ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΥΔΑΤΙΚΩΝ ΠΟΡΩΝ ΚΑΙ ΠΕΡΙΒΑΛΛΟΝΤΟΣ

Extreme-oriented rainfall modelling on global scale using knowable moments

---

Μοντελοποίηση ακραίων βροχοπτώσεων σε παγκόσμια κλίμακα με τη χρήση εύγνωστων ροπών

Αγκαθέρης Νικόλαος
Επιβλέπων καθηγητής: Δημήτριος Κουτσογιάννης

Αθήνα, Οκτώβριος 2019

"If you can look into the seeds of time, and
say which grain will grow and which will not,
speak then unto me. "

-- William Shakespeare

"Prediction is very difficult, especially if it's
about the future."

-- Nils Bohr

Ευχαριστίες | Acknowledgments

Ένας σημαντικός κύκλος της ζωής μου κλείνει με την περάτωση αυτής της διπλωματικής εργασίας. Μαζί με αυτό ανοίγει η πόρτα για καινούργιες ευκαιρίες και αναμνήσεις. Στο σημείο αυτό θα ήθελα να ευχαριστήσω σημαντικά πρόσωπα τα οποία το καθένα με τον δικό του ρόλο με ώθησε να φτάσω στο σημείο που βρίσκομαι σήμερα.

Αρχικά, μεγάλο ευχαριστώ στον καθηγητή μου και πρώην κοσμήτορα της σχολής Πολιτικών Μηχανικών ΕΜΠ κ. Δημήτριο Κουτσογιάννη ο οποίος αποτέλεσε και αποτελεί έμπνευση με τις ιδέες και το ερευνητικό του έργο. Τον ευχαριστώ που μου εμπιστεύθηκε την ιδέα και το θέμα αυτής της διπλωματικής εργασίας, δείχνοντας πίστη σε μένα και τις ικανότητές μου. Με δική του παρότρυνση έλαβα μέρος στο συνέδριο της EGU 2019 παρουσιάζοντας ένα τμήμα από τον παρόν θέμα. Ήταν εκεί σε κάθε βήμα, από τον Μάρτιο μέχρι και σήμερα, ώστε να μου λύνει όλες τις απορίες που εμφανίζονταν. Χωρίς την συνεχή καθοδήγηση του θα ήταν αδύνατο να έρθει σε πέρας η εργασία.

Ακόμη ένα μεγάλο ευχαριστώ στον Παναγιώτη Δημητριάδη και την Άνυ Ηλιοπούλου, οι οποίοι μαζί με τον κ. Κουτσογιάννη μου κίνησαν το ενδιαφέρον προς τις στοχαστικές μεθόδους και την παραγωγή μοντέλων. Η συμβολή τους ειδικά για την προετοιμασία της εργασίας για το συνέδριο της EGU ήταν καθοριστική και ανεκτίμητη, διότι από την πρώτη έως την τελευταία μέρα με βοηθούσαν αδιάκοπα για την ολοκλήρωση της.

Τέλος, το μεγαλύτερο ευχαριστώ το οφείλω συγκεντρωτικά στην οικογένεια μου για την αδιάκοπη στήριξη τους σε κάθε βήμα αυτής της πορείας και στους φίλους μου οι οποίοι όλα αυτά τα χρόνια μου προσέφεραν αμέτρητες αναμνήσεις, μοναδικές στιγμές, και γνωρίζω ότι θα είναι εκεί και στο μέλλον για οτιδήποτε χρειαστώ.

Αγκαθέρης Νικόλας

Αθήνα, 2019

Abstract

Assessment of extremes in hydrological processes is crucial in a variety of tasks from engineering design to risk management. Using classical moments to express important attributes of such assessment, proves to be efficient only for low order of moments. However, extreme rainfall events are better modelled using high-order moments. Whilst L – moments can be reliably estimated even for those higher orders, they fail in accounting for long-term dependence bias which exists in most large hydrological records. Thus, the newly introduced *knowable* (K) moments are used to model extremes, as they provide better grounds for prediction based on high orders, whilst retaining precision of classical moments for low orders. This study's findings may improve knowledge on how to correctly model and predict such extreme rainfall events, providing comparison between the effectiveness of K – moments and classic methods. As this is a global study using data from the GHCN – Daily database, an attempt is made at constructing the basic framework for correlating a distribution's fitting parameters and regional climatic characteristics.

Εκτενής Περίληψη στα Ελληνικά | Extended Abstract in Greek

Οι ανησυχίες σχετικά με τις ακραίες καιρικές συνθήκες διογκώνονται συνεχώς, σε μια εποχή όπου η κλιματική μεταβλητότητα βρίσκεται στο προσκήνιο. Η αξιολόγηση αυτών των ακραίων συνθηκών, ιδίως όταν αναφερόμαστε σε ακραία φαινόμενα υδρολογικών διεργασιών, είναι καίριας σημασίας για μια ποικιλία εφαρμογών από τον σχεδιασμό των έργων υποδομής μέχρι τη διαχείριση του κινδύνου.

Κύριος στόχος της μελέτης είναι να επιτευχθεί η δημιουργία ενός γενικού πλαισίου για τη μοντελοποίηση ακραίων βροχοπτώσεων με τη χρήση της νεοεισαχθείσας μεθόδου των εύγνωστων ροπών (K – moments / K – ροπές). Επίσης, οι δύο πιο κλασικές μέθοδοι μοντελοποίησης χρησιμοποιούνται, προκειμένου να συγκριθούν και να αξιολογηθούν συγκριτικά με την νέα μέθοδο για την ισχύ πρόβλεψής τους. Δεδομένου ότι η μελέτη αποτελεί παγκόσμια ανάλυση, θα χρησιμοποιηθεί μια καθιερωμένη βάση δεδομένων μέτρησης βροχόπτωσης των σταθμών από όλο τον κόσμο, η οποία στην περίπτωση αυτή είναι η GHCN – Daily από τον *National Oceanic and Atmospheric Administration* των Ηνωμένων Πολιτειών της Αμερικής. Με τη χρήση παγκόσμιων δεδομένων, η μελέτη στοχεύει στο να ερευνήσει την αξιοπιστία και τη συνέπεια των K - ροπών για τα κλιματικά χαρακτηριστικά κάθε περιοχής.

Μετά την καθιέρωση των όποιων πλεονεκτημάτων έχει η χρήση K – ροπών, είναι υποχρεωτικό να εκτιμηθεί η επίδραση της μακροπρόθεσμης εμμονής που υπάρχει στα περισσότερα δείγματα βροχόπτωσης. Οι K – ροπές παρέχουν το πλαίσιο για την εκτίμηση αυτής της μεροληψίας. Συνεπώς, η ποσοτικοποιημένη μεροληψία που προκαλείται από την μακροχρόνια εμμονή προστίθεται στο τελικό μοντέλο και αυτό με τη σειρά του συγκρίνεται με το ίδιο μοντέλο αγνοώντας την εμμονή. Αυτό έχει ως αποτέλεσμα να φανεί η σημαντική επιρροή που κατέχει η μεροληψία αυτή στις τελικές προβλεπόμενες τιμές βροχόπτωσης.

Τέλος, υπάρχει προοπτική στην ανάλυση της κατανομής των παραμέτρων που προκύπτουν από τη θεωρητική συνάρτηση κατανομής πιθανότητας. Η ανάλυση αυτή γίνεται σε ολόκληρο τον κόσμο για την εξεύρεση συσχετισμού μεταξύ των τιμών που λαμβάνουν και των κλιματικών χαρακτηριστικών της κάθε περιοχής. Στο πέρας της μελέτης αυτής, μπορεί να καθοριστεί ένα γενικό πλαίσιο αναμενόμενων τιμών παραμέτρων για μελλοντική αναφορά.

Όσον αφορά τις ακραίες βροχοπτώσεις, όπως και όλα τα υπόλοιπα ακραία καιρικά φαινόμενα αποτελούν φυσικό μέρος του κλιματικού συστήματος της Γης. Οι ακραίες εκφάνσεις της κατακρήμνισης θα πρέπει να αναμένονται, και εκφράζονται είτε σαν μακροχρόνιες ξηρασίες, είτε σαν συχνές εμφανίσεις βροχών. Ωστόσο, αυτά τα ακραία γεγονότα έχουν σημαντικό αντίκτυπο στην καθημερινή ανθρώπινη ζωή, στις υποδομές, καθώς και στο περιβάλλον.

Στην εποχή της εμφανούς κλιματικής μεταβλητότητας οι ανησυχίες σχετικά με τις ακραίες καιρικές συνθήκες οξύνονται. Σύμφωνα με τη Διακυβερνητική Επιτροπή για την κλιματική αλλαγή (IPCC, 2012), πρέπει να δοθεί η μεγαλύτερη προσοχή στην αξιόπιστη πρόβλεψη ακραίων εμφανίσεων οποιουδήποτε είδους φυσικής διαδικασίας. Η έκθεση αυτή ανέφερε

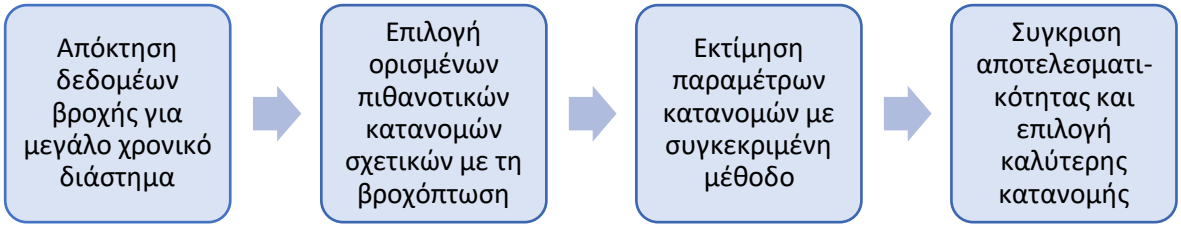
επίσης έρευνες που δείχνουν ότι τα μοντέλα που αξιολογούν παρελθόντα γεγονότα υποδεικνύουν μια ελαφρά αύξηση των ακραίων φυσικών γεγονότων.

Πιο συγκεκριμένα, ζημιές στην ιδιοκτησία και το περιβάλλον (loss events) που αποδίδονται σε ακραία φαινόμενα υδρολογικής φύσης εμφανίζουν συνεχή αυξανόμενη εμφάνιση από τη δεκαετία του 1990 μέχρι σήμερα. Ενώ η επιφανειακή ανάλυση υποδεικνύει συσχετισμό μεταξύ του αριθμού των πλημμυρών και των γεγονότων απώλειας, είναι σημαντικό να ληφθούν υπόψη οι αυξημένες εκτάσεις γης που χρησιμοποιούνται τώρα για στέγαση και βιομηχανική υποδομή, οι οποίες μπορούν κάλλιστα να αποτελέσουν τον λόγο για την αύξηση αυτή των γεγονότων απώλειας. Οποιοδήποτε από τα δύο και να ισχύει, το συμπέρασμα είναι το ίδιο.

Ο σημαντικότερος κίνδυνος από τις ακραίες βροχοπτώσεις είναι οι πλημμύρες. Σύμφωνα με τον Οργανισμό Οικονομικής Συνεργασίας και Ανάπτυξης (Organization for Economic Cooperation and Development), οι πλημμύρες προκαλούν ετησίως $40 δισ. εκατ. ζημίες, τόσο σε κατοικημένες περιοχές όσο και σε υποδομές. Από το 1995, οι πλημμύρες αποτελούν το 43% όλων των φυσικών καταστροφών που σχετίζονται με τον καιρό και επηρεάζουν συνολικά 2.3 δις. εκατ. άτομα. Σε συνδυασμό με την καταστροφή των ανθρώπινων περιουσιακών στοιχείων, οι ζημίες στη γεωργία οφείλονται κατά κύριο λόγο από πλημμύρες, πράγμα που σημαίνει ότι υπονομεύεται η φυτική παραγωγή, προκαλώντας ζημίες και στον χρηματοπιστωτικό τομέα.

Από επιστημονική σκοπιά, η γενικευμένη μελέτη της κατανομής των βροχοπτώσεων με την πάροδο του χρόνου για μια συγκεκριμένη περιοχή είναι ζωτικής σημασίας για την αξιολόγηση της ποσότητας νερού που διατίθεται για την κάλυψη των απαιτήσεων της βιομηχανίας, της γεωργίας ή άλλων ανθρώπινων δραστηριοτήτων. Ωστόσο, η ακρίβεια στην πρόβλεψη ακραίων γεγονότων είναι επίσης σημαντική, δεδομένου ότι χρησιμοποιούνται για τον σχεδιασμό και την κατασκευή έργων που προορίζονται για σκοπούς διαχείρισης του νερού, όπως φράγματα, έργα διαχείρισης πλημμυρικού κινδύνου και υδροηλεκτρικές μονάδες ηλεκτροπαραγωγής. Υποεκτιμώντας το μέγεθος της βροχόπτωσης, είναι βέβαιο ότι θα οδηγήσει σε αποτυχίες ή ανεπαρκή άμβλυνση των πλημμυρών, θέτοντας σε κίνδυνο κατοικημένες περιοχές και ανθρώπινες ζωές. Αντίθετα, η υπερεκτίμηση του οδηγεί σε οικονομικές ζημίες, δεδομένου ότι θα χρησιμοποιηθούν περιττοί πόροι για την κατασκευή και τη συντήρηση του εκάστοτε έργου.

Παρόλο που η χρήση της ντετερμινιστικής προσέγγισης για την βραχυπρόθεσμη πρόβλεψη βροχοπτώσεων είναι εφικτή με τα σημερινά τεχνολογικά πρότυπα και τα μετεωρολογικά μοντέλα, δεν είναι δυνατή η χρήση της όσον αφορά τις μακροπρόθεσμες προβλέψεις, που ενδιαφέρουν το σχεδιασμό και τη κατασκευή των μεγάλων έργων υποδομής. Σε αυτό το πλαίσιο, οι βροχοπτώσεις πρέπει να αντιμετωπίζονται ως μια τυχαία μεταβλητή που ακολουθεί μια καθορισμένη συνάρτηση κατανομής πιθανότητας, η οποία παρέχει τη δυνατότητα σύνδεσης περιόδων επαναφοράς σε τιμές βροχόπτωσης (Papalexiou et al, 2012). Η σύνοψη της γενικής διαδικασίας μοντελοποίησης παρατίθεται παρακάτω:



Εικόνα 1: Σύνοψη διαδικασίας επιλογής κατάλληλης κατανομής

Με βάση τα ευρήματα των Papalexiou et al (2012), οι βαριές (όσον αφορά την ουρά) κατανομές είναι πιο κατάλληλες για την περιγραφή των μακροπρόθεσμων χαρακτηριστικών της βροχόπτωσης και ιδιαίτερα τις ακραίες τιμές της. Έτσι, βαριές κατανομές χρησιμοποιούνται για τη μοντελοποίηση και πιο συγκεκριμένα η Γενικευμένη Κατανομή Pareto (Generalized Pareto Distribution) και η Pareto-Burr-Feller (PBF), η οποία είναι μια ειδική περίπτωση της κατανομής Burr που αποδείχθηκε μαθηματικά από τον Feller (Dimitriadis, 2017).

Η Γενικευμένη Κατανομή Pareto (GPD) μετά την συμβολή και του Pickands (1975) έκτοτε έχει χρησιμοποιηθεί εκτενώς σε πολλούς τομείς της επιστημονικής έρευνας. Ορισμένες από τις εφαρμογές της καλύπτουν ανάλυση ακραίων γεγονότων ή μοντελοποίηση μεγάλων ασφαλιστικών διεκδικήσεων (Hosking & Wallis, 1987). Αποτελεί μια οικογένεια συνεχών κατανομών πυκνότητας πιθανότητας, και εκφράζεται από τρεις παραμέτρους: Δείκτη ουράς κ, κλίμακα λ (ή b), και θέση ψ. Για τη μελέτη αυτή:

A. Παρόλο που η χρήση και των τριών παραμέτρων θα έχει ως πιθανό αποτέλεσμα μεγαλύτερη συνολική ακρίβεια, η παράμετρος θέσης έχει οριστεί $\psi = 0$, ώστε να είναι φυσικά συνεπής με το χαμηλότερο όριο της βροχόπτωσης το οποίο είναι το μηδέν.

B. Για $\kappa = 0$ η κατανομή Pareto μετατρέπεται στην εκθετική κατανομή.

C. Για $\kappa < 0$ η ουρά τείνει ταχύτερα στο μηδέν και θεωρείται «ελαφρά» άρα ακατάλληλη για τη διαδικασία της βροχόπτωσης. Ταυτόχρονα, κατανομές με αρνητικό κ είναι άνω φραγμένες το οποίο είναι φυσικά λανθασμένο.

Ομοίως με τη Γενικευμένη Κατανομή Pareto, η Pareto-Burr-Feller (PBF) είναι μια παρόμοια κατανομή με τρεις παραμέτρους. Είναι μια κατανομή που χρησιμοποιείται κυρίως στην Οικονομετρία (Singh & Maddala, 1976), και εμφανίζεται και ως Pareto IV ή Burr XII. Η παραγωγή της PBF μελετήθηκε από τον Burr (1942) και αποδείχθηκε μαθηματικά από τον Feller (1970) ο οποίος τη συνέδεσε με την συνάρτηση Βήτα και την αντίστοιχη κατανομή της. Η χρησιμότητά της σε μια ποικιλία πεδίων μελετήθηκε από τον Brouers (2015). Σε αυτή τη μελέτη, χρησιμοποιείται σε συνδυασμό με την GPD για τη μοντελοποίηση ακραίων βροχοπτώσεων, δεδομένου ότι η προσθήκη μιας τρίτης παραμέτρου μπορεί να αποδειχθεί επωφελής για την ακρίβεια του τελικού μοντέλου.

Η χρήση των στατιστικών ροπών προσφέρει τη δυνατότητα περιγραφής των κατανομών πιθανότητας με μεγάλη ευκολία (Feller, 1968). Κατά την ανάλυση ενός μετρήσιμου μεγέθους για διαφορετικές χρονικές κλίμακες, στέκονται ως το βασικό εργαλείο για το στοχαστικό χαρακτηρισμό της αλλαγής και της μεταβλητότητας, τα οποία αποτελούν

σημαντικά χαρακτηριστικά κάθε φυσικής διεργασίας. Ωστόσο, τόσο οι κλασικές όσο και η L – ροπές, οι δύο βασικές μέθοδοι για τον χαρακτηρισμό μιας κατανομής, έχουν μειονεκτήματα.

Οι κλασικές ροπές, κεντρικές ή μη κεντρικές, δεν μπορούν να εκτιμηθούν αξιόπιστα από μεγάλα δείγματα για τάξεις μεγαλύτερες από 2 ή 3 (Lombardo et al, 2014). Όπως εξετάστηκε από τον Koutsoyiannis (2019), για υψηλές τάξεις (*p*), η εκτιμήτρια κλασικής ροπής απεικονίζει μία ακραία ποσότητα και παρουσιάζει σημαντικά αργή σύγκλιση στη θεωρητική τιμή. Αυτό σε συνδυασμό με το γεγονός ότι οι περισσότερες γεωφυσικές και υδρολογικές διεργασίες δεν ακολουθούν την κανονική κατανομή, σημαίνει ότι οι κλασικές μέθοδοι δεν είναι ιδανικές για να χαρακτηριστούν οι κατανομές αξιόπιστα.

Αντίθετα, οι L – ροπές, εκτιμώνται ακόμη και αν μόνο η πρώτη κλασική ροπή (μέσος όρος) είναι πεπερασμένη. Το πιο σημαντικό μειονέκτημα τους είναι η ανικανότητα να χαρακτηρίζουν και να εκτιμούν την εμμονή στοχαστικών διεργασιών. Η εμμονή, όπως προαναφέρθηκε, αποτελεί σημαντικό χαρακτηριστικό των περισσότερων γεωφυσικών διεργασιών και είναι απαραίτητο να υπολογίζεται η επιρροή της.

Οι ακραίες τιμές βρίσκονται στην ουρά της συνάρτησης κατανομής, έτσι συσχετίζονται στενά με ροπές υψηλής τάξης. Οι K – ροπές, συνδυάζουν τα πλεονεκτήματα των κλασικών και L – ροπών, επιτρέποντας την αξιόπιστη εκτίμηση και περιγραφή των στατιστικών χαρακτηριστικών υψηλών τάξεων, ενώ παράλληλα παρέχουν το πλαίσιο για την εκτίμηση της μακροπρόθεσμης εμμονής.

Οι εκτιμήτριες των *μη κεντρικών* και *κεντρικών* αμερόληπτων K – ροπών ορίζονται ως:

$$K'_{pq} = \sum_{i=1}^{n} b_{i,n,p-q+1} \underline{x}^{q}_{(i:n)} \tag{1}$$

$$K'_{p1} = \sum_{i=1}^{n} b_{i,n,p} \left(\underline{x}_{(i:n)} - \mu \right) \tag{2}$$

με τον b_{inp} να αντιστοιχεί σε:

$$b_{inp} = \begin{cases} 0, & i < p \\ \dfrac{p}{n} \dfrac{\Gamma(n-p+1)}{\Gamma(n)} \dfrac{\Gamma(i)}{\Gamma(i-p+1)}, & i \geq p \geq 0 \end{cases} \tag{3}$$

όπου:

- $x_{i:n}$ είναι το ταξινομημένο σύνολο των παρατηρήσεων κατά αύξοντα αριθμό.
- Γ είναι η συνάρτηση Γάμα και *p* ορίζει την τάξη της ροπής και μπορεί να είναι οποιοσδήποτε θετικός αριθμός, συνήθως ακέραιος, αλλά αυτό δεν είναι απαραίτητο.
- μ είναι η μέση τιμή του δείγματος
- Η αμεροληψία της κεντρικής K – ροπής επιτυγχάνεται μόνο για $q = 1$ αλλά αυτό επαρκεί για τους σκοπούς της μελέτης.

Η πρακτική σημασία του όρου b_{inp} είναι το γεγονός ότι καθώς η ροπή αυξάνει σε τάξη (p) όλο και λιγότερα δεδομένα από το δείγμα καθορίζουν την τελική εκτίμηση της Κ – ροπής. Αυτό αποδεικνύεται από το γεγονός ότι $b_{inp} = 0$ για $i < p$. Άρα, μεγαλύτερη έμφαση δίνεται σε υψηλότερες τιμές του δείγματος, από ό, τι σε χαμηλότερες, το οποίο με τη σειρά του ανοίγει το δρόμο για πιο ακριβή εκτίμηση ροπών υψηλής τάξης με ελάχιστη υπολογιστική ισχύ, έτσι η μοντελοποίηση ακραίων φαινομένων επιτυγχάνεται με μεγαλύτερη ταχύτητα και ακρίβεια.

Η μοντελοποίηση με τις παραπάνω μεθόδους στοχεύει πρακτικά στον ακριβή ορισμό τιμών βροχόπτωσης σε συγκεκριμένες περιόδους επαναφοράς. Η αντιστοίχιση περιόδων επαναφοράς σε παρατηρούμενες τιμές είναι σημαντική για την στοχαστική μοντελοποίηση των ακραίων τιμών. Από τον αρχικό ορισμό τους από τον Fuller (1914), η έννοια της περιόδου επαναφοράς είναι κρίσιμη για τον σχεδιασμό και την αξιολόγηση των κινδύνων των περισσότερων κατασκευών, παρέχοντας τα μέσα αξιολόγησης της συχνότητας των ακραίων γεγονότων. Σε όρους πιθανοτήτων, η περίοδος επαναφοράς συνδέεται αντιστρόφως με την πιθανότητα υπέρβασης μιας συγκεκριμένης αξίας μιας μεταβλητής άρα έχει άμεση σχέση με στατιστική ανάλυση ταξινομημένου δείγματος (order statistics).

Με την ίδια νοοτροπία, αφού οι Κ – ροπές είναι κατασκευασμένες με βάση τις θεωρητικές ιδιότητες της στατιστικής ταξινομημένου δείγματος, είναι προφανές ότι μπορούν να αντιστοιχιστούν και περίοδοι επαναφοράς σε αυτές. Αυτό εδραιώνεται με τον ορισμό των Λ – συντελεστών οι οποίοι χρησιμοποιούνται για την αντιστοίχιση εμπειρικών περιόδων επαναφοράς σε κάθε υπολογιζόμενη Κ – ροπή.

Σημαντική ιδιότητα των Κ – ροπών είναι και η ικανότητα υπολογισμού της μεροληψίας λόγω μακροχρόνιας εμμονής. Η θεωρητική έννοια της μακροχρόνιας εμμονής, η οποία υπάρχει στις περισσότερες φυσικές διεργασίες, συμπεριλαμβανομένων των βροχόπτωσης, ανακαλύφθηκε από τα έργα του Η.Ε. Hurst (1951) που σπούδαζε τις μακροπρόθεσμες χωρητικότητες ταμιευτήρων. Πριν από αυτό, ο Α. Kolmogorov (1941) έδωσε μαθηματικό ορισμό σε αυτή την έννοια, ενώ ανέλυσε τα χαρακτηριστικά της τύρβης. Σήμερα, αναγνωρίζεται ως το φαινόμενο Hurst ή συμπεριφορά Hurst-Kolmogorov (ΗΚ) και ποσοτικοποιείται από τον συντελεστή Hurst (H).

Ο υπολογισμός του πραγματοποιείται μέσω της κλίσης του θεωρητικού Κ – κλιμακογράμματος. Το τελευταίο δείχνει τις αμερόληπτες κεντρικές Κ – ροπές σε σχέση με την χρονική κλίμακα, ενώ το κλασικό κλιμακόγραμμα παρουσιάζει τη διασπορά σε σχέση με την χρονική κλίμακα.

Πέρα από το θεωρητικό υπόβαθρο των μεθόδων που θα χρησιμοποιηθούν, πρέπει να επιλεγεί κατάλληλη βάση δεδομένων βροχόπτωσης η οποία θα πληροί τις ακόλουθες απαιτήσεις:

- Ημερήσιες μετρήσεις βροχόπτωσης πρέπει να χρησιμοποιηθούν (Min et al, 2011). Μεγαλύτερες χρονικές κλίμακες, δεν δείχνουν αξιόπιστα τη συχνότητα και την ένταση των ακραίων τιμών.
- Δεδομένου ότι ο απώτερος σκοπός είναι η αξιόπιστη πρόβλεψη των γεγονότων ακόμη και για ορίζοντα άνω των 1000 ετών, τα ιστορικά δεδομένα πρέπει να έχουν μήκος άνω των 30 ετών για τον επαρκή προσδιορισμό των επιπτώσεων της

μακροχρόνιας εμμονής στο δείγμα, η οποία αποτελεί αναπόσπαστο μέρος της διαδικασίας μοντελοποίησης.
- Για να αποδειχθεί η αποτελεσματικότητα της διαδικασίας μοντελοποίησης, πρέπει να χρησιμοποιηθούν σταθμοί που υπάγονται σε όλα τα είδη κλιμάτων.
- Η βάση δεδομένων πρέπει να παρέχει όσους περισσότερους σταθμούς γίνεται, ώστε το τελικό αποτέλεσμα να αναδεικνύει σε μεγαλύτερο βαθμό την αποτελεσματικότητα της μεθόδου.

Η καταλληλότερη βάση δεδομένων κρίθηκε η GHCN – Daily διαθέσιμη δωρεάν από τον ιστότοπο του NOAA. Παρέχει μέχρι στιγμής ημερήσιες μετρήσεις κατακρημνίσεων για 112,777 σταθμούς από το ξεδιάλεγμα των οποίων τελικά απομένουν 34,784 προς χρήση στην παρούσα μελέτη.

Η διαδικασία μοντελοποίησης ξεκινάει με ένα παράδειγμα ενός επιλεγμένου σταθμού και μετά θα γενικευθεί στο ευρύτερο σύνολο της βάσης δεδομένων. Για το παράδειγμα επιλέγεται ο σταθμός «SZ000002220» με συντεταγμένες [47,250, 9,340]. Βρίσκεται στην επαρχία Απενζέλ, η οποία είναι βορειοανατολική περιοχή της Ελβετίας. Πιο συγκεκριμένα βρίσκεται σε μια κορυφή βουνού των Άλπεων του Απενζέλ, που συνήθως ονομάζεται Säntis. Όλα τα δεδομένα καιρού από την αρχή της λειτουργίας του μέχρι σήμερα, συγκεντρώνονται στην GHCN – Daily. Έπειτα από ποιοτικό έλεγχο, η βάση δεδομένων μέχρι σήμερα περιέχει 43,276 ημερήσιες παρατηρήσεις, που ανέρχονται συνολικά σε 119 έτη συνεχών μετρήσεων.

Η διαδικασία μοντελοποίησης είναι ξεκάθαρη για τις κλασικές μεθόδους και περιλαμβάνει την χρήση των πρώτων 2 ροπών και τις θεωρητικές σχέσεις υπολογισμού των παραμέτρων της προαναφερθείσας κατανομής Pareto, αντίστοιχα. Όσον αφορά τις K – ροπές, η διαδικασία μπορεί να οργανωθεί στα εξής βήματα:

A. Με τη χρήση όλων των τιμών του δείγματος εκτός των μηδενικών, εκτιμώνται οι αμερόληπτες κεντρικές K – ροπές (2), για $q = 1$ και για p έως 1/10 το μέγεθος του δείγματος.
B. Το K – κλιμακόγραμμα κατασκευάζεται με βάση τις προαναφερθείσες κεντρικές K – ροπές και για κλίμακες έως 1/10 του μεγέθους του δείγματος. Από την κλίση του για μεγάλη χρονική κλίμακα εκτιμάται ο συντελεστής Hurst.
C. Εκτιμώνται οι μη κεντρικές αμερόληπτες K – ροπές (1), για $q = 1$ και για p μέχρι το μέγεθος του δείγματος n.
D. Ανάλογα με το μέγεθος του συντελεστή Hurst εκτιμάται και λαμβάνεται υπόψη η μεροληψία λόγω μακροχρόνιας εμμονής (3) στις μη κεντρικές αμερόληπτες K – ροπές (4).

$$\theta^{HK}(n,H) \approx \frac{2H(1-H)}{n-1} - \frac{1}{2(n-1)^{2-2H}} \tag{4}$$

$$K_{p1}^{d} = K_{p'1} = (1+\theta)K_{p1} \tag{5}$$

E. Με τις θεωρητικές εξισώσεις των συντελεστών Λ για την κατανομή Pareto, οι εμπειρικές περίοδοι επαναφοράς αποδίδονται στις μη κεντρικές K – ροπές.

F. Καθορίζοντας ένα σημείο εκκίνησης των κ και λ (παραμέτρων της κατανομής Pareto), εκτιμώνται οι θεωρητικές περίοδοι επαναφοράς.
G. Χρησιμοποιώντας έναν αλγόριθμο βελτιστοποίησης, η καλύτερη θεωρητική προσαρμογή παράγεται με την ελαχιστοποίηση του σφάλματος μεταξύ των εμπειρικά αντιστοιχισμένων περιόδων επαναφοράς και των αντίστοιχων θεωρητικών. Σε αυτή την περίπτωση, χρησιμοποιούνται τα ελάχιστα τετράγωνα (LSE). Δεδομένου ότι σκοπός αυτής της μελέτης είναι να μοντελοποιήσει αποτελεσματικά τις ακραίες τιμές, καθορίζοντας ένα κατώτατο όριο στις εμπειρικές περιόδους επαναφοράς ($T > 1\ \chi\rho\acute{o}\nu o$) και ελαχιστοποιώντας το LSE σε αυτό το εύρος επιτυγχάνεται η βέλτιστη προσαρμογή για αυτά. Η ευελιξία της μεθόδου είναι προφανής, καθώς το μοντέλο μπορεί να βαθμονομηθεί για να αναφέρεται σε οποιοδήποτε εύρος περιόδων επαναφοράς.
H. Παράγεται λογαριθμικό διάγραμμα με κατακόρυφο άξονα τις εντάσεις βροχής και οριζόντιο τις περιόδους επαναφοράς.

Ακολουθώντας αυτή τη μέθοδο, γίνεται αντιληπτό ότι δίνεται μεγαλύτερη έμφαση στην ουρά της κατανομής. Έτσι, ορισμένες φορές, ενώ ελαχιστοποιείται το LSE σε ένα συγκεκριμένο εύρος περιόδων επαναφοράς, το χαμηλότερο μέρος της κατανομής δεν προσαρμόζεται με τον καλύτερο τρόπο. Ενώ η ακρίβεια θυσιάζεται στις χαμηλότερες τιμές, η ακρίβεια στις ακραίες τιμές είναι πιο σημαντική, δεδομένου εκεί δίνεται προσοχή από τις περισσότερες μελέτες σχεδιασμού έργων και εκτίμησης κινδύνου. Η προσαρμογή όλων των μεθόδων παρουσιάζεται παρακάτω.

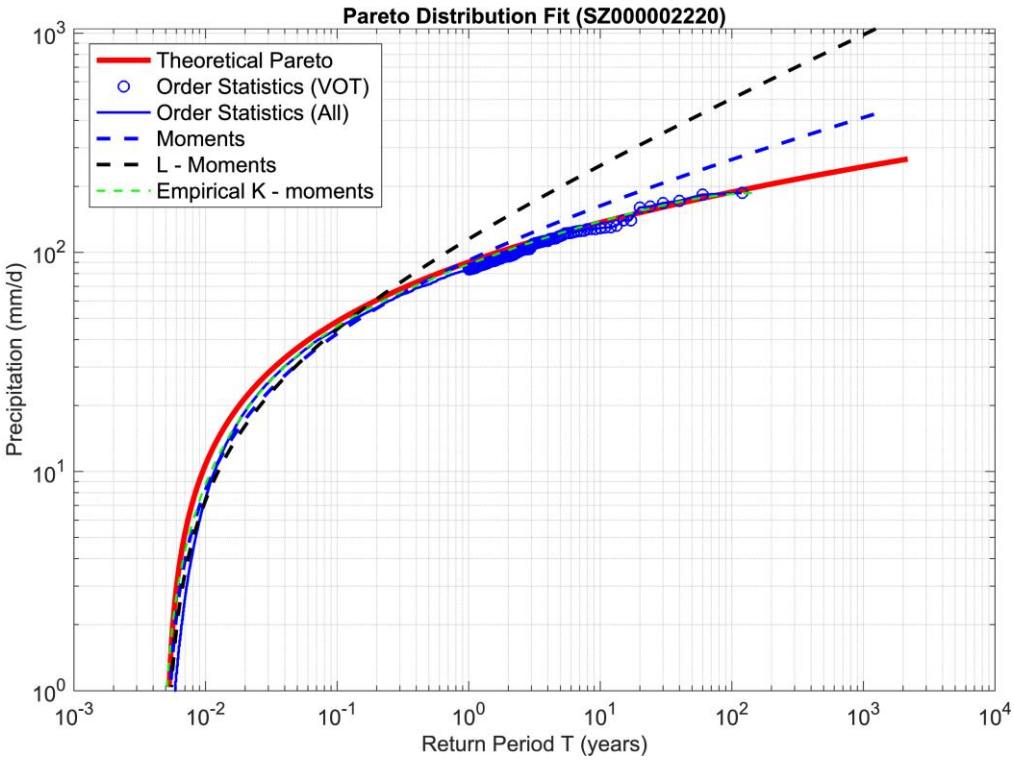
Η μέθοδος των Κ – ροπών υπερισχύει σημαντικά των κλασικών όταν αναφερόμαστε σε ακραίες τιμές βροχόπτωσης. Παρόλα αυτά, δείχνει μια σχετική αδυναμία σε σχέση με τις κλασικές μεθόδους στις χαμηλές τιμές. Η μεταγενέστερη χρήση της Pareto-Burr-Feller, με την προσθήκη της επιπλέον παραμέτρου, λύνει αυτό το πρόβλημα.

Πίνακας 1: Αποτελέσματα προσαρμογής κατανομής και απόδοση μεθόδων για τις ακραίες και τις χαμηλές τιμές, αλλά και συνολικά

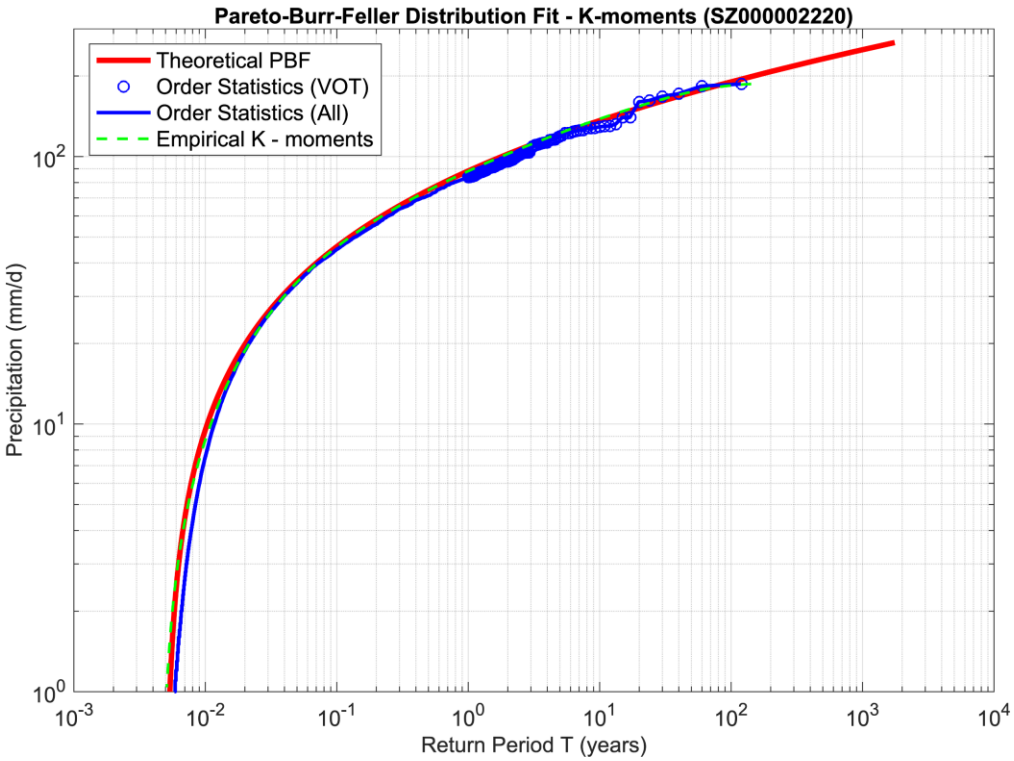
Method	κ	λ	RMSE High	RMSE Low	RMSE Total	NRMSE High	NRMSE Low	NRMSE Total
Classic moments	0.158	11.047	39.413	2.620	27.257	-0.177	0.895	0.521
L - moments	0.278	9.474	163.476	9.858	113.001	-3.883	0.603	-0.998
K - moments	0.046	15.000	5.921	3.820	4.933	0.823	0.846	0.913

Πίνακας 2: Αναμενόμενες τιμές βροχόπτωσης για διάφορες περιόδους επαναφοράς (σε χρόνια), με ή χωρίς την επιρροή της μεροληψίας λόγω μακροπρόθεσμης εμμονής

Κατάσταση Μεροληψίας	Αναμενόμενη Βροχόπτωση (mm/d)	
	$T = 100$	$T = 1000$
Αμερόληπτο	178.85	230.21
Με μεροληψία	187.86	245.58



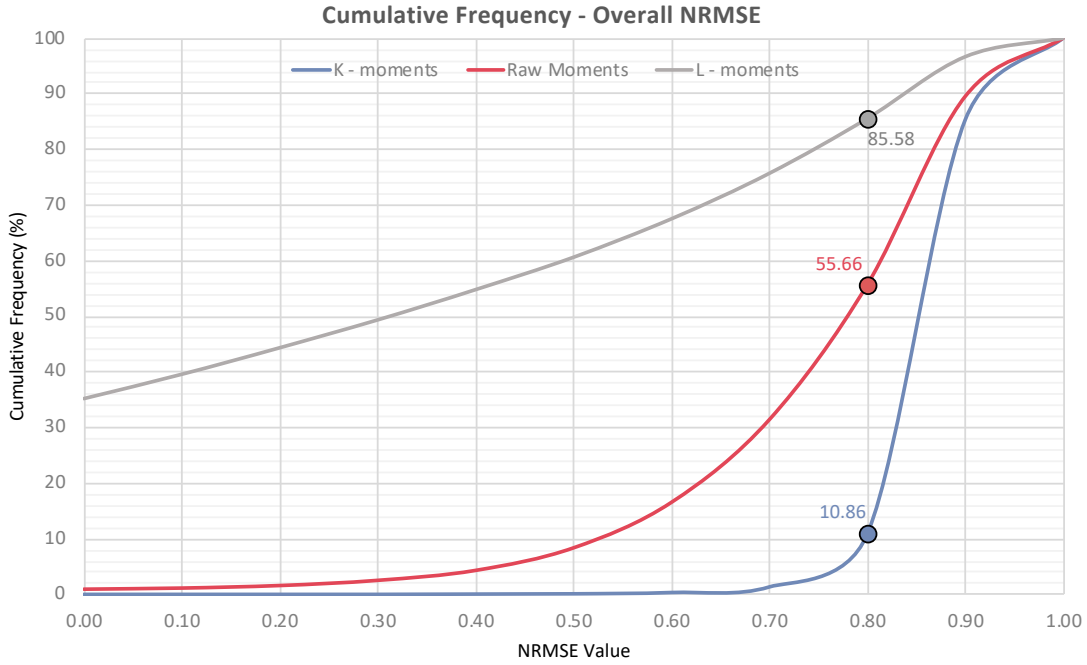
Διάγραμμα 1: Τελική προσαρμογή Γενικευμένης Κατανομής Pareto για όλες τις μεθόδους



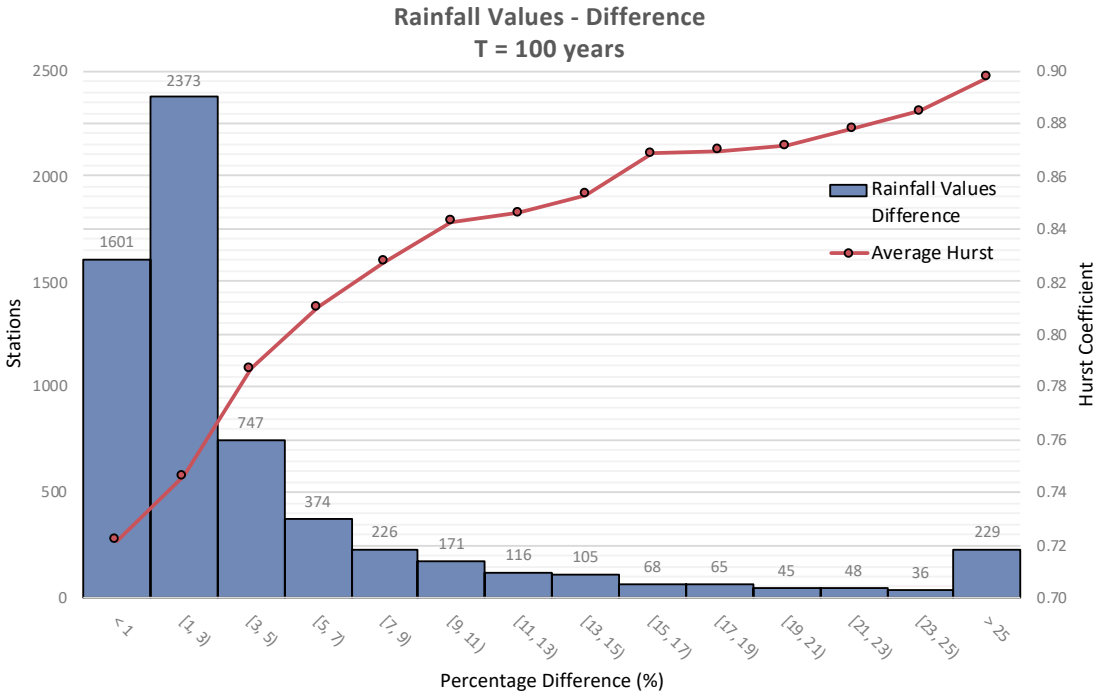
Διάγραμμα 2: Τελική προσαρμογή Pareto-Burr-Feller για τη μέθοδο των Κ – ροπών

Στη συνέχεια η διαδικασία παραγωγής μοντέλου γίνεται για κάθε σταθμό της βάσης δεδομένων και παρουσιάζονται στατιστικά στοιχεία για την απόδοση των μεθόδων χρησιμοποιώντας τα εργαλεία NRMSE (Normalized Root Squared Mean Error) και RMSE. Το σφάλμα NRMSE δείχνει την απόδοση της προσαρμογής του θεωρητικού μοντέλου στις

παρατηρούμενες τιμές. Όσο πιο κοντά στη 1, τόσο καλύτερο το μοντέλο. Όπως παρουσιάζεται και στο Διάγραμμα 3, οι Κ – ροπές εμφανίζουν το μικρότερο ποσοστό σταθμών με τιμή NRMSE χαμηλότερη από 0.8. Το ίδιο ισχύει και για τις ακραίες τιμές, ενώ για τις μικρότερες τιμές υπερτερούν ελάχιστα οι κλασικές μέθοδοι.



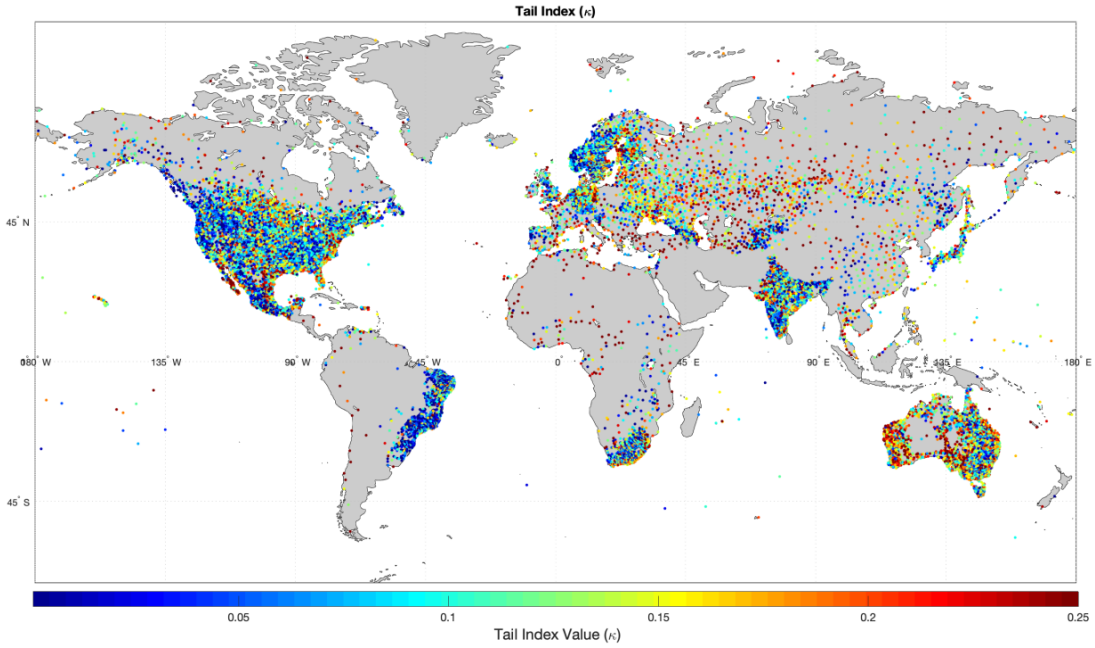
Διάγραμμα 3: Αθροιστική συχνότητα συνολικών NRMSE για όλες τις μεθόδους. Οι ετικέτες δεδομένων δείχνουν το ποσοστό των σταθμών όπου η εκτιμώμενη τιμή του NRMSE είναι χαμηλότερη από 0,8.



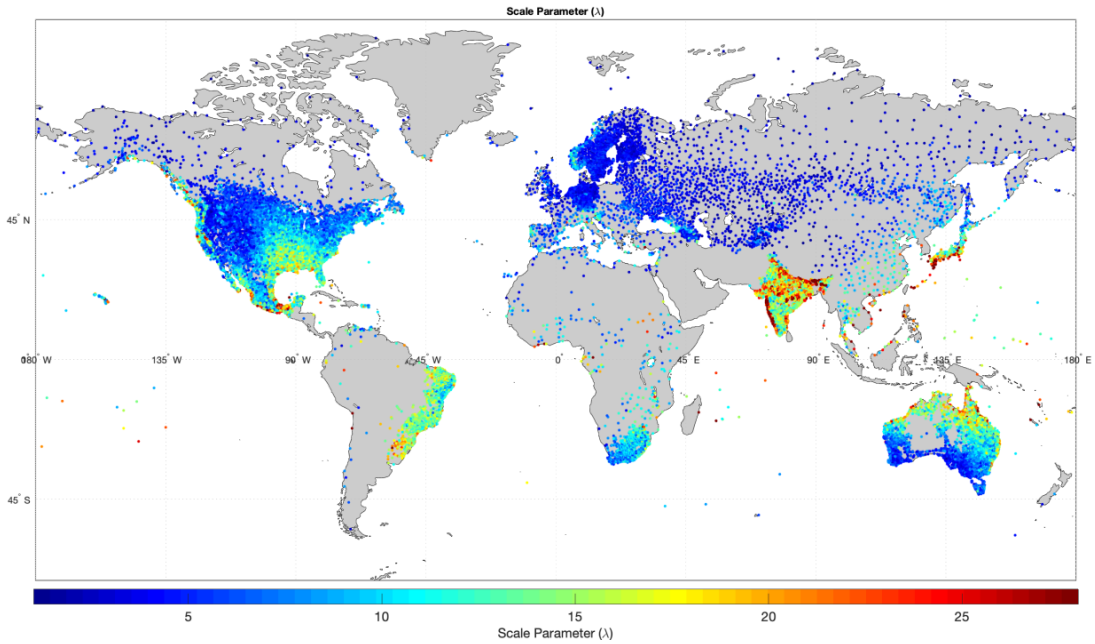
Διάγραμμα 4: Ποσοστιαίες διαφορές μεταξύ προσαρμογής με μεροληψία και αμεροληψία. Η γραμμή αντιπροσωπεύει τη μέση τιμή Hurst για κάθε περιοχή ποσοστών.

Το Διάγραμμα 4 δείχνει την ποσοστιαία διαφορά μεταξύ των τιμών βροχόπτωσης για την εν λόγω περίοδο επαναφοράς. Οι τιμές της βροχής, δείχνουν σαφή επιρροή της δομής εξάρτησης σε ακραία γεγονότα. Τα δεδομένα του ιστογράμματος απεικονίζουν ότι οι περισσότεροι σταθμοί δεν υφίστανται μεγάλη συνολική αλλαγή στην τελική τιμή, αλλά η αλλαγή αυτή συσχετίζεται στενά με την τιμή του συντελεστή Hurst.

Ενώ ο συντελεστής Hurst δεν είναι η μόνη παράμετρος που επηρεάζει την ποσοτικοποίηση της πραγματικής διαφοράς μεταξύ των τελικών τιμών του μοντέλου, είναι παρόλα αυτά η πιο ισχυρή. Όσο υψηλότερη είναι η ποσοστιαία αλλαγή, τόσο υψηλότερος είναι ο μέσος συντελεστής Hurst, άρα παρουσιάζουν υψηλή θετική συσχέτιση μεταξύ τους.



Εικόνα 2: Παγκόσμια κατανομή συντελεστή ουράς (*κ*)



Εικόνα 3: Παγκόσμια κατανομή συντελεστή κλίμακας (*λ*)

Τέλος, γίνεται προσπάθεια συσχέτισης των τιμών των παραμέτρων της κατανομής Pareto με τα κλιματικά χαρακτηριστικά της εκάστοτε περιοχής. Όσον αφορά τον συντελεστή ουράς (*κ*) παρατηρείται ότι εμφανίζει χαμηλές τιμές σε περιοχές που καταγράφεται συχνή και σημαντική βροχόπτωση καθόλη τη διάρκεια του χρόνου, όπως περιοχές κοντά στον Ισημερινό (π.χ. Βραζιλία, Ινδία, Μεξικό) (Εικόνα 2).

Ομοίως, για τον συντελεστή κλίμακας (*λ*) παρατηρείται ότι για περιοχές τροπικού κλίματος παίρνει μεγαλύτερες τιμές, εφόσον έχει θετική συσχέτιση με το μέγεθος του μέσου όρου βροχόπτωσης για κάθε σταθμό (Εικόνα 3)

Τα συμπεράσματα που προκύπτουν από την χρήση της μεθόδου των Κ – ροπών συγκριτικά με τις κλασικές μεθόδους είναι:

- Η μέθοδος των Κ – ροπών είναι αποτελεσματικότερη από τις κλασικές, προβλέποντας αξιόπιστα τα ακραία γεγονότα στις περισσότερες περιπτώσεις για υψηλές περιόδους επαναφοράς. Ωστόσο, δεδομένου ότι η διαδικασία προσαρμογής πραγματοποιείται με αλγόριθμο βελτιστοποίησης, εστίαση δίνεται στην καλύτερη προσαρμογή για ακραίες τιμές, έτσι υπάρχει ελαφρά απώλεια ακρίβειας στις αντίστοιχες χαμηλές, με τις κλασικές μεθόδους να δείχνουν ελάχιστα καλύτερη εφαρμογή.

- Η κατανομή Pareto-Burr-Feller, με την χρήση της επιπλέον παραμέτρου, διατηρεί την τέλεια προσαρμογή στην ουρά, ενώ παράλληλα τη βελτιώνει για χαμηλές περιόδους επαναφοράς, επιτυγχάνοντας γενικότερα την καλύτερη δυνατή προσαρμογή στα δεδομένα.

- Η μακροπρόθεσμη εμμονή έχει μεγάλο αντίκτυπο στα τελικά αποτελέσματα. Η ποσοστιαία διαφορά στις υψηλές περιόδους επαναφοράς είναι μη αμελητέα για σταθμούς με συντελεστή Hurst πάνω από 0.70. Χωρίς να συμπεριλαμβάνεται η μεροληψία λόγω εμμονής, οι ακραίες βροχοπτώσεις υποεκτιμώνται. Ταυτόχρονα, αποδεικνύεται η ισχυρή θετική συσχέτιση μεταξύ του συντελεστή Hurst και της διαφοράς βροχής για μεγάλες περιόδους επαναφοράς.

- Από την περαιτέρω διερεύνηση της συσχέτισης μεταξύ των κλιματολογικών χαρακτηριστικών και του δείκτη ουράς, κλίματα με σταθερά αυξημένες τιμές βροχόπτωσης, όπως το τροπικό (Ισημερινός), παράγουν κυρίως χαμηλές τιμές. Αντίθετα, οι σταθμοί που βρίσκονται σε άνυδρο ή μεσογειακό κλίμα, οι οποίοι λαμβάνουν κατά μέσο όρο χαμηλές βροχοπτώσεις με τις σπάνιες ακραίες τιμές να είναι σημαντικά υψηλότερες από το κανονικό, δείχνουν τις υψηλότερες τιμές του δείκτη μεταξύ όλων.

- Εφαρμόζοντας την ίδια διαδικασία στην εξεύρεση συσχετισμού μεταξύ της παραμέτρου κλίμακας και των κλιματολογικών χαρακτηριστικών μιας περιοχής, διαπιστώνεται ότι τα τροπικά και πλήρως υγρά εύκρατα κλίματα απεικονίζουν υψηλές τιμές, αντίθετα με τα ξηρά κλίματα και το χιόνι που συνδέονται με χαμηλές τιμές.

Table of Contents

Table of Figures

Table of Graphs

Table of Tables

1 Introduction

1.1 General Context

In an era where climate variability is becoming more and more significant, concerns about extreme weather conditions are at peak. Assessment of such extremes, especially when referring to extremes in hydrological processes, is crucial in a variety of tasks from engineering design of infrastructure projects to risk management.

Failing to model extremes reliably can lead to catastrophic consequences, depending on the magnitude of that failure. According to the COP21 Weather Disaster Report 2015 (CRED, 2015) floods are considered as the most prominent natural disaster, accounting for 43% of total disasters during the period between 1995 to 2015. Underestimation of extremes, is destined to lead to dam failures or insufficient flood mitigation and consequently pose threatening consequences to residential areas and human lives. On the other hand, overestimation, especially severe, can lead to financial losses and overbudgeting of a project, since more unnecessary resources will be used in construction and maintenance. Thus, the fabrication of a consistently reliable method for extreme-oriented rainfall modelling is deemed as paramount.

For achieving reliable long-term predictions, deterministic methods fail to produce credible results. Consequently, rainfall has to be treated as a random variable bound to a probability distribution function. Statistical moments are the fundamental tool used to express the important attributes of probability distributions of natural processes, and in this case rainfall.

Classical moments whilst having the advantage of being simple in their calculation, are proven to be efficient only in expressing attributes for orders up to 2 and in most large samples can't be estimated for orders higher than 3 (Lombardo, et al., 2014). However, extreme rainfall events are better modelled using high-order moments, since they are closely correlated with each other.

On the other hand, L – moments can be reliably estimated for high-orders if only the first moment is known. However, their most significant disadvantage is their inability to characterise and model dependence of stochastic processes. Long-term dependence bias, when not taken into account can lead to severe underestimation of extremes, especially for the higher return periods needed in designing and constructing engineering projects.

In order to overcome the issues portrayed by classic methods, newly introduced *knowable* moments (K – moments) (Koutsoyiannis, 2019) combine both methods' advantages and provide a sound basis for reliably estimating high-order moments and statistical characteristics of marginal and joint distributions, whilst retaining precision in low-orders. Moreover, they create a reliable framework for estimation of long-term dependence important for any study of natural processes.

1.2 Scope of Work

The main objective of this study is to achieve in providing a general framework for extreme-oriented rainfall modelling using the newly introduced *knowable* moments (K – moments) method. Classic methods are also used for the modelling process in order to compare and assess the prediction power of all three. Since this is a global study, an established precipitation database of stations from around the world will be used, namely the GHCN – Daily database from NOAA (NOAA, 2019). By using global data, the study aims in proving the reliability and consistency of K – moments for every regions' climatic characteristics.

All methods estimate the parameters of a specified probability distribution function and are compared to each other for their efficiency in fitting such distribution to observed data. Distributions from the Pareto family are the most optimal and concurrent with the rainfall process (Papalexiou, et al., 2013). While comparative results are being shown for the whole distribution (body and tail), in this study focus is mostly given in the fitting power for extreme values. Extremes are considered values for return periods higher than 1 year. Thus, more comprehensive analysis is done for such extremes, which in statistical terms, are located in the distribution tail. All comparisons are being made using goodness-of-fit statistics between observed and theoretically modelled data.

After establishing the fitting advantages of the knowable moments' method, it is obligatory to assess the influence of long-term dependence bias existing in most rainfall samples. Knowable moments provide the framework for estimating this bias. Consequently, the effects of long-term dependence are estimated and infused in the fitting process and are being compared to an otherwise general sample independence in order to show the magnitude this bias holds in the final predicted rainfall values.

Finally, with consistent modelling results for most worldwide stations, prospect exists in analysing them by the distribution of parameters across the globe and finding correlation between them and each region's climatic characteristics. This is done while using the aforementioned K – moment approach with the dependence structure of each station, if present, taken into account. In this regard, a general framework of expected distribution's parameter values can be established for future reference.

1.3 Work Structure

This study is split into eight (8) chapters all with their distinct value.

The **first chapter** gives a general overview over rainfall extremes, while also highlighting the goals this study aims to achieve and what the analysis will showcase.

The **second chapter** provides a bibliographic analysis of the differences between weather and climate, while also defining extreme rainfall and showcasing the importance of acquiring a reliable model for modelling such extremes.

The **third chapter** is an extensive analysis of all probabilistic and stochastic theory used in the modelling process with increased focus on describing the notion of knowable moments. Moreover, usage of Pareto distributions for best description of the rainfall process is

justified combined with the hydrometeorological significance of using the whole dataset in the fitting process.

The **fourth chapter** provides information about the chosen database containing global precipitation data. Next, a general distribution of all its provided stations is shown along with whichever remain from the elimination process.

In the **fifth chapter** emphasis is given on the computer-based modelling tools that were used. Specifically, the MATLAB and Python programming environments are analysed along with any of their toolboxes used for specific applications in this study. Also, their advantages in respect to other programming languages are also explained for justifying their use case.

The **sixth chapter** is an analytic description of the modelling procedure. It is analysed in the context of a step by step guideline, with deeper analysis on the K – moments method. Moreover, details on the initial file processing are given along with definitions of all the goodness-of-fit statistic tools used for later evaluation purposes.

In the **seventh chapter** an application of all methods is presented in a specific station. Every method is analysed and compared to each other for its modelling power and consistency, especially in the extremes, by providing comparative figures and goodness-of-fit values. This assessment is being made for the overall fit, as well as for the distribution's body and tail fit, separately. Moreover, a depiction of the long-term dependence bias effect is provided for showing the importance of accounting for a sample's dependence structure. Also, an alternative distribution for fitting with K – moments is showcased which further improves the fitting result.

The **eighth chapter** contains comparative results from the generalization of the process followed in modelling of the aforementioned sample station. The process is now universally applied to the entirety of the database and overall results from the extreme-oriented fitting process are being produced, showcasing every method's predicting power. Conclusions on the effectiveness of knowable moments are drawn and with them an extensive analysis on the influence of long-term dependence when fitting stations with high estimated dependence bias. Finally, the distribution's parameters are assessed for their correlation with each region's climatic characteristics, in order to draw conclusions on representative parameter values for different climate zones.

The **ninth chapter** is a review of the whole research focusing on its preliminary objectives and the conclusions drawn from the resulting extreme-oriented rainfall modelling procedure, evaluating knowable moments for their overall effectiveness. Finally, perspectives on future research along the lines of this study are also given.

2 Theoretical Analysis

2.1 Weather and Climate

Starting from the top, the difference between the concepts of regional "weather" and "climate" has to be clarified. Both notions are closely related and since they are required in understanding the proceedings of this study they should be explained before moving to deeper analysis.

Weather, in its simplest form is defined as the way the atmosphere behaves in respect to day-to-day effects on human activities (American Meteorological Society, 2015).

Climate, on the other hand is in short, the description of the long-term patterns of weather in a specific area of interest. Therefore, as climate is a long-term characteristic of an area, climate zones have been established through the Köppen Climate Classification (Figure 2.1) describing the average climatic features on any place on Earth (Rubel & Kottek, 2010). A more established definition by the IPCC is provided below:

"Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or millions of years. The classical period for averaging these variables is 30 years, as defined by the World Meteorological Organization. The relevant quantities are most often surface variables such as temperature, precipitation, and wind. Climate in a wider sense is the state, including a statistical description, of the climate system." (IPCC, 2014)

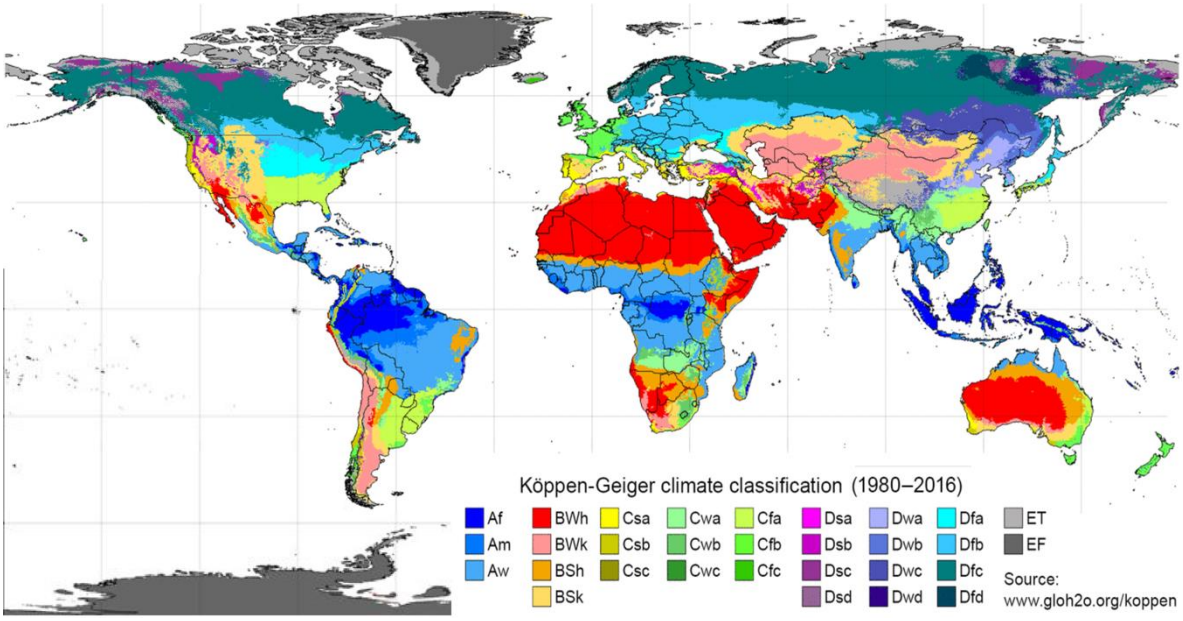


Figure 2.1: Köppen climate zones classification (1980-2016) (Beck, et al., 2018)

2.2 Weather Forecasting

The importance of knowing and anticipating weather conditions spanning a short period of time in the future was acknowledged even before the modern era. For centuries, even millennia, people have been trying to forecast weather. Ancient Greeks such as Aristotle and Theophrastus described weather patterns in *Meteorologica* and *Book of Signs*, respectively. The Babylonians predicted weather from astrology signs and cloud patterns, while the Chinese are assumed to have been attempting to predict the weather since 300 BC. Their methods largely relied on recognizing specific patterns of events and most of them don't prove to have reliable outcomes by todays' standards.

In recent times, advances in the deeper comprehension of atmospheric physics followed by technological innovations in the 20th century led to the founding of Numerical Weather Prediction. Its practical use started in 1955 with the emergence of programmable electronic computers (Wikipedia, 2019)

The core principle behind numerical weather prediction is sampling the fluid state at a specific time and with the use of fluid dynamics and thermodynamics, construct a model that estimates the fluid state in the near or far future. Inputs of this system are real-time observable quantitative weather data such as precipitation, temperature, and barometric pressure. A week-long rainfall prediction output of a model, is shown below in Figure 2.2.

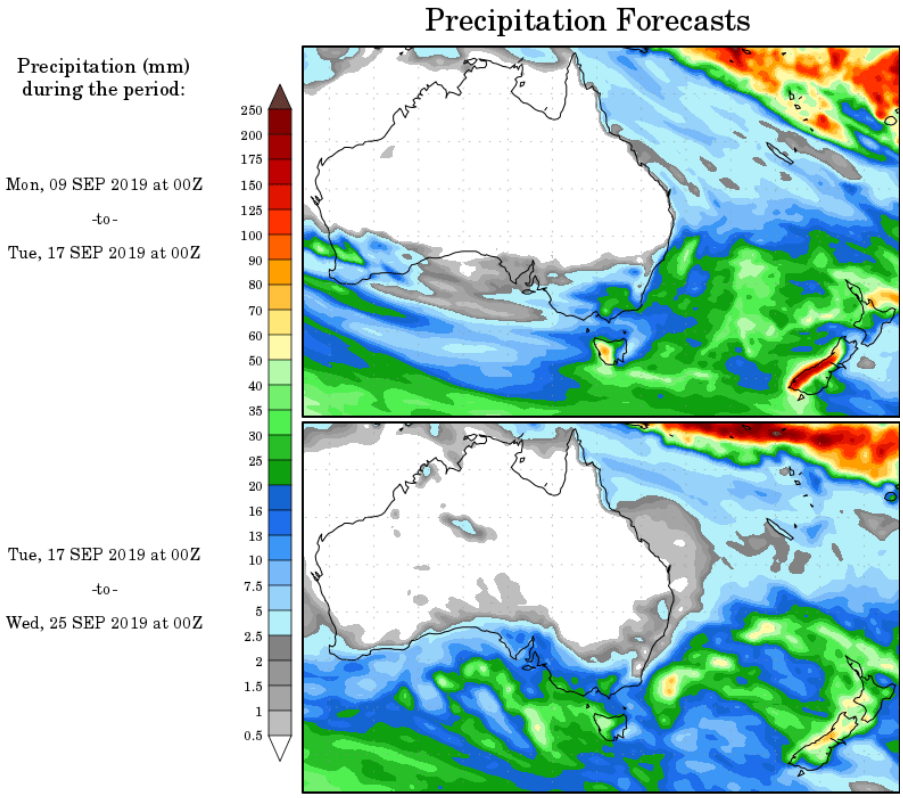


Figure 2.2: 7-day rainfall forecast in Australia and New Zealand (NOAA, 2019)

Despite the successful implementation of those prediction systems, the chaotic nature of weather cannot be ignored. Minute errors in the initial conditions of a model grow quickly, hindering the forecasting power of the model, while similarly, errors in approximating the simulation of atmospheric processes leads to limited predictability.

2.3 Precipitation

Based on the American Meteorological Society (American Meteorological Society, 2015) precipitation is defined as:

"All liquid or solid phase aqueous particles that originate in the atmosphere and fall to the earth's surface."

The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail. In theory, precipitation occurs when air temperature falls below the dew point, which refers to the temperature to which a parcel of air has to be cooled in order to become saturated, and condense into water. Raindrops have dimensions ranging from 0.1 millimetres to 9 millimetres mean diameter, above which they tend to separate into smaller sizes (Wikipedia, 2019).

Throughout history, long-term annual averages of precipitation have been fluctuating ever so slightly. Although this is the case, expectations on rainfall patterns are still consequent to an area's climate characteristics throughout the year i.e. areas near the Equator receive heavier rainfall annually than areas with temperate climate, such as Europe and North America. In Figure 2.3 this consistency of average rainfall values in different regions of the world is showcased.

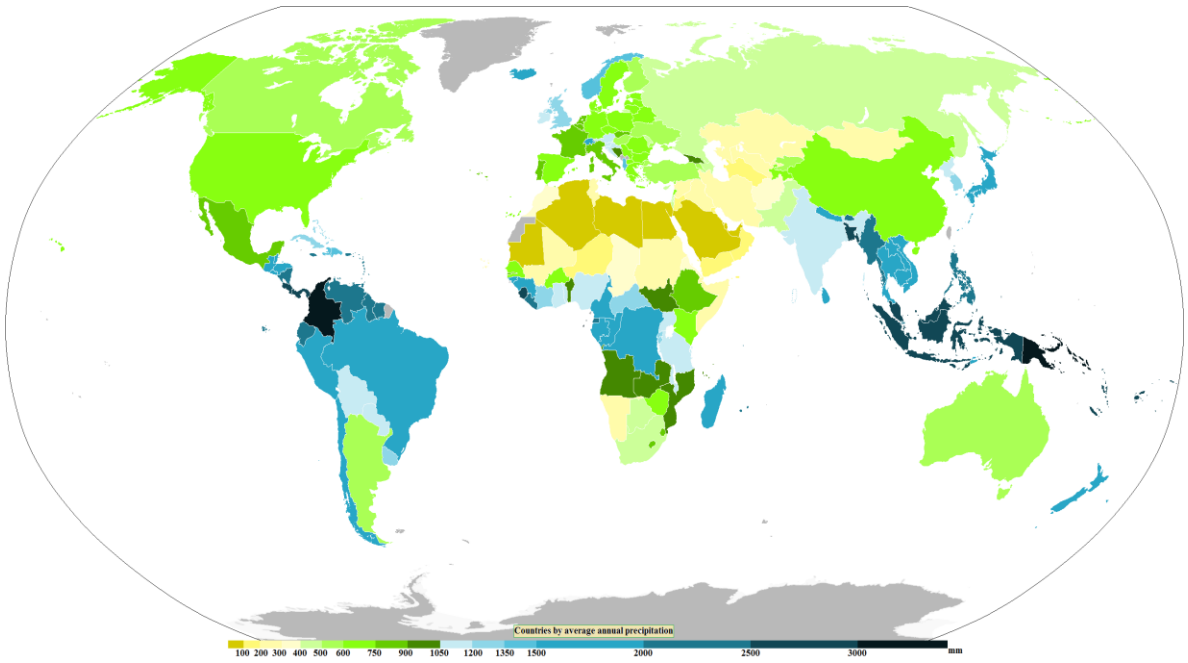


Figure 2.3: Average annual precipitation (mm) by country (Wikipedia, 2019)

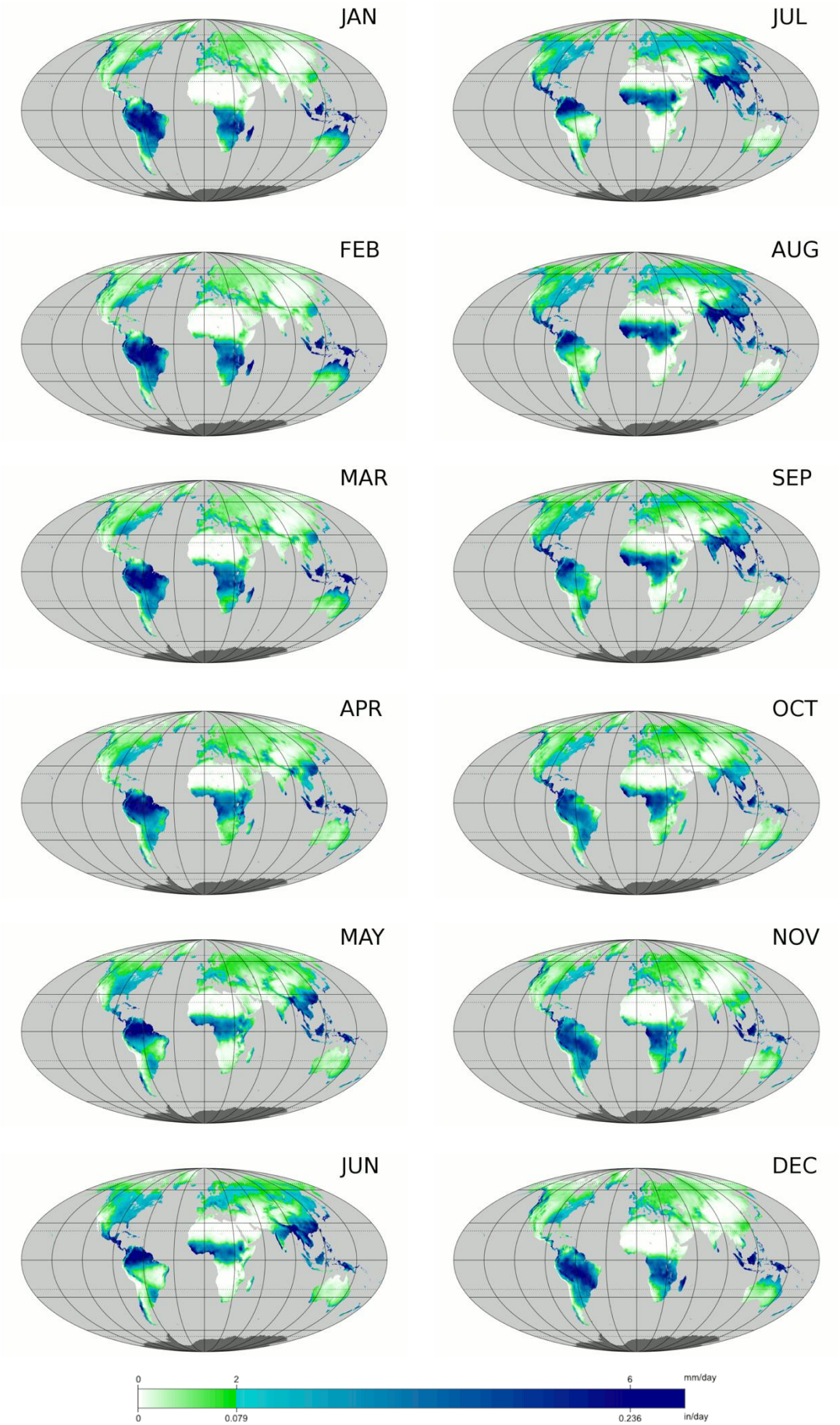


Figure 2.4: Average global monthly precipitation patterns (mm/d) (Wikipedia, 2019)

Focusing on short-term annual averages of any natural process, might fabricate the notion that there is no statistically significant change between years. However, applying that same study on long-term annual averages spanning more than 50 years, it is evident that even on the annual scale, there exists evident variability. Also referred to as periodicity or cyclostationarity, this characteristic suggests that rainfall or any other natural process cannot be considered as a random variable, but rather should be attributed a stochastic nature. This seasonal periodicity is also evident in smaller time scales such as months, as seen from Figure 2.4

Given the importance of weather forecasting and the extent of applications in human activities, achieving accuracy in measurements has become a science in its own. There exist numerous weather measuring devices ranging from ground weather stations and radars to unmanned aircrafts and satellite atmospheric imagery. The most common are shown in Figure 2.5.

In order to obtain a reliable set of data, most of the time there is crossover and validation between techniques for each region. Ground monitoring is used in conjunction with aerial monitoring, were one is used as the primary measurement tool and the other as a validation method.

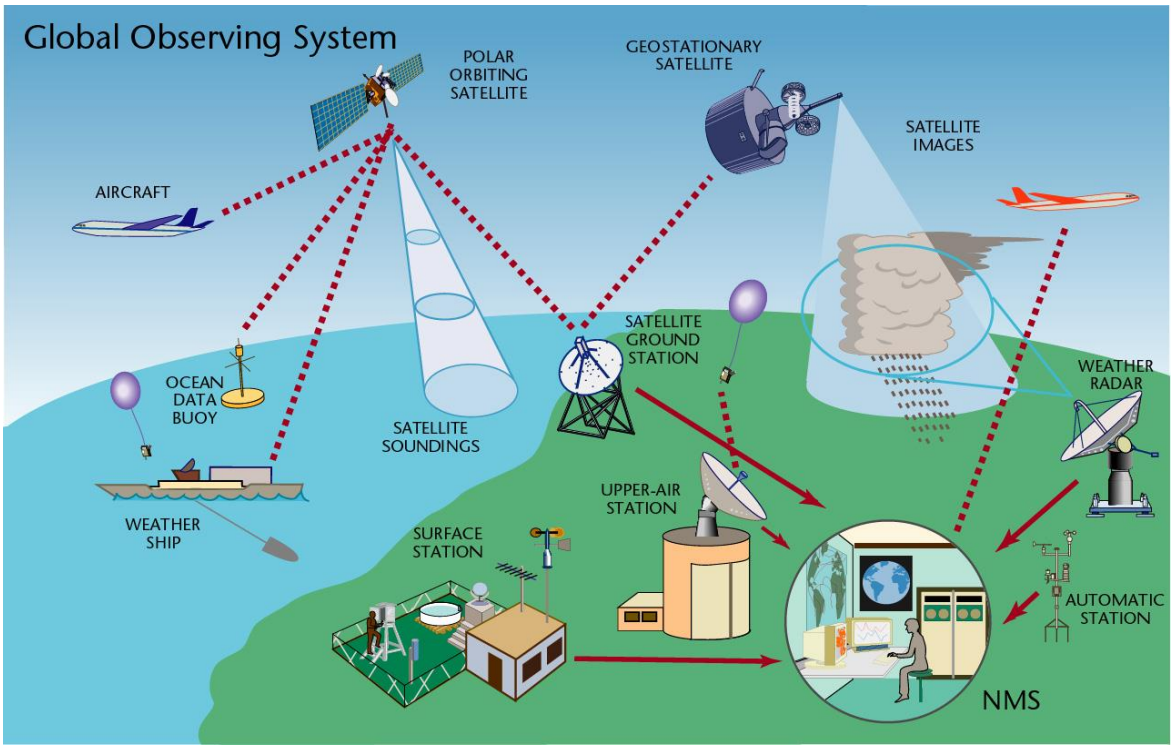


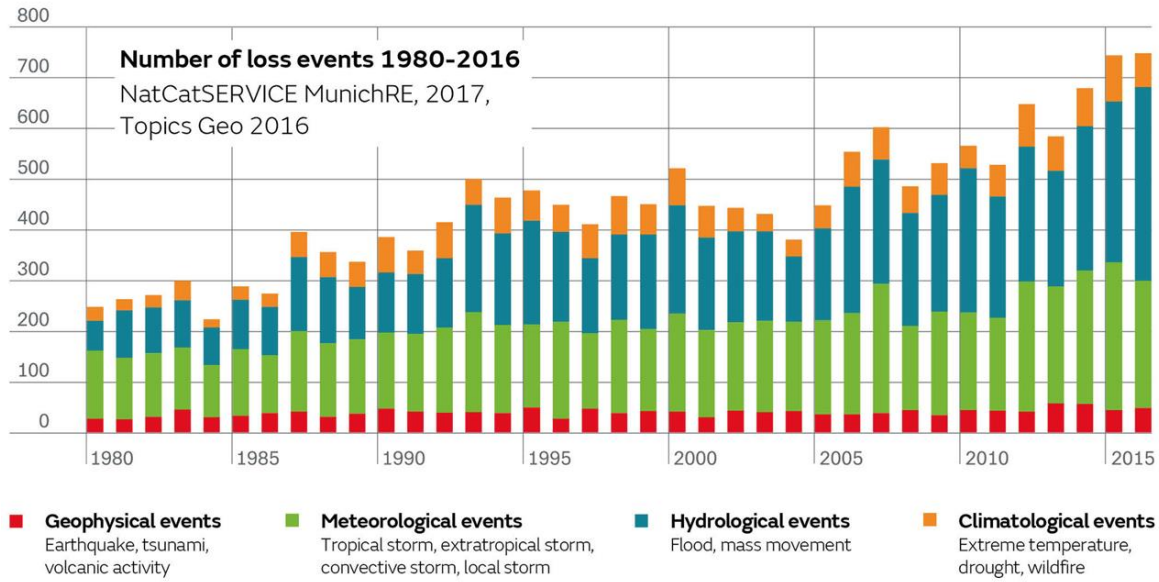
Figure 2.5: Most common weather collection methods (World Meteorological Organization, 2016)

2.4 Extreme Precipitation

As is with any natural process, extreme weather is a natural part of the Earth's climate system. Extreme precipitation events should be expected, either by long lasting droughts, or by severe rainfall occurrences. Nonetheless, these extreme events have significant impact on everyday human life, infrastructure, as well as on the environment. Assuming the climate is not changing, these events sustain an annual constant frequency and thus are expected and dealt with efficiency and resilience, not being intrusive and disastrous to humanity and the environment.

However, in an era where climate variability is becoming more and more significant, concerns about extreme weather conditions are at peak. According to the Intergovernmental Panel on Climate Change (IPCC, 2012) the uppermost attention must be given in reliably predicting extremes of any kind of natural process. This report also showcased some research showing that models assessing past events produced results which hinted at a slight increase of extreme natural events.

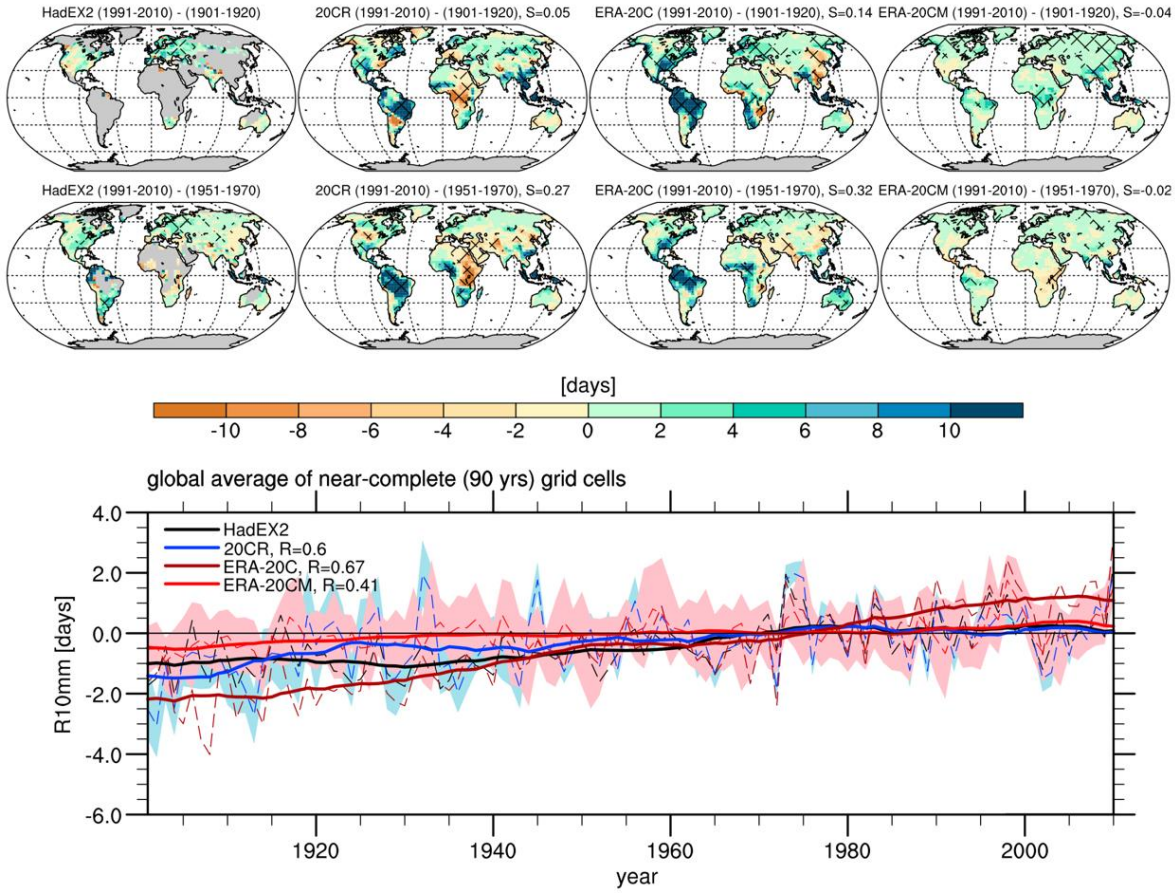
More specifically, damages to property and the environment (loss events) attributed to hydrological extremes, show continuous increasing occurrence from the 1990s until today, with their frequency caused by floods and mass movements more than tripling during this timeframe (Graph 2.1). While a superficial look of this graph shows correlation between floods numbers and loss events it is important to take note of the increased land areas now used for housing and industrial infrastructure which ultimately might be the reason for the increase in loss events. Either way, the resulting fact is the same.



Graph 2.1: Extreme weather loss events' occurrences from 1980 to 2016 (Met Office UK, 2017)

Donat, et al. (2016) studied different models which assessed the fluctuation of heavy rainfall events and their intensity through the past century. Findings showed probability that:

- **Northern Europe and central Eurasia** → slight increase over the past century
- **Eastern North & South America** → slight increase since the 1950s
- **Eastern Africa** → slight decrease over the past century
- **Tropical Africa** → slight increase over the past century
- **Southeast Asia & Indonesia** → increase over the past century



Graph 2.2: Multiple models output on global fluctuation of heavy rainfall days [R10mm] from 1901 to 2010 (Donat, et al., 2016)

These findings (Graph 2.2) show that on global average, extremes have fluctuated through the past century with some areas hinting at increases and others at decreases. Whichever the case, monitoring and analysing them is an important component of assessing the climate system, since it is vital to know how their characteristics are evolving, and will change in the future, in order to facilitate appropriate adaptation.

2.5 Importance of Extreme Rainfall Modelling

Extreme precipitation events, even though in later years technology provides the means of predicting them with more certainty, can hardly be stopped from disrupting human activities as well as damaging the environment.

The most significant hazard from extreme rainfall are floods. According to the Organization for Economic Cooperation and Development floods cause annually $40 billion in damages, both on residential areas and infrastructure (CRED, 2015). Since 1995, floods make up 43% of all weather-related natural disasters, affecting 2.3 billion people in total. In conjunction to human property destruction, agriculture losses are mostly liable to floods, meaning that essential crop production is undermined, producing losses to the financial sector as well.

From an engineering point of view, studying the overall distribution of rainfall over time in a specific region is vital for evaluating the amount of water available for meeting the demands of industry, agriculture, or other human activities. However, accuracy in the prediction of extreme events is also important, since they are being used in the design and construction of projects that are destined for water management purposes, such as dams, flood mitigation works, and hydroelectric power plants. Underestimating extremes, is bound to lead to dam failures or insufficient flood mitigation, placing at risk residential areas and human lives. On the other hand, overestimation leads to financial losses and overbudgeting, since more unnecessary resources will be used in construction and maintenance.

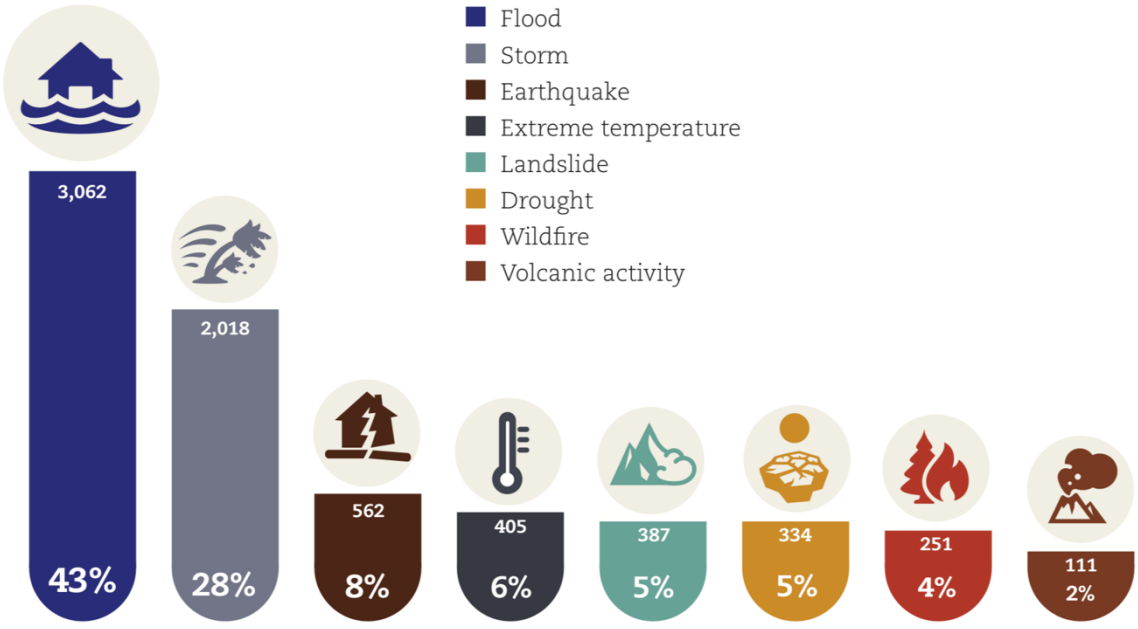


Figure 2.6: Extreme events percentage by type (CRED, 2015)

3 Stochastic Framework Analysis

3.1 Modelling Process

While today's technology standards and meteorological models allow the prediction of short-term precipitation events using the deterministic approach, long-term predictions are not possible considering deterministic methods. In this regard, rainfall has to be treated as a random variable that follows a specified probability distribution function, which is the mediator of the all-important assignment of return periods to rainfall values. The selection of this distribution can be generally summarised in four steps (Papalexiou, et al., 2013):

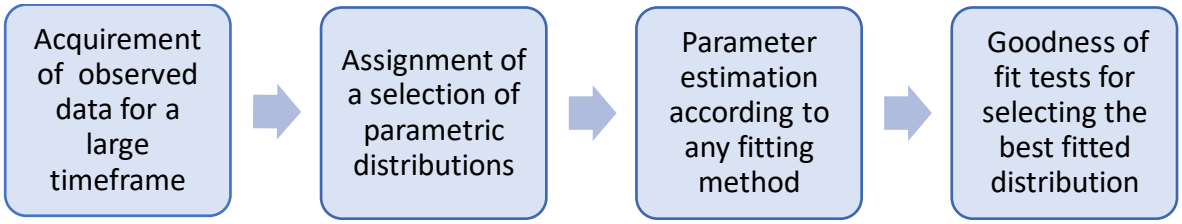


Figure 3.1: General procedure on probability distribution function selection

3.2 Heavy and Light Tailed Distributions

One way of classifying distributions, is by the nature of the asymptotic behaviour of their tails. The tail of a distribution is responsible for the magnitude and frequency of extreme values, thus distinguishing distributions by this factor is an essential starting stage. Distribution tails are one of two kinds depending on their relation to the behaviour of an exponential tail (Teugels, 1975):

A. Heavy Tailed (Sub exponential class) → Referring to distributions which converge to zero slower than an exponential tail.
B. Light Tailed (Hyper exponential class) → Referring to distributions which converge to zero faster than an exponential tail.

Mathematically the definition of a heavy tailed distribution is given by:

$$\lim_{x\to\infty} \frac{1-F(x)}{\exp\left(-\frac{x}{\beta}\right)} = \infty\,, \qquad \forall\,\beta > 0 \tag{3.1}$$

The norm when modelling rainfall is to apply a light-tailed distribution model (e.g. Gamma distribution) and fitting to the whole sample of observed data. The typical procedure of applying a distribution law to rainfall is a provides the best fit on the whole spectrum of observations and does not guarantee efficiency when trying to model for the extremes. As extremes are located on the tail-end of the distribution and usually only a fraction of the empirical data is also located there, all traditional light-tailed methods are biased against extreme values.

Furthermore, the distinct characteristic of heavy tails is that they predict more frequent larger magnitude rainfall occurrences compared to light tails. Consequently, when using a light-tailed model, there is great risk of underestimating extreme events putting human lives at risk.

3.3 Distribution Function for Rainfall Modelling

Based on the findings by Papalexiou, et al. (2013) heavy tailed distributions are more suited in describing the long-term characteristics of rainfall and especially its extremes. Thus, heavy tailed distributions are used for modelling and more specifically the Generalized Pareto Distribution (GPD) and the Pareto-Burr-Feller (PBF).

3.3.1 Generalized Pareto Distribution

The classic Pareto distribution is a power-law probability distribution used extensively in many observable natural phenomena, as well as in socioeconomic research. It was originally applied by Vilfredo Pareto to model the distribution of wealth among a society, and nowadays is most known and associated by the famous Pareto principle or the "80-20 rule".

The Generalized Pareto Distribution, after the contribution of Pickands (1975), has since been used extensively in many sectors of research. Some of its applications cover analysis of extreme events or modelling of large insurance claims (Hosking & Wallis, 1987). It consists of a family of continuous probability distributions, stated by originally three parameters: tail index κ, scale λ (or b), and location ψ. In this study:

A. Though, using all three parameters will result in greater overall accuracy, the location parameter is set to $\psi = 0$, in order to be naturally consistent with the rainfall process's zero lower bound.

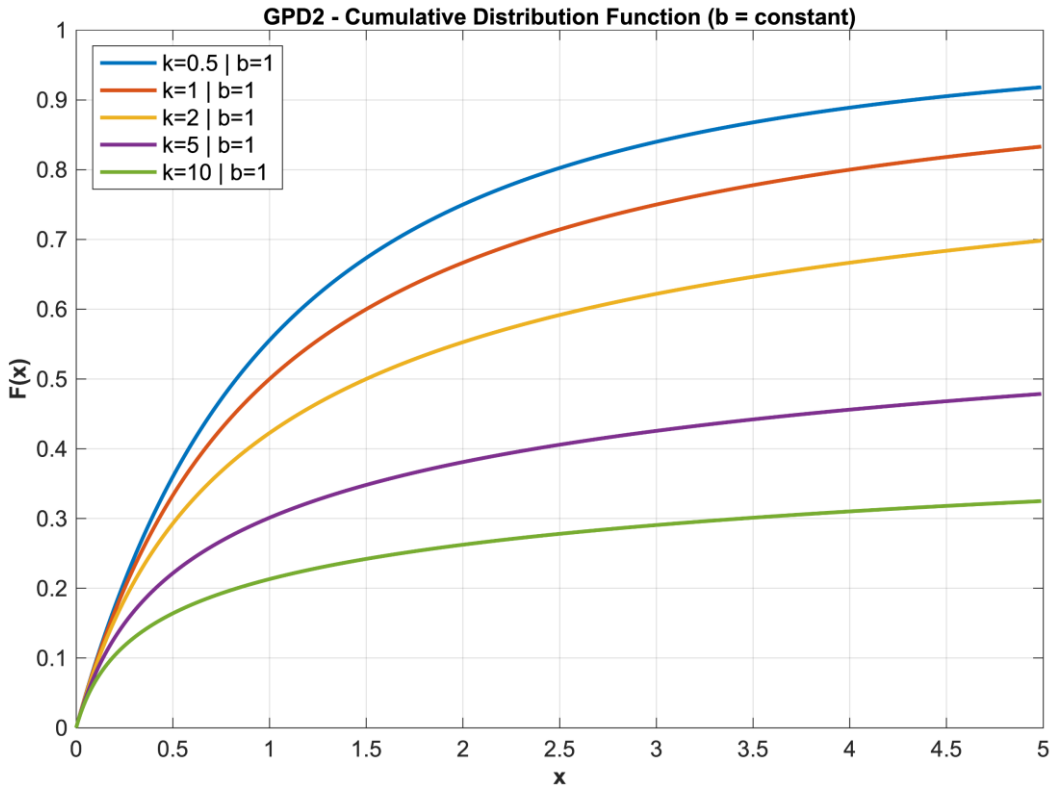
B. For $\kappa = 0$ the Pareto distribution specializes into the exponential distribution

C. For $\kappa < 0$ the tail converges faster to zero, thus it is a light tailed distribution and not suitable for this study's modelling purposes. Especially in this case, negative tail index gives the distribution an upper bound which contradicts the rainfall process.

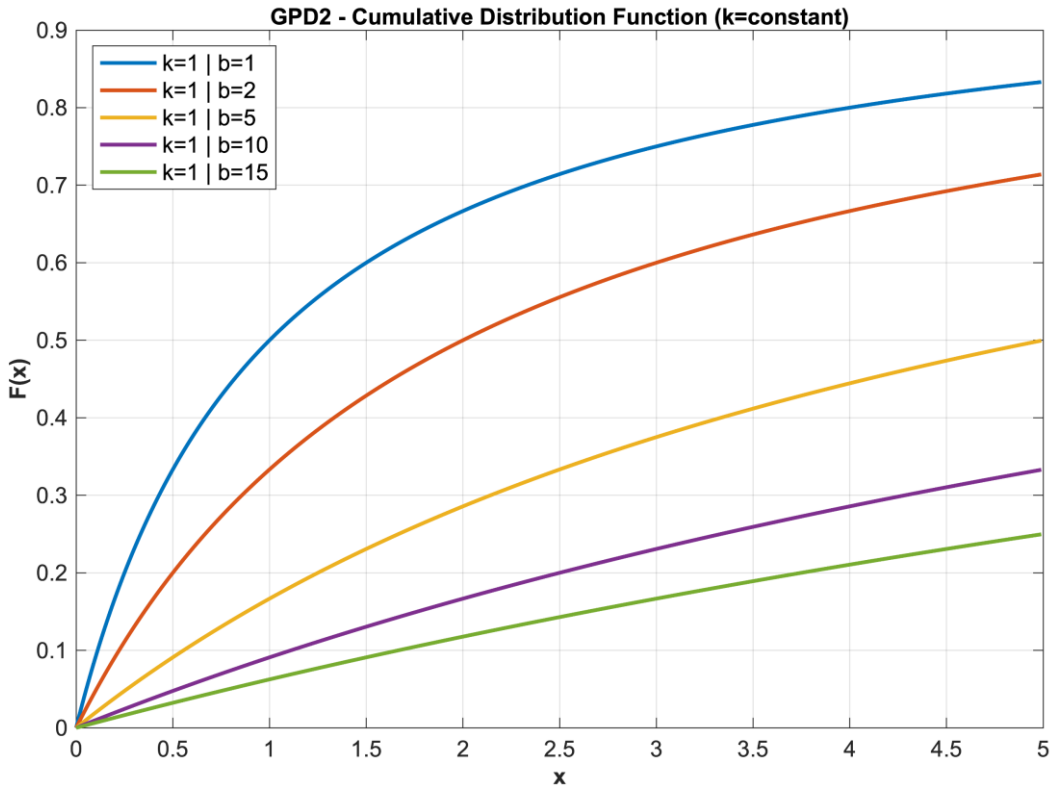
Thus, the two parameter Generalized Pareto Distribution's (GPD2) probability and cumulative distribution functions are given below. In Graph 3.1 and Graph 3.2 the behaviour of the GPD2 with changing tail index and scale parameter is showcased.

$$F(x) = \begin{cases} 1 - \left(1 + \kappa \frac{x}{\lambda}\right)^{-\frac{1}{\kappa}}, & \kappa \neq 0 \\ 1 - \exp\left(-\frac{x}{\lambda}\right), & \kappa = 0 \end{cases} \tag{3.2}$$

$$f(x) = \begin{cases} \frac{1}{\lambda}\left(1 + \kappa \frac{x}{\lambda}\right)^{-\frac{1}{\kappa - 1}}, & \kappa \neq 0 \\ \frac{1}{\lambda}\exp\left(-\frac{x}{\lambda}\right), & \kappa = 0 \end{cases} \tag{3.3}$$



Graph 3.1: Generalized Pareto Distribution (GPD2) for different tail index *κ*



Graph 3.2: Generalized Pareto Distribution (GPD2) for different scale parameter *λ* or *b*

3.3.2 Pareto-Burr-Feller Distribution

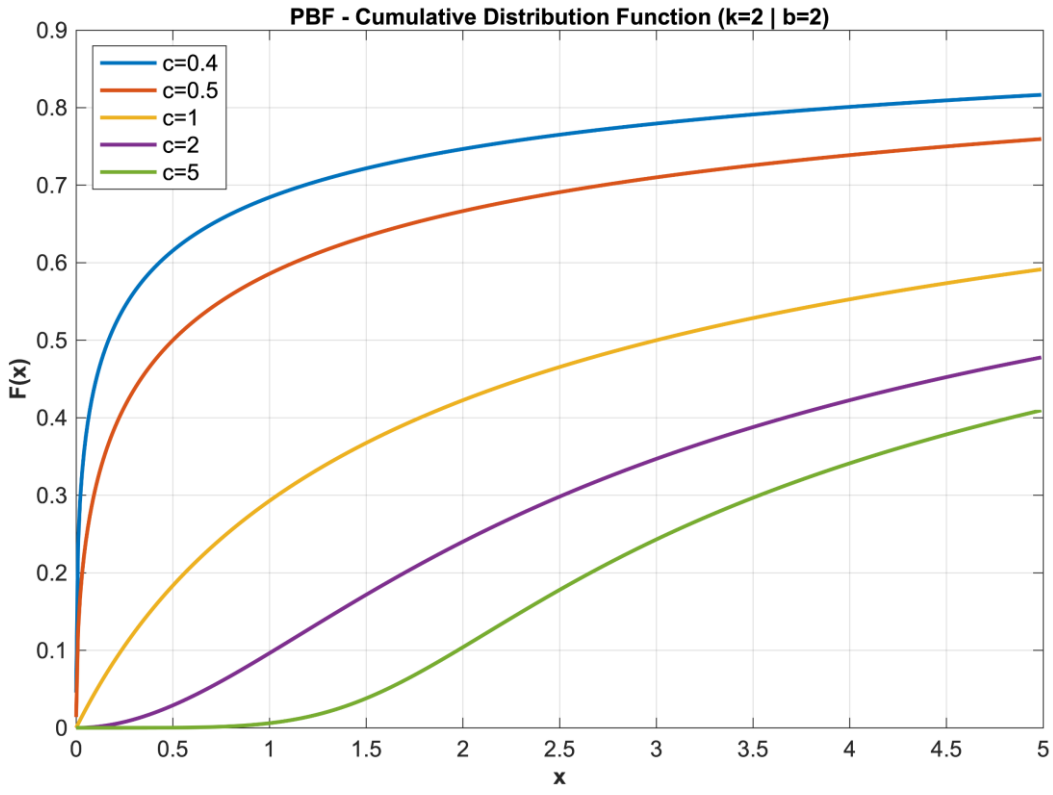
In similarity to the Generalized Pareto Distribution, the Pareto-Burr-Feller (PBF) is a heavy tailed power-law probability distribution with three parameters. It is a distribution used mostly in econometrics (Singh & Maddala, 1976), and is more commonly named the Pareto IV or the Burr Type XII. The derivation of the PBF was studied by Burr (1942) and given mathematical justification from Feller (1970) who linked it to the Beta function and distribution. Its usefulness in a variety of fields is shown in Brouers (2015).

In this study, it is used in conjunction with the GPD2 in modelling extreme rainfall, considering its two different asymptotic properties, that of a Weibull distribution for low precipitation values, and that of a Pareto distribution in the tail. Furthermore, the addition of a third parameter of the GPD2 may prove advantageous in the accuracy of the final model. The importance of these properties will be displayed subsequently. The Pareto-Burr-Feller is defined as:

$$F(x) = 1 - \left(1 + \kappa c \left(\frac{x}{\lambda}\right)^{c}\right)^{-\frac{1}{ck}} \tag{3.4}$$

With a probability density function:

$$f(x) = \frac{c\left(\frac{x}{\lambda}\right)^{c-1}}{\lambda}\left(1 + \kappa c \left(\frac{x}{\lambda}\right)^{c}\right)^{-\frac{1}{ck-1}} \tag{3.5}$$



Graph 3.3: Pareto-Burr-Feller cumulative distribution for different *c*

3.4 Definitions of Moments / Estimators

In statistics, the expected value is the foundation of producing moments. Moments are quantitative measures that portray the shape and characteristics of a distribution function. If *X* is a random variable and *g(X)* is a function of *X*, then the expectation of *g(X)* is given by:

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx \tag{3.6}$$

Or for a discreet random variable *X*:

$$E[g(X)] = \sum_{i=1}^{+\infty} g(x_i)P(X = x_i) \tag{3.7}$$

From the theoretical expected value, moments can be defined as:

A. Non-central Moment:

$$m_x^{(r)} := E[X^r] \tag{3.8}$$

B. Central Moment:

$$\mu_x^{(r)} := E[(X - m_x)^r] \tag{3.9}$$

In hydrology and most natural sciences, the moments and central moments up to the fourth order are consistently used to describe characteristics in distributions. Their practical estimators are presented in Table 3.1.

Table 3.1: First four order estimators of classical moments

Order (r)	Estimator	Characteristic
1	$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (3.10)	Mean
2	$Var = \sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ (3.11)	Variance
3	$s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\sigma^3}$ (3.12)	Skewness
4	$k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\sigma^4}$ (3.13)	Kurtosis

Where:

- n is the size of the sample
- σ is the standard deviation and,
- $\overline{x}$ is the first non-central moment, or the mean

3.5 L – moments / Estimators

Comparably with classic moments, L – moments are statistic tools aiming to describe the shape and characteristics of a probability distribution. For a random variable *X*, the r^{th} order L – moment is given by (Hosking, 1990):

$$\lambda_r = r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} E_{r-k:r} \tag{3.14}$$

Where:

- $X_{k:n}$ denotes the kth smallest value (order statistic) from an independent sample of *n* observations from the *X* distribution and,
- E is the expected value

The first four population L – moments are:

$$\lambda_1 = EX \tag{3.15}$$

$$\lambda_2 = (EX_{2:2} - EX_{1:2})/2 \tag{3.16}$$

$$\lambda_3 = (EX_{3:3} - 2EX_{2:3} + EX_{1:3})/3 \tag{3.17}$$

$$\lambda_4 = (EX_{4:4} - 3EX_{3:4} + 3EX_{2:4} - EX_{1:4})/4 \tag{3.18}$$

L – moments can be derived from Probability Weighted Moments (PWM) first discovered by Greenwood, et al., 1979. They are connected by the probability weighted average:

$$\beta_r = n^{-1} \sum_{i=1}^{n} \left(1 - p_{j:n}\right)^r x_{j:n} \tag{3.19}$$

And in the unbiased form:

$$b_r = n^{-1} \sum_{i=1}^{n} \frac{(n-j)(n-j-1)\dots(n-j-r+1)}{(n-1)(n-2)\dots(n-r)} x_{j:n} \tag{3.20}$$

Where:

- $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$ is the ordered sample based on order statistics
- n is the sample size
- $p_{j:n} = \frac{j+\gamma}{n+\delta}$, where γ and δ are suitable constants. Based on the findings of Landwehr, et al., 1979 for the Wakeby distribution (of which the Pareto is a special case), the recommended values are: $\gamma = -0.35$ and $\delta = 0$.

In practice the estimators of the first four L – moments, based on PWM, are provided below:

Table 3.2: First four order estimators of L - moments

Order (r)	Estimator
1	$\lambda_1 = \beta_0$ (3.21)
2	$\lambda_2 = 2\beta_1 - \beta_0$ (3.22)
3	$\lambda_3 = 6\beta_2 - 6\beta_1 + \beta_0$ (3.23)
4	$\lambda_4 = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0$ (3.24)

The use of L – moments instead of classical ones proves to have some major advantages. Firstly, because of their linear nature in estimation, they have higher robustness in dealing with outliers in a sample, meaning less sensitivity in extreme values. Moreover, their existence is dependent only in a finite sample mean, thus higher order L – moments exist, even if classic ones don't.

In order to obtain statistical data for the shape of the distribution using L – moments the following coefficients can be used:

A. L – moment mean → the equivalent of the sample mean.

$$\tau_1 = \lambda_1 \quad (3.25)$$

B. L – moment coefficient of variation → equivalent to the standard coefficient of variance, showing in percentage the variation of values without accounting for the sample mean.

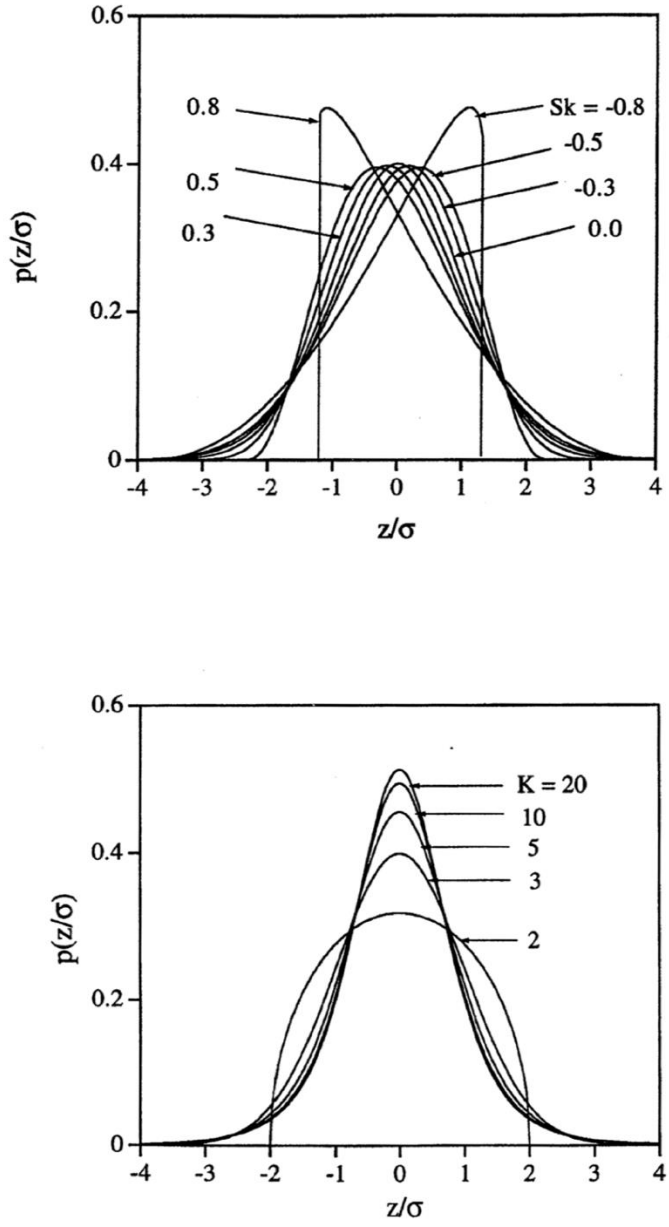
$$\tau_2 = \frac{\lambda_2}{\lambda_1} \quad (3.26)$$

C. L – moment coefficient of skew → equivalent to the standard skewness coefficient, showing the symmetry of the distribution with respect to its mean and median.

$$\tau_3 = \frac{\lambda_3}{\lambda_2} \tag{3.27}$$

D. L – moment of Kurtosis → equivalent to the standard kurtosis measure, showing the density of values around the mean (sharpness of the top).

$$\tau_4 = \frac{\lambda_4}{\lambda_2} \tag{3.28}$$



Graph 3.4: Skewness and kurtosis coefficients values and representation

3.6 Order Statistics

If *X* is a stochastic variable and $x_1, x_2, \dots, x_n$ are copies of it, independent and identically distributed they form a sample. By rearranging them in ascending order of magnitude order statistics are formed:

$$x_{(1:n)} \leq x_{(2:n)} \leq \cdots \leq x_{(n:n)} \tag{3.29}$$

The minimum and the maximum of the ordered sample are special cases of order statistics and are defined as:

$$\min\{\mathrm{X}\} = \mathrm{x}_{(1:\mathrm{n})} \tag{3.30}$$

$$\max\{\mathrm{X}\} = \mathrm{x}_{(\mathrm{n}:\mathrm{n})} \tag{3.31}$$

Order statistics can be a useful tool for stochastics since they take into account both the magnitude and the relative position to other observations. Furthermore, it is important to note that from all ordered samples there can arise efficient estimators. Order statistics are usually used by many modelling methodologies and could prove as a valuable tool in extremes modelling too, especially on the assignment of return periods to sample values, which are examined in the next chapter.

3.7 Sample Return Period

Assigning return periods to sample values is crucial in stochastic modelling of extremes. The concept of the return period is crucial in the designing and risk assessment of most engineering works, providing with the means of evaluating the frequency of extreme events (Volpi, et al., 2015).

In probability terms, the return period is inversely related to the probability of exceedance of a specific value of a variable (e.g. precipitation). Another definition from probability theory indicates that for a specific event A, which is a subset of some certain event Ω, return period *T* is defined as the mean time between consecutive occurrences of event A. This notion is not deterministic by any case and simply suggests that the time between consecutive occurrences of event A is a stochastic variable with *T* as the mean (Koutsoyiannis, 2019).

With the use of order statistics, sample return periods are assigned to precipitation values based on the *Weibull plotting position* by Weibull (1939).

$$\frac{T_{(i:n)}}{D} = \frac{n+1}{n+1-i} \tag{3.32}$$

With maximum return period at the highest value of the ordered sample:

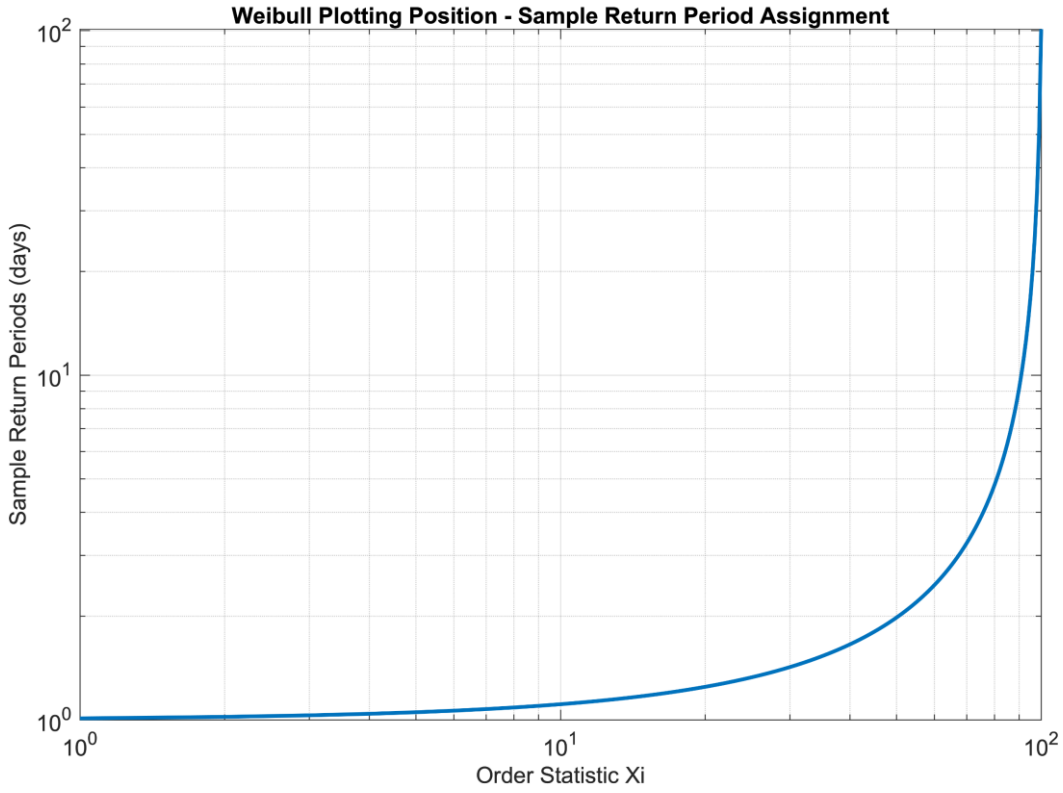
$$\frac{T_{(n:n)}}{D} = n+1 \tag{3.33}$$

A variety of methods for assigning sample return periods exist, but in this case the *Weibull plotting position* method is used since:

A. It's implementation in very simple and sufficient for the modelling process

B. It isn't susceptible to distribution function changes

C. It is unbiased for $F(\underline{x}_{(i:n)})$.



Graph 3.5: Weibull plotting position - sample return periods for sample size of *n*=100

3.8 K – moments

The use of statistical moments offers the ability of describing probability distributions with reasonable simplicity (Feller, 1968). When analysing an observable process among different time scales, they stand as the basic tool for stochastic characterization of change and variability, both important features when studying natural processes. Nonetheless, both classical and L – moments, the two basic methods of characterising a distribution, have disadvantages.

Classical moments, central or non-central, cannot be reliably estimated from large samples for orders beyond two or three (Lombardo, et al., 2014). As examined in Koutsoyiannis (2019) for high orders (*p*) the standard moment estimator portrays an estimator of an extreme quantity and converges considerably slowly to the theoretical value (Equation 3.34). This combined with the fact that most geophysical and hydrological processes don't follow the normal distribution, means that two moment statistics aren't enough to characterise their distributions reliably.

$$\hat{\mu}'_p = \frac{1}{n}\sum_{i=1}^{n} x_i^p \approx \frac{1}{n}\left(\max_{1\le i\le n}(x_i)\right)^p \tag{3.34}$$

On the other hand, L – moments, as mentioned before, exist even if only the first order classical moment is finite. However, because of their fundamental linearity, they are all first order in terms of the process of interest. Their most significant disadvantage is their inability to characterise and model dependence of stochastic processes. Dependence is an important characteristic of most geophysical processes and will be defined later in the study.

Extremes are located in the tail-end of a distribution function, thus are closely correlated with high-order moments. Consequently, using classical moments to model extremes (i.e. in rainfall), proves to be efficient only for low-order of moments. Thus, the newly introduced *knowable* moments (K – moments) (Koutsoyiannis, 2019) will be used in the modelling process, as they provide better grounds for prediction based on high orders, whilst retaining precision of classical moments for low orders.

K – moments, combine the advantages of using classical or L – moments, allowing reliable estimation and description of high-order statistics, imperative for marginal and joint distributions of stochastic processes, whilst also providing the framework for estimation of long-term dependence.

3.8.1 Definitions of K – moments

With the use of order statistics, if $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ are copies of stochastic variable that form a sample, then the expected maximum of order *p* of *x* (i.e. the expected value of $\underline{x}_{(p)}$) defines a statistical moment:

$$K'_{p1} := E\left[\max\left(\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p\right)\right] = pE\left[\left(F\left(\underline{x}\right)\right)^{p-1}\underline{x}\right], \qquad p \ge 1 \tag{3.35}$$

Non-central knowable moments of order (p, q) form with the generalisation of equation (3.35) and are defined as:

$$K'_{pq} := (p-q+1)E\left[\left(F\left(\underline{x}\right)\right)^{p-q}\underline{x}^q\right], \qquad p \ge q \tag{3.36}$$

In the same manner, *central knowable* moments of order (p, q) are defined as:

$$K_{pq} := (p-q+1)E\left[\left(F\left(\underline{x}\right)\right)^{p-q}\left(\underline{x}-\mu\right)^q\right], \qquad p \ge q \tag{3.37}$$

Finally, *hypercentral knowable* moments of order (p, q) are defined as:

$$K^+_{pq} := (p-q+1)E\left[\left(2F\left(\underline{x}\right)-1\right)^{p-q}\left(\underline{x}-\mu\right)^q\right], \qquad p \ge q \tag{3.38}$$

From equation, it is clear that K – moments are by definition connected to maxima. This holds true for all other definitions of K – moments, both central and hypercentral, providing with the means of a reliable estimation of expected extreme values.

3.8.2 Biased Estimators of K – moments

In order to estimate K – moments, the values of $(F(x))^{p-q}$ and $(2F(x)-1)^{p-q}$ have to be known. Their quantities can be estimated if order statistics are involved, thus by arranging the sample $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_p$ in ascending order.

$$\underline{x}_{(1:n)} \leq \underline{x}_{(2:n)} \leq \cdots \leq \underline{x}_{(n:n)} \tag{3.39}$$

It is worth noting that the sample is arranged in order of $\underline{x}$ and not $\underline{x}^q$, thus making the estimation more reliable. In this regard, $F\left(x_{(i:n)}\right)$ and $2F\left(x_{(i:n)}\right)-1$ are estimated as:

$$F\left(x_{(i)}\right) = \frac{i-1}{n-1} \tag{3.40}$$

With a value range of $0 \leq F\left(x_{(i)}\right) \leq 1$. In the same manner:

$$2F\left(x_{(i)}\right) - 1 = \frac{2i-n-1}{n-1} \tag{3.41}$$

With a value range of $-1 \leq 2F\left(x_{(i)}\right) \leq 1$.

By the use of equations (3.40) and (3.41) , K – moments estimators become:

$$\underline{K}'_{pq} = \frac{p-q+1}{n} \sum_{i=1}^{n} \left(\frac{i-1}{n-1}\right)^{p-q} \underline{x}^{q}_{(i:n)} \tag{3.42}$$

$$\underline{K}_{pq} = \frac{p-q+1}{n} \sum_{i=1}^{n} \left(\frac{i-1}{n-1}\right)^{p-q} \left(\underline{x}_{(i:n)} - \mu\right)^{q} \tag{3.43}$$

$$\underline{K}^{+}_{pq} = \frac{p-q+1}{n} \sum_{i=1}^{n} \left(\frac{2i-n-1}{n-1}\right)^{p-q} \left(\underline{x}_{(i:n)} - \mu\right)^{q} \tag{3.44}$$

3.8.3 Unbiased Estimators of K – moments

Deriving the estimators of $F(x_{(i:n)})$ and $2F(x_{(i:n)}) - 1$ from simple uses of order statistics yields biased results, especially for high orders of *p*. Thus, the estimators are biased and insufficient for high orders. Unbiased estimators can be produced by denoting the non-stochastic variable $\left(\frac{p}{n}\right)\left(F(x_{(i:n)})\right)^{p-1}$, which depends only on the values of *i, n* and *p*, as b_{inp}. Thus, the estimators of the *non-central* and *central* K - moments becomes:

$$K'_{pq} = \sum_{i=1}^{n} b_{i,n,p-q+1} \underline{x}^{q}_{(i:n)} \tag{3.45}$$

$$K'_{p1} = \sum_{i=1}^{n} b_{i,n,p} \left(\underline{x}_{(i:n)} - \mu\right) \tag{3.46}$$

Which in turn becomes unbiased if:

$$b_{inp} = \begin{cases} 0, & i < p \\ \dfrac{p}{n} \dfrac{\Gamma(n-p+1)}{\Gamma(n)} \dfrac{\Gamma(i)}{\Gamma(i-p+1)}, & i \geq p \geq 0 \end{cases} \tag{3.47}$$

Where:

- Γ is the gamma function and *p* defines the moment order and can be any positive number, usually an integer, but that's not necessary.
- For the central K – moment, complete unbiasedness is achieved only for $q = 1$, but that is sufficient for this study's purposes.
- The unbiasedness of the estimator can be easily substantiated by the fact that:

$$\sum_{i=1}^{n} b_{inp} = 1 \tag{3.48}$$

The significance of b_{inp} in extremes modelling stands in the fact that as moment order *p* increases, lesser data from the sample determine the final K - moment estimate. This is due to $b_{inp} = 0$ for $i < p$. This means that for high *p*, more emphasis is given in higher sample values, than in lower ones, which in turn paves the way for accurate high order moment estimation with minimal computing power and better precision in extremes modelling.

3.8.4 Statistical Significance and Relation to Other Moments

Using specific combinations of *p* and *q*, K – moments provide the basis for estimating basic statistical characteristics otherwise produced by the classic method of moments or L – moments. Consequently, both classic methods can be derived from special cases of K – moments, meaning that they can be fully replaced. Classic moments can be derived as:

$$K'_{pp} \equiv \mu'_{p}, \qquad K_{pp} \equiv \mu_{p} \tag{3.49}$$

While L – moments are derived using their direct relationship with Probability Weighted Moments (PWM), as the *non-central* K – moment with $q = 1$. They are defined as:

$$K'_{p1} = p\beta_{p-1} \tag{3.50}$$

With the use of equations (3.47) and (3.48), a table of the customary and more statistically useful moments can be created:

Table 3.3: K - moments relationship to classic moments

Order (p)	Relationship	Characteristic
1	$K'_{11} = \mu$	Mean
2	$K^{+}_{22} = K_{22} = \mu_2 = \sigma^2$	Variance
3	$\frac{K_{33}}{K_{22}^{3/2}} = \frac{\mu_3}{\sigma^3}$	Skewness (Dimensionless)
4	$\frac{K_{44}}{K_{22}^{2}} = \frac{\mu_4}{\sigma^4}$	Kurtosis (Dimensionless)

Table 3.4: K - moments relationship to L - moments

Order (p)	Relationship	Characteristic
1	$K'_{11} = \lambda_1$	Mean
2	$K^{+}_{21} = 2K_{21} = 2(K'_{21} - \mu) = 2\lambda_2$	Variance
3	$\frac{K^{+}_{31}}{K^{+}_{21}} = 2\frac{K_{31}}{K_{21}} - 3 = \frac{\lambda_3}{\lambda_2}$	Skewness (Dimensionless)
4	$\frac{K^{+}_{41}}{K^{+}_{21}} = 4\frac{K_{41}}{K_{21}} - 8\frac{K_{31}}{K_{21}} + 6 = 0.8\frac{\lambda_4}{\lambda_2} + 1.2$	Kurtosis (Dimensionless)

3.8.5 Return Periods of K – moments

Assigning return periods at an observed sample is closely related to the use of order statistics. With the same mindset, since K – moments are constructed upon the theoretical properties of order statistics, it is evident that they can be assigned return periods as well. The general rule applied for non-central K – moments with $q = 1$ is of the form (Koutsoyiannis, 2019):

$$\frac{T\left(K'_{p1}\right)}{D} = \Lambda_p p \tag{3.51}$$

Where:

- *D* is a time reference for the specific return period and,
- Λ_p is a coefficient dependent on moment order *p* and the distribution function associated with the specific sample

Solving for the Λ_p coefficient with the theoretical definition of a return period and time reference of $D = 1$:

$$\Lambda_p := \frac{1}{p\left(1 - F\left(K'_{p1}\right)\right)} \tag{3.52}$$

In order to determine the variation of the return period between different moment orders, firstly an exact relationship between *p* and Λ_p should be constructed. This relationship can be extracted by first estimating the lower Λ_1 and upper Λ_∞ boundaries of Λ_p. Since Λ_p is also dependent on the distribution function, in this study focus is given on defining the coefficient for the Pareto (GPD2) and the Pareto-Burr-Feller distribution (PBF), which are the ones used for modelling extremes as mentioned earlier.

For the Pareto distribution:

$$T(x) = \left(1 + \kappa \frac{x}{\lambda}\right)^{\frac{1}{\kappa}} \tag{3.53}$$

For the Pareto-Burr-Feller:

$$T(x) = \left(1 + \kappa c \left(\frac{x}{\lambda}\right)^{c}\right)^{\frac{1}{c\kappa}} \tag{3.54}$$

For the different distributions, Λ_1 and Λ_∞ values can be calculated respectively as:

For the Pareto distribution:

$$\Lambda_1 = \left(\frac{1}{1 - \kappa}\right)^{\frac{1}{\kappa}} \tag{3.55}$$

$$\Lambda_\infty = \left(\Gamma(1-\kappa)\right)^{1-\kappa} \tag{3.56}$$

For the Pareto-Burr-Feller distribution:

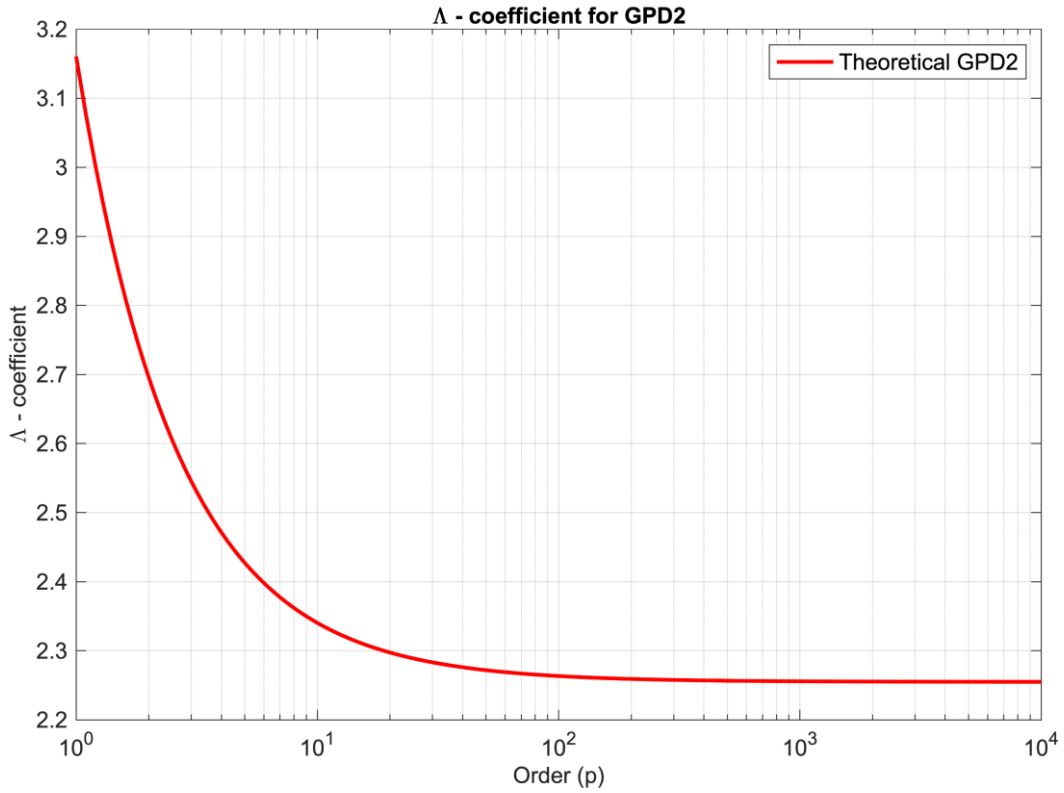
$$\Lambda_1 = \left(1 + \left(\frac{B\left(\frac{1}{\kappa c} - \frac{1}{c}, \frac{1}{c}\right)}{c}\right)^c\right)^{\frac{1}{\kappa c}} \tag{3.57}$$

$$\Lambda_\infty = \left(\Gamma(1-\kappa)\right)^{1-\kappa} \tag{3.58}$$

With the boundaries known, an approximate relationship of Λ_p and p can be formulated below. Despite the fact, that this approximate relationship can be applied reliably for a number of distributions, in the case of the Pareto distribution, an exact theoretical relationship exists too. This exact relationship is the one used in this study.

$$\Lambda_p \approx \Lambda_\infty + (\Lambda_1 - \Lambda_\infty)\frac{1}{p} \tag{3.59}$$

$$\Lambda_p = \frac{\left((p+1-\kappa)B(1-\kappa, p+1)\right)^{\frac{1}{\kappa}}}{p} \tag{3.60}$$



Graph 3.6: Λ - coefficients for the GPD2 using the theoretical relationship from Equation (3.60)

The nature of Λ – coefficients enhances them with important properties which arise from a direct observation of their immediate definition and relation with return periods. As shown in Graph 3.6, the coefficient varies in a narrow range for the Pareto distribution, this also being the case for many other common distributions. This allows for reliable assessment of the whole series by just two estimates (i.e. $\Lambda_1, \Lambda_\infty$), which can be approximated by generic functions, independent of the distribution function for which they are needed.

In conclusion, by assigning empirical return periods to K – moments and using them instead of the standard practice used in classic order statistics, the modelling procedure profits with some significant advantages.

A. In terms of expected values and uncertainty, both methods are identical.

B. With the classic method of order statistics, one can assign return periods only to values in the sample thus only *n* values of return periods. However, with K – moments as the return period is dependent on the moment order *p*, which can be assigned any value up to the size of the sample, one can empirically produce them for any quantile up to the size of the sample.

C. Because of the most accurate estimation formulas of assigning empirical return periods to K – moments, compared to return periods with order statistics, the former method is bound to be more accurate or at least equivalent to the later.

D. In the classic approach, each return period is dependent only on one sample value. With the K – moments method, each return period is a weighted average of several observations. Consequently, this enhances overall accuracy in the estimation.

3.8.6 Climacogram / Persistence and Long-term Dependence / HK Behaviour

The theoretical concept of persistence or long-term dependence, which exists in most natural processes including rainfall, was discovered by the works of H.E. Hurst (1951) who was studying the long-term capacities of reservoirs. Before that, A. Kolmogorov (1941) gave mathematical significance to this concept while analysing turbulence characteristics. Nowadays, it is recognised as the Hurst phenomenon or Hurst-Kolmogorov (HK) behaviour and is quantified by the Hurst coefficient *H*. In order to calculate the Hurst coefficient, the most accurate method is by formulating the climacogram (Dimitriadis & Koutsoyiannis, 2015), which is defined as the plot of variance of an averaged process versus averaged time scale. The Hurst coefficient is equal to half the slope of the climacogram plus 1 in a log-log plot. Based on its value it is assumed that:

- $0 \leq H < 0.5$ → the process is antipersistent (or anticorrelated) and it is not common in natural processes.

- $H = 0.5$ → the process is equivalent to white noise, meaning that there is no long-term change (dependence) or persistence in the sample.

- $0.5 < H \leq 1$ → the process has enhanced long-term persistence (or positively correlated), which is the most common behaviour on hydroclimatic processes.

Before estimating the Hurst parameter, the climacogram should be constructed. The theoretical definition of the climacogram for a stochastic process is given in the equation below:

$$\gamma(k) := \mathrm{var}\left[\frac{x(k)}{k}\right] \tag{3.61}$$

To construct it (Graph 3.7) the subsequent procedure has to be followed:

A. A range of time scales is created, ranging from 1 to 1/10 of the sample size *n*.

B. For each time scale an average of consecutive items in the time series is made. For example, for scale two (2):

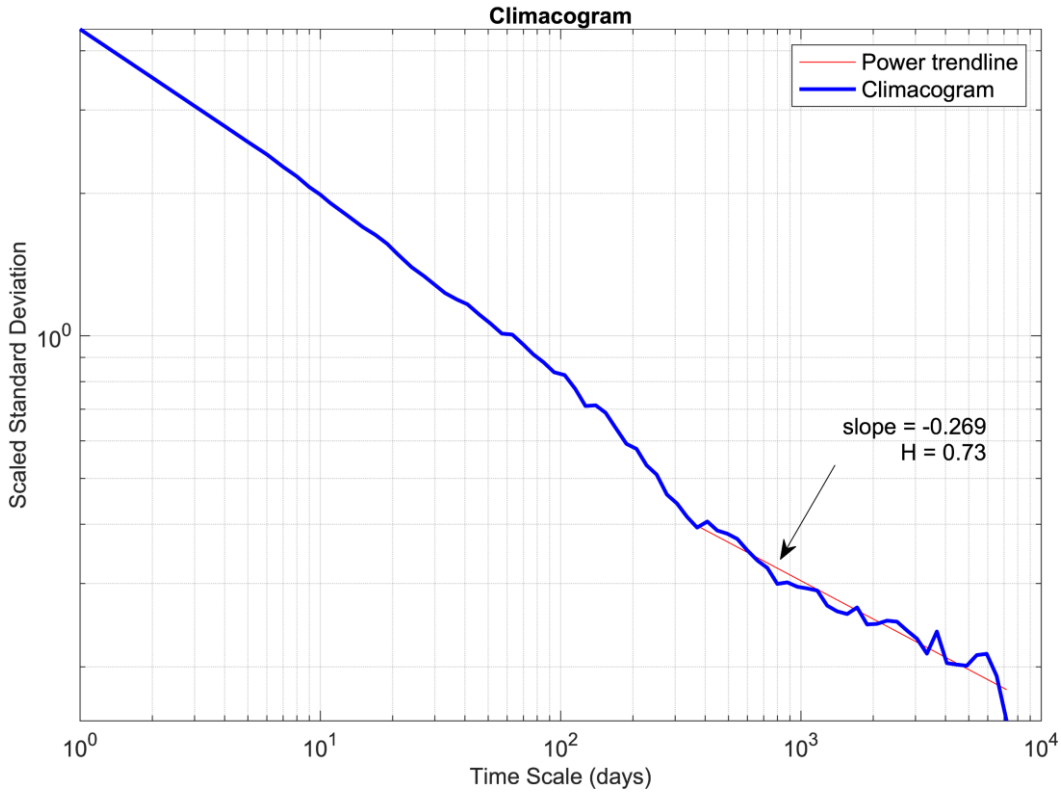
$$x^{(2)}{}_1 := \frac{x_1 + x_2}{2}, x^{(2)}{}_2 := \frac{x_3 + x_4}{2}, \dots, x^{(2)}{}_{n/2} := \frac{x_{n-1} + x_n}{2} \tag{3.62}$$

C. For each constructed averaged time series, the variance is calculated. For the same example for scale two (2):

$$\hat{\gamma}(2) := \frac{(x_1 - \mu)^2 + \cdots + \left(x_{n/2} - \mu\right)^2}{2} \tag{3.63}$$

D. By plotting the variances and time scales in a log-log plot, the climacogram is built and the Hurst parameter is estimated from the slope (*s*) in high time scales:

$$H = 1 + s \tag{3.64}$$



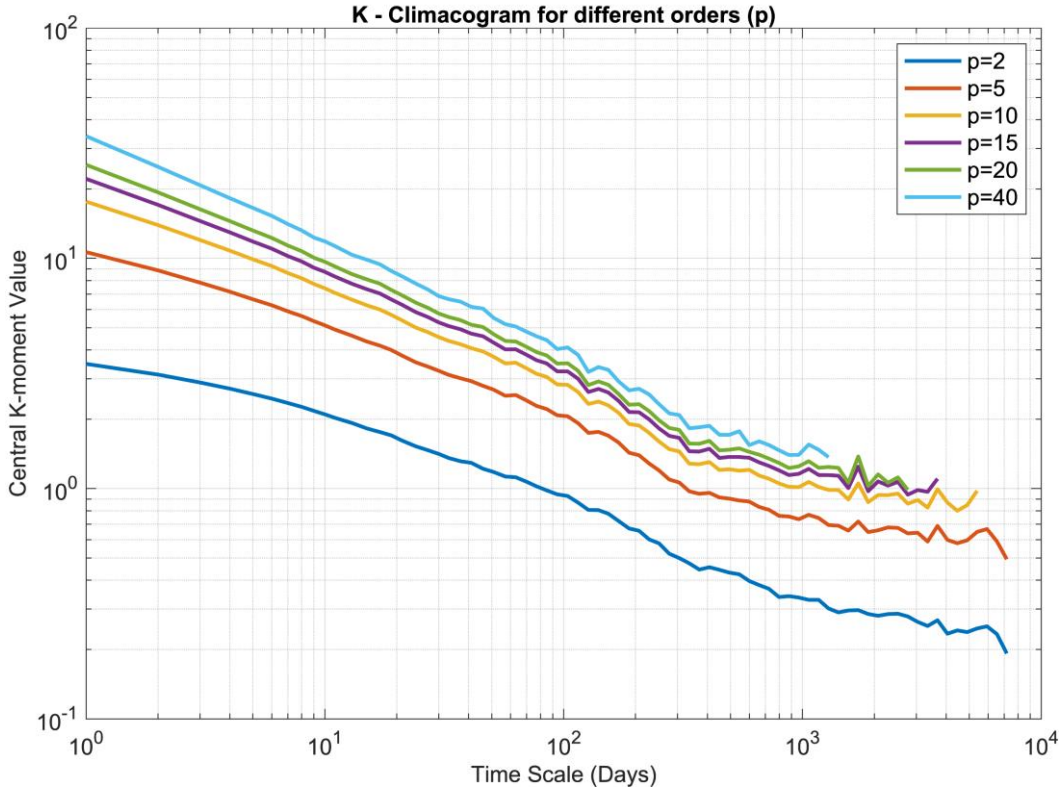
Graph 3.7: Climacogram for station "NLE00100503" with 55708 total observations and Hurst estimation.

3.8.7 K – climacogram

Long-term dependence or change, is considered a second-order characteristic of a stochastic process. In order to determine characteristics that are of higher order the standard method of using the covariance function equation, requires many variables whose estimation is difficult. This is due to the fact that using classic methods for estimating moments for orders higher than two or three is proven inaccurate. To overcome this, the K – climacogram is defined (Koutsoyiannis, 2019), using the standard climacogram idea and expanding it with the use of hypercentral K – moments.

$$\gamma_{pq}(k) := (p-q+1)E\left[\left(2F\left(\frac{\underline{x}(k)}{k}\right)-1\right)^{p-q}\left(\frac{\underline{x}(k)}{k}-\mu\right)^{q}\right] \tag{3.65}$$

The K – climacogram is versatile in the description of high-order statistics. In this study, the K – climacogram is used in the same manner as the standard climacogram, which is to investigate long-term dependence in the rainfall time series. Like standard methods, by using the K – climacogram with $q = 1$ and $p = 2$ the Hurst parameter is calculated as in equation (3.64). An interesting characteristic is that for different orders of *p,* the plots are similar and parallel to each other (Graph 3.8). However, the statistical significance of this, if any, is not part of this study.



Graph 3.8: K - climacograms for different orders (*p*) and *q* = 1 for station "NLE00100503"

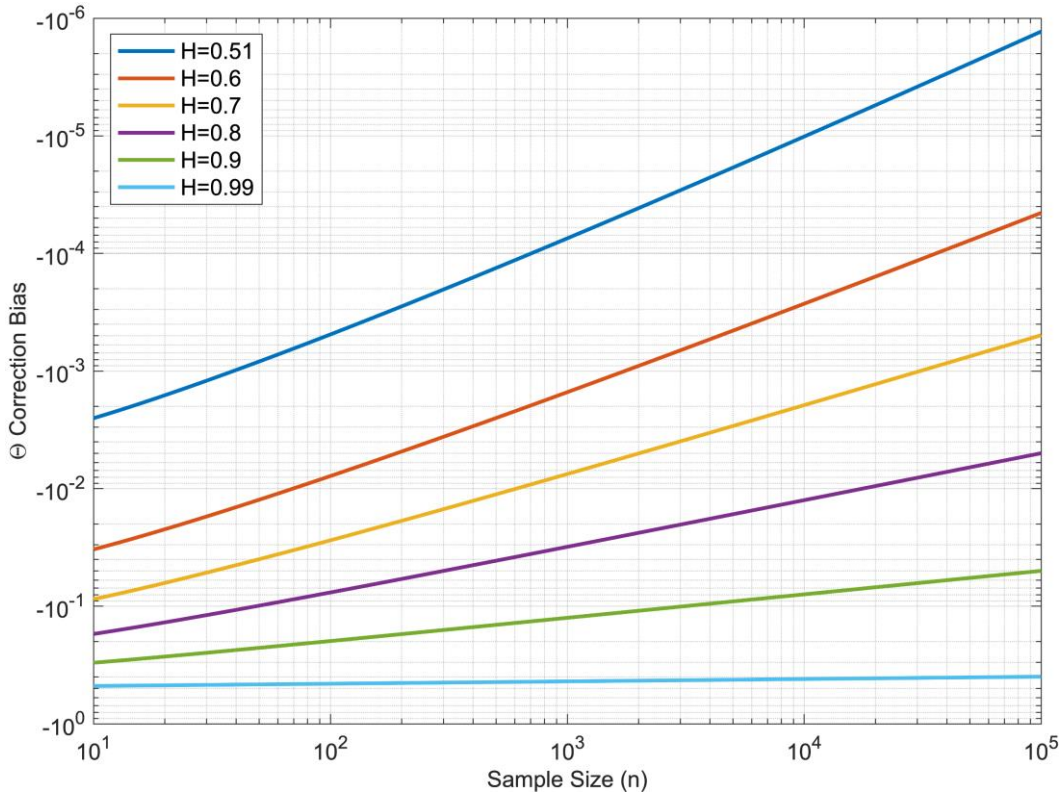
3.8.8 Long-term Dependence Bias in K – moments

Unlike the theoretical K – moment definition which is a first order characteristic of a marginal distribution thus dependence is not a differentiating factor, K – moments estimators like in Equation (3.45), contain long-term persistence bias. This bias can be estimated and removed using the procedure outlined in Koutsoyiannis (2019). Depending on the process structure the bias estimator takes different forms. For a Markov process with autocorrelation function $r(t) = r^t$ and sample size *n*:

$$\theta^M(n,r) \approx \frac{2r}{(1-r)(n-1)} \tag{3.66}$$

Alike, for an HK process with theoretical autocorrelation function $r(t) \approx H(2H-1)t^{2H-2}$:

$$\theta^{HK}(n,H) \approx \frac{2H(1-H)}{n-1} - \frac{1}{2(n-1)^{2-2H}} \tag{3.67}$$



Graph 3.9: Bias correction factor *Θ* for different Hurst parameter (*H*) and sample sizes (*n*) (MATLAB)

By plotting the bias estimation for different n, H (Graph 3.9) it is evident that the more long-term persistence exists in the observations, the greater the bias, while for *H* values closer to white noise behaviour the bias is practically negligible. Sample size on the other hand seems to affect the bias estimation, but not as much as the dependence structure of the process. The bias can be redefined as:

$$\theta = \theta^{HK}(n,H) = \frac{K_{p1}^d - K_{p1}}{K_{p1}} \tag{3.68}$$

And the final K – moments are defined by solving equation:

$$K_{p1}^{d} = K_{p'1} = (1+\theta)K_{p1} \tag{3.69}$$

Where $K_{p'1}$ corresponds to the unbiased K – moment with modified order *p'*, which is derived from:

$$p' \approx 2\theta + (1-2\theta)p^{(1+\theta)^2} \tag{3.70}$$

The final K – moment value is almost the same either using equation or equation. Dependence is important to be taken into account when trying to model natural processes. Combining the results from Graph 3.9 and Equation (3.70) it is clear that the difference between removing the bias and neglecting it is not negligible in most cases.

As moment order increases, since the bias is defined as a percentage of the final K – moment value, the difference will increase too. Consequently, dependence biased high-order moments needed to successfully model extremes will be significantly inaccurate, leading to underestimation of extreme events. From another perspective, since empirical K – moment return periods are assigned and reliant on *p* value, it is evident from Equation (3.59), that there will be consequences in their assignment to extreme values.

3.9 Hydrometeorological Analysis Methods – Use of Complete Record

Nowadays, analysis of hydrometeorological records used for modelling extremes takes place with two main methods. Block maxima and values over threshold.

The block maxima approach introduced by Gumbel (1958), mainly used in extreme value theory (EVT), consists of dividing the sample into equal time periods and choosing to use only the highest observation from each one. The final statistical sample is called “block maxima” with size equal to the number of periods (blocks) and is then used to model an extreme value distribution. The significant disadvantage of this method is that it misses high value observations that simply aren’t the highest in their block, but may well be higher than other maximums in other blocks.

By assigning a value threshold and forming a sample containing all values above that threshold, the “block maxima” disadvantage is alleviated. This method is known as Values Over Threshold (VOT) or more commonly called Peaks Over Threshold (POT). Another advantage in using POT is that the sample used includes most high values, thus it focuses the modelling process on the distribution tail. But, an important disadvantage of using POT is that time dependence and especially in the long-term cannot be discovered. Dependence especially when studying extremes is crucial in order to model correctly. Ignoring it will most probably lead to severe underestimation of extremes.

Consequently, modelling with the whole record as the statistical sample is the most reliable method to proceed with, as no observation is omitted. This means that the modelling result will prove to be more accurate, with dependence correctly estimated. In this study, all observations are used in the modelling process for every method.

4 Precipitation Database

4.1 Data Collection Requirements

The main purpose of this study, as mentioned in the introduction, is to effectively model extreme rainfall from historical observations and comparing the different methods. For this goal, the chosen data set needs to meet the following requirements:

- Daily or sub-daily time scales of rainfall observations need to be used in order to effectively model extremes (Min, et al., 2011). Higher time scales than daily, don't reliably show the frequency and intensity of extremes.

- As the purpose is to provide the means to reliably predict events even in the horizon of 1000 years or more, the historical data should be of length higher than 30 years. Data sets with length lower than 30 years, don't often provide sufficient means of determining the effects of long-term persistence in the sample, which as shown in section (3.8.8) is an integral part of the modelling process.

- In order to show the effectiveness and adaptation of the modelling process, data with precipitation patterns from different climatic regions should be used. Thus, the data set needs to contain worldwide weather stations.

- Aiming to provide a reliable method for estimating extremes, the data set except from its versatility in climate patterns, needs to be in bulk. The more stations provided, the more substantiated the final result.

4.2 The GHCN – Daily Database

After extensive research, the well-established Global Historical Climatology Network (GHCN) – Daily database is chosen. The GHCN – Daily contains daily data from over 100,000 ground weather stations in 180 countries worldwide, from which about two-thirds are used exclusively for precipitation measurements. Like its counterpart for monthly data, GHCN – Daily is composed by numerous daily weather reports from different sources merged together and subjected to rigorous quality assurance (QA) reviews for ensuring their reliability. While the database is mostly focused on precipitation and temperature, many stations also provide measurements for snow, snowfall depth, and other important weather variables (Menne, et al., 2012).

The initiative in creating the GHCN database was made a few decades ago, when a reliable procedure in archiving global weather observation data was needed and had not yet been initialized, since most data was handled by individual state organizations. The largest collections, now fully integrated into GHCN were created by the Global Daily Climatology Network (GDCN) (Gleason, et al., 2002) containing numerous international stations and the National Oceanic and Atmospheric Administration / National Climatic Data Centre (NOAA/NCDC) which contains mostly data from the US and South America. Another important asset is the Global Climate Observing System (GCOS) program, which works to

facilitate the free exchange of daily data from GCOS surface stations. The final GHCN – Daily database contains data from all these organizations and more.

An important aspect of the usage of GHCN – Daily, is the fact that the database is up to this day updated and continuously undergoes QA assessments. Consequently, reliability in the weather observations is well established. The quality tests are mainly comprised of record integrity checks of which some are aimed to flag (Menne, et al., 2012):

- Stations with missing data between days
- Nearby stations with significant differences between each other
- Duplication of data records
- Climatological characteristics inconsistent with location

Comparing it to other databases, GHCN – Daily is most likely the most comprehensive global archive of global weather observations. Up until 2012 the number of total elements in the dataset was over 2 billion, containing nearly 300 million maximum and minimum temperature reads and 800 million daily precipitation measurements. The complete database can be accessed by visiting the NOAA website.

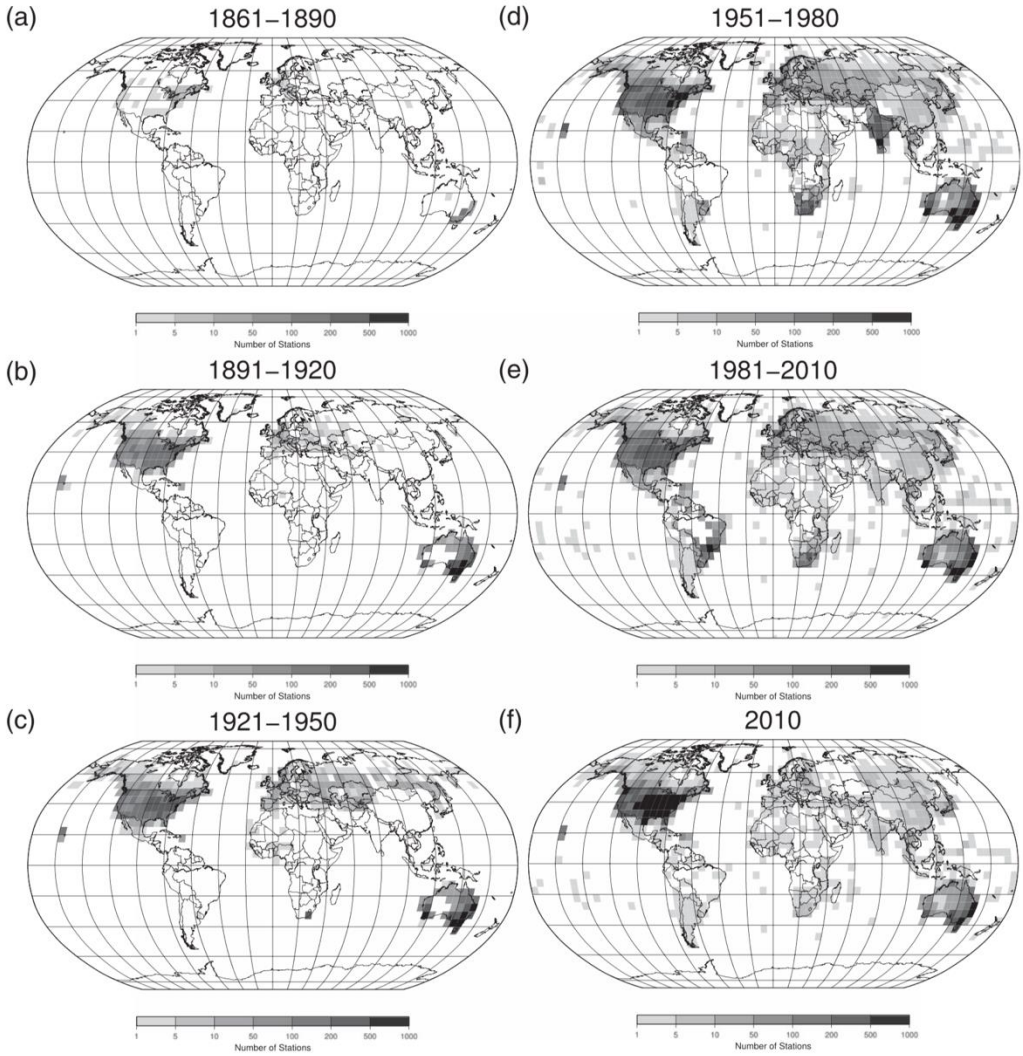


Figure 4.1: Density of GHCN stations measuring precipitation (Menne, et al., 2012)

4.3 Elimination Process and Final Dataset

Data assimilated from NOAA's website containing the newest GHCN – Daily database inquiry accounted to 112,777 total precipitation measuring stations. Their distribution based on their coordinates is shown in Figure 4.2.

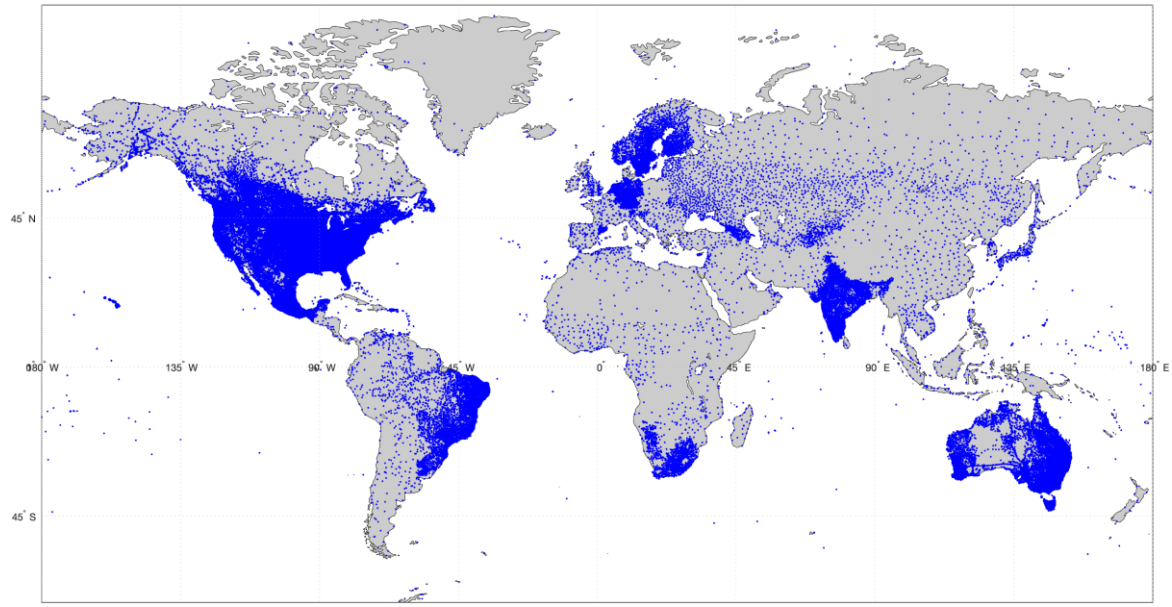


Figure 4.2: World map with total GHCN - Daily stations' distribution

However, they do not constitute the final dataset. In extremes modelling, as mentioned above, it is imperative to use stations whose observations span for over 30 years of continuous recording. Thus, from the total stations acquired, the ones which don't satisfy this requirement are eliminated from the final dataset. The total stations remaining account to 34,782. Their distribution is again shown in Figure 4.3.

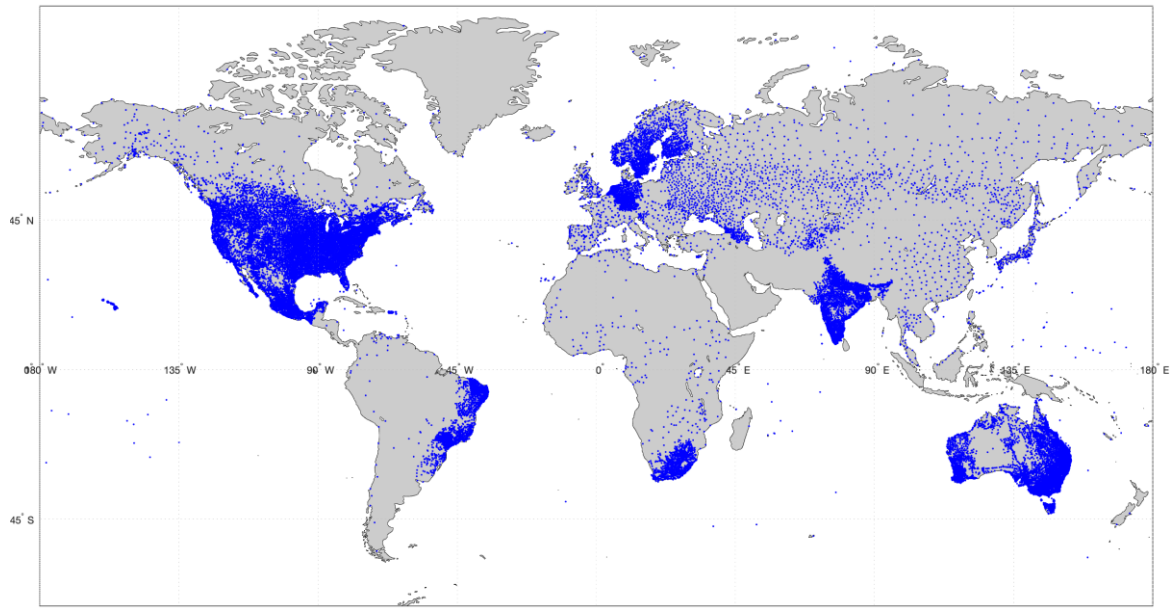
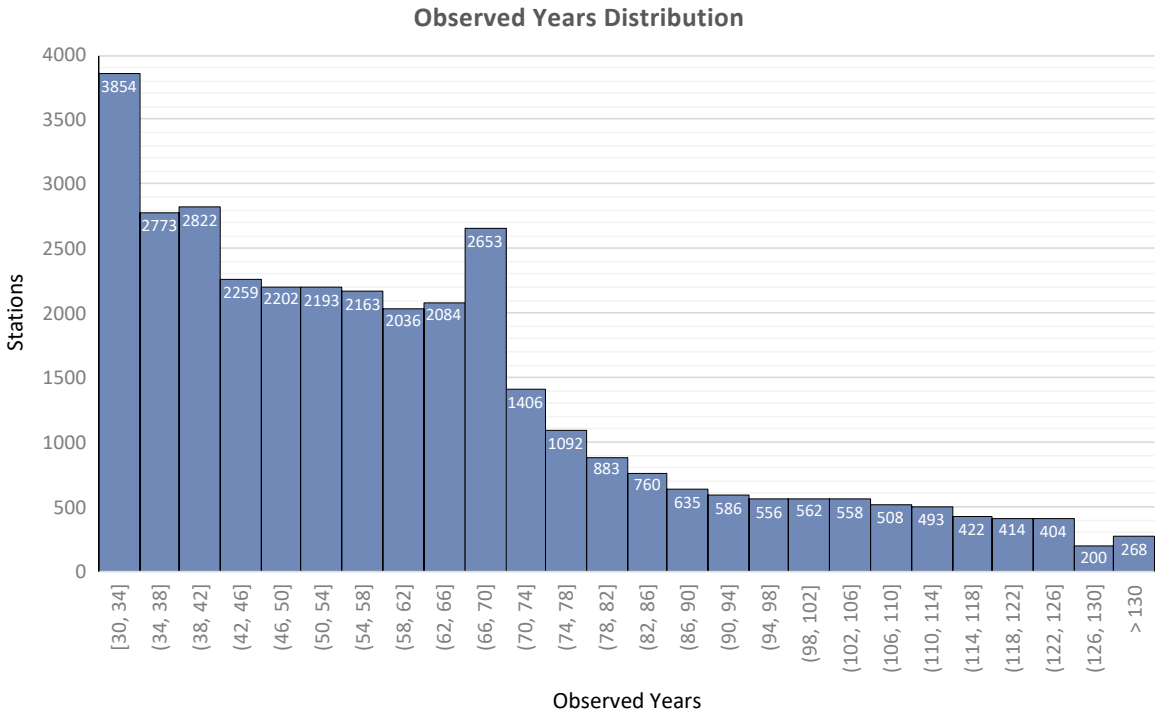


Figure 4.3: World map with GHCN - Daily remaining stations' distribution

Based on the quality assurance assessments made directly into the observations when they are integrated into the GHCN – Daily database, no further action is necessary in eliminating stations. The sample features looked upon when aiming for reliability in the observations and in the final modelling results, are continuity in time without major disruptions, and mitigation of outliers not consistent with stations' regional climatic characteristics. These features are already dealt with in the aforementioned QA. Below are presented; a histogram depicting the distribution of observed years for remaining stations (Graph 4.1) and a heat map of the total observed years for each station respectively (Figure 4.4)



Graph 4.1: Distribution of stations depending on total years observed (>30 years)

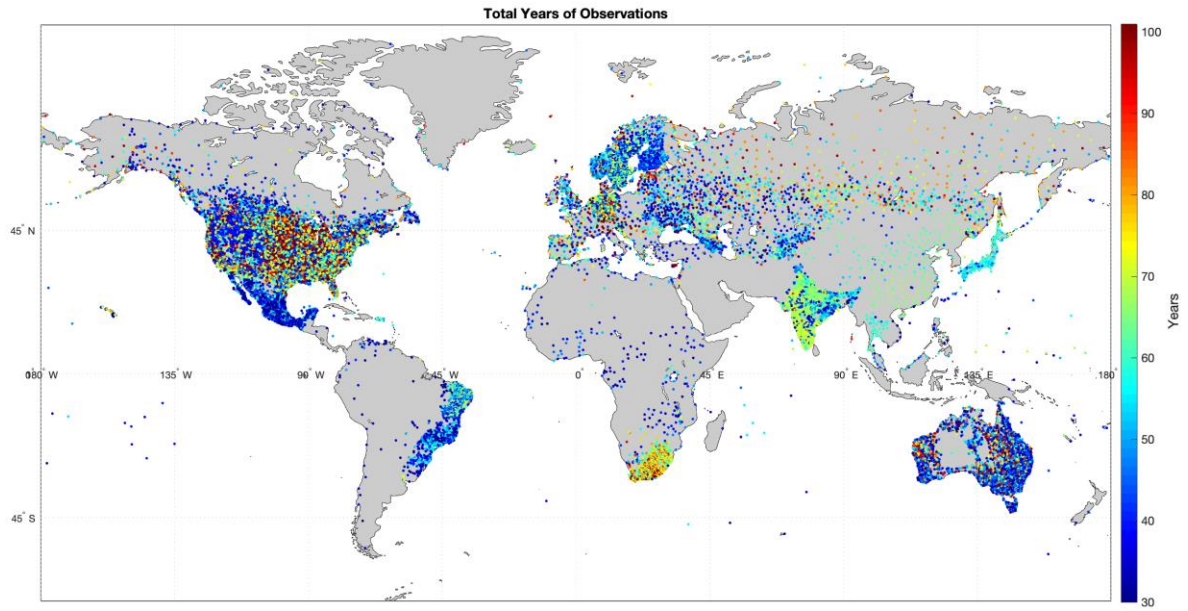


Figure 4.4: Heat map of total years observed from each station

5 Modelling Tools

5.1 Microsoft Excel

For analysing precipitation data and in order to provide the means of reliable modelling and extraction of useful statistical characteristics multiple platforms can be used. Microsoft Excel offers a wide range of statistical tools, while the exclusivity in use of the GRG non-linear solver is an important tool in the fitting process of distributions with complex equations and multiple parameters. Moreover, MS Excel offers the ability to code user specific functions with the Visual Basic Suite meaning that it is versatile and doesn't limit the user to only Excel's functions libraries.

However, when studying bulk data, and in this case over 34,000 stations, it becomes clearly impossible to use Excel in modelling for each one of them, since it can't be completely automated in analysing data. For this reason, the next step was to provide with an equally reliable interface for complex statistics and fitting methods, that could also loop between all stations. The solution was found in the MATLAB programming language which will be analysed below.

MS Excel is still going to be used to statistically analyse the final modelling results and assess the fittings provided by MATLAB. The main chart production in the results section is also produced by Excel's extensive chart possibilities.

5.2 MATLAB

The MATLAB programming language is becoming more and more famous in the engineering community over the past few years. This is due to its versatility and ease of use in a variety of engineering, financial, or statistical analysis. In this study, the core development is being produced in the MATLAB environment and ranges from initial data extraction of GHCN – Daily database, to the final fittings on extreme rainfall modelling and finally exporting the results in Excel (.xlsx) format for further analysis.

It was first conceived by Cleve Moler in the 1970s as an alternative to having to learn Fortran in order to use the mathematical libraries of LINPACK and EISPACK. MATLAB is nowadays being developed by MathWorks (https://www.mathworks.com) and was rewritten in C, C++ and Java in order to improve its versatility. It is considered as a high-level multi paradigm numerical computing environment which aims at solving complex problems with an ease-of-use approach to the user. Some of its main characteristic applications are:

- Algorithm development
- Image processing
- Math and computation
- Modelling and simulation
- Data analysis
- General application development including General User Interface (GUI) production

- Scientific graphs production
- Artificial intelligence designing
- Computational finance support

Standing as an acronym for MATrix LABoratory, its core data element is an array that doesn't require dimensioning, allowing to use simple logic in order to solve even the more complex problems. Evolving over the years, MATLAB has gained popularity in scientific and engineering applications and is considered as the basic programming language for data analytics and high-productivity research. Moreover, it has one of the largest scientific communities online, where the user can download purpose specific code or solve any question regarding the software.

The user-friendly interface (Figure 5.1), other than providing with the basic operations for the script, offers a group of application-specific bundles termed toolboxes. These toolboxes allow the user to easily learn and apply study-specific technology and are provided as basic *.m* function files solving problems covering a large branch of engineering studies (i.e. control systems, neural networks, simulation, statistics, finance etc.). From the abundance of toolboxes this study will use the optimization and curve fitting toolboxes for further validation and quality assurance tests on the fitting process.

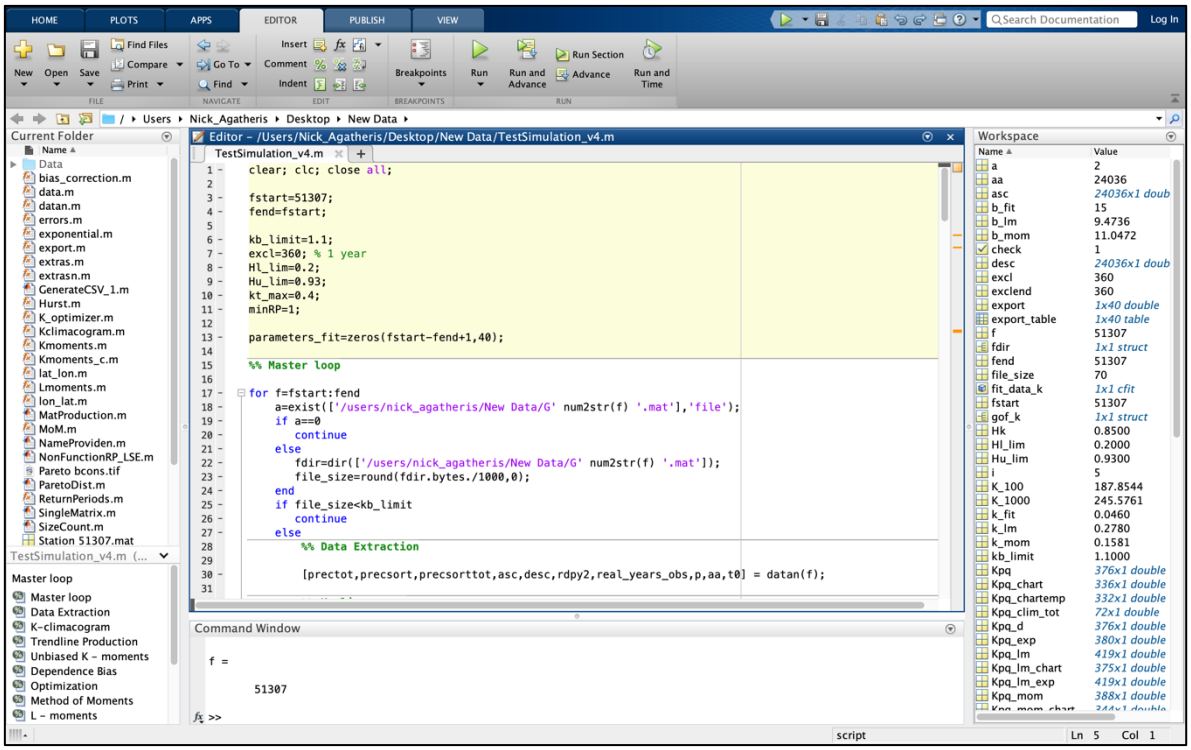


Figure 5.1: MATLAB user interface (version R2018a)

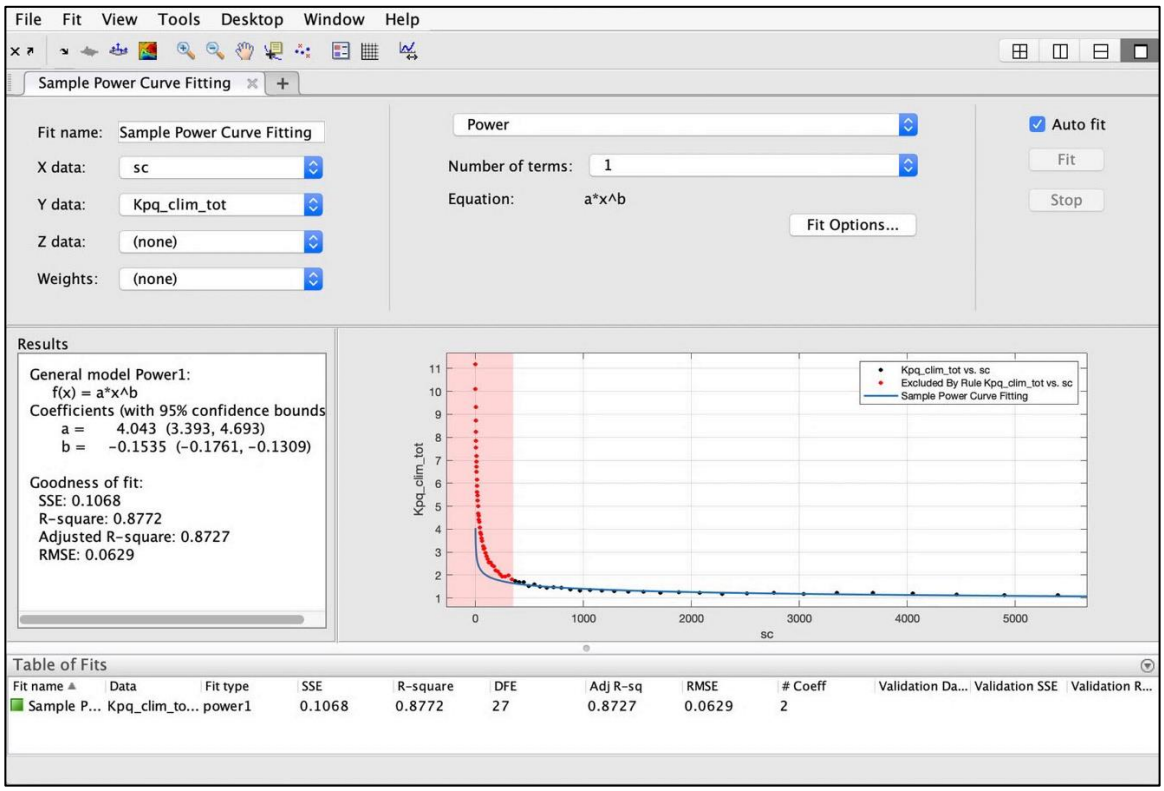


Figure 5.2: Demonstration of MATLAB's Curve Fitting tool used for evaluation of the Hurst parameter from the slope of the fitted power curve to K – climacogram's values (version R2018a)

5.3 Python

In spite of its variability, MATLAB or its aforementioned toolboxes can't provide everything needed for this study's purposes. Instead, for the completion of less observable, but similarly important tasks, the Python programming language was recruited. The choice of using Python was mainly justified for its ease of use and extensive community and function libraries compared to other engineering-oriented programming languages.

In this study, Python is used for manipulating the GHCN – Daily database's files and structuring them effectively for increased run speed and compatibility with MATLAB. Furthermore, each station's metadata contains information such as its unique code name (providing information on its location) as well as its raw geographical coordinates. Thus, Python is used to corroborate these coordinates with the location provided from the station's name based on Federal Information Processing Standard (FIPS) codes, aiming at improving reliability (Table 5.1).

Table 5.1: Station metadata region validation example

Station Name	Latitude	Longitude	Country (FIPS code)	Country (Coordinates)	Check
SZ000002220	47.250	9.350	SZ	Switzerland	✓

Python was first conceived in 1991 by Guido van Rossum as a substitute to the ABC language. It consists of a general use, object-oriented, high-level language, that is dynamically typed and garbage collected, with central focus in code readability. The object-oriented approach helps programmers, engineers and scientists produce clear cut code for projects of every scale or form. To this day it is one of the most used programming languages worldwide (Wikipedia, 2019).

The purpose of using it in this study is its high level of modularity. Specific sets of functions (modules) can be installed to its core, depending on the user's needs. They can be downloaded from a large library of packages (https://pypi.org) thus making Python a very modular programming environment.

In this study, Python version 3.7 is used with code written using the free PyCharm Integrated Development Environment (IDE) made by JetBrains (https://www.jetbrains.com). The main packages needed for the purposes of this study are Pandas, NumPy and Geopy. They are in short analysed below:

- NumPy → main package for scientific and numerical computing
- Pandas → extensive support for data structures and data analysis tools
- Geopy → support for geocoding (or reverse geocoding) services. Receiving as input geographical coordinates, they output information about the respective location (or in the opposite manner).

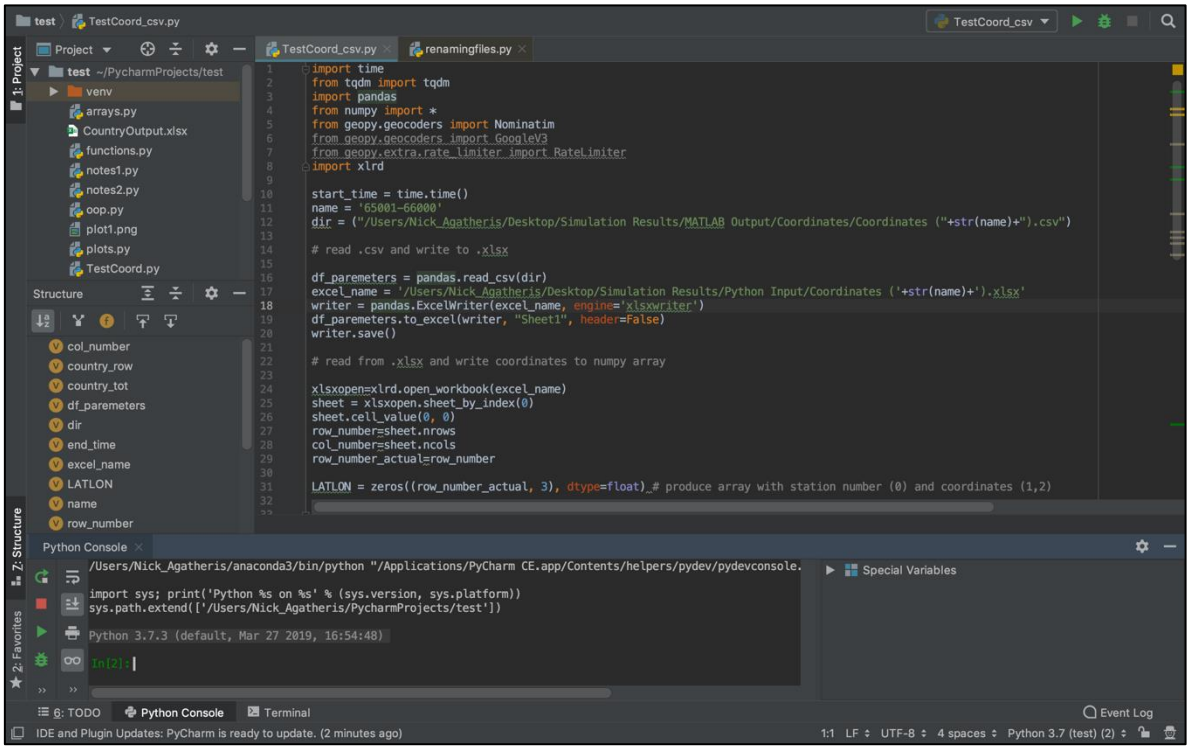


Figure 5.3: PyCharm CE interface (Python 3.7)

6 Modelling Methodology

6.1 Methods Used

With precipitation data provided from GHCN – Daily the modelling process mainly consists of fitting a suitable distribution, in this case the GPD2 and PBF, and providing with goodness-of-fit parameters to judge the accuracy and reliability of the resulting fit. The concept of "fitting a distribution" refers to estimating the theoretical distributions' parameters. In this study focus is given in showcasing the significant advantages of using the newly introduced K – moments for modelling extremes, compared to classic and L – moments. For this reason, all three methods are analysed and compared to each other for their modelling power and consistency. Classic and L – moments fit the GPD2, while K – moments fit both the GPD2 and PBF. This choice will be analysed further in chapter.

6.2 Goodness of Fit Comparison

The aforementioned comparison between fitting methods materialises by calculating the goodness-of-fit between the theoretical distribution and observed data. In practice, the goal is to measure the divergence between observed values and values produced by the model in hand. This can be done with a number of different statistical tests (e.g. Chi-squared test, Kolmogorov-Smirnoff test, Anderson-Darling test), however, in order to provide with a simple yet accurate method to compare models, this study uses the standard Root Mean Squared Error.

In statistics, Root Mean Square Error (RMSE) or Root Mean Square Deviation (RMSD) is a measure of the root squared differences of model produced values with observed values. These differences are more commonly called residuals and in simple terms it measures the accuracy of a model. The estimator of RMSE is (MathWorks, 2019):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{theor} - X_{obs})^2}{n}} \tag{6.1}$$

Where:

- X_{theor} are the theoretical model values
- X_{obs} are the observed data values
- n is the sample size

RMSE always takes positive values, while a zero (0) value means that there is perfect fit to the observed data. Thus, when comparing methods, a lower RMSE value means better overall fit. However, since the error is being squared, higher scales produce disproportionally larger errors than lower scales, which means that high error in large scales doesn't mean that the fitted model is always unreliable.

In order to solve this scaling issue, the Normalised Root Mean Square Error (NRMSE) is also used. The benefit of normalising data is that all scaling properties of the model are omitted, consequently creating a more efficient and more straightforward fitting evaluation process. NRMSE is a slight variation from the traditional RMSE, allowing this valuable data normalisation. NRMSE outputs values in the range $(-\infty, 1]$ with 1 showing perfect fit. It is defined as (MathWorks, 2019):

$$NRMSE = 1 - \sum_{i=1}^{n} \frac{X_{obs} - X_{theor}}{X_{obs} - \overline{X}_{obs}} \tag{6.2}$$

Where:

- $\overline{X}_{obs}$ is the observed sample mean
- n is the sample size

6.3 Modelling Procedure

6.3.1 Initial File Processing

In the ensuing subsections, the procedure followed in rainfall extremes modelling with each of the three methods is analysed step by step. Firstly, the raw files from the GHCN – Daily database have to be modified to work in the MATLAB environment. The steps followed for this purpose are:

A. Raw database files (.dly) containing measurement data for a number of variables (i.e. precipitation, temperature, snow cover), as well as station metadata, are accessed and converted to MATLAB compatible files (.mat) containing only the whole set of precipitation data and important metadata such as station name, coordinates and starting-ending date of measurements (Jaffrés, 2019).

B. Each individual file is checked for satisfying the requirement for more than 30 years of observed data. Any station below this threshold is ignored from the core modelling process.

C. The mainframe for order statistics is built by sorting each station's data in descending order, while excluding zero values since they don't contribute in the fitting process. Order statistics are essential for assigning return periods to sample values.

D. Other important statistical characteristics of the sample are extracted (e.g. average, standard deviation, rain days per year)

6.3.2 Modelling with Method of Moments

The first method used is the classic method of moments (MoM). For fitting the Generalised Pareto Distribution with two (2) parameters the following process is used:

A. The first and second order moments for the whole data set are estimated through Equations 3.10 and 3.11. The number of moments used is equivalent to the number of the distributions' parameters. While similarly third and fourth order moments could be used, since higher moment orders are better for modelling extremes, the counterargument is that moment estimation becomes significantly unreliable for orders higher than two. Consequently, this study uses the simplistic and more common approach of estimating the *mean* ($\overline{x}$) and *variance* (s^2).

B. Providing that mean and variance are finite, parameters κ and λ are estimated using equations produced by Hosking & Wallis (1987) for the method of moments and GPD2:

$$\lambda = \frac{1}{2}\overline{x}\left(\frac{\overline{x}^2}{s^2} + 1\right) \tag{6.3}$$

$$\kappa = \frac{1}{2}\left(\frac{\overline{x}^2}{s^2} - 1\right) \tag{6.4}$$

C. With known GPD2 parameters, return periods (T_{MoM}) are assigned by using Equation 3.53 and converted to yearly values by dividing with rain days per year ($rdpy$).

D. Observed data are sorted in ascending order and are assigned sample return periods (T_{emp}) from Equation 3.32 and similarly converted to yearly values by dividing with $rdpy$.

E. In a log-log plot sample return periods and theoretical return periods are plotted against rainfall values.

F. Goodness-of-fit is estimated with RMSE and NRMSE by using equations (6.1) and (6.2) respectively, between theoretical and observed plots.

6.3.3 Modelling with L – moments

A similar procedure to classic moments is once again used in modelling with L – moments, for the Generalised Pareto Distribution with two (2) parameters:

A. Observed data without zero values is sorted in ascending order to facilitate the use of order statistics in L – moments estimators.

B. Either from Equation 3.19 or 3.20, the first two Probability Weighted Averages (β_0, β_1) are estimated. Constants are attributed their respective recommended values $\gamma = 0.35$ and $\delta = 0$.

C. Based on equations produced by Hosking & Wallis (1987) the parameter estimators for GPD2 using L – moments are:

$$\kappa = \frac{\beta_o}{\beta_0 - 2\beta_1} - 2 \tag{6.5}$$

$$\lambda = \frac{2\beta_o\beta_1}{\beta_0 - 2\beta_1} \tag{6.6}$$

Then the same process for plotting and error calculation is applied.

6.3.4 Modelling with K – moments

Moving on from the classic methods, the new concept of rainfall modelling using knowable moments holds several advantages as seen from the theoretical analysis. Especially for extreme-oriented distribution fitting they prove to have several unique advantages not present in any other method.

Firstly, unlike classic moments, K – moments are knowable with reliable and unbiased estimators for orders up to the size of the sample. With increasing order, more weight is given in higher values of the sample, thus their estimation for high orders is greatly focused on extremes.

Long-term persistence bias existing in most rainfall records is taken into account when using K – moments and the whole data set. In classic methods using Peaks Over Threshold, dependence is omitted thus producing a significant probability of severely underestimating extreme values. Lastly, K – moments can be directly assigned return periods, with the use of Λ – coefficients.

The following framework was used in order to successfully implement the use of K – moments for extreme-oriented rainfall modelling:

A. While using all data, unbiased central K – moments are estimated from Equation 3.47, for $q = 1$ and for *p* up to 1/10 the size of the sample.

B. The K – climacogram is constructed following the procedure in section 3.8.7, using the aforementioned central K – moments and for scales up to 1/10 the size of the sample. From its slope, the Hurst coefficient is estimated from Equation 3.64.

C. Non-central unbiased K – moments are estimated from Equation 3.46, for $q = 1$ and for *p* up to the size of the sample *n*.

D. Depending on the value of the Hurst coefficient dependence bias is estimated from Equation 3.67 and taken into account in the non-central K – moments according to Equation 3.69.

E. By the theoretical equations of Λ – coefficients for the Pareto distribution, empirical return periods are assigned to the non-central K – moments.

F. Setting a starting point of κ and λ, theoretical return periods based on Equation 3.53, are estimated.

G. Using an optimization algorithm, the best theoretical fit is produced by minimizing the error between empirically assigned return periods and theoretical ones. In this case, Least Squares (LSE), as in Equation 6.7, are used. Since the purpose of this study is to efficiently model extremes, by setting a threshold on empirical return periods ($T > 1\ year$) and minimizing the LSE on that range an optimal fit on extremes is achieved. The flexibility of the method is obvious, as the model can be calibrated to fit effectively specific ranges of return periods thus specific intensities of rainfall.

$$LSE = \sum_{i=1}^{n} \left(\frac{\ln(T_{\text{theor}})}{\ln(T_{\text{emp}})} \right)^2 \tag{6.7}$$

D. In a log-log plot sample return periods, empirical K – moments return periods and theoretical return periods are plotted against rainfall values.

E. Goodness-of-fit is estimated with RMSE and NRMSE.

By following this method, as it is evident more emphasis is given in extremes. Thus, sometimes while minimizing the LSE in a specific range of return periods, the lower part of the distribution will not be fitted as accurately as possible. While accuracy is sacrificed in the lower end, precision on extreme values is significantly more important since they are the focus in most aspects of engineering design and risk assessment studies.

However, complete fitting accuracy can be achieved by adding one more parameter to the theoretical distribution function. In this study, an evolution of the two (2) parameter Generalized Pareto distribution is the Pareto-Burr-Feller (PBF) distribution which was analysed in section 3.3.2. The extra parameter of the PBF is vital in combining accuracy in lower and high return period values. The process of fitting is the same as with the classic GPD2 and should provide with the best overall fit of observed data. The only difference is that since a best overall fit is needed, the LSE minimizing process is done without setting a threshold (quantile weight). While this method will most probably provide better results, by adding an extra parameter, the model experiences higher uncertainty. Consequently, it is advised to use the more parsimonious model of GPD2 (Koutsoyiannis, 2019).

In the following chapter, an application of all methods is presented in a specific station, which will then be generalised in the whole of the database for evaluating both classic and K – moments methods.

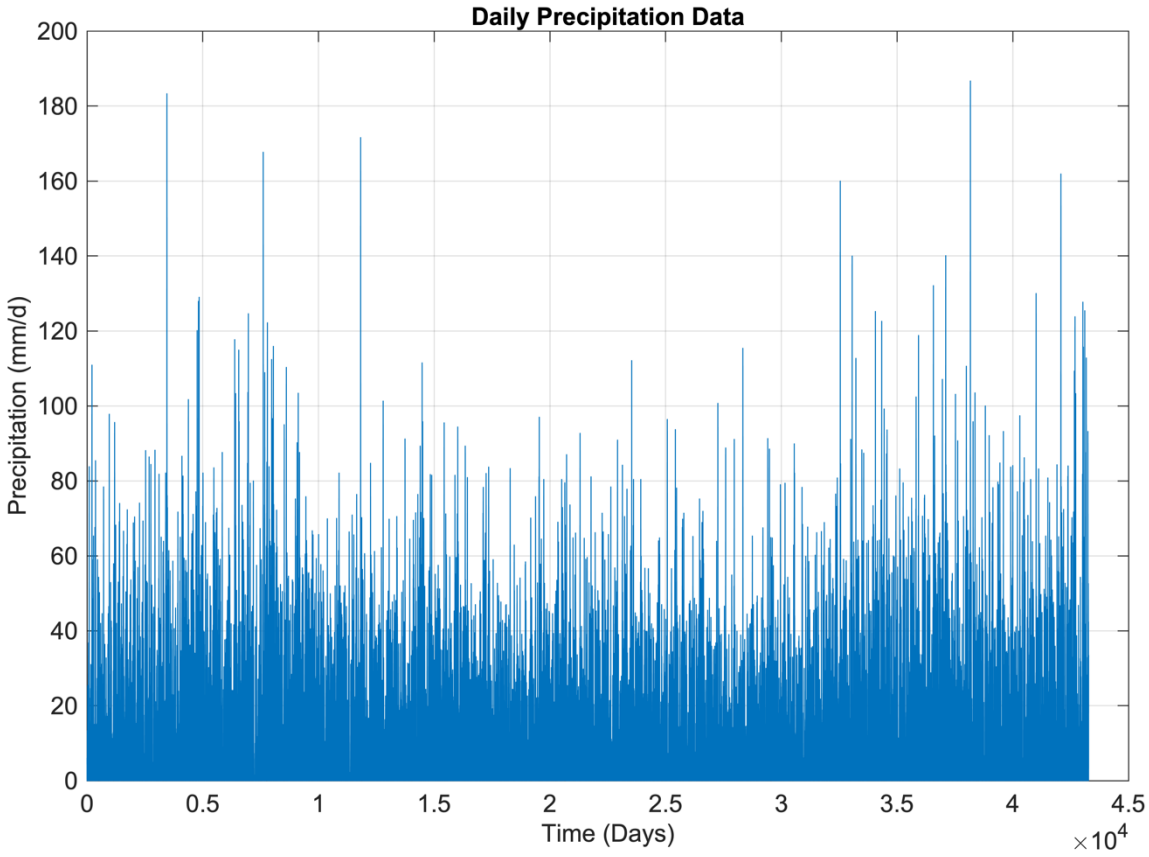
7 Sample Station for Extreme-oriented Rainfall Modelling

7.1 Station Characteristics

For the methods demonstration, station named “SZ000002220” is chosen. The reason for this selection is the evident portrayal of the effectiveness of the K – moments method in comparison to classic and L – moments. Furthermore, it satisfies all expected requirements for reliability in the modelling process.

Station “SZ000002220” with coordinates [47.250, 9.340] is situated in the Appenzell Innerrhodden province, in the North-East region of Switzerland. More specifically it is located in the peak of the highest mountain of the Appenzell alps, most commonly called Säntis. Elevation at the weather station’s position is 2502m. The station was built by order of the International Meteorological Congress of Rome in 1879 in which it was deemed necessary to build weather stations across the most accessible mountain peaks of Europe. The general location and a bird’s eye view of the station are provided in Figure 7.1 and Figure 7.2.

All weather data from its commissioning until today, are compiled into the GHCN – Daily database. By filtering it through quality assessment the database to this day contains 43,276 total daily observations, amounting to a total of 119 years of measurements, far above the required minimum of 30 years. From the total number of observations, precipitation days amount to 24,036, and by dividing with the 119 years, results in 202 rain days per year. The complete time series is presented below in Graph 7.1.



Graph 7.1: Rainfall observations of station “SZ000002220”

Figure 7.1: Säntis weather station bird's eye view (Wikipedia)

Figure 7.2: Station's location in respect to Western Europe (Google Earth)

7.2 Classic Methods Evaluation

Using the whole data set of 24,036 precipitation days and following the process from sections 6.3.2 and 6.3.3, the first step is to estimate the *mean, variance,* and the first two *probability weighted averages.* Then, from equations 5.3 and 5.4, the GPD2 parameters are estimated for classic moments, and from equations 5.5 and 5.6, for L – moments respectively. The results are presented below:

Table 7.1: Classic moments parameter estimation and goodness-of-fit statistics

Classic Moments					
Mean	**Variance**	**κ**	**λ**	**RMSE**	**NRMSE**
13.122	251.812	0.158	11.047	27.257	0.521

As for classic moments, because of the station's position and the fact that on average the region receives 202 rain days per year, the high *mean* value is not unusual. The fitted GPD2 with the method of moments is shown in Graph 7.2. While, for low orders the fit shows perfect results, for high orders where extremes are located, the fitted distribution tail slightly overestimates observed values. As this is a log – log plot it is evident that with increasing return periods, the difference between observed and theoretical values increases considerably. Mathematically, in this case, the tail index *κ* should have had a lower value in order not to overestimate extreme values.

Table 7.2: L - moments parameter estimation and goodness-of-fit statistics

L – moments					
β_0	**β_1**	**κ**	**λ**	**RMSE**	**NRMSE**
13.122	2.751	0.278	9.474	113.009	-0.988

On the other hand, L – moments fit estimates higher tail index value than classic moments and as seen from Graph 7.2 this leads to even higher overestimation of extremes, though similarly, for lower return periods, the fit is almost perfect. The significant unreliability in fitting is also confirmed by the substantial RMSE value and the negative value on NRMSE.

RMSE and the variant NRMSE as mentioned before, are steadfast goodness-of-fit determinants for the whole fitted distribution. In order to facilitate the means for evaluating and comparing fitting methods separately on the distribution's body and tail, three RMSE and NRMSE values will be calculated in each method.

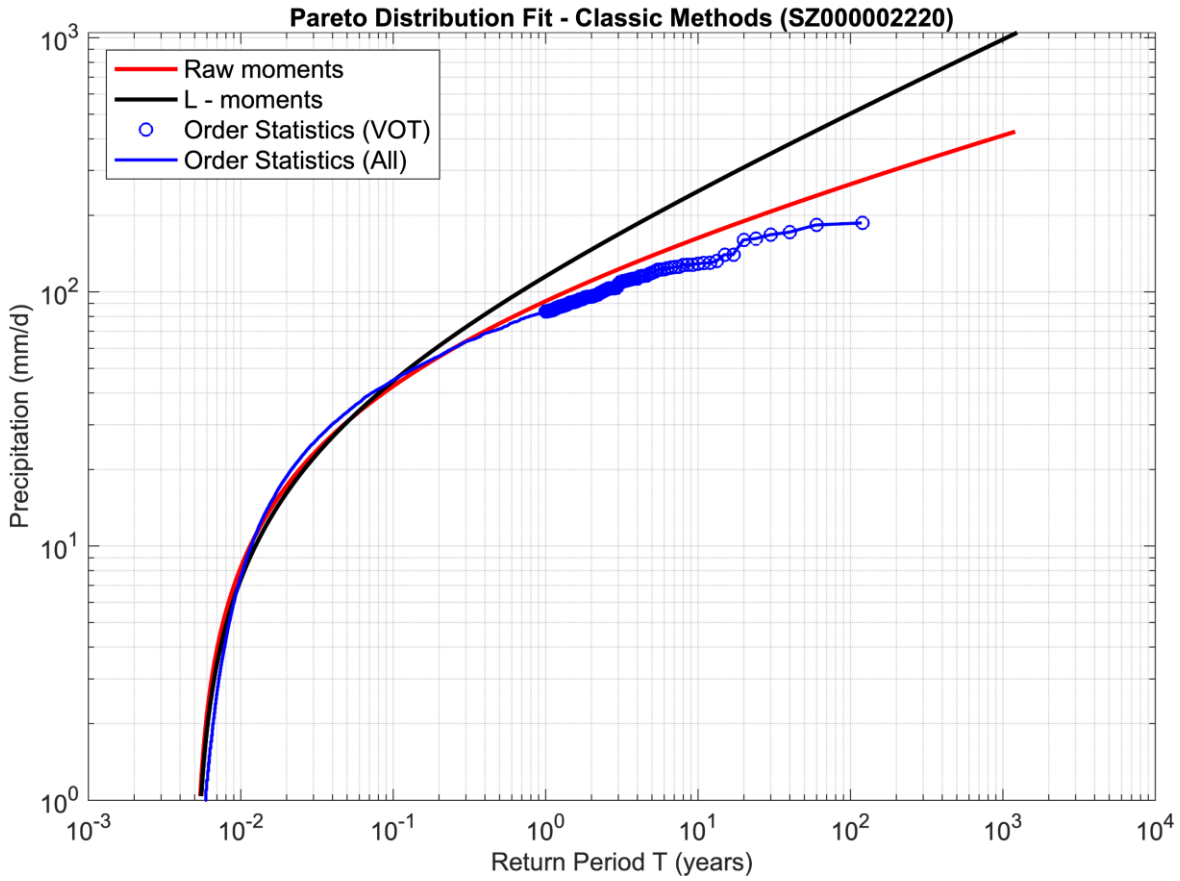
A. GoF – Total → error of the whole fitted distribution
B. GoF – Low → error for $T < 1$ year, which portrays goodness of fit on low orders.
C. GoF – High → error for $T \geq 1$ year, which portrays goodness of fit on high orders (i.e. extreme values).

Table 7.3: RMSE and NRMSE values for different parts of the fitted distribution

Method	RMSE		NRMSE	
	Classic Moments	L - moments	Classic Moments	L - moments
High	39.413	163.476	-0.177	-3.883
Low	2.620	9.858	0.895	0.603
Total	27.257	113.009	0.521	-0.998

For the classic methods showcased, the specific RMSE and NRMSE values are shown in Table 7.3. Confirming the visual representation from Graph 7.2, RMSE is minimal in low and significant in high orders. In the same way, NRMSE is closer to 1 in low and lower in high orders. By splitting the goodness of fit in high and low return period values, all methods effectiveness on modelling extremes can be quantified reliably. By only using the standard total RMSE and NRMSE values, a comparison specifically on extremes fitting can't be made, as from a single number there is no way of knowing where the error is produced from.

The unsatisfactory results portrayed are produced by using the classic methods of distribution fitting, while using every precipitation observation ($x_i > 0$) of the sample. While a statistically better fit on extremes can be achieved if peaks over threshold are used, this is done in expense to not taking into account long-term dependence, an important characteristic of most rainfall samples, as mentioned in section 3.8.8. The purpose should be to try and find a method that is both statistically and naturally consistent.



Graph 7.2: Modelling results for classic methods. Values over threshold (T>1 year) and the whole sample are also plotted for reference

7.3 K – moments Method Evaluation

7.3.1 Assuming Sample Independence

In order to showcase the difference in taking into account the long-term dependence bias, the procedure in fitting with K – moments is applied twice. At first, the sample is assumed as independent, while on the second trial the dependence bias is estimated and a comparison between the two is produced.

For ignoring dependence modelling results are provided in Table 7.4. Since the fitting process is based on minimizing the Least Squared Error (LSE) between empirical return periods assigned to K – moments and theoretical return periods the error value is also provided.

Table 7.4: Independent sample - K - moments parameter estimation and goodness-of-fit

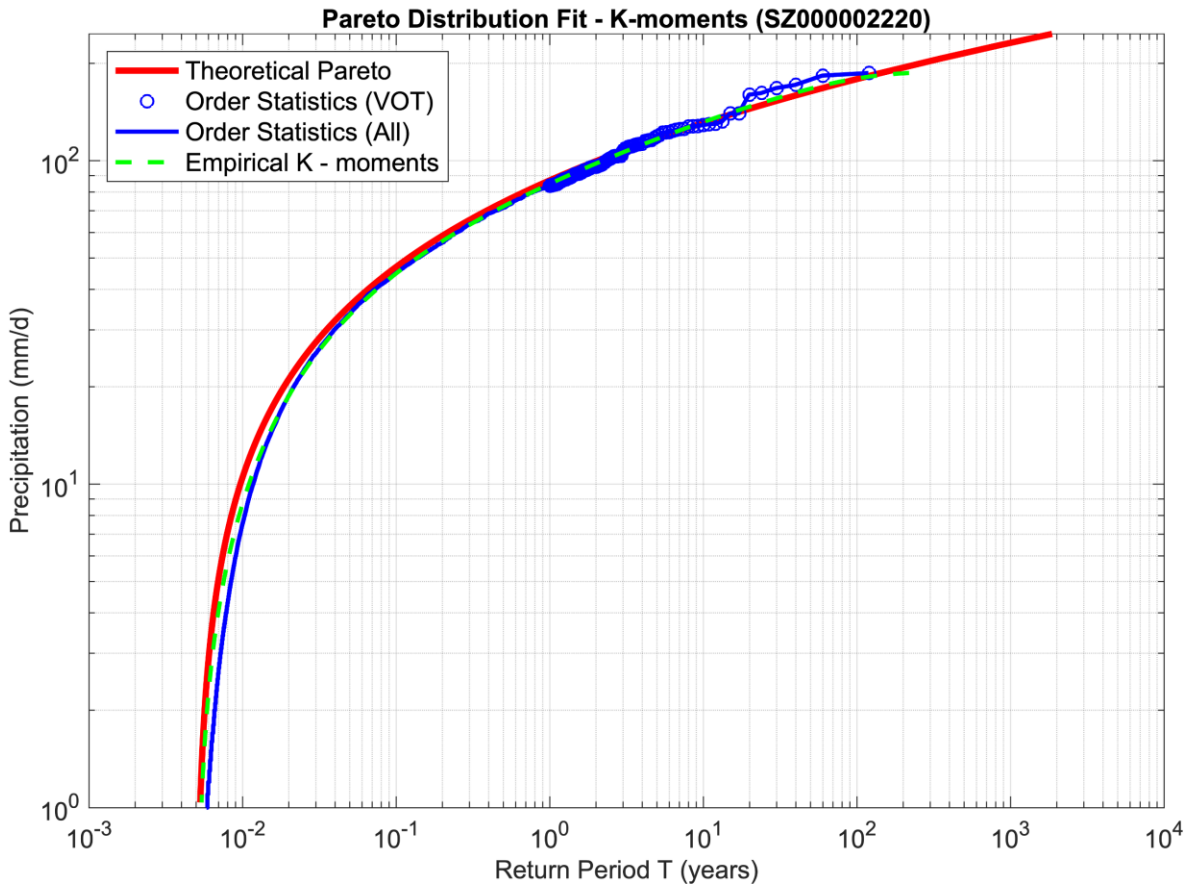
K – moments (Independent)				
κ	**λ**	**LSE**	**RMSE**	**NRMSE**
0.040	14.700	2.083	5.842	0.897

The low LSE value shows almost perfect fit of the GPD2 to empirical return periods, which in turn validates the selected distribution as the appropriate for the modelling process. Moreover, the goodness of fit statistics for the theoretical and observed values are respectively significantly better than the classic methods, thus meaning that the fit is close to perfect for all orders.

The fitted distribution, combined with empirical return periods are showcased in Graph 7.3. It is evident that the fit is perfect for high orders where extremes are located, and almost as good for low orders. Moreover, by using the same splitting process in evaluating the RMSE and NRMSE, the results in Table 7.5, show low error on both high and low return periods, as expected.

Table 7.5: RMSE and NRMSE values for different parts of the fitted distribution

Method	**K – moments (Independent)**	
	RMSE	**NRMSE**
High	8.098	0.758
Low	2.358	0.905
Total	5.842	0.897



Graph 7.3: Modelling results with the K - moments method for assumed sample independence. Empirical K - moments return periods are also plotted.

7.3.2 Long-term Dependence Bias Effect

For taking into account dependence bias, the Hurst parameter has to estimated first. By constructing the K – climacogram (Graph 7.4) for scales up to 5,000 days ($\sim 1/10$ of the sample) the estimated Hurst parameter is $H = 0.85$ which indicates significant long-term persistence. Afterwards, the order (p) values that correspond to K – moments estimated from the assumed independence method are corrected and the modelling process continues as before. The final fitting results are shown in Table 7.6.

Table 7.6: Dependence biased sample - K - moments parameter estimation and goodness-of-fit

K – moments (Dependent)				
κ	**λ**	**LSE**	**RMSE**	**NRMSE**
0.046	15.000	1.872	4.933	0.913

Table 7.7: RMSE and NRMSE values for different parts of the fitted distribution

Method	RMSE	NRMSE
	K – moments (Dependent)	K – moments (Dependent)
High	5.921	0.823
Low	3.820	0.846
Total	4.933	0.913

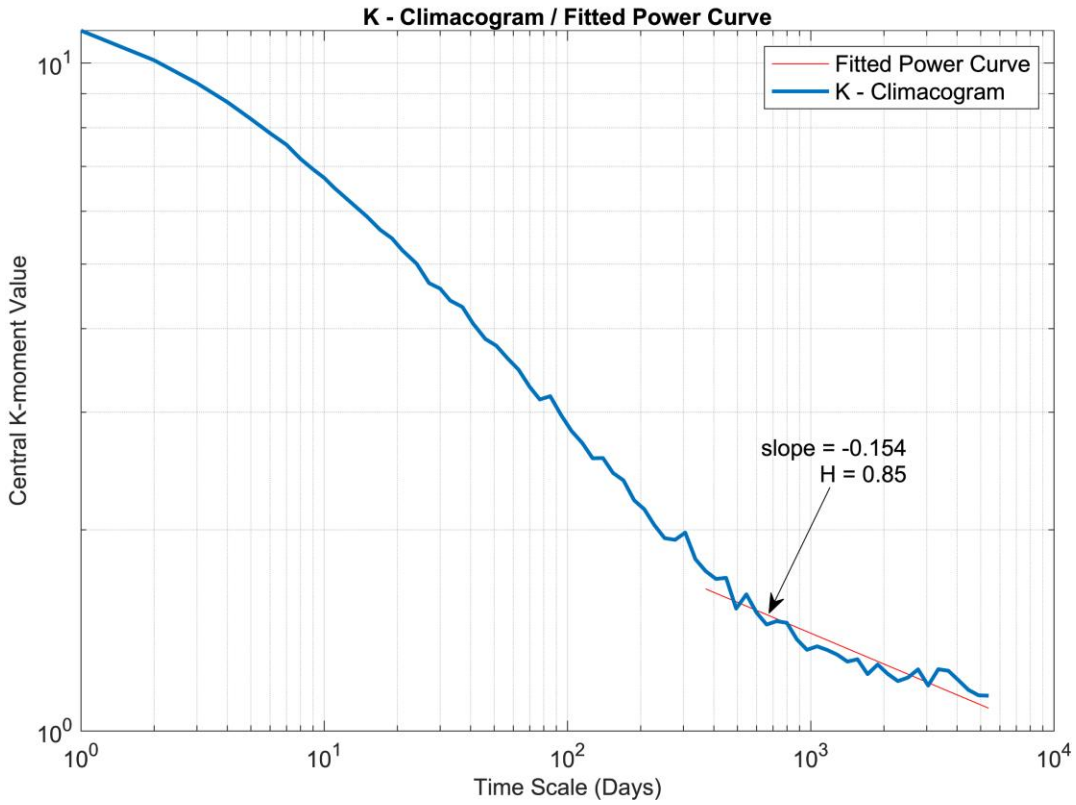
The results themselves don't show significant differences in terms of parameter estimation of the GPD2. While LSE and RMSE errors are now lower (and NRMSE better) than in the previous trial, this doesn't mean that if they were slightly higher the fit would be less effective. This comes down to how much the bias affects empirical return periods, which in turn depict the behaviour of theoretical GPD2 return periods. Again, RMSE and NRMSE is split and the individual errors are provided in Table 7.7.

By comparing pure goodness of fit statistics and GPD2 parameters the true magnitude of long-term persistence bias is not portrayed. Table 7.8, provides with interpolated rainfall values from the two fitted GPD2 for return periods $T = 100$ and $T = 1000$ years which are mostly associated with designing engineering works. The difference in expected rainfall is significant, with the first trial underestimating values by a non-negligible margin for either return period. Consequently, bias is necessary to be acknowledged when modelling extremes.

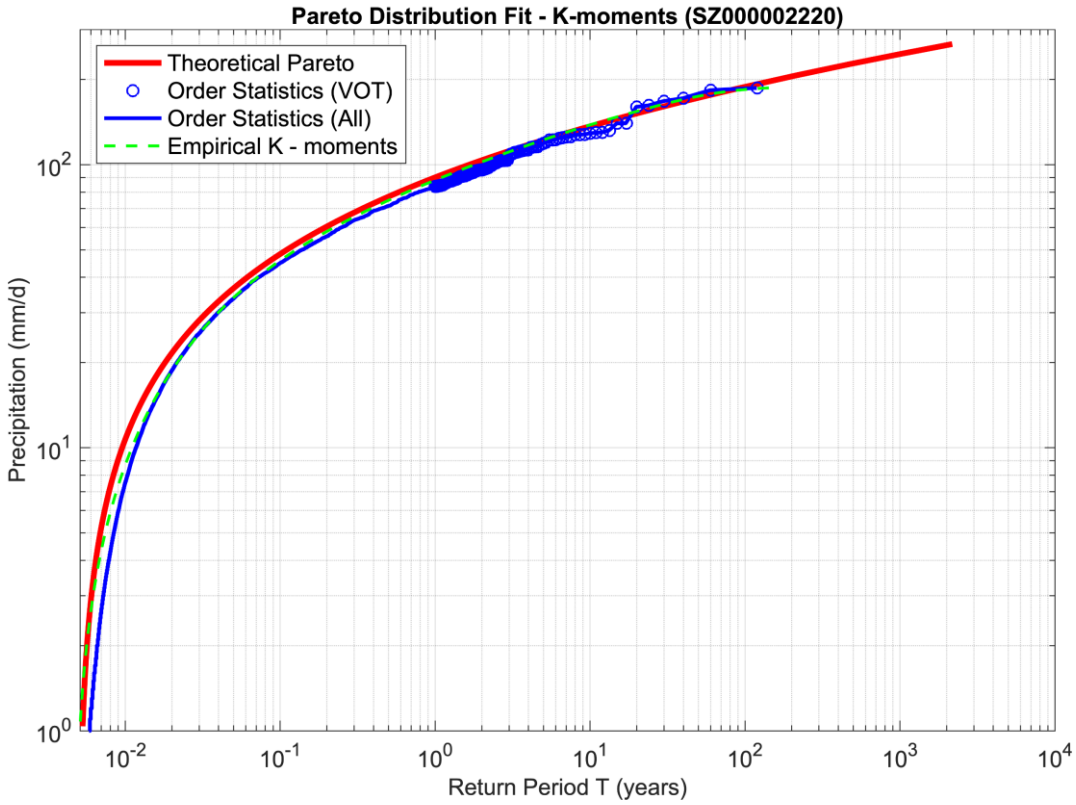
Moreover, from Graph 7.3 and Graph 7.5, by purely comparing empirical return periods assigned to K – moments from both methods, one can easily notice the difference in the positioning of the empirical curve. For example, the highest empirically set K – moment value $K_{pq} = 186.7$ is assigned a return period (in years) of $T_{emp(indep)} = 219.17$ assuming independence, while with dependence accounted for, this value is $T_{emp(dep)} = 142.48$. In risk analysis terms and engineering design studies this is a significant difference which should always be accounted for.

Table 7.8: Effect of long-term dependence bias on rainfall values for large return periods

Sample Structure	Rainfall Expectation (mm/d)	
	$T = 100$	$T = 1000$
Independent	178.85	230.21
Dependent	187.86	245.58



Graph 7.4: K - climacogram from unbiased central K - moments (p=2) and fitted trendline to measure Hurst parameter in large time scales.



Graph 7.5: Modelling results with K - moments method plus long-term dependence bias estimation. Empirical K - moments return periods are also plotted.

7.3.3 Methods Comparison

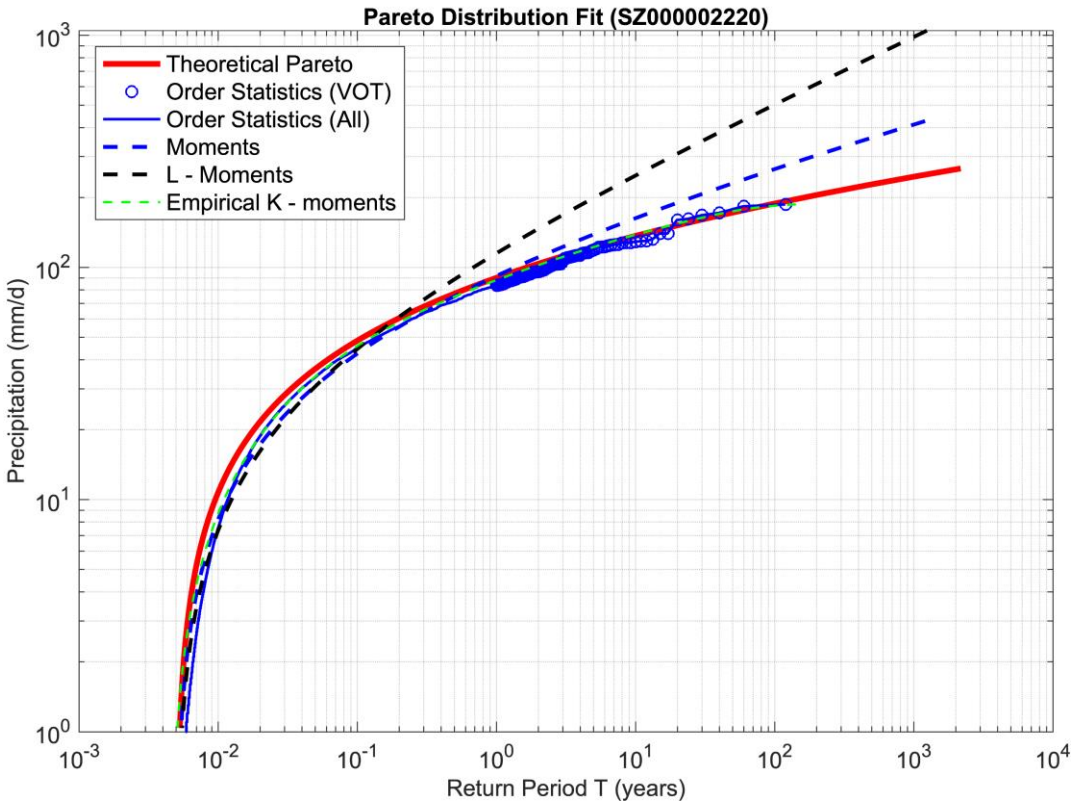
In the following tables, a summary of the fitting methods parameters and errors is provided along with rainfall expectations at large return periods, in order to showcase the superior performance of using K – moments for modelling extremes. Furthermore, in Graph 7.6 all classic methods are plotted together with the bias dependent K – moments method.

Table 7.9: Modelling results parameters and goodness-of-fit comparison for all methods

Method	κ	λ	RMSE High	RMSE Low	RMSE Total	NRMSE High	NRMSE Low	NRMSE Total
Classic moments	0.158	11.047	39.413	2.620	27.257	-0.177	0.895	0.521
L - moments	0.278	9.474	163.476	9.858	113.001	-3.883	0.603	-0.998
K - moments	0.046	15.000	5.921	3.820	4.933	0.823	0.846	0.913

Table 7.10: Rainfall expectation comparison for all methods used

Method	Rainfall Expectation (mm/d)	
	$T = 100$ years	$T = 1000$ years
Classic moments	264.85	411.78
L - moments	502.10	983.20
K - moments	187.86	245.58



Graph 7.6: Final fitting with all methods for comparison

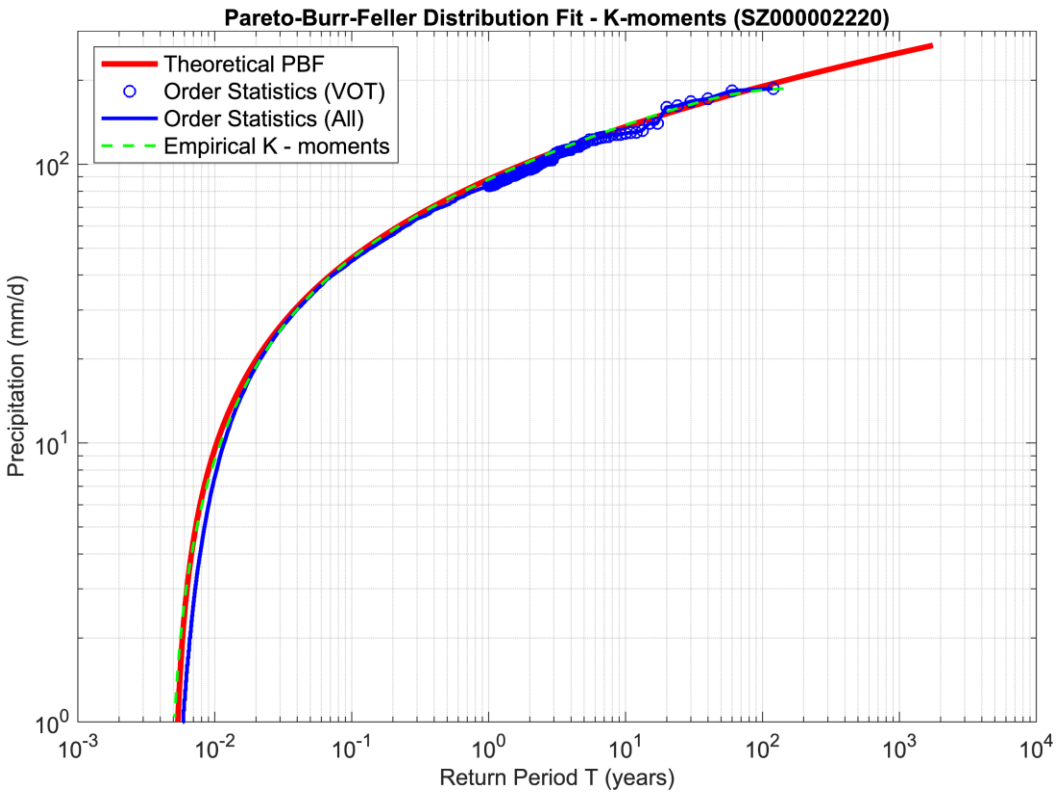
7.4 Overall Fit Improvement

The dominance of using K – moments when focusing on extremes is evident, while for low orders there is a slight tendency of overestimating observed values, best explained on section 6.3.4. Although the focus remains in proving reliability on predicting extremes, an overall better fit can be achieved with the use of K – moments and the PBF distribution. The same fitting process is applied as before except now Least Squares are measured in the whole spectrum of the distribution, not just for $T > 1$ year. Moreover, in order to maintain consistency in tail accuracy, the tail index from the standard K – moments modelling is preserved. The variables are now only the scale parameter λ and the new parameter c. Results are provided in table and figure.

Although better overall fitting is achieved, in this study the main focus is on extremes, while using the simplest model possible for providing the most consistent results. A simpler distribution suggests less overall model uncertainty. With either distribution fitting of high orders accomplishes almost the same great reliability. Thus, the PBF distribution is shown here as an example and will not be present in the final fitting results for the whole database.

Table 7.11: Comparison of different distributions used for modelling with K - moments

Distribution	κ	λ	c	RMSE High	RMSE Low	RMSE Total	NRMSE High	NRMSE Low	NRMSE Total
GPD2	0.046	15.000	1	5.921	3.820	4.933	0.823	0.846	0.913
PBF	0.046	13.510	0.953	5.176	1.974	3.847	0.845	0.921	0.932



Graph 7.7: Fitting result with PBF distribution

8 Cumulative Results

8.1 General Overview

This section contains the results produced by the generalization of the process followed in fitting the theoretical distribution to observed results for the aforementioned sample station. The process is now universally applied to the entirety of the GHCN – Daily database selected from section 4.3. The procedure is exactly the same as in station "SZ000002220".

By automating the procedure through MATLAB ease of use in monitoring for potential errors during each station's fitting is achieved. The final fitting results are assembled in an Excel spreadsheet (Table 8.1) produced by MATLAB, as it enables easier evaluation and comparison of gathered data. This spreadsheet contains:

A. Geographical coordinates and official name of each station, along with a station specific identification code based on the computer handling MATLAB scripts and the GHCN — Daily database. This code is used as an easier reference for locating a station in the file system, if necessary.

B. Basic statistical characteristics for each station including total number of observations, total rainfall observations, total years observed, averages and standard deviations for all data and only rainfall data accordingly.

C. Parameter estimation results according to the process followed in sections 6.3.2, 6.3.3, 6.3.4. These results include the estimation of the tail index κ and the scale parameter λ. Specifically for the K — moments method, as it is based on a minimization algorithm, the Least Squared Error (LSE) between theoretical and empirical return periods is also printed along with the parameters.

D. Goodness-of-fit statistics (6.2. This means that total RMSE and NRMSE values are given, combined with the individual RMSE and NRMSE values for different parts of the distribution.

E. Rainfall values following the K— moments method assuming long-term dependence, for large return periods, and specifically for T = 100 years and T = 1000 years, from MATLAB's interpolation algorithm.

F. The percentage difference between interpolated rainfall values according to K – moments and classical methods to showcase the impact on rainfall value inconsistency between methods. From these percentages, if needed, actual rainfall values for the other methods can be estimated.

G. Data assimilated from rerun of the script for stations with high Hurst parameter, now assuming sample independence.

After the production of the Excel spreadsheet, analysis on the fitting results can begin. First of all, a general evaluation and comparison is made between all methods used in the modelling process. An overview of the fitting parameters composition is given in order to show the general disparity between estimation methods, while also histograms are produced to better showcase the result.

In order to showcase the effectiveness of knowable moments against other methods, separate histograms are created depicting goodness-of-fit statistical parameters. Greater density of high NRMSE values shows better fit, while the same is valid for greater density of low RMSE values. Since in this study the focus is in modelling extremes, again histograms are produced, now showing the effectiveness of K – moments for high return periods, while also evaluating their efficiency in low return periods.

An integral part of this study is evaluating the importance of accounting for long-term dependence in the observed sample when modelling extremes. Thus, the different results produced for presumed structural independence or long-term dependence are compared to each other. For this reason, a density graph is created showing which stations are affected the most by their dependence structure and if the Hurst coefficient is correlated to this change. The aforementioned comparison is made for the most prominent stations.

Moreover, more comparisons between other characteristics from the fitting process are produced. In specific, since, the GHCN – Daily database provides globally distributed data, this study attempts at constructing effective estimation of average rainfall modelling characteristics based on each region’s climatic attributes.

Table 8.1: Indicative sample of fitting results for the first 20 stations. Each method is represented by its first letter; "K" for knowable moments, "M" for classic moments, and "L" for L – moments.

N	Name	Lat	Lon	Country	Obs	Rain Obs	Years Obs	Avrg	Avrg Tot	StDev	StDev Tot	Hurst	k (K)	λ (K)	LSE	k (M)	λ (M)	k (L)	λ (L)	KM% (T100)	KM% (T1000)	KL% (T100)	KL% (T1000)
11	AG000060390	36.717	3.250	Algeria	25839	7122	71	7.091	1.954	10.260	6.249	0.73	0.114	8.550	2.984	0.261	5.239	0.396	4.283	45.8	90.6	192.2	405.9
12	AG000060590	30.567	2.867	Algeria	31313	780	86	4.504	0.112	7.157	1.329	0.65	0.306	3.300	0.032	0.302	3.144	0.325	3.039	-6.5	-7.1	0.7	4.7
13	AG000060611	28.050	9.633	Algeria	19559	373	54	4.061	0.077	5.988	0.995	0.5	0.226	3.500	0.509	0.27	2.964	0.298	2.851	1.4	10.5	7.8	26
14	AG000060680	22.800	5.433	Algeria	21768	698	60	4.437	0.142	6.078	1.339	0.58	0.034	5.800	3.052	0.234	3.4	0.24	3.372	35.9	79.3	40	85.8
15	AGE00135039	35.730	0.650	Algeria	26848	4233	74	7.132	1.124	9.367	4.537	0.75	0.056	8.900	0.468	0.21	5.633	0.268	5.218	40.3	81.8	81.3	162.9
17	AGE00147705	36.780	3.070	Algeria	21287	5828	59	7.200	1.971	10.319	6.281	0.47	0.114	8.150	0.205	0.257	5.352	0.403	4.295	50.8	95.5	220.9	464.8
20	AGE00147708	36.720	4.050	Algeria	24415	6112	67	9.278	2.323	11.317	6.944	0.59	0.108	9.300	1.261	0.164	7.757	0.189	7.525	13.5	24.7	27	46.3
21	AGE00147709	36.630	4.200	Algeria	19744	5080	55	10.700	2.753	12.392	7.835	0.59	0.001	13.250	4.889	0.127	9.338	0.135	9.254	35.3	58.1	40.1	66.3
23	AGE00147711	36.370	6.620	Algeria	16343	3782	45	6.884	1.593	8.270	4.925	0.62	0.092	7.100	0.108	0.153	5.828	0.159	5.789	15.2	27	18.4	31.1
24	AGE00147712	36.170	1.340	Algeria	18101	3837	50	5.517	1.169	6.789	3.854	0.44	0.02	7.050	0.798	0.17	4.579	0.222	4.29	39.8	76	78.9	146.1
25	AGE00147713	36.180	5.400	Algeria	14734	3511	41	5.188	1.236	7.254	4.174	0.56	0.12	5.750	0.509	0.244	3.921	0.34	3.426	41.3	75.3	126.3	240.5
27	AGE00147715	35.420	8.120	Algeria	12886	2054	36	5.663	0.903	7.158	3.530	0.7	0.196	4.250	1.861	0.187	4.603	0.226	4.38	2.4	1.5	22	29.8
28	AGE00147716	35.100	-1.850	Algeria	25809	4085	71	7.646	1.210	9.336	4.646	0.69	0.016	9.700	1.588	0.165	6.388	0.192	6.178	36.3	74.1	53.2	105.7
29	AGE00147717	35.200	0.630	Algeria	16103	2844	45	6.329	1.118	7.332	3.914	0.52	0.02	7.450	0.797	0.127	5.522	0.133	5.488	28.2	48.1	30.1	52.4
30	AGE00147718	34.850	5.720	Algeria	26775	2084	74	5.782	0.450	10.125	3.221	0.65	0.374	3.500	0.038	0.337	3.833	0.384	3.561	-11.2	-17.8	7.8	10.1
31	AGE00147719	33.800	2.890	Algeria	16716	1403	46	5.449	0.457	7.320	2.603	0.48	0.224	4.300	0.412	0.223	4.234	0.225	4.222	-1.9	-2	-1.1	-0.9
32	AGE00147720	33.680	1.000	Algeria	14558	2544	40	5.701	0.996	8.619	4.203	0.76	0.098	8.300	8.389	0.281	4.097	0.307	3.953	36	90.7	53.1	125.8
36	AGM00060360	36.822	7.809	Algeria	11489	3350	32	6.927	2.020	9.538	6.036	0.69	0.118	7.950	0.393	0.236	5.291	0.3	4.852	34	64.1	83.9	156
40	AGM00060402	36.712	5.070	Algeria	10958	3050	31	8.349	2.324	11.955	7.333	0.42	0.118	9.550	0.831	0.256	6.21	0.38	5.175	46.1	87.8	175.5	357.5
46	AGM00060419	36.276	6.620	Algeria	10800	2958	30	5.818	1.593	8.557	5.175	0.82	0.178	6.450	3.261	0.269	4.253	0.339	3.845	14.9	36.5	63.7	125.3

N	RMSE (K)	RMSE (M)	RMSE (L)	RMSE High (K)	RMSE High (M)	RMSE High (L)	RMSE Low (K)	RMSE Low (M)	RMSE Low (L)	NRMSE (K)	NRMSE (M)	NRMSE (L)	NRMSE High (K)	NRMSE High (M)	NRMSE High (L)	NRMSE Low (K)	NRMSE Low (M)	NRMSE Low (L)	Rainfall (T100)	Rainfall (T1000)
11	5.735	11.420	63.375	7.264	16.356	91.266	3.798	1.715	2.591	0.864	0.729	-0.502	0.737	0.407	-2.309	0.731	0.879	0.817	138.747	202.719
12	5.142	6.785	5.304	6.330	8.345	6.520	0.531	0.506	0.501	0.786	0.718	0.779	0.713	0.622	0.705	0.836	0.843	0.845	78.500	169.348
13	3.067	3.654	2.780	3.665	4.431	3.367	1.011	0.514	0.460	0.793	0.754	0.813	0.707	0.645	0.731	0.501	0.747	0.773	52.582	97.886
14	2.565	2.557	2.595	2.510	3.174	3.225	2.654	0.795	0.787	0.822	0.823	0.820	0.736	0.666	0.661	0.272	0.782	0.784	44.638	64.045
15	3.631	7.879	19.486	3.895	10.861	27.096	3.325	1.535	0.961	0.879	0.738	0.353	0.785	0.401	-0.494	0.710	0.866	0.916	98.939	132.841
17	5.304	11.633	64.022	7.365	16.793	93.151	2.218	2.048	2.434	0.862	0.697	-0.665	0.709	0.336	-2.684	0.848	0.859	0.833	132.388	193.383
20	5.510	2.795	6.344	7.666	3.985	9.100	1.927	0.487	0.540	0.871	0.935	0.852	0.731	0.860	0.680	0.872	0.968	0.964	144.450	208.704
21	4.514	7.462	8.873	5.174	10.815	12.883	3.835	1.130	0.948	0.875	0.794	0.755	0.695	0.362	0.240	0.777	0.934	0.945	118.241	151.972
23	2.765	1.101	1.326	3.709	1.540	1.904	1.546	0.465	0.389	0.899	0.960	0.952	0.774	0.906	0.884	0.859	0.958	0.964	98.634	140.345
24	2.381	4.242	10.659	2.181	6.111	15.441	2.549	0.642	0.424	0.882	0.789	0.470	0.786	0.399	-0.519	0.712	0.927	0.952	67.541	87.537
25	3.149	4.471	20.137	4.300	6.437	29.746	1.636	1.378	1.036	0.876	0.824	0.205	0.730	0.596	-0.869	0.834	0.860	0.895	92.204	137.485
27	4.021	4.099	6.329	5.473	5.858	9.191	1.989	1.136	0.850	0.813	0.810	0.706	0.655	0.630	0.420	0.774	0.871	0.904	96.594	163.372
28	3.310	6.179	10.495	3.373	8.545	14.614	3.243	1.158	0.901	0.880	0.776	0.620	0.781	0.445	0.052	0.714	0.898	0.921	89.101	112.682
29	2.206	2.033	2.385	2.431	2.887	3.407	1.975	0.510	0.480	0.892	0.900	0.883	0.769	0.726	0.676	0.778	0.943	0.946	69.128	89.595
30	13.414	17.728	10.303	17.876	23.645	13.736	0.898	0.547	0.544	0.742	0.659	0.802	0.648	0.534	0.729	0.876	0.924	0.925	175.192	427.132
31	7.439	7.935	7.778	10.229	10.910	10.694	0.235	0.306	0.299	0.727	0.709	0.715	0.568	0.539	0.548	0.963	0.952	0.953	95.727	174.783
32	8.001	8.028	8.663	8.576	11.679	12.605	7.453	0.376	0.332	0.762	0.761	0.742	0.566	0.409	0.362	0.203	0.960	0.964	115.141	164.010
36	4.758	2.915	14.392	6.086	4.183	21.958	3.439	1.318	0.706	0.864	0.917	0.589	0.710	0.801	-0.047	0.748	0.904	0.948	132.117	195.729
40	6.618	5.416	39.265	9.413	7.887	59.848	3.174	2.146	2.553	0.843	0.871	0.067	0.642	0.700	-1.279	0.809	0.871	0.846	158.269	232.853
46	7.174	7.167	11.770	9.608	10.916	17.915	4.574	0.829	1.474	0.801	0.802	0.674	0.614	0.562	0.281	0.586	0.925	0.867	149.667	243.759

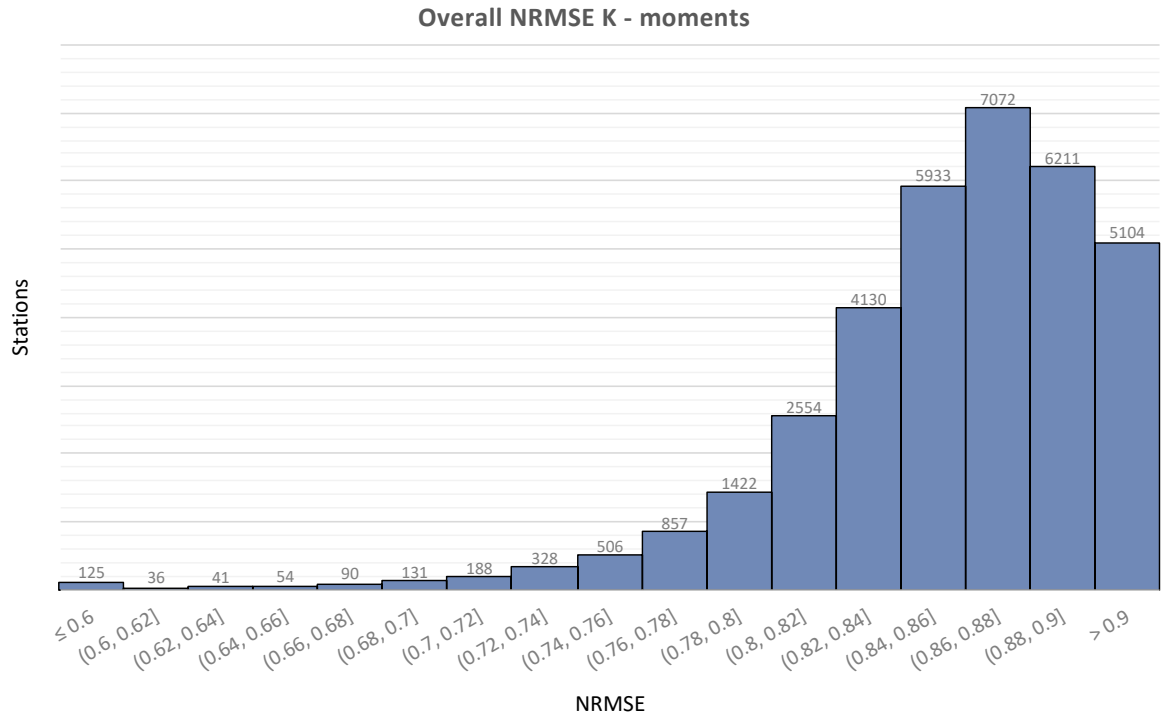
8.2 Fitting Methods Comparative Performance – Goodness-of-fit

8.2.1 Overall Performance

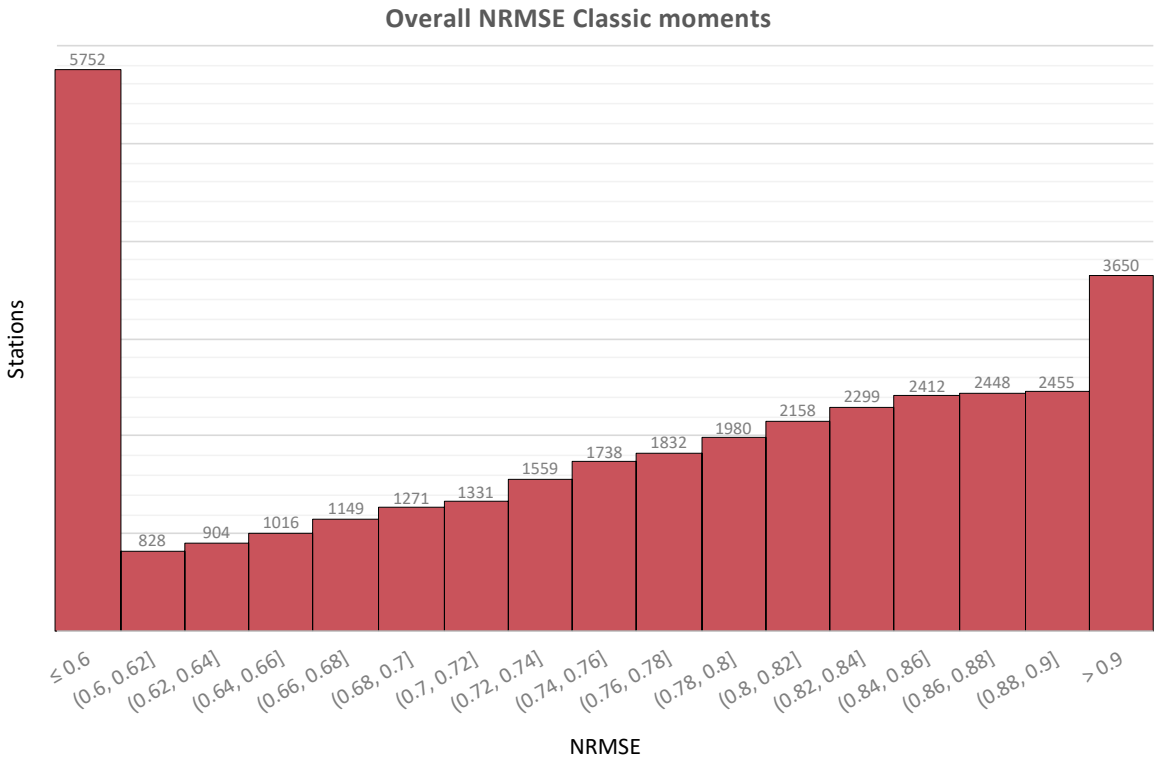
As mentioned before, the evaluation of a method's performance is done by comparing goodness-of-fit statistic tools. In this study the used tools are the RMSE and NRMSE. In this chapter overall method performance is analysed.

Firstly, NRMSE charts are presented. The provided charts are modified to show values in the same range for each method, in order to make the comparison more evident. It is observed from random station tests that an overall NRMSE value over 0.7 shows acceptable compatibility between observed values and the theoretical distribution. This NRMSE value should not be confused with explicit measurement of reliability in the distribution tail, but is an overall indicator of the whole distribution fit. However, again through sampling of different stations, a value over 0.8 suggests reliability in both low and high orders.

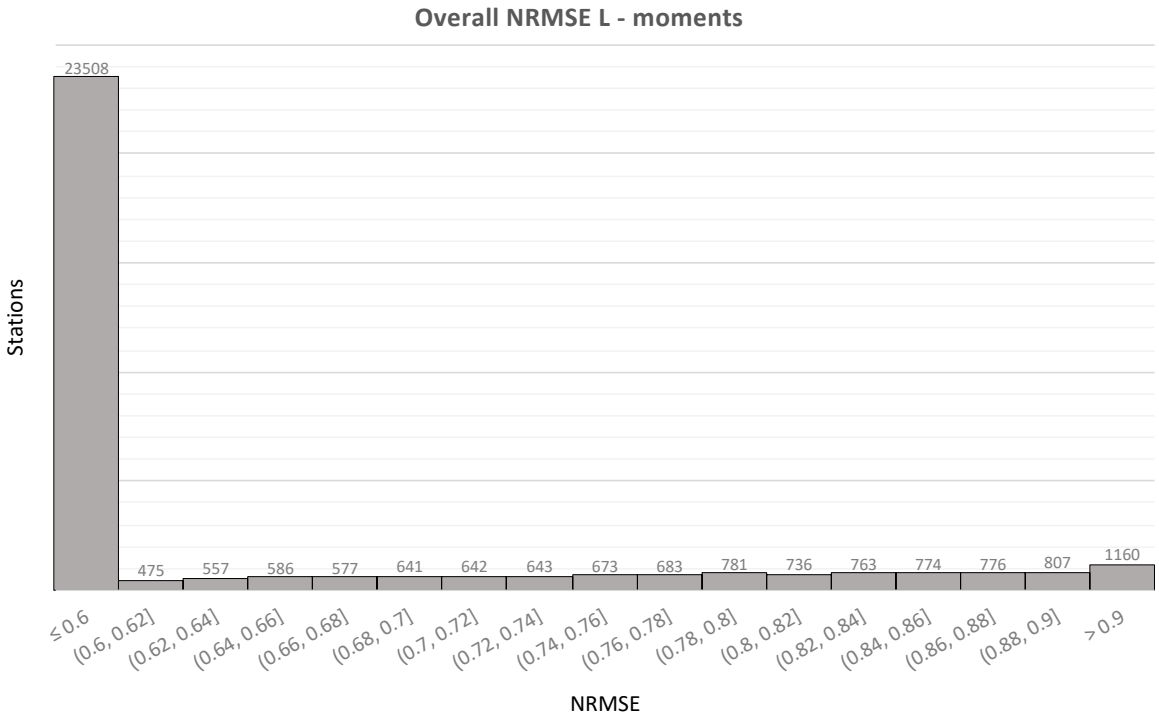
From graphs Graph 8.1, Graph 8.2, and Graph 8.3 it is evident that fitting with the K – moments method using the two parameter Generalized Pareto Distribution proves to be the most efficient overall. Most of the stations are well over the 0.7 range, with most of them even above 0.8 suggesting great performance overall. On the other hand, from the classic methods, classic moments achieve second best performance with L – moments achieving the worst result. As seen from Graph 8.2 and Graph 8.3 over 5,500 and 23,500 stations respectively are below the 0.6 mark, with many of them even on the negative range.



Graph 8.1: Overall NRMSE values - Knowable moments



Graph 8.2: Overall NRMSE values - Classic moments

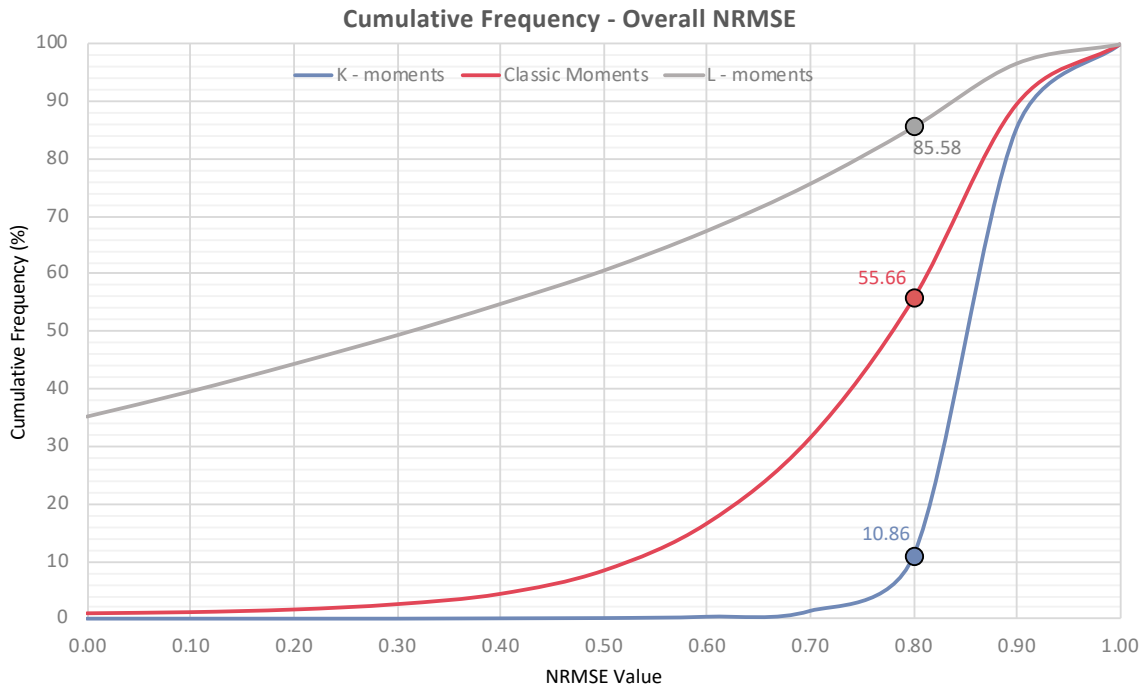


Graph 8.3: Overall NRMSE values - L-moments

In order to better illustrate the comparison between methods, Graph 8.4 provides the cumulative frequencies of NRMSE values for all methods for every tested station. As shown, K – moments NRMSE frequencies start the steep climb after the 0.7 mark. However, this is not the case for classic and L – moments which show significant frequency even for low NRMSE values.

The 0.8 threshold is presented for all methods in the same graph. The data label shows the percentage of stations with NRMSE values below 0.8. K – moments are in the range of 11%, while L – moments on the other end of the spectrum show a calculated percentage of 86%. Classic moments give an in-between result of 56%.

Since this section provides the overall fitting result, it still can't be assumed that K – moments show the best extreme-oriented distribution fitting. For pointing to this conclusion, a deeper analysis on how the fit performs specifically in high orders is needed.

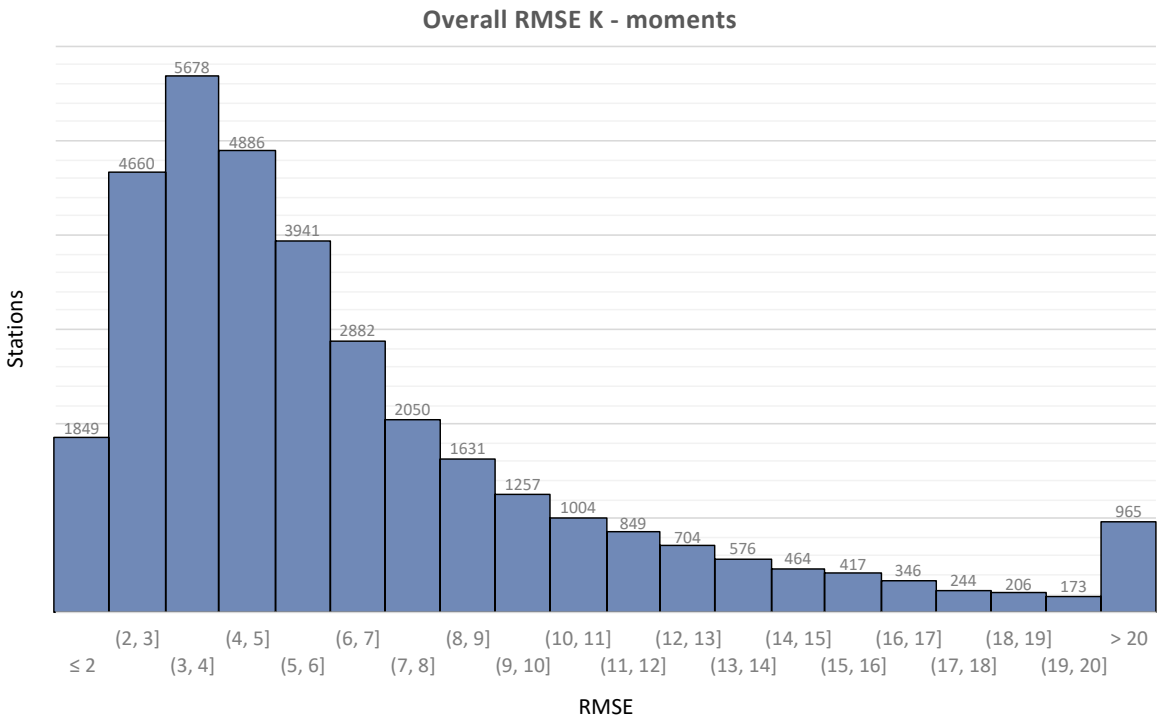


Graph 8.4: Cumulative frequency of overall NRMSE for all methods. The data labels show the percentage of stations where the estimated NRMSE value is below 0.8.

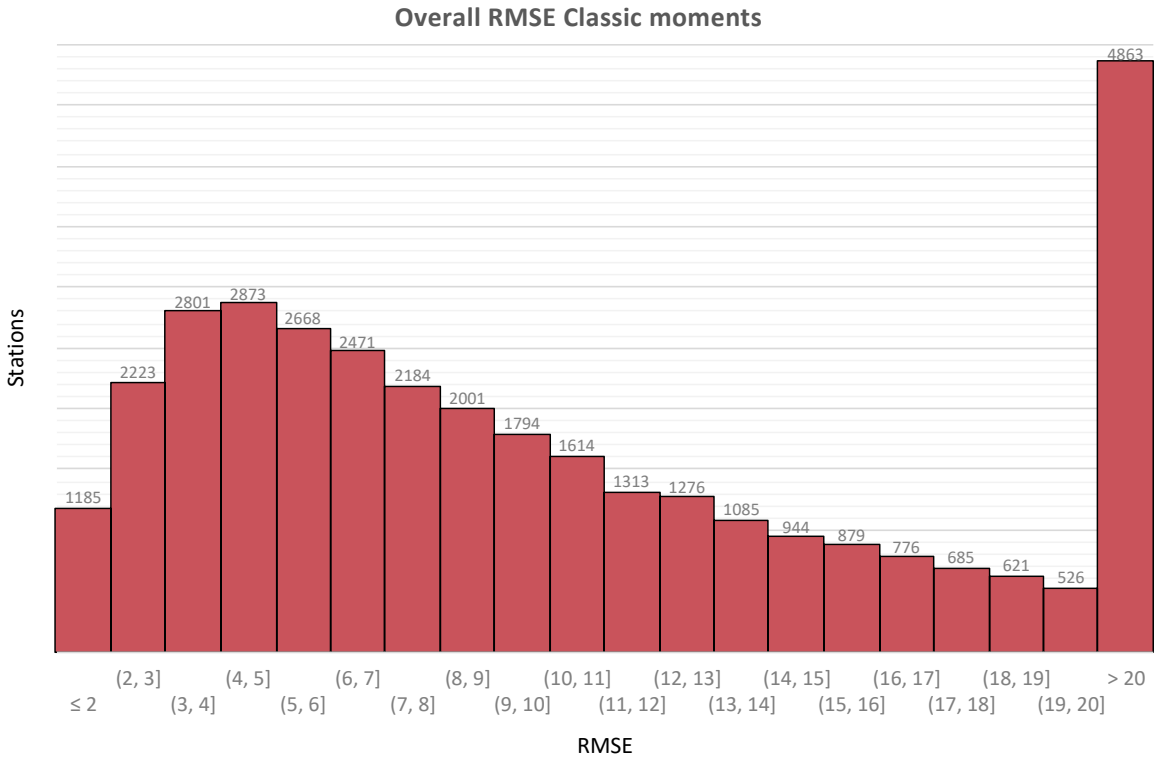
These comparisons were made using the NRMSE tool. Now, the same results are presented using the standard RMSE. While, RMSE shows reliability by how close to zero the error is, maximum values aren't theoretically defined. As with NRMSE, from trial tests in different stations, RMSE values are considered reliable enough if below 4-6.

The problem with RMSE is that while it gives a great representation of the fit, it is biased for high values, meaning that equally sufficient fit is achieved for stations with different valued extremes. A station with higher observed extreme values, compared to another with lower extremes, whilst may have identical fit results, will most certainly show a greater RMSE value which is not representative. Thus, main focus is given in the normalised NRMSE value which is independent of the sample size and intensity of extreme values. Although this is

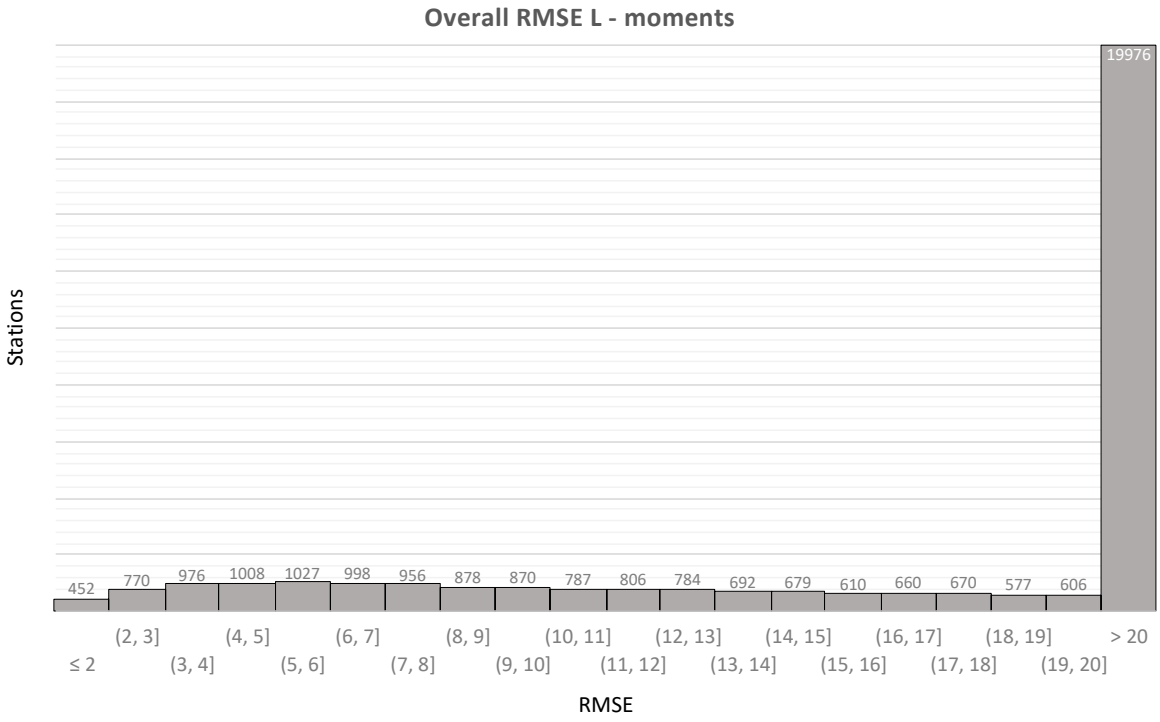
the case, RMSE results follow the same pattern as with the NRMSE, suggesting overall better fit for K – moments (Graph 8.5), followed by classic moments (Graph 8.6), and finally L – moments (Graph 8.7).



Graph 8.5: Overall RMSE values - Knowable moments

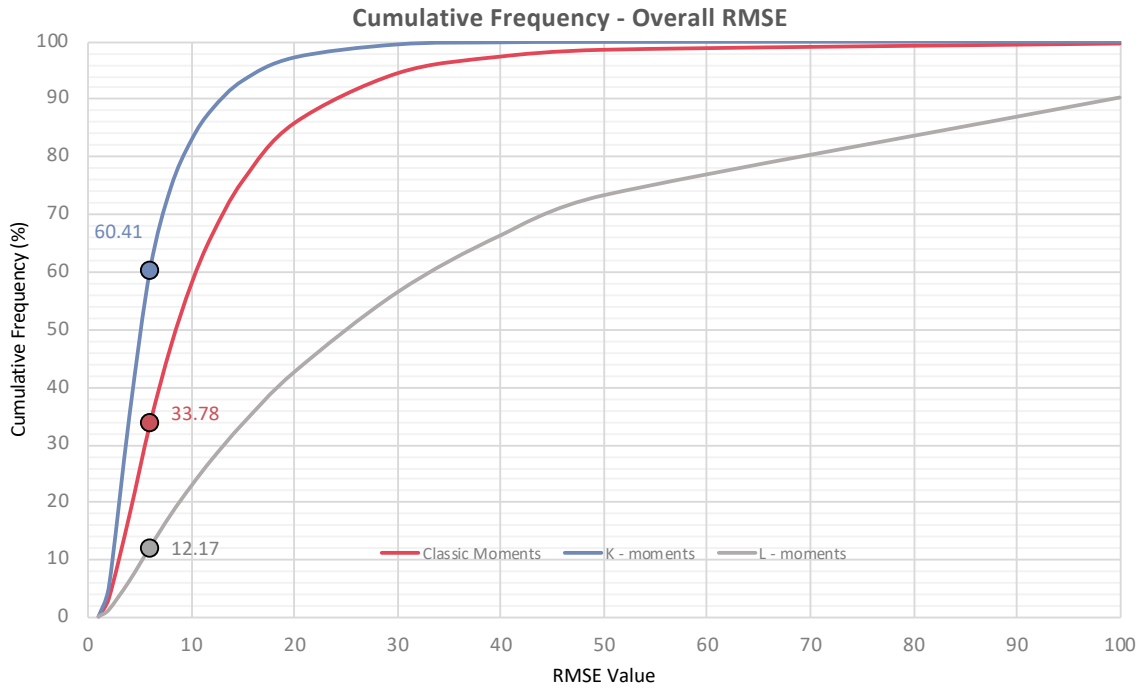


Graph 8.6: Overall RMSE values - Classic moments



Graph 8.7: Overall RMSE values - L-moments

Again, for better comparison between methods, a cumulative frequency graph for RMSE values is shown below. Depicted is the percentage of stations below RMSE value of 6, which is equivalent to an overall reliable fit. As with NRMSE, here again, knowable moments prevail over classic methods, with almost double target frequency against classic moments.

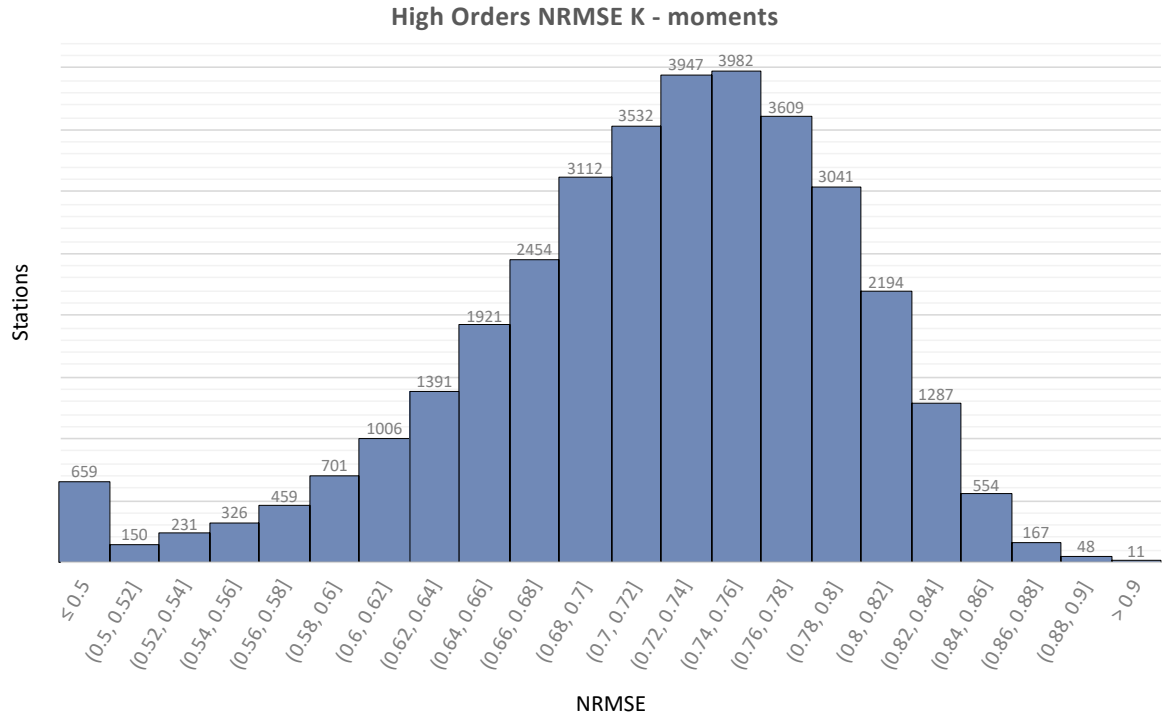


Graph 8.8: Cumulative frequency of overall RMSE for all methods. The data labels show the percentage of stations where the estimated RMSE value is below 6.

8.2.2 High Order Performance - Extremes

While an overall performance statistic is most of the time sufficient to judge a distribution's power to suitably match observed data, in this study the main focus as mentioned before is extreme-oriented rainfall modelling. For this purpose, as already tested in Chapter 7 NRMSE and RMSE statistics are separated and used upon different parts of the distribution, in order to showcase performance in both the tail and the body. Goodness-of-fit values are estimated as in section 7.2.

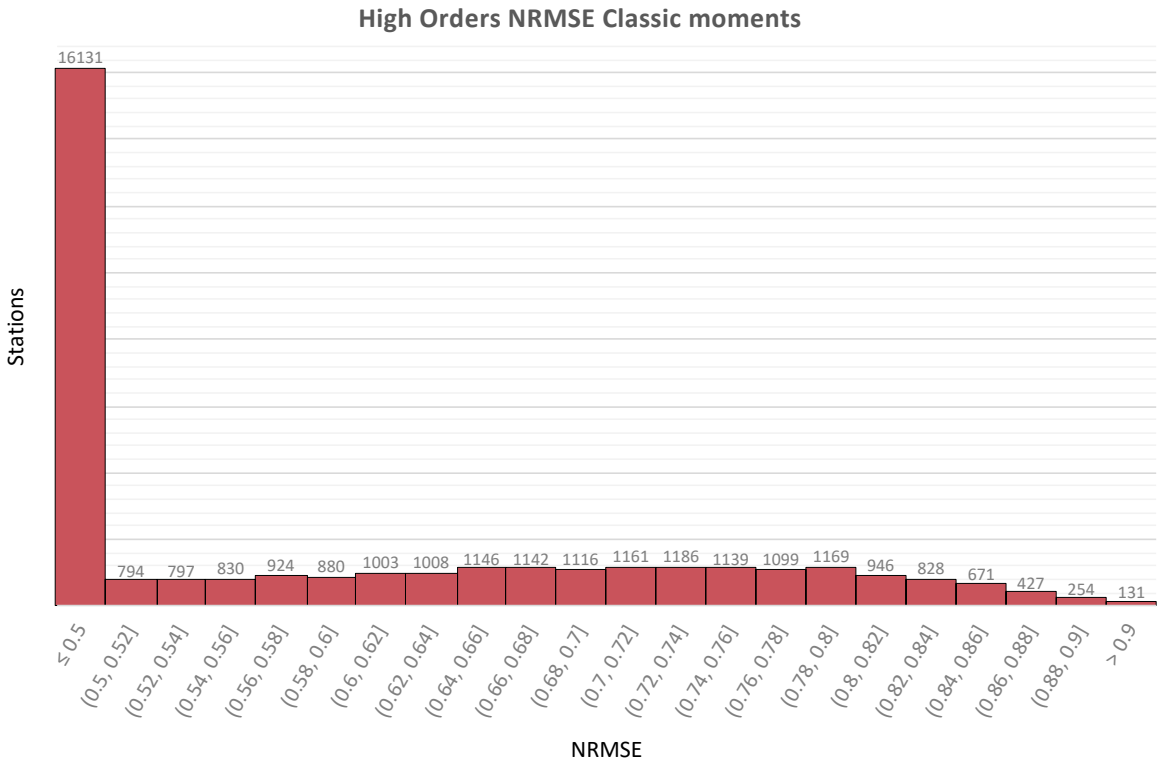
For high moment orders, thus as extremes are concerned, the results are provided below in the same format as the overall goodness-of-fit values. From Graph 8.9, Graph 8.10, and Graph 8.11, it is evident that using the K – moments method gives significantly more reliable results. Classic methods fail to accurately model the tail of the distribution, while classic moments are the best between the two.



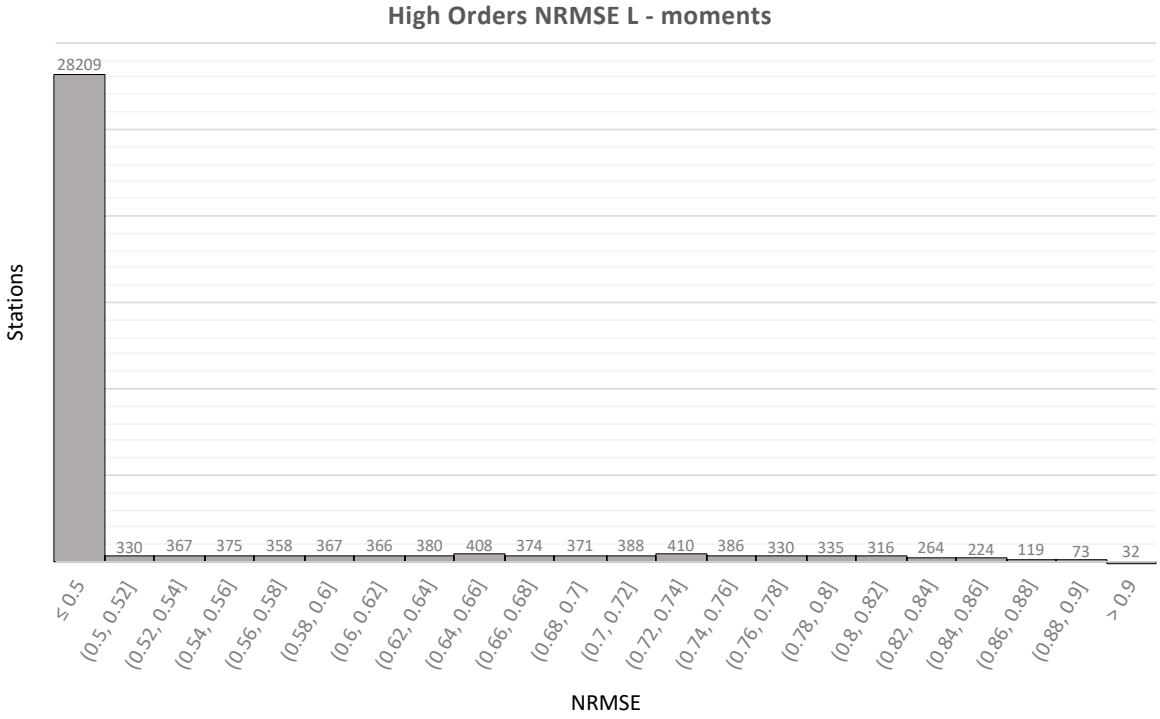
Graph 8.9: High Orders (T > 1 year) NRMSE values - Knowable moments

Using the same principle as before (Section 8.2.1) values over 0.7 consist of reliable tail fits. Thus, by comparing the three methods, it is clear that with K – moments there is a significant advantage. Most stations give a result around 0.75 with K – moments which is considerably higher than classic and L – moments which show average results below 0.5.

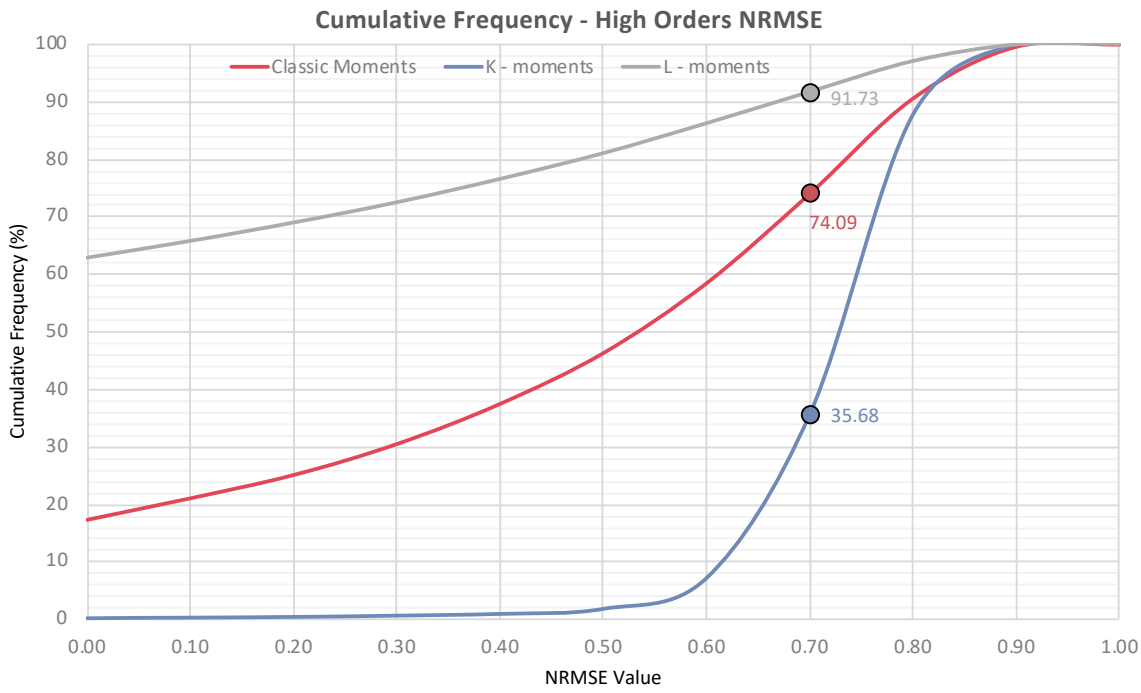
A better representation of this is again shown in the cumulative frequency chart (Graph 8.12), where the threshold 0.7 is overpassed by almost 65% of stations modelled with K – moments, whereas the next closest in reliability are classic moments with 25% of stations overpassing 0.7 in high-order NRMSE.



Graph 8.10: High orders (T > 1 year) NRMSE values - Classic moments



Graph 8.11: High Orders (T > 1 year) NRMSE values - L-moments

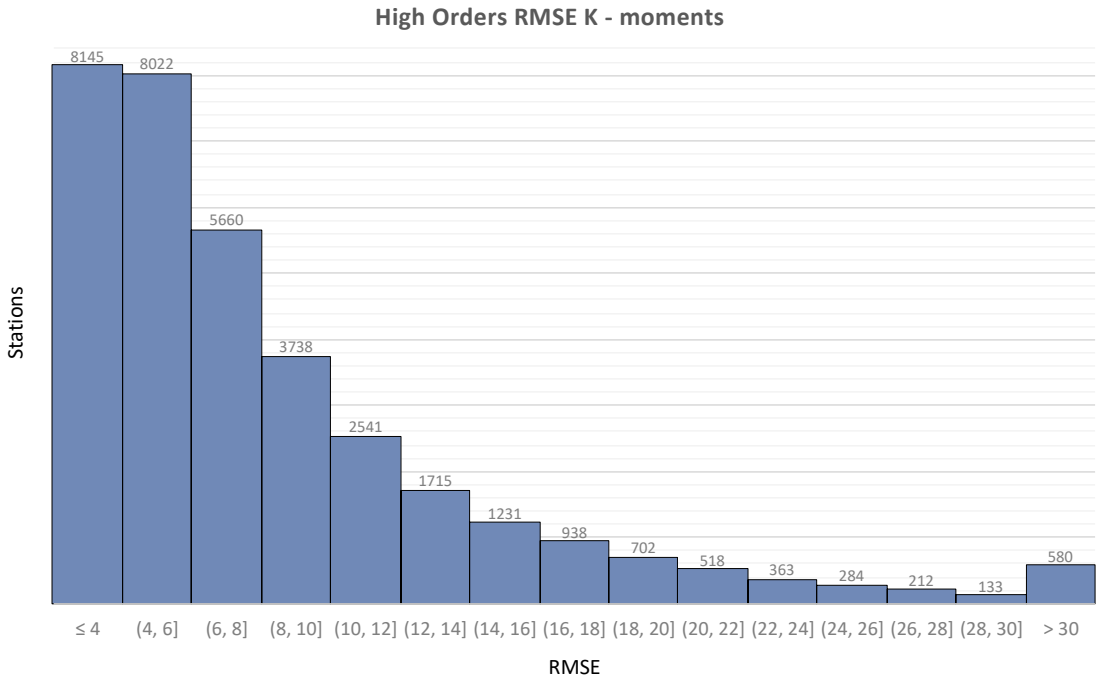


Graph 8.12: Cumulative frequency of high-order NRMSE values for all methods. The data labels show the percentage of stations where the estimated NRMSE value is below 0.7.

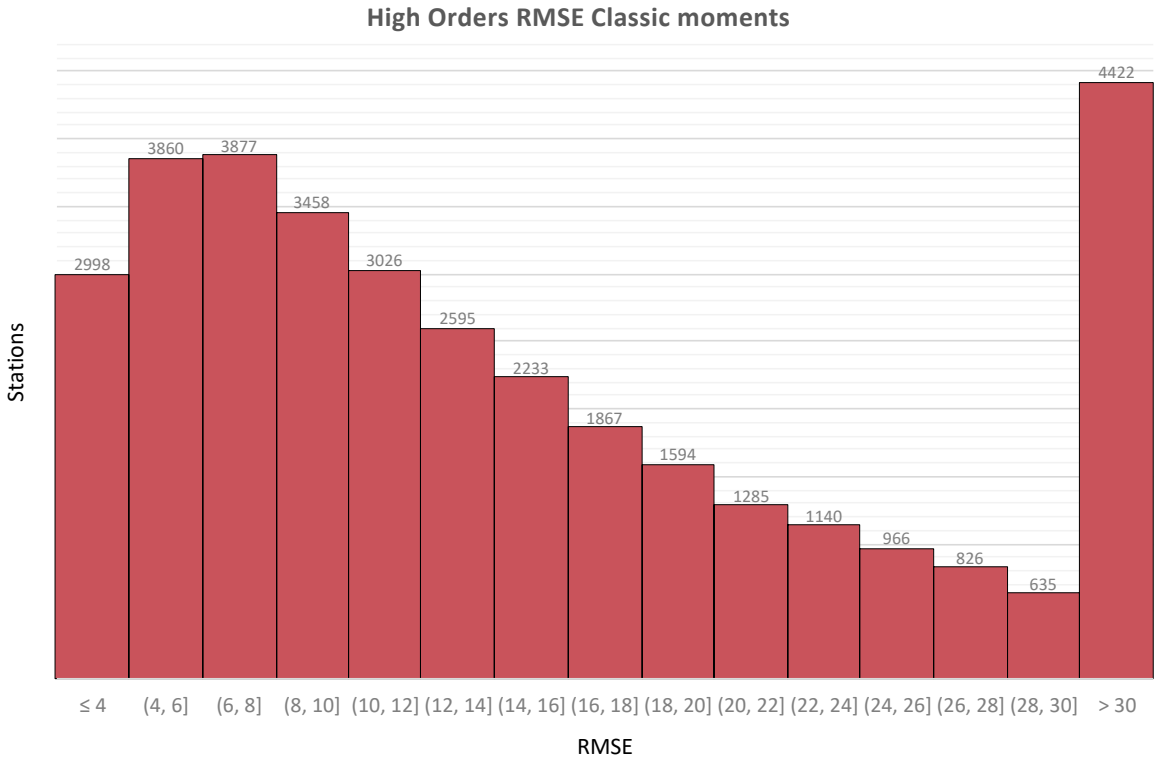
Again, for validation purposes, RMSE values are also estimated and shown below in Graph 8.13, Graph 8.14, and Graph 8.15. The same pattern arises from both evaluation methods, showing better fit of the K – moments method for extreme values. Using K – moments, most stations give estimated RMSE in the range of 1-6. As for classic moments, RMSE values are distributed among the spectrum provided, but the number of stations above 30 is significantly greater. Finally, L – moments show again the worst result with most stations showing RMSE above 30.

Moreover, the cumulative frequency chart for RMSE (Graph 8.16) gives comparable results to the NRMSE (Graph 8.12). Specifically, it estimates that 46% of stations are lower than the applied threshold (equal to 6, as before), while for classic and L – moments, this percentage is in the range of 20% and 7%, respectively.

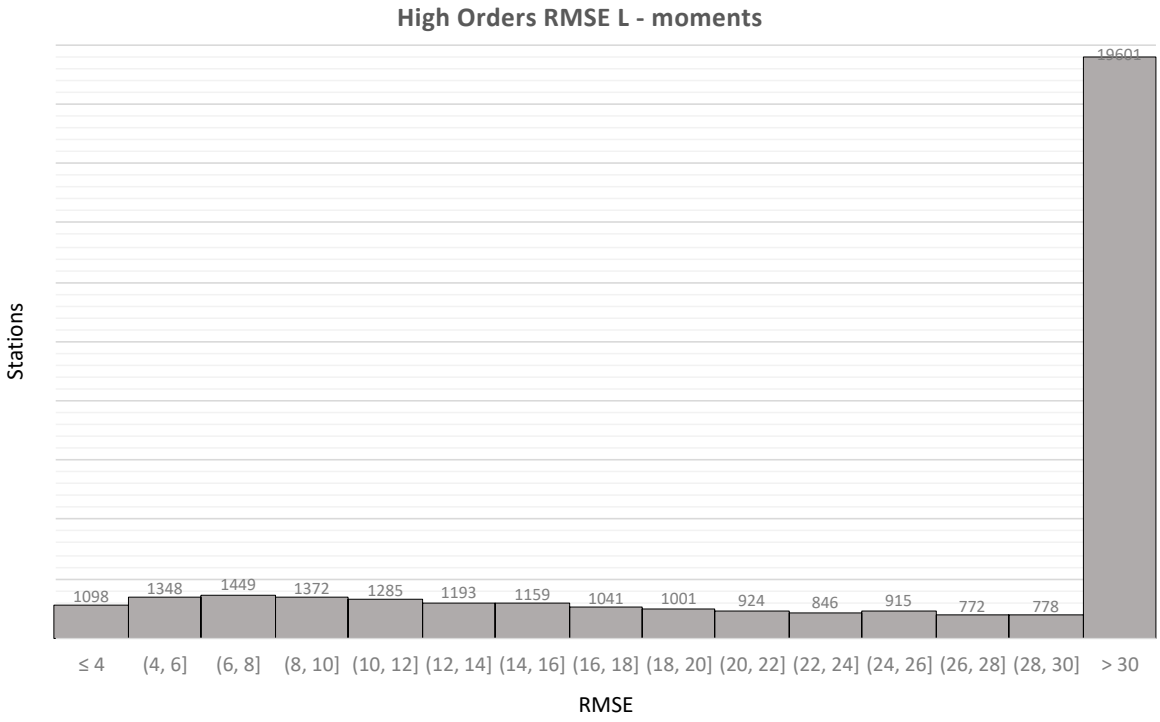
Both goodness-of-fit statistics, show clear preference of the K – moments method as the best for extreme-oriented rainfall modelling. The results provided analytically for the sample station are validated throughout the entirety of the dataset, specifically when focusing on high-orders, which is the primary goal of this study.



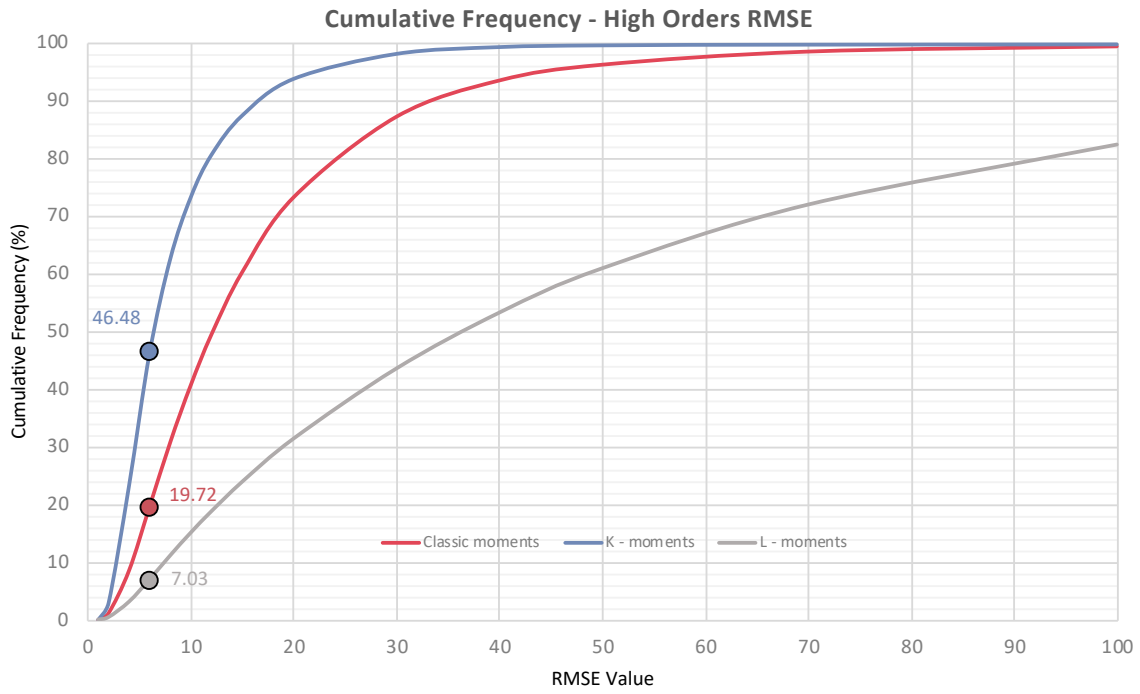
Graph 8.13: High Orders (T > 1 year) RMSE values - Knowable moments



Graph 8.14: High Orders (T > 1 year) RMSE values - Classic moments



Graph 8.15: High Orders (T > 1 year) RMSE values - L-moments



Graph 8.16: Cumulative frequency of high-order RMSE values for all methods. The data labels show the percentage of stations where the estimated RMSE value is below 6.

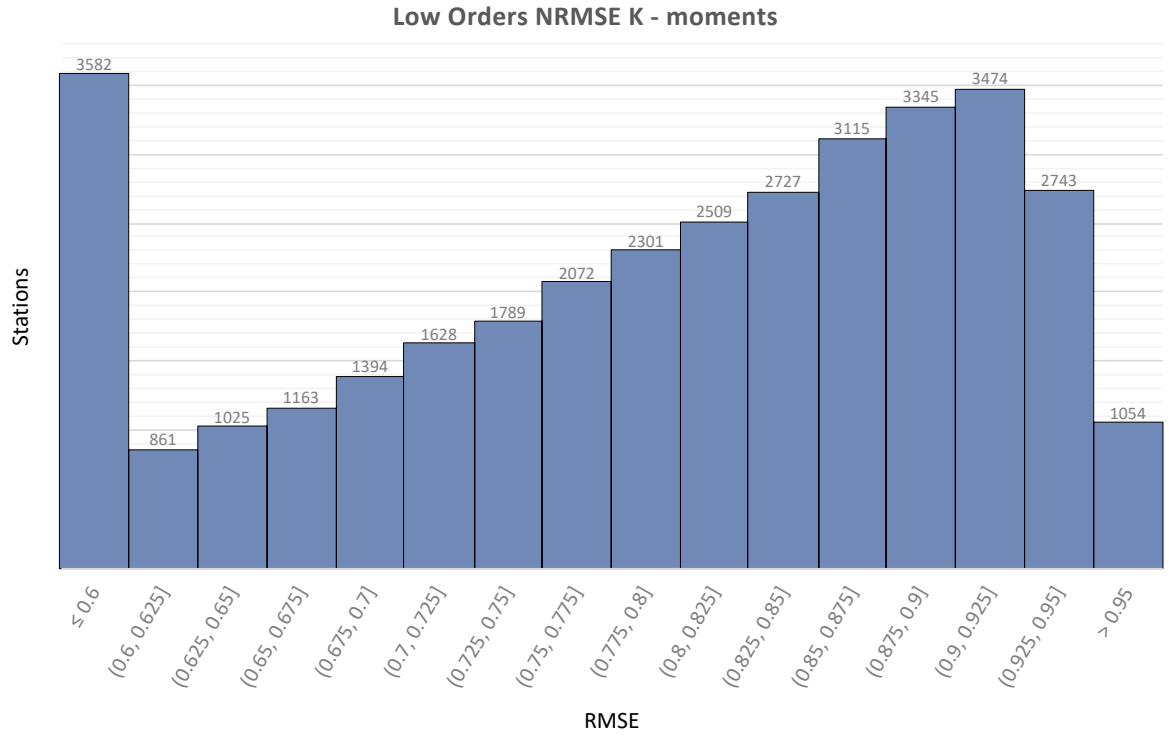
8.2.3 Low Order Performance

While extremes modelling efficiency is achieved with the K – moments method, the ultimate goal is to find a reliable modelling method for overall best fitting results. Since preliminary analysis indicated overall superiority of K – moments compared to classic methods, a definitive conclusion can be reached if superiority is also attained for low-order moments (i.e. the distribution's body). For this purpose, the same procedure as in Section 8.2.2 will be followed, showing goodness-of-fit statistics for low-order moments.

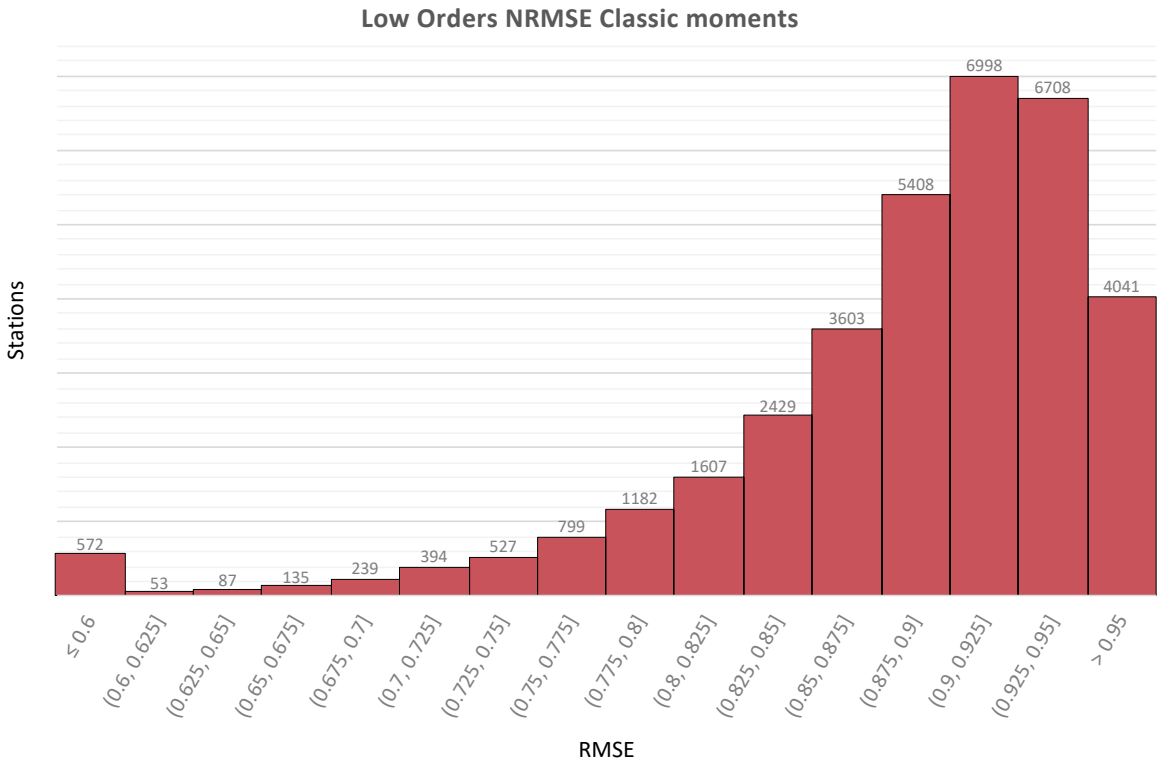
Firstly, NRMSE histograms depict each methods behaviour for low orders, which consist of values for return periods lower than 1 year (T < 1 year). All methods show similarly good fitting performance to each other, with K – moments (Graph 8.17) achieving NRMSE values of 0.7 in most stations. However, over 3,000 stations have estimated NRMSE below 0.6.

On the other hand, both classic methods (Graph 8.18, Graph 8.19) show slightly better performance than K – moments, with NRME values more densely compacted over 0.85 for classic moments and over 0.8 for L – moments. Using classic moments, stations with NRMSE below 0.6 are almost 600, while the same number for L – moments is just above 1,300, both significantly lower the K – moments method.

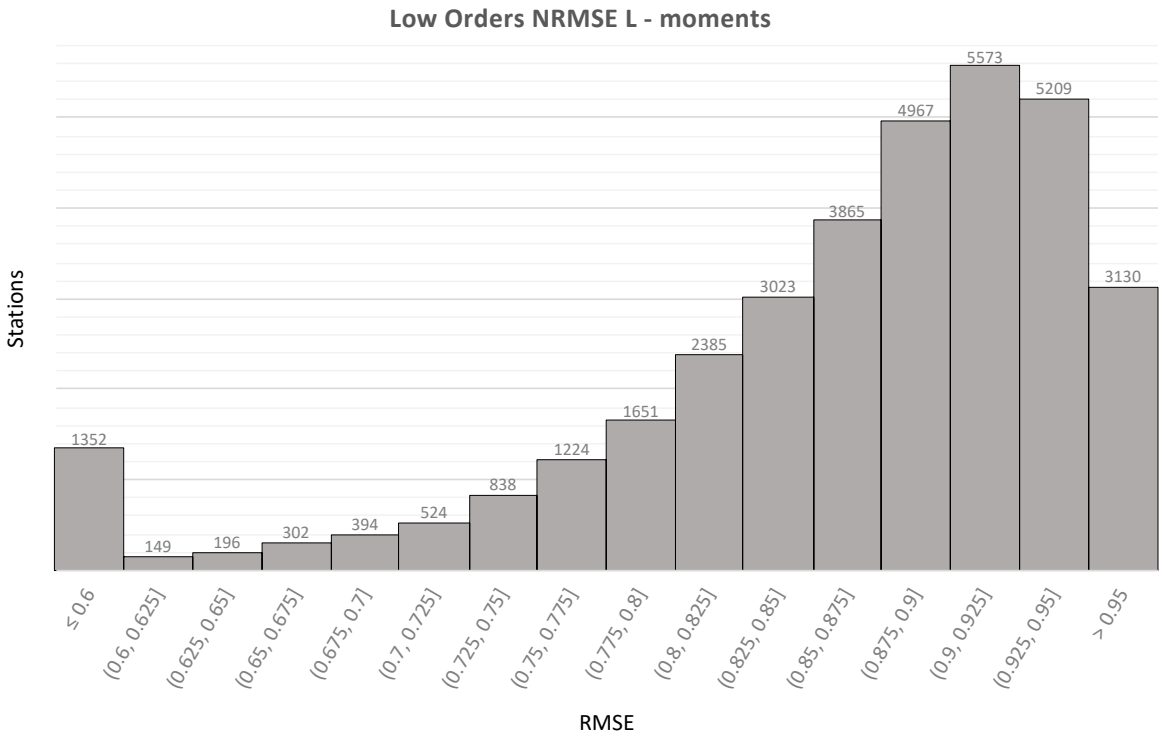
This result is expected when comparing methods, since K – moments are specifically used in focusing the modelling process on extremes, rather than in the distribution's body. While this is the case, the difference in the reliability between knowable and classic methods is minimal.



Graph 8.17: Low Orders (T < 1 year) NRMSE values - Knowable moments

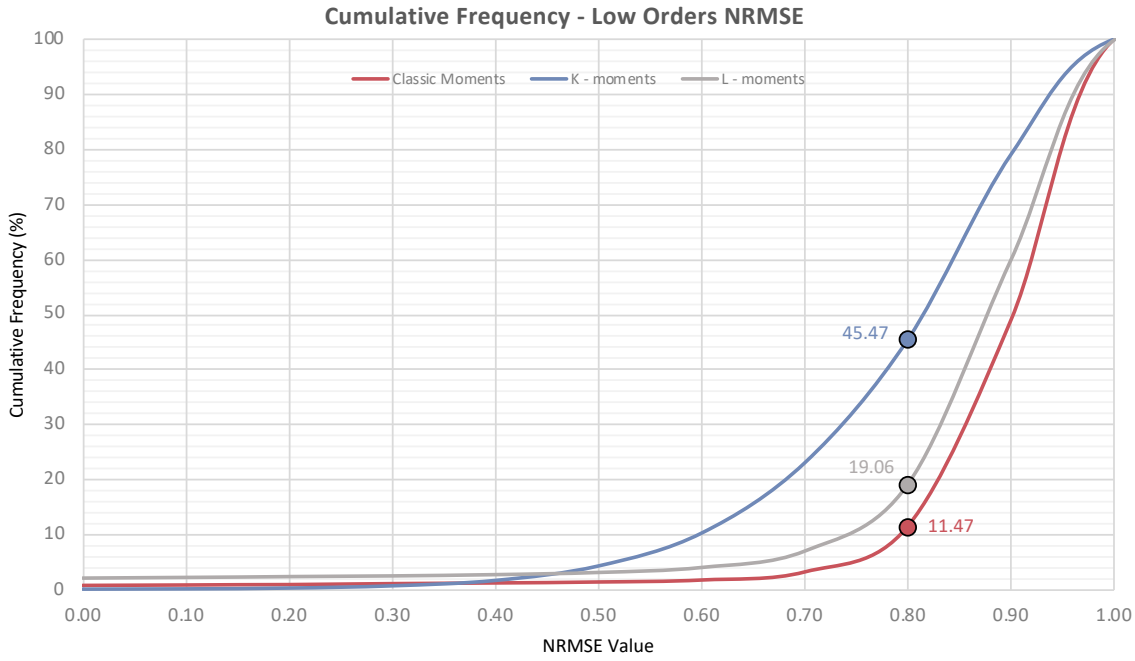


Graph 8.18: Low Orders (T < 1 year) NRMSE values - Classic moments



Graph 8.19: Low Orders (T < 1 year) NRMSE values - L-moments

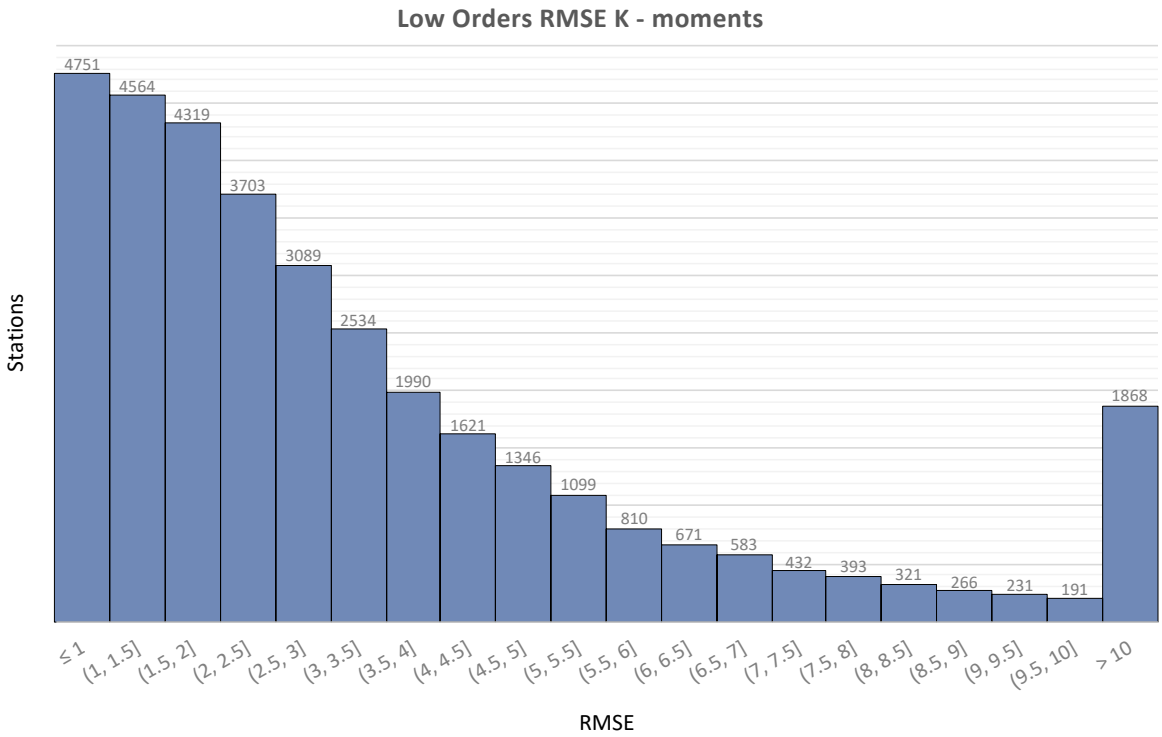
Using the cumulative frequency chart for low-order NRMSE values it is evident that the best performer are classic moments, followed by L – moments, while K – moments provide with the least good results. Classic moments show 11% of stations with low-order NRMSE value below 0.8, L – moments show 19%, and finally K – moments display 45%. While this seems like a significantly higher number compared to the best performer, this threshold gives stations with almost perfect results. Good reliability is achieved even from NRMSE in above 0.7, where K – moments achieve almost 20% stations below that range.



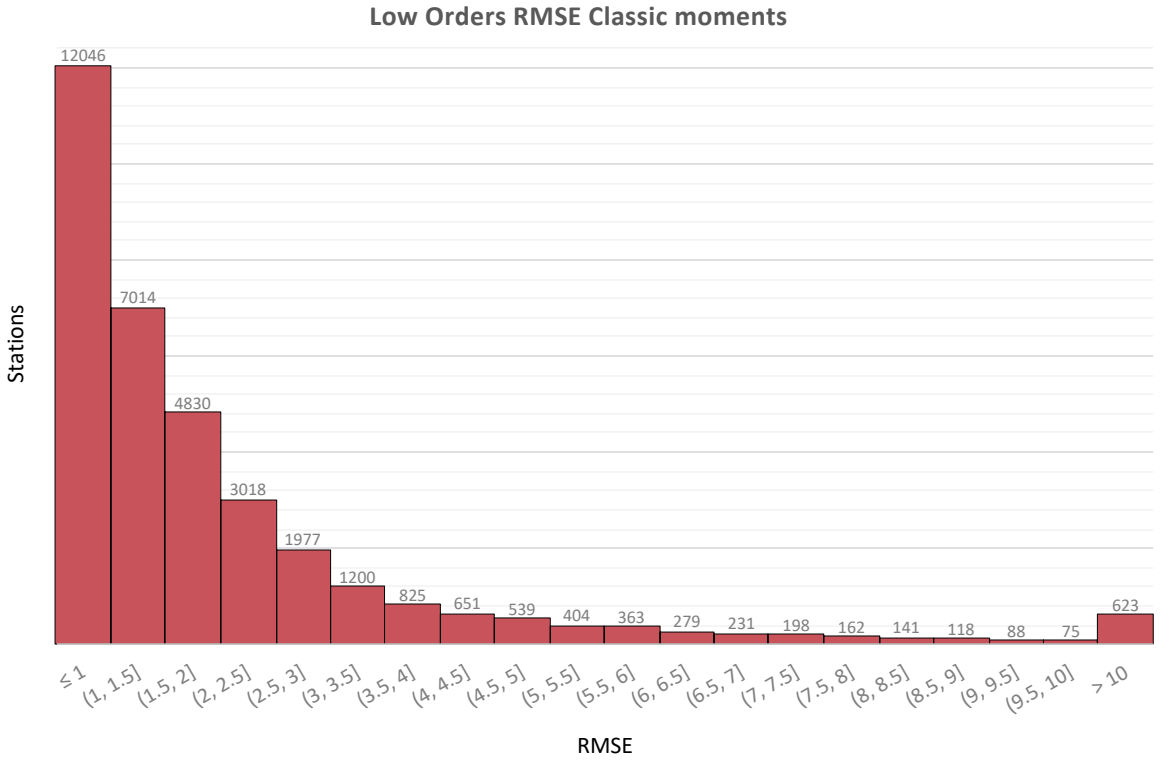
Graph 8.20: Cumulative frequency of low-order NRMSE values for all methods. The data labels show the percentage of stations where the estimated NRMSE value is below 0.8.

The same behaviour is depicted from RMSE estimation of low-order moments, which are provided below, again for validation purposes (Graph 8.21, Graph 8.22, Graph 8.23). K – moments show a tendency for low RMSE values with most of them below the 4 value threshold mark, while classic methods tend to achieve RMSE values concentrated below 2. Moreover, almost triple the stations with over 10 RMSE value are achieved with the use of knowable, rather than classic moments. This again shows the slight advantage of classic methods.

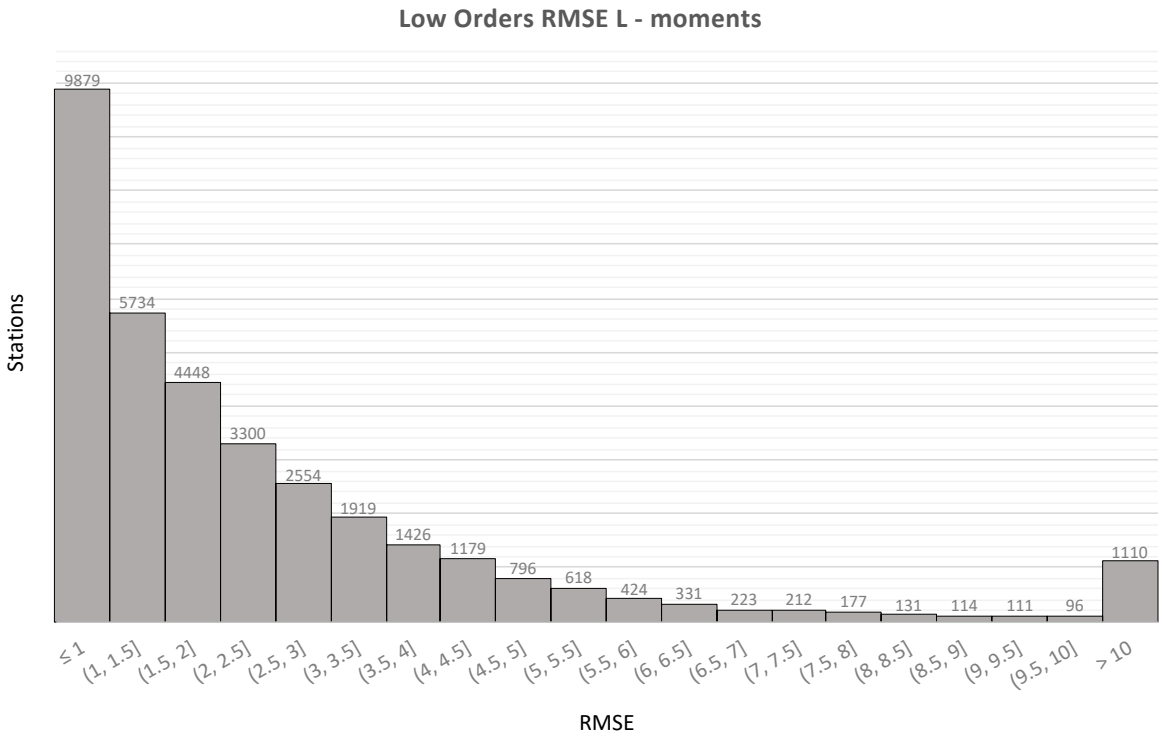
In more detail, from the cumulative frequency Graph 8.24, the slight advantage is again evident. With 4 as threshold, stations below it using K – moments show overall percentage of 71%, classic moments 89%, and L – moments 84%.



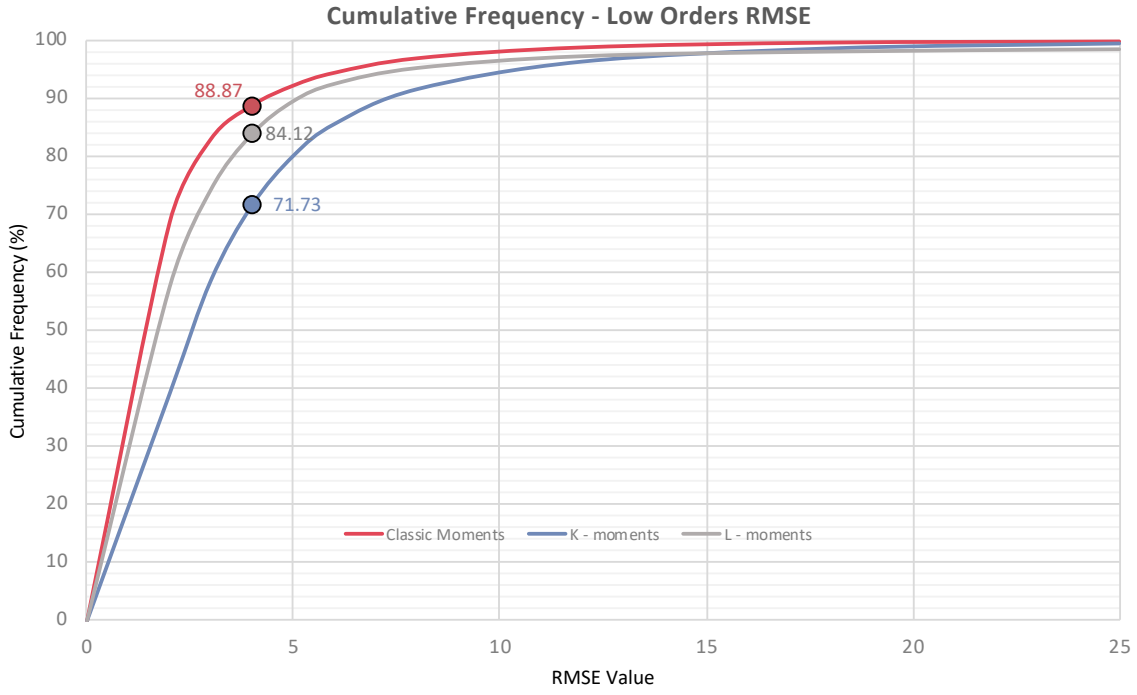
Graph 8.21: Low Orders (T < 1 year) values - Knowable moments



Graph 8.22: Low Orders (T < 1 year) RMSE values - Classic moments



Graph 8.23: Low Orders (T < 1 year) RMSE values - L-moments



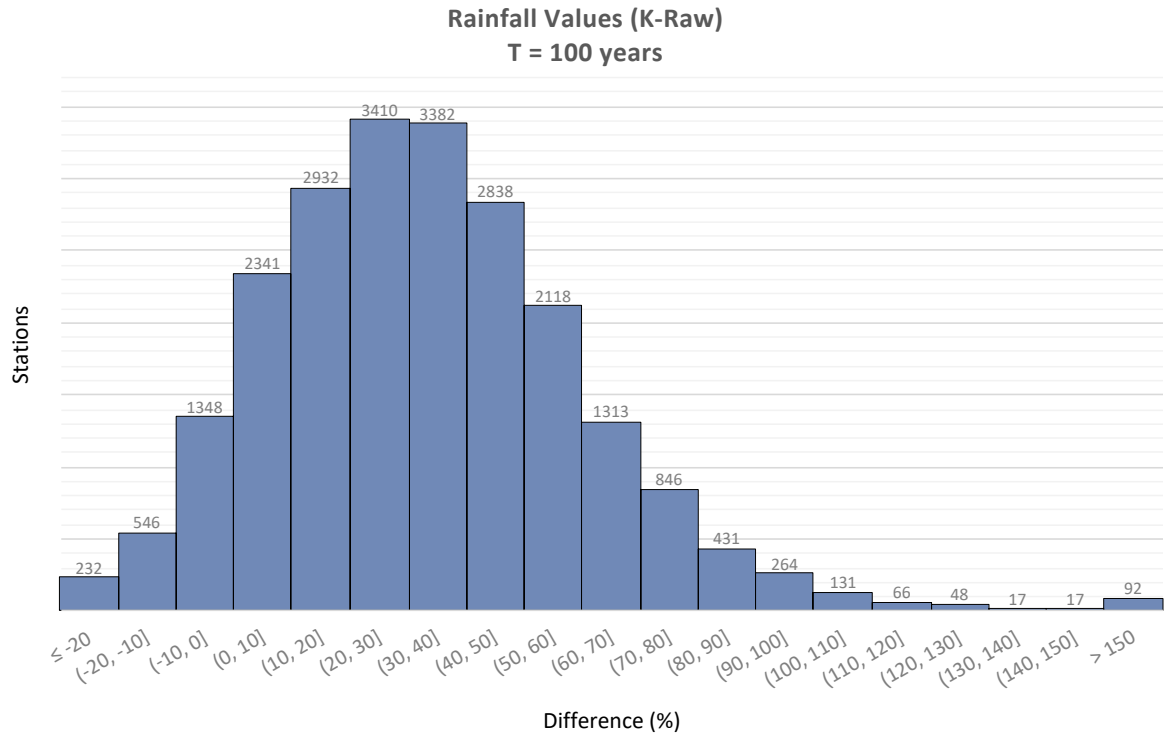
Graph 8.24: Cumulative frequency of low-order RMSE values for all methods. The data labels show the percentage of stations where the estimated RMSE value is below 4.

8.2.4 Rainfall Value Comparison Between K – moments and Classic Methods

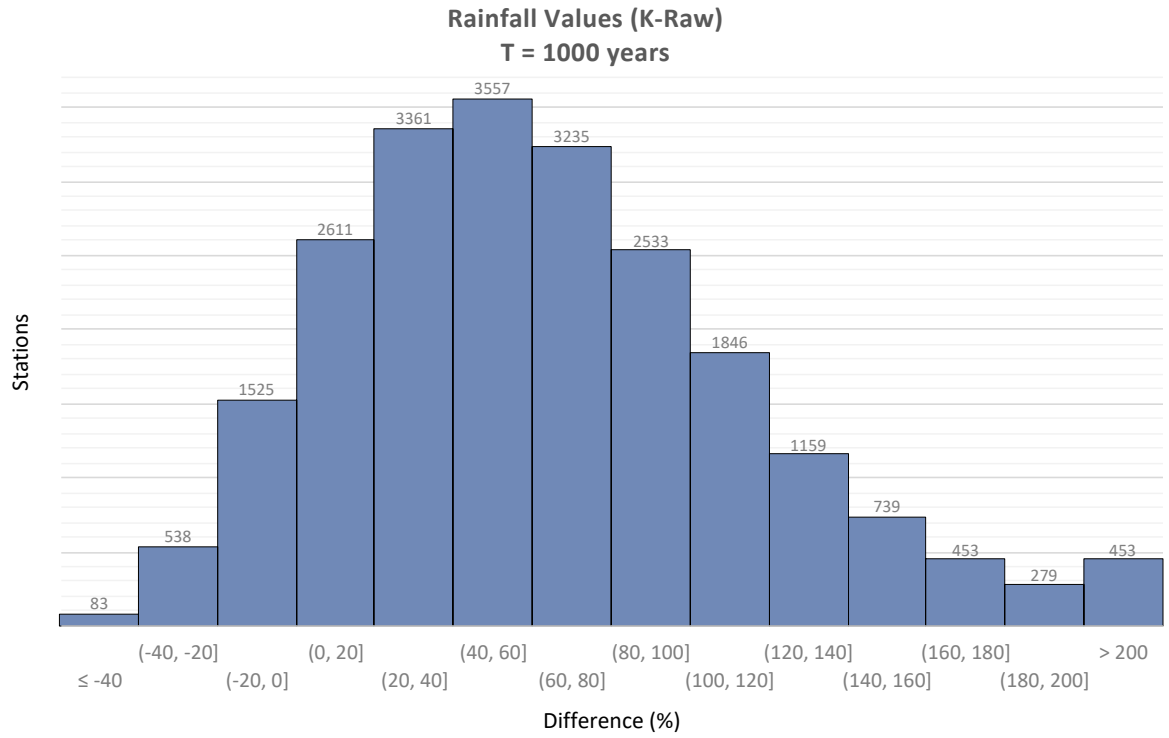
As with the sample station (Chapter 7) where rainfall values at different high return periods are showcased for more direct and clear comparison between methods, the same is applied for the whole dataset. Since the data is now in bulk the most appropriate way of approaching analysing it is through rainfall value percentage differences between fitting with K – moments and classic methods.

The goodness-of-fit comparison (8.2) portrayed the dominance of K – moments in extreme-oriented modelling. For this purpose, rainfall values predicted from K – moments for each station are compared against the remaining inferior methods at specific high order return periods, namely for $T = 100$ and $T = 1000$ years. Both return periods are consistently used as standards in the design of most hydraulic engineering works, thus important estimation of these values is paramount. Even slight variance in their estimation can prove to cause disastrous consequences for infrastructure works and consequently for the population affected by them.

In Graph 8.25, Graph 8.26, Graph 8.27, and Graph 8.28, the percentage difference depicted is positive for higher classic method value than the K – moment one and negative otherwise. More importantly, the graphs depict station that achieved NRMSE for high-order moments over 0.7, in order for the comparison to be concurrent with increased reliability in modelling extreme values. Choosing to use the whole dataset is invalid, since not all stations showed perfect fit while using K – moments. Thus, only the 22,373 stations who achieved to be over this threshold are used in this analysis.



Graph 8.25: Rainfall value percentage comparison between K - moments and Classic moments for return periods of T = 100 years.



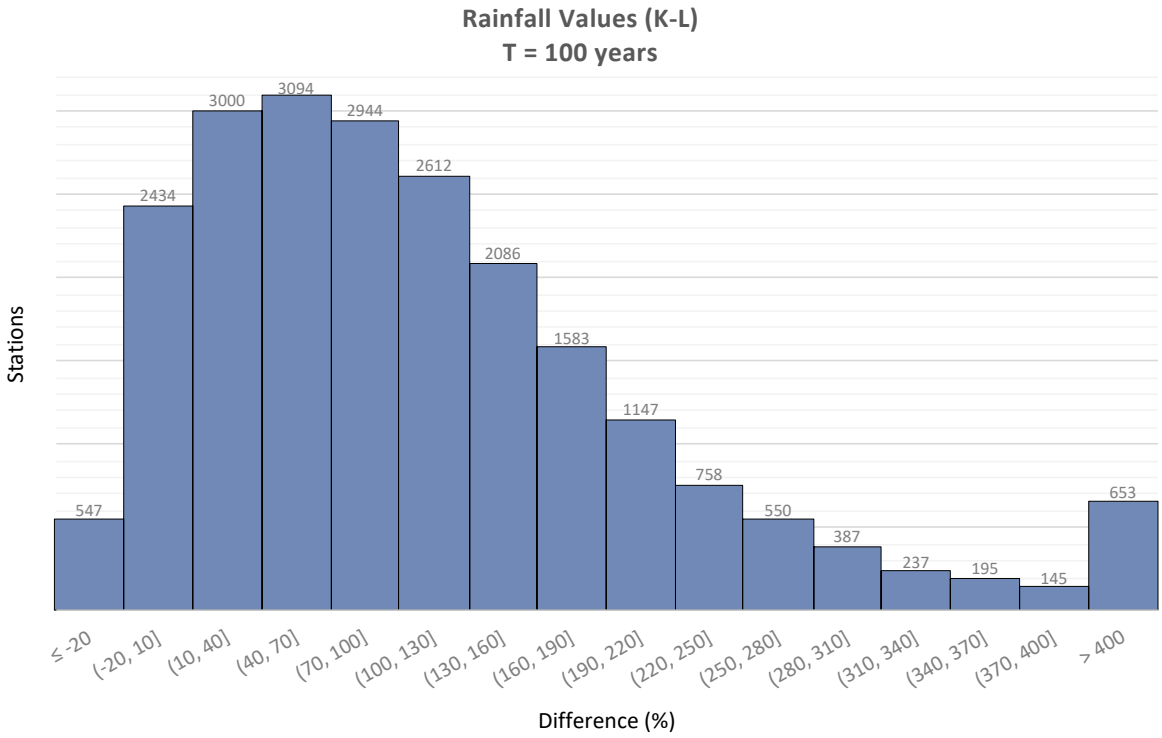
Graph 8.26: Rainfall value percentage comparison between K - moments and Classic moments for return periods of T = 1000 years.

The first comparison is between knowable and classic moments. As it is evident, for 100 years (Graph 8.25), rainfall values are slightly overestimated by using classic moments. Moreover, as expected for 1000 years (Graph 8.26) the overestimation continues and at a higher rate than before. However, there are cases where there exists minor underestimation of observed values, but as shown, these are exceptions.

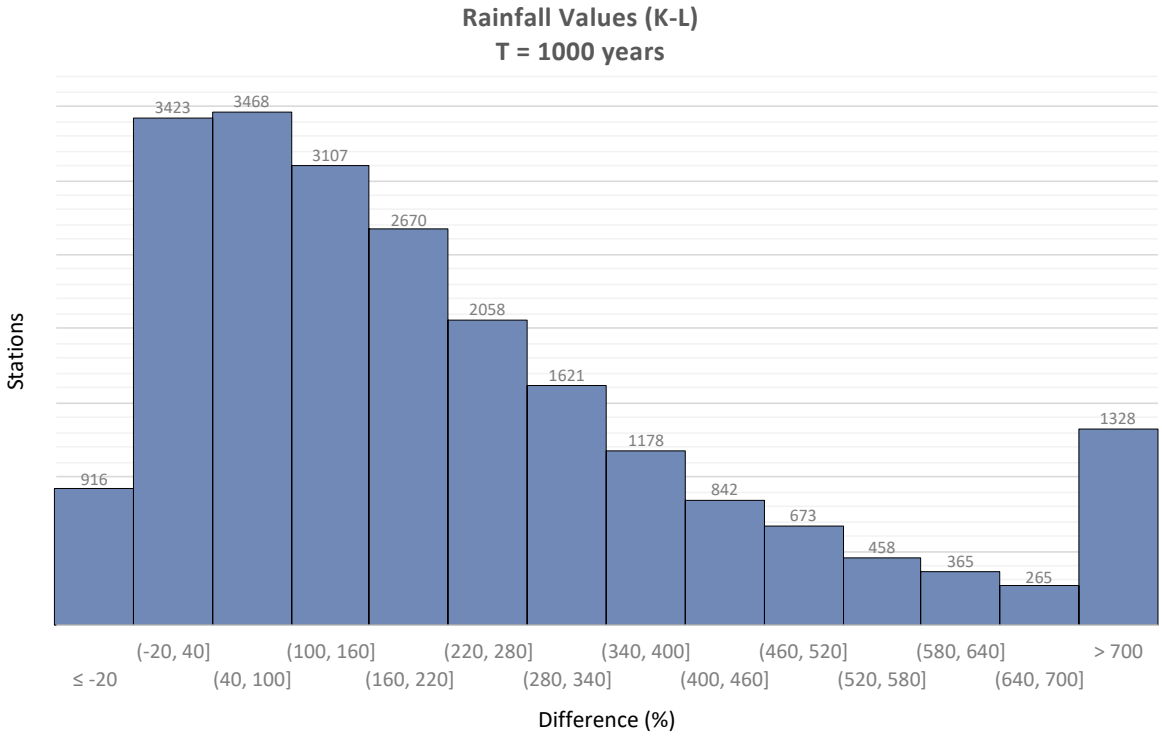
Since, only high reliability stations are plotted, K – moment rainfall value is close to the actual observed value. Thus, the overestimation is not only attributed to comparing to K – moments, but also to real observed data. An average value in the range of 39% for $T = 100$ years and 95% for $T = 1000$ years, it is safe to say that classic moments overestimate observed values by a great margin.

As for knowable moments and L – moments the analysis proves significantly worse results compared to classic moments. This is in sync with goodness-of-fit statistics which portrayed worse performance of L – moments in most cases. Again, theoretical values are overestimated, now with greater difference between them in both 100 and 1000 years rainfall estimation (Graph 8.27, Graph 8.28).

This significant overestimation experienced from classic methods, is detrimental to the designing of especially large engineering works, since it can cause inconsistent risk analyses, significant financial losses and increased resources usage where there is no need to.



Graph 8.27: Rainfall value percentage comparison between K - moments and L - moments for return periods of T = 100 years.



Graph 8.28: Rainfall value percentage comparison between K - moments and L - moments for return periods of T = 1000 years.

8.3 Extreme-Oriented Modelling Effectiveness using K – moments

As a general conclusion for the performance of all methods used in rainfall modelling, it seems that the K – moments approach is the most effective one. Using different distribution regions to study fitting effectiveness for each method it is concluded that:

A. Classic Moments → provide reliable results only for the distribution body, while extremes values are not successfully modelled showing moderate overestimation of extremes for most stations.

B. L – moments → again like classic moments, show reliable results only for the body of the distribution. This method is the least effective for modelling extremes, with significantly low goodness-of-fit statistics and considerably high extremes overestimation patterns for most stations.

C. K – moments → show best overall results. Since they are constructed to focus on extreme values the fitting for high moment orders is the best from the three methods, showing general consistency for most stations. For the same reason, giving emphasis in extremes means that reliability in lower values is sacrificed. Thus, K – moments show slightly worse results in low order moments from classic methods. However, overall, they still provide the best results from the comparison of goodness-of-fit statistics.

Table 8.2, and Table 8.3 depict a general overview of goodness-of-fit statistics for all methods and for all tested distribution regions. As shown, knowable moments are on average the most reliable for overall distribution and extreme-oriented fitting as depicted by both the NRMSE and RMSE. While they are worse for low orders compared to classic moments, their difference is insignificant. Thus, K – moments appear to be the most appropriate for modelling rainfall extremes.

Table 8.2: Average NRMSE values in every distribution region for all methods

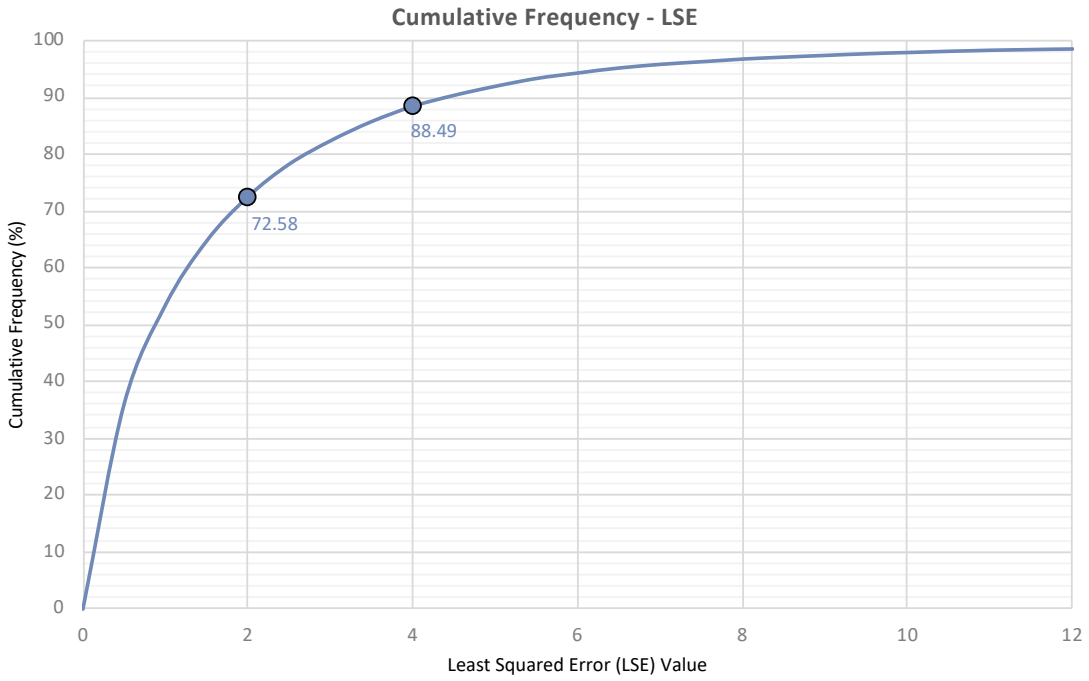
NRMSE	Classic	L	Knowable
Overall	0.722	-0.226	0.854
High	0.303	-2.218	0.713
Low	0.870	0.768	0.783

Table 8.3: Average RMSE values in every distribution region for all methods

RMSE	Classic	L	Knowable
Overall	11.80	46.805	6.638
High	16.907	68.236	8.427
Low	2.112	3.104	3.585

By establishing that knowable moments are the most suitable for modelling extremes, the study can now focus more on the fitting process's effectiveness and efficiency. Since fitting with K – moments rely on an optimization process through minimizing the least squares error (LSE) between theoretical return periods obtained from Equation 3.53 and empirical return periods assigned to K – moments from Equation 3.60, evaluation for the error parameter is provided and analysed throughout the whole dataset (Graph 8.29).

Graph 8.29: Optimization Least Squared Error (LSE) used in the fitting process between return periods.



Graph 8.30: Cumulative frequency of LSE fitting values. The data labels show the percentage of stations where the estimated LSE value is below 2 and 4.

The average value of LSE through the whole dataset is estimated at 1.84. Since, the optimization method is based on least squares, the optimum solution is achieved for the lowest LSE value, and it will always be higher than 0. While this is the case, it is observed that LSE values below 2 depict an almost perfect fit, and stations below 4 are quite reliable.

From Graph 8.29 it is evident that most stations are optimized with an LSE lower than 4. In more detail, from Graph 8.30 the cumulative frequency of stations with LSE below 2 is around 72% and for those below 4 is 88%. Consequently, there is great compatibility between empirical equations through Λ – coefficients for estimating return periods and the theoretical ones estimated from the Pareto distribution's definition. Stations

At this moment, it is important to note that the LSE optimization tool doesn't directly showcase the reliability of the modelling process between the fitted distribution and observed values. These two concepts are linked through the empirical return periods assigned to K – moments through Λ – coefficients. In practice, this means that low LSE value doesn't guarantee that the fitted model correctly describes extremes (i.e. high NRMSE or low RMSE value). The fitted model performance compared to observed data is displayed from goodness-of-fit statistics as discussed in 8.2, which show great results especially for high-order moments.

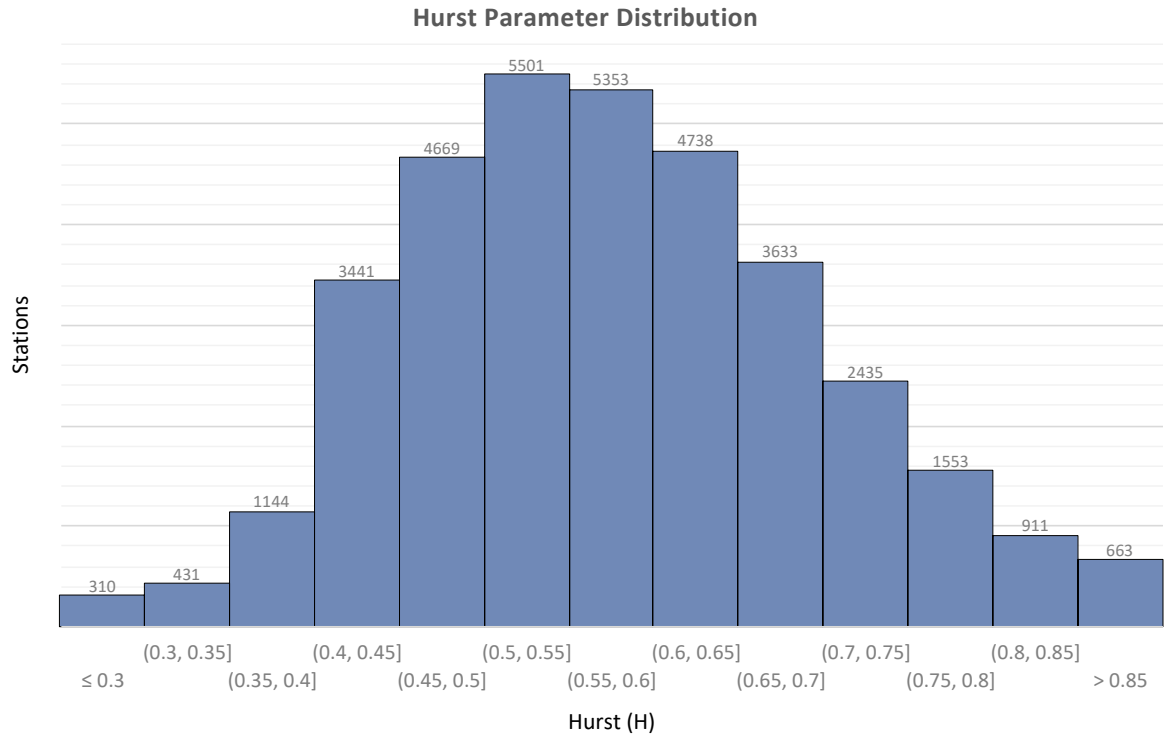
Therefore, it is safe to conclude that both empirical return periods and the final fitted Pareto model are effective for directly describing rainfall extremes. This assumption is validated by comparing the cumulative frequencies from NRMSE for high orders (Graph 8.12) and LSE values (Graph 8.30). As depicted from both charts, about 65% of stations have estimated high-order NRMSE above 0.7, while 72% have an optimization LSE below 2. Both thresholds show a suitable respective fitting and since those percentages are only slightly different from each other, the effectiveness and indirect correlation between empirical return periods and final model fit is confirmed. The 7% difference, with LSE achieving the highest percentage between the two confirms the point discussed in the previous paragraph and is caused from the incapability of the standard two parameter Pareto distribution (GPD2) to successfully describe the given data.

8.4 Impact of Long-term Dependence on Modelling Results

As proven in modelling of the sample station with K - moments (7.3), taking into account the effects of persistence or long-term dependence, yields significant difference in the fitting results. Failing to account for dependence can lead to great underestimation of rainfall values for high return periods, which are of the most interest.

For this reason, all modelling results produced and analysed in the previous chapters implement the effects of long-term dependence. Thus, in order to solidify and showcase its impact in the totality of the database, a part of it is remodelled without accounting for dependence.

Hurst parameter depicting the magnitude of positive long-term dependence takes values over 0.50. However, as seen in Graph 3.9 the effects are significant for values over 0.70. Thus, stations with $H \geq 0.70$ are remodelled now ignoring the dependence structure and only those results are presented. While, all stations with $0.50 < H < 0.70$ will be affected from positive persistence the difference in the end is minimal.

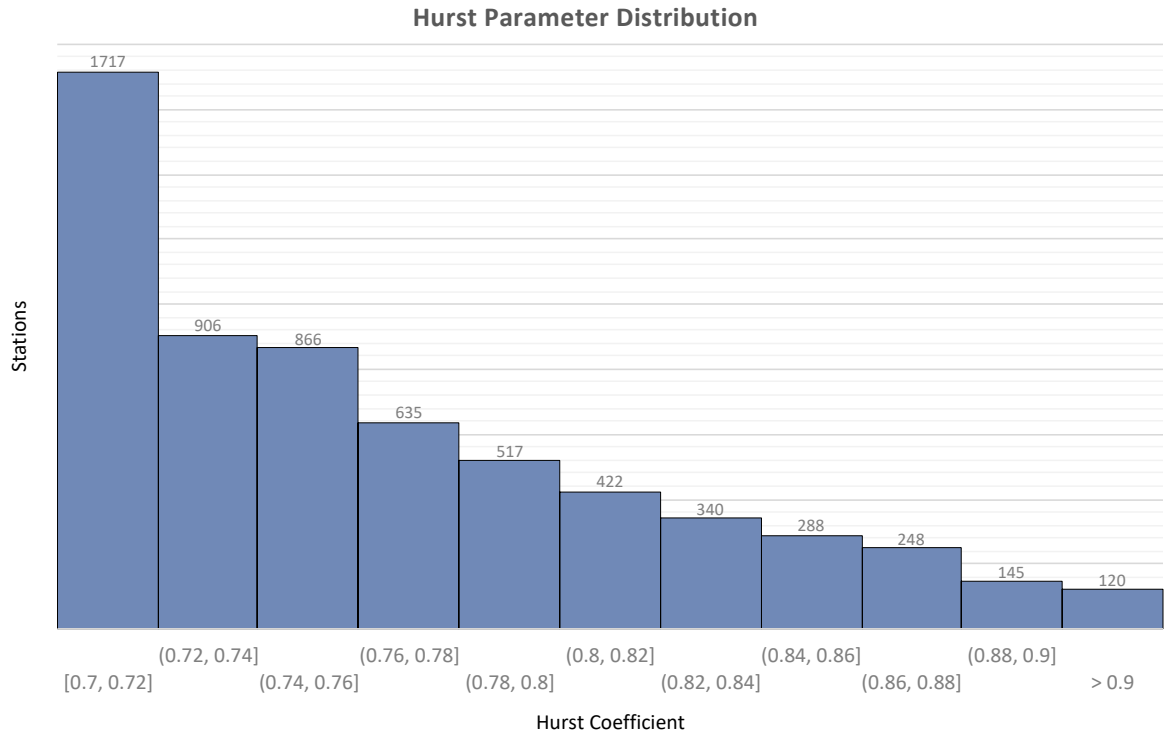


Graph 8.31: Hurst coefficient distribution for the complete database

The Hurst coefficient's distribution from the whole database gives an average of 0.58. While, rainfall data is mostly associated with high long-term correlation, many stations show either a lack of long-term dependence at around the $0.4 < H < 0.6$ mark and those with $H < 0.4$ signify negative long-term persistence, but these are limited to about 2,000 stations of the total (Graph 8.31).

The fact that some stations are showing moderate anti-persistent behaviour is not considered typical for the rainfall process, and it may be attributed to lack of sufficient years of observed data or to specific extreme values that skew the curve fitting result of the K-climacogram, from which the Hurst coefficient is estimated.

Remodelled stations account for 6,204 of the total and their distribution is shown in Graph 8.32. It is evident that most stations are found below the 0.8 mark. However, for Hurst values $H \geq 0.8$ station density is still significant. Although, the Hurst coefficient plays a significant role in quantifying the effect of long-range dependence, as seen from Graph 3.9, observed sample size is also a contributor, but with less influence on the resulting quantification.



Graph 8.32: Hurst coefficient distribution for stations with H > 0.70

For evaluating the effects of long-term persistence on stations prone to show such behaviour, the difference between modelled rainfall values for large return periods (100 and 1000 years) will be calculated with the same process as the sample station (7.3.2). Moreover, correlation between the Hurst coefficient values and the magnitude on the results is also investigated.

By accounting for long-term dependence, modification on the distribution tail is being made, with a tendency to upscale extreme events for the same values of return period, compared to sample independence. In other words, the distribution tail shifts upwards on the *y* axis (rainfall values) thus estimating larger extreme events for a given time period.

The magnitude of this upward shift cannot be determined beforehand, even if the Hurst coefficient and the dependence bias are estimated. This is due to the fact that Pareto tail behaviour is predominantly dependent on extreme rainfall values of the specific observed

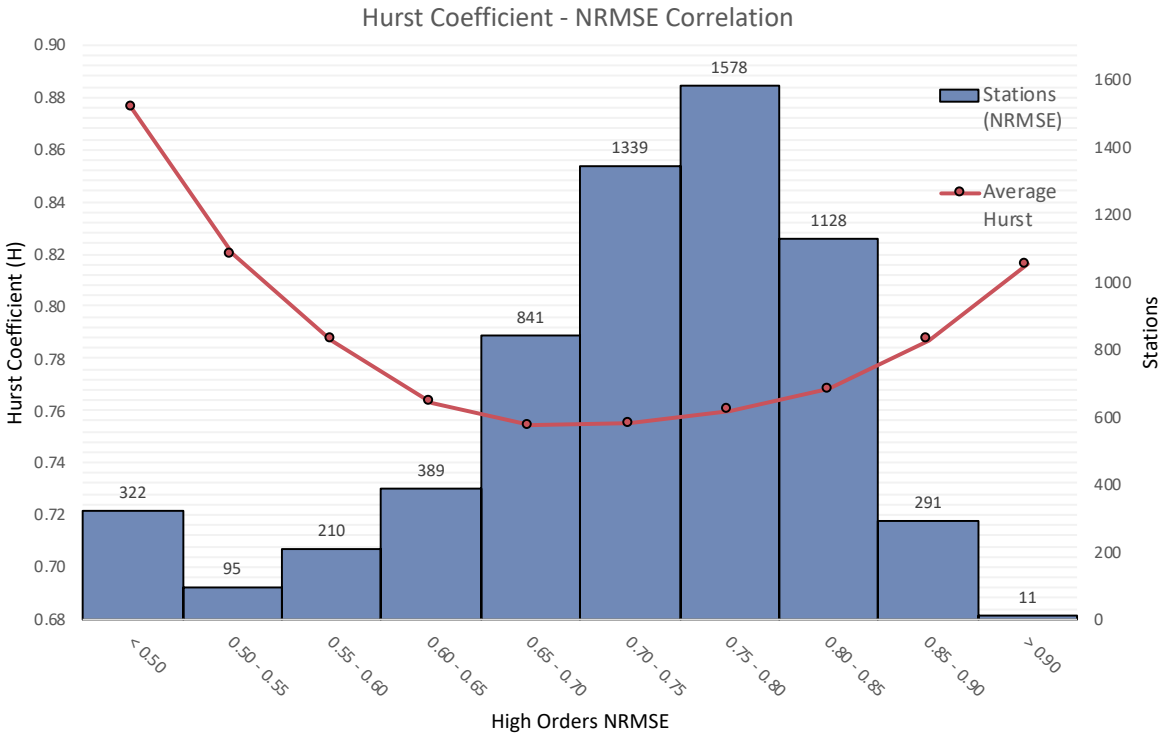
data set and less on the dependence bias. However, slight changes on the tail index κ are bound to be noticed between the two different assumed dependence structures.

For this reason, when there is great influence of long-term dependence in the sample, while the fit between return periods empirically assigned to K – moments and theoretical ones might be perfect (minimal LSE), goodness-of-fit parameters may not show a great overall fit to observed values, especially for high values ($T > 1$ year). Empirical return periods now aren't only bound by observed values, but also by the bias from the sample's dependence structure. The theoretical Pareto distribution is fitted by means of the empirical return periods thus the bias transfers to it in the end.

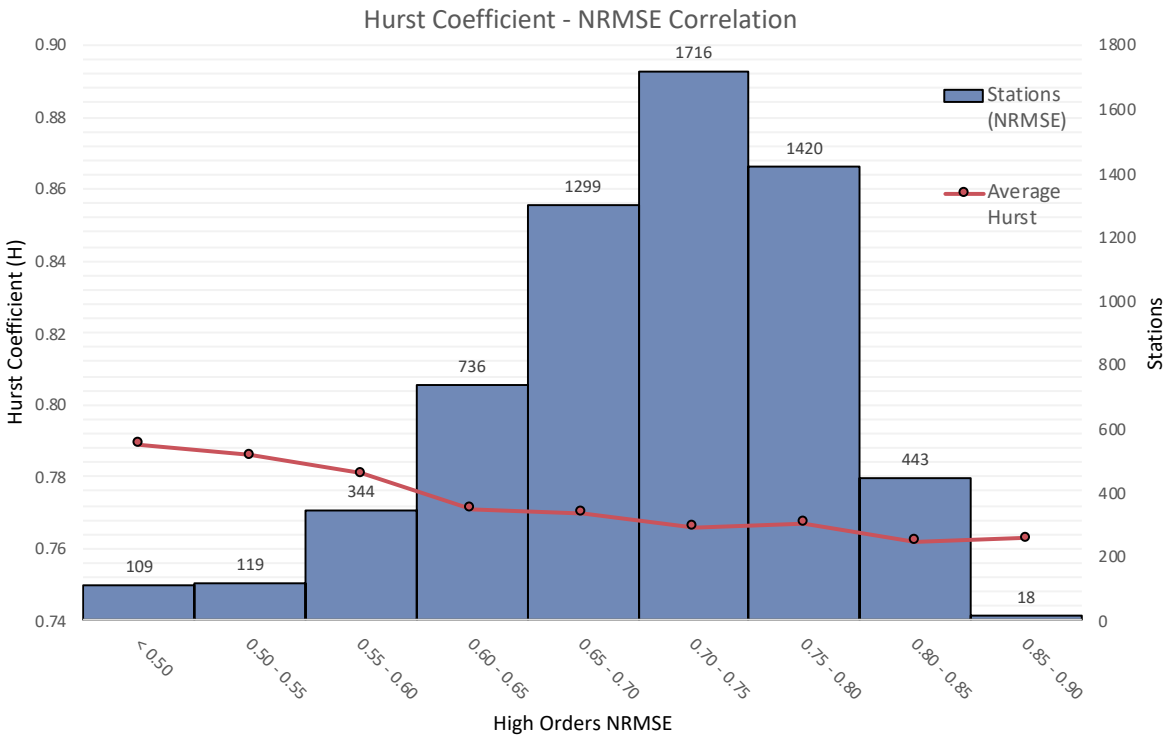
In order to showcase this discrepancy, comparative results for high-order NRMSE values for each added or ignored dependence bias are also provided (Graph 8.33, Graph 8.34). Depicted is the distribution of high-orders NRMSE for stations with $H \geq 0.70$ while the line represents the average Hurst coefficient value of each error range.

It is clear that when Hurst average is high, NRMSE is low, showing "poor" fit between observed values and the theoretical distribution. As mentioned before, this doesn't portray unreliability, but shows the dependence bias effect to the final modelling result. Furthermore, the trend is downward up to $NRMSE = 0.7$, while after it slightly increases. This is again due to the fact that in some cases the upscaling effects the bias ensues can achieve positive influence on the overall fitting result, meaning greater NRMSE value.

On the other hand, by ignoring bias the average Hurst coefficient stays practically constant for each NRMSE bin.



Graph 8.33: Correlation of Hurst coefficient and NRMSE value while accounting for long-term dependence

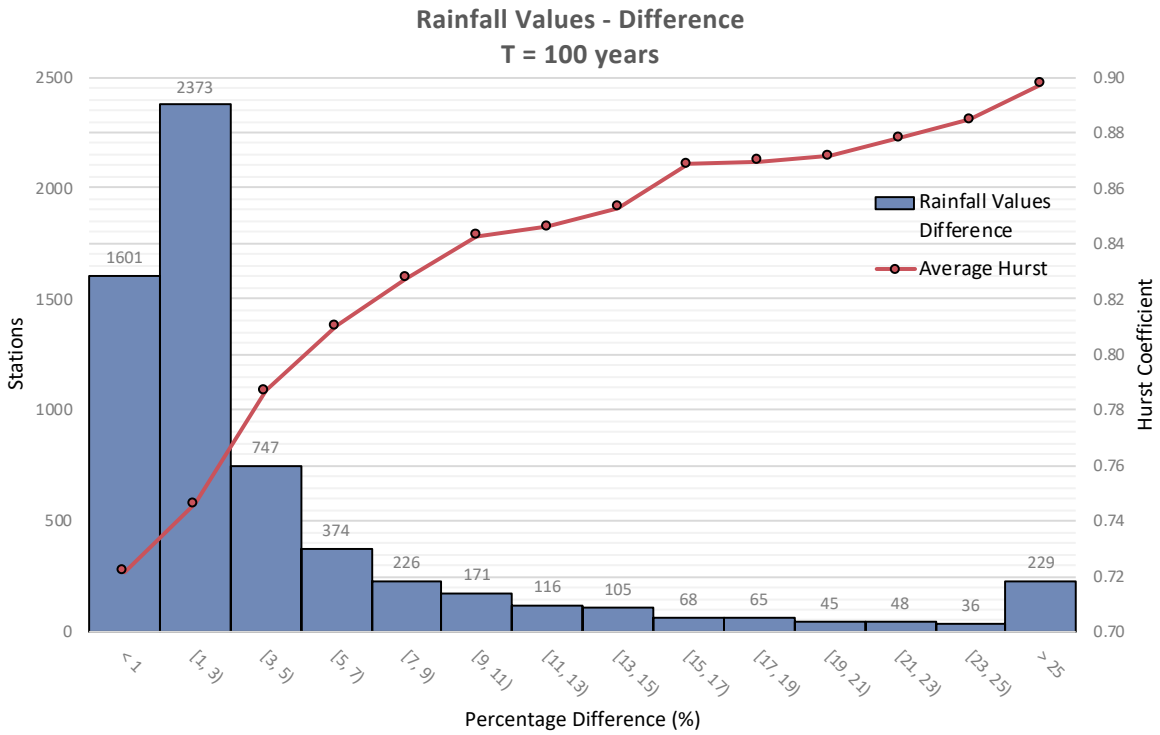


Graph 8.34: Correlation of Hurst coefficient and NRMSE value while ignoring long-term dependence

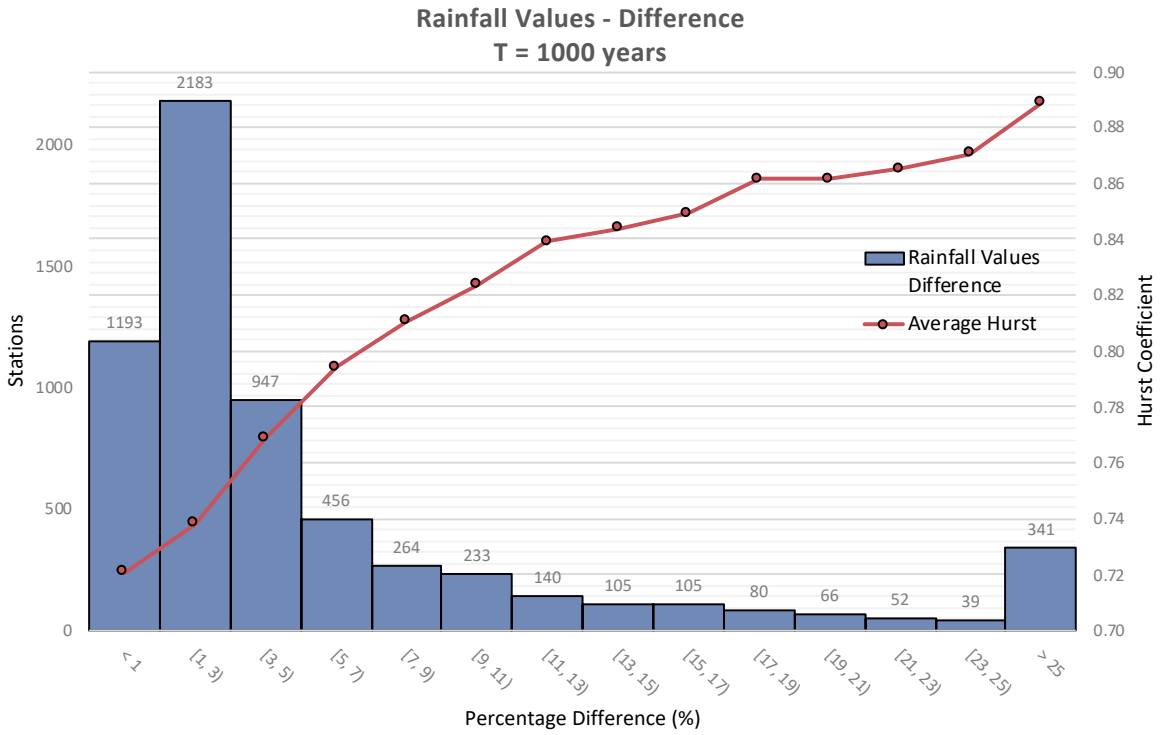
Moving to the results, Graph 8.35 and Graph 8.36 and show the percentage difference between rainfall values for said return periods. Also, the average Hurst coefficient for each bin is also plotted in order to showcase the positive correlation between dependence bias and value alteration. Rainfall values for $T = 100$ years, show clear influence of the dependence structure on extreme events. The histogram data depict that most stations don't suffer great overall change in their rainfall value, but this is closely correlated with the Hurst coefficient value.

While the coefficient isn't the only parameter in quantifying the true disparity between assumed sample independence and accounted dependence bias, it is the most influential one. The correlation presented proves this fact. The higher the percentage change, the higher the average Hurst coefficient. As shown before, many stations achieve Hurst of below 0.8, thus it is normal for the histogram to depict higher density for low difference values.

As for rainfall values for $T = 1000$ years, the same behaviour is noticed. However, the effect is slightly upscaled due to the increased time period investigated, while the average Hurst line produces the same upward trend as before.

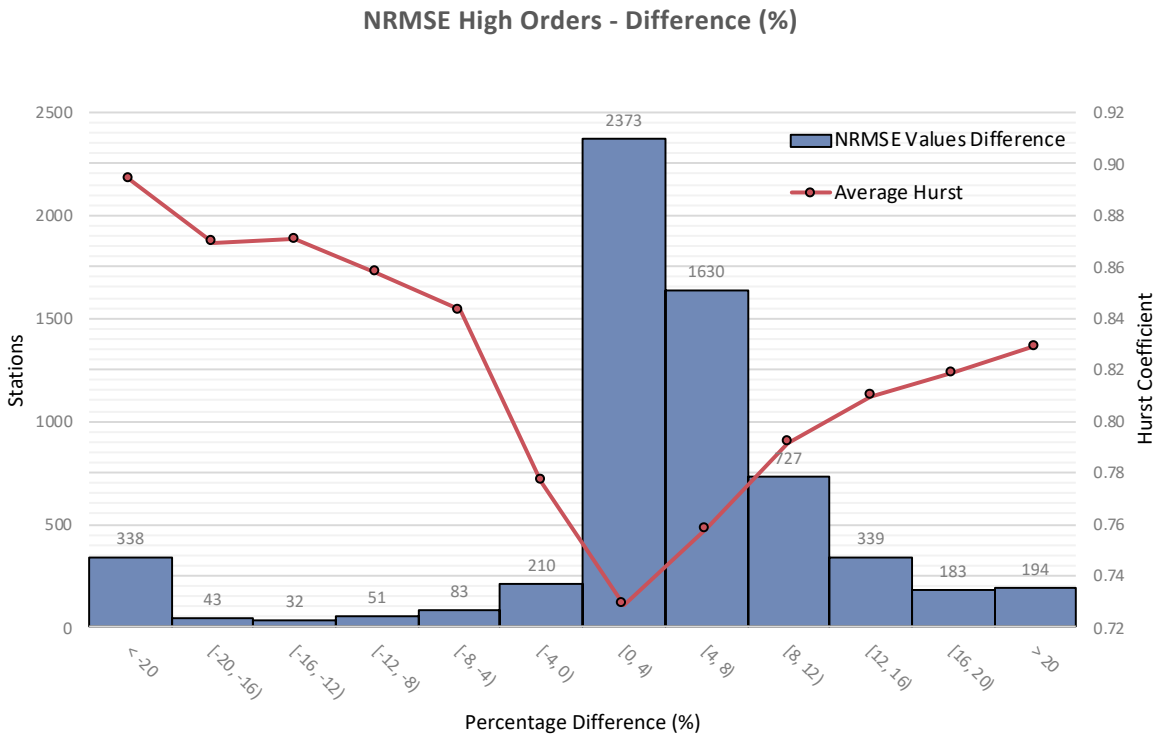


Graph 8.35: Distribution of stations depicting the percentage difference of rainfall values (T = 100 years) between ignored and added dependence bias. Line represents the average Hurst value for each bin.



Graph 8.36: Distribution of stations depicting the percentage difference of rainfall values (T = 1000 years) between ignored and added dependence bias. Line represents the average Hurst value for each bin.

Graph 8.37 proves the notion that goodness-of-fit for high return periods is slightly worse when stations with moderate to significant dependence bias are compared with the same stations when ignoring the dependence structure. While this is true for some stations, the graph also shows performance increase for a significant number of stations, which means that the added bias is a benefit to the overall fitting result. For either case, the influence of the Hurst parameter in the outcome is again clear. The lowest average point is for the [0,4) bin which signifies no major difference in NRMSE value, while the highest averages are found for the highest absolute differences.



Graph 8.37: Distribution of stations depicting the percentage difference of high-order NRMSE values between ignored and added dependence bias. Line represents the average Hurst value for each bin.

Despite of the fluctuations due to the dependence bias, the fit to observed values should still be considered reliable for the lower NRMSE values and is most likely attributed to inconsistencies of the K – moments approach. Being naturally consistent, thus accounting for the long-term persistence of a rainfall data set, is more important than a perfect goodness-of-fit parameter. If this priority in modelling is not followed, then the final model would underestimate reality and in many cases by a significant margin.

In conclusion, dependence bias greatly affects the outcome of the fitting result and should be taken into account for every station. Specifically, for those with high estimated long-range dependence, it is even more important since as proven above develop the greatest overall differences in estimating rainfall extremes.

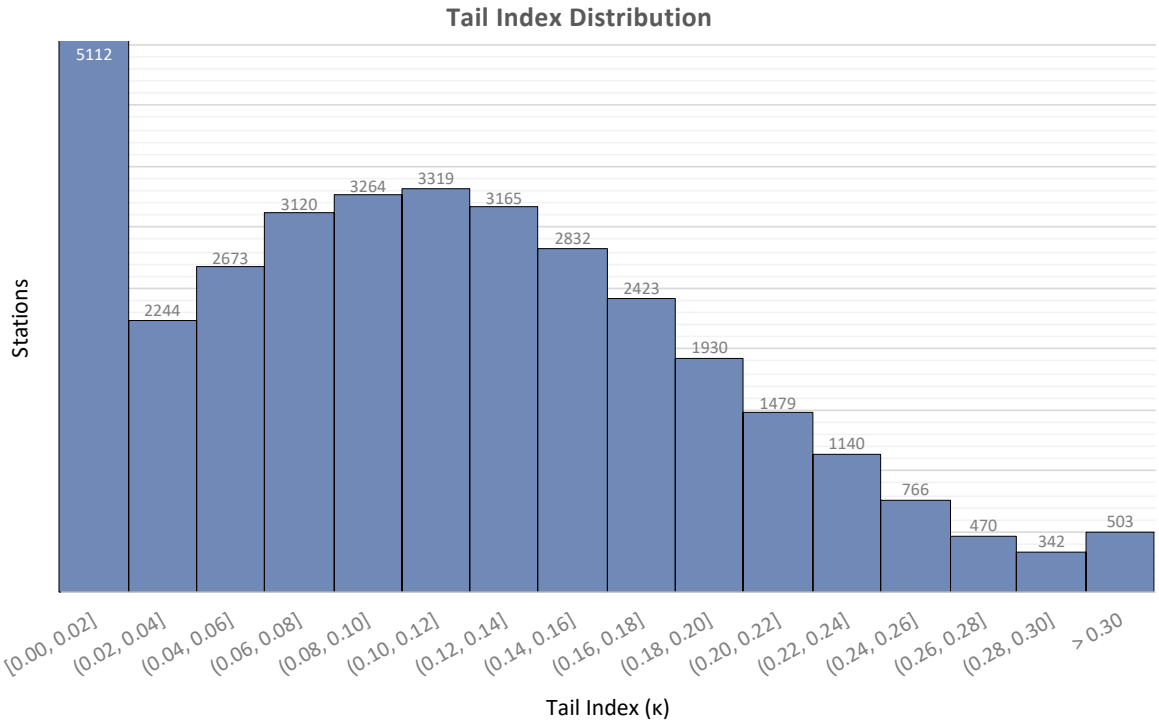
8.5 Global Results of Fitting Parameters

In this chapter focus is given in analysing the modelling results for all stations, showing distributions of parameters across the globe and finding correlation between certain results, all while using the aforementioned K – moment approach with the dependence structure of each station, if present, taken into account.

8.5.1 Tail Index

The first parameter analysed is the tail index (κ). Its practical use is to control the behaviour of the distribution's tail. This is clarified from Graph 3.2 where the tail index is kept constant, providing with same tail behaviour despite the changes of the scale parameter. In other words, it gives a representation of the tail's slope. An important attribute to note is that, while for the cumulative distribution (Equation 3.3) higher tail index means lower slope, for plotting return periods (Equation 3.53), the opposite is true; lower κ suggests lower slope, with zero transforming it to an exponential distribution.

Connecting it with the results from this study, the tail index is an indicator of how quickly rainfall values increase over a specific range of high-order return periods and consequently depicts the degree of this increase (usually for $T \geq 1$ year). In order to prove this, with the already provided analysis over rainfall values for the large return periods such as $T = 100$ and $T = 1000$ years, by calculating the percentage difference between them the overall increase in this time range is shown.



Graph 8.38: Tail index (κ) distribution for all modelled stations

Graph 8.38 shows the distribution of the tail index among all stations. The results show a significant number of stations, approximately 15% of the total, with $\kappa \leq 0.02$ and almost 3,000 of them valued at the lower threshold set in the fitting process, which is $\kappa = 0.001$. For those stations, the distribution's tail fit might improve with the parameter being even lower than the threshold, but since having index values below zero is not considered naturally consistent the results remain as is.

Nonetheless, the fitting error even for these stations still remains low, suggesting reliability, just not as perfect as could otherwise have been ($r = -0.274$). In this case, solutions can be found by using the scale parameter of the GPD which was previously set as zero for consistency reasons or alternative theoretical distributions with more parameters, like the PBF or the Dagum which contain one more parameter. However, they are not put to the test in this study. Despite of this, most tail index values are in the range of 0.04 to 0.2 which is considered normal for the rainfall process (Koutsoyiannis, 2004).

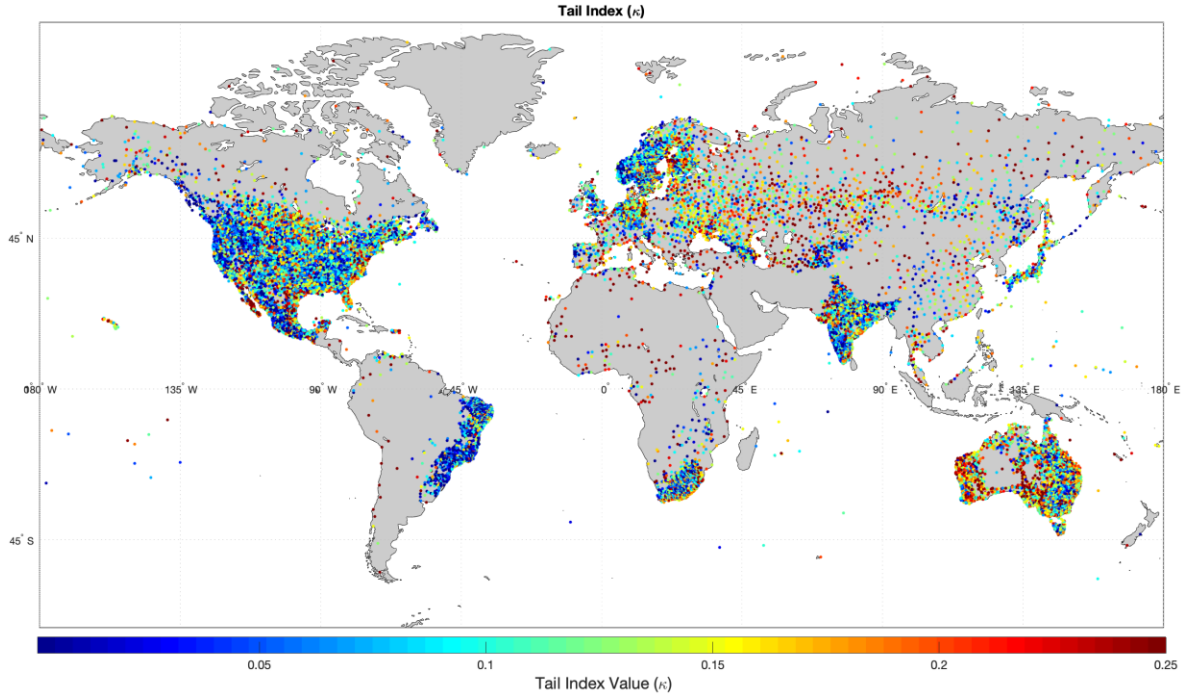


Figure 8.1: Global map showing tail index distribution.

While Graph 8.38 shows a general overview of how the tail index values are distributed, it doesn't provide with information regarding the locations that these values arise. Thus, a heat map with all stations showing the tail index value is provided in Figure 8.1 and a more detailed in Figure 8.2 . The main observations can be summarised to:

A. Major regions with low values ($\kappa \leq 0.02$) are Brazil, India, Mexico, north-western North America, (and the Scandinavian countries). Connecting them to their climate classification (Figure 2.1), the greatest contributors being Brazil, India, and Mexico, have either monsoonal or dry winter equatorial climate. This suggests high rainfall values, but stable extremes throughout the years, thus producing low tail slope in the modelling process due to the predictability of extremes. The same is acceptable for the other regions where climate is regarded as snow (fully humid) and in some areas polar.

B. As for high values ($\kappa \geq 0.2$) are western and central Australia, central Africa, central Eurasia, and Mexico's Gulf of California region. Applying the same logic as before, all regions now are known to have variations of the arid climate. This means that there is little precipitation throughout the year and not many high rainfall values, which in practice makes the GPD2 reach that "low" extreme value really fast, thus producing high tail index value. High index values, aside from arid regions, are also observed in the Mediterranean where exists a certain sub-category of warm temperate climate, in which dry and hot summers are the main distinctive factors.

C. Values in-between ($0.02 < \kappa < 0.2$) are scattered throughout and in general are found in other variations of the temperate and snow climates, especially those with fully humid seasons and warm summers.

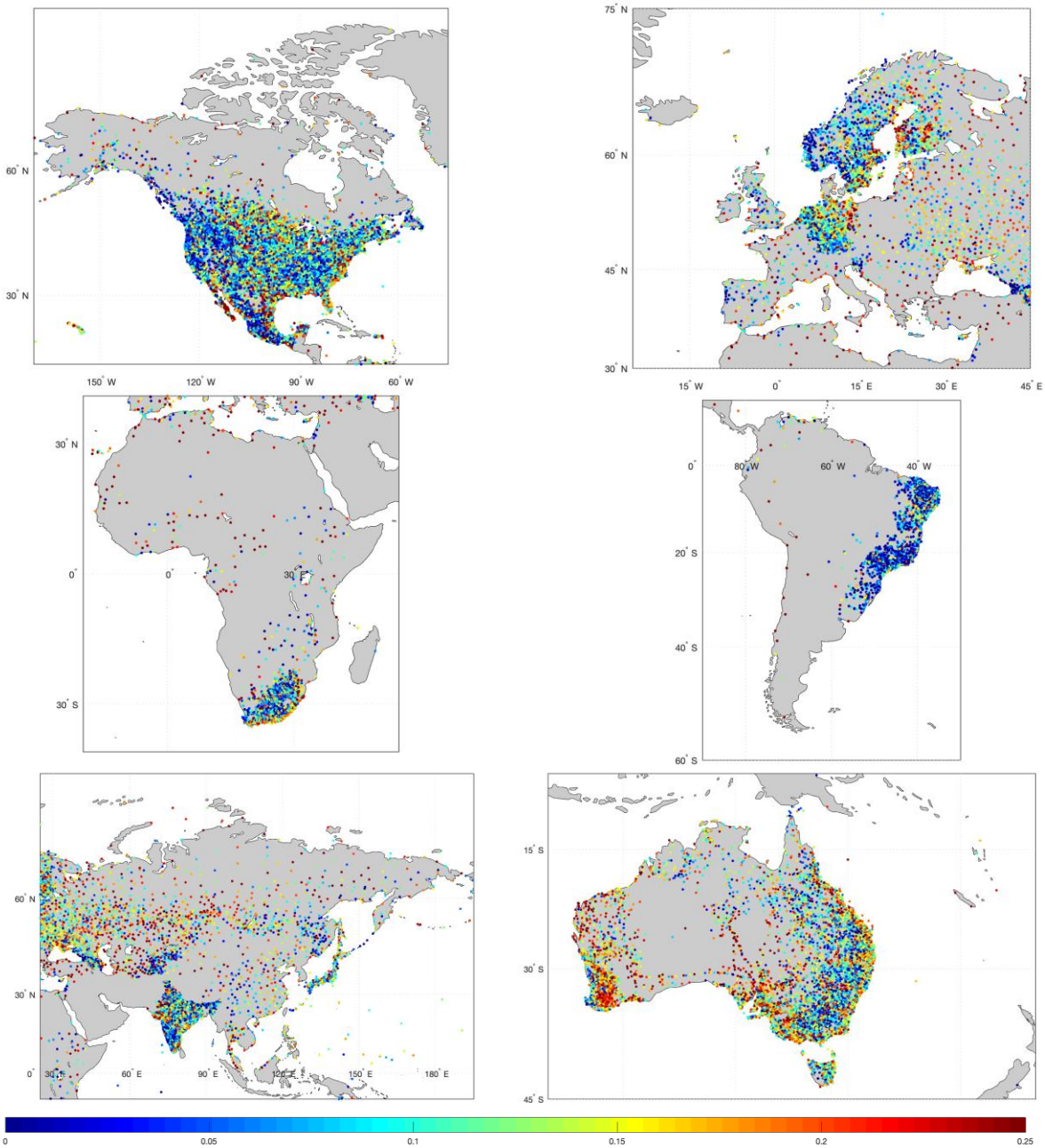
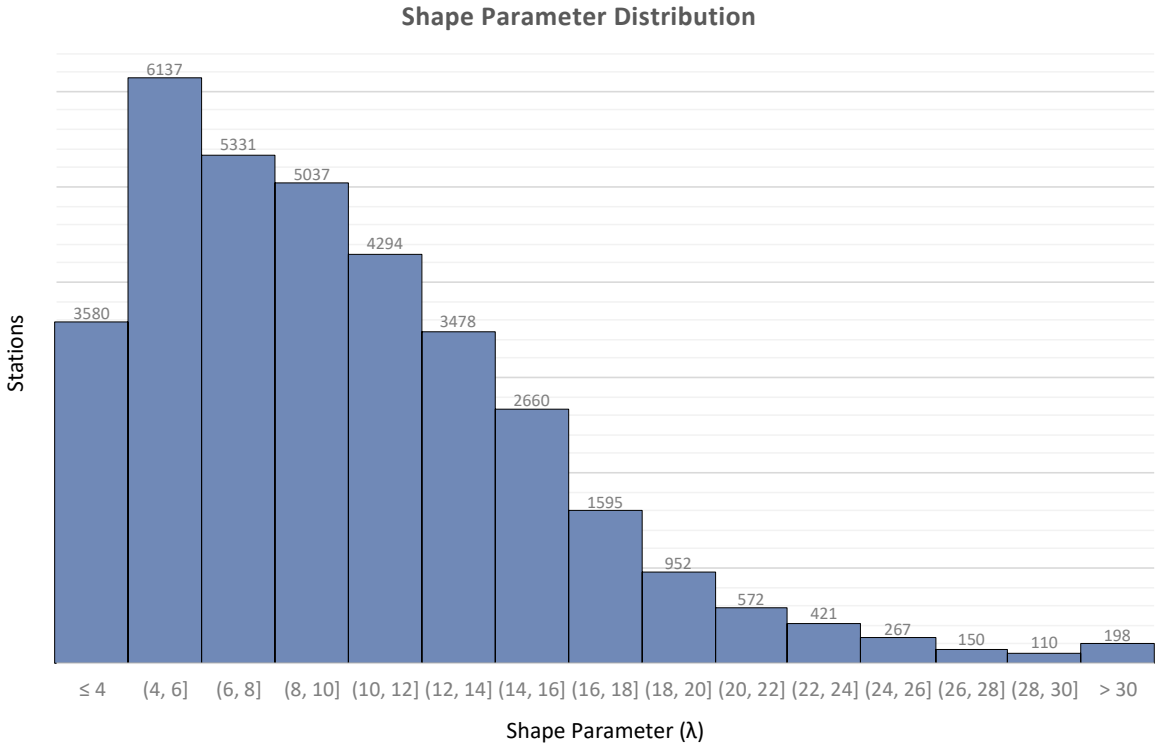


Figure 8.2: Continental distribution of the tail index. From top to bottom and left to right; Europe, Africa, Asia, North America, South America, Australia

8.5.2 Scale Parameter

The second parameter of the GPD2 (λ) controls the overall scale of the distribution. More specifically, it is responsible for the behaviour of the distribution's body and its curvature characteristics (Graph 3.2). When plotting theoretical return periods, higher parameter values produce steeper increase of rainfall values in the distribution's body, and higher overall in the extremes range. With the tail index kept unchanged and for high return periods, the figure consists of parallel lines with higher λ values producing greater overall rainfall.

The scale parameter despite controlling the behaviour of the body, it plays an important role in modelling extremes, since it indirectly depicts the magnitude of extreme values, unlike the tail index which is responsible for the incremental change of such extremes. This characteristic is proven by comparing the parameter's values with the average rainfall value in each station. Achieving a Pearson correlation coefficient of $r = 0.878$ the correlation between rainfall intensity and the scale parameter is evident. The results are presented exactly as for the tail index above.



Graph 8.39: Scale parameter (λ) distribution for all modelled stations

Graph 8.39 depicts the distribution of the scale parameter, with most stations producing values in the range $4 \leq \lambda \leq 12$. The distribution is predominantly skewed towards lower values which is more consistent with a general representation of the rainfall process, but high values aren't correlated with high fitting errors, as the correlation coefficient between them is $r = 0.158$ which is low enough to assume independence. In order to show dependence to climatic characteristics heat maps are again presented.

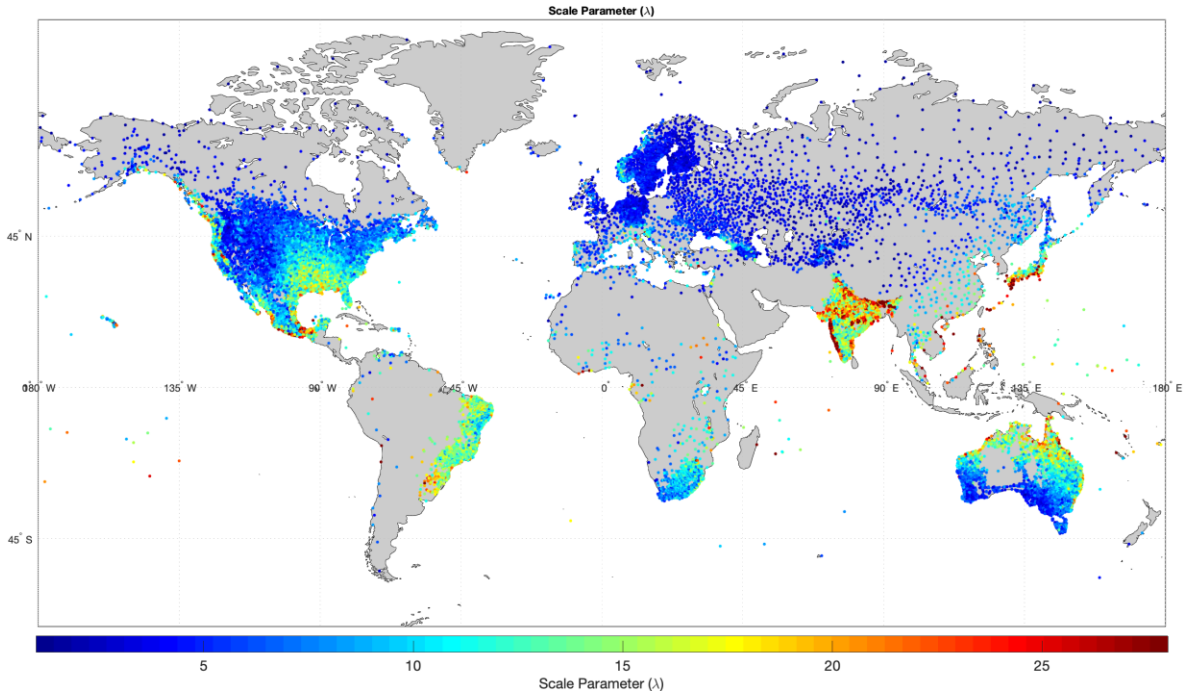


Figure 8.3: Global map showing scale parameter distribution.

The worldwide heat map (Figure 8.3) produces clear results in terms of parameter distribution. The main points taken from it are:

A. Most high values ($\lambda > 0.15$) are concentrated around low absolute latitude values. Regions like north Australia, Brazil, India, Mexico, south-eastern USA, most of Japan and the few modelled stations in Indonesia, all show high scale parameter values. All these regions have either equatorial (tropical) or fully humid warm temperate climate, which as mentioned before are prone to delivering frequent rain days and to a great intensity, since tropical storms and monsoons are a typical seasonal phenomenon.

B. For higher absolute latitudes, values begin to diminish. Regions like Russia, Europe, South Africa, south Australia, northern USA, and Canada, all produce low scale parameter values consistently. In connection to their climate characteristics, all of them belong to either an arid, warm temperate with dry warm summers, or snow climate. All of those classifications show less rain and with less intensity throughout the year, thus achieving lower extreme values.

From these observations it is evident that climate plays a significant role in the value the scale parameter receives from the modelling process and comes in direct correlation with the intensity of precipitation along those regions. This is a reason why the correlation coefficient between λ and the average values of rainfall for each station is so high.

In conclusion, the scale parameter is low for low extremes and low precipitation in general and high otherwise. Continental maps are also provided for further clarity on the results described (Figure 8.4).

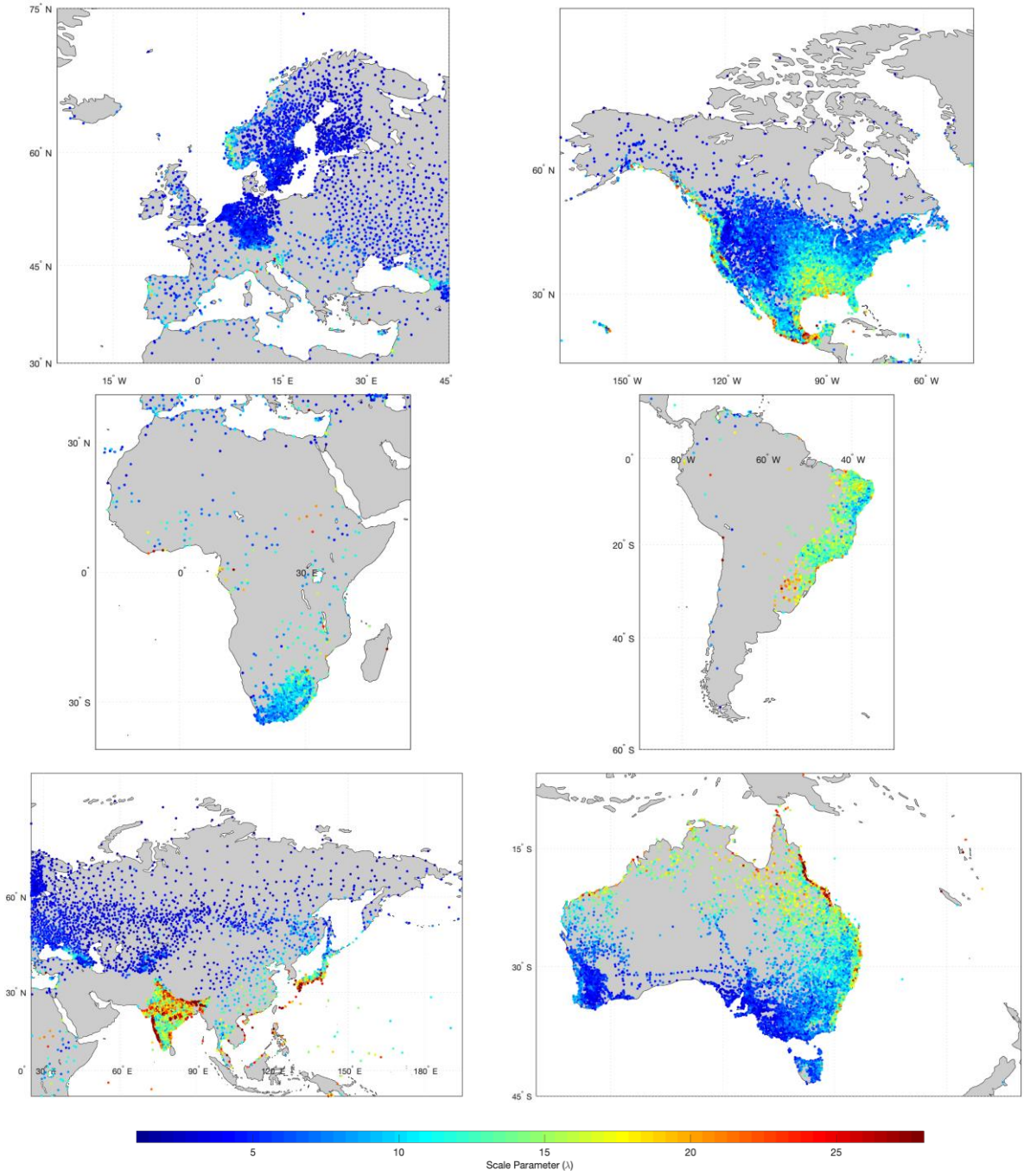


Figure 8.4: Continental distribution of the scale parameter. From top to bottom and left to right; Europe, Africa, Asia, North America, South America, Australia

While it is proven that each of the GPD2 parameters are connected to the climatic characteristics of the area they describe independently from each other, in the end both influence the final behaviour of the distribution and especially its tail. This comes in conjunction with the fact that there is not absolute correlation between one characteristic and a parameter, but simply a stronger influence to it.

In order to showcase this, rainfall values for $T = 100$ years are plotted on the heatmap. This characteristic is chosen, since it is the effective resulting product of the extreme-oriented rainfall modelling and it concerns an extreme value. If independent parameter

influence on the rainfall value is assumed based on the previous results, the only parameter affecting the magnitude of the result is the scale parameter. By observing Figure 8.5 it is evident that high-order rainfall values are influenced mainly by scale parameter as the same patterns arise, suggesting high rainfall values in the same regions as was for the scale parameter, but this effect is not valid for all stations, with some breaking the pattern especially in Australia, central Europe, and Brazil. As a quantitative measure again the Pearson coefficient between the scale parameter and said rainfall values is $r = 0.663$ which suggests positive correlation, but not as strong so as to assume clear independence from other sources in the estimated result.

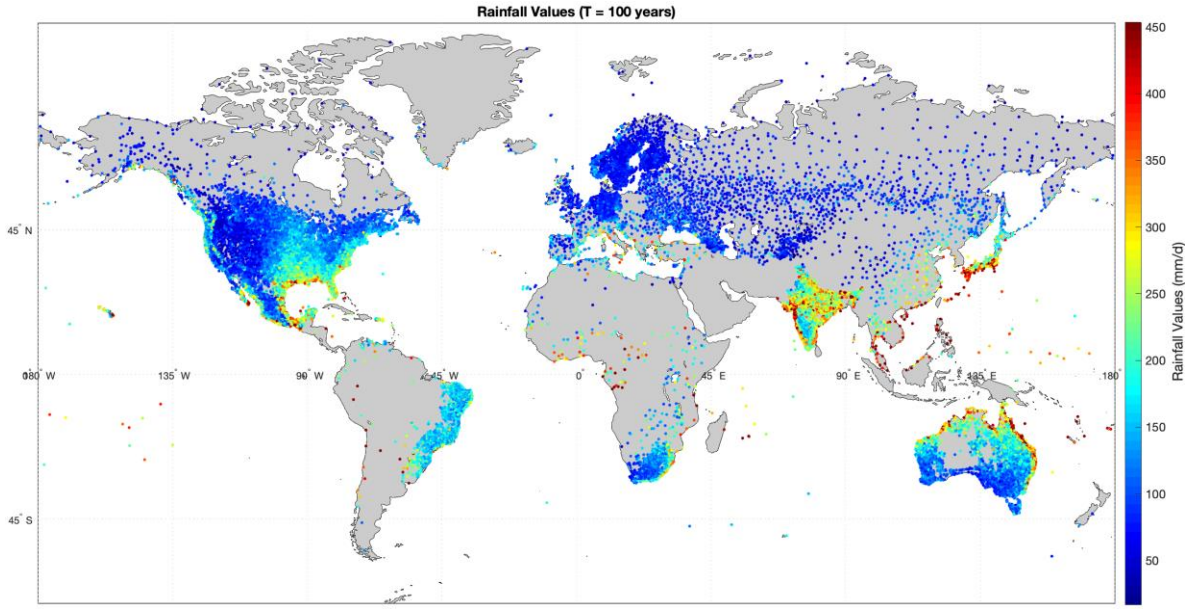


Figure 8.5: Rainfall values (mm/d) for T = 100 years

9 Conclusions

9.1 Research Objectives

This study focused in providing the means of effective extreme-oriented rainfall modelling in a global scale using the newly introduced knowable (K) moments method. While classical moments are simple in their application and are widely used in modelling different natural processes, when focusing on extremes, they fail to produce credible results. This is due to the fact that extremes are closely related to high-order moments which can't be reliably estimated from typical rainfall samples while using classic moments. The L – moments method is also studied, and despite it having the theoretical capabilities of evaluating high-order moments, it fails to take into account the dependence structure existing in almost every natural process and more so in rainfall.

On the other hand, knowable moments combine the advantages of both these methods allowing reliable estimation and description of high-order statistics, whilst retaining classic methods' low-order precision and solving inherent handicaps by conveying the framework for evaluation of long-term dependence bias.

A common yet simple distribution that can reliably model the rainfall process is the three parameter Generalized Pareto Distribution (GPD). However, for being naturally consistent its scale parameter is set to zero, thus the final distribution used is the GPD with two parameters (GPD2). While using three parameters proves to achieve greater accuracy in the final fitting process, it is better to be consistent with the rainfall process' characteristics. For improving the fitting where needed, a variation of the Burr distribution is also showcased, known as the Pareto-Burr-Feller, which uses an extra parameter which in turn assists the fitting process.

For producing comparative results the GHCN – Daily database is used providing with 34,782 stations that met the reliability requirements set. Those stations underwent fitting with all three methods by using the whole hydrometeorological record and the aforementioned GPD2. It is common to set thresholds in order to focus the modelling process on extremes, but by doing so hinders the discovery of any long-term dependence on the sample which affects the final model.

The results of the model are then compared to each other for their effectiveness and efficiency firstly in reliably estimating extremes, and secondly in an overall reliable fit. Moreover, comparison is being made between K – moments fittings accounting for long-term dependence and fittings assuming sample independence. Comparisons are made by use of goodness-of-fit statistic tools namely the RMSE and NRMSE which can showcase the effectiveness of the model in either the distribution's body or the tail.

Finally, the distribution's parameters are assessed for their behaviour in influencing the modelling results. Also, any correlation between them and their respective station's regional climate characteristics is also investigated, consequently discovering similarities between certain climates and certain parameter values.

9.2 Conclusions

From the preliminary fittings using all three methods for all eligible stations of the GHCN – Daily database it is concluded that:

A. While using each station's total hydrometeorological sample, classic methods fail to accurately describe extreme values highly overestimating rainfall values for return periods higher than 1 year. For lower return periods, the fitting is almost perfect to observed values.

B. Classic moments perform better than L – moments in modelling extremes, which consistently show significant overestimation of rainfall values.

C. Knowable moments outperform classic methods, reliably predicting extreme events in most cases for high return periods. However, since the fitting process is focused on an optimization algorithm, focus is given in fitting best for extreme values, thus there is slight loss of accuracy for low orders, with classic methods showing marginal greater fit.

D. The Pareto-Burr-Feller distribution, with the implementation of the extra parameter, keeps the perfect tail fit while also improving it for low return periods, achieving best fit for all return periods. These results are showcased for the sample station.

E. Goodness-of-fit parameters clearly show the effectiveness of knowable moments for the overall fitting process, with only 11% of stations below an NRMSE value of 0.8.

From the overall process of extreme-oriented fitting with knowable moments, many insights on the method can be concluded:

A. As for its optimization process using Least Squares, the overall average value is 1.84, with 89% of stations achieving error below 4 which allows the fit to be deemed as reliable considering the relationship between empirically assigned and theoretical return periods.

B. The effectiveness and indirect correlation between empirical return periods and final model fit is confirmed by the almost equal percentages of good LSE fits and their respective NRMSE values.

C. Long-term dependence bias has a great impact in the final results while using the K-moments approach. The total difference in high return periods by assuming an independent structure and accounting for long-term dependence bias is non-negligible for stations with Hurst coefficient over 0.70. Not including the bias, extremes are underestimated which poses a great risk.

D. Strong positive correlation between the Hurst parameter and the rainfall difference for large return periods is shown with coefficients over 0.85 achieving more than 25% change in rainfall results. Accounting for dependence is vital for getting reliable results and the K – moments method provides the means for accomplishing that.

E. Goodness-of-fit NRMSE values are correlated with the Hurst parameter, since by including the dependence bias the fit is shifted upwards for all return periods, thus worsening the error value in some cases or improving it in others. Lower NRMSE values should not be considered as a flaw of the process, because being naturally consistent is more essential than attaining perfect fit.

From the analysis of the Pareto distribution's parameters it is determined that:

A. Pareto tail index (κ) controls the tail's behaviour where extremes are located and acts as a gauge of how rapidly rainfall values increase over a specific range of high-order return periods and consequently depicts the degree of this increase.

B. From further investigation of the correlation between climatic characteristics and the tail index, it is suggested that climates with consistently high precipitation climates such as tropical (equatorial) and fully humid snow, show mostly low index values. On the contrary, stations situated in arid or Mediterranean climates which receive on average low rainfall with rare extremes being significantly higher than normal, show the highest index values among all.

C. Pareto scale parameter (λ) controls the behaviour of the distribution's body, while indirectly having an important role in modelling extremes, since it depicts the magnitude of extreme values. Correlation between the scale parameter and the station's rainfall average is $r = 0.878$.

D. Applying the same process in finding correlation between the scale parameter and climatic characteristics of a region, it is established that tropical and fully humid warm temperate climates depict high scale parameter values, opposite to arid and snow climates which are connected to low λ values.

E. While both parameters control some aspect of the distribution, their influence in the final attained extreme rainfall values are produced from a combination of theirs. From generalizing this fact, in other words, there is no unconditional correlation between one characteristic and a parameter, but simply a greater influence to it.

9.3 Future Research Potential

While this study elaborates on most aspects of using the K – moments method with the two-parameter Pareto distribution for extreme-oriented rainfall modelling, it still is a preliminary analysis.

In upcoming studies first of all the Pareto distribution, while being parsimonious since having only two parameters, it isn't perfect for every kind of sample provided. In some stations low rainfall values are poorly modelled, while in others, the tail index alone isn't enough to correctly model extremes, as seen from many stations in tropical climates achieving very low tail index and even then, the model wasn't perfect. Thus, other distributions should be studied for their effectiveness like the showcased Pareto-Burr-Feller for improving low-order values.

Furthermore, deeper investigation of the correlation between the distribution's parameters and climatic characteristics can be made, showing in more detail the effect for each region including analysis for the third dimension being altitude.

To further strengthen the reliability of the K – moments method, other rainfall databases should be studied providing with more stations for countries that didn't contribute much to the GHCN – Daily especially countries in Africa or south America.

Knowable moments with the use of the K – climacogram can be used for further analysis of rainfall events at a finer scale, namely for the production of ombrian curves.

10 References

A.N., K., 1941. Dissipation energy in locally isotropic turbulence. *Dokl. Akad. Nauk,* Volume 32, pp. 16-18.

American Meteorological Society, 2015. *Glossary of Meteorology.* [Online]
Available at: http://glossary.ametsoc.org/wiki/

Beck, H. et al., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data,* 5(180214).

Brouers, F., 2015. The Burr XII distribution family and the maximum entropy principle: power-law phenomena are not necessarily nonextensive. *Open J Stat,* Volume 5, pp. 730-741.

Burr, I., 1942. Cumulative Frequency Functions. *Annals of Mathematical Statistics,* Volume 13, pp. 215-235.

Climate Change & Infectious Diseases Group, 2019. *World Maps of Koppen Geiger Classification.* [Online]
Available at: http://koeppen-geiger.vu-wien.ac.at
[Accessed 2 September 2019].

CRED, 2015. *The Human Cost of Weather Related Disasters 1995 - 2015,* s.l.: UNISDR.

Dimitriadis, P., 2017. *Hurst-Kolmogorov dynamics in hydroclimatic processes and in the microscale of turbulence.* Athens: National Technical University of Athens.

Dimitriadis, P. & Koutsoyiannis, D., 2015. processes, Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov. *Stoch Environ Res Risk Assess,* Volume 29, pp. 1649-1669.

Donat, M. G., Alexander, L. V., Herold, N. & Dittus, A. J., 2016. Temperature and precipitation extremes in century-long gridded observations, reanalyses, and atmospheric model simulations. *J. Geophys. Res. Atmos,* 10(1002).

E.J., G., 1958. *Statistics of extremes.* New York: Columbia University Press.

Feller, W., 1970. *An Introduction to Probability and its Applications.* 2nd ed. Chichester, UK: John Wiley & Sons.

Gleason, B. E. et al., 2002. *A new global daily tem- perature and precipitation data set.* Orlando, FL, 13th AMS symposium on global change studies.

Greenwood, J., Landwehr, J., Matalas, N. & Wallis, J., 1979. Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water resources research,* 15(5), pp. 1049-1054.

H.E., H., 1951. Long term storage capacities of reservoirs. *Trans. Am. Soc. Civil Engrs,* Volume 116, pp. 776-808.

Hosking, J., 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society: Series B (Methodological),* 52(1), pp. 105-124.

Hosking, J. & Wallis, J., 1987. Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics,* 29(3), pp. 339-349.

IPCC, 2012. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation,* Cambridge, UK: Cambridge University Press.

IPCC, 2014. *Glossary of terms. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation.* [Online]
Available at:
https://www.ipcc.ch/site/assets/uploads/2018/02/AR5_SYR_FINAL_Annexes.pdf
[Accessed 2019].

Jaffrés, J., 2019. GHCN-Daily: a treasure trove of climate data awaiting discovery. *Computers & geosciences,* Volume 122, pp. 35-44.

Jet Brains, 2019. *PyCharm Edu.* [Online]
Available at:
https://www.jetbrains.com/education/?fromMenu#lang=python&role=learner
[Accessed 2019].

Kottek, M. et al., 2016. World Map of the Köppen-Geiger climate classification updated. *Meteorologische Zeitschrift,* 15(3), pp. 259-263.

Koutsoyiannis, 2019. Knowable moments for high-order stochastic characterization and modelling of hydrological processes (solicited). *Hydrological Sciences Journal,* pp. 19-33.

Koutsoyiannis, D., 2019. *Extreme-oriented selection and fitting of probability distributions.* Vienna, Austria, European Geosciences Union General Assembly 2019, Geophysical Research Abstracts.

Koutsoyiannis, D., Dimitriadis, P., Lombardo, F. & Stevens, S., 2018. From Fractals to Stochastics: Seeking Theoretical Consistency in Analysis of Geophysical Data. In: *Advances in Nonlinear Geosciences.* Cham: Springer.

Landwehr, J., Matalas, N. & Wallis, J., 1979. Probability weighted moments compared with some traditional techniques in estimating Gumbel parameters and quantiles. *Water Resources Research,* 15(5), pp. 1055-1064.

Lombardo, F., Volpi, E., Koutsoyiannis, D. & Papalexiou, S., 2014. Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology. *Hydrol. Earth Syst.,* p. 243–255.

MathWorks, 2019. [Online]
Available at: https://www.mathworks.com
[Accessed 2019].

MathWorks, 2019. *goodnessOfFit.* [Online]
Available at: https://www.mathworks.com/help/ident/ref/goodnessoffit.html
[Accessed 2019].

Menne, M. et al., 2012. An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology,* 29(7), pp. 897-910.

Met Office UK, 2017. *How is climate linked to extreme weather?.* [Online]
Available at: https://www.metoffice.gov.uk/weather/learn-about/climate-and-climate-change/climate/what-affects-climate/extreme-weather
[Accessed 2019].

Min, S., Zhang, X., Zwiers, F. & Hegerl, G., 2011. Human contribution to more-intense precipitation extremes. *Nature,* 470(7334), p. 378.

NOAA, 2019. *COLA Weather and Climate Data.* [Online]
Available at: http://wxmaps.org
[Accessed 2019].

NOAA, 2019. *Global Historical Climate Network Daily - Description.* [Online]
Available at: https://www.ncdc.noaa.gov/ghcn-daily-description
[Accessed 10 August 2019].

Papalexiou, S. M., Koutsoyiannis, D. & Makropoulos, C., 2013. How extreme is extreme? An assessment of daily rainfall distribution tails. *Hydrology and Earth System Sciences,* 17(2), pp. 851-862.

Pickands III, J., 1975. Statistical inference using extreme order statistics. *the Annals of Statistics,* 3(1), pp. 119-131.

PyPi, 2019. *Python Package Index.* [Online]
Available at: https://pypi.org
[Accessed 2019].

Rubel, F. & Kottek, M., 2010. Observed and projected climate shifts 1901–2100 depicted by world maps of the Ko¨ppen-Geiger climate classification. *Meteorologische Zeitschrift,* 19(2), pp. 135-141.

Singh, S. & Maddala, G., 1976. A function for size distribution of incomes. *Econometrica,* Volume 44, pp. 963-970.

Teugels, J., 1975. The class of subexponential distributions. *The Annals of Probability,* pp. 1000-1011.

Volpi, E. et al., 2015. One hundred years of return period: Strengths and limitations. *Water Resources Research,* 51(10), pp. 8570-8585.

Wikipedia, 2019. *Wikipedia.* [Online]
Available at: https://www.wikipedia.org
[Accessed 2019].

World Meteorological Organization, 2016. *Observation components of the Global Observing System.* [Online]
Available at: https://www.wmo.int/pages/prog/www/OSY/Gos-components.html
[Accessed 2019].

11 Appendix

11.1 MATLAB Scripts

Below all scripts and functions from the MATLAB programming interface are provided for insight into the modelling process.

A. General script used for the production of the fitting results for all methods.

```
clear; clc; close all;

fstart=1;
fend=1;

years_limit=30; kb_limit=1.1; excl=360; Hl_lim=0.2; Hu_lim=0.93; kt_max=0.4;
minRP=1;

parameters_fit=zeros(fstart-fend+1,49);

%% Master loop

for f=fstart:fend
    a=exist(['/users/nick_agatheris/New Data/G' num2str(f) '.mat'],'file');
    if a==0
       continue
    else
       fdir=dir(['/users/nick_agatheris/New Data/G' num2str(f) '.mat']);
       file_size=round(fdir.bytes./1000,0);
    end
    if file_size<kb_limit
       continue
    else
       l=load(['/users/nick_agatheris/New Data/G' num2str(f) '.mat']);
       ryo_check=ceil(length(l.TSprec(:,1))/365);
    end
    if ryo_check<years_limit
        continue
    else
        %% Data Extraction

        [prectot,precsort,precsorttot,asc,desc,rdpy2,real_years_obs,p,aa,t0] =
datan(f);
        ch=find(precsort>0);
        %% K-climacogram

        [sc,prec_scaled,Kpq_clim_tot] = Kclimacogram(2,prectot);

        %% Trendline Production

        [Hk,slope_k,gof_k,fit_data_k,exclend] =
Hurst(sc,Kpq_clim_tot,excl,Hl_lim,Hu_lim);
        clc;
        Station=f
        %% Unbiased K - moments

        [Kpq] = Kmoments(aa,desc,p,1,precsort);

        %% Dependence Bias

        [Kpq_d,check,p_d] = bias_correction(Hk,desc,Kpq,p);
```

```
        %% Optimization

        [k_fit,b_fit,total_lse,minTotal_lse,~]
K_optimizer(p_d,Kpq,rdpy2,kt_max,minRP);

        %% Method of Moments

        [k_mom,b_mom,Tmom,Kpq_mom] = MoM(precsort,Kpq,rdpy2);

        %% L - moments

        [k_lm,b_lm,Tlm,Kpq_lm] = Lmoments(precsort,asc,Kpq,rdpy2);

        %% Climacogram

        varprec=var(prec_scaled,1,'omitnan');
        logsc=log(sc);
        stdev_scaled=sqrt(varprec)./sc;
        [Hc,slopec,gofc,fit_datac] = Hurst(sc,stdev_scaled,360,Hl_lim,Hu_lim);

        %% Chart Production

        ylim=1;
        [Tovthr_fit,Tnothr_fit,Tempy_fit,Ttheory_fit]=ReturnPeriods...
            (k_fit,b_fit,p_d,aa,real_years_obs,rdpy2,asc,Kpq);
       % y axis

        Kpq_chart=Kpq(Kpq>=ylim);
        Kpq_chartemp=Kpq(Kpq>=ylim);
        Kpq_mom_chart=Kpq_mom(Kpq_mom>=ylim);
        Kpq_lm_chart=Kpq_lm(Kpq_lm>=ylim);
        precsort_chart=precsort(precsort>=ylim);

        % x axis

        Ttheory_fit_chart=Ttheory_fit(1:length(Kpq_chart));
        if length(Tempy_fit)>=length(Kpq_chartemp)
            Tempy_fit_chart=Tempy_fit(1:length(Kpq_chartemp));
        else
            Tempy_fit_chart=Tempy_fit;
        end
        Tmom_chart=Tmom(1:length(Kpq_mom_chart));
        Tlm_chart=Tlm(1:length(Kpq_lm_chart));

        i=1;
        while Ttheory_fit_chart(1,1)<=1200 % produce adequate T until T=1000y + a
margin
            Kpq_chart=[20+Kpq_chart(1,1);Kpq_chart];

Ttheory_fit_chart=[((1+k_fit.*(Kpq_chart(1,1)./b_fit)).^(1./k_fit))./rdpy2;Ttheor
y_fit_chart];
            i=i+1;
        end

        % Fit Result (Pareto Distribution)

        figure(1) % y axis should start at 1
        loglog(Ttheory_fit_chart,Kpq_chart,'r','linewidth',3) % Fitted Pareto
with K-moments
        hold on
        loglog(Tovthr_fit,precsort(1:length(Tovthr_fit),1),'ob','linewidth',1) %
Values over threshold >=1
        loglog(Tnothr_fit(1:length(precsort_chart)),precsort_chart,'-
b','linewidth',1.4) % All values
```

```
        loglog(Tmom_chart,Kpq_mom_chart,'--b','linewidth',2) % Fitted Pareto with
MoM
        loglog(Tlm_chart,Kpq_lm_chart,'--k','linewidth',2) % Fitter Pareto with
LM
        loglog(Tempy_fit_chart,Kpq_chartemp(1:length(Tempy_fit_chart)),'--
g','linewidth',1.2), % K-moments empirical RP
        xlabel('Return Period T (years)')
        ylabel('Precipitation (mm/d)')
        title(['Pareto Distribution Fit (' num2str(t0.name) ')'])
        grid on
        legend({'Theoretical Pareto','Order Statistics (VOT)','Order Statistics
(All)',...
            'Moments','L - Moments','Empirical K -
moments'},'Location','Northwest','FontSize',11)

        % Climacogram

        figure(2)
        plot(fit_datac)
        set(gca,'XScale','log')
        set(gca,'YScale','log')
        hold on
        loglog(sc,stdev_scaled,'b','linewidth',2)
        xlabel('Time Scale (days)')
        ylabel('Scaled Standard Deviation')
        title('Climacogram')
        legend('Power trendline','Climacogram')
        grid on

        % Daily Precipitation Data

        figure(2)
        plot(prectot,'b')
        xlabel('Time (Days)')
        ylabel('Precipitation (mm/d)')
        title('Daily Precipitation Data')
        grid on

        % K - climacogram / Fitted Power Curve

        figure(3)
        plot(fit_data_k)
        set(gca,'XScale','log')
        set(gca,'YScale','log')
        grid on
        hold on
        loglog(sc,Kpq_clim_tot,'linewidth',2)
        xlabel('Time Scale (Days)')
        ylabel('Central K-moment Value')
        title('K - Climacogram / Fitted Power Curve')
        legend({'Fitted Power Curve','K - Climacogram'},'FontSize',11)

        figure(4) % MoM - LM
        loglog(Tmom_chart,Kpq_mom_chart,'-r','linewidth',2)
        hold on
        loglog(Tlm_chart,Kpq_lm_chart,'-k','linewidth',2)
        loglog(Tovthr_fit,precsort(1:length(Tovthr_fit),1),'ob','linewidth',1) %
        loglog(Tnothr_fit(1:length(precsort_chart)),precsort_chart,'-
b','linewidth',1.5)
        xlabel('Return Period T (years)')
        ylabel('Precipitation (mm/d)')
        title(['Pareto Distribution Fit - Classic Methods (' num2str(t0.name)
')'])
        grid on
        legend({'Raw moments','L - moments','Order Statistics (VOT)','Order
Statistics (All)',...
            },'Location','Northwest','FontSize',11)
```

```
        figure(5) % K - moments
        loglog(Ttheory_fit_chart,Kpq_chart,'r','linewidth',3)
        hold on
        loglog(Tovthr_fit,precsort(1:length(Tovthr_fit),1),'ob','linewidth',1) %
        loglog(Tnothr_fit(1:length(precsort_chart)),precsort_chart,'-
b','linewidth',2)
        loglog(Tempy_fit_chart,Kpq_chartemp(1:length(Tempy_fit_chart)),'--
g','linewidth',1.5),
        xlabel('Return Period T (years)')
        ylabel('Precipitation (mm/d)')
        axis([0 10000 1 300])
        title(['Pareto Distribution Fit - K-moments (' num2str(t0.name) ')'])
        grid on
        legend({'Theoretical Pareto','Order Statistics (VOT)','Order Statistics
(All)'...
            ,'Empirical K - moments'},'Location','Northwest','FontSize',11)

         %% Errors

        [LSE,RMSE,NRMSE,perc,Ter,Xer,Ttheory_fit_exp,Tmom_exp,Tlm_exp,...
            Kpq_exp,Kpq_mom_exp,Kpq_lm_exp,K_100,K_1000] =...
            errors(precsort,Kpq,Kpq_mom,Kpq_lm,rdpy2,...
            k_fit,k_mom,k_lm,b_fit,b_mom,b_lm,Tnothr_fit,Ttheory_fit,Tmom,Tlm);

        %% Data Takeoff

        parameters_fit(f-fstart+1,1)=f;
        parameters_fit(f-fstart+1,2)=file_size;
        parameters_fit(f-fstart+1,3)=t0.latlon_TS(1,1);
        parameters_fit(f-fstart+1,4)=t0.latlon_TS(1,2);
        parameters_fit(f-fstart+1,5)=length(prectot);
        parameters_fit(f-fstart+1,6)=Hk;
        parameters_fit(f-fstart+1,7)=k_fit;
        parameters_fit(f-fstart+1,8)=b_fit;
        parameters_fit(f-fstart+1,9)=round(minTotal_lse,3);
        parameters_fit(f-fstart+1,10)=round(k_mom,3);
        parameters_fit(f-fstart+1,11)=round(b_mom,3);
        parameters_fit(f-fstart+1,12)=round(k_lm,3);
        parameters_fit(f-fstart+1,13)=round(b_lm,3);
        parameters_fit(f-fstart+1,14)=perc(1);
        parameters_fit(f-fstart+1,15)=perc(2);
        parameters_fit(f-fstart+1,16)=perc(3);
        parameters_fit(f-fstart+1,17)=perc(4);
        parameters_fit(f-fstart+1,18)=perc(5);
        parameters_fit(f-fstart+1,19)=perc(6);
        parameters_fit(f-fstart+1,20)=LSE(1,1);
        parameters_fit(f-fstart+1,21)=LSE(1,2);
        parameters_fit(f-fstart+1,22)=LSE(1,3);
        parameters_fit(f-fstart+1,23)=LSE(2,1);
        parameters_fit(f-fstart+1,24)=LSE(2,2);
        parameters_fit(f-fstart+1,25)=LSE(2,3);
        parameters_fit(f-fstart+1,26)=LSE(3,1);
        parameters_fit(f-fstart+1,27)=LSE(3,2);
        parameters_fit(f-fstart+1,28)=LSE(3,3);
        parameters_fit(f-fstart+1,29)=RMSE(1,1);
        parameters_fit(f-fstart+1,30)=RMSE(1,2);
        parameters_fit(f-fstart+1,31)=RMSE(1,3);
        parameters_fit(f-fstart+1,32)=RMSE(2,1);
        parameters_fit(f-fstart+1,33)=RMSE(2,2);
        parameters_fit(f-fstart+1,34)=RMSE(2,3);
        parameters_fit(f-fstart+1,35)=RMSE(3,1);
        parameters_fit(f-fstart+1,36)=RMSE(3,2);
        parameters_fit(f-fstart+1,37)=RMSE(3,3);
        parameters_fit(f-fstart+1,38)=NRMSE(1,1);
        parameters_fit(f-fstart+1,39)=NRMSE(1,2);
        parameters_fit(f-fstart+1,40)=NRMSE(1,3);
```

```
        parameters_fit(f-fstart+1,41)=NRMSE(2,1);
        parameters_fit(f-fstart+1,42)=NRMSE(2,2);
        parameters_fit(f-fstart+1,43)=NRMSE(2,3);
        parameters_fit(f-fstart+1,44)=NRMSE(3,1);
        parameters_fit(f-fstart+1,45)=NRMSE(3,2);
        parameters_fit(f-fstart+1,46)=NRMSE(3,3);
        parameters_fit(f-fstart+1,47)=K_100;
        parameters_fit(f-fstart+1,48)=K_1000;
        parameters_fit(f-fstart+1,49)=exclend;

    end
end

%% Exportable Arrays

[names,export,export_table]=takeoff(parameters_fit);
[extra_table] = extrasn(fstart,fend,kb_limit,years_limit);

writetable(export_table,...
    ['/users/nick_agatheris/desktop/Simulation Results/matlab output/fit
results/Fit Results (G'...
    num2str(fstart) ' - G' num2str(fend) ').xlsx'])
```

B. General script for heat map production

```
clear; clc; close all;

a=load('extras_used.mat');
b=load('Wmpk.mat');
c=load('Wmpl.mat');
d=load('Wmplse.mat');
e=load('YearsObs.mat');
par=load('Parameters.mat');
r=load('Rainfall.mat');
h=load('Hurst.mat');

hurst=h.FitResultsGNewv2;
rain=r.FitResultsGNewv2;
parameters=par.FitResultsGNewv2;
yo=e.FitResultsGNewv2;
wmp_totk=b.WorldMapParametersS1;
wmpk=wmp_totk(all(wmp_totk,2),:);
wmp_totl=c.WorldMapParametersS2;
wmpl=wmp_totl(all(wmp_totl,2),:);
wmp_totlse=d.WorldMapParametersS3;
wmplse=wmp_totlse(all(wmp_totlse,2),:);

ms=input('Marker Size: ');
mc=0.8;
mapcolor=[mc mc mc];
ss=[100, 400, 200, 200];

figure(4)
f=worldmap([-65 85],[-180 180]);
setm(gca,'mapprojection','miller','Frame','on','FlineWidth',0.7)
geoshow('landareas.shp', 'FaceColor',mapcolor,'DefaultEdgeColor', 'k')
PointLatLon = [yo(:,1) yo(:,2)];
mValue = parameters(:,1);
plotm(PointLatLon(:,1),PointLatLon(:,2),'w.');
markerSize = ms;
scatterm(PointLatLon(:,1), PointLatLon(:,2), markerSize, mValue,'Filled');
colormap(f);
set(gca,'CLim',[min(mValue),max(mValue)-0.15]);
cb=colorbar('FontSize',12);
```

```
set(0,'DefaultFigureColormap',feval('jet'));
tightmap;

make_it_tight = true;
subplot = @(m,n,p) subtightplot (m, n, p, [0.04 0.01], [0.02 0.01], [0.1 0.01]);
if ~make_it_tight,  clear subplot;  end

figure(5)
for i=1:input('Subplots: ')
    s1=subplot(2,2,i);
    region=input('Region: ','s');
    f=worldmap(region);
    setm(gca,'mapprojection','miller','Frame','on','FlineWidth',0.7)
    geoshow('landareas.shp', 'FaceColor',mapcolor,'DefaultEdgeColor', 'k')
    PointLatLon = [yo(:,1) yo(:,2)];
    mValue = parameters(:,1);
    plotm(PointLatLon(:,1),PointLatLon(:,2),'w.');
    markerSize = ms;
    scatterm(PointLatLon(:,1), PointLatLon(:,2), markerSize, mValue,'Filled');
    colormap(f);
    set(gca,'CLim',[min(mValue),max(mValue)-0.15]);
    tightmap;
end
```

C. Initial data processing

```
function [prectot,precsort,precsorttot,asc,desc,rdpy,real_years_obs,p,aa,t0] =
datan(f)
% Used for extracting data of .mat files from precipitation stations
% library. Files have been renamed in ascending order with numbers in order
% to make loading easier. Provides mainframe for making data accesible for
% other functions in script (order statistics)

t0=load(['/users/nick_agatheris/New Data/G' num2str(f) '.mat']);

prectot=t0.TSprec(:,1);
precsorttot=sort(prectot,'descend'); % all sorted data
precsort=precsorttot(precsorttot~=0); % non-zero sorted data

% calculations are for non-zero data

desc=(length(precsort):-1:1)'; % descending numbers from last to first
observation
aa=length(precsort); % size of non-zero sample
asc=(1:aa)';
real_years_obs=ceil(length(prectot)/365); % real years observed with zeros from
total observations

r=aa; % starting p=pmax
% rdpy=ceil(length(precsort)/years_obs); % rain days per year
rdpy=ceil(length(precsort)/real_years_obs);

i=1;
p(1,i)=aa;
while p(1,i)>0.01 % produces adequate p until p=0.01
    p(1,i+1)=p(1,i)/1.04;
    i=i+1;
end

end
```

D. Production of K – climacogram

```
function [sc,prec_scaled,Kpq_clim_tot] = Kclimacogram(pclim,prectot)
% Production of sorted climacogram array and K-moments from this scaled
% array used for plotting the K-climacogram of central K-moments on time
% scales up to 1/10 of the total sample observations for investigating long term
persistence with the use
% of the Hurst parameter in a later stage.

qclim=1; sc=1; i=1;
while sc(1,i)<ceil(1/10*length(prectot))+1000 % 1/10*length(prectot)
    sc(1,i+1)=ceil(1.1.*sc(1,i)); % scale array for producing K-clim
    i=i+1;
end

[~,csc]=size(sc);
for i=1:csc
    numz(1,i)=floor(length(prectot)./sc(i)); % number of elements in each scale
(column)
end

prec_scaled=zeros(length(prectot),csc);
prec_scaled(:,1)=prectot;
for j=2:size(numz,2)
    for i=1:numz(1,j)
        l=(i-1)*sc(1,j)+1:i*sc(1,j);
        y=prectot(l);
        prec_scaled(i,j)=sum(y);
    end
end

for w=1:length(sc)
    for e=1:max(numz)
        if e>numz(w)
           prec_scaled(e,w)=NaN;
        end
    end
end

precsort_scaled=sort(prec_scaled./sc);
Kpq_clim_tot=zeros(length(sc),1);

for i=1:length(sc)
    asc_clim=(1:numz(:,i))';
    aak=numz(i);
    [Kpq_clim]=Kmoments_c(aak,asc_clim,pclim,1,precsort_scaled(:,i));
    Kpq_clim_tot(i,1)=2.*Kpq_clim;

end

end
```

E. Estimation of Hurst parameter from fitted trendline power curve

```
function [H,slope,gof,fit_data,excl,sc_trend,Kpq_trend] =
Hurst(sc,Kpq_clim_tot,excl,l_lim,u_lim)
% Fitting of power trendline to K-moments produced from Kclimacogram in
% order to estimate the Hurst parameter of the sample. excl input gives
% freedom in choosing the min scale which is considered important in long
% term dependence. Results provide the raw scale of the curve, the goodness
% of fit statistics, combined with the Hurst parameter.

Kpq_trend=Kpq_clim_tot(Kpq_clim_tot~=0);
Kpq_trend=Kpq_trend(~isnan(Kpq_trend));
sc_trend=sc(1:length(Kpq_trend));
[xData, yData] = prepareCurveData( sc_trend, Kpq_trend );
```

```
r=(find(sc_trend>=360));
if length(r)>=20
    excludedPoints = xData <= excl;
else
    excl=50;
    excludedPoints = xData <= excl;
end

% Set up fittype and options.

ft = fittype( 'power1' );
opts = fitoptions( 'Method', 'NonlinearLeastSquares' );
opts.Display = 'Off';
opts.Robust = 'Bisquare';
opts.Exclude = excludedPoints;

% Fit model to data.

[fit_data,gof] = fit( xData, yData, ft, opts);
coef=coeffvalues(fit_data);
slope=coef(1,2);
H=round(1+slope,2);

if H>=u_lim || H<=l_lim
    excl=0;
    [xData, yData] = prepareCurveData( sc_trend, Kpq_trend );
    excludedPoints = xData <= excl;

    % Set up fittype and options.

    ft = fittype( 'power1' );
    opts = fitoptions( 'Method', 'NonlinearLeastSquares' );
    opts.Display = 'Off';
    opts.Robust = 'Bisquare';
    opts.Exclude = excludedPoints;

    % Fit model to data.

    [fit_data,gof] = fit( xData, yData, ft, opts);
    coef=coeffvalues(fit_data);
    slope=coef(1,2);
    H=round(1+slope,2);
else
    H=H;
end

end
```

F. Production of unbiased non-central K – moments

```
function [Kpq1] = Kmoments(aa,desc,p,q,precsort)
% Production of Unbiased non-central K-moments. Moments are calculated
% using the theoretical (exact) estimator with the denoted binp for q=1 and p up
to
% the size of the sample. Non-central K-moments are used for the fitting
% process.

forKpq1=gammaln(desc)-gammaln(aa)-log(aa);
Kpq1=zeros(length(p),1);

for i=1:length(p)
    pm1=p(1,i);
    j=1:aa;
    ispos=desc(j,1)-pm1+1>=0;
    nonzero=find(ispos~=0);
```

```
    Kpqstart1=(precsort(nonzero,1)).^q.*exp(gammaln(aa-pm1+1)-
gammaln(desc(nonzero,1)-pm1+1)...
        +log(pm1)+forKpq1(nonzero,1));
    Kpq1(i,1)=sum(Kpqstart1);
end

end
```

G. Long-term dependence bias correction to moment orders

```
function [Kpq_d,check,p_d] = bias_correction(Hk,desc,Kpq,p)
% Estimation of dependence bias for calculating non-central K-moments.
% Theta parameter is calculated and depends solely on the Hurst parameter
% and the size of the sample (non-zero). If n is too high or H is 0.5 theta
% approximates to zero and bias is non existent.

theta=(2*Hk*(1-Hk))/(desc(1)-1)-1/(2*(desc(1)-1)^(2-2*Hk));
check=abs(theta)>0.001;
if abs(theta)<0.001
    Kpq_d=Kpq;
else
    Kpq_d=(1+theta).*Kpq;
end
p_d=2*theta+(1-2*theta).*p.^((1+theta)^2);
p_d=p_d(p_d>=0.01);
end
```

H. Optimization process for K – moments

```
function [k_fit1,b_fit1,total_lse1,minTotal_lse1,Tempy_d] =
K_optimizer(p,Kpq,rdpy,kt_max,minRP)
% Optimizer for calculating the minimum Least Squared Error for the
% best Pareto distributions parameters k & b. Use for unbiased non-central K-
moments
% comparing theoretical RP to empirical RP obtained by the L parameter.
% Error minimization occurs for RPs with p>1 to focus

kt=[0.001,0.002,0.004,0.005,0.006:0.002:kt_max];
bt=1:0.05:38;

total_lse1=zeros(length(kt),length(bt));

lexact1=zeros(length(kt),length(p));
for i=1:length(kt)
    kl1=kt(i);
    lexact1(i,:)=((pi./(sin(pi.*kl1).*beta(kl1,p+1-kl1))).^(1./kl1))./p; % L
factor for every p value
end

for i=1:length(kt)
     kl1=kt(i);
    for j=1:length(bt)
        bl1=bt(j);
%         [~,~,Tempy,Ttheory]=ReturnPeriods(kl,bl,p,aa,ryo,rdpy,asc,Kpq);
        Tempd1=(lexact1(i,:).*p)'; % Empirical Daily Return Periods using L
factor for Pareto dist
        Tempy1=Tempd1/rdpy; % Empirical Yearly Return Periods using L factor for
Pareto dist
        Ttheord1=(1+kl1.*(Kpq./bl1)).^(1/kl1); % Theoretical Daily Return Periods
using Pareto dist
        Ttheory1=Ttheord1/rdpy; % Theoretical Yearly Return Periods using Pareto
dist
        [row_Tmin,~]=find(Tempy1<=minRP,1);
            for l1=1:(row_Tmin-1)
```

```
                lse1(l1,1)=(log(Ttheory1(l1,1)/Tempy1(l1,1))).^2;
            end
        total_lse1(i,j)=sum(lse1);
    end
end

minTotal_lse1=min(total_lse1(:)); % Total error - parameters
[k_posfit1,b_posfit1]=find(total_lse1==minTotal_lse1);
k_fit1=kt(k_posfit1);
b_fit1=bt(b_posfit1);

lexact=((pi./(sin(k_fit1*pi).*beta(k_fit1,p+1-k_fit1))).^(1/k_fit1))./p;
Tempd=(lexact.*p)'; % Empirical Daily Return Periods using L factor for Pareto
dist
Tempy_d=Tempd/rdpy; % Empirical Yearly Return Periods using L factor for Pareto
dist

end
```

I. Parameter estimation using classic moments

```
function [k_mom,b_mom,Tmom,Kpq_mom] = MoM(precsort,Kpq,rdpy)
% Parameter estimation using classic method of moments (MoM). Used with
% conjuction with already found K-moments as it is more convenient in later
% chart production. Equations used for estimation are theoretical for the Pareto
% distribution using MoM.

avrg=mean(precsort);
vrnc=var(precsort);
stdev=sqrt(vrnc);

k_mom=abs(0.5*(avrg.^2/vrnc-1)); % parameter production from MoM equations for
Pareto dist (2P)
b_mom=0.5*avrg*(avrg.^2/vrnc+1);

Kpq_mom=Kpq;
i=1;
Tmom_test=0;
while Tmom_test<=1200
    Kpq_mom=[20+Kpq_mom(1,1);Kpq_mom];
    Tmom_test=((1+k_mom.*(Kpq_mom(1,1))./b_mom).^(1/k_mom))/rdpy;
    i=i+1;
end

Tmom=((1+k_mom.*(Kpq_mom)./b_mom).^(1/k_mom))/rdpy;

end
```

J. Parameter estimation using L – moments

```
function [k_lm,b_lm,Tlm,Kpq_lm] = Lmoments(precsort,asc,Kpq,rdpy)
% Parameter estimation using method of L-moments. Used with
% conjuction with already found K-moments as it is more convenient in later
% chart production. Equations used for estimation are theoretical for the Pareto
% distribution using L-moments (or PWM).

precrev=sort(precsort,'ascend');

bsi=[(1-(asc-0.35)./asc(end)).*precrev,(1-(asc-0.35)./asc(end)).^2.*precrev];
bi=[mean(precrev),1/asc(end)*sum(bsi(:,1))];
li=[bi(1),(2*bi(2)-bi(1))];

k_lm=abs(bi(1)/(bi(1)-2*bi(2))-2); % parameter production from PWM equations for
Pareto dist (2P)
```

```
b_lm=2*(bi(1)*bi(2))/(bi(1)-2*bi(2));

Kpq_lm=Kpq;
i=1;
Tlm_test=0;
while Tlm_test<=1200
    Kpq_lm=[20+Kpq_lm(1,1);Kpq_lm];
    Tlm_test=((1+k_lm.*Kpq_lm(1,1)./b_lm).^(1/k_lm))/rdpy;
    i=i+1;
end

Tlm=((1+k_lm.*Kpq_lm./b_lm).^(1/k_lm))/rdpy;

end
```

K. Return periods estimation

```
function [Tovthr,Tnothr,Tempy,Ttheory] = ReturnPeriods(k,b,p,aa,ryo,rdpy,asc,Kpq)

lexact=((pi./(sin(k*pi).*beta(k,p+1-k))).^(1/k))./p; % L factor for every p value
i=1;
Tovthr(i,1)=(ryo+1)/asc(1,1); % Observed Return Periods until T=1
if asc(end)>ryo
    while Tovthr(i,1)>1
        Tovthr(i+1,1)=(ryo+1)/asc(i+1,1);
        i=i+1;
    end
else
    for i=2:asc(end)
        Tovthr(i,1)=(ryo+1)/asc(i,1);
    end
end

Tnothr=(aa+1)./(asc.*rdpy); % Observed Return Period for all non-zero values

Tempd=(lexact.*p)'; % Empirical Daily Return Periods using L factor for Pareto
dist
Tempy=Tempd/rdpy; % Empirical Yearly Return Periods using L factor for Pareto
dist

Ttheord=(1+k.*(Kpq./b)).^(1/k); % Theoretical Daily Return Periods using Pareto
dist
Ttheory=Ttheord/rdpy; % Theoretical Yearly Return Periods using Pareto dist

end
```

L. Error evaluation framework

```
function
[LSE,RMSE,NRMSE,perc,Ter,Xer,Ttheory_fit_exp,Tmom_exp,Tlm_exp,Kpq_exp,Kpq_mom_exp
,Kpq_lm_exp,K_100,K_1000] = errors(...
    precsort,Kpq_d,Kpq_mom,Kpq_lm,rdpy,...
    k_fit,k_mom,k_lm,b_fit,b_mom,b_lm,Tnothr_fit,Ttheory_fit,Tmom,Tlm)
% Framework for calculating differences and errors between empirical,
% theoretical, and fitted data curves produced with each method.

% Expand RP to reach T=1000y

Ttheory_fit_exp=Ttheory_fit;
Tmom_exp=Tmom;
Tlm_exp=Tlm;
Kpq_exp=Kpq_d;
Kpq_mom_exp=Kpq_mom;
Kpq_lm_exp=Kpq_lm;
```

```
i=1;
while Ttheory_fit_exp(1,1)<=1200 % K-moments
    Kpq_exp=[20+Kpq_exp(1,1);Kpq_exp];

Ttheory_fit_exp=[((1+k_fit.*(Kpq_exp(1,1)./b_fit)).^(1./k_fit))./rdpy;Ttheory_fit
_exp];
    i=i+1;
end
while Tmom_exp(1,1)<=1200 % Moments
    Kpq_mom_exp=[20+Kpq_mom_exp(1,1);Kpq_mom_exp];
    Tmom_exp=[((1+k_mom.*(Kpq_mom_exp(1,1)./b_mom)).^(1./k_mom))./rdpy;Tmom_exp];
    i=i+1;
end
while Tlm_exp(1,1)<=1200 % L-moments
    Kpq_lm_exp=[20+Kpq_lm_exp(1,1);Kpq_lm_exp];
    Tlm_exp=[((1+k_lm.*(Kpq_lm_exp(1,1)./b_lm)).^(1./k_lm))./rdpy;Tlm_exp];
    i=i+1;
end

% Precipitation value for each method in either T=100y and T=1000y

K_100=interp1(Ttheory_fit_exp,Kpq_exp,100);
K_1000=interp1(Ttheory_fit_exp,Kpq_exp,1000);
M_100=interp1(Tmom_exp,Kpq_mom_exp,100);
M_1000=interp1(Tmom_exp,Kpq_mom_exp,1000);
L_100=interp1(Tlm_exp,Kpq_lm_exp,100);
L_1000=interp1(Tlm_exp,Kpq_lm_exp,1000);

% Percentage difference between methods for T=100y and T=1000y

KM_100=((M_100-K_100)/K_100)*100;
KM_1000=((M_1000-K_1000)/K_1000)*100;
KL_100=((L_100-K_100)/K_100)*100;
KL_1000=((L_1000-K_1000)/K_1000)*100;
ML_100=((L_100-M_100)/M_100)*100;
ML_1000=((L_1000-M_1000)/M_1000)*100;

% Calculate Prec with same RP

j=0;
for i=-3:0.01:3 % RP from 10^-3 to 10^3
    j=j+1;
    Ter(j,1)=10^i; % same step RP
end


for i=1:length(Ter)
    Xobs(i,1)=interp1(Tnothr_fit,precsort,Ter(i));
    Xk(i,1)=interp1(Ttheory_fit_exp,Kpq_exp,Ter(i));
    Xm(i,1)=interp1(Tmom_exp,Kpq_mom_exp,Ter(i));
    Xl(i,1)=interp1(Tlm_exp,Kpq_lm_exp,Ter(i));
end

Xobslog=isnan(Xobs);
Xklog=isnan(Xk);
Xmlog=isnan(Xm);
Xllog=isnan(Xl);

% Generate arrays w/o NaN for RMSE calculation

j=1; % Xk
for i=1:length(Xobslog)
    if Xobslog(i)==0 && Xklog(i)==0
       Xobskrmse(j,1)=Xobs(i);
       Xkrmse(j,1)=Xk(i);
```

```
        Terk(j,1)=Ter(i);
        j=j+1;
    else
        continue
    end
end
j=1; % Xm
for i=1:length(Xobslog)
    if Xobslog(i)==0 && Xmlog(i)==0
        Xobsmrmse(j,1)=Xobs(i);
        Xmrmse(j,1)=Xm(i);
        Term(j,1)=Ter(i);
        j=j+1;
    else
        continue
    end
end
j=1; % Xl
for i=1:length(Xobslog)
    if Xobslog(i)==0 && Xllog(i)==0
        Xobslrmse(j,1)=Xobs(i);
        Xlrmse(j,1)=Xl(i);
        Terl(j,1)=Ter(i);
        j=j+1;
    else
        continue
    end
end

% Least Squares Error between curves

LSE_k=sum((log(Xobs./Xk)).^2,'omitnan');
LSE_m=sum((log(Xobs./Xm)).^2,'omitnan');
LSE_l=sum((log(Xobs./Xl)).^2,'omitnan');

% Least Squares Error for T>1y

[rhigh,~]=find(Ter>=1,1);
LSE_khigh=sum((log(Xobs(rhigh:end,1)./Xk(rhigh:end,1))).^2,'omitnan');
LSE_mhigh=sum((log(Xobs(rhigh:end,1)./Xm(rhigh:end,1))).^2,'omitnan');
LSE_lhigh=sum((log(Xobs(rhigh:end,1)./Xl(rhigh:end,1))).^2,'omitnan');

% Least Squares Error for T<1y

LSE_klow=LSE_k-LSE_khigh;
LSE_mlow=LSE_m-LSE_mhigh;
LSE_llow=LSE_l-LSE_lhigh;

% Root Mean Square Error

method='MSE';
RMSE_k=sqrt(goodnessOfFit(Xkrmse,Xobskrmse,method));
RMSE_m=sqrt(goodnessOfFit(Xmrmse,Xobsmrmse,method));
RMSE_l=sqrt(goodnessOfFit(Xlrmse,Xobslrmse,method));

% Root Mean Square Error for T>1y

[rhighk,~]=find(Terk>=1,1); [rhighm,~]=find(Term>=1,1);
[rhighl,~]=find(Terl>=1,1);
RMSE_khigh=sqrt(goodnessOfFit(Xkrmse(rhighk:end,1),Xobskrmse(rhighk:end,1),method
));
RMSE_mhigh=sqrt(goodnessOfFit(Xmrmse(rhighm:end,1),Xobsmrmse(rhighm:end,1),method
));
RMSE_lhigh=sqrt(goodnessOfFit(Xlrmse(rhighl:end,1),Xobslrmse(rhighl:end,1),method
));
```

```
% Root Mean Square Error for T<1y

RMSE_klow=sqrt(goodnessOfFit(Xkrmse(1:rhighk-1,1),Xobskrmse(1:rhighk-
1,1),method));
RMSE_mlow=sqrt(goodnessOfFit(Xmrmse(1:rhighm-1,1),Xobsmrmse(1:rhighm-
1,1),method));
RMSE_llow=sqrt(goodnessOfFit(Xlrmse(1:rhighl-1,1),Xobslrmse(1:rhighl-
1,1),method));

% Normalised Root Mean Square Error

method='NRMSE';
NRMSE_k=goodnessOfFit(Xkrmse,Xobskrmse,method);
NRMSE_m=goodnessOfFit(Xmrmse,Xobsmrmse,method);
NRMSE_l=goodnessOfFit(Xlrmse,Xobslrmse,method);

% Normalised Root Mean Square Error for T>1y

[rhighk,~]=find(Terk>=1,1); [rhighm,~]=find(Term>=1,1);
[rhighl,~]=find(Terl>=1,1);
NRMSE_khigh=goodnessOfFit(Xkrmse(rhighk:end,1),Xobskrmse(rhighk:end,1),method);
NRMSE_mhigh=goodnessOfFit(Xmrmse(rhighm:end,1),Xobsmrmse(rhighm:end,1),method);
NRMSE_lhigh=goodnessOfFit(Xlrmse(rhighl:end,1),Xobslrmse(rhighl:end,1),method);

% Normalised Root Mean Square Error for T<1y

NRMSE_klow=goodnessOfFit(Xkrmse(1:rhighk-1,1),Xobskrmse(1:rhighk-1,1),method);
NRMSE_mlow=goodnessOfFit(Xmrmse(1:rhighm-1,1),Xobsmrmse(1:rhighm-1,1),method);
NRMSE_llow=goodnessOfFit(Xlrmse(1:rhighl-1,1),Xobslrmse(1:rhighl-1,1),method);

% Takeoff array

perc=[round(KM_100,1),round(KM_1000,1),round(KL_100,1),...
    round(KL_1000,1),round(ML_100,1),round(ML_1000,1)];
LSE=[LSE_k,LSE_m,LSE_l;LSE_khigh,LSE_mhigh,LSE_lhigh;LSE_klow,LSE_mlow,LSE_llow];
if isnan(RMSE_klow)
    RMSE=[RMSE_k,RMSE_m,RMSE_l;RMSE_khigh,RMSE_mhigh,RMSE_lhigh;0,0,0];
else

RMSE=[RMSE_k,RMSE_m,RMSE_l;RMSE_khigh,RMSE_mhigh,RMSE_lhigh;RMSE_klow,RMSE_mlow,R
MSE_llow];
end

if isnan(NRMSE_klow)
    NRMSE=[NRMSE_k,NRMSE_m,NRMSE_l;NRMSE_khigh,NRMSE_mhigh,NRMSE_lhigh;0,0,0];
else

NRMSE=[NRMSE_k,NRMSE_m,NRMSE_l;NRMSE_khigh,NRMSE_mhigh,NRMSE_lhigh;NRMSE_klow,NRM
SE_mlow,NRMSE_llow];
end
Xer=[Xobs,Xk,Xm,Xl];
Xerrmsek=[Xobskrmse,Xkrmse];
Xerrmsem=[Xobsmrmse,Xmrmse];
Xerrmsel=[Xobslrmse,Xlrmse];

varNames2={'KM_100y','KM_1000y','KL_100y','KL_1000y','ML_100y','ML_1000y'};
```

M. Data takeoff

```
function [varNames,parameters_fit_export,parameters_fit_export_table] =
takeoff(parameters_fit)
% Framework for exporting data in a matrix form factor. csv_comp can be
% used in Python but is unecessary.

varNames={'N','KB','Lat','Lon','Obs','Hurst','k_fit','b_fit','LSE',...
```

```
'k_MoM','b_MoM','k_LM','b_LM','KM_100y','KM_1000y','KL_100y','KL_1000y','ML_100y'
,'ML_1000y',...

'LSE_k','LSE_m','LSE_l','LSE_khigh','LSE_mhigh','LSE_lhigh','LSE_klow','LSE_mlow'
,'LSE_llow'...

'RMSE_k','RMSE_m','RMSE_l','RMSE_khigh','RMSE_mhigh','RMSE_lhigh','RMSE_klow','RM
SE_mlow','RMSE_llow',...

'NRMSE_k','NRMSE_m','NRMSE_l','NRMSE_khigh','NRMSE_mhigh','NRMSE_lhigh','NRMSE_kl
ow','NRMSE_mlow','NRMSE_llow',...
    'K_100','K_1000','Exclim'};
parameters_fit_export=parameters_fit;
parameters_fit_export(all(~parameters_fit_export,2),:)=[]; % remove zero rows
from exported .xlsx
parameters_fit_export_table=array2table(parameters_fit_export,'VariableNames',var
Names);
```

11.2 Python Scripts

Below all scripts from the PyCharm Python programming interface are provided for insight into the evaluation of location data provided from the GHCN – Daily database (coordinate geocoding).

```
import time
from tqdm import tqdm
import pandas
from numpy import *
from geopy.geocoders import Nominatim
from geopy.geocoders import GoogleV3
from geopy.extra.rate_limiter import RateLimiter
import xlrd

start_time = time.time()
name = '1-112777'
dir = ("/Users/Nick_Agatheris/Desktop/Simulation Results/MATLAB Output/Coordinates/Coordinates
("+str(name)+").csv")

# read .csv and write to .xlsx

df_paremeters = pandas.read_csv(dir)
excel_name = '/Users/Nick_Agatheris/Desktop/Simulation Results/Python Input/Coordinates
('+str(name)+').xlsx'
writer = pandas.ExcelWriter(excel_name, engine='xlsxwriter')
df_paremeters.to_excel(writer, "Sheet1", header=False)
writer.save()

# read from .xlsx and write coordinates to numpy array

xlsxopen=xlrd.open_workbook(excel_name)
sheet = xlsxopen.sheet_by_index(0)
sheet.cell_value(0, 0)
row_number=sheet.nrows
col_number=sheet.ncols
row_number_actual=row_number
```

```
LATLON = zeros((row_number_actual, 3), dtype=float) # produce array with station number (0) and
coordinates (1,2)

for i in range(0, row_number_actual, 1):
    LATLON[i, 0] = int(sheet.cell_value(i, 1))
    LATLON[i, 1] = sheet.cell_value(i, 2)
    LATLON[i, 2] = sheet.cell_value(i, 3)

country_row = []
state_row = []

# find location using coordinates (Reverse Geocoding) - Nominatim geocoder

for i in tqdm(range(0, row_number_actual, 1)):
    geolocrev = Nominatim(user_agent="nick"+str(i), timeout=900)
    coordinates = geolocrev.reverse(str(LATLON[i, 1])+','+str(LATLON[i, 2]))
    # print(coordinates.address)
    raw_ID = coordinates.raw
    try:
        country_name = raw_ID['address']['country']
        # state_name = raw_ID['address']['state']
        country_row.append(country_name)
        pass
    except KeyError:
        country_row.append('-')


# .xlsx file generation with countries found

country_tot = array([LATLON[:, 0], country_row])
pandas.DataFrame(country_tot).to_excel('/users/nick_agatheris/desktop/Simulation Results/Python
Output/Country Output ('+str(name)+').xlsx', header=False, index=False)

# total runtime of script

end_time = time.time()
total_time = int(end_time-start_time)
print(str(total_time)+'s for '+str(row_number_actual)+' locations')
```