National Technical University of Athens

INTERDISCIPLINARY - INTERDEPARTMENTAL
POSTGRADUATE PROGRAM
"SCIENCE & TECHNOLOGY OF WATER RESOURCES"

Large-scale evaluation of machine learning & conceptual approaches in Rainfall-Runoff modeling

George Koutsos
Supervisor: Professor Ch. Makropoulos

Athens, June 2024

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

INTERDISCIPLINARY – INTERDEPARTMENTAL
POST GRADUATE PROGRAM
"SCIENCE & TECHNOLOGY OF WATER RESOURCES

MSc Thesis

Large-scale evaluation of machine learning & conceptual approaches in Rainfall-Runoff modeling

GEORGE KOUTSOS

Supervisor: Professor Ch. Makropoulos

The content of this thesis is the result of my own intellectual effort. The incorporation of third-party material, whether published or unpublished, is done with proper citation of the sources, ensuring there are no ambiguities or misinterpretations

Athens, June 2024

to my nieces and nephews Ioanna, Alexandros, Konstantinos and Nefeli.

ACKNOWLEDGEMENTS

The current Master Thesis was conducted under the interdisciplinary-interdepartmental program of postgraduate studies in Science and Technology of Water Resources by the National Technical University of Greece.

First and foremost, I would like to express my very great appreciation to my thesis supervisor Professor Ch. Makropoulos, not only for trusting me to complete this thesis but also for enabling my academic journey to continue.

I would also like to extend my gratitude to P. Kossieris for his support throughout this journey. His guidance and suggestions have significantly impacted me, not only in relation to this thesis but also in a broader context. I also feel the need to mention that the trust he has shown in me means a great deal to me.

I would also like to thank G. Tsoukalas for his supportive ideas about the topic of this thesis. His productive and sober-minded thinking made our conversation highly engaging and insightful.

I am also grateful for all the professors and academic staff to the program for their continuous effort in shearing their knowledge and expertise. Looking back over the past two years, many memories stand out reminding me that this educational journey has been both enjoyable and fruitful with their guidance.

I could not forget to thank all my friends inside and outside the university but in particular my childhood friends N. K., A. Z., and G. D. for their unconditional support. Additionally, a big thank you goes to the city of Paris, where a significant portion of this thesis was conducted.
Last but not least, I would like to thank my family, the invisible heroes who have supported me all that time. I would probably not be here without the support of my parents Ploutarchos and Ioanna and my siblings Serafim, Popi and Manolis.

Finally, I would like to dedicate this work to my lovely nieces and nephews Ioanna, Alexandros, Konstantinos and Nefeli.

TABLE OF CONTENTS

TABLES & FIGURES

ABSTRACT

This work focuses on investigating the use of machine learning models as rainfall-runoff models. For this purpose, the development of a Long Short-Term Memory, a Transformer, and a conceptual model was chosen. The Long Short-Term Memory model was selected due to its widespread use in numerous time series prediction problems and specifically runoff prediction. The Transformer model was chosen because, despite the extensive discussion around its capabilities as a natural language processing model, its use as a time series prediction model is still quite limited. For comparing the above models, the conceptual rainfall-runoff model Zygos was developed, which has been initially developed by the ITIA research team of National Technical University of Athens.

The training and application of these models were carried out separately for 164 catchment basins from the CARAVAN dataset. These basins are located in the United States and were selected to cover a wide range of hydrological conditions. The analysis of the final models was performed at the catchment basin level, where the models were trained, calibrated, and tested for the same period of time.

The results show that machine learning models can be effectively used as rainfall-runoff models, as they outperform traditional models in performance criteria.

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία εστιάζει στη διερεύνηση της χρήσης μοντέλων μηχανικής μάθησης ως μοντέλων βροχής-απορροής. Για τον σκοπό αυτό επιλέχθηκε η ανάπτυξη ενός Long Short-Term Memory, ενός Transformer και ενός εννοιολογικού μοντέλου. Η επιλογή του μοντέλου Long Short-Term Memory έγινε δεδομένης της ευρείας χρήσης του σε πλήθος προβλημάτων πρόβλεψης χρονοσειρών και συγκεκριμένα σε προβλήματα πρόβλεψης απορροής. Το μοντέλο Transformer επιλέχθηκε διότι, παρά την εκτενή χρήση ως μοντέλου επεξεργασίας φυσικής γλώσσας (Natural language processing), η χρήση του ως μοντέλου πρόβλεψης χρονοσειρών είναι ακόμα αρκετά περιορισμένη. Για τη σύγκριση των παραπάνω μοντέλων, εφαρμόστηκε το εννοιολογικό μοντέλο Zygos, που έχει αναπτυχθεί από την ερευνητική ομάδα ΙΤΙΑ του Εθνικού Μετσόβιου Πολυτεχνείου.

Η ανάπτυξη και εφαρμογή αυτών των μοντέλων πραγματοποιήθηκε ξεχωριστά για 164 λεκάνες απορροής που περιέχονται στο σετ δεδομένων του CARAVAN. Αυτές οι λεκάνες βρίσκονται στις Ηνωμένες Πολιτείες και επιλέχθηκαν για να καλύψουν ένα ευρύ φάσμα υδρολογικών συνθηκών. Η ανάλυση των τελικών μοντέλων έγινε σε επίπεδο λεκάνης απορροής, όπου τα μοντέλα εκπαιδεύτηκαν - βαθμονομήθηκαν και αξιολογήθηκαν για την ίδια χρονική περίοδο.

Τα αποτελέσματα δείχνουν ότι τα μοντέλα μηχανικής μάθησης μπορούν να χρησιμοποιηθούν αποτελεσματικά ως μοντέλα βροχής-απορροής, καθώς υπερέχουν στα κριτήρια επίδοσης σε σύγκριση με το παραδοσιακό εννοιολογικό μοντέλο.

ΕΚΤΕΝΗΣ ΠΕΡΙΛΗΨΗ

Εισαγωγή

Η αναπαράσταση των φυσικών διεργασιών που συμβαίνουν κατά τη διάρκεια μιας βροχόπτωσης σε μια υδρολογική λεκάνη αποτελεί την κύρια και πιο διαχρονική πρόκληση των υδρολόγων μηχανικών. Η προσομοίωση των υδρολογικών συνιστωσών κατά την διάρκεια εισροής νερού σε μια υδρολογική λεκάνη, υπό την μορφή κατακρημνισμάτων και ο υπολογισμός του νερού που απορρέει στα υδάτινα σώματα ονομάζεται προσομοίωση βροχής απορροής. Η μοντελοποίηση βροχής-απορροής είναι κρίσιμη για τη διαχείριση των υδατικών πόρων και για τη λήψη αποφάσεων (decision-making). Ακριβείς και προηγμένες προβλέψεις ροής από προσομοιώσεις βροχής-απορροής μπορούν να βοηθήσουν στην αντιμετώπιση ζητημάτων διαχείρισης υδατικών πόρων και στον μετριασμό των επιπτώσεων πλημμυρών και ξηρασιών (Beven 2012).

Υπάρχουν διάφορες προσεγγίσεις για τη μοντελοποίηση βροχής-απορροής, που κυμαίνονται από φυσικά ή εννοιολογικά μοντέλα έως μοντέλα που βασίζονται στην μηχανική μάθηση. Τα φυσικά μοντέλα, τα οποία βασίζονται σε λεπτομερή κατανόηση των φυσικών διεργασιών, σπάνια χρησιμοποιούνται για προβλέψεις ροής λόγω της πολυπλοκότητάς τους. Αντίθετα, τα εννοιολογικά μοντέλα, που είναι γενικά απλούστερα και απαιτούν λιγότερα δεδομένα, χρησιμοποιούνται πιο συχνά για αυτό το σκοπό (Beven 2012). Στην εποχή των μεγάλων δεδομένων (big data), τα μοντέλα που μπορούν να αξιοποιήσουν μεγάλες ποσότητες δεδομένων διερευνώνται ευρέως σε διάφορα επιστημονικά πεδία. Αυτά τα μοντέλα χρησιμοποιούν τεχνικές μηχανικής μάθησης για να εξάγουν εξαρτήσεις και σχέσεις που προέχονται από τα δεδομένα εισόδου. Παρόλο που τα μοντέλα μηχανικής μάθησης για υδρολογικούς σκοπούς έχουν εξερευνηθεί αρκετά την τελευταία δεκαετία, εξακολουθεί να υπάρχει μια υστέρηση συγκριτικά με άλλα επιστημονικά πεδία.

Ο σκοπός αυτής της μελέτης είναι να διερευνήσει την εφαρμοσιμότητα δύο μοντέλων μηχανικής μάθησης σε αντίθεση με ένα κλασικό εννοιολογικό μοντέλο βροχής-απορροής. Συγκεκριμένα, η μελέτη δοκιμάζει τα μοντέλα μηχανικής μάθησης Long Short-Term Memory (LSTM) και Transformer έναντι του εννοιολογικού μοντέλου Zygos. Η σύγκριση αξιολογεί όχι μόνο την αξιοπιστία των προβλέψεων κάθε μοντέλου, αλλά και την πολυπλοκότητα των αρχιτεκτονικών τους καθώς και τις απαιτήσεις που έχει το καθένα σε υπολογιστικούς πόρους. Τα μοντέλα εκπαιδεύτηκαν στο σετ δεδομένων Caravan (Kratzert et al. 2023), που περιλαμβάνει σαράντα έτη (1981-2020) ημερήσιων μετεωρολογικών δεδομένων και

τριανταπέντε έτη (1981-2015) ημερήσιων δεδομένων απορροής. Ενώ το σετ δεδομένων Caravan περιέχει δεδομένα από 6.830 λεκάνες παγκοσμίως, η παρούσα μελέτη εστιάζει σε 164 λεκάνες από το υποσύνολο δεδομένων CAMELS (ΗΠΑ). Η επιλογή αυτών έγινε με κριτήρια τις διαφορετικές υδρολογικές συνθήκες και τους αντικειμενικούς υπολογιστικούς περιορισμούς.

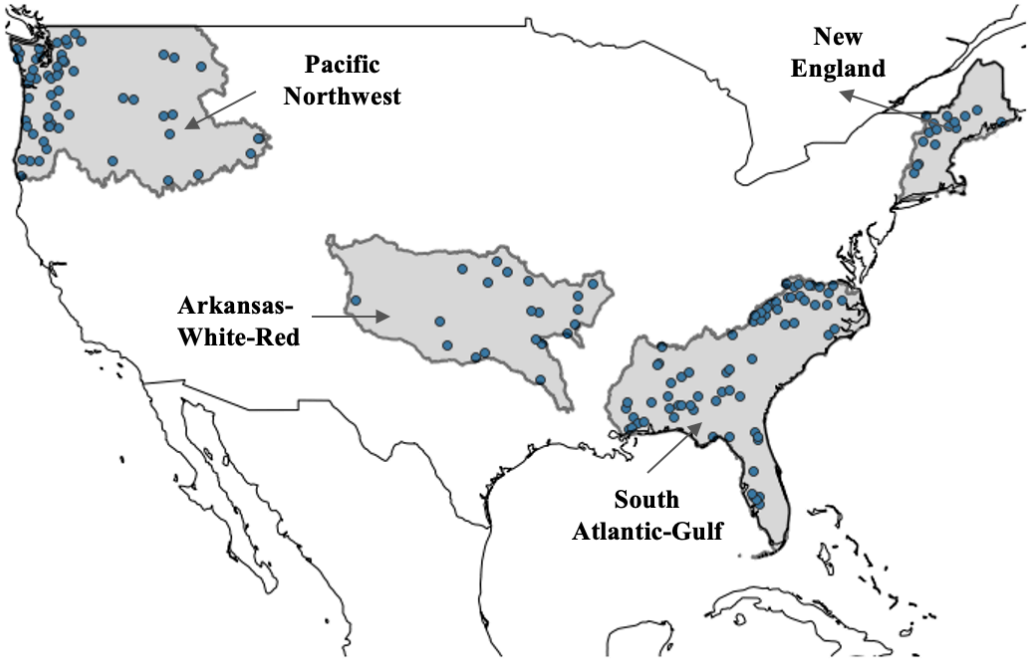
Συμπερασματικά, 164 μοντέλα LSTM, Transformer και Zygos αναπτύχθηκαν σε αντίστοιχο αριθμό λεκανών απορροής του Caravan. Αυτά τα μοντέλα προβλέπουν την ημερήσια απορροή χρησιμοποιώντας υδρομετεωρολογικά δεδομένα ως δεδομένα εισόδου.

Όλα τα μοντέλα αναπτύχθηκαν σε Python, χρησιμοποιώντας βασικές βιβλιοθήκες όπως οι Pandas, NumPy και TensorFlow για την επεξεργασία δεδομένων, αριθμητικούς υπολογισμούς και διαδικασίες βαθιάς μάθησης (deep learning).

Περιοχή Μελέτης και Σετ Δεδομένων

Το Caravan (Kratzert et al. 2023) είναι ένα ανοιχτό σετ δεδομένων που περιλαμβάνει μετεωρολογικά δεδομένα, χαρακτηριστικά λεκανών και δεδομένα απορροής για λεκάνες από όλο τον κόσμο. Τα μετεωρολογικά δεδομένα προήλθαν από το ERA5-Land (Muñoz-Sabater et al. 2021), τα χαρακτηριστικά των λεκανών ελήφθησαν από το ERA5-Land και το HydroATLAS (Linke et al. 2019), και τα δεδομένα απορροής προήλθαν από επτά ανοιχτά σετ δεδομένων.

Σε αυτή τη μελέτη χρησιμοποιήθηκαν 164 λεκάνες από το υποσύνολο των δεδομένων CAMELS (US) (Newman et al. 2015) του Caravan. Η παρούσα εργασία ακολούθησε τη μεθοδολογία που προτάθηκε από τους Kratzert et al. 2018, η οποία προτείνει τη χρήση 4 από τις συνολικά 18 υδρολογικές μονάδες, που οριοθετούνται από τον χάρτη της U.S. Geological Survey (Seaber, Kapinos, and Knapp 1987) για κάθε υδρολογική μονάδα (Hydrological Unit Map), για την κάλυψη ενός μεγάλου φάσματος υδρολογικών συνθηκών και τη μείωση του υπολογιστικού κόστους. Συγκεκριμένα εξετάστηκαν, οι υδρολογικές μονάδες: (01) New England, (03) South Atlantic-Gulf, (11) Arkansas-White-Red και (17) Pacific Northwest, όπως φαίνεται και στην παρακάτω εικόνα.



***Εικόνα 1**. Οι 164 υδρολογικές λεκάνες του CARAVAN.*

Αυτές οι τέσσερις υδρολογικές μονάδες περιέχουν 172 λεκάνες, αλλά 8 από αυτές έχουν ελλιπή δεδομένα που δεν μπορούσαν να συμπληρωθούν. Επομένως, αναπτύχθηκαν 164 μοντέλα ως μοντέλα βροχής-απορροής.

Μέθοδοι & Εργαλεία

Το μοντέλο LSTM, που αρχικά αναπτύχθηκε από τους Hochreiter and Schmidhuber το 1997, είναι ένας τύπος αναδρομικού νευρωνικού δικτύου με ειδική αρχιτεκτονική σχεδιασμένη να ξεπερνά την αδυναμία του παραδοσιακού Recurrent Neural Network (RNN) να εξάγει μακροχρόνιες εξαρτήσεις (Goodfellow, Bengio, and Courville 2016). Τα βασικά στοιχεία ενός LSTM κυττάρου (LSTM cell) είναι η κατάσταση του κυττάρου (Cell State), οι πύλες λήθης (Forget Gate), οι πύλες εισόδου (Input Gate) και οι πύλες εξόδου (Output Gate). Η κατάσταση του κυττάρου (Cell State) λειτουργεί ως η μνήμη του δικτύου, μεταφέροντας πληροφορίες στο επόμενο χρονικό βήμα. Η πύλη λήθης (Forget Gate) ελέγχει τις πληροφορίες που θα απορριφθούν από την κατάσταση του κυττάρου (Cell State). Η πύλη εισόδου (Input Gate) ελέγχει τις νέες πληροφορίες που θα προστεθούν στην κατάσταση του κυττάρου (Cell State), ενώ η πύλη εξόδου (Output Gate) ελέγχει τις πληροφορίες που περνούν από την κατάσταση του κυττάρου (Cell State) στην επόμενη κρυφή κατάσταση (Hidden State).

Η αρχιτεκτονική του δικτύου LSTM, όπως σχεδιάστηκε και χρησιμοποιήθηκε από τους Kratzert et al. 2018, έχει αποδειχθεί αποτελεσματική, οδηγώντας στην υιοθέτησή της σε αυτή τη μελέτη.

Το μοντέλο Transformers, έχει αναδειχθεί πρόσφατα από τους Vaswani et al. 2017, και πραγματοποιείται εκτεταμένη έρευνα γύρω από την δυναμική και τις δυνατότητες του τα τελευταία χρόνια. Το μοντέλο αυτό χρησιμοποιεί έναν μηχανισμό που ονομάζεται scaled dot-product attention, ο οποίος επιτρέπει στο μοντέλο να εντοπίζει και να καταγράφει μακροχρόνιες εξαρτήσεις. Το μοντέλο χρησιμοποιεί μια δομή κωδικοποιητή-αποκωδικοποιητή (Encoder - Decoder) όπου ο κωδικοποιητής (Encoder) επεξεργάζεται τα δεδομένα εισόδου ενώ ο αποκωδικοποιητής (Decoder) χρησιμοποιεί τα δεδομένα εξόδου του κωδικοποιητή (Encoder) για να δημιουργήσει τη σειρά εξόδου. Οι Yin et al. 2022 πρότειναν το RR-Former το οποίο έχει αποδειχθεί αρκετά αποτελεσματικό ως μοντέλο βροχής-απορροής και επομένως στην παρούσα μελέτη χρησιμοποιήθηκε παρόμοια αρχιτεκτονική.

Τέλος, το μοντέλο Zygos είναι ένα ντετερμινιστικό εννοιολογικό μοντέλο που αρχικά αναπτύχθηκε από ερευνητές της ομάδας ΙΤΙΑ στο Εθνικό Μετσόβιο Πολυτεχνείο. Εφαρμόζει ένα εννοιολογικό σχήμα λογιστικής υγρασίας εδάφους, βασισμένο σε μια γενίκευση του τυπικού μοντέλου Thornthwaite, επεκτεινόμενο με μία δεξαμενή υπόγειων υδάτων (Kozanis and Efstratiadis 2006). Για την προσαρμογή μιας απαραίτητης ρουτίνας χιονοκάλυψης, αναπτύχθηκε σε αυτή τη μελέτη μια ενημερωμένη έκδοση του αρχικού μοντέλου Zygos, προτεινόμενη από τους Efstratiadis, Nalbantis, and Koutsoyiannis 2015.

Η καταλληλότητα των μοντέλων αξιολογήθηκε χρησιμοποιώντας το κριτήριο επίδοσης Nash-Sutcliffe (NSE) (Nash and Sutcliffe 1970) καθώς και το κριτήριο Kling-Gupta (KGE) (Hoshin Vijai Gupta and Kling 2011). Επιπλέον, η μεροληψία (BIAS), το μέσο απόλυτο ποσοστιαίο σφάλμα (Mean Absolute Percentage Error) και το ριζικό μέσο τετραγωνικό σφάλμα (Rout Mean Squared Error) χρησιμοποιήθηκαν για να αξιολογήσουν την επίδοση των μοντέλων στις 5% παρατηρούμενες μέγιστες τιμές απορροής.

Αποτελέσματα

Τα αποτελέσματα των μέτρων επίδοσης NSE και KGE για τα 164 μοντέλα βροχής-απορροής LSTM, Transformer και Zygos απεικονίζονται στον ακόλουθο πίνακα (Πίνακας 1). Σημειώνεται, ότι τόσο τα μοντέλα LSTM όσο και τα Transformer παρουσιάζουν καλύτερη απόδοση σε σύγκριση με το μοντέλο Zygos. Συγκεκριμένα, οι μέσες τιμές NSE στο σύνολο των λεκανών είναι 0.50 για το LSTM, 0.71 για το Transformer και 0.37 για το Zygo, τονίζοντας την αποτελεσματικότητα των αρχιτεκτονικών LSTM και Transformer στην πρόβλεψη της

απορροής. Ο δείκτης KGE ακολουθεί τη λογική του NSE όσον αφορά την απόδοση των μοντέλων. Και τα δύο μοντέλα μηχανικής μάθησης υπερτερούν του πρότυπου μοντέλου με μέσες τιμές KGE 0.23, 0.69 και 0.07 για τα μοντέλα LSTM, Transformer και Zygos, αντίστοιχα.

***Πίνακας 1**. Μέτρα επίδοσης των μοντέλων LSTM, Transformer και Zygos.*

Model	Dataset	Max NSE	Max KGE	Mean NSE	Mean KGE	Median NSE	Median KGE	Min NSE	Min KGE
LSTM	training	0.957	0.875	0.635	0.364	0.650	0.524	0.067	-3.574
	validation	0.927	0.955	0.571	0.379	0.571	0.492	-0.119	-2.928
	testing	0.880	0.910	0.501	0.234	0.449	0.403	**-0.053**	-2.924
Transformer	training	0.993	0.994	0.843	0.795	0.908	0.892	0.353	-0.761
	validation	0.989	0.987	0.777	0.761	0.829	0.832	0.278	-0.223
	testing	**0.987**	**0.990**	**0.712**	**0.690**	**0.786**	**0.813**	-0.418	**-1.141**
Zygos	training	0.884	0.905	0.415	0.122	0.373	0.269	-0.024	-2.612
	testing	0.844	0.851	0.374	0.075	0.354	0.234	-0.184	-3.071

Επιπλέον, για την διερεύνηση της απόδοσης των μοντέλων κατά τη διάρκεια των μέγιστων απορροών, υπολογίστηκαν τα κριτήρια μεροληψίας (BIAS), μέσου απόλυτου ποσοστιαίου σφάλματος (MAPE) και ριζικού μέσου τετραγωνικού σφάλματος (RMSE) για το 5% των μεγίστων ροών. Ο ακόλουθος πίνακας (Πίνακας 2) δείχνει ότι το μοντέλο Transformer επιτυγχάνει καλύτερη απόδοση σε σύγκριση με τα μοντέλα LSTM και Zygos.

***Πίνακας 2**. Τιμές των μέτρων επίδοσης BIAS, MAPE και RMSE για το 5% των μεγίστων απορροών.*

Model	Mean BIAS	Median BIAS	Mean MAPE	Median MAPE	Mean RMSE	Median RMSE
LSTM	-0.33	-0.34	0.47	0.40	6.02	5.31
Transformer	-0.11	-0.09	0.29	0.24	4.75	4.08
Zygos	-0.42	-0.40	0.52	0.46	7.22	5.91

Συμπεράσματα & Συζήτηση

Ο κύριος στόχος αυτής της μελέτης είναι να διερευνήσει τις δυνατότητες δύο πρωτοποριακών μοντέλων μηχανικής μάθησης, των LSTM (Long Short-Term Memory) και Transformer, ως μοντέλων βροχής-απορροής. Αναπτύχθηκαν 164 μοντέλα LSTM και 164 μοντέλα Transformer και η απόδοση αυτών συγκρίθηκε με τα αντίστοιχα εννοιολογικά μοντέλα Zygos. Συγκεκριμένα, οι διαφορές στις μέσες τιμές των κριτηρίων επίδοσης Nash-Sutcliffe (NSE) και

Kling-Gupta (KGE) μεταξύ των μοντέλων Transformer και Zygos είναι 0.338 και 0.615, αντίστοιχα ενώ οι διαφορές για τα μοντέλα LSTM και Zygos είναι 0.127 και 0.159, αντίστοιχα. Αυτό υποδεικνύει ότι, τα μοντέλα Transformer υπερέχουν των μοντέλων LSTM και του μοντέλου Zygos ως μοντέλων βροχής-απορροής. Επιπλέον, τα κριτήρια επίδοσης για την πρόβλεψη μέγιστων απορροών, όπως η μεροληψία (BIAS), το μέσο απόλυτο ποσοστιαίο σφάλμα (MAPE) και το ριζικό μέσο τετραγωνικό σφάλμα (RMSE), που παρουσιάζονται στον Πίνακα 2, δείχνουν ότι το μοντέλο Transformer επιτυγχάνει καλύτερη απόδοση στην πρόβλεψη των μέγιστων ροών.

Τα κύρια συμπεράσματα αυτής της μελέτης είναι τα εξής:

- Τα μοντέλα μηχανικής μάθησης μπορούν να χρησιμοποιηθούν ως μοντέλα βροχής-απορροής.
- Τα μοντέλα μηχανικής μάθησης είναι ικανά να προβλέψουν την απορροή σε λεκάνες για διάφορες υδρολογικές συνθήκες, γεγονός που σημαίνει ότι μπορούν να “μάθουν” εξαρτήσεις που συσχετίζονται με ποικίλες υδρολογικές διεργασίες.
- Οι υπολογιστικοί πόροι σε όρους χρόνου είναι ουσιώδεις τόσο για την βαθμονόμηση των κλασικών εννοιολογικών μοντέλων όσο και για την εκπαίδευση των μοντέλων μηχανικής μάθησης.

1 INTRODUCTION

1.1 Research scope

Hydrologists and engineers have long sought to represent the physical processes occurring during rainfall events due to the significant importance of water in human life over the centuries. These processes are described within a framework known as the water cycle, which encompasses many diverse and complex interactions. Because of this complexity, it is almost infeasible to represent water cycle processes in a purely physically based manner. Consequently, water engineers have developed alternative methods to model the water cycle system.

One key component of the water cycle that is crucial to define, due to its significant impact on social and economic life and the environment, is runoff. The runoff variable plays an important role in water resource management, flood and drought mitigation, and environmental protection. Rainfall-runoff relationships describe how basin discharge responds to mass inputs like precipitation and energy inputs like radiation. A hydrological model is defined as a set of mathematical transformations that use field data and reasonable assumptions about the processes of the hydrological cycle and their interactions, with the aim of quantitatively estimating the variables of interest (A. Efstratiadis 2008).

Hydrologist have developed various approaches in rainfall-runoff modeling in order to predict runoff. The most common approach among them, is the rainfall-runoff conceptual modeling, translating complex non-linear processes of the water cycle in a simple and understandable way. The conceptual model may be more or less complex, ranging from the use of simple mass balance equations for components representing storage in the catchment to coupled nonlinear partial differential equations (Beven 2012).

Another approach that has recently emerged for rainfall-runoff modeling is data-driven modeling utilizing artificial intelligence (AI) techniques. Machine learning (ML), a subset of artificial intelligence, tries to mimic the functioning of the human brain by acquiring knowledge through a learning process. Machine learning models have the ability to learn and generalize 'knowledge' from data pairs, enabling them to solve large-scale, complex problems (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000). Many machine learning models have been developed in recent decades, with Recurrent Neural Networks (RNNs) being among the most common for addressing regression problems. A special type of RNN is the Long Short-Term Memory (LSTM) network, which proposed by

Hochreiter and Schmidhuber 1997. Recently, (Vaswani et al. 2017), in their work *'Attention is All You Need*,' proposed a highly promising model called Transformer and has been gaining a lot of attention since its release.

The purpose of this research is to investigate the applicability of two machine learning models compared to a classical conceptual model for rainfall-runoff modeling. This investigation is motivated by the ongoing discussion among scientists about the potential capabilities of data-driven models. Specifically, 164 rainfall-runoff LSTMs, Transformers and Zygos models were trained – calibrated and tested utilizing catchments from the Caravan dataset. The comparison evaluates not only the reliability of each model's predictions but also the complexity of their architectures and their computational cost demands.

The main four research questions that this study tries to answer are as follows:

- Can machine learning models be used as rainfall-runoff models?
- Can machine learning models outperform the classical conceptual one?
- Do machine learning models are more complex in terms of its architecture?
- Do machine learning models need more computational resources to be trained?

1.2 Work structure

The study is organized as follows:

- Chapter 1 introduces the subject of the thesis and its research objectives.
- Chapter 2 presents the advances in hydrology and specifically in rainfall-runoff modeling utilizing machine learning methods. Moreover, introduces the subject of Large-sample hydrology.
- Chapter 3 provides a detailed overview of the Caravan dataset and the catchments utilized in this thesis.
- Chapter 4 presents the model architectures used in this thesis as well as the model evaluation protocols.
- Chapter 5 presents the experimental results and provides a comparison of the models.
- Chapter 6 gives conclusion and discusses further future work.

2 ADVANCES IN HYDROLOGICAL MODELING

2.1 Introduction

In general, a model is a simplified representation of real-world system which means that the best model is the one that gives results close to reality with the use of least parameters and model complexity. Models are mainly used for predicting system behavior and understanding various hydrological processes and can be classified into two main categories. The first one is whether the model is deterministic or stochastic. Deterministic models produce a single output for each iteration given a specific set of inputs and parameter values. In contrast, stochastic models account for uncertainty in input variables, boundary conditions, or model parameters, allowing for some randomness in the outcomes (Beven 2012). The second one is whether the models would be physically-based, conceptual, statistical-stochastic or data-driven (A. Efstratiadis 2008). The first one is based on theoretical equations or semi-empirical equations from experimental data. Conceptual models are based on parametric relationships that represent the basic processes of the system. Furthermore, statistical or stochastic models reproduce the basic statistical structure of the observed samples. Finally, data-driven models transform the input data to derive complex cause-and-effect relationships.

The most common approach among hydrologist to represent the complex system of water cycle is the conceptual modeling. Many studies have been carried out using different conceptual approaches for rainfall-runoff modeling. Among the most prominent and widely used models among else are the TOPMODEL (Beven and Kirkby 1979), MIKESHE (www.dhigroup.com), Soil and Water Assessment Tool (SWAT) (Arnold et al. 1998) and Sacramento Soil Moisture Accounting model (SAC-SMA) (Burnash 1973). Models are typically created to address specific questions and therefore, they cannot be compared in a general way (Gehlert and Pfeiffer 2005). However, many studies have been undertaken comparing the performance of them. For example, the study conducted by the Word Meteorological Organization (1975) applied 10 rainfall–runoff models to six different catchments, and compared them in terms of the physical concepts used, data and computer requirements, and level of accuracy under different hydro-climatic conditions.

In general, conceptual models are based on two criteria: firstly, the structure of the model is specified prior to any modeling being undertaken, and secondly not all of the model parameters have a direct physical interpretation. Therefore at least some conceptual model parameters have to be estimated through calibration against observed data. The calibration problem for

hydrological modeling, despite being widely studied for over thirty years, has not yet been fully addressed and remains topical due to the complexity of modern models. Moreover, this issue is common across all types of models due to the necessity of adjusting the modeled data to match the ground truth (A. Efstratiadis 2008).

The common calibration approach for a conceptual rainfall-runoff model involves using an automatic optimization technique, where an objective function evaluates how well the modeled data matches the observed data.

The automatic calibration of a hydrological model can be mathematically addressed as an optimization problem of the form:

$$max g(e) = g[y - h(s_0, x, \theta)], \;\; s.t. \theta \in \Theta \tag{2.1}$$

where g(.) is a set of goodness-of-fit measures, $\Theta \subset Rp$ is the feasible space, and e is the error vector or residual of the model, defined as the difference between observed and simulated responses, specifically:

$$e = y - y' \tag{2.2}$$

Typically, the feasible space is defined by two vectors of extreme values θ_{min} and θ_{max}, which expresses the allowable range of parameter value.

The goodness-of-fit g(.) function is a numerical measure of the difference between the model simulated output and the observed output (Schaefli and Gupta 2007). There are many objective functions that can be found in the literature with the most common ones based on the standard least squares methods and maximum likelihood methods (Pechlivanidis et al. 2011). Calibration techniques relying on a single objective function often fail to capture all key characteristics of the modeled system. The need for multi-objective calibration stems from the limitations of single objective calibration in accurately characterizing and constraining model behaviors, as well as advancements in optimization technology (Khu and Madsen 2005).

It should be noted that the optimization problem, as formulated in equation 2.2, is a multi-objective one since the function g is vector-valued. To reduce it to a single-objective problem, so that it can be tackled with standard extremum search methods, a unified numerical expression must be formulated in terms of the errors e, which describes an overall criterion of the model's goodness-of-fit to the measured responses y.

The process of estimating the parameters of a hydrological model (known as the inverse hydrological problem) can be automated as follows (A. Efstratiadis 2008):

- A sample of measured (observed) responses is selected.

- A measure of the model's fit to the observations is chosen.
- The problem of global optimization is formulated (stochastic function, control variables, feasible parameter limits).
- A suitable algorithm is selected to search for the most appropriate parameter values, with a reasonable number of trials.

The calibration problem is resource-intensive, requiring significant computational power to determine the most suitable parameters.

In recent years, a lot of research has been devoted to developing automated calibration routines or procedures based on numerical optimization techniques such as genetic algorithms (Goldberg and Holland 1988) and the shuffled complex evolution (SCE) algorithm (Duan, Gupta, and Sorooshian 1993). The need for automatic calibration routines in hydrologic models has also been widely recognized over many years as demonstrated by the amount of work done in this area. A comprehensive study about multi-objective calibration approaches was conducted by Andreas Efstratiadis and Koutsoyiannis 2010.

After calibration, the performance of the optimized model parameters is always checked against an independent time period. This process evaluates the predictive capacity of the model.

Regardless of the strategy adopted, the calibration of a hydrological, model is considered reliable if:

- The model has sufficient predictive capability, meaning it can reproduce the entire range of responses of a basin with satisfactory accuracy.
- The optimized parameters of the model can be attributed some physical meaning, so they are considered compatible with the characteristics of the natural system.

2.2 Machine learning approaches in Rainfall-Runoff modeling

Artificial intelligence (AI) based models have recently emerged as powerful tools to enhance hydrological modeling, offering new approaches to handle large datasets, capture non-linear relationships, and improve predictive accuracy. Many machine learning techniques have been applied for addressing various regression and classification problems. Popular algorithms include linear regression, logistic regression, classification and regression tree (CART), naive Bayes model (NB), support vector machine (SVM), K-nearest neighbor (KNN), random forest (RF), and artificial neural networks (ANN).

Support vector machines for regression were first introduced by Cortes and Vapnik 1995, and the first applications were reported in the late 1990s. A comprehensive review by Raghavendra. N and Deka 2014 highlights the application of SVMs in the field of hydrology. In their paper the authors list nearly 40 SVMs models developed for various hydrological application. SVMs have been successfully applied for rainfall-runoff modeling as evidenced by studies such as Dibike et al. 2001; Bray and Han 2004; Asefa et al. 2006; Ch et al. 2013.

Random forest is a supervised machine learning algorithm which use decision trees as base learners. Random forests introduced by Breiman 2001, have been applied to several scientific fields and associated research areas such as agriculture, land cover classification, remote sensing, wetland classification and ecology. Tyralis, Papacharalampous, and Langousis 2019 highlight that although the practical value of random forest, it remains obscure with limited use in hydrological applications. Iorgulescu and Beven 2004 are perhaps the first authors to cite Breiman 2001 in a water resources journal for rainfall-runoff application. Several comparative studies with a hydrological focus have shown that random forest can outperform to other machine learning techniques such as artificial neural networks, support vector machines, and regression models (Erdal and Karakurt 2013; Li et al. 2016; Bachmair et al. 2017). Many studies have been conducted using random forest for rainfall-runoff modeling (see Galelli and Castelletti 2013; Gudmundsson and Seneviratne 2016; Shortridge, Guikema, and Zaitchik 2016; Worland, Farmer, and Kiang 2018; Chang and Chen 2018).

ANNs are a fundamental and essential component of many deep learning architectures, also known as dense layers or Multilayer Percepton or fully connected network. In the early 1990s, researchers began investigating the potential of neural networks for modeling watershed runoff based on rainfall inputs (see French, Krajewski, and Cuykendall 1992). In the first year of this century a task committee of American Society of Civil Engineers (ASCE) has discussed thoroughly and established the role of ANN in hydrology and also compared it with the other modelling methods (ASCE Task Committee on Application of Artificial Neural Networks in Hydrology 2000). Rakesh Tanty, Tanweer S. Desmukh, and Manit Bhopal 2015 conducted a review on the application of ANN to hydrological related problems. Their review highlights the development of ANN models in various areas, including rainfall-runoff modeling, streamflow modeling, water quality modeling and groundwater modeling applications (see Kalteh 2008; Goyal and Ojha 2010; Chen, Wang, and Tsou 2013).

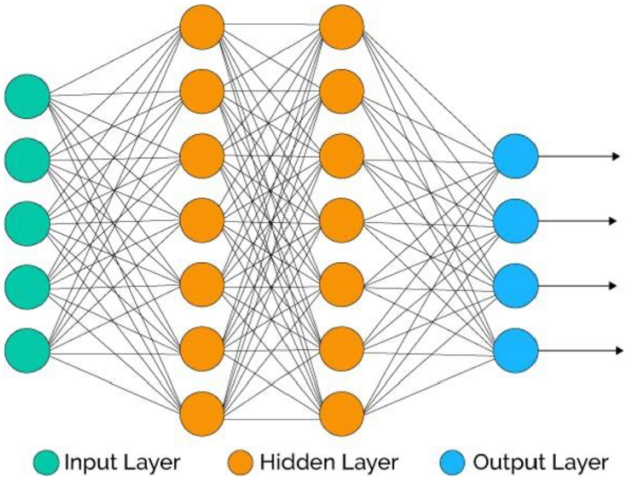


***Figure 2**. Diagram of an artificial network architecture (Goodfellow, Bengio, and Courville 2016).*

However, a drawback of feed-forward ANNs is that any information about the sequential order of the inputs is lost. Recurrent neural networks (RNNs) are a special type of neural network architecture that have been specifically designed to understand temporal dynamics by processing the input in its sequential order (Rumelhart, Hintont, and Williams 1986). Although, RNNs can detect patterns in sequential data they face difficulties when the sequence of data is long enough. The Long Short Term Memory (LSTM) models have been developed by Hochreiter and Schmidhuber 1997, as a type of RNN to address the vanishing gradient and exploding challenges in long sequences of data (Goodfellow, Bengio, and Courville 2016). The use of LSTM for modeling runoff has recently increased and more studies have been immersed over the last years (see Kratzert et al. 2018; 2019; Frame et al. 2021; Sanjay Potdar et al. 2021; Yin et al. 2021; Nevo et al. 2022; Yin, Wang, et al. 2022; Shrestha and Pradhanang 2023). Moreover, LSTM has shown great ability to handle long dependencies which is desirable for modeling processes like snow accumulation, seasonal vegetation patterns or other processes that have long timescales (Kratzert et al. 2018) and play significant role in rainfall-runoff modeling.

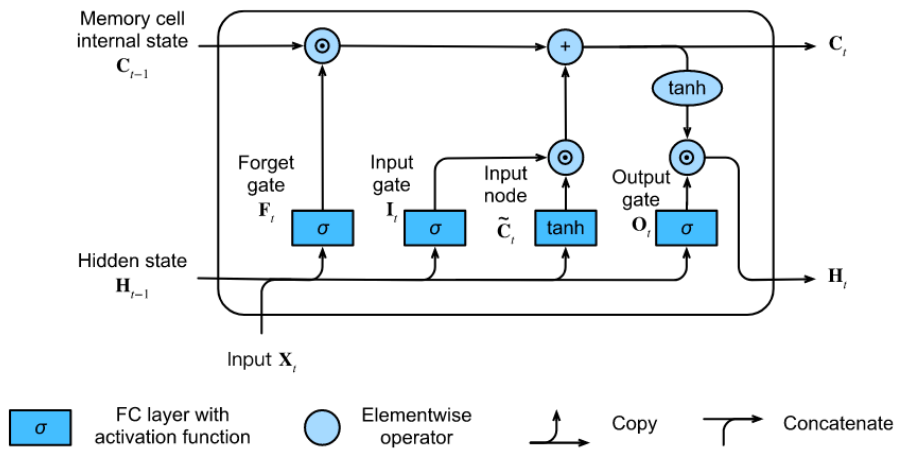


Figure 3. *Diagram of LSTM unit architecture (Calzone, 2022).*

Another machine learning model, proposed by Vaswani et al. 2017, is the Transformer model. This model utilizes the attention mechanism, instead of recurrency to capture relationships and patterns in the input data. Originally Transformer was designed for language processing, particularly for language translation tasks. Since then, there has been widespread implementation of Transformer models, with the most prominent applications being Chatbots like Chat-GPT. Despite the extensive research focusing on natural language processing (NLP) a significant research have been recently applied for various timeseries tasks. Wen et al. 2023 have summarized the recent studies being applied for in forecasting, anomaly detection and classification problems. Although, Transformers are increasingly explored in timeseries (see Kitaev, Kaiser, and Levskaya 2020; H. Zhou et al. 2021; Liu et al. 2022; T. Zhou et al. 2022; Wu et al. 2022; Shen and Wang 2022; Zhang and Yan 2023) the implementation of them in hydrology is scares. Yin et al. 2022 investigate the use of Transformer model for rainfall-runoff modeling and Amanambu, Mossa, and Chen 2022 for hydrological drought forecasting prepose.

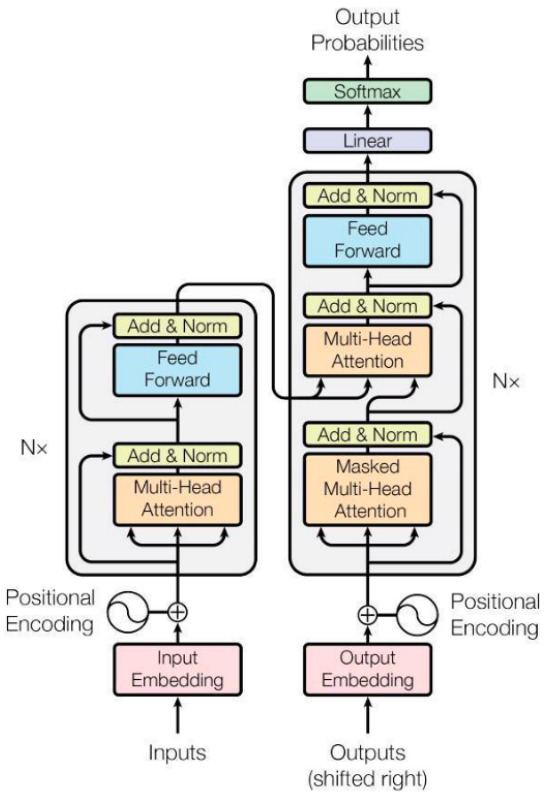


***Figure 4**. Transformer architecture (Vaswani et al., 2017).*

2.3 Large-sample hydrology

Gridded meteorological data sets have become increasingly prevalent, and with the availability of streamflow records and computing resources, large-sample hydrology (LSH) studies have been gaining significant traction over the past decade (Newman et al. 2015). Furthermore, to achieve better hydrological modeling across a variety of hydrological settings at multiple spatiotemporal scales and under changing environmental conditions, it is crucial to fully understand catchment processes and that cannot be achieved only through placed-based investigation or heavily instrumented catchments. To that extend, the use of LSH has been actively promoted for rainfall-runoff modeling (H. V. Gupta et al. 2014).

H. V. Gupta et al. 2014, provide a comprehensive list of studies with a focus on rainfall-runoff modeling in more than 30 catchments and summarized the reasons for using LSH as follows:

- To draw conclusions that require data from more than one catchment.
- To establish the range of applicability or the expected level of efficiency of methods/models.

- To ensure sufficient information to enable statistically significant relationships to be established.

LSH datasets can provide data classified into three categories: streamflow observations, hydrometeorological timeseries and landscape and hydroclimatic attributes. Addor et al. 2020, have summarized key LSH datasets that are available and cover different parts of the world, ranging from basins within a single country to those across the entire globe. The Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) dataset (Newman et al. 2015; Addor et al. 2017) uses recent datasets to provide up-to-date hydrometeorological variables and a variety of landscape attributes for 671 catchments across the United States of America. CARAVAN (a series of CAMELS) dataset (Kratzert et al. 2023) standardizes and aggregates seven existing LSH datasets. CARAVAN is both a dataset, containing 6830 catchments, and open-source software that allows members of the hydrology community to extend the dataset to new location (see also Chapter-3).

Large-sample hydrology has become an indispensable tool in modern hydrology, offering comprehensive insights into water-related processes and their management on a global scale. LSH has significantly applied in applications in flood and drought prediction, climate change impact assessment and water resources management. Many studies have applied LSH for rainfall-runoff modeling especially nowadays with the exploitation of data-driven model (Addor et al. 2017; Kratzert et al. 2018; 2019; Flamig, Vergara, and Gourley 2020; Frame et al. 2021; Yin, Guo, et al. 2022).

3 STUDY AREA AND DATASET

Caravan (Kratzert et al. 2023) is an open community dataset of meteorological forcing data, catchment attributes and discharge data for catchments around the world (Figure 5).

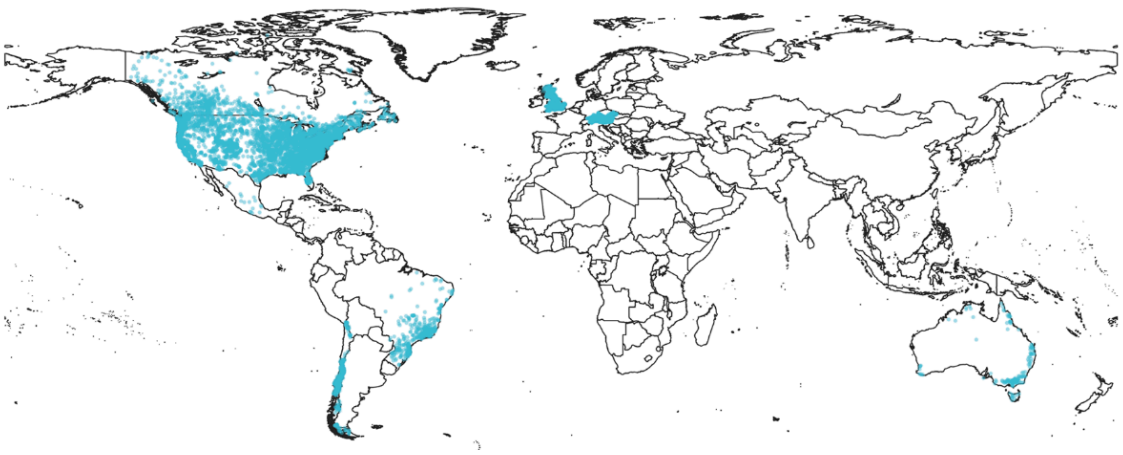


***Figure 5**. Caravan basins distribution (Kratzert et al. 2023).*

The meteorological forcing data was derived from the ERA5-Land product (Muñoz-Sabater et al. 2021), the basin attributes were taken from ERA5-Land and HydroATLAS (Linke et al. 2019), and the discharge data was sourced from seven open datasets:

- 482 basins from CAMELS (US)
- 150 basins from CAMELS-AUS
- 376 basins from CAMELS-BR
- 314 basins from CAMELS-CL
- 408 basins from CAMELS-GB
- 4621 basins from HYSETS
- 479 basins from LamaH-CE

In this study 164 basins were utilized from the CAMELS (US) (Newman et al. 2015) sub-dataset in Caravan. Following the methodology proposed by (Kratzert et al. 2018), which suggest using 4 out of the 18 hydrological units, delineated by the U.S. Geological Survey's HUC map (Seaber, Kapinos, and Knapp 1987), to cover a wide range of hydrological conditions and to reduce computational costs, a similar approach was adopted. Specifically, the hydrological units (01) New England, (03) South Atlantic-Gulf, (11) Arkansas-White-Red and (17) Pacific Northwest were considered in this study (Figure 6).

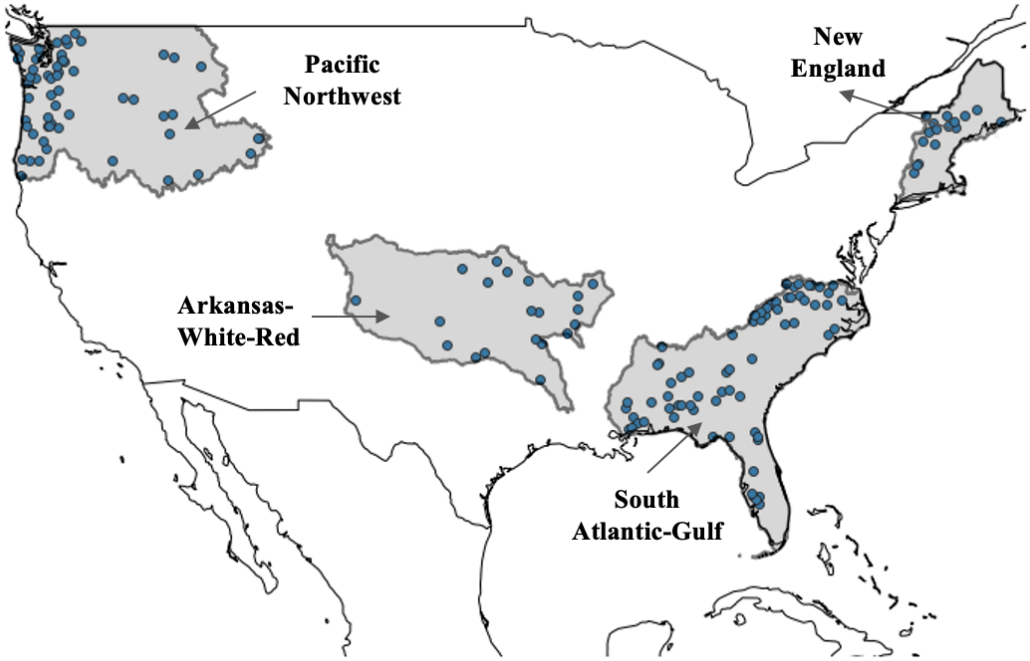


***Figure 6**. The 164 Basins in the 4 hydrological units.*

The New England region in the northeast comprises 16 basins that are relatively homogeneous, particularly in terms of snow influence and aridity. Arkansas-White-Red region, located centrally in the United States, has 20 basins with significantly different characteristics. This region exhibits a substantial variance in aridity and mean annual precipitation, with a pronounced gradient from east to west. Similarly sized but with diverse hydro-climatic conditions are the South Atlantic-Gulf region, containing 70 basins, and the Pacific Northwest region, with 58 basins. The Pacific Northwest stretches from the Pacific coast to the Rocky Mountains and shows a wide range of attributes across its basins, much like the Arkansas-White-Red region. For instance, some catchments near the Pacific coast receive over 3000 mm of precipitation annually, while areas in the southeast are extremely arid. In contrast, the relatively flat South Atlantic-Gulf region has more uniform basins. Unlike New England, this region is not affected by snow. Additionally, the South Atlantic-Gulf has a higher mean aridity (5.08 ± 1.2) compared to New England (3.69 ± 0.43), and its mean altitude (189 ± 179 m) is lower than that of New England (316 ± 182 m).

Table 3. *[1]PET/P, see Addor et al. 2017. [2]Fraction of precipitation failing on days with temperature below 0^0C. [3]Positive values indicate that precipitation peaks in summer, negative values that precipitation peaks in the winter month and values close to 0 that the precipitation is uniform throughout the year, see Addor et al. 2017.*

HUC	Region name	No. of basins	Mean precipitation (mm day $^{-1}$)	Mean aridity[1] (-)	Mean snow frac.[2] (-)	Mean seasonality[3] (-)	Mean altitude (m)
01	New England	16	3.39 ± 0.21	3.69 ± 0.43	0.31 ± 0.04	0.63 ± 0.17	316 ± 182
03	South Atlantic-Gulf	70	3.36 ± 0.31	5.08 ± 1.2	0.00 ± 0.00	0.25 ± 0.15	189 ± 179
11	Arkansas-White-Red	20	2.88 ± 0.61	4.74 ± 1.89	0.01 ± 0.05	0.27 ± 0.11	613 ± 713
17	Pacific Northwest	58	4.46 ± 1.55	3.09 ± 1.51	0.28 ± 0.26	1.6 ± 0.14	1077 ± 589

These four hydrological units contain 172 basins, but 8 of them have missing streamflow values that could not be infilled. Therefore, 164 models were trained and tested for rainfall-runoff modeling.

Table 4. *Basins with more than 10 missing values.*

Basin_id	**Missing values**
02178400	365
02202600	92
02231342	48
02235200	746
02310947	365
12025000	365
12141300	366
13310700	1767

Streamflow timeseries having less than 10 missing values were filled using the linear interpolation method.

4 MATERIALS AND METHODS

All models described below were developed in Python programming language using libraries such as NymPy, Pandas, TensorFlow, Matplotlib, Seaborn and GeoPandas for data manipulation, numerical computation, deep learning techniques and visualization.

4.1 Long Short – Term Memory

LSTMs models are a type of recurrent neural network with a special architecture designed to overcome the weakness of the traditional RNN to learn long dependencies. The clever idea of introducing self-loops to produce paths where the gradient can flow for long durations is a core contribution of the initial long short-term memory model (Goodfellow, Bengio, and Courville 2016). LSTMs has a chain structure with four interacting neural network layers.
The LSTM have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation.

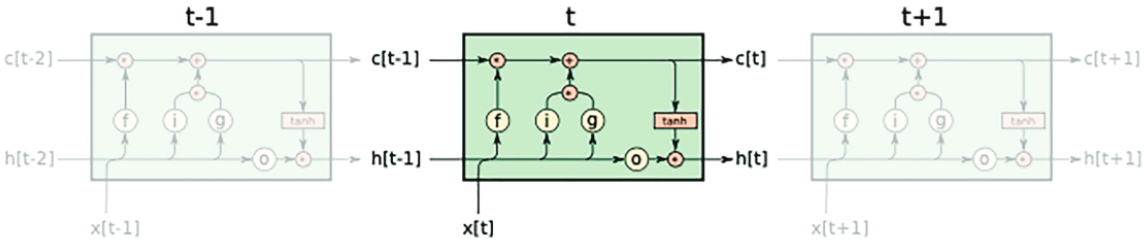


***Figure** 7. Visualization of the LSTM cell (Kratzert et al. 2019).*

In the above diagram, each line carries an entire vector, from the output of one node to the inputs of others. The cell state (c) is the memory component of the LSTM cell. It allows information to flow along the cells unchanged. It considers to be the key component in an LSTM cell due to the ability to retain information over long sequences. The forget gate (f) takes the input data from the current timestep and decides what information should be discarded from the cell state. The input gate (i) takes the input data from the current timestep and the previous hidden and decides what information is going to be stored in the cell state. The cell input (g) is a vector of new candidate values that should be added to the state. The output gate (o) takes input from the current input data, the previous hidden state and the updated cell state and

determines the information to be the output of the LSTM cell. The hidden state is the output of the LSTM.

The LSTM network is described by the following equations:

$$x[t] + U_i h[t-1] + b_i$$

$$f[t] = \sigma(W_f x[t] + U_f h[t-1] + b_f$$

$$g[t] = \tanh\,(W_g x[t] + U_g h[t-1] + b_g$$

$$o[t] = \sigma\,(W_o x[t] + U_o + b_o)$$

$$c[t] = f[t] \circ c[t-1] + i[t] \circ g[t]$$

$$h[t] = o[t] \circ \tanh\,(c[t])$$

where, x *is network input,*

f *is the forget gate,*

g *is the cell input,*

o *is the output gate,*

c *is the cell state,*

i *is the input gate,*

h *is the hidden state,*

W, U *and* b *are calibrated parameters and*

$tanh\,(\cdot)$ *and* $\sigma(\cdot)$ *are the hyberparabolic tangent and the sigmoid activation functions*

The LSTM network architecture, as designed by Kratzert et al. 2018, has been proven effective, leading to its adoption in this study. Table 5 shows the values of the hyperparameters for the LSTM model.

***Table 5**. LSTM model hyperparameters.*

Hyperparameters	Value
Sequence Length	365
Batch Size (train)	256
Batch Size (val)	2048
LSTM Units	20
Number of Layers	2
Dropout Rate	0.1
Epochs (Patience)	50 (5)
Loss	MSE
Optimizer	Adam

4.2 Transformer

Transformers models have shown superior performance in capturing long-range dependency than RNN models (H. Zhou et al. 2021). As depicted in Figure 4, the Transformer utilizes an encoder-decoder structure. The encoder processes the input sequence through multiple layers of multi – head attention and fully connected feed-forward network. The final output of the encoder is a set of vectors that captures the relevant information from the input sequence. This output is then passed to the decoder for generating the output sequence. The decoder processes the input sequence in a non-autoregressive manner which is different by the originally decoder block suggested by Vaswani et al. 2017. A key distinction in the decoder block is the incorporation of a masked self-attention layer, which prevents information leakage from future values during training.

Transformer models lack recurrence and instead of processing the input sequentially, they handle the entire sequence simultaneously using the scaled dot-product attention mechanism. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values and outputs are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query layers running in parallel.

The scaled dot-product attention mechanism combines the query and key vectors to determine how well they match, the “attention score” as described in the following equation.

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where, Q *is the query vectors,*

K *is the key vectors,*

V *is the value vectors and*

d_k *is the dimensionality of the key vectors.*

The scaling factor $\sqrt{d_k}$ prevents the dot products to grow large in magnitude.

Instead of performing a single attention function with d_{model}-dimensional keys, values and queries Vaswani et al. 2017 found it beneficial to linearly project the queries, keys and values h times with different learned linear projections to d_k, d_k and d_u dimensions, respectively. On each of these projected versions of queries, keys and values the attention function can be performed in parallel. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions.

$$MultiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h) \cdot W^0$$

$$\text{Where, } head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$

Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{model} x d_k}, W_i^K \in \mathbb{R}^{d_{model} x d_k}, W_i^V \in \mathbb{R}^{d_{model} x d_k}$ *and* $W_i^0 \in \mathbb{R}^{d_{model} x d_k}$.

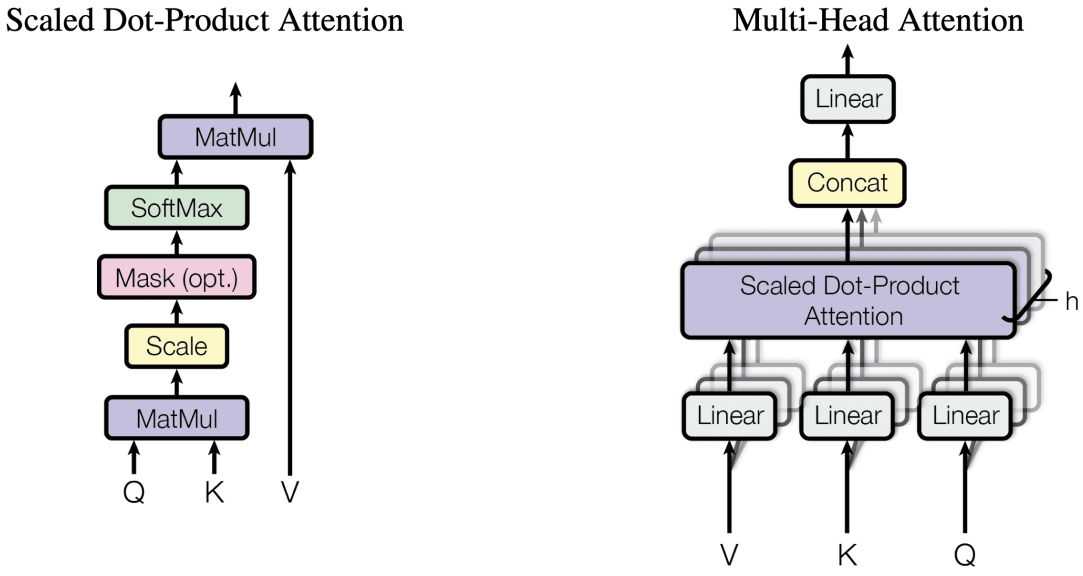


Figure 8*. Scaled Dot-Product Attention (left) - Multi-Head Attention (right) (Vaswani et al. 2017).*

The Transformer uses multi-head attention in both the encoder and decoder blocks. In the encoder block, all of the keys, values and queries come from the same place and therefore it is called self-attention layer.

Since transformer has neither recurrence nor convolution, in order for the model to make use of the order of the sequence, some information about the relative and absolute position must be injected to the model. This is done by using sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right)$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{model}}\right)$$

where, pos is the position and

i is the dimension.

This process is called positional embedding and it is implemented right after the linear transformation with a fully connected layer both in the encoder and decoder block.

Yin et al. 2022 proposed the RR-Former which has been proven quite effective in rainfall-runoff modeling and therefore the same model architecture was used. The architecture of the model is depicted in Figure 9.

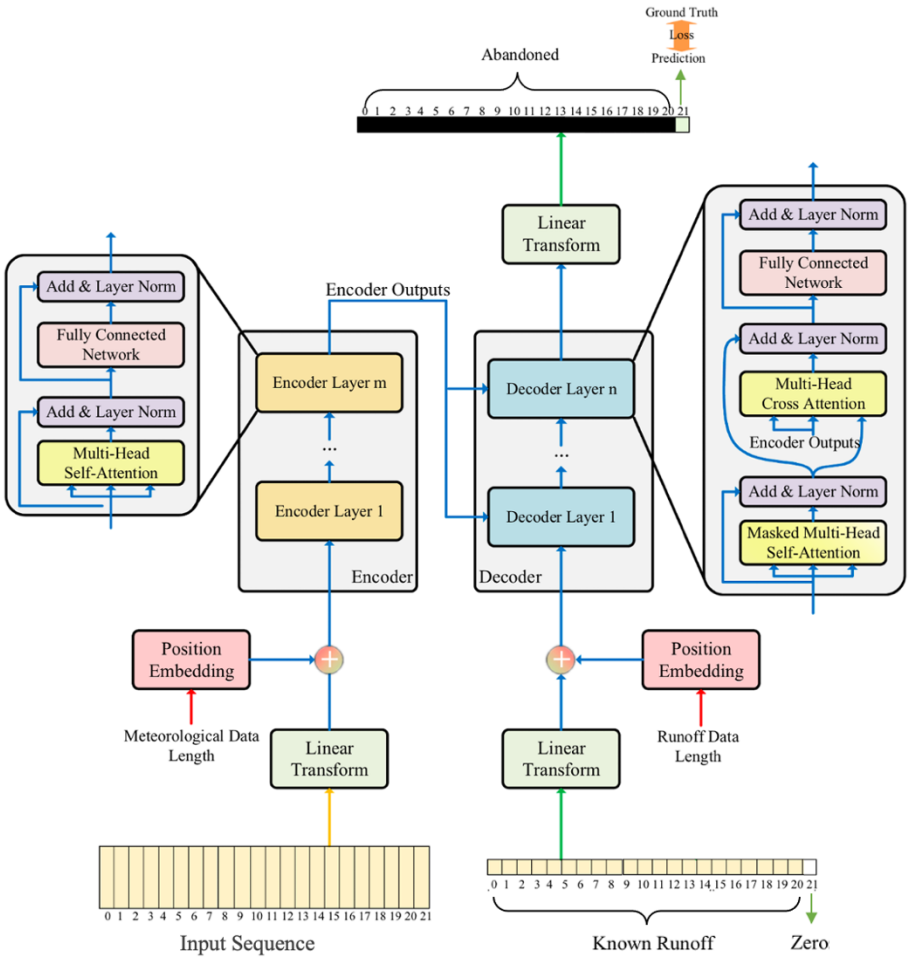


Figure 9. *Architecture of the transformer model.*

In order to fit the inputs to the models dimension a linear operation with a dense layer is applied converting the inputs to vectors with a dimension of the model. Then the output is added with the output of a learnable embedding which is implemented to give relative position information to the model. The final output is the same dimension with the model dimension and is passed to the encoder and decoder.

Each encoder layer has self-attention layers and position-wise fully connected networks. Residual connection and layer normalization are employed sequentially around each of the sub-layers. The encoder output is then passed to the decoder. Each decoder layer has masked self-attention layers, encoder-decoder attention layers and position-wise fully connected networks. Just like the encoder, residual connection and layer normalization are employed sequentially around each of the sub-layers.

It is worth mentioning, that due to the quadratic structure of the attention mechanism the sequence length could not be the same with that in the LSTM and thus was kept to three weeks.

The values of the Transformer model hyperparameters are shown in table 6.

***Table 6**. Transformer model hyperparameters.*

Hyperparameters	Value
Sequence length	21
Batch Size	256
Number of heads in multi-head attention	4
Number of encoder/decoder layers	4
Dropout rate	0.1
Model dimension	256
Dimension of the Position-wise fully connected layer	256
Learning Rate	0.0001
Optimizer	Adam
Loss Function	MSE
Epochs (Patience)	200 (10)

4.3 Zygos

The Zygos lumped conceptual model was selected as the benchmark model. Several variants of the Zygos model have been implemented for rainfall-runoff modeling. The scheme used in this study is based on the scheme introduced by Efstratiadis, Nalbantis, and Koutsoyiannis 2015. The model uses 11 parameters and the model inputs are the precipitation, the mean temperature and the potential evapotranspiration. Despite that Caravan dataset provides the mean daily potential evapotranspiration, it has been identified as not applicable for hydrological application due to the systematically overestimation of it (Clerc-Schwarzenbach et al. 2024). Hence, potential evapotranspiration is computed with Hargreaves method using the maximum and minimum daily temperature derived from the Caravan dataset.

As depicted in Figure 10, the basin is vertically subdivided into three storage elements or tanks that represent the snowpack, soil water and groundwater.

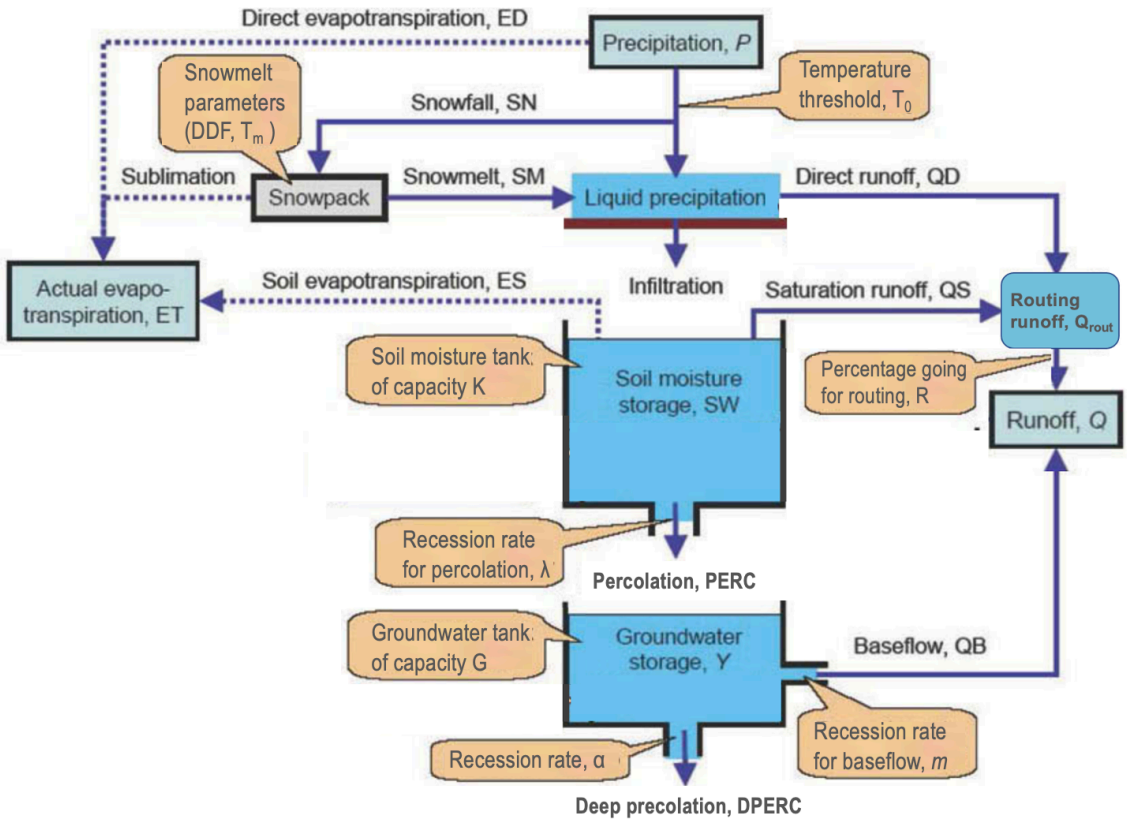


Figure 10. *Sketch of the Zygos model, illustrating the modeled processes and associated parameters.*

First, precipitation is considered as snowfall if temperature is below a certain threshold. In that case, precipitation is added to snowpack tank, otherwise precipitation is considered as liquid and fulfils, by priority the potential evapotranspiration. Snowfall and sublimation allow to update the water equivalent of the snowpack tank via the snowpack water balance. Then, the snowmelt (SM) is estimated through the degree-day factor.

$$SM = DDF(T - T_m)$$

where, SM, *is the snowmelt (mm)*

DDF, *is the degree-day factor (mm/d/ °C)*

T_m, *is the temperature threshold (°C).*

The snowmelt is added to the available liquid precipitation and the sum is contributed to direct runoff (QD).

$$QD = DP^2/(DP - SW + K)$$

where, QD, *is the direct runoff (mm)*

DP, *is the available liquid precipitation (mm)*

SW, *is the current soil water storage (mm)*

K, *is the capacity of the soil water storage (mm).*

The rest of available liquid fulfils the soil moisture tank. The soil water tank loses water through actual evapotranspiration, percolation and saturation excess runoff. The soil evapotranspiration (ES) is calculated as,

$$ES = SW * \varphi(2 - SW/K)/(1 + \varphi(1 - SW/K))$$

where, ES, *is the actual evapotranspiration (mm)*

φ, *is calculated as* $tanh((PETi\ - sublimationi - Etdi)/K)$

The percolation quantity to groundwater is defined as a fraction of the current soil water storage.

$$PERC = \lambda * SW$$

where, λ, is the recession rate for percolation.

After the actual evapotranspiration and percolation losses, if the remaining quantity exceeds the soil moisture capacity, overland flow is occurring.

$$QS = \max(0, SW - K)$$

The percolation quantity is added to the ground water tank that loses water through baseflow and deep percolation. Baseflow (QB) is defined as a fraction of the overflow if the current water quantity exceeds the ground water capacity.

$$QB = \max[0, m(Y - G)]$$

where, QB, is the baseflow (mm)

Y, *is the current groundwater storage (mm)*

G, *is the capacity of groundwater tank (mm)*

m, *is the recession rate parameter for baseflow*

The new water quantity of the groundwater storage contributes to the deep percolation process. Similarly to percolation, deep percolation is obtained as a fraction of the current groundwater storage.

$$DPERC = \alpha * SW$$

where, α, is the recession rate parameter for deep percolation.

The saturation runoff is combined with direct runoff, with a portion directed to the routing process and the remainder carried over to the next time step. The total runoff is computed as the sum of baseflow and a contribution from the routing runoff over the past four days. More precisely, the daily contribution comprises 3.52% from the runoff of the fourth day prior, 5.54% from the third day, 12.3% from the second day, and 87.11% from the current day.
For example, let's assume that at time step j, Q_0 represent the sum of saturation runoff, direct runoff and remaining runoff of the routing process in the previous timestep, $j-1$. The runoff goes for routing is $Q_0 * R$ and the $Q_0 * (1-R)$ will contribute to the Q_0 as the remainder runoff at the $j+1$ timestep.

4.4 The evaluation protocols

The model performances were evaluated using the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970), and the Kling–Gupta efficiency (Hoshin V. Gupta et al. 2009). The NSE and KGE metrics are widely used among modelers to evaluate rainfall-runoff model performance. These metrics provide valuable information on the accuracy and reliability of the models' discharge predictions and can be used to compare different models or determine the most suitable model for a specific application.

NSE evaluation metric ranges from minus infinity to 1.0 and 1.0 is the best agreement and mathematically is calculated as:

$$NSE = 1 - \frac{\sum_{i=1}^{N} (O_i - P_i)^2}{\sum_{i=1}^{N} (O_i - \bar{O}_i)^2}$$

where, Pi is the calculated flow,

Oi is the observed flow,

$\bar{O}_i$ is the mean observed flow and

N is the length of the timeseries.

KGE evaluates the hydrological model performance like NSE does and it was developed based on the limitation of NSE. Hoshin V. Gupta et al. 2009 decomposed NSE into three distinct components, the correlation, the bias and a measure of relative variability in the simulated and observed values. KGE is formulated by computing the Euclidian distances of the components

from the ideal point. KGE ranges from minus infinity to 1.0 and 1.0 indicates the best performance. KGE is described as:

$$KGE = 1 - \sqrt{\left(r - 1)^2 + \left(\frac{P}{O} - 1\right)^2 + (\frac{\bar{P}}{\bar{O}} - 1)^2\right)}$$

Where r is the Pearson correlation between simulation and observation runoff and $\bar{P}_i$ is the mean calculated flow.

Moreover bias, mean absolute percentage error (MAPE) and the root mean square error (RMSE) evaluate the performance of the 5% peak flows. The mathematical expression of those metrics is:

$$Bias = \frac{(\bar{P} - \bar{O})}{\bar{O}}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i - O_i|}{O_i}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - P_i)^2}$$

5 RESULTS

5.1 Input data and model training/calibration

The calibration classical procedure for models is to subdivide the data into three parts, referred to as training, validation and test data (Goodfellow, Bengio, and Courville 2016). The first two splits are used to derive the parametrization of the network and the remainder of the data to diagnose the actual performance of the model. Table 7, shows the periods of the data splitting used in this study.

***Table 7**. Time period for each sud-dataset.*

Dataset	Period	Percentage (%)
Training	01/10/1981 - 30/09/1996	45
Validation	01/10/1996 - 30/09/2001	15
Testing	01/10/2001 - end date	40

For the calibration of parameters in the Zygos model, the differential evolutionary optimization algorithm (Storn and Price, 1997) was employed, with the Nash-Sutcliffe Efficiency as the objective function. For the LSTM and Transformer models, the Adam (Kingma and Ba 2017) optimization algorithm, a commonly used approach in machine learning regression problems, was utilized alongside the Mean Squared Error as the loss function. Details regarding the number of parameters for each model are presented in Table 8.

To ensure consistent computational time for training the Zygos model, the following technique was employed. First, the maximum number of iterations and the population size were determined using a trial-and-error method, based on the time consumption and NSE values achieved for 10 basins. After this, the 164 Zygos models were trained. The models that achieved an NSE value below 0.1 were retrained following the same procedure.

***Table 8**.Number of parameters for LSTM, Transformer and Zygos models.*

Model	Number of Parameters
LSTM	5,381
Transformer	1,052,672
Zygos	11

LSTM and Transformer, as machine learning models, undergo training in epochs. An epoch signifies a complete pass through the entire dataset during the training phase. During this process, the model forwards the input data through the network to compute the error, followed by backward propagation to update the network's parameters. For instance, the LSTM model underwent 50 epochs using the training data, meaning the model iteratively adjusted its parameters based on the entire training dataset for 50 cycles. Subsequently, the validation dataset was utilized to assess the performance of these parameters. The best parameter set was then evaluated using the testing dataset. To mitigate computational costs and prevent overfitting, the early stopping technique was employed. This technique halts the training process if the model's performance on the validation dataset does not improve for a specified consecutive epoch.

The input features for each model are detailed in Table 9. The transformer model requires absolute positional information to be explicitly conveyed to the model. Consequently, the input feature structure slightly differs from that of the LSTM model. It includes a fixed time series incorporating the month of the year, as the month significantly influences runoff variance.

***Table 9**. Input data for LSTM, Transformer and Zygos models.*

Model	input Feature
LSTM	Average daily precipitation
	Surface-incident solar radiation
	2 m daily maximum air temperature
	2 m daily minimum air temperature
	Near-surface daily average vapor pressure
Transformer	Average daily precipitation
	Surface-incident solar radiation
	2 m daily maximum air temperature
	2 m daily minimum air temperature
	Near-surface daily average vapor pressure
	Month of the year
Zygos	Average daily precipitation
	2 m daily mean air temperature
	Potential evapotranspiration*

**Potential evapotranspiration is computed using the Hargreaves method.*

For efficient learning in machine learning models, all input and output features are normalized and serialized. The goal of normalization is to transform features to be on a similar scale. This improves the performance and training stability of the model (Bishop 2006). The z-score technique was employed, involving the normalization of data by subtracting the mean and dividing by the standard deviation. Serialization is essential for preparing the input data in a suitable format for learning purpose (Figure 11).

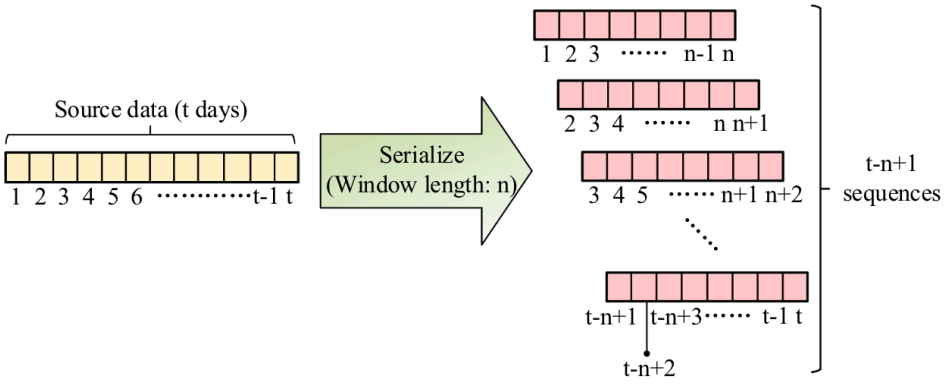


***Figure 11**. An illustration of data serialization (Yin, Guo, et al. 2022).*

5.2 Results

The model performances were evaluated using the Nash-Sutcliffe efficiency (NSE), and the Kling–Gupta efficiency (KGE). Moreover bias, mean absolute percentage error and the root mean square error evaluate the performance of the 5% peak flows (Chapter 4.4).

Because LSTM modeling approach needs 365 days of meteorological data as input for predicting one time step, while Transformer needs 20 the evaluation period is shifted by one year. Moreover, the Zygos model does not use the validation dataset because of the different calibration-training approach.

The NSE and KGE results for 164 LSTM, Transformer, and Zygos rainfall-runoff models are shown in Figures 12-14. Both the LSTM and Transformer models demonstrate superior performance compared to the benchmark Zygos models. Specifically, the mean NSE values on the testing dataset are 0.50 for LSTM, 0.71 for Transformer, and 0.37 for Zygos, underscoring the efficacy of the LSTM and Transformer architectures in runoff predicting.

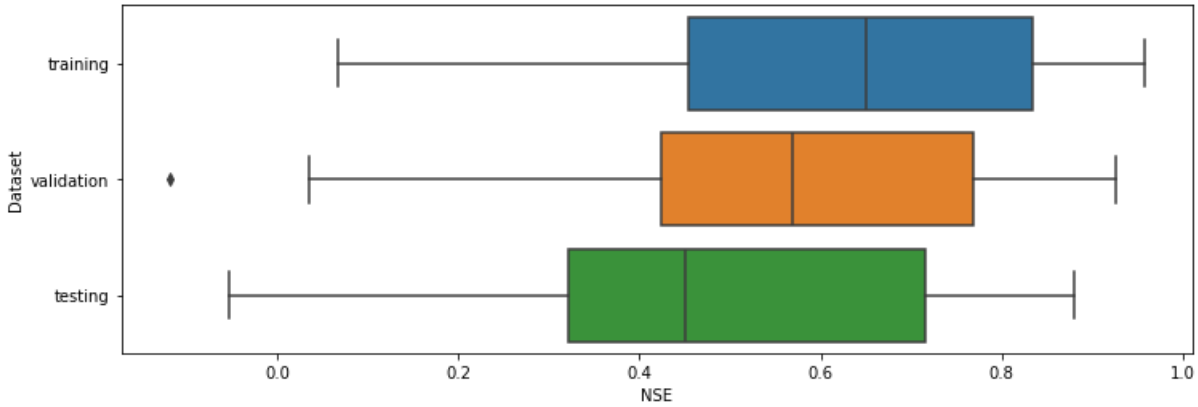


Figure 12. *Boxplot of the NSE for LSTM models.*

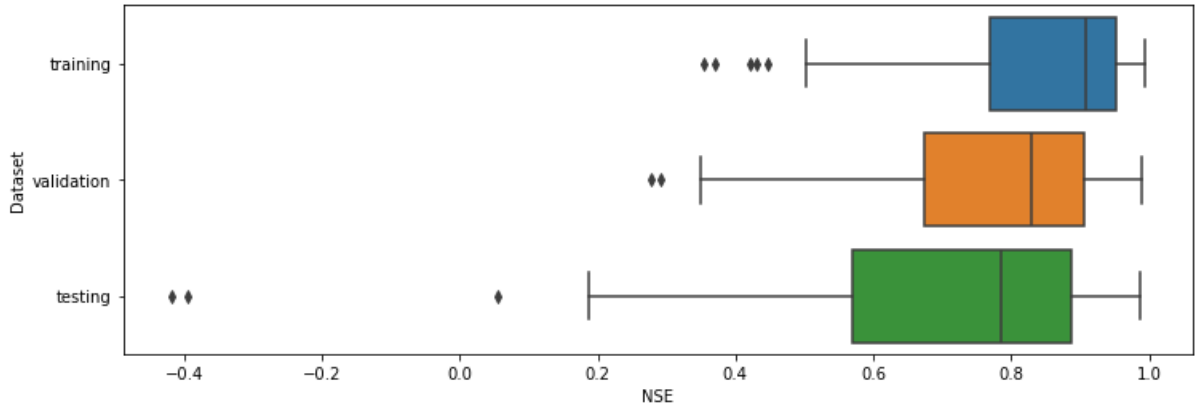


Figure 13. . *Boxplot of the NSE for Transformer models.*

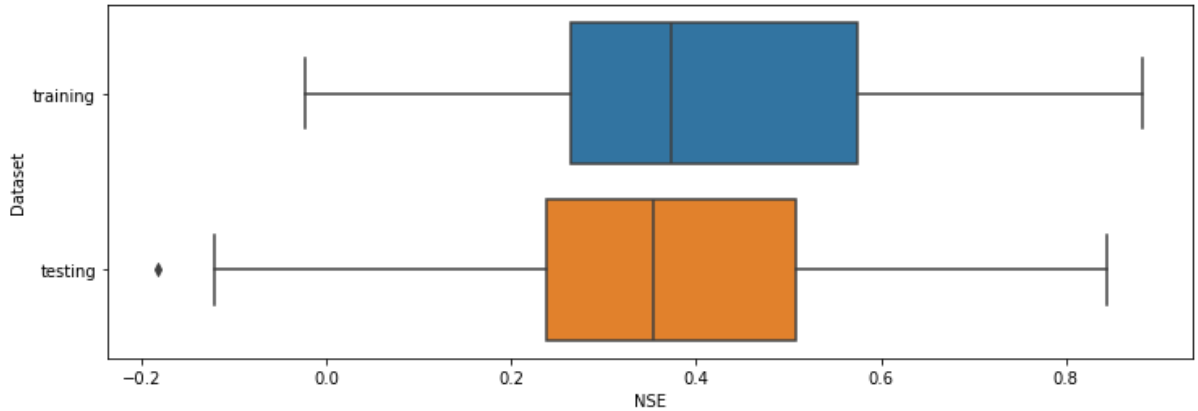


Figure 14. . *Boxplot of the NSE for Zygos model.*

The KGE follows the logic of the NSE as far as models' performance is concerned. Both of the machine learning models outperform the benchmark model with a mean KGE values 0.23, 0.69 and 0.07 for the LSTM, the Transformer and the Zygos models, respectively (Figures 15-17).

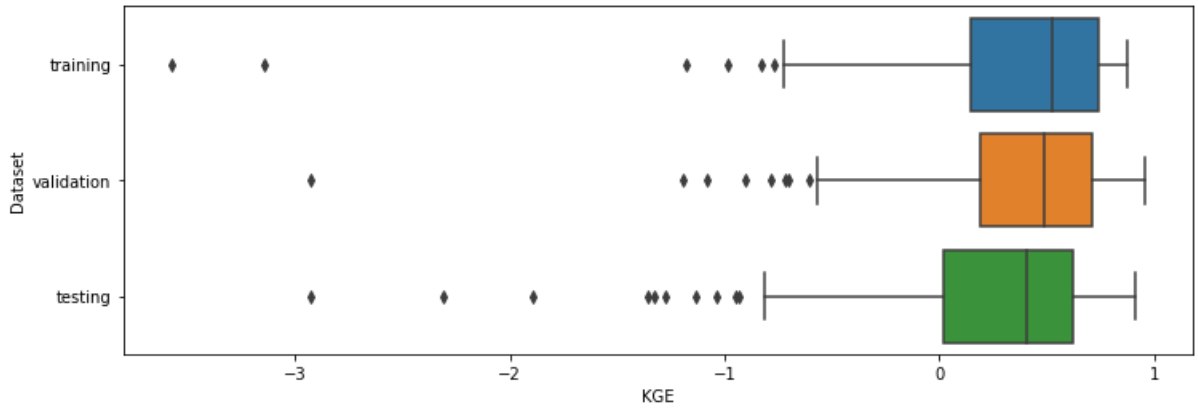


***Figure 15**. Boxplot of the KGE for LSTM models.*

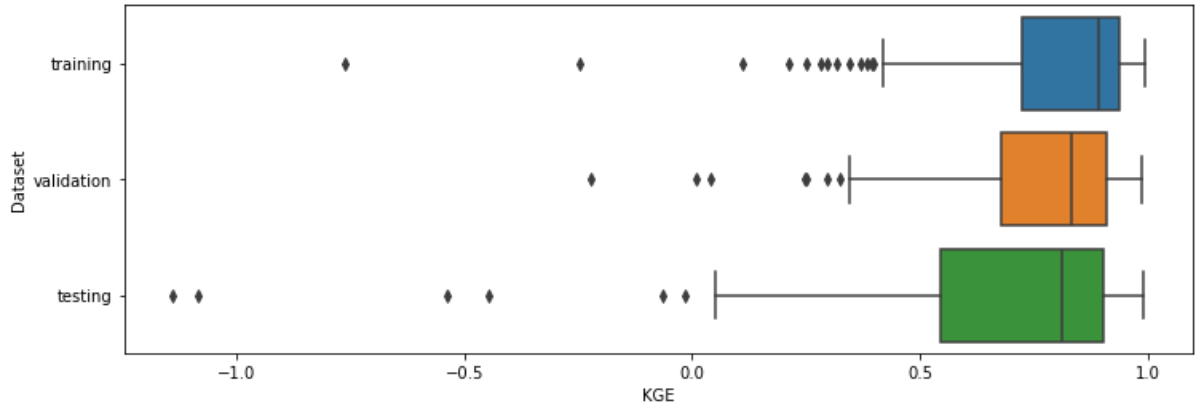


***Figure 16**. Boxplot of the KGE for Transformer models.*

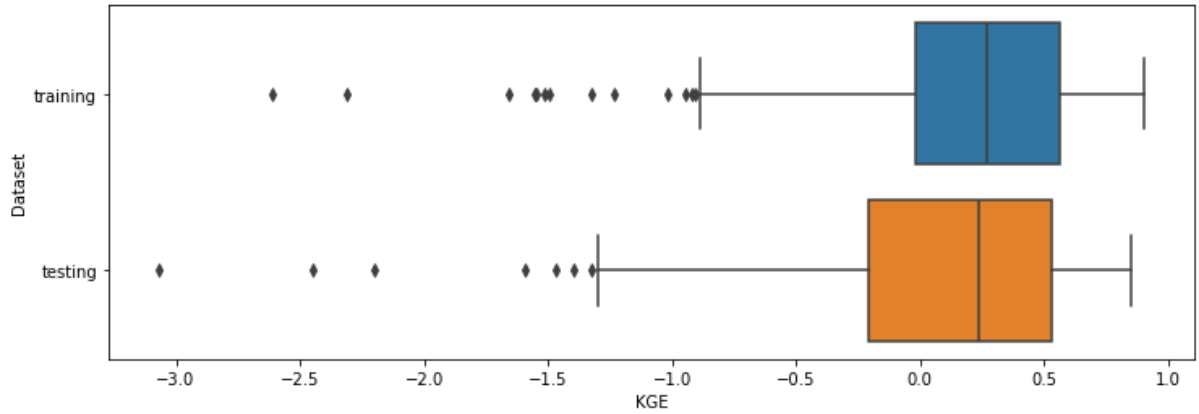


***Figure 17**. Boxplot of the KGE for Zygos models.*

***Table 10**. NSE and KGE evaluation metrics for LSTM, Transformer and Zygos models*

Model	Dataset	Max NSE	Max KGE	Mean NSE	Mean KGE	Median NSE	Median KGE	Min NSE	Min KGE
LSTM	training	0.957	0.875	0.635	0.364	0.650	0.524	0.067	-3.574
	validation	0.927	0.955	0.571	0.379	0.571	0.492	-0.119	-2.928
	testing	0.880	0.910	0.501	0.234	0.449	0.403	**-0.053**	-2.924
Transformer	training	0.993	0.994	0.843	0.795	0.908	0.892	0.353	-0.761
	validation	0.989	0.987	0.777	0.761	0.829	0.832	0.278	-0.223
	testing	**0.987**	**0.990**	**0.712**	**0.690**	**0.786**	**0.813**	-0.418	**-1.141**
Zygos	training	0.884	0.905	0.415	0.122	0.373	0.269	-0.024	-2.612
	testing	0.844	0.851	0.374	0.075	0.354	0.234	-0.184	-3.071

Transformer models exhibit consistent performance across all datasets, showcasing high values for both NSE and KGE. However, in only four basins, the NSE values for Transformer models are inferior to the benchmark Zygos model, while the corresponding number of basins for KGE evaluation metrics is five. Despite LSTM models performing better than the benchmark models, they do not surpass Transformer models. Additionally, 36 Zygos models exhibit superior NSE values compared to LSTM models, while 66 models demonstrate better KGE values. The comparison of the NSE and KGE metrics among the models are visualized in Figure 18.

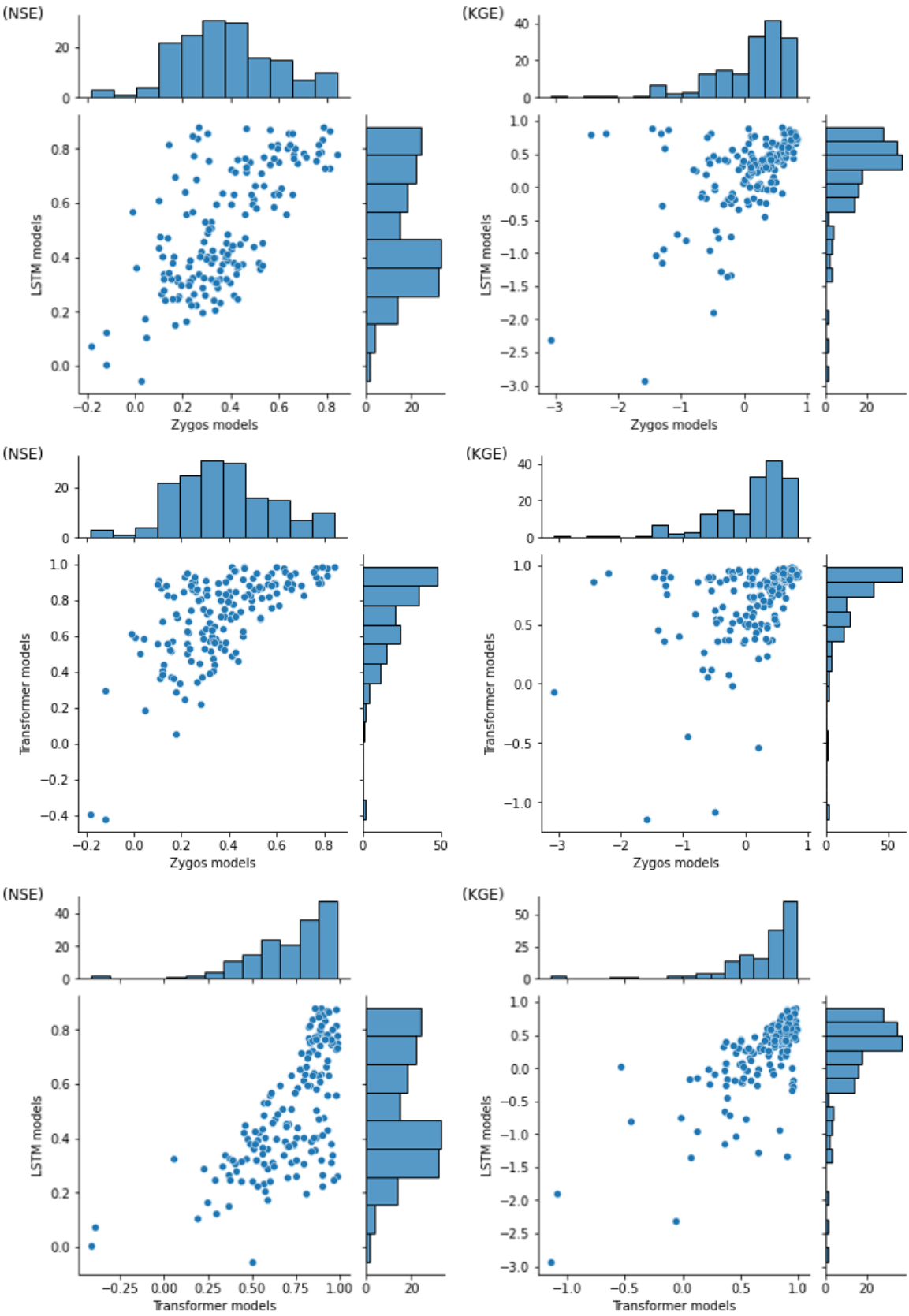


Figure 18. *NSE & KGE comparison for LSTM, Transformer and Zygos models.*

To analyze the behavior of each model in more detail, a number of evaluation metrics were measured. Specifically, the evaluation metrics Root Mean Squared Error (RMSE), the coefficient of determination R^2, Mean Absolute Error, Max Absolute Error were used to further evaluate the models. As illustrated in Figure 19, the Transformer models outperform in all these metrics, while the performance of the Zygos model is inferior to the others.

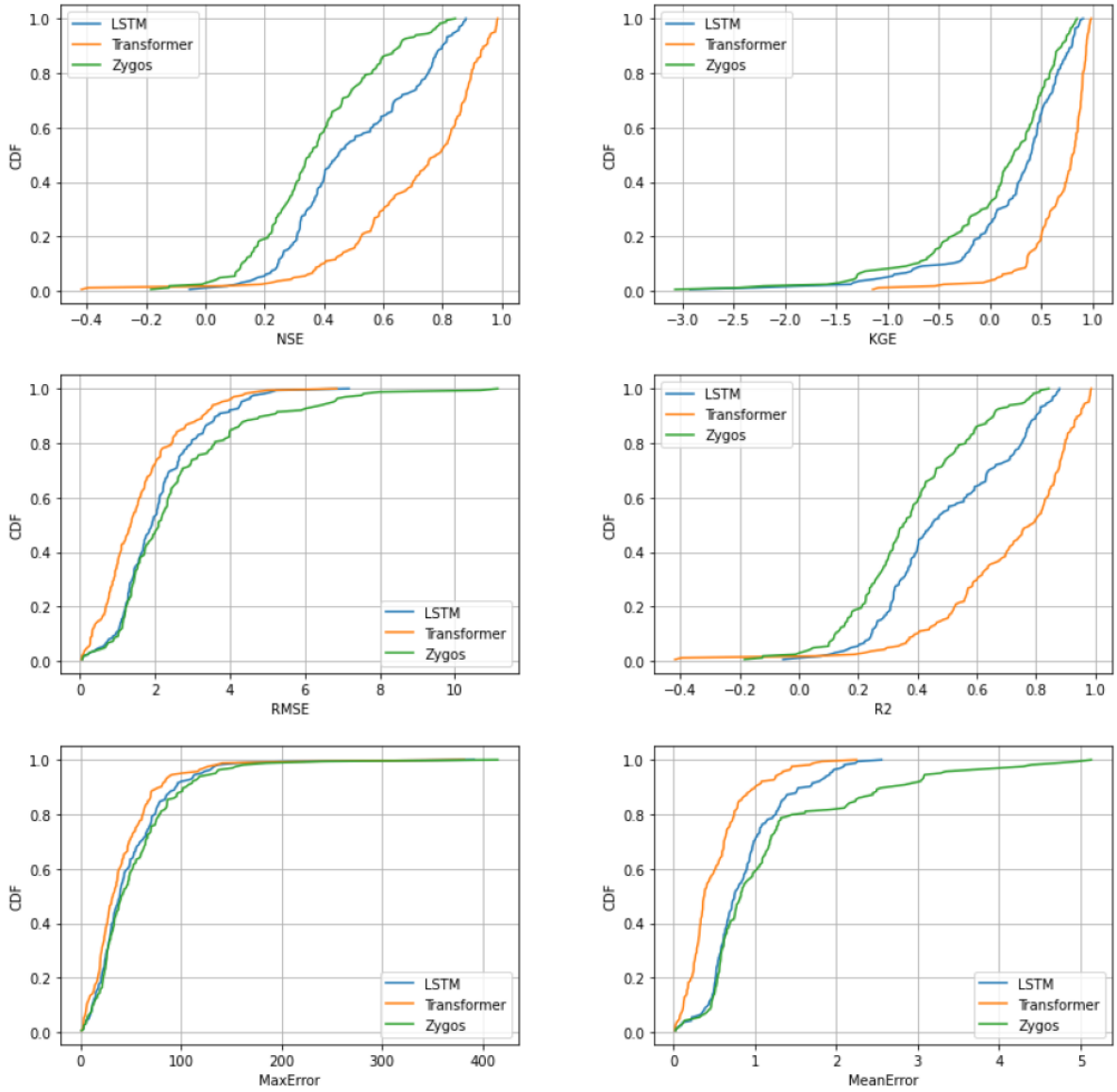


***Figure 19**. Cumulative density functions for various metrics of the testing period.*

Figures 20 and 21 depict the NSE and KGE values, respectively, for all models across the operational basins within the four hydrological units in the United States of America. The markers on the figures represent the centroids of the basins along with their corresponding area sizes.

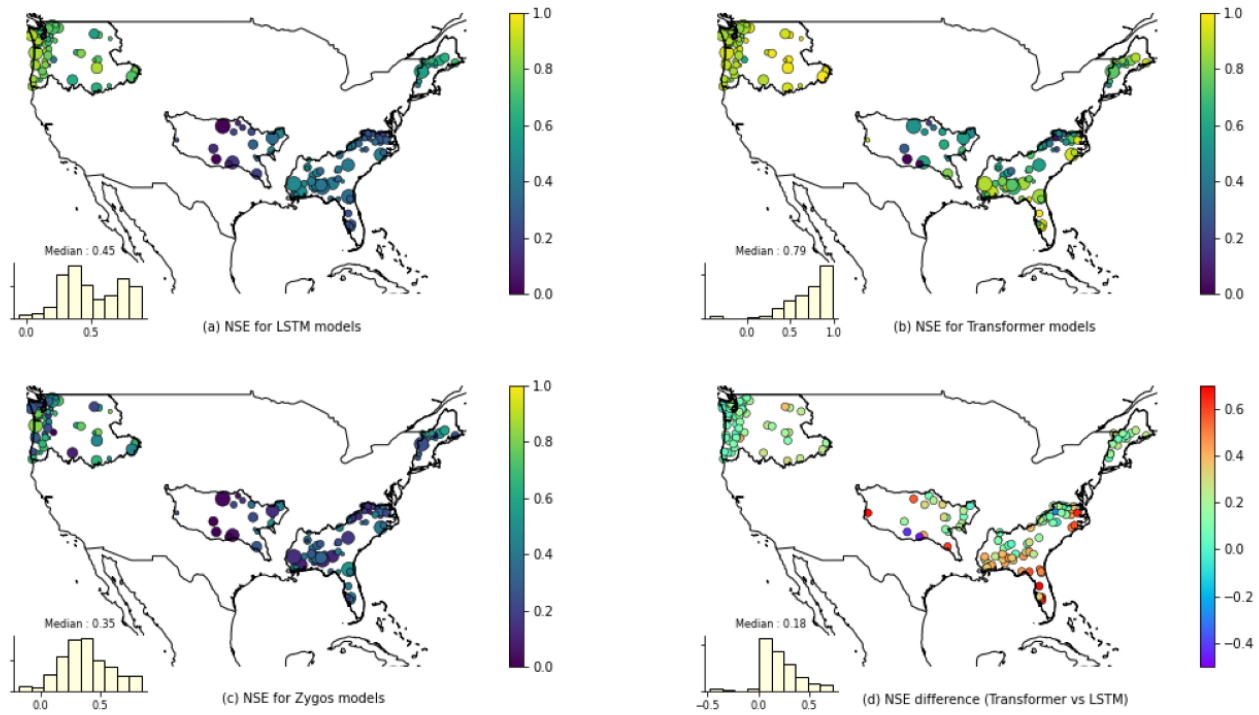


Figure 20. *(a)-(b)-(c) NSE evaluation metrics of the testing period for LSTM, Transformer and LSTM model. (d) Difference of the NSE between Transformer and LSTM models (red color >0 indicates that the Transformer performs better).*

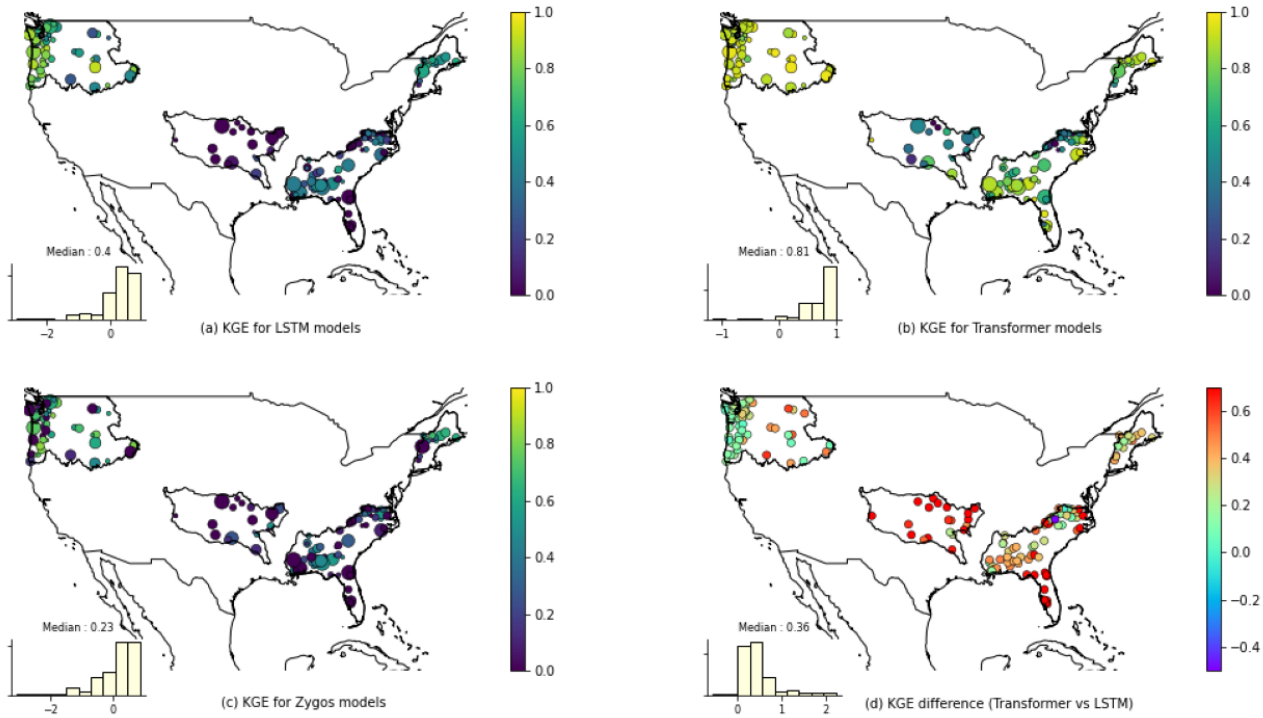


Figure 21. *(a)-(b)-(c) KGE evaluation metrics of the testing period for LSTM, Transformer and LSTM model. (d) Difference of the KGE between Transformer and LSTM models (red color >0 indicates that the Transformer performs better).*

To analyze how machine learning models capture dependencies on different hydrological conditions, the mean and median NSE values were computed separately for each of the four hydrological units. Table11 shows that the Transformer model consistently outperforms both Zygos and LSTM models across all four hydrological units in terms of mean and median NSE values. Additionally, there's a clear variation in model performance across different hydrological units, with some units exhibiting higher mean NSE and KGE values compared to others.

***Table 11**. Evaluation metrics NSE and KGE for the four hydrological units.*

HUC	Model	Mean NSE	Mean KGE
New England	Zygos	0.381	0.348
	LSTM	0.579	0.477
	Transformer	0.751	0.817
South Atlantic-Gulf	Zygos	0.297	0.011
	LSTM	0.358	0.061
	Transformer	0.647	0.610
Arkansas-White-Red	Zygos	0.199	-0.499
	LSTM	0.226	-0.617
	Transformer	0.418	0.300
Pacific Northwest	Zygos	0.526	0.274
	LSTM	0.746	0.668
	Transformer	0.880	0.886

The comparison across hydrological units reveals that static attributes such as precipitation, aridity, snow fraction, seasonality, and altitude significantly influence the performance of hydrological models. The Transformer model consistently outperforms the LSTM and Zygos models across all regions, particularly excelling in areas with high precipitation and seasonality such as the Pacific Northwest. Conversely, regions like the Arkansas-White-Red with lower precipitation and high-altitude variability present greater challenges for accurate modeling.

The best and worst benchmark Zygos models for each hydrological unit were chosen for an illustrative comparison for the hydrological year 2009-2010. As seen in Figure 22, all three models perform well, with the Transformer model being the best. On the other hand, in Figure 23, although the Zygos models are not able to accurately represent the system, the LSTM and Transformer models perform fairly well. For the catchment 07315200, none of the models can achieve good performance.

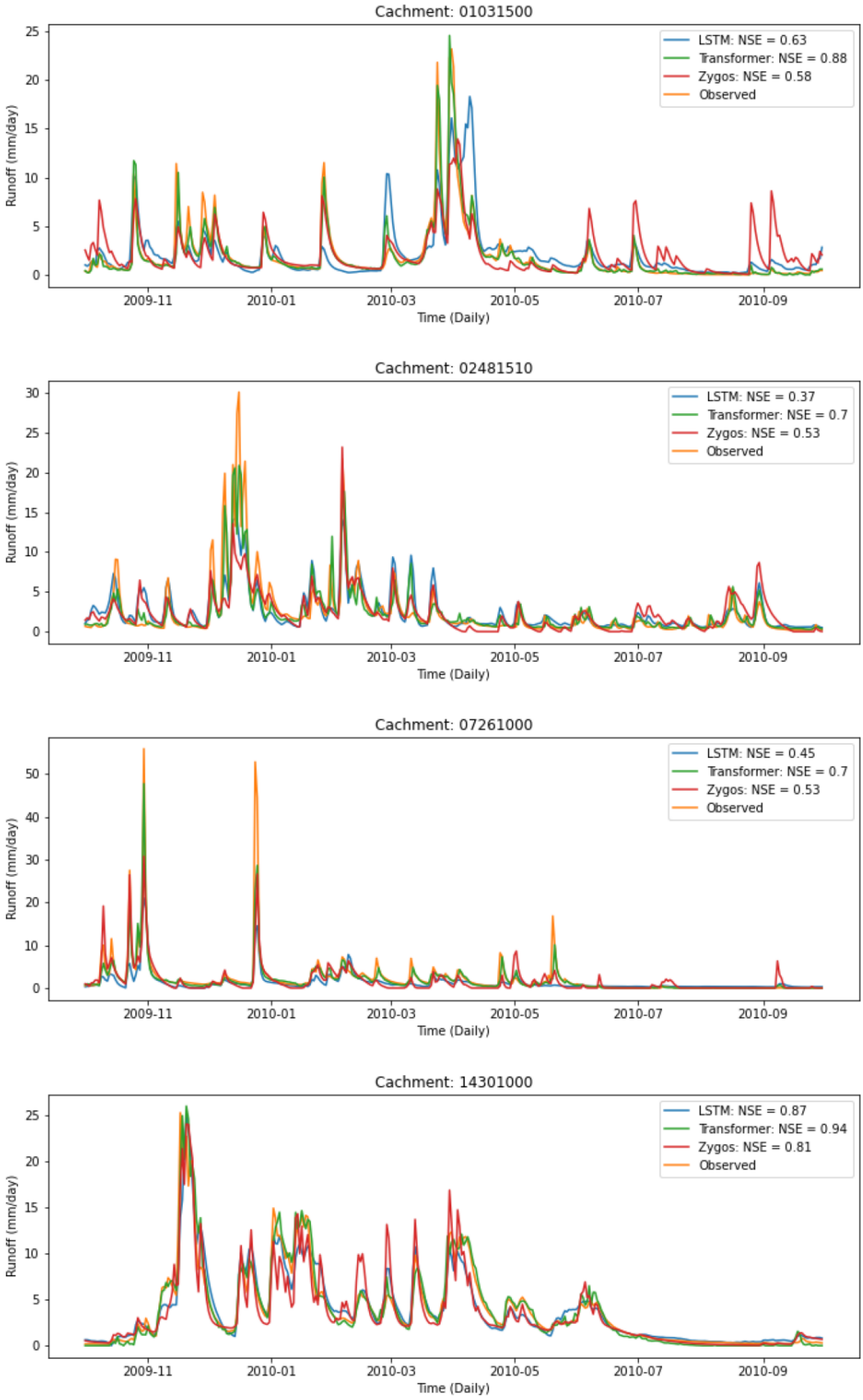


Figure 22. *An illustration comparing of LSTM, Transformer and Zygos models for basins where Zygos achieve the best performance.*

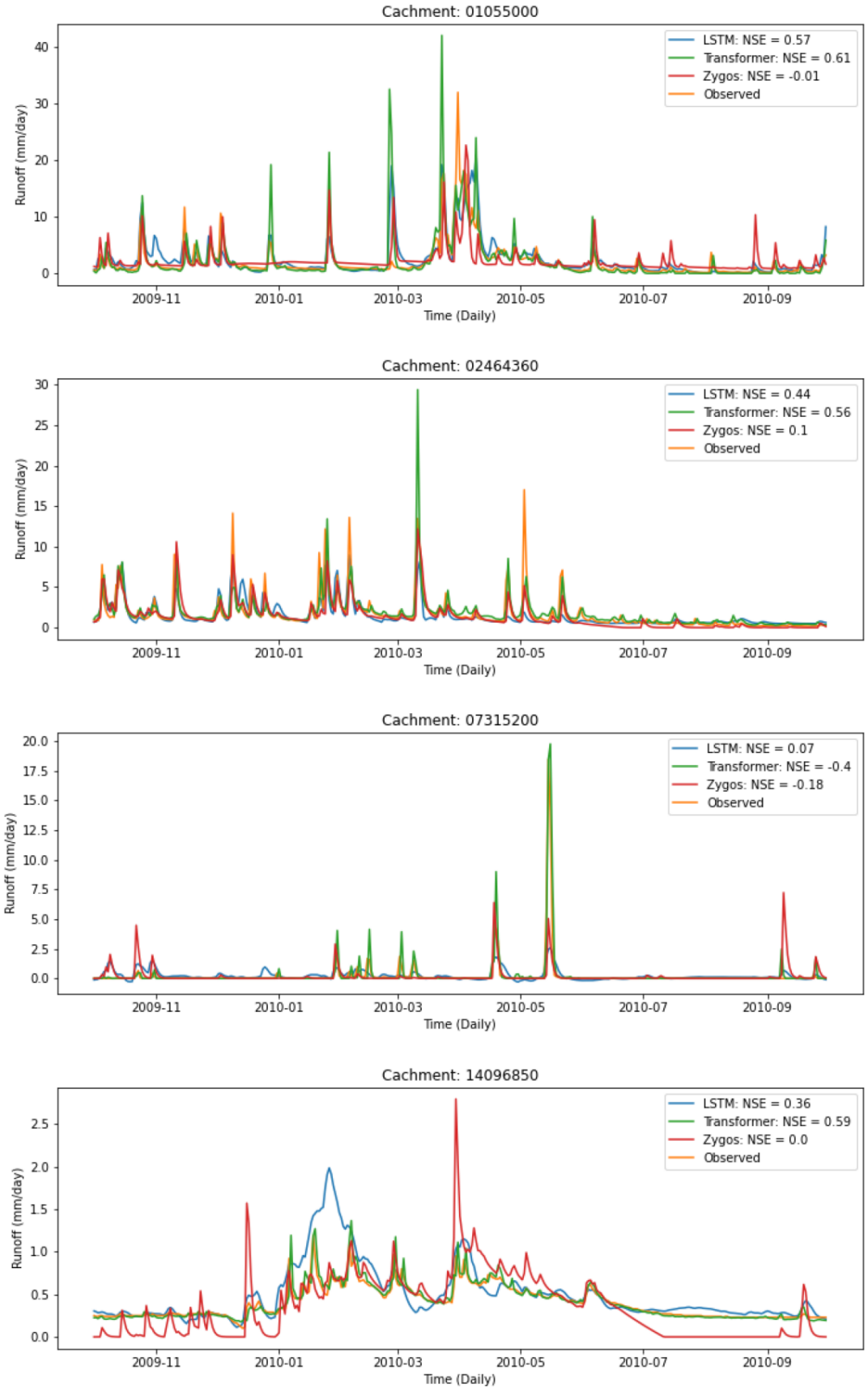


Figure 23. *An illustration comparing of LSTM, Transformer and Zygos models for basins where Zygos achieve the worst performance.*

Furthermore, to explore the performance of the models during peak flows, the bias (BIAS), mean absolute percentage error (MAPE), and root mean square error (RMSE) criteria were calculated for the top 5% peak flows. The following figures indicate that the Transformer model achieves better performance on predicting peak flows compared to the LSTM and Zygos models.

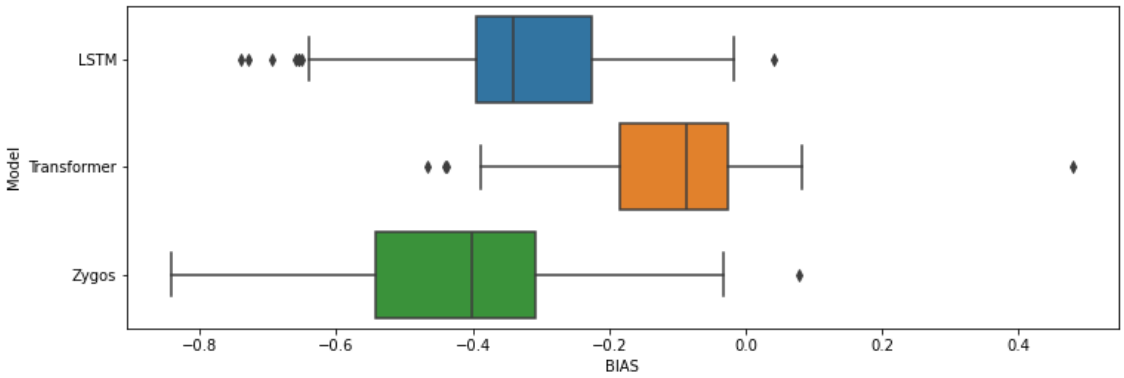


Figure 24. *Boxplot of the BIAS of the testing period for the 5% peak flows.*

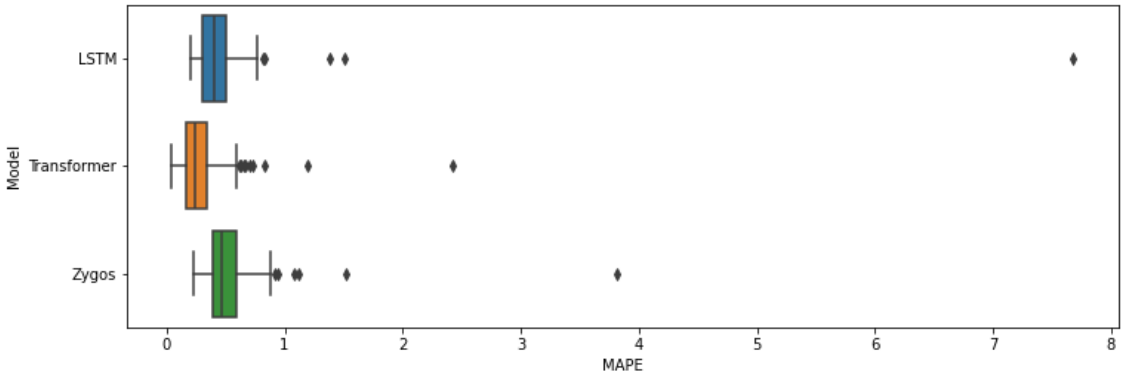


Figure 25. *Boxplot of the mean absolute percentage error for the 5% peak flows.*

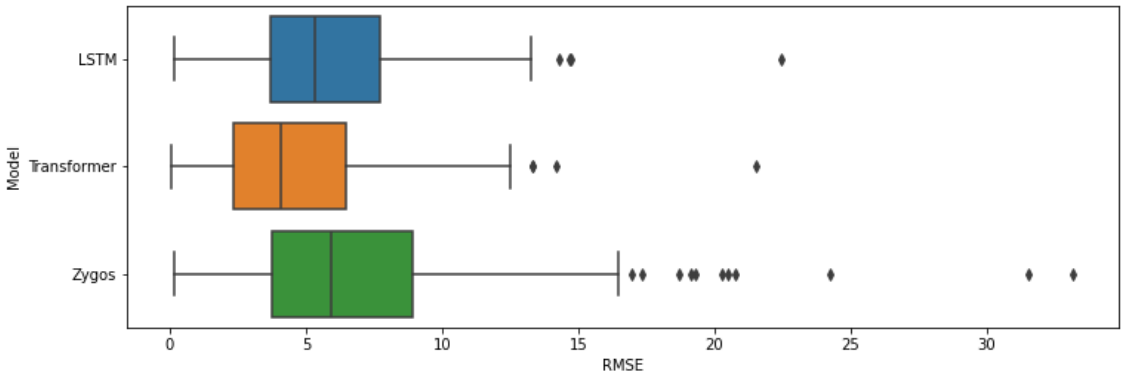


***Figure 26**. Boxplot of the root mean square error for the 5% peak flows.*

Table 12 shows that all the models are negatively biased, which means that they underestimate the peak flows. Transformer models perform the best for all metrics with the smallest mean bias -0.11, mean absolute percentage error 0.29 and rout mean square error 4.75. Zygos models demonstrate the highest errors and bias, indicating it is the least accurate among the three models for peak flow prediction.

***Table 12**. Mean and Median BIAS, MAPE and RMSE values for the 5% peak flows.*

Model	Mean BIAS	Median BIAS	Mean MAPE	Median MAPE	Mean RMSE	Median RMSE
LSTM	-0.33	-0.34	0.47	0.40	6.02	5.31
Transformer	**-0.11**	**-0.09**	**0.29**	**0.24**	**4.75**	**4.08**
Zygos	-0.42	-0.40	0.52	0.46	7.22	5.91

6 CONCLUSIONS

6.1 Thesis conclusions

The primary objective of this study is to investigate the potential of two state-of-the-art machine learning models, Long Short-Term Memory (LSTM) and Transformer, in rainfall-runoff modeling. The performance of each model is compared with the Zygos conceptual rainfall-runoff model across 164 basins. Specifically, the differences in the Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) evaluation metrics between the Transformer and Zygos models are 0.338 and 0.615, respectively. In comparison, the differences for LSTM and Zygos are 0.127 and 0.159, respectively. This indicates that Transformer models outperform LSTM models and the Zygos model in rainfall-runoff modeling. Furthermore, evaluation metrics for peak flow predictions, such as bias, mean absolute percentage error (MAPE), and root mean squared error (RMSE show that both of machine learning models achieve better performance in predicting peak flows than the conceptual one.

The variation in evaluation metrics across the four hydrological units highlights that machine learning models can effectively capture a range of dependencies under diverse hydrological conditions. For instance, in the Pacific Northwest hydrological unit, which includes 58 basins, the LSTM and Transformer models achieve mean NSE values of 0.75 and 0.88, respectively. This high performance suggests that both machine learning models can "learn" complex hydrological relationships of a region with high precipitation, significant portion of which falls as snow, along with high aridity and seasonality. The varied performance also highlights the importance of considering regional characteristics when selecting and applying hydrological models for predictive purposes.

Despite the Transformer models outperforming the others, there is a significant drawback in their architecture. This drawback is that the decoder component functions in a non-autoregressive manner, meaning that the ground truth values are used as inputs for the decoder block, and errors do not accumulate as the output moves to the next timesteps. This indicates that machine learning models lack of interpretability and therefore cannot be implemented without considerations about their reliability.

It is worth mentioning that the architecture of the Transformer model is far more complex than that of the LSTM, as it utilizes a more sophisticated learning method. However, this complexity comes at the cost of increased training time and memory usage. Transformers utilize self-

attention mechanisms that scale quadratically with input length, making them computationally intensive. To ensure a fair comparison among the models, the computational time for training each model was kept within comparable limits. The results indicate that, even with the same computational cost, both of the machine learning models can achieve superior performance compared to Zygos traditional models. This underscores the potential of advanced machine learning techniques to enhance hydrological predictions, provided that sufficient computational resources are available.

In summary the main conclusions of this study are as follows:

1. Machine learning models can be used as rainfall-runoff models.
2. Although machine learning models can achieve sufficient performance and outperforms the traditional conceptual models it is still not clear whether conceptual models can fully be replaced by machine learning models.
3. Machine learning models are able to predict the runoff in basins with diverse hydrological conditions meaning that they can "learn" dependencies associated with various hydrological processes.
4. Computational resources in terms of time are essential for training both to classical conceptual models and to machine learning models.

6.2 Future research

The Transformer architecture developed in this study has demonstrated superior performance compared to LSTM and Zygos models. However, it should not be considered the definitive superior model. Instead, it should be viewed as a highly promising architecture with significant potential. Zhou et al. 2021 mention that vanilla Transformer architecture cannot be directly applicable for long sequence timeseries forecasting due to its quadratic time complexity and high memory usage. This comes in contradiction to the Transformer model implemented in this study. Therefore, while the Transformer model shows great promise as a rainfall-runoff machine learning model, much research is still needed to refine its architecture.

Moreover, both LSTM and Transformer hyperparameters were pre-defined, and no hyperparameter tuning was done. Hence, a systematic sensitivity analysis of the effects of different hyperparameters would boost the performance and reliability of these models.

Although identifying machine learning models as black-box models is considered a myth by Maier et al. 2023 because the mathematical relationship between the model's input and output is known, it is still not feasible to fully analyze the behavior of such complex models. This suggests that a systematic interpretation of the network's internals would enhance our understanding on those models leading to error diagnosis and model improvement. Hence, comprehensive research utilizing explainable AI techniques to dress model outcomes with interpretability, expandability and transparency is a well promising field of study.

Finally, LSTM and Transformer models have only been applied on limited large-scale datasets and the majority of them in the CAMELS dataset. To expand our understanding of the uncertainty associated with input data, further research should focus on testing rainfall-runoff machine learning models utilizing other datasets.

7 REFERENCES

Addor, Nans, Hong X. Do, Camila Alvarez-Garreton, Gemma Coxon, Keirnan Fowler, and Pablo A. Mendoza. 2020. “Large-Sample Hydrology: Recent Progress, Guidelines for New Datasets and Grand Challenges.” *Hydrological Sciences Journal* 65 (5): 712–25. https://doi.org/10.1080/02626667.2019.1683182.

Addor, Nans, Andrew J Newman, Naoki Mizukami, and Martyn P Clark. 2017. “The CAMELS Data Set: Catchment Attributes and Meteorology for Large-Sample Studies.” *Hydrol. Earth Syst. Sci.*

Amanambu, Amobichukwu C., Joann Mossa, and Yin-Hsuen Chen. 2022. “Hydrological Drought Forecasting Using a Deep Transformer Model.” *Water* 14 (22): 3611. https://doi.org/10.3390/w14223611.

Arnold, J. G., R. Srinivasan, R. S. Muttiah, and J. R. Williams. 1998. “Large Area Hydrologic Modeling and Assessment Part I: Model Development1.” *JAWRA Journal of the American Water Resources Association* 34 (1): 73–89. https://doi.org/10.1111/j.1752-1688.1998.tb05961.x.

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology. 2000. “Artificial Neural Networks in Hydrology. II: Hydrologic Applications.” *Journal of Hydrologic Engineering* 5 (2): 124–37. https://doi.org/10.1061/(ASCE)1084-0699(2000)5:2(124).

Asefa, Tirusew, Mariush Kemblowski, Mac McKee, and Abedalrazq Khalil. 2006. “Multi-Time Scale Stream Flow Predictions: The Support Vector Machines Approach.” *Journal of Hydrology* 318 (1–4): 7–16. https://doi.org/10.1016/j.jhydrol.2005.06.001.

Bachmair, Sophie, Cecilia Svensson, Ilaria Prosdocimi, Jamie Hannaford, and Kerstin Stahl. 2017. “Developing Drought Impact Functions for Drought Risk Management.” *Natural Hazards and Earth System Sciences* 17 (11): 1947–60. https://doi.org/10.5194/nhess-17-1947-2017.

Beven, K. J. 2012. *Rainfall-Runoff Modelling: The Primer*. 2nd ed. Chichester, West Sussex ; Hoboken, NJ: Wiley-Blackwell.

Beven, K. J., and M. J. Kirkby. 1979. “A Physically Based, Variable Contributing Area Model of Basin Hydrology / Un Modèle à Base Physique de Zone d’appel Variable de l’hydrologie Du Bassin Versant.” *Hydrological Sciences Bulletin* 24 (1): 43–69. https://doi.org/10.1080/02626667909491834.

Bishop, Christopher M. 2006. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer.

Bray, Michaela, and Dawei Han. 2004. “Identification of Support Vector Machines for Runoff Modelling.” *Journal of Hydroinformatics* 6 (October):265–80. https://doi.org/10.2166/hydro.2004.0020.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (1): 5–32. https://doi.org/10.1023/A:1010933404324.

Burnash, Robert J. C. 1973. *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers*. U. S. Department of Commerce, National Weather Service, and State of California, Department of Water Resources.

Ch, Sudheer, Nitin Anand, B.K. Panigrahi, and Shashi Mathur. 2013. “Streamflow Forecasting by SVM with Quantum Behaved Particle Swarm Optimization.” *Neurocomputing* 101 (February):18–23. https://doi.org/10.1016/j.neucom.2012.07.017.

Chang, Won, and Xi Chen. 2018. “Monthly Rainfall-Runoff Modeling at Watershed Scale: A Comparative Study of Data-Driven and Theory-Driven Approaches.” *Water* 10 (9): 1116. https://doi.org/10.3390/w10091116.

Chen, S M, Y M Wang, and I Tsou. 2013. “Using Artificial Neural Network Approach for Modelling Rainfall–Runoff Due to Typhoon.” *Journal of Earth System Science* 122 (2): 399–405. https://doi.org/10.1007/s12040-013-0289-8.

Clerc-Schwarzenbach, Franziska Maria, Giovanni Selleri, Mattia Neri, Elena Toth, Ilja Van Meerveld, and Jan Seibert. 2024. “HESS Opinions: A Few Camels or a Whole Caravan?” https://doi.org/10.5194/egusphere-2024-864.

Cortes, Corinna, and Vladimir Vapnik. 1995. “Support-Vector Networks.” *Machine Learning* 20 (3): 273–97. https://doi.org/10.1007/BF00994018.

Dibike, Yonas, Slavco Velickov, Dimitri Solomatine, and Michael Abbott. 2001. “Model Induction With Support Vector Machines: Introduction and Applications.” *Journal of Computing in Civil Engineering - J COMPUT CIVIL ENG* 15 (July). https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208).

Duan, Q. Y., V. K. Gupta, and S. Sorooshian. 1993. “Shuffled Complex Evolution Approach for Effective and Efficient Global Minimization.” *Journal of Optimization Theory and Applications* 76 (3): 501–21. https://doi.org/10.1007/BF00939380.

Efstratiadis, A. 2008. “Μη γραμμικές μέθοδοι σε πολυκριτηριακά προβλήματα βελτιστοποίησης υδατικών πόρων, με έμφαση στη βαθμονόμηση υδρολογικών μοντέλων.” Εθνικό Μετσόβιο Πολυτεχνείο (ΕΜΠ), Σχολή Πολιτικών Μηχανικών, Τομέας Υδατικών Πόρων και Περιβάλλοντος. https://doi.org/10.12681/eadd/16954.

Efstratiadis, A., I. Nalbantis, and D. Koutsoyiannis. 2015. “Hydrological Modelling of Temporally-Varying Catchments: Facets of Change and the Value of Information.” *Hydrological Sciences Journal* 60 (7–8): 1438–61. https://doi.org/10.1080/02626667.2014.982123.

Efstratiadis, Andreas, and Demetris Koutsoyiannis. 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.

Erdal, Halil Ibrahim, and Onur Karakurt. 2013. “Advancing Monthly Streamflow Prediction Accuracy of CART Models Using Ensemble Learning Paradigms.” *Journal of Hydrology* 477 (January):119–28. https://doi.org/10.1016/j.jhydrol.2012.11.015.

Flamig, Zachary L., Humberto Vergara, and Jonathan J. Gourley. 2020. “The Ensemble Framework For Flash Flood Forecasting (EF5) v1.2: Description and Case Study.” *Geoscientific Model Development* 13 (10): 4943–58. https://doi.org/10.5194/gmd-13-4943-2020.

Frame, Jonathan, Frederik Kratzert, Daniel Klotz, Martin Gauch, Guy Shelev, Oren Gilon, Logan M. Qualls, Hoshin V. Gupta, and Grey S. Nearing. 2021. “Deep Learning Rainfall-Runoff Predictions of Extreme Events.” https://doi.org/10.5194/hess-2021-423.

French, Mark, Witold Krajewski, and Robert Cuykendall. 1992. “Rainfall Forecasting in Space and Time Using a Neural Network.” *Journal of Hydrology* 137 (August):1–31. https://doi.org/10.1016/0022-1694(92)90046-X.

Galelli, S., and A. Castelletti. 2013. “Assessing the Predictive Capability of Randomized Tree-Based Ensembles in Streamflow Modelling.” *Hydrology and Earth System Sciences* 17 (7): 2669–84. https://doi.org/10.5194/hess-17-2669-2013.

Gehlert, Andreas, and Daniel Pfeiffer. 2005. “A Framework for Comparing Conceptual Models.” In , 108–22.

Goldberg, David E., and John H. Holland. 1988. “Genetic Algorithms and Machine Learning.” *Machine Learning* 3 (2): 95–99. https://doi.org/10.1023/A:1022602019183.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press.

Goyal, Manish Kumar, and Chandra Shekhar Prasad Ojha. 2010. “Analysis of Mean Monthly Rainfall Runoff Data of Indian Catchments Using Dimensionless Variables by Neural Network.” *Journal of Environmental Protection* 01 (02): 155–71. https://doi.org/10.4236/jep.2010.12020.

Gudmundsson, Lukas, and Sonia I. Seneviratne. 2016. “Observation-Based Gridded Runoff Estimates for Europe (E-RUN Version 1.1).” *Earth System Science Data* 8 (2): 279–95. https://doi.org/10.5194/essd-8-279-2016.

Gupta, H. V., C. Perrin, G. Blöschl, A. Montanari, R. Kumar, M. Clark, and V. Andréassian. 2014. “Large-Sample Hydrology: A Need to Balance Depth with Breadth.” *Hydrology and Earth System Sciences* 18 (2): 463–77. https://doi.org/10.5194/hess-18-463-2014.

Gupta, Hoshin V., Harald Kling, Koray K. Yilmaz, and Guillermo F. Martinez. 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.

Gupta, Hoshin Vijai, and Harald Kling. 2011. “On Typical Range, Sensitivity, and Normalization of Mean Squared Error and Nash-Sutcliffe Efficiency Type Metrics.” *Water Resources Research* 47 (10). https://doi.org/10.1029/2011WR010962.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. “Long Short-Term Memory.” *Neural Computation* 9 (8): 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.

Iorgulescu, I., and K. J. Beven. 2004. “Nonparametric Direct Mapping of Rainfall-runoff Relationships: An Alternative Approach to Data Analysis and Modeling?” *Water Resources Research* 40 (8): 2004WR003094. https://doi.org/10.1029/2004WR003094.

Kalteh, A M. 2008. “Rainfall-Runoff Modelling Using Artificial Neural Networks (ANNs): Modelling and Understanding.” *Caspian Journal of Environmental Sciences* 6:53–58.

Khu, Soon Thiam, and Henrik Madsen. 2005. “Multiobjective Calibration with Pareto Preference Ordering: An Application to Rainfall-runoff Model Calibration.” *Water Resources Research* 41 (3): 2004WR003041. https://doi.org/10.1029/2004WR003041.

Kingma, Diederik P., and Jimmy Ba. 2017. “Adam: A Method for Stochastic Optimization.” arXiv. http://arxiv.org/abs/1412.6980.

Kitaev, Nikita, Łukasz Kaiser, and Anselm Levskaya. 2020. “Reformer: The Efficient Transformer.” arXiv. http://arxiv.org/abs/2001.04451.

Kozanis, Stefanos, and Andreas Efstratiadis. 2006. “‘Zygos’: A Basin Processes Simulation Model.” *21st European Conference for ESRI*.

Kratzert, Frederik, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. 2018. “Rainfall–Runoff Modelling Using Long Short-Term Memory (LSTM) Networks.” *Hydrology and Earth System Sciences* 22 (11): 6005–22. https://doi.org/10.5194/hess-22-6005-2018.

Kratzert, Frederik, Daniel Klotz, Mathew Herrnegger, Alden K. Sampson, Sepp Hochreiter, and Grey S. Nearing. 2019. “Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning.” *Water Resources Research* 55 (12): 11344–54. https://doi.org/10.1029/2019WR026065.

Kratzert, Frederik, Grey Nearing, Nans Addor, Tyler Erickson, Martin Gauch, Oren Gilon, Lukas Gudmundsson, et al. 2023. “Caravan - A Global Community Dataset for Large-Sample Hydrology.” *Scientific Data* 10 (1): 61. https://doi.org/10.1038/s41597-023-01975-w.

Li, Bing, Guishan Yang, Rongrong Wan, Xue Dai, and Yanhui Zhang. 2016. “Comparison of Random Forests and Other Statistical Methods for the Prediction of Lake Water Level:

A Case Study of the Poyang Lake in China." *Hydrology Research* 47 (S1): 69–83. https://doi.org/10.2166/nh.2016.264.

Linke, Simon, Bernhard Lehner, Camille Ouellet Dallaire, Joseph Ariwi, Günther Grill, Mira Anand, Penny Beames, et al. 2019. "Global Hydro-Environmental Sub-Basin and River Reach Characteristics at High Spatial Resolution." *Scientific Data* 6 (1): 283. https://doi.org/10.1038/s41597-019-0300-6.

Liu, Shizhan, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2022. "PYRAFORMER: LOW-COMPLEXITY PYRAMIDAL AT- TENTION FOR LONG-RANGE TIME SERIES MODELING AND FORECASTING."

Maier, Holger R., Stefano Galelli, Saman Razavi, Andrea Castelletti, Andrea Rizzoli, Ioannis N. Athanasiadis, Miquel Sànchez-Marrè, Marco Acutis, Wenyan Wu, and Greer B. Humphrey. 2023. "Exploding the Myths: An Introduction to Artificial Neural Networks for Prediction and Forecasting." *Environmental Modelling & Software* 167 (September):105776. https://doi.org/10.1016/j.envsoft.2023.105776.

Muñoz-Sabater, Joaquín, Emanuel Dutra, Anna Agustí-Panareda, Clément Albergel, Gabriele Arduini, Gianpaolo Balsamo, Souhail Boussetta, et al. 2021. "ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications." *Earth System Science Data* 13 (9): 4349–83. https://doi.org/10.5194/essd-13-4349-2021.

Nash, J.E., and J.V. Sutcliffe. 1970. "River Flow Forecasting through Conceptual Models Part I — A Discussion of Principles." *Journal of Hydrology* 10 (3): 282–90. https://doi.org/10.1016/0022-1694(70)90255-6.

Nevo, Sella, Efrat Morin, Adi Gerzi Rosenthal, Asher Metzger, Chen Barshai, Dana Weitzner, Dafi Voloshin, et al. 2022. "Flood Forecasting with Machine Learning Models in an Operational Framework." *Hydrology and Earth System Sciences* 26 (15): 4013–32. https://doi.org/10.5194/hess-26-4013-2022.

Newman, A. J., M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, et al. 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 (1): 209–23. https://doi.org/10.5194/hess-19-209-2015.

Organization, World Meteorological. 1975. *Intercomparison of Conceptual Models Used in Operational Hydrological Forecasting*. Hydrological Network Design and Information Transfer: Proceedings of the International Seminar, τ. 6-8. Secretariat of the World Meteorological Organization. https://books.google.gr/books?id=d-QHAQAAIAAJ.

Pechlivanidis, Ilias G., Bethanna Jackson, Neil McIntyre, and Howard S. Wheater. 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* 13:193–214.

Raghavendra. N, Sujay, and Paresh Chandra Deka. 2014. "Support Vector Machine Applications in the Field of Hydrology: A Review." *Applied Soft Computing* 19 (June):372–86. https://doi.org/10.1016/j.asoc.2014.02.002.

Rakesh Tanty, Tanweer S. Desmukh, and Manit Bhopal. 2015. "Application of Artificial Neural Network in Hydrology- A Review." *International Journal of Engineering Research And* V4 (06): IJERTV4IS060247. https://doi.org/10.17577/IJERTV4IS060247.

Rumelhart, David E, Geoffrey E Hintont, and Ronald J Williams. 1986. "Learning Representations by Back-Propagating Errors."

Sanjay Potdar, Akhil, Pierre-Emmanuel Kirstetter, Devon Woods, and Manabendra Saharia. 2021. "Towards Predicting Flood Event Peak Discharge in Ungauged Basins by Learning Universal Hydrological Behaviors with Machine Learning." *Journal of Hydrometeorology*, August. https://doi.org/10.1175/JHM-D-20-0302.1.

Schaefli, Bettina, and Hoshin V. Gupta. 2007. "Do Nash Values Have Value?" *Hydrological Processes* 21 (15): 2075–80. https://doi.org/10.1002/hyp.6825.

Seaber, Paul. R, F. Paul Kapinos, and George L. Knapp. 1987. "Hydrologic Unit Maps."

Shen, Li, and Yangzhu Wang. 2022. "TCCT: Tightly-Coupled Convolutional Transformer on Time Series Forecasting." *Neurocomputing* 480 (April):131–45. https://doi.org/10.1016/j.neucom.2022.01.039.

Shortridge, Julie E., Seth D. Guikema, and Benjamin F. Zaitchik. 2016. "Machine Learning Methods for Empirical Streamflow Simulation: A Comparison of Model Accuracy,

Interpretability, and Uncertainty in Seasonal Watersheds." *Hydrology and Earth System Sciences* 20 (7): 2611–28. https://doi.org/10.5194/hess-20-2611-2016.

Shrestha, Shiva Gopal, and Soni M. Pradhanang. 2023. "Performance of LSTM over SWAT in Rainfall-Runoff Modeling in a Small, Forested Watershed: A Case Study of Cork Brook, RI." *Water* 15 (23): 4194. https://doi.org/10.3390/w15234194.

Tyralis, Hristos, Georgia Papacharalampous, and Andreas Langousis. 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.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. "Attention Is All You Need." arXiv. http://arxiv.org/abs/1706.03762.

Wen, Qingsong, Tian Zhou, Chaoli Zhang, Weiqi Chen, Ziqing Ma, Junchi Yan, and Liang Sun. 2023. "Transformers in Time Series: A Survey." arXiv. http://arxiv.org/abs/2202.07125.

Worland, Scott C., William H. Farmer, and Julie E. Kiang. 2018. "Improving Predictions of Hydrological Low-Flow Indices in Ungaged Basins Using Machine Learning." *Environmental Modelling & Software* 101 (March):169–82. https://doi.org/10.1016/j.envsoft.2017.12.021.

Wu, Haixu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2022. "Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting." arXiv. http://arxiv.org/abs/2106.13008.

Yin, Hanlin, Zilong Guo, Xiuwei Zhang, Jiaojiao Chen, and Yanning Zhang. 2022. "RR-Former: Rainfall-Runoff Modeling Based on Transformer." *Journal of Hydrology* 609 (June):127781. https://doi.org/10.1016/j.jhydrol.2022.127781.

Yin, Hanlin, Fandu Wang, Xiuwei Zhang, Yanning Zhang, Jiaojiao Chen, Runliang Xia, and Jin Jin. 2022. "Rainfall-Runoff Modeling Using Long Short-Term Memory Based Step-Sequence Framework." *Journal of Hydrology* 610 (July):127901. https://doi.org/10.1016/j.jhydrol.2022.127901.

Yin, Hanlin, Xiuwei Zhang, Fandu Wang, Yanning Zhang, Runliang Xia, and Jin Jin. 2021. "Rainfall-Runoff Modeling Using LSTM-Based Multi-State-Vector Sequence-to-Sequence Model." *Journal of Hydrology* 598 (July):126378. https://doi.org/10.1016/j.jhydrol.2021.126378.

Zhang, Yunhao, and Junchi Yan. 2023. "CROSSFORMER: TRANSFORMER UTILIZING CROSS-DIMENSION DEPENDENCY FOR MULTIVARIATE TIME SERIES FORECASTING."

Zhou, Haoyi, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. 2021. "Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting." arXiv. http://arxiv.org/abs/2012.07436.

Zhou, Tian, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. 2022. "FEDformer: Frequency Enhanced Decomposed Transformer for Long-Term Series Forecasting." arXiv. http://arxiv.org/abs/2201.12740.