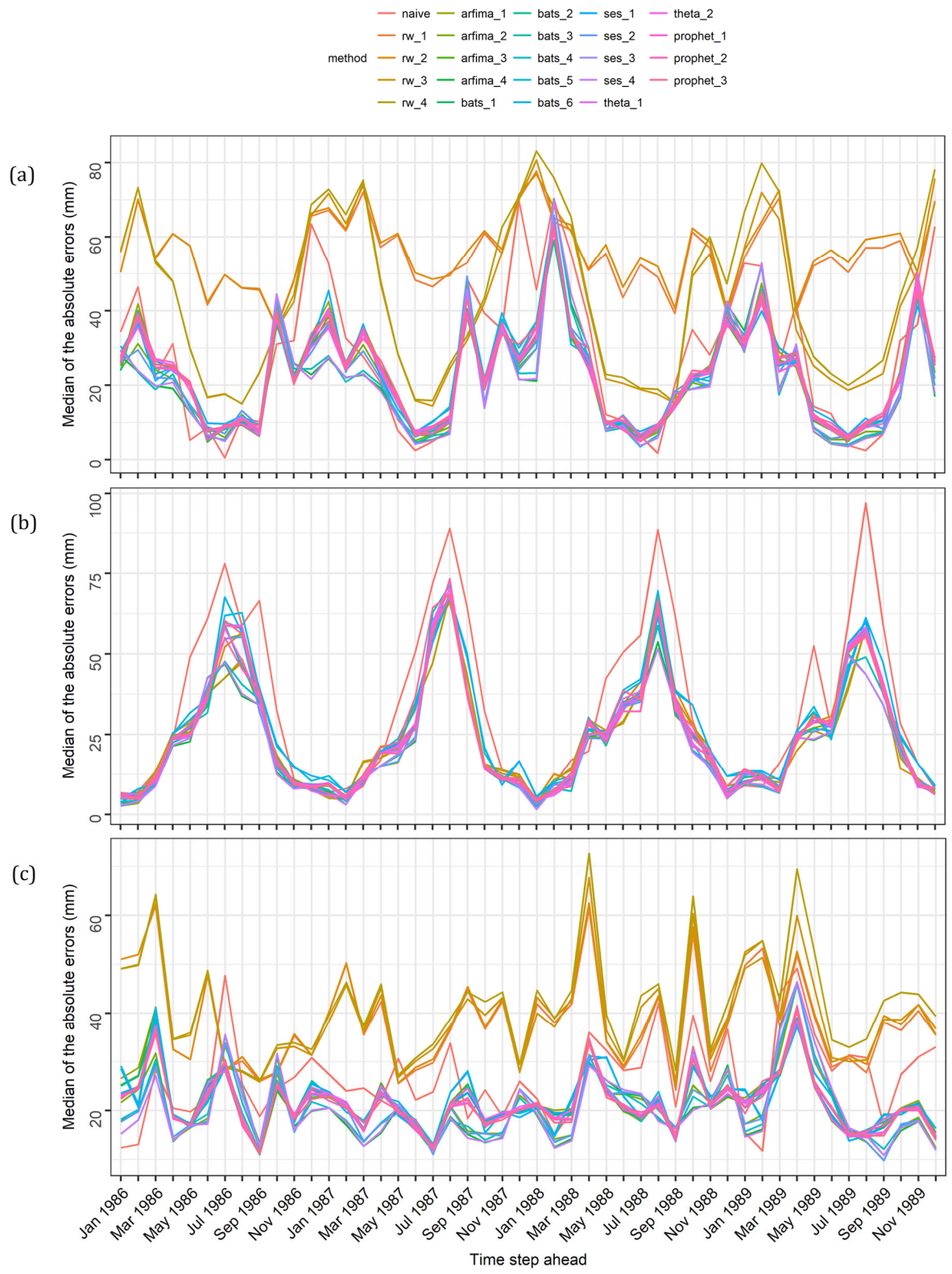


Figure 5.12. Medians of the absolute errors at each time step of the forecast horizon for the precipitation time series observed in: (a) South Africa, (b) East Asia and (c) Australia.

In Figures 5.13 and 5.14, we present in more detail the medians of the absolute errors for two special cases, i.e. those of North and South Africa respectively. In these figures we individually compare the {rw_1, rw_2, rw_3, rw_4}, {arfima_1, arfima_2, arfima_3, arfima_4}, {bats_1, bats_2 bats_3, bats_4, bats_5, bats_6}, {ses_1, ses_2, ses_3, ses_4}, {theta_1, theta_2}, {prophet_1, prophet_2, prophet_3} sets of variants. As illustrated in Figure 5.13, in North Africa the tested

choices for handling the seasonality seem to result in different seasonality patterns in the median values of the absolute errors, a fact not applying to the choices for handling the non-normality. Particularly for this specific geographical region the use of the additive model results in larger absolute errors than the multiplicative model from October to April and to smaller absolute errors for the rest of the year. These differences are more visible for the BATS, simple exponential smoothing, Theta and Prophet variants, but also exist for the random walk and ARFIMA ones. In South Africa two discrete seasonality patterns are observed only for the random walk variants.

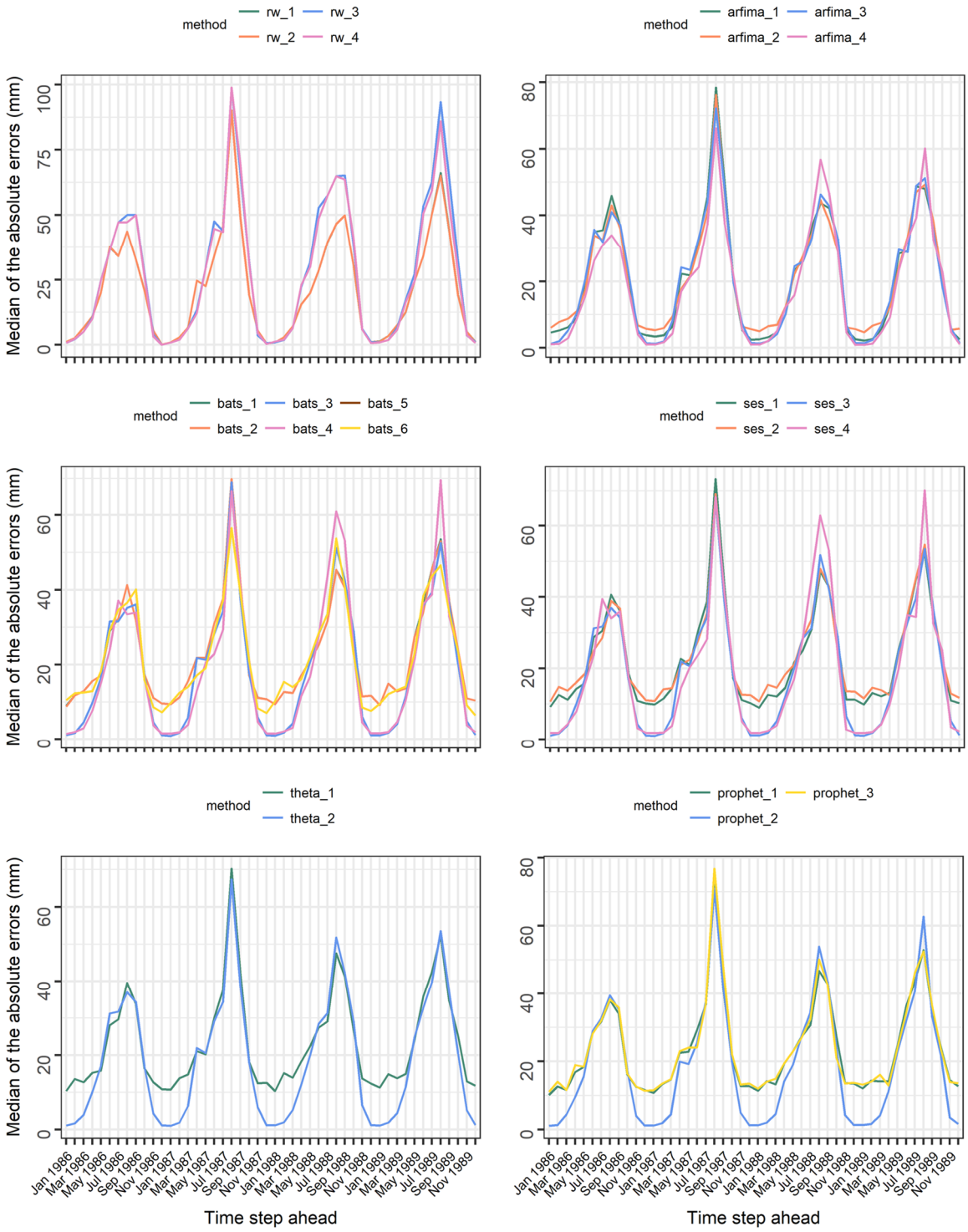


Figure 5.13. Medians of the absolute errors at each time step of the forecast horizon for the precipitation time series observed in North Africa: comparison among the methods using the same model.

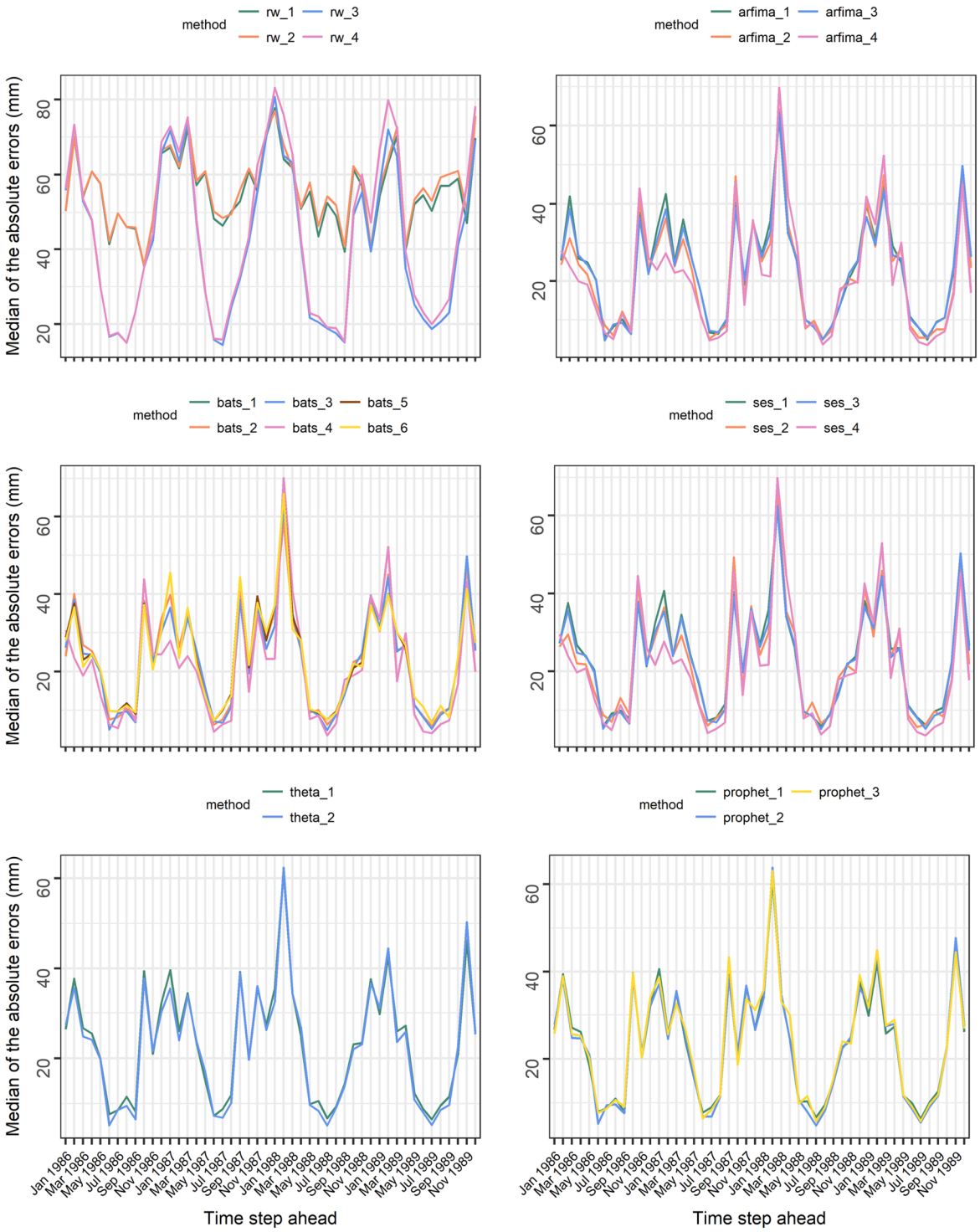


Figure 5.14. Medians of the absolute errors at each time step of the forecast horizon for the precipitation time series observed in South Africa: comparison among the methods using the same model.

Additionally, in Tables 5.9 and 5.10 we present the medians of the measured RMSE and NSE values of the global dataset and for each of the groups of stations of Table 5.2, while the side-by-side boxplots of the RMSE values are presented in Figure 5.15. In the latter we notice the similarity in the performance of the ARFIMA, BATS, simple exponential smoothing, Theta and Prophet variants, which is also reported for the analysis of Section 5.3.1. Some (absolute and relative) differences in the forecasting performance of the methods are also evident. For example, for the case of North Africa rw_1 and rw_2 are more accurate than rw_3 and rw_4, while the opposite

applies to the case of North Europe albeit to a smaller extend. By the examination of Table 5.9 we observe that for the total of the precipitation time series the use of all the automatic methods apart from the random walk variants (offering median values between 41.67 mm and 42.39 mm) instead of naïve (offering a median value of 53.74 mm) leads to about 21–22% more accurate forecasts in terms of RMSE. For the time series observed in North America, North Europe and East Asia these percentages are 26–29%, 22–24% and 32–38% respectively, while for those observed in North Africa, South Africa and Australia 18–25%, 15–18% and 19–22% respectively.

Table 5.9. Medians of the RMSE values (mm) of the forecasts for each group of precipitation stations. The best performance for each model is in bold.

Method	Globe	North America	North Europe	North Africa	South Africa	East Asia	Australia
naïve	**53.74**	**63.20**	**47.89**	**59.91**	**58.84**	**75.61**	**46.38**
rw_1	**53.65**	**55.35**	48.46	**47.82**	69.97	**48.69**	**51.70**
rw_2	54.04	55.39	48.50	47.82	70.17	48.70	53.18
rw_3	56.00	56.68	**46.89**	61.11	**66.18**	54.13	55.36
rw_4	56.53	56.53	47.47	58.88	67.33	55.22	57.46
arfima_1	41.75	**45.16**	**36.65**	46.18	48.34	48.57	36.51
arfima_2	42.07	45.29	37.26	**45.27**	48.81	47.98	37.49
arfima_3	**41.67**	45.61	36.65	45.36	**48.20**	**47.79**	**36.25**
arfima_4	42.01	45.34	37.09	46.19	49.60	49.07	37.26
bats_1	**41.88**	45.78	36.59	46.02	48.85	**47.56**	**36.21**
bats_2	41.90	45.62	36.59	45.75	48.85	**47.56**	**36.21**
bats_3	41.98	46.15	**36.54**	**45.15**	**48.56**	48.01	36.51
bats_4	42.06	**45.28**	36.69	47.69	50.04	48.85	36.75
bats_5	42.39	45.80	36.78	47.50	49.99	51.71	37.28
bats_6	42.34	45.56	37.52	47.50	49.99	50.77	37.13
ses_1	41.88	45.54	36.60	45.77	48.83	48.07	36.35
ses_2	42.23	45.49	36.78	46.16	49.17	48.03	37.45
ses_3	**41.79**	45.90	**36.30**	**45.17**	**48.40**	**47.87**	**36.16**
ses_4	42.13	**45.39**	36.95	48.93	50.29	49.12	37.26
theta_1	42.08	46.24	36.87	46.09	48.95	47.22	36.50
theta_2	**41.79**	**45.90**	**36.30**	**45.17**	**48.40**	**47.87**	**36.16**
prophet_1	42.16	**46.22**	37.06	**46.03**	**48.70**	47.18	36.56
prophet_2	**41.85**	46.72	**36.84**	46.26	48.89	**47.08**	**36.31**
prophet_3	42.34	46.54	36.90	46.19	49.26	51.21	36.56

Table 5.10. Medians of the NSE values of the forecasts for each group of precipitation stations. The best performance for each model is in bold.

Method	Globe	North America	North Europe	North Africa	South Africa	East Asia	Australia
naïve	**−0.38**	**−0.55**	**−0.45**	**0.54**	**−0.04**	**0.05**	**−0.44**
rw_1	**−0.17**	−0.28	−0.14	**0.68**	−0.20	**0.49**	**−0.33**
rw_2	−0.17	**−0.24**	−0.15	0.68	−0.20	0.49	−0.34
rw_3	−0.20	−0.34	**−0.09**	0.49	**0.01**	0.42	−0.36
rw_4	−0.20	−0.30	−0.11	0.51	−0.02	0.44	−0.37
arfima_1	**0.15**	**0.10**	**0.12**	0.69	0.29	0.49	0.04
arfima_2	0.11	0.09	0.08	0.69	0.25	0.50	−0.04
arfima_3	0.14	0.09	0.12	0.70	**0.30**	0.49	**0.05**
arfima_4	0.08	0.08	0.05	**0.71**	0.18	**0.51**	−0.04
bats_1	0.14	0.08	**0.13**	0.70	**0.29**	0.49	**0.04**
bats_2	**0.14**	0.08	**0.13**	0.70	**0.29**	0.49	**0.04**
bats_3	0.14	0.08	0.13	**0.71**	0.29	0.50	0.04
bats_4	0.11	**0.10**	0.11	0.69	0.18	**0.52**	−0.01
bats_5	0.11	0.08	0.10	0.70	0.25	0.45	0.01
bats_6	0.10	0.07	0.05	0.70	0.25	0.44	0.01
ses_1	**0.14**	0.09	**0.12**	0.70	0.25	0.49	0.04
ses_2	0.10	**0.09**	0.08	0.69	0.22	0.50	−0.04
ses_3	0.14	0.08	0.12	**0.71**	0.29	0.49	**0.05**
ses_4	0.09	0.09	0.05	0.68	0.16	**0.51**	−0.04
theta_1	0.14	0.08	**0.13**	0.69	0.29	**0.49**	0.04
theta_2	**0.14**	**0.08**	0.12	**0.71**	**0.29**	0.49	**0.05**
prophet_1	**0.14**	**0.07**	0.13	0.69	0.28	0.50	0.04
prophet_2	0.14	0.07	**0.13**	**0.70**	**0.29**	**0.50**	**0.04**
prophet_3	0.13	0.07	0.11	0.68	0.28	0.49	0.03

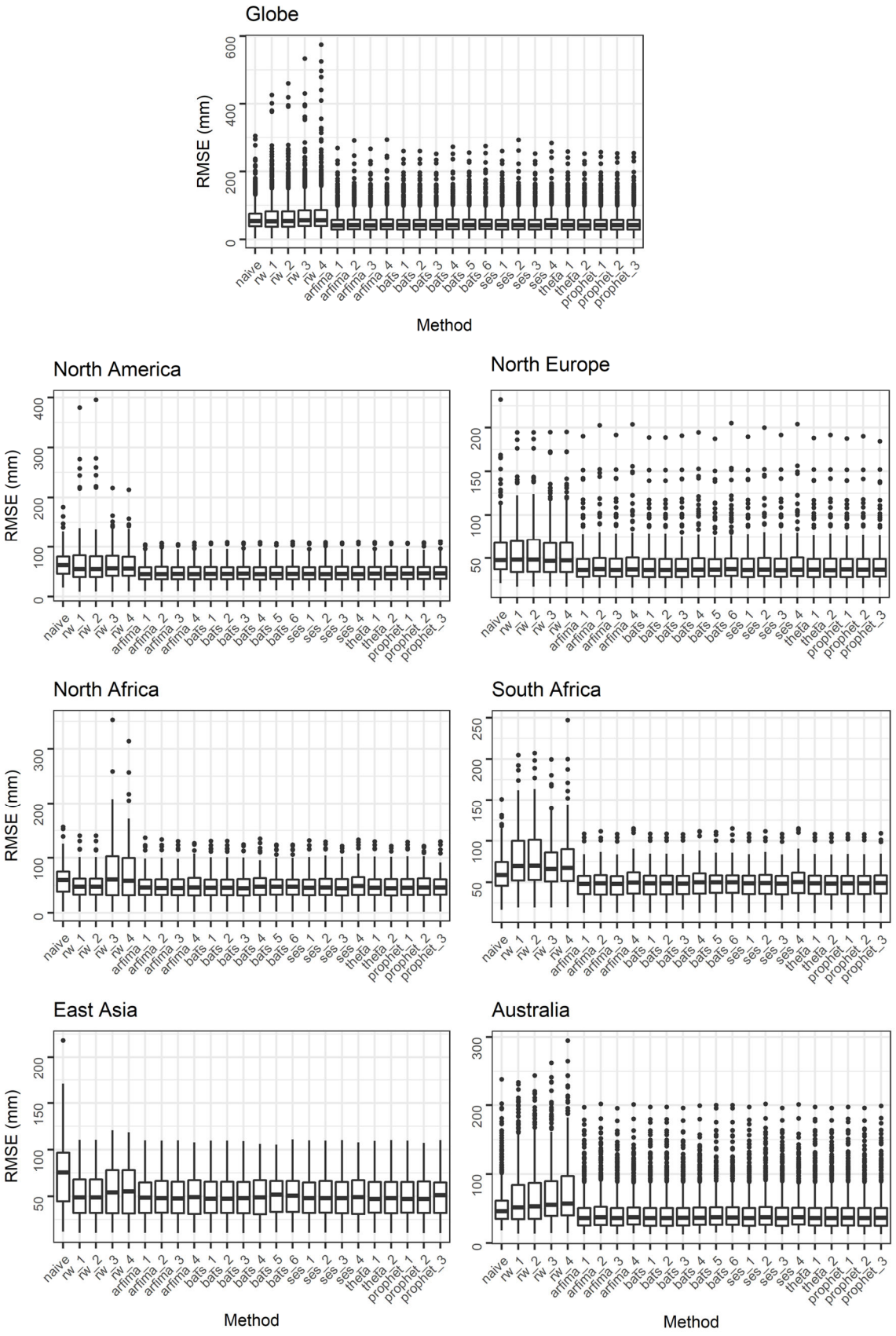


Figure 5.15. RMSE for the precipitation time series.

Here as well, the Prophet model is competitive to ARFIMA, BATS, simple exponential smoothing and Theta models, offering from 16% up to 38% (depending on the examined set of time series) better results than the naïve method, while exhibiting the smallest RMSE amongst all

the methods for the precipitation forecasts for East Asia. The median values of the best-performing automatic methods are close to the median of the estimated standard deviations of the deseasonalized time series (see Figure 5.2), as also applying to the temperature forecasts. Furthermore, the variants belonging to each of the {arfima_1, arfima_2, arfima_3, arfima_4}, {bats_1, bats_2, bats_3, bats_4}, {ses_1, ses_2, ses_3, ses_4}, {theta_1, theta_2}, {prophet_1, prophet_2} sets cannot be ranked either using the results of the experiments on precipitation time series, while the use of bats_5, bats_6 and prophet_3 has mostly (but not in all cases) led to less accurate forecasts. Regarding the different seasonality patterns illustrated in Figure 5.13, these do not result in some dramatic difference in the numerical results in contrast to those illustrated in Figure 5.14 for the case of the random walk variants. Finally, the NSE values of Table 5.10 are far worse than those corresponding to the temperature forecasts. Still, most of them are greater than zero and, thus, indicate acceptable performances, while for the geographical regions of East Asia and North Africa the performances could be characterized as moderate.

5.4 Summary and discussion

We investigate the predictability of monthly temperature and precipitation by applying seven automatic univariate time series forecasting methods to 985 and 1 552 monthly time series of temperature and precipitation, respectively. The methods include a naïve one based on the monthly values of the last year, while the rest are based on the random walk (with drift), ARFIMA, BATS, simple exponential smoothing, Theta and Prophet models. Prophet is a recently introduced model inspired by the nature of time series forecasted at Facebook and it has not been applied to hydrometeorological time series before. The ARFIMA model, on the other hand, is widely used in a non-automatic way in the hydrological literature, while the rest of the models have been rarely implemented in hydrology, e.g., in Chapters 3 and 4 herein, although they are very common in the forecasting literature. In the latter studies, no investigation is provided on how different choices of handling the seasonality and non-normality affect the performance of the models. This investigation constitutes one of the main aims of the present Chapter (therefore, proper variants of the methods are examined), together with the quantification of the performance of the selected models on monthly hydrometeorological time series and the comparison of the Prophet model to the rest. The used time series are 480 months long with no missing values, observed between January 1950 and December 1989 in stations covering a significant part of the Earth's surface and, therefore, including various real-world process behaviours. The models are fitted in the first 36 years of data (432 months) and subsequently tested in performing multi-step ahead forecasts for the last four years of data (48 months). The results are summarized in global scores, while their examination by group of stations leads to five individual scores for temperature and six for precipitation. The groups are formed according to the geographical vicinity of the stations.

The results indicate that all the examined methods apart from the naïve and random walk ones are accurate enough to be used in long-term forecasting applications. Even the simple exponential smoothing and Theta models, which exhibit a rather moderate performance in terms of RMSE and NSE in the simulation experiments of Chapters 3, in this Chapter are found to be equally competitive with the ARFIMA and BATS models, which are the most accurate in terms of RMSE and NSE in the above-mentioned experiments. This may be explained by the fact that these specific experiments use non-seasonal simulated and real-world processes, with different predictability than the monthly temperature and precipitation processes. Seasonality can be assumed to be the deterministic term of a process and its proper handling leads to a significant improvement of the forecasts. Seasonality is also the reason why patterns of error evolution, investigated in Papacharalampous et al. (2018c), are not revealed within the experiments of the present Chapter, although the forecasting horizon is long enough here as well. The above-stated qualitative outcome is consistent with the 50 single-case studies of Chapter 6, which also use monthly temperature and precipitation data. In this latter Chapter, the seasonality term is estimated using the multiplicative model for the temperature time series and the additive model for the precipitation time series. Regarding the investigation of the present Chapter on how different choices of handling seasonality and non-normality affect the performance of the models,

the results do not suggest any specific combination of choices for the external handling of seasonality and non-normality as best. Nevertheless, the handling of seasonality through the BATS and Prophet models (the only models that offer this possibility amongst the used ones) mostly leads to less accurate forecasts than the external handling, especially for the former model.

Admittedly, the quantitative information provided by the present Chapter is also important, since it directly expresses the predictability of monthly temperature and precipitation. The minimum and maximum medians of the absolute errors of the temperature forecasts are found to be around 0.25 K and 8.2 K respectively. Furthermore, a zero median of the absolute errors is computed for the precipitation forecasts produced for the dry months in geographical regions with relatively regular variability in precipitation, while the maximum median computed is around 100 mm. These values could be viewed in comparison with the minimum and maximum medians of absolute errors for annual temperature and precipitation, as derived in Chapter 4 of this thesis using two real-world datasets of 297 time series in total, which are approximately equal to 0.23 K and 1.10 K, and 68 mm and 189 mm, respectively. Moreover, the computed RMSE values range between 1.01 K and 3.65 K for temperature, and 36.16 mm and 70.17 mm for precipitation, while the respective NSE values are 0.79 and 0.98 for temperature, and −0.55 and 0.71 for precipitation.

Excluding the naïve method and the variants using the random walk model, the respective RMSE values range between 1.01 K and 2.84 K for temperature, and 36.16 mm and 51.71 mm for precipitation. In more detail, for the total of the temperature time series the use of an ARFIMA, BATS, simple exponential smoothing, Theta or Prophet model, instead of the naïve method, leads to about 19–29% more accurate forecasts in terms of RMSE, or even in about 30–32% more accurate forecasts specifically for the temperature time series observed in North Europe. For the total of the precipitation time series the use of all these automatic methods leads to about 21–22% better forecasts than the use of the naïve method, while for the geographical regions of North America, North Europe and East Asia these percentages are 26–29%, 22–24% and 32–38% respectively. This higher degree of accuracy is non-ignorable and particularly important in a long run perspective. Importantly, the Prophet model is found to offer from 13% up to 32% and from 16% up to 38% better results than the naïve method for the temperature and precipitation time series respectively. Moreover, the minimum and maximum NSE medians for the ARFIMA, BATS, simple exponential smoothing, Theta and Prophet models are 0.89 and 0.98 for temperature, and −0.04 and 0.71 for precipitation. The former NSE values indicate good forecasting performances and the latter acceptable to moderate. The higher predictability of the monthly temperature compared to the monthly precipitation is expected already from the comparison of their corresponding standard deviation values of the seasonally decomposed time series, which have a median around 1.7 K and 42 mm respectively. We think that the level of the forecasting accuracy can barely be improved using other methods, as the experiments of Chapter 3 suggest.

5.5 Conclusions

We have investigated the predictability of monthly temperature and precipitation, and simultaneously assessed the multi-step ahead performance of seven automatic univariate time series forecasting methods by applying the latter to the largest sample of hydrometeorological time series ever used for such purposes. The implemented methods are a naïve one based on the monthly values of the last year, as well as random walk (with drift), ARFIMA (acronym for AutoRegressive Fractionally Integrated Moving Average), BATS (acronym for Box–Cox transform, ARMA errors, Trend, and Seasonal components), simple exponential smoothing, Theta and Prophet. The latter is a recently introduced model, inspired by the nature of time series forecasted at Facebook and never applied to geophysical processes in the past, while most of the remaining methods are rarely used in hydrology. Proper variants of the methods have been examined to further investigate how different choices of handling the seasonality and non-normality affect the performance of the models. The results indicate that (a) the last five models perform well, better than the naïve and random walk methods, (b) monthly temperature and precipitation can be forecasted to a level of accuracy which can barely be improved using other methods, (c) the

externally applied classical seasonal decomposition results mostly in better forecasts compared to the automatic seasonal decomposition and (d) the Prophet forecasting method is competitive, especially when it is combined with externally applied classical seasonal decomposition.

6. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece

In this Chapter, we provide contingent empirical evidence on the solutions to three problems associated with univariate time series forecasting using machine learning (ML) algorithms by conducting an extensive multiple-case study. These problems are: (a) lagged variable selection, (b) hyperparameter handling, and (c) comparison between ML and classical algorithms. The multiple-case study is composed by 50 single-case studies, which use time series of mean monthly temperature and total monthly precipitation observed in Greece. We focus on two ML algorithms, i.e. neural networks and support vector machines, while we also include four classical algorithms and a naïve benchmark in the comparisons. We apply a fixed methodology to each individual case and, subsequently, we perform a cross-case synthesis to facilitate the detection of systematic patterns. We fit the models to the deseasonalized time series. We compare the one- and multi-step ahead forecasting performance of the algorithms. Regarding the one-step ahead forecasting performance, the assessment is based on the absolute error of the forecast of the last monthly observation. For the quantification of the multi-step ahead forecasting performance we compute five metrics on the test set (last year's monthly observations), i.e., the root mean square error, the Nash-Sutcliffe efficiency, the ratio of standard deviations, the coefficient of correlation and the index of agreement. The evidence derived by the experiments can be summarized as follows: (a) the results mostly favour using less recent lagged variables, (b) hyperparameter optimization does not necessarily lead to better forecasts, (c) the ML and classical algorithms seem to be equally competitive.

6.1 Introduction

6.1.1 Background information

Machine learning (ML) algorithms are widely used for the forecasting of univariate geophysical time series as an alternative to classical algorithms. Popular ML algorithms are the rather well-established Neural Networks (NN) and the new-entrant in most scientific fields Support Vector Machines (SVM). The latter algorithm has been presented in its current form by Cortes and Vapnik (1995; see also Vapnik 1995, 1999). The large number and wide range of the relevant applications is apparent in the review papers of Maier and Dandy (2000), and Raghavendra and Deka (2014), respectively. The competence of ML algorithms in univariate time series forecasting has been empirically proven in Chapters 3 and 4, and in Tyralis and Papacharalampous (2017) through extensive simulation experiments and large-scale real-world investigations.

Nevertheless, univariate time series forecasting using ML algorithms also implies the handling of specific factors that may improve or deteriorate the performance of the algorithms, i.e., the lagged variables and the hyperparameters. In contrast to the typical regression problem, in a forecasting problem the set of predictor variables is a set of lagged variables, formed using observed past values of the process to be forecasted and, consequently, holding information about the temporal dependence. Although the amount of the available historical information taken into account increases when using a large number of lagged variables, the length of the fitting set concomitantly decreases; for more details, see Tyralis and Papacharalampous (2017). While there is a wide literature on applications of ML algorithms in hydrological univariate time series forecasting, mainly comprising single- or few-case studies that particularly focus on details about the model structure (e.g., Atiya et al. 1999; Guo et al. 2011; Hong 2008; Kumar et al. 2004; Moustris et al. 2011; Ouyang and Lu 2017; Sivapragasam et al. 2001; Wang et al. 2006), studies explicitly stating information concerning the variable selection issue, such as Belayneh et al. (2014), Nayak et al. (2004), Hung et al. (2009) and Yaseen et al. (2016), are less. Tyralis and Papacharalampous (2017) have investigated the effect of a sufficient number of lagged variable selection choices on the performance of the Breiman's random forests algorithm (Breiman 2001a) in one-step ahead univariate time series forecasting.

On the other hand, information on the hyperparameter selection is usually emphasized in the hydrological literature (see e.g., Belayneh et al. 2014; Hung et al. 2009; Koutsoyiannis et al. 2008; El-Shafie et al. 2007; Tongal and Berndtsson 2017; Valipour et al. 2013; Yu et al. 2004). An example of a hyperparameter is the number of hidden nodes within a neural networks structure. Hyperparameters are distinguished from the basic parameters, because they are usually optimized or tuned with the aim to improve the performance of a ML algorithm. Hyperparameter optimization can be performed using a single validation set extracted from the fitting set or k-fold cross-validation, which involves multiple set divisions and tests. The optimal hyperparameter values are most frequently searched heuristically, either using grid search or random search, while ML or Bayesian methods can be adopted for this task as well (Witten et al. 2017). However, non-tuned ML models are also used in hydrology (see e.g., Yaseen et al. 2016). Finally, a popular problem arising when using ML forecasting algorithms is the comparison between ML and classical algorithms. This problem is mostly examined within single-case studies (see e.g., Ballini et al. 2001; Koutsoyiannis et al. 2008; Tongal and Berndtsson 2017; Valipour et al. 2013; Yu et al. 2004; see also Tables 3.1 and 4.1 herein), as also applying to the problems of lagged variable and hyperparameter selection.

6.1.2 Main contribution and research questions

The main contribution of this Chapter is the exploration in geoscientific concepts of the problems problems presented in detail in Section 6.1.1 and summarized here below, together with their related research questions of focus:

- Problem 1: Lagged variable selection in time series forecasting using ML algorithms

  Research question 1: *Should we select less recent lagged variables or a large number of lagged variables in time series forecasting using ML algorithms?*

- Problem 2: Hyperparameter selection in time series forecasting using ML algorithms

  Research question 2: *Does hyperparameter optimization necessarily lead to a better performance in time series forecasting using ML algorithms?*

- Problem 3: Comparison between ML and classical algorithms

  Research question 3: *Do the ML algorithms exhibit better (or worse) performance than the classical ones?*

In fact, exploration is indispensable for understanding the phenomena involved in a specific problem and, therefore, it constitutes an essential part within every theory-development process.

6.1.3 Research method and implementation

We adopt the multiple-case study research method (presented in detail in Yin 2003). This method embraces the examination of more than one individual cases, facilitating the observation of specific phenomena from multiple perspectives or within different contexts (Dooley 2002). For the detection of systematic patterns across the individual cases a cross-case synthesis can be performed (Larsson 1993). Given the fact that the boundaries between the phenomena and the context are not clear (thus, it is meaningful to consider a case study design, as explained in Baxter and Jack 2008), it is important that each individual case keeps its identity within the multiple-case study, so that one can specifically focus on it. This exploration within and across the individual cases can provide interesting insights into the phenomena under investigation, as well as a form of generalization named "contingent empirical generalization", while retaining the immediacy of the single-case study method (Achen and Snidal 1989).

We conduct an extensive multiple-case study composed by 50 single-case studies. The latter use temperature and precipitation time series observed in Greece. We examine these two geophysical processes, because they exhibit different properties, which may affect differently the results within the explorations. We focus on two ML algorithms, i.e. NN and SVM, for an analogous reason. Moreover, the explorations are conducted for the one- and a multi-step ahead horizons, as their corresponding forecasting attempts are not of the same difficulty. We apply a fixed

methodology to each individual case. This fixed methodology provides the common basis to further perform a cross-case synthesis for the detection of systematic patterns across the individual cases. The latter is the novelty of this Chapter.

6.2 Data and methods

In this Section, we present the data and methods of the Chapter. Basic information on the methods' implementation is also provided, while the total of the exploited `R` packages is independently listed in Section 2.9.4. All `R` functions are used with their predefined values, unless specified differently. Hereafter, to specify an implemented `R` function, we state its name accompanied by the name of the `R` package. The latter name is given between curly brackets (`{}`). To imply that we implement a built-in-`R` function, we accompany its name with "`{stats}`".

6.2.1 Methodology outline

We conduct 50 single-case studies by applying a fixed methodology to each of the 50 time series presented in Section 6.2.2, as explained subsequently. First, we split the time series into a fitting set and a test set. The latter is the last monthly observation for the one-step ahead forecasting experiments and the last year's monthly observations for the multi-step ahead forecasting experiments. Second, we fit the models to the seasonally decomposed fitting set, within the context described in Section 6.2.3, and make predictions corresponding to the test set. Third, we recover the seasonality in the predicted values and compare them to their corresponding observed using the metrics of Section 6.2.4. Finally, we perform a cross-case synthesis to demonstrate similarities and differences between the single-case studies conducted. We present the results per category of tests, which is determined by the set {set of methods, process, forecast horizon}, and further summarize them, as discussed in Section 6.2.4. The sets of methods are defined in Section 6.2.3, while the total number of categories is 20. We place emphasis on the exploration of the three problems summarized in Section 6.1, but we also present quantitative information about the produced forecasts and search for evidence regarding the existence of a possible relationship between the forecast quality, and the standard deviation (σ), coefficient of variation (cv) and Hurst parameter (H) estimates of the fractional Gaussian noise process (see Section 2.1.6) for the deseasonalized time series. These estimates are presented in Section 6.2.2.

6.2.2 Temperature and precipitation time series

We use 50 time series of mean monthly temperature and total monthly precipitation observed in Greece. These time series are sourced from Lawrimore et al. (2011), and Peterson and Vose (1997), respectively. We select only those with few missing values (blocks with length equal or less than one). Subsequently, we use a seasonal Kalman filter (see Section 2.1.9), implemented through the `R` function `na.StructTS {zoo}`, for filling in the missing values. The basic information about the time series is provided in Table 6.1, while Figure 6.1 presents the locations of the stations at which the data has been recorded. We use the deseasonalized fitting sets for fitting the forecasting models, as suggested in Taieb et al. (2012) for the improvement of the forecast quality (see also Section 2.2.6). The time series decomposition is performed exclusively on the fitting sets by using the multiplicative model (see Section 2.1.7) for the temperature time series and the additive model (see Section 2.1.7) for the precipitation ones. The reason for this differentiation is that the use of the multiplicative model on the precipitation time series results in zero forecasts for some methods, as a result of zero precipitation observations in the summer months. Both seasonal decomposition models are implemented throught the `R` function `decompose {stats}`.

Table 6.1. Time series investigated in the Chapter.

S/n	Process	Code name	Location	Station details			Reference	Start	End	Length (months)
				ID	Latitude	Longitude				
1	Temperature	temp_1	Araxos	16687001	38.20	21.40	Lawrimore et al. (2011)	Jan 1951	Dec 1980	360
2		temp_2	Athens	16714000	37.97	23.72		Jan 1858	Dec 1975	1416
3		temp_3	Athens	16714000	37.97	23.72		Jan 1989	Dec 2001	156
4		temp_4	Athens	16716000	37.90	23.73		Jan 1951	Dec 2012	744
5		temp_5	Heraklion	16754000	35.33	25.18		Jan 1950	Dec 2015	792
6		temp_6	Kalamata	16726000	37.07	22.02		Jan 1956	Dec 2015	720
7		temp_7	Kerkyra	16641000	39.62	19.92		Jan 1951	Dec 2016	792
8		temp_8	Larissa	16648000	39.63	22.42		Jan 1899	Dec 2016	1416
9		temp_9	Lemnos	16650000	39.92	25.23		Jan 1951	Dec 1998	576
10		temp_10	Methoni	16734000	36.83	21.70		Jan 1951	Dec 1972	264
11		temp_11	Methoni	16734000	36.83	21.70		Jan 1975	Dec 2000	312
12		temp_12	Patra	16689000	38.25	21.73		Jan 1951	Dec 1989	468
13		temp_13	Samos	16723000	37.70	26.92		Jan 1955	Dec 1969	180
14		temp_14	Samos	16723000	37.70	26.92		Jan 1974	Dec 2003	360
15		temp_15	Souda	16746000	35.48	24.12		Jan 1961	Dec 2015	660
16		temp_16	Thessaloniki	16622000	40.52	22.97		Jan 1892	Dec 2016	1500
17		temp_17	Thessaloniki	16622001	40.52	23.02		Jan 1961	Dec 1970	120
18	Precipitation	prec_1	Agrinion	16672000	38.60	21.70	Peterson and Vose (1997)	Jan 1956	Dec 1987	384
19		prec_2	Alexandroupoli	16627000	40.80	25.90		Jan 1951	Dec 1990	480
20		prec_3	Aliartos	16674000	38.40	23.10		Jan 1907	Dec 1990	1008
21		prec_4	Anogeia	16754001	35.30	24.90		Jan 1919	Dec 1939	252
22		prec_5	Anogeia	16754001	35.30	24.90		Jan 1950	Dec 1979	360
23		prec_6	Araxos	16687000	38.20	21.40		Jan 1949	Dec 2000	624
24		prec_7	Athens	16714000	38.00	23.70		Jan 1860	Dec 1881	264
25		prec_8	Athens	16714000	38.00	23.70		Jan 1887	Dec 2005	1428
26		prec_9	Athens	16716000	37.90	23.70		Jan 1929	Dec 1945	204
27		prec_10	Fragma	16715001	38.20	23.90		Jan 1926	Dec 1990	780
28		prec_11	Heraklion	16754000	35.30	25.10		Jan 1946	Dec 1990	540
29		prec_12	Igoumenitsa	16641001	39.50	20.30		Jan 1951	Dec 1990	480
30		prec_13	Ioannina	16642000	39.70	20.80		Jan 1951	Dec 1990	480
31		prec_14	Kalamata	16726000	37.00	22.10		Jan 1956	Dec 1970	180
32		prec_15	Kalo Chorio	16756001	35.10	25.70		Jan 1950	Dec 1984	420
33		prec_16	Kastelli	16760001	35.20	25.30		Jan 1949	Dec 1976	336
34		prec_17	Kerkyra	16641000	39.60	19.90		Jan 1952	Dec 1996	540
35		prec_18	Kythira	16743000	36.30	23.00		Jan 1951	Dec 1973	276
36		prec_19	Kos	16742000	36.80	27.10		Jan 1958	Dec 1990	396
37		prec_20	Kozani	16632000	40.30	21.80		Jan 1955	Dec 1987	396
38		prec_21	Larissa	16648000	39.60	22.40		Jan 1951	Dec 1997	564
39		prec_22	Lemnos	16650001	39.90	25.30		Jan 1951	Dec 2000	600
40		prec_23	Methoni	16734000	36.80	21.70		Jan 1951	Dec 1991	492
41		prec_24	Milos	16738000	36.70	24.50		Jan 1951	Dec 1990	480
42		prec_25	Mytilene	16667000	39.10	26.60		Jan 1952	Dec 1990	468
43		prec_26	Naxos	16732000	37.10	25.50		Jan 1955	Dec 1971	204
44		prec_27	Patra	16689000	38.20	21.70		Jan 1901	Dec 1984	1008
45		prec_28	Sitia	16757000	35.20	26.10		Jan 1960	Dec 1983	288
46		prec_29	Skyros	16684000	38.90	24.60		Jan 1955	Dec 1987	396
47		prec_30	Thessaloniki	16622000	40.60	23.00		Jan 1931	Dec 1997	804
48		prec_31	Thessaloniki	16622002	40.50	22.90		Jan 1961	Dec 1970	120
49		prec_32	Trikala	16645001	39.60	21.80		Jan 1951	Dec 1990	480
50		prec_33	Tripoli	16710000	37.50	22.40		Jan 1951	Dec 1985	420

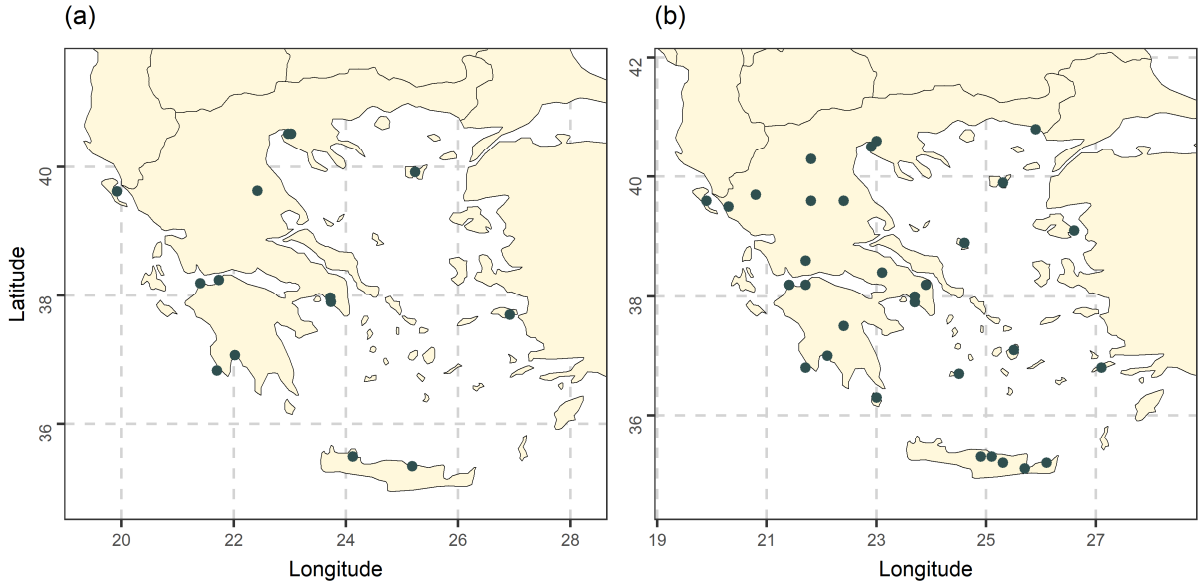


Figure 6.1. Maps of the locations of the (a) temperature and (b) precipitation stations; their sources are Lawrimore et al. (2011), and Peterson and Vose (1997), respectively.

We also apply the time series decomposition models to the entire time series to deseasonalize them. We then estimate the mean (μ), σ and H parameters of the fractional Gaussian noise process (see Section 2.1.6) for each of the seasonally decomposed entire time series by using the `R` function `mleHK {HKprocess}`. We further estimate the coefficient of variation (cv) of the fractional Gaussian noise process, as detailed in Section 2.1.6. The μ, σ, cv and H estimates are presented in Tables 6.2 and 6.3.

Table 6.2. Mean (μ), standard deviation (σ), coefficient of variation (cv) and Hurst parameter (H) estimates for the deseasonalized temperature time series.

Time series	μ estimate (°C)	σ estimate (°C)	cv estimate	H estimate
temp_1	17.95	1.25	0.07	0.66
temp_2	17.86	1.93	0.11	0.67
temp_3	18.51	1.81	0.10	0.68
temp_4	18.70	1.62	0.09	0.65
temp_5	18.97	1.18	0.06	0.69
temp_6	17.90	1.42	0.08	0.74
temp_7	17.75	1.47	0.08	0.67
temp_8	15.91	2.75	0.17	0.64
temp_9	16.36	2.11	0.13	0.74
temp_10	18.24	1.07	0.06	0.59
temp_11	17.83	1.20	0.07	0.61
temp_12	17.71	1.41	0.08	0.69
temp_13	18.21	1.46	0.08	0.64
temp_14	18.38	1.64	0.09	0.64
temp_15	18.63	1.47	0.08	0.71
temp_16	16.21	2.59	0.16	0.67
temp_17	16.13	2.16	0.13	0.48

Table 6.3. Mean (μ), standard deviation (σ), coefficient of variation (cv) and Hurst parameter (H) estimates for the deseasonalized precipitation time series.

Time series	μ estimate (mm)	σ estimate (mm)	cv estimate	H estimate
prec_1	81.09	56.61	0.70	0.47
prec_2	46.50	37.30	0.80	0.56
prec_3	55.52	42.14	0.76	0.53
prec_4	93.61	78.01	0.83	0.57
prec_5	95.62	74.42	0.78	0.48
prec_6	57.59	43.65	0.76	0.54
prec_7	33.44	30.45	0.91	0.56
prec_8	32.79	29.44	0.90	0.53
prec_9	29.65	27.87	0.94	0.53
prec_10	47.30	37.03	0.78	0.53
prec_11	40.02	35.27	0.88	0.50
prec_12	88.81	66.22	0.75	0.56
prec_13	94.36	60.85	0.64	0.57
prec_14	66.19	45.58	0.69	0.46
prec_15	42.12	35.65	0.85	0.50
prec_16	60.14	47.45	0.79	0.52
prec_17	92.53	65.00	0.70	0.56
prec_18	47.10	39.39	0.84	0.52
prec_19	58.63	53.36	0.91	0.57
prec_20	43.94	32.23	0.73	0.54
prec_21	36.46	30.90	0.85	0.54
prec_22	40.84	36.72	0.90	0.55
prec_23	60.59	44.00	0.73	0.50
prec_24	35.08	32.84	0.94	0.47
prec_25	56.00	49.39	0.88	0.51
prec_26	27.61	22.43	0.81	0.53
prec_27	60.23	44.64	0.74	0.52
prec_28	40.39	35.38	0.88	0.46
prec_29	38.55	32.86	0.85	0.56
prec_30	37.15	27.98	0.75	0.54
prec_31	35.24	24.94	0.71	0.55
prec_32	62.91	47.51	0.76	0.61
prec_33	68.45	44.77	0.65	0.47

6.2.3 Forecasting algorithms and methods

All the algorithms used herein (see Table 6.4) are well-grounded in the literature; thus, in their subsequent presentation we place emphasis on implementation information. Their theoretical documentation can be found in Section 2.2 (see also the references therein). We focus on two ML forecasting algorithms, i.e., NN and SVM. The NN algorithm is implemented through the `R` function `mlp {nnet}`, while the SVM algorithm is implemented through the `R` function `ksvm {kernlab}`. These algorithms implement a single-hidden layer Multilayer Perceptron (MLP), and the Radial Basis kernel "Gaussian" function with C = 1 and epsilon = 0.1, respectively. Their application is made by using the `R` functions `CasesSeries {rminer}`, `fit {rminer}` and `lforecast {rminer}`. We also include four classical algorithms, i.e., AR(1), auto_ARFIMA, BATS and Theta, and the seasonal naïve benchmark in the comparisons. We apply the classical algorithms by using the `R` functions `Arima {forecast}`, `arfima {forecast}`, `bats {forecast}`, `forecast {forecast}` and `thetaf {forecast}`. The auto_ARFIMA algorithm considers the long-range dependence observed in the time series through the d parameter. The AR(1), auto_ARFIMA and BATS algorithms apply Box-Cox transformation to the input data before fitting a model to them.

Table 6.4. Forecasting algorithms and their corresponding models from Table 2.3.

S/n	Role in this work	Code name	Corresponding model from Table 2.3	General category	Description
1	Main	NN	Neural networks (NN)	NN	Section 2.3.2
2		SVM	Support vector machines (SVM)	SVM	Section 2.3.4
3	Additional	Naive	Seasonal naïve	Simple	Section 2.2.1
4		AR(1)	Fixed-order autoregressive moving average (ARMA)	ARIMA	
5		auto_ARFIMA	Optimum-order autoregressive fractionally integrated moving average (ARFIMA)	ARFIMA	Section 2.2.2
6		BATS	Exponential smoothing state space with Box-Cox transformation, ARMA errors correction, trend and seasonal components (BATS)	Innovations state space	Section 2.2.3
7		Theta	Theta	Exponential smoothing	

While the classical methods are simply defined by the classical algorithm, the ML methods are defined by the set {ML algorithm, hyperparameter selection procedure, lags}. We compare 21 regression matrices, each using the first n time lags, n = 1, 2, …, 21, and two procedures for hyperparameter selection, i.e., predefined hyperparameters (default values of the algorithms) or defined after optimization. The symbol * in the name of a ML method is used in this Chapter to denote that the model's hyperparameters have been optimized. The hyperparameter optimization is performed with the grid search method using a single validation set (last 1/3 of the deseasonalized fitting set). The hyperparameters optimized are the number of hidden nodes and the number of variables randomly sampled as candidates at each split of the NN and SVM models respectively. For the NN* method the hyperparameter optimization procedure is described subsequently. First, we fit 16 different NN models (defined by the grid values 0, …, 15) to the fist 2/3 of the deseasonalized fitting set. Second, we use these models to produce forecasts corresponding to the validation set. Third, we select the one exhibiting the smallest root mean square error (RMSE) on the validation set. To produce the forecast corresponding to the test set we further fit the selected model to the whole deseasonalized fitting set. For the SVM* method the procedure is the same, except that the candidate models are five (defined by the grid values 1, … 5). Hereafter, we consider that the ML models are used with predefined hyperparameters and that the regression matrix is built using only the first lag, unless mentioned differently. We use the sets of methods defined in Table 6.5. Each of them has a specific utility within our experiments, which is also reported in Table 6.5. A secondary utility of set of methods no 5 is the investigation of the existence of a possible relationship between the forecast quality and the parameter estimates for the deseasonalized time series.

Table 6.5. Sets of forecasting methods and their main utility within the Chapter. The forecasting algorithms are defined in Table 6.4. The symbol * in the name of a machine learning method is used to denote that the model's hyperparameters have been optimized.

S/n	Set of methods	Number of included methods	Main utility
1	{NN given a regression matrix formed using the first n lags, n = 1, 2, …, 21}	21	Exploration of Problem 1 for the NN algorithm
2	{SVM given a regression matrix formed using the first n lags, n = 1, 2, …, 21}	21	Exploration of Problem 1 for the SVM algorithm
3	{NN, NN*}	2	Exploration of Problem 2 for the NN algorithm
4	{SVM, SVM*}	2	Exploration of Problem 2 for the SVM algorithm
5	{Naïve, AR(1), auto_ARFIMA, BATS, Theta, NN, SVM}	7	Exploration of Problem 3 for the NN and SVM algorithms

6.2.4 Metrics and summary statistics

The one-step ahead forecasting performance is assessed by computing the absolute error (AE) of the forecast, while the multi-step ahead forecasting performance by computing the RMSE, the Nash-Sutcliffe efficiency (NSE), the ratio of standard deviations (rSD), the index of agreement (*d*) and the coefficient of correlation (Pr). The definitions of these metrics are given in Section 2.8.2 (see also Table 2.6). To summarize the results of the multiple-case study, we compute some summary statistics for the values of each metric, i.e., the minimum, median and maximum, separately for each algorithm. For the ML ones, these summary statistics are computed by aggregating the total of the values of each metric computed for methods that are based on each specific ML algorithm (tested for the exploration of Problems 1, 2 or 3). We also compute the linear regression coefficient (LRC) for each method per category of tests. The definition of the LRC statistic is provided in Section 2.8.2 (see also Table 2.6).

6.3 Results and discussion

In Section 6.3, we present and discuss the results of our multiple-case study. We place emphasis on the qualitative presentation of the results, because of its importance in the exploration of the research questions of Section 6.1. Especially the heatmap visualization adopted herein allows the examination of each single-case study alone and in comparison to the rest simultaneously. Quantitative information, derived by our multiple-case study and particularly significant for the case of Greece, is also presented. Regarding this type of information, the present Chapter could be viewed as an expansion of Moustris et al. (2011). The latter study has focused on four long precipitation time series observed in Alexandroupoli, Athens, Patra and Thessaloniki (a subset of the time series examined within our multiple-case study), with the aim to present forecasts for the monthly maximum, minimum, mean and cumulative precipitation totals using NN methods.

6.3.1 Explorations on lagged variable selection

This Section is devoted to the exploration of Problem 1. In Figures 6.2 and 6.3, we visualize the one and twelve-step ahead temperature forecasts respectively, produced for this exploration for the NN and SVM algorithms, in comparison to their corresponding target values. We observe that, for a specific target value, the forecasts are more scattered (in the vertical direction) for the NN algorithm than they are for the SVM algorithm. This fact indicates that the performance of the SVM algorithm is affected less than the performance of the NN algorithm by changes in the lagged regression matrix used in the fitting process. The effect under discussion may result in more or less accurate NN forecasts (laying closer or farther from the 1:1 line included in the scatterplots of Figures 6.2 and 6.3) than the ones produced by the SVM algorithm. Evidence that the NN algorithm is more prone to changes in the regression matrix than the SVM one is provided by the tests conducted using the precipitation time series as well. In Figure 6.4, we present the twelve-step ahead precipitation forecasts in comparison to their corresponding target values.

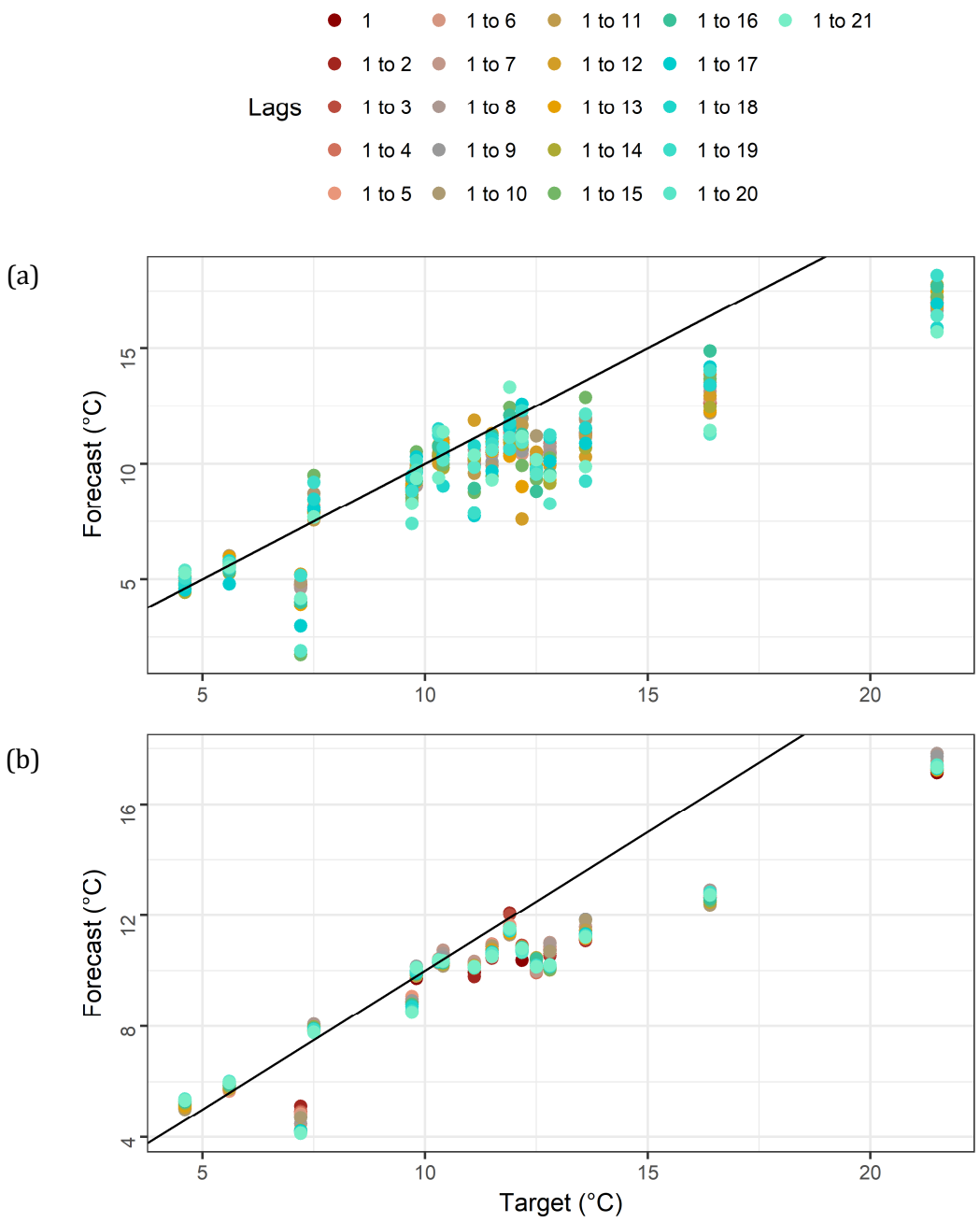


Figure 6.2. One-step ahead temperature forecasts, produced for the exploration of Problem 1 for the (a) NN and (b) SVM algorithms, in comparison to their corresponding target values.

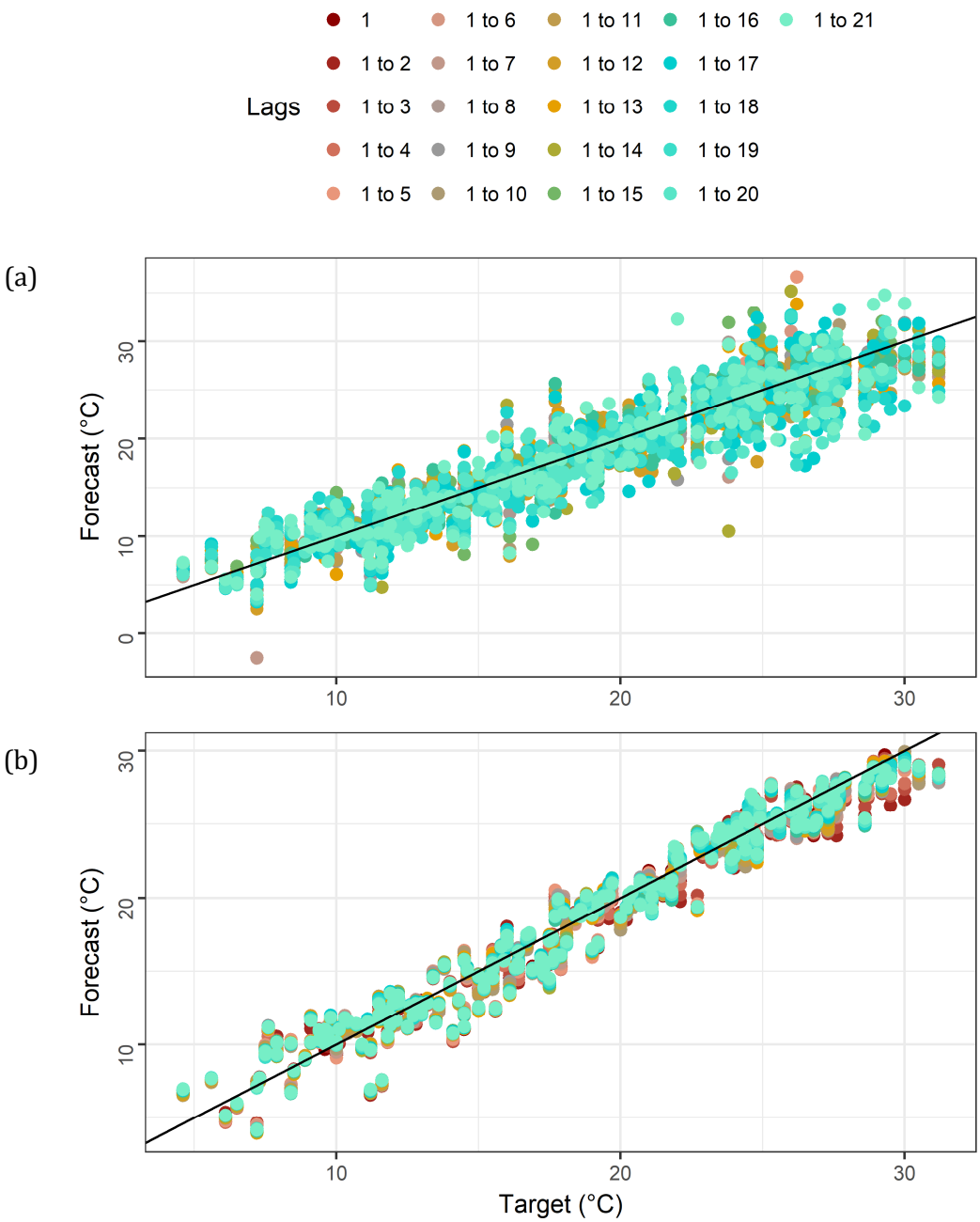


Figure 6.3. Twelve-step ahead temperature forecasts, produced for the exploration of Problem 1 for the (a) NN and (b) SVM algorithms, in comparison to their corresponding target values.

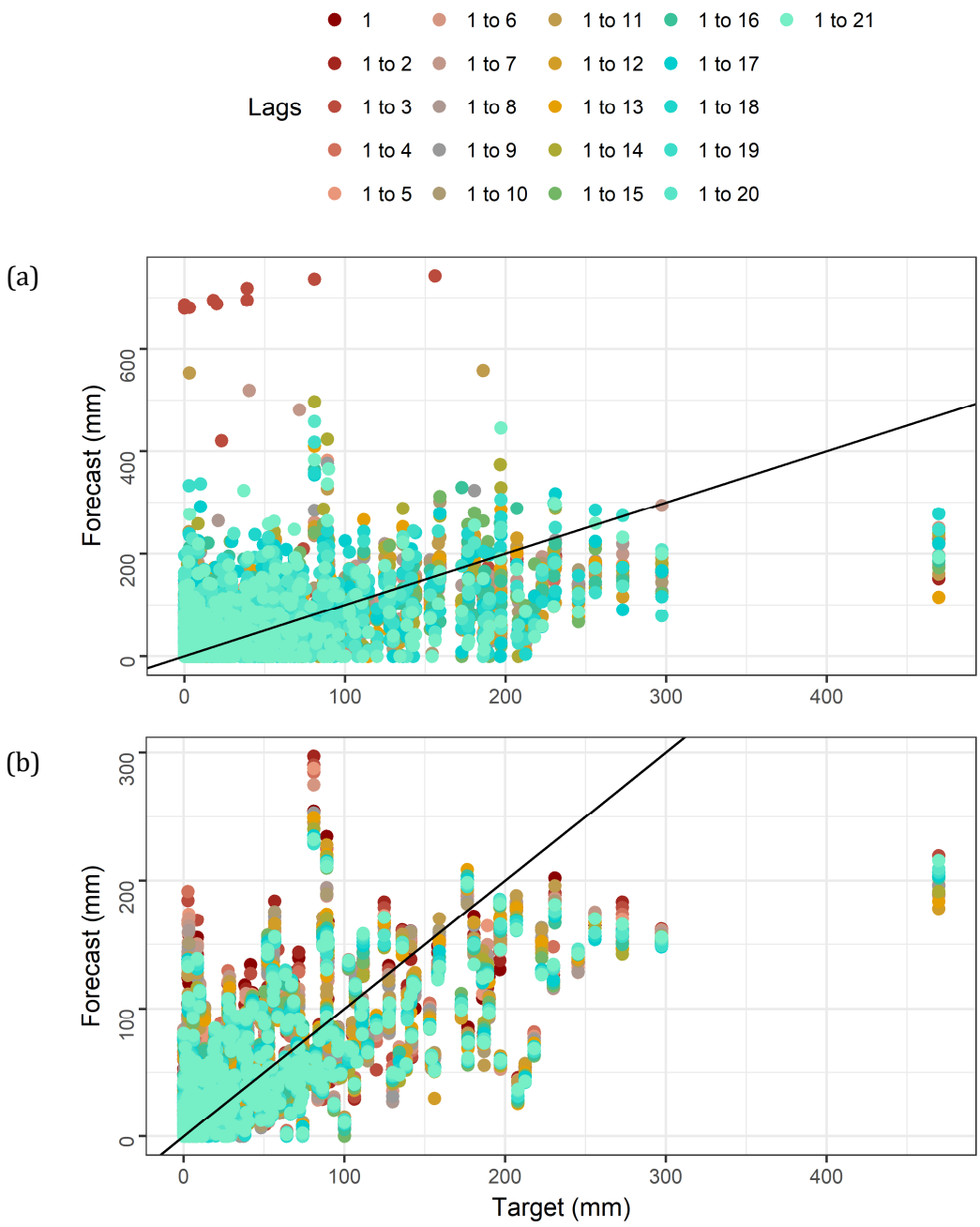


Figure 6.4. Twelve-step ahead precipitation forecasts, produced for the exploration of Problem 1 for the (a) NN and (b) SVM algorithms, in comparison to their corresponding target values.

More importantly, in Figures 6.5 and 6.6 we comparatively present the AE, RMSE, NSE and *d* values computed for the temperature forecasts, produced for the exploration of Problem 1 for the NN and SVM algorithms, for each individual case examined. By the examination of these two figures we observe the following:

(a) There are variations in the results across the individual cases, to an extent that it is impossible to decide on a best or worst method. Therefore, no evidence is provided by the respective categories of tests that any of the compared lagged regression matrices systematically leads to better forecasts than the rest, either for the NN or the SVM algorithms.

(b) The heatmaps formed for the SVM algorithm are smoother in the row direction than those formed for the NN algorithm, a fact rather expected from Figures 6.2 and 6.3. In other words, the variations within each single-case study are of small magnitude for the case of the SVM algorithm, while they are significant for the NN algorithm.

(c) For the SVM algorithm there are no systematic patterns and the small variations seem to be rather random.

(d) For the NN algorithm and especially for the twelve-step ahead forecasts the left parts of the heatmaps are smoother with no white cells. Alternatively worded, it seems that is more likely that the forecasts are better when using less recent lagged variables in conjunction with this algorithm.

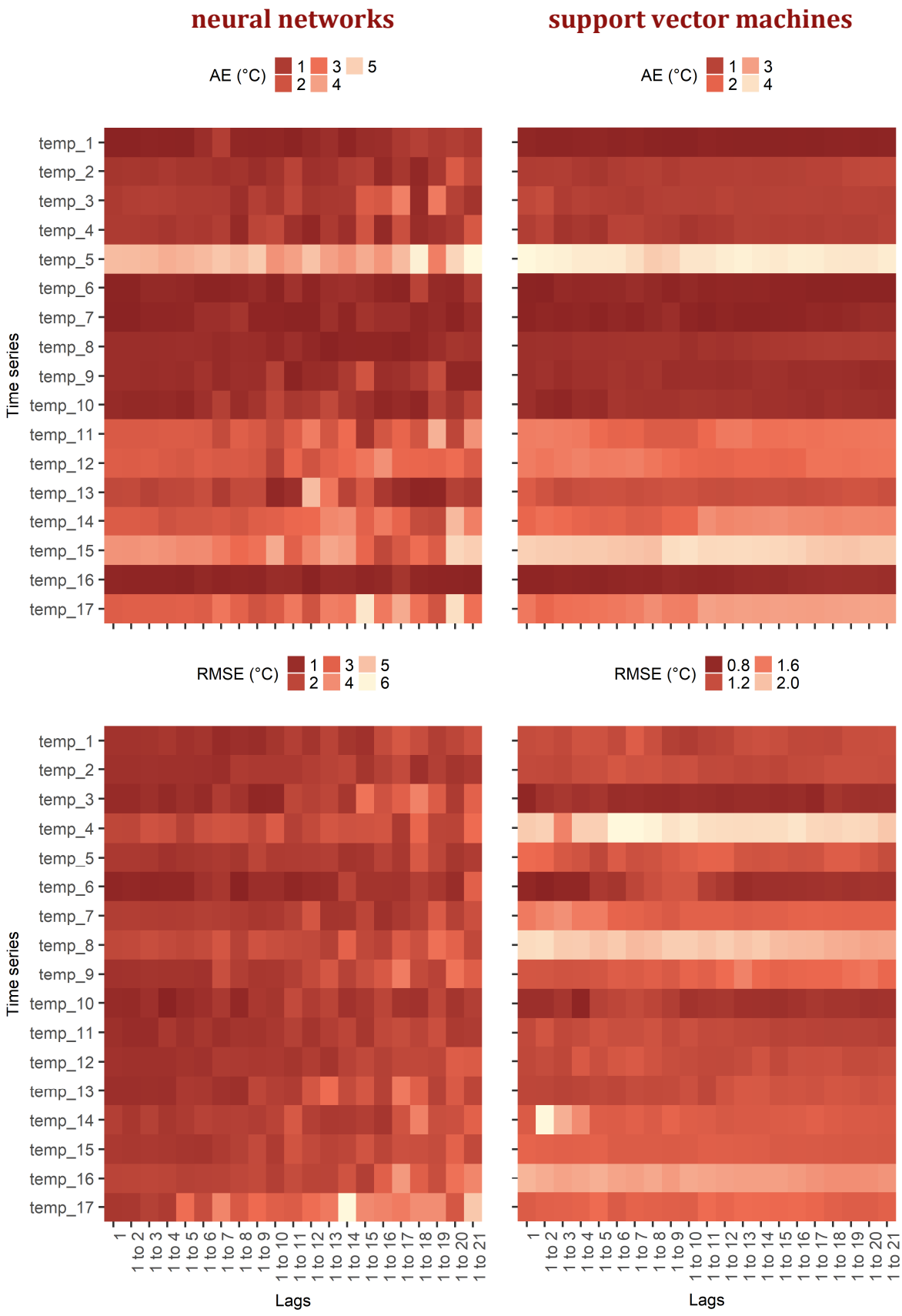


Figure 6.5. Cross-case synthesis for the exploration of Problem 1 for the NN and SVM algorithms using the temperature time series (part 1).

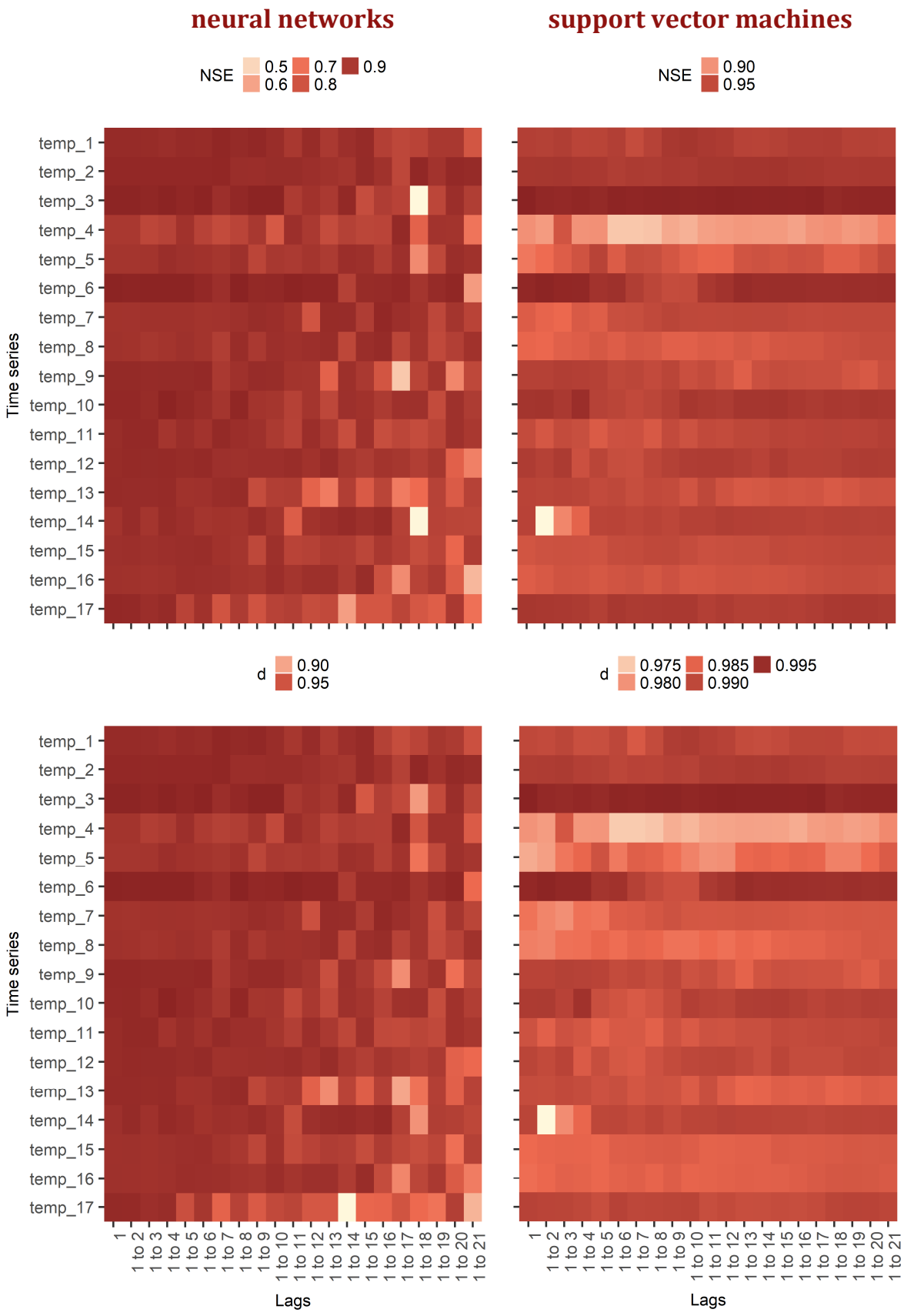


Figure 6.6. Cross-case synthesis for the exploration of Problem 1 for the NN and SVM algorithms using the temperature time series (part 2).

Observation (a) is particularly important, because it reveals that the forecast quality is subject to limitations. Each forecasting method has some specific theoretical properties and, due to the latter, it performs better or worse than other forecasting methods, depending on the case examined. Even forecasting methods based on the same algorithm can produce forecasts with very different quality, as indicated by the results obtained for the NN algorithm. Observation (d), on the other hand, provides some interesting evidence, which however is contingent and, therefore, should be further investigated within larger forecast-comparing studies, such as Tyralis and Papacharalampous (2017). Furthermore, in Figure 6.7 we present the AE and RMSE values computed for the precipitation forecasts, produced for the exploration of Problem 1 for the NN and SVM algorithms, within each single-case study. Observations (a) and (b) apply here as well. Moreover, both the ML algorithms, seem to perform rather better, to a small extent though, when given a lagged regression matrix using less recent lags.

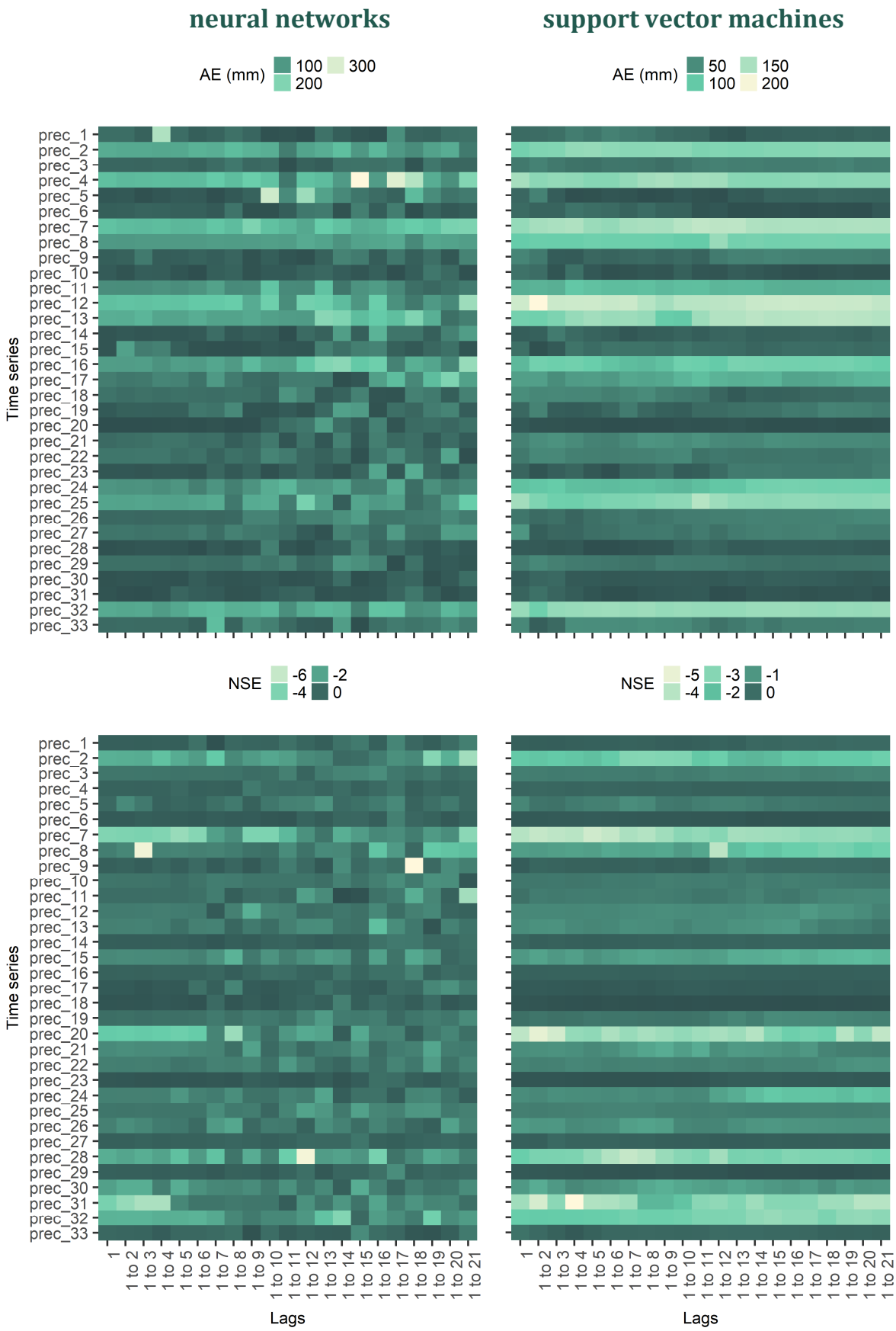


Figure 6.7. Cross-case synthesis for the exploration of Problem 1 for the NN and SVM algorithms using the precipitation time series.

6.3.2 Explorations on hyperparameter selection

This Section is devoted to the exploration of Problem 2. In Figure 6.8, we present the twelve-step ahead precipitation forecasts, produced for this exploration for the NN and SVM algorithms, in comparison to their corresponding target values. Figure 6.8 could be studied alongside with Figure 6.4, providing contingent evidence that hyperparameter optimization affects less the performance of these two ML algorithms than lagged variable selection does. The latter observation applies more to the NN algorithm. Furthermore, in Figure 6.9 we comparatively present the AE, RMSE, rSD and *d* values computed for the one- and twelve-step ahead temperature forecasts, produced for the exploration of Problem 2, within each single-case study. By the examination of Figure 6.9 we observe the following:

(a) Here as well, none of the compared methods seems to be systematically better across the individual cases examined. In other words, the results do not systematically favour any of the two tested hyperparameter selection procedures and, therefore, we can state that hyperparameter optimization does not necessarily lead to better forecasts than the use of the default values of the algorithms.

(b) For both the ML algorithms the observed variations within each of the single-case studies are of smaller magnitude for the one-step ahead forecasts than they are for the twelve-step ahead ones.

(c) For the case of the NN algorithm the twelve-step ahead forecasts seem to be rather better when hyperparameter optimization precedes the fitting process, while the opposite applies to the case of the SVM algorithm.

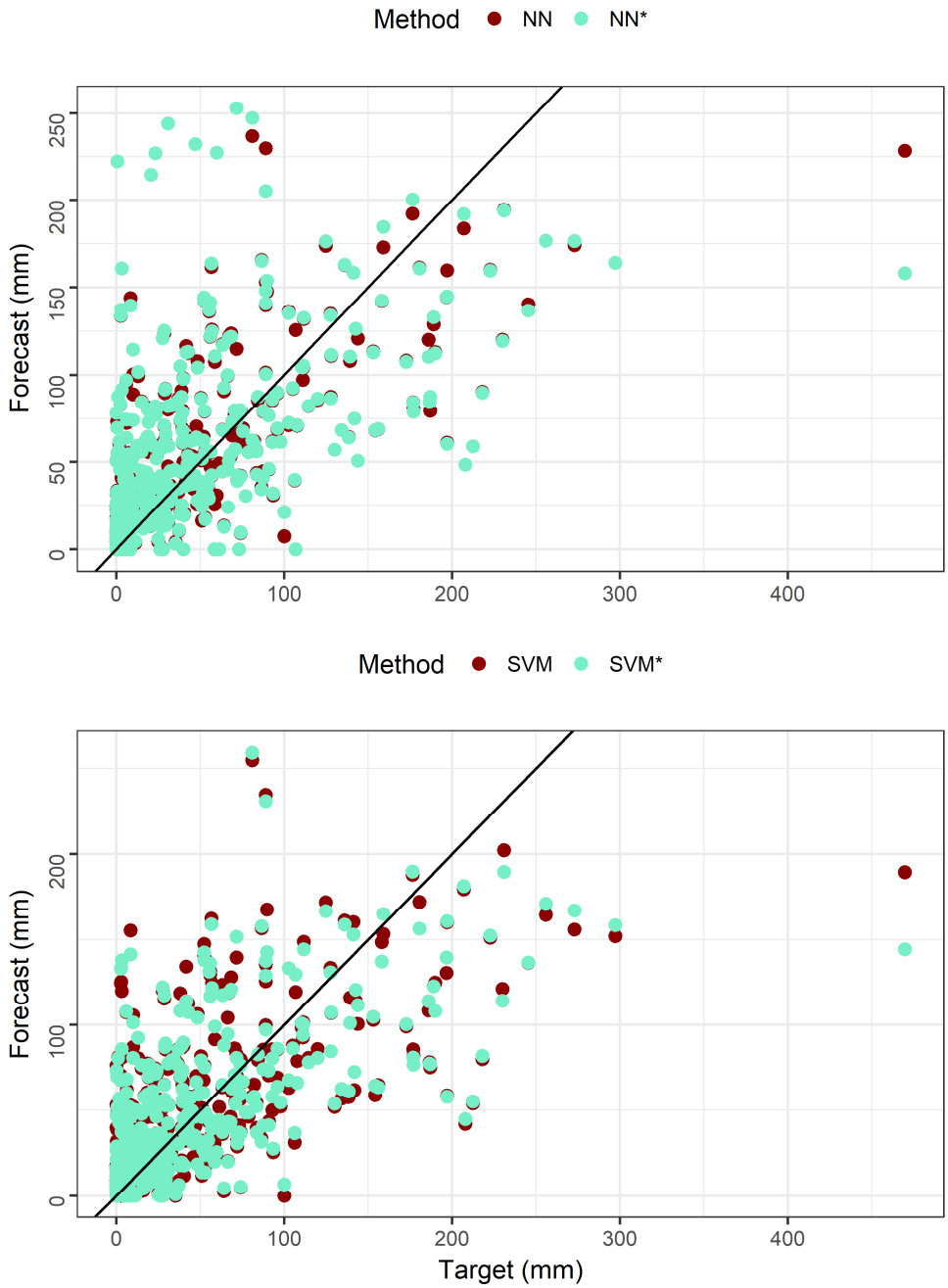


Figure 6.8. Twelve-step ahead precipitation forecasts, produced for the exploration of Problem 2 for the NN and SVM algorithms, in comparison to their corresponding target values.

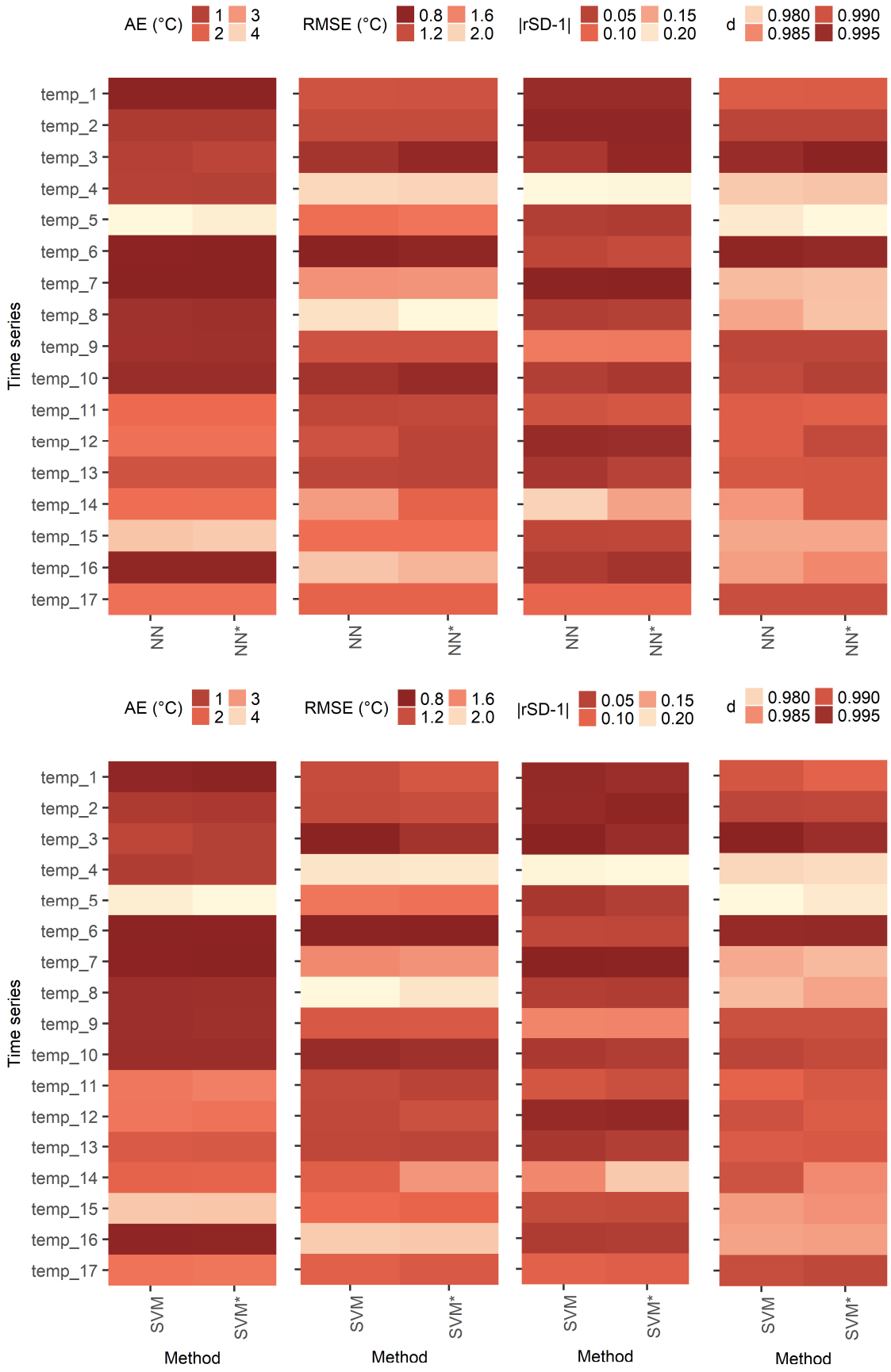


Figure 6.9. Cross-case synthesis for the exploration of Problem 2 for the NN and SVM algorithms using the temperature time series.

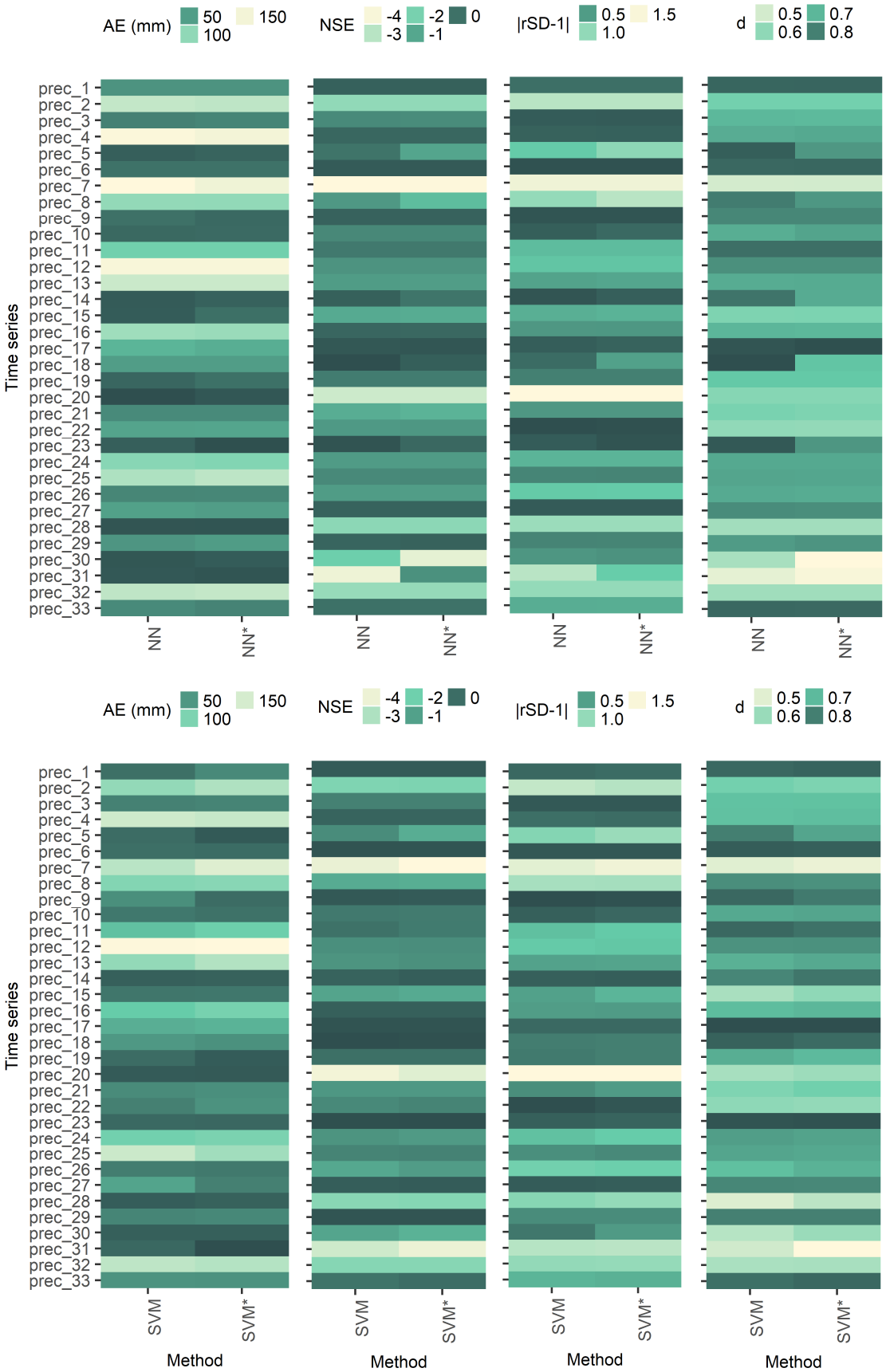


Figure 6.10. Cross-case synthesis for the exploration of Problem 2 for the NN and SVM algorithms using the precipitation time series.

Finally, in Figure 6.10 we present the AE, NSE, rSD and *d* values computed for the one- and twelve-step ahead precipitation forecasts, produced for the exploration of Problem 2, within each single-case study. Observation (a) also applies to the precipitation forecasts, while the variations can be significant for both the one- and twelve-step ahead forecasts. For the latter it seems that hyperparameter optimization mostly leads to less accurate forecasts. This may be explained by the fact that the default values of the algorithms are usually set based on tests performed by their developers or in the scientific literature, so that the performance of the algorithms is mostly maximized for a variety of problems.

6.3.3 Explorations on the comparison of different algorithms

This Section is devoted to the exploration of Problem 3. In Figure 6.11, we present the one- and twelve-step ahead temperature forecasts, produced for this exploration, in comparison to their corresponding target values, while in Figure 6.12 we present an analogous visualization for the precipitation forecasts serving the same purpose. Moreover, in Figures 6.13 and 6.14 we comparatively present all the metric values computed for the temperature forecasts and the AE, RMSE and *d* values computed for the precipitation forecasts respectively within each single-case study. By the examination of these four figures we observe the following:

(a) Here as well, the results of the single-case studies vary significantly.

(b) The best method within a specific single-case study depends on the criterion of interest. In fact, even within a specific single-case study, we cannot decide on one best (or worst) method regarding all the criteria set simultaneously.

(c) Observations (a) and (b) apply equally to the ML and the classical methods. In fact, it seems that both categories can rather perform equally well, under the same limitations.

(d) We observe that the Naïve benchmark, competent as well, frequently produces far different forecasts than those produced by the ML or classical algorithms.

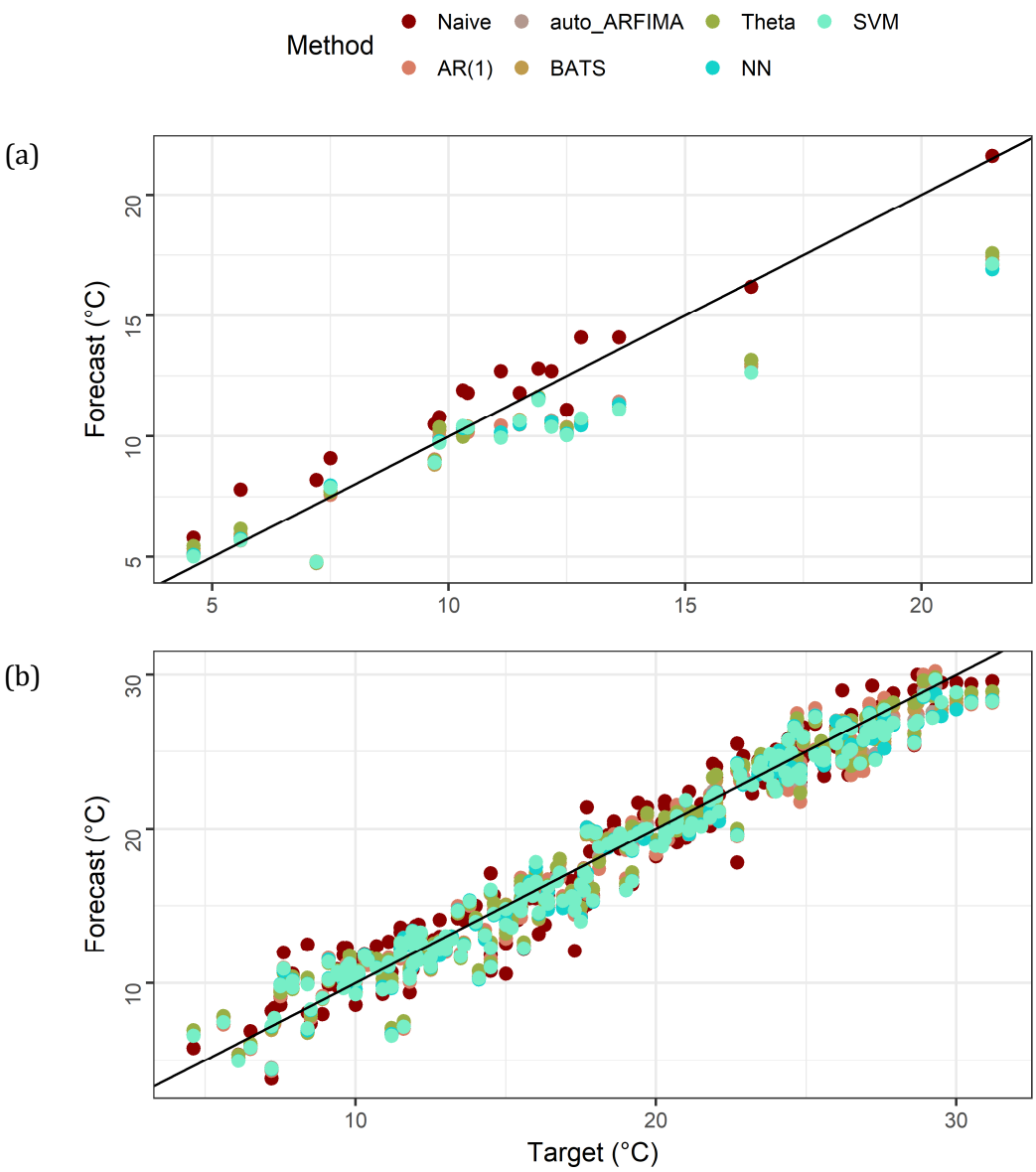


Figure 6.11. (a) One- and (b) twelve-step ahead temperature forecasts, produced for the exploration of Problem 3, in comparison to their corresponding target values.

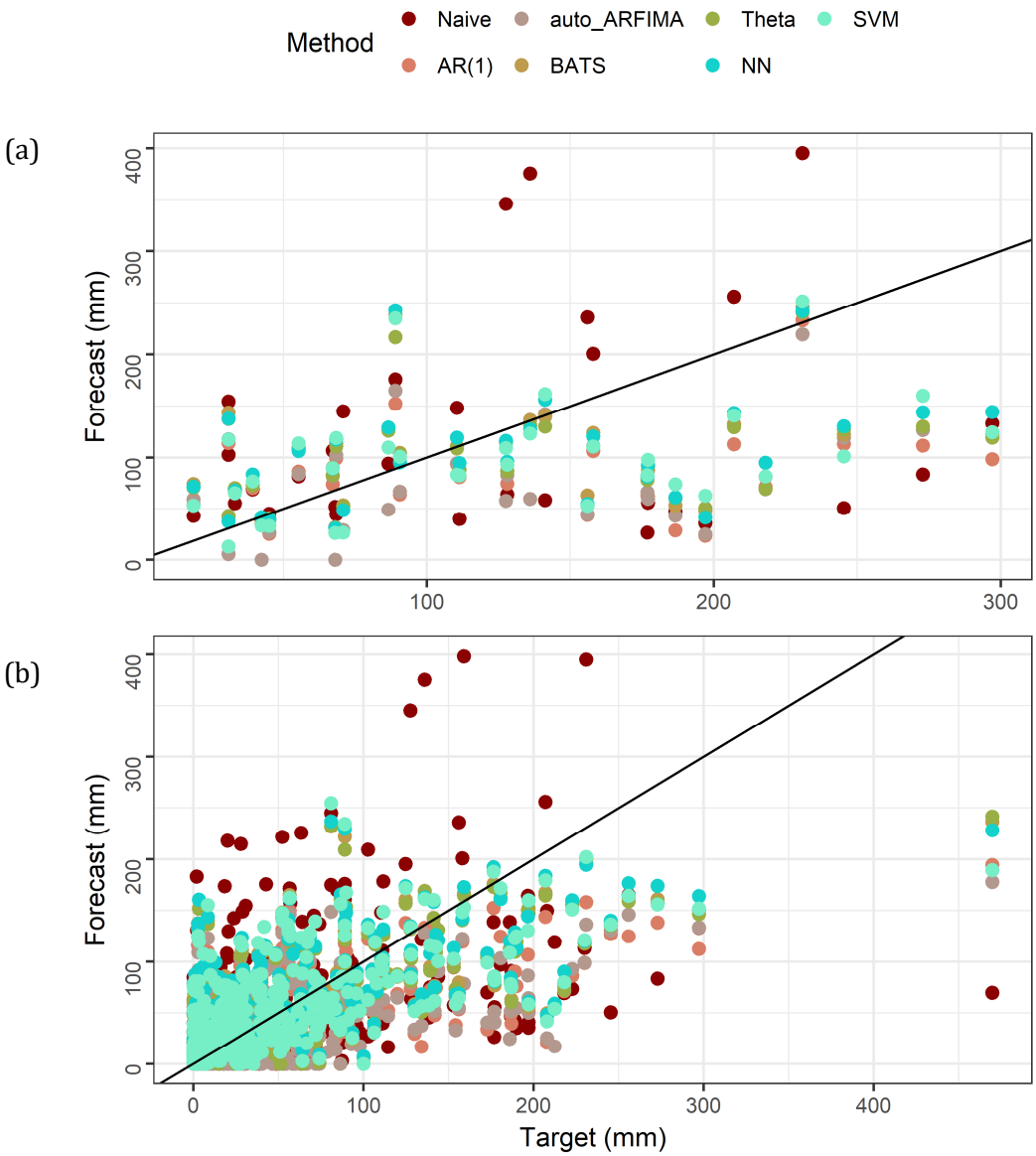


Figure 6.12. (a) One- and (b) twelve-step ahead precipitation forecasts, produced for the exploration of Problem 3, in comparison to their corresponding target values.

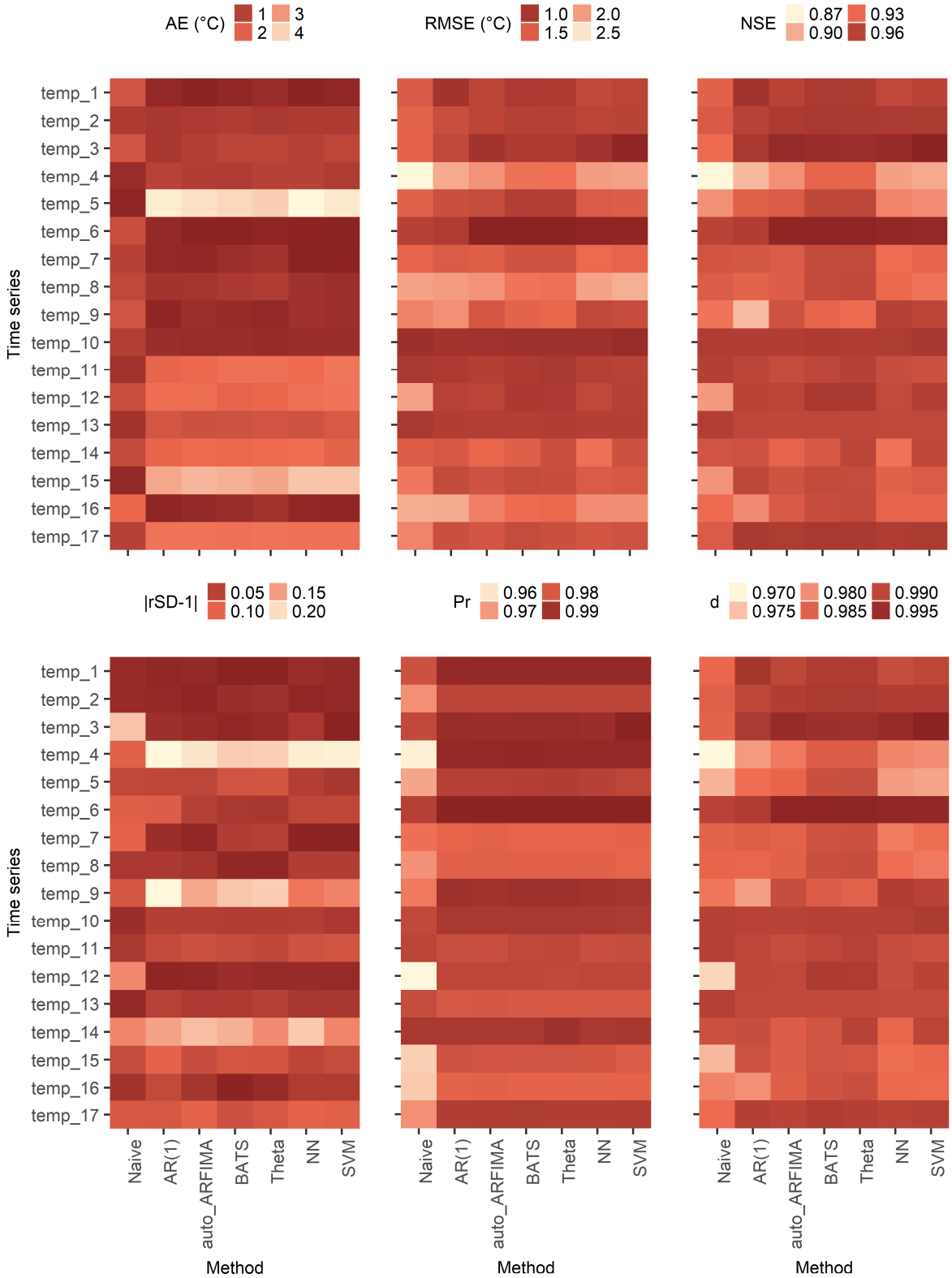


Figure 6.13. Cross-case synthesis for the exploration of Problem 3 using the temperature time series.

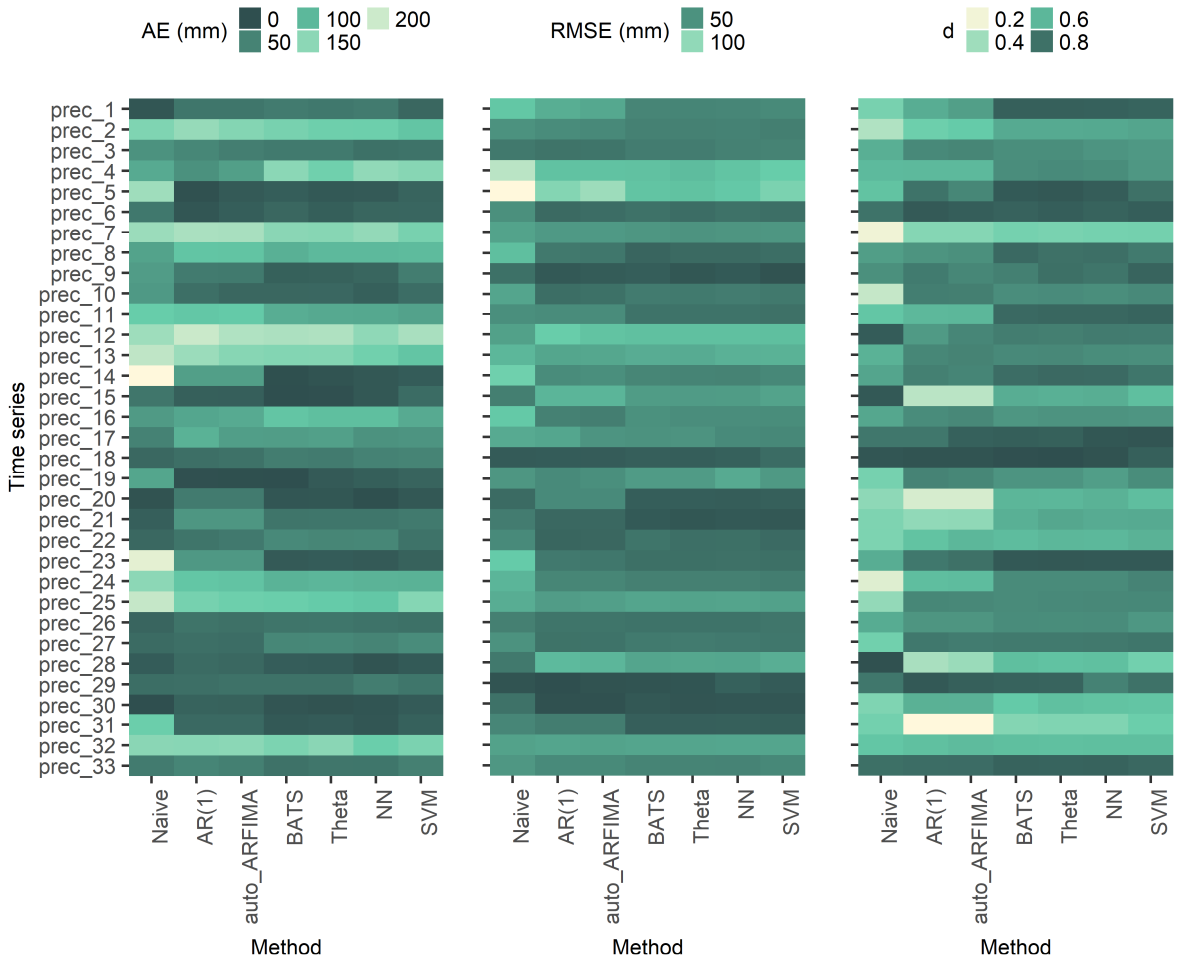


Figure 6.14. Cross-case synthesis for the exploration of Problem 3 using the precipitation time series.

If we further compare Figures 6.11(a), 6.11(b) and 6.12 with Figures 6.2, 6.3 and 6.4 respectively, we observe that the performance of the NN algorithm (when given the 21 regression matrices examined in the present Chapter) can vary more than the performance of the here compared ML and classical methods. This observation does not apply to the case of the SVM algorithm. Finally, we note that the exploration presented in Section 6.3.3 and Chapter 3 effectively complement each other. In fact, the former illustrates and provides evidence on important points by presenting real-world results, while the latter confirms the evidence derived by the former by conducting simulation experiments of large scale. Both illustration and confirmation are integral parts of every theory-building process.

6.3.4 Additional information extracted from the experiments

This Section is devoted to some additional worth-discussed information derived by our multiple-case study. In fact, the results produced mainly for the exploration of Problems 1, 2 and 3 can also be examined from different points of view, which are considered of secondary importance in this Chapter. In Tables 6.6 and 6.7, we present the summary statistics of the metric values, separately for each algorithm, and in Table 6.8 the LRC values for each category of tests. This information stands as a summary of the quantitative information provided by our multiple-case study and, together with Figures 6.2–6.14, can facilitate the below discussion in a satisfactory manner. Regarding an overall assessment of the algorithms, they are all found to mostly have a better average-case forecasting performance than the Naïve benchmark, with the NN algorithm being the worst. This is due to the reported high effect of the lagged regression matrix on the performance of this algorithm. On the contrary, the SVM algorithm has a better average-case

performance, (almost) as good as the one of the best-performing classical algorithms, i.e. BATS, Theta and auto_ARFIMA.

Table 6.6. Summary statistics of the metric values computed for the temperature forecasts. The values reported for the NN and SVM algorithms are computed for the total of the NN and SVM methods implemented in the Chapter respectively.

Metric	Algorithm	Summary statistic		
		Minimum	Median	Maximum
AE (°C)	Naïve	0.10	1.00	2.20
	AR(1)	0.08	0.66	4.41
	auto_ARFIMA	0.02	0.88	4.22
	BATS	0.00	0.86	4.07
	Theta	0.11	1.00	3.92
	NN	0.00	0.98	5.79
	SVM	0.01	0.90	4.52
RMSE (°C)	Naïve	0.92	1.60	2.62
	AR(1)	0.96	1.32	2.12
	auto_ARFIMA	0.74	1.28	1.95
	BATS	0.74	1.14	1.75
	Theta	0.74	1.14	1.73
	NN	0.63	1.70	6.05
	SVM	0.73	1.31	2.30
NSE	Naïve	0.87	0.94	0.97
	AR(1)	0.89	0.96	0.97
	auto_ARFIMA	0.91	0.95	0.98
	BATS	0.93	0.96	0.99
	Theta	0.93	0.96	0.99
	NN	0.44	0.93	0.99
	SVM	0.85	0.95	0.99
rSD	Naïve	0.87	1.01	1.18
	AR(1)	0.90	1.01	1.22
	auto_ARFIMA	0.90	1.01	1.21
	BATS	0.92	1.00	1.19
	Theta	0.92	0.99	1.19
	NN	0.89	1.01	1.24
	SVM	0.89	1.02	1.24
Pr	Naïve	0.96	0.97	0.99
	AR(1)	0.98	0.99	0.99
	auto_ARFIMA	0.98	0.99	0.99
	BATS	0.98	0.99	0.99
	Theta	0.98	0.99	0.99
	NN	0.79	0.98	1.00
	SVM	0.97	0.99	0.99
d	Naïve	0.97	0.98	0.99
	AR(1)	0.98	0.99	0.99
	auto_ARFIMA	0.98	0.99	1.00
	BATS	0.99	0.99	1.00
	Theta	0.98	0.99	1.00
	NN	0.86	0.98	1.00
	SVM	0.97	0.99	1.00

Table 6.7. Summary statistics of the metric values computed for the precipitation forecasts. The values reported for the NN and SVM algorithms are computed for the total of the NN and SVM methods implemented in the Chapter respectively.

Metric	Algorithm	Summary statistic		
		Minimum	Median	Maximum
AE (mm)	Naïve	0	72	239
	AR(1)	2	52	199
	auto_ARFIMA	1	45	178
	BATS	0	41	175
	Theta	2	40	178
	NN	0	51	340
	SVM	0	39	206
RMSE (mm)	Naïve	17	52	147
	AR(1)	15	46	94
	auto_ARFIMA	16	45	105
	BATS	17	41	76
	Theta	18	41	75
	NN	17	47	588
	SVM	11	41	101
NSE	Naïve	−13.20	−0.21	0.48
	AR(1)	−46.17	−0.90	0.64
	auto_ARFIMA	−46.17	−1.01	0.61
	BATS	−4.46	−0.35	0.69
	Theta	−5.07	−0.30	0.70
	NN	−7.55	−0.42	0.86
	SVM	−5.44	−0.44	0.76
rSD	Naïve	0.35	1.05	3.59
	AR(1)	0.55	1.60	4.10
	auto_ARFIMA	0.56	1.55	4.10
	BATS	0.53	1.47	2.53
	Theta	0.53	1.46	2.71
	NN	0.19	1.10	2.60
	SVM	0.48	1.38	2.71
Pr	Naïve	−0.09	0.46	0.93
	AR(1)	0.09	0.62	0.92
	auto_ARFIMA	0.09	0.62	0.93
	BATS	0.21	0.60	0.91
	Theta	0.24	0.60	0.91
	NN	−0.74	0.54	0.96
	SVM	−0.37	0.62	0.92
d	Naïve	0.20	0.59	0.89
	AR(1)	0.17	0.70	0.89
	auto_ARFIMA	0.17	0.73	0.89
	BATS	0.46	0.73	0.90
	Theta	0.47	0.73	0.90
	NN	0.01	0.67	0.97
	SVM	0.25	0.71	0.93

Table 6.8. LRC values computed for each category of tests.

Set of methods (see Table 6.5)	Process	One-step ahead forecasts		Twelve-step ahead forecasts	
		Minimum	Maximum	Minimum	Maximum
1	Temperature	0.62	0.79	0.88	0.97
2		0.70	0.75	0.93	0.96
3		0.69	0.70	0.94	0.94
4		0.70	0.70	0.94	0.95
5		0.69	0.88	0.94	0.96
1	Precipitation	0.00	0.43	0.41	0.56
2		0.21	0.29	0.49	0.52
3		0.25	0.27	0.48	0.52
4		0.25	0.29	0.49	0.51
5		0.21	0.29	0.40	0.52

The reported values of the summary statistics, as well as Figures 6.2–6.4, 6.8, 6.11 and 6.12, reveal that the temperature forecasts are remarkably better than the precipitation ones. This may be explained by the cv estimates presented in Tables 6.2 and 6.3. Finally, in Figure 6.15 we visualize the AE values computed for the one-step ahead temperature forecasts, produced using the set of methods no 5 of Table 6.5, in comparison to their corresponding σ, cv and H estimates for the deseasonalized time series (presented in Table 6.2), while in Figures 6.16 and 6.17 we present an analogous visualization for the AE values computed for the one-step ahead precipitation forecasts and the RMSE values computed for the twelve-step ahead precipitation forecasts respectively, produced for the exploration of Problem 3. The estimated parameters for the deseasonalized precipitation time series are presented in Table 6.3. These figures are representative of the conducted investigation of the existence of a possible relationship between the forecast quality and the estimated parameters for the deseasonalized time series, and provide no evidence of such existence either for temperature or precipitation. This fact may be related to our methodological framework and, in particular, to the way that we handle seasonality to produce better forecasts. These negative results could be viewed in comparison with the results delivered through analogous investigations by Papacharalampous and Tyralis (2020). In this latter work, it is shown that, as the magnitudes of autocorrelation and long-term persistence increase, its gets more likely for a naïve forecast (therein the last year's observation) to be better than forecasts produced by sophisticated forecasting methods. This outcome has been obtained by exploiting approximately 600 annual river flow time series.

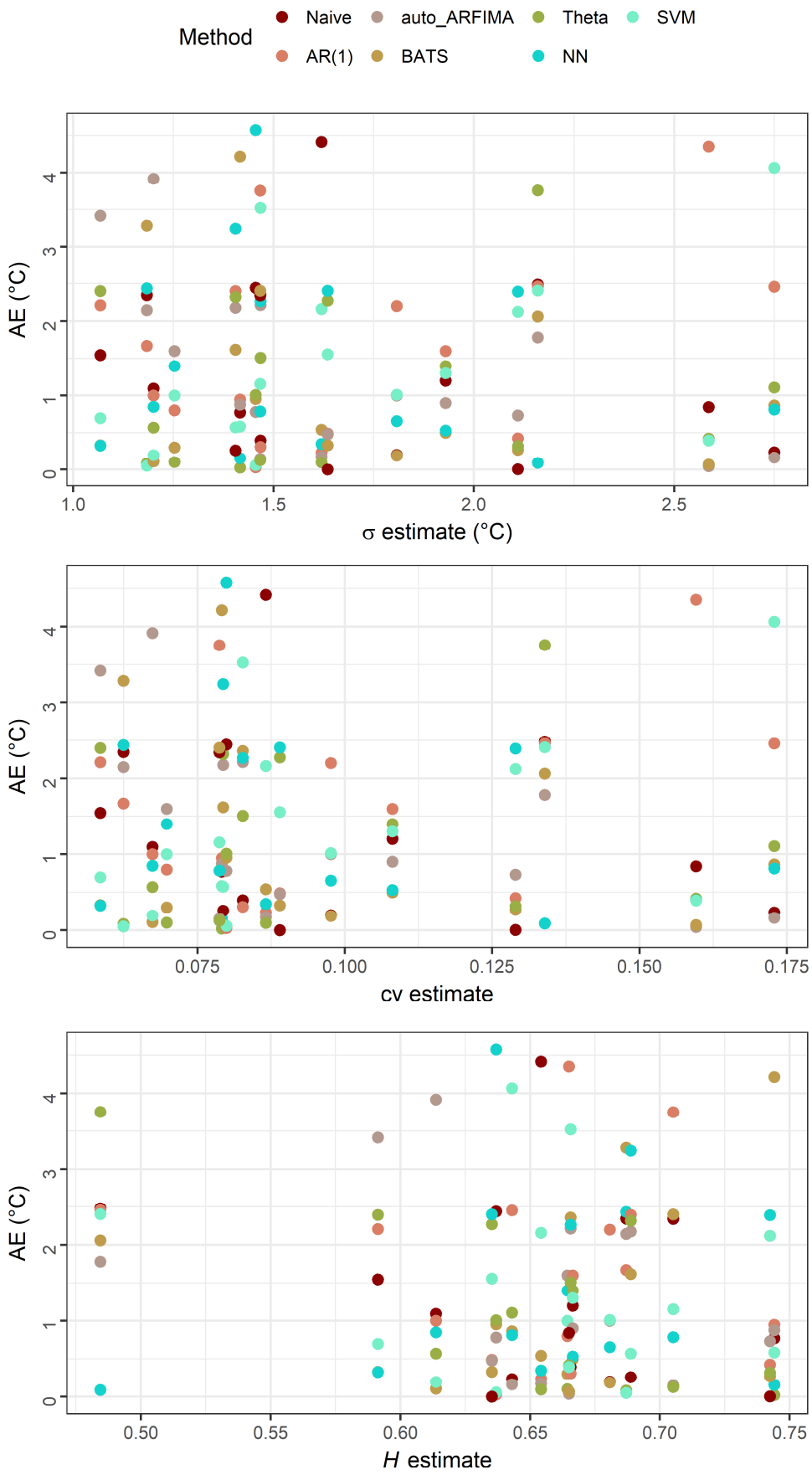


Figure 6.15. AE values of the one-step ahead temperature forecasts, produced by set of methods no 5 (see Table 6.5), in comparison to the σ, cv and H estimates.

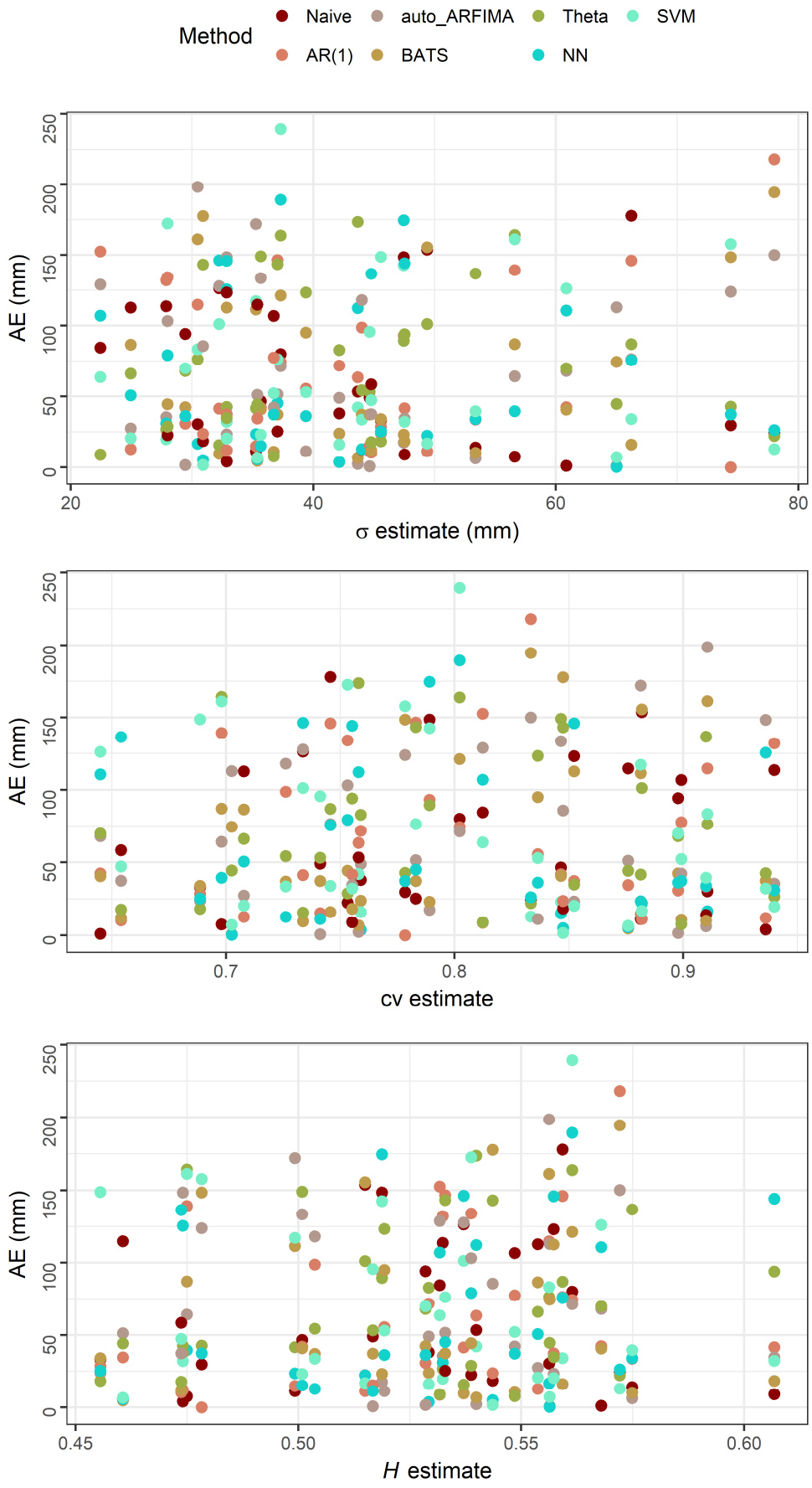


Figure 6.16. AE values of the one-step ahead precipitation forecasts, produced by set of methods no 5 (see Table 6.5), in comparison to the σ, cv and H estimates.

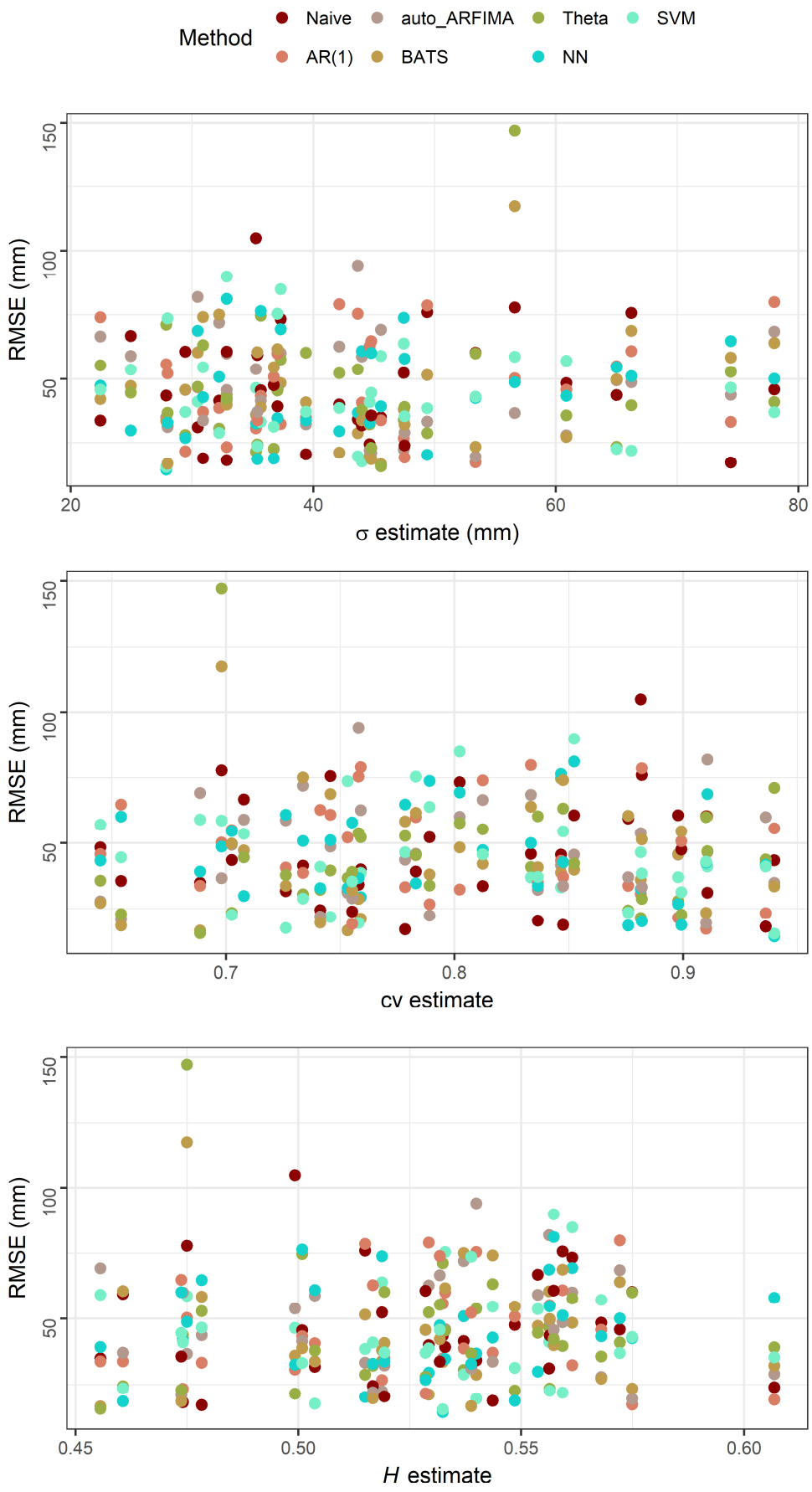


Figure 6.17. RMSE values of the twelve-step ahead precipitation forecasts, produced by set of methods no 5 (see Table 6.5), in comparison to the σ, cv and H estimates.

6.4 Summary and conclusions

We have examined 50 mean monthly temperature and total monthly precipitation time series observed in Greece by applying a fixed methodology to each of them and, subsequently, by performing a cross-case synthesis. The main aim of this multiple-case study is the exploration of three problems associated with univariate time series forecasting using machine learning algorithms, i.e., the (a) lagged variable selection, (b) hyperparameter selection, and (c) comparison between machine learning and classical algorithms. We also present quantitative information about the quality of the forecasts (particularly important for the case of Greece) and search for evidence regarding the existence of a possible relationship between the forecast quality, and the standard deviation, coefficient of variation and Hurst parameter estimates for the deseasonalized time series (used for model-fitting). We have focused on two machine learning algorithms, i.e. neural networks and support vector machines, while we have also included four classical algorithms and a naïve benchmark in the comparisons. We have assessed the one- and twelve-step ahead forecasting performance of the algorithms.

The findings suggest that forecasting methods based on the same machine learning algorithm may exhibit very different performance, to an extent mainly depending on the algorithm and the individual case. In fact, the neural networks algorithm can produce forecasts of many different qualities for a specific individual case, in contrast to the support vector machines one. The performance of the former algorithm seems to be more affected by the selected lagged variables than by the adopted hyperparameter selection procedure (use of predefined hyperparameters or defined after optimization). While no evidence is provided that any of the compared lagged regression matrices systematically leads to better forecasts than the rest, either for the neural networks or the support vector machines algorithms, the results mostly favour using less recent lagged variables. Furthermore, for the algorithms used in the present Chapter hyperparameter optimization does not necessarily lead to better forecasts than the use of the default hyperparameter values of the algorithms. Regarding the comparisons performed between machine learning and classical algorithms, the results indicate that methods from both categories can perform equally well, under the same limitations. The best method depends on the case examined and the criterion of interest, while it can be either machine learning or classical. Some information of secondary importance derived by our experiments is subsequently reported. The average-case performance of the algorithms used to produce one- and twelve-step ahead monthly temperature forecasts ranges between 0.66 °C and 1.00 °C, and 1.14 °C and 1.70 °C, in terms of absolute error and root mean square error respectively. For the monthly precipitation forecasts the respective values are 39 mm and 72 mm, and 41 mm and 52 mm. Finally, no evidence is provided by our multiple-case study that there is any relationship between the forecast quality and the estimated parameters for the deseasonalized time series.

7. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models

In this Chapter, we introduce an ensemble learning post-processing methodology for probabilistic hydrological modelling. This methodology generates numerous point predictions by applying a single hydrological model, yet with different parameter values drawn from the respective simulated posterior distribution. We call these predictions "sister predictions". Each sister prediction extending in the period of interest is converted into a probabilistic prediction using information about the hydrological model's errors. This information is obtained from a preceding period for which observations are available, and is exploited using a flexible quantile regression model. All probabilistic predictions are finally combined via simple quantile averaging to produce the output probabilistic prediction. The idea is inspired by the ensemble learning methods originating from the machine learning literature. The proposed methodology offers larger robustness in performance than basic post-processing methodologies using a single hydrological point prediction. It is also empirically proven to "harness the wisdom of the crowd" in terms of average interval score, i.e., the obtained quantile predictions score no worse –usually better– than the average score of the combined individual predictions. This proof is provided within toy examples, which can be used for gaining insight on how the methodology works and under which conditions it can optimally convert point hydrological predictions to probabilistic ones. A large-scale hydrological application is made in Chapter 8.

7.1 Introduction

Hydrological models are routinely applied for flood forecasting, water resources management and other environmental engineering applications (Montanari 2011). Their history, tracing back to 1850, can be found in Todini (2007), while their optimal design and exploitation (towards uncertainty reduction in hydrological modelling) remain since their early beginnings at the forefront of the hydrological research activity, currently consisting one of the only two modelling challenges included in the 23 major open problems in hydrology, as these problems were identified by Blöschl et al. (2019).

Based on their structure, hydrological models can be primarily classified as follows (see e.g., Solomatine and Wagener 2011; Pechlivanidis et al. 2011): (a) data-driven models, (b) conceptual models and (c) physically-based models. Models of categories (b) and (c) are also jointly called "process-based" (Montanari and Koutsoyiannis 2012). This specific term is largely associated with deterministic models by several authors (see e.g., Beven and Kirkby 1979; Makhlouf and Michel 1994; Perrin et al. 2003; Mouelhi et al. 2006a,b; Efstratiadis et al. 2008; Makropoulos et al. 2008; see also the applications by Madsen 2000; Nayak et al. 2013; Kaleris and Langousis 2017; Széles et al. 2018; Khatami et al. 2019, and the review by Efstratiadis and Koutsoyiannis 2010). On the contrary, models of category (a) are purely statistical. They are mostly borrowed from the statistical learning or machine learning literature (see e.g., Alpaydin 2010; Hastie et al. 2009; James et al. 2013; Witten et al. 2017) to be implemented in the hydrological literature with selected configurations and inputs (see e.g., Minns and Hall 1996; Dibike and Solomatine 2001; Solomatine and Dulal 2003; Nayak et al. 2013; Taormina and Chau 2015; Papacharalampous and Tyralis 2018a; Tyralis and Papacharalampous 2018). In what follows, we use the terms "statistical learning" and "machine learning" interchangeably.

Data-driven and process-based models are, in fact, known to represent two different cultures or schools of thought in hydrological modelling, which need to be compromised in a way that will allow an optimal exploitation of predictability and uncertainty quantification (Todini 2007). In search of such a compromise, optimum (i.e., minimum error) point hydrological predictions (including forecasts) may result by post-processing the outcome of process-based models using statistical point prediction models (see e.g., Brath et al. 2002; Toth et al. 1999; Toth and Brath 2002; Abebe and Price 2003; Toth and Brath 2007). Hydrological post-processing methodologies aiming to convert point hydrological predictions, mostly predictions provided by process-based

models, into probabilistic predictions are also available. These probabilistic methodologies utilize proper statistical models (i.e., probabilistic prediction or simulation models) complementary to the process-based ones. The statistical models used in post-processing are hereafter referred to under the term "error models", as they usually focus on the modelling of the hydrological model's error conditional on selected variables.

Here the interest is in probabilistic hydrological post-processing methodologies, in which the error model is estimated conditional upon the point prediction(s) of the hydrological model by using an independent segment (with respect to the one used for estimating the parameters of the hydrological model) extracted from the historical dataset. Various methodologies of this category are currently available (see e.g., Bock et al. 2018; Bourgin et al. 2015; Dogulu et al. 2015; Farmer and Vogel 2016; López López et al. 2014; Montanari and Brath 2004; Montanari and Grossi 2008; Montanari and Koutsoyiannis 2012; Solomatine and Shrestha 2009; Tyralis et al. 2019a; Wani et al. 2017; see also the methodologies of Chapter 9 herein), amongst other probabilistic hydrological modelling and hydrological forecasting methodologies based on the idea of integrating process-based models and statistical approaches (see e.g., Beven and Binley 1992; Hernández-López and Francés 2017; Kavetski et al. 2002, 2006a; Krzysztofowicz 1999, 2001b, 2002; Krzysztofowicz and Kelly 2000, Krzysztofowicz and Herr 2001; Kuczera et al. 2006; Todini 2008; see also the review by Montanari 2011). Hereafter, we use the comprehensive term "two-stage" by Evin et al. (2014) to imply that the parameters of a probabilistic hydrological post-processing methodology are estimated within two subsequent stages.

Relying on the concept of ensemble simulations and opposed to "basic two-stage post-processing methodologies" utilizing a single point hydrological prediction (see e.g., Dogulu et al. 2015; Farmer and Vogel 2016; López López et al. 2014; Montanari and Brath 2004; Montanari and Grossi 2008; see also the methodologies of Chapter 9 herein), the two-stage post-processing methodology by Montanari and Koutsoyiannis (2012) (hereafter referred to as "MK blueprint methodology") generates a large number of point hydrological predictions by using a single hydrological model (in its basic form; with different parameter values and ensemble inputs). These point predictions are hereafter referred to as "sister predictions" using the terminology of Nowotarski et al. (2016), Wang et al. (2016) and Liu et al. (2017). Different variants of the MK blueprint methodology can be found in Sikorska et al. (2015) and Quilty et al. (2019). The flexibility of the MK blueprint methodology is proved by the latter study, which focuses on probabilistic water demand forecasting using exogenous variables. Its main objective is converting point water demand forecasts produced by machine learning algorithms into probabilistic forecasts. The MK blueprint is outlined in Section 2.7.3.

Here we introduce three novel variants of the MK blueprint methodology. These variants (hereafter collectively referred to as "proposed methodology") are inspired by the ensemble learning methods originating from the machine learning literature, while they are based on the concept of combining probabilistic predictions via simple quantile averaging from the forecasting field. Simple averaging (or equally weighted averaging or averaging) is a special form of linear combination (or linear pooling or weighted averaging) of predictions, in which all weights are equal (see e.g., Granger 1989; Wallis 2011; Lichtendahl et al. 2013; Winkler 2015). According to Granger (1989), (point) prediction combination can be traced back in the study of Barnard (1963), in which two point forecasts were averaged to form an outperforming forecast. Although having its roots in 1963 and more sophisticated combination approaches have been developed since then, this combination in simple fashion is even today suggested by Winkler (2015; see also Lichtendahl et al. 2013), because of its:

- Interpretability.
- Simplicity in modelling.
- Better performance than weighted linear (or other) combinations in many cases.

In fact, as it is quoted from O'Hagan et al. (2006, p. 190) in Lichtendahl et al. (2013), "simple, equally weighted opinion pool is hard to beat in practice". Moreover, it is the most common way of combining point or probability distribution function (PDF) forecasts (Lichtendahl et al. 2013;

Wallis 2011). Especially when we are interested in combining a large number of predictions, as it is the case herein, simple averaging is rather the only reasonable option, also reminding us of several ensemble learning methods (see Hastie et al. 2009), e.g., the bagging by Breiman (1996) and random forests by Breiman (2001a), originating from the machine learning literature. These two examples of ensemble learning methods produce a large number of individual predictions and compute their average to finally produce the output prediction. This averaging leads to more accurate predictions, as it reduces their variance (Hastie et al. 2009, pp. 282–288). Similarly, the average of quantile predictions may offer stability in performance, among other advantages. Quantile averaging has the following distinguishing features (Lichtendahl et al. 2013, Section 5; see also the interpretations provided by Winkler 2015):

- Under specific conditions (see e.g., the stylized versions examined in Lichtendahl et al. 2013) a predictor based on quantile averaging is robust. The same applies to a predictor based on PDF averaging.
- Under specific conditions (see e.g., the stylized versions examined in Lichtendahl et al. 2013) the average of quantile predictions scores no worse –usually better– than the average of scores of the combined individual predictions. This property (also applying to PDF averaging) is referred to as "ability to harness the wisdom of the crowd". Still, it has to be empirically proven for the problem and scores of interest.
- Quantile averaging can be convenient in practice, in contrast to PDF averaging.
- Quantile averaging is as useful as (or even more useful than) PDF averaging.

The proposed methodology has been developed in light of the above by also conducting a set of toy experiments (see e.g., Hartmann 1995; Frigg and Hartmann 2006; Klein and Romero 2007; Goldfarb and Ratner 2008; Luczak 2017; Reutlinger et al. 2017). Examples of toy experiments from the probabilistic hydrological modelling literature are available in Krzysztofowicz (1999), Beven and Freer (2001), Stedinger et al. (2008), Farmer and Vogel (2016), and Volpi et al. (2017). Toy models have also been exploited for other modelling situations in geoscience (see e.g., Koutsoyiannis 2006, 2010; see also the references in Koutsoyiannis 2006), while falling into the broader category of simulation or synthetic experiments, which are increasingly conducted within various hydrological contexts, including some more relevant to the present Chapter (see e.g., Kavetski et al. 2002; Vrugt et al. 2003, 2005; Montanari 2005; Vrugt and Robinson 2007; Vrugt et al. 2008; Renard et al. 2010; Schoups and Vrugt 2010; Montanari and Koutsoyiannis 2012; Montanari and Di Baldassarre 2013; Tyralis et al. 2013; Vrugt et al. 2013; Sadegh and Vrugt 2014; Sadegh et al. 2015; Sikorska et al. 2015; Vrugt 2016; Tyralis and Papacharalampous 2017; see also Chapters 3 and 4 herein). Discussions on the significance of this type of experiments can be found in Montanari (2007). In fact, simplified modelling situations can be useful as starting points for achieving effective real-world modelling, especially in cases where analytical solutions exist (see e.g., Volpi 2012).

The aims of the Chapter are to:

1) Introduce a two-stage probabilistic hydrological modelling methodology that exploits in an optimal way (from a predictive modelling perspective) key concepts of the MK blueprint methodology.
2) Inspect the performance of the proposed methodology under known conditions and demonstrate how it works. In particular, we aim at testing whether and under which conditions this methodology can optimally convert point hydrological predictions to probabilistic ones.
3) Illustrate in simple fashion why and when it is meaningful for someone to select the proposed methodology over basic two-stage post-processing methodologies.
4) Increase the understanding on two-stage hydrological post-processing.

As implied by aims 2–4 above and made e.g., by Krzysztofowicz (1999) and Stedinger et al. (2008), we herein present toy examples only. Chapter 8 of this thesis is devoted to the validation

of the herein introduced methodology by using real-world data. In particular, in this latter work a different set of research questions is addressed by conducting a large-scale experiment at monthly time scale. This experiment comprises 270 rainfall-runoff problems, which are found to be well-solved by the proposed methodology, while the larger robustness in performance of this methodology compared to basic two-stage post-processing methodologies is illustrated for all the examined problems. In the same experiment, we also clearly demonstrate the ability of the proposed methodology to harness the wisdom of the crowd.

7.2 An ensemble methodology for probabilistic hydrological modelling

In this Section, we introduce a new methodology for probabilistic hydrological modelling, inspired by the MK blueprint methodology on the one hand and ensemble learning methodologies (see e.g., the review by Sagi and Rokach 2018) on the other hand.

7.2.1 Proposed methodology (with three variants)

In this Section, we present the proposed methodology. The presentation is made in a more formal and systematic manner with respect to Section 2.7, in which a considerable part of the methodological background of the Chapter is summarized. Therefore, we here also set the largest part of the notations used throughout the Chapter. The formal presentation is accompanied by Figure 7.1, which summarizes in a compact way the methodological contribution of the Chapter. In what follows, random variables are underscored, following the Dutch convention.

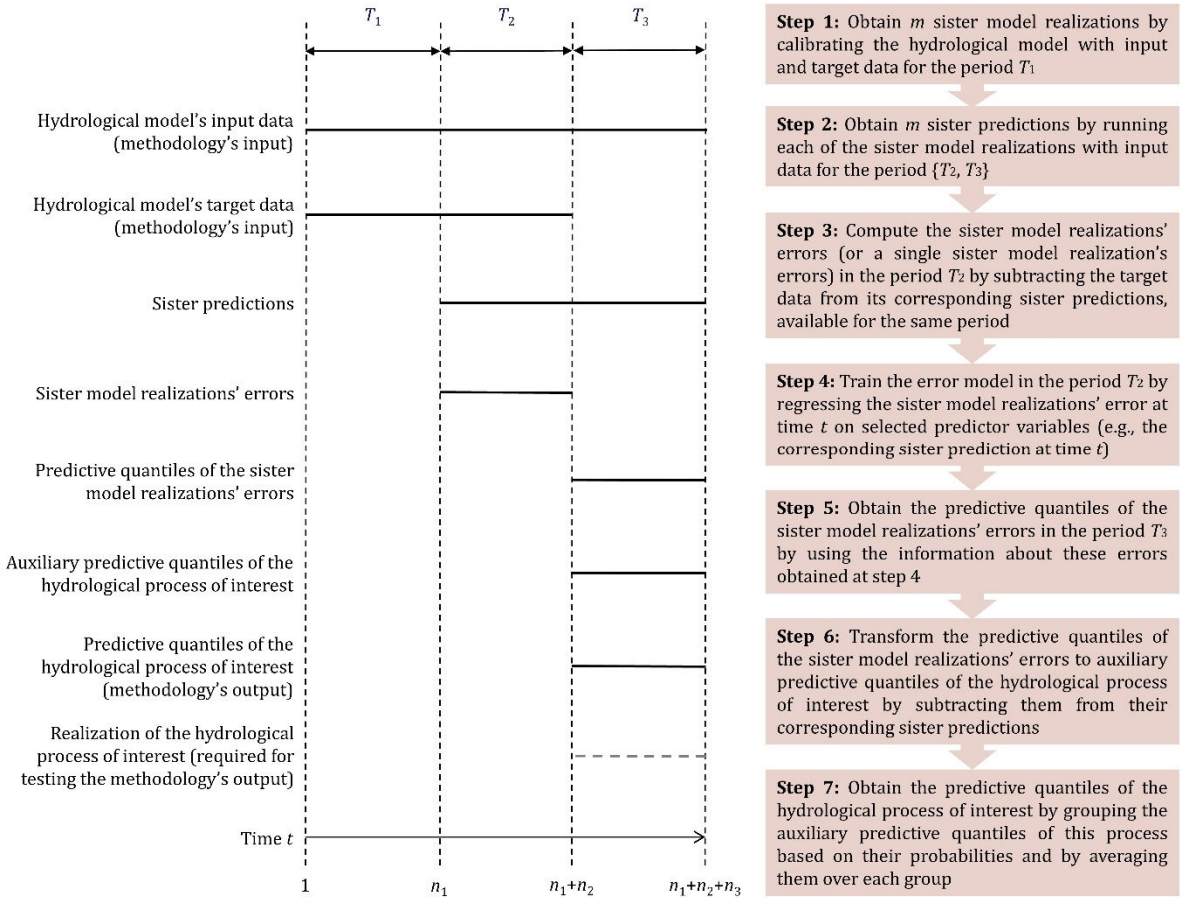


Figure 7.1. Schematic summarizing the proposed methodology. The sister model realizations are defined as variants of a single hydrological model, each using different parameter values. The latter are herein drawn from the respective simulated posterior distribution of model parameters, while they could be also obtained by using informal calibration schemes. Each sister model realization is used for obtaining a single point prediction, referred to as "sister prediction". The number of sister model realizations m should be adequately large. The realization of the hydrological process of interest, considered unknown at the time of the prediction, is denoted with a light grey dashed line.

Let $\underline{\boldsymbol{y}}$ be a stochastic process (typically a hydrological process, e.g., a streamflow or river discharge process), which is expressed in discrete time by Equation (7.1). In the following notations, the subscript of the variables $\underline{\boldsymbol{y}}$ indicates the time t or the time period. We wish to probabilistically predict the stochastic process $\underline{\boldsymbol{y}}_{T_3}$ (hereafter referred to as "hydrological process of interest"), the realization of which is considered unknown at the time of the prediction. At this end, we assume the stochastic processes $\underline{\boldsymbol{x}}_i$, where $i \in \{1, \ldots, n_0\}$ (denoting the sequential number assigned to each), and $\underline{\boldsymbol{x}}$, which are informative about $\underline{\boldsymbol{y}}$, and are expressed in discrete time by Equations (7.2) and (7.3), respectively. In the following notations, the subscript of the variables $\underline{\boldsymbol{x}}$ and second subscript of the variables $\underline{\boldsymbol{x}}_i$ (separated by a comma from the first subscript) indicate the time t or the time period. Let us also assume that the observations $\boldsymbol{x}_{T_3}$ are known at the time of the prediction.

$$\underline{\boldsymbol{y}} := \underline{\boldsymbol{y}}_T := (\underline{y}_1, \ldots, \underline{y}_{n_1}, \underline{y}_{(n_1+1)}, \ldots, \underline{y}_{(n_1+n_2)}, \underline{y}_{(n_1+n_2+1)}, \ldots, \underline{y}_{(n_1+n_2+n_3)})^{\mathrm{T}}: (n_1+n_2+n_3) \times 1 \tag{7.1}$$

$$\underline{\boldsymbol{x}}_i := \underline{\boldsymbol{x}}_{i,T} := (\underline{x}_{i,1}, \ldots, \underline{x}_{i,n_1}, \underline{x}_{i,(n_1+1)}, \ldots, \underline{x}_{i,(n_1+n_2)}, \underline{x}_{i,(n_1+n_2+1)}, \ldots, \underline{x}_{i,(n_1+n_2+n_3)})^{\mathrm{T}}: (n_1+n_2+n_3) \times 1 \tag{7.2}$$

$$\underline{\boldsymbol{x}} := (\underline{\boldsymbol{x}}_1, \ldots, \underline{\boldsymbol{x}}_i, \ldots, \underline{\boldsymbol{x}}_{n_0}): (n_1+n_2+n_3) \times n_0 \tag{7.3}$$

Let S be an arbitrary hydrological model, typically a (deterministic) point prediction model (e.g., a process-based hydrological model) that is suitable for predicting a variable $\underline{y}_t$ given the

observations $\boldsymbol{x}_t$. The equations of such models may involve a variety of parameters, inputs (e.g., precipitation and temperature data at given time steps) and state variables (e.g., soil moisture levels, water table levels, snow cover). State variables are internal variables that describe the state of the catchment during simulation and change as a result of the modelling process (Beven 2012, pp. 5, 67, 176). Note also that S could be used for forecasting $\underline{y}_t$ given forecasts instead of observations (see Klemeš 1986). Under this modelling approach, $\underline{y}_t$ is the dependent or response variable and $\underline{\boldsymbol{x}}_t$ are the predictor variables at time t (as assumed here), both expressed in stochastic terms. Let also $\underline{\boldsymbol{\theta}}$ represent in stochastic terms the parameters of S, defined by Equation (7.4).

$$\underline{\boldsymbol{\theta}} := (\underline{\theta}_1, \ldots, \underline{\theta}_j, \ldots, \underline{\theta}_n): 1 \times n \tag{7.4}$$

Moreover, let us define m variants of S, each using different parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$, where m is adequately large (as large as our computational resources permit). These variants are hereafter referred to as "sister model realizations". The parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$ are obtained by exploiting information from the period T_1. This exploitation can take various forms, such as simulation of the posterior distribution of $\underline{\boldsymbol{\theta}}$ (by using Bayesian methods) or artificial simulation of $\underline{\boldsymbol{\theta}}$ by using some type of randomization applied to a "best parameter estimate". The latter could be obtained by optimizing an objective function of our preference. Random selection of the parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$ could also be an option. Herein, we follow the Bayesian approach (see Section 2.5), as described in detail in Section 7.3.4.

Once the sister model realizations are defined, they are all applied in the period $\{T_2, T_3\}$. The resulted m sister predictions also extend in the period $\{T_2, T_3\}$. Let $\zeta_{k,t}$ be the point prediction at time $t \in \{T_2, T_3\}$ provided by the sister model realization that is defined by $\boldsymbol{\theta}_k$ (hereafter referred to as "k^{th} sister model realization"). This point prediction is hereafter referred to as "k^{th} sister prediction" at time t to be distinguished from the remaining $m-1$ sister predictions at time t. In this case, $\zeta_{k,t}$ is obtained under the single-value transformation expressed by Equation (7.5), where $\boldsymbol{x}_t$ are the inputs to the model and $\boldsymbol{s}_t$ the values of the state variables at time t. We should note here again that the assumptions expressed through Equation (7.5) may vary from model to model. The input to S could also include information from preceding time steps (e.g., $\boldsymbol{x}_{t-1}$, $\mathbf{x}_{t-2}$, $\boldsymbol{x}_{t-3}$, …), while the toy hydrological models used herein do not involve state variables $\boldsymbol{s}_t$ in their equations.

$$\zeta_{k,t} = S(\boldsymbol{\theta}_k, \boldsymbol{x}_t, \boldsymbol{s}_t) \tag{7.5}$$

At time $t \in \{T_2, T_3\}$, the k^{th} sister prediction $\zeta_{k,t}$ deviates from its target observation y_t, as expressed by Equation (7.6). The deviation $\varepsilon_{k,t}$, ignored by convention in the output of any point prediction model, is hereafter referred to as "k^{th} sister model realization's error" at time t and can be assumed as a realization of a random variable $\underline{\varepsilon}_k$. Such realizations are assumed to be informative about the uncertainty of the predictand y_t conditional upon the k^{th} sister prediction. Under this view, the sister model realizations' errors in the period T_2, i.e., $\boldsymbol{\varepsilon}_{k,T_2}\ \forall\ k \in \{1, \ldots, m\}$, computed using the sister predictions $\boldsymbol{\zeta}_{k,T_2}\ \forall\ k \in \{1, \ldots, m\}$ alongside with their targeted observations $\boldsymbol{y}_{T_2}$ (available), consist historical information that can be exploited for quantifying the predictive uncertainty in the period T_3.

$$\varepsilon_{k,t} := \zeta_{k,t} - y_t \tag{7.6}$$

The proposed methodology is subdivided into three alternative variants, which differ to each other only in the exploitation of this historical information. For variants 1 and 2, we subsequently compute $\boldsymbol{\varepsilon}_{k,T_2}\ \forall\ k \in \{1, \ldots, m\}$, while for variant 3 we compute $\boldsymbol{\varepsilon}_{k_o,T_2}$ for a randomly selected sister prediction $\boldsymbol{\zeta}_{k_o,T_2}$ with $k_o \in \{1, \ldots, m\}$. The exploitation of the related information is made by using an error model M, which falls into the category of statistical learning regression models that are suitable for predicting quantiles (see e.g., the quantile regression model detailed in Section 2.6.2). Let $e_{p,k,t}$ be the prediction of the conditional quantile with probability p of the k^{th} sister model realization's error at time t, obtained by using a trained version of M. Under this modelling approach, $\varepsilon_{k,t}$ is assumed to depend on selected informative variable(s). For reasons of simplicity, $\zeta_{k,t}$ is the only predictor variable considered herein for all three variants. The latter differ in the training of M. Specifically:

- Variant 1 trains M separately for each sister model realization. The training is, therefore, made m times, each time on a different dataset formed by using a different sister prediction $\boldsymbol{\zeta}_{k,T_2}$ and its corresponding errors $\boldsymbol{\varepsilon}_{k,T_2}$, where $k \in \{1, \ldots, m\}$;
- Variant 2 trains M collectively for all sister model realizations. The training is, therefore, made once on a single dataset formed by using all sister predictions $\{\boldsymbol{\zeta}_{k,T_2}, k = 1, \ldots, m\}$ and their corresponding errors $\{\boldsymbol{\varepsilon}_{k,T_2}, k = 1, \ldots, m\}$;
- Variant 3 also trains M once; however, the training here is made for an arbitrary sister model realization, i.e., on a dataset formed by using a randomly selected sister prediction $\boldsymbol{\zeta}_{k_o,T_2}$ and its corresponding errors $\boldsymbol{\varepsilon}_{k_o,T_2}$, where $k_o \in \{1, \ldots, m\}$, under the assumption that $\boldsymbol{\varepsilon}_{k_o,T_2}$ are informative about $\underline{\varepsilon}_k$ in general.

In what follows, the presentation is made for a single central prediction interval $(1 - \alpha)$, where $\alpha \in (0, 1)$, while the generalization to obtaining multiple central prediction intervals is straightforward. Let also $z_{p,k,t}$ be the obtained quantile with probability $p \in \{\alpha/2, 1 - \alpha/2\}$ of a variable of interest $\underline{y}_t$ conditional upon $\zeta_{k,t}$, hereafter referred to as "k^{th} predictive quantile with probability p" of a variable of interest. Moreover, let $v_{p,t}$ be the finally delivered quantile with probability p of a variable of interest $\underline{y}_t$, hereafter referred to simply as "predictive quantile with probability p" of this variable.

For each sister prediction $\boldsymbol{\zeta}_{k,T_3}$, where $k \in \{1, \ldots, m\}$, we (a) predict the quantiles of the sister model realization's errors $\{\boldsymbol{e}_{p,k,T_3}, p = \alpha/2, 1 - \alpha/2\}$ by using the information obtained in the preceding step, and (b) transform these predictive quantiles to "auxiliary predictive quantiles" of the hydrological process of interest $\{\boldsymbol{z}_{p,k,T_3}, p = \alpha/2, 1 - \alpha/2\}$ by subtracting them from their corresponding sister prediction $\boldsymbol{\zeta}_{k,T_3}$. At step (a), each trained version of M is applied to predict the error quantiles of its corresponding sister prediction for variant 1, while for variants 2 and 3 the same trained version of M is applied to predict the error quantiles of all sister predictions. Finally, at each time $t \in T_3$ we group the auxiliary predictive quantiles of the hydrological process of interest based on their corresponding probability p (e.g., probability 0.95) to average them over each group. The resulted time series are the delivered quantile predictions $\{\boldsymbol{v}_{p,T_3}, p = \alpha/2, 1 - \alpha/2\}$.

For the sake of completeness, variants 1–3 are algorithmically formulated in Tables 7.1–7.3. We note that these variants reduce to the same method in the case that a single point prediction is generated, i.e., for $m = 1$. In this case, the proposed methodology would fall into the category of basic two-stage post-processing methodologies using statistical learning regression models for quantile prediction (see e.g., López López et al. 2014; Dogulu et al. 2015; see also Chapter 9 herein). The presentation is made for a single central prediction interval $(1 - \alpha)$, where $\alpha \in (0, 1)$, formed by the predictive quantiles with probability p, where $p \in \{\alpha/2, 1 - \alpha/2\}$. The generalization to obtaining multiple central prediction intervals is straightforward.

Table 7.1. Algorithmic formulation of the proposed methodology (variant 1). The presentation is made for a single central prediction interval $(1 - \alpha)$, where $\alpha \in (0, 1)$, while the generalization to obtaining multiple central prediction intervals is straightforward. The repeated procedures are reported with different text alignment. Note that (i) the parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$ could be alternatively obtained through informal calibration schemes, and (ii) more predictors could be exploited in regression.

Step	Procedure
1	Simulate the posterior distribution of $\underline{\boldsymbol{\theta}}$ using information for the time period T_1, i.e., obtain $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$, for m sufficiently large
	Repeat steps 2–6 $\forall\, k \in \{1, \ldots, m\}$
2	Obtain the k^{th} sister prediction for the time period $\{T_2, T_3\}$, i.e., obtain $\boldsymbol{\zeta}_{k,\{T_2, T_3\}}$ according to: $\boldsymbol{\zeta}_{k,\{T_2, T_3\}} = S(\boldsymbol{\theta}_k, \boldsymbol{x}_{\{T_2, T_3\}})$
3	Compute the k^{th} sister model realization's error for the time period T_2, i.e., obtain $\boldsymbol{\varepsilon}_{k,T_2}$ according to: $\boldsymbol{\varepsilon}_{k,T_2} = \boldsymbol{\zeta}_{k,T_2} - \boldsymbol{y}_{T_2}$
4	Regress the k^{th} sister model realization's error $\varepsilon_{k,t}$ on the k^{th} sister prediction $\zeta_{k,t}$ for the time period T_2, i.e., train M between $\boldsymbol{\varepsilon}_{k,T_2}$ and $\boldsymbol{\zeta}_{k,T_2}$
5	Obtain the predictive quantiles of the k^{th} sister model realization's error for the time period T_3 using the trained M, i.e., obtain $\boldsymbol{e}_{p,k,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, according to: $\boldsymbol{e}_{p,k,T_3} = M(\boldsymbol{\zeta}_{k,T_3})$
6	Obtain the k^{th} predictive quantiles of the process of interest, i.e., obtain $\boldsymbol{z}_{p,k,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, according to: ○ $\boldsymbol{z}_{(\alpha/2),k,T_3} = \boldsymbol{\zeta}_{k,T_3} - \boldsymbol{e}_{(1-\alpha/2),k,T_3}$ ○ $\boldsymbol{z}_{(1-\alpha/2),k,T_3} = \boldsymbol{\zeta}_{k,T_3} - \boldsymbol{e}_{(\alpha/2),k,T_3}$
7	Obtain the predictive quantiles of the process of interest, i.e., obtain $\boldsymbol{v}_{p,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, by averaging separately $\forall\, t \in T_3$ the predictive quantiles $\{z_{p,k,t}, k = 1, \ldots, m\}$ according to: $v_{p,t} = \sum_{k=1}^{m} z_{p,k,t}$

Table 7.2. Algorithmic formulation of the proposed methodology (variant 2). The presentation is made for a single central prediction interval $(1 - \alpha)$, where $\alpha \in (0, 1)$, while the generalization to obtaining multiple central prediction intervals is straightforward. The repeated procedures are reported with different text alignment. Note that (i) the parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$ could be alternatively obtained through informal calibration schemes, and (ii) more predictors could be exploited in regression.

Step	Procedure
1	Simulate the posterior distribution of $\underline{\boldsymbol{\theta}}$ using information for the time period T_1, i.e., obtain $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$, for m sufficiently large
	Repeat steps 2, 3 $\forall\, k \in \{1, \ldots, m\}$
2	Obtain the k^{th} sister prediction for the time period $\{T_2, T_3\}$, i.e., obtain $\boldsymbol{\zeta}_{k,\{T_2, T_3\}}$ according to: $\boldsymbol{\zeta}_{k,\{T_2, T_3\}} = S(\boldsymbol{\theta}_k, \boldsymbol{x}_{\{T_2, T_3\}})$
3	Compute the k^{th} prediction error for the time period T_2, i.e., obtain $\boldsymbol{\varepsilon}_{k,T_2}$ according to: $\boldsymbol{\varepsilon}_{k,T_2} = \boldsymbol{\zeta}_{k,T_2} - \boldsymbol{y}_{T_2}$
4	Regress the k^{th} sister model realization's error $\varepsilon_{k,t}$ on the k^{th} sister prediction $\zeta_{k,t}$ for the time period T_2, i.e., training of M between $\boldsymbol{\varepsilon}_{k,T_2}$ and $\boldsymbol{\zeta}_{k,T_2}$. The training is performed collectively for all $k \in \{1, \ldots, m\}$.
	Repeat steps 5, 6 $\forall\, k \in \{1, \ldots, m\}$
5	Obtain the predictive quantiles of the k^{th} sister model realization's error for the time period T_3 using the trained M, i.e., obtain $\boldsymbol{e}_{p,k,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, according to: $\boldsymbol{e}_{p,k,T_3} = M(\boldsymbol{\zeta}_{k,T_3})$
6	Obtain the k^{th} predictive quantiles of the process of interest, i.e., obtain $\boldsymbol{z}_{p,k,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, according to: ○ $\boldsymbol{z}_{(\alpha/2),k,T_3} = \boldsymbol{\zeta}_{k,T_3} - \boldsymbol{e}_{(1-\alpha/2),k,T_3}$ ○ $\boldsymbol{z}_{(1-\alpha/2),k,T_3} = \boldsymbol{\zeta}_{k,T_3} - \boldsymbol{e}_{(\alpha/2),k,T_3}$
7	Obtain the predictive quantiles of the process of interest, i.e., obtain $\boldsymbol{v}_{p,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, by averaging separately $\forall\, t \in T_3$ the predictive quantiles $\{z_{p,k,t}, k = 1, \ldots, m\}$ according to: $v_{p,t} = \sum_{k=1}^{m} z_{p,k,t}$

Table 7.3. Algorithmic formulation of the proposed methodology (variant 3). The presentation is made for a single central prediction interval $(1 - \alpha)$, where $\alpha \in (0, 1)$, while the generalization to obtaining multiple central prediction intervals is straightforward. The repeated procedures are reported with different text alignment. Note that (i) the parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$ could be alternatively obtained through informal calibration schemes, and (ii) more predictors could be exploited in regression.

Step	Procedure
1	Simulate the posterior distribution of $\underline{\boldsymbol{\theta}}$ using information for the time period T_1, i.e., obtain $\{\boldsymbol{\theta}_k, k = 1, \ldots, m\}$, for m sufficiently large
	Repeat step 2 $\forall\, k \in \{1, \ldots, m\}$
2	Obtain the k^{th} sister prediction for the time period $\{T_2, T_3\}$, i.e., obtain $\boldsymbol{\zeta}_{k,\{T_2, T_3\}}$ according to: $\boldsymbol{\zeta}_{k,\{T_2, T_3\}} = S(\boldsymbol{\theta}_k, \boldsymbol{x}_{\{T_2, T_3\}})$
3	Select a random $k_o \in \{1, \ldots, m\}$
4	Compute the k_o^{th} sister model realization's error for the time period T_2, i.e., obtain $\boldsymbol{\varepsilon}_{k_o,T_2}$ according to: $\boldsymbol{\varepsilon}_{k_o,T_2} = \boldsymbol{\zeta}_{k_o,T_2} - \boldsymbol{y}_{T_2}$
5	Regress the k_o^{th} sister model realization's error $\varepsilon_{k_o,t}$ on the k_o^{th} sister prediction $\zeta_{k_o,t}$ for the time period T_2, i.e., train M between $\boldsymbol{\varepsilon}_{k_o,T_2}$ and $\boldsymbol{\zeta}_{k_o,T_2}$
	Repeat steps 6, 7 $\forall\, k \in \{1, \ldots, m\}$
6	Obtain the predictive quantiles of the k^{th} sister model realization's error for the time period T_3 using the trained M, i.e., obtain of $\boldsymbol{e}_{p,k,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, according to: $\boldsymbol{e}_{p,k,T_3} = M(\boldsymbol{\zeta}_{k,T_3})$
7	Obtain the k^{th} predictive quantiles of the process of interest, i.e., obtain $\boldsymbol{z}_{p,k,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, according to: ○ $\boldsymbol{z}_{(\alpha/2),k,T_3} = \boldsymbol{\zeta}_{k,T_3} - \boldsymbol{e}_{(1-\alpha/2),k,T_3}$ ○ $\boldsymbol{z}_{(1-\alpha/2),k,T_3} = \boldsymbol{\zeta}_{k,T_3} - \boldsymbol{e}_{(\alpha/2),k,T_3}$
8	Obtain the predictive quantiles of the process of interest, i.e., obtain $\boldsymbol{v}_{p,T_3}$, $\forall\, p \in \{\alpha/2, 1 - \alpha/2\}$, by averaging separately $\forall\, t \in T_3$ the predictive quantiles $\{z_{p,k,t}, k = 1, \ldots, m\}$ according to: $v_{p,t} = \sum_{k=1}^{m} z_{p,k,}$

7.2.2 Remarks on the proposed methodology

The following remarks on the proposed methodology are important:

- The proposed methodology relies on the use of error models that by construction quantify predictive uncertainty, i.e., the total uncertainty of the predictand (parameter uncertainty included). This is why these error models have been exploited within basic hydrological post-processing methodologies. For instance, see the large-sample investigations conducted in Chapter 9 herein.
- The use of numerous parameter sets for the hydrological model is, thus, not a condition for properly considering parameter uncertainty. This is why the hydrological model's parameters can be obtained through informal calibration schemes.
- Simple quantile averaging is a novel methodological step compared to the original blueprint by Montanari and Koutsoyiannis (2012), and its variants by Sikorska et al. (2015) and Quilty et al. (2019). It is introduced herein to allow the accommodation of statistical learning regression models that are suitable for predicting quantiles into the methodology.
- Simple quantile averaging does not harm predictive uncertainty quantification. In fact, it works in the same way as simple PDF averaging. The latter has been exploited in hydrological post-processing concepts, e.g., by Vrugt (2018; see also 2019). We should note here again that, according to Lichtendahl et al. (2013), simple quantile averaging is as useful as (or even more useful than) simple PDF averaging. In Vrugt (2018, 2019), various point hydrological predictions are obtained by using different hydrological models (under a multi-model approach) and not by using a single hydrological model, as it is the case in the proposed methodology. These point hydrological predictions are first converted to PDF hydrological predictions (via post-processing) and then combined via (simple) PDF averaging.

7.2.3 Differences from other two-stage post-processing methodologies

Since the proposed methodology can be regarded as a set of variants of the MK blueprint methodology, some key changes with respect to the precursor methods should be underlined. These are the following:

- The proposed methodology is formulated to work with given data, i.e., it does not explicitly consider input data uncertainty (stemming e.g., from measurement errors; under the assumption of error-free data). Note that input data uncertainty could be considered (in a similar way to the one adopted in the precursor methods) if enough information is available to characterize it.
- The error models adopted in the precursor variants, i.e., the meta-Gaussian bivariate distribution model used in simulation mode by Montanari and Koutsoyiannis (2012), and the kNN model used by Sikorska et al. (2015) and Quilty et al. (2019), are here replaced by a statistical learning regression model that is suitable for predicting quantiles.
- Alternative options for the modelling of the sister model realizations' errors are provided. Additionally to variant 3, which extracts this type of information from a single sister prediction (as made in the MK blueprint methodology), we also include variants 1 and 2. These variants extract information about the hydrological model's error from all sister predictions.
- Ensemble predictions (i.e., individual predictions to be combined within an ensemble learning methodology; instead of ensemble simulations, i.e., individual simulations collectively composing an ensemble) are obtained and ensemble prediction averaging is involved. In fact, the proposed methodology falls into the category of ensemble learning methods (see e.g., Hastie et al. 2009, Chapter 16), while the original variant, and the variants by Sikorska et al. (2015) and Quilty et al. (2019) are ensemble simulation methods.

Some key differences from other two-stage post-processing methodologies are also summarized subsequently:

- In contrast to basic two-stage post-processing methodologies using flexible quantile regression models (see e.g., Solomatine and Shrestha 2009; López López et al. 2014; Dogulu et al. 2015; Wani et al. 2017; see also Chapter 9 herein), the proposed methodology is an ensemble learning methodology, as it combines multiple predictions to offer improved predictive performance.
- In contrast to multi-model ensemble learning post-processing methodologies, the proposed methodology utilizes a single hydrological model.
- In contrast to ensemble learning post-processing methodologies using multiple error models (see e.g., Tyralis et al. 2019a for the first stacked generalization approach to hydrological post-processing, and Chapter 9 herein for an equal-weight combiner of six error models), the proposed methodology utilizes a single error model.

7.3 Experimental methodology

Here we present the experimental methodology adopted for the conducted toy model investigation. Statistical software information is independently summarized in Section 2.9.4.

7.3.1 Toy data simulation

We simulate the three large toy datasets presented in Figure 7.2. Each of these datasets includes 12 000 pairs of (x_t, y_t) values, drawn i.i.d. from the populations described in Table 7.4. Benchmark remarks on the selection of the simulating models are also provided in Table 7.4.

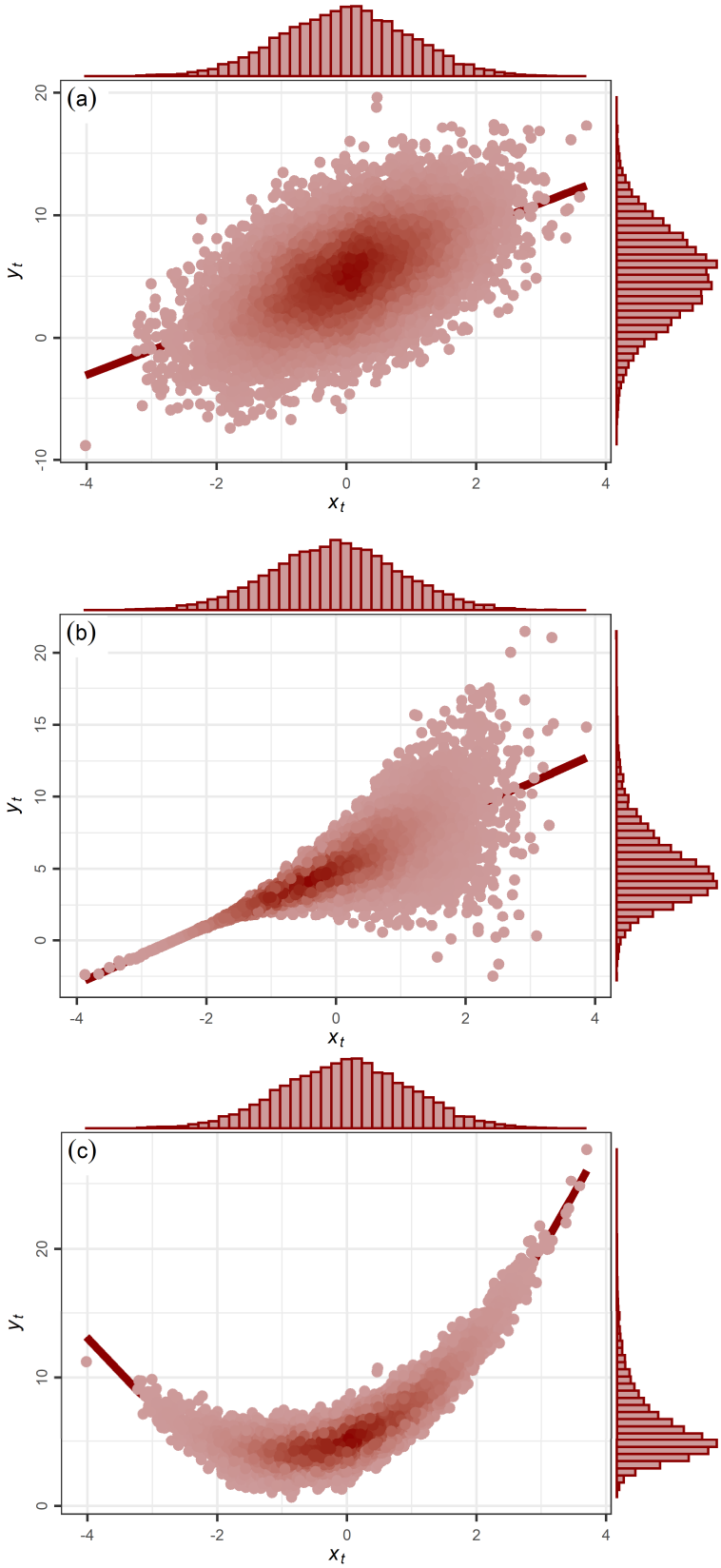


Figure 7.2. Toy datasets (a–c) 1–3. Details about their simulation are presented in Table 7.4. The pairs (x_t, y_t) are depicted with coloured bubbles (pink for low density and red for high density), while the red lines are the plots of the functions $y_t = f(x_t)$, i.e., the deterministic parts of the simulating models. The deviation in the vertical direction of a red line from any bubble is a realization of $\underline{u}_t$.

Table 7.4. Information about toy data simulation. The simulating models' types and parameters are selected to ensure a clear demonstration of the proposed methodology. The toy datasets are depicted in Figure 7.2. The function f and the random variables $\underline{x}_t$, $\underline{u}_t$ and $\underline{y}_t$, where t denotes the time, are defined as follows for each simulating model.

Toy dataset	Simulating model	Remarks (see also Sections 2.3.1 and 2.6.2)
1	$\underline{x}_t \sim N(\mu = 0, \sigma^2 = 1^2)$ $f(x) := 5 + 2x$ $\underline{u}_t \sim N(\mu = 0, \sigma^2 = 3^2)$ $\underline{y}_t := f(\underline{x}_t) + \underline{u}_t$	i) There exists an analytical solution to delivering prediction intervals for this dataset, and ii) the assumptions of the simple linear regression model are proper for this dataset.
2	$\underline{x}_t \sim N(\mu = 0, \sigma^2 = 1^2)$ $f(x) := 5 + 2x$ $\underline{u}_t \sim N(\mu = 0, \sigma^2 = (0.2f(\underline{x}_t))^2)$ $\underline{y}_t := f(\underline{x}_t) + \underline{u}_t$	The assumption of homoscedasticity of the error term (made by the simple linear regression model) is not proper for this dataset.
3	$\underline{x}_t \sim N(\mu = 0, \sigma^2 = 1^2)$ $f(x) := 5 + 2x + x^2$ $\underline{u}_t \sim N(\mu = 0, \sigma^2 = 1^2)$ $\underline{y}_t := f(\underline{x}_t) + \underline{u}_t$	The assumption of linearity in the relationship between the predictor and the response (made by the simple linear regression model) is not proper for this dataset.

7.3.2 Statistical learning models

We implement three statistical learning regression models. Following the suggestions by Abrahart et al. (2008), we subsequently emphasize on reproducibility and not on exhaustive descriptions of these models. Some related theoretical information is independently given in Chapter 2. The first regression model exploited in this Chapter is the linear regression model (see Section 2.3.1). The assumptions of this model might not be efficient for real-world hydrological modelling applications; however, it offers the advantage of being interpretable (Hastie et al. 2009, p. 43). We use it as described in Sections 7.3.2–7.3.5, particularly focusing on the violation of the homoscedasticity assumption and on how its replacement with more flexible models under a predictive modelling view could result in improved probabilistic predictions. The second regression model is the quadratic regression one (see Section 2.3.1). Lastly, the quantile regression model (see Section 2.6.2) is the third regression model implemented in this Chapter. This model offers a good compromise between interpretability (offered by the linear regression model) and flexibility (offered by more sophisticated statistical learning methods). In this Chapter, it represents all regression models that can directly provide the predictive quantiles of the response variable, while they are also appropriate for modelling heteroscedasticity.

7.3.3 Toy experiments, prediction schemes and expected outcomes

We conduct four toy experiments. Within each of these experiments we assess six ensemble schemes in obtaining interval predictions. The ensemble schemes are based on the proposed methodology, while they are defined by their underlying variants of this methodology and their adopted error models, as prescribed by Table 7.5. They can be applied by using any point prediction model as (toy) hydrological model. Depending on the toy experiment, we adopt either the linear regression model or the quadratic regression model as toy hydrological models (see Table 7.6). These regression models are described in Section 2.3.1. They are utilized by the ensemble schemes to generate point predictions of y_t given x_t; therefore, $\underline{y}_t$ is the response variable and $\underline{x}_t$ is the predictor variable, both expressed in stochastic terms. A factor defining a toy experiment, together with the adopted toy hydrological model by all ensemble schemes, is the examined toy dataset (see Table 7.6). The toy datasets are presented in Section 7.3.1.

Table 7.5. Ensemble schemes assessed within the toy experiments.

Ensemble scheme	Variant of the proposed methodology	Outlined algorithm	Error model (from Table 2.3)	Description
1	1	Table 7.1	Linear regression model	Section 2.3.1
2	2	Table 7.2		
3	3	Table 7.3		
4	1	Table 7.1	Quantile regression model	Section 2.6.2
5	2	Table 7.2		
6	3	Table 7.3		

Table 7.6. Toy experiments. The toy datasets are presented in Section 7.3.1. The toy hydrological models are described in Section 2.3.1.

Toy experiment	Toy dataset	Toy hydrological model (from Table 2.3) for all ensemble schemes
1	1	Linear regression model
2	2	
3	3	
4	3	Quadratic regression model

For the application of the ensemble schemes, detailed in Section 7.3.4, and following the definitions and notations provided in Section 7.2, for each toy dataset we define the periods T_1 = {1, …, 1 000}, T_2 = {1 001, …, 2 000} and T_3 = {2 001, …, 12 000}. We include a large amount of information in the period T_3 to facilitate proper testing. To benchmark the toy results obtained using the proposed methodology, we also apply two basic probabilistic prediction schemes, namely the linear regression and quantile regression schemes. Their application is made according to Section 7.3.5. In particular for the case of toy experiment 1, we also consider the analytical solution provided by a Bayesian regression scheme, when the latter is applied under specific assumptions (see Section 7.3.5).

The only a priori theoretically expected outcomes in the conducted toy experiments are the following (see also Table 7.4; outcomes that need to be empirically proven are presented in Section 7.4):

- All three benchmark schemes are expected to perform well within toy experiment 1, in which the simple linear regression problem is solved for a large dataset. This problem is, in fact, the inverse problem with respect to the simulation of the therein utilized dataset for the linear regression model.
- The problem examined within toy experiment 2 is expected to be well-solved by the quantile regression model, while the solution provided by the linear regression model for the same problem is expected to be suboptimal. This problem could be viewed as an extension of the simple linear regression problem (James et al. 2013).
- The linear regression and quantile regression schemes are not the ideal models (when used with a single predictor) to be used for modelling a quadratic relationship. However, their predictions when both applied to toy dataset 3 are expected to not be equivalent.
- Ensemble schemes 1 and 4 are expected to provide the exact same solution with the basic post-processing methodologies using the linear regression and quantile regression models as error models respectively, when applied to toy datasets 1–3 with the simple linear regression model as toy hydrological model. The reason is theoretical; the problem solved by the error model for any point prediction provided by the simple linear regression model is practically the exact same one.
- This equivalence does not hold for any other toy hydrological model, e.g., the quadratic regression one. Therefore, within toy experiment 4 ensemble schemes 1 and 4 are expected to not be equivalent to basic two-stage post-processing methodologies.

7.3.4 Application of ensemble schemes

We describe the application of the ensemble schemes for one toy experiment, as all toy experiments are made in the same manner. The following steps are carried out once for all ensemble schemes:

1) We define 1 000 sister model realizations by obtaining the parameters $\{\boldsymbol{\theta}_k, k = 1, \ldots, 1\,000\}$ of the toy hydrological model. Specifically, we obtain 1 000 random samples of the joint posterior distribution of the toy hydrological model's parameters $\underline{\boldsymbol{\theta}}$ and the variance of its error term $\underline{\sigma}^2$ conditional on the observations of the period T_1. The joint posterior distribution is obtained by using a uniform prior distribution and an inverse prior distribution for $\underline{\boldsymbol{\theta}}$ and $\underline{\sigma}^2$ respectively, as detailed in Section 2.5.1. Figure 7.3 summarizes information about the obtained $\{\boldsymbol{\theta}_k, k = 1, \ldots, 1\,000\}$ for toy experiment 1.

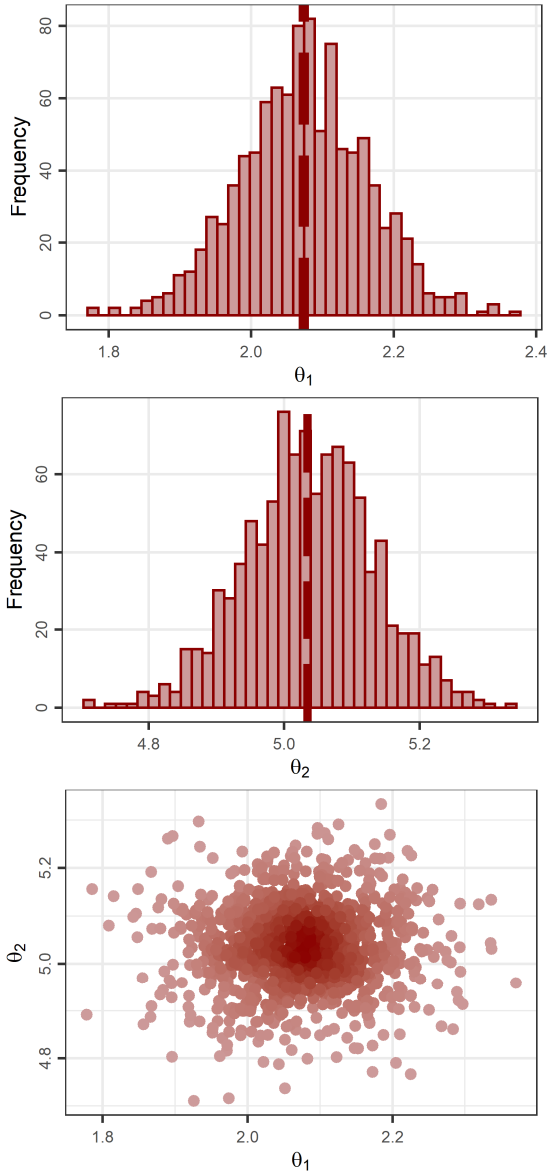


Figure 7.3. Simulated parameter values obtained using information from the period T_1 within toy experiment 1. The median θ_1 and θ_2 values are denoted with red thick dashed line on the presented histograms.

2) We obtain 1 000 sister predictions for the period $\{T_2, T_3\}$. Each sister prediction contains 11 000 values, while it is obtained by implementing a different sister model realization given the same information, i.e., input information for the period $\{T_2, T_3\}$.

3) By using the resulted sister predictions extending in the period T_2 alongside with their corresponding target values, we compute the sister model realizations' errors in the same period. The total number of the computed error values is 1 000 × 1 000 = 1 000 000. These values are considered informative about the sister model realization's errors in the period T_3 under the stationarity and ergodicity assumptions; therefore, they are used at the next step.

The following steps are carried out independently by each ensemble scheme:

4) We train the error model in the period T_2. Specifically, we regress the sister model realizations' error at time t (response variable) on the sister prediction at time t (predictor variable). The error model (linear regression or quantile regression), the number of the error model trainings (1 or 1 000) and the size of the training dataset(s) (1 000 000 or 1 000 pairs of values) depend on the ensemble scheme (see Section 7.2.1). We train the quantile regression model by using the algorithmic routine fully documented in Koenker and d'Orey (1987, 1994). Examples of training datasets are presented in Figure 7.4;

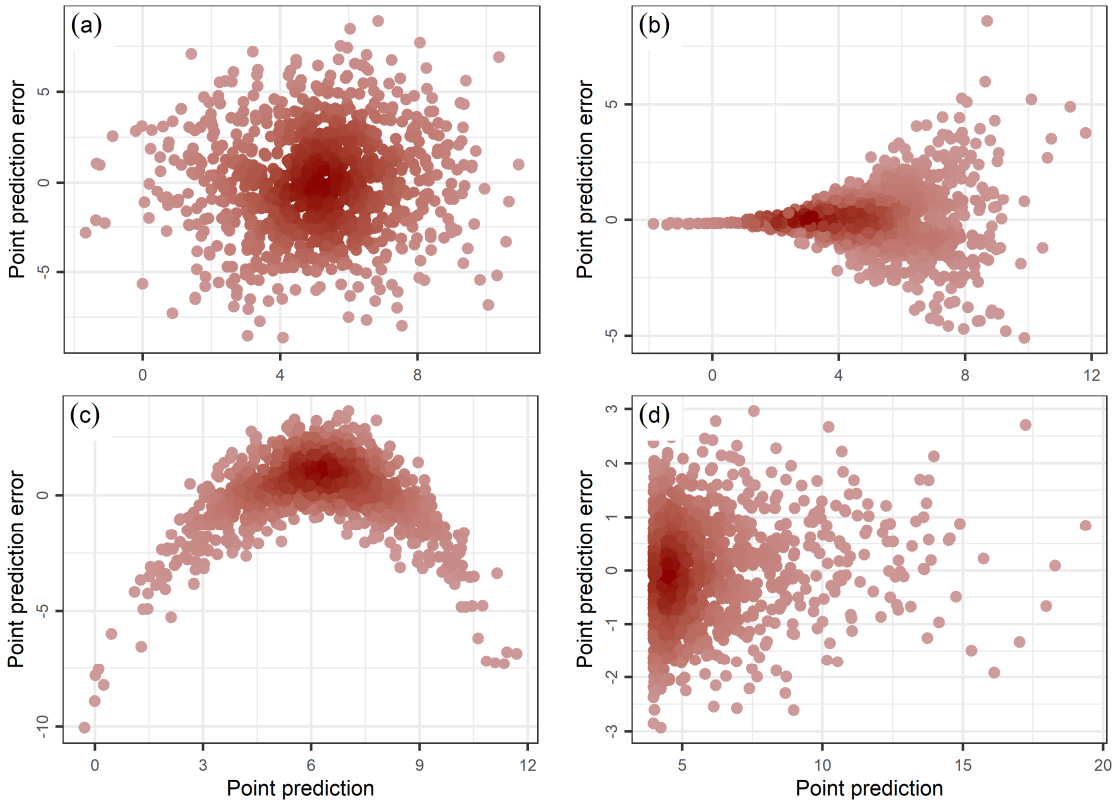


Figure 7.4. Error model training datasets for the ensemble schemes 3 and 6 within the toy experiments (a–d) 1–4.

5) We use each sister prediction extending in the period T_3 to predict a set of selected quantiles, specifically the quantiles with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$, of its corresponding sister model realization's error. The predictions are made by exploiting information obtained in the preceding step (for details see Section 7.2.1). The result of this step is 1 000 probabilistic predictions for 10 000 data points, each consisting of 10 quantile predictions.

6) By subtracting each of these 1 000 × 10 = 10 000 quantile predictions from its corresponding sister prediction, we obtain 1 000 auxiliary probabilistic predictions of the process of interest, each consisting of 10 quantile predictions.

7) Finally, we separately average, for each p (as defined at point 5 above) and at each time $t \in T_3$, all the auxiliary predictive quantiles with probability p, i.e., 1 000 in number predictive quantiles, to obtain the finally delivered predictive quantile with probability p at time t. The

finally delivered predictive quantiles of the process of interest form the 99%, 97.5%, 95%, 90% and 80% central prediction intervals.

7.3.5 Application of benchmark schemes

The linear regression and quantile regression schemes are implemented by (a) training the linear regression and the quantile regression models respectively directly on the data from the period $\{T_1, T_2\}$ and, subsequently, by (b) applying the trained regression model in the period T_3 to predict the quantiles with probability p (as defined at point 5 of Section 7.3.4) of the process of interest. The obtained predictive quantiles are then used to form the 99%, 97.5%, 95%, 90% and 80% central intervals. The predictor variable in regression is $\underline{x}_t$, expressed in stochastic terms.

The Bayesian regression scheme (benchmark within toy experiment 1) is trained by obtaining 1 000 random samples of the joint posterior distribution of the toy hydrological model's parameters $\underline{\boldsymbol{\theta}}$ and the variance of its error term $\underline{\sigma}^2$ conditional on the observations of the period $\{T_1, T_2\}$. A uniform prior distribution and an inverse prior distribution are used for $\underline{\boldsymbol{\theta}}$ and $\underline{\sigma}^2$ respectively, as detailed in Section 2.5.1. Based on this joint posterior distribution of $\underline{\boldsymbol{\theta}}$ and $\underline{\sigma}^2$, the posterior predictive distribution for the period T_3 is obtained, according to the definition of prediction intervals.

7.3.6 Performance assessment

We assess the reliability and sharpness of the obtained interval predictions by computing their coverage probabilities and average widths, respectively. To simultaneously assess both these desired properties of the predictions, we also compute their average interval scores. For benchmarking purposes we also compute the relative improvements, obtained when using a prediction interval of level $(1 - \alpha)$ (provided by a predictor of interest) with respect to another prediction interval of the same level (provided by a benchmark predictor) in terms of average interval score. All computations are made for the period T_3, as detailed in Section 2.8.3.

7.4 Experimental results, interpretations and illustrations

This section is devoted to the toy model investigation of the proposed methodology. This investigation is conducted within a purely statistical framework, while complementing Section 7.2 by largely facilitating the methodology's interpretation. The larger robustness in performance of this methodology compared to basic two-stage post-processing methodologies and its ability to harness the wisdom of the crowd are also illustrated using the obtained results.

7.4.1 Overall interpretation of the proposed methodology

In this section, we answer the following research questions (related to aims 2 and 4 of the Chapter; see Section 7.1): (i) How does the proposed methodology and other two-stage hydrological post-processing methodologies work, and (ii) under which conditions these methodologies work well? In fact, although we focus on the proposed methodology, the presented toy examples can also be used to gain insight into two-stage hydrological post-processing in general. In Table 7.7, we present the coverage probabilities, average widths and average interval scores computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals obtained for all prediction schemes within toy experiment 1, while the respective results obtained for the toy experiments 2–4 are presented in Tables 7.8–7.10 respectively.

Table 7.7. Metric values computed for the prediction intervals delivered by the compared schemes for the period T_3 within the toy experiment 1.

Metric	Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Coverage probability	Bayesian regression	0.988	0.973	0.949	0.895	0.798
	Linear regression	0.989	0.973	0.948	0.897	0.798
	Quantile regression	0.986	0.971	0.945	0.891	0.802
	Ensemble scheme 1	0.989	0.973	0.948	0.895	0.797
	Ensemble scheme 2	0.989	0.972	0.947	0.895	0.797
	Ensemble scheme 3	0.989	0.973	0.948	0.895	0.797
	Ensemble scheme 4	0.987	0.967	0.951	0.890	0.805
	Ensemble scheme 5	0.986	0.968	0.949	0.891	0.804
	Ensemble scheme 6	0.987	0.967	0.951	0.890	0.805
Average width	Bayesian regression	15.29	13.35	11.69	9.82	7.65
	Linear regression	15.40	13.40	11.71	9.83	7.66
	Quantile regression	15.09	13.31	11.62	9.73	7.71
	Ensemble scheme 1	15.36	13.36	11.68	9.80	7.63
	Ensemble scheme 2	15.31	13.32	11.65	9.78	7.62
	Ensemble scheme 3	15.36	13.36	11.68	9.80	7.63
	Ensemble scheme 4	14.98	12.88	11.87	9.70	7.81
	Ensemble scheme 5	14.94	13.03	11.81	9.73	7.76
	Ensemble scheme 6	14.98	12.88	11.87	9.70	7.81
Average interval score	Bayesian regression	17.63	15.69	14.13	12.47	10.62
	Linear regression	17.47	15.67	14.11	12.46	10.61
	Quantile regression	17.77	15.81	14.23	12.52	10.61
	Ensemble scheme 1	17.49	15.68	14.14	12.47	10.61
	Ensemble scheme 2	17.49	15.69	14.14	12.47	10.61
	Ensemble scheme 3	17.49	15.69	14.14	12.47	10.61
	Ensemble scheme 4	17.56	15.82	14.18	12.52	10.65
	Ensemble scheme 5	17.59	15.81	14.18	12.52	10.64
	Ensemble scheme 6	17.57	15.82	14.18	12.52	10.65

Table 7.8. Metric values computed for the prediction intervals delivered by the compared schemes for the period T_3 within the toy experiment 2.

Metric	Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Coverage probability	Linear regression	0.975	0.962	0.946	0.919	0.864
	Quantile regression	0.994	0.986	0.967	0.918	0.824
	Ensemble scheme 1	0.972	0.958	0.939	0.911	0.856
	Ensemble scheme 2	0.972	0.958	0.939	0.911	0.856
	Ensemble scheme 3	0.972	0.958	0.940	0.911	0.856
	Ensemble scheme 4	0.994	0.982	0.960	0.905	0.819
	Ensemble scheme 5	0.994	0.982	0.962	0.905	0.821
	Ensemble scheme 6	0.994	0.982	0.961	0.906	0.819
Average width	Linear regression	7.74	6.73	5.89	4.94	3.84
	Quantile regression	7.75	6.45	5.21	3.99	2.92
	Ensemble scheme 1	7.44	6.47	5.65	4.74	3.70
	Ensemble scheme 2	7.41	6.45	5.64	4.73	3.69
	Ensemble scheme 3	7.44	6.47	5.65	4.74	3.70
	Ensemble scheme 4	7.61	6.06	4.93	3.75	2.86
	Ensemble scheme 5	7.65	6.06	4.97	3.75	2.87
	Ensemble scheme 6	7.63	6.08	4.94	3.77	2.87
Average interval score	Linear regression	14.08	10.50	8.54	6.90	5.40
	Quantile regression	8.86	7.56	6.37	5.31	4.31
	Ensemble scheme 1	14.54	10.64	8.57	6.86	5.36
	Ensemble scheme 2	14.60	10.66	8.57	6.87	5.36
	Ensemble scheme 3	14.54	10.64	8.57	6.86	5.36
	Ensemble scheme 4	8.86	7.48	6.33	5.33	4.31
	Ensemble scheme 5	8.88	7.46	6.34	5.33	4.31
	Ensemble scheme 6	8.88	7.49	6.33	5.33	4.31

Table 7.9. Metric values computed for the prediction intervals delivered by the compared schemes for the period T_3 within the toy experiment 3.

Metric	Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Coverage probability	Linear regression	0.979	0.970	0.959	0.933	0.865
	Quantile regression	0.989	0.975	0.950	0.909	0.813
	Ensemble scheme 1	0.977	0.968	0.956	0.928	0.858
	Ensemble scheme 2	0.977	0.968	0.956	0.927	0.857
	Ensemble scheme 3	0.977	0.968	0.956	0.928	0.858
	Ensemble scheme 4	0.990	0.977	0.945	0.903	0.802
	Ensemble scheme 5	0.990	0.976	0.946	0.902	0.805
	Ensemble scheme 6	0.990	0.976	0.945	0.902	0.801
Average width	Linear regression	9.04	7.87	6.88	5.77	4.50
	Quantile regression	10.88	8.73	6.93	5.53	3.99
	Ensemble scheme 1	8.86	7.71	6.74	5.65	4.40
	Ensemble scheme 2	8.84	7.69	6.72	5.64	4.40
	Ensemble scheme 3	8.86	7.71	6.74	5.65	4.40
	Ensemble scheme 4	10.83	8.57	6.48	5.37	3.87
	Ensemble scheme 5	10.84	8.56	6.51	5.38	3.89
	Ensemble scheme 6	10.83	8.57	6.49	5.37	3.87
Average interval score	Linear regression	16.13	11.90	9.60	7.72	6.10
	Quantile regression	12.73	10.47	8.90	7.45	5.98
	Ensemble scheme 1	16.58	12.06	9.65	7.72	6.09
	Ensemble scheme 2	16.68	12.09	9.66	7.73	6.09
	Ensemble scheme 3	16.51	12.02	9.62	7.71	6.08
	Ensemble scheme 4	12.46	10.56	8.98	7.44	5.98
	Ensemble scheme 5	12.49	10.54	8.98	7.45	5.98
	Ensemble scheme 6	12.48	10.57	8.99	7.45	5.99

Table 7.10. Metric values computed for the prediction intervals delivered by the compared schemes for the period T_3 within the toy experiment 4. The results of the linear regression and quantile regression schemes are repeated with respect to Table 7.9 for consistency in the presentation.

Metric	Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Coverage probability	Linear regression	0.979	0.970	0.959	0.933	0.865
	Quantile regression	0.989	0.975	0.950	0.909	0.813
	Ensemble scheme 1	0.989	0.973	0.947	0.895	0.798
	Ensemble scheme 2	0.989	0.972	0.946	0.894	0.798
	Ensemble scheme 3	0.990	0.972	0.947	0.893	0.798
	Ensemble scheme 4	0.986	0.968	0.949	0.893	0.802
	Ensemble scheme 5	0.987	0.969	0.949	0.894	0.801
	Ensemble scheme 6	0.987	0.968	0.950	0.892	0.803
Average width	Linear regression	9.04	7.87	6.88	5.77	4.50
	Quantile regression	10.88	8.73	6.93	5.53	3.99
	Ensemble scheme 1	5.12	4.46	3.90	3.27	2.55
	Ensemble scheme 2	5.11	4.44	3.89	3.26	2.54
	Ensemble scheme 3	5.12	4.46	3.90	3.27	2.55
	Ensemble scheme 4	5.00	4.34	3.93	3.25	2.57
	Ensemble scheme 5	4.99	4.35	3.93	3.26	2.57
	Ensemble scheme 6	5.01	4.32	3.95	3.24	2.59
Average interval score	Linear regression	16.13	11.90	9.60	7.72	6.10
	Quantile regression	12.73	10.47	8.90	7.45	5.98
	Ensemble scheme 1	5.83	5.23	4.72	4.16	3.54
	Ensemble scheme 2	5.83	5.23	4.72	4.16	3.54
	Ensemble scheme 3	5.84	5.23	4.72	4.16	3.54
	Ensemble scheme 4	5.86	5.27	4.72	4.16	3.54
	Ensemble scheme 5	5.87	5.27	4.72	4.16	3.54
	Ensemble scheme 6	5.86	5.27	4.72	4.16	3.54

Two considerations applying to each of toy experiments 1–3 (see Tables 7.7–7.9) are the following: (a) ensemble schemes 1–3, as well as ensemble schemes 4–6, are equivalent to each

other on the examined normal data, and (b) each of the tested ensemble schemes is equivalent to its corresponding benchmark, i.e., ensemble schemes 1–3 and 4–6 perform as well as the linear regression and quantile regression schemes respectively. These two types of equivalence hold in terms of all three criteria examined. Consideration (a) also applies to the case of toy experiment 4 (see Table 7.10), while indicating that the three variants of the proposed methodology are equivalent in solving the examined problems. Moreover, consideration (b) can be viewed as an empirical proof that these problems are well-solved by the proposed methodology. The reason behind consideration (b) may become perceivable to some extent by comparing the original datasets (see Figure 7.2) with the datasets formed and used for training the incorporated quantile prediction models by the ensemble schemes (see Figure 7.4). Segments of the former datasets are used for training the benchmark schemes. In fact, the problems solved by each of the ensemble schemes and its corresponding benchmark seem to be of the same difficulty for toy experiments 1–3.

We have also tested the prediction schemes using shorter series (see e.g., the investigations of Section 7.5 and the large-sample real-world experiment conducted in Chapter 8 herein). In that particular case for which the provided historical information is much less, the prediction schemes differentiate with each other in terms of performance. Nevertheless, by repeating the procedure an essentially large number of times with varying seed in the simulation of the datasets, we may observe long-run equivalence between specific prediction schemes, depending on the attributes of the datasets. For related discussions, the interested reader is referred to Section 7.5.

One of the most important outcomes of the conducted toy model investigation is related to the satisfying coverage probabilities computed for all the ensemble schemes within toy experiment 1. Moreover, their good performance (equivalent to the performance of the Bayesian regression scheme and the two remaining benchmark schemes) in terms of average width of the prediction intervals and average interval score, observed within the same toy experiment, is important from an engineering point of view, as it points out that the proposed methodology does not lead to excessively precautionary design; see also the three criteria identified in Murphy (1993) for assessing the quality of predictions and the related discussions in Weijs et al. (2010), Ramos et al. (2010) and Chapter 3 herein.

The performances of all prediction schemes differ for the toy implementations made on toy datasets 2 and 3, both in terms of coverage probability and average width; therefore, this differentiation is also manifested in the average interval scores. In fact, while both benchmark schemes are theoretically expected to be equally well-performing within toy experiment 1, quantile regression is theoretically expected to be better than linear regression within toy experiment 2, because of its advantage in modelling heteroscedasticity. We herein show that we can obtain equally good probabilistic predictions on normal data, by integrating the same model within the proposed methodology as error model. The interpretation of this outcome is straightforward; the incorporation of flexible models, such as the herein adopted quantile regression model may be the key to obtain efficient probabilistic predictions in specific modelling situations, including the hydrological modelling ones (see the comments on the violation of the homoscedasticity assumption in hydrological modelling, e.g., in Schoups and Vrugt 2010; Montanari and Koutsoyiannis 2012; Evin et al. 2013, 2014).

In greater detail, the numerical results of Table 7.8 can be summarized as follows. When using quantile regression instead of linear regression (within the proposed methodology) the average interval score is largely improved by around 40%, 30%, 35%, 30% and 25% for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively. The respective relative improvements in terms of average width are around −3%, 6%, 13%, 22% and 22%, while the coverage probabilities computed for the predictions of the ensemble schemes 4–6 are essentially better than the coverage probabilities computed for the predictions of the ensemble schemes 1–3 for the 99%, 97.5%, 90% and 80% prediction intervals. The coverage probabilities of all ensemble schemes are comparable for the 95% prediction intervals. Typical performance differences observed within the Chapter between probabilistic prediction schemes that use perfect and imperfect error models (with respect to modelling heteroscedasticity) are presented in Figure 7.5.

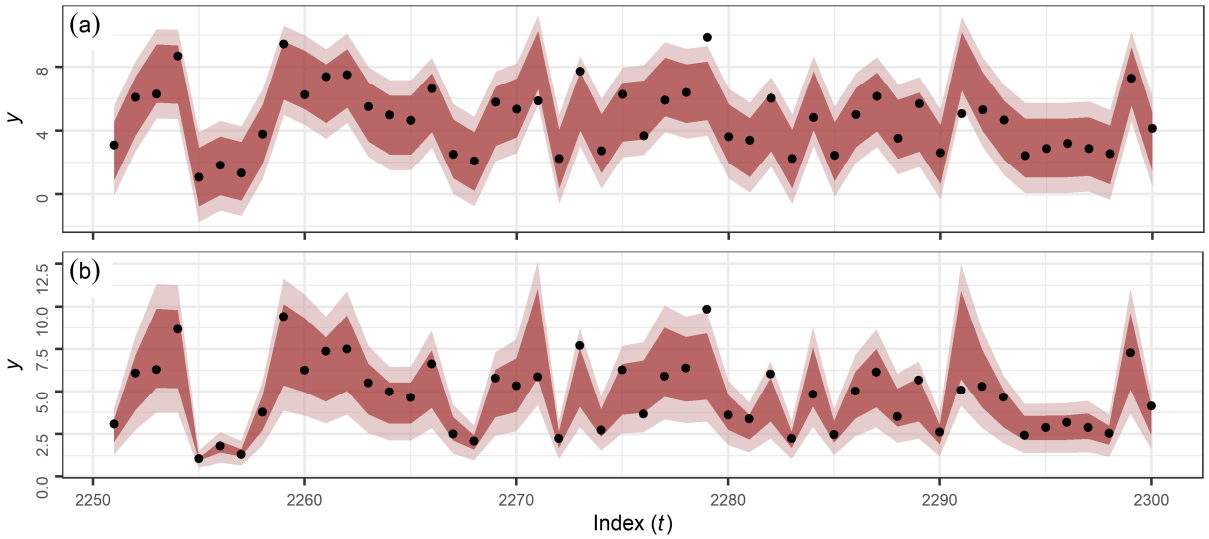


Figure 7.5. Toy solutions provided by ensemble schemes (a) 2 and (b) 5 within toy experiment 2 for a common 50-point sub-period of the period T_3. Black dots denote the targeted points, while light pink and dark pink ribbons denote the 95% and 80% prediction intervals respectively.

Moreover, within toy experiment 3 (see Table 7.9) we show that we can get probabilistic predictions with satisfactory coverage probabilities by using the proposed methodology, even when the incorporated toy hydrological model is imperfect (linear toy hydrological model for a quadratic relationship). This outcome is particularly important if we consider that process-based hydrological models are also imperfect. Specifically, we obtain perfect coverage probabilities by incorporating the quantile regression model within the proposed methodology, while the relative improvements in terms of average interval scores are around 25%, 13%, 8%, 4% and 2% for the 99%, 97.5%, 95%, 90% and 80% prediction intervals. The average widths, on the other hand, are better for ensemble schemes 4–6 only for the 95%, 90% and 80% prediction intervals, while they are much larger than those produced by ensemble schemes 1–3 for the 99% and 97.5% prediction intervals.

However, the average widths and average interval scores computed within toy experiment 3 for all ensemble schemes are found to be far from optimal, when contrasted to the results obtained within toy experiment 4 (see Table 7.10). The replacement of the imperfect (for toy dataset 3) toy hydrological model with a perfect one, has led to around 54%, 49%, 39%, 39% and 34% better average widths, and around 53%, 50%, 47%, 44% and 41% better average interval scores for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively, as the latter are provided by ensemble schemes 4–6. In fact, the quality of the obtained probabilistic solution largely depends on the adopted toy hydrological model. A toy illustration of typical performance differences between probabilistic prediction schemes that use perfect and imperfect toy hydrological models is made in Figure 7.6.

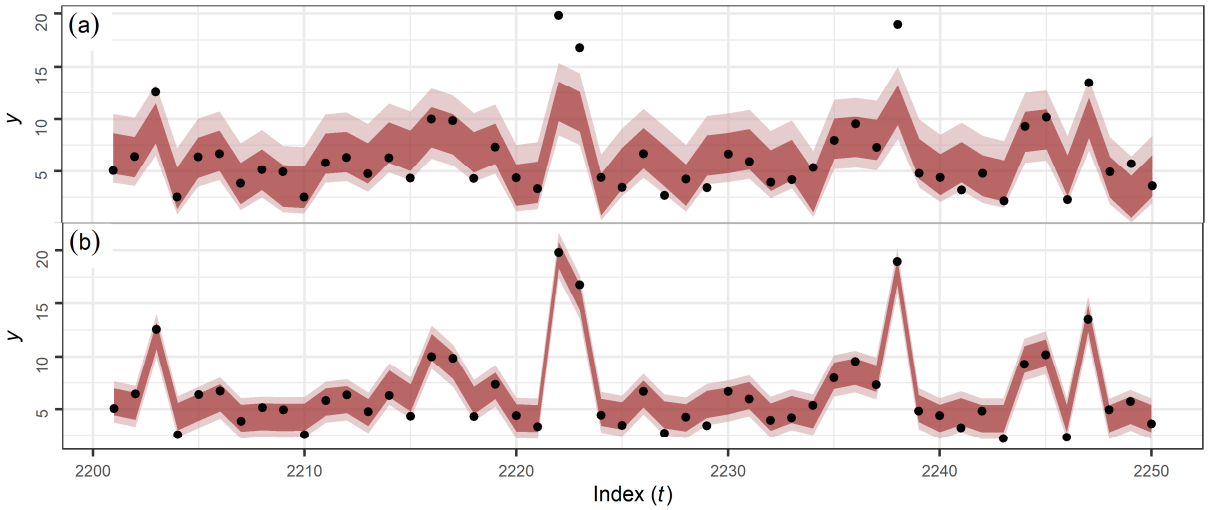


Figure 7.6. Toy solutions provided by ensemble scheme 5 within toy experiments (a) 3 and (b) 4 for a common 50-point sub-period of the period T_3. Black dots denote the targeted points, while light pink and dark pink ribbons denote the 95% and 80% prediction intervals respectively.

7.4.2 Improved robustness in hydrological post-processing

Here we present illustrative examples that can be used to gain further insight on how the proposed methodology works (aim 2 of the Chapter; see Section 7.1) and to answer the following research question (related to aim 3 of the Chapter): Why and when is it meaningful for someone to choose the proposed methodology over a basic two-stage post-processing methodology utilizing the same error model? In Figure 7.7, we present the relative improvements resulted within toy experiment 4 in terms of average interval score, when using the output of ensemble scheme 4, instead of each of the combined individual predictions, while in Figure 7.8 we present the respective relevant improvements provided by ensemble scheme 5. We observe that these relative improvements can be either positive or negative, while their mean is slightly higher than zero. Specifically, the average relative improvements computed for the histograms displayed in Figure 7.7 (Figure 7.8) are equal to 0.10%, 0.06%, 0.05%, 0.06% and 0.06% (0.20%, 0.10%, 0.13%, 0.14% and 0.12%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively, with the results being analogous for the remaining ensemble schemes.

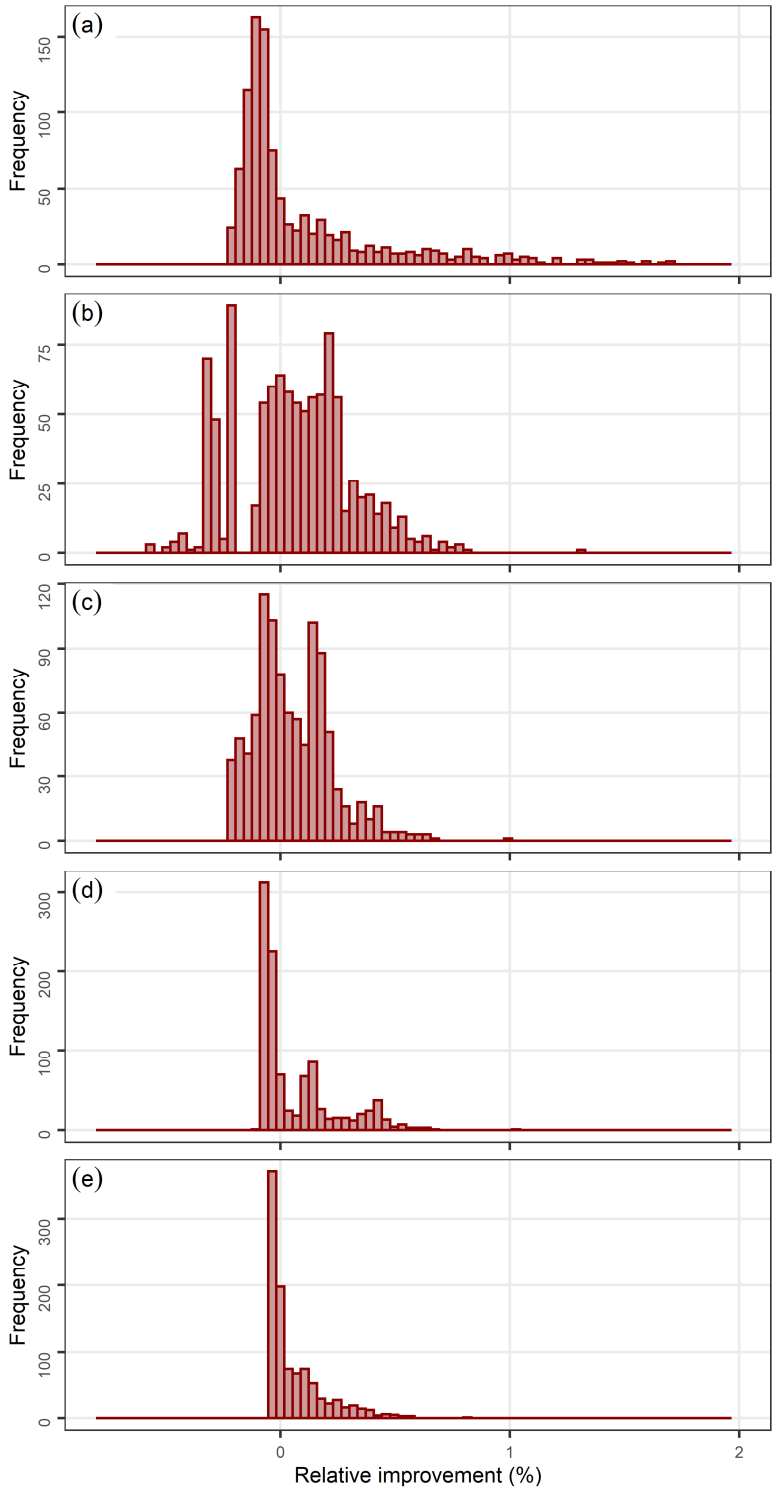


Figure 7.7. Relative improvements in terms of average interval score when using the output of ensemble scheme 4, i.e., the average of 1 000 probabilistic predictions, instead of each of the combined individual predictions. The relative improvements are computed for the (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80% prediction intervals obtained for the period T_3 within toy experiment 4. The horizontal axis has been truncated at −0.8% and 2%. Each histogram summarizes 1 000 values.

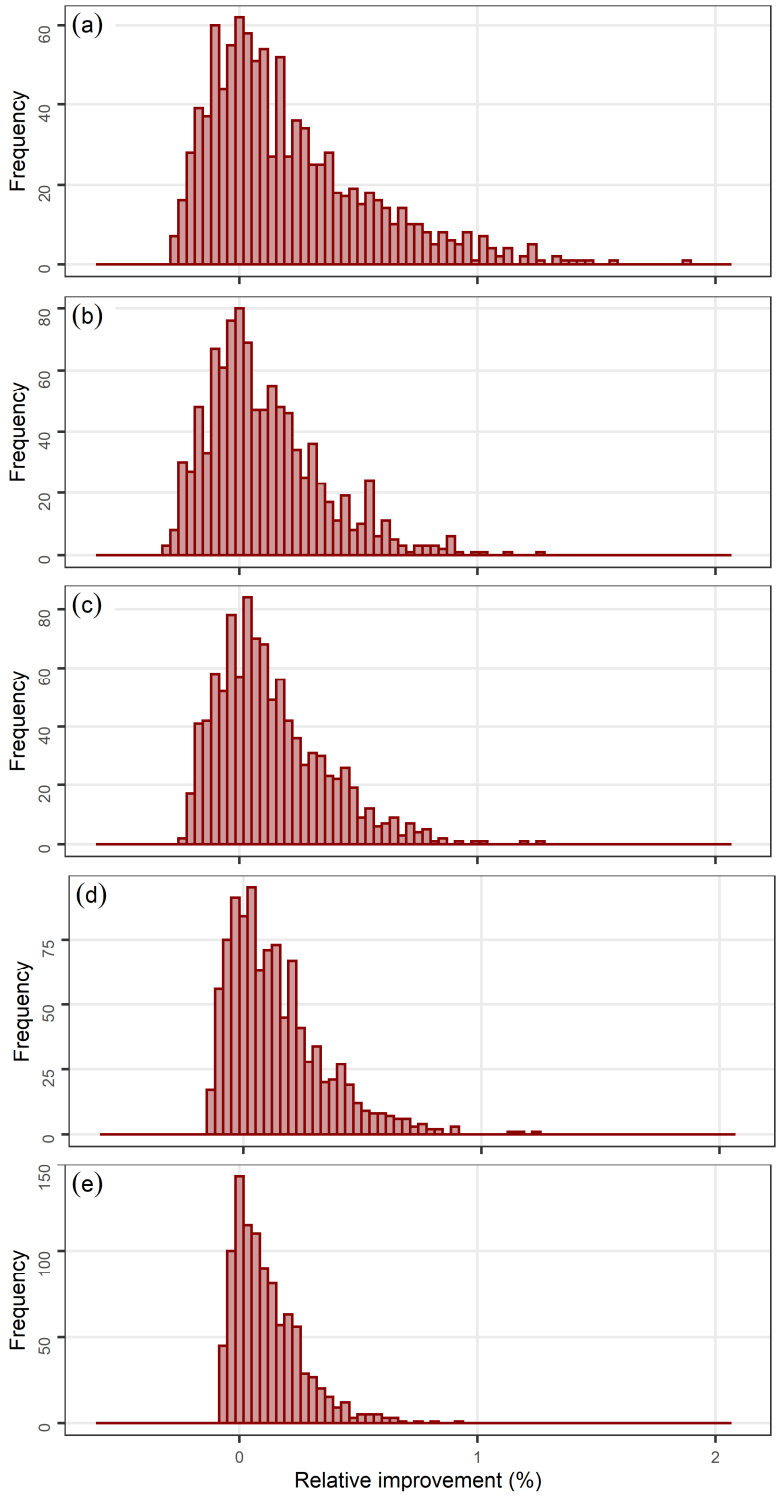


Figure 7.8. Relative improvements in terms of average interval score when using the output of ensemble scheme 5, i.e., the average of 1 000 probabilistic predictions, instead of each of the combined individual predictions. The relative improvements are computed for the (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80% prediction intervals obtained for the period T_3 within toy experiment 4. The horizontal axis has been truncated at −0.6% and 2%. Each histogram summarizes 1 000 values.

These average relative improvements computed for ensemble schemes 1 and 4 could be viewed as a direct comparison in terms of robustness in performance between the proposed methodology and a basic two-stage post-processing methodology (generating a single point

hydrological prediction and, therefore, using a single set of hydrological model's parameters), the latter using the linear regression and quantile regression models respectively as error models. In fact, although many of the individual probabilistic predictions (obtained using different sister model realizations and, therefore, multiple sets of hydrological model's parameters) score better than the finally delivered one in the period T_3, we cannot know in advance which sister model realizations can be used for obtaining these better results and, therefore, should be preferred over the remaining ones within a basic post-processing methodology. By averaging numerous probabilistic predictions (obtained using the same number of different sister model realizations) we simply reduce the risk of delivering a probabilistic prediction of bad quality for the period T_3.

An important remark to be highlighted is that this risk can be high or low depending on the problem. In the toy problems examined herein the risk of delivering a probabilistic prediction of bad quality for the period T_3 (manifested in the magnitude of the relative improvements presented e.g., in Figures 7.7 and 7.8) is much lower than the respective risk that was found to be present in the rainfall-runoff problems examined in Chapter 8 herein. In this latter Chapter the computed relative improvements in terms of average interval score when using the output of the proposed methodology, instead of using each of the individual predictions combined for obtaining this output, range from about −330% to about 90%. (Negative relative improvements are computed for predictions that perform better that the prediction combination). Therefore, while it would not be that cost-efficient to use the proposed methodology for problems, such as the simple ones solved (for illustrative purposes) herein, it is cost-efficient from a risk management perspective to use this methodology (instead of a basic post-processing methodology) for probabilistic hydrological modelling applications.

Finally, for all ensemble schemes and all prediction intervals, the output of the proposed methodology is herein found to score slightly better than the average of the scores computed for each of the combined individual predictions in terms of average interval score. This latter information stands as an empirical proof that this methodology harnesses the wisdom of the crowd for the examined problem and in terms of average interval score. This useful property of the proposed methodology is further investigated by using rainfall-runoff datasets in Chapter 8 of this thesis.

7.5 Additional investigations and derived interpretations

7.5.1 Large-scale variant of the basic toy experiment using shorter toy datasets

Toy experiment 1 is particularly important because there exists an analytical solution to it, thereby allowing us to extensively explore under which conditions the data-driven solutions provided by the remaining schemes are adequate. This analytical solution is provided by the herein implemented Bayesian regression scheme. To further facilitate the interpretation of the proposed methodology (by answering questions related to aims 2 and 4 of the Chapter; see Section 7.1), we here conduct the "type 1 additional investigations" by repeating toy experiment 1 using shorter toy datasets. We run 500 repetitions, each time using a different toy dataset comprising 300 pairs of (x_t, y_t) values. These toy datasets result by following the same simulation procedure that was previously adopted for obtaining toy dataset 1 (see Table 7.4). Multiple runs are important in this case, because randomness can largely affect the results when relying on few data.

For each of the resulted toy datasets, we i) define the periods T_1 = {1, …, 100}, T_2 = {101, …, 200} and T_3 = {201, …, 300}, ii) run the three benchmark schemes according to Section 3.2.3, iii) run the six ensemble schemes according to Section 7.3.4 by adopting the linear regression model as toy hydrological model, and iv) compute the metric values for each delivered prediction according to Section 7.3.6. Finally, we compute the average metric values for each combination of prediction scheme and prediction interval. The coverage probability, average width and average interval score values are presented in Figures 7.9–7.11 respectively, while the average metric values are presented in Table 7.11. Note that the here examined 500 toy datasets are all in the same scale; therefore, the average metric values are highly informative.

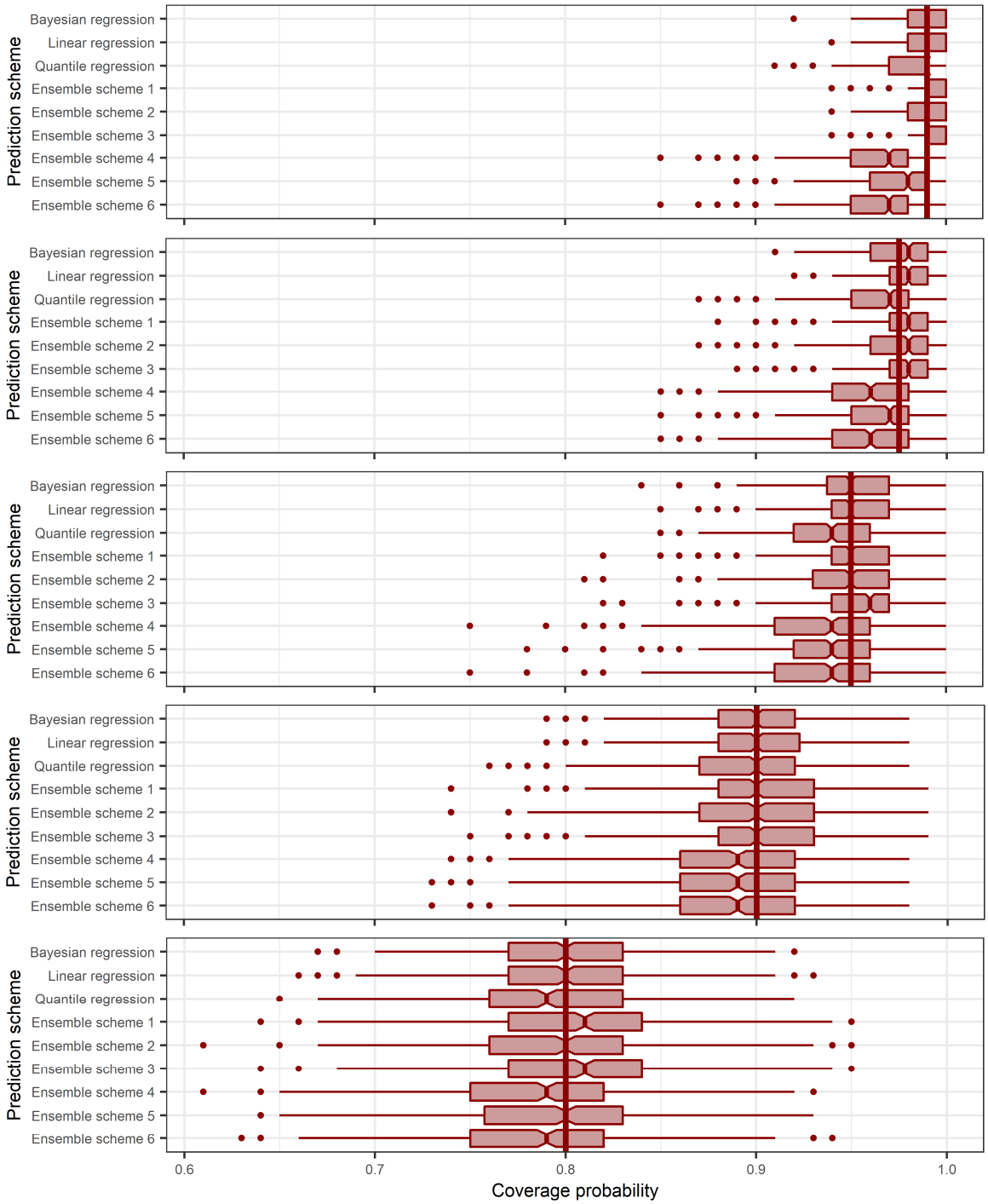


Figure 7.9. Coverage probabilities computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 within the type 1 additional investigations. Each boxplot summarizes 500 values. The optimal values are denoted with red thick vertical lines.

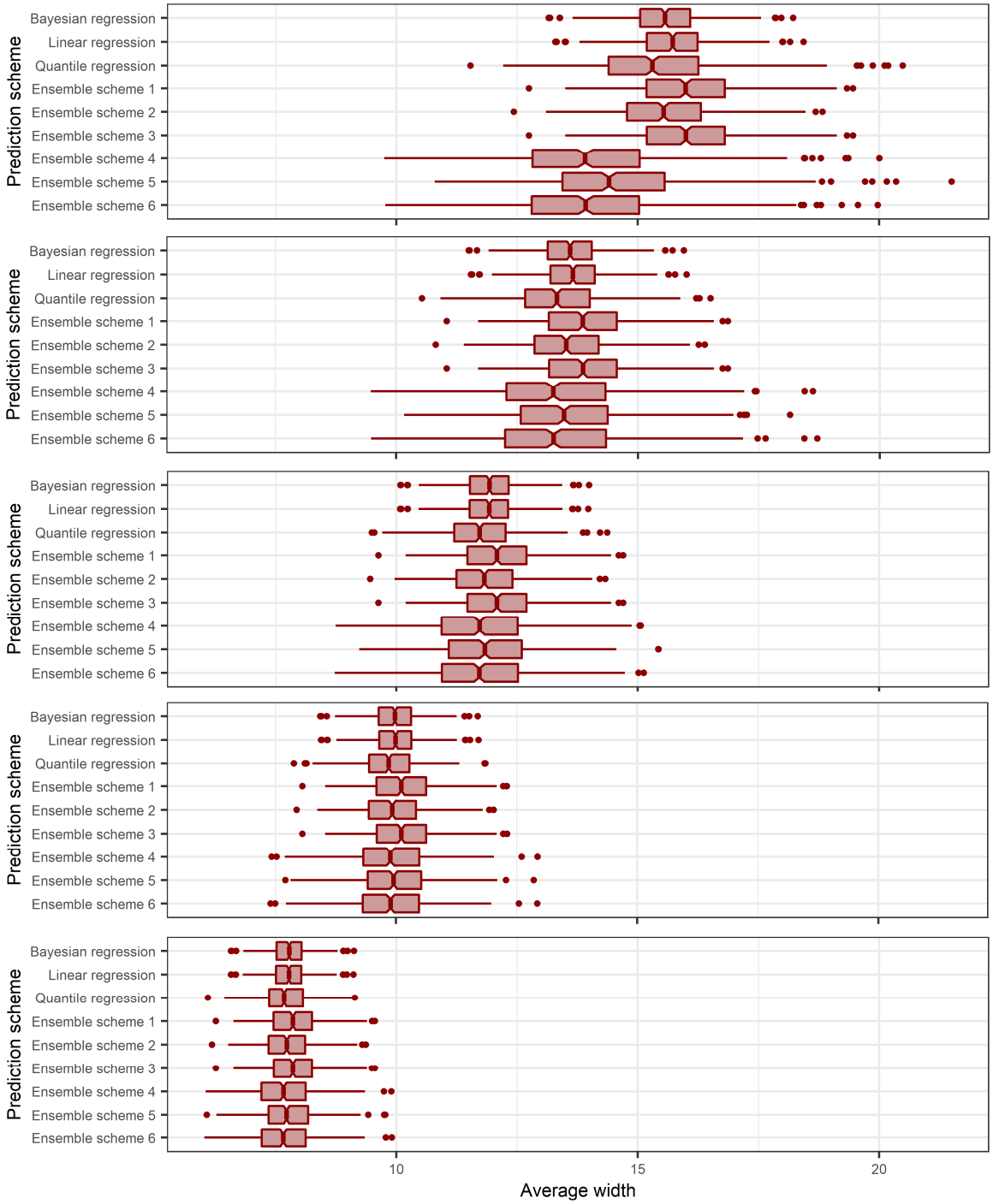


Figure 7.10. Average widths computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 within the type 1 additional investigations. Each boxplot summarizes 500 values.

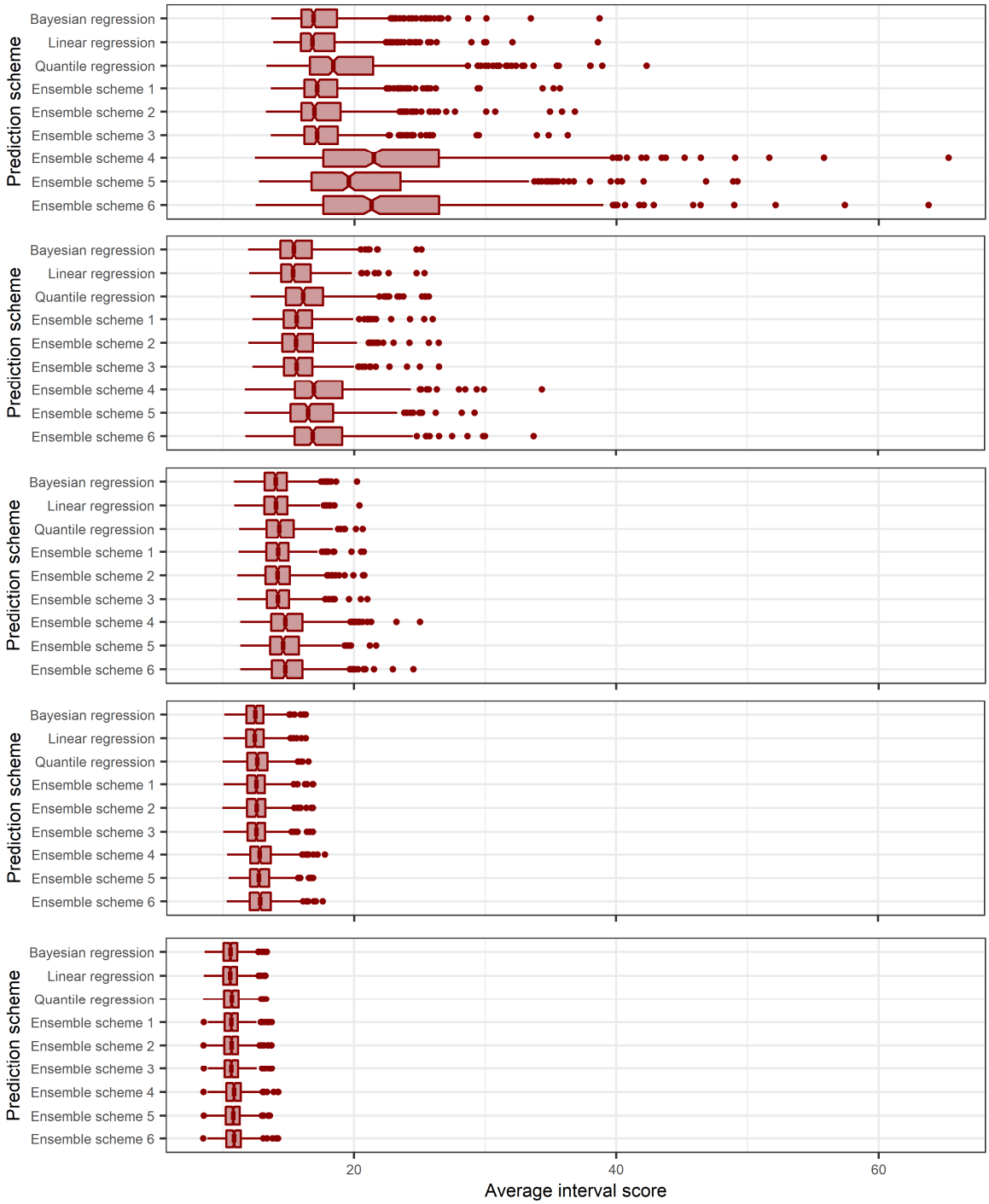


Figure 7.11. Average interval scores computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 within the type 1 additional investigations. Each boxplot summarizes 500 values.

Table 7.11. Average metric values computed for the prediction intervals delivered by the compared schemes for the period T_3 within the type 1 additional investigations. Each presented value summarizes 500 metric values.

Metric	Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Coverage probability	Bayesian regression	0.989	0.975	0.949	0.899	0.801
	Linear regression	0.990	0.976	0.950	0.900	0.800
	Quantile regression	0.981	0.966	0.941	0.892	0.793
	Ensemble scheme 1	0.991	0.976	0.952	0.902	0.803
	Ensemble scheme 2	0.989	0.973	0.947	0.896	0.796
	Ensemble scheme 3	0.991	0.976	0.952	0.902	0.803
	Ensemble scheme 4	0.964	0.957	0.933	0.886	0.785
	Ensemble scheme 5	0.973	0.962	0.939	0.891	0.792
	Ensemble scheme 6	0.964	0.957	0.933	0.885	0.786
Average width	Bayesian regression	15.57	13.60	11.93	9.98	7.78
	Linear regression	15.72	13.65	11.92	9.99	7.77
	Quantile regression	15.43	13.36	11.75	9.86	7.70
	Ensemble scheme 1	16.00	13.87	12.09	10.12	7.86
	Ensemble scheme 2	15.54	13.53	11.83	9.93	7.73
	Ensemble scheme 3	16.00	13.87	12.09	10.12	7.86
	Ensemble scheme 4	13.98	13.32	11.71	9.90	7.67
	Ensemble scheme 5	14.53	13.46	11.86	9.97	7.75
	Ensemble scheme 6	13.99	13.32	11.71	9.90	7.67
Average interval score	Bayesian regression	17.72	15.72	14.18	12.51	10.62
	Linear regression	17.61	15.69	14.16	12.50	10.62
	Quantile regression	19.66	16.46	14.51	12.66	10.70
	Ensemble scheme 1	17.88	15.91	14.31	12.60	10.69
	Ensemble scheme 2	17.89	15.91	14.31	12.60	10.69
	Ensemble scheme 3	17.88	15.91	14.31	12.60	10.69
	Ensemble scheme 4	22.85	17.49	15.07	12.93	10.86
	Ensemble scheme 5	20.97	16.97	14.83	12.83	10.80
	Ensemble scheme 6	22.76	17.47	15.06	12.92	10.86

The main observations extracted from these investigations can be summarized as follows: (a) The linear regression scheme is equivalent to the Bayesian regression scheme in the long run, (b) ensemble schemes 1–3 perform almost as well as the two best-performing benchmarks (since their error modelling procedures are benefited from proper assumptions, i.e., prior knowledge on the system to be modelled), (c) ensemble schemes 4–6 are the worst-performing, (d) the quantile regression scheme exhibits a moderate performance, (e) ensemble schemes 1–3 are almost equivalent to each other, and (f) ensemble scheme 5 performs better than ensemble schemes 4 and 6. By comparing these observations with those extracted from toy experiment 1 (see Section 7.4.1), we understand that the quantile regression model needs to be "fed" with more data to reach its best performance, which in the case of the here examined data type is as good as the performance of the linear regression model. Note that this "data consuming" consideration stems from the statistical learning nature of the modelling process and, therefore, it applies to the linear regression model as well, yet to a smaller extent. It could also be viewed as a limitation of two-stage hydrological post-processing in general (see Section 8.1).

7.5.2 Large-scale toy regression experiment with non-informative predictors

We repeat the type 1 additional investigations by using different toy datasets of the same number and length. The new datasets result from a simulating model that implies no dependence of $\underline{\boldsymbol{y}}$ on $\underline{\boldsymbol{x}}$, specifically the following: $\underline{x}_t \sim N(\mu = 0, \sigma^2 = 1^2)$ and $\underline{y}_t \sim N(\mu = 0, \sigma^2 = 1^2)$. These investigations are hereafter referred to as "type 2 additional investigations". The analytical solution to the examined toy problem is provided by the "Bayesian non-regression" benchmark. This scheme assumes that $\underline{y}_t \sim N(\mu, \sigma^2)$, prior independence of μ and σ, and a uniform prior distribution on (μ, logσ). Then, the posterior distribution is a Student-t distribution with location $\bar{y} = (1/n) \sum_{t=1}^{n} y_t$, scale $(1 + 1/n)^{0.5} ((1/(n-1)) \sum_{t=1}^{n} (y_t - \bar{y})^2)^{0.5}$, and $n - 1$ degrees of freedom, where n is the number

of data points included in the fitting sample (Gelman et al. 2004). In our case, the fitting sample is consisted of the first 200 data points of each simulated series, i.e., $n = 200$.

The main observations extracted from the type 2 additional investigations can be summarized as follows (see Table 7.12): (a) The Bayesian non-regression, Bayesian regression and linear regression benchmarks are equivalent in the long run, (b) ensemble schemes 1 and 2 perform almost as well as the three best-performing benchmarks, (c) ensemble schemes 3 and 6 are the worst-performing (mostly because of outliers), (d) ensemble scheme 5 and the quantile regression benchmark exhibit a moderate performance, and (e) ensemble scheme 4 exhibits better performance than ensemble scheme 6 and worse than ensemble scheme 5.

Table 7.12. Average metric values computed for the prediction intervals delivered by the compared schemes for the period T_3 within the type 2 additional investigations. Each presented value summarizes 500 metric values.

Metric	Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Coverage probability	Bayesian non-regression	0.990	0.976	0.950	0.901	0.801
	Bayesian regression	0.989	0.975	0.949	0.899	0.801
	Linear regression	0.990	0.976	0.950	0.900	0.800
	Quantile regression	0.981	0.966	0.941	0.892	0.793
	Ensemble scheme 1	0.991	0.976	0.952	0.902	0.803
	Ensemble scheme 2	0.988	0.972	0.947	0.895	0.795
	Ensemble scheme 3	0.994	0.982	0.962	0.920	0.832
	Ensemble scheme 4	0.964	0.957	0.933	0.886	0.785
	Ensemble scheme 5	0.980	0.964	0.943	0.894	0.794
	Ensemble scheme 6	0.893	0.895	0.877	0.835	0.736
Average width	Bayesian non-regression	5.23	4.54	3.96	3.32	2.58
	Bayesian regression	5.19	4.53	3.98	3.33	2.59
	Linear regression	5.24	4.55	3.97	3.33	2.59
	Quantile regression	5.14	4.45	3.92	3.29	2.57
	Ensemble scheme 1	5.33	4.62	4.03	3.37	2.62
	Ensemble scheme 2	5.14	4.48	3.91	3.28	2.56
	Ensemble scheme 3	8.72	7.56	6.59	5.51	4.28
	Ensemble scheme 4	4.66	4.44	3.90	3.30	2.56
	Ensemble scheme 5	5.02	4.37	3.92	3.30	2.57
	Ensemble scheme 6	4.95	5.01	3.98	3.51	2.61
Average interval score	Bayesian non-regression	5.85	5.22	4.71	4.16	3.53
	Bayesian regression	5.91	5.24	4.73	4.17	3.54
	Linear regression	5.87	5.23	4.72	4.17	3.54
	Quantile regression	6.55	5.49	4.84	4.22	3.57
	Ensemble scheme 1	5.96	5.30	4.77	4.20	3.56
	Ensemble scheme 2	5.93	5.28	4.75	4.18	3.55
	Ensemble scheme 3	9.15	8.06	7.16	6.19	5.10
	Ensemble scheme 4	7.62	5.83	5.02	4.31	3.62
	Ensemble scheme 5	6.43	5.48	4.84	4.21	3.56
	Ensemble scheme 6	188.33	48.85	25.36	12.38	6.66

7.6 Summary, discussion and conclusions

We have focused on the problem of probabilistically predicting hydrological variables, such as river discharge variables, by incorporating hydrological point prediction models, mainly falling into the category of deterministic process-based models, within stochastic modelling approaches. We have presented three novel variants of the blueprint methodology by Montanari and Koutsoyiannis (2012), also relying on the seminal work by Lichtendahl et al. (2013). In summary, the proposed methodology generates a large number of point predictions by utilizing a single hydrological model, yet with different parameter values. By solving a typical regression problem, these "sister predictions" are converted into auxiliary probabilistic predictions (consisted of quantile predictions), which are finally combined via simple quantile averaging. To the best of our knowledge, this is the first quantile averaging hydrological post-processing methodology that

creates and exploits different information sets using a single model with different parameter values.

It is relevant to highlight that both the original blueprint and the herein introduced methodology fall into the family of two-stage probabilistic hydrological post-processing methodologies. Being mostly characterized by algorithmic-modelling-culture features (defined and analysed e.g., in Breiman 2001b and Shmueli 2010), and concomitant advantages and disadvantages, these methodologies aspire to achieve optimality in predictive performance in a fundamentally different way with respect to Bayesian joint inference methodologies for hydrological post-processing. Related information is provided in Section 8.1 (see also Evin et al. 2014). In light of this information, the present Chapter has been mostly devoted to finding modelling 'tricks' and concepts for maximizing predictive performance in two-stage hydrological post-processing by building on the original blueprint by Montanari and Koutsoyiannis (2012). An additional advantage offered by the latter with respect to other two-stage hydrological post-processing methodologies is its larger flexibility by perception, in the sense that it allows the formation and testing of various alternative configurations. This advantage is particularly important from a predictive modelling perspective.

A key improvement achieved herein compared to the original work by Montanari and Koutsoyiannis (2012), and the variants by Sikorska et al. (2015) and Quilty et al. (2019) in terms of flexibility in modelling is the use of statistical learning regression models that can directly provide predictive quantiles of the response variable, while they are also appropriate for modelling heteroscedasticity, such as the six machine learning algorithms examined in Chapter 9 of this thesis. These are quantile regression, generalized random forests for quantile regression, generalized random forests for quantile regression emulating quantile regression forests, gradient boosting machine, model-based boosting with linear models as base learners and quantile regression neural networks. Allowing the exploitation of the possibilities provided by this model category should, in fact, be regarded as a primary strength of the proposed methodology from a predictive modelling perspective.

Herein, we have demonstrated the usefulness of the proposed methodology and how our understanding of the system to be modelled can guide us to achieve better predictive modelling when using this methodology by conducting a toy model investigation. Within this investigation we have focused on the unsuitability of the homoscedasticity assumption, when the latter is made in the modelling of the hydrological model's error, and on how the selection of an appropriate regression model for this task results in improved probabilistic predictions. We have also demonstrated the significance of using a better hydrological model for delivering probabilistic predictions that are simultaneously reliable and as sharp as possible. Finally, we have used the obtained toy results to show how the proposed methodology increases its robustness in performance by averaging many quantile predictions.

In spite of focusing on the introduced methodology, some of the obtained results can be used for gaining insight in general on how two-stage hydrological post-processing methodologies work and under which conditions their performance is maximized. The presented toy examples, demonstrating the key roles of both the statistical learning regression model and the hydrological model within a hydrological post-processing methodology, go beyond of some few exemplary (yet basic) toy tests that have already been made for the interpretation of methodologies for the quantification of the predictive hydrological uncertainty. Such tests mostly assume homoscedasticity and a perfect toy hydrological model, while here we are also inspired by recent simulation experiments that do not rely on these assumptions (see e.g., Vrugt et al. 2005; Renard et al. 2010; Evin et al. 2014).

The present work is accompanied by the work presented in Chapter 8. The latter work is devoted to validating the herein introduced methodology and its key properties using a large amount of real-world data. Two simultaneously attractive and useful properties of this methodology that are extensively tested therein are its larger robustness in performance compared to the combined individual predictors and, by extension, compared to basic two-stage

post-processing methodologies (which produce a single probabilistic prediction and, therefore, no prediction combination is made in their case), and its ability to "harness the wisdom of the crowd". The latter is defined in Lichtendahl et al. (2013, Section 5) as the property of some prediction combinations to score no worse –usually better– than the average score of the combined individual predictions. In fact, the larger the number of the combined quantile predictions (equal to the number of the generated sister predictions), the more robust the ensemble predictor and the more harnessed the wisdom of the crowd.

The proposed methodology is characterized by some additional strengths that are also particularly important from a predictive modelling point of view. First, it is computationally convenient in the sense that it can be easily expressed in algorithmic form (see Section 7.2.1) and programmed using open source routines (see Section 2.9.4). Second, it offers certain modelling options that could be exploited to maximize predictive performance, as detailed in Section 8.6. For instance, variants 1 and 2 allow the exploitation by the error model of a large number of different information sets, instead of a single one (exploited by variant 3), thereby facilitating the enlargement of the sample space of the hydrological model's observed errors. This enlargement could be particularly important for modelling these errors using methods which do not extrapolate beyond the values of the training dataset, such as the quantile regression forests model (see the related theoretical information summarized by Tyralis et al. 2019b). Lastly, it allows the exploitation of the total amount of available information, in the sense that each sister prediction is herein converted into a probabilistic prediction (consisted of several quantile predictions) instead of a single simulation (randomly extracted from its predictive PDF; see the utilization of the meta-Gaussian bivariate distribution model in Montanari and Koutsoyiannis 2012; see also Kelly and Krzysztofowicz 1997).

Some limitations of the proposed methodology should also be considered. These include limitations implied by its two-stage nature (see Section 8.1), such as its shortcoming in terms of interpretability in modelling (especially in terms of producing interpretable parameter estimates) and its significant data length requirements (revealed e.g., in Section 7.5). Although this latter limitation should be acknowledged herein and perhaps taken into consideration in real-world applications, (daily) datasets are usually essentially large. Moreover, in Chapter 8 herein it is empirically proven that, in practice, even when the available historical information is little, the proposed methodology is well-performing when implemented using the quantile regression model as error model.

Furthermore, the computational requirements of the proposed methodology are (at the moment) high when (i) computationally intensive procedures (e.g., Markov Chain Monte Carlo simulation sampling) are preferred for calibrating the hydrological model, and/or (ii) the error model is trained as implied by variant 1 or variant 2, unless the application is restricted to considering a small number of sister predictions. Note that a computationally convenient and simple algorithm is not necessarily computationally fast. It is also important to clarify that the above-outlined limitation holds only for applications to hundreds of catchments and timescales finer than the monthly one, and for implementations through regular personal computers. It does not hold for applications to a small number of catchments, and applications at the monthly and annual timescales. Still, large-scale applications at the daily timescale can be supported by variant 3, when this variant is implemented by using computationally fast algorithms for calibrating the hydrological model (see e.g., the calibration scheme tested in Section 8.4).

In addition to the above-discussed considerations and in contrast to several statistical methodologies for probabilistic prediction, such as the Bayesian methodology by Tyralis and Koutsoyiannis (2014), a well-known drawback of flexible statistical learning models for quantile prediction is their inappropriateness for modelling long-range dependence (see also Cox et al. 2018). Modelling this dependence when solving prediction problems is a frequently met concern in applied hydrology (see e.g., the large-scale investigations in Chapters 3–5 herein; see also the comparative case study in Chapter 6). Nonetheless, empirical evidence (see e.g., Evin et al. 2014) suggests that the AR(1) assumption (in some sense allowed by the proposed methodology by using as a predictor variable in regression the hydrological model's prediction at time $t-1$) is

adequate when modelling hydrological models' errors. In general, by including more than one predictor variables (e.g., the hydrological model's predictions at times t, $t-1$, $t-2$, etc.) in the regression settings we can increase the amount of the available information exploited and improve predictive performance, as it is empirically proven for rainfall-runoff modelling problems in Chapter 9 of this thesis.

Overall, the main trade-off to be considered when selecting between the proposed methodology and basic two-stage post-processing methodologies (utilizing the same error model) is the one between (a) the increased robustness in performance and the ability to harness the wisdom of the crowd, both offered by the former methodology, and (b) the significantly less computational requirements of a basic post-processing methodology. We believe that from a risk management standpoint this trade-off is worthy, as the large-sample experiment of Chapter 8 suggests.

8. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale

Predictive hydrological uncertainty can be quantified by using ensemble methods. If properly formulated, these methods can offer improved predictive performance by combining multiple predictions. In this Chapter, we use 50-year-long monthly time series observed in 270 catchments in the United States to explore the performances provided by the ensemble learning post-processing methodology introduced in Chapter 7. This methodology allows the utilization of flexible quantile regression models for exploiting information about the hydrological model's error. Its key differences with respect to basic two-stage hydrological post-processing methodologies using the same type of regression models are that (a) instead of a single point hydrological prediction it generates a large number of "sister predictions" (yet using a single hydrological model), and that (b) it relies on the concept of combining probabilistic predictions via simple quantile averaging. A major hydrological modelling challenge is obtaining probabilistic predictions that are simultaneously reliable and associated to prediction bands that are as narrow as possible; therefore, we assess both these desired properties of the predictions by computing their coverage probabilities, average widths and average interval scores. The results confirm the usefulness of the proposed methodology and its larger robustness with respect to basic two-stage post-processing methodologies. Finally, this methodology is empirically proven to harness the "wisdom of the crowd" in terms of average interval score, i.e., the average of the individual predictions combined by this methodology scores no worse –usually better– than the average of the scores of the individual predictions.

8.1 Introduction

Uncertainty is a subject of ongoing discussions in hydrology (see e.g., Beven 1993; Vogel 1999; Beven 2000, 2001; Beven and Feer 2001; Krzysztofowicz 2001b; Pappenberger and Beven 2006; Koutsoyiannis and Montanari 2007; Montanari 2007; Koutsoyiannis et al. 2009; Koutsoyiannis 2010; Kuczera et al. 2010; Ramos et al. 2010; Weijs et al. 2010; Koutsoyiannis 2011; Juston et al. 2012; Ramos et al. 2013; Nearing at al. 2016). Hydrological modelling uncertainty is traditionally recognised within the model calibration and validation phases (Montanari 2011) in the context of the widely accepted evaluation framework proposed by Klemeš (1986). Within this framework "uncertainty treatment" serves the verification of hydrological model's reliability (Montanari 2011). The large number of relevant studies and their high significance are summarised, for instance, in the review papers by Efstratiadis and Koutsoyiannis (2010), and Pechlivanidis et al. (2011).

As discussed in Koutsoyiannis (2010), an appropriate modelling approach for any uncertain hydrological system should necessarily include quantification of its uncertainty within a stochastic framework. Uncertainty is naturally quantified using the probability theory, i.e., in terms of probability distribution function (PDF; Todini 2007; see also Todini 2004, 2008). Todini (2007; quoting Krzysztofowicz 1999) emphasizes the fact that in engineering applications the targeted uncertainty quantification should be no other than the quantification of the predictive uncertainty, i.e., the total uncertainty of the predictand. Along with this strong engineering-oriented interest of hydrologists (which might be underestimated in some cases but is of vital significance for hydrology, as for any applied science; Shmueli 2010), understanding of predictive performance and uncertainty in hydrological modelling is undoubtedly a major science-oriented target (see e.g., Clark et al. 2008; Renard et al. 2010; Montanari 2011; Pechlivanidis et al. 2011; Renard et al. 2011; Beven 2012; Montanari and Koutsoyiannis 2012; Clark et al. 2015; Farmer and Vogel 2016; Széles et al. 2018; Khatami et al. 2019).

The preference for process-based (including conceptual) hydrological models (over the data-driven ones; Toth et al. 1999), along with both the practical relevance of predictive uncertainty quantification in hydrology and the attentiveness of hydrologists towards increasing understanding in (probabilistic) hydrological modelling, has led to the development of a wide

range of methodologies for the integration of process-based and statistical models. This range includes (but is not limited to) various types of methodologies that statistically post-process the output of process-based models (hereafter referred to as "post-processing" methodologies). Considering information from deterministic models within uncertainty assessment frameworks (instead of exclusively using statistical methods) is a state-of-the-art methodological approach that is also adopted in contiguous fields (see e.g., Tyralis and Koutsoyiannis 2017). This approach holds a prominent position in the field of probabilistic hydrological modelling, in contrast to purely statistical probabilistic methodologies, which are rarely preferred; therefore, the below-provided outline exclusively focuses on it.

Perhaps the most frequently exploited methodology for predictive uncertainty quantification in hydrological modelling is the Generalized Likelihood Uncertainty Estimation (GLUE; Beven and Binley 2014). This approach has been proposed by Beven and Binley (1992), and is based on the concept of equifinality (see e.g., Beven 2006; Khatami 2019). It has been discussed, for example, in Montanari (2005), Mantovan and Todini (2006), Stedinger et al. (2008), Vrugt et al. (2009b), and Sadegh and Vrugt (2013); see also the related comments in Todini (2007).

Another predictive uncertainty quantification methodology that has received attention both by researchers and practitioners is the Bayesian Forecasting System (BFS). The BFS has been introduced by Krzysztofowicz (1999, 2001b, 2002), Krzysztofowicz and Kelly (2000), and Krzysztofowicz and Herr (2001) for producing probabilistic river stage forecasts. It consists of three discrete components, namely the Precipitation Uncertainty Processor (PUB), the Hydrologic Uncertainty Processor (HUP) and the INTegrator (INT). Information about these components can be found in Kelly and Krzysztofowicz (2000), Krzysztofowicz and Kelly (2000), and Krzysztofowicz (2001a) respectively. This Bayesian methodology is conceived for real-time forecasting and relies on the assumption that uncertainty is mainly introduced by rainfall forecast errors.

There are also Bayesian post-processing methodologies that explicitly consider the contribution of input and output data uncertainty (which also affects the quantification of parameter uncertainty; see Di Baldassarre and Montanari (2009), McMillan et al. (2010), Di Baldassarre et al. (2012), McMillan et al. (2012), Kauffeldt et al. (2013), Montanari and Di Baldassarre (2013), Tomkins (2014) and Coxon et al. (2015) for information on rainfall-runoff data errors). Perhaps the most characteristic example of such a methodology is the Bayesian Total Error Analysis (BATEA) framework by Kavetski et al. (2002; see also Kavetski et al. 2006a, Kuczera et al. 2006), implemented, for instance, in Thyer et al. (2009) and Renard et al. (2010, 2011). This Bayesian framework facilitates the joint modelling of parameter uncertainty, data uncertainties, and model error, i.e., of all sources of uncertainty that are often assumed to collectively compose the predictive uncertainty. Other Bayesian post-processing methodologies introduced for parameter and predictive uncertainty quantification are described by Kuczera (1983), Schoups and Vrugt (2010), Evin et al. (2013; see also Evin et al. 2014), Hernández-López and Francés (2017) and Romero-Cuellar et al. (2019); see also the literature review in Hernández-López and Francés (2017).

Non-Bayesian post-processing methodologies that in their majority focus on the modelling of a single error term conditional on hydrological point predictions and historical information are also available in the hydrological modelling literature (see e.g., Montanari and Brath 2004; Montanari and Grossi 2008; Solomatine and Shrestha 2009; López López et al. 2014; Dogulu et al. 2015; Bourgin et al. 2015; Farmer and Vogel 2016; Wani et al. 2017; Bock et al. 2018). Adopting the terminology by Evin et al. (2014), such methodologies are hereafter referred to as "two-stage" post-processing methodologies, as their hydrological and error models are estimated in two subsequent stages. It is relevant to note at this point that Bayesian and two-stage post-processing methodologies are rather not directly comparable, since they are characterized by different statistical-modelling-culture traits and distinguishing features, which in their turn lead to different advantages and disadvantages (see Tables 8.1–8.3). For extensive discussions on the statistical modelling cultures, the reader is referred to Breiman (2001b) and Shmueli (2010).

Table 8.1. Advantages and disadvantages of Bayesian hydrological post-processing methodologies (see also Evin et al. 2014). These post-processing methodologies jointly infer (within a Bayesian framework) the parameters of the hydrological and error models by using the entire historical dataset.

Advantages	○ If their assumptions are proper, they produce optimal probabilistic predictions by theory. This could be possible in principle, since the hydrological literature presents generalized findings on the distributions of hydrological variables with increasing frequency and reliability. ○ They can largely facilitate interpretability in modelling, since they allow the inspection of the impact of their assumptions on both parameter and predictive uncertainty. ○ Their performance depends less on the length of the historical dataset than the performance of two-stage post-processing methodologies (see Table 8.2), since their fitting does not require sample splitting.
Disadvantages	○ Their predictive performance largely depends on the appropriateness of their assumptions. ○ They might get over-parameterized in an effort to ensure the adoption of proper assumptions. ○ Their use is accompanied by computational limitations.

Table 8.2. Advantages and disadvantages of two-stage hydrological post-processing methodologies (see also Evin et al. 2014; Chapter 9 herein). These post-processing methodologies estimate their error models conditional on the predictions provided by their hydrological models. The latter have been calibrated by using an independent segment of the historical dataset.

Advantages	○ They can be nearly assumption-free (i.e., their performance does not necessarily depend on the appropriateness of assumptions) when implemented with flexible machine learning quantile regression algorithms as error models. The advantages of these algorithms are listed independently in Table 8.3. ○ Computational requirements and limitations are mostly few in their case. Therefore, their automation and application to big datasets is feasible. This is one of the main reasons why two-stage hydrological post-processing is popular in forecasting applications. This popularity is emphasized e.g., by Evin et al. (2014). ○ In light of the two points above, their performance can be maximized by adopting algorithmic strategies and well-established guidelines from the machine learning literature (see e.g., the experiment presented herein). The role of big datasets for achieving optimal modelling solutions under this new-era approach is emphasized e.g., in Tyralis et al. (2019b).
Disadvantages	○ They largely lack interpretability by perception. Interactions between the hydrological model parameters and the trained version of the error model are ignored; therefore, their hydrological model parameter estimates are only auxiliary to predictive uncertainty quantification and cannot be used in any case for understanding parameter uncertainty. ○ Their performance depends more on the length of the historical dataset than the performance of Bayesian post-processing methodologies (see Table 8.1), since their fitting requires sample splitting. ○ The adoption of flexible machine learning quantile regression algorithms as error models has an additional cost in terms of interpretability and further increases the large-sample requirements (see the disadvantages of Table 8.3). These requirements are revealed and discussed e.g., in Section 7.5 herein.

Table 8.3. Advantages and disadvantages of statistical learning (or machine learning) quantile regression algorithms (see also Waldmann 2018; Sections 2.6 and 9.5.3 herein). Quantile regression algorithms issue quantile predictions instead of PDF predictions.

Advantages	○ They are ideal when the conditional distribution of the dependent variable is not known or is hard to deduce. ○ They model heteroscedasticity by perception and construction. ○ In light of the above point, they are also straightforward to apply, as they do not need to be fitted separately for each season (or month), in contrast to distribution-based modelling approaches (e.g., conditional-distribution models). ○ They are robust with respect to outliers in the observations of the dependent variable. ○ They are available in open source and mostly optimally programmed.
Disadvantages	○ They are trained separately for each quantile probability; therefore, the more the quantiles (or prediction intervals) we are interested in issuing, the more computationally costly the training process. ○ Quantile crossing is possible. ○ Parameter estimation is harder than in standard regression. ○ Their performance depends to some extent on the sample size. ○ They lack interpretability. Only their linear variant, i.e., the quantile regression model implemented herein, offers interpretability to some extent.

In the context described so far, Montanari and Koutsoyiannis (2012) introduced a flexible two-stage post-processing methodology (hereafter referred to as "MK blueprint methodology") that facilitates both probabilistic modelling and understanding from a stochastic perspective of rainfall-runoff (and other stochastic) relationships. In its basic configuration (for its outline, see Section 2.7.3), this methodology utilizes a single hydrological model to generate a large number of point predictions (hereafter referred to as "sister predictions"; adopting a similar terminology to the one by Nowotarski et al. 2016, Wang et al. 2016 and Liu et al. 2017). As implied by its post-processing nature, it also utilizes a second −necessarily statistical− model for modelling the error of the hydrological model (hereafter referred to as "error model").

Different variants of the MK blueprint methodology can be found in Sikorska et al. (2015), Quilty et al. (2019) and Chapter 7 of this thesis. The original blueprint and the variant by Sikorska et al. (2015) are formulated to explicitly consider input data uncertainty, while in both related papers a large number of hydrological model parameters are obtained by using the DREAM algorithm by Vrugt et al. (2009a; see also Vrugt 2016). This algorithm (see e.g., Schoups and Vrugt 2010; Laloy and Vrugt 2012; Vrugt et al. 2013; Sadegh and Vrugt 2014) is a popular Markov chain Monte Carlo (MCMC) algorithm for sampling from the posterior parameter distribution of hydrological models (see also the related implementations in Sadegh et al. 2015; Hernández-López and Francés 2017; Vrugt et al. 2008; Volpi et al. 2017). Other (non-Bayesian) methodologies could also be used for obtaining a large number of hydrological model parameters (Montanari and Koutsoyiannis 2012), while in absence of relevant information the MK blueprint methodology can also be applied without explicitly considering input data uncertainty (see e.g., the implementations in Quilty et al. 2019 and the formulations of the variants of Chapter 7 herein). Quilty et al. (2019) perform probabilistic water demand forecasting using exogenous variables; therefore, their variants constitute integrations within the MK blueprint framework of concepts particularly useful and/or popular for this task, such as bootstrapping, variable selection and wavelet decomposition.

In spite of their (larger or smaller) differences in terms of conceptualization, underlying modelling cultures and inherent modelling assumptions, all the above-outlined state-of-the-art techniques aim at filling a common knowledge gap that currently exists in the probabilistic hydrological modelling and forecasting literatures, specifically at answering the following research question: How to reduce modelling uncertainty as much as possible? Risk reduction in (probabilistic) hydrological modelling is the 20th of the 23 major "unsolved" hydrological problems, as posed by Blöschl et al. (2019, Section 3) through a community-based process. The present Chapter aspires to contribute to the large efforts made towards solving this problem.

We extensively test the hydrological modelling capabilities provided by the variants of the MK blueprint methodology introduced in Chapter 7 herein (hereafter collectively referred to as "working methodology"), when these variants are applied using the quantile regression model by Koenker and Bassett (1978; see also Koenker 2005) as error model. The quantile regression model is a balanced choice between interpretable and more flexible algorithms from the statistical learning literature. It has already been applied for post-processing hydrological predictions within hydrological modelling case studies (see e.g., Solomatine and Shrestha 2009; López López et al. 2014; Dogulu et al. 2015; Wani et al. 2017), while its use is more common in the field of hydrological forecasting (see e.g., Tyralis et al. 2019a and the references therein); see also the references in Dogulu et al. (2015), and Abbas and Xuan (2019) for applications of this model in other geoscience concepts.

For benchmarking purposes, we also apply the working methodology using the linear regression model (see e.g., James et al. 2013; Hastie et al. 2009) as error model, and two naïve probabilistic data-driven schemes. For the merits of using benchmarks in hydrological modelling, the reader is referred to Pappenberger et al. (2015); see also benchmarking examples in Montanari and Brath (2004), Evin et al. (2014), Sikorska et al. (2015), Tyralis and Papacharalampous (2017), Papacharalampous and Tyralis (2018), Quilty et al. (2019), Tyralis and Papacharalampous (2018), Tyralis et al. (2018, 2019a,c), Xu et al. (2018), as well as Chapters 3–7 and 9 of this thesis.

The working methodology is implemented within a large-sample real-world experiment. In the latter, we probabilistically solve monthly rainfall-runoff modelling problems for 270 catchments in the United States (US). Large-sample hydrological studies are increasingly carried out in the literature (see e.g., Perrin et al. 2001; Mouelhi et al. 2006a,b; Sawicz et al. 2011; Papalexiou and Koutsoyiannis 2013; Weijs et al. 2013; Bourgin et al. 2015; Coxon et al. 2015; Farmer and Vogel 2016; Langousis et al. 2016; Ren et al. 2016; Tyralis and Koutsoyiannis 2017; Tyralis and Papacharalampous 2017, 2018; Tyralis et al. 2018; Bock et al. 2018; Xu et al. 2018; Tyralis et al. 2019a,d; Xu et al. 2019; see also Chapters 3–5, 9 herein), while this is the first work performing a large-scale assessment of the MK blueprint methodology.

The aims of the Chapter (that can be addressed only within a large-sample hydrological study) are to:

1) Validate the working methodology.
2) Compare its variants both in terms of predictive performance and computational requirements.
3) Quantify the improvement in performance when using the quantile regression model instead of the linear regression model as error model. In contrast to the latter model, the former model is known to be appropriate for modelling heteroscedasticity (Koenker and Hallock 2001; Koenker 2005).
4) Demonstrate in real-world applications the larger robustness in performance of the working methodology compared to two-stage post-processing methodologies producing a single point hydrological prediction (hereafter referred to as "basic" two-stage post-processing methodologies).
5) Provide an empirical proof of the ability of the working methodology to harness the wisdom of the crowd. This ability stems from the concept of combining probabilistic predictions via simple quantile averaging, on which this methodology relies, while in Lichtendahl et al. (2013, Section 5) it is defined as follows: The average of predictions scores no worse −usually better− than the average of the scores of the combined predictions. According to the same study, this ability has to be empirically proven for the problem and scores of interest, since the proofs in Lichtendahl et al. (2013) are made for stylized versions.

8.2 Experimental data and methods

In this section, we present the experimental methodology of the Chapter by emphasizing implementation details, as it is suggested by the guidelines by Abrahart et al. (2008). Statistical software information is independently summarized in Section 2.9.4. The working methodology is outlined in Section 8.2.1, while the reader is referred to Section 7.2.1 for its detailed and formal presentation.

8.2.1 Working methodology

This Section aims at summarizing the working methodology. For this summary, we first define the time period $T = \{1, \ldots, (n_1+n_2+n_3)\}$, and its three distinct sub-periods $T_1 = \{1, \ldots, n_1\}$, $T_2 = \{(n_1+1), \ldots, (n_1+n_2)\}$ and $T_3 = \{(n_1+n_2+1), \ldots, (n_1+n_2+n_3)\}$. We also define the sister model realizations as variants of a single hydrological model, each using different parameter values. The latter are obtained by calibrating the hydrological model in the period T_1. The calibration could be made by using either Bayesian schemes (e.g., Markov Chain Monte Carlo simulation sampling; see e.g., the procedures described in Section 8.2.6) or informal calibration schemes (see e.g., the procedures described in Section 8.4). Let us assume that we obtain m sister model realizations, where m is adequately large. Each sister model realization is then applied in the period $\{T_2, T_3\}$. The m resulted sister predictions also extend in the period $\{T_2, T_3\}$. We subsequently compute the sister model realizations' errors in the period T_2 by using the sister predictions alongside with their corresponding target values.

Information about the sister model realizations' error is then obtained by training a statistical learning regression model that is suitable for predicting quantiles (hereafter referred to as "error model"; see e.g., the error models exploited in Chapter 9 herein) in the period T_2. In particular, we regress the sister model realizations' error at time t (response variable) on selected predictor variables (e.g., the sister prediction at time t). For each sister prediction extending in the period T_3, we (a) predict a set of quantiles (with selected probabilities) of the sister model realization's errors using the information obtained at the preceding step, and (b) transform these predictive quantiles to auxiliary predictive quantiles of the hydrological process of interest (by subtracting them from their corresponding sister prediction). Finally, at each time $t \in T_3$ we group the auxiliary predictive quantiles of the hydrological process of interest based on their corresponding probability (e.g., probability 0.95) to average them over each group. The resulted time series are the output quantile predictions.

The basic steps adopted within the working methodology are also summarized in Figure 7.1.

The working methodology is subdivided into three alternative variants. These variants differ in the error model's training only. Specifically:

- Variant 1 trains the error model m times, each time on a different dataset formed by using a different sister prediction;
- Variant 2 trains the error model on a single dataset formed by using all sister predictions;
- Variant 3 also trains the regression model once; however, the training here is made on a dataset formed by using one randomly selected sister prediction.

We note that the three variants reduce to the same method in the case that a single point hydrological prediction is generated. In this case, the working methodology would fall into the category of basic two-stage post-processing methodologies using regression models.

8.2.2 Rainfall-runoff dataset

We use the US Model Parameter Estimation Experiment (MOPEX) dataset, which is documented in Schaake et al. (2006; see also Schaake et al. 2000, Duan et al. 2006, Wagener et al. 2006). This dataset comprises hydrometeorological and land-surface-characteristic data originating from US catchments of intermediate size, and has been extensively used in hydrological studies (see e.g., Kavetski et al. 2006b; Sawicz et al. 2011; Huang et al. 2013; Evin et al. 2014; Weijs et al. 2013; Ye

et al. 2014; Ren et al. 2016; Hernández-López and Francés 2017). All included catchments are unregulated; therefore, the modelling assumption of stationarity is reasonable on these real-world data (see e.g., Koutsoyiannis 2011; Montanari and Koutsoyiannis 2014; Koutsoyiannis and Montanari 2015).

From the original dataset we retrieve daily information about mean areal precipitation, climatic potential evaporation and streamflow discharge for 431 US catchments. The retrieved data span from January 1st, 1948 to December 31st, 2003, thus covering a 56-year period, yet containing a large amount of missing and negative (unrealistic) values. We process the retrieved data aiming to simultaneously achieve two objectives, i.e., (a) extracting time series blocks covering a long common period of complete historical information (with no missing or unreliable values), and (b) retaining historical information for a large number of catchments. A satisfactory compromise between these two objectives is reached when using as sampling period each of the periods 1950–1999 and 1949–1998. Both these samplings result in 50 (calendar) years of complete daily time series data for 270 catchments. We adopt the former option, as it offers (slightly) more recent data compared to the alternative one. The retained time series data are aggregated to produce total monthly precipitation, potential evaporation and streamflow discharge time series, each comprising 600 values. The resulted total monthly time series constitute the herein examined dataset. The locations of the examined MOPEX catchments are depicted in Figure 8.1. A wide range of climate regimes is well-represented by this sample set of catchments (see Kottek et al. 2006).

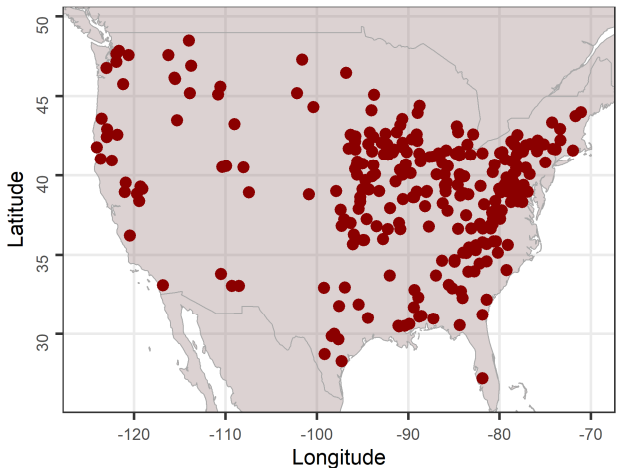


Figure 8.1. Locations of the 270 MOPEX catchments examined within the large-sample experiment of the Chapter. The data are sourced from Schaake et al. (2006).

8.2.3 Overview of modelling methodology

The monthly data (see Section 8.2.2) are handled as described in Section 8.2.4. We use these data to assess two basic and six ensemble schemes in obtaining interval predictions. Two statistical learning regression models (see Section 8.2.5) and one hydrological model (see Section 8.2.6) are utilized for this assessment. We define the prediction problem to be solved as the problem of predicting the quantiles with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ of monthly streamflow discharge in the period T_3 (hereafter referred to as "quantiles of interest") given monthly precipitation and monthly potential evaporation observations for the period $\{T_0, T_1, T_2, T_3\}$ and monthly streamflow discharge observations for the period $\{T_0, T_1, T_2\}$. These periods are defined in Section 8.2.4.

The basic schemes are "linear regression" and "quantile regression". Both of them are implemented by training the regression model directly on monthly data for the period $\{T_0, T_1, T_2\}$ and, subsequently, by using the trained regression model to predict the quantiles of interest (for the period T_3). The predictor variables in regression are monthly precipitation at time t and monthly potential evaporation at time t, while the response variable is monthly streamflow

discharge at time t. We note that these benchmark implementations of the regression models can only be viewed as naïve data-driven approaches to probabilistic hydrological modelling (because of the small number of predictor variables utilized). For more sophisticated implementations (which are outside of the scope of the Chapter), more predictor variables could be used.

On the other hand, the ensemble schemes can be perceived as different configurations of the working methodology (allowing us to address the aims of the Chapter). Ensemble schemes 1–3 (4–6) are based on variants 1–3 respectively of this methodology. Moreover, ensemble schemes 1–3 utilize a different statistical learning regression model as error model with respect to ensemble schemes 4–6. Specifically, ensemble schemes 1–3 utilize the linear regression model, while ensemble schemes 4–6 utilize the quantile regression model. The same ensemble schemes are also implemented in Chapter 7 of this thesis; however, their implementation therein is made by using toy hydrological models.

We describe here below the application of the ensemble schemes for a single catchment; the extension to all catchments is straightforward. The following steps are made once for all ensemble schemes:

1) We use monthly precipitation, potential evaporation and streamflow discharge observations for the period T_1 to obtain 600 sets of the hydrological model's parameters, as detailed in Section 8.2.6. This number of parameter sets offers a good compromise between computational requirements and predictive performance. We use these parameters to define 600 sister model realizations.

2) We obtain 600 sister predictions for the period $\{T_2, T_3\}$. Each sister prediction is obtained by implementing a different sister model realization given the monthly precipitation and potential evaporation observations for the same period. Each sister prediction contains 444 values.

3) We compute the sister model realizations' errors in the period T_2 by using the parts of the sister predictions extending in the same period alongside with their corresponding target values. The total number of the computed error values is 600 × 144 = 86 400.

The following steps are made independently by each ensemble scheme:

4) We train the error model in the period T_2. Specifically, we regress the sister model realizations' error at time t (response variable) on the sister prediction at time t (predictor variable). Ensemble schemes 1 and 4 train the error model 600 times, each time using a different sister prediction and its corresponding sister model realization's errors (use of 600 training datasets of size 144). Ensemble schemes 2 and 5 train the error model once by using all sister predictions and their corresponding sister model realizations' errors (use of one training dataset of size 86 400). Ensemble schemes 3 and 6 train the error model once by using a randomly selected sister prediction and its corresponding sister model realization's errors (use of one training dataset of size 144). The result of this step is 600 trained versions of the error model (each corresponding to a specific sister prediction) for each of the ensemble schemes 1 and 4, and one trained version of the error model for each of the ensemble schemes 2, 3, 5 and 6.

5) We apply the trained versions of the error models, obtained in the preceding step, to predict the quantiles with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ of each sister model realization's errors in the period T_3 given their corresponding sister prediction. For each ensemble scheme, the result of this step is 600 probabilistic predictions, each consisting of 10 quantile predictions.

6) We obtain 600 auxiliary probabilistic predictions of the process of interest, each consisting of 10 quantile predictions, by subtracting each of the 600 × 10 = 6 000 quantile predictions from its corresponding sister prediction.

7) The finally delivered predictive quantile with probability $p \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ at time $t \in T_3$ is the average over all auxiliary predictive quantiles with the same probability p at time t, i.e., the average of 600 in number auxiliary

predictive quantiles. The finally delivered predictive quantiles of the process of interest form the 99%, 97.5%, 95%, 90% and 80% central prediction intervals.

The total number of sister predictions produced herein is 270 × 600 = 162 000, each containing 444 values, while the total number of auxiliary quantile predictions is 270 × 600 × 10 × 6 = 9 720 000, each containing 300 values, and the finally delivered quantile predictions are 270 × 10 × 8 = 21 600, each containing 300 values. For addressing aim 2 of the Chapter, we measure the computational time consumed by each ensemble scheme.

8.2.4 Data handling and related remarks

Following the notations provided in Section 8.2.1, we define the periods T_1 = {13, …, 156}, T_2 = {157, …, 300} and T_3 = {301, …, 600} (corresponding to years 1951–1962, 1963–1974 and 1975–1999 respectively). We include a large amount of the available information in the period T_3 to facilitate proper testing. We also define period T_0 = {1, …, 12} (corresponding to year 1950). This period is used for warming-up the hydrological model (see Section 8.2.6). For a justification on this choice the reader is referred to Section 2.4.3.

We note that the data are used without any transformation applied to it (see Section 2.1.8). We attempted to apply the linear regression and quantile regression schemes to river discharge data that were pre-processed by using the square-root transformation. Nevertheless, this pre-processing (not presented here for reasons of brevity) had a negative effect on the quality of the naïve probabilistic predictions, mainly to those delivered by the linear regression scheme; therefore, it was abandoned. Moreover, a logarithmic transformation was not feasible, due to some zero monthly values of river discharge. We also attempted to apply the Yeo-Johnson and ordered quantile normalization transformations on the response, when solving the error modelling problems outlined in Section 8.2.3 (steps 4–5 of the application of the ensemble schemes). These transformations were also abandoned due to infinite predicted values. The square-root and logarithmic transformations on the response variable, i.e., the error of the hydrological model at time t, are not feasible due to the existence of negative error values.

8.2.5 Regression models and related procedures

We implement the linear regression and quantile regression models. We use these two models to solve the regression problems described in Section 8.2.3. Koenker and Hallock (2001) comprehensively discuss the difference in rationale behind these two models, as summarized subsequently. The training outcome in linear regression (i.e., least-squares regression with i.i.d. Gaussian errors with zero mean and constant variance; James et al. 2013) is a conditional mean function. The latter is a function describing how the mean of the response variable changes with the changes of the predictor variables. This function is obtained by minimizing a sum of squared residuals. On the contrary, the training outcome in quantile regression is a set of conditional quantile functions, obtained by minimizing the average quantile score. While in linear regression the PDF of the response variable is assumed to have the exact same variance and distributional shape independently of the values of the predictors, quantile regression does not make any particular assumption about this PDF; therefore, allowing a more representative description of the relationship between predictors and predictand. Related technical remarks can be found in Section 2.6.2.

8.2.6 Hydrological model and related procedures

We implement the monthly GR2M (see Section 2.4.1). This model has two parameters that are hereafter denoted with θ_1 and θ_2. We simulate the posterior distribution of these parameters conditional on the observations of the period T_1 within a Bayesian MCMC framework (see Section 2.5.2). We use flat priors for both the parameters θ_1 and θ_2. The likelihood error function is defined with Equation (8.1). In this Equation, y_t is the monthly streamflow discharge observations at time t, $u_t(\theta_1, \theta_2)$ is the prediction of the GR2M model at time t and $|T_1|$ is the number of target data points included in the period T_1.

$$L(\theta_1, \theta_2) \propto (\sum_t (y_t - u_t(\theta_1, \theta_2))^2)^{-|T_1|/2} \quad (8.1)$$

We run 3 parallel Markov chains with different initial values, each comprising 2 000 iterations. We assess the approximate convergence of these chains by adopting the algorithm detailed in Section 2.5.2. The simulation process is repeated until a point estimate smaller than 1.10 is delivered. Once the simulation is over, we retain the last 200 values of each chain, i.e., 600 values in total for each catchment. An example of simulated and retained parameters is presented in Figure 8.2.

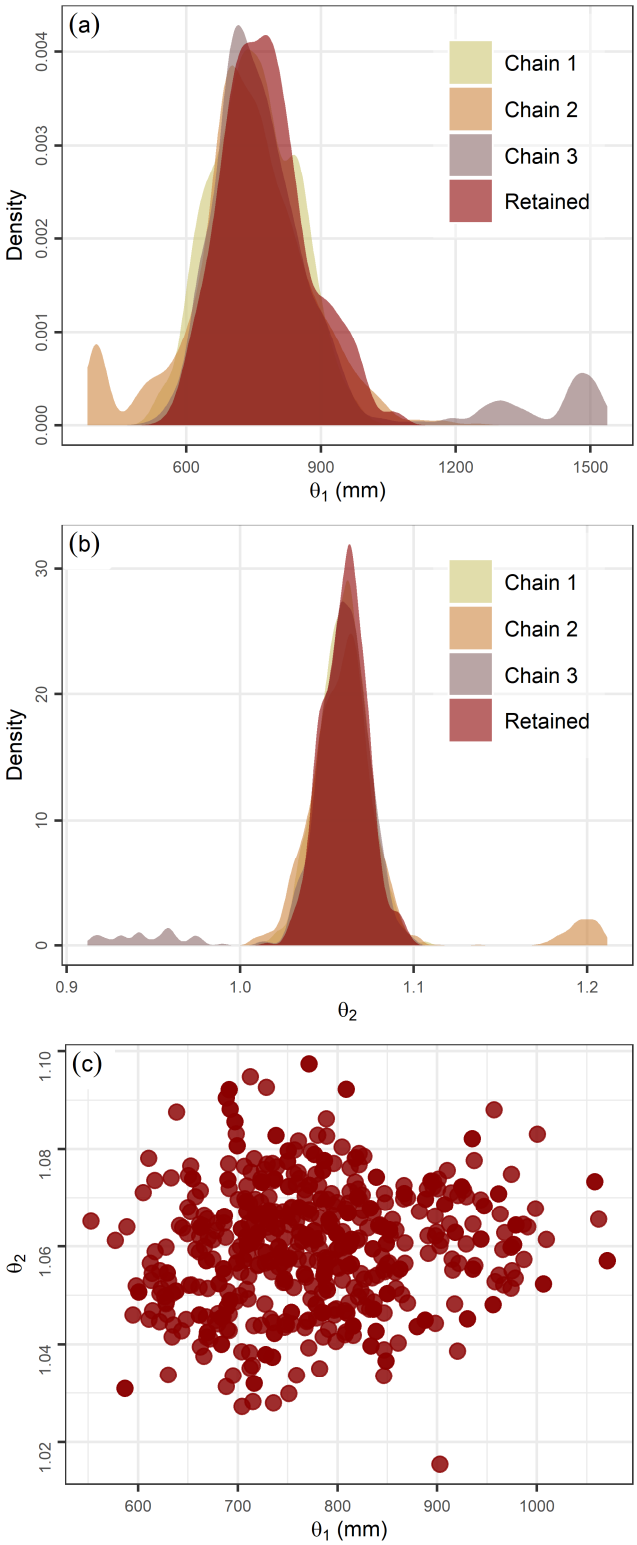


Figure 8.2. Simulated chains in (a–b), and retained parameter values in (a–c) obtained using precipitation, potential evaporation and streamflow discharge information for the period T_1 (years 1951–1962) for a randomly selected catchment.

8.2.7 Prediction interval assessment

We assess the quality of the interval predictions by computing their coverage probabilities, average widths and average interval scores. All computations are made for the period T_3, as detailed in Section 2.8.3. The computed metrics are used according to Table 8.4 to assess two

desired properties in probabilistic modelling, i.e., the reliability and sharpness of interval predictions. For illustrative purposes, we also present examples of prediction intervals. We do not present QQ-plots for the following two reasons: i) we deliver predictive quantiles with probabilities that are either equal or smaller than 0.10, or equal or larger than 0.90 (since we are interested in specific prediction intervals; see Section 8.2.3), while QQ-plots are ideal when PDF predictions (or at least sets of predictive quantiles with probabilities running on a grid from 0 to 1) are delivered, and ii) we are interested in objectively assessing on a massive scale the predictive performance of several prediction schemes (separately for each of them) in 270 catchments, while QQ-plots are particularly useful for assessments made on a smaller scale.

Table 8.4. Metrics used for assessing the prediction interval $(1 - \alpha)$, $0 < \alpha < 1$. Their definitions are given in Section 2.8.2 (see also Table 2.6).

Metric	Preferred values	Criterion/criteria
Coverage probability (CP_α)	Smaller $\lvert CP_\alpha - (1 - \alpha)\rvert$	Reliability
Average width (AW_α)	Smaller AW_α	Sharpness
Average interval score (AIS_α)	Smaller AIS_α	Reliability, sharpness

Since the magnitude of the average interval score largely depends on the examined dataset, we mostly base our conclusions on relative improvements in terms of average interval score (see Section 2.8.2). Specifically, for addressing aims 1–3 of the Chapter we compute the relative improvements provided all prediction schemes with respect to the linear regression and quantile regression schemes, and the relative improvements provided by ensemble schemes 4–6 with respect to ensemble schemes 1–3. For addressing aim 4 of the Chapter, we use all auxiliary quantile predictions (9 720 000 in number) and the finally delivered quantile predictions (21 600 in number) to compute the relative improvements in terms of average interval score, when using the output of each ensemble scheme instead of each of the auxiliary interval predictions combined to obtain this output, according to Equation (8.2). In this equation, AIS_{OUT} denotes the average interval score of the output interval prediction (obtained by using the method), AIS_{IN_i} the average interval score of one from the auxiliary interval predictions $\{IN_i, i = 1, \ldots, 600\}$ that are averaged by the method to obtain the output interval prediction (with average interval score equal to AIS_{OUT}), and RI_{OUT,IN_i} the relative improvement of interest.

$$RI_{OUT,IN_i} := (AIS_{IN_i} - AIS_{OUT})/AIS_{IN_i} \tag{8.2}$$

Finally, for addressing aim 5 of the Chapter we use the same quantile predictions used for addressing aim 4 to compute the relative differences between the average interval score computed for the outputs of the ensemble schemes, i.e., the average of 600 probabilistic predictions (denoted with AIS_{OUT}; see above), and the average of the average interval scores computed for each of the combined auxiliary interval predictions $\{AIS_{IN_i}, i = 1, \ldots, 600\}$ (denoted with $AAIS_{IN}$; see also Equation (8.3) for its definition), the latter used as reference for the former. The computation of these relative differences is made using an equation analogous to Equations (2.92) and (8.2), specifically Equation (8.4), where $RD_{OUT,AAIS_{IN}}$ denotes the relative difference of interest.

$$AAIS_{IN} := \sum_{i=1}^{600}(AIS_{IN_i})/600 \tag{8.3}$$

$$RD_{OUT,AAIS_{IN}} := (AAIS_{IN} - AIS_{OUT})/AAIS_{IN} \tag{8.4}$$

8.3 Results and discussions

In this Section, we present and discuss the results of our large-sample experiment. For reasons of brevity, we present only a representative sample of the conducted Figures and Tables, while in Papacharalampous et al. (2019c; hereafter referred to as "Chapter's supplement") the interested reader can find some additional ones.

8.3.1 Overall assessment of the working methodology

This section is devoted to addressing aims 1–3 of the Chapter (see Section 8.1). The presentation is mostly made in an aggregated form across all the examined catchments, while emphasis is placed on the average interval scores computed for the obtained prediction intervals and on the relative improvements provided by the ensemble schemes with respect to the basic schemes in terms of the same metric. This choice is implied by the fact that an objective co-assessment regarding reliability and sharpness provided, for instance, by the interval score is of the most practical relevance in technical applications. In spite of this placed emphasis and keeping pace with several works available in the literature (e.g., in Renard et al. 2010, 2011; Evin et al. 2013, 2014; Tyralis et al. 2019a; Chapter 9 herein), we separately summarize information that is purely related to the assessment of reliability and information that is purely related to the assessment of sharpness. In this way, we facilitate an adequate degree of interpretability and understanding of what follows.

In Figure 8.3, we present several examples of prediction intervals, all delivered by ensemble scheme 5, in comparison to the targeted data points. As extracted from Figure 8.3, this scheme offers a (rather) high degree of reliability, i.e., it delivers prediction intervals that mostly contain the desired percentage of data points. The same applies to the remaining prediction schemes. Herein the related information is objectively summarized with Figure 8.4 and Table 8.5. In Figure 8.4, we comparatively present the boxplots of the coverage probabilities computed for all delivered and assessed solutions to the 270 examined rainfall-runoff problems. These coverage probabilities are rather good (than bad). The latter characterization holds, especially if we consider that the examined monthly time series are of only 600 values. In particular, the coverage probabilities for the 95% prediction intervals are comparable to those computed for the probabilistic predictions of Bock et al. (2018).

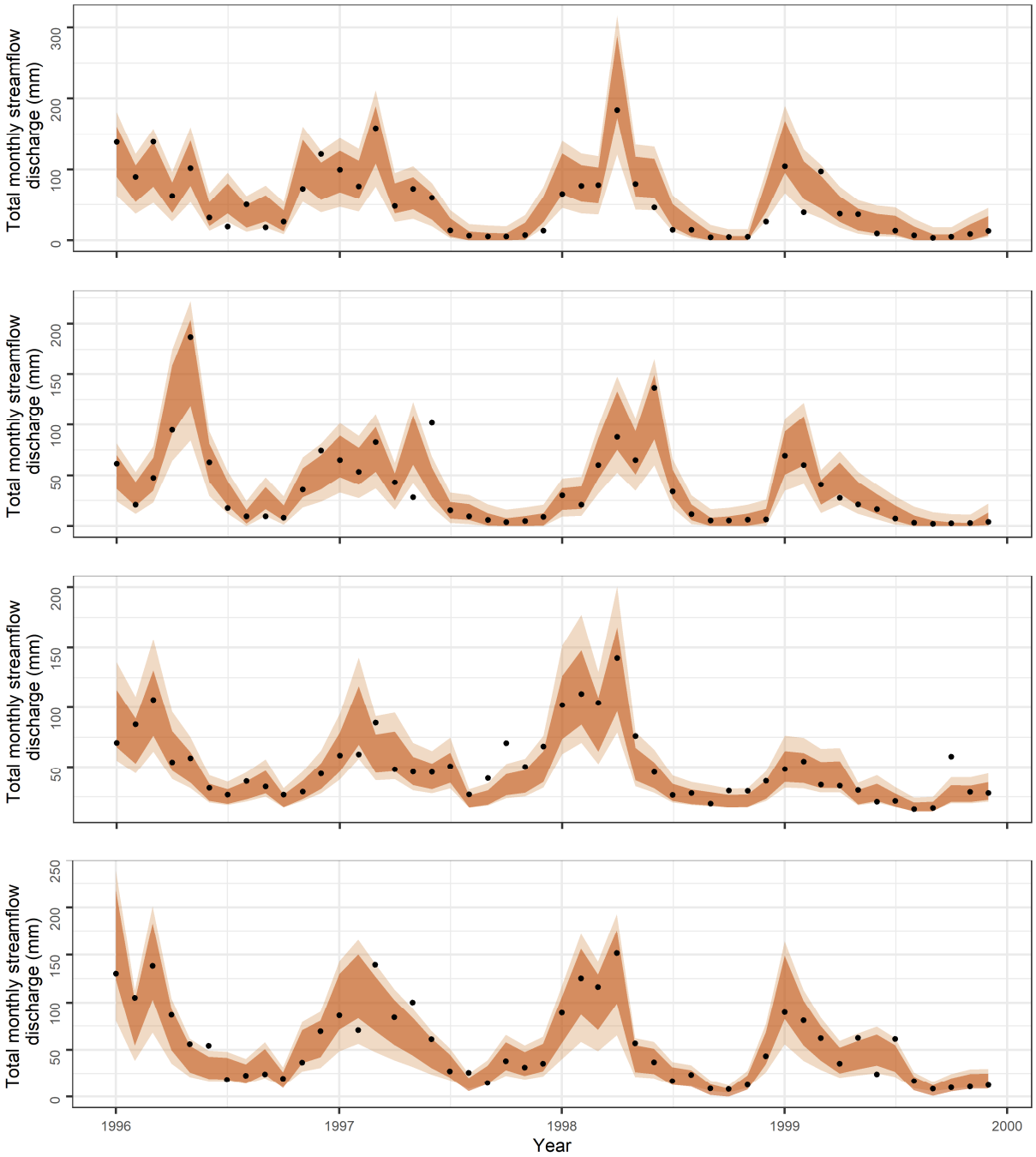


Figure 8.3. Prediction intervals provided by ensemble scheme 5 for four arbitrary catchments and a common 4-year sub-period of the period T_3 (years 1996–1999). Black dots denote the targeted points, while light orange and dark orange ribbons denote the 95% and 80% prediction intervals respectively.

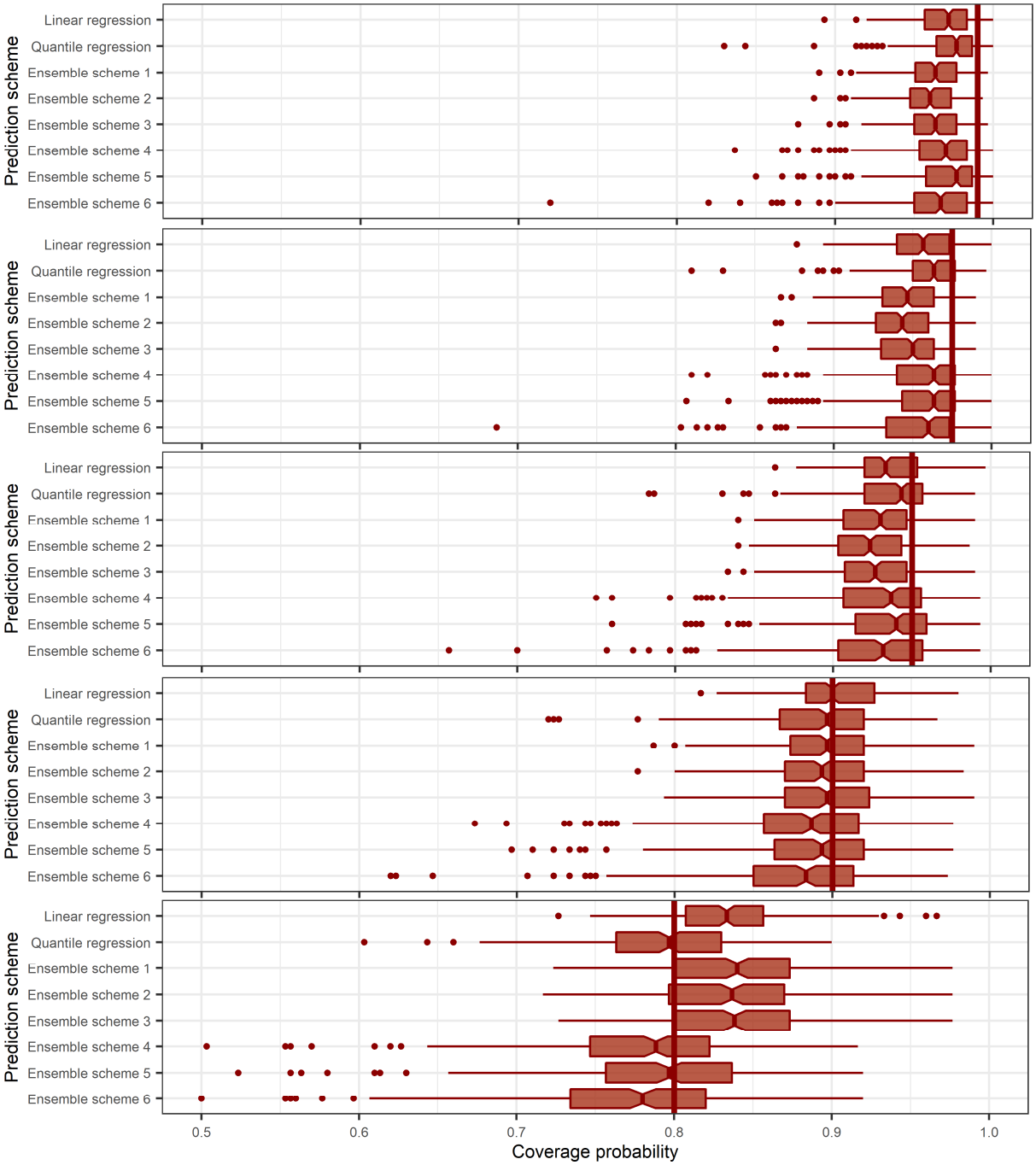


Figure 8.4. Coverage probabilities computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values. The optimal values are denoted with red thick vertical lines.

Table 8.5. Average coverage probabilities computed for the prediction intervals delivered by the compared schemes for the period T_3 (years 1975–1999). Each presented value summarizes 270 metric values.

Prediction scheme	99% prediction intervals	97.5% prediction intervals	95% prediction intervals	90% prediction intervals	80% prediction intervals
Linear regression	0.969	0.955	0.937	0.904	0.835
Quantile regression	0.973	0.961	0.936	0.889	0.793
Ensemble scheme 1	0.962	0.946	0.926	0.895	0.834
Ensemble scheme 2	0.959	0.943	0.923	0.892	0.834
Ensemble scheme 3	0.962	0.946	0.926	0.895	0.837
Ensemble scheme 4	0.965	0.953	0.928	0.881	0.781
Ensemble scheme 5	0.969	0.956	0.932	0.886	0.789
Ensemble scheme 6	0.961	0.948	0.923	0.874	0.773

While the average-case reliability of all prediction schemes is remarkably high (see Table 8.5), the performance of the prediction schemes in terms of coverage probabilities varies from catchment to catchment (see Figure 8.4). The observed differences in performance become larger, e.g., in terms of interquartile range of the formed datasets, as we move from the 99% to the 80% prediction intervals. Moreover, although differentiations are observed between prediction schemes, the overall performance of most schemes is rather of the same quality (in particular for the outer prediction intervals), with the quantile regression scheme and ensemble scheme 5 to be the best-performing, especially the former one.

The average widths, on the other hand, clearly favour the ensemble schemes over the basic schemes (see Figure 8.5), with ensemble schemes 4–6 providing sharper predictions than ensemble schemes 1–3. In terms of the same criterion, ensemble schemes from the former (latter) category exhibit remarkably close performance to each other. The same applies in terms of coverage probabilities. As already expected because of the large differences observed in the river discharge regimes of the examined catchments, the average widths of the prediction intervals may differ significantly from catchment to catchment. These differences become smaller, as we move from the outer to the inner prediction intervals, i.e., from the 99% to the 80% prediction intervals.

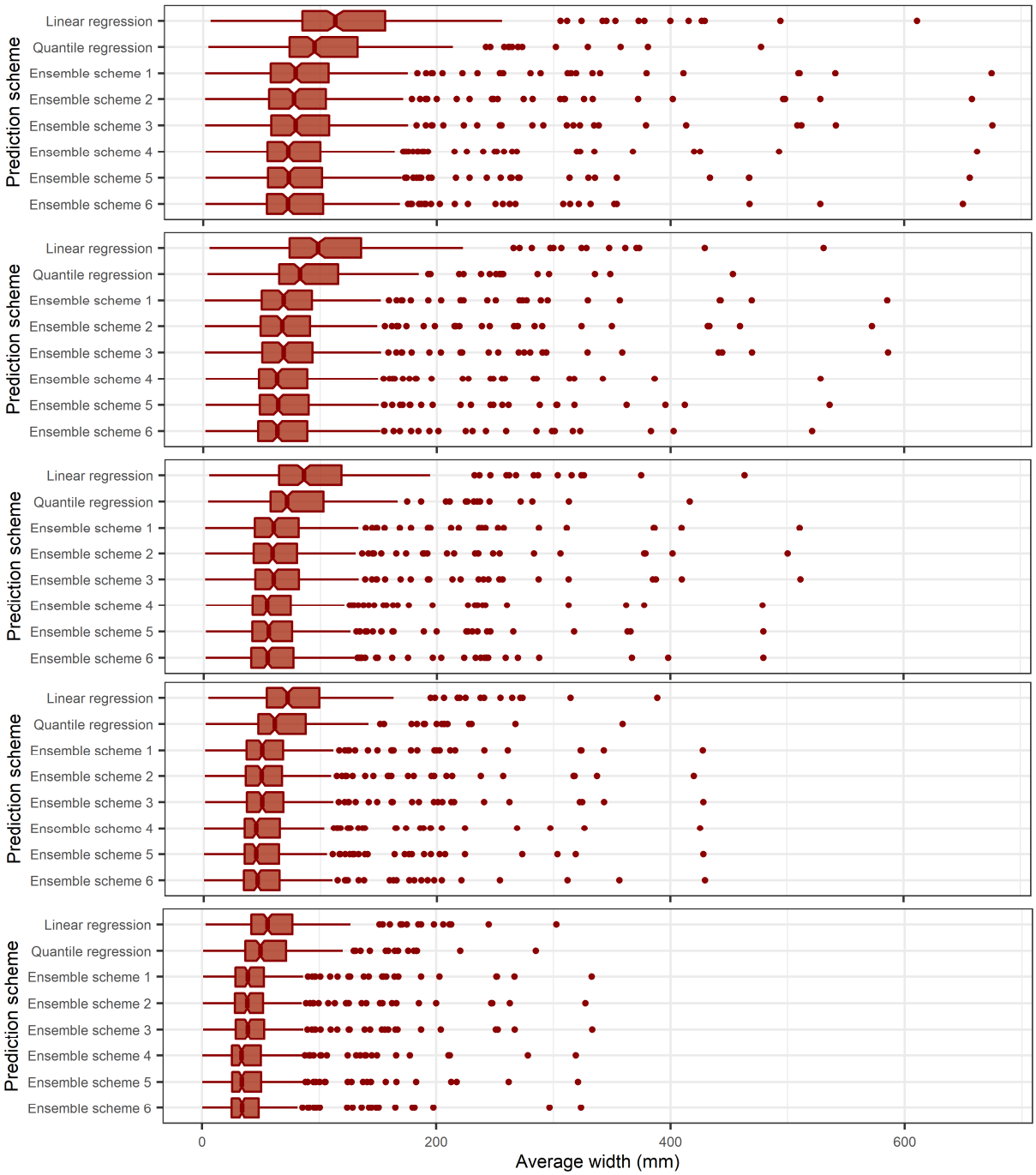


Figure 8.5. Average widths computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values.

The above-outlined information is objectively summarized in the average interval scores. The latter are collectively presented in Figure 8.6. The main information extracted from this figure is that (a) ensemble schemes 1–3, as well as ensemble schemes 4–6, exhibit very close performance to each other, (b) each ensemble scheme exhibits a better overall performance than its corresponding basic scheme, and (c) ensemble schemes 1–3 perform better than the quantile regression scheme for the 90% and 80% prediction intervals. Observation (b) indicates that the herein adopted implementations of the working methodology have an advantage over the naïve implementations of the data-driven (or purely statistical) models. This advantage should be further investigated before any generalization is made; nevertheless, this additional investigation

involving, for instance, utilization of more predictor variables, goes beyond the aim of the present Chapter.

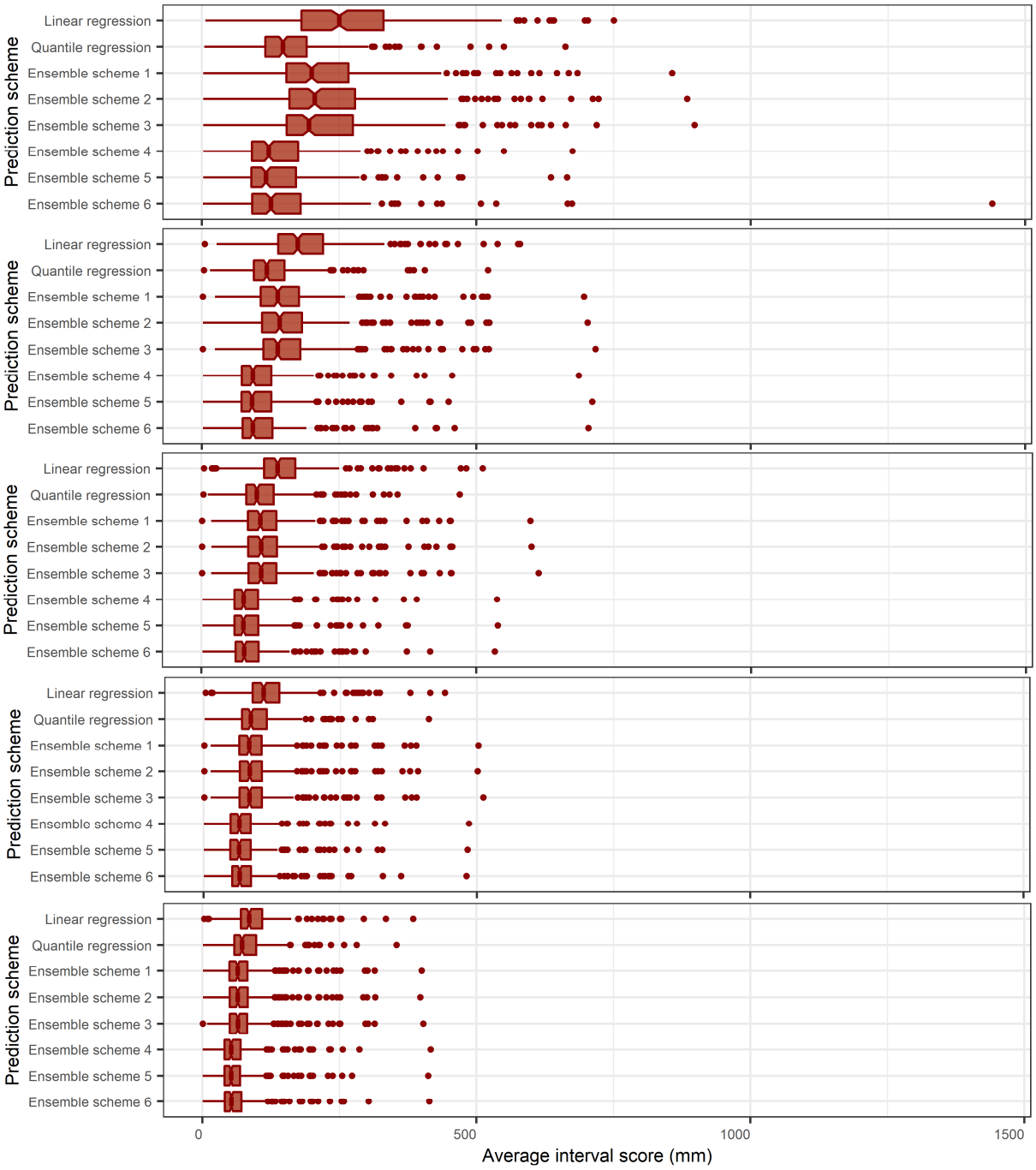


Figure 8.6. Average interval scores computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values.

We also note that both observations (a) and (b) are roughly expected already from the examination of Figures 8.4 and 8.5. By examining the aggregated average interval scores we additionally observe that the differences with respect to this metric are in average smaller for the inner prediction intervals than for the outer ones (as expected; see Section 8.2.7). Some small differences in the performance of ensembles schemes 1–3, favouring to a small extent ensemble schemes 1 and 3 over ensemble scheme 2, are mostly noticeable for the 99% and 97.5% prediction intervals. Similarly, ensemble scheme 5 seems to perform slightly better than ensemble scheme 4

for the same prediction intervals. It is also more effective than ensemble scheme 6 for all five prediction intervals.

To further inspect all differences, both the smaller and larger ones, in terms of rankings, the latter resulted for each catchment and for each examined prediction interval according to the computed average interval scores, we present Figures 8.7 and 8.8. The maps displayed in the former figure correspond to the upper side-by-side boxplots displayed on Figure 8.6, while allowing the examination of the rankings resulted both per catchment and per prediction scheme. From these maps we perceive that ensemble scheme 5 is ranked in a better average position than the remaining prediction schemes for the 99% prediction intervals, closely followed by ensemble schemes 4 and 6. Moreover, the quantile regression scheme is mostly ranked above the linear regression scheme and ensemble schemes 1−3. These schemes are mostly ranked in the last four positions. Importantly, there is not a fixed ranking position for any of the prediction schemes across the various catchments, while there are also some few catchments in which the four less competitive ones perform better than some the remaining. The quantile regression scheme is also ranked in the first three positions for a sufficient number of catchments. These latter observations provide us with a good reason to always perform large-scale benchmark experiments instead of (or alongside with) case studies.

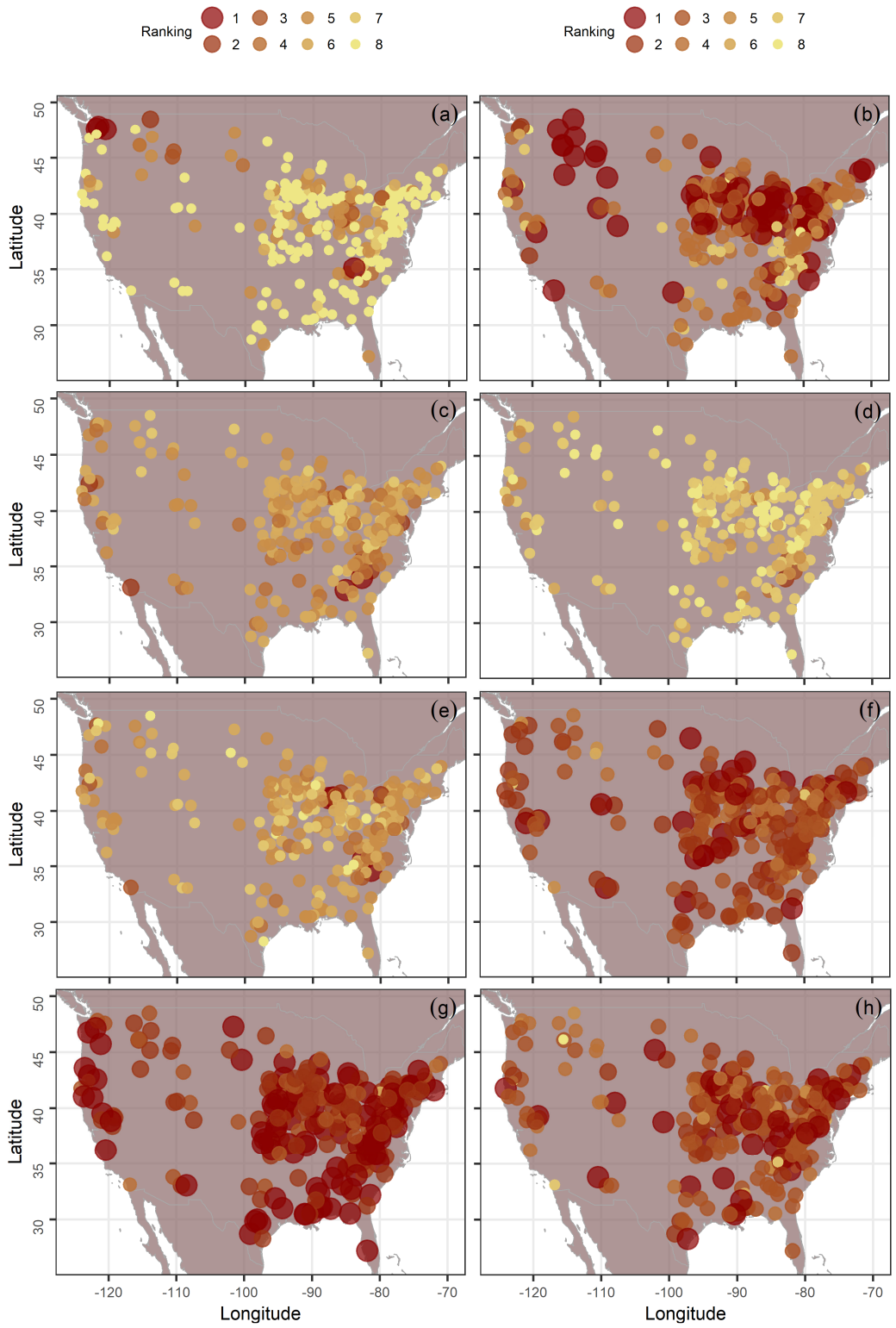


Figure 8.7. Rankings of (a) linear regression, (b) quantile regression and ensemble schemes (c–h) 1–6 according to the average interval scores computed for the 99% prediction intervals delivered for the period T_3 (years 1975–1999). The prediction schemes are ranked from best (1st) to worst (8th).

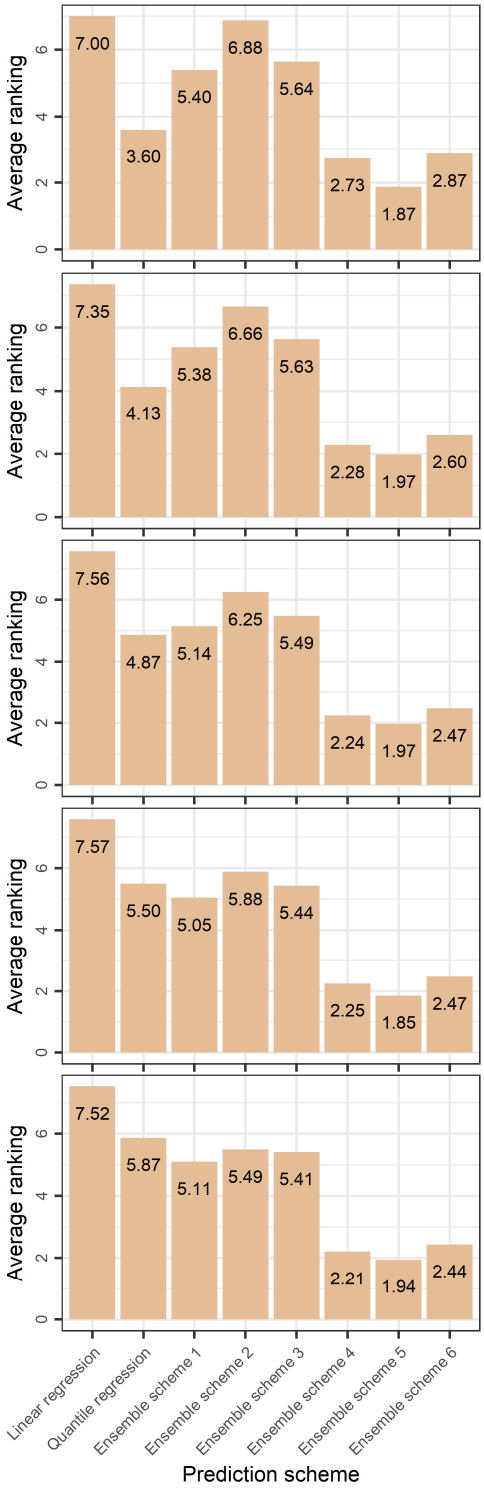


Figure 8.8. Average rankings of the prediction schemes according to the average interval scores computed for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). The prediction schemes are ranked from best (1st) to worst (8th). Each bar summarizes 270 values.

Overall, the image depicted in Figure 8.7 is rather neat when contrasted with its corresponding image in a similar visualization by Tyralis and Papacharalampous (2018); see Figure 4 therein. The latter study presents a large-scale comparison of point prediction methods that are equivalent to each other in a long run; therefore, no pattern is observed in their performance when the latter is depicted in maps. The pattern clearly observed in Figure 8.7, favouring the quantile regression model over the linear regression one, is due to the suitability of

the former algorithm for modelling heteroscedasticity. Thus, it is our knowledge on the examined problem and the difference in the appropriateness of the adopted methodologies that created this pattern rather than anything else.

As also emphasized in Chapter 3, only our knowledge on the system could make a tangible difference in (predictive) modelling in a long run. In fact, the homoscedasticity assumption is known to be inefficient when made during the probabilistic modelling of hydrological variables, such as the monthly river discharge variables that are of interest herein (see the comments, e.g., in Schoups and Vrugt 2010; Montanari and Koutsoyiannis 2012; Evin et al. 2013, 2014). Therefore, more flexible algorithms not assuming homoscedasticity are a reasonable choice to be made in such cases, while the same algorithms do not offer anything in comparison with less flexible algorithms in modelling cases where the homoscedasticity assumption is reasonable; see also Chapter 7 herein, in particular the results displayed in Tables 7.7 and 7.8 for an illustration-justification of this fact.

The greatest part of the ranking-related information extracted from Figure 8.7 applies as well to the remaining prediction intervals, while a summary of this information for the 99%, 97.5%, 95%, 90% and 80% prediction intervals, presented in Figure 8.8, provides additional observations. The latter effectively complement those obtained from Figure 8.6. In fact, for all prediction intervals ensemble scheme 5 exhibits the best average-case ranking, closely followed by ensemble schemes 4 and 6. Moreover, the quantile regression scheme exhibits a significantly better (comparable) average-case ranking than (with) ensemble schemes 1–3 for the 99% and 97.5% (95%, 90% and 80%) prediction intervals, while the linear regression scheme is the worst performing in terms of average rankings, as it could be expected already from Figure 8.6.

To obtain a more faithful image of the gain or loss in performance when using each prediction scheme over the remaining ones, in Figure 8.9 we present the side-by-side boxplots of the relative improvements in terms of average interval score with respect to the linear regression scheme, while in Figure 8.10 we present the respective information using the quantile regression scheme as a reference. The closeness in the performance of ensembles schemes 1–3 is also perceivable by the examination of these figures. The same applies to the closeness in the performance of ensemble schemes 4–6. Nevertheless, the small differences favouring ensemble schemes 1 and 3 over ensemble scheme 2, and ensemble scheme 5 over ensemble schemes 4 and 6 are also highlighted. Additionally, we observe that the differences in the relative performance of a specific prediction scheme can be large, while there are cases in which the ensemble schemes are (far) worse than their respective basic schemes. However, the long-run image clearly favours the former over the latter, as already expected from the preceding visualizations.

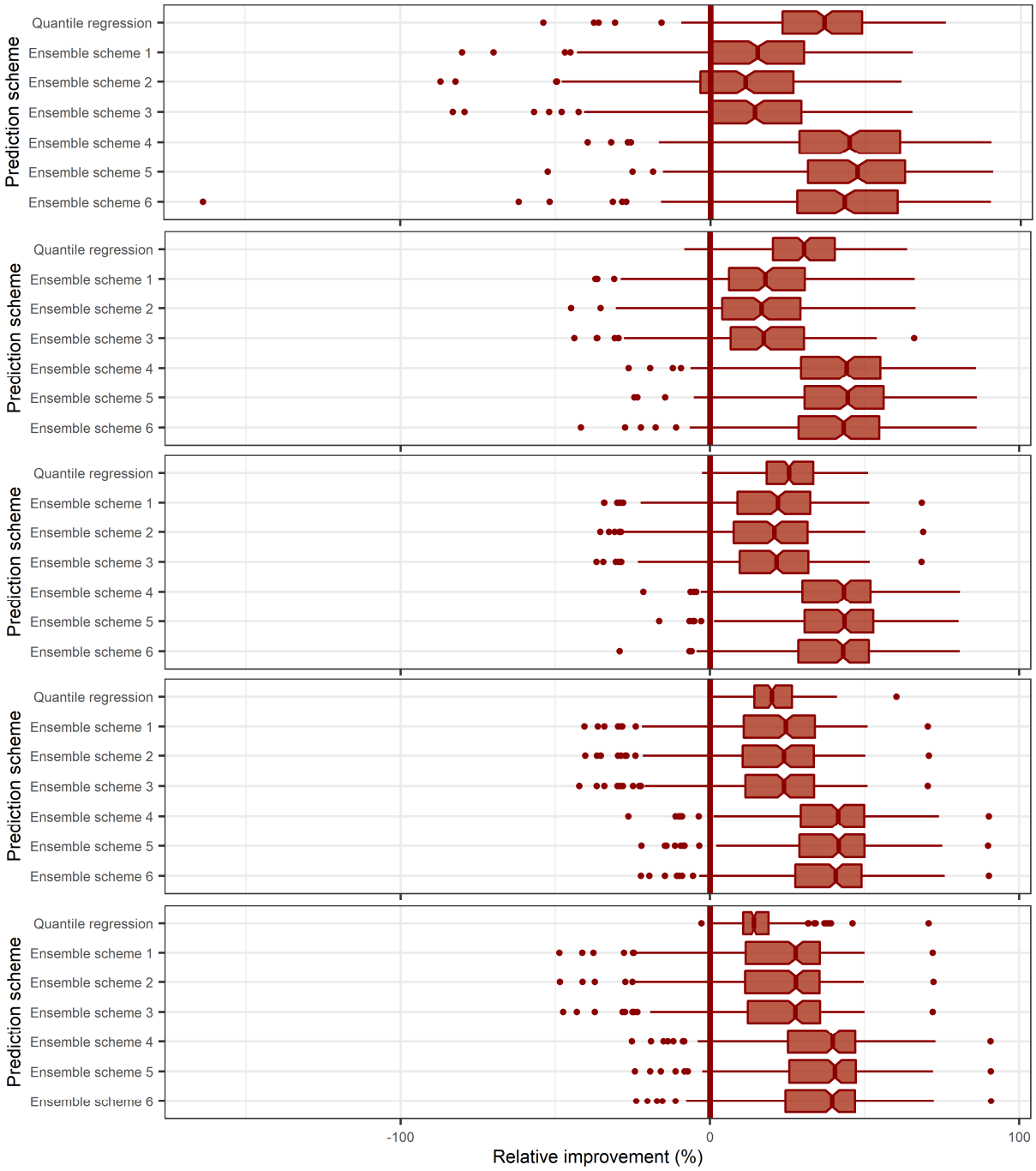


Figure 8.9. Relative improvements in terms of average interval score with respect to the linear regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values. The reference values (zero values) are denoted with red thick vertical lines.

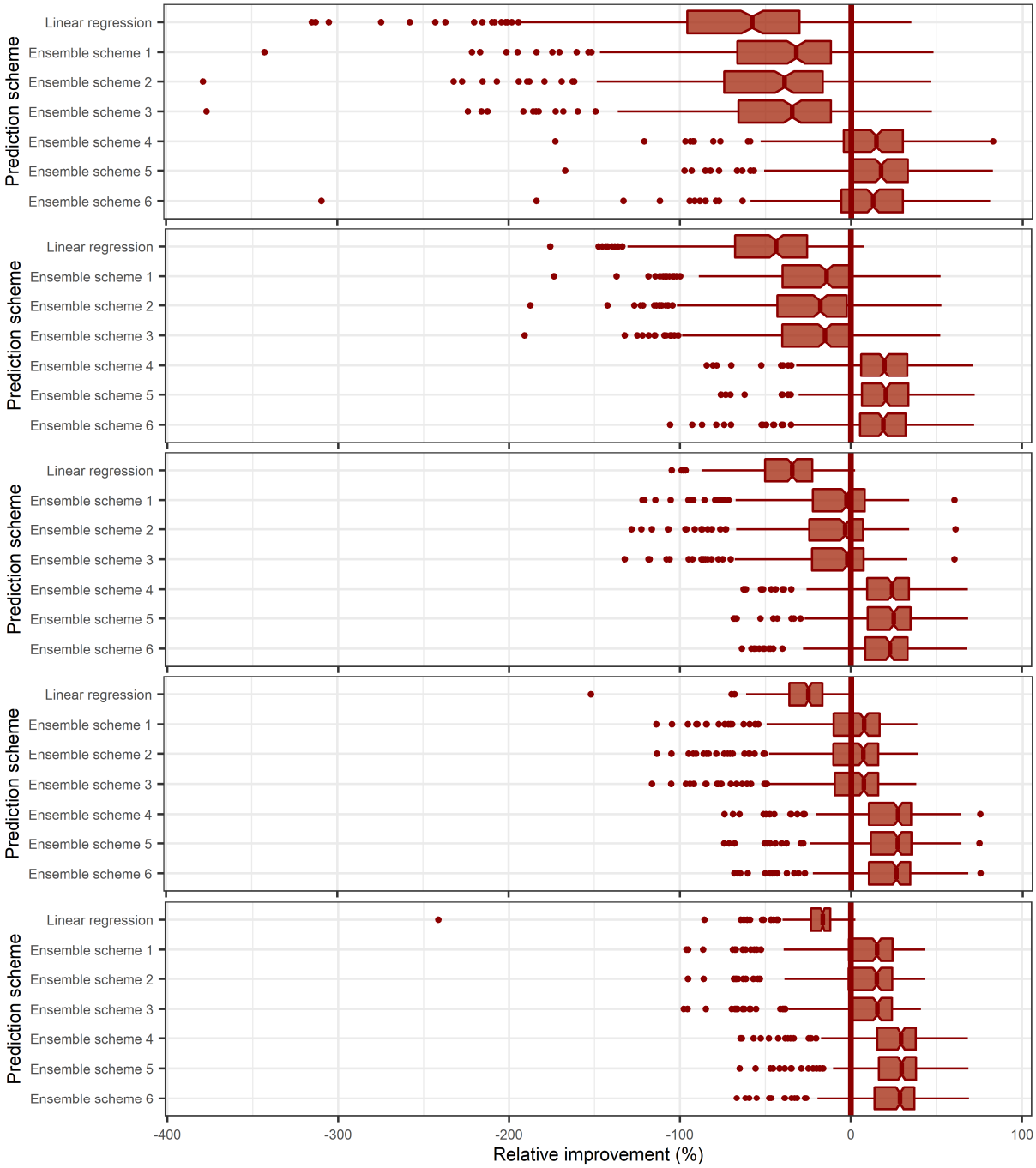


Figure 8.10. Relative improvements in terms of average interval score with respect to the quantile regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each boxplot summarizes 270 values. The reference values (zero values) are denoted with red thick vertical lines.

We subsequently provide a numerical summary of the gain in performance when using specific schemes over others, as extracted from the real-world experiment of the Chapter. In Figures 8.11 and 8.12 we present the average-case relative improvements in terms of average interval score with respect to the linear regression and the quantile regression schemes respectively. These two figures objectively summarize the information presented in Figures 8.9 and 8.10, while they are particularly useful in assessing how small the differences between ensemble schemes 1–3, as well as between ensembles schemes 4–6, are; see also Figures S.1 and S.2 of Chapter's supplement for inspecting these differences in terms of median relative

improvements. For the former category of ensemble schemes, we observe that the difference in the average-case improvements is at maximum 3.65%. The latter difference is computed for ensemble schemes 1 and 2 for the 99% prediction intervals, while it is smoothened to 1.94%, 1.07%, 0.48% and 0.13% for the 97.5%, 95%, 90% and 80% prediction intervals respectively. The average relative improvements when using ensemble scheme 1 instead of ensemble scheme 2 are 4.24%, 2.39%, 1.36%, 0.63% and 0.18% for the obtained 99%, 97.5%, 95%, 90% and 80% prediction intervals. The respective median improvements are 3.75%, 2.18%, 1.20%, 0.53% and 0.15%, while the cost in terms of computational time is about 12 min for all 270 catchments. Ensemble scheme 3 offers comparable profit in performance alongside with a 28-minute profit in terms of computational time compared to ensemble scheme 1.

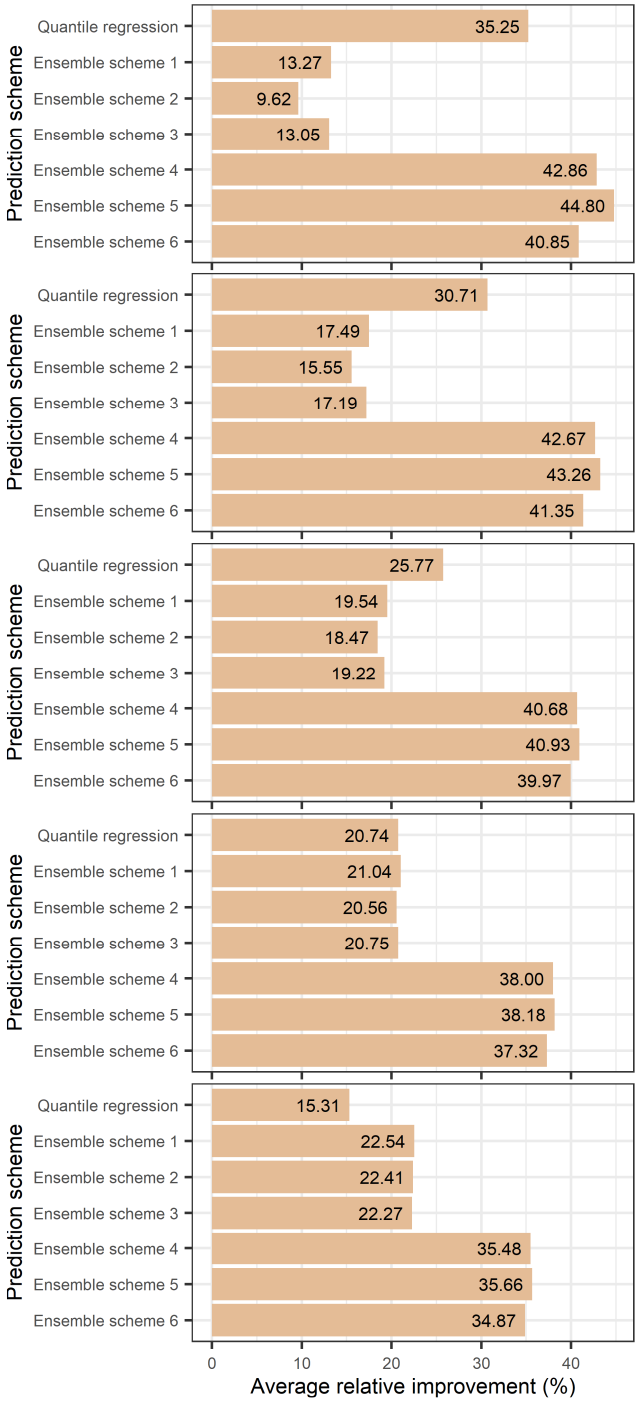


Figure 8.11. Average relative improvements in terms of average interval score with respect to the linear regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each bar summarizes 270 values.

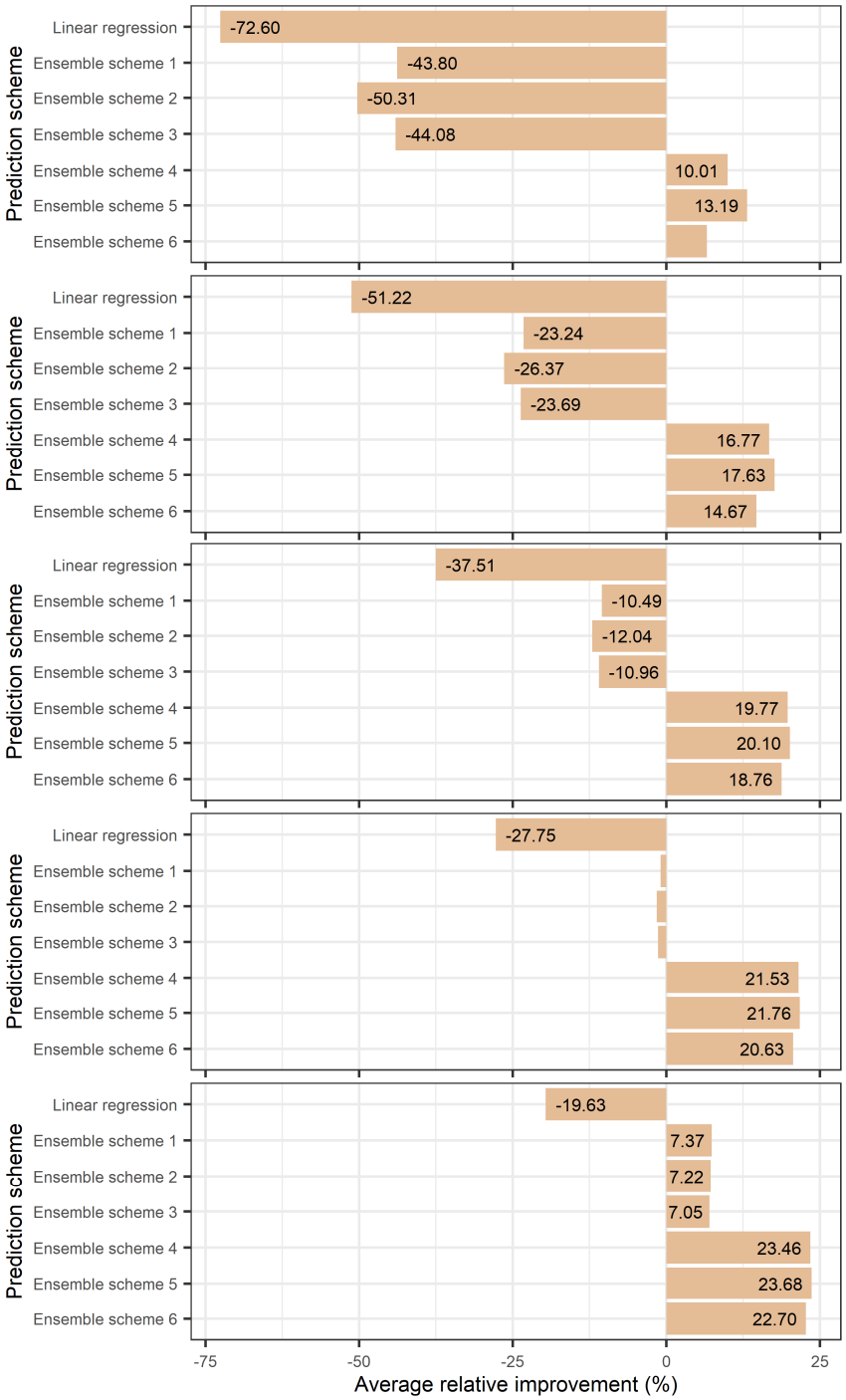


Figure 8.12. Average relative improvements in terms of average interval score with respect to the quantile regression scheme for the 99%, 97.5%, 95%, 90% and 80% prediction intervals (from top to bottom) delivered by the compared schemes for the period T_3 (years 1975–1999). Each bar summarizes 270 values.

Moreover, the mean (median) profit when using ensemble scheme 5 instead of ensemble scheme 4 is found to be 3.09%, 0.99%, 0.48%, 0.34% and 0.25% (2.07%, 0.54%, 0.32%, 0.27% and 0.18%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively, while the concomitant cost in terms of computational time is about 36 min. The respective profit when using ensemble scheme 6 over ensemble scheme 4 is about 12 min. Nonetheless, the use of the latter scheme instead of the former scheme offers an average (median) relative improvement equal to 2.23%, 1.77%, 1.11%, 1.00% and 0.85% (0.31%, 0.47%, 0.24%, 0.28% and 0.31%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively. Moreover, the respective average

(median) relative improvements provided by ensemble scheme 5 with respect to ensemble scheme 6 are 5.46%, 2.74%, 1.60%, 1.36%, 1.10% (3.39%, 1.44%, 0.73%, 0.57%, 0.45%). The gain in performance from the incorporation into the working methodology of the quantile regression model instead of the linear regression model can be summarized by the average-case (median) relative improvements in terms of average interval score provided when using ensemble scheme 5 instead of ensemble scheme 1. These are 37.00%, 31.62%, 26.82%, 22.10% and 17.22% (37.97%, 31.32%, 25.85%, 20.95% and 15.84%) for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively.

8.3.2 Harnessing the wisdom of the crowd in probabilistic hydrological modelling

Two key properties of the working methodology, as identified in Chapter 7 of this thesis based on the seminal work by Lichtendahl et al. (2013, Section 5), are its larger robustness in performance compared to basic two-stage post-processing methodologies and its ability to harness the wisdom of the crowd, both stemming from the concept of prediction averaging. These properties can also be considered as the result of an optimal exploitation of the possibilities offered by the MK blueprint methodology. The demonstration of these properties has only been made so far within toy examples, while it is still pending for rainfall-runoff problems. This section is devoted to empirically proving these two properties of the working methodology using the results of the herein conducted real-world experiment, i.e., to addressing aims 4−5 of the Chapter. These aims are of particular importance in justifying the conceptualization and rationale behind the working methodology.

In Figure 8.13, we present the relative improvements when using the output of ensemble scheme 5, i.e., the average of 600 quantile predictions, instead of separately using each of them (i.e., the relative improvements $\{RI_{OUT,IN_i}, i = 1, \ldots, 600\}$, defined with Equation (8.2), for ensemble scheme 5), computed for all catchments and for all prediction intervals. We observe that these relative improvements are approximately symmetric around zero, in average slightly higher than zero. Specifically, the average relative improvements corresponding to Figure 8.13 are found to be equal to 0.82%, 0.83%, 0.74%, 0.70% and 0.71% for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively (see Table S.1 in Chapter's supplement). The interpretation of this outcome is straightforward, while indicating an advantage in terms of robustness of the working methodology over basic two-stage post-processing methodologies using a single probabilistic prediction.

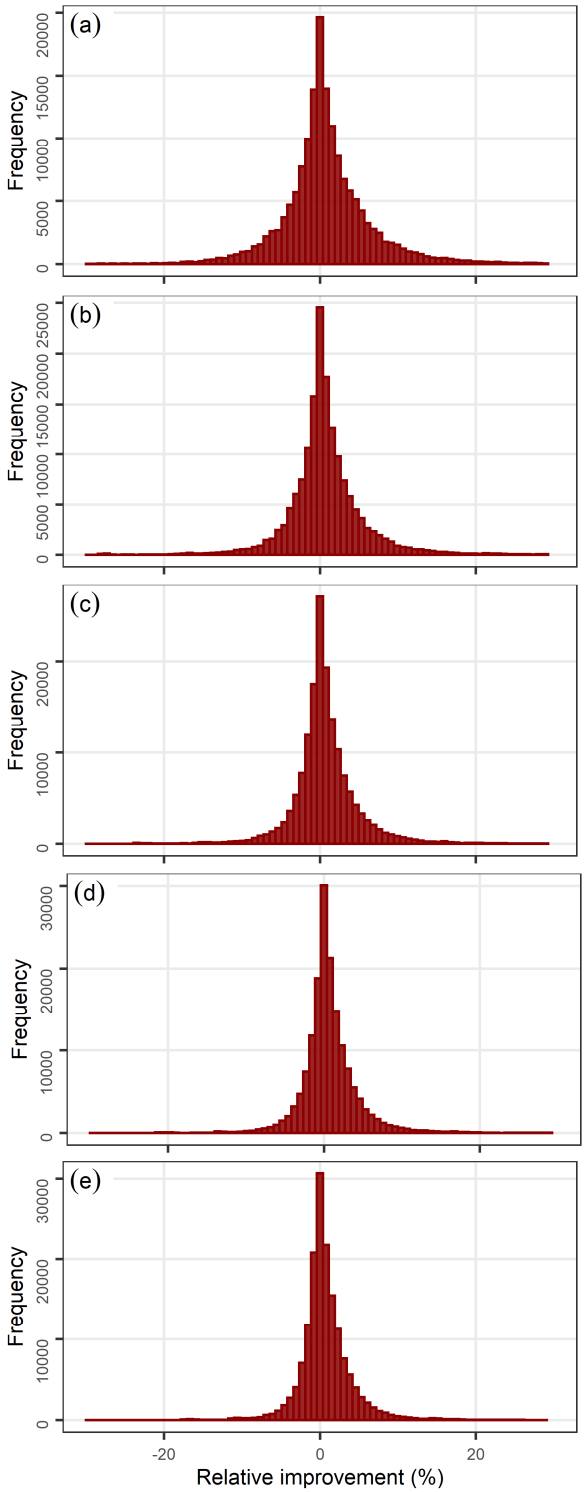


Figure 8.13. Relative improvements in terms of average interval score when using the output of ensemble scheme 5, i.e., the average of 600 probabilistic predictions, instead of each of the combined individual predictions. The relative improvements are computed for all catchments, and for the (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80% prediction intervals obtained for the period T_3 (years 1975–1999). The horizontal axis has been truncated at −30% and 30%. Each histogram summarizes 270 × 600 = 162 000 values.

In fact, while approximately half of the probabilistic predictions score better (or worse) than the finally delivered by the working methodology probabilistic prediction, there is no way to know in advance which hydrological model's parameters will lead in better average interval score in the period T_3. While this lack of knowledge could significantly affect (in terms of performance) the

delivered probabilistic prediction for a basic two-stage post-processing methodology, this effect is largely reduced by the working methodology.

Moreover, by comparing the degree of spread in the five histograms displayed in Figure 13, we also perceive that the degree of the offered stabilization in performance seems to become larger as we move from the inner prediction intervals to the more outer ones. Nevertheless, even for the 80% prediction intervals the provided stabilization is significant.

Furthermore, in Figure 8.14 we present the relative differences between the average interval score of the output of ensemble scheme 5 and the average of the average interval scores of each of the combined (for obtaining this output) individual predictions, the latter used as reference for the former (i.e., the relative differences $RD_{OUT,AAIS_{IN}}$, defined with Equation (8.4), for ensemble scheme 5), computed for all catchments and for all prediction intervals. Importantly, all computed relative differences are positive (or approximately zero) with no exception; therefore, the average of quantile predictions scores no worse than the average score of the combined individual predictions, i.e., the working methodology harnesses the wisdom of the crowd in terms of average interval score when applied for solving monthly rainfall-runoff problems (see also Lichtendahl et al. 2013, Section 5). The average relative differences corresponding to Figure 8.14 are 1.30%, 1.12%, 0.94%, 0.85% and 0.84% for the 99%, 97.5%, 95%, 90% and 80% prediction intervals respectively (see Table S.2 in Chapter's supplement).

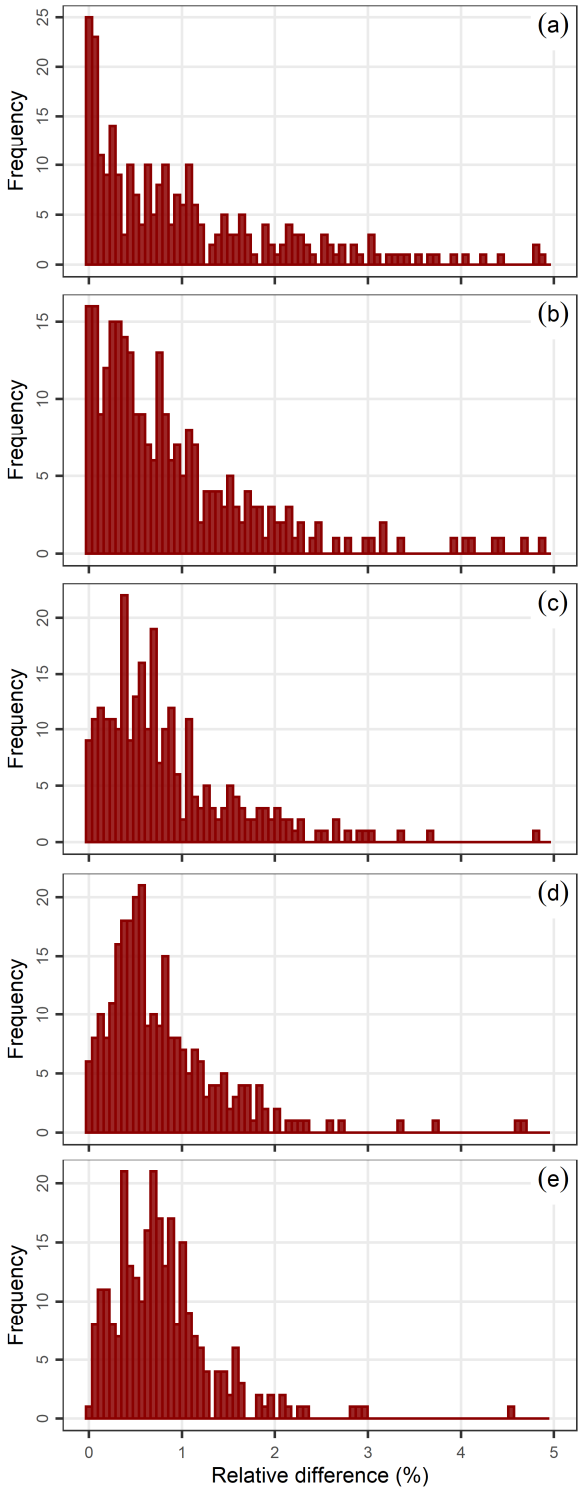


Figure 8.14. Relative differences favouring the average interval score computed for the output of ensemble scheme 5, i.e., the average of 600 probabilistic predictions, over the average of the average interval scores computed for each of the combined individual predictions. The relative differences are computed for all catchments, and for the (a) 99%, (b) 97.5%, (c) 95%, (d) 90% and (e) 80% prediction intervals obtained for the period T_3 (years 1975–1999). The horizontal axis has been truncated at 5%. Each histogram summarizes 270 values.

Analogous observations are extracted from analogous investigations for all remaining ensemble schemes (see Figures S.3–S.12 and Tables S.1–S.2 in Chapter's supplement). In summary, the relative improvements when using the output of an ensemble scheme, i.e., the average of 600 quantile predictions, instead of separately using each of these predictions range

from −327.10% to 91.42%. The average of these relative improvements ranges between 0.13% and 1.13%. Similarly, the average relative differences favouring the average interval score computed for the output of an ensemble scheme over the average of the average interval scores computed for each of the combined (for obtaining this output) individual predictions range between 0.19% and 1.83%. The average relative improvement (difference) is in general larger for the outer prediction intervals than for the inner ones, while its magnitude also depends on the ensemble scheme.

As also emphasized in Chapter 7, the overall trade-off to be considered when someone has to choose between the working methodology and a basic two-stage post-processing methodology allowing the utilization of the same type of flexible error models (see e.g., López López et al. 2014; Dogulu et al. 2015; see also Chapter 9 herein) is the one between (a) the larger robustness in performance offered by the former methodology (demonstrated in Figures 13, S.3, S.5, S.7, S.9 and S.11, and Table S.1 of Chapter's supplement) and the ability of this methodology to harness the wisdom of the crowd (empirically proven based on Figures 8.14, S.4, S.6, S.8, S.10 and S.12, and Table S.2 of Chapter's supplement), and (b) the significantly less computational requirements of the latter methodologies.

8.4 Additional investigations and outcomes

So far, we have validated the working methodology (aim 1 of the Chapter; see Section 8.1) only for the case in which Bayesian schemes are adopted for obtaining a large number of hydrological model's parameters. To investigate the possibility of replacing the Bayesian schemes with informal calibration schemes, in this Section we repeat the large-sample experiment of the Chapter (only for the ensemble schemes) by using different parameter values for the hydrological model within the working methodology. Specifically, for each catchment we retain the first 200 parameter values from each simulated chain (see Section 8.2.6) that have not converged to the posterior distribution of the parameters, instead of the last 200 values that were previously retained (for the application presented in Section 8.3). Hereafter, let us refer to the calibration scheme adopted for obtaining the parameters of the hydrological model in the original large-sample experiment of the Chapter (presented in Section 8.3) and the calibration scheme that is adopted in this appendix as "Bayesian calibration scheme" and "informal calibration scheme" respectively. The remaining components of the ensemble schemes are retained as detailed in Sections 8.2.3–8.2.6.

Once we have obtained the interval predictions, we compute their interval scores and the relative improvements provided in terms of average interval score by the informal calibration scheme with respect to the Bayesian calibration scheme, when both these schemes are exploited as components of ensemble schemes 1–6. The computations are made as detailed in Section 8.2.7, while the related information is presented in Figure 8.15. We mainly observe that (a) the relative improvements can be either positive or negative, and (b) the results favour the Bayesian calibration scheme to some extent, mostly due to outliers. These outliers may become fewer with increasing the length of the period T_2. To objectively summarize the derived information, we also compute the mean and median relative improvements in terms of the same score. These are presented in Figures 8.16 and 8.17, respectively.

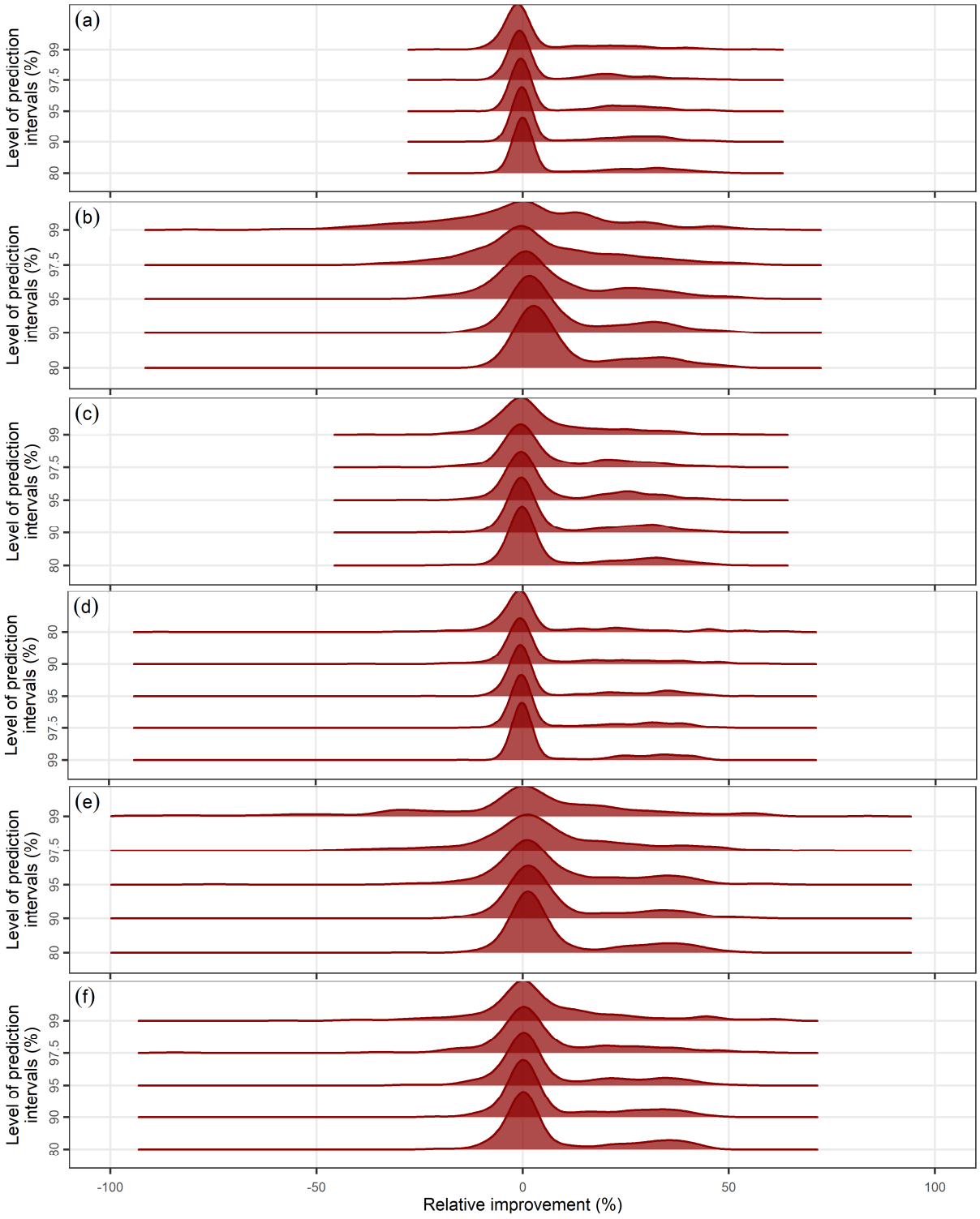


Figure 8.15. Densities of the relative improvements in terms of average interval score provided by the Bayesian calibration scheme with respect to the informal calibration scheme, when both these schemes are used as components of (a–f) ensemble schemes 1–6. The latter are implemented with their remaining components and parameters set common. The horizontal axis has been truncated at −100% and 100%. Each density summarizes 270 values.

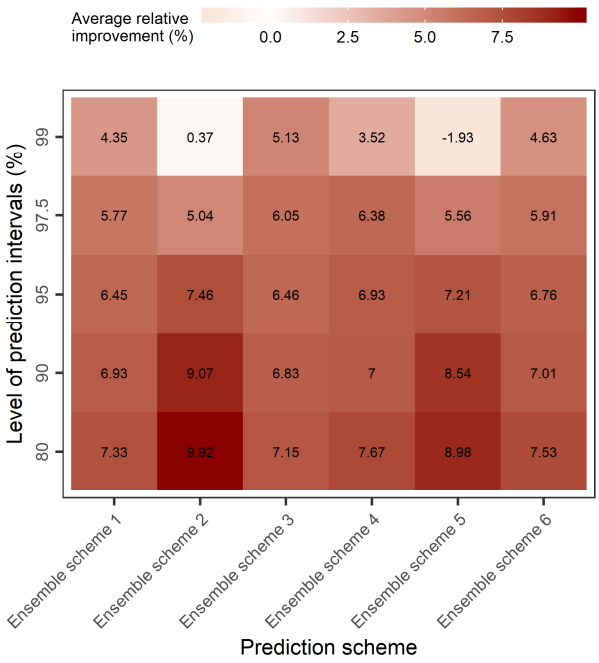


Figure 8.16. Average relative improvements in terms of average interval score provided by the Bayesian calibration scheme with respect to the informal calibration scheme, when both these schemes are used as components of ensemble schemes 1–6. The latter are implemented with their remaining components and parameters set common. The legend limits are common for Figures 8.16 and 8.17. Each presented value summarizes 270 values.

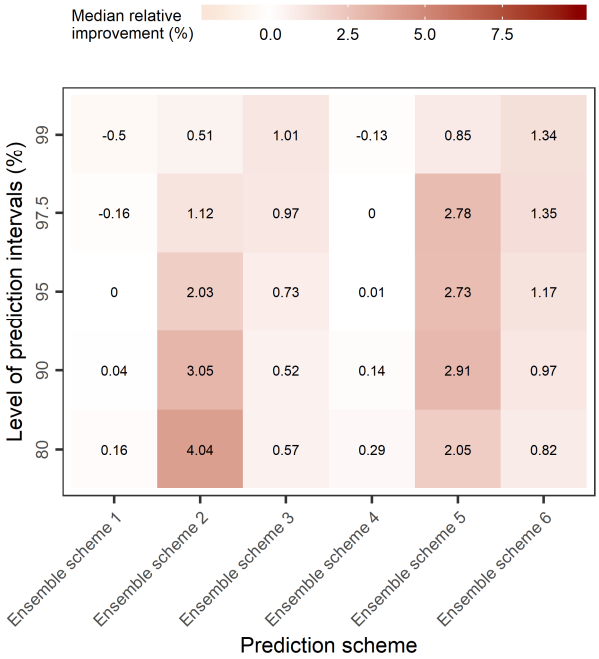


Figure 8.17. Median relative improvements in terms of average interval score provided by the Bayesian calibration scheme with respect to the informal calibration scheme, when both these schemes are used as components of ensemble schemes 1–6. The latter are implemented with their remaining components and parameters set common. The legend limits are common for Figures 8.16 and 8.17. Each presented value summarizes 270 values.

8.5 Concluding remarks

We have validated the probabilistic hydrological modelling methodology proposed in Chapter 7. This methodology adopts key concepts from the ensemble post-processing methodology by Montanari and Koutsoyiannis (2012), while also relying on the concept of probabilistic prediction combination from the forecasting field. It applies a single hydrological model using a large number of different parameter values to generate the same number of "sister predictions". The parameters of the hydrological model can be obtained by using either Bayesian calibration schemes or informal calibration schemes (see the related investigations in Section 8.4). Therefore, this methodology does not have any particular relationship with Bayesian methods by construction, as it also applies to its precursor. A statistical learning (or machine learning) regression model that is suitable for predicting quantiles (see e.g., the models exploited in Chapter 9 of this thesis) is then used to obtain information about the hydrological model's error. This information is used to convert the sister predictions into probabilistic predictions, which are finally combined in simple fashion to obtain the output probabilistic predictions. The assessed methodology is subdivided into three alternative variants, which differ only in the training of the regression model.

We have conducted a large-sample real-world experiment at monthly timescale, set up using complete 50-year daily information for 270 catchments in the United States. Aiming to increase the understanding in probabilistic hydrological modelling, we have insisted on interpretability and benchmarking within all conducted tests. We have used the parsimonious GR2M hydrological model and two (largely) interpretable regression models, specifically the linear regression and the quantile regression ones, to implement six ensemble schemes, all of them based on the assessed methodology. Those ensemble schemes implemented using the linear model (three in number) have been used as benchmarks for the remaining schemes (also three in number). Those ensemble schemes using the same regression model rely on different variants of the assessed methodology. The performance of the ensemble schemes has been assessed by computing the coverage probabilities, average widths and average interval scores of the obtained interval predictions, and by also benchmarking their results using naïve probabilistic data-driven models.

The obtained numerical results (metric values computed for 4 870 800 interval predictions) suggest the usefulness of the assessed methodology in obtaining probabilistic predictions of hydrological quantities. The best-performing variant, offering a mean relative improvement up to 5.46% with respect to its alternative variants, when implemented using the quantile regression model, is variant 2. This variant trains the regression model on a single large dataset formed by using information from all sister predictions. The average-case relevant improvements when using the quantile regression model instead of the linear regression one range up to about 37% in terms of average interval score. This latter numerical result should be appraised on the basis that only the former of these models can model heteroscedasticity. The homoscedasticity assumption is often made in the literature when modelling the hydrological model's error.

Finally, we have demonstrated the increased robustness of the assessed methodology with respect to the combined (by this methodology) individual predictors and, by extension, to basic two-stage post-processing methodologies. The ability to "harness the wisdom of the crowd" has also been empirically proven. The quantile predictions obtained by all ensemble predictors are found to score no worse –usually better– than the average of the individual scores of the combined individual predictions in terms of average interval score. This outcome is in line with demonstrations for stylized cases by Lichtendahl et al. (2013). The computed relative differences favour the former quantity over the latter up to about 37%, while their mean values range between 0.19% and 1.83%, depending both on the prediction interval and the variant of the assessed methodology. For the best-performing ensemble scheme the respective average relative differences are around 1%. Overall, the robustness and the ability to harness the wisdom of the crowd are identified as two key properties of the working methodology.

8.6 Suggestions for future research

We have extensively explored through benchmark tests the modelling possibilities provided by the working methodology, when this methodology is applied for solving monthly rainfall-runoff problems using the quantile regression model as error model. Our benchmark experiment is of large-scale; nevertheless, it could not highlight all aspects of the working methodology. For exploiting this methodology in an optimal way, the following key adjustments to its components and parameters could be made:

- The historical dataset can be divided in various ways, i.e., different proportions of the available information could be devoted to hydrological model calibration and error model training. This adjustment could be made to maximize predictive performance by exploiting evidence extracted from properly designed large-sample investigations. It could also be made for reducing the computational requirements, also depending on our choices on the remaining components and parameters. Applications to hundreds of catchments at timescales finer than the monthly one may require achieving a balance between predictive performance and computational requirements (when our computational resources are limited).
- Any hydrological model (e.g., a process-based hydrological model of our preference) can be selected. Predictive performance improvements may be achieved by selecting one hydrological model over another or by adopting multi-model approaches (as proposed in Vrugt 2018, 2019, yet with the interest being in producing and combining quantile predictions instead of PDF predictions), thereby extending the working methodology, as suggested by Montanari and Koutsoyiannis (2012) for the original blueprint. Properly designed large-sample investigations could effectively guide our related choices.
- The parameters of the hydrological model can be obtained by using a large variety of calibration schemes, including informal calibration schemes. (Note that random selection of the parameters, i.e., no period T_1, could also be an option). This point may be particular important for reducing the computational requirements. In Section 8.4, we present large-sample investigations (on the monthly rainfall-runoff data exploited in the Chapter) focusing on the comparison between Bayesian and informal calibration schemes for obtaining a large number of hydrological model parameters within the working methodology.
- The number of sister predictions can be selected based on the available computational resources. Nonetheless, the larger this number the larger the advantage of the methodology in terms of robustness (compared to basic two-stage post-processing methodologies). Properly designed benchmark experiments could also focus on optimizing this parameter of the working methodology (separately for the various timescales).
- Any statistical learning regression model that is suitable for predicting quantiles (e.g., the error models exploited in Chapter 9 herein) can be selected as error model. This point may be particularly important for maximizing predictive performance (see also the key remarks in Section 8.5).
- Any set of predictor variables (e.g., the hydrological model predictions at times t, $t-1$, $t-2$, etc.) can be used in the application of the error model. This point may be important for maximizing predictive performance for timescales finer than the monthly one (see e.g., the findings of Chapter 9 herein).
- All the above adjustments and modelling choices can be made separately for each of the three variants and for each level of prediction interval (or level of predictive quantile).

9. Probabilistic hydrological post-processing at scale: Why and how to apply machine learning quantile regression algorithms

In this Chapter, we conduct a large-scale benchmark experiment aiming to advance the use of machine learning quantile regression algorithms for probabilistic hydrological post-processing "at scale" within operational contexts. The experiment is set up using 34-year-long daily time series of precipitation, temperature, evapotranspiration and streamflow for 511 catchments over the contiguous United States. Point hydrological predictions are obtained using the GR4J hydrological model and exploited as predictor variables within quantile regression settings. Six machine learning quantile regression algorithms and their equal-weight combiner are applied to predict conditional quantiles of the hydrological model errors. The individual algorithms are quantile regression, generalized random forests for quantile regression, generalized random forests for quantile regression emulating quantile regression forests, gradient boosting machine, model-based boosting with linear models as base learners and quantile regression neural networks. The conditional quantiles of the hydrological model errors are transformed to conditional quantiles of daily streamflow, which are finally assessed using proper performance scores and benchmarking. The assessment concerns various levels of predictive quantiles and central prediction intervals, while it is made both independently of the flow magnitude and conditional upon this magnitude. Key aspects of the developed methodological framework are highlighted and practical recommendations are formulated. In technical hydro-meteorological applications, the algorithms should be applied preferably in a way that maximizes the benefits and reduces the risks from their use. This can be achieved by (i) combining algorithms (e.g., by averaging their predictions) and (ii) integrating algorithms within systematic frameworks (i.e., by using the algorithms according to their identified skills), as our large-scale results point out.

9.1 Introduction

Issuing useful hydrological predictions (e.g., river flow predictions) is one of the most important challenges in hydrology. Dealing with this challenge involves answering numerous research questions, but also putting research into practice by exploiting research advancements in operational contexts. This additional consideration introduces some extra requirements for the prediction methodologies, mostly related to their appropriateness for what we call prediction "at scale". Issuing hydrological predictions "at scale" is a major theme in the present Chapter. The term "at scale" is here used according to Taylor and Letham (2018), i.e., to imply several notions of scale, mostly (i) a large number of required predictions, and (ii) a large variety of prediction problems to be solved. The latter are created, e.g., under different climate and catchment conditions.

The present Chapter is primarily founded upon the premise that (operational) hydrological predictions can be most useful when expressed in probabilistic terms (see e.g., Krzysztofowicz 1999, 2001b; Todini 2007; Koutsoyiannis 2010; Montanari and Koutsoyiannis 2012), i.e., in terms of probability distribution function (PDF) (Todini 2007; see also Todini 2004, 2008) or in terms of prediction intervals (or predictive quantiles). Delivering probabilistic hydrological predictions is a relatively new practice (Todini 2004, 2008; Montanari 2011; Montanari and Koutsoyiannis 2012) considering the much longer history of hydrological modelling, comprehensively summarized by Todini (2007). This practice is also referred to in the related literature as "global uncertainty" quantification (see e.g., Montanari 2011) or "predictive uncertainty" quantification (see e.g., Todini 2007), while its technical implications are under consideration and ongoing discussions (see e.g., Krzysztofowicz 2001b; Sivakumar 2008b; Ramos et al. 2010; Montanari 2011; Ramos et al. 2013).

The background of the present Chapter lies in the tremendous and growing progress made in two distinct research fields whose advancements can be exploited in hydrological contexts for predictive modelling (contrasted to explanatory and descriptive modelling in Shmueli 2010). These are the field of "process-based" hydrological modelling (term used here as defined in Montanari and Koutsoyiannis 2012; see e.g., Beven and Kirkby 1979; Todini 1996; Jayawardena

and Zhou 2000; Perrin et al. 2001, 2003; Mouelhi et al. 2006b; Fiseha e al. 2013; Kaleris and Langousis 2017) and the field of machine learning (see e.g., Hastie et al. 2009; Alpaydin 2010; James et al. 2013; Witten et al. 2017). The former includes various modelling approaches spanning from distributed to lumped conceptual approaches, which also aim (besides prediction) at supporting some sort of "physical interpretation" of the catchment-scale hydrological phenomena (Todini 2007), and describing the catchment's behaviour as a whole (Perrin et al. 2003), respectively. Moreover, the machine learning field includes a large variety of multi-purpose algorithmic techniques, potentially useful in various applied fields, such as hydrology. Amongst its latest advancements are ensemble learning methods (e.g., the bagging by Breiman 1996 and random forests by Breiman 2001a), i.e., methods that combine the results of individual learning algorithms to improve predictive performance Sagi and Rokach (2018). Machine learning algorithms are often referred to in the hydrological literature under the more general term "data-driven models".

Process-based hydrological models and data-driven algorithmic approaches are regarded as two different "streams of thought" in predictive hydrological modelling that need to be harmonized "for the sake of hydrology" (Todini 2007). In fact, machine learning techniques can be perceived as manifestations of the algorithmic modelling culture, a statistical modelling culture that is grounded on the premise that the mechanism behind the data generation is completely unknown and, therefore, obtaining predictions by exploiting the data does not require its prior description through an analytical model (Breiman 2001b). This culture fundamentally deviates from what is called "process-based modelling".

Often perceived to represent tradition, experience and lessons-learnt knowledge (from a "physical process-oriented" modeller's point of view; Todini 2007), process-based models are mostly preferred by the hydrological modellers and hydro-meteorological forecasters (Toth et al. 1999). Among the plethora of the currently available process-based hydrological models, few exemplary ones are more trustable than others (e.g., the GR hydrological models by Perrin et al. 2003, Mouelhi et al. 2006b, and others, which are also available in open source by Coron et al. 2017, 2019), as it is evident from the literature that they are the result of decades of continuous and labour-intensive hydrological research focusing on better overall prediction, better prediction of low and high flows, and model parsimony, among others (see e.g., the related comments in Perrin et al. 2003).

On the other hand, "engineering-oriented" modellers report on (unexploited) opportunities for high predictive performance stemming from the use of data-driven hydrological models (Todini 2007). Machine learning regression algorithms are regularly implemented in the data-driven hydrological literature for solving a vast amount of technical problems, and for building confidence in predictive and explanatory modelling (see e.g., Jayawardena and Fernando 1998; Sivakumar et al. 2002; Koutsoyiannis et al. 2008; Sivakumar and Berndtsson 2010; Quilty et al. 2019; Tyralis et al. 2019c; see also Chapters 3, 4 and 6 herein). Yet, their potential has been realized and exploited only to a limited extent, and mostly for obtaining "point" predictions (term used here as opposed to "probabilistic"). Nonetheless, this potential includes the possibility of delivering probabilistic hydrological predictions (including forecasts; see e.g., the relevant practical suggestions for using random forests in water-related applications by Tyralis et al. 2019b), in spite of the widespread misconception existing in the minds of hydrologists that machine learning algorithms are by nature deterministic (i.e., not statistical). Actually, machine learning methods are all statistical (therefore, "machine learning" and "statistical learning" are terms interchangeably used beyond hydrology), while some of them (e.g., the quantile regression ones, on which this Chapter focuses) are ideal for predictive uncertainty quantification.

Advancing the implementation of machine learning regression algorithms by conducting large-sample (and in-depth) hydrological investigations has been gaining prominence recently (see e.g., Tyralis and Papacharalampous 2017; Xu et al. 2018; Tyralis et al 2019a; see also Chapters 3 and 4), perhaps following a more general tendency for embracing large-scale hydrological analyses and model evaluations (see e.g., Mamassis and Koutsoyiannis 1996; Langousis et al. 2016; Papalexiou and Koutsoyiannis 2016; Sivakumar et al. 2019; see also Chapters 5 and 8). The

key significance of such studies in improving the modelling of hydrological phenomena, especially when the modelling is data-driven, has been emphasized by several experts in the field (see e.g. Perrin et al. 2003; Andréassian et al. 2006, 2007, 2009; Gupta et al. 2014).

In the present Chapter, we exploit a large dataset for advancing the use of machine learning algorithms within broader methodological approaches for quantifying the predictive uncertainty in hydrology. The hydrological modelling and hydro-meteorological forecasting literatures include a large variety of such methodologies (see e.g., Beven and Binley 1992; Krzysztofowicz and Kelly 2000; Kavetski et al. 2002; Krzysztofowicz 2002; Montanari and Brath 2004; Kuczera et al. 2006; Montanari and Grossi 2008; Schoups and Vrugt 2010; Montanari and Koutsoyiannis 2012; López López et al. 2014; Dogulu et al. 2015; Bogner et al. 2016, 2017; Hernández-López and Francés 2017; Tyralis et al. 2019a; see also Chapters 7 and 8), reviewed in detail by Montanari (2011) and Li et al. (2017). Deterministic "process-based" hydrological models are usually and preferably a core ingredient of probabilistic approaches of this family. In this context, statistical models are applied to convert the point predictions provided by hydrological models to probabilistic predictions. Such methodologies are hereafter referred to under the term "probabilistic hydrological post-processing" methodologies.

We are explicitly interested in probabilistic hydrological post-processing methodologies whose model parameters are estimated sequentially in more than one stage (hereafter referred to as "multi-stage probabilistic hydrological post-processing methodologies"; see also the relevant background information in Section 2.7) and machine learning quantile regression algorithms, since the former can accommodate the latter naturally and effectively. The effectiveness of this accommodation has already been proven, for example, with the large-scale results of Chapter 8 and those by Tyralis et al. (2019a) for the monthly and daily timescales, respectively. Aiming at combining the advantages from both the above-outlined "streams of thought" in predictive hydrological modelling, these works and a few earlier ones (to the best of our knowledge, those mentioned in Table 9.1) have integrated process-based hydrological models and data-driven algorithmic approaches (spanning from conditional distribution modelling approaches to regression algorithms) within multi-stage probabilistic hydrological post-processing methodologies for predictive uncertainty quantification purposes.

Table 9.1. List of statistical models implemented within multi-stage hydrological post-processing methodologies.

Statistical model	Classification	Works
Meta-Gaussian bivariate distribution model	Parametric; conditional distribution	Montanari and Brath (2004); Montanari and Grossi (2008); Montanari and Koutsoyiannis (2012)
Generalized additive models (GAMLSS)	Parametric; machine learning	Rigby and Stasinopoulos (2005); Yan et al. (2014)
Quantile regression	Non-parametric; machine learning; quantile regression	López López et al. (2014); Dogulu et al. (2015); Tyralis et al. (2019a); Weerts et al. (2011); Chapters 7, 8 herein
Quantile regression forests		Taillardat et al. (2016); Tyralis et al. (2019a)
Quantile regression neural networks		Taylor (2000); Bogner et al. (2016, 2017)

As summarized in Table 9.1, multi-stage (mostly two-stage) probabilistic hydrological post-processing has been implemented both using parametric and non-parametric statistical models. Machine learning quantile regression algorithms do not make assumptions about the probability distribution function (PDF) of the predictand; therefore, they fall into the broader class of non-parametric techniques. Their output is a set of predictive quantiles of selected levels (e.g., the predictive quantiles of levels $\alpha/2$ and $1-\alpha/2$, which form the $(1-\alpha)$ 100% central prediction interval), instead of predictive PDFs of the hydrological processes of interest. While (three) algorithms from this category have already been incorporated into multi-stage probabilistic hydrological post-processing methodologies (mostly for solving technical problems within case studies; see Table 9.1), there is no extensive study focusing on formalizing and framing this incorporation. We aspire to fill this gap by conducting the largest and most systematic assessment of machine learning algorithms for probabilistic post-processing in hydrology.

We aim at answering the following research question: Why and how to apply machine learning quantile regression algorithms for probabilistic hydrological post-processing? As implied by our aim, our contribution in the literature includes the inspection and appraisal of both quantitative and qualitative aspects of the application of the algorithms. Although our benchmark experiment holds a prominent position in this Chapter, the theoretical and practical information on the proposed methodologies and framework, also provided herein, is rather equally important for answering the above-stated research question. Specifically, we:

1) Explore through benchmark tests the modelling possibilities provided by the integration of process-based models and machine learning quantile regression algorithms for probabilistic hydrological modelling. This exploration encompasses the:
   - ✓ comparative assessment of a representative sample set of machine learning quantile regression algorithms in a two-stage probabilistic hydrological post-processing with emphasis on delivering probabilistic predictions "at scale" (an important aspect within operational settings);
   - ✓ identification of the properties of these algorithms, as well as the properties of the broader algorithmic approaches, by investigating their performance in delivering predictive quantiles and central prediction intervals of various levels; and
   - ✓ exploration of the performance of these algorithms for different flow magnitudes, i.e., in conditions characterized by different levels (i.e., magnitudes) of predictability.
2) Explore through benchmark tests the modelling possibilities provided by simple quantile averaging. Simple quantile averaging is the simplest way to combine multiple quantile predictions (by averaging them), but also "hard to beat in practice" (Lichtendahl et al. 2013; Winkler 2015).
3) Formulate practical recommendations and technical advice on the implementation of the algorithms for solving the problem of interest (and other problems of technical nature). An important remark to be made is that these recommendations are not meant in any case to be limited to selecting a single algorithm for all tasks and under all conditions. Each algorithm has its strengths and limitations, which have to be identified so that it finds its place within a broader framework (provided that the algorithm is a good fit for solving the problem of interest). This point of view is in accordance with the ''no free lunch theorem'' by Wolpert (1996).
4) Justify and interpret key aspects of the developed methodological framework and its high appropriateness for progressing our understanding on how machine learning quantile regression algorithms should be used to maximize benefits and minimize risks from their implementation.

The algorithms assessed herein can be accommodated by ensemble learning probabilistic hydrological post-processing methodologies, e.g., the methodology of Chapters 7 and 8 (built on the work by Montanari and Koutsoyiannis 2012), and the one by Tyralis et al. (2019a). Ensemble learning methods, i.e., methods combining the predictions obtained by multiple learning algorithms (e.g., the equal-weight combiner tested herein), are increasingly adopted in many engineering and applied science fields, since they frequently provide improved predictive performance with respect to each of the individual learning algorithms (see e.g., the review by Sagi and Rokach 2018). The results of the present Chapter also advocate the value of ensemble learning for probabilistic hydrological post-processing.

9.2 Two-stage hydrological post-processing methodology

We adopt a typical two-stage probabilistic hydrological post-processing methodology (see Section 2.7.2) using a single machine learning quantile regression algorithm for modelling the hydrological model errors, as summarized with Figure 9.1. This methodology is flexible and can be used with various machine learning quantile regression algorithms and predictior variables. The predictions provided by multiple machine learning quantile regression algorithms can be

further combined, for example, via simple quantile averaging through an equal-weight combiner (see Section 2.8.1). Simple quantile averaging is exclusively performed on quantiles of the same level.

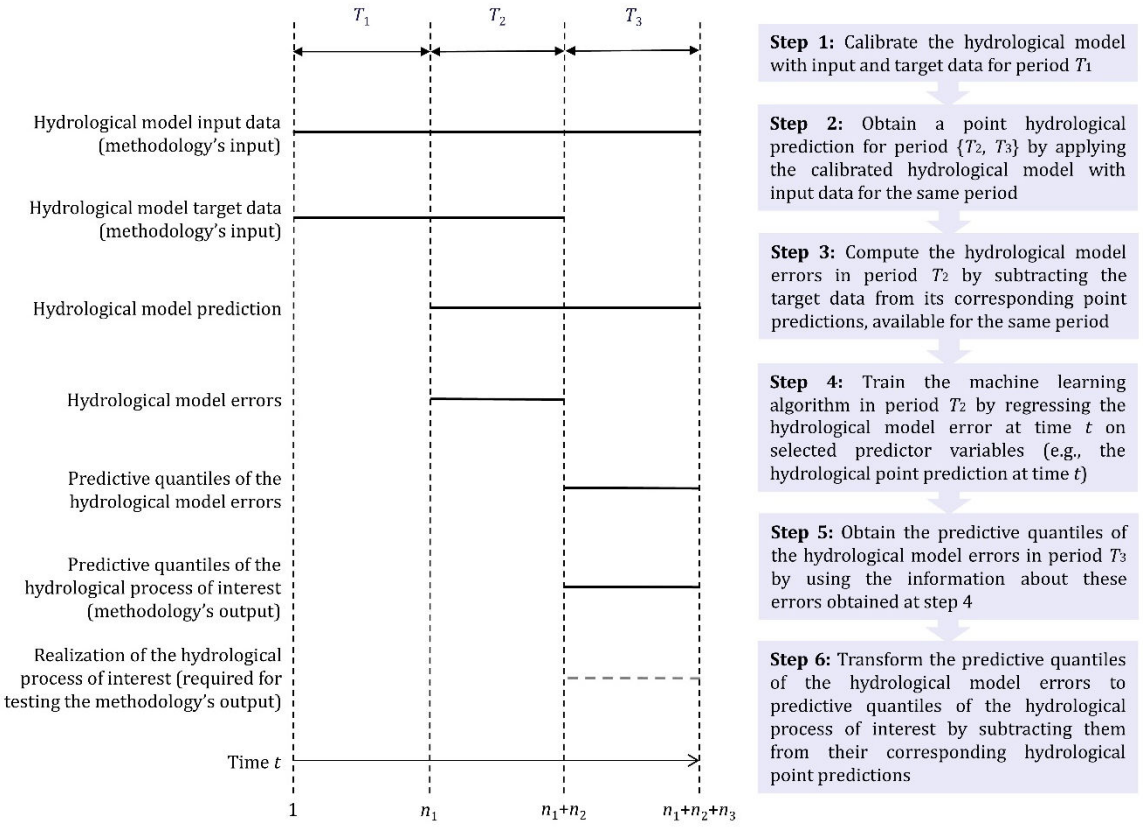


Figure 9.1. Schematic summarizing a typical two-stage probabilistic hydrological post-processing methodology using a single machine learning quantile regression algorithm for modelling the hydrological model errors. The latter are defined as the deviations of the target values from the point predictions provided by the hydrological model.

9.3 Experimental data and methodology

In this Section, we present the experimental data and methodology adopted in the Chapter. Statistical software information is independently provided in Section 2.9.4.

9.3.1 Rainfall-runoff data and time periods

We use data originating from 511 catchments in the contiguous United States. The locations of the stations are presented in Figure 9.2. These catchments are minimally affected by human activities. The data are sourced from the Catchment Attributes and MEteorology for Large sample Studies (CAMELS) dataset (Newman et al. 2014; Addor et al. 2017a), which is fully documented in Newman et al. (2015) and Addor et al. (2017b). The dataset includes complete daily precipitation, temperature and streamflow information over a 34-year period of 1980–2013. Daily precipitation and temperature data were originally made available by Thornton et al. (2014). We estimate daily potential evapotranspiration using the Oudin's formula (see Section 2.4.3).

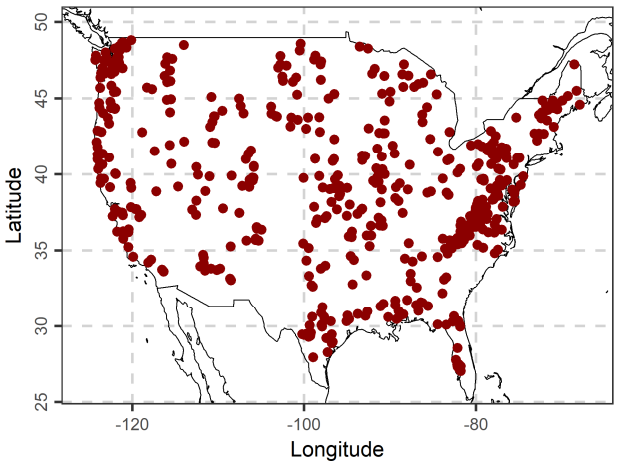


Figure 9.2. Locations of the 511 CAMELS catchments examined in the Chapter. The data are sourced from Newman et al. (2014) and Addor et al. (2017a).

We divide the entire 34-year time period T = {1980-01-01, ..., 2013-12-31} into sub-periods T_0 = {1980-01-01, ..., 1980-12-31} (1-year period), T_1 = {1981-01-01, ..., 1991-12-31} (11-year period), T_2 = {1992-01-01, ..., 2002-12-31} (11-year period) and T_3 = {2003-01-01, ..., 2013-12-31} (11-year period). We use data from these sub-periods as detailed in Sections 3.2–3.4 (see also Section 2.1).

9.3.2 Implemented hydrological model

We implement the GR4J model (see its brief description in Section 2.4.2). We note that implementation of this hydrological model is auxiliary herein. Specifically, this model is used to form the regression problem solved by the machine learning algorithms, as explained in Section 9.2. Therefore, while possible, implementation of other hydrological models is out of the scope of the Chapter.

9.3.3 Assessed and combined machine learning algorithms

The assessed machine learning quantile regression algorithms are listed in Table 9.2 together with their abbreviations. To ensure the reproducibility of these algorithms, in Tables 9.3 and 9.4, we present detailed information on their implementation herein. The predictand and predictor variables in the regression are defined in Section 9.3.5.

Table 9.2. Machine learning quantile regression algorithms assessed in the Chapter. Their software implementation is detailed in Tables 9.3 and 9.4.

S/n	Corresponding machine learning algorithm from Table 2.3	Abbreviation	Description
1	Quantile regression	qr	Section 2.6.2
2	Generalized random forests for quantile regression	qrf	Section 2.6.3
3	Generalized random forests for quantile regression emulating quantile regression forests	qrf_meins	
4	Gradient boosting machine with trees as base learners	gbm	Section 2.6.4
5	Model-based boosting with linear models as base learners	mboost_bols	
6	Quantile regression neural networks	qrnn	Section 2.6.5
7	Equal-weight combiner of the above six algorithms implemented with the same predictor variables	ensemble	Section 2.8

Table 9.3. Details on the implementation of the machine learning quantile regression algorithms (part 1). All R functions are implemented with their arguments set to the default values unless specified differently. The variables of the regression and the levels of the predictive quantiles are defined in Section 9.3.5.

Machine learning algorithm	Training R function	Implementation notes	R package
Quantile regression	`rq`	-	`quantreg`
Generalized random forests for quantile regression	`quantile_forest`	-	`grf`
Generalized random forests for quantile regression emulating quantile regression forests	`quantile_forest`	`(regression.splitting = TRUE)`	`grf`
Gradient boosting machine with trees as base learners	`gbm`	`(distribution = list(name = "quantile", alpha = 0.005), weights = NULL, n.trees = 2000, keep.data = FALSE)`	`gbm`
Model-based boosting with linear models as base learners	`mboost`	`(family = QuantReg(tau = τ, qoffset = τ), baselearner = "bols", control = boost_control(mstop = 2000, risk = "inbag"))`	`mboost`
Quantile regression neural networks	`qrnn.fit`	`(n.hidden = 1, n.trials = 1)`	`qrnn`

Table 9.4. Details on the implementation of the machine learning quantile regression algorithms (part 2). All R functions are implemented with their arguments set to the default values.

Machine learning algorithm	Predicting R function	R package
Quantile regression	`predict`	`quantreg`
Generalized random forests for quantile regression	`predict`	`quantreg`
Generalized random forests for quantile regression emulating quantile regression forests	`predict`	`grf`
Gradient boosting machine with trees as base learners	`predict.gbm`	`gbm`
Model-based boosting with linear models as base learners	`predict`	`mboost`
Quantile regression neural networks	`qrnn.predict`	`qrnn`

9.3.4 Hydrological model application

We apply the selected hydrological model (see Section 9.3.2) to obtain a point prediction of daily streamflow for each catchment through the following steps:

- Data from period T_0 are used to warm up the hydrological model.
- Data from period T_1 are used to calibrate the hydrological model. For the calibration, we implement the Michel's algorithm (see Section 2.4.3) for maximizing the Nash–Sutcliffe efficiency criterion (see Section 2.8.2). The latter is a well-established criterion for hydrological model calibration. While both are possible, implementation of other optimization algorithms and objective functions are out of the scope of the Chapter.
- The calibrated hydrological model is used with daily precipitation and potential evapotranspiration data from period $\{T_2, T_3\}$ to predict daily streamflow for the same period.

9.3.5 Solved regression problem and assessed configurations

The hydrological model predictions for period T_2 are used together with their respective target values to obtain the hydrological model errors for the same period. For each catchment, the hydrological model errors and the hydrological model predictions for period T_2 are used to train each of the assessed machine learning algorithms in predicting the quantiles of level $\tau \in \{0.005, 0.0125, 0.025, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.95, 0.975, 0.9875, 0.995\}$ of the hydrological model errors. The response variable of the regression is the hydrological model error at time t, while the predictor variables are presented in Table 9.5, together with the assessed configurations of the machine learning algorithms that they define. We do not try configurations using a single predictor variable (specifically the hydrological model prediction at time t only),

because some of the machine learning algorithms (e.g., the generalized random forests for quantile regression) do not work without a second predictor. While possible, we also do not try configurations using precipitation and/or evapotranspiration (or temperature) variables, because (i) such variables are already considered by the hydrological model, and (ii) their consideration is less common in the literature than the consideration of hydrological model predictions.

Table 9.5. Configurations of the machine learning quantile regression algorithms assessed in the Chapter. The primal algorithms are presented in Section 2.3.

Abbreviations of assessed configurations	Predictor variables of the regression
qr_2, qrf_2, qrf_meins_2, gbm_2, mboost_bols_2, qrnn_2, ensemble_2	Hydrological model predictions at times $t-1$ and t
qr_3, qrf_3, qrf_meins_3, gbm_3, mboost_bols_3, qrnn_3, ensemble_3	Hydrological model predictions at times $t-2$, $t-1$ and t
qr_4, qrf_4, qrf_meins_4, gbm_4, mboost_bols_4, qrnn_4, ensemble_4	Hydrological model predictions at times $t-3$, $t-2$, $t-1$ and t

9.3.6 Performance assessment

The predictive quantiles of the hydrological model errors are transformed to predictive quantiles of daily streamflow for period T_3 (hereafter referred to as "predictive quantiles of interest") by being subtracted from their corresponding hydrological model predictions. The predictive quantiles of interest are processed using the following subsequent steps: (i) Negative values of predictive quantiles of level 0.005 are censored to zero; and (ii) quantile crossing is handled in an ad hoc manner (if present), i.e., by replacing predictive quantiles of level τ_{k+1} (where k is the sequential number of the quantile levels of interest starting from 1 for quantile level 0.005) with the predictive quantiles of level τ_k delivered by the same algorithm for the same target random variable, if the former predictive quantiles are predicted to be smaller than the latter predictive quantiles.

We assess the quality of the processed predictive quantiles of interest using daily streamflow data for period T_3. The performance assessment is made by computing the scores presented in Table 9.6, as detailed in Section 2.8.2. Note that computing point prediction performance metrics (e.g., the root mean square error; RMSE) is irrelevant to the targeted assessment and, therefore, out of the scope of this Chapter. Nevertheless, the information provided by the average quantile score, when this score is computed for the predictive quantiles of level 0.5, is equivalent to the information provided by the mean absolute error (MAE). For the overall assessment of the algorithms, we compute (a) the four scores (CP_α, AW_α, AIS_α and AQS_τ) conditional upon the algorithm and the catchment; and (b) the relative decreases provided by all algorithms in terms of AW_α, AIS_α and AQS_τ with respect to qr_2 (benchmark). We compute the relative decreases instead of the relative increases, since the former can be interpreted as relative improvements (see Table 9.6). Moreover, for each 34-year-long time series of daily streamflow (i.e., from 511 catchments), we define 100 quantile ranges corresponding to 100 quantile level ranges of equal size, i.e., levels (0, 0.01), [0.01, 0.02), …, [0.99, 1), to also compute the employed scores conditional upon the algorithm, the catchment and the range of observed flow quantiles, and the corresponding relative decreases in terms of AW_α, AIS_α and AQS_τ with respect to qr_2. These latter computations allow us to inspect the performance of the algorithms for different flow magnitudes.

Table 9.6. Scores computed for assessing a prediction interval of level $(1 - \alpha)$, $0 < \alpha < 1$, or a predictive quantile of level τ, $0 < \tau < 1$. The scores are defined in Section 2.8.2 (see also Table 2.6).

Score	Units	Preferred values	Criterion/criteria
Coverage probability (CP_α)	–	Smaller $\lvert CP_\alpha - (1 - \alpha)\rvert$	Reliability
Average width (AW_α)	mm/day	Smaller AW_α	Sharpness
Average interval score (AIS_α)	mm/day	Smaller AIS_α	Reliability, sharpness
Average quantile score (AQS_τ)	mm/day	Smaller AQS_τ	Reliability, sharpness

As stemming from the above-outlined methodological information, the quantile regression algorithm has been selected as the reference algorithm in the experiment. Since this algorithm is linear in parameters (see Section 2.6.2), fast to implement and already exploited in the literature to a significant extent for solving the problem of interest (see Table 9.1), it is a befitting benchmark for non-linear, more computationally demanding and rarely or never-used before (for the problem of interest) algorithms. A last remark to be highlighted concerning the performance assessment is that, while benchmarking is undoubtedly the only available means for characterizing an algorithm as "good enough" in terms of any score, the AW_{α}, AIS_{α} and AQS_{τ} values can only be properly interpreted when presented comparatively (using benchmarking). In fact, the widths of the prediction intervals (and the related components in the interval and quantile scores) largely depend on the flow magnitude, in contrast to the RS_{α} values that are bounded within the range [0, 1].

9.4 Experimental results and interpretations

9.4.1 Overall assessment of the machine learning algorithms

In this Section, we present and discuss summary results of the overall assessment of the machine learning algorithms, when these algorithms are accommodated within two-stage probabilistic hydrological post-processing methodologies. The assessment refers to how well the algorithms deliver various central prediction intervals and predictive quantiles of several levels, while it is here collectively made for all observed flow magnitudes. Some additional visualizations (Figures S.1–S.41), resulted for the same investigations, are presented in Papacharalampous et al. (2019e). In these visualizations, the interested reader can find information about differences in predictive performance from catchment to catchment and related patterns revealed for the machine learning algorithms through the investigations of the Chapter. This information is herein omitted for reasons of brevity.

A comparison of the machine learning algorithms with respect to their average-case reliability (i.e., the average coverage across all catchments) when delivering the 20%, 40%, 60%, 80%, 90%, 95%, 97.5% and 99% central prediction intervals is well supported by Figure 9.3. In Figure 9.3, we present the mean absolute deviations of the coverage probabilities from their nominal values, as computed conditionally on the algorithm and the prediction interval. This figure can be interpreted according to the following example: A mean absolute deviation equal to 0.05 for the 90% prediction intervals means that the absolute deviation of the 511 coverage probabilities (computed for the 511 catchments) from 0.90 (nominal value for the 90% prediction intervals) is on average equal to 0.05. This mean absolute deviation could, for instance, be computed for the case in which the absolute deviations (always positive or zero) are equal to 0.02 for 255 catchments, equal to 0.05 for one catchment and equal to 0.08 for 255 catchments, since $(0.02\times255 + 0.05\times1 + 0.08\times255)/511 = (5.1 + 0.05 + 20.4)/511 = 0.05$. In summary, qr and qrnn are found to mostly perform on average better than the remaining algorithms, while mboost_bols also stands out because of its good average-case performance for the 95%, 97.5% and 99% prediction intervals. With respect to the same criterion, the worst performing algorithm is mostly gbm. For the 60%, 80%, 90%, 95% and 97.5% prediction intervals, all gbm configurations exhibit the smallest average-case reliability. The ensemble learner, i.e., the equal-weight combiner of all the algorithms (when these algorithms are implemented with the same predictor variables), exhibits performance that could be characterized similar or even better (for the 20% prediction intervals) than the performance of the individual algorithms combined. Another remark to be highlighted here is that the mean absolute deviations can be less informative about the quality of the outer prediction intervals (e.g., for the 95%, 97.5%, and 99% prediction intervals). In fact, even an algorithm that always produces prediction intervals from $-\infty$ to $+\infty$, would offer mean absolute deviations equal to 0.05, 0.025 and 0.01 for these intervals, respectively.

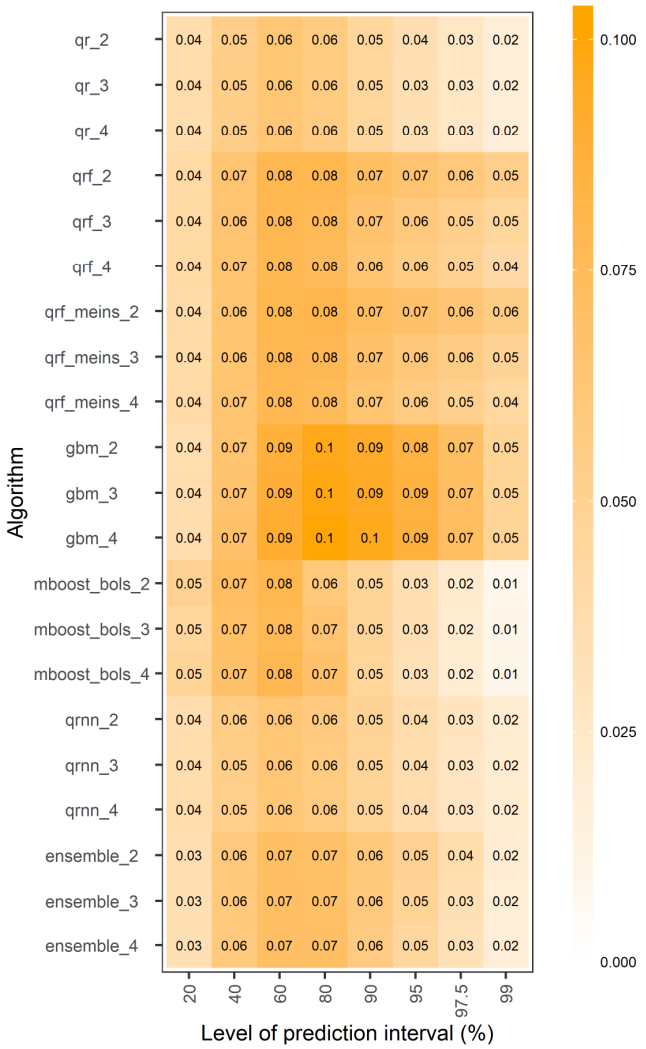


Figure 9.3. Mean absolute deviations of the computed coverage probabilities from their nominal values. The smaller the displayed values, the larger the average-case reliability of the algorithms.

Furthermore, as opposed to the whole picture, only relatively small average-case differences in reliability (differences up to 0.01) are observed across the various configurations of the same algorithms. Larger differences are observed from one algorithm to the other (differences up to 0.05) and for the various prediction intervals of the same algorithm (differences up to 0.06). The interpretation of this observation is straightforward: the two additional predictors do not add as much information as switching from one algorithm to another does, while the predictive performance also largely depends on the prediction task. It is relevant and important to note that, even when we focus on a single criterion (here the average-case reliability), we cannot identify a best performing algorithm for all tasks, i.e., we cannot identify a best performing algorithm in delivering all prediction intervals. For example, if we were only interested in delivering the four outer prediction intervals (i.e., 90%, 95%, 97.5% and 99%), mboost_bols would be the safest choice.

The degree of sharpness characterizing the delivered prediction intervals is also relevant when we are interested in applying the machine learning algorithms for technical purposes. In Figure 9.4, we present the median relative decreases (i.e., the median values of relative decreases computed across all catchments) in terms of average width of the prediction intervals provided by each of the assessed algorithms with respect to qr_2. In more precise terms, these median relative decreases can be interpreted according to the following example: A median relative decrease in terms of average width equal to 9.25%, provided by the gbm_3 algorithm for the 90% prediction intervals, means that the gbm_3 algorithm produces 90% prediction intervals that are,

in the median case across the 511 catchments, narrower than the 90% prediction intervals provided by qr_2 by 9.25%. The median relative decreases are mostly positive, i.e., the algorithms provide narrower prediction intervals compared to the benchmark. Only mboost_bols delivers wider prediction intervals at the 97.5% and 99% prediction levels. Overall, the sharpest prediction intervals are the ones delivered by gbm, followed by those delivered by qrf and qrf_meins. Regarding the behaviour of the various algorithms from a comparative perspective, different patterns characterize the displayed relative decreases for the various algorithms.

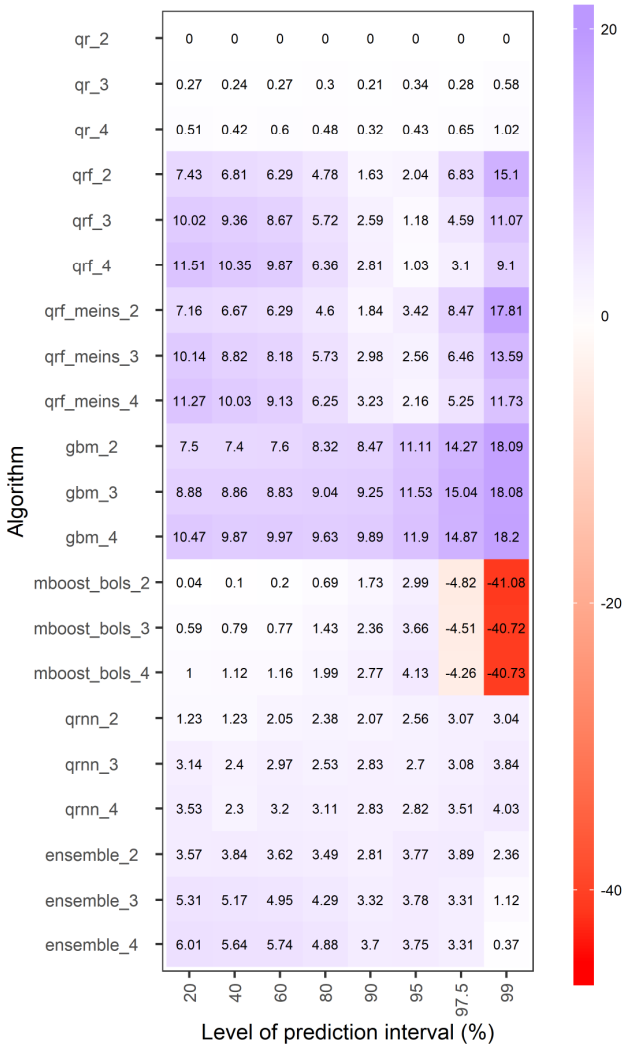


Figure 9.4. Median relative decreases (%) in terms of average width of the prediction intervals with respect to qr_2. The larger the displayed values, the larger the median-case relative sharpness of the delivered prediction intervals.

We should note here again that relatively sharp prediction intervals are only desired when accompanied by a good performance in terms of reliability, and vice versa. Therefore, some interesting observations could be drawn from Figures 9.3 and 9.4. For instance, qrf and qrf_meins seem to exhibit comparable average-case reliability with qr for the 20% prediction intervals, and at the same time to be offering a larger degree of sharpness. Moreover, qrnn and ensemble offer significant median-case decreases in terms of average widths with respect to the benchmark, while they are also quite reliable compared to it. Such observations are important for gaining insight on how the algorithms behave in comparison to one another while solving the problem of interest. Nevertheless, from a practical point of view, we are most interested in collectively assessing reliability and sharpness in an objective manner.

This objective co-assessment with respect to reliability and sharpness is herein allowed by Figures 9.5 and 9.6, which display the median relative decreases (which can be interpreted as median relative improvements) with respect to qr_2 in terms of average interval score and average quantile score, respectively. In more precise terms, these median relative decreases can be interpreted according to the following example: A median relative decrease in terms of average interval score (average quantile score) equal to 1.58% (2.54%) is provided by the ensemble_2 algorithm for the 99% prediction intervals (quantiles of level 0.995). This result means that the ensemble_2 algorithm delivers prediction intervals (predictive quantiles) that are, in the median case across the 511 catchments, better than those delivered by qr_2 by 1.58% (2.54%) in terms of average interval score (average quantile score). The following observations are important:

- More predictor variables result in mostly improved performance for the tree-based methods (qrf, qrf_meins, gbm) and the equal-weight combiner of all algorithms, and slightly less pronounced improvements for qrnn.
- The performance of qr and mboost_bols is found to not be significantly affected by the number of predictor variables.
- The overall best performing algorithm is the equal-weight combiner of all algorithms, offering up to about 3.5% decrease in terms of both average interval and quantile scores with respect to qr_2.
- For all prediction intervals, qr performs mostly better than mboost_bols, while it is also better than gbm for the 60%, 80%, 90%, 95%, 97.5% and 99% prediction intervals. Only for the predictive quantiles of levels 0.4, 0.5, 0.6, 0.7 and 0.8, gbm performs better than qr. Still, gbm is not the best-performing algorithm either for these quantiles.
- For the 90%, 95%, 97.5% and 99% prediction intervals, qr performs better than most of the remaining algorithms, while the equal-weight combiner is the best. The latter offers decreases from about 1.5% to about 2.5% with respect to the former in terms of average interval score, and up to about 3.5% decrease in terms of average quantile score. The equal-weight combiner is worse than qr only for the two lower levels of predictive quantiles tested herein.
- For the 90%, 95%, 97.5% and 99% prediction intervals, the tree-based methods are performing poorly, probably because they cannot extrapolate beyond the observed values of the training set.
- For the predictive quantiles of levels 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8, and the 20%, 40% and 60% prediction intervals, qrf and qrf_meins are comparable with (or even better performing than) the equal-weight combiner of all algorithms.
- For all tested levels of predictive quantiles except for 0.005 and 0.0125, and the 20%, 40%, 60%, 80% and 90% prediction intervals, qrnn perform better than qr.
- Different patterns are observed regarding the performance of the algorithms in predicting the targeted quantiles.
- The performance of qrf and qrf_meins could be characterized as symmetric with respect to the predictive quantile of level 0.5, i.e., these machine learning algorithms show comparably low skill in predicting the upper and lower quantiles that form a specific central prediction interval.
- The same observation does not apply to the remaining machine learning algorithms. Specifically, gbm is less skilful in predicting the lowest quantiles than the highest ones, probably because of the technical settings of the Chapter, i.e., because we predict the quantiles of the error of the hydrological model and later transform these quantiles to quantiles of daily streamflow.
- The same holds for qrnn and the equal-weight combiner, yet these latter algorithms are more skilful, while mboost_bols is less effective in predicting quantiles of the highest levels.

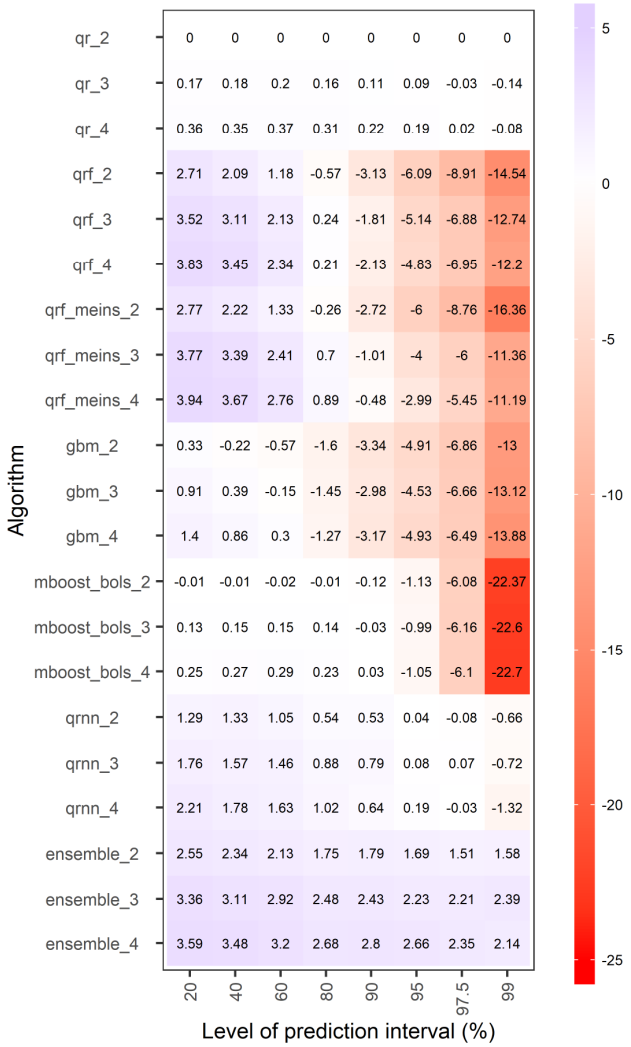


Figure 9.5. Median relative decreases (%) in terms of average interval score with respect to qr_2. The larger the displayed values, the larger the median-case relative skill of the algorithms in delivering the specific prediction intervals.

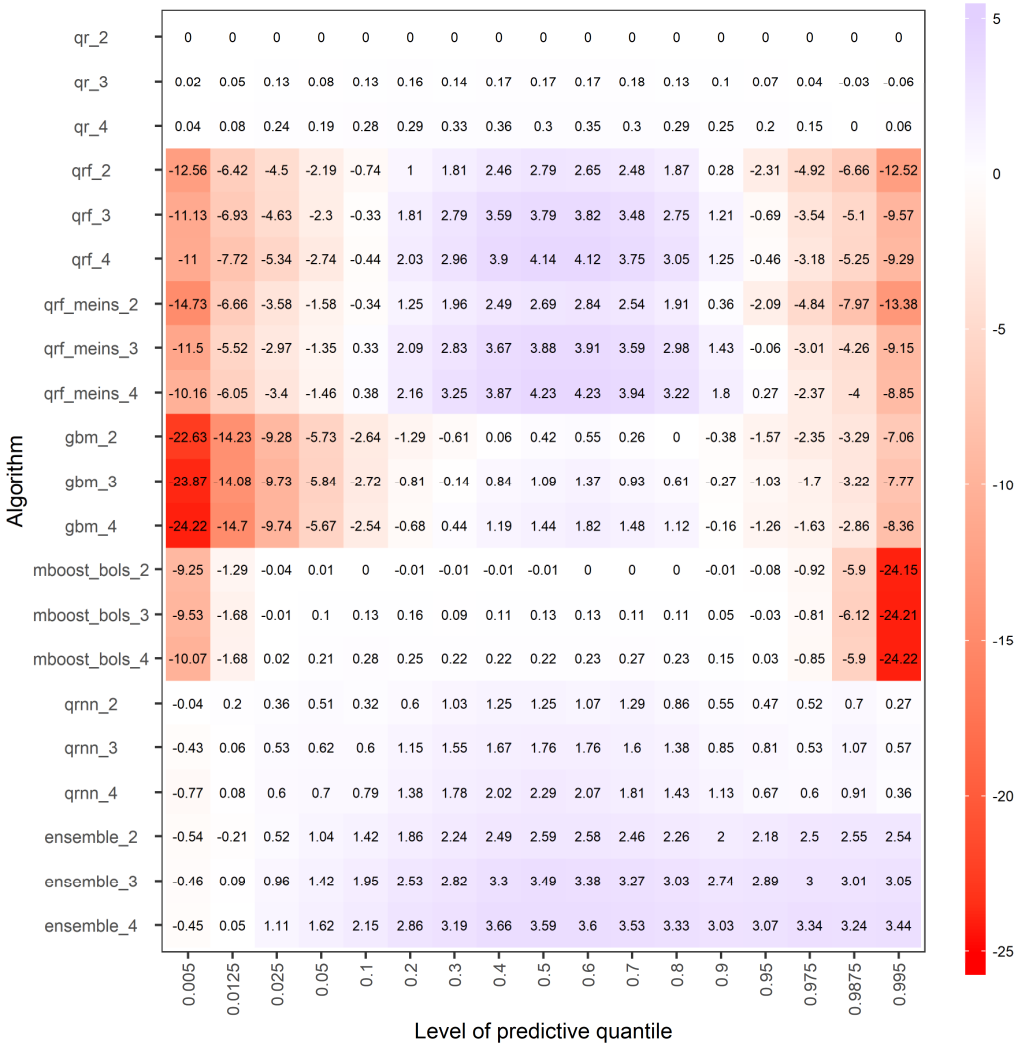


Figure 9.6. Median relative decreases (%) in terms of average quantile score with respect to qr_2. The larger the displayed values, the larger the median-case relative skill of the algorithms in delivering the specific predictive quantiles.

An important remark to be made, at this point, is that the figures presented both herein and in the supplementary material could not highlight all the important details extracted from the conducted tests. Notably, the qrnn algorithms were found to produce significant outliers in terms of predictive performance for 10 of the 511 investigated catchments. These outliers largely affect the respective widths of the prediction intervals provided by these algorithms and, thus, can be easily identified using benchmarking by comparing the widths of the prediction intervals provided by qrnn with the widths of the prediction intervals provided by the benchmark (although the realization of the process of interest will be unknown at the time of the prediction). In fact, they result in relative increases of average widths with respect to the qr algorithms in the order of thousands. Their effect is also manifested in the widths of the prediction intervals provided by the ensemble algorithms (yet in a less-pronounced degree), and in the interval and quantile scores computed for both types of algorithms. The median relative decreases in terms of average widths, average interval score and average quantile score (that are presented herein) are not affected by this limitation of qrnn, while the average relative decreases in terms of average widths, average interval score and average quantile score would be.

Lastly, since we are foremost interested in providing information that could be useful within operational contexts, some tangible information on the computational requirements of the algorithms is also essential. In Figure 9.7, we present the total computational time consumed by each of the assessed machine learning algorithms within the experiments of the Chapter. The least

time-consuming algorithm is by far qr. The remaining algorithms can be ordered from the least to the most time consuming as follows: qrf_meins, qrf, gbm, mboost_bols, qrnn and ensemble. The ensemble algorithm requires more than 10 times the computational time required for qrf to run. Nevertheless, this computational cost may be tolerable in many cases; e.g., when using workstations and/or computer clusters.

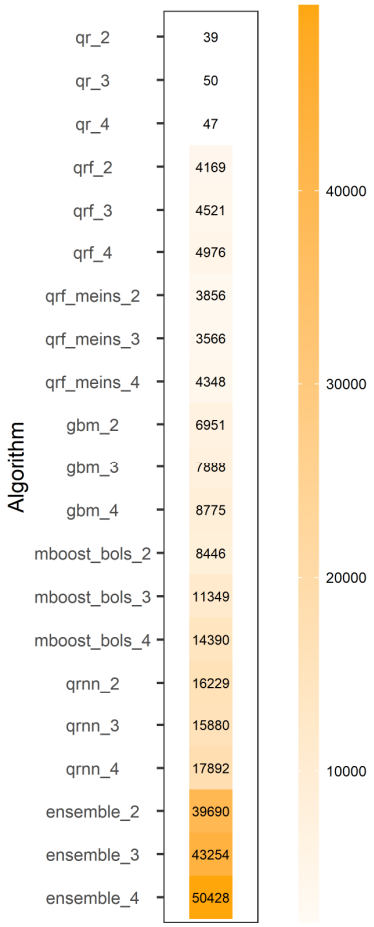


Figure 9.7. Total computational time (in seconds) consumed by the machine learning algorithms within the experiments of the Chapter. The numbers were rounded up to the nearest integer. The computations were performed on a regular personal computer.

9.4.2 Investigations for different flow magnitudes

This section is devoted to summarizing the results of the investigations conducted for different flow magnitudes. These investigations complement the overall assessment of the machine learning algorithms, which is made independently of the flow magnitude, as presented in the preceding section. Due to resolution differences, the results presented in the previous section are not comparable to these in this section.

Figure 9.8 presents the mean absolute deviations of the coverage probabilities from their nominal values, computed per level of observed flow quantile and prediction interval. This information can be exploited to comparatively assess the machine learning algorithms with respect to their average-case reliability for various levels of predictability. In more precise terms, this figure can be interpreted according to the following example: A mean absolute deviation equal to 0.08 for the 20% prediction intervals and the quantile range [0.49, 0.50) means that the absolute deviation of the 511 coverage probabilities computed for the flow magnitude defined by this quantile range from 0.20 (nominal value for the 20% prediction intervals) is on average equal to 0.08. For all prediction intervals, the algorithms are more reliable for the middle half of the sample quantiles of observed flow, while the delivered probabilistic predictions are quite

unreliable for the highest and lowest flows. Regarding this latter point, we also observe that the algorithms are, on average, less reliable for the lowest flows (level of observed flow quantile lower than 0.25) than they are for the highest flows (level of observed flow quantile higher than 0.75), although there is a rough symmetry in the performance of the machine learning algorithms with respect to the observed flow quantiles of levels close to 0.5. This symmetry is perhaps the most characteristic observed pattern, stemming from limitations implied by the nature of the solved problem (in the sense that low and high flows are less predictable than moderate flows).

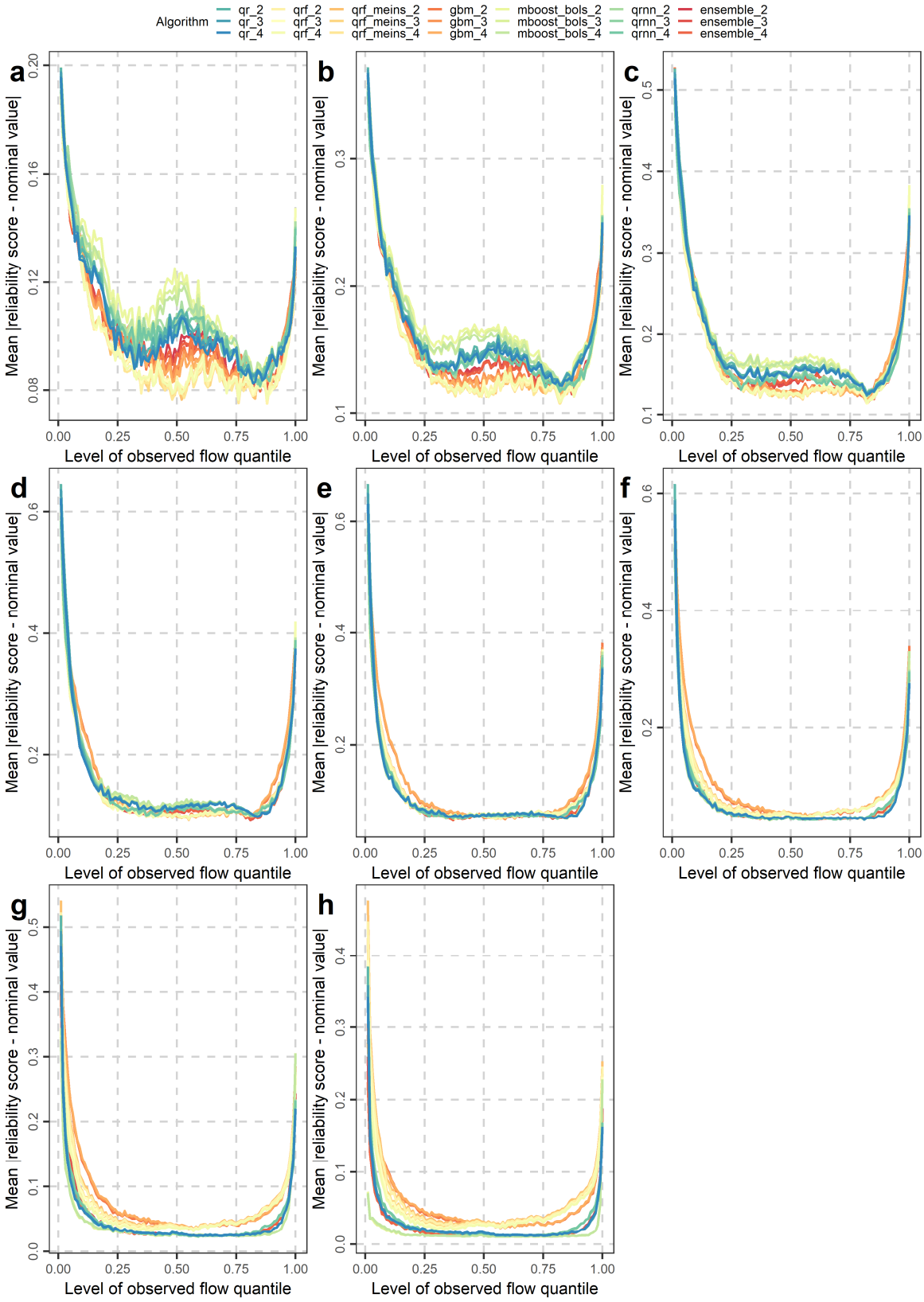


Figure 9.8. Mean absolute deviation of the computed coverage probabilities from their nominal values presented conditional upon the level of observed flow quantile for the (a) 20%, (b) 40%, (c) 60%, (d) 80%, (e) 90%, (f), 95%, (g) 97.5% and (h) 99% prediction intervals delivered by the assessed algorithms.

For the 20%, 40% and 60% prediction intervals and for the middle half of the sample quantiles of observed flow, the qrf, qrf_meins, gbm and ensemble algorithms mostly produce probabilistic predictions that are in better statistical agreement with the observations than qr, while qrnn is mostly comparable to the same algorithm and mboost_bols is the least reliable. For the same prediction intervals and the outer quantiles (level of observed flow quantile lower than 0.25 or larger than 0.75), the differences between the algorithms are slight. For the 80%, 90% and 95% prediction intervals, the performance of all algorithms is mostly similar, with some significant differences being present for the outer quantiles. The algorithms differentiate more for all quantile levels for the 97.5% and 99% prediction intervals.

Moreover, Figure 9.9 presents the median relative decreases in terms of average widths provided by the assessed algorithms with respect to qr_2. This information is presented per level of observed flow quantile and prediction interval and, therefore, it can be exploited to comparatively assess the machine learning algorithms with respect to the median-case sharpness of the delivered prediction intervals for different flow magnitudes. In more precise terms, the presented median relative decreases can be interpreted according to the following example: A median relative decrease in terms of average width equal to ~10%, provided by the gbm_2 algorithm for the 95% prediction intervals and the quantile range [0.49, 0.50), means that, for the flow magnitude defined by this quantile range, the gbm_2 algorithm produces 95% prediction intervals that are, in the median case across the 511 catchments, narrower than the 95% prediction intervals provided by qr_2 by ~10%. In summary, qr produces the wider prediction intervals for all quantiles with some exceptions mostly observed for the lowest and highest flows. Some interesting related patterns should be discussed. The first is related to mboost_bols that produces, on average, much narrower 95% prediction intervals than the benchmark for the lowest half of the observed flows, and 97.5% and 99% prediction intervals for all levels of observed flow quantiles except for the highest (about) 10%. The second pattern is related to the ensemble learner, which is largely affected by mboost_bols for 99% prediction intervals. For the latter and for the lowest 75% of observed flow quantiles, the prediction intervals provided by ensemble are, on average, narrower than those provided by the benchmark, but still much wider than those provided by mboost_bols.

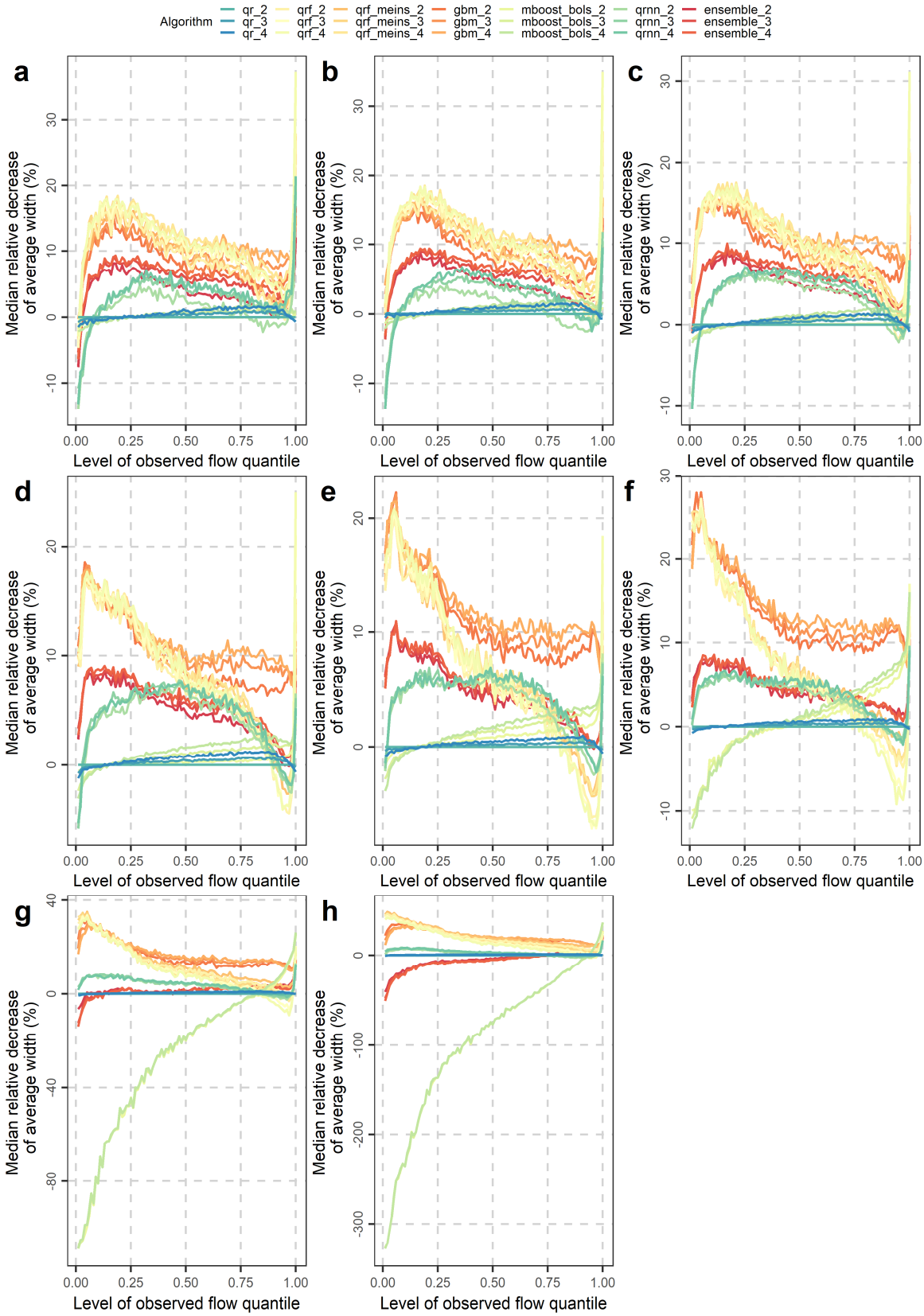


Figure 9.9. Median relative decrease (%) of average widths per level conditional upon the observed flow quantile for the (a) 20%, (b) 40%, (c) 60%, (d) 80%, (e) 90%, (f) 95%, (g) 97.5% and (h) 99% prediction intervals delivered by the assessed algorithms.

To comparatively assess the machine learning algorithms with respect to both reliability and sharpness for different flow magnitudes, we present, in Figure 9.10, their relative improvements in terms of average interval score with respect to qr_2, computed per observed flow quantile and prediction interval. These median relative decreases can be interpreted according to the following example: A median relative decrease in terms of average interval score equal to ~5%, provided by the qrnn_2 algorithm for the 95% prediction intervals and the quantile range [0.49, 0.50), means that for the flow magnitude defined by this quantile range the qrnn_2 algorithm produces 95% prediction intervals that are, in the median case across the 511 catchments, better than the 95% prediction intervals delivered by qr_2 by ~5% in terms of average interval score. For the sample quantiles of observed flow of level (mostly) higher than 0.75, qr is mostly the best performing algorithm, while for the lower half of the sample quantiles of observed flow, qrf and qrf_meins are mostly the best performing algorithms. For the middle half of the sample quantiles of observed flow, qrnn is amongst the best performing algorithms. Moreover, some similar patterns can be observed between Figure 9.9 and Figure 9.10. For instance, mboost_bols delivers 95%, 97.5% and 99% prediction intervals that offer negative median-case decreases (median-case deteriorations) in terms of average interval scores. These decreases follow the respective negative decreases presented in Figure 9.9. Furthermore, qrnn reach their best performance, both in terms of average width and average interval score, for the middle levels of observed flow quantiles, while their performance seems quite symmetric around this highest value for most levels of prediction intervals.

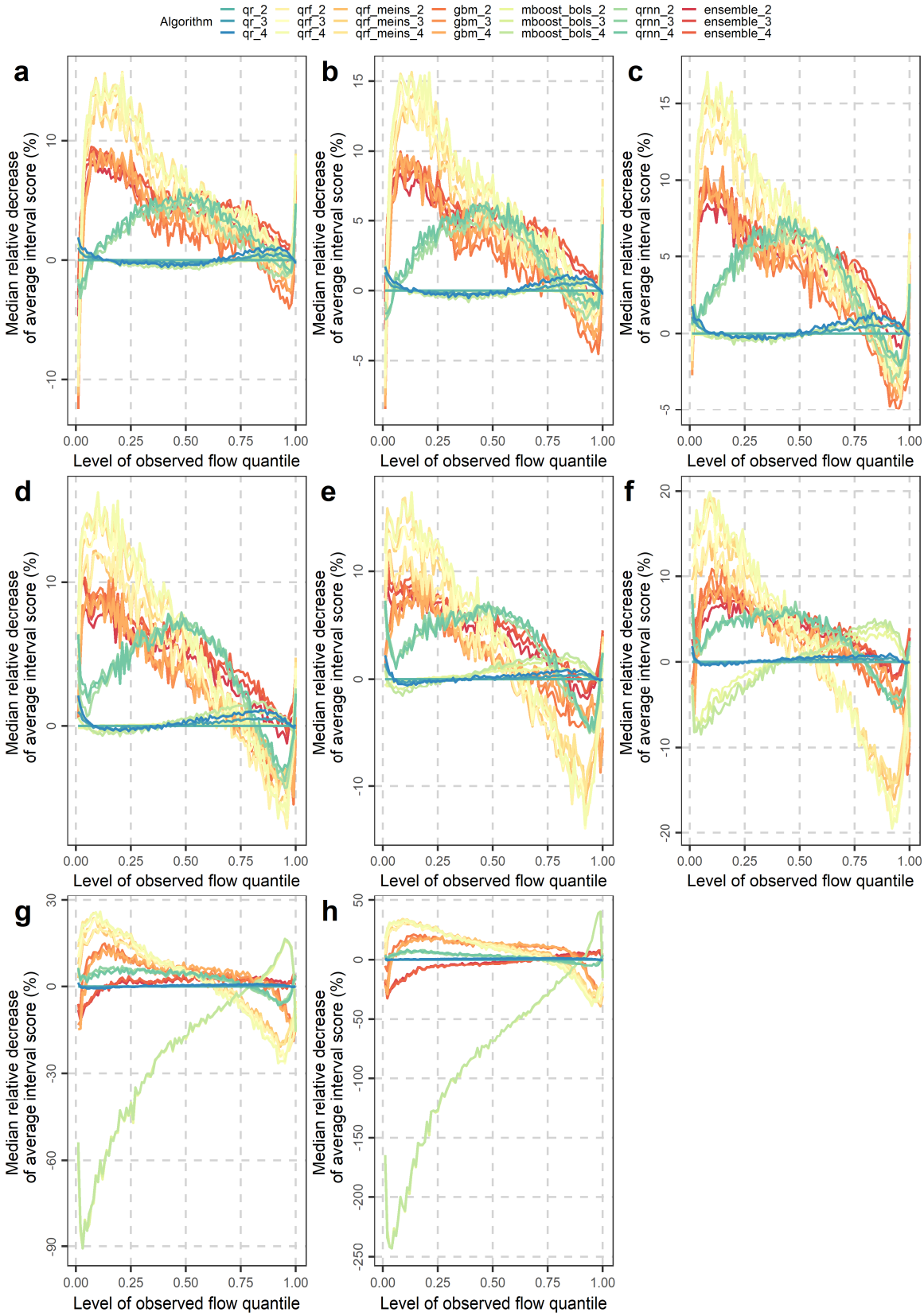


Figure 9.10. Median relative decrease (%) of average interval score conditional upon the level of observed flow quantile for the (a) 20%, (b) 40%, (c) 60%, (d) 80%, (e) 90%, (f) 95%, (g) 97.5% and (h) 99% prediction intervals delivered by the assessed algorithms.

Lastly, in Figures 9.11 and 9.12, we present the relative decreases provided by the machine learning algorithms in terms of average quantile score with respect to qr_2, computed conditional upon the algorithm, the observed flow quantile and the level of predictive quantile. These median relative decreases can be interpreted according to the following example: A median relative decrease in terms of average quantile score equal to ~5%, provided by the qrnn_4 algorithm for the predictive quantiles of level 0.7 and the quantile range [0.49, 0.50), means that for the flow magnitude defined by this quantile range the qrnn_4 algorithm deliver predictive quantiles of level 0.7 that are, in the median case across the 511 catchments, better than the predictive quantiles of level 0.7 delivered by qr_2 by ~5% in terms of average quantile score. As stems from the above, Figures 9.11 and 9.12 can provide tangible information about the skill of the algorithms in delivering a predictive quantile of interest for different flow magnitudes. They can also be used to inspect the contribution of the quality of each predictive quantile in the quality of the central prediction intervals, as well as to assess the machine learning methods in predicting the median of the targeted PDFs per observed flow quantile (see Figure 9.11i). Regarding this latter task, the relative skills of the machine learning methods seem to follow a pattern that is similar to the patterns observed, for instance, for the 40% and 60% prediction intervals (see Figure 9.10b,c).

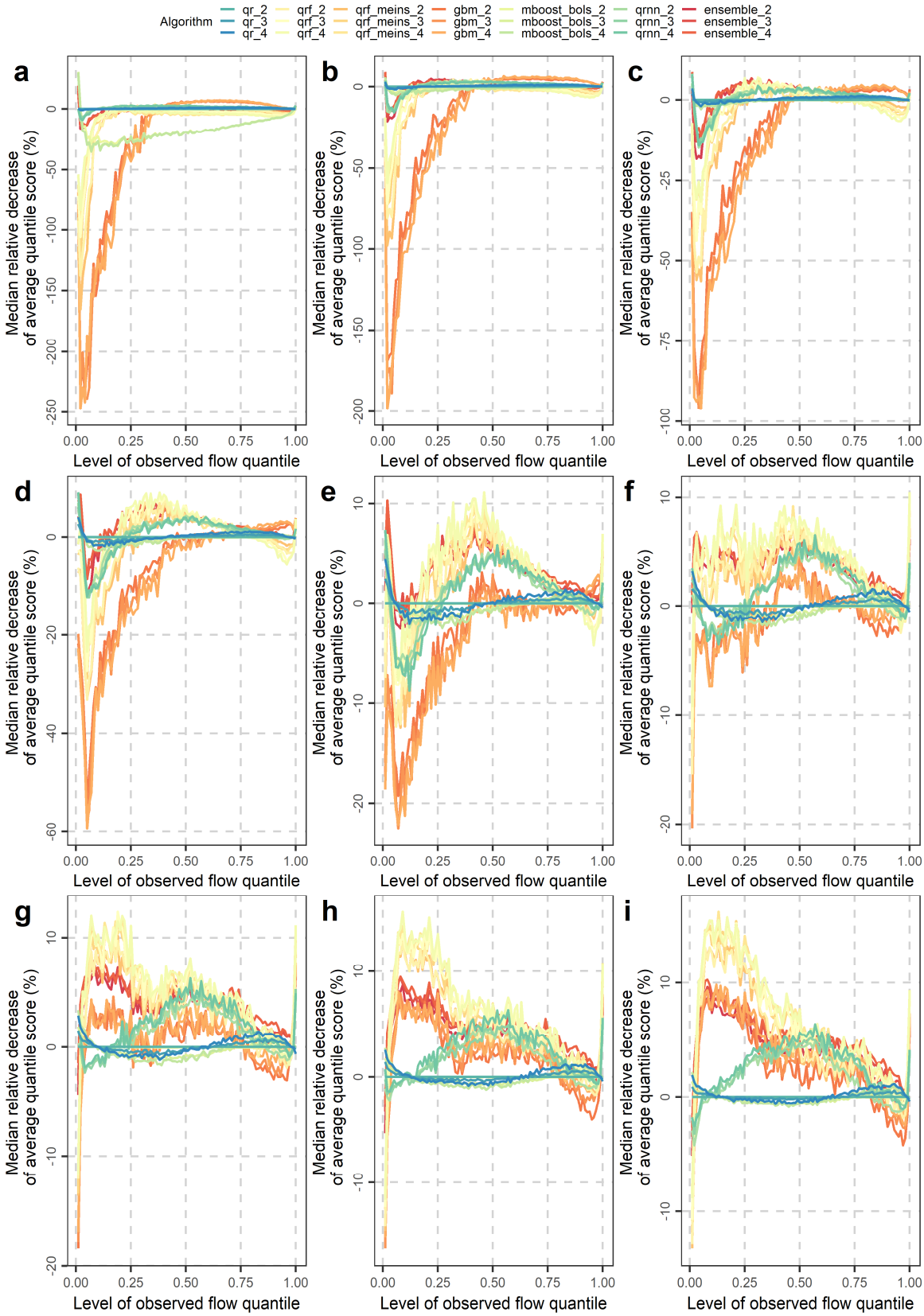


Figure 9.11. Median relative decrease (%) of average quantile score conditional upon the level of observed flow quantile for the predictive quantiles of level (a) 0.005, (b) 0.0125, (c) 0.025, (d) 0.05, (e) 0.1, (f) 0.2, (g) 0.3, (h) 0.4 and (i) 0.5 delivered by the assessed algorithms.

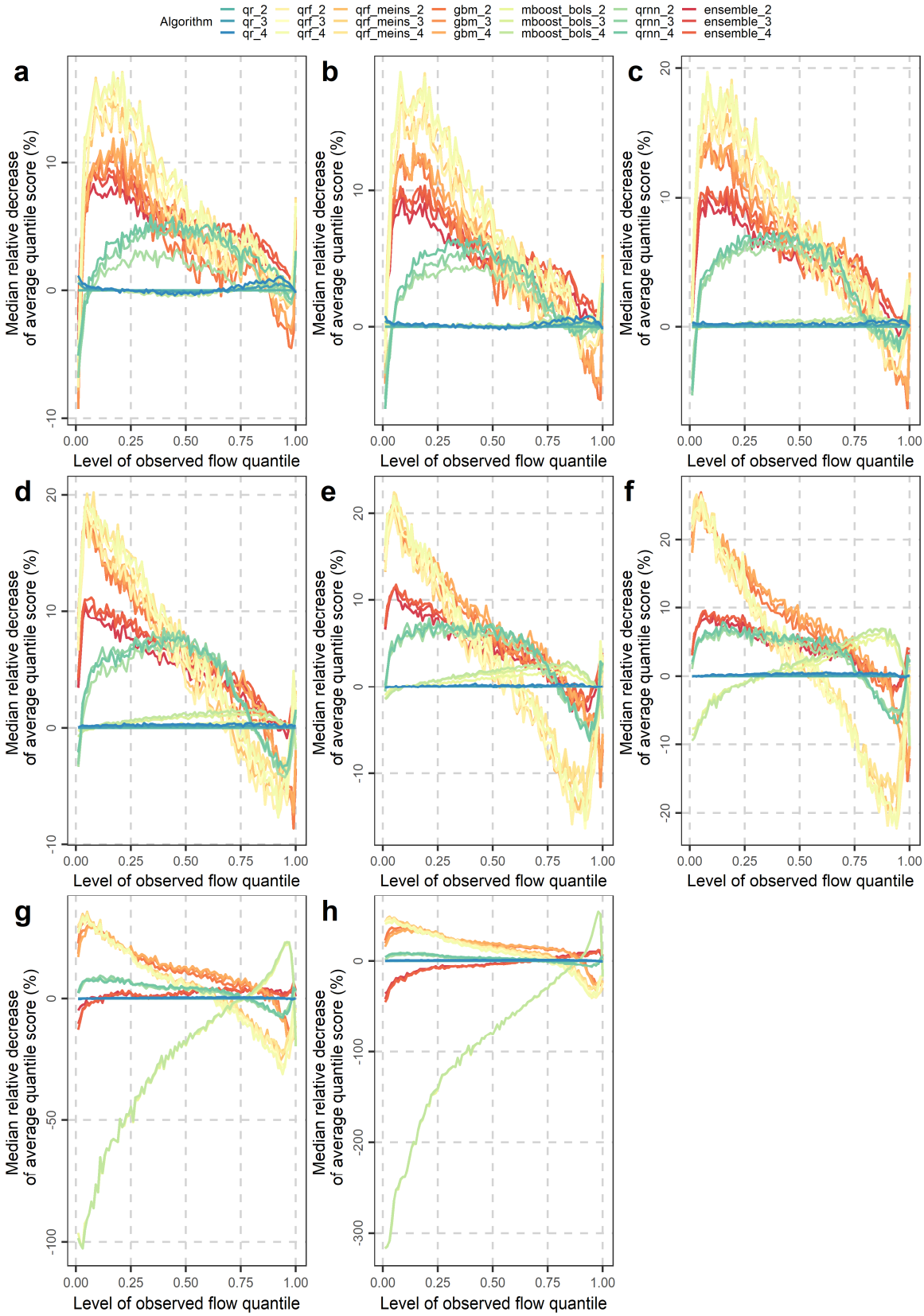


Figure 9.12. Median relative decrease (%) of average quantile score conditional upon the level of observed flow quantile for the predictive quantiles of level (a) 0.6, (b) 0.7, (c) 0.8, (d) 0.9, (e) 0.95, (f) 0.975, (g) 0.9875 and (h) 0.995 delivered by the assessed algorithms.

9.5 Literature-driven and evidence-based discussions

9.5.1 Innovations and highlights in light of the literature

Some key innovations characterizing the present Chapter are the following:

1) It is amongst the very few large-sample works presently available in both the fields of probabilistic hydrological modelling and the field of hydro-meteorological forecasting (see e.g., Tyralis et al. 2019a; Farmer and Vogel 2016; Bock et al. 2018; see also Chapter 8 herein).
2) It includes the largest range of methods ever compared in such concepts and a detailed quantitative assessment, using proper scores, and performing investigations for various prediction intervals and flow magnitudes.
3) Three of the assessed machine learning quantile regression algorithms, specifically generalized regression forests, gradient boosting machine and gradient boosting with linear models as base learners, are implemented for the first time to solve the practical problem of interest.
4) It deviates from the mainstream culture of "model overselling" (Andréassian et al. 2007) or proving that "my model is better than yours" to "justify model development" (Sivakumar 2008a), since it does not aim at promoting the use of any single algorithm. Instead, it formulates practical recommendations, which highlight the need of making the most of all the assessed algorithms (see the related comments in Sivakumar 2008a).
5) It is one of the very few studies that aim at attracting attention to ensemble learning post-processing methodologies in probabilistic hydrological modelling and hydro-meteorological forecasting.

It is important to highlight that most of the above-outlined innovations apply beyond hydrology as well. A large-sample regional study by Bakker et al. (2019), conducted in a different field and under a different approach, has focused on post-processing solar radiation forecasts at hourly timescale for 30 stations in the Netherlands. The study is, in general, of large scale, since it examines two parametric and five non-parametric machine learning algorithms, together with a large number of predictor variables; therefore, it provides generalized results for the case of the Netherlands.

9.5.2 Contributions and challenges from an uncertainty reduction perspective

The challenging character of probabilistic hydrological modelling has been widely acknowledged in the literature (see e.g., Montanari 2007; Sivakumar 2008b; Montanari and Koutsoyiannis 2012). Assumptions are certainly unavoidable when it comes to modelling (Montanari and Koutsoyiannis 2012), and probabilistic predictions are not (and should not be expected to be) perfect (Sivakumar 2008b). What matters the most, from an engineering point of view, is to deliver predictions that are useful. To increase this usefulness (which implies an adequate degree of reliability), one can (i) increase the amount of available information and its quality; and/or (ii) improve its exploitation, i.e., the usefulness of the contributing models, methodologies and frameworks.

These two ways to increase the usefulness of predictions are often collectively referred to under the umbrella term "uncertainty reduction" (or "risk reduction"), while, perhaps, they should be pursued to an extent that is simultaneously feasible, beneficial (e.g., in terms of interval and/or quantile scores that are appropriate for quantifying usefulness) and cost-effective. Point (ii) above can be, in principle, achieved, for example, by (a) reaching a better (physical) understanding of the system to be modelled; (b) (developing or) identifying better models and better predictor variables for each predictive task; (c) developing methodologies that combine different models (and/or algorithms) in an effective manner; and (d) developing unifying frameworks that maximize the benefits from using various methodologies.

By embracing and studying uncertainty, as suggested, for example, in Koutsoyiannis (2010), one can also reduce uncertainty. Uncertainty reduction in (probabilistic) hydrological modelling is one of the 23 major unsolved problems in hydrology identified by Blöschl et al. (2019; see also the related discussions in Montanari 2011, and Montanari and Koutsoyiannis 2012). In this Chapter, we are explicitly interested in contributing towards point (ii), mostly towards points (b–d), conditional on the available data quantity and quality offered by the CAMELS dataset, and the information provided by the GR4J hydrological model. Hydrological understanding is assumed to be encompassed in the latter, under the justification provided in the following subsection. We believe that the investigations conducted herein and the proposed methodological framework should be accounted as a tangible step towards a new era in (operational) probabilistic hydrological modelling and forecasting.

9.5.3 A culture-integrating approach to probabilistic hydrological modelling

By seeing opportunities (rather than threats) in the integration of process-based and data-driven models within multi-stage probabilistic hydrological post-processing methodologies, new fruitful avenues could open up in the field of hydrological modelling. In the following, we discuss some key benefits stemming from this integration, as understood from an uncertainty reduction point of view. We also discuss the practical advantages exploited by this integrating approach, well-supported by their large-scale application made herein.

Hydrological research has been focusing for decades on uncertainty reduction in point hydrological modelling (Montanari 2011). All the related knowledge and experience gained through the years until today has been encompassed in what is called process-based hydrological modelling (Todini 2007; quoting Krzysztofowicz 1999; see e.g., the review by Efstratiadis and Koutsoyiannis 2010). By incorporating process-based hydrological models into probabilistic hydrological post-processing methodologies, we benefit from this experience (therefore, uncertainty is reduced to some extent) and simultaneously quantify predictive hydrological uncertainty. Moreover, we facilitate the straightforward incorporation and exploitation of any future advancement in the field of process-based hydrological modelling, embedded either within new distributed/lumped hydrological models or within frameworks dedicated to boosting the application of such models, as soon as this advancement is achieved (see the related comments in Montanari and Koutsoyiannis 2012).

To further reduce uncertainty, one has to optimize the statistical modelling part of the probabilistic methodology, which is commonly related to the modelling of the hydrological model errors (see e.g., Montanari and Brath 2004; Montanari and Grossi 2008; Montanari and Koutsoyiannis 2012; López López et al. 2014; Dogulu et al. 2015; see also Chapters 7 and 8 of this thesis). These errors are known to be heteroscedastic and correlated (see e.g., Montanari and Koutsoyiannis 2012; Montanari 2011; Evin et al. 2014). Based on the below-discussed properties of machine learning quantile regression algorithms, we believe that their use for solving the problem of interest could further reduce uncertainty (to some extent) by increasing the amount of information gained from the available historical records. In fact, these algorithms are not only a suitable (of course, not the only suitable), but also a direct and straightforward-to-apply, option for modelling hydrological model errors.

From a theoretical point of view, machine learning quantile regression algorithms are expected to be optimal in offering a satisfactory compromise between reliability and sharpness (targeted in technical applications), since they (most of them) are trained by minimizing the quantile score (see Section 2.6.1). They are also appropriate for modelling heteroscedasticity by perception and construction without requiring multiple fittings (i.e., a different fitting for each season or month), as it would be required for modelling heteroscedasticity using conditional distribution models. Some related technical illustrations on the appropriateness of (machine learning) quantile regression algorithms for probabilistic hydrological post-processing can be found in Chapter 7. Furthermore, additionally to using the hydrological model predictions at time t as predictor variable in the regression setting, one can also use the hydrological model predictions at times $t-1$, $t-2$, etc. (see e.g., the implementations herein), and/or precipitation and

potential evapotranspiration (or temperature) variables, to increase the amount of exploited information.

To further support our reasoning and rationale behind the selection of machine learning quantile regression algorithms as statistical post-processing models within methodologies for predictive uncertainty quantification in hydrology, we subsequently discuss some additional practical advantages stemming from their use. First, algorithms from this family are available in open source; therefore, their reproducibility is fully assured. Reproducibility is needed in hydrology, for example, according to Abrahart et al. (2008), Ceola et al. (2015), and Tyralis et al. (2019b), while only very few statistical post-processing models by hydrologists are made available in open source (see e.g., Vrugt 2016, 2018). Moreover, machine learning algorithms are well-tested (e.g., in forecasting competitions) in solving many practical problems and mostly optimally programmed (by computer scientists). This latter point is particularly important when one is interested in the operational use of the post-processing methodology, since it assures its fast implementation.

Some last, but certainly not least, practical advantages, as identified based on preliminary investigations, are also worth-discussing. In contrast to a few parametric (machine learning) models tried for this Chapter, these algorithms were found (a) to be highly reliable, in the sense that their (satisfactory) fitting was (almost) always possible; and (b) to (mostly) produce reasonable results with respect to the whole picture. Only quantile regression neural networks were found to produce significant outliers in terms of predictive performance, probably due to fitting quality problems. Specifically, this algorithm produced significant outliers for 10 of the 511 investigated catchments in the contiguous United States.

Another sound practical advantage, stemming from point (a) above, is related to what is called "automatic modelling", i.e., modelling that does not require human intervention during the whole process (see e.g., Chatfield 1988; Hyndman and Khandakar 2008; Taylor and Letham 2018). In light of this latter point, one could understand that automatic methodologies are the heart of operational hydrology, since they can effectively support large-sample hydrological applications, even at a global level (see e.g., Chapter 5 herein). The preference of these algorithms can indeed facilitate the complete automation of the probabilistic hydrological modelling process and, therefore, can effectively support probabilistic hydrological post-processing "at scale". An important clarification to be made here, is that complete automation is possible even in the case where quantile regression neural networks are exploited, as their rare failures significantly affect the widths of the prediction intervals and, therefore, can be foreseen using benchmarking. However, in such a case additional attention should be paid, by introducing an extra algorithmic step to detect extreme relative differences (usually relative increases) in terms of average width with respect to a performance stability benchmark (e.g., the quantile regression algorithm used herein). Such detection should be followed by the discard of the respective prediction, and its non-consideration by the equal-weight combiner. This automation has not been applied herein.

In summary, the integration of process-based models and machine learning quantile regression algorithms is considered highly meaningful, mainly due to the diverse backgrounds and specializations of the experts involved in the model development process for the two mother research fields, and not because these two model categories "simply exist" (see Sivakumar 2008a). It is also in line with the compromise between process-based and data-driven models proposed by Todini (2007). Inspired by this latter study, one would characterize the related approach to the problem of quantifying predictive hydrological uncertainty as "culture-integrating".

9.5.4 Value of ensemble learning hydrological post-processing methodologies

A certainly worth-of-attention way to reduce uncertainty in probabilistic hydrological modelling is to (optimally) exploit information provided by different hydrological models and/or different statistical post-processing models. The former type of model combination is more frequently applied and suggested in the literature (see e.g., the relevant suggestions by Montanari and Koutsoyiannis 2012). A concise and to-the-point presentation of several hydrological model

combination approaches, varying in terms of conceptualization and theory-driven reasoning, can be found in Vrugt (2019). Among the methods discussed therein that are appropriate for probabilistic hydrological modelling are PDF combination methods. Simple PDF averaging has been exploited to some degree in hydrological contexts (see e.g., Okoli et al. 2018).

In the present Chapter, we have exploited information from different machine learning quantile regression algorithms through quantile combination approaches. The latter are known to be more convenient in practice than PDF combination approaches (for reasons already reported in the above sub-section) and equally (or even more) useful in terms of predictive performance (Lichtendahl et al. 2013). To the best of our knowledge, such approaches have only been exploited so far for solving hydrological modelling and forecasting problems in Chapters 7 and 8 herein, and Tyralis et al. (2019a), while different machine learning quantile regression algorithms have been only combined for such purposes in Tyralis et al. (2019a). These three works emphasize the value of ensemble learning in general and equal-weight ensemble learning in particular (see also Lichtendahl et al. 2013; Winkler 2015), which is also well-supported by the large-scale empirical results delivered herein. In fact, the equal-weight combiner of the six machine learning algorithms of the present Chapter has been found to be an outstanding modelling choice with respect to several criteria.

Further improvements may result by adopting optimally unequal-weight stacked generalization approaches, such as the methodology introduced and validated by Tyralis et al. (2019a); see also Wang et al. (2019) for a similar approach applied within a different context). In Tyralis et al. (2019a), these improvements (with respect to the equal-weight combiner) have been quantified to be up to 2% in terms of average interval score, when adopting quantile regression and quantile regression forests for probabilistic hydrological post-processing in one-step ahead prediction problems. Such improvements are larger than one would think they are based on comparisons within single-case studies, since a case-specific improvement can be extremely better (or worse) than the average-case and median-case improvements (see the related comments, for instance, in Andréassian et al. (2007), Sivakumar (2008a) and Chapter 3 of this thesis), and should be pursued, especially for specific categories of applications, for which cost-effectiveness of the performance-improving methods also applies.

9.5.5 Grounds and implications of the proposed methodological framework

Understanding how the algorithms behave to improve predictive performance and reduce uncertainty in predictive modelling needs much more than inspecting their regular application and comparison to alternative approaches in some cases (Sivakumar 2008a). It needs properly conceptualized benchmark experiments (that, in turn, rely on data of adequate quantity and quality; see e.g., related comments by Andréassian et al. 2007 and Todini 2007), while toy experiments can also provide valuable insight into methodologies (see e.g., Krzysztofowicz 1999; Volpi et al. 2017). Andréassian et al. (2007) reported on a "lack of standardized procedures in model testing" in hydrology, emphasizing the fact that gaining end users' trust necessarily requires filling this methodological gap. We contribute towards this direction by developing a detailed framework for assessing statistical post-processing models in hydrological contexts. This framework is grounded on key suggestions made, for instance, by Sivakumar (2005), Andréassian et al. (2007), Todini (2007) and Sivakumar (2008a), and on empirical evidence derived from large-scale assessments, as summarized in the following.

The proposed framework produces trustable (or generalized) results. The fundamental role of large datasets in building trust in predictive hydrological modelling (which cannot be completely theory-driven) has been extensively pointed out and exploited by experts in the field (see e.g., the comments by Andréassian et al. 2006 and the model assessment by Perrin et al. 2003). This usefulness of large datasets holds, provided that they also represent a "wide range of climate and catchment conditions" (Perrin et al. 2003). As emphasized in Andréassian et al. (2006), operational hydrologists only trust models that perform well in a wide range of cases. Related comments can be found in Sivakumar (2008a), who underlines the fact that any model could be proven better than a competitive one in specific cases. This latter fact is consistent with

the "no free lunch" theorem by Wolpert (1996), which has been first put in a hydrological context in Chapter 3 of this thesis. This large-sample work and its companions (e.g., Tyralis and Papacharalampous 2017; Chapters 4–6 herein) have empirically proven the validity of the above comment by Sivakumar (2008) in hydrological forecasting, when endogenous variables are exclusively used.

Moreover, the proposed framework allows us to find optimized solutions to the following well-posed practical problem: How should we integrate different algorithms (or statistical post-processing models in general) within unifying frameworks or combine different algorithms, aiming at maximizing the benefits and reducing the risks from their use? This research question arises in light of key comments by Sivakumar (2008a); see also the related comments by Todini (2007). As pointed out in this latter study, the most useful comparative evaluations are those aimed at revealing the strengths and limitations of the various approaches to facilitate their optimal exploitation by answering research questions such as the above-stated one. It is relevant to highlight that posing research questions of this type requires us to first and foremost embrace the fact that a specific algorithm (or model) can be either useful or useless depending on its intended use, as it is also discussed in Chapter 3.

Furthermore, finding reliable answers to such practical questions also requires keeping the scale of our experiments as large as possible in general, i.e., by means besides the exploitation of large datasets as well. In fact, implementing an adequate number of algorithms (and/or models) and contrasting their predictive performance in various modelling situations can help in identifying well-performing algorithms for several prediction tasks that might be of interest. These tasks could be determined, for example, by specific prediction intervals and/or specific ranges of flow magnitudes, which therefore are separately examined within the introduced framework. By only reporting the performance of the algorithms in predicting the entire PDF (e.g., by computing the continuous ranked probability score – CRPS, as made, for example, in Bakker et al. 2019, and by relying our practical recommendations on it) and independently of the flow magnitude, a large amount of information (that would be useful in hydrological modelling and forecasting contexts) would remain unrevealed and unstudied.

A single score is mostly enough for properly quantifying the usefulness in performance. In our case, this single score could be the interval or quantile score, depending on the exact application of interest. Nonetheless, a multi-faced presentation of the results is also essential, since it (a) strengthens our understanding on how the various algorithms work by allowing related interpretations; and (b) provides some clues as to how to integrate these algorithms. Such multi-faced presentation is allowed, for instance, by those scores computed in Bourgin et al. (2015), Bock et al. (2018), Tyralis et al. (2019a), and Chapters 7 and 8 of this thesis, and the set of scores proposed in this Chapter. For instance, even when we are interested in delivering central prediction intervals, historical quantile scores can guide us towards delivering better probabilistic predictions by facilitating an optimal integration of two algorithms for forming the targeted prediction interval. Within this integration, each algorithm is used to predict quantiles of different level. Finally, we would like to highlight the appropriateness of the proposed framework in facilitating the selection of flow magnitude thresholds for the application of the various algorithms, based on the comparative performance of these algorithms for various flow magnitudes. Sivakumar (2005) underlines the role of such thresholds in hydrological modelling and forecasting. As pointed out by Sivakumar (2005), a single model should not be expected to model equally well high, medium and low values.

In summary, by applying the framework introduced herein, one can reliably gain insight on (i) which algorithm to select for each prediction task; and/or (ii) how to combine algorithms (also by testing various combinations), to maximize the benefits and minimize the risks from their use, thus facilitating a tangible contribution to the problem of uncertainty reduction. In light of this fact, the introduced framework could be further exploited in the future for:

- identifying the advantages and limitations of more statistical post-processing approaches, utilizing other machine learning quantile regression algorithms and ensemble learning

approaches (implemented with various sets of predictor variables) and/or other hydrological models, provided that these approaches are computationally fast and can be applied in a fully automatic way;

- solving related technical problems at different timescales (e.g., the monthly or seasonal timescales); and
- assessing statistical post-processing approaches in forecasting mode, i.e., by running the hydrological model using forecasts as inputs (instead of using observations).

Some final remarks should be made on our above-expressed suggestion for implementing different hydrological models within the broader methodologies exploited herein. As illustrated in Tyralis et al. (2019a, Figure 3), the GR4J hydrological model (implemented herein) successfully "pre-processes" the regression datasets (exploited by the machine learning quantile regression algorithms in probabilistic hydrological post-processing) by linearizing them. The smaller differences found between the machine learning algorithms of the present Chapter in predicting the median of daily streamflow compared to those found in Tyralis et al. (2020b) for point forecasting of daily streamflow by exclusively using machine learning algorithms could perhaps be attributed to this linearization (which seems to ease the regression problem to be solved). Under this view, the relative differences in the predictive performance of the machine learning algorithms would perhaps become larger or smaller (to some extent) for potential exploitations of the methodologies of the Chapter with different hydrological models, depending on how well these models perform.

9.6 Summary and take-home messages

We contribute with large-scale results and best practices to the problem of quantifying predictive uncertainty in hydrology, when the problem is examined from a predictive modelling perspective. We have made a detailed assessment of six machine learning quantile regression algorithms (i.e., quantile regression, generalized random forests for quantile regression, generalized random forests for quantile regression emulating quantile regression forests, gradient boosting machine, model-based boosting with linear models as base learners and quantile regression neural networks) and their equal-weight combiner in solving probabilistic hydrological modelling problems for 511 catchments in the contiguous United States. The examined catchments represent divergent climatic and catchment characteristics and, therefore, are appropriate for benchmarking purposes. By taking a quick glance at our large-scale results, one can immediately identify which algorithm should be selected (among the assessed ones) for maximizing the benefits and minimizing the risks from their use. The findings can be used in technical applications. The algorithms could be applied as detailed herein or within ensemble learning probabilistic hydrological post-processing methodologies.

In the following, we summarize the practical and methodological contributions of the Chapter in the form of take-home messages and recommendations:

- Preliminary large-sample investigations should focus on identifying a useful set of statistical post-processing models, such as the one composed by the six machine learning quantile regression algorithms of the Chapter.
- Machine learning quantile regression algorithms can effectively serve as statistical post-processing models, since they model heteroscedasticity by perception and construction without requiring multiple fittings, i.e., a different fitting for each season, as applying for the case of conditional distribution models.
- These algorithms are also straightforward-to-apply, fully automatic (i.e., their implementation does not require human intervention), available in open source, and computationally convenient and fast, and thus are highly appropriate for large-sample hydrological studies, while machine learning methods, in general, are known to be ideal for exploiting computers' brute force.

- Once a useful set of statistical post-processing models is identified, making the most of it, through model integrations and combinations, should be our target.
- Quantifying both the algorithms' overall performance (independently of the flow magnitude) and the algorithms' performance conditional upon the flow magnitude is of practical interest.
- Useful results are mostly those presented per level of prediction interval or predictive quantile, while those summarizing the quality of the entire predictive density (e.g., the continuous ranked probability score – CRPS) might also be of interest.
- Although the separate quantification of reliability and sharpness could be useful (mainly for increasing our understanding on how the algorithms work), what is most useful is computing scores that facilitate an objective co-assessment of these two criteria, such as the (rarely used in the literature) interval and quantile scores.
- The computational requirements might also be an important criterion for selecting an algorithm over others.
- In most cases, finding a balance between computational time and predictive performance is required. In any case, the criteria for selecting a statistical post-processing model should be clear.
- If we are foremost interested in obtaining results fast, then we probably should select quantile regression. This selection should be made keeping in mind that this algorithm is up to about 3.5% worse in terms of average quantile score than using the equal-weight combiner of all six algorithms of the Chapter.
- The equal-weight combiner of all six algorithms in this Chapter is identified as the best-performing algorithm overall, confirming the value of ensemble learning in general and ensemble learning via simple quantile averaging in particular. This value is well-recognized in the forecasting literature, but has not received much attention yet in the hydrological modelling and hydro-meteorological forecasting literature, in contrast to the popular concepts of ensemble simulation and ensemble prediction (e.g., via Bayesian model averaging) by exploiting information from multiple hydrological models.
- In spite of its outstanding performance, the equal-weight combiner of the six algorithms of the Chapter is, in turn, expected to perform worse than some of the individual algorithms in many modelling situations.
- In general, no algorithm should be expected to be (or presented as) the best performing with respect to every single criterion.
- By using different algorithms for delivering each predictive quantile (or prediction interval), the risk of producing a probabilistic prediction of bad quality is reduced. Related information on the predictive performance of the algorithms was extensively given in Section 9.4.1, while a summary is given below:
  - ✓ The equal-weight combiner is the best choice or amongst the best choices in terms of predictive performance for delivering predictive quantiles of level that is higher than 0.0125; however, it is also the most computationally demanding choice.
  - ✓ Quantile regression is the best choice in terms of predictive performance for predicting low-level quantiles (practically predictive quantiles of level lower than 0.0125) and the third-best choice for predicting high-level quantiles (practically predictive quantiles of level higher than 0.9).
  - ✓ Generalized random forests for quantile regression and generalized random forests for quantile regression emulating quantile regression forests are identified as the best choices or amongst the best choices in terms of predictive performance, when one is interested in delivering predictive quantiles of levels between 0.2 and 0.8. Since they are less computationally intensive than the equal-weight combiner, they would probably be preferred over the latter for relevant modelling applications.

- ✓ Improvements up to about 1.5% may be achieved for the generalized random forests for quantile regression and the generalized random forests for quantile regression emulating quantile regression forests by using as predictor variables of the regression the hydrological model predictions at times $t-3$, $t-2$, $t-1$ and t instead of using the hydrological model predictions only at times $t-1$ and t. By switching from the former set of predictors to the latter one, the improvements for the equal-weight combiner may reach an improvement of about 1%.
- ✓ Quantile regression neural networks is also a well-performing algorithm with respect to the whole picture and less computationally demanding than the equal-weight combiner; nevertheless, it is also the only individual algorithm among the assessed ones that was found to produce significant outliers (for ~2% of the investigated catchments). These performance issues were also manifested in the equal-weight combiner, yet in a less-pronounced degree.

- The overall performance improvements expressed in terms of average interval or quantile score are mostly up to 3%, while only for some extreme cases these improvements may reach up to about 20%. These cases concern some predictive quantiles of the lowest and highest levels, for which the tree-based methods, i.e., generalized random forests for quantile regression and generalized random forests for quantile regression emulating quantile regression forests and gradient boosting machine, do not work at their best.
- Unrealistic improvements in the order of 50% and 60%, even up to more than 100%, in terms of overall performance (often appearing in the literature) may result either by chance or by design when using small datasets, while they are highly unlikely to result on a regular basis when using large datasets. Only large-sample studies can produce trustable quantitative results in predictive modelling.
- Conducting large-sample studies is feasible nowadays, due to both the tremendous evolution of personal computers over the past few years and the fact that large datasets (e.g., the CAMELS dataset) are increasingly made available.
- Performance improvements may also be obtained by selecting algorithms according to their skill in predicting low, medium or high flows for the various quantiles (or central prediction intervals). Related information was extensively given in Section 9.4.2.
- Since we are mostly interested in obtaining results that are useful within operational settings, we have not performed hyperparameter optimization (which would require significantly higher computational time). The results could differ, if such optimization was performed.
- An alternative to hyperparameter optimization is ensemble learning, in the sense that both these procedures aim at improving probabilistic predictions. Here, we have extensively studied this alternative and showed that the improvements achieved are worth-of-attention.

This work is one of the very few large-scale ones in probabilistic hydrological post-processing and the even fewer ones conducted at daily timescale. We hope it will trigger interest and future research on the use of machine learning quantile regression algorithms in probabilistic hydrological post-processing "at scale" and on ways to maximize the benefits from their use.

10. Extended summary, innovations and contributions

10.1 Overall summary and considerations

This thesis falls into the scientific areas of stochastic hydrology, hydrological modelling and hydroinformatics. It contributes with new practical solutions, new methodologies and large-scale results to predictive modelling of hydrological processes, specifically to solving two technical predictive modelling problems. The latter are:

1) hydrological time series forecasting by exclusively using endogenous predictor variables (hereafter, referred to simply as "hydrological time series forecasting"); and

2) stochastic process-based modelling of hydrological systems via probabilistic post-processing (hereafter, referred to simply as "probabilistic hydrological post-processing").

These two technical problems are interrelated, and have been extensively investigated in Chapters 3–6 and Chapters 7–9, respectively. Moreover, in Chapter 2 the interested reader can find a brief overview of the theoretical, methodological and technical background of the conducted original works. In the same Chapter, we have outlined the predictive modelling and benchmarking toolbox, formed and exploited within the context of the thesis. This toolbox is consisted of (i) approximately 6 000 hydrological time series (sourced from larger freely available datasets), (ii) over 45 ready-made automatic models and algorithms mostly originating from the four major families of stochastic, (machine learning) regression, (machine learning) quantile regression, and conceptual process-based models, (iii) seven flexible methodologies (which together with the ready-made automatic models and algorithms consist the basis of our modelling solutions), and (iv) approximately 30 predictive performance evaluation metrics. Novel model combinations coupled with different algorithmic argument choices have resulted in numerous model variants, many of which could be perceived as new methods.

All of the exploited models and algorithms (see point (ii) above) are flexible, computationally convenient and fast; thus, they are appropriate for large-sample (even global-scale) hydrological investigations. Conducting such investigations has been of major priority herein, together with the introduction of new methodologies and new practical solutions. This priority is implied by the practical and algorithmic orientation of the thesis, and the (mainly) algorithmic nature of its methodologies. It is also relevant to note that most of the models or algorithms of point (ii) incorporate several others, thereby making it difficult for this thesis to explicitly describe (or even count) all the individual models exploited within its context. A key note to be made, in this regard, is that the understanding from a theoretical point of view of most (but not all) of the exploited models could hardly help in interpreting the algorithmically obtained outcomes of the present thesis. In light of the above, a strength (and limitation) characterizing the thesis (implied by its aims) is its algorithmic nature. In spite of this nature and the main orientation of our methodological frameworks, this thesis has also provided innovative theoretical supplements to its practical and methodological contribution.

In what follows, we summarize the content of Chapters 3–9 by emphasizing their main innovations in light of the literature and the technical know-how that they provide. We also discuss the way that these Chapters build on each other to (a) provide new technical solutions and novel methodologies, (b) answer practical and theoretical research questions, and (c) deliver new insights into the investigated technical problems by conducting large-scale comparisons and model evaluations.

10.2 Hydrological time series forecasting

10.2.1 Stochastic versus machine learning methods in multi-step ahead forecasting

Chapter 3 has overall aspired to promote large-scale comparisons in the area of hydrological time series forecasting. The Chapter begins with a brief overview (along with a critical view) of the hydrological time series forecasting literature. This literature often focuses on the comparison

between stochastic and machine learning forecasting methods, and on the validation of new "hybrid" data-driven methodologies, by conducting case studies. Case studies can ideally serve illustrative purposes; thus, they can be considerably useful when accompanying analytical investigations or large-scale empirical works. Analytical investigations have been conducted in the literature for several forecasting methods (mostly for the less flexible ones); nonetheless, they can be highly demanding (to nearly impossible) for many others (mostly for the most flexible machine learning ones). We have, therefore, argued that meaningful assessments and comparisons of hydrological time series forecasting methods would necessarily require the examination of a sufficiently large and representative sample of forecasting cases.

We have focused on the following research question: *Does the stochastic-machine learning dipole actually correspond to a clear difference in the forecasting performance of the methods?* To address this question, we have developed and exploited a detailed framework for assessing forecasting techniques in hydrology. Complying with the principles of forecasting, the introduced framework incorporates large-scale benchmarking. The latter relies on big hydrological datasets, large-scale time series simulation by using classical stationary stochastic models, many automatic forecasting models and algorithms (including benchmarks), and many forecast quality metrics. Our specific aim is to provide large-scale results and useful insights on the comparison of stochastic and machine learning forecasting methods for the case of hydrological time series forecasting at large temporal scales (e.g., the annual and monthly ones), with an emphasis on annual river discharge processes.

We have compared 11 stochastic and nine machine learning methods regarding their multi-step ahead forecasting properties. The stochastic methods include simple models, models from the frequently used families of autoregressive moving average and autoregressive fractionally integrated moving average, and innovations state space and exponential smoothing models, while the machine learning ones are neural networks, random forests and support vector machines. Among these categories of models, only the autoregressive (fractionally integrated) moving average, the neural network and the support vector machine ones are widely used in hydro-meteorological forecasting contexts; yet, methods from these categories are usually applied in a non-automatic form. Most of the remaining methods have been only exploited in (some of) the large-scale companions of this work (see e.g., Chapters 4, 5 and 6). We have used ready-made automatic time series forecasting algorithms with selected algorithmic argument choices and have also combined different algorithms to automate new ones. For the machine learning methods, we have proposed three objective lagged variable selection methods (among which one is inspired by a ready-made automatic algorithm) and three sets of grid hyperparameter values for optimization via grid search. The proposed set of methods could be used to benchmark the performance of any new time series forecasting method in hydrology. It has also been made available in code form.

We have conducted 12 large-scale computational experiments based on simulations, each using a different stationary stochastic model. The selected simulating models correspond to different types of autocorrelation. We have conducted each simulation experiment twice; the first time by using time series of 100 values and the second time by using time series of 300 values. Additionally, we have conducted a real-world experiment by using 405 mean annual river discharge time series of 100 values. The total number of forecasts is 858 480, among which 6 480 are produced within the real-world experiment. We have quantified the forecasting performance of the methods by using 18 metrics. These metrics do not share one-to-one relationships with each other, emphasizing –more or less– different aspects of the same information. They have been selected to provide a multi-faced assessment in multi-step ahead hydrological time series forecasting.

Our large-scale results suggest that stochastic and machine learning methods do not differ dramatically, as it is usually asserted in the literature. In fact, methods from both these categories have been found to be equally useful in short hydrological time series forecasting at large temporal scales. This outcome is especially interesting, given the claims that machine learning methods are more likely to be superior in "non-linear situations". The latter are often asserted to

characterize river discharge processes. In general, we cannot decide on a universally best or worst forecasting method, neither we can rank the forecasting methods based on our large-scale results. Any ranking of the forecasting methods would require the a priori selection of an experiment and a criterion of interest, as well as the application of a simplification procedure and, therefore, would not be general. However, the grouping of the forecasting methods based on their similar or contrasting performance with respect to the various metrics is possible, though only to some extent.

Another important contribution of the Chapter is related to the "no free lunch theorem". According to this theorem, in the space of all possible problem instances, there is not a model that will always perform better than other models in the absence of significant information for the problem at hand. Our large-scale results are consistent with this theorem, albeit the theorem refers to an infinite space of problem instances, while here we have examined a finite space of problems, formed by simulated and annual river discharge time series. In fact, finding the best algorithm mostly depends on our knowledge of the system, which apparently is deeper than the knowledge of its statistical properties (e.g., the mean, variance and autocorrelation function). Regarding the extent to which the conclusions could be generalizable for the forecasting of short hydrological time series at large time scales, we note that the stationarity assumption and the reasoning of its appropriateness for the modelling of geophysical properties is consistent with the no free lunch theorem. In particular, if we cannot explain the behaviour of a geophysical process based on a deterministic mechanism, then the most appropriate models are stationary. This is a frequently met case in the modelling of geophysical processes (i.e., there is not an adequate explanation for the behaviour of the geophysical process), proving that our conclusions could be generalizable.

10.2.2 One-step ahead predictability of annual temperature and precipitation

Chapter 4 has overall aspired to promote traditional forecasting methods in geoscience. The Chapter begins by providing detailed methodological information on several works using statistical methods for issuing hydro-meteorological time series forecasts, thereby complementing the introductory section of Chapter 3. It has specifically aimed to examine the fundamental problem of one-step ahead forecasting within a purely statistical framework (justified by forecasting experts) in geoscience, and hopefully to establish the results obtained by the examination of standardized real-world datasets as rough benchmarks for the one-step ahead predictability of geophysical processes. The establishment of forecasting benchmarks is meaningful, especially for the one-step ahead attempts, as the latter constitute the most simple ones and their accuracy can be quantified using a single metric, i.e., the absolute error.

To reach the above-outlined aims, we have expanded the work presented in Chapter 3 by exploring the one-step ahead forecasting properties of its methods, when the latter are applied to geophysical time series. Emphasis has been put on the examination of two real-world datasets, a precipitation dataset and a temperature dataset, together containing 297 annual time series of 91 values. These datasets have been examined in both their original and standardized forms. We have further performed large-scale experiments on 12 simulated datasets. In total, these datasets contain 24 000 time series of 91 values. The conducted simulation experiments complement the real-world ones by allowing the examination of a large variety of process behaviours, while they are also controlled to some extent, facilitating generalizations and increasing the understanding on the examined problem. We have used the first 50, 60, 70, 80 and 90 data points for model fitting and model validation, and have made predictions corresponding to the 51st, 61st, 71st, 81st and 91st data points, respectively. The number of forecasts produced in the same Chapter is 2 177 520, among which 47 520 are obtained using the real-world datasets. The assessment has been based on eight error metrics and accuracy statistics.

The simulation experiments have revealed the most and least accurate methods for long-run one-step ahead forecasting applications, also suggesting that the simple methods may be competitive in specific cases. They have also shown that the relative performance of the forecasting methods is slightly affected by the time series length (when considering time series of

51, 61, 71, 81, 91 values), while it strongly depends on the process. Regarding the results of the real-world experiments using the original (standardized) time series, the minimum and maximum medians of the absolute errors have been found equal to 68 mm (0.55) and 189 mm (1.42), respectively, for precipitation, and 0.23°C (0.33) and 1.10°C (1.46), respectively, for temperature. Since there is an absence of relevant information in the literature, the numerical results obtained using the standardized real-world datasets could be used as rough benchmarks for the one-step ahead predictability of annual precipitation and temperature.

10.2.3 Multi-step ahead predictability of monthly temperature and precipitation

Chapter 5 has overall aspired to promote the use of automatic time series forecasting methods in geoscience. The non-automatic or subjective approach to the problem of time series forecasting, often adopted in the geoscientific (including the hydrological) literature, requires the prior conduct of an exploratory data analysis for each specific individual case to be predicted and human intervention during the forecasting process. Therefore, its implementation can be significantly limited by scale-dependent factors. Automatic time series forecasting is essential, for example, when a large number of time series forecasts is required.

We have conducted two global-scale investigations. We have quantified the predictability of monthly temperature and precipitation by applying 24 automatic time series forecasting methods to 985 and 1 552 monthly time series of temperature and precipitation, respectively. This sample is the largest used in hydrology for assessing the performance of time series forecasting methods. We have exploited ready-made automatic models with different algorithmic argument choices and have also combined different models to automate new ones. The exploited automatic methods are (a) the seasonal naïve (based on the monthly values of the last year), (b) four methods based on random walk with drift, (c) four methods based on an automatic autoregressive fractionally integrated moving average model, (d) six methods based on the exponential smoothing state space model with Box-Cox transformation, autoregressive moving average error correction, trend and seasonal components, (e) four methods based on simple exponential smoothing, (f) two methods based on Theta, and (g) three methods based on Prophet.

Prophet is a recently introduced model inspired by the nature of time series forecasted at Facebook, which in this work has been applied for the first time to hydrometeorological time series. The automatic autoregressive fractionally integrated moving average model, on the other hand, is widely used in a non-automatic way in the hydrological literature, while the rest of the models have been rarely implemented in hydrology, e.g., in Chapters 3 and 4, although they are very common in the forecasting literature. In the latter studies, no investigation is provided on how different choices of handling the seasonality and non-normality affect the performance of the models. This investigation constitutes one of the main aims of the present Chapter (therefore, proper variants of the methods are examined), together with the quantification of the performance of the selected models on monthly hydrometeorological time series and the comparison of the Prophet model to the rest. The exploited time series are 480 months long with no missing values, observed between January 1950 and December 1989 in stations covering a significant part of the Earth's surface and, therefore, including various real-world process behaviours. The models are fitted in the first 36 years of data (432 months) and subsequently tested in performing multi-step ahead forecasts for the last four years of data (48 months). The results has been summarized in global scores, while their examination by group of stations has led to five individual scores for temperature and six for precipitation. The groups have been formed according to the geographical vicinity of the stations.

The results indicate that all the examined methods apart from the naïve and random walk ones are accurate enough to be used in long-term forecasting applications. Even the simple exponential smoothing and Theta models, which exhibit a rather moderate performance in terms of root mean square error and Nash-Sutcliffe efficiency in the simulation experiments of Chapter 3, here have been found to be equally competitive with the autoregressive fractionally integrated moving average model and the exponential smoothing state space model with Box-Cox transformation, autoregressive moving average error correction, trend and seasonal components.

These latter models are the most accurate in terms of root mean square error and Nash-Sutcliffe efficiency in the above-mentioned Chapter. This may be explained by the fact that the simulation experiments of Chapter 3 examine non-seasonal simulated processes, with different predictability than the monthly temperature and precipitation processes. Seasonality can be assumed to constitute a deterministic component of a process and its proper handling leads to a significant improvement of the forecasts. The above-stated qualitative outcome is consistent with the 50 single-case studies of Chapter 6. These case studies use monthly temperature and precipitation data as well. In the same work, the seasonality term is estimated using the multiplicative and additive model, for the temperature and precipitation time series, respectively. Regarding the investigations on how different choices of handling seasonality and non-normality affect the performance of the models, the results do not suggest any specific combination of choices for the external handling of seasonality and non-normality as best. Nevertheless, the handling of seasonality through the exponential smoothing state space model with Box-Cox transformation, autoregressive moving average error correction, trend and seasonal components and the Prophet model (the only models that offer this possibility amongst the used ones) mostly leads to less accurate forecasts than the external handling, especially for the former model.

Admittedly, the quantitative information provided by Chapter 5 is also important, since it directly expresses the predictability of monthly temperature and precipitation. The minimum and maximum medians of the absolute errors of the temperature forecasts have been found to be approximately equal to 0.25 K and 8.2 K, respectively. Furthermore, a zero median of the absolute errors has been computed for the precipitation forecasts produced for the dry months in geographical regions with relatively regular variability in precipitation, while the maximum median computed has been approximately equal to 100 mm. These values could be viewed in comparison with the minimum and maximum medians of absolute errors for annual temperature and precipitation, as derived in Chapter 4 using two real-world datasets of 297 time series in total. These are approximately equal to 0.23 K and 1.10 K, and 68 mm and 189 mm, respectively. Moreover, the computed RMSE values range between 1.01 K and 3.65 K for temperature, and 36.16 mm and 70.17 mm for precipitation, while the respective NSE values are 0.79 and 0.98 for temperature, and −0.55 and 0.71 for precipitation.

Excluding the naïve method and the variants using the random walk model, the respective RMSE values range between 1.01 K and 2.84 K for temperature, and 36.16 mm and 51.71 mm for precipitation. In more detail, for the total of the temperature time series the use of an ARFIMA, BATS, simple exponential smoothing, Theta or Prophet model, instead of the naïve method, leads to about 19–29% more accurate forecasts in terms of RMSE, or even in about 30–32% more accurate forecasts specifically for the temperature time series observed in North Europe. For the total of the precipitation time series the use of all these automatic methods leads to about 21–22% better forecasts than the use of the naïve method, while for the geographical regions of North America, North Europe and East Asia these percentages are 26–29%, 22–24% and 32–38% respectively. This higher degree of accuracy is non-ignorable and particularly important in a long run perspective. Importantly, the Prophet model has been found to offer from 13% up to 32% and from 16% up to 38% better results than the naïve method for the temperature and precipitation time series, respectively. Moreover, the minimum and maximum NSE medians for the ARFIMA, BATS, simple exponential smoothing, Theta and Prophet models are 0.89 and 0.98 for temperature, and −0.04 and 0.71 for precipitation. The former NSE values indicate good forecasting performances and the latter acceptable to moderate. The higher predictability of the monthly temperature compared to the monthly precipitation is expected already from the comparison of their corresponding standard deviation values of the seasonally decomposed time series, which have a median around 1.7 K and 42 mm respectively. We think that the level of the forecasting accuracy can barely be improved using other methods, as the experiments of Chapter 3 suggest.

10.2.4 A multiple-case study focusing on machine learning algorithms

Chapter 6 has overall aspired to promote the multiple-case study research strategy –in its large-scale version– as an innovative and more comprehensive alternative to conducting single-case studies in the field of hydrological time series forecasting. This strategy embraces the examination of more than one individual cases, thereby facilitating the observation of specific phenomena from multiple perspectives or within different contexts. For the detection of systematic patterns across the individual cases, a cross-case synthesis can be performed. Given the fact that the boundaries between the phenomena and the context are not clear, it is important that each individual case keeps its identity within a multiple-case study, so that one can specifically focus on it. This exploration within and across the individual cases can provide interesting insights into the phenomena under investigation, as well as a form of generalization named "contingent empirical generalization", while retaining the immediacy of the single-case study method.

We have conducted an extensive multiple-case study composed by 50 single-case studies. The latter have used monthly temperature and precipitation time series observed in Greece. We have examined these two geophysical processes, because they exhibit different properties, which may affect differently the results within the explorations. The main aim of this multiple-case study has been the exploration of three problems associated with hydrological time series forecasting using machine learning algorithms. The investigated problems are: (a) lagged variable selection, (b) hyperparameter handling, and (c) comparison of machine learning and stochastic algorithms. We have also presented quantitative information about the quality of the forecasts (particularly important for the case of Greece), and searched for evidence regarding the existence of possible relationships between the forecast quality, and the maximum likelihood estimates of the standard deviation, coefficient of variation and Hurst parameter of the fractional Gaussian noise process for the deseasonalized time series (used for model fitting).

We have focused on two machine learning algorithms, i.e., neural networks and support vector machines, and have also included four stochastic methods and a seasonal naïve benchmark in the comparisons. The stochastic methods are (i) the autoregressive order one model, (ii) an automatic algorithm from the family of autoregressive fractionally integrated moving average models, (iii) the exponential smoothing state space algorithm with Box-Cox transformation, autoregressive moving average errors, trend and seasonal components, and (iv) the Theta algorithm. We have applied a fixed methodology to each individual case and, subsequently, we have performed a cross-case synthesis to facilitate the detection of systematic patterns. We have fitted the models to the deseasonalized time series. We have compared the one- and twelve-step ahead forecasting performance of the algorithms. The assessment of the one-step ahead forecasting performance is based on the absolute error of the forecast of the last monthly observation. For the quantification of the multi-step ahead forecasting performance, we have computed five metrics on the test set (last year's monthly observations), i.e., the root mean square error, the Nash-Sutcliffe efficiency, the ratio of standard deviations, the coefficient of correlation and the index of agreement.

The findings suggest that forecasting methods based on the same machine learning algorithm may exhibit very different performance, to an extent mainly depending on the algorithm and the individual case. In fact, the neural networks algorithm can produce forecasts of many different qualities for a specific individual case, in contrast to the support vector machines one. The performance of the former algorithm seems to be more affected by the selected lagged variables than by the adopted hyperparameter selection procedure (use of predefined hyperparameters or defined after optimization). While no evidence has been provided that any of the compared lagged regression matrices systematically leads to better forecasts than the rest, either for the neural networks or the support vector machines algorithms, the results mostly favour using less recent lagged variables. Furthermore, hyperparameter optimization does not necessarily lead to better forecasts than the use of the default hyperparameter values of the examined algorithms. Regarding the comparisons performed between machine learning and classical algorithms, the results indicate that methods from both categories can perform equally well, under the same

limitations. The best method depends on the case examined and the criterion of interest, while it can be either machine learning or classical. Some information of secondary importance derived by our experiments is subsequently reported. The average-case performance of the algorithms used to produce one- and twelve-step ahead monthly temperature forecasts ranges between 0.66°C and 1.00°C, and 1.14°C and 1.70°C, in terms of absolute error and root mean square error respectively. For the monthly precipitation forecasts the respective values are 39 mm and 72 mm, and 41 mm and 52 mm. Finally, no evidence has been provided by our multiple-case study that there is any relationship between the forecast quality and the estimated parameters of the fractional Gaussian noise process for the deseasonalized time series.

10.3 Probabilistic hydrological post-processing

10.3.1 An ensemble learning methodology and its toy model investigation

Chapter 7 has introduced a novel probabilistic hydrological post-processing methodology by using a theoretically consistent probabilistic hydrological modelling blueprint as a starting point. The proposed methodology is subdivided into three alternative variants. In summary, it generates a large number of point predictions by utilizing a single hydrological model, yet with different parameter values. By solving a typical regression problem using a quantile regression algorithm (hereafter referred to as the "error model" of the methodology), these "sister predictions" are converted into auxiliary probabilistic predictions (consisted of quantile predictions), which are finally combined via simple quantile averaging. To the best of our knowledge, this is the first quantile averaging hydrological post-processing methodology that creates and exploits different information sets using a single model with different parameter values.

A key improvement achieved in terms of flexibility in modelling (compared to the original work and the precursor variants) is the use of statistical learning regression models that can directly provide predictive quantiles of the response variable, while they are also appropriate for modelling heteroscedasticity. Such models are the quantile regression, generalized random forests for quantile regression, generalized random forests for quantile regression emulating quantile regression forests, gradient boosting machine, model-based boosting with linear models as base learners and quantile regression neural networks (see Chapter 9). Allowing the exploitation of the possibilities provided by this model category should, in fact, be regarded as a primary strength of the proposed methodology from a predictive modelling perspective.

We have demonstrated the usefulness of the proposed methodology and how our understanding of the system to be modelled can guide us to achieve better predictive modelling when using this methodology by conducting a toy model investigation. Within this investigation, we have focused on the unsuitability of the homoscedasticity assumption, when the latter is made in the modelling of the hydrological model's error, and on how the selection of an appropriate regression model for this task results in improved probabilistic predictions. We have also demonstrated the significance of using a better hydrological model for delivering probabilistic predictions that are simultaneously reliable and as sharp as possible. Finally, we have used the obtained toy results to show how the proposed methodology increases its robustness in performance by averaging many quantile predictions.

In spite of focusing on the introduced methodology, some of the obtained results can be used for gaining insight in general on how two-stage hydrological post-processing methodologies work and under which conditions their performance is maximized. The presented toy examples, demonstrating the key roles of both the statistical learning regression model and the hydrological model within a hydrological post-processing methodology, go beyond of some few exemplary (yet basic) toy tests that have already been made for the interpretation of methodologies for the quantification of the predictive hydrological uncertainty. Such tests mostly assume homoscedasticity and a perfect toy hydrological model, while here we are also inspired by recent simulation experiments that do not rely on these assumptions.

Two simultaneously attractive and useful properties of this methodology (extensively tested in Chapter 8) are its larger robustness in performance compared to the combined individual predictors and, by extension, compared to basic two-stage post-processing methodologies (which produce a single probabilistic prediction and, therefore, no prediction combination is made in their case), and its ability to "harness the wisdom of the crowd". The latter is defined in the forecasting literature as the property of some prediction combinations to score no worse –usually better– than the average score of the combined individual predictions. In fact, the larger the number of the combined quantile predictions (equal to the number of the generated sister predictions), the more robust the ensemble predictor and the more harnessed the wisdom of the crowd.

The proposed methodology is characterized by some additional strengths that are also particularly important from a predictive modelling point of view. First, it is computationally convenient in the sense that it can be easily expressed in algorithmic form and programmed using open source routines. Second, it offers certain modelling options that could be exploited to maximize predictive performance. For instance, variants 1 and 2 allow the exploitation by the error model of a large number of different information sets, instead of a single one (exploited by variant 3), thereby facilitating the enlargement of the sample space of the hydrological model's observed errors. This enlargement could be particularly important for modelling these errors using methods which do not extrapolate beyond the values of the training dataset, such as the quantile regression forests model. Lastly, it allows the exploitation of the total amount of available information, in the sense that each sister prediction is herein converted into a probabilistic prediction (consisted of several quantile predictions) instead of a single simulation (as implied by the original work and the precursor variants).

Some limitations of the proposed methodology should also be considered. These include limitations implied by its two-stage nature, such as its shortcoming in terms of interpretability in modelling (especially in terms of producing interpretable parameter estimates) and its significant data length requirements. Although this latter limitation should be acknowledged herein and perhaps taken into consideration in real-world applications, (daily) datasets are usually essentially large. Moreover, in Chapter 8 it is empirically proven that, in practice, even when the available historical information is little, the proposed methodology is well-performing when implemented using the quantile regression model as error model.

Furthermore, the computational requirements of the proposed methodology are (at the moment) high when (i) computationally intensive procedures (e.g., Markov Chain Monte Carlo simulation sampling) are preferred for calibrating the hydrological model, and/or (ii) the error model is trained as implied by variant 1 or variant 2, unless the application is restricted to considering a small number of sister predictions. Note that a computationally convenient and simple algorithm is not necessarily computationally fast. It is also important to clarify that the above-outlined limitation holds only for applications to hundreds of catchments and timescales finer than the monthly one, and for implementations through regular personal computers. It does not hold for applications to a small number of catchments, and applications at the monthly and annual timescales. Still, large-scale applications at the daily timescale can be supported by variant 3, when this variant is implemented by using computationally fast algorithms for calibrating the hydrological model.

In addition to the above-discussed considerations and in contrast to several statistical methodologies for probabilistic prediction, a well-known drawback of flexible statistical learning models for quantile prediction is their inappropriateness for modelling long-range dependence. Modelling this dependence when solving prediction problems is a frequently met concern in applied stochastic hydrology (see e.g., the large-scale investigations in Chapters 3–5 and the comparative case study in Chapter 6). Nonetheless, empirical evidence suggests that the AR(1) assumption (in some sense allowed by the proposed methodology by using as a predictor variable in regression the hydrological model's prediction at time $t-1$) is adequate when modelling hydrological models' errors. In general, by including more than one predictor variables (e.g., the hydrological model's predictions at times t, $t-1$, $t-2$, etc.) in the regression settings we can

increase the amount of the available information exploited and improve predictive performance, as it is empirically proven for rainfall-runoff modelling problems in Chapter 9 of this thesis.

Overall, the main trade-off to be considered when selecting between the proposed methodology and basic two-stage post-processing methodologies (utilizing the same error model) is the one between (a) the increased robustness in performance and the ability to harness the wisdom of the crowd, both offered by the former methodology, and (b) the significantly less computational requirements of a basic post-processing methodology. We believe that from a risk management standpoint this trade-off is worthy, as the large-sample experiment of Chapter 8 suggests.

10.3.2 Large-sample investigations emphasizing on robustness assessment

Chapter 8 has been devoted to validating the probabilistic hydrological modelling methodology proposed in Chapter 7. This methodology adopts key concepts from a flexible probabilistic hydrological modelling methodology, while also relying on the concept of probabilistic prediction combination from the forecasting field. It applies a single hydrological model using a large number of different parameter values to generate the same number of "sister predictions". The parameters of the hydrological model can be obtained by using either Bayesian calibration schemes or informal calibration schemes. Therefore, this methodology does not have any particular relationship with Bayesian methods by construction, as it also applies to its precursor. A statistical learning (or machine learning) regression model that is suitable for predicting quantiles (see e.g., the models exploited in Chapter 9 of this thesis) is then used to obtain information about the hydrological model's error. This information is used to convert the sister predictions into probabilistic predictions, which are finally combined in simple fashion to obtain the output probabilistic predictions. The assessed methodology is subdivided into three alternative variants, which differ only in the training of the regression model.

We have conducted a large-sample real-world experiment at monthly timescale, set up using complete 50-year monthly information for 270 catchments in the United States. Aiming to increase the understanding in probabilistic hydrological modelling, we have insisted on interpretability and benchmarking within all conducted tests. We have used the parsimonious GR2M hydrological model and two (largely) interpretable regression models, specifically the linear regression and the quantile regression ones, to implement six ensemble schemes, all of them based on the assessed methodology. Those ensemble schemes implemented using the linear model (three in number) have been used as benchmarks for the remaining schemes (also three in number). Those ensemble schemes using the same regression model rely on different variants of the assessed methodology. The performance of the ensemble schemes has been assessed by computing the coverage probabilities, average widths and average interval scores of the obtained interval predictions, and by also benchmarking their results using naïve probabilistic data-driven models.

The obtained numerical results (metric values computed for 4 870 800 interval predictions) suggest the usefulness of the assessed methodology in obtaining probabilistic predictions of hydrological quantities. The best-performing variant, offering a mean relative improvement up to 5.46% with respect to its alternative variants, when implemented using the quantile regression model, is variant 2. This variant trains the regression model on a single large dataset formed by using information from all sister predictions. The average-case relevant improvements when using the quantile regression model instead of the linear regression one range up to about 37% in terms of average interval score. This latter numerical result should be appraised on the basis that only the former of these models can model heteroscedasticity. The homoscedasticity assumption is often made in the literature when modelling the hydrological model's error.

Finally, we have demonstrated the increased robustness of the assessed methodology with respect to the combined (by this methodology) individual predictors and, by extension, to basic two-stage post-processing methodologies. The ability to "harness the wisdom of the crowd" has also been empirically proven. The quantile predictions obtained by all ensemble predictors are

found to score no worse –usually better– than the average of the individual scores of the combined individual predictions in terms of average interval score. The computed relative differences favour the former quantity over the latter up to about 37%, while their mean values range between 0.19% and 1.83%, depending both on the prediction interval and the variant of the assessed methodology. For the best-performing ensemble scheme the respective average relative differences are around 1%. Overall, the robustness and the ability to harness the wisdom of the crowd are identified as two key properties of the working methodology.

10.3.3 Why and how to combine process-based and machine learning models

Chapter 9 has overall aspired to (i) promote the use of machine learning algorithms in the fields of probabilistic hydrological modelling and hydro-meteorological forecasting, (ii) pass on the message in hydrology that machine learning methods can deliver probabilistic predictions, (iii) attract attention to ensemble learning post-processing methodologies, and (iv) promote the use of large datasets and benchmarking when using machine learning methods in hydrology. The Chapter has introduced the largest range of probabilistic hydrological modelling methods ever introduced in a single work and has additionally conducted the largest benchmark experiment ever conducted on the use of machine learning quantile regression algorithms for probabilistic hydrological post-processing. We have focused on the following research question: *Why and how to combine process-based models and machine learning quantile regression algorithms for probabilistic hydrological post-processing?* Therefore, our contribution includes the inspection and appraisal of both quantitative and qualitative aspects of the application of the algorithms.

We have discussed some key benefits stemming from the integration of process-based and machine learning models, as understood from an uncertainty reduction point of view. We have also discussed some sound practical advantages stemming from the same integration. In summary, by incorporating process-based hydrological models into probabilistic hydrological post-processing methodologies, we benefit from the hydrological modellers' experience (therefore, uncertainty is reduced to some extent) and simultaneously quantify predictive hydrological uncertainty. Moreover, machine learning quantile regression algorithms can effectively serve as statistical post-processing models, since they model heteroscedasticity by perception and construction, thereby contributing further to our uncertainty reduction aim. They are also straightforward-to-apply, fully automatic (i.e., their implementation does not require human intervention), available in open source, and computationally convenient and fast. Thus, they are highly appropriate for large-sample hydrological studies.

Our benchmark experiment has been set up using 34-year-long daily time series of precipitation, temperature, evapotranspiration and streamflow for 511 catchments over the contiguous United States. Point hydrological predictions have been obtained using the GR4J hydrological model and exploited as predictor variables within quantile regression settings. Six machine learning quantile regression algorithms and their equal-weight combiner are applied to predict conditional quantiles of the hydrological model errors. The selected individual algorithms are quantile regression, generalized random forests for quantile regression, generalized random forests for quantile regression emulating quantile regression forests, gradient boosting machine, model-based boosting with linear models as base learners and quantile regression neural networks. The conditional quantiles of the hydrological model errors have been transformed to conditional quantiles of daily streamflow, which have been finally assessed using proper performance scores and benchmarking. The assessment has concerned various levels of predictive quantiles and central prediction intervals, while it has been made both independently of the flow magnitude and conditional upon this magnitude.

The findings can be used in technical applications. The algorithms should be applied in a way that maximizes the benefits and reduces the risks from their use. This can be achieved by combining algorithms (e.g., by exploiting the methodology of Chapters 7 and 8), and by integrating algorithms within systematic frameworks (e.g., by using different algorithms for delivering each predictive quantile of interest or by selecting algorithms according to their skill in predicting low, medium or high flows for the various quantiles). If we are foremost interested in obtaining results

fast, then we probably should select quantile regression. This selection should be made keeping in mind that this algorithm is up to about 3.5% worse in terms of average quantile score than using the equal-weight combiner of all six algorithms. This combiner has been identified as the best-performing algorithm overall, confirming the value of ensemble learning in general and ensemble learning via simple quantile averaging in particular. This value is well-recognized in the forecasting literature, but has not received much attention yet in the hydrological modelling and hydro-meteorological forecasting literature. In spite of its outstanding performance, the equal-weight combiner of the six algorithms is, in turn, expected to perform worse than some of the individual algorithms in many modelling situations. In general, no algorithm should be expected to be (or presented as) the best performing with respect to every single criterion.

References

[1] Abbas SA, Xuan Y (2019) Development of a new quantile-based method for the assessment of regional water resources in a highly-regulated river basin. Water Resources Management 33(8):3187–3210. https://doi.org/10.1007/s11269-019-02290-z

[2] Abebe AJ, Price RK (2003) Managing uncertainty in hydrological models using complementary models. Hydrological Sciences Journal 48(5):679–692. https://doi.org/10.1623/hysj.48.5.679.51450

[3] Abrahart RJ, See LM, Dawson CW (2008) Neural network hydroinformatics: Maintaining scientific rigour. In: Abrahart RJ, See LM, Solomatine DP (eds) Practical Hydroinformatics. Springer-Verlag Berlin Heidelberg, pp 33–47. https://doi.org/10.1007/978-3-540-79881-1_3

[4] Abrahart RJ, Anctil F, Coulibaly P, Dawson CW, Mount NJ, See LM, Shamseldin AY, Solomatine DP, Toth E, Wilby RL (2012) Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Progress in Physical Geography: Earth and Environment 36(4):480–513. https://doi.org/10.1177/0309133312444943

[5] Abudu S, Cui C, King JP, Abudukadeer K (2010) Comparison of performance of statistical models in forecasting monthly streamflow of Kizil River, China. Water Science and Engineering 3(3):269–281. https://doi.org/10.3882/j.issn.1674-2370.2010.03.003

[6] Achen CH, Snidal D (1989) Rational deterrence theory and comparative case studies. World Politics 41(2):143–169. https://doi.org/10.2307/2010405

[7] Addor N, Newman AJ, Mizukami N, Clark MP (2017a) Catchment attributes for large-sample studies. Boulder, CO: UCAR/NCAR. https://doi.org/10.5065/D6G73C3Q

[8] Addor N, Newman AJ, Mizukami N, Clark MP (2017b) The CAMELS data set: Catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21:5293–5313. https://doi.org/10.5194/hess-21-5293-2017

[9] Ahmed NK, Atiya AF, GayarAn NE, El-Shishiny H (2010) An empirical comparison of machine learning models for time series forecasting. Econometric Reviews 29(5–6):594–621. https://doi.org/10.1080/07474938.2010.481556

[10] Akaike H (1974) A new look at statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723. https://doi.org/10.1109/TAC.1974.1100705

[11] Allaire JJ, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, Iannone R (2019) `rmarkdown`: Dynamic Documents for `R`. `R` package version 1.14. https://CRAN.R-project.org/package=rmarkdown

[12] Alpaydin E (2010) Introduction to Machine Learning, second edition. MIT Press

[13] Anctil F, Perrin C, Andréassian V (2004) Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models. Environmental Modelling and Software 19(4):357–368. https://doi.org/10.1016/S1364-8152(03)00135-X

[14] Anctil F, Filion M, Tournebize J (2009) A neural network experiment on the simulation of daily nitrate-nitrogen and suspended sediment fluxes from a small agricultural catchment. Ecological Modelling 220(6):879–887. https://doi.org/10.1016/j.ecolmodel.2008.12.021

[15] Andrawis RR, Atiya AF, El-Shishiny H (2011) Forecast combinations of computational intelligence and linear models for the NN5 time series forecasting competition. International Journal of Forecasting 27(3):672–688. https://doi.org/10.1016/j.ijforecast.2010.09.005

[16] Andréassian V, Hall A, Chahinian N, Schaake J (2006) Introduction and synthesis: Why should hydrologists work on a large number of basin data sets?. IAHS publication 307:1

[17] Andréassian V, Lerat J, Loumagne C, Mathevet T, Michel C, Oudin L, Perrin C (2007) What is really undermining hydrologic science today?. Hydrological Processes 21(20):2819–2822. https://doi.org/10.1002/hyp.6854

[18] Andréassian V, Perrin C, Berthet L, Le Moine N, Lerat J, Loumagne C, Oudin L, Mathevet T, Ramos M-H, Valéry A (2009) HESS Opinions "Crash tests for a standardized evaluation of hydrological models". Hydrology and Earth System Sciences 13:1757–1764. https://doi.org/10.5194/hess-13-1757-2009

[19] Arcuri A, Fraser G (2013) Parameter tuning or default values? An empirical investigation in search-based software engineering. Empirical Software Engineering 18(3):594–623. https://doi.org/10.1007/s10664-013-9249-9

[20] Armstrong JS (2001) Evaluating forecasting methods. In: Armstrong JS (ed) Principles of Forecasting. International Series in Operations Research & Management Science, vol 30. Springer, Boston, MA, pp 443–472. https://doi.org/10.1007/978-0-306-47630-3_20

[21] Armstrong JS, Collopy F (1992) Error measures for generalizing about forecasting methods: Empirical comparisons. International Journal of Forecasting 8(1):69–80. https://doi.org/10.1016/0169-2070(92)90008-W

[22] Armstrong JS, Fildes R (2006) Making progress in forecasting. International Journal of Forecasting 22(3):433-441. https://doi.org/10.1016/j.ijforecast.2006.04.007

[23] Assimakopoulos V, Nikolopoulos K (2000) The theta model: A decomposition approach to forecasting. International Journal of Forecasting 16(4):521–530. https://doi.org/10.1016/S0169-2070(00)00066-2

[24] Athey S, Tibshirani J, Wager S (2019) Generalized random forests. The Annals of Statistics 47(2):1148–1178. https://doi.org/10.1214/18-AOS1709

[25] Atiya AF, El-Shoura SM, Shaheen SI, El-Sherif MS (1999) A comparison between neural-network forecasting techniques-case study: river flow forecasting. IEEE Transactions on Neural Networks 10(2):402–409. https://doi.org/10.1109/72.750569

[26] Attali D (2018) `ggExtra`: Add Marginal Histograms to `'ggplot2'`, and More `'ggplot2'` Enhancements. `R` package version 0.8. https://CRAN.R-project.org/package=ggExtra

[27] Babu CN, Reddy BE (2012) Predictive data mining on Average Global Temperature using variants of ARIMA models. 2012 International Conference on Advances in Engineering, Science and Management (ICAESM)

[28] Bakker K, Whan K, Knap W, Schmeits M (2019) Comparison of statistical post-processing methods for probabilistic NWP forecasts of solar radiation. https://doi.org/10.1016/j.solener.2019.08.044

[29] Ballini R, Soares S, Andrade MG (2001) Multi-step-ahead monthly streamflow forecasting by a neurofuzzy network model. Joint 9th IFSA World Congress and 20th NAFIPS International Conference:992–997. https://doi.org/10.1109/NAFIPS.2001.944740

[30] Bărbulescu A (2016) Studies on Time Series Applications in Environmental Sciences. Springer International Publishing, Cham, Switzerland. https://doi.org/10.1007/978-3-319-30436-6

[31] Barnard GA (1963) New methods of quality control. Journal of the Royal Statistical Society. Series A (General) 126(2):255–258. https://doi.org/10.2307/2982365

[32] Baxter P, Jack S (2008) Qualitative case study methodology: Study design and implementation for novice researchers. The Qualitative Report 13(4):544–559

[33] Belayneh A, Adamowski J, Khalil B, Ozga-Zielinski B (2014) Long-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet neural network and wavelet support vector regression models. Journal of Hydrology 508:418–429. https://doi.org/10.1016/j.jhydrol.2013.10.052

[34] Bengtsson H (2018) `matrixStats`: Functions that Apply to Rows and Columns of Matrices (and to Vectors). `R` package version 0.54.0. https://CRAN.R-project.org/package=matrixStats

[35] Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. Advances in Water Resources 16(1):41–51. https://doi.org/10.1016/0309-1708(93)90028-E

[36] Beven KJ (2000) Uniqueness of place and process representations in hydrological modelling. Hydrology and Earth System Sciences 4:203–213. https://doi.org/10.5194/hess-4-203-2000
[37] Beven KJ (2001) How far can we go in distributed hydrological modelling?. Hydrology and Earth System Sciences 5:1–12. https://doi.org/10.5194/hess-5-1-2001
[38] Beven KJ (2006) A manifesto for the equifinality thesis. Journal of Hydrology 320(1–2):18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007
[39] Beven KJ (2012) Rainfall-runoff modelling: The primer, second edition. John Wiley and Sons Ltd, Chichester
[40] Beven KJ, Binley AM (1992) The future of distributed models: Model calibration and uncertainty prediction. Hydrological Processes 6(3):279–298. https://doi.org/10.1002/hyp.3360060305
[41] Beven KJ, Binley AM (2014) GLUE: 20 years on. Hydrological Processes 28(24):5897–5918. https://doi.org/10.1002/hyp.10082
[42] Beven KJ, Freer JE (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Journal of Hydrology 249(1–4):11–29. https://doi.org/10.1016/S0022-1694(01)00421-8
[43] Beven KJ, Kirkby MJ (1979) A physically based, variable contributing area model of basin hydrology. Hydrological Sciences Bulletin 24(1):43–69. https://doi.org/10.1080/02626667909491834
[44] Biau G (2012) Analysis of a random forests model. Journal of Machine Learning Research 13(Apr):1063–1095
[45] Biau G, Scornet E (2016) A random forest guided tour. TEST 25(2):197–227. https://doi.org/10.1007/s11749-016-0481-7
[46] Billah B, Hyndman RJ, Koehler AB (2005) Empirical information criteria for time series forecasting model selection. Journal of Statistical Computation and Simulation 75(10):831–840. https://doi.org/10.1080/00949650410001687208
[47] Blöschl G, et al. (2019) Twenty-three Unsolved Problems in Hydrology (UPH) – A community perspective. Hydrological Sciences Journal 64(1):1141–1158. https://doi.org/10.1080/02626667.2019.1620507
[48] Bock AR, Farmer WH, Hay LE (2018) Quantifying uncertainty in simulated streamflow and runoff from a continental-scale monthly water balance model. Advances in Water Resources 122:166–175. https://doi.org/10.1016/j.advwatres.2018.10.005
[49] Bogner K, Liechti K, Zappa M (2016) Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. Water 8(4):115. https://doi.org/10.3390/w8040115
[50] Bogner K, Liechti K, Zappa M (2017) Technical note: Combining quantile forecasts and predictive distributions of streamflows. Hydrology and Earth System Sciences 21:5493–5502. https://doi.org/10.5194/hess-21-5493-2017
[51] Bontempi G (2013) Machine Learning Strategies for Time Series Prediction. European Business Intelligence Summer School, Hammamet, Lecture. 2013. Available online: https://pdfs.semanticscholar.org/f8ad/a97c142b0a2b1bfe20d8317ef58527ee329a.pdf (accessed on 12 September 2018)
[52] Bourgin F, Andréassian V, Perrin C, Oudin L (2015) Transferring global uncertainty estimates from gauged to ungauged catchments. Hydrology and Earth System Sciences 19:2535–2546. https://doi.org/10.5194/hess-19-2535-2015
[53] Box GEP, Cox DR (1964) An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological) 26(2):211–252. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x
[54] Box GEP, Jenkins GM (1968) Some recent advances in forecasting and control. Journal of the Royal Statistical Society. Series C (Applied Statistics) 17(2):91–109. https://doi.org/10.2307/2985674

[55] Brath A, Montanari A, Toth E (2002) Neural networks and non-parametric methods for improving real-time flood forecasting through conceptual hydrological models. Hydrology and Earth System Sciences 6(4):627–639. https://doi.org/10.5194/hess-6-627-2002
[56] Breiman L (1996) Bagging predictors. Machine Learning 24(2):123–140. https://doi.org/10.1007/BF00058655
[57] Breiman L (2001a) Random Forests. Machine Learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324
[58] Breiman L (2001b) Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical Science 16(3):199–231
[59] Brooks SP, Gelman A (1998) General methods for monitoring convergence of iterative simulations. Journal of Computational and Graphical Statistics 7(4):434–455
[60] Brown RG (1959) Statistical forecasting for inventory control. McGraw-Hill, New York
[61] Brownrigg R, Minka TP, Deckmyn A (2018) `maps`: Draw Geographical Maps. `R` package version 3.3.0. https://CRAN.R-project.org/package=maps
[62] Bühlmann P, Hothorn T (2007) Boosting algorithms: Regularization, prediction and model fitting. Statistical Science 22(4):477–505. https://doi.org/10.1214/07-STS242
[63] Cannon AJ (2011) Quantile regression neural networks: Implementation in `R` and application to precipitation downscaling. Computers and Geosciences 37(9):1277–1284. https://doi.org/10.1016/j.cageo.2010.07.005
[64] Cannon AJ (2019) `qrnn`: Quantile Regression Neural Network. `R` package version 2.0.4. https://cran.r-project.org/web/packages/qrnn
[65] Carlson RF, MacCormick AJA, Watts DG (1970) Application of linear random models to four annual streamflow series. Water Resources Research 6(4):1070–1078. https://doi.org/10.1029/WR006i004p01070
[66] Ceola S, Arheimer B, Baratti E, Blöschl G, Capell R, Castellarin A, Freer J, Han D, Hrachowitz M, Hundecha Y, et al. (2015) Virtual laboratories: New opportunities for collaborative water science. Hydrology and Earth System Sciences 19(4):2101–2117. https://doi.org/10.5194/hess-19-2101-2015
[67] Chatfield C (1988) What is the 'best' method of forecasting?. Journal of Applied Statistics 15(1):19–38. https://doi.org/10.1080/02664768800000003
[68] Chau KW, Wu CL (2010) A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. Journal of Hydroinformatics 12(4):458-473. https://doi.org/10.2166/hydro.2010.032
[69] Chawsheen TA, Broom M (2017) Seasonal time-series modeling and forecasting of monthly mean temperature for decision making in the Kurdistan Region of Iraq. Journal of Statistical Theory and Practice 11(4):604-633. https://doi.org/10.1080/15598608.2017.1292484
[70] Chen XY, Chau KW, Busari AO (2015) A comparative study of population-based optimization algorithms for downstream river flow forecasting by a hybrid neural network model. Engineering Applications of Artificial Intelligence 46(Part A):258-268 https://doi.org/10.1016/j.engappai.2015.09.010
[71] Cheng CT, Xie JX, Chau KW, Layeghifard M (2008) A new indirect multi-step-ahead prediction model for a long-term hydrologic prediction. Journal of Hydrology 361(1–2):118–130. https://doi.org/10.1016/j.jhydrol.2008.07.040
[72] Cheng KS, Lien YT, Wu YC, Su YF (2017) On the criteria of model performance evaluation for real-time flood forecasting. Stochastic Environmental Research and Risk Assessment 31(5):1123–1146. https://doi.org/10.1007/s00477-016-1322-7
[73] Chevillon G (2007) Direct multi-step estimation and forecasting. Journal of Economic Surveys 21(4):746–785. https://doi.org/10.1111/j.1467-6419.2007.00518.x
[74] Clark MP, Slater AG, Rupp DE, Woods RA, Vrugt JA, Gupta HV, Wagener T, Hay LE (2008) Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research 44(12):W00B02. https://doi.org/10.1029/2007WR006735

[75] Clark MP, Nijssen B, Lundquist JD, Kavetski D, Rupp DE, Woods RA, Freer GF, Gutmann ED, Wood AW, Brekke LD, Arnold JR, Gochis DJ, Rasmussen RM (2015) A unified approach for process-based hydrologic modeling: 1. Modeling concept. Water Resources Research 51(4):2498–2514. https://doi.org/10.1002/2015WR017198

[76] Coron L, Thirel G, Delaigue O, Perrin C, Andréassian V (2017) The suite of lumped GR hydrological models in an `R` package. Environmental Modelling and Software 94:166–171. https://doi.org/10.1016/j.envsoft.2017.05.002

[77] Coron L, Delaigue O, Thirel G, Perrin C, Michel C (2019) `airGR`: Suite of GR Hydrological Models for Precipitation-Runoff Modelling. `R` package version 1.3.2.23. https://CRAN.R-project.org/package=airGR

[78] Cortez P (2010) Data mining with neural networks and support vector machines using the `R`/`rminer` tool. In: Perner P (ed) Advances in Data Mining. Applications and Theoretical Aspects. Springer Berlin Heidelberg, pp 572–583. https://doi.org/10.1007/978-3-642-14400-4_44

[79] Cortez P (2016) `rminer`: Data Mining Classification and Regression Methods. `R` package version 1.4.2. https://CRAN.R-project.org/package=rminer

[80] Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3):273–297. https://doi.org/10.1007/BF00994018

[81] Cox DR, Kartsonaki C, Keogh RH (2018) Big data: Some statistical issues. Statistics and Probability Letters 136:111–115. https://doi.org/10.1016/j.spl.2018.02.015

[82] Coxon G, Freer J, Westerberg IK, Wagener T, Woods R, Smith PJ (2015) A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations. Water Resources Research 51(7):5531–5546. https://doi.org/10.1002/2014WR016532

[83] Criss RE, Winston WE (2008) Do Nash values have value? Discussion and alternate proposals. Hydrological Processes 22:2723–2725. https://doi.org/10.1002/hyp.7072

[84] De Gooijer JG, Hyndman RJ (2006) 25 years of time series forecasting. International Journal of Forecasting 22(3):443–473. https://doi.org/10.1016/j.ijforecast.2006.01.001

[85] De Gooijer JG, Klein A (1992) On the cumulated multi-step-ahead predictions of vector autoregressive moving average processes. International Journal of Forecasting 7(4):501–513. https://doi.org/10.1016/0169-2070(92)90034-7

[86] De Gooijer JG, Kumar K (1992) Some recent developments in non-linear time series modelling, testing, and forecasting. International Journal of Forecasting 8(2):135–156. https://doi.org/10.1016/0169-2070(92)90115-P

[87] De Livera AM, Hyndman RJ, Snyder RS (2011) Forecasting time series with complex seasonal patterns using exponential smoothing. Journal of the American Statistical Association 106(496):1513–1527. https://doi.org/10.1198/jasa.2011.tm09771

[88] De Vos NJ (2013) Echo state networks as an alternative to traditional artificial neural networks in rainfall–runoff modelling. Hydrology and Earth System Sciences 17:253–267. https://doi.org/10.5194/hess-17-253-2013

[89] Di Baldassarre G, Montanari A (2009) Uncertainty in river discharge observations: A quantitative analysis. Hydrology and Earth System Sciences 13:913–921. https://doi.org/10.5194/hess-13-913-2009

[90] Di Baldassarre G, Laio F, Montanari A (2012) Effect of observation errors on the uncertainty of design floods. Physics and Chemistry of the Earth, Parts A/B/C 42–44:85–90. https://doi.org/10.1016/j.pce.2011.05.001

[91] Dibike YB, Solomatine DP (2001) River flow forecasting using artificial neural networks. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 26(1):1–7. https://doi.org/10.1016/S1464-1909(01)85005-X

[92] Dogulu N, López López P, Solomatine DP, Weerts AH, Shrestha DL (2015) Estimation of predictive hydrologic uncertainty using the quantile regression and UNEEC methods and their comparison on contrasting catchments. Hydrology and Earth System Sciences 19:3181–3201. https://doi.org/10.5194/hess-19-3181-2015

[93] Dooley LM (2002) Case study research and theory building. Advances in Developing Human Resources 4(3):335–354. https://doi.org/10.1177/1523422302043007
[94] Dowle M, Srinivasan A (2019) `data.table`: Extension of `'data.frame'`. `R` package version 1.12.2. https://CRAN.R-project.org/package=data.table
[95] Duan Q, Schaake J, Andreassian V, Franks S, Gotetie G, Gupta HV, Gusev YM, Habets F, Hall A, et al. (2006) Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. Journal of Hydrology 320(1–2):3–17. https://doi.org/10.1016/j.jhydrol.2005.07.031
[96] Dunsmore IR (1968) A Bayesian approach to calibration. Journal of the Royal Statistical Society. Series B (Methodological) 30(2):396–405. https://doi.org/10.1016/j.rser.2018.05.038
[97] Edijatno, Nascimento NO, Yang X, Makhlouf Z, Michel C (1999) GR3J: A daily watershed model with three free parameters. Hydrological Sciences Journal 44(2):263–277. https://doi.org/10.1080/02626669909492221
[98] Efron B, Hastie T (2016) Computer age statistical inference, first edition. Cambridge University Press: New York, ISBN 9781107149892
[99] Efstratiadis A, Koutsoyiannis D (2010) One decade of multi-objective calibration approaches in hydrological modelling: A review. Hydrological Sciences Journal 55(1):58–78. https://doi.org/10.1080/02626660903526292
[100] Efstratiadis A, Nalbantis I, Koukouvinos A, Rozos E, Koutsoyiannis D (2008) HYDROGEIOS: A semi-distributed GIS-based hydrological model for modified river basins. Hydrology and Earth System Sciences 12:989–1006. https://doi.org/10.5194/hess-12-989-2008
[101] El-Shafie A, Taha MR, Noureldin A (2007) A neuro-fuzzy model for inflow forecasting of the Nile river at Aswan high dam. Water Resources Management 21(3):533–556. https://doi.org/10.1007/s11269-006-9027-1
[102] Emiliano PC, Vivanco MJ, De Menezes FS (2014) Information criteria: How do they behave in different models? Computational Statistics & Data Analysis 69:141–153. https://doi.org/10.1016/j.csda.2013.07.032
[103] Evin G, Kavetski D, Thyer M, Kuczera G (2013) Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. Water Resources Research 49(7):4518–4524. https://doi.org/10.1002/wrcr.20284
[104] Evin G, Thyer M, Kavetski D, McInerney D, Kuczera G (2014) Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resources Research 50(3):2350–2375. https://doi.org/10.1002/2013WR014185
[105] Farmer WH, Vogel RM (2016) On the deterministic and stochastic use of hydrologic models. Water Resources Research 52(7):5619–5633. https://doi.org/10.1002/2016WR019129
[106] Fildes R (1992) The evaluation of extrapolative forecasting methods. International Journal of Forecasting 8(1):81–98. https://doi.org/10.1016/0169-2070(92)90009-X
[107] Fildes R, Kourentzes N (2011) Validation and forecasting accuracy in models of climate change. International Journal of Forecasting 27(4):968-995. https://doi.org/10.1016/j.ijforecast.2011.03.008
[108] Fiseha BM, Setegn SG, Melesse AM, Volpi E, Fiori A (2013) Hydrological analysis of the Upper Tiber River Basin, Central Italy: A watershed modelling approach. Hydrological Processes 27(16):2339–2351. https://doi.org/10.1002/hyp.9234
[109] Fraley C, Leisch F, Maechler M, Reisen V, Lemonte A (2012) `fracdiff`: Fractionally differenced ARIMA aka ARFIMA(*p*,*d*,*q*) models. `R` package version 1.4-2. https://CRAN.R-project.org/package=fracdiff
[110] Franses PH, Legerstee R (2010) A unifying view on multi-step forecasting using an autoregression. Journal of Economic Surveys 24(3):389–401. https://doi.org/10.1111/j.1467-6419.2009.00581.x
[111] Friedman JH (2001) Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5):1189–1232. https://doi.org/10.1214/aos/1013203451

[112] Frigg R, Hartmann S (2018) Models in science. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy (Summer 2018 Edition)
[113] Gagolewski M (2019) `stringi`: Character String Processing Facilities. `R` package version 1.4.3. https://CRAN.R-project.org/package=stringi
[114] Gardner Jr ES (1985) Exponential smoothing: The state of the art. Journal of Forecasting 4(1):1–28. https://doi.org/10.1002/for.3980040103
[115] Gardner Jr ES (2006) Exponential smoothing: The state of the art—Part II. International Journal of Forecasting 22(4):637–666. https://doi.org/10.1016/j.ijforecast.2006.03.005
[116] Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Statistical Science 7(4):457–472
[117] Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis, second edition. Chapman and Hall/CRC
[118] Gholami V, Chau KW, Fadaee F, Torkaman J, Ghaffari A (2015) Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. Journal of Hydrolology 529(Part 3):1060–1069. https://doi.org/10.1016/j.jhydrol.2015.09.028
[119] Giunta G, Salerno R, Ceppi A, Ercolani G, Mancini M (2015) Benchmark analysis of forecasted seasonal temperature over different climatic areas. Geoscience Letters 2(9). https://doi.org/10.1186/s40562-015-0026-z
[120] Gneiting T, Katzfuss M (2014) Probabilistic forecasting. Annual Review of Statistics and Its Application 1:125–51. http://dx.doi.org/10.1146/annurev-statistics-062713-085831
[121] Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and stimation. Journal of the American Statistical Association 102(477):359–378. https://doi.org/10.1198/016214506000001437
[122] Gneiting T, Balabdaoui F, Raftery AE (2007) Probabilistic forecasts, calibration and sharpness. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69(2):243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x
[123] Goldfarb RS, Ratner J (2008) "Theory" and "models": Terminology through the looking glass. Econ Journal Watch 5(1):91–108
[124] Granger CWJ (1989) Invited review combining forecasts—Twenty years later. Journal of Forecasting 8(3):167–173. https://doi.org/10.1002/for.3980080303
[125] GRDC (2017) Long-term statistics and annual characteristics of GRDC timeseries data. Online provided by the Global Runoff Data Centre of WMO. Koblenz: Federal Institute of Hydrology (BfG). [Date of retrieval:2018-01-06]. http://www.bafg.de/GRDC/EN/03_dtprdcts/32_LTMM/longtermstat_node.html
[126] Green KC, Armstrong JS (2007) Global warming: Forecasts by scientists versus scientific forecasts. Energy Environ 18(7):997-1021. https://doi.org/10.1260/095830507782616887
[127] Green KC, Armstrong JS, Soon W (2009) Validity of climate change forecasting for public policy decision making. International Journal of Forecasting 25(4):826–832. https://doi.org/10.1016/j.ijforecast.2009.05.011
[128] Greenwell B, Boehmke B, Cunningham J, GBM Developers (2019) `gbm`: Generalized Boosted Regression Models. `R` package version 2.1.5. https://cran.r-project.org/web/packages/gbm
[129] Guerrero VM (1993) Time-series analysis supported by power transformations. Journal of Forecasting 12(1):37–48. https://doi.org/10.1002/for.3980120104
[130] Guo J, Zhou J, Qin H, Zou Q, Li Q (2011) Monthly streamflow forecasting based on improved support vector machine model. Expert Systems with Applications 38(10):13073–13081. https://doi.org/10.1016/j.eswa.2011.04.114
[131] Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology 377(1–2):80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003
[132] Gupta HV, Perrin C, Blöschl G, Montanari A, Kumar R, Clark MP, Andréassian V (2014) Large-sample hydrology: A need to balance depth with breadth. Hydrology and Earth System Sciences 18:463–477. https://doi.org/10.5194/hess-18-463-2014

[133] Haario H, Laine M, Mira A, Saksman E (2006) DRAM: Efficient adaptive MCMC. Statistics and Computing 16(4):339–354. https://doi.org/10.1007/s11222-006-9438-0
[134] Hartmann S (1995) Models as a tool for theory construction: Some strategies of preliminary physics. In: Herfel W, Krajewski W, Niiniluoto I, Wójcicki R (eds) Theories and Models in Scientific Processes, pp. 49–67
[135] Harvey AC (1984) A unified view of statistical forecasting procedures. Journal of Forecasting 3(3):245–275. https://doi.org/10.1002/for.3980030302
[136] Harvey A, Peters S (1990) Estimation procedures for structural time series models. Journal of Forecasting 9(2):89–108. https://doi.org/10.1002/for.3980090203
[137] Haslett J, Raftery AE (1989) Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource. Journal of the Royal Statistical Society, Series C (Applied Statistics) 38(1):1–50. https://doi.org/10.2307/2347679
[138] Hastie T, Tibshirani R (1987) Generalized additive models: Some applications. Journal of the American Statistical Association 82(398):371-386. https://doi.org/10.1080/01621459.1987.10478440
[139] Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: Data mining, inference and prediction, second edition. Springer, New York. https://doi.org/10.1007/978-0-387-84858-7
[140] He Z, Wen X, Liu H, Du J (2014) A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flown in the semiarid mountain region. Journal of Hydrology 509:379–386. https://doi.org/10.1016/j.jhydrol.2013.11.054
[141] Hernández-López MR, Francés F (2017) Bayesian joint inference of hydrological and generalized error models with the enforcement of Total Laws. Hydrology and Earth System Sciences Discussions. https://doi.org/10.5194/hess-2017-9
[142] Hofner B, Mayr A, Robinzonov N, Schmid M (2014) Model-based boosting in R: A hands-on tutorial using the R package mboost. Computational Statistics 29(1–2):3–35. https://doi.org/10.1007/s00180-012-0382-5
[143] Holt CC (2004) Forecasting seasonals and trends by exponentially weighted moving averages. International Journal of Forecasting 20(1):5–10. https://doi.org/10.1016/j.ijforecast.2003.09.015
[144] Hong WC (2008) Rainfall forecasting by technological machine learning models. Applied Mathematics and Computation 200(1):41–57. https://doi.org/10.1016/j.amc.2007.10.046
[145] Hong T, Fan S (2016) Probabilistic electric load forecasting: A tutorial review. International Journal of Forecasting 32(3):914–938. https://doi.org/10.1016/j.ijforecast.2015.11.011
[146] Hothorn T, Leisch F, Zeileis A, Hornik K (2005) The design and analysis of benchmark experiments. Journal of Computational and Graphical Statistics 14(3):675–699. https://doi.org/10.1198/106186005X59630
[147] Hothorn T, Buehlmann P, Kneib T, Schmid M, Hofner B (2018) mboost: Model-Based Boosting. R package version 2.9-1. https://cran.r-project.org/web/packages/mboost
[148] Htike KK, Khalifa OO (2010) Rainfall forecasting models using focused time-delay neural networks. 2010 International Conference on Computer and Communication Engineering (ICCCE). https://doi.org/10.1109/ICCCE.2010.5556806
[149] Hu J, Liu J, Liu Y, Gao C (2001) EMD-KNN model for annual average rainfall forecasting. Journal of Hydrologic Engineering 18(11):1450–1457. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000481
[150] Huang M, Hou Z, Leung LR, Ke Y, Liu Y, Fang Z, Sun Y (2013) Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins. Journal of Hydrometeorology 14(6):1754–1772. https://doi.org/10.1175/JHM-D-12-0138.1

[151] Huard D, Mailhot A (2008) Calibration of hydrological model GR2M using Bayesian uncertainty analysis. Water Resources Research 44(2):W02424. https://doi.org/10.1029/2007WR005949
[152] Humphrey GB, Maier HR, Wu W, Mount NJ, Dandy GC, Abrahart RJ, Dawson CW (2017) Improved validation framework and R-package for artificial neural network models. Environmental Modelling and Software 92:82–106. https://doi.org/10.1016/j.envsoft.2017.01.023
[153] Hung NQ, Babel MS, Weesakul S, Tripathi NK (2009) An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrology and Earth System Sciences 13:1413–1425. https://doi.org/10.5194/hess-13-1413-2009
[154] Hurvich CM, Tsai CL (1993) A corrected Akaike information criterion for vector autoregressive model selection. Journal of Time Series Analysis 14(3):271–279. https://doi.org/10.1111/j.1467-9892.1993.tb00144.x
[155] Hutter F, Lücke J, Schmidt-Thieme L (2015) Beyond manual tuning of hyperparameters. KI - Künstliche Intelligenz 29(4):329–337. https://doi.org/10.1007/s13218-015-0381-0
[156] Hyndman RJ, Athanasopoulos G (2018) Forecasting: Principles and Practice. OTexts: Melbourne, Australia. https://otexts.org/fpp2
[157] Hyndman RJ, Billah B (2003) Unmasking the Theta method. International Journal of Forecasting 19(2):287–290. https://doi.org/10.1016/S0169-2070(01)00143-1
[158] Hyndman RJ, Khandakar Y (2008) Automatic time series forecasting: The forecast package for R. Journal of Statistical Software 27(3):1–22. https://doi.org/10.18637/jss.v027.i03
[159] Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. International Journal of Forecasting 22(4):679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001
[160] Hyndman RJ, Koehler AB, Snyder RD, Grose S (2002) A state space framework for automatic forecasting using exponential smoothing methods. International Journal of Forecasting 18(3):439–454. https://doi.org/10.1016/S0169-2070(01)00110-8
[161] Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2005) Prediction intervals for exponential smoothing using two new classes of state space models. Journal of Forecasting 24(1):17-37. https://doi.org/10.1002/for.938
[162] Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) Forecasting with exponential smoothing: The state space approach. Springer - Verlag Berlin Heidelberg, pp 3–7. https://doi.org/10.1007/978-3-540-71918-2
[163] Hyndman RJ, Athanasopoulos G, Bergmeir C, Caceres G, Chhay L, O'Hara-Wild M, Petropoulos F, Razbash S, Wang E, Yasmeen F (2018) forecast: Forecasting Functions for Time Series and Linear Models. R package version 8.4. https://CRAN.R-project.org/package=forecast
[164] Jain SK, Das A, Srivastava DK (1999) Application of ANN for reservoir inflow prediction and operation. Journal of Water Resources Planning and Management 125(5):263–271. https://doi.org/10.1061/(ASCE)0733-9496(1999)125:5(263)
[165] James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer-Verlag New York. https://doi.org/10.1007/978-1-4614-7138-7
[166] Jayawardena AW, Fernando DAK (1998) Use of radial basis function type artificial neural networks for runoff simulation. Computer-Aided Civil and Infrastructure Engineering 13(2):91–99. https://doi.org/10.1111/0885-9507.00089
[167] Jayawardena AW, Zhou MC (2000) A modified spatial soil moisture storage capacity distribution curve for the Xinanjiang model. Journal of Hydrology 227(1–4):93–113. https://doi.org/10.1016/S0022-1694(99)00173-0
[168] Juston JM, Kauffeldt A, Montano BQ, Seibert J, Beven KJ, Westerberg IK (2012) Smiling in the rain: Seven reasons to be positive about uncertainty in hydrological modelling. Hydrological Processes 27(7):1117–1122. https://doi.org/10.1002/hyp.9625
[169] Kaleris V, Langousis A (2017) Comparison of two rainfall–runoff models: Effects of conceptualization on water budget components. Hydrological Sciences Journal 62(5):729–748. https://doi.org/10.1080/02626667.2016.1250899

[170] Kalman RE (1960) A new approach to linear filtering and prediction problems. Journal of Basic Engineering 82(Series D):35–45. https://doi.org/10.1115/1.3662552
[171] Karatzoglou A, Smola A, Hornik K, Zeileis A (2004) `kernlab` – An S4 Package for Kernel Methods in `R`. Journal of Statistical Software 11(9):1–20. https://doi.org/10.18637/jss.v011.i09
[172] Karatzoglou A, Smola A, Hornik K (2018) `kernlab`: Kernel-Based Machine Learning Lab. `R` package version 0.9-27. https://CRAN.R-project.org/package=kernlab
[173] Kassambara A (2019) `ggpubr`: '`ggplot2`' Based Publication Ready Plots. `R` package version 0.2.1. https://cran.r-project.org/web/packages/ggpubr
[174] Kashyap RL (1982) Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Transactions on Pattern Analysis and Machine Intelligence 4(2):99–104. https://doi.org/10.1109/TPAMI.1982.4767213
[175] Kauffeldt A, Halldin S, Rodhe A, Xu C-Y, Westerberg IK (2013) Disinformative data in large-scale hydrological modelling. Hydrology and Earth System Sciences 17:2845–2857. https://doi.org/10.5194/hess-17-2845-2013
[176] Kavetski D, Franks SW, Kuczera G (2002) Confronting input uncertainty in environmental modelling. In: Duan Q, Gupta HV, Sorooshian S, Rousseau AN, Turcotte R (eds) Calibration of Watershed Models. AGU, pp 49–68. https://doi.org/10.1029/WS006p0049
[177] Kavetski D, Kuczera G, Franks SW (2006a) Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research 42(3):W03407. https://doi.org/10.1029/2005WR004368
[178] Kavetski D, Kuczera G, Franks SW (2006b) Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis. Journal of Hydrology 320(1–2):187–201. https://doi.org/10.1016/j.jhydrol.2005.07.013
[179] Keenlyside NS (2011) Commentary on "Validation and forecasting accuracy in models of climate change". International Journal of Forecasting 27(4):1000-1003. https://doi.org/10.1016/j.ijforecast.2011.07.002
[180] Kelly KS, Krzysztofowicz R (1997) A bivariate meta-Gaussian density for use in hydrology. Stochastic Hydrology and Hydraulics 11(1):17–31. https://doi.org/10.1007/BF02428423
[181] Kelly KS, Krzysztofowicz R (2000) Precipitation uncertainty processor for probabilistic river stage forecasting. Water Resources Research 36:2643–2653. https://doi.org/10.1029/2000WR900061
[182] Khan MS, Coulibaly P (2006) Application of support vector machine in lake water level prediction. Journal of Hydrologic Engineering 11(3):199–205. https://doi.org/10.1061/(ASCE)1084-0699(2006)11:3(199)
[183] Khatami S, Peel MC, Peterson TJ, Western AW (2019) Equifinality and Flux Mapping: A new approach to model evaluation and process representation under uncertainty. Water Resources Research 55(11):8922–8941. https://doi.org/10.1029/2018WR023750
[184] Kim KB, Kwon HH, Han D (2018) Exploration of warm-up period in conceptual hydrological modelling. Journal of Hydrology 556:194–210. https://doi.org/10.1016/j.jhydrol.2017.11.015
[185] Kim TW, Valdés JB (2003) Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. Journal of Hydrologic Engineering 8(6):319–328. https://doi.org/10.1061/(ASCE)1084-0699(2003)8:6(319)
[186] Kişi Ö (2004) River flow modeling using artificial neural networks. Journal of Hydrologic Engineering 9(1):60–63. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:1(60)
[187] Kişi Ö (2007) Streamflow forecasting using different artificial neural network algorithms. Journal of Hydrologic Engineering 12(5):532–539. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(532)
[188] Kişi Ö, Cimen M (2011) A wavelet-support vector machine conjunction model for monthly streamflow forecasting. Journal of Hydrology 399(1–2):132–140. https://doi.org/10.1016/j.jhydrol.2010.12.041

[189] Kişi Ö, Cimen M (2012) Precipitation forecasting by using wavelet-support vector machine conjunction model. Engineering Applications of Artificial Intelligence 25(4):783–792. https://doi.org/10.1016/j.engappai.2011.11.003
[190] Kişi Ö, Shiri J, Nikoofar B (2012) Forecasting daily lake levels using artificial intelligence approaches. Computers and Geosciences 41:169–180. https://doi.org/10.1016/j.cageo.2011.08.027
[191] Kitanidis PK, Bras RL (1980) Real time forecasting with a conceptual hydrologic model: 2. Applications and results. Water Resources Research 16(6):1034–1044. https://doi.org/10.1029/WR016i006p01034
[192] Klein DB, Romero PP (2007) Model building versus theorizing: The paucity of theory in the Journal of Economic Theory. Econ Journal Watch 4(2):241–271
[193] Klemeš V (1986) Operational testing of hydrological simulation models. Hydrological Sciences Journal 31:13–24. https://doi.org/10.1080/02626668609491024
[194] Koenker RW (2005) Quantile regression. Cambridge University Press, Cambridge, UK
[195] Koenker RW (2017) Quantile regression: 40 years on. Annual Review of Economics 9(1):155–176. https://doi.org/10.1146/annurev-economics-063016-103651
[196] Koenker RW (2019) `quantreg`: Quantile Regression. `R` package version 5.51. https://CRAN.R-project.org/package=quantreg
[197] Koenker RW, Bassett Jr G (1978) Regression quantiles. Econometrica 46(1):33–50. https://doi.org/10.2307/1913643
[198] Koenker RW, D'Orey V (1987) Computing regression quantiles. Journal of the Royal Statistical Society: Series C (Applied Statistics) 36(3):383–393. https://doi.org/10.2307/2347802
[199] Koenker RW, D'Orey V (1994) A remark on algorithm AS 229: Computing dual regression quantiles and regression rank scores. Journal of the Royal Statistical Society: Series C (Applied Statistics) 43(2):410–414. https://doi.org/10.2307/2986030
[200] Koenker RW, Hallock K (2001) Quantile regression: An introduction. Journal of Economic Perspectives, 15(4):43–56
[201] Koenker RW, Machado JAF (1999) Goodness of fit and related inference processes for quantile regression. Journal of the American Statistical Association 94(448):1296–1310. https://doi.org/10.1080/01621459.1999.10473882
[202] Kohavi R, John GH (1997) Wrappers for feature subset selection. Artificial Intelligence 97(1–2):273–324. https://doi.org/10.1016/S0004-3702(97)00043-X
[203] Komorník J, Komorníková M, Mesiar R, Szökeová D, Szolgay J (2006) Comparison of forecasting performance of nonlinear models of hydrological time series. Physics and Chemistry of the Earth, Parts A/B/C 31(18):1127-1145. https://doi.org/10.1016/j.pce.2006.05.006
[204] Kottek M, Grieser J, Beck C, Rudolf B, Rubel F (2006) World map of the Köppen-Geiger climate classification updated. Meteorologische Zeitschrift 15(3):259–263. https://doi.org/10.1127/0941-2948/2006/0130
[205] Koutsoyiannis D (2006) A toy model of climatic variability with scaling behaviour. Journal of Hydrology 322(1–4):25–48. https://doi.org/10.1016/j.jhydrol.2005.02.030
[206] Koutsoyiannis D (2008) Probability and statistics for geophysical processes. https://doi.org/10.13140/RG.2.1.2300.1849/1
[207] Koutsoyiannis D (2010) HESS Opinions "A random walk on water". Hydrology and Earth System Sciences 14:585–601. https://doi.org/10.5194/hess-14-585-2010
[208] Koutsoyiannis D (2011) Hurst-Kolmogorov Dynamics and Uncertainty. Journal of American Water Resources Association 47(3):481–495. https://doi.org/10.1111/j.1752-1688.2011.00543.x
[209] Koutsoyiannis D, Montanari A (2007) Statistical analysis of hydroclimatic time series: Uncertainty and insights. Water Resources Research 43(5):W05429, https://doi.org/10.1029/2006WR005592

[210] Koutsoyiannis D, Montanari A (2015) Negligent killing of scientific concepts: The stationarity case. Hydrological Sciences Journal 60(7–8):1174–1183. https://doi.org/10.1080/02626667.2014.959959
[211] Koutsoyiannis D, Yao H, Georgakakos A (2008) Medium-range flow prediction for the Nile: A comparison of stochastic and deterministic methods. Hydrological Sciences Journal 53(1):142–164. https://doi.org/10.1623/hysj.53.1.142
[212] Koutsoyiannis D, Makropoulos C, Langousis A, Baki S, Efstratiadis A, Christofides A, Karavokiros G, Mamassis N (2009) HESS Opinions: "Climate, hydrology, energy, water: recognizing uncertainty and seeking sustainability". Hydrology and Earth System Sciences 13:247–257. https://doi.org/10.5194/hess-13-247-2009
[213] Krause P, Boyle DP, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. Advances in Geosciences 5:89–97. https://doi.org/10.5194/adgeo-5-89-2005
[214] Krzysztofowicz R (1999) Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resources Research 35(9):2739–2750. https://doi.org/10.1029/1999WR900099
[215] Krzysztofowicz R (2001a) Integrator of uncertainties for probabilistic river stage forecasting: Precipitation-dependent model. Journal of Hydrology 249:69–85. https://doi.org/10.1016/S0022-1694(01)00413-9
[216] Krzysztofowicz R (2001b) The case for probabilistic forecasting in hydrology. Journal of Hydrology 249(1–4):2–9. https://doi.org/10.1016/S0022-1694(01)00420-6
[217] Krzysztofowicz R (2002) Bayesian system for probabilistic river stage forecasting. Journal of Hydrology 268:16–40. https://doi.org/10.1016/S0022-1694(02)00106-3
[218] Krzysztofowicz R, Herr HD (2001) Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model. Journal of Hydrology 249:46–68. https://doi.org/10.1016/S0022-1694(01)00412-7
[219] Krzysztofowicz R, Kelly KS (2000) Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resources Research 36:3265–3277. https://doi.org/10.1029/2000WR900108
[220] Kuczera G (1983) Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. Water Resources Research 19(5):1151–1162. https://doi.org/10.1029/WR019i005p01151
[221] Kuczera G, Kavetski D, Franks S, Thyer M (2006) Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology 331(1–2):161–177. https://doi.org/10.1016/j.jhydrol.2006.05.010
[222] Kuczera G, Renard B, Thyer M, Kavetski D (2010) There are no hydrological monsters, just models and observations with large uncertainties!. Hydrological Sciences Journal 55(6):980–991. https://doi.org/10.1080/02626667.2010.504677
[223] Kumar DN, Raju KS, Sathish T (2004) River flow forecasting using recurrent neural networks. Water Resources Management 18(2):143–161. https://doi.org/10.1023/B:WARM.0000024727.94701.12
[224] Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics 54(1–3):159–178. https://doi.org/10.1016/0304-4076(92)90104-Y
[225] Laloy E, Vrugt JA (2012) High-dimensional posterior exploration of hydrologic models using multiple-try DREAM$_{(ZS)}$ and high-performance computing. Water Resources Research 48(1):W01526. https://doi.org/10.1029/2011WR010608
[226] Lambrakis N, Andreou AS, Polydoropoulos P, Georgopoulos E, Bountis T (2000) Nonlinear analysis and forecasting of a brackish karstic spring. Water Resources Research 36(4):875–884. https://doi.org/10.1029/1999WR900353
[227] Lanc TL (1992) The importance of input variables to a neural network fault-diagnostic system for nuclear power plants. MSc thesis. https://lib.dr.iastate.edu/rtd/208

[228] Langousis A, Mamalakis A, Puliga M, Deida R (2016) Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. Water Resources Research 52(4):2659–2681. https://doi.org/10.1002/2015WR018502
[229] Larsson R (1993) Case survey methodology: Quantitative analysis of patterns across case studies. Academy of Management Journal 36(6):1515–1546. https://doi.org/10.2307/256820
[230] Lawrimore JH, Menne MJ, Gleason BE, Williams CN, Wuertz DB, Vose RS, Rennie J (2011) An overview of the Global Historical Climatology Network monthly mean temperature data set, version 3. Journal of Geophysical Research Atmospheres 116(D1912). https://doi.org/10.1029/2011JD016187
[231] Lebecherel L, Andréassian V, Perrin C (2016) On evaluating the robustness of spatial-proximity-based regionalization methods. Journal of Hydrology 539:196–203. https://doi.org/10.1016/j.jhydrol.2016.05.031
[232] Legates DR, McCabe Jr GJ (1999) Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resources Research 35(1):233–241. https://doi.org/10.1029/1998WR900018
[233] Li W, Duan Q, Miao C, Ye A, Gong W, Di Z (2017) A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. Wiley Interdisciplinary Reviews: Water 4(6):e1246. https://doi.org/10.1002/wat2.1246
[234] Liaw A (2018) `randomForest`: Breiman and Cutler's Random Forests for Classification and Regression. `R` package version 4.6-14. https://CRAN.R-project.org/package=randomForest
[235] Liaw A, Wiener M (2002) Classification and Regression by `randomForest`. `R` News 2(3):18–22
[236] Lichtendahl Jr KC, Grushka-Cockayne Y, Winkler RL (2013) Is it better to average probabilities or quantiles?. Management Science 59(7):1594–1611. https://doi.org/10.1287/mnsc.1120.1667
[237] Lin JY, Cheng CT, Chau KW (2006) Using support vector machines for long-term discharge prediction. Hydrological Sciences Journal 51(4):599–612. https://doi.org/10.1623/hysj.51.4.599
[238] Lippmann R (1987) An introduction to computing with neural nets. IEEE ASSP Magazine 4(2):4–22. https://doi.org/10.1109/MASSP.1987.1165576
[239] Liong SY, Sivapragasam C (2002) Flood stage forecasting with support vector machines. Journal of American Water Resources Association 38(1):173–186. https://doi.org/10.1111/j.1752-1688.2002.tb01544.x
[240] Liu B, Nowotarski J, Hong T, Weron R (2017) Probabilistic load forecasting via quantile regression averaging on sister forecasts. IEEE Transactions on Smart Grid 8(2):730–737. https://doi.org/10.1109/TSG.2015.2437877
[241] López López P, Verkade JS, Weerts AH, Solomatine DP (2014) Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: A comparison. Hydrology and Earth System Sciences 18:3411–3428. https://doi.org/10.5194/hess-18-3411-2014
[242] Louvet S, Paturel JE, Mahé G, Rouché N, Koité M (2016) Comparison of the spatiotemporal variability of rainfall from four different interpolation methods and impact on the result of GR2M hydrological modeling—Case of Bani River in Mali, West Africa. Theoretical and Applied Climatology 123(1–2):303–319. https://doi.org/10.1007/s00704-014-1357-y
[243] Lu K, Wang L (2011) A novel nonlinear combination model based on support vector machine for rainfall prediction. Fourth International Joint Conference on Computational Sciences and Optimization:1343–1346. https://doi.org/10.1109/CSO.2011.50
[244] Luczak J (2017) Talk about toy models. Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 57:1–7. https://doi.org/10.1016/j.shpsb.2016.11.002

[245] Luo G (2016) A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Network Modeling Analysis in Health Informatics and Bioinformatics 5:18. https://doi.org/10.1007/s13721-016-0125-6
[246] Madsen H (2000) Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. Journal of Hydrology 235(3–4):276–288. https://doi.org/10.1016/S0022-1694(00)00279-1
[247] Maier HR, Dandy GC (2000) Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling and Software 15(1):101–124. https://doi.org/10.1016/S1364-8152(99)00007-9
[248] Maier HR, Jain A, Dandy GC, Sudheer KP (2010) Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. Environmental Modelling and Software 25(8):891–909. https://doi.org/10.1016/j.envsoft.2010.02.003
[249] Makhlouf Z, Michel C (1994) A two-parameter monthly water balance model for French watersheds. Journal of Hydrology 162(3–4):299–318. https://doi.org/10.1016/0022-1694(94)90233-X
[250] Makridakis S, Hibon M (2000) The M3-Competition: Results, conclusions and implications. International Journal of Forecasting 16(4):451–476. https://doi.org/10.1016/S0169-2070(00)00057-1
[251] Makridakis S, Hibon M, Lusk E, Belhadjali M (1987) Confidence intervals: An empirical investigation of the series in the M-competition. International Journal of Forecasting 3(3–4):489–508. https://doi.org/10.1016/0169-2070(87)90045-8
[252] Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE 13(3):e0194889. https://doi.org/10.1371/journal.pone.0194889
[253] Makropoulos C, Koutsoyiannis D, Stanić M, Djordjević S, Prodanović D, Dašić T, Prohaskad S, Maksimović Č, Wheater H (2008) A multi-model approach to the simulation of large scale karst flows. Journal of Hydrology 348(3–4):412–424. https://doi.org/10.1016/j.jhydrol.2007.10.011
[254] Mamassis N, Koutsoyiannis D (1996) Influence of atmospheric circulation types in space-time distribution of intense rainfall. Journal of Geophysical Research-Atmospheres 101(D21):26267–26276. https://doi.org/10.1029/96JD01377
[255] Mantovan P, Todini E (2006) Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. Journal of Hydrology 330(1–2):368–381. https://doi.org/10.1016/j.jhydrol.2006.04.046
[256] Marsland S (2011) Machine learning: an algorithmic perspective, second edition. Chapman and Hall/CRC, New York
[257] Mayer J, Khairy K, Howard J (2010) Drawing an elephant with four complex parameters. American Journal of Physics 78(6). https://doi.org/10.1119/1.3254017
[258] Mayr A, Binder H, Gefeller O, Schmid M (2014) The evolution of boosting algorithms. Methods of Information in Medicine 53(06):419–427. https://doi.org/10.3414/ME13-01-0122
[259] McMillan H, Freer J, Pappenberger F, Krueger T, Clark MP (2010) Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. Hydrological Processes 24(10):1270–1284. https://doi.org/10.1002/hyp.7587
[260] McMillan H, Krueger T, Freer J (2012) Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. Hydrological Processes 26(26):4078–4111. https://doi.org/10.1002/hyp.9384
[261] McSharry PE (2011) Validation and forecasting accuracy in models of climate change: Comments. International Journal of Forecasting 27(4):996-999. https://doi.org/10.1016/j.ijforecast.2011.07.003
[262] Meinshausen N (2006) Quantile regression forests. Journal of Machine Learning Research 7:983–999

[263] Michel C (1991) Hydrologie appliquée aux petits bassins ruraux. Cemagref, Antony, France
[264] Millard SP (2013) `EnvStats`: An `R` Package for Environmental Statistics. Springer, New York
[265] Millard SP (2018) `EnvStats`: Package for Environmental Statistics, Including US EPA Guidance. `R` package version 2.3.1. https://CRAN.R-project.org/package=EnvStats
[266] Mills TC (2011) The Foundations of Modern Time Series Analysis. Palgrave Macmillan, UK. https://doi.org/10.1057/9780230305021
[267] Minns AW, Hall MJ (1996) Artificial neural networks as rainfall-runoff models. Hydrological Sciences Journal 41(3):399–417. https://doi.org/10.1080/02626669609491511
[268] Mishra AK, Desai VR, Singh VP (2007) Drought forecasting using a hybrid stochastic and neural network model. Journal of Hydrologic Engineering 12(6):626–638. https://doi.org/10.1061/(ASCE)1084-0699(2007)12:6(626)
[269] Moisen GG (2008) Classification and regression trees. In: Jørgensen SE, Fath BD (eds) Encyclopedia of Ecology, vol 1. Elsevier, Oxford, UK, pp 582-588
[270] Montanari A (2005) Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. Water Resources Research 41(8):W08406. https://doi.org/10.1029/2004WR003826
[271] Montanari A (2007) What do we mean by 'uncertainty'? The need for a consistent wording about uncertainty assessment in hydrology. Hydrological Processes 21(6):841–845. https://doi.org/10.1002/hyp.6623
[272] Montanari A (2011) Uncertainty of hydrological predictions. In: Wilderer PA (ed) Treatise on water science 2. Elsevier, pp 459–478. https://doi.org/10.1016/B978-0-444-53199-5.00045-2
[273] Montanari A, Brath A (2004) A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. Water Resources Research 40(1):W01106. https://doi.org/10.1029/2003WR002540
[274] Montanari A, Di Baldassarre G (2013) Data errors and hydrological modelling: The role of model structure to propagate observation uncertainty. Advances in Water Resources 51:498–504. https://doi.org/10.1016/j.advwatres.2012.09.007
[275] Montanari A, Grossi G (2008) Estimating the uncertainty of hydrological forecasts: A statistical approach. Water Resources Research 44(12):W00B08. https://doi.org/10.1029/2008WR006897
[276] Montanari A, Koutsoyiannis D (2012) A blueprint for process-based modeling of uncertain hydrological systems. Water Resources Research 48(9):W09555. https://doi.org/10.1029/2011WR011412
[277] Montanari A, Koutsoyiannis D (2014) Modeling and mitigating natural hazards: Stationarity is immortal!. Water Resources Research 50(12):9748–9756. https://doi.org/10.1002/2014WR016092
[278] Montanari A, Rosso R, Taqqu MS (1997) Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation. Water Resources Research 33(5):1035–1044. https://doi.org/10.1029/97WR00043
[279] Montanari A, Taqqu MS, Teverovsky V (1999) Estimating long-range dependence in the presence of periodicity: An empirical study. Mathematical and Computer Modelling 29 (10–12):217–228. https://doi.org/10.1016/S0895-7177(99)00104-1
[280] Montanari A, Rosso R, Taqqu MS (2000) A seasonal fractional ARIMA Model applied to the Nile River monthly flows at Aswan. Water Resources Research 36(5):1249–1259. https://doi.org/10.1029/2000WR900012
[281] Mouelhi S, Michel C, Perrin C, Andréassian V (2006a) Linking stream flow to rainfall at the annual time step: The Manabe bucket model revisited. Journal of Hydrology 328(1–2):283–296. https://doi.org/10.1016/j.jhydrol.2005.12.022
[282] Mouelhi S, Michel C, Perrin C, Andréassian V (2006b) Stepwise development of a two-parameter monthly water balance model. Journal of Hydrology 318(1–4):200–214. https://doi.org/10.1016/j.jhydrol.2005.06.014

[283] Moustris KP, Larissi IK, Nastos PT, Paliatsos AG (2011) Precipitation forecast using artificial neural networks in specific regions of Greece. Water Resources Management 25(8):1979–1993. https://doi.org/10.1007/s11269-011-9790-5
[284] Murphy AM (1993) What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting 8:281–293. https://doi.org/10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2
[285] Murtagh F (1991) Multilayer perceptrons for classification and regression. Neurocomputing 2(5–6):183-197. https://doi.org/10.1016/0925-2312(91)90023-5
[286] Narayanan P, Basistha A, Sarkar S, Kamna S (2013) Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India. Comptes Rendus Geoscience 345(1):22-27. https://doi.org/10.1016/j.crte.2012.12.001
[287] Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. Journal of Hydrology 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6
[288] Natekin A, Knoll A (2013) Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 7:21. https://doi.org/10.3389/fnbot.2013.00021
[289] Nayak PC, Sudheer KP, Ranganc DM, Ramasastrid KS (2004) A neuro-fuzzy computing technique for modeling hydrological time series. Journal of Hydrology 291(1–2):52–66. https://doi.org/10.1016/j.jhydrol.2003.12.010
[290] Nayak PC, Venkatesh B, Krishna B, Jain SK (2013) Rainfall-runoff modeling using conceptual, data driven, and wavelet based computing approach. Journal of Hydrology 493:57–67. https://doi.org/10.1016/j.jhydrol.2013.04.016
[291] Nearing GS, Tian Y, Gupta HV, Clark MP, Harrison KW, Weijs SV (2016) A philosophical basis for hydrological uncertainty. Hydrological Sciences Journal 61(9):1666–1678. https://doi.org/10.1080/02626667.2016.1183009
[292] Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1983) Applied linear statistical models. Richard D. Irwin, Inc., Homewood, Illinois
[293] Newman AJ, Sampson K, Clark MP, Bock A, Viger RJ, Blodgett D (2014) A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. Boulder, CO: UCAR/NCAR. https://doi.org/10.5065/D6MW2F4D
[294] Newman AJ, Clark MP, Sampson K, Wood A, Hay LE, Bock A, Viger RJ, Blodgett D, Brekke L, Arnold JR, Hopson T, Duan Q (2015) Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. Hydrology and Earth System Sciences 19:209–223. https://doi.org/10.5194/hess-19-209-2015
[295] Niel H, Paturel JE, Servat E (2003) Study of parameter stability of a lumped hydrologic model in a context of climatic variability. Journal of Hydrology 278(1–4):213–230. https://doi.org/10.1016/S0022-1694(03)00158-6
[296] Nowotarski J, Liu B, Weron R, Hong T (2016) Improving short term load forecast accuracy via combining sister forecasts. Energy 98(1):40–49. https://doi.org/10.1016/j.energy.2015.12.142
[297] O'Hagan AO, Buck CE, Daneshkhah A, Eiser JR, Garthwaite PH, Jenkinson DJ, Oakley JE, Rakow T (2006) Uncertain judgements: Eliciting experts' probabilities. John Wiley & Sons
[298] Okoli K, Breinl K, Brandimarte L, Botto A, Volpi E, Di Baldassarre G (2018) Model averaging versus model selection: Estimating design floods with uncertain river flow data. Hydrological Sciences Journal 63(13–14):1913–1926. https://doi.org/10.1080/02626667.2018.1546389
[299] Oudin L, Hervieu F, Michel C, Perrin C, Andréassian V, Anctil F, Loumagne C (2005) Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. Journal of Hydrology 303(1–4):290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026
[300] Oudin L, Perrin C, Mathevet T, Andréassian V, Michel C (2006) Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. Journal of Hydrology 320(1–2):62–83. https://doi.org/10.1016/j.jhydrol.2005.07.016

[301] Oudin L, Kay A, Andréassian V, Perrin C (2010) Are seemingly physically similar catchments truly hydrologically similar?. Water Resources Research 46(11):W11558. https://doi.org/10.1029/2009WR008887
[302] Ouyang Q, Lu W (2017) Monthly rainfall forecasting using echo state networks coupled with data preprocessing methods. Water Resources Management 32(2):659–674. https://doi.org/10.1007/s11269-017-1832-1
[303] Pai PF, Hong WC (2007) A recurrent support vector regression model in rainfall forecasting. Hydrological Processes 21:819–827. https://doi.org/10.1002/hyp.6323
[304] Palma W (2007) Long-Memory Time Series. John Wiley & Sons, Hoboken, New Jersey
[305] Papacharalampous GA, Tyralis H (2018a) Evaluation of random forests and Prophet for daily streamflow forecasting. Advances in Geosciences 45:201–208. https://doi.org/10.5194/adgeo-45-201-2018
[306] Papacharalampous GA, Tyralis H (2018b) One-step ahead forecasting of geophysical processes within a purely statistical framework: Supplementary material. Figshare. https://doi.org/10.6084/m9.figshare.5357359.v1
[307] Papacharalampous GA, Tyralis H (2018c) Supplementary material for the paper "Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes". Figshare. https://doi.org/10.6084/m9.figshare.7092824
[308] Papacharalampous GA, Tyralis H (2020) Hydrological time series forecasting using simple combinations: Big data testing and investigations on one-year ahead river flow predictability. Journal of Hydrology 590:125205. https://doi.org/10.1016/j.jhydrol.2020.125205
[309] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2017a) Comparison between stochastic and machine learning methods for hydrological multi-step ahead forecasting: All forecasts are wrong!. European Geosciences Union General Assembly 2017, Vienna, Austria: EGU2017-3068-2. https://doi.org/10.13140/RG.2.2.17205.47848
[310] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2017b) Forecasting of geophysical processes using stochastic and machine learning algorithms. European Water 59:161-168
[311] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2017c) Large scale simulation experiments for the assessment of one-step ahead forecasting properties of stochastic and machine learning point estimation methods. Asia Oceania Geosciences Society 14th Annual Meeting, Singapore: HS06-A002. https://doi.org/10.13140/RG.2.2.33273.77923
[312] Papacharalampous GA, Koutsoyiannis D, Montanari A (2018a) Toy models for increasing the understanding on stochastic process-based modelling. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-1900-1. https://www.itia.ntua.gr/1813
[313] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018b) A step further from model-fitting for the assessment of the predictability of monthly temperature and precipitation. European Geosciences Union General Assembly 2018, Vienna, Austria: EGU2018-864. https://doi.org/10.6084/m9.figshare.7325783.v1
[314] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018c) Error evolution patterns in multi-step ahead streamflow forecasting. 13th International Conference on Hydroinformatics, Palermo, Italy:1598–1607. https://doi.org/10.29007/84k6
[315] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018d) One-step ahead forecasting of geophysical processes within a purely statistical framework. Geoscience Letters 5(1):12. https://doi.org/10.1186/s40562-018-0111-1
[316] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018e) Predictability of monthly temperature and precipitation using automatic time series forecasting methods. Acta Geophysica 66(4):807–831. https://doi.org/10.1007/s11600-018-0120-7
[317] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2018f) Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. Water Resources Management 32(15):5207–5239. https://doi.org/10.1007/s11269-018-2155-6

[318] Papacharalampous GA, Tyralis H, Koutsoyiannis D (2019a) Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. Stochastic Environmental Research and Risk Assessment 33(2):481–514. https://doi.org/10.1007/s00477-018-1638-6

[319] Papacharalampous GA, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019b) Large-scale comparison of machine learning regression algorithms for probabilistic hydrological modelling via post-processing of point predictions. European Geosciences Union General Assembly 2019, Vienna, Austria: EGU2019-3576. https://doi.org/10.6084/m9.figshare.8018342.v1

[320] Papacharalampous GA, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019c) Probabilistic hydrological post-processing at scale: Why and how to apply machine learning quantile regression algorithms. Water 11(10):2126. https://doi.org/10.3390/w11102126

[321] Papacharalampous GA, Tyralis H, Koutsoyiannis D, Montanari A (2019d) Supplementary material for the paper "Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale". Figshare. https://doi.org/10.6084/m9.figshare.7959473.v2

[322] Papacharalampous G, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D (2019e) Supplementary material for the paper "Probabilistic hydrological post-processing at scale: Why and how to apply machine learning quantile regression algorithms". Figshare. https://doi.org/10.6084/m9.figshare.9496262.v1

[323] Papacharalampous GA, Koutsoyiannis D, Montanari A (2020a) Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: Methodology development and investigation using toy models. Advances in Water Resources 136:103471. https://doi.org/10.1016/j.advwatres.2019.103471

[324] Papacharalampous GA, Tyralis H, Koutsoyiannis D, Montanari A (2020b) Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. Advances in Water Resources 136:103470. https://doi.org/10.1016/j.advwatres.2019.103470

[325] Papalexiou SM, Koutsoyiannis D (2013) Battle of extreme value distributions: A global survey on extreme daily rainfall. Water Resources Research 49(1):187–201. https://doi.org/10.1029/2012WR012557

[326] Papalexiou SM, Koutsoyiannis D (2016) A global survey on the seasonal variation of the marginal distribution of daily precipitation. Advances in Water Resources 94:131–145. https://doi.org/10.1016/j.advwatres.2016.05.005

[327] Pappenberger F, Beven KJ (2006) Ignorance is bliss: Or seven reasons not to use uncertainty analysis. Water Resources Research 42(5):W05302. https://doi.org/10.1029/2005WR004820

[328] Pappenberger F, Ramos MH, Cloke HL, Wetterhall F, Alfieri L, Bogner K, Mueller A, Salamon P (2015) How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. Journal of Hydrology 522:697–713. https://doi.org/10.1016/j.jhydrol.2015.01.024

[329] Patel SS, Ramachandran P (2015) A comparison of machine learning techniques for modeling river flow time series: the case of upper Cauvery river basin. Water Resources Management 29(2):589–602. https://doi.org/10.1007/s11269-014-0705-0

[330] Paturel JE, Servat E, Vassiliadis A (1995) Sensitivity of conceptual rainfall-runoff algorithms to errors in input data—Case of the GR2M model. Journal of Hydrology 168(1–4):111–125. https://doi.org/10.1016/0022-1694(94)02654-T

[331] Pechlivanidis IG, Jackson BM, McIntyre NR, Wheater HS (2011) Catchment scale hydrological modelling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Global NEST Journal 13(3):193–214

[332] Pemberton J (1987) Exact least squares multi-step prediction from nonlinear autoregressive models. Journal of Time Series Analysis 8(4):443-448. https://doi.org/10.1111/j.1467-9892.1987.tb00007.x
[333] Perrin C, Michel C, Andréassian V (2001) Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. Journal of Hydrology 242(3–4):275–301. https://doi.org/10.1016/S0022-1694(00)00393-0
[334] Perrin C, Michel C, Andréassian V (2003) Improvement of a parsimonious model for streamflow simulation. Journal of Hydrology 279(1–4):275–289. https://doi.org/10.1016/S0022-1694(03)00225-7
[335] Peterson RA (2017) Estimating normalization transformations with `bestNormalize`. https://github.com/petersonR/bestNormalize
[336] Peterson RA (2019) `bestNormalize`: Normalizing Transformation Functions. `R` package version 1.4.0. https://CRAN.R-project.org/package=bestNormalize
[337] Peterson TC, Vose RS (1997) An Overview of the Global Historical Climatology Network Temperature Database. Bulletin of the American Meteorological Society 78:2837–2849. https://doi.org/10.1175/1520-0477(1997)078<2837:AOOTGH>2.0.CO;2
[338] Plummer M, Best N, Cowles K, Vines K (2006) `CODA`: convergence diagnosis and output analysis for MCMC. `R` news 6(1):7–11
[339] Plummer M, Best N, Cowles K, Vines K, Sarkar D, Bates D, Almond R, Magnusson A (2019) `coda`: Output Analysis and Diagnostics for MCMC. `R` package version 0.19-3. https://CRAN.R-project.org/package=coda
[340] Quilty J, Adamowski J, Boucher MA (2019) A stochastic data-driven ensemble forecasting framework for water resources: A case study using ensemble members derived from a database of deterministic wavelet-based models. Water Resources Research 55(1):175–202. https://doi.org/10.1029/2018WR023205
[341] `R` Core Team (`2019`) `R`: A language and environment for statistical computing. `R` Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org
[342] Raghavendra NS, Deka PC (2014) Support vector machine applications in the field of hydrology: A review. Applied Soft Computing 19:372–386. https://doi.org/10.1016/j.asoc.2014.02.002
[343] Ramos MH, Mathevet T, Thielen J, Pappenberger F (2010) Communicating uncertainty in hydro-meteorological forecasts: Mission impossible?. Meteorological Applications 17(2):223–235. https://doi.org/10.1002/met.202
[344] Ramos MH, Van Andel SJ, Pappenberger F (2013) Do probabilistic forecasts lead to better decisions?. Hydrology and Earth System Sciences 17:2219–2232. https://doi.org/10.5194/hess-17-2219-2013
[345] Remesan R, Mathew J (2015) Hydrological Data Driven Modelling. Springer International Publishing. https://doi.org/10.1007/978-3-319-09235-5
[346] Ren H, Hou Z, Huang M, Bao J, Sun Y, Tesfa T, Leung LR (2016) Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins. Journal of Hydrology 536:92–108. https://doi.org/10.1016/j.jhydrol.2016.02.042
[347] Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW (2010) Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. Water Resources Research 46(5):W05521. https://doi.org/10.1029/2009WR008328
[348] Renard B, Kavetski D, Leblois E, Thyer M, Kuczera G, Franks SW (2011) Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. Water Resources Research 47(11):W11516. https://doi.org/10.1029/2011WR010643
[349] Reutlinger A, Hangleiter D, Hartmann S (2017) Understanding (with) toy models. The British Journal for the Philosophy of Science 69(4):1069–1099. https://doi.org/10.1093/bjps/axx005

[350] Rigby RA, Stasinopoulos DM (2005) Generalized additive models for location, scale and shape. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54(3):507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x
[351] Ripley B (2016) `nnet`: Feed-Forward Neural Networks and Multinomial Log-Linear Models. `R` package version 7.3-12. https://CRAN.R-project.org/package=nnet
[352] Ripley B (2019) `MASS`: Support Functions and Datasets for Venables and Ripley's MASS. `R` package version 7.3-51.4. https://CRAN.R-project.org/package=MASS
[353] Romero-Cuellar J, Abbruzzo A, Adelfio G, Francés F (2019) Hydrological post-processing based on approximate Bayesian computation (ABC). Stochastic Environmental Research and Risk Assessment 33(7):1361–1373. https://doi.org/10.1007/s00477-019-01694-y
[354] Sadegh M, Vrugt JA (2013) Bridging the gap between GLUE and formal statistical approaches: Approximate Bayesian computation. Hydrology and Earth System Sciences 17:4831–4850. https://doi.org/10.5194/hess-17-4831-2013
[355] Sadegh M, Vrugt JA (2014) Approximate Bayesian Computation using Markov Chain Monte Carlo simulation: DREAM$_{(ABC)}$. Water Resources Research 50(8):6767–6787. https://doi.org/10.1002/2014WR015386
[356] Sadegh M, Vrugt JA, Xu C, Volpi E (2015) The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM$_{(ABC)}$. Water Resources Research 51(11):9207–9231. https://doi.org/10.1002/2014WR016805
[357] Sagi O, Rokach L (2018) Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1249. https://doi.org/10.1002/widm.1249
[358] Sapankevych NI, Sankar R (2009) Time series prediction using support vector machines: a survey. IEEE Computational Intelligence Magazine 4(2):24–38. https://doi.org/10.1109/MCI.2009.932254
[359] Savel'ev E, Miroshnikov A, Conlon E (2015) BayesSummaryStatLM: MCMC Sampling of Bayesian Linear Models via Summary Statistics. R package version 1.0-1. https://CRAN.R-project.org/package=BayesSummaryStatLM
[360] Sawicz K, Wagener T, Sivapalan M, Troch PA, Carrillo G (2011) Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA. Hydrology and Earth System Sciences 15:2895–2911. https://doi.org/10.5194/hess-15-2895-2011
[361] Schaake J, Cong S, Duan Q (2006) US MOPEX data set. IAHS Publication 307:9–28
[362] Schaake JC, Duan Q, Smith M, Koren V (2000) Criteria to select basins for hydrologic model development and testing. Preprints in: 15th Conference on Hydrology (Long Beach, California, USA, Am. Met. Soc., 10–14 January 2000), Paper P1.8
[363] Schaefli B, Gupta HV (2007) Do Nash values have value?. Hydrological Processes 21(15):2075–2080. https://doi.org/10.1002/hyp.6825
[364] Schoups G, Vrugt JA (2010) A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resources Research 46(10):W10531. https://doi.org/10.1029/2009WR008933
[365] Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464. https://doi.org/10.1214/15-AOS1321
[366] Scornet E, Biau G, Vert JP (2015) Consistency of random forests. The Annals of Statistics 43(4):1716–1741
[367] Shabri A, Suhartono (2012) Streamflow forecasting using least-squares support vector machines. Hydrological Sciences Journal 57(7):1275–1293. https://doi.org/10.1080/02626667.2012.714468
[368] Shi Z, Han M (2007) Support vector echo-state machine for chaotic time-series prediction. IEEE Transactions on Neural Networks 18(2):359–372. https://doi.org/10.1109/TNN.2006.885113
[369] Shmueli G (2010) To explain or to predict?. Statistical Science 25(3):289–310. https://doi.org/10.1214/10-STS330

[370] Sikorska AE, Montanari A, D Koutsoyiannis (2015) Estimating the uncertainty of hydrological predictions through data-driven resampling techniques. Journal of Hydrologic Engineering 20(1):A4014009. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000926
[371] Silver D, Huang A, Maddison C, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, et al. (2016) Mastering the game of Go with deep neural networks and tree search. Nature 529:484–489. https://doi.org/10.1038/nature16961
[372] Singh M, Singh R, Shinde V (2011) Application of software packages for monthly stream flow forecasting of Kangsabati River in India. International Journal of Computer Applications 20(3):7–14. https://doi.org/10.5120/2416-3231
[373] Sivakumar B (2004) Chaos theory in geophysics: past, present and future. Chaos, Solitons and Fractals 19(2):441–462. https://doi.org/10.1016/S0960-0779(03)00055-9
[374] Sivakumar B (2005) Hydrologic modeling and forecasting: Role of thresholds. Environmental Modelling and Software 20(5):515–519. https://doi.org/10.1016/j.envsoft.2004.08.006
[375] Sivakumar B (2008a) The more things change, the more they stay the same: The state of hydrologic modelling. Hydrological Processes 22(21):4333–4337. https://doi.org/10.1002/hyp.7140
[376] Sivakumar B (2008b) Undermining the science or undermining Nature?. Hydrological Processes 22(6):893–897. https://doi.org/10.1002/hyp.7004
[377] Sivakumar B (2017) Chaos in hydrology: Bridging determinism and stochasticity. Springer Netherlands. https://doi.org/10.1007/978-90-481-2552-4
[378] Sivakumar B, Berndtsson R (2010) Advances in data-based approaches for hydrologic modeling and forecasting. World Scientific Publishing Company, Singapore. https://doi.org/10.1142/7783
[379] Sivakumar B, Jayawardena AW, Fernando TMKG (2002) River flow forecasting: Use of phase-space reconstruction and artificial neural networks approaches. Journal of Hydrology 265(1–4):225–245. https://doi.org/10.1016/S0022-1694(02)00112-9
[380] Sivakumar B, Woldemeskel FM, Vignesh R, Jothiprakash V (2019) A correlation–scale–threshold method for spatial variability of rainfall. Hydrology 6(1):11. https://doi.org/10.3390/hydrology6010011
[381] Sivapragasam C, Liong SY, Pasha MFK (2001) Rainfall and runoff forecasting with SSA-SVM approach. Journal of Hydroinformatics 3(3):141–152. https://doi.org/10.2166/hydro.2001.0014
[382] Smola AJ, Schölkopf B (2004) A tutorial on support vector regression. Statistics and Computing 14(3):199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88
[383] Soetaert K, Petzoldt T (2010) Inverse modelling, sensitivity and Monte Carlo analysis in R using package FME. Journal of Statistical Software 33(3):1–28. https://doi.org/10.18637/jss.v033.i03
[384] Soetaert K, Petzoldt T (2016) FME: A Flexible Modelling Environment for Inverse Modelling, Sensitivity, Identifiability and Monte Carlo Analysis. R package version 1.3.5. https://CRAN.R-project.org/package=FME
[385] Solomatine DP, Dulal KN (2003) Model trees as an alternative to neural networks in rainfall—runoff modelling. Hydrological Sciences Journal 48(3):399–411. https://doi.org/10.1623/hysj.48.3.399.45291
[386] Solomatine DP, Ostfeld A (2008) Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics 10(1):3–22. https://doi.org/10.2166/hydro.2008.015
[387] Solomatine DP, Shrestha DL (2009) A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research 45(12):W00B11. https://doi.org/10.1029/2008WR006839
[388] Solomatine DP, Wagener T (2011) Hydrological modeling. In: Wilderer PA (eds) Treatise on Water Science 2. Elsevier, pp 435–458

[389] Stedinger JR, Vogel RM, Lee SU, Batchelder R (2008) Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resources Research, 44(12). https://doi.org/10.1029/2008WR006822
[390] Stoica P, Nehorai A (1989) On multistep prediction error methods for time series models. Journal of Forecasting 8(4):357–368. https://doi.org/10.1002/for.3980080402
[391] Sutton CD (2005) Classification and regression trees, bagging, and boosting. Handbook of Statistics 24:303–329. https://doi.org/10.1016/S0169-7161(04)24011-1
[392] Széles B, Broer M, Parajka J, Hogan P, Eder A, Strauss P, Blöschl G (2018) Separation of scales in transpiration effects on low flows: A spatial analysis in the Hydrological Open Air Laboratory. Water Resources Research 54(9):6168–6188. https://doi.org/10.1029/2017WR022037
[393] Taieb SB, Atiya AF (2016) A Bias and Variance Analysis for Multistep-Ahead Time Series Forecasting. IEEE Transactions on Neural Networks and Learning Systems 27(1):62–76. https://doi.org/10.1109/TNNLS.2015.2411629
[394] Taieb SB, Bontempi G, Atiya AF, Sorjamaa A (2012) A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. Expert Systems with Applications 39(8):7067–7083. https://doi.org/10.1016/j.eswa.2012.01.039
[395] Taillardat M, Mestre O, Zamo M, Naveau P (2016) Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. Monthly Weather Review 144:2375–2393. https://doi.org/10.1175/MWR-D-15-0260.1
[396] Taormina R, Chau KW (2015) Data-driven input variable selection for rainfall–runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines. Journal of Hydrology 529(Part 3):1617–1632. https://doi.org/10.1016/j.jhydrol.2015.08.022
[397] Taylor JW (2000) A quantile regression neural network approach to estimating the conditional density of multiperiod returns. Journal of Forecasting 19(4):299–311. https://doi.org/10.1002/1099-131X(200007)19:4<299::AID-FOR775>3.0.CO;2-V
[398] Taylor SJ, Letham B (2018) Forecasting at scale. The American Statistician 72(1):37–45. https://doi.org/10.1080/00031305.2017.1380080
[399] Taylor SJ, Letham B (2017) `prophet`: Automatic Forecasting Procedure. `R` package version 0.2. https://CRAN.R-project.org/package=prophet
[400] Thornton PE, Thornton MM, Mayer BW, Wilhelmi N, Wei Y, Devarakonda R, Cook RB (2014) Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2. ORNL DAAC, Oak Ridge, Tennessee, USA. Date accessed: 2016/01/20. https://doi.org/10.3334/ORNLDAAC/1219
[401] Thissen U, Van Brakel R, De Weijer AP, Melssena WJ, Buydens LMC (2003) Using support vector machines for time series prediction. Chemometrics and Intelligent Laboratory Systems 69(1–2):35–49. https://doi.org/10.1016/S0169-7439(03)00111-4
[402] Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S (2009) Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research 45(12):W00B14. https://doi.org/10.1029/2008WR006825
[403] Tian Y, Xu YP, Zhang XJ (2013) Assessment of climate change impacts on river high flows through comparative use of GR4J, HBV and Xinanjiang models. Water Resources Management 27(8):2871–2888. https://doi.org/10.1007/s11269-013-0321-4
[404] Tibshirani J, Athey S (2019) `grf`: Generalized Random Forests (Beta). `R` package version 0.10.3. https://CRAN.R-project.org/package=grf
[405] Todini E (1996) The ARNO rainfall—runoff model. Journal of Hydrology 175(1–4), 339–382. https://doi.org/10.1016/S0022-1694(96)80016-3
[406] Todini E (2004) Role and treatment of uncertainty in real-time flood forecasting. Hydrological Processes 18:2743–2746. https://doi.org/10.1002/hyp.5687
[407] Todini E (2007) Hydrological catchment modelling: past, present and future. Hydrology and Earth System Sciences 11:468–482. https://doi.org/10.5194/hess-11-468-2007

[408] Todini E (2008) A model conditional processor to assess predictive uncertainty in flood forecasting. International Journal of River Basin Management 6(2):123–137. https://doi.org/10.1080/15715124.2008.9635342
[409] Tomkins KM (2014) Uncertainty in streamflow rating curves: methods, controls and consequences. Hydrological Processes 28(3):464–481. https://doi.org/10.1002/hyp.9567
[410] Tongal H, Berndtsson R (2017) Impact of complexity on daily and multi-step forecasting of streamflow with chaotic, stochastic, and black-box models. Stochastic Environmental Research and Risk Assessment 31(3):661–682. https://doi.org/10.1007/s00477-016-1236-4
[411] Toth E, Brath A (2002) Flood forecasting using artificial neural networks in black-box and conceptual rainfall-runoff modelling. International Congress on Environmental Modelling and Software:166–171
[412] Toth E, Brath A (2007) Multistep ahead streamflow forecasting: Role of calibration data in conceptual and neural network modeling. Water Resources Research 43(11):W11405. https://doi.org/10.1029/2006WR005383
[413] Toth E, Montanari A, Brath A (1999) Real-time flood forecasting via combined use of conceptual and stochastic models. Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere 24(7):793–798. https://doi.org/10.1016/S1464-1909(99)00082-9
[414] Tyralis H (2016) `HKprocess`: Hurst-Kolmogorov Process. `R` package version 0.0-2. https://CRAN.R-project.org/package=HKprocess
[415] Tyralis H, Koutsoyiannis D (2011) Simultaneous estimation of the parameters of the Hurst–Kolmogorov stochastic process. Stochastic Environmental Research and Risk Assessment 25(1):21–33. https://doi.org/10.1007/s00477-010-0408-x
[416] Tyralis H, Koutsoyiannis D (2014) A Bayesian statistical model for deriving the predictive distribution of hydroclimatic variables. Climate Dynamics 42(11–12):2867–2883. https://doi.org/10.1007/s00382-013-1804-y
[417] Tyralis H, Koutsoyiannis D (2017) On the prediction of persistent processes using the output of deterministic models. Hydrological Sciences Journal 62(13):2083–210 2. https://doi.org/10.1080/02626667.2017.1361535
[418] Tyralis H, Papacharalampous GA (2017) Variable selection in time series forecasting using random forests. Algorithms 10(4):114. https://doi.org/10.3390/a10040114
[419] Tyralis H, Papacharalampous GA (2018) Large-scale assessment of Prophet for multi-step ahead forecasting of monthly streamflow. Advances in Geosciences 45:147–153. https://doi.org/10.5194/adgeo-45-147-2018
[420] Tyralis H, Papacharalampous GA (2020) Boosting algorithms in energy research: A systematic review. https://arxiv.org/abs/2004.07049
[421] Tyralis H, Dimitriadis P, Koutsoyiannis D, O'Connell PE, Tzouka K, Iliopoulou T (2018) On the long-range dependence properties of annual precipitation using a global network of instrumental measurements. Advances in Water Resources 111:301–318. https://doi.org/10.1016/j.advwatres.2017.11.010
[422] Tyralis H, Koutsoyiannis D, Kozanis S (2013) An algorithm to construct Monte Carlo confidence intervals for an arbitrary function of probability distribution parameters. Computational Statistics 28(4):1501–1527. https://doi.org/10.1007/s00180-012-0364-7
[423] Tyralis H, Papacharalampous GA, Burnetas A, Langousis A (2019a) Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS. Journal of Hydrology 577:123957. https://doi.org/10.1016/j.jhydrol.2019.123957
[424] Tyralis H, Papacharalampous GA, Langousis A (2019b) A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11(5):910. https://doi.org/10.3390/w11050910

[425] Tyralis H, Papacharalampous GA, Tantanee S (2019c) How to explain and predict the shape parameter of the generalized extreme value distribution of streamflow extremes using a big dataset. Journal of Hydrology 574:628–645. https://doi.org/10.1016/j.jhydrol.2019.04.070
[426] Tyralis H, Papacharalampous GA, Langousis A (2020a) Streamflow forecasting at large time scales using statistical models. In: Sharma P, Machiwal D (eds) Advances in Streamflow Forecasting, Elsevier. In press
[427] Tyralis H, Papacharalampous GA, Langousis A (2020b) Super ensemble learning for daily streamflow forecasting: Large-scale demonstration and comparison with multiple machine learning algorithms. Neural Computing and Applications. https://doi.org/10.1007/s00521-020-05172-3
[428] Valipour M, Banihabib ME, Behbahani SMR (2013) Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. Journal of Hydrology 476(7):433–441. https://doi.org/10.1016/j.jhydrol.2012.11.017
[429] Vapnik VN (1995) The nature of statistical learning theory, first edition. Springer-Verlag New York. https://doi.org/10.1007/978-1-4757-3264-1
[430] Vapnik VN (1999) An overview of statistical learning theory. IEEE Transactions on Neural Networks 10(5):988–999. https://doi.org/10.1109/72.788640
[431] Venables WN, Ripley BD (2002) Modern Applied Statistics with S, fourth edition. Springer-Verlag New York. https://doi.org/10.1007/978-0-387-21706-2
[432] Vitolo C (2017) `hddtools`: hydrological data discovery tools. The Journal of Open Source Software 2(9). https://doi.org/10.21105/joss.00056
[433] Vitolo C (2018) `hddtools`: Hydrological Data Discovery Tools. `R` package version 0.8.2. https://CRAN.R-project.org/package=hddtools
[434] Vogel RM (1999) Stochastic and deterministic world views. Journal of Water Resources Planning and Management 125(6):311–313
[435] Volpi E, Di Lazzaro M, Fiori A (2012) A simplified framework for assessing the impact of rainfall spatial variability on the hydrologic response. Advances in Water Resources 46:1–10. https://doi.org/10.1016/j.advwatres.2012.04.011
[436] Volpi E, Schoups G, Firmani G, Vrugt JA (2017) Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling. Water Resources Research 53(7):6133–6158. https://doi.org/10.1002/2016WR020167
[437] Vrugt JA (2016) Markov chain Monte Carlo simulation using the `DREAM` software package: Theory, concepts, and `MATLAB` implementation. Environmental Modelling and Software 75:273–316. https://doi.org/10.1016/j.envsoft.2015.08.013
[438] Vrugt JA (2018) `MODELAVG`: A `MATLAB` Toolbox for postprocessing of model ensembles [preprint made available by the author]
[439] Vrugt JA (2019) Merging models with data. Topic 6: Model averaging [presentation made available by the author]
[440] Vrugt JA, Robinson BA (2007) Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. Water Resources Research 43(1):W01411. https://doi.org/10.1029/2005WR004838
[441] Vrugt JA, Gupta HV, Bouten W, Sorooshian S (2003) A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resources Research 39(8):1201. https://doi.org/10.1029/2002WR001642
[442] Vrugt JA, Diks CGH, Gupta HV, Bouten W, Verstraten JM (2005) Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. Water Resources Research 41(1):W01017. https://doi.org/10.1029/2004WR003059

[443] Vrugt JA, Ter Braak CJF, Clark MP, Hyman JM, Robinson BA (2008) Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resources Research 44(12):W00B09. https://doi.org/10.1029/2007WR006720
[444] Vrugt JA, Ter Braak CJF, Diks CGH, Robinson BA, Hyman JM, Higdon D (2009a) Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. International Journal of Nonlinear Sciences and Numerical Simulation 10(3). https://doi.org/10.1515/IJNSNS.2009.10.3.273
[445] Vrugt JA, Ter Braak CJF, Gupta HV, Robinson BA (2009b) Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?. Stochastic Environmental Research and Risk Assessment 23(7):1011–1026. https://doi.org/10.1007/s00477-008-0274-y
[446] Vrugt JA, Ter Braak CJF, Diks CGH, Schoups G (2013) Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. Advances in Water Resources 51:457–478. https://doi.org/10.1016/j.advwatres.2012.04.002
[447] Wagener T, Hogue T, Schaake J, Duan Q, Gupta H, Andreassian V, Hall A, Leavesley G (2006) The Model Parameter Estimation Experiment (MOPEX): Its structure, connection to other international initiatives and future directions. IAHS Publication Series 307:339–346. https://www.osti.gov/servlets/purl/898007
[448] Waldmann E (2018) Quantile regression: A short story on how and why. Statistical Modelling 18(3–4):203–218. https://doi.org/10.1177/1471082X18759142
[449] Wallis KF (2011) Combining forecasts–Forty years later. Applied Financial Economics 21(1–2):33–41. https://doi.org/10.1080/09603107.2011.523179
[450] Wang W, Van Gelder PH, Vrijling JK, Ma J (2006) Forecasting daily streamflow using hybrid ANN models. Journal of Hydrology 324(1–4):383–399. https://doi.org/10.1016/j.jhydrol.2005.09.032
[451] Wang WC, Chau KW, Cheng CT, Qiu L (2009) A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. Journal of Hydrology 374(3–4):294–306. https://doi.org/10.1016/j.jhydrol.2009.06.019
[452] Wang QJ, Shrestha DL, Robertson DE, Pokhrel P (2012) A log-sinh transformation for data normalization and variance stabilization. Water Resources Research 48(5):W05514. https://doi.org/10.1029/2011WR010973
[453] Wang S, Feng J, Liu G (2013) Application of seasonal time series model in the precipitation forecast. Mathematical and Computer Modelling 58(3-4):677-683. https://doi.org/10.1016/j.mcm.2011.10.034
[454] Wang W, Chau K, Xu D, Chen XY (2015) Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. Water Resources Management 29(8):2655-2675. https://doi.org/10.1007/s11269-015-0962-6
[455] Wang P, Liu B, Hong T (2016) Electric load forecasting with recency effect: A big data approach. International Journal of Forecasting 32(3):585–597. https://doi.org/10.1016/j.ijforecast.2015.09.006
[456] Wang Y, Zhang N, Tan Y, Hong T, Kirschen DS, Kang C (2019) Combining Probabilistic Load Forecasts. IEEE Transactions on Smart Grid 10(4):3664–3674. https://doi.org/10.1109/TSG.2018.2833869
[457] Wani O, Beckers JVL, Weerts AH, Solomatine DP (2017) Residual uncertainty estimation using instance-based learning with applications to hydrologic forecasting. Hydrology and Earth System Sciences 21:4021–4036. https://doi.org/10.5194/hess-21-4021-2017
[458] Warnes GR, Bolker B, Gorjanc G, Grothendieck G, Korosec A, Lumley T, MacQueen D, Magnusson A, Rogers J, et al. (2017) `gdata`: Various `R` Programming Tools for Data Manipulation. `R` package version 2.18.0. https://CRAN.R-project.org/package=gdata

[459] Weerts AH, Winsemius HC, Verkade JS (2011) Estimation of predictive hydrological uncertainty using quantile regression: Examples from the National Flood Forecasting System (England and Wales). Hydrology and Earth System Sciences 15:255–265. https://doi.org/10.5194/hess-15-255-2011
[460] Wei WWS (2006) Time Series Analysis, Univariate and Multivariate Methods, second edition. Pearson Addison Wesley, Boston
[461] Weijs SV, Schoups G, Van de Giesen N (2010) Why hydrological predictions should be evaluated using information theory. Hydrology and Earth System Sciences 14:2545–2558. https://doi.org/10.5194/hess-14-2545-2010
[462] Weijs SV, van de Giesen N, Parlange MB (2013) HydroZIP: How hydrological knowledge can be used to improve compression of hydrological data. Entropy 15(4):1289–1310; https://doi.org/10.3390/e15041289
[463] Wickham H (2007) Reshaping data with the `reshape` package. Journal of Statistical Software 21(12):1–20. https://doi.org/10.18637/jss.v021.i12
[464] Wickham H (2011) The split-apply-combine strategy for data analysis. Journal of Statistical Software 40(1):1–29
[465] Wickham H (2016a) `ggplot2`: Elegant Graphics for Data Analysis, 2nd ed. Springer International Publishing, Switzerland. https://doi.org/10.1007/978-3-319-24277-4
[466] Wickham H (2016b) `plyr`: Tools for Splitting, Applying and Combining Data. `R` package version 1.8.4. https://cran.r-project.org/web/packages/plyr
[467] Wickham H (2017) `reshape2`: Flexibly Reshape Data: A Reboot of the `reshape` Package. `R` package version 1.4.3. https://CRAN.R-project.org/package=reshape2
[468] Wickham H (2018) `reshape`: Flexibly Reshape Data. `R` package version 0.8.8. https://CRAN.R-project.org/package=reshape
[469] Wickham H (2019) `stringr`: Simple, Consistent Wrappers for Common String Operations. `R` package version 1.4.0. https://CRAN.R-project.org/package=stringr
[470] Wickham H, Henry L (2019) `tidyr`: Easily Tidy Data with `'spread()'` and `'gather()'` Functions. `R` package version 0.8.3. https://CRAN.R-project.org/package=tidyr
[471] Wickham H, Hester J, Francois R (2018) `readr`: Read Rectangular Text Data. `R` package version 1.3.1. https://CRAN.R-project.org/package=readr
[472] Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, Woo K, Yutani H (2019a) `ggplot2`: Create Elegant Data Visualisations Using the Grammar of Graphics. `R` package version 3.2.1. https://CRAN.R-project.org/package=ggplot2
[473] Wickham H, François R, Henry L, Müller K (2019b) `dplyr`: A Grammar of Data Manipulation. `R` package version 0.8.3. https://CRAN.R-project.org/package=dplyr
[474] Wickham H, Hester J, Chang W (2019c) `devtools`: Tools to Make Developing `R` Packages Easier. `R` package version 2.1.0. https://CRAN.R-project.org/package=devtools
[475] Wilke CO (2018) `ggridges`: Ridgeline Plots in `'ggplot2'`. `R` package version 0.5.1. https://CRAN.R-project.org/package=ggridges
[476] Winkler RL (1972) A decision-theoretic approach to interval estimation. Journal of the American Statistical Association 67(337):187–191. https://doi.org/10.1080/01621459.1972.10481224
[477] Winkler RL (2015) Equal versus differential weighting in combining forecasts. Risk Analysis 35(1):16–18. https://doi.org/10.1111/risa.12302
[478] Witten IH, Frank E, Hall MA, Pal CJ (2017) Data Mining: Practical machine learning tools and techniques, fourth edition. Elsevier Inc. https://doi.org/10.1016/C2015-0-02071-8
[479] Witthoft C (2015) `cgwtools`: Miscellaneous Tools. `R` package version 3.0. https://cran.r-project.org/src/contrib/Archive/cgwtools
[480] Wolpert DH (1996) The lack of a priori distinctions between learning algorithms. Neural Computation 8(7):1341–1390. https://doi.org/10.1162/neco.1996.8.7.1341

[481] Wu CL, Chau KW, Fan C (2010) Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques. Journal of Hydrology 389(1-2):146-167. https://doi.org/10.1016/j.jhydrol.2010.05.040
[482] Xie Y (2014) `knitr`: A comprehensive tool for reproducible research in `R`. In: Stodden V, Leisch F, Peng RD (eds) Implementing Reproducible Computational Research. Chapman and Hall/CRC
[483] Xie Y (2015) Dynamic Documents with R and knitr, second edition. Chapman and Hall/CRC, Boca Raton, Florida
[484] Xie Y (2019) `knitr`: A General-Purpose Package for Dynamic Report Generation in `R`. `R` package version 1.24. https://CRAN.R-project.org/package=knitr
[485] Xu C (2001) Statistical analysis of parameters and residuals of a conceptual water balance model–methodology and case study. Water Resources Management 15(2):75–92. https://doi.org/10.1023/A:1012559608269
[486] Xu L, Chen N, Zhang X, Chen Z (2018) An evaluation of statistical, NMME and hybrid models for drought prediction in China. Journal of Hydrology 566:235–249. https://doi.org/10.1016/j.jhydrol.2018.09.020
[487] Xu L, Chen N, Zhang X, Chen Z, Hu C, Wang C (2019) Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. Climate Dynamics 53(1–2):601–615. https://doi.org/10.1007/s00382-018-04605-z
[488] Yan J, Liao GY, Gebremichael M, Shedd R, Vallee DR (2014) Characterizing the uncertainty in river stage forecasts conditional on point forecast values. Water Resources Research 48(12):W12509. https://doi.org/10.1029/2012WR011818
[489] Yaseen ZM, Allawi MF, Yousif AA, Jaafar O, Hamzah FM, El-Shafie A (2016) Non-tuned machine learning approach for hydrological time series forecasting. Neural Computing and Applications 30(5):1479–1491. https://doi.org/10.1007/s00521-016-2763-0
[490] Yapo PO, Gupta HV, Sorooshian S (1996) Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. Journal of Hydrology 181(1–4):23–48. https://doi.org/10.1016/0022-1694(95)02918-4
[491] Ye M, Meyer PD, Neuman SP (2008) On model selection criteria in multimodel analysis. Water Resources Research 44(3):W03428. https://doi.org/10.1029/2008WR006803
[492] Ye M, Neuman SP, Meyer PD (2004) Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resources Research 40(5):W05113. https://doi.org/10.1029/2003WR002557
[493] Ye A, Duan Q, Yuan X, Wood EF, Schaake J (2014) Hydrologic post-processing of MOPEX streamflow simulations. Journal of Hydrology 508:147–156. https://doi.org/10.1016/j.jhydrol.2013.10.055
[494] Yevjevich VM (1987) Stochastic models in hydrology. Stochastic Hydrology and Hydraulics 1(1):17–36. https://doi.org/10.1007/BF01543907
[495] Yin RK (2003) Case study research: Design and methods, third edition. Sage Publications, Inc
[496] Yu X, Liong SY (2007) Forecasting of hydrologic time series with ridge regression in feature space. Journal of Hydrology 332(3-4):290-302. https://doi.org/10.1016/j.jhydrol.2006.07.003
[497] Yu X, Liong SY, Babovic V (2004) EC-SVM approach for real-time hydrologic forecasting. Journal of Hydroinformatics 6(3):209–223. https://doi.org/10.2166/hydro.2004.0016
[498] Zambrano-Bigiarini M (2017a) `hydroGOF`: Goodness-of-Fit Functions for Comparison of Simulated and Observed Hydrological Time Series. `R` package version 0.3-10. https://CRAN.R-project.org/package=hydroGOF
[499] Zambrano-Bigiarini M (2017b) `hydroTSM`: Time Series Management, Analysis and Interpolation for Hydrological Modelling. `R` package version 0.5-1. https://github.com/hzambran/hydroTSM

[500] Zeileis A, Grothendieck G (2005) zoo: S3 infrastructure for regular and irregular time series. Journal of Statistical Software 14(6):1–27. https://doi.org/10.18637/jss.v014.i06
[501] Zeileis A, Grothendieck G, Ryan JA (2019) zoo: S3 Infrastructure for Regular and Irregular Time Series (Z's Ordered Observations). R package version 1.8-6. https://CRAN.R-project.org/package=zoo
[502] Zhang GP (2001) An investigation of neural networks for linear time-series forecasting. Computers and Operations Research 28(12):1183–1202. https://doi.org/10.1016/S0305-0548(00)00033-2
[503] Zhang GP, Patuwo BE, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting 14(1):35–62. https://doi.org/10.1016/S0169-2070(97)00044-7