
**NATIONAL TECHNICAL UNIVERSITY OF ATHENS**

SCHOOL OF CIVIL ENGINEERING

# MAXIMUM ENTROPY PROBABILITY DISTRIBUTIONS AND STATISTICAL-STOCHASTIC MODELLING OF RAINFALL

**Simon Michael Papalexiou**

**Athens, June 2013**

**NATIONAL TECHNICAL UNIVERSITY OF ATHENS**
**SCHOOL OF CIVIL ENGINEERING**
**DEPARTMENT OF WATER RESOURCES AND ENVIRONMENTAL ENGINEERING**

# MAXIMUM ENTROPY PROBABILITY DISTRIBUTIONS AND STATISTICAL-STOCHASTIC MODELLING OF RAINFALL

**Ph.D. Thesis**

**Simon Michael Papalexiou**

**Athens, June 2013**

**Thesis committee**

**Thesis supervisor**

D. Koutsoyiannis, Prof. N.T.U.A

**Advisory committee**

1. D. Koutsoyiannis, Prof. N.T.U.A (Supervisor)
2. M. Mimikou, Prof. N.T.U.A
3. N. Mamasis, Asst. Prof. N.T.U.A

**Evaluation committee**

1. D. Koutsoyiannis, Prof. N.T.U.A (Supervisor)
2. M. Mimikou, Prof. N.T.U.A
3. N. Mamasis, Ass. Prof. NTUA
4. A. Loukas, Prof. U.Th
5. M. Vafeiadis, Prof. A.U.Th
6. H. Pavlopoulos, Assoc. Prof. A.U.E.B
7. I. Nalbantis, Asst. Prof. N.T.U.A

This research was partially funded by the National Technical University of Athens.

*Κάποτε, με χρόνο ή χωρίς*
*Κάπου, με χώρο ή χωρίς*
*Κάπως, με μορφή ή χωρίς*
*Θα ξανανταμώσουμε*
*Και θα ξέρεις πως είμαι εγώ*
*Και θα ξέρω πως είσαι εσύ*

*Για σένα πατέρα…*

*... Και τώρα θα προσθέσω:*
*Όσοι από σας πια βαρεθήκατε στον κόσμο αυτόν τον άδικον και τον βλακώδη να άγεσθε και να φέρεσθε από τους ψεύτες, από τους σοφιστάς και λαοπλάνους, όσοι πια βαρεθήκατε οι δεσμοφύλακές σας σαν τόπια ταλαίπωρα να σας εξαποστέλλουν εις τον Καϊάφα και πριν απ αυτόν στον ’Αννα, προσμένοντας να έλθη η Ώρα η χρυσαυγής, η πολυύμνητος και ευλογημένη, όσοι πιστοί, όσοι ζεστοί, όσοι την σημερινήν ελεεινήν πραγματικότητα να αλλάξετε ποθείτε, προσμένοντας να έλθη η Ώρα, όσοι πιστοί, όσοι ζεστοί, ελάτε και ως ανακράξωμεν μαζί (νυν και αεί, νυν και αεί) σαν προσευχή και σαν παιάνα, ας ανακράξωμεν μαζί, με μια ψυχή, με μια φωνή ΟΚΤΑΝΑ!*

*ΑΝΔΡΕΑΣ ΕΜΠΕΙΡΙΚΟΣ*

# Πρελουδιο

Δε νιώθεις τίποτα. Αλλιώς το ’χες φανταστεί, κάπως αλλιώς. Κλείνει στροφή η σπείρα με μια απροσδιόριστη στοναχή, κάτι σαν γεύση πικρή μιας νύχτας βαριάς και άχαρης που με ανακούφιση τελειώνει. Ίσως νιώθεις πως έχεις ανάγκη κάτι να πεις, πως πρέπει κάποιες λέξεις να στραβώσεις, να πεις κάτι σαν μουσική, κάτι σαν ρέκβιεμ για τα χρόνια. Αλλά τι να πεις, πώς να το πεις; Σε δυο γραμμές μονάχα θα ψελλίσεις.

Ίσως να φταιν τα κόκαλα που φέρνουν πάντα σκύλο. Ίσως και συ. Δεν έχεις όμως άλλη οφειλή, πρέπει σαν δέντρο σαν σταθείς, ατάραχος και δίκαιος όσο μπορείς. Δεν ξέρεις αν μπορείς, μυρίζεις ακόμη πυρετό, ίλιγγο και φλόγα. Φοβάσαι αυτή τη ζυγαριά, που θα τη γείρουν τόσα χρόνια; Τι έχει απομείνει εδώ, τι έχει νόημα χωρίς το βλέμμα του Οδυσσέα; Μένεις ενεός μπρος τα χαμένα χρόνια. Το κεφάλι σου πονά. Ποιος θα σου πει τι σε όρισε και τι ορίζεις; Τι είναι αυτό που ρήμαξε και τι είναι αυτό που ανθίζει; Επικαλέστηκες τα τέσσερα στοιχειά, πνεύμα, καρδιά, ψυχή και σώμα και είπες στη μνήμη να σε σώσει. Κάτι δεν πάει καλά, η άλγεβρά σου είναι γεμάτη καταπέλτες.

Δεν αντέχεις, πνίγεσαι, τι είναι όλα αυτά τριγύρω; Που είναι ο κύκλος των χαμένων ποιητών; Ένας σωρός μοιρογνωμόνια που πρέπει να περάσεις. Μα είσαι στραβός, γωνίες γεμάτος, χτίσου παράλληλος αν θέλεις να περάσεις. Γίνε βορά γλυκιά στους τιμαριώτες των ιδεών. Να συμφωνείς, πώς αλλιώς θα βαφτιστούν όλοι οι δημοκράτες; Έχει κι άλλα λεν για τα Πεδία η συνταγή, λίγα ακόμη και θα ’ναι ανθοφόρα και η οδός. Αλλά μη λες πολλά. Κράτα το στόμα σου κλειστό κι άσε το καλοκαίρι να σύρει αυτό το ερπετό έξω απ’ τη φωλιά του. Κράτα το στόμα σου κλειστό, όταν με τη φρικαλέα ειρωνία των περιδεών θα σου χαμογελάνε. Ίδε ο άνθρωπος «που τον πατάν στ’ αλήθεια τα πόδια του τα ίδια». Τίποτα απ’ όλα τούτα δε σ’ αρέσει. Δεν αντέχεις, πνίγεσαι. Κάπως αλλιώς, κάπως αλλιώς το ’χες φανταστεί. «Αυτοί που δραπετεύουν, έμαθαν τι γυρεύουν». Γίνε αυτό που είσαι αν μπορείς τα δαιμόνια σου ολολύζουν, μα έχε κατά νου δε σπάει η πέτρα με τα δόντια.

Αρκετά. Το βέλος του χρόνου έσπασε και δε δείχνει πουθενά. Κλείνεις τα μάτια. Όλα ξεμακραίνουν, όλα είναι ξένα και όλα ομορφαίνουν. Είναι κάτι σαν πάλη με φωτόνια στο σκοτάδι, ένας πυρρίχιος χορός, κάπως έτσι είναι αυτό που κάνεις. Εδώ σ’ αντέχει ο σκελετός, εδώ ξεχνάς. Προσωπικός μυστικισμός, μοναχικό, πυρηνικό και ανέσπερο δρομάκι. Μόνο εδώ ο κόπος σου αξίζει. Εδώ ξέρεις τι σε μετρά. Μόνο η ιδέα στ’ αλήθεια σε μετρά και υπηρέτησες σωστά με ωραίο πάθος καθαρό. Άλλα δε θέλεις να θυμάσαι.

Τι ήθελα με αυτά να πω, ίσως να αναρωτιέσαι; Πώς να στο πω, δεν έχει σημασία. Είναι πράγματα που αν δεν τα «νιώθεις» χωρίς να στα εξηγήσω δε θα τα «νιώσεις» ούτε αν στα εξηγήσω. Να πω όμως για ανθρώπους, που χωρίς αυτούς μάλλον δε θα ’γραφα «εδώ».

Μεγάλωσα μες στην πελώρια καρδιά του. Με βλέμμα καθαρό έδινε το ρυθμό σε όλα τα βήματά μου. Και τώρα, πάντα απών και πάντα παρών, με το αόρατο γιγάντιό του χέρι με σηκώνει όταν λυγίζω. Δημήτρη Παπαλεξίου ’λεγαν τον πατέρα μου. Ήταν σπουδαίος ιατρός, μέγιστος άνθρωπος, ευθύς, απλός, αυθεντικός, βαθιά σοφός, ανενδεής και πάντοτε «ωραίος». Θα προσπαθώ να γίνω αυτός. Με δυο τρεις λέξεις αν με ορίσω, τότε είμαι ο γιός του και τίποτα άλλο. Αλλιώς πώς να το πω, είναι για μένα η αρχή των πάντων.

Είναι η θεά των μικρών πραγμάτων, βρίσκει το νόημα στο πιο μικρό, στο πιο απλό και αυτό είναι το πιο «μεγάλο». Μαζί της μαθαίνω τι είναι το ουσιαστικό και το αληθινά σπουδαίο. Με κρατά στο φώς και κάθε φορά που την κοιτώ νιώθω πως κάποιο νόημα έχει αυτός ο θαυμαστός παραλογισμός που είπαμε ζωή. Κάνει κάτι μαγικό, με κάθε χαμόγελό της γίνομαι καλύτερος άνθρωπος. Είναι η γυναίκα μου, η Ευτυχία μου. Περάσαμε μέσα απ’ τις φωτιές, όλα μαζί, πάντα μαζί.

Μικρή σημασία έχει αν διαφωνείς ή συμφωνείς μαζί του, πάντα κάτι θα πάρεις. Είχαμε σπουδαίες αλλά και ανάποδες στιγμές, όλες εξίσου σημαντικές. Πιστεύω, αν είμαι άξιος να κρίνω, πως είναι πολύ σημαντικός επιστήμονας, έχει προσφέρει πολλά και θα προσφέρει άλλα τόσα. Δεν είναι άλλος φυσικά από τον Δημήτρη Κουτσογιάννη.

Με χαρά θα αναφερθώ και στους λοιπούς που έπρεπε τις έρευνές μου να «εγκρίνουν». Στο Νίκο Μαμάση, ωραίος τύπος, αλλιώτικος, άλλα δε θα πω. Στη Μαρία Μιμίκου που με χαρά βοήθησε όποτε χρειάστηκα κάτι. Και βεβαίως στους ανθρώπους της εξεταστικής επιτροπής και για τα ωραία και για τα στραβά που βρήκαν αλλά προπαντός γιατί είναι όλοι τους άνθρωποι εξαιρετικοί που μες των αναβρασμό κείνων των ημερών ’καναν ότι μπορούσαν για να βοηθήσουν.

Η αγάπη είναι τυφλή γι’ αυτό η φιλία έχει τα μάτια της πάντα κλειστά, κάτι τέτοιο είπε ένας μεγάλος. Πώς θα μαλάκωναν χωρίς αυτούς άγριες μέρες; Μοιραστήκαμε ίδιο χώρο και ίδιες «εκπνοές». Να ’ναι και το μέλλον, να ’ναι κι η Οκτάνα. Είναι οι φίλοι μου ο Ανδρέας Ευστρατιάδης, ο Παναγιώτης Δημητριάδης, ο Γιάννης Μαρκόνης και ο Παναγιώτης Κοσσιέρης. Μνεία και σ’ όλα τα άλλα άξια παιδιά του «ορόφου». Σπάνια συγκέντρωση τόσο ωραίων ανθρώπων. Το ευχαριστώ δεν είναι για μένα λέξη μονάχα.

Αθήνα, αργά μια νύχτα,
Σίμων Μιχαήλ Παπαλεξίου

# ΕΚΤΕΝΗΣ ΠΕΡΙΛΗΨΗ

Στην παρούσα Διατριβή εξετάζονται τρία κυρίως θέματα: (α) η δυνατότητα να χρησιμοποιηθεί μια θεωρητική αρχή, συγκεκριμένα η αρχή της μέγιστης εντροπίας, ως βάση για τη διαμόρφωση και την επιλογή πιθανοτικών κατανομών κατάλληλων για τη βροχόπτωση και εν δυνάμει και για άλλες γεωφυσικές μεταβλητές, (β) η πιθανοτική-στατιστική ανάλυση σε παγκόσμια κλίμακα της ημερήσιας βροχόπτωσης καθώς και της ακραίας ημερήσιας βροχόπτωσης και (γ) η στοχαστική δομή της βροχόπτωσης σε πολύ μικρή χρονική κλίμακα (10 s). Βασικός στόχος της έρευνας είναι να διατυπώσει απλά αλλά θεμελιώδη και ευρέος ενδιαφέροντος ερωτήματα σχετικά με τη στατιστική-στοχαστική φύση της βροχόπτωσης και να δώσει απαντήσεις όχι μόνο θεωρητικής αλλά κυρίως πρακτικής αξίας.

## Σχετικά με την αρχή της μέγιστης εντροπίας

Η έμφαση δίνεται στη διαμόρφωση και στη λογική και θεωρητική τεκμηρίωση απλών περιορισμών που σε συνδυασμό με τον κλασικό ορισμό της εντροπίας, δηλαδή της εντροπίας Boltzmann-Gibbs-Shannon (BGS) (Εξ. (1)), θα οδηγούν σε ευέλικτες και απλές κατανομές κατάλληλες για την πιθανοτική περιγραφή της βροχόπτωσης αλλά και άλλων γεωφυσικών μεταβλητών.

$$S_X = -\int_0^\infty f_X(x) \ln f_X(x) \mathrm{d}x \tag{1}$$

Συνοπτικά η αρχή της μέγιστης εντροπίας [*E. T. Jaynes*, 1957a, 1957c] είναι ένα εργαλείο για την εξαγωγή συμπερασμάτων υπό συνθήκες αβεβαιότητας ή ελλιπούς γνώσης και στοχεύει στην εξεύρεση της πλέον κατάλληλης κατανομής πιθανοτήτων σύμφωνα με την διαθέσιμη πληροφορία, η οποία εκφράζεται ως ένα σύνολο περιορισμών που σχηματίζονται ως αναμενόμενες τιμές συναρτήσεων $g_j(\ )$ της τυχαίας μεταβλητής $X$, ήτοι,

$$E\left(g_j(X)\right) = \int_0^\infty g_j(x) f_X(x) \mathrm{d}x = c_j, \qquad j = 1,...,n \tag{2}$$

Η κατανομή μέγιστης εντροπίας προκύπτει από τη μεγιστοποίηση της εντροπίας (Εξ. (1)) θέτοντας περιορισμούς συμφώνα με την Εξ. (2) και πραγματοποιείται με τη μέθοδο των

πολλαπλασιαστών Lagrange. Η γενική λύση που προκύπτει για αυθαίρετους περιορισμούς είναι

$$f_X(x) = \exp\left(-\lambda_0 - \sum_{j=1}^{n} \lambda_j g_j(x)\right) \tag{3}$$

όπου $f_X(x)$ η πυκνότητα πιθανότητας, $\lambda_j$, με $j$ =1,…, $n$, οι πολλαπλασιαστές Lagrange που συνδέονται με τους περιορισμούς της Εξ. (2). Ο πολλαπλασιαστής Lagrange $\lambda_0$ προκύπτει από τον περιορισμό $\int_0^\infty f_X(x)\mathrm{d}x = 1$.

- ***Γιατί και πως μπορεί η αρχή της μέγιστης εντροπίας να συμβάλει στο σχηματισμό ή στην επιλογή κατάλληλων πιθανοτικών κατανομών για μια τυχαία μεταβλητή;***

Οι γνωστές πιθανοτικές κατανομές είναι μερικές δεκάδες, ενώ από μαθηματικής απόψεως ο συνολικός αριθμός των κατανομών είναι άπειρος καθώς άπειρος αριθμός συναρτήσεων μπορεί να οριστεί με τις ιδιότητες μιας πιθανοτικής κατανομής. Η κοινή τεχνική για την επιλογή μιας κατανομής βασίζεται συνήθως σε μεθόδους δοκιμής-σφάλματος, δηλαδή, προσαρμόζεται συνήθως ένας μικρός αριθμός κατανομών στα εμπειρικά δεδομένα και επιλέγεται η κατανομή με την καλύτερη προσαρμογή που προκύπτει σύμφωνα με κάποιο κριτήριο σφάλματος ή τα αποτελέσματα στατιστικών ελέγχων. Θεωρητικά, η διαδικασία αυτή δεν έχει τέλος, εφόσον άπειρες κατανομές μπορούν κατασκευαστούν και συνεπώς να δοκιμαστούν ως προς την καταλληλότητά τους. Αντίθετα η αρχή της μέγιστης εντροπίας προσφέρει ένα ισχυρό θεωρητικό υπόβαθρο για να προσδιοριστεί ένα πιθανοτικό μοντέλο βάσει της διαθέσιμης πληροφορία. Ωστόσο, η επιτυχής χρήση αυτής της αρχής προϋποθέτει την ενσωμάτωση όλης της διαθέσιμης πληροφορίας με τη μορφή μαθηματικών περιορισμών.

- ***Ποια πρέπει να είναι η μορφή αυτών των περιορισμών για γεωφυσικές μεταβλητές όπως η βροχή;***

Η βασική παραδοχή σχετικά με τη μορφή των περιορισμών είναι ότι οι περιορισμοί πρέπει να είναι όσο το δυνατόν λιγότεροι και απλοί καθώς και να ενσωματώνουν την όποια εκ των προτέρων διαθέσιμη πληροφορία. Αυτή η πληροφορία, για παράδειγμα, μπορεί να αφορά τις γενικές ιδιότητες του σχήματος της συνάρτησης πυκνότητας πιθανότητας της υπό μελέτη μεταβλητής και θα μπορούσε να έχει προκύψει από εμπειρικές αναλύσεις. Οι τρεις περιορισμοί που μελετήθηκαν και τεκμηριώθηκαν βάσει λογικών και μαθηματικών επιχειρημάτων σχετίζονται με τη λογαριθμική συνάρτηση και τη συνάρτηση δύναμης, οι οποίες, όπως προκύπτει, είναι κατάλληλες για θετικά ορισμένες, έντονης μεταβλητότητας

και ασύμμετρες τυχαίες μεταβλητές, χαρακτηριστικά τα οποία εντοπίζονται συνήθως σε γεωφυσικές διεργασίες, π.χ. όπως οι βροχοπτώσεις και οι απορροές των ποταμών. Συγκεκριμένα, οι περιορισμοί είναι οι αναμενόμενες τιμές των παρακάτω συναρτήσεων: (α) $\ln x$, (β) $x^q$ και (γ) $\ln(1+px^q)/p$. Ο τελευταίος περιορισμός, οι $p$-ροπές, αποτελούν μια γενίκευση των κλασσικών ροπών καθώς για $p = 0$ ισοδυναμούν με τις κλασικές ροπές $x^q$ αφού $x_0^q = \lim_{p\to 0} \ln(1+px^q)/p = x^q$.

- ***Τι κατανομές προκύπτουν με τη χρήση αυτών των περιορισμών;***

Η μεγιστοποίηση της εντροπίας BGS συνδυάζοντας τους περιορισμούς (α)-(β) και (α)-(γ) οδηγεί σε δύο ευέλικτες κατανομές, συγκεκριμένα, μια τριπαραμετρική εκθετικού τύπου (Εξ. (4)), γνωστή ως Generalized Gamma (GG) [*Stacy*, 1962] και μια τετραπαραμετρική τύπου δύναμης (Εξ. (5)), γνωστή ως Generalized Beta of the Second Kind (GB2) [*Mielke Jr and Johnson*, 1974] με την πρώτη κατανομή να είναι μια ειδική (οριακή) περίπτωση της δεύτερης.

$$f_X(x) = \frac{\gamma_2}{\beta\Gamma(\gamma_1/\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma_2}\right), \qquad x \geq 0 \tag{4}$$

$$f_X(x) = \frac{\gamma_3}{\beta\mathrm{B}(\gamma_1,\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1\gamma_3-1}\left(1+\left(\frac{x}{\beta}\right)^{\gamma_3}\right)^{-(\gamma_1+\gamma_2)}, \qquad x \geq 0 \tag{5}$$

Για πρακτικούς σκοπούς πάντως προτείνεται η χρήση μιας τριπαραμετρικής κατανομής (Εξ.(6)), γνωστής ως Burr τύπου XII (BrXII) [*Burr*, 1942], η οποία προκύπτει εύκολα ως απλοποίηση της κατανομής GB2 για $\gamma_1 = 1$.

$$f_X(x) = \frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\left(1+\gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right)^{-\frac{1}{\gamma_1\gamma_2}-1}, \qquad x \geq 0 \tag{6}$$

Τόσο η GG όσο και η BrXII είναι πολύ ευέλικτες κατανομές διότι εκτός από μια παράμετρο κλίμακας, περιλαμβάνουν και δύο παραμέτρους σχήματος που ελέγχουν τόσο τη δεξιά όσο και την αριστερή ουρά της κατανομής.

- ***Είναι απαραίτητες οι γενικεύσεις της εντροπίας για την διαμόρφωση κατανομών με «χοντρές» ουρές, όπως για παράδειγμα οι κατανομές τύπου δύναμης;***

Η μεγιστοποίηση της εντροπίας BGS «παραδοσιακά» πραγματοποιείται με την χρήση περιορισμών που οδηγούν σε κατανομές με εκθετικές ή υπερεκθετικές ουρές, όπως η εκθετική ή η κανονική κατανομή. Η εμπειρική ανάλυση όμως διαφόρων φαινομένων υποδεικνύει ότι αυτές οι κατανομές σε πολλές περιπτώσεις είναι ανεπαρκείς να περιγράψουν την πραγματικότητα, καθώς απαιτούνται κατανομές με υποεκθετικές ουρές, π.χ. ουρές τύπου δύναμής, για να εκφράσουν ορθά τα ακραία γεγονότα. Αυτό οδήγησε στην εισαγωγή γενικευμένων μέτρων εντροπίας τα οποία όμως έχουν δεχτεί κριτική αναφορικά με την εγκυρότητά τους, σε σύγκριση με την κλασική και ισχυρά θεμελιωμένη εντροπία BGS. Με την «επιστράτευση» όμως των προαναφερθέντων περιορισμών η χρήση των γενικευμένων μέτρων εντροπίας δεν είναι απαραίτητη, καθώς οι συγκεκριμένοι περιορισμοί, ιδίως οι $p$-ροπές, οδηγούν αβίαστα, σε συνδυασμό με την κλασική BGS εντροπία, σε κατανομές τύπου δύναμης.

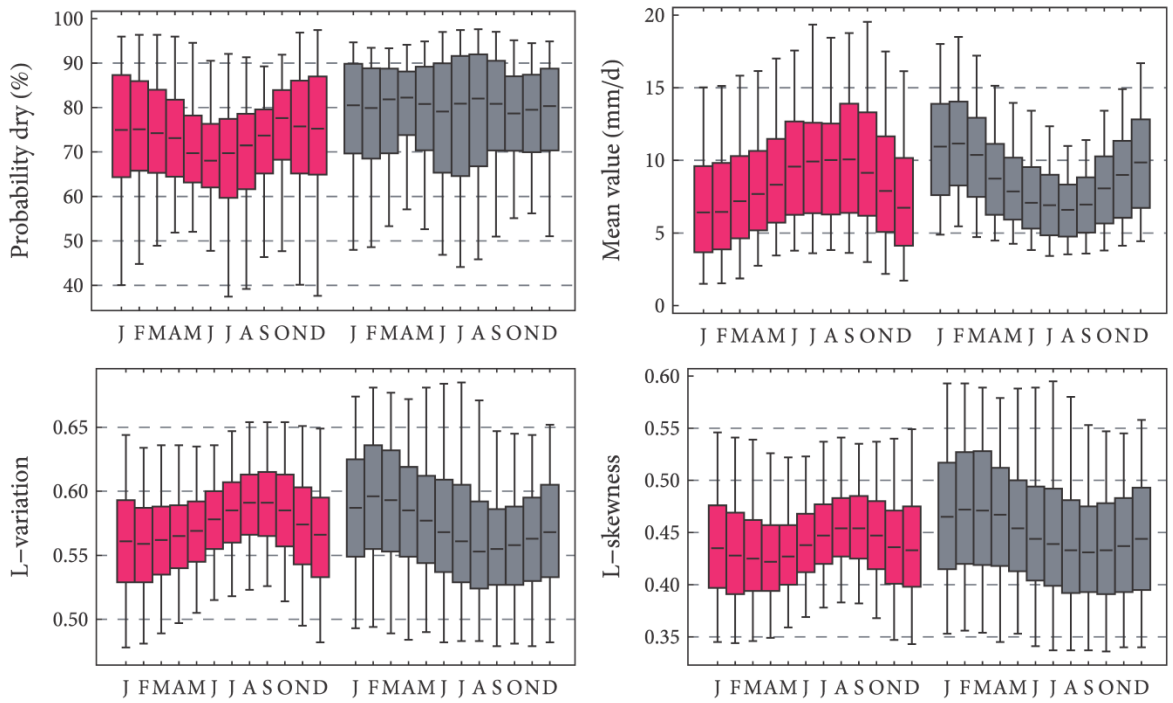
## Σχετικά με την περιθώρια κατανομή της ημερήσιας βροχόπτωσης

Εκπονήθηκε μια μαζική εμπειρική ανάλυση περισσότερων από 170 000 μηνιαίων χρονοσειρών βροχόπτωσης σε περισσότερους από 14 000 σταθμούς σε όλο τον κόσμο με στόχο να απαντηθούν δύο βασικά ερωτήματα: (α) ποια στατιστικά χαρακτηριστικά της ημερήσιας βροχόπτωσης παρουσιάζουν τη μεγαλύτερη εποχιακή διακύμανση, και (β) κατά πόσον υπάρχει ή όχι ένα σχετικά απλό πιθανοτικό μοντέλο ικανό να περιγράψει τη θετική ημερήσια βροχόπτωση για κάθε μήνα και σε κάθε περιοχή του κόσμου.

- ***Ποια χαρακτηριστικά της περιθώριας κατανομής της ημερήσιας βροχόπτωσης παρουσιάζουν εποχιακή μεταβλητότητα;***

Η μηνιαία εμπειρική ανάλυση, ανά ημισφαίριο αρχικά (Σχήμα 1), της πιθανότητας ξηρασίας, της μέσης τιμής και δύο στατιστικών μέτρων του σχήματος της κατανομής της μη μηδενικής βροχόπτωσης, δηλαδή, της L-μεταβλητότητας και της L-ασυμμετρίας, αποκαλύπτει σε γενικές γραμμές ημιτονοειδή μοτίβα για όλα τα στατιστικά μέτρα που αναλύθηκαν υποδεικνύοντας συνεπώς εποχιακή διακύμανση αυτών των χαρακτηριστικών. Επιπλέον, για την ακριβέστερη ανάλυση κατασκευάστηκε μια στατιστική δοκιμή που ελέγχει την εποχιακή διακύμανση (SV-Test) και τα αποτελέσματα της εφαρμογής του δείχνουν μια σαφή μηνιαία διακύμανση της πιθανότητας ξηρασίας και της μέσης τιμής της μη μηδενικής ημερήσιας βροχόπτωσης σε 95.1% και 91.7%, αντίστοιχα, των σταθμών που αναλύθηκαν, ενώ τα αντίστοιχα ποσοστά για τα δυο χαρακτηριστικά σχήματος, δηλαδή,

της L-μεταβλητότητας και L-ασυμμετρίας, είναι 66.1% και 54.2%, αντίστοιχα. Αυτά τα αποτελέσματα, αν συνδυαστούν με τη γενική εικόνα που προκύπτει από την ανάλυση των σταθμών ανά ημισφαίριο δείχνουν ότι όχι μόνο η πιθανότητα ξηρασίας και η μέση τιμή της μη μηδενικής βροχόπτωσης παρουσιάζουν εποχιακή διακύμανση αλλά επίσης και το σχήμα της κατανομής.




**Σχήμα 1.** Εκτιμήσεις στατιστικών χαρακτηριστικών της μηνιαίας ημερήσιας βροχόπτωσης των σταθμών που αναλύθηκαν (κόκκινα και γκρι θηκογράμματα για το Βόρειο και Νότιο ημισφαίριο αντίστοιχα).

- ***Ποια χαρακτηριστικά παρουσιάσουν την εντονότερη εποχιακή διακύμανση;***

Η μηνιαία διακύμανση αυτών των στατιστικών στοιχείων σε κάθε σταθμό που αναλύθηκε ποσοτικοποιήθηκε με διάφορα μέτρα απόκλισης σε σχέση με το μέσο όρο όλων των μηνών. Η ανάλυση έδειξε ότι η υψηλότερη μηνιαία διακύμανση παρατηρείται στη μέση τιμή της μη μηδενικής βροχόπτωσης ενώ έπονται κατά σειρά η πιθανότητα ξηρασίας, η L-ασυμμετρία και τέλος, η L-μεταβλητότητα, υποδεικνύοντας ότι η εποχιακή διακύμανση των χαρακτηριστικών σχήματος, αν και υπαρκτή, δεν είναι πολύ υψηλή.

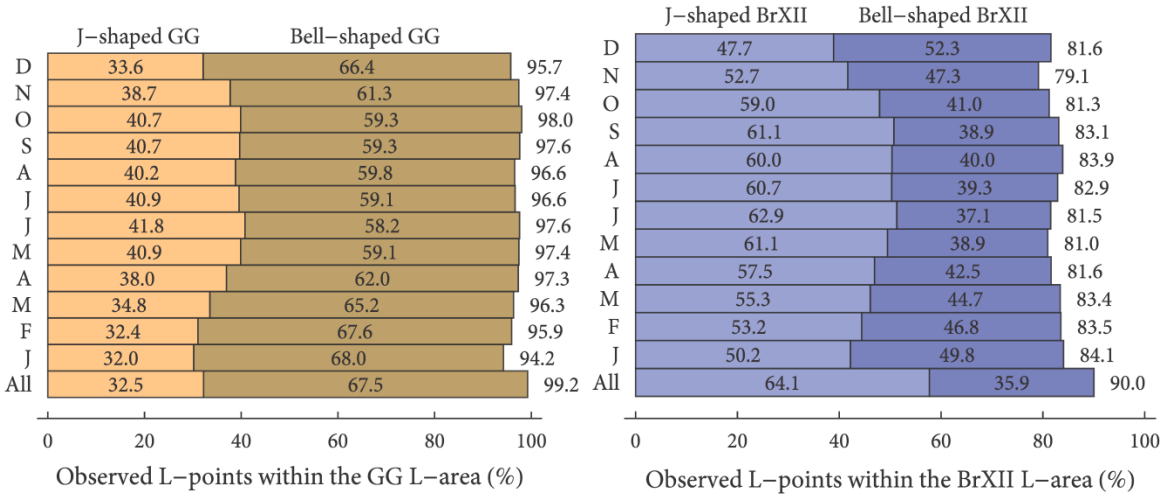
- ***Ποια είναι τα γενικά χαρακτηριστικά του σχήματος της κατανομής της βροχόπτωσης;***

Η μεταβλητότητα των στατιστικών μέτρων που μελετήθηκαν, καθώς και οι τιμές των παραμέτρων των κατανομών που προσαρμόστηκαν στα δεδομένα, δείχνουν ότι η συνάρτηση πυκνότητας της μη μηδενικής βροχόπτωσης μπορεί να διαφέρει σημαντικά από

σταθμό σε σταθμό. Η διαφοροποίηση αυτή δεν εντοπίζεται μόνο στη γενική μορφή του σχήματος κατανομής, δηλαδή αν είναι σχήματος J (J-shaped) ή κωδωνοειδούς μορφής (bell-shaped) (τα ποσοστά παρουσιάζονται στο Σχήμα 2), αλλά επίσης και στη συμπεριφορά της ουράς της που συνεπάγεται διαφορετική συμπεριφορά στα ακραία γεγονότα.

- ***Μπορούν τα δημοφιλή διπαραμετρικά μοντέλα να περιγράψουν επαρκώς την ημερήσια βροχόπτωση;***

Η εποχιακή και η χωρική μεταβλητότητα που παρατηρήθηκε στα χαρακτηριστικά του σχήματος υποδεικνύουν πως τα δημοφιλή διπαραμετρικά μοντέλα όπως η Gamma, η Weibull, η Lognormal, η Pareto, κ.λπ., δεν μπορούν να χρησιμεύσουν ως «καθολικά» μοντέλα για τη μοντελοποίηση της ημερήσιας βροχόπτωσης καθώς η ευελιξία τους είναι περιορισμένη και ως εκ τούτου δεν μπορούν να περιγράψουν επαρκώς το κύριο σώμα της κατανομής και συγχρόνως και την αριστερή και τη δεξιά ουρά της.




**Σχήμα 2.** Ποσοστό εμπειρικών σημείων L-ροπών (L-ασυμμετρία συναρτήσει L-μεταβλητότητας) που ανήκουν μέσα στο θεωρητικό χώρο που σχηματίζουν οι κατανομές.

- ***Υπάρχει ένα "καθολικό" μοντέλο ικανό να περιγράψει την ημερήσια βροχόπτωση σε όλες τις εποχές και σε όλες τις περιοχές του κόσμου;***

Ένα "καθολικό" πιθανοτικό μοντέλο για την ημερήσια βροχόπτωση πρέπει να έχει τουλάχιστον δύο παραμέτρους σχήματος, όπου η μία θα ελέγχει την αριστερή ουρά και η άλλη την δεξιά. Δύο κατανομές με τα ανωτέρω χαρακτηριστικά που προέκυψαν από την εφαρμογή της αρχής της μέγιστης εντροπίας είναι η BrXII και η GG. Η επίδοση αμφότερων των κατανομών είναι πολύ καλή με την GG να αποδίδει ακόμη καλύτερα από την BrXII προσφέροντας έτσι μια εξαιρετική επιλογή. Κάποια από τα χαρακτηριστικά αυτών των δύο

κατανομών αλληλοσυμπληρώνονται, έτσι η GB2 κατανομή, η οποία περιλαμβάνει και τις δύο ως ειδικές περιπτώσεις, μπορεί να χρησιμοποιηθεί για να μοντελοποιήσει ολόκληρο το σύνολο δεδομένων για όλους τους μήνες και όλους τους σταθμούς. Το Σχήμα 2 παρουσιάζει το ποσοστό των εμπειρικών σημείων L-ροπών (L-ασυμμετρία συναρτήσει L-μεταβλητότητας) των χρονοσειρών που αναλύθηκαν σε μηνιαία βάση που ανήκουν μέσα στο θεωρητικό χώρο που σχηματίζουν οι κατανομές. Αν ένα σημείο ανήκει μέσα στο θεωρητικό χώρο της κατανομής σημαίνει πως η κατανομή μπορεί να προσαρμοστεί διατηρώντας τις πρώτες τρεις L-ροπές.

- ***Τι υποδεικνύουν οι τιμές των παραμέτρων της κατανομής με την καλύτερη προσαρμογή στα δεδομένα;***

Η παράμετρος σχήματος $\gamma_2$ GG κατανομής, η οποία ελέγχει τη δεξιά ουρά και συνεπώς τις ακραίες τιμές, για τη συντριπτική πλειονότητα των δειγμάτων που αναλύθηκαν ισχύει $\gamma_2 < 1$, τιμή που αντιστοιχεί σε υποεκθετικές ουρές, ενώ για $\gamma_2 = 1$ η GG απλοποιείται στην κατανομή Gamma. Αυτό συνεπάγεται ότι μερικά από τα ευρέως χρησιμοποιούμενα μοντέλα με εκθετική ουρά όπως η Εκθετική, η Gamma ή μικτά μοντέλα με εκθετικές ουρές εν δυνάμει αποτελούν επικίνδυνη επιλογή και δεν πρέπει να χρησιμοποιούνται αδικαιολόγητα στην πράξη, δεδομένου ότι μπορούν να υποτιμήσουν σοβαρά το μέγεθος και τη συχνότητα των ακραίων βροχοπτώσεων σε ημερήσια κλίμακα.

## Σχετικά με την ουρά της ημερήσιας βροχόπτωσης

Εξετάζεται η δεξιά ουρά της κατανομής της ημερήσιας βροχόπτωσης, δηλαδή, το μέρος της κατανομής που περιγράφει τα ακραία γεγονότα. Αναλύθηκαν ακραίες βροχοπτώσεις σε περισσότερους από 15 000 σταθμούς σε όλο τον κόσμο και συγκρίθηκε η απόδοση τεσσάρων κοινών και μονοπαραμετρικών πιθανοτικών μοντέλων ουράς που αντιστοιχούν στις κατανομές Pareto type II, Weibull, Lognormal και Gamma με συναρτήσεις υπέρβασης πιθανότητας (exceedance probability function) που δίνονται, αντίστοιχα, από τις

$$\overline{F}_{\mathrm{PII}}(x) = \left(1 + \gamma \frac{x}{\beta}\right)^{-\frac{1}{\gamma}} \tag{7}$$

$$\overline{F}_{\mathrm{LN}}(x) = \frac{1}{2}\mathrm{erfc}\left(\ln\left(\frac{x}{\beta}\right)^{1/\gamma}\right) \tag{8}$$

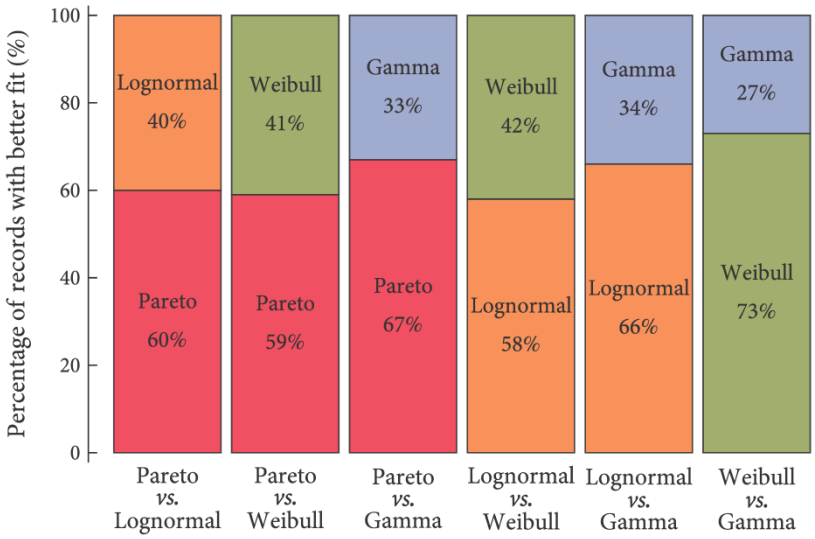
$$\overline{F}_{\mathrm{W}}(x) = \exp\left(-\left(\frac{x}{\beta}\right)^{\gamma}\right) \tag{9}$$

$$\overline{F}_{\mathrm{G}}(x) = \Gamma\left(\gamma, \frac{x}{\beta}\right) / \Gamma(\gamma) \tag{10}$$

Σκοπός ήταν να αποκαλυφθεί ποιος τύπος ουράς περιγράφει καλύτερα τη συμπεριφορά των ακραίων γεγονότων. Η μέθοδος προσαρμογής ήταν άμεση, δηλαδή, προσαρμογή (με ελαχιστοποίηση μιας τροποποιημένης νόρμας ελαχίστων τετραγώνων) των τεσσάρων ουρών στην εμπειρική ουρά κάθε δείγματος η οποία ορίστηκε για ένα δείγμα $N$ ετών ως οι $N$ μεγαλύτερες τιμές του δείγματος.

- ***Ποίος τύπος ουράς κατανομής περιγράφει καλύτερα την ακραία ημερήσια βροχόπτωση άνω κατωφλίου (above threshold);***

Η ανάλυση δείχνει πως οι πιο «χοντρές» ουρές, ή αλλιώς οι κατανομές με υποεκθετικές ουρές έχουν καλύτερες επιδόσεις σε σχέση με τις «λεπτές» ουρές. Συγκεκριμένα, στο 72.6% των σταθμών που μελετήθηκαν, οι υποεκθετικού τύπου ουρές προσαρμόστηκαν καλύτερα, ενώ οι εκθετικές-υπερεκθετικές ουρές είχαν καλύτερη προσαρμογή μόνο στο 27.4% των σταθμών. Η κατάταξη από την καλύτερη προς τη χειρότερη επίδοση σχετικά με την προσαρμογή των ουρών είναι: (α) η Pareto, (β) η Λογαριθμοκανονική (Lognormal), (γ) η Weibull, και (δ) η Γάμα (Gamma). Στο Σχήμα 3 παρουσιάζεται μια σύγκριση των αποτελεσμάτων προσαρμογής των ουρών ανά ζεύγη. Όπως προκύπτει μεταξύ των δύο κατανομών που συγκρίνονται η κατανομή με την πιο «χοντρή» ουρά αποδίδει καλύτερα.




**Σχήμα 3.** Σύγκριση των προσαρμοσμένων ουρών σε ζευγάρια βάσει του τετραγωνικού σφάλματος.

- ***Μπορούν τα πιο κοινά μοντέλα να περιγράψουν αξιόπιστα την ακραία βροχόπτωση;***

Η ανάλυση αποκάλυψε ότι το πιο δημοφιλές μοντέλο που χρησιμοποιείται στην πράξη, η κατανομή Γάμα, είχε τη χειρότερη επίδοση, πράγμα που σημαίνει ότι η κατανομή αυτή υποεκτιμά τόσο τη συχνότητα όσο και το μέγεθος των ακραίων φαινομένων. Αυτό οδηγεί στο συμπέρασμα ότι οι υποεκθετικού τύπου κατανομές είναι προτιμότερες για τη μοντελοποίηση των ακραίων γεγονότων βροχόπτωσης.

- ***Ποίες είναι οι συνέπειες για τον υδρολογικό σχεδιασμό;***

Ένα γενικό συμπέρασμα που προκύπτει από αυτή την ανάλυση είναι ότι η συχνότητα και το μέγεθος των ακραίων φαινομένων έχουν γενικά υποτιμηθεί στο παρελθόν, δεδομένου ότι οι πιο συχνά χρησιμοποιούμενες κατανομές για την ακραία ημερήσια βροχόπτωση έχουν «λεπτή» ουρά όπως της κατανομής Γάμα. Αυτό σημαίνει ότι ο υδρολογικός σχεδιασμός βάσει αυτών των κατανομών είναι μια επικίνδυνη πρακτική και ως εκ τούτου πρέπει να αναθεωρηθεί αναγνωρίζοντας ότι τα ακραία γεγονότα δεν είναι τόσο σπάνια όσο έχουν θεωρηθεί στο παρελθόν. Εν κατακλείδι, για την ορθότερη μοντελοποίηση των ακραίων βροχοπτώσεων προτείνεται η χρήση κατανομών με υποεκθετικές ουρές.

## Σχετικά με τις κατανομές ακραίων τιμών

Αναλύονται οι χρονοσειρές της ετήσιας μέγιστης ημερήσιας βροχόπτωσης σε 15 137 σταθμούς από όλο τον κόσμο με στόχο να απαντηθεί ίσως το βασικότερο ερώτημα της στατιστικής υδρολογίας, δηλαδή, ποια εκ των τριών κατανομών ακραίων τιμών περιγράφει καλύτερα τα ετήσια μέγιστα της ημερήσιας βροχόπτωσης .

Οι τρεις κατανομές ακραίων τιμών είναι οι τύπου I ή Gumbel (G), η τύπου II ή Fréchet (F) και η τύπου III ή ανάστροφη Weibull (reversed Weibull; RW) με συναρτήσεις κατανομής, που δίνονται, αντίστοιχα, από τις

$$G_{\mathrm{G}}(x) = \exp\left(-\exp\left(-\frac{x-\alpha}{\beta}\right)\right), \qquad x \in \mathbb{R} \tag{11}$$

$$G_{\mathrm{F}}(x) = \exp\left(-\left(\frac{x-\alpha}{\beta}\right)^{-1/\gamma}\right), \qquad x \geq \alpha \tag{12}$$

$$G_{\mathrm{RW}}(x) = \exp\left(-\left(-\frac{x-\alpha}{\beta}\right)^{1/\gamma}\right), \qquad x \leq \alpha \tag{13}$$

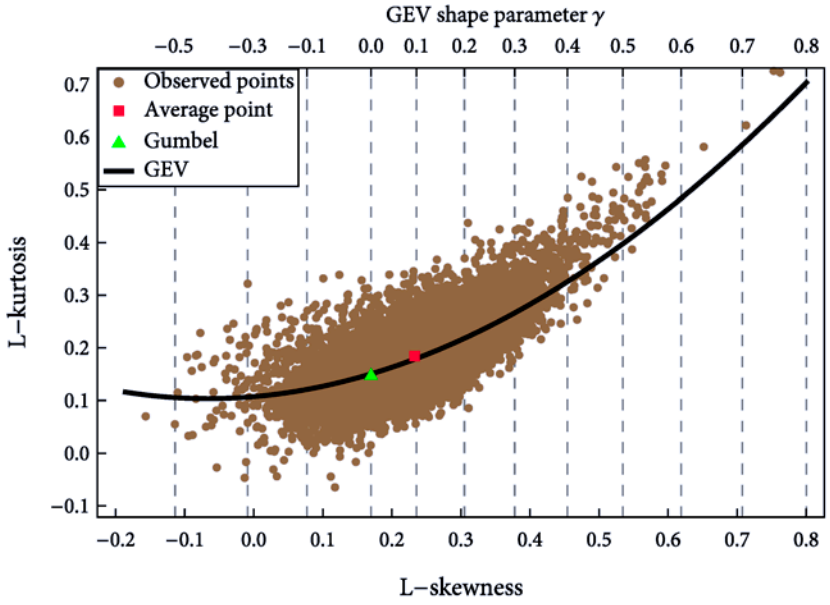
Οι τρείς αυτοί τύποι κατανομών μπορούν να ενοποιηθούν σε μια ενιαία έκφραση γνωστή ως Γενικευμένη Κατανομή Ακραίων Τιμών (Generalized Extreme Value; GEV) με συνάρτηση κατανομής

$$G_{\mathrm{GEV}}(x) = \exp\left( -\left( 1+\gamma\frac{x-\alpha}{\beta} \right)^{-1/\gamma} \right), \quad 1+\gamma\frac{x-\alpha}{\beta} \geq 0 \tag{14}$$

Η τιμή της παραμέτρου σχήματος της GEV αποκαλύπτει και τον τύπο της κατανομής ακραίων τιμών, ήτοι, για $\gamma < 0$ αντιστοιχεί στην RW, για $\gamma \to 0$ στην G και για $\gamma > 0$ στην F. Γι' αυτό και στην ανάλυση η έμφαση δόθηκε στη εκτίμηση αυτής της παραμέτρου.

- ***Ποια εκ των τριών τύπων κατανομών περιγράφει καλύτερα τη μέγιστη ημερήσια βροχόπτωση του έτους;***

Ξεκινώντας με κάποια θεωρητική τεκμηρίωση σημειώνεται ότι η RW προϋποθέτει μια μητρική κατανομή για την ημερήσια βροχόπτωση με άνω όριο το οποίο είναι φυσικά ασυνεπές, γεγονός που ενισχύεται λαμβάνοντας υπόψιν ότι άνω φραγμένες κατανομές δεν έχουν χρησιμοποιηθεί για την ημερήσια βροχόπτωση σε αξιόπιστες μελέτες. Συγκρίνοντας την Fréchet εναντίον της Gumbel προκύπτει, όσο και αν φαίνεται αντιφατικό, πως τα ετήσια μέγιστα ακόμη και αν προέρχονται από μητρική κατανομή που ανήκει στο πεδίο έλξης Gumbel περιγράφονται καλύτερα την κατανομή τύπου Fréchet. Αυτό συμβαίνει για δύο λόγους: πρώτον, ο ρυθμός σύγκλισης των μητρικών υποεκθετικών κατανομών στην κατανομή Gumbel είναι εξαιρετικά αργός, και δεύτερον, η παράμετρος σχήματος της κατανομής Fréchet επιτρέπει στην κατανομή να προσαρμόζεται αρκετά καλά όχι μόνο σε κατανομές με ουρές τύπου δύναμης, αλλά και σε άλλες υποεκθετικές ουρές. Όσον αφορά τα εμπειρικά στοιχεία που προκύπτουν από την ανάλυση των χρονοσειρών η «ετυμηγορία» είναι σαφής, δηλαδή, η κατανομή Fréchet επικρατεί έναντι των άλλων δυο ασυμπτωτικών κατανομών. Στο Σχήμα 4 παρουσιάζονται τα εμπειρικά σημεία των L-ροπών σε σύγκριση με τη θεωρητική καμπύλη της GEV και παρατηρείται πως το νέφος των σημείων είναι μετατοπισμένο δεξιά του σημείου της Gumbel υποδηλώνοντας πως το μεγαλύτερο μέρος των σημείων, για την ακρίβεια το 80%, περιγράφεται καλύτερα από την κατανομή Fréchet.



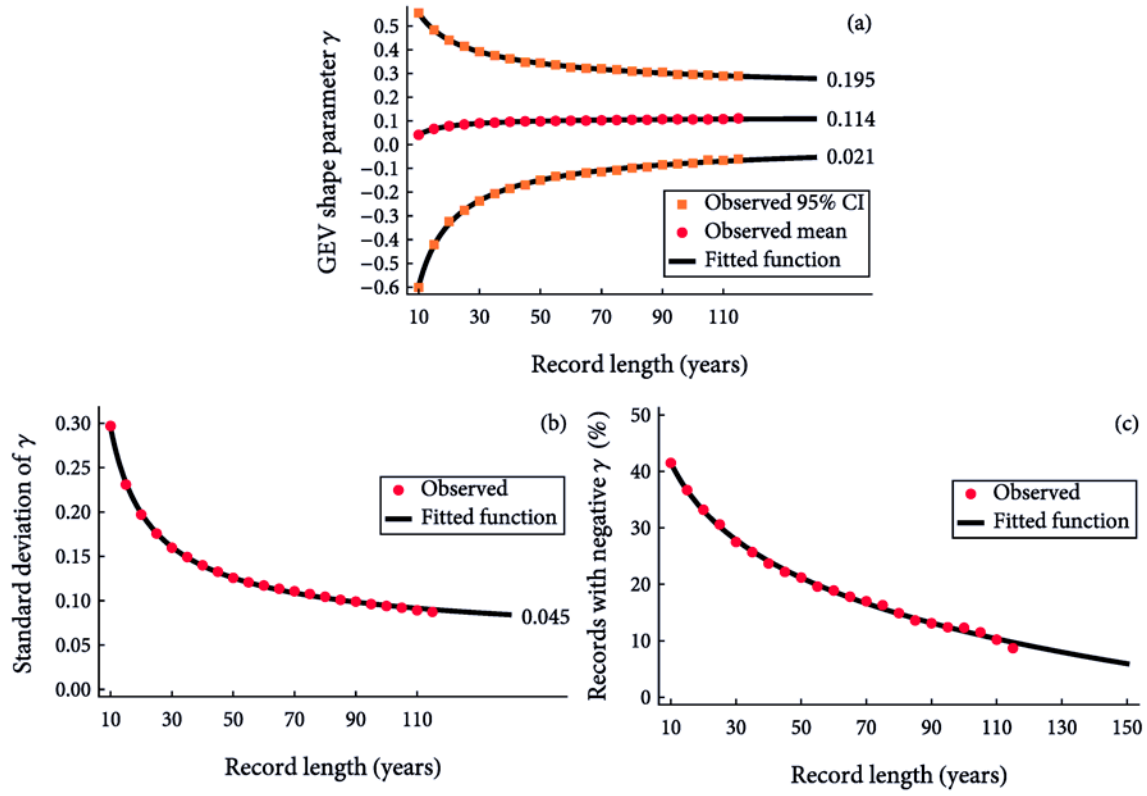

**Σχήμα 4.** Εμπειρικά σημεία L-κύρτωσης συναρτήσει L-ασυμμετρίας των 15 137 χρονοσειρών ετήσιας μέγιστης βροχόπτωσης σε σύγκριση με τη θεωρητική καμπύλη της GEV και του σημείου της Gumbel.

- ***Επηρεάζεται η εκτίμηση της παραμέτρου σχήματος της κατανομής GEV από το μήκος του δείγματος;***

Η ανάλυση αποκαλύπτει μια σαφή σχέση μεταξύ της τιμής της παραμέτρου σχήματος της κατανομής GEV και του μήκους δείγματος, γεγονός που σημαίνει ότι μόνο πολύ μεγάλα δείγματα μπορούν να αποκαλύψουν την πραγματική τιμή αυτής της παραμέτρου ή αλλιώς την πραγματική συμπεριφορά των ακραίων βροχοπτώσεων. Ενδεικτικά το Σχήμα 5 παρουσιάζει προσαρμοσμένες θεωρητικές καμπύλες σε εμπειρικά σημεία.

- ***Ποια είναι η πραγματική κατανομή της παραμέτρου σχήματος της GEV;***

Η «ασυμπτωτική» ανάλυση που πραγματοποιήθηκε, βάσει των συναρτήσεων που προσαρμόστηκαν (Σχήμα 5) στη μέση τιμή και στην τυπική απόκλιση της παραμέτρου σχήματος της GEV σε σχέση με το μήκος του δείγματος, αποκαλύπτει ότι η κατανομή της παραμέτρου σχήματος της GEV όπως θα προέκυπτε αν εξαιρετικά μεγάλα δείγματα ήταν διαθέσιμα είναι περίπου κανονική με μέση τιμή 0.114 και τυπική απόκλιση 0.045.




**Σχήμα 5.** (α) Μέση τιμή, ποσοστημόρια $Q_5$ και $Q_{95}$ όπως έχουν εκτιμηθεί για διάφορα μήκη χρονοσειρών και προσαρμοσμένες θεωρητικές καμπύλες; (β) τυπική απόκλιση; (γ) ποσοστό σταθμών με αρνητική τιμή της παραμέτρου σχήματος.

- ***Σε ποιο εύρος είναι αναμενόμενο να κυμαίνεται η πραγματική τιμή της παραμέτρου σχήματος της GEV; Μπορούμε να έχουμε τυφλή εμπιστοσύνη στις εκτιμήσεις που προκύπτουν από τα δεδομένα;***

Η παράμετρος σχήματος της GEV αναμένεται να ανήκει σε ένα στενό εύρος, περίπου από το 0 έως 0.23 με αξιοπιστία 99%. Ουσιαστικά, η ανάλυση δείχνει ότι δε μπορεί να εμπιστευτεί κανείς τυφλά τα δεδομένα καθώς τα μικρά κυρίως δείγματα μπορούν να παραμορφώσουν την πραγματική εικόνα. Στην κατεύθυνση αυτή, η Εξ. (15) διορθώνει την εκτιμήσεις της παραμέτρου σχήματος της GEV που βασίζονται στις L-ροπές αφαιρώντας τη μεροληψία λόγω του περιορισμένου μεγέθους του δείγματος. Η εξίσωση προκύπτει συνδυάζοντας την ασυμπτωτική κατανομή της παραμέτρου σχήματος $\gamma$ που αναμένεται να είναι η $N(\mu_\gamma, \sigma_\gamma^2)$ και την κατανομή για συγκεκριμένο μήκος χρονοσειράς $n$ που αναμένεται να είναι η $N(\mu_\gamma(n), \sigma_\gamma^2(n))$. Όπου $\mu_\gamma(n) = \mu_\gamma - 0.69\, n^{-0.98}$ και $\sigma_\gamma(n) = \sigma_\gamma + 1.27\, n^{-0.70}$ είναι οι καμπύλες που έχουν προσαρμοστεί στη μέση τιμή και στην τυπική απόκλιση (Σχήμα 5). Η αμερόληπτη εκτιμήτρια $\tilde{\gamma}(n)$ που προκύπτει δίνεται από τη σχέση

$$\tilde{\gamma}(n) = \frac{\sigma_\gamma}{\sigma_\gamma(n)}\left(\hat{\gamma} - \mu_\gamma(n)\right) + \mu_\gamma \tag{15}$$

όπου $n$ το μήκος του δείγματος (σε έτη), $\hat{\gamma}$ η κλασική εκτιμήτρια των L-ροπών και $\mu_\gamma \simeq 0.114$ και $\sigma_\gamma \simeq 0.045$.

- ***Είναι δόκιμη η χρήση της GEV με αρνητική τιμή της παραμέτρου σχήματος (άνω φραγμένη κατανομή);***

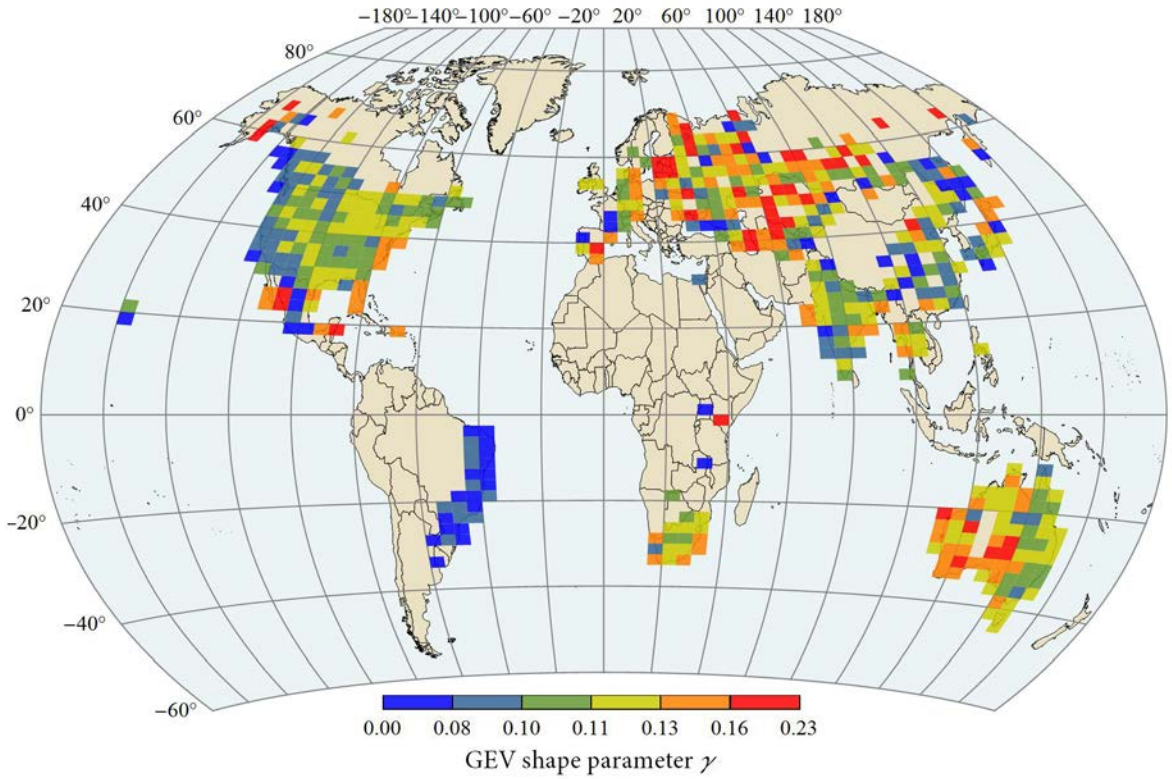
Σε ένα μικρό ποσοστό των σταθμών που αναλύθηκαν (20%) η αρχική εκτίμηση της παραμέτρου σχήματος της GEV ήταν αρνητική (reversed Weibull), ωστόσο η ανάλυση αποκαλύπτει ότι το ποσοστό αυτό μειώνεται ταχύτατα καθώς μέγεθος του δείγματος αυξάνεται, ενώ η συνάρτηση που έχει προσαρμοστεί και εκφράζει τη σχέση με το μέγεθος δείγματος δείχνει ότι για το μήκος δείγματος μεγαλύτερο από 226 χρόνια το ποσοστό αυτό θα είναι μηδέν (Σχήμα 5). Επιπλέον, κανένα από τα 16 δείγματα με μήκος μεγαλύτερο από 140 χρόνια δεν αντιστοιχεί σε αρνητική τιμή της παραμέτρου σχήματος. Επιπρόσθετα, η πιθανότητα να εμφανιστεί αρνητική παράμετρος σχήματος σύμφωνα με την κατανομή που έχει προκύψει είναι μόνο 0.005 και συνδυάζοντας αυτό το συμπέρασμα με τα προηγούμενα ευρήματα προκύπτει ότι μια κατανομή GEV με αρνητική παράμετρο σχήματος (άνω φραγμένη) είναι εντελώς ακατάλληλη για τη βροχόπτωση.

- ***Υπάρχει γεωγραφική διαφοροποίηση της παραμέτρου σχήματος της GEV;***

Η μελέτη της μέσης τιμής της παραμέτρου σχήματος της GEV σε περιοχές που ορίζονται από διαφορά γεωγραφικού πλάτους $\Delta\varphi = 2.5°$ και διαφορά γεωγραφικού μήκους $\Delta\lambda = 5°$ και η κατασκευή αντίστοιχου χάρτη (Σχήμα 6) αποδεικνύει ότι μεγάλες περιοχές του κόσμου μοιράζονται περίπου την ίδια τιμή της παραμέτρου σχήματος, ωστόσο είναι προφανές πως διαφορετικές περιοχές του πλανήτη παρουσιάζουν διαφορετική συμπεριφορά στην ακραία βροχόπτωση.

- ***Ποια η σημασία αυτών των ευρημάτων και τι θα μπορούσε να προταθεί ως πρακτικός κανόνας;***

Η κατανομή ακραίων τιμών Fréchet, ή αλλιώς η GEV με θετική παράμετρο σχήματος, υπερισχύει της κατανομής Gumbel και πρωτίστως της reversed Weibull, με την τελευταία να αποτελεί επικίνδυνη επιλογή για τον υδρολογικό σχεδιασμό. Ως γενικός κανόνας προκύπτει πως ακόμη και στην περίπτωση όπου τα δεδομένα υποδεικνύουν μια κατανομή GEV με αρνητική παράμετρο σχήματος το συμπέρασμα αυτό δεν πρέπει να θεωρηθεί αξιόπιστο, αντ' αυτού, προτείνεται η κατανομή Gumbel ή για πρόσθετη ασφάλεια η κατανομή GEV με τιμή παραμέτρου σχήματος ίση με 0.114.




**Σχήμα 6.** Γεωγραφική κατανομή της μέσης τιμής της παραμέτρου σχήματος της GEV. Οι εκτιμήσεις έχουν γίνει βάσει της αμερόληπτης εκτιμήτριας από την εξίσωση (15).

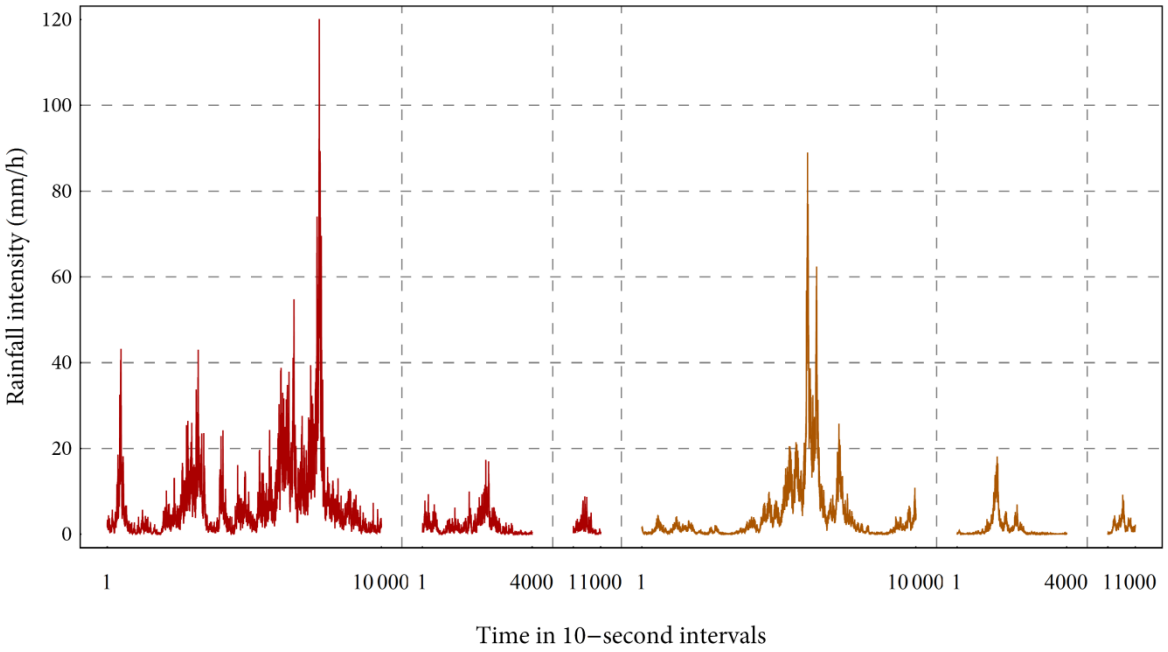
## Σχετικά με τα στοχαστικά χαρακτηριστικά της βροχόπτωσης σε μικρή χρονική κλίμακα

Εξετάζονται οι στοχαστικές ιδιότητες της βροχόπτωσης σε λεπτή χρονική κλίμακα, μελετώντας ένα μοναδικό σύνολο δεδομένων που περιλαμβάνει μετρήσεις επτά επεισοδίων βροχόπτωσης με χρονική διακριτοποίηση 5-10 δευτερόλεπτων [*Georgakakos et al.*, 1994]. Το ερώτημα που τίθεται και επιχειρείται να απαντηθεί είναι αν είναι δυνατόν ένα μοναδικό και απλό στοχαστικό μοντέλο να αναπαράγει τη μεγάλη στατιστική διαφοροποίηση που παρατηρήθηκε στα επεισόδια αυτά, καθώς και να εντοπιστούν τα κύρια χαρακτηριστικά του.

- ***Μπορεί ένα απλό στοχαστικό μοντέλο να παραγάγει χρονοσειρές βροχόπτωσης που διαφέρουν δραστικά μεταξύ τους;***

Είναι εφικτό ένα μοναδικό και σχετικά απλό στοχαστικό μοντέλο να παραγάγει επεισόδια βροχόπτωσης σε λεπτή χρονική κλίμακα με στατιστικά χαρακτηριστικά που διαφέρουν πάρα πολύ μεταξύ τους. Το αποτέλεσμα είναι τα παραγόμενα επεισόδια βροχόπτωσης να "φαίνονται" πολύ διαφορετικά μεταξύ τους, όπως ακριβώς έχει παρατηρηθεί και σε καταγεγραμμένα επεισόδια αυτής της χρονικής κλίμακας. Στο Σχήμα 7 παρουσιάζονται

συνθετικά επεισόδια βροχόπτωσης από δυο στοχαστικά μοντέλα με την ίδια περιθώρια κατανομή αλλά διαφορετική δομή αυτοσυσχέτισης.




**Σχήμα 7.** Συνθετικά επεισόδια βροχόπτωσης που έχουν παραχθεί από μοντέλο με δομή αυτοσυσχέτισης τύπου δύναμης (τα τρία πρώτα) και εκθετικού τύπου (τα τρία τελευταία).

- ***Ποιά είναι τα χαρακτηριστικά ενός τέτοιου μοντέλου;***

Ένα τέτοιο μοντέλο χαρακτηρίζεται από «ισχυρή» δομή αυτοσυσχέτισης, που μειώνεται δηλαδή σιγά-σιγά με τη χρονική υστέρηση, καθώς επίσης και από ουρά κατανομής που μειώνεται σιγά-σιγά με την ένταση της βροχής. Η αυτοσυσχέτιση αυτού του τύπου μπορεί να παράγει τεράστιες διαφοροποιήσεις στη χρονική δομή των διαφόρων επεισοδίων, ενώ μια περιθώρια κατανομή με τέτοιου τύπου ουρά μπορεί να παράγει εξαιρετικά υψηλές εντάσεις βροχής. Τα δύο αυτά χαρακτηριστικά είναι ακριβώς αντίθετα με τις πιο γνωστές στοχαστικές ανελίξεις που μοιάζουν με Γκαουσιανό λευκό θόρυβο, οι οποίες θα παρήγαγαν πολύ "ομαλά" επεισόδια με εξαιρετικά σπάνια την εμφάνιση μεγάλων εντάσεων. Από την άποψη αυτή, τόσο η «ισχυρή» αυτοσυσχέτιση όσο και οι «χοντρές» ουρές μπορούν να ιδωθούν ως ιδιότητες που αυξάνουν την τυχαιότητα και την αβεβαιότητα (ή την εντροπία), μιας και οι παραγόμενες χρονοσειρές από ανελίξεις με αυτά τα χαρακτηριστικά μπορούν να διαφέρουν δραματικά μεταξύ τους και συνεπώς είναι λιγότερο «προβλέψιμες» συγκριτικά με χρονοσειρές που προκύπτουν από μοντέλα τύπου Markov με περιθώρια κατανομή με «λεπτή» ουρά.

# CONTENTS

# CHAPTER 1

*"One day I will find the right words, and they will be simple."*

JACK KEROUAC

# INTRODUCTION

## 1.1 Motivation

The great philosopher Henry David Thoreau once said that "*Our life is frittered away by detail. Simplify, simplify.*" Maybe, science suffers too from detail and complexity, and although "axiomatically" has to go down to detail and deal with complexity, this does not imply that the more general, the simpler, and the more fundamental questions have been answered. A fundament question is usually simple but never simplistic and not necessarily easy to answer. Setting this kind of questions, even if the answers can be found, is not always enough as these answers have to be useful, of wide interest, and of theoretical and practical value too.

This is exactly the motivation of the research presented in this thesis, i.e., to try to reply to some fundamental and of wide interest questions, mainly regarding the statistical properties of daily rainfall, that from the author's perception have not been clearly answered. The fundamental questions explored here regard:

- the seasonal variation of the marginal distribution of daily rainfall;
- the existence or not of a "universal" model capable to probabilistically describe rainfall at all areas of the world and for every season;
- the possibility to apply well-established and theoretically justified principles like the Principle of Maximum Entropy to derive these models;
- the probabilistic nature of the extreme daily rainfall, i.e., what type of distribution tail better describes extremes above a threshold value or which one of the three extreme value distributions better describes annual daily maxima;

- and last but not least the stochastic nature or rainfall at very fine temporal scales.

Most of these questions cannot be answered solely based on theoretical considerations, and even if they would, empirical verification would still be necessary in order to verify the theory and to shift from theory to practice. Moreover, these questions cannot either be answered based on empirical analyses of limited datasets. Yet in our digital era very large datasets exist that may help provide answers to these questions and also to others too. In this direction and trying to exploit these datasets, massive analyses of empirical data were performed (among the largest ever conducted in statistical hydrology as far as the author is aware) from thousands of stations all around the globe, hoping that the derived answers are scientifically sound, empirically justified, and also are consistent with common sense—because as Pierre-Simon Laplace phrased it, almost two centuries ago, "*Probability theory is nothing but common sense reduced to calculation.*"

## 1.2 Outline

This thesis is designed so as each chapter can be read independently from the others. It is organized as follows:

**Chapter 2** explores the prospect to use the Principle of Maximum Entropy with the Boltzmann-Gibbs-Shannon entropy in order to derive suitable probability distributions for rainfall, or more generally, for geophysical processes. The emphasis is on formulating and logically justifying the constraints used with entropy maximization.

**Chapter 3** investigates the seasonal variation of daily rainfall focusing on the properties of its marginal distribution. A massive empirical analysis is performed of more than 170 000 monthly daily rainfall records from more than 14 000 stations from all over the globe aiming to answer two major questions: (a) which statistical characteristics of daily rainfall vary the most over the months and how much, and (b) whether or not there is a relatively simple probability model that can describe the nonzero daily rainfall at every month and every area of the world.

**Chapter 4** focuses on the distribution tail of daily rainfall, i.e., the distribution's part that describes the extreme events. More than 15 000 daily rainfall records are examined in order to test the performance of four common distribution tails that correspond to the Pareto, the Weibull, the Lognormal and the Gamma distributions aiming to find out which of them better describes the behaviour of extreme events.

**Chapter 5** regards the analysis of annual maxima of daily rainfall. The annual maxima time series from more than 15 000 stations from all over the world are extracted and analysed in order to answer one of the most basic question in statistical hydrology, i.e.,

which one of the three Extreme Value distributions better describes the annual maximum daily rainfall.

**Chapter 6** examines the stochastic properties of rainfall at fine temporal scales by studying a unique dataset comprising measurements of seven storm events at a temporal resolution of 5-10 seconds. The question raised and attempted to be answered is if it is possible for a single and simple stochastic model to generate a plethora of temporal rainfall patterns, as well as to detect the major characteristics of such a model.

**Chapters 7** completes the thesis with a brief summary and the conclusions.

## 1.3 Innovation points

***Application of the principle of maximum entropy***

The principle of maximum entropy is a well-established tool to make inference under uncertainty or to find the most suitable probability distribution under the available information. Entropy maximization is traditionally performed using classical moments as constraints. This practice, along with the classical definition of entropy, leads to exponential type distributions with light tails that are in contradiction with empirical evidence, i.e., many natural phenomena cannot be probabilistically described by these distributions. To tackle this problem several generalizations of entropy measures emerged that, however, have been criticized for their theoretical consistency, and additionally, from the author's perspective, still do not result in distributions flexible enough to describe most of geophysical random variables. In this direction:

i. A new rationale is formed regarding the application of the principle of maximum entropy that is based on using the classical and well-justified definition of entropy, i.e., the Boltzmann-Gibbs-Shannon entropy (BGS), with suitable constraint that lead to flexible distributions appropriate for positive and skewed random variables.
ii. The constraints formed and used in the maximization of the BGS entropy are theoretically or rationally justified and differ from those that have been commonly used. Particularly, the constraints formed are the expected value of the logarithmic function, the classical moments but of unspecified order, and a generalization of the classical moments.
iii. The generalization of the classical moments proposed here, named $p$-moments, is justified and leads naturally to power type distribution avoiding thus the use of generalized entropy measures.

iv. The BGS entropy maximization under two basic combinations of the aforementioned constraints leads to two distribution which are by far more flexible than those emerging using the commonly used constraints along with the BGS entropy or even along with generalized entropy measures. These distributions are: (a) the Generalized Gamma distribution (GG) which is of exponential form yet its right tail can be heavy depending on its parameter values, and (b) the Generalized Beta of the Second Kind (GB2) which is an extremely flexible four-parameter power type distribution. For practical reasons instead of the GB2 the use of a three-parameter simplification is proposed named the Burr type XII (BrXII).

***The marginal distribution of daily rainfall***

Literature reveals that numerous different probability models, some of them completely different to each other, have been used to describe daily rainfall, depending on the season or the area of the world. Two major questions are explored here: (a) whether the marginal distribution of daily rainfall varies or not markedly over the months and how much, and (b) if there is a simple probability model capable to describe the nonzero daily rainfall at every month and every area of the world. In this direction:

i. An unprecedented massive empirical analysis was performed of more than 170 000 monthly daily rainfall records from more than 14 000 stations from all over the globe.
ii. In order to verify the seasonal variation of some important statistical characteristics of daily rainfall an original test, named the SV-Test, was formed and applied indicating that the shape characteristic of the marginal distribution, generally, vary over the months.
iii. The efficacy of the distributions derived previously from entropy maximization, i.e., the GG and the BrXII, was tested. These distributions have not been used systematically before to describe daily rainfall yet the analysis revealed that they both performed very well with the GG distribution performing exceptionally well.
iv. Analytical equations of the first three L-moments were derived for the BrXII distribution. Additionally, the theoretical L-skewness *vs.* L-variation space was formed for the GG and the BrXII distributions that proved a valuable tool providing insights on the performance of those distributions and of many other that are special cases of them.
v. An ad hoc fitting method was constructed based on L-moments in order to fit these distributions fast and with the accuracy that L-moments provide.

vi. It was revealed that the most commonly used models for rainfall, e.g., like the Gamma distribution or other exponential-tail distribution consist a dangerous choice, as these model can severely underestimate the frequency and the magnitude of extreme events. Additionally, none of the commonly used two-parameter models can serve as a "universal" model for daily rainfall as these models cannot match the variation in shape that the empirical analysis revealed.

***Distribution tails of the daily rainfall***

The upper part of the distribution, commonly named as tail, is of great importance in hydrological design as it describes the extreme events. Yet identifying the type of tail that better describes the daily rainfall is not trivial as the tail constitutes that part of the distribution for which, usually, empirical data are not available. For this reason, previous studies that analysed a limited number of records may offer a blur picture as the extreme behaviour, by definition, needs a lot of information to be revealed. On the contrary, this information can be found by analysing large datasets from all over the world. In this direction:

i. A massive analysis of more than 15 000 daily rainfall records was performed in order to draw conclusions regarding the nature of the tail, i.e., if it is heavy or light and more specifically to find out which one of the common type of tails better describes extremes.
ii. The method proposed here differs from the classical peak above threshold (POT) analysis which essentially is based on the generalized Pareto distribution. Specifically, the performance of four common distribution tails that correspond to the Pareto, the Weibull, the Lognormal and the Gamma distributions was tested by directly fitting these distributions only to the tail data.
iii. The fitting of the distribution to the empirical tail data was accomplished by introducing and using a modified least square norm that proved to be better than the commonly used and almost unbiased. The fitting method was verified using original and intensive Monte Carlo schemes.
iv. The analysis revealed that the tail of the most commonly used model, i.e., the Gamma distribution, performed the worst while that with the heavier tail, i.e., the Pareto distribution, performed the best.
v. A world map was constructed depicting the variation of the percentage of best fitted subexponential tails indicating thus the areas where subexponential tails prevail.

***Distribution of annual maximum daily rainfall***

Probably, the most basic question in statistical hydrology is which one of the three Extreme Value distributions better describes the annual maximum daily rainfall. Literally hundreds of studies exist using extreme value distributions and arguing against or for one of them. Despite the importance and the popularity of the subject, most studies are of local character, i.e., limited to specific areas, or analysing a limited number of records that fails to provide a clear answer to the aforementioned question. To provide an answer to this question:

i. A massive analysis of annual maxima time series from more than 15 000 stations was performed.
ii. In general, most of the existing records of daily rainfall contain missing values. Obviously extracting the annual maximum value form a record with missing values is not completely reliable as a larger value may have occurred during the missing days. For this reason an original method was formed for extracting the annual maxima values from incomplete records. The method uses a combination of two simple criteria, i.e., the percentage of missing values per year and the rank of the year's maximum value. The method was verified by an original Monte Carlo scheme and proved very robust and reliable.
iii. The Generalized Extreme Value (GEV) distribution comprises all three extreme value distributions and "switches" to one or the other depending on the value of its shape parameter stressing thus the parameter's importance. The analysis indicated a clear relationship between this parameter with the record length while a new kind of asymptotic analysis was performed and revealed the true distribution of this parameter. The reliability of this distribution was verified based on original Monte Carlo simulations.
iv. An original and practical formula that corrects the L-moments estimation bias, induced by small or finite length records, was created.
v. World maps with the mean value of the GEV shape parameter were constructed and revealed a clear geographical variation of this parameter, yet large areas can be found that share approximately the same parameter value.
vi. The massive number of records analysed indicated clearly that the Fréchet law prevails over the Gumbel law and over the reversed Weibull law with the latter two laws, in general, consisting a dangerous choice in hydrological design.

***Stochastic properties of rainfall at fine temporal scales***

The study of a unique dataset comprising measurements of seven storm events at a fine temporal resolution of 5-10 seconds shows that these storm events differ significantly to each other with some of them being completely different in terms of their statistical properties. The question raised and attempted to be answered here is if this complexity can emerge by a simple underlying stochastic process and if it is possible to construct a single and simple stochastic model capable of reproducing this complexity by generating various storm patterns that differ markedly from one another.

i. An original and effective normalizing transformation was invented able to normalize the original dataset having a marginal distribution that deviated severely from the normal distribution.
ii. An original stochastic model or else a stochastic simulation scheme was created by using the reverse of the aforementioned normalizing transformation and incorporating bias correction formulas.
iii. The assumption that all these storm events are the outcome of a sole process cannot be rejected by the analysis. On the contrary, the simulation and the synthetic storm events produced by a single model fortify this possibility. Although it seems counterintuitive that such a model has a very strong autocorrelation structure, as someone would expect strong autocorrelation to generate similar events, it is exactly this feature, combined with a marginal distribution with heavy tails, which creates rich and different storm patterns.

# CHAPTER 2

*"Only entropy comes easy"*

*ANTON CHEKHOV*

# ENTROPIC DISTRIBUTIONS

**ABSTRACT**

The principle of maximum entropy, along with empirical considerations, can provide consistent basis for constructing a consistent probability distribution model for highly varying geophysical processes. This study examines the potential of using this principle with the Boltzmann-Gibbs-Shannon entropy definition in order to derive suitable probability distributions for rainfall or more generally for geophysical processes. Specific simple and general entropy maximization constraints are defined and theoretically justified which lead to two flexible distributions, i.e., the three-parameter Generalized Gamma (GG) and the four-parameter Generalized Beta of the second kind (GB2), with the former being a particular limiting case of the latter.

## 2.1 Introduction

Even though long-term predictions of rainfall are not possible in deterministic terms (e.g., weather forecasts are skilful for no more than a week ahead), in probabilistic terms it is possible to assign a stochastic model or a probabilistic law and to any rainfall amount assign a return period or a probability of exceedance. Actually, most infrastructures affected by rainfall and flood are designed this way. Rainfall is generally characterized as an intermittent stochastic process (for fine timescales), with a mixed-type marginal distribution, partly discrete and partly continuous. The discrete part is concentrated at zero and defines the probability dry, while the rest is continuously spread over the positive real axis and determines the nonzero rainfall distribution. The discrete part of the rainfall distribution can be easily estimated as the ratio of the number of dry days to total number of days. On the contrary, the continuous part of the distribution cannot be easily assessed.

Rainfall is usually studied in many different timescales, e.g., from sub-hourly to yearly, yet, the daily timescale is one of the most convenient and important in hydrological design. Specifically, it is the smallest timescale for which thousands of records exist with some of them being more than a century long. Nevertheless, and although daily rainfall has been extensively studied over the years, a search in the literature reveals that a universally accepted model for the wet-day daily rainfall distribution does not exist. On the contrary, many distributions have been proposed in specific studies for specific locations of the world including, e.g., the two-parameter Gamma, which is probably the prevailing model, the two- and three-parameter Lognormal, the Generalized Logistic, the Pearson Type III, the Pareto and the Generalized Pareto, the three- and four-parameter Kappa distributions, and many more.

The common method to construct an appropriate probability distribution model for describing one or more samples is to try a variety of different models and choose the best fitted using a particular mathematical norm, e.g., a least square error or a likelihood norm. Nevertheless, this approach is rather naïve and laborious; first, there are (at least theoretically) infinitely many different models to try, and second, this method does not offer any theoretical justification for the final choice, thus making it an ad hoc empirical choice. This practice explains why so numerous models have been proposed. Here, the principle of maximum entropy is used as a solid theoretical background for constructing an appropriate probability distribution for rainfall and for geophysical processes in general. These theoretically derived results, i.e., the resulting probability distributions, are tested for their validity in the next chapter by using more than 180 000 daily rainfall records across

the world aiming also to assess whether a single generalized model could be appropriate for all rainfall records worldwide.

## 2.2 Entropy measures

The concept of entropy dates back to the works of Rudolf Clausius in 1850, yet, it was Ludwig Boltzmann around 1870 who gave entropy a statistical meaning and related it to statistical mechanics. The concept of entropy was advanced later in the works of J. Willard Gibbs in thermodynamics and Von Neumann in quantum mechanics, and was reintroduced in information theory by *Claude Shannon* [1948], who showed that entropy is a purely probabilistic concept, a measure of the uncertainty related to a random variable (RV).

The most famous and well justified measure of entropy for continuous RVs, is the Boltzmann-Gibbs-Shannon (BGS) entropy, which for a non-negative RV $X$ is

$$S_X = -\int_0^\infty f_X(x) \ln f_X(x) \mathrm{d}x \tag{2.1}$$

where $f_X(x)$ is the probability density function of $X$. The BGS entropy is not the only entropy measure. A search in the literature reveals that more than twenty different entropy measures have been proposed, mainly generalizations of BGS entropy (for a summary of entropy measures see [*Esteban and Morales*, 1995]). Among those measures, it is worth noting the Rényi entropy, introduced by the Hungarian mathematician Alfréd Rényi in 1961, which have been used in many different disciplines, e.g., ecology and statistics. It is also worth noting another entropy measure that has gained much popularity in the last decade, the Havrda-Charvat-Tsallis (HTC) entropy. It was initially proposed by *Havrda and Charvat* [1967] and was reintroduced and applied to physics by *Tsallis* [1988]. Apart from its use in physics, the HTC entropy has also been used more recently in hydrology as it gives rise to power-type distributions. The HTC entropy is a generalization of the BGS entropy given by

$$S_X(q) = \frac{1 - \int_0^\infty \left( f_X(x) \right)^q \mathrm{d}x}{q-1} \tag{2.2}$$

It is easy to verify that for $q = 1$ it becomes identical to the BGS entropy.

## 2.3 The principle of maximum entropy

The principle of maximum entropy was established, as a tool for inference under uncertainty, by *Edwin Jaynes* [1957a, 1957b]. In essence, the principle of maximum entropy relies in finding the most suitable probability distribution under the available information. As *Jaynes* [1957a] expressed it, the resulted maximum entropy distribution "is the least biased estimate possible on the given information; i.e., it is maximally noncommittal with regard to missing information".

In a mathematical frame, the given information used in the principle of maximum entropy, is expressed as a set of constraints formed as expectations of functions $g_j(\ )$ of $X$, i.e.,

$$E\Big(g_j(X)\Big) = \int_0^\infty g_j(x) f_X(x)\mathrm{d}x = c_j, \qquad j = 1,...,n \tag{2.3}$$

The resulting maximum entropy distributions emerge by maximizing the selected form of entropy with constraints (2.3), and with the obvious additional constraint

$$\int_0^\infty f_X(x)\mathrm{d}x = 1 \tag{2.4}$$

The maximization is accomplished by using calculus of variation and the method of Lagrange multipliers. Particularity, the general solution of the maximum entropy distributions resulting from the maximization of BGS entropy and the HCT entropy, assuming arbitrary constraints are, respectively,

$$f_X(x) = \exp\left(-\lambda_0 - \sum_{j=1}^{n} \lambda_j g_j(x)\right) \tag{2.5}$$

$$f_X(x) = \left(1 + (1-q)\left(\lambda_0 + \sum_{j=1}^{n} \lambda_j g_j(x)\right)\right)^{-\frac{1}{1-q}} \tag{2.6}$$

where $\lambda_j$, with $j = 1,\ldots, n$, are the Lagrange multipliers linked to the constraints (2.3) and $\lambda_0$ is the multiplier linked to the constraint (2.4), i.e., $\lambda_0$ guarantees the legitimacy of the distribution.

## 2.4 Justification of the constraints

It becomes clear from the above discourse that the resulting maximum entropy distribution is uniquely defined by the choice of the imposed constraints. This implies that this choice is the most important and determinative part of the method. Constraints express our state of knowledge concerning a RV and should summarize all the available information from observations or from theoretical considerations. Nevertheless, choosing constraints is not trivial; they are introduced as expectations of RV functions without any intrinsic limitation on the form of those functions.

So, how would one choose the appropriate constraints among an infinite number of choices? In classical statistical mechanics, these constraints are imposed by physical principles such as the mass, momentum and energy conservation. However, in complex geophysical processes, these principles cannot help. In geophysical processes, the standard procedure to assign a probability law is to study the available observations and infer the underlying distribution without entropy considerations. However, whatever is inferred in this way, is in fact based on a small portion of the past (the available record), which may (or may not) change in the future. Nevertheless, it can reasonably assumed that some RV features may be more likely to be approximately preserved in the future than others, e.g., coarse features like the mean and the variance are less likely to change in the future [*Jaynes*, 2003] than finer features based on higher moments (e.g., it is well known that the kurtosis coefficient is extremely sensitive to observations and additional observations may radically alter it). Therefore, as a first rule, constraints should be simple and express those features that are likely to be preserved in the future.

The previous rule is rather subjective in the sense that is difficult to distinguish between simple and not simple constraints or to foresee what RV quantities will be preserved. Furthermore, the use of a particular set of "simple" constraints may lead to a distribution that is not supported by the empirical data. Obviously, it is difficult to reject or verify the detailed shape features of a distribution based on a small sample which apparently does not provide the sufficient amount of information needed. Nonetheless, many geophysical processes, even if long records do not exist for particular regions, are extensively recorded worldwide e.g., thousands of stations record precipitation, temperature, etc. Thus, the study of this massive amount of information may lead in determining some important prior characteristics of the underlying distribution that should be preserved, e.g., a J- or bell-shaped distribution or a heavy- or light-tailed

distribution. Therefore, constraints should be chosen not only based on simplicity, but also on the appropriateness of the resulting distribution given the empirical evidence.

Commonly used constraints in maximizing entropy assume known mean and variance, i.e., known first and second moments, which are clearly two very simple constraints. Particularly, entropy maximization assuming known first two moments leads: (a) to the celebrated normal distribution in the BGS entropy case, or, to the truncated normal if the mandatory constraint of non-negativity for geophysical processes is imposed, and (b) to a symmetric bell-shaped distribution with power-type tails in the HCT entropy case, or, its truncated version for a non-negative RV. The distribution arising in the HCT case for zero mean is now known as the Tsallis distribution. For non-zero mean the resulting distribution is the Pearson type VII introduced by *Pearson* in 1916, whose special case is the Tsallis distribution. Both these distributions are symmetric bell-shaped, in which asymmetry can only emerge by truncation at zero. As a consequence, those distributions may likely fail to describe sufficiently many geophysical processes that exhibit a rich pattern of asymmetries (e.g., it is well known that the rainfall in small time scales is heavily skewed and likely heavy tailed).

Accordingly, this study aims to define some simple and general constraints alternative to those of the first two moments that lead to suitable probability distributions for geophysical processes, particularly for rainfall. Additionally, another aim is to use only the BGS entropy, which is theoretically justified and widely accepted, avoiding the use of generalized entropy measures.

The mean is one of the most commonly used constraints, as it is a classical measure of central tendency. Another useful measure of central tendency, exhibiting the convenient property for geophysical processes to be defined only for positive values, is the geometric mean $\mu_G$. An estimate of this, from a sample of size $n$, is given by

$$\mu_G = \left( \prod_{i=1}^{m} x_i \right)^{1/n} = \exp\left( \frac{1}{n} \sum_{i=1}^{n} \ln x_i \right) = \exp\left( \overline{\ln x} \right) \tag{2.7}$$

where the overbar stands for the sample average. The sample geometric mean (also referred as a constraint in [*Kapur*, 1989]) is smaller than the arithmetic mean. Intuitively, this leads to the formulation the following constraint for entropy maximization

$$E(\ln X) = \ln \mu_G \tag{2.8}$$

The expectation of $\ln X$, apart from its relationship to the geometric mean and its simplicity, makes an essential constraint for positively skewed RVs. To clarify, samples drawn from positively skewed distributions and, even more so, drawn from heavy-tailed distributions, exhibit values located on the right area very far from the mean value; in a sense, those values act like outliers and consequently strongly influence the sample moments, especially those of higher order. Therefore, it is not rational to assume that sample moments, especially based on samples drawn from heavy-tailed distributions, are likely to be preserved. On the contrary, the function $\ln x$ applied to this kind of samples eliminates the influence of those "extreme" values and offers a very robust measure that is more likely to be preserved than the estimated sample moments. Essentially for this reason, the logarithmic transformation is probably the most common transformation used in hydrology as it tends to normalize positively skewed data.

As stated above, the link of the mean and variance with the physical principles of momentum and energy conservation is invalid in geophysical processes. For example, the mean of the rainfall is not its momentum and its variance it is not its energy. Even in these processes, mean and variance (as measures of central tendency and dispersion) provide useful information, which can at least explain general behaviours and shapes of probability density functions [*Koutsoyiannis*, 2005a]. However, this information is good only for explanatory purposes and does not enable detailed and accurate modelling. For, there do not exist theoretical arguments (apart from simplicity and conceptual meaning as measures of central tendency and dispersion) which to favour mean and variance against, e.g., fractional moments of small order or even negative. For example, if the second moment is likely to be preserved, then probably the square root moment is more likely to be preserved as it is more robust in outliers. Additionally, low order fractional moments can be related with the $\ln x$ function, as it is well known that

$$\lim_{q\to 0}\frac{x^q-1}{q}=\ln x \tag{2.9}$$

Thus, it could be said that the function $x^q$ for small values of $q$ behaves similar to $\ln x$, thus exhibiting properties similar to those of the logarithmic function described above.

Based on this reasoning it is deemed that, instead of choosing the order of moments a priori, it is better to let the order unspecified, so that any value can be a posteriori chosen, including small fractional values. This leads in imposing as a constraint any moment $m_q$ of order $q$, i.e.,

$$m_q = E(X^q) = \int_0^\infty x^q f_X(x)\mathrm{d}x \tag{2.10}$$

One reason that many entropy generalizations have emerged was to explain many empirically detected deviations from exponential type distributions that arise from the BGS entropy using standard moment constraints. Yet, generalized entropy measures have been criticized for lacking theoretically consistency and for being arbitrary, a reasonable argument considering the large number of entropy generalizations available in the literature. Here, instead of using generalized entropy measures that might result in power-law distributions, the important notion of moments is generalized inspired by the limiting definition of the exponential function, i.e., $\exp(x^q) = \lim_{p\to 0}(1+px^q)^{1/p}$. First the function $x_p^q$ is defined as

$$x_p^q := \ln(1+px^q)/p \tag{2.11}$$

which for $p = 0$ becomes the familiar power function $x^q$ as $x_0^q = \lim_{p\to 0}\ln(1+px^q)/p = x^q$. Thus, a generalization of the classical moments can be defined, given the name $p$-moments, by

$$m_q(p) = E(X_p^q) = \frac{1}{p}\int_0^\infty \ln(1+px^q) f_X(x)\mathrm{d}x \tag{2.12}$$

Arguably, this generalization is arbitrary and many other moment generalizations can be (and in fact are) constructed. Nonetheless, is deemed that there is a rationale that supports the use of $p$-moments, which can be summarized as follows: (a) if generalized entropy measures, considered by many as arbitrary, have been successfully used, then there is no reason to avoid using generalized moments; (b) maximization of the BGS entropy using $p$-moments leads, as will become apparent in the next section, to flexible power-type distributions (including the Pareto and Tsallis distributions for $q$ = 1 and $q$ = 2, respectively); (c) $p$-moments are simple and, for $p$ = 0, become identical to the ordinary moments; and (d) they are based on the $x_p^q$ function that exhibits all the desired properties, like those of the $\ln x$ function described above, and thus are suitable for positively skewed RVs; additionally, compared to $E(\ln X)$ they are always positive.

## 2.5 The resulting entropy distributions

Entropy optimization can be accomplished in many different combinations of the previously defined constraints (see Table 2.1); however, here, two simple combinations of the aforementioned constraints are used based on the type and the generality of the distributions emerging. Particularly, the $E(\ln X)$ constraint is combined, first, with the classical moments, and second, with the $p$-moments, letting in both cases the moment order arbitrary.

**Table 2.1.** The resulting maximum entropy distributions for various imposed constraints.

| Constraints | Distribution Name | Density function | Ref. No. |
|---|---|---|---|
| $m_1$ | Exponential | $f_X(x) = C\exp(-\lambda_1 x)$ | (2.13) |
| $m_2$ | Half-Normal | $f_X(x) = C\exp(-\lambda_1 x^2)$ | (2.14) |
| $m_1$ and $m_2$ | Normal | $f_X(x) = C\exp(-\lambda_1 x - \lambda_2 x^2)$ | (2.15) |
| $m_q$ | Generalized Exponential | $f_X(x) = C\exp(-\lambda_1 x^q)$ | (2.16) |
| $m_1$ and $E(\ln x)$ | Gamma | $f_X(x) = Cx^{-\lambda_1}\exp(-\lambda_2 x)$ | (2.17) |
| $m_q$ and $E(\ln x)$ | Generalized Gamma | $f_X(x) = Cx^{-\lambda_1}\ \exp(-\lambda_2 x^q)$ | (2.18) |
| $m_1(p)$ | Pareto type II | $f_X(x) = C(1 + x/p)^{-\lambda_1 p}$ | (2.19) |
| $m_2(p)$ | Tsallis | $f_X(x) = C\left(1 + (x/p)^2\right)^{-\lambda_1 p^2}$ | (2.20) |
| $m_1(p)$ and $m_2(p)$ | Not named | $f_X(x) = C\left(1 + x/p\right)^{-\lambda_1 p}\left(1 + (x/p)^2\right)^{-\lambda_2 p^2}$ | (2.21) |
| $m_q(p)$ | Not named | $f_X(x) = C\left(1 + (x/p)^q\right)^{-\lambda_1 p^q}$ | (2.22) |
| $m_1(p)$ and $E(\ln x)$ | Beta of the second kind | $f_X(x) = C\ x^{-\lambda_1}(1 + x/p)^{-\lambda_2 p}$ | (2.23) |
| $m_q(p)$ and $E(\ln x)$ | Generalized Beta of the second kind | $f_X(x) = C\ x^{-\lambda_1}\left(1 + (x/p)^q\right)^{-\lambda_2 p^q}$ | (2.24) |

where $C = \exp(-\lambda_0)$ is the integration constant so that $\int_0^\infty f_X(x) = 1$.

In the first case, the maximization of the BGS entropy, given in (2.1), with constraints (2.8) and (2.10) results in the density function

$$f_X(x) = \exp(-\lambda_0 - \lambda_1 \ln x - \lambda_2 x^q) \tag{2.25}$$

which after algebraic manipulations and parameter renaming (please see Appendix A for details) can be written as

$$f_X(x) = \frac{\gamma_2}{\beta\Gamma(\gamma_1/\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma_2}\right), \qquad x \geq 0 \tag{2.26}$$

corresponding to the distribution function

$$F_X(x) = 1-\Gamma\left(\frac{\gamma_1}{\gamma_2},\left(\frac{x}{\beta}\right)^{\gamma_2}\right)/\Gamma\left(\frac{\gamma_1}{\gamma_2}\right), \qquad x \geq 0 \tag{2.27}$$

where $\Gamma(a) = \int_0^\infty t^{a-1}\exp(-t)\mathrm{d}t$ is the Gamma function and $\Gamma(a,x) = \int_x^\infty t^{a-1}\exp(-t)\mathrm{d}t$ is the upper incomplete Gamma function.

This distribution, commonly attributed to Stacy [1962] appeared much earlier in the literature in the works of Amoroso around 1920, and seems to have been rediscovered many times under different forms [see e.g., *Kleiber and Kotz*, 2003].Here, a slightly different form is used compared to the one proposed by Stacy. Essentially, it is a generalization of the Gamma distribution and will be denoted by $\mathrm{GG}(\beta,\gamma_1,\gamma_2)$, or simply GG. It is a very flexible distribution that includes many other well-known distributions as particular cases, e.g., the Gamma, the Weibull, the Exponential, or even the Chi-squared distributions and others.

The distribution includes the scale parameter $\beta > 0$, and the shape parameters $\gamma_1 > 0$ and $\gamma_2 > 0$. The parameter $\gamma_1$ controls the behaviour of the left tail, i.e., if $0 < \gamma_1 < 1$ the density function is J-shaped and for $x \to 0$, $f_X(x) \to \infty$; if $\gamma_1 > 1$ the density function is bell-shaped and mainly positively skewed; yet, for certain values of $\gamma_1$ and $\gamma_2$ it can be symmetric or even negatively skewed, and for $x = 0$, $f_X(x) = 0$; finally, for $\gamma_1 = 1$ the distribution degenerates to a generalized exponential function and for $x = 0$, $f_X(0) < \infty$. The parameter $\gamma_2$ is very important as for fixed $\gamma_1$ it controls the behaviour of the right tail, i.e., it determines the frequency and the magnitude of the extreme events. Generally and loosely speaking, for $\gamma_2 < 1$ the distribution can be characterized as sub-exponential or heavy-tailed, and for $\gamma_2 > 1$ as hyper-exponential or light-tailed [for a classification of distribution tails see *Goldie and Klüppelberg*, 1998]. Figure 2.1, where several probability density functions of the Generalized Gamma distribution are depicted, clearly,

demonstrates its flexibility in terms of shape. Notably, the distribution is also valid if the shape parameters are simultaneously negative (a generalized inverse Gamma distribution); however, the distribution loses some important shape characteristics and seems not suitable for geophysical RV like rainfall, thus, here the distribution is only considered for positive shape parameters.

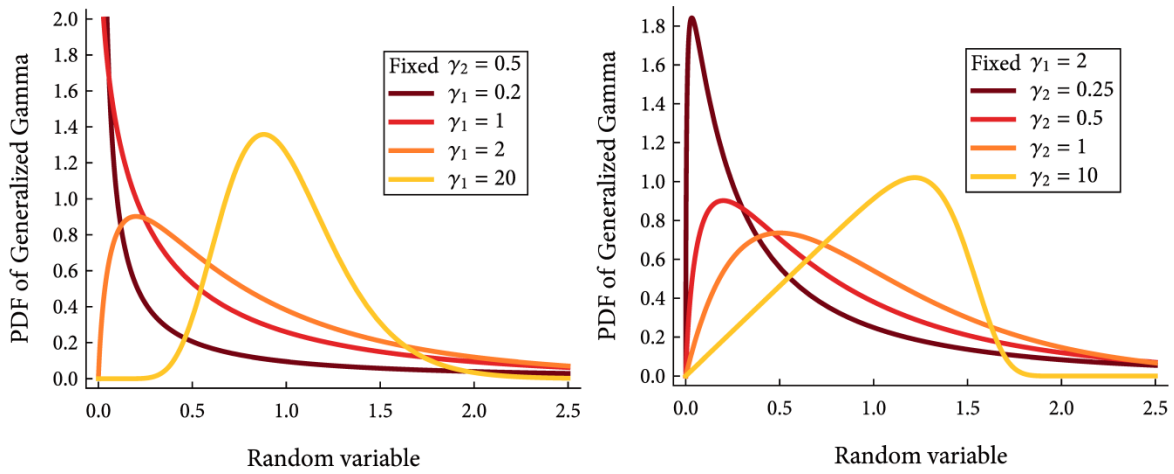



**Figure 2.1.** Probability density functions of the Generalized Gamma distribution for various shape parameter values. The values of scale parameter $\beta$ were chosen so that mean value of each distribution equals 1.

In the second case, the maximization of the BGS entropy with constraints (2.8) and (2.12) results in the density function

$$f_X(x) = \exp(-\lambda_0 - \lambda_1 \ln x - \lambda_2 \ln(1 + px^q)/p) \tag{2.28}$$

which after algebraic manipulations and parameter renaming (please see appendix A for details) can be written as

$$f_X(x) = \frac{\gamma_3}{\beta \mathrm{B}(\gamma_1, \gamma_2)} \left(\frac{x}{\beta}\right)^{\gamma_1\gamma_3 - 1} \left(1 + \left(\frac{x}{\beta}\right)^{\gamma_3}\right)^{-(\gamma_1+\gamma_2)}, \qquad x \geq 0 \tag{2.29}$$

corresponding to the distribution function

$$F_X(x) = \mathrm{B}_z(\gamma_1, \gamma_2)/\mathrm{B}(\gamma_1, \gamma_2), \qquad \text{where } z = \left(1 + (x/\beta)^{-\gamma_3}\right)^{-1} \tag{2.30}$$

where $\mathrm{B}(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the Beta function and $\mathrm{B}_x(a,b)=\int_0^x t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the incomplete Beta function.

This distribution has not been formed earlier on a similar rationale, yet, a search in the literature reveals that it has been rediscovered many times under different names and parameterizations. It is most commonly known as the Generalized Beta of the second kind—hereafter denoted as $\mathrm{GB2}(\beta,\gamma_1,\gamma_2,\gamma_3)$, or simply GB2. It seems that Milke and Johnson [1974] were the first that formed this distribution, and proposed it for describing hydrological and meteorological variables. It has also been used in different disciplines, e.g., McDonald [1984] used the GB2 as an income distribution. Nevertheless, the distribution can be considered as a simple generalization of many well-known and much earlier introduced distributions, e.g., the *F*-distribution or the Pearson type VI of the celebrated Pearson system.

The GB2 distribution is a very flexible four-parameter distribution with $\beta>0$ being the scale parameter, and $\gamma_1>0$, $\gamma_2>0$ and $\gamma_3>0$ being the three shape parameters, allowing the distribution to form very many different shapes. The GB2 distribution includes as special or limiting cases many of the well-known distributions, e.g., the Beta of the second kind, the Pareto type II, the Loglogistic, the Burr type XII, even the Generalized Gamma [*McDonald*, 1984; *Kleiber and Kotz*, 2003].

Obviously, the flexibility of the GB2 distribution makes it a good model for describing rainfall—the GB2 has already been used under the name JH distribution, to describe the rainfall in a large range of timescales [*Papalexiou and Koutsoyiannis*, 2008b] and to construct theoretically consistent IDF curves [*Papalexiou and Koutsoyiannis*, 2008a]. Nonetheless, as a general rule based on the principle of parsimony, a three-parameter model is preferable than a four-parameter model, provided that the simpler model describes the data adequately. Additionally, it is not reasonable to compare the performance of the GG distribution, which is a three-parameter model, with GB2, which is a four-parameter model. Thus, a simpler form of the GB2 distribution is selected based on its flexibility and its simple analytical expression of the distribution function, and consequently, of the quantile function.

A simple three-parameter form of GB2 is derived by setting $\gamma_1=1$ in Eq. (2.29). By renaming the parameters and after algebraic manipulations a distribution is obtained known as the Burr type XII [*Burr*, 1942] (denoted hereafter as BrXII), which was

introduced by Burr in 1942 in the framework of a distribution system similar to Pearson's. Its probability density function is

$$f_X(x) = \frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\gamma_1 - 1}\left(1 + \gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right)^{-\frac{1}{\gamma_1\gamma_2}-1}, \qquad x \geq 0 \tag{2.31}$$

and its distribution function is

$$F_X(x) = 1 - \left(1 + \gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right)^{-\frac{1}{\gamma_1\gamma_2}}, \qquad x \geq 0 \tag{2.32}$$

The BrXII distribution is a flexible power-type distribution that comprises the scale parameter $\beta > 0$ and the shape parameters $\gamma_1 > 0$ and $\gamma_2 \geq 0$. The shape flexibility of the Burr type XII distribution is demonstrated in Figure 2.2 where several probability density functions, for various combinations of the shape parameters, are depicted.

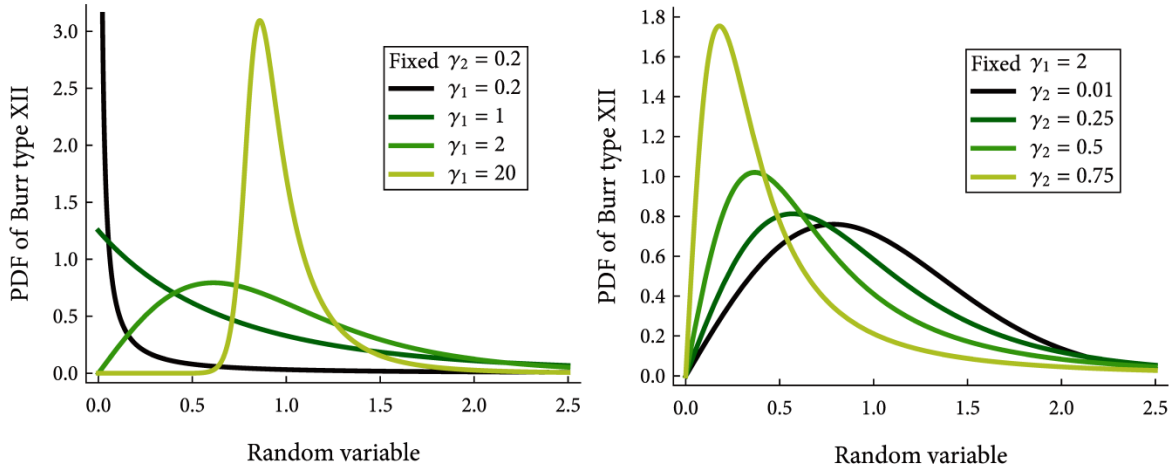



**Figure 2.2.** Probability density functions of the Burr type XII distribution for various shape parameter values. The values of scale parameter $\beta$ were chosen so that mean value of each distribution equals 1.

The form of the BrXII distribution used here is not the one found in the literature [see e.g., *Tadikamalla*, 1980]. The expression (2.31) is preferred because it is suggestive of a generalization of the familiar Weibull distribution (for $\gamma_2 \to 0$) and also because the asymptotic behaviour of the right tail is solely controlled by the parameter $\gamma_2$ (for large values of $X$, $P(X > x) \sim \gamma_2 \beta^{1/\gamma_2} x^{-1/\gamma_2}$). The distribution has a finite variance distribution for

$0 \le \gamma_2 < 0.5$ and finite mean for $0 \le \gamma_2 < 1$. Finally, the shape parameter $\gamma_1$ controls the left tail as for $0 < \gamma_1 < 1$ the distribution is J-shaped, for $\gamma_1 > 1$ bell-shaped and for $\gamma_1 = 1$ degenerates to the familiar Pareto type II distribution.

## 2.6 Summary and conclusions

In order to derive statistical distributions suitable for geophysical processes, and particularly for rainfall, a rationale is proposed for defining and selecting constraints within a BGS entropy maximization framework. Entropy maximization offers a solid theoretical basis for identifying a probabilistic law based on the available information, in contrast to the common technique of choosing a distribution from a repertoire based on trial-and-error methods. This rationale is based on the premises that the constraints should be as few and simple as possible and incorporate prior information on the process of interest. This prior information may concern the general shapes of densities and could be obtained by studying the process worldwide. Three particular constraints are studied and conceptually justified that are related to the logarithmic and the power functions, which are suitable for positive, highly varying and asymmetric RVs. Namely, the constraints are the expected values of (a) $\ln x$; (b) $x^q$; and (c) $\ln(1+px^q)/p$. The last constraint generalizes the classical moments and naturally leads to power-type distributions avoiding generalized entropy measures.

The BGS entropy maximization under two combinations of these constraints leads to two flexible distributions, i.e., a three-parameter exponential type, known as the Generalized Gamma (GG), and, a four-parameter power type, known as the Generalized Beta of the second kind (GB2)—the former is a particular limiting case of the latter. Another three-parameter model, known as the Burr type XII (power type), easily derived from the GB2, proves to be also useful.

# CHAPTER 3

*"O, wind, if winter comes, can spring be far behind?"*

*PERCY BYSSHE SHELLEY*

# A WORLDWIDE SURVEY ON THE DISTRIBUTION OF DAILY RAINFALL

**ABSTRACT**

To characterize the seasonal variation of the marginal distribution function of daily rainfall, it is important to find which statistical characteristics of daily rainfall actually vary the most from month to month and which could be regarded to be invariant. Relevant to the latter issue is the question whether there is a single model capable to describe effectively the nonzero daily rainfall for every month and at every area of the world. To study these questions a massive analysis is performed of more than 170 000 monthly daily rainfall records at more than 14 000 stations from all over the globe. The analysis indicates that: (a) the shape characteristics of the marginal distribution of daily rainfall, generally, vary over the months, (b) commonly used distributions like the Exponential, the Gamma, the Weibull, the Lognormal, or the Pareto, etc. are incapable to describe "universally" the daily rainfall, (c) exponential-tail distributions like the Exponential, mixed Exponentials or the Gamma can severely underestimate the magnitude of extreme events and thus they constitute a dangerous choice, and (d) the Burr type XII and the Generalized Gamma distributions are two good models, with the latter performing exceptionally well.

## 3.1 Introduction

Most geophysical processes exhibit seasonal variation, which implies an underlying regular pattern, which potentially enables a degree of predictability, utilizing the periodic changes of the process's coarse behaviour with time. This is exactly why it is important to correctly characterize the seasonal variability of geophysical processes. Among those, rainfall is one of the most important, as it highly affect human lives. For example, agricultural, irrigation and water supply planning, and more generally water resources management, in order to be efficient and competent, has to take seasonality into account. Seasonality does not necessarily refer to the four standard seasons of the temperate zones, but it generally describes the within year variability. An effective scale to characterize seasonality is the monthly scale.

Rainfall, if perceived as a stochastic process, is determined by two components: its marginal probability distribution and its dependence structure. It is reasonably expected these components to vary periodically if rainfall is studied at any subannual time scale. Furthermore, it is rational to assume that the daily time scale is the finest time scale in which the seasonality could be studied without complications, because rainfall at subdaily scales may also be affected by earth's daily rotation (the daily cycle). In practice, estimating and trying to reproduce the statistical characteristics of rainfall on a daily basis can be a laborious task and, most importantly, can have questionable reliability as the estimation of the various characteristics will be based on small samples. For this reason, daily rainfall is typically studied and modelled on a monthly basis assuming that within a specific month its statistical characteristics remain essentially invariant. Consequently, the daily rainfall process can be decomposed into 12 different processes with fixed monthly autocorrelation structure and fixed monthly marginal distribution. This study does not concern with the autocorrelation structure [see e.g., *Haan et al.*, 1976; *Waymire and Gupta*, 1981; *Mimikou*, 1983, 1984; *Schoof and Pryor*, 2008] but it is focused on the monthly variation of the marginal distribution of the daily rainfall.

The marginal distribution of daily rainfall belongs to the so-called mixed type distributions and comprises two parts: a discrete part describing the probability dry and mathematically expressed as a probability mass concentrated at zero, and a continuous part spread over the positive real numbers describing probabilistically the amount or the intensity of nonzero rainfall. The probability dry, in general, can be easily assessed from empirical data as the as the dry-days to total-days ratio, while the continuous part is usually modelled by a parametric continuous distribution fitted to nonzero values. Yet this

distribution is not unique and in practice, as a literature review reveals, various distributions have been used for the nonzero daily rainfall. For example the Exponential distribution [e.g., *Smith and Schreiber*, 1974; *Todorovic and Woolhiser*, 1975], mixed Exponentials [e.g., *Woolhiser and Roldán*, 1982; *Wilks*, 1998, 1999], the Gamma distribution [e.g., *Buishand*, 1978a; *Bruhn et al.*, 1980; *Geng et al.*, 1986], the Weibull distribution [e.g., *Swift and Schreuder*, 1981; *Wilson and Toumi*, 2005], the Lognormal distribution [e.g., *Biondini*, 1976; *Swift and Schreuder*, 1981], mixed Lognormals [*Shimizu*, 1993], power-type distributions like the two-, three- and four-parameter Kappa distributions [*Mielke Jr*, 1973; *Mielke Jr and Johnson*, 1973; *Hosking*, 1994; *Park et al.*, 2009], generalized Beta distributions [*Mielke Jr and Johnson*, 1974], as well as the Generalized Pareto [e.g., *Fitzgerald*, 1989] for peaks over threshold, and probably many more.

A question that can be raised based on the aforementioned studies and on many more is whether or not all of these distributions, some completely different with each other in structure, are indeed suitable to probabilistically describe the (nonzero) daily rainfall or if they have prevailed and become popular for technical reasons, e.g., simplicity in their form. Additionally, most of these studies are of local character, i.e., they are based on the analysis of a limited number of rainfall records and from specific areas of the world. The exceptions are very few, e.g. in a study by *Papalexiou and Koutsoyiannis* [2012] daily rainfall was analysed in more than 10 000 stations worldwide. In practice, in most cases rainfall in modelled using exponential-type distributions like the Exponential distribution, the Gamma or mixed Exponentials [see e.g., *Foufoula-Georgiou and Lettenmaier*, 1987]. These, however, might be a very dangerous choice if the actual distribution of nonzero rainfall has a significantly heavier tail than those light-tail distributions that may severely underestimate the magnitude and the frequency of extreme events. Actually, two recent studies [*Papalexiou and Koutsoyiannis*, 2013; *Papalexiou et al.*, 2013], where daily rainfall extremes were analysed in more than 15 000 stations worldwide, revealed that most of the records cannot be described by exponential-tail distributions but rather by distributions with heavier tails.

In order to characterize the seasonal variation of the marginal distribution function of daily rainfall, the study aims in finding which statistical characteristics of daily rainfall actually vary the most from month to month and which could be regarded to be invariant. Relevant to the latter issue is the question whether there is a single model capable to describe effectively the nonzero daily rainfall for every month and at every area of the world. Obviously these questions cannot be answered by local analyses. Therefore, a

massive analysis is performed of more than 170 000 monthly daily rainfall records from more than 14 000 stations from all over the globe.

## 3.2 The data

The original database used here is the Global Historical Climatology Network-Daily database (version 2.60, www.ncdc.noaa.gov/oa/climate/ghcn-daily) which comprises thousands of daily rainfall records from stations all around the globe. Nevertheless, only a part of these records is used as many of them are very short in length, contain a large percentage of missing values, or have values of questionable accuracy which are assigned with various quality flags (details on quality flags can be found in the website given above). For these reasons and in order to create a robust subset of records with ensured quality, the records finally chosen fulfil the following criteria: (a) record length larger than 50 years, (b) missing values less than 20% and, (c) values assigned with quality flags less than 0.1%. As an additional measure to ensure the quality of the data all values assigned with flags "G" (failed gap check) or "X" (failed bounds check) were deleted as these flags are used for unrealistically large values. Fortunately, only 594 records in total had such values and typically no more than one or two values per record. The resulting subset comprises 15 137 stations (for further details on the dataset please Appendix B).

Although this study concerns the monthly daily rainfall, the daily rainfall of all months is also analysed as in some cases, especially for design purposes, the focus is not on the month that an event occurs but just on its exceedance probability or else on its return period. In this case monthly daily values can be merged and treated as represented by a single random variable (note that the term "daily rainfall" refers to daily rainfall values of all months while the term "monthly daily rainfall" refers to the daily rainfall values of individual months). From each station 13 different records were formed, one for all daily values and 12 for the monthly daily values, resulting in a total of 196 781 different records. Nevertheless, some months at stations located in very dry areas have very few nonzero rainfall values or even none so that estimation of the various important statistics would be highly uncertain or even impossible (e.g., estimation of L-skewness needs at least three values). To overcome this problem the minimum sample size of monthly nonzero rainfall values was constrained; so among the 15 137 records initially chosen were finally selected those having at least 20 nonzero values for each month resulting in a total of 14 157 stations and consequently 169 884 monthly daily records were formed. The locations of these stations and their corresponding lengths in years are given in the map of Figure 3.1. Note that in some areas the map cannot provide the clear picture of the record length

distribution. For example in the USA, the network of stations is very dense and inevitably points overlap, so that, below the layer of points representing high record lengths, other points exist representing smaller records lengths.

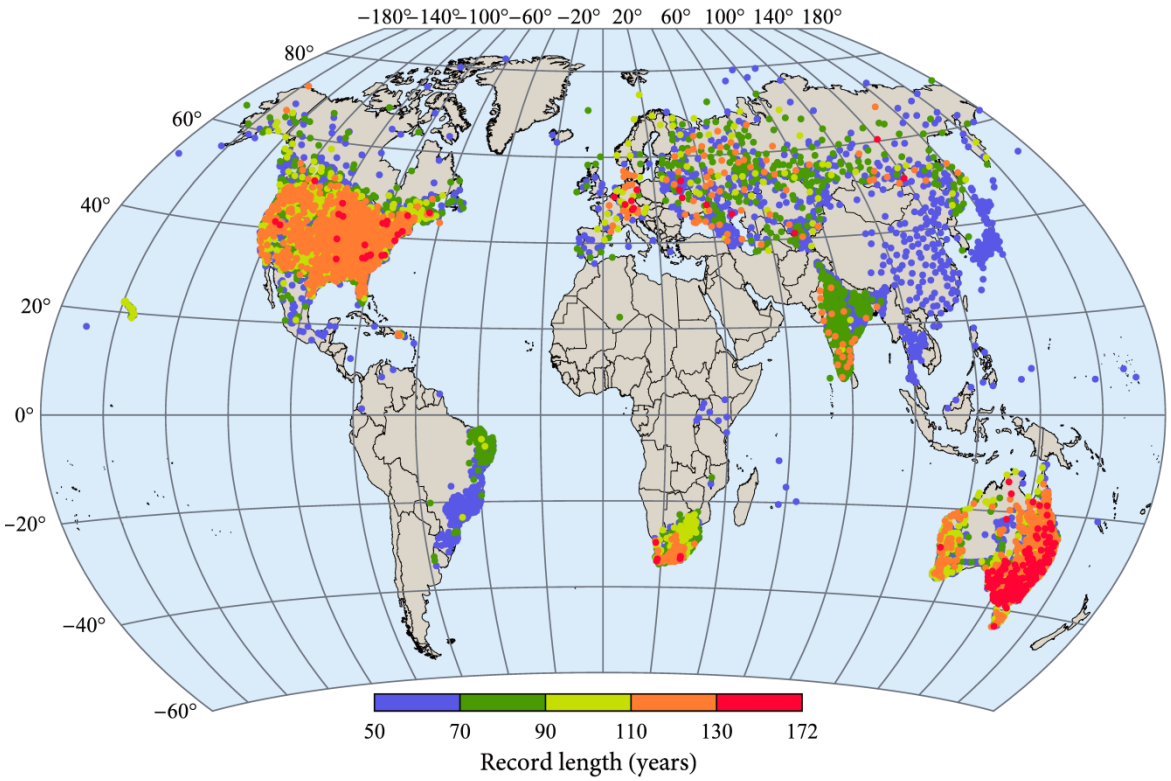



**Figure 3.1.** Locations of the 14 157 stations studied.

## 3.3 Seasonal variation

### 3.3.1 Statistics studied

To assess the seasonal variation of daily rainfall representative statistics of the marginal distribution are studied on a monthly basis. Additionally, in order for the study to be more complete as well as for comparison purposes these statistics were also estimated for the daily rainfall values of all months too (indicated with "All" in the figures). Particularly, the statistics studied are: (a) the probability dry, (b) the mean value, (c) the L-variation, and (d) the L-skewness. The probability dry expresses the discrete part of the marginal distribution and is simply estimated as the ratio of dry days to total days. The latter three are statistics for the continuous part of the marginal distribution describing thus the nonzero rainfall and obviously are calculated using only nonzero rainfall values.

The mean value of nonzero rainfall is a classical measure of central tendency while L-variation $\tau_2 = \lambda_2/\lambda_1$ and L-skewness $\tau_3 = \lambda_3/\lambda_2$, defined as ratios of L-moments $\lambda_i$ [*Hosking*, 1990], are dimensionless measures of the distributional shape. L-ratios are preferable over

ratios based on the classical moments like the coefficients of skewness and kurtosis as they exhibit better statistical properties, e.g., they are more robust [see e.g., *Hosking*, 1992]. Additionally, L-kurtosis (defined as $\tau_4 = \lambda_4/\lambda_2$) is also commonly used as a measure of shape, yet for positive random variables L-variation is well defined and actually is more robust and more convenient as it is bounded in [0,1]. Usually, L-variation or even the classical coefficient of variation (defined as the ratio of standard deviation to the mean value) are interpreted as standardized measures of variance; indeed, they express, respectively, the value of the second L-moment $\lambda_2$ and the value of the standard deviation of a distribution having mean value equal to 1. Yet for positive random variables, where actually these coefficients are meaningful, both depend on the distribution's shape parameters only or are constants if the distribution does not have shape parameters, and thus, they are essentially measures of distributional shape.

As already noted, it is anticipated from our experience the probability dry to vary over the months in most areas of the world. Additionally, it may seem obvious that the monthly mean value of daily rainfall (including zero values) will vary too as it is directly related to probability dry, e.g., a larger number of rainy days on average in a month logically will increase the monthly mean (estimated as the record's total monthly rainfall divided by the total number of month's days). However, it is not that evident that the mean value of the monthly nonzero daily rainfall (estimated as the record's total monthly rainfall divided by the total number of the month's rainy days) will vary over the months (during rainy days it could be possible to rain on average the same amount irrespective of the month). Finally, our perception on rainfall may lead to assume that extreme rainfall varies with season, e.g., it is well-known that specific weather mechanisms, responsible for extreme rainfall, are linked with specific seasons. Consequently, this may imply that the shape characteristics of rainfall distribution change over seasons, as the distribution's shape, particularly the right tail, controls the frequency and the magnitude of extreme events. Yet this assumption may be false as extreme rainfall may emerge by a change in the scale or else in the variance of rainfall and not necessarily by a change in its shape characteristics. For these reasons whether or not the distributional shape characteristics vary with season needs to be investigated and verified.

### 3.3.2 Variation in the hemispheres

Northern Hemisphere (NH) and Southern Hemisphere (SH) have opposite seasons and thus, it is reasonable to assume that natural processes under seasonal variation exhibit different behaviour between the two hemispheres. This may be generally valid, especially

for processes like the surface temperature, yet rainfall is a more complex process that may be affected more by regional climate conditions. For example, the celebrated Köppen climate classification [see e.g., *Kottek et al.*, 2006; *Peel et al.*, 2007], which classifies climate according to the annual and monthly average temperature and precipitation, defines several different types and subtypes of climate for each hemisphere. Thus, different rainfall patterns may appear even in adjacent areas of the same hemisphere.

Nevertheless, a first coarse approach that could provide a general picture is to present the seasonal variation of the statistics by hemisphere. Among the 14 157 stations analysed, 8447 belong in the NH and 5710 in the SH. The aforementioned statistics, i.e., the probability dry, mean value, L-variation and L-skewness, were calculated for the monthly daily rainfall of each station; their averages and standard deviations are given, for each hemisphere and additionally for the whole globe, in Table 3.1.

**Table 3.1.** Mean values and standard deviation values of the four statistics studied.

| | | All | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Northern hemisphere | | | | | | | | | | | | |
| $P_{dry}$ | $\mu$ | 72.03 | 73.55 | 74.23 | 74.03 | 73.18 | 71.05 | 68.49 | 67.80 | 68.97 | 71.37 | 74.70 | 73.68 | 73.65 |
| | $\sigma$ | 11.19 | 16.74 | 15.10 | 14.24 | 13.28 | 12.71 | 13.48 | 15.95 | 15.30 | 12.78 | 13.50 | 16.45 | 17.36 |
| $\mu$ | $\mu$ | 9.52 | 7.08 | 7.18 | 7.80 | 8.28 | 8.99 | 9.95 | 10.21 | 10.11 | 10.47 | 10.04 | 8.73 | 7.58 |
| | $\sigma$ | 4.67 | 4.31 | 4.26 | 4.25 | 4.14 | 4.31 | 4.86 | 5.22 | 4.70 | 4.94 | 5.20 | 4.93 | 4.57 |
| $\tau_2$ | $\mu$ | 0.59 | 0.56 | 0.56 | 0.56 | 0.57 | 0.57 | 0.58 | 0.58 | 0.59 | 0.59 | 0.59 | 0.57 | 0.57 |
| | $\sigma$ | 0.04 | 0.05 | 0.05 | 0.05 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.05 | 0.05 |
| $\tau_3$ | $\mu$ | 0.46 | 0.44 | 0.43 | 0.43 | 0.43 | 0.43 | 0.44 | 0.45 | 0.46 | 0.46 | 0.45 | 0.44 | 0.44 |
| | $\sigma$ | 0.05 | 0.06 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.06 | 0.06 |
| | | Southern hemisphere | | | | | | | | | | | | |
| $P_{dry}$ | $\mu$ | 77.91 | 77.73 | 76.80 | 78.29 | 79.69 | 78.32 | 76.50 | 76.84 | 77.91 | 78.77 | 77.74 | 77.85 | 78.05 |
| | $\sigma$ | 10.60 | 14.38 | 14.34 | 12.96 | 11.62 | 13.35 | 15.99 | 17.32 | 16.79 | 14.25 | 12.22 | 12.06 | 13.37 |
| $\mu$ | $\mu$ | 9.27 | 11.09 | 11.46 | 10.54 | 9.06 | 8.34 | 7.71 | 7.21 | 6.81 | 7.15 | 8.21 | 9.01 | 10.08 |
| | $\sigma$ | 3.70 | 4.56 | 4.47 | 4.22 | 3.55 | 3.22 | 3.19 | 2.98 | 2.62 | 2.74 | 3.15 | 3.53 | 4.04 |
| $\tau_2$ | $\mu$ | 0.58 | 0.59 | 0.59 | 0.59 | 0.58 | 0.58 | 0.58 | 0.57 | 0.56 | 0.56 | 0.56 | 0.56 | 0.57 |
| | $\sigma$ | 0.05 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.05 | 0.05 |
| $\tau_3$ | $\mu$ | 0.46 | 0.47 | 0.47 | 0.47 | 0.46 | 0.46 | 0.45 | 0.45 | 0.44 | 0.44 | 0.44 | 0.44 | 0.45 |
| | $\sigma$ | 0.06 | 0.07 | 0.07 | 0.07 | 0.07 | 0.07 | 0.08 | 0.08 | 0.08 | 0.07 | 0.07 | 0.07 | 0.07 |
| | | Global | | | | | | | | | | | | |
| $P_{dry}$ | $\mu$ | 74.40 | 75.24 | 75.27 | 75.75 | 75.80 | 73.99 | 71.72 | 71.44 | 72.58 | 74.36 | 75.92 | 75.36 | 75.42 |
| | $\sigma$ | 11.33 | 15.97 | 14.85 | 13.90 | 13.04 | 13.45 | 15.06 | 17.10 | 16.51 | 13.87 | 13.08 | 14.98 | 16.01 |
| $\mu$ | $\mu$ | 9.42 | 8.70 | 8.91 | 8.90 | 8.60 | 8.73 | 9.05 | 9.00 | 8.78 | 9.13 | 9.30 | 8.85 | 8.59 |
| | $\sigma$ | 4.31 | 4.83 | 4.83 | 4.45 | 3.93 | 3.92 | 4.41 | 4.69 | 4.31 | 4.50 | 4.57 | 4.42 | 4.53 |
| $\tau_2$ | $\mu$ | 0.58 | 0.57 | 0.57 | 0.57 | 0.57 | 0.57 | 0.58 | 0.58 | 0.58 | 0.58 | 0.57 | 0.57 | 0.57 |
| | $\sigma$ | 0.04 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| $\tau_3$ | $\mu$ | 0.46 | 0.45 | 0.45 | 0.45 | 0.44 | 0.44 | 0.45 | 0.45 | 0.45 | 0.45 | 0.44 | 0.44 | 0.44 |
| | $\sigma$ | 0.05 | 0.07 | 0.07 | 0.07 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.06 | 0.07 |

Furthermore, a better picture is provided by the box plots given in Figure 3.2 which present these statistics on a monthly basis and for each hemisphere. The left (red) box plots are for the NH while the right (grey) are for the SH while the box plot's inner lower and upper fences that define the box indicate, respectively, the 25% and 75% empirical quantile points and thus define the empirical interquartile range (IQR) or the 50% of the central values. The line within the box indicates the median, while the lower and upper fences of the whiskers indicate, respectively, the 5% and 95% empirical quantile points or else they define the 90% empirical confidence interval (ECI) of the studied statistics.

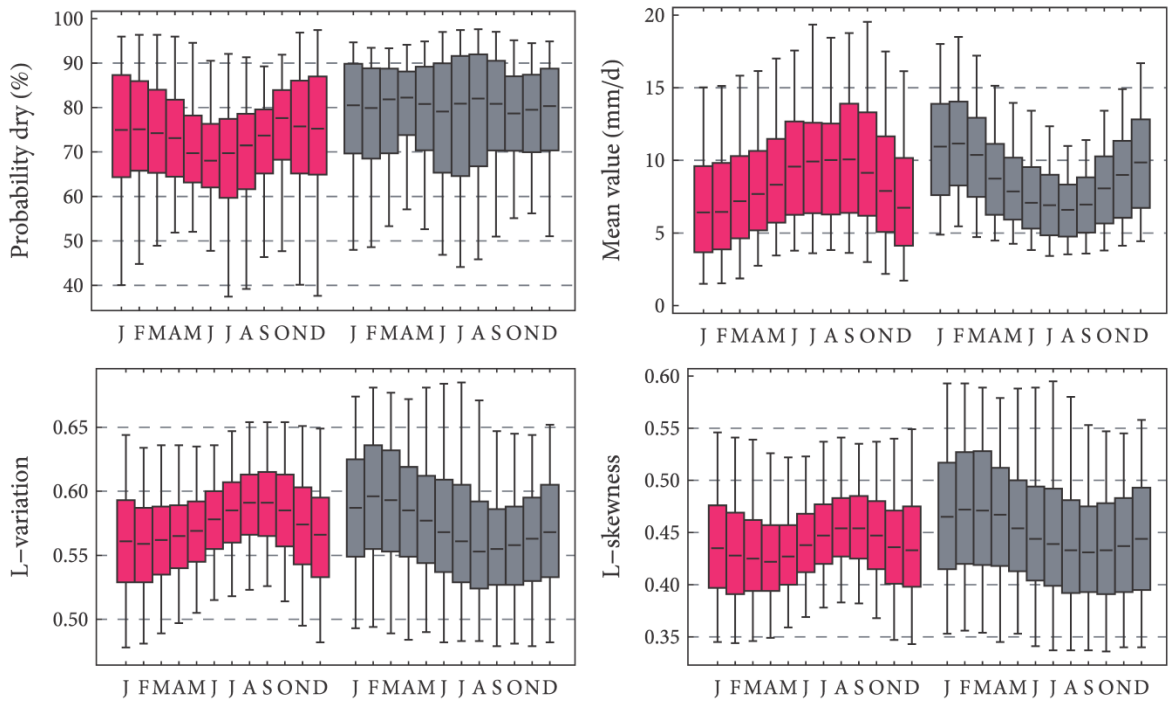



**Figure 3.2.** Estimated statistics of the monthly daily records analysed; red box plots on the left are for the NH; grey boxplots on the right are for the SH; outer fences indicate the 90% ECI.

As Figure 3.2 shows, the probability dry in NH exhibits the typical anticipated behaviour, i.e., dry summer months and wet winter months. Particularly, the median of each box plot exhibits a sinusoidal-like variation, so it seems that most stations in NH have this pattern. Surprisingly, the corresponding pattern in SH is not clear at all; a focus on the median does not reveal, although it resembles a sinusoidal-like function, the familiar or the anticipated behaviour as the median has three "local" peaks, i.e., in January, April and August. It is also noted that the IQR seems to vary irregularly and does not follow the variation of the median. Of course, this does not imply the absence of seasonality in probability dry in the SH, as this result can easily emerge assuming several different

patterns for the studied stations. Also, it is interesting that the variation of the median in both hemispheres is not very large, especially in the SH, yet the range of the 90% ECI is very wide expressing the large variation of probability dry around the world.

The mean value of the nonzero rainfall in both hemispheres, as Figure 3.2 shows, exhibits a clear seasonal pattern, which reminds that of the surface temperature. Specifically, NH and SH show essentially a contrasting behaviour to each other, yet in terms of seasons the behaviour is the same, i.e., the warm months in both hemispheres are those with the highest average nonzero daily rainfall. This behaviour though is not in full correspondence as in NH the minimum and the maximum mean values (comparing the medians) are, respectively, in January and in September, while the corresponding values in the SH are observed, respectively, in August and in February. Remarkably, for the NH the average nonzero daily rainfall pattern is in contrast with probability dry implying greater rainfall depths in rainy days of dry months than of wet months. Yet this is not absolutely precise as the driest moths are from June to August while those with the highest average of nonzero daily rainfall are from July to September; additionally, the lowest value in probability dry is in July while the peak average value is in September. This contrast seems not to be valid for the SH as the probability dry exhibits an irregular pattern.

Figure 3.2 also reveals a marked monthly variation pattern for L-variation and L-skewness. Similarly to the average of nonzero daily rainfall, both statistics exhibit a contrasting behaviour between the two hemispheres; but again, comparing the medians, high and low values are observed, respectively, at warm and cold months. A comparison between the two shape statistics shows that L-variation and L-skewness in SH show an almost identical pattern with the only difference being in the lowest value which is observed one month later for L-skewness. Additionally, L-variation in NH takes its lower values around February while L-skewness around April. Generally, the monthly variation of both statistics (based on their medians) is small, i.e., in both hemispheres L-variation and L-skewness range, respectively, from 0.55 to 0.6 and from 0.42 to 0.47. However, the IQR or the 90% ECI is much wider in the SH compared to NH. The comparison of the shape statistics with the mean value of daily rainfall indicates an agreement in the general pattern in SH, while in NH especially for L-skewness the difference in the patterns is significant.

### 3.3.3 A simple test to identify seasonal variation

All previous comparisons based on the monthly box plots of the statistics indicate clear seasonal variation patterns; a surprising exception is the probability dry of the SH.

Nevertheless, both the IQR and the 90% ECI of all those statistics are much wider allowing at least theoretically a portion of the stations studied to have different patterns than the characteristic one indicated by the medians in Figure 3.2.

As mentioned, it is intuitively anticipated some characteristics of daily rainfall like the probability dry to vary with season, yet this it is not self-evident, e.g., for distributional shape measures like L-variation and L-skewness. When dealing with a small number of records it is relatively easy to assess if a statistic varies with season using simple means, e.g., a plot of the statistic *vs.* month would reveal the variation pattern. Yet when dealing with thousands of stations, an "eyeball" technique would be insufficient or even subjective. For this reason a simple test is formed here to assess and quantify the seasonal variation of the various statistics investigated.

Seasonal variation evokes sinusoidal-like functions; however, even if a statistic is expected to obey a sinusoidal-like law, its sample counterpart may deviate significantly from the anticipated law due to sample variability commonly caused either by sampling uncertainty, particularly for small samples, or by non-robust estimators, or even from local weather characteristics modifying the expected behaviour in some months. This implies that a precise sinusoidal variation may not be common to observe and thus a test based on these characteristics would be inflexible and probably with doubtful efficacy. For this reason, a non-parametric test is proposed allowing for the statistic under investigation to deviate from the exact sinusoidal form.

The seasonal variation test (SV-Test) is described in the following steps: (a) the desired statistic is calculated for each month, (b) the numbers 1 and −1 are assigned, respectively, to monthly values smaller and larger than the median of all months (c) this sequence is rotated until the first and the last value have different signs, (d) this sequence is split into sub-sequences consisting of identical-value runs (SIVR), (e) the number of SIVR is calculated. It is noted that due to step (c) the number of feasible SIVR that a sequence consisting 1 and −1 can be split is 2, 4, 6, 8, 10 or 12; an odd number of SIVR indicates that the first and the last value have the same sign and thus step (c) can be applied; also, step (c) ensures that the resulting number of SIVR is the minimum.

The resulting number of SIVR quantifies seasonality. If the considered statistic exhibits a sinusoidal-like seasonal variation the SV-Test will result exactly in two SIVR. Figure 3.3 depicts an explanatory sketch of the SV-Test showing the monthly values of a statistic after rotation so that the first and the last value are in opposite sides of the median; even though the statistic does not resemble exactly a sinusoidal law, the application of the

test results in two SIVR revealing the seasonality that is visually apparent. It should also be expected that four SIVR still reveal seasonal variation as they could easily emerge if the statistic's sample estimates are sensitive, e.g., if the December's value in the graph of Figure 3.3 was above the median, then four SIVR would result. It seems reasonable to assume that a larger resulting number of SIVR indicates random variation or a variation that does not resemble the "familiar" seasonal variation.

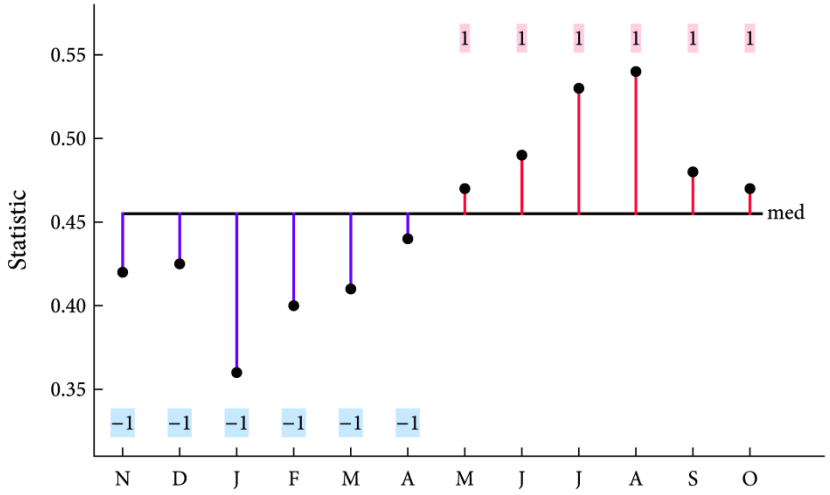



**Figure 3.3.** Explanatory sketch of the seasonal variation test; values above and below the median are denoted, respectively, with 1 and −1.

One could argue that the previous interpretation of the resulting number of SIVR is subjective, e.g., it could be assumed that two or four SIVR could easily emerge even if there is no seasonal variation due to randomness. Thus, in order to make the SV-Test complete benchmark values are necessary for reference and comparison. The idea is to find the probability for each feasible number of SIVR to emerge in the case where the variation of a statistic is random. Theoretically, this problem can be solved analytically using combinatorics, yet it is not that easy; in contrast a Monte Carlo approach can easily provide the answer. In this direction, a Monte Carlo simulation is performed summarized in three simple steps: (a) generation of $10^6$ samples consisting of 12 random numbers each, (b) application of the SV-Test to estimate the resulting number of SIVR for each sample, and (c) estimation of the probability for each feasible number of SIVR as the ratio of the times that this number of SIVR emerged to total number of samples ($10^6$).

The results are graphically depicted in Figure 3.4 where the first number above the bars indicates the probability for a specific SIVR number to occur and the second number above the bars indicates the cumulative probability, e.g., the probability for up to four SIVR to occur is 17.6%. Accordingly, if a statistic varies randomly the probability for two SIVR is

only 1.3% and for four is 16.3%, while the most probable numbers of SIVR are six and eight with probabilities 43.3% and 32.5%, respectively. This implies that if the studied statistic does not exhibit seasonal variation then application of the test will result in more than two SIVR with probability 98.7% and in more than four SIVR with probability 82.4%, and thus, it can be safely assumed that not only two but also four SIVR indicate seasonal variation.

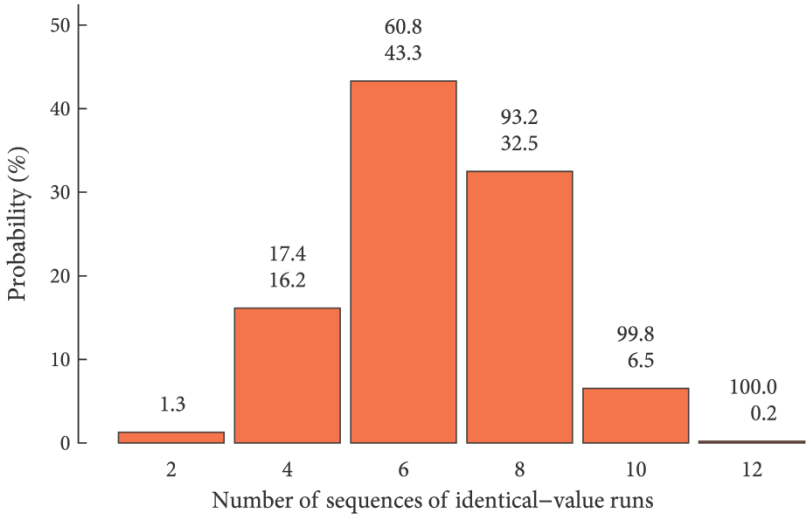



**Figure 3.4.** Benchmark values for the SV-Test; the bars indicate the probabilities (the upper number is cumulative) corresponding to specific number of SIVR in the case of 12 randomly generated numbers (no seasonality).

### 3.3.4 Application of the test

The SV-Test was applied for each station and for the four aforementioned statistics with the results presented in Figure 3.5. The SV-Test verifies, as Figure 3.5a shows, that indeed probability dry exhibits seasonal variation with 64.1% of the stations resulting in two SIVR and with only 4.9% of the stations resulting in more than four SIVR indicating random variation. Similar results are obtained for the mean value of the nonzero daily rainfall, given in Figure 3.5b, with only 8.3% of the stations resulting in more than four SIVR.

The results of the SV-Test regarding the shape characteristics of the nonzero daily rainfall, i.e., the L-variation and the L-skewness are depicted, respectively, in Figure 3.5c and Figure 3.5d. The first observed is that the profile of the two graphs is completely different from the "benchmark" graph describing the random case in Figure 3.3; however, the results are not as clear as for the probability dry or for the mean value case. It is observed that the most common SIVR number is four, both for L-variation and for L-skewness, with 36.9% and 34.5%, respectively. Nevertheless, two or four SIVR (numbers indicating seasonal variation) emerge at 66.2% of stations for L-variation and at 54.5% of

stations for L-skewness, while the corresponding value for the random case is much smaller, i.e., 17.6%. Additionally, two SIVR are observed in 29.3% and 19.7% of the records for L-variation and L-skewness, respectively. These percentages are much larger than 1.3%, which corresponds to the random case. Finally, the seasonality signal is it is much stronger for L-variation than for L-skewness, a difference that may attributed in the fact that estimation of L-variation is more robust than L-skewness.

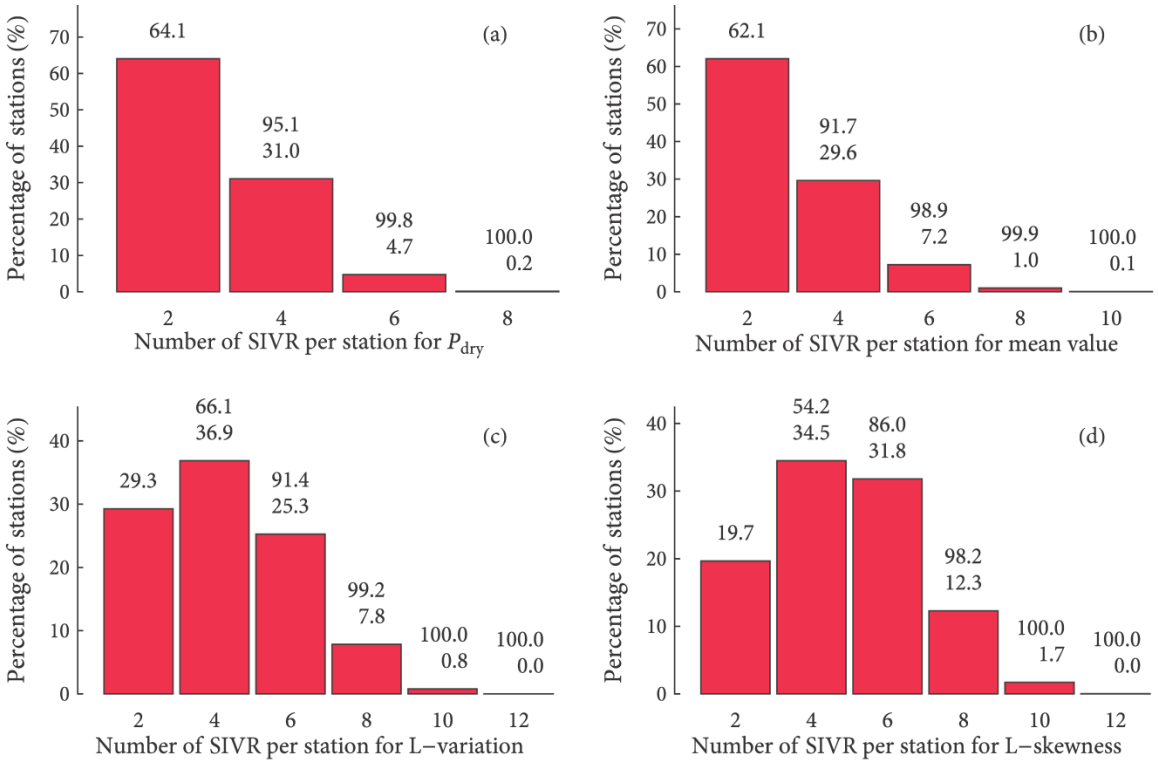



**Figure 3.5.** Results of the SV-Test applied to: (a) the probability dry, (b) mean value (c) L-variation and (d) L-skewness.

### 3.3.5 Why and how much statistics vary?

The analysis of the statists by hemisphere as well as the results of the SV-Test revealed that seasonal variation occurs not only in probability dry and in the mean value of nonzero rainfall but also in the shape characteristics. This implies that the marginal distribution varies over the months, yet the mechanism of this variation is not clear. Particularly, different aspects of the rainfall process are interrelated. For example, the distributional shape variation may be affected by seasonal variation of the average storm duration. To clarify by an example, let us consider the random variables $X$ and $Y$ representing, respectively, the amount of nonzero rainfall at the daily and at a much finer time scale, e.g., the one-minute scale, and let us assume that the marginal distribution of $Y$ does not have seasonal variation; then the distribution function of $X$ emerges by the $n$-term sum of $Y$

variables where $n$ corresponds to the storm duration in minutes in that particular day. Cleary, if the average storm duration varies per month, then the "average" $n$-term sum will vary too and hence the distribution of $X$. This issue raised can only be answered by an analysis of fine temporal scale data which is not the subject of this particular study.

In order to quantify the seasonal variation of the studied statistics per station, four difference measures relative to the statistic's average value of all months are defined. These measures are illustrated in the sketch of Figure 3.6 depicting the monthly variation of a statistic. Particularly, the $i$-th monthly difference $D_i = V_i - \mu$ is defined as the difference between the $i$-th month statistic's value $V_i$ and the average of all $V_i$ denoted as $\mu$. Negative differences (blue lines in the graph) are denoted with $D_{\mathrm{N}}$ and their average with $\overline{D}_{\mathrm{N}}$; likewise, $D_{\mathrm{P}}$ denotes positive differences (red lines in the graph) and $\overline{D}_{\mathrm{P}}$ denotes their average. Additionally, $D_{\min}$ and $D_{\max}$ denote, respectively, the minimum and the maximum difference with reference to $\mu$. Note that this analysis in performed for each individual station and does not provide any comparison between different stations.

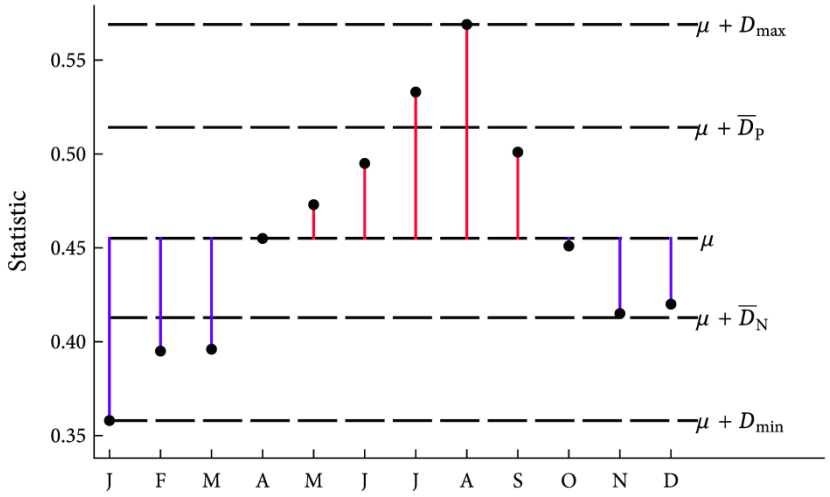



**Figure 3.6.** Explanatory sketch of the four difference measures studied.

The difference measures $\overline{D}_{\mathrm{N}}$, $\overline{D}_{\mathrm{P}}$, $D_{\min}$ and $D_{\max}$ are calculated in terms of percentage change (PC) in respect to the average $\mu$, i.e., $\mathrm{PC} = 100\, D / \mu$ with $D$ being any of the four difference measures. The first two measures can be interpreted as the "expected" or the average negative or positive percentage change in reference to the monthly average while the latter two indicate the minimum or maximum percentage change in reference to the monthly average. The percentage change of these measures was calculated for each station and for the four statistics studied. The results are given in Figure 3.7 in the form of box

plots (note that the PC of the negative differences $\overline{D}_N$ and $D_{min}$ is given in absolute values for better presentation).

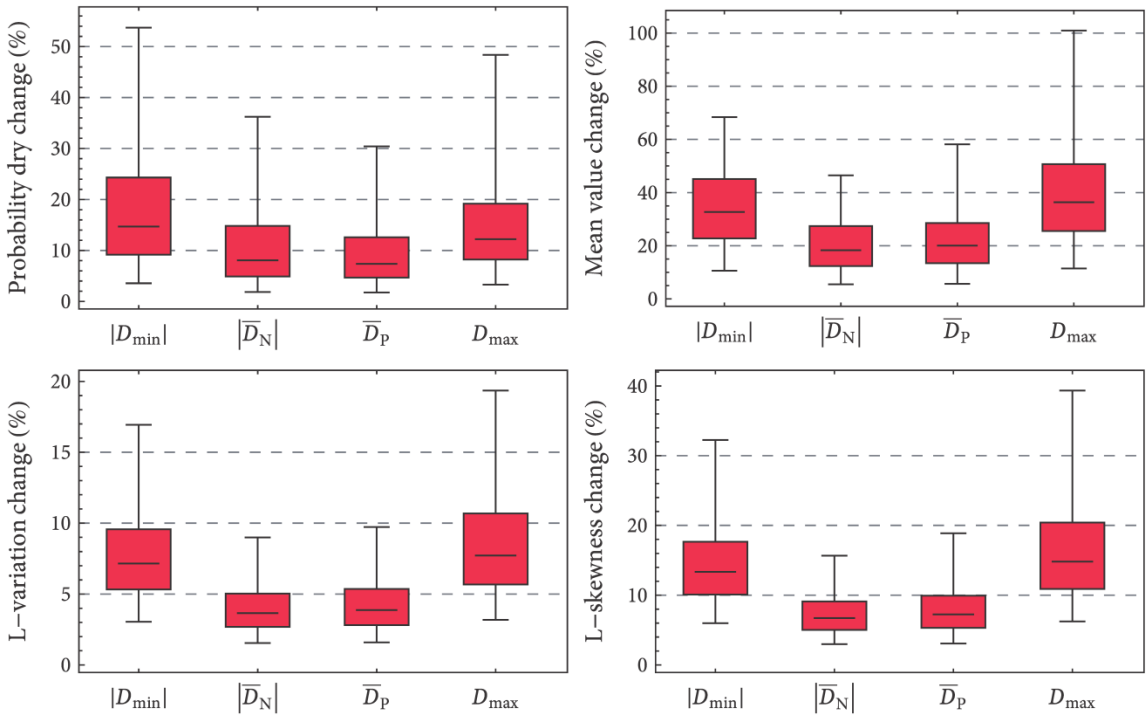



**Figure 3.7.** Box plots depicting the percentage change of the difference measures relative to the average of all months for the four statistics studied. Each box plot is constructed by the values determined from the stations studied. Outer fences indicate the 95% ECI.

A first look in the box plots indicates that the largest monthly variation is observed in the mean value of the nonzero rainfall, followed by the probability dry, next by L-skewness and last by L-variation exhibiting the lowest variability. Particularly, the IQR of the nonzero rainfall mean value, which represents the 50% of the central values, for $D_{min}$ and $D_{max}$ ranges, respectively, from −45.2% to −22.8% and from 25.5% to 50.6%; these values indicate a large variability around the average. These ranges are lower for the probability dry where the IQR of $D_{min}$ and $D_{max}$ ranges, respectively, from −24.3% to −9.2%) and from 8.2% to 19.2%. Regarding L-skewness it is observed that 75% of the records have percentage change of $D_{min}$ and $D_{max}$ less than −17.7% and 20.4%, respectively, while the corresponding percentages for the L-variation are −9.5% and 10.7%. Comparing the box plots of the distributional shape measures, i.e., the L-variation and L-skewness, with the box plots of the probability dry and of the mean value it is observed that in the first two cases $\overline{D}_N$ and $\overline{D}_P$ vary at a lower level relative to $D_{min}$ and $D_{max}$ than in the former two

cases. This may indicate that the "expected" difference from the monthly average, expressed by $\overline{D}_{\mathrm{N}}$ and $\overline{D}_{\mathrm{P}}$, for L-variation and L-skewness for most of the months is "small"; yet the "extreme" differences, expressed by $D_{\min}$ and $D_{\max}$, are relatively large; or else, this indicates that the marginal distribution of nonzero daily rainfall for most of the months does not vary much in terms of shape.

## 3.4 In search for the "universal" rainfall model

### 3.4.1 Candidate models

The shape characteristics of nonzero daily rainfall, as empirical evidence suggests, vary not only with location but also by month; this implies that the consistent probabilistic modelling of nonzero daily rainfall demands different models for different areas and possibly for different months. So it would be of paramount importance if a single parametric distribution can be used for nonzero daily rainfall for all months and for the whole world. The fact that distributional shape varies excludes, in principle, distributions with fixed shape, thus favouring those with great shape flexibility. Additionally, it is reasonable to assume that a competitive model should also be physically consistent with rainfall, i.e., defined in the positive real axis, and if possible having a theoretical basis. In this direction, in a previous study [*Papalexiou and Koutsoyiannis*, 2012] the principle of maximum entropy was used to derive consistent distributions for geophysical random variables. These entropy derived distribution were tested in their ability to describe the nonzero daily rainfall (but not in a monthly basis) using more than 10 000 stations with very good results.

The distributions derived in the aforementioned study, and also used here are the Burr type XII distribution (BrXII) [*Burr*, 1942; *Tadikamalla*, 1980] and the Generalized Gamma distribution (GG) [*Stacy*, 1962]. Their probability density functions are given, respectively, by

$$f_{\mathrm{BrXII}}(x) = \frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\left(1+\gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right)^{-\frac{1}{\gamma_1\gamma_2}-1} \qquad x \geq 0 \tag{3.1}$$

$$f_{\mathrm{GG}}(x) = \frac{\gamma_2}{\beta\Gamma(\gamma_1/\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma_2}\right) \qquad x \geq 0 \tag{3.2}$$

Note that the parameterization used here for the BrXII is different from the most typical found in the literature; first, it clearly shows its asymptotic behaviour (for $\gamma_2 \to 0$ the Weibull distribution emerges) and second, the two shape parameters are directly related to each of the distribution tails (left and right). Regarding the parameterization of GG distribution it is mentioned that other forms also exist but this is one of the commonly used.

Both distributions are very flexible, each comprising one scale parameter $\beta > 0$, and two shape parameters. The shape parameter $\gamma_1 > 0$ controls the behaviour of the left tail, i.e., for $\gamma_1 < 1$ the distributions are J-shaped while for $\gamma_1 > 1$ they are bell-shaped; the parameter $\gamma_2 > 0$ controls the asymptotic behaviour of the right tail, i.e., the "heaviness" of tail and thus the frequency and the magnitude of extreme events. It is noted that although these two distributions have a structural similarity in terms of their parameters, in principle, they differ, i.e., the BrXII distribution is a power-type distribution having finite moments up to order $1/\gamma_2$ while the GG distribution is of exponential form with all of its moments finite. Some well-known special cases worth mentioning for the BrXII distribution are the Pareto type II and the Weibull distributions (limiting case), while for the GG distribution, special cases are the Weibull, the Gamma and the Exponential distributions.

### 3.4.2 A first approach based on L-moments

There are some useful graphical tools, especially when dealing with a large number of records, which help to provide an overall and general picture of the studied variable from a statistical point-of-view. Such a tool for identifying suitable distributions for the variable under investigation is the L-moments ratio diagram [see e.g., *Vogel and Fennessey*, 1993; *Peel et al.*, 2001]. Essentially, this diagram provides a comparison between observed statistics calculated from the records and the theoretical ones emerging by the distribution under investigation. Practically, any pair of L-ratios could be used to form an L-ratio diagram; yet the most common pairs are the L-skewness *vs*. L-variation or the L-kurtosis *vs*. L-skewness, with the latter being more popular in the literature as L-variation is not well defined for some distributions, e.g., for distributions with mean value zero or negative. Nevertheless, as noted, L-variation is well defined for positive random variables and is more robust than L-kurtosis.

L-ratios as functions of the distribution's shape parameters are essentially measures of shape. Thus, in an L-ratio diagram a distribution with none, one or two shape parameters forms, respectively, a point, a line or an area. Consequently, the

aforementioned distributions, in any L-ratio diagram, form an area (denoted as L-area) whose extent is finite (does not cover the entire plane). Here the L-skewness *vs*. L-variation diagram is used aiming to form the theoretical L-area of the BrXII and the GG distributions and calculate the percentage of the observed L-points that lie within the L-area of each distribution and for each month. An observed point that lies within the distribution's theoretical L-area implies that specific parameter values exist so the distribution can reproduce the first three L-moments. Practically, the theoretical L-area of a distribution is formed using equations of $\tau_2$ and $\tau_3$. Unfortunately, analytical L-moment expressions for the GG distribution do not exist; exception is the first L-moment (identical with the mean value) and is given by

$$\lambda_1 = \beta\Gamma\left(\frac{1+\gamma_1}{\gamma_2}\right)/\Gamma\left(\frac{\gamma_1}{\gamma_2}\right) \tag{3.3}$$

where $\Gamma(a)=\int_0^\infty t^{a-1}\exp(-t)\mathrm{d}t$ is the Gamma function. In contrast, solving the L-moments definition integrals [see e.g., *Hosking*, 1990] for the BrXII distribution results in the following expressions:

$$\lambda_1 = \frac{\beta\gamma_2^{-1/\gamma_1}}{\gamma_1}\mathrm{B}\left(\frac{1}{\gamma_1},\frac{1-\gamma_2}{\gamma_1\gamma_2}\right) \tag{3.4}$$

$$\tau_2 = 1-\mathrm{B}\left(\frac{1}{\gamma_1},\frac{2-\gamma_2}{\gamma_1\gamma_2}\right)/\mathrm{B}\left(\frac{1}{\gamma_1},\frac{1-\gamma_2}{\gamma_1\gamma_2}\right) \tag{3.5}$$

$$\tau_3 = 1-2\frac{\mathrm{B}\left(\frac{1}{\gamma_1},\frac{2-\gamma_2}{\gamma_1\gamma_2}\right)-\mathrm{B}\left(\frac{1}{\gamma_1},\frac{3-\gamma_2}{\gamma_1\gamma_2}\right)}{\mathrm{B}\left(\frac{1}{\gamma_1},\frac{1-\gamma_2}{\gamma_1\gamma_2}\right)-\mathrm{B}\left(\frac{1}{\gamma_1},\frac{2-\gamma_2}{\gamma_1\gamma_2}\right)} \tag{3.6}$$

where $\mathrm{B}(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the Beta function. The two parametric equations $\tau_i = g_i(\gamma_1,\gamma_2)$ given in Eq. (3.5) and Eq. (3.6) can be used to implicitly determine the L-area. Functions of this form, and in this particular case, can be easily plotted by fixing one parameter to a specific value, varying the other in a dense grid and plotting the resulting ($\tau_2$, $\tau_3$) points. The method for determining the theoretical L-area covered by the GG

distribution is exactly the same, with the only difference that ($\tau_2$, $\tau_3$) points are calculated by the numerical integration of the L-moments integrals.

The theoretical BrXII and GG L-areas are depicted in Figure 3.8, with several fixed-value parameter lines also plotted. For the BrXII distribution values ranging from 1 to 10 (lower bound) denote fixed $\gamma_1$ parameter values while those ranging from 0.1 to 0.9 (upper bound) denote fixed $\gamma_2$ parameter values. Similarly, for the GG distribution values ranging from 0.5 to 6 (lower bound) denote fixed $\gamma_1$ parameter values while those ranging from 0.5 to 10 (within the area) denote fixed $\gamma_2$ parameter values. The observed L-points of the nonzero daily rainfall for the month of January are also shown in Figure 3.8, superimposed over the L-areas (graphs for individual months as well as for the nonzero daily rainfall of all months are given in Appendix C). At each plot empirical points are colored in three ways; the red-colored points lie outside the area; the dark-colored indicate a Bell-shaped distribution; the light-colored indicate a J-shaped distribution. Interestingly, the GG and the BrXII distributions are complementary in the sense that the observed L-points not belonging to one's area belong to the other's, implying that just these two distributions can describe all records analysed here. Note that both distributions are special cases of the Generalized Beta of the second kind distribution [see e.g., *Mielke Jr and Johnson*, 1974; *Papalexiou and Koutsoyiannis*, 2012], but this distribution is more complicated as it comprises one scale and three shape parameters.

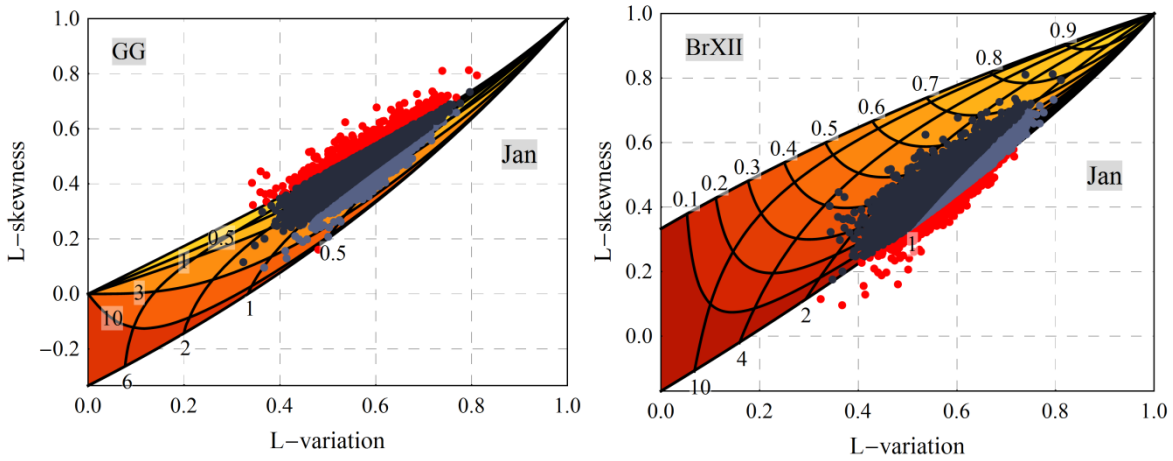



**Figure 3.8.** Observed L-points for the month of January of the 14 157 daily rainfall records studied in comparison to the theoretical L-areas of (a) the BrXII distribution and (b) the GG distribution. Red-colored L-points lie outside the L-area; dark-colored indicate a Bell-shaped distribution; light-colored indicate a J-shaped distribution.

Particularly, Figure 3.9 shows the estimated percentages of the observed L-points of monthly daily rainfall lying within the area as well as the percentages of J- and Bell-shaped distributions that would emerge if the distributions were actually fitted. It is apparent that both distributions, especially the GG distribution, perform very well. For example, the GG distribution describes 99.2% of the observed L-points for the values of all months, while the lowest percentage, observed in January, remains very high, i.e., 94.2%. The BrXII distribution also performs well by managing to describe 90.0% of the observed L-points for the values of all months and with its lowest percentage observed in May with 81.0%. It is noted that the actual percentages of the observed points that lie within the theoretical areas are expected to be even higher if larger samples were available.

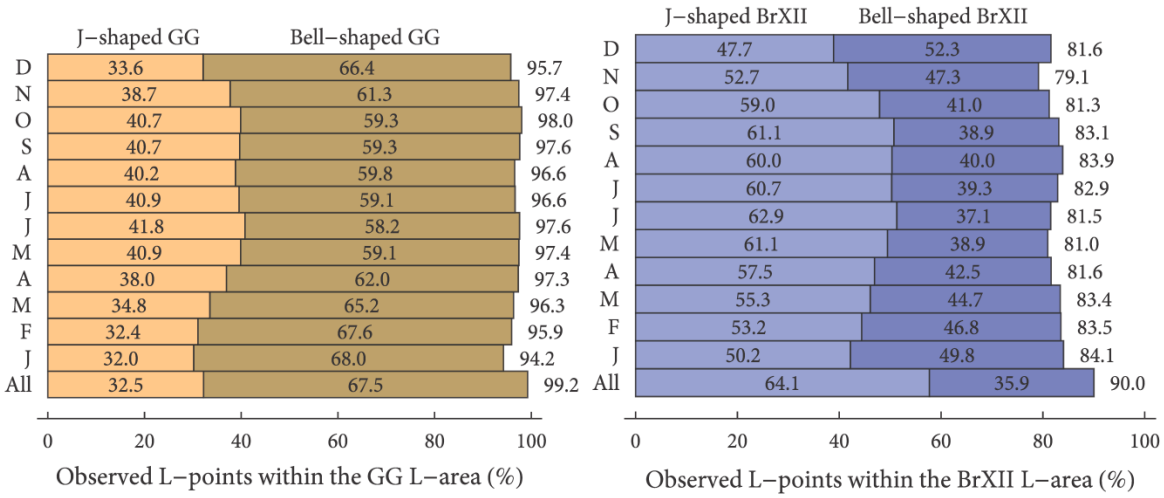



**Figure 3.9.** Percentage of empirical L-points lying within the L-areas of the GG and the BrXII distributions.

Clearly, the variability of the statistics decreases with increasing sample size and thus many points that lie outside the area actually would not if the sample was larger. Actually, this is the reason why the percentage of the observed L-points for the values of all months is higher than those of individual months. Finally, it may seem peculiar that the percentages of J-shaped GG distributions are significantly lower (almost half) compared to those of the BrXII distributions. This implies that for the same record a J- and a Bell-shaped distribution may be fitted equally well in terms of L-moments. Note that a density function $f(x)$ is called J-shaped if the value of $f(x)$ at its lower bound (zero for positive random variables) is the maximum, i.e., $f(0) = \max\big(f(x)\big)$; otherwise, the distribution is called Bell-shaped. This simple criterion may however be meaningless in several practical situations, e.g., two GG distributions with $\gamma_1$ values a little less and a little more than 1

would be characterized, respectively, as J- and Bell-shaped, yet apart from this difference they are almost identical.

The previous analysis gave a clear indication that both the GG and the BrXII distributions are very good models for describing rainfall. Yet an important and more specific question that naturally arises is if a single distribution can be used to describe all months within the same station; in order to answer this question an analysis by record has to be performed. To clarify, each record has 12 L-points, one for each month, so the idea is to estimate the number of monthly L-points per station that lie within the theoretical L-area. For example, if all monthly points of a station lie within the distribution's area, then this distribution could be used for all months in this particular station. The results are shown in Figure 3.10. Evidently, in this test the GG distribution performs much better than the BrXII, as it can be used as an all-month model for 78.8% of the stations, a percentage almost double than the corresponding one to the BrXII distribution which is 43.2%. Additionally, the percentage of record in which the GG distribution is suitable for more than ten months is very high, i.e., 95.6% while the corresponding one for the BrXII it has significantly increased to 69.5%.

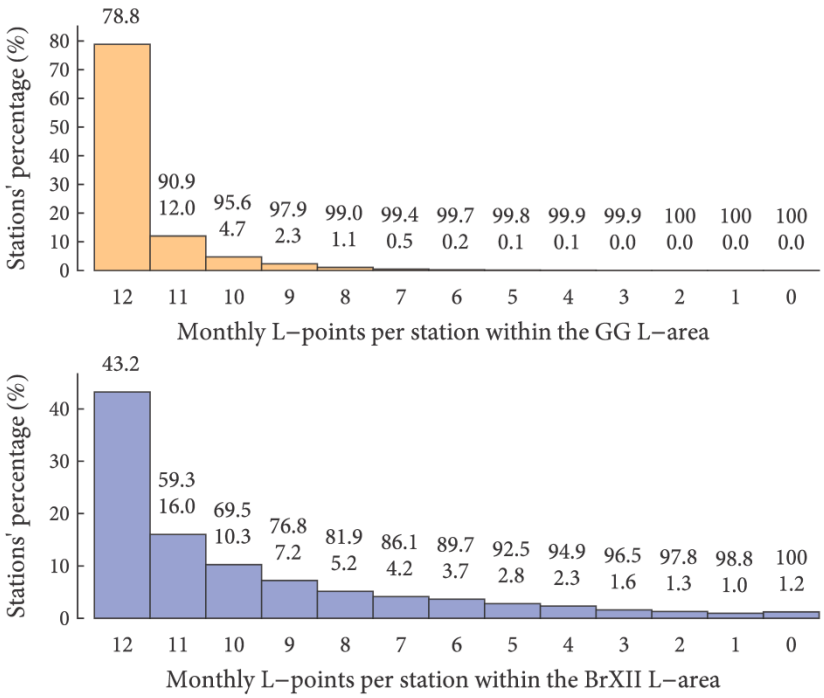



**Figure 3.10.** Percentage of records *vs.* the number of monthly L-points per station lying within the theoretical L-areas of the GG and the BrXII distributions.

### 3.4.3 The actual fitting

The previous analysis showed that both distributions can describe a very large percentage of the records in terms of the first three L-moments. Additionally, it is very important to study the actual values of the shape parameters, especially of the parameter $\gamma_2$ as it controls the extreme behaviour. As noted though, the GG distribution does not have analytical L-moments equations while in the BrXII case, where analytical formulas exist, the resulting system of equations between theoretical and sample estimates can only be solved numerically. So it is clear that explicit functions, easily applicable, of the form $\theta = g(\lambda_1,\tau_2,\tau_3)$ that relate any of the distribution's parameter $\theta$ with the first three L-moments measures cannot be formed.

To overcome this issue and in order to create an accurate and fast fitting method for both distributions, based on L-moments, a solution is inspired by the way engineers and statisticians used to practice in the past (or even at present) using the "good-old" graphical tools (e.g., nomograms). For example, the shape parameters $\gamma_1$ and $\gamma_2$ can be approximately estimated by placing an observed $(\tau_2,\tau_3)$ point within the L-ratio diagram in Figure 3.8 and do an "eyeball" linear regression using the nearest fixed-value parameter lines surrounding the observed point. Essentially, our approach is an accurate and computerized version of this technique, i.e., the algorithmic "translation" of a $(\tau_2,\tau_3)$ point to a $(\gamma_1,\gamma_2)$ point. The basic idea is to "replace" the initial functions of L-variation and L-skewness, which are highly nonlinear and without analytical expressions in the GG case, with simple linear interpolation functions that can be more easily handled. First, the $\tau_2 = g_2(\gamma_1,\gamma_2)$ and $\tau_3 = g_3(\gamma_1,\gamma_2)$ are calculated from the initial expressions ($g_2$ and $g_3$ are analytical expressions or integrals numerically estimated) in a very dense and appropriately selected grid of $(\gamma_1,\gamma_2)$ points; and second, from the $(\gamma_1,\gamma_2,\tau_2)$ and $(\gamma_1,\gamma_2,\tau_3)$ points two bivariate linear interpolation functions are formed, i.e., $\tau_2 = h_2(\gamma_1,\gamma_2)$ and $\tau_3 = h_3(\gamma_1,\gamma_2)$ (note that any mathematical software creates easily bivariate interpolation functions). Replacing $\tau_2$ and $\tau_3$ in these equations with their counterpart estimates $\hat{\tau}_2$ and $\hat{\tau}_3$ a square error norm can be formed that can be numerically minimized. Particularly, the estimated shape parameters $\gamma_1$ and $\gamma_2$ are those emerging by the following expression

$$(\gamma_1,\gamma_2) = \underset{\gamma_1,\gamma_2}{\arg\min} \sum_{j=2}^{3} \left( h_j(\gamma_1,\gamma_2) - \hat{\tau}_j \right)^2 \tag{3.7}$$

Once the parameters $\gamma_1$ and $\gamma_2$ are estimated for either distribution the trivial scale parameter $\beta$ can be directly estimated from the corresponding expression of the first L-

moment $\lambda_1$ given in Eq. (3.3) and Eq. (3.4). As a final technical detail it is noted that the fitting method was tested to millions of random points to assess its accuracy and to define the parameters' range where the method works essentially without estimation error. It was observed that for the GG distribution these ranges are $0.2 \le \gamma_1 \le 10$ and $0.1 \le \gamma_2 \le 10$, while for the BrXII distribution they are $0.2 \le \gamma_1 \le 10$ and $0.001 \le \gamma_2 \le 0.9$. If the fitting procedure resulted in parameters outside these ranges it was considered inaccurate.

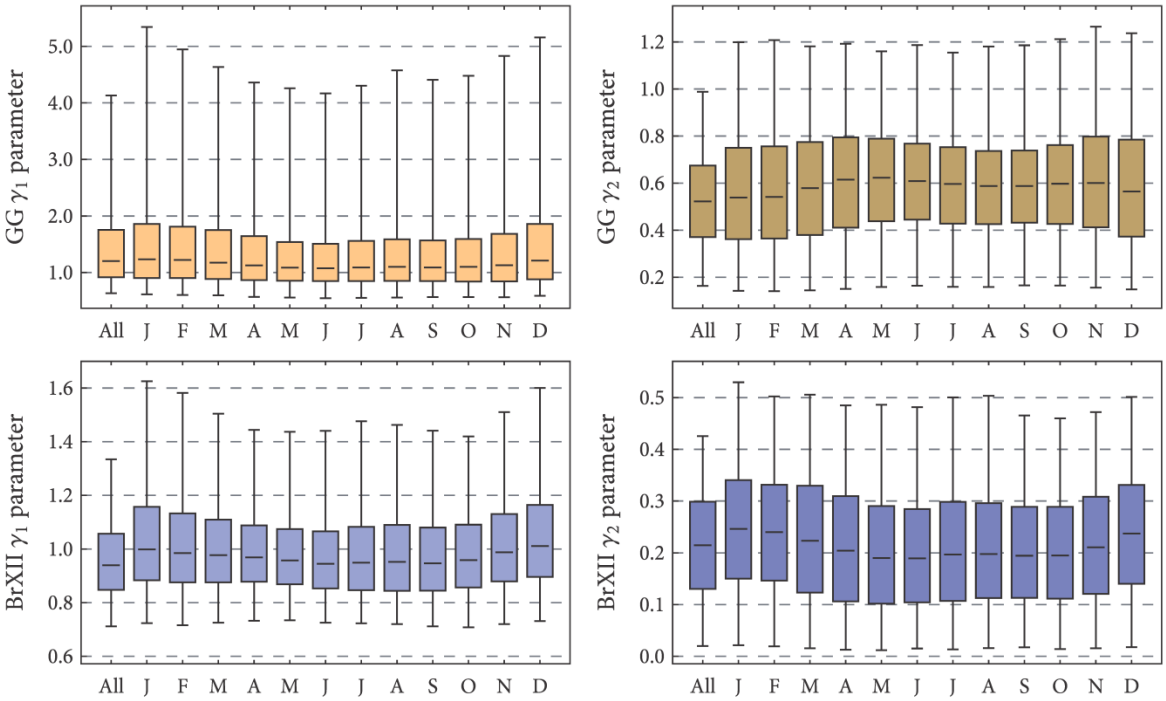



**Figure 3.11.** Estimated shape parameters of the GG and BrXII distributions using the method of L-moments.

The estimated values of the shape parameters for both distributions are presented in the form of box plots in Figure 3.11 while some of their basic summary statistics are given in Appendix C in Table C.1. Considering the theoretical range of the parameters, i.e., (0, ∞), of both parameters and for both distributions it is apparent that they actually vary in a narrow range as the 95% empirical confidence intervals indicate in Figure 3.11 (outer fences of the whiskers). For the GG distribution the median of the parameter $\gamma_1$ for all months ranges from 1.08 to 1.23 while for all month and for most of the records $\gamma_1 > 1$ indicating bell-shaped densities. The average of all monthly medians of the parameter $\gamma_2$ is approximately 0.59 with the majority of records having $\gamma_2 < 1$ indicating a heavier tail than the exponential or the Gamma tail [see also *Papalexiou et al.*, 2013]. The median values of the BrXII $\gamma_1$ parameter for all months are close to 1; actually the average of all monthly

medians is 0.97, a value very close to the Pareto type II value, i.e., $\gamma_1 = 1$. Additionally, it is noted that more than 50% of the records have $\gamma_1 < 1$ indicating J-shape densities and verifying also the results presented in Figure 3.9. Finally, the monthly median values of the $\gamma_2$ parameter vary in a narrow range, i.e., form 0.19 to 0.25, while the upper limit in the 95% ECI is for all months (except January) less than 0.5, indicating finite variance distributions.

### 3.4.4 Performance of the models

The GG distribution as the analysis showed is able to describe more records than the BrXII. Yet as the two distributions differ significantly in the behaviour of the tail, as the former is of exponential form and the latter is power type, it is useful to compare them in terms of some fitting error measures. Obviously, the comparison is possible only for the samples in which both distributions were fitted. For example Figure 3.12 presents a probability plot of the fitted distributions to the (nonzero) daily rainfall values of a station (station code CA006158350). Clearly, both distributions fit well and it is evident that the BrXII distribution has a heavier tail and thus for small exceedance probabilities (large return periods) predicts larger values.

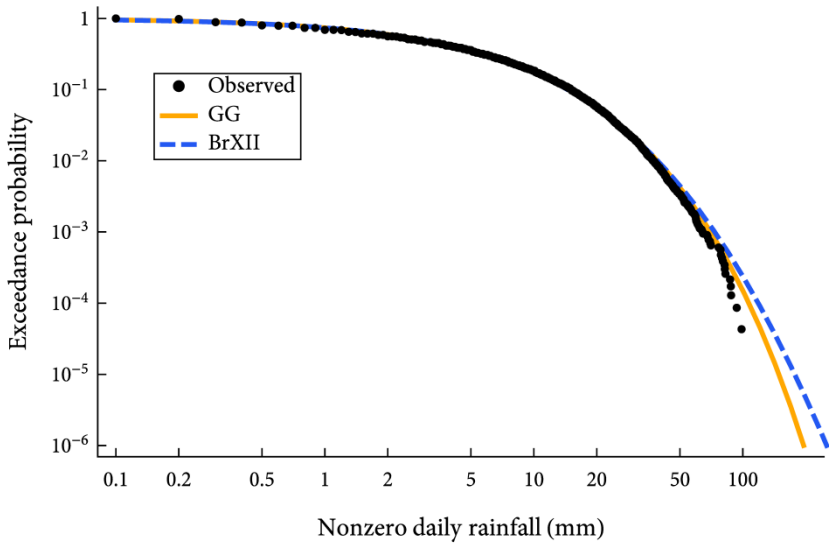



**Figure 3.12.** Probability plot of the fitted distributions to a specific station (station code CA006158350) using the method of L-moments.

In order to evaluate and compare the fitting performance of the distributions the following four error measures are defined:

$$\text{ER-I} = \frac{1}{n}\sum_{i=1}^{n}\left|\Delta x_{(i)}\right| \tag{3.8}$$

$$\text{ER-II} = \frac{1}{m}\sum_{i=n-m+1}^{n}\left|\Delta x_{(i)}\right| \tag{3.9}$$

$$\text{ER-III} = \max\left(\left|\Delta x_{(1)}\right|,\ldots,\left|\Delta x_{(n)}\right|\right) \tag{3.10}$$

$$\text{ER-IV} = \frac{\Delta x_{(n)}}{\hat{x}_{(n)}}100 \tag{3.11}$$

where $\Delta x_{(i)} = x_{(i)} - \hat{x}_{(i)}$ is the difference between the predicted value $x_{(i)}$ and its corresponding observed one $\hat{x}_{(i)}$ with the index $i$ indicating the position in the ordered sample, i.e., $\hat{x}_{(1)} \leq,\ldots,\leq \hat{x}_{(n)}$. The predicted value is estimated by the quantile function of each distribution, i.e., $x_{(i)} = Q_X(p_i)$, using the corresponding empirical probability according to the Weibull plotting position, i.e., $p_i = i/(n+1)$.

Thus, ER-I is the mean value of the absolute differences of all sample values and provides an overall measure of fitting performance; ER-II is focused on the last $m$ largest sample values and may be seen as a fitting measure to the extreme values or to the tail (here $m = 10$); ER-III is the absolute maximum difference identified between observed and predicted values and does not necessarily correspond to the sample's maximum value; ER-IV is focused on the percentage difference between the predicted maximum value and the maximum observed value with negative and positive differences implying, respectively, underestimation or overestimation of the maximum value by the fitted distribution.

The results are presented in Figure 3.13 (box plots of the four error measures for the values of all months) and in Figure 3.14 (box plots for the individual months). Additionally, Table 3.2 shows, for all months and for individual months, the number of records that were actually compared (both distributions fitted) as well as the averages of the error measures.

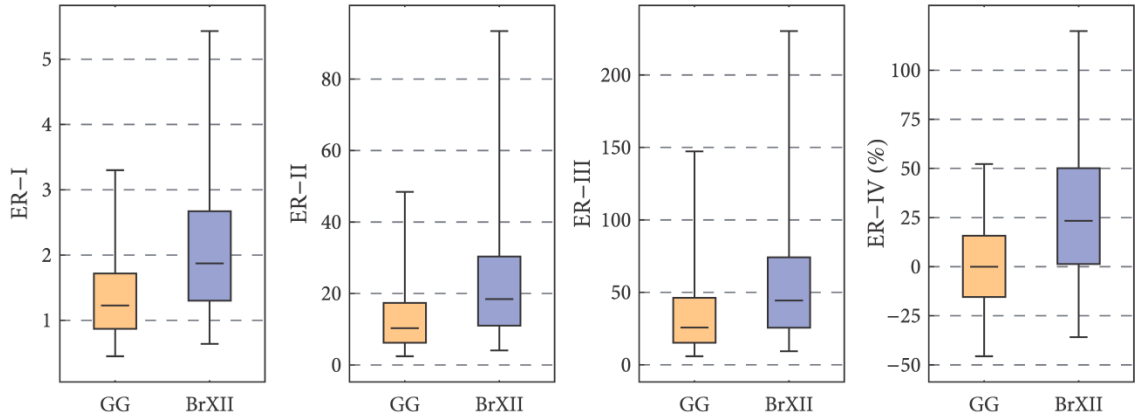



**Figure 3.13.** Box plots of the error measures that evaluate the fitting performance of the GG and BrXII distributions to daily rainfall of all months.

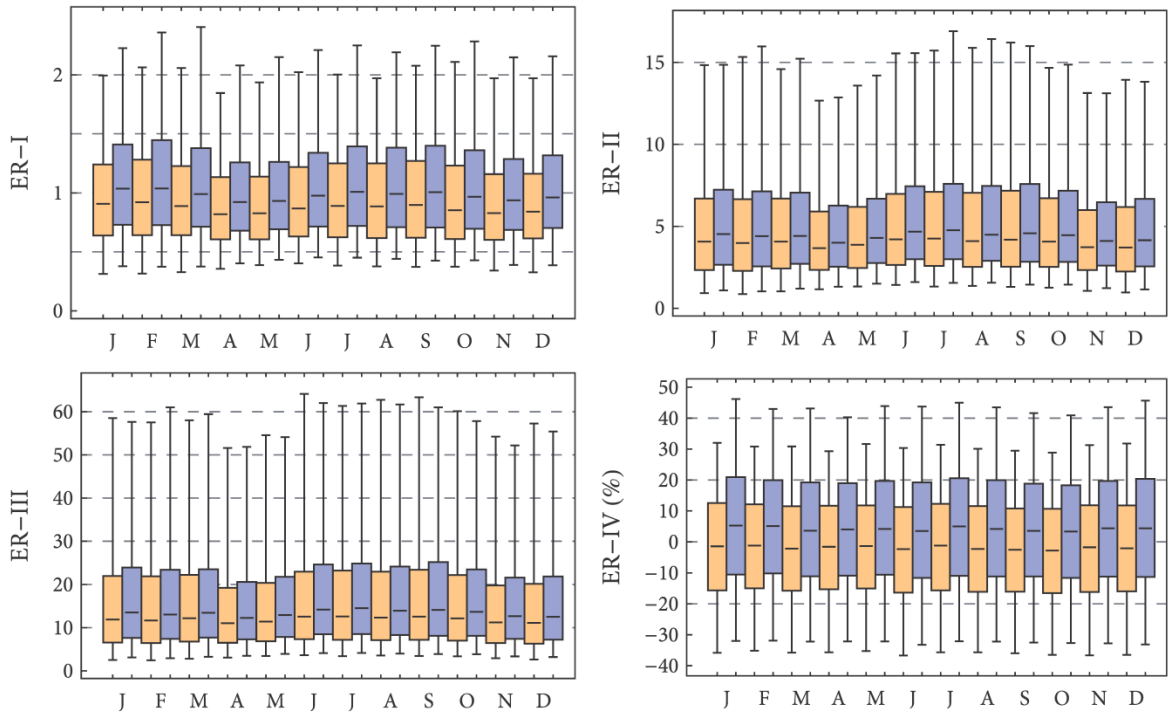



**Figure 3.14.** Box plots of the error measures of the fitting of the GG and BrXII distributions to the monthly daily rainfall records.

In general, as the box plots and the values of Table 3.2 reveal, the GG distribution according to all error measures performs better than the BrXII. A focus on the ER-IV, which estimates the percentage difference between the predicted and the observed maximum value, indicates that the GG distribution performs exceptionally well. For example for all months (Figure 3.13) this estimate is essentially unbiased while the 95% ECI

is between −45.6% and 52.2%; in contrast, the BrXII overestimates the maximum on average 28.2% (see Table 3.2) while the 95% ECI is much wider, i.e., from −35.9% to 120.0%. Yet the performance of the BrXII distribution improves for each specific month separately (Figure 3.14) where the average overestimation per month for the BrXII is 4.7% (estimated form the values of Table 3.2) while the GG distribution underestimates on average the maximum value by −2.2%.

**Table 3.2.** Mean values of the error measures evaluating the fitting performance of the distributions, as well as percentage values of records in which the GG was better fitted compared to Burr XII.

| | All | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Fit No. | 12413 | 10474 | 10684 | 10769 | 10754 | 10750 | 10879 | 10877 | 11041 | 11124 | 10967 | 10457 | 10396 |
| | Mean values of the error measures for the GG distribution | | | | | | | | | | | | |
| ER-I | 1.4 | 1.0 | 1.0 | 1.0 | 0.9 | 0.9 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| ER-II | 14.1 | 5.5 | 5.5 | 5.5 | 4.9 | 5.2 | 5.7 | 5.9 | 5.8 | 5.9 | 5.5 | 5.0 | 5.1 |
| ER-III | 38.2 | 18.9 | 18.6 | 19.0 | 17.0 | 17.8 | 20.2 | 20.1 | 20.0 | 20.3 | 19.5 | 17.5 | 17.8 |
| ER-IV | 0.7 | -1.6 | -1.6 | -2.2 | -2.1 | -1.7 | -2.7 | -1.7 | -2.4 | -2.7 | -3.1 | -2.2 | -2.2 |
| | Mean values of the error measures for Burr XII distribution | | | | | | | | | | | | |
| ER-I | 2.2 | 1.1 | 1.2 | 1.1 | 1.0 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 | 1.1 |
| ER-II | 25.4 | 5.8 | 5.9 | 5.9 | 5.2 | 5.6 | 6.1 | 6.3 | 6.2 | 6.1 | 5.8 | 5.3 | 5.4 |
| ER-III | 62.0 | 19.8 | 19.9 | 20.1 | 17.9 | 18.8 | 21.0 | 21.3 | 20.9 | 20.9 | 20.1 | 18.2 | 18.6 |
| ER-IV | 28.2 | 5.8 | 5.2 | 4.5 | 4.2 | 4.9 | 4.2 | 5.5 | 4.7 | 4.1 | 3.6 | 4.6 | 5.0 |
| | Percentage values that the GG distribution was better fitted compared to Burr XII (%) | | | | | | | | | | | | |
| ER-I | 87.0 | 80.8 | 80.9 | 77.9 | 77.8 | 76.2 | 74.6 | 79.6 | 77.6 | 75.4 | 77.1 | 78.0 | 78.4 |
| ER-II | 79.2 | 65.8 | 66.1 | 62.9 | 62.5 | 63.3 | 61.3 | 65.2 | 63.2 | 59.3 | 60.6 | 63.6 | 64.9 |
| ER-III | 69.5 | 59.9 | 60.2 | 56.6 | 56.4 | 56.8 | 55.1 | 58.7 | 56.8 | 54.1 | 54.1 | 58.3 | 58.6 |
| ER-IV | 67.0 | 55.8 | 55.8 | 53.1 | 53.9 | 53.3 | 52.5 | 55.5 | 54.2 | 52.5 | 51.7 | 54.9 | 55.3 |

Finally, the percentage of the records in which the GG distribution was better fitted according to the four error measures are also given in Table 3.2 while a side-by-side comparison of the two distributions is presented in Figure 3.15. Apparently, the GG distribution performs better especially according to ER-I which evaluates the overall fitting. Comparing the percentages of the two distributions, shown in Figure 3.15, it is observed that the GG distribution improves even more its performance over the BrXII distribution at the daily rainfall compared to the monthly daily rainfall. This might be an extra argument for the GG distribution as the daily rainfall samples are much larger in size than the monthly samples and thus the parameter estimation is more accurate in this case.

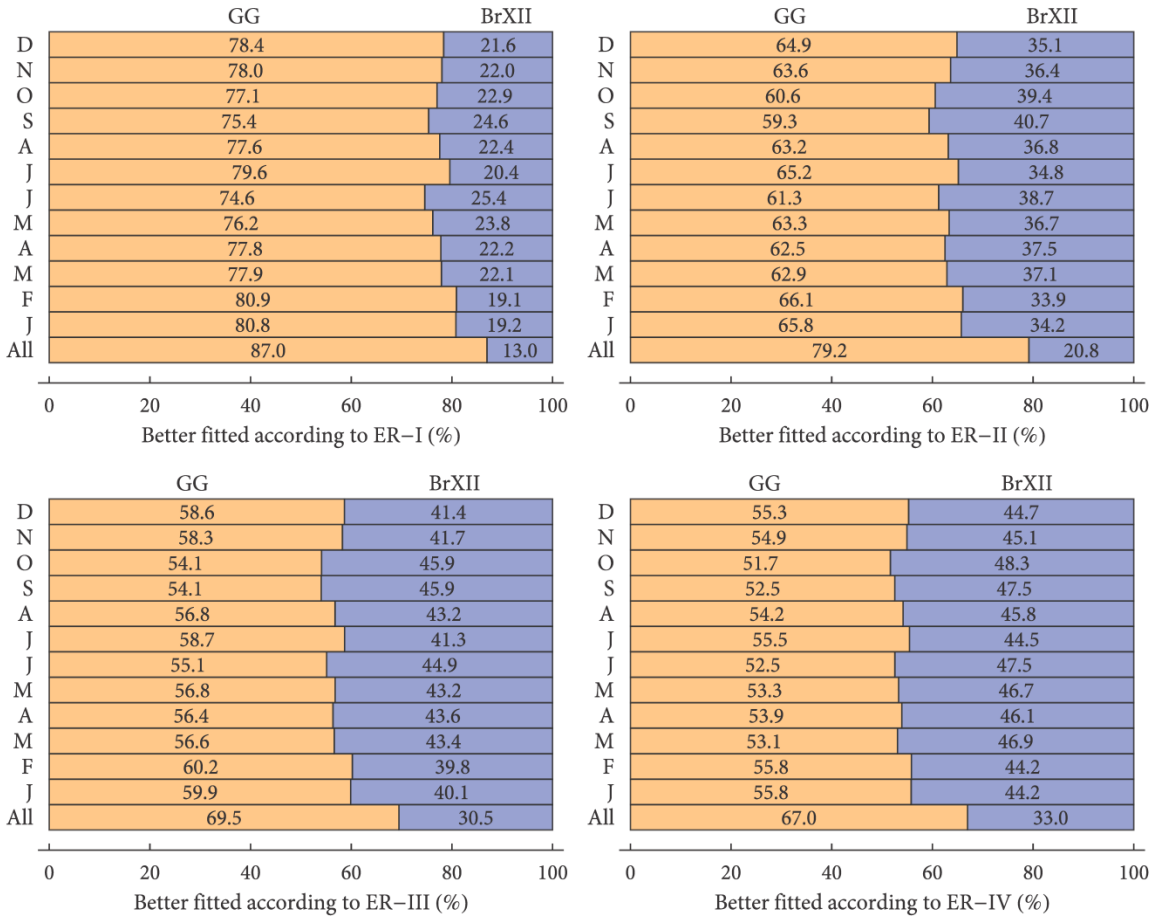



**Figure 3.15.** Comparison of the fitting performance of the two distributions; the values within the bars indicate the percentage of stations in which each distribution was better fitted according to the error measures.

## 3.5 Summary and conclusions

This study investigates the seasonal variation of daily rainfall focusing on the properties of its marginal distribution. The two major questions set are: (a) which statistical characteristics of daily rainfall vary the most over the months and how much, and (b) whether or not there is a relatively simple probability model that can describe the nonzero daily rainfall at every month and every area of the world. In order to treat these questions a massive analysis is performed of more than 170 000 monthly daily rainfall records from more than 14 000 stations from all over the globe.

Regarding the first question, the variation in the two hemispheres of four statistics is investigated; specifically, of probability dry and of three representative characteristics of the marginal distribution of nonzero daily rainfall, i.e., the mean value, the L-variation and the L-skewness. In general, a typical sinusoidal-like pattern was revealed (see Figure 3.2) for all statistics and for both hemispheres, with a surprising exception in the probability dry of the

SH where a more complicated picture is observed. Additionally, to explore the monthly variation in detail at each record a test for seasonality is proposed and applied, i.e., the SV-Test. Application of the SV-Test revealed a clear monthly variation in probability dry and in the mean value of nonzero daily rainfall in 95.1% and in 91.7%, respectively, of the stations studied (see Figure 3.5); the corresponding percentages of the shape characteristics, i.e., of L-variation and L-skewness, were 66.1% and 54.2%, respectively, these results if combined with the general picture obtained by the analysis in the hemispheres indicate that, in general, the shape characteristic vary too. The monthly variation of those statistics at each station was quantified by various deviation measures with respect to the average of all months (see Figure 3.7). The analysis showed that the highest monthly variation is observed in the mean value of nonzero rainfall followed by probability dry, L-skewness and finally by L-variation, implying that although the shape characteristics vary, their variability is much less than that of the mean value and the probability dry.

Regarding the second question the performance of two flexible distributions was assessed; specifically, one power-type, the Burr type XII distribution, and one of exponential form, the Generalized Gamma. In order to check the suitability of these distributions for the nonzero daily rainfall, first, L-moments ratio diagrams were used to evaluate their potential to describe or reproduce the observed shape characteristics of all records; and second, these distributions were actually fitted and the parameters were estimated for all records. For the huge number of records analysed both distributions performed very well. Particularly, the Burr type XII in the worst case, i.e., in November, managed to describe 79.1% of the records (see Figure 3.9); the corresponding value for the Generalized Gamma distribution was observed in January and was 94.2% while this distribution was able to describe the shape characteristics for all months in 78.8% of the stations (see Figure 3.10). Finally, the two distributions were compared to each other using various error measures and the Generalized Gamma performed better in most of the cases (see Figure 3.15).

The implications of this study are: (a) the marginal distribution of daily rainfall varies over the months and over location suggesting the necessity for a flexible probability model; (b) the seasonal and the spatial variability observed in the shape characteristics points out that the commonly used two-parameter models, e.g., the Gamma, the Weibull, the Lognormal, the Pareto, etc. cannot serve as “universal” models for the daily rainfall; (c) the density function of daily rainfall may significantly differ not only in its general shape, i.e., J-

shaped or Bell-shaped, but also in its tail behaviour; this dictates that a "universal" probability model for daily rainfall must have at least two shape parameters, one to control the left tail and one to control the right tail; (d) two simple models with the above characteristics that perform very well are the Burr type XII distribution and the Generalized Gamma distribution with the latter performing even better than the former providing thus an excellent model choice; (e) using only these two distributions, having some of their characteristics complementary to each other, the entire dataset can be modelled for all months and all stations; and (f) the shape parameter $\gamma_2$ of the Generalized Gamma distribution, which controls the right tail and thus the extreme values, for the vast majority of records analysed is $\gamma_2 < 1$, with 1 corresponding to the Gamma distribution; this implies that some of the most commonly used exponential-tail distributions like the Exponential, the Gamma or mixed Exponentials may constitute a dangerous choice and should not be used unjustifiably in practice as they can severely underestimate the magnitude and the frequency of the extreme daily rainfall.

# CHAPTER 4

*"...the premise of probability simultaneously postulates the existence of the improbable."*

CARL GUSTAV JUNG

# A FOCUS ON THE DISTRIBUTION TAILS OF DAILY RAINFALL

**ABSTRACT**

The upper part of a probability distribution, usually known as the tail, governs both the magnitude and the frequency of extreme events. The tail behaviour of all probability distributions may be, loosely speaking, categorized in two families: heavy-tailed and light-tailed distributions, with the latter generating "milder" and less frequent extremes compared to the former. This emphasizes how important for hydrological design it is to assess the tail behaviour correctly. Traditionally, the wet-day daily rainfall has been described by light-tailed distributions like the Gamma distribution, although heavier-tailed distributions have also been proposed and used, e.g., the Lognormal, the Pareto, the Kappa, and others. This study investigates the distribution tails for daily rainfall by comparing the upper part of empirical distributions of thousands of records with four common theoretical tails: those of the Pareto, Lognormal, Weibull and Gamma distributions. Specifically, 15 029 daily rainfall records are used from around the world with record lengths from 50 to 172 years. The analysis shows that heavier-tailed distributions are in better agreement with the observed rainfall extremes than the more often used lighter tailed distributions. This result has clear implications on extreme event modelling and engineering design.

## 4.1 Introduction

Heavy rainfall may induce serious infrastructure failures and may even result in loss of human lives. It is common then to characterize such rainfall with adjectives like "abnormal", "rare" or "extreme". But what can be considered "extreme" rainfall? Behind any discussion on the subjective nature of such pronouncements, there lies the fundamental issue of infrastructure design, and the crucial question of the threshold beyond which events need not be taken into account as they are considered too rare for practical purposes. This question is all the more pertinent in view of the EU Flooding Directive's requirement to consider "extreme (flood) event scenarios" [*European Commission*, 2007].

Although short term prediction of rainfall is possible to a degree (and useful for operational purposes), long term prediction, on which infrastructure design is based, is infeasible in deterministic terms. Thus, rainfall is treated here in a probabilistic manner, i.e., it is considered as a random variable (RV) governed by a distribution law. Such a distribution law enables to assign a return period to any rainfall amount, so that it could be then reasonably argued that a rainfall event, e.g., with return period 1000 years or more, is indeed an extreme. Yet, which distribution law is the appropriate is still a matter of debate.

The typical procedure for selecting a distribution law for rainfall is to (a) try some of many, a priori chosen, parametric families of distributions, (b) estimate the parameters according to one of many existing fitting methods, and (c) choose the one best fitted according to some metric or fitting test. Nevertheless, this procedure does not guarantee that the selected distribution will model adequately the tail, which is the upper part of the distribution that controls both the magnitude and frequency of extreme events. On the contrary, as only a very small portion of the empirical data belongs to the tail (unless a very large sample is available), all fitting methods will be "biased" against the tail, since the estimated fitting parameters will point towards the distribution that best describes the largest portion of the data (by definition not belonging to the tail). Clearly, an ill-fitted tail may result in serious errors in terms of extreme event modelling with potentially severe consequences for hydrological design. For example, in Figure 4.1 where four different distributions are fitted to the empirical distribution tail, it can be observed that the predicted magnitude of the 1000-year event varies significantly.

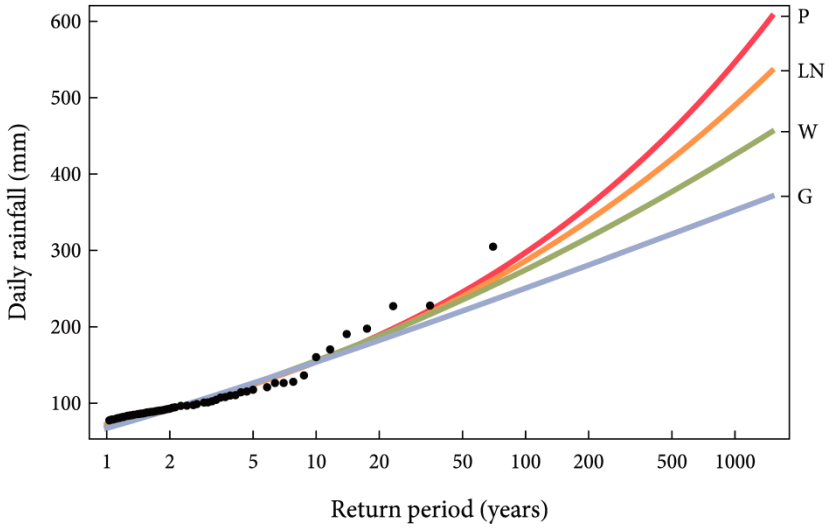



**Figure 4.1**. Four different distribution tails fitted to an empirical tail (P, LN, W and G stands for the Pareto, the Lognormal, the Weibull and the Gamma distribution). A wrong choice may lead to severely underestimated or overestimated rainfall for large return periods.

The distributions can be classified according to the asymptotic behaviour of their tail in two general classes: (a) the subexponential class with tails tending to zero less rapidly than an exponential tail (here the term "exponential tail" is used to describe the tail of the exponential distribution), and (b) the hyperexponential or the superexponential class, with tails approaching zero more rapidly than an exponential tail [*Teugels*, 1975; *Klüppelberg*, 1988, 1989]. Mathematically, this "intuitive" definition of the subexponential class for a distribution function $F$ is expressed as

$$\lim_{x\to\infty}\frac{1-F(x)}{\exp(-x/\beta)}=\infty, \quad \forall\,\beta>0 \tag{4.1}$$

while several equivalent mathematical conditions in order to classify a distribution as subexponential have been proposed [see e.g., *Embrechts et al.*, 1997; *Goldie and Klüppelberg*, 1998]. Furthermore, this is not the only classification, as several other exist [see e.g., *El Adlouni et al.*, 2008 and references therein]. In addition, many different terms have been used in the literature to refer to tails "heavier" than the exponential, e.g., "heavy tails", "fat tails", "thick tails", or, "long tails", that may lead to some ambiguity: see for example the various definitions that exist for the class of heavy-tailed distributions discussed by *Werner and Upper* [2004]. Here, the term "heavy tail" is used in an intuitive and general way, i.e., to refer to tails approaching zero less rapidly than an exponential tail.

The practical implication of a heavy tail is that it predicts more frequent larger magnitude rainfall compared to light tails. Hence, if heavy tails are more suitable for modelling extreme events, the usual approach of adopting light-tailed models (e.g., the Gamma distribution) and fitting them on the whole sample of empirical data would result in a significant underestimation of risk with potential implications for human lives. However, there are significant indications that heavy tailed distributions may be more suitable. For example, in a pioneering study *Mielke* [1973] proposed the use of the Kappa distribution, a power-type distribution, to describe daily rainfall. Today there are large databases of rainfall records that allow us to investigate the appropriateness of light or heavy tails for modelling extreme events. This is the subject in which this paper aims to contribute.

## 4.2 The dataset

The data used in this study are daily rainfall records from the Global Historical Climatology Network-Daily database (version 2.60, www.ncdc.noaa.gov/oa/climate/ghcn-daily) which includes over 40 000 stations worldwide. Many of the records, however, are too short, have many missing data, or, contain data suspect in terms of quality (for details regarding the quality flags refer to the Network's website above).

Thus, only records fulfilling the following criteria were selected for the analysis: (a) record length greater or equal than 50 years, (b) missing data less than 20% and, (c) data assigned with "quality flags" less than 0.1%. Among the several different quality flags assigned to measurements, the data were screened against two: values with quality flags "G" (failed gap check) or "X" (failed bounds check) which are used to flag suspiciously large values, i.e., a sample value that is orders of magnitude larger than the second larger value in the sample. Whenever such a value existed in the records it was deleted (this however occurred in only 594 records in total, and in each of these records typically one or two values had to be deleted). Screening with these criteria resulted in 15 137 stations. The locations of these stations as well as their record lengths can be seen in Figure 4.2 while Table 4.1 presents some basic summary statistics of the nonzero daily rainfall of those records (for further details on the dataset please Appendix B).

It is noted that none of the missing values was filled because this would be meaningless for this study which focuses on extreme rainfall as any regression-type technique would underestimate the real extreme values. Missing values only affect the effective record length and, given the relatively high lower limit of record length set (50 years, while much smaller records are often used in hydrology, e.g. 10-30 years), the

resulting problem is not serious. Additionally, the percentage 20% of missing daily values refers to the worst case and actually it is much smaller in the majority of the records; thus missing values cannot alter or modify the conclusions drawn.

**Table 4.1.** Some basic statistics of the 15 137 records of daily rainfall. For each record the statistics of the first row were estimated. Apart from probability dry ($P_{dry}$) these statistics are for the nonzero daily rainfall.

| | $P_{dry}$ (%) | Nonzero values No. | Median (mm) | Mean (mm) | SD (mm) | Skew |
|---|---|---|---|---|---|---|
| min | 15.11 | 320 | 0.40 | 1.00 | 1.76 | 1.37 |
| $Q_5$ | 53.92 | 2 121 | 1.70 | 3.61 | 5.01 | 2.36 |
| $Q_{25}$ | 68.55 | 4 038 | 3.00 | 6.18 | 8.28 | 2.85 |
| $Q_{50}$ | 76.35 | 5 973 | 4.80 | 9.27 | 12.08 | 3.28 |
| $Q_{75}$ | 83.65 | 8 497 | 6.90 | 12.65 | 16.42 | 3.94 |
| $Q_{95}$ | 91.36 | 13 060 | 10.20 | 17.75 | 24.25 | 5.38 |
| max | 98.25 | 27 867 | 25.70 | 83.96 | 158.02 | 26.31 |
| Mean | 75.13 | 6 604 | 5.18 | 9.77 | 12.97 | 3.56 |
| SD | 11.46 | 3 508 | 2.70 | 4.60 | 6.20 | 1.31 |
| Skew | −0.74 | 1.12 | 1.03 | 1.16 | 1.88 | 5.58 |

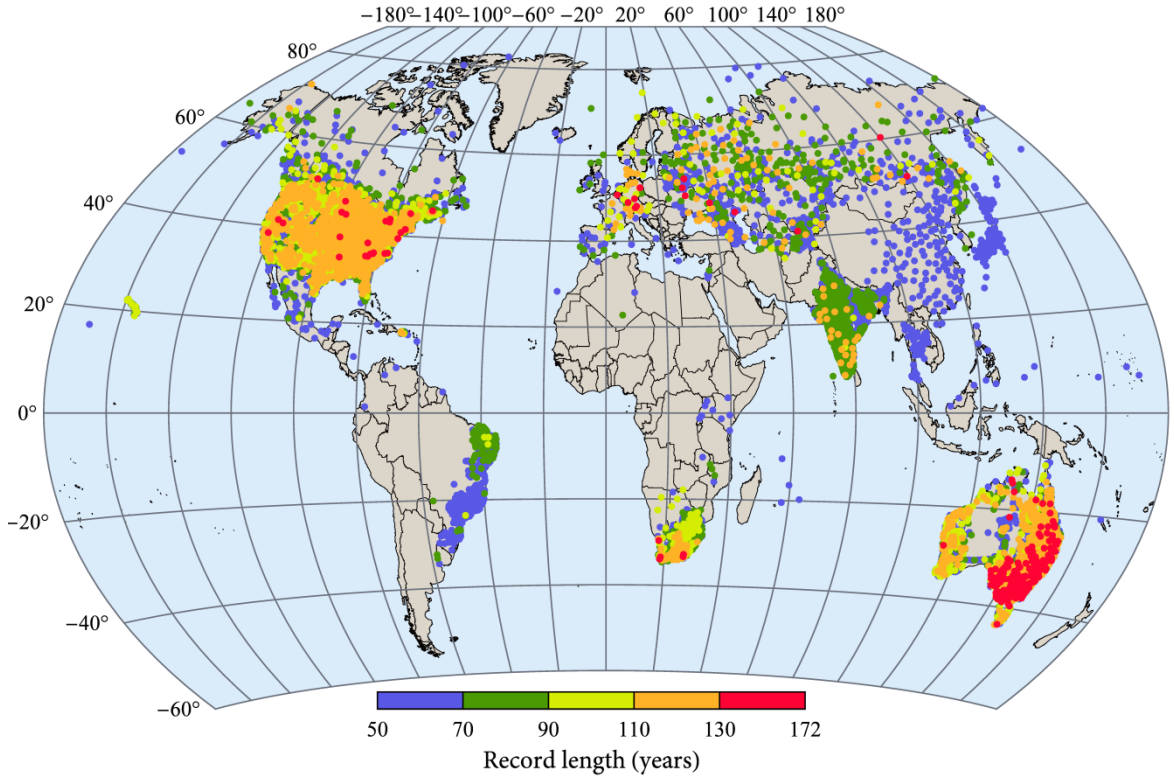



**Figure 4.2.** Locations of the stations studied (a total of 15 137 daily rainfall records with time series length greater than 50 years). Note that there are overlaps with points corresponding to high record lengths shadowing (being plotted in front of) points of lower record lengths.

Finally, it is noted that the statistical procedure, described next, failed in a few records, for reasons of algorithmic convergence or time limits. Excluding these records, the total number of records where the analysis was applied is 15 029.

## 4.3 Defining and fitting the tail

The marginal distribution of rainfall, particularly at small time scales like the daily, belongs to the so-called mixed type distributions, with a discrete part describing the probability of zero rainfall, or the probability dry, and a continuous part expressing the magnitude of the nonzero (wet-day) rainfall. As suggested earlier, studying extreme rainfall requires focusing on the behaviour of the distribution's right tail which governs the frequency and the magnitude of extremes.

If rainfall is denoted with $X$, and the nonzero rainfall with $X \mid X > 0$, then the exceedance probability function (EPF; also known as survival function, complementary distribution function, or tail function) of the nonzero rainfall, using common notation, is defined as

$$P(X > x \mid X > 0) = \overline{F}_{X|X>0}(x) = 1 - F_{X|X>0}(x) \tag{4.2}$$

where $F_{X|X>0}(x)$ is any valid probability distribution function chosen to describe nonzero rainfall. It should be clear that the unconditional EPF is easily derived if the probability dry $p_0$ is known: $\overline{F}_X(x) = (1-p_0)\overline{F}_{X|X>0}(x)$. Since the focus is on the continuous part of the distribution, and more specifically on the right tail, from this point on, for notational simplicity the subscript in $\overline{F}_{X|X>0}(x)$ is omitted, denoting thus the conditional EPF function simply as $\overline{F}(x)$. To avoid ambiguity due to the term "tail function" for EPF, it is clarified that the term "tail" is used to refer only to the upper part of the EPF, i.e., the part that describes the extremes.

At this point, however, it is necessary to define what can be considered as the upper part. A common practice is to set a lower threshold value $x_L$ [see e.g., *Cunnane*, 1973; *Tavares and Da Silva*, 1983; *Ben-Zvi*, 2009] and study the behaviour for values greater than $x_L$. Yet, there is no universally accepted method to choose this lower value. A commonly accepted method (known as partial duration series method) is to determine the threshold indirectly based on the empirical distribution, in such a way that the number of values above the threshold equals the number of years $N$ of the record [see e.g., *Cunnane*, 1973]. The resulting series, defined in this way, is known in the literature as annual exceedance

series and is a standard method for studying extremes in hydrology [see e.g., *Chow*, 1964; *Gupta*, 2011].

This may look similar to another common method, in which the *N* annual maxima of the *N* years are extracted and studied. However, the method of annual maxima, by selecting the maximum value of each year may distort the tail behaviour (e.g., when the three largest daily values occur within a single year, it only takes into account the largest of them). For this reason, instead of studying the *N* daily annual maxima, the focus is on the *N* largest daily values of the record assuming that these values are representative of the distribution's tail and can provide information for its behaviour. Thus, the method adopted here has the advantage of better representing the exact tail of the parent distribution.

It is worth noting that a common method of studying series above a threshold value is based on the results obtained by *Balkema and de Haan* [1974] and *Pickands III* [1975]. According to these results, loosely speaking, as this threshold tends to infinity, the conditional distribution above the threshold converges to the Generalized Pareto which includes, as a special case, the Exponential distribution. It is noted though, that these results are asymptotic results, i.e., valid (or providing a good approximation) if this threshold value tends to infinity (or if it is very large). In the case where the parent distribution is of power type or of exponential type, the theory is applicable even for not so large threshold values because the convergence of the tail is fast. In other cases, e.g., Lognormal or stretched exponential distributions, the convergence is very slow. The same applies to the classical extreme value theory (EVT), which predicts that the distribution of maxima converges to one of the three extreme value distributions. For some examples illustrating the slow convergence to the asymptotic distributions of EVT (the same philosophy applies for Balkema-de Haan-Pickands theorem) see, e.g., *Papalexiou and Koutsoyiannis* [2013] and *Koutsoyiannis* [2004a].

Given that each station has an *N*-year record of daily values and a total number *n* of nonzero values, the empirical EPF $\overline{F}_N(x_i)$, conditional on nonzero rainfall, is defined as the empirical probability of exceedance (according to the Weibull plotting position)

$$\overline{F}_N(x_i) = 1 - \frac{r(x_i)}{n+1} \tag{4.3}$$

where $r(x_i)$ is the rank of the value $x_i$, i.e., the position of $x_i$ in the ordered sample $x_{(1)} \le \ldots \le x_{(n)}$ of the nonzero values. Thus the empirical tail is determined by the *N* largest nonzero rainfall values of $\overline{F}_N(x_i)$ with $n - N + 1 \le i \le n$ (note that $x_{\mathrm{L}} = x_{(n-N+1)}$). Some basic

summary statistics of the series of the $N$ largest nonzero rainfall values are presented in Table 4.2.

**Table 4.2.** Some basic statistics of the 15 137 tail-samples defined for an $N$-year record as the $N$ largest nonzero values. For each tail-sample the statistics of the first row were estimated.

| | Tail values No. | Median | Mean | SD | Max |
|---|---|---|---|---|---|
| min | 50 | 8.90 | 10.42 | 3.01 | 21.50 |
| $Q_5$ | 52 | 28.30 | 31.71 | 8.61 | 68.60 |
| $Q_{25}$ | 61 | 43.55 | 48.24 | 13.85 | 110.00 |
| $Q_{50}$ | 70 | 62.75 | 69.12 | 19.01 | 152.40 |
| $Q_{75}$ | 97 | 85.30 | 93.72 | 27.59 | 218.40 |
| $Q_{95}$ | 122 | 130.30 | 144.70 | 47.48 | 357.60 |
| max | 172 | 977.00 | 1 041.02 | 395.96 | 1 750.00 |
| Mean | 79 | 68.78 | 76.01 | 22.50 | 175.06 |
| SD | 23 | 34.84 | 38.20 | 13.21 | 93.42 |
| Skew | 0.80 | 2.73 | 2.58 | 3.55 | 1.79 |

Obviously the number of nonzero daily rainfall values is $n = (1 - p_0) n_d N$ where $n_d = 365.25$ is the average number of the days in a year. According to the Weibull plotting position given in Eq. (4.3) the exceedance probability $\overline{p}(x_L)$ of $x_L$ will be

$$\overline{p}(x_L) = 1 - \frac{n - N + 1}{n + 1} = \frac{N}{(1 - p_0) n_d N + 1} \approx \frac{1}{(1 - p_0) n_d} \tag{4.4}$$

This shows that the exceedance probability of the threshold $x_L$ depends only on the probability dry $p_0$. Interestingly, the average $p_0$ of the records analysed in this study is approximately 0.75 which implies that the exceedance probability of $x_L$ is on average as low as 0.01, while even for $p_0 = 0.95$ its value is 0.055. It is reasonable to assume that values above this threshold can be assumed that belong to the tail of the distribution. It is noted that there are studies [see e.g., *Beguería et al.*, 2009] where the threshold value was chosen to correspond to the 90th percentile, a value much smaller than the one corresponding to our choice of threshold. Section 4.6 refers further to the selection of the threshold, also in comparison with different methods of selection.

The fitting method followed here is straightforward, i.e., directly fitting and comparing the performance of different theoretical distribution tails to the empirical tails estimated from the daily rainfall records previously described. The theoretical tails are fitted to the empirical ones by minimizing numerically a modified mean square error (MSE) norm N1 defined as

$$\mathrm{N1} = \frac{1}{N} \sum_{i=n-N+1}^{n} \left( \frac{\overline{F}(x_{(i)})}{\overline{F}_N(x_{(i)})} - 1 \right)^2 \tag{4.5}$$

A complete verification of the method and a comparison with other norms is presented in section 4.6. At this point it is only noted that its rationale (and advantage over classical square error norms) as it properly "weights" each point that contributes in the sum. Namely, it considers the relative error between the theoretical and the empirical values rather than using the $x$ values themselves. For example, considering the classical square error, i.e., $(x_i - x_u)^2$, with $x_u$ denoting the quantile value for probability $u$ equal to the empirical probability of the value $x_i$, then large values would contribute much more to the total error than the smaller ones. This may be a problem especially for rainfall records where the values usually differ more than one order of magnitude is, e.g., from 0.1 mm to more than 100 mm. Obviously, the best fitted tail for a specific record is considered to be the one with the smallest MSE.

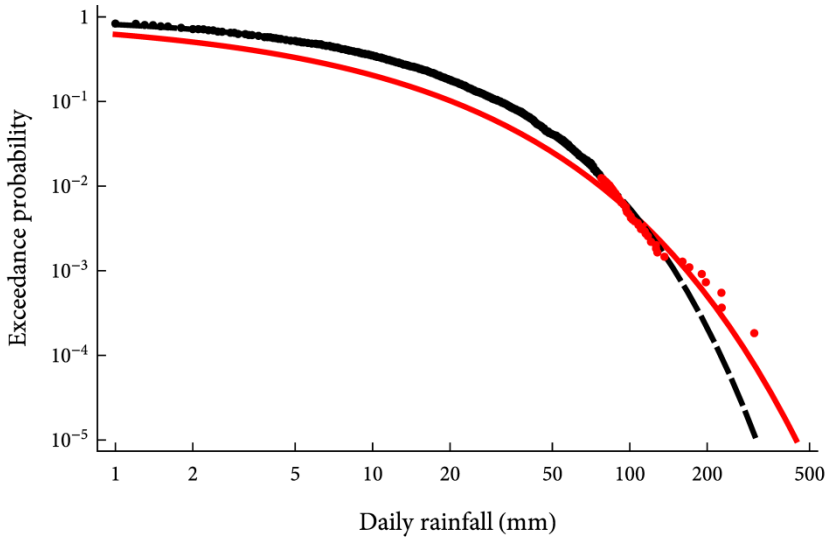



**Figure 4.3**. Explanatory diagram of the fitting approach followed. The dashed line depicts a Weibull distribution fitted to the whole empirical distribution points while the solid red line depicts the distribution fitted only to the tail points.

The proposed approach, which fits the theoretical distribution only to the $N$ largest points of each dataset, ensures that the fitted distribution provides the best possible description of the tail and is not affected by lower values. As an example of the fitting method, Figure 4.3 depicts the Weibull distribution fitted to an empirical sample (the station was randomly selected and has code IN00121070) by minimizing the norm given by (4.5) in two ways, (a) in all the points of the empirical distribution and (b) in only the

largest $N$ points. It is clear that the first approach (dashed line) does not adequately describe the tail.

It is well known that several other methods have been extensively used to estimate the parameters of candidate distributions, e.g., the lognormal maximum likelihood and the log-probability plot regression [*Kroll and Stedinger*, 1996], and more recently the log partial probability weighted moments and the partial L-moments [*Wang*, 1996; *Bhattarai*, 2004; *Moisello*, 2007]. Yet, the advantage of the proposed method is that any tail can be fitted in the same manner and can be directly compared with other fitted tails since the resulting MSE value can clearly indicate the best fitted; in the aforementioned methods an additional measure has to be estimated in order to compare the performance of the fitted distributions.

## 4.4 The fitted distribution tails

It is clear from the previous section that any tail can be fitted to the empirical ones. Nevertheless, here four different and common distribution tails, i.e., the tails of the Pareto type II (PII) the Lognormal (LN), the Weibull (W), and the Gamma (G) distributions, are fitted and compared in terms of their performance. These distributions were chosen for their simplicity, popularity, as well as for being tail-equivalent (or for having similar asymptotic behaviour) with many other more complicated distributions. It is reminded that two distribution functions $F$ and $G$ with support unbounded to the right are called tail-equivalent if $\lim_{x\to\infty} \overline{F}(x)/\overline{G}(x) = c$ with $0 < c < \infty$.

The Pareto and the Lognormal distributions belong to the subexponential class and are considered heavy-tailed distributions; the Weibull can belong to both classes depending on the values of its shape parameter, while the Gamma distribution has essentially an exponential tail but not precisely (see below). From a practical point of view, the ordering of these distributions, from heavier to lighter tail, is: Pareto, Lognormal, Weibull with shape parameter < 1, Gamma and Weibull with shape parameter > 1 [see e.g., *El Adlouni et al.*, 2008]. Note that Pareto is the only power-type distribution while the rest three are of exponential form.

Specifically, the Pareto type II distribution is the simplest power-type distribution defined in [0,∞). Its probability density function (PDF) and EPF are given, respectively, by

$$f_{\mathrm{PII}}(x) = \frac{1}{\beta}\left(1+\gamma\frac{x}{\beta}\right)^{-\frac{1}{\gamma}-1} \tag{4.6}$$

$$\overline{F}_{\mathrm{PII}}(x) = \left(1+\gamma\frac{x}{\beta}\right)^{-\frac{1}{\gamma}} \tag{4.7}$$

and it is defined by the scale parameter $\beta > 0$, and the shape parameter $\gamma \geq 0$ that controls the asymptotic behaviour of the tail. Namely, as the value of $\gamma$ increases, the tail becomes heavier and consequently extreme values occur more frequently. For $\gamma = 0$ it degenerates to the exponential tail while for $\gamma \geq 0.5$ the distribution has infinite variance. Many other power-type distributions are tail-equivalent, i.e., exhibiting asymptotic behaviour similar to $x^{-1/\gamma}$ with the Pareto type II tail, e.g., the Burr type XII [*Burr*, 1942; *Tadikamalla*, 1980] the two- and three-parameter Kappa [*Mielke Jr*, 1973], the Log-Logistic [e.g., *Ahmad et al.*, 1988] and the Generalized Beta of the second kind [*Mielke Jr and Johnson*, 1974].

Another very common distribution used in hydrology is the Lognormal with PDF and EPF, respectively,

$$f_{\mathrm{LN}}(x) = \frac{1}{\sqrt{\pi}\,\gamma\,x}\exp\left(-\ln^2\left(\frac{x}{\beta}\right)^{1/\gamma}\right) \tag{4.8}$$

$$\overline{F}_{\mathrm{LN}}(x) = \frac{1}{2}\mathrm{erfc}\left(\ln\left(\frac{x}{\beta}\right)^{1/\gamma}\right) \tag{4.9}$$

where $\mathrm{erfc}(x) = 2\pi^{-1/2}\int_x^\infty \mathrm{e}^{-t^2}\,\mathrm{d}t$. The distribution comprises the scale parameter $\beta > 0$ and the parameter $\gamma > 0$ that controls the shape and the behaviour of the tail. Lognormal is also considered a heavy-tailed distribution (it belongs to the subexponential family) and can approximate power-law distributions for a large portion of the distribution's body [*Mitzenmacher*, 2004]. Notice that the notation in Eq. (4.8) and Eq. (4.9) differs from the common one and illustrates more clearly the kind of the two parameters (scale and shape).

The Weibull distribution, which can be considered as a generalization of the exponential distribution, is another common model in hydrology [*Heo et al.*, 2001a,b] and its PDF and EPF are given, respectively, by

$$f_{\mathrm{W}}(x) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma}\right) \tag{4.10}$$

$$\overline{F}_{\mathrm{W}}(x) = \exp\left(-\left(\frac{x}{\beta}\right)^{\gamma}\right) \tag{4.11}$$

The parameter $\beta > 0$ is a scale parameter, while the shape parameter $\gamma > 0$ governs also the tail's asymptotic behaviour. For $\gamma < 1$ the distribution belongs to the subexponential family with a tail heavier than the exponential one, while for $\gamma > 1$ the distribution is characterized as hyperexponential with a tail thinner than the exponential. Many distributions can be assumed tail-equivalent with the Weibull for a specific value of the parameter $\gamma$, e.g., the Generalized Exponential, the Logistic and the Normal.

Finally, one of the most popular models for describing daily rainfall is the Gamma distribution [e.g., *Buishand*, 1978b], which like the Weibull distribution belongs to the exponential family. Its PDF and EPF are given, respectively, by

$$f_{\mathrm{G}}(x) = \frac{1}{\beta\Gamma(\gamma)}\left(\frac{x}{\beta}\right)^{\gamma-1}\exp\left(-\frac{x}{\beta}\right) \tag{4.12}$$

$$\overline{F}_{\mathrm{G}}(x) = \Gamma\left(\gamma, \frac{x}{\beta}\right)/\Gamma(\gamma) \tag{4.13}$$

where $\Gamma(a,x) = \int_x^{\infty} t^{a-1}\exp(-t)\mathrm{d}t$ is the upper incomplete Gamma function and $\Gamma(a) = \int_0^{\infty} t^{a-1}\exp(-t)\mathrm{d}t$ the Gamma function. Generally, it can be assumed that the Gamma tail behaves similar to the exponential tail. Yet, this is only approximately correct as the Gamma distribution belongs to a class of distributions [denoted as S($\gamma$); see e.g., *Embrechts and Goldie*, 1982; *Klüppelberg*, 1989; *Alsmeyer and Sgibnev*, 1998] that irrespective of its parameter values cannot be classified as subexponential, while it is not tail-equivalent with the exponential. This can be seen from the fact that the $\lim_{x\to\infty} \overline{F}_{\mathrm{G}}(x)/\overline{G}(x)$ is 0 for $\beta < \beta_{\mathrm{E}}$ and $\infty$ for $\beta > \beta_{\mathrm{E}}$, where $\overline{G}(x) = \exp(-x/\beta_{\mathrm{E}})$ is the exponential tail. Yet, it is noted that if compared with an exponential tail with $\beta = \beta_{\mathrm{E}}$, then

$$\lim_{x\to\infty}\frac{\overline{F}(x)}{\overline{G}(x)} = \begin{cases} 0 & 0 < \gamma < 1 \\ 1 & \gamma = 1 \\ \infty & \gamma > 1 \end{cases} \tag{4.14}$$

Therefore, in this case, practically speaking, for $0 < \gamma < 1$ the Gamma distribution has a "slightly lighter" tail than the exponential tail as it decreases faster, while for $\gamma > 1$ it exhibits a "slightly heavier" tail as it decreases more slowly than the exponential tail.

Finally, it is worth noting that the distributions compared here, and consequently their tails have similarities in their structure as all have two parameters and specifically one scale parameter and one shape parameter. Nevertheless, among the various distributions with the same parameter structure, inevitably, some are more flexible than others. One way to quantify this flexibility is by comparing them in terms of various shape measures (e.g., skewness, kurtosis, etc.). For example, the feasible ranges of skewness for the Pareto, Lognormal, Weibull and Gamma are, respectively, $(2, \infty)$, $(0, \infty)$, $(-1.14, \infty)$ and $(0, \infty)$. Therefore, the Weibull distribution seems to be the most "flexible" distribution among them and the Pareto the less. Yet, this argument is not valid when the focus is on the tail because the general shape of the tail is basically similar and what differs is the rate at which the tail approaches zero.

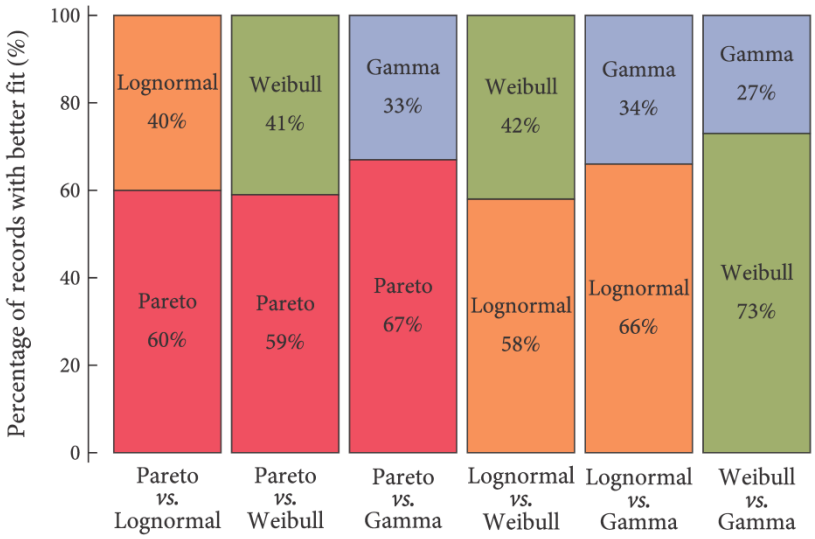



**Figure 4.4.** Comparison of the fitted tails in couples in terms of the resulting MSE. The heavier tail of each couple is better fitted to the empirical points in a higher percentage of the records.

## 4.5 Results and discussion

The basic statistical results from fitting the four distribution tails, following the methodology described, to the 15 029 daily rainfall records are given in Table 4.3. In order to assess which tail has the best fit, the four tails were compared in couples in terms of the resulting MSE, i.e., the tail with the smaller MSE is considered better fitted. As shown in Figure 4.4, the Pareto tail, when compared with the other three distributions, was better fitted in about 60% of the stations. Interestingly, the distribution with the heavier tail of

each couple, in all cases, was better fitted in a higher percentage of the stations, which implies a rule of thumb of the type "the heavier, the better"!

**Table 4.3.** Summary statistics from the fitting of the four distribution tails into the 15 029 tail-samples of daily rainfall.

| | Pareto | | | Lognormal | | |
|---|---|---|---|---|---|---|
| | MSE | $\beta$ | $\gamma$ | MSE | $\beta$ | $\gamma$ |
| Min | 0.002 | 0.42 | 0.001 | 0.002 | 1.22 | 0.531 |
| Mode† | 0.011 | 7.54 | 0.134 | 0.012 | 8.78 | 1.060 |
| Mean | 0.017 | 8.80 | 0.140 | 0.018 | 9.46 | 1.087 |
| Median | 0.021 | 9.51 | 0.145 | 0.022 | 10.59 | 1.107 |
| Max | 0.336 | 54.79 | 0.797 | 0.322 | 76.74 | 2.284 |
| SD | 0.015 | 4.92 | 0.076 | 0.015 | 6.44 | 0.214 |
| Skew | 2.910 | 1.23 | 0.495 | 2.755 | 1.73 | 0.561 |
| | Weibull | | | Gamma | | |
| | MSE | $\beta$ | $\gamma$ | MSE | $\beta$ | $\gamma$ |
| Min | 0.002 | 0.02 | 0.230 | 0.002 | 3.79 | 0.010 |
| Mode | 0.013 | 4.33 | 0.661 | 0.015 | 17.50 | 0.092 |
| Mean | 0.019 | 5.91 | 0.678 | 0.023 | 23.15 | 0.219 |
| Median | 0.022 | 6.88 | 0.692 | 0.032 | 28.18 | 0.294 |
| Max | 0.298 | 52.72 | 1.491 | 0.482 | 120.00 | 2.433 |
| SD | 0.015 | 4.69 | 0.139 | 0.034 | 17.30 | 0.269 |
| Skew | 2.151 | 1.82 | 0.668 | 4.377 | 1.65 | 2.567 |

†The mode was estimated from the empirical density function (histogram) after smoothing.

Another comparison revealing the overall performance of the fitted tails was based on their average rank. That is, the fitted tails in each record were ranked according to their MSE, i.e., the tail with the smaller MSE was ranked as 1 and the one with the largest as 4. Figure 4.5 depicts the average rank of each tail for all stations. Again, the Pareto performed best, while the most popular model for rainfall, the Gamma distribution, performed the worst. The percentages of each distribution tail that was best fitted are: 30.7% for Pareto, 29.8%, for Lognormal, 13.6% for Weibull and 25.8% for Gamma. Again the Pareto distribution is best according to these percentages; interestingly however, the Gamma distribution has a relatively high percentage, higher than the Weibull. This does not contradict the conclusion derived by the average rank. The explanation is that the Gamma distribution was ranked as best in some cases, but when it was not the best fitted, it was probably the worst fitted.

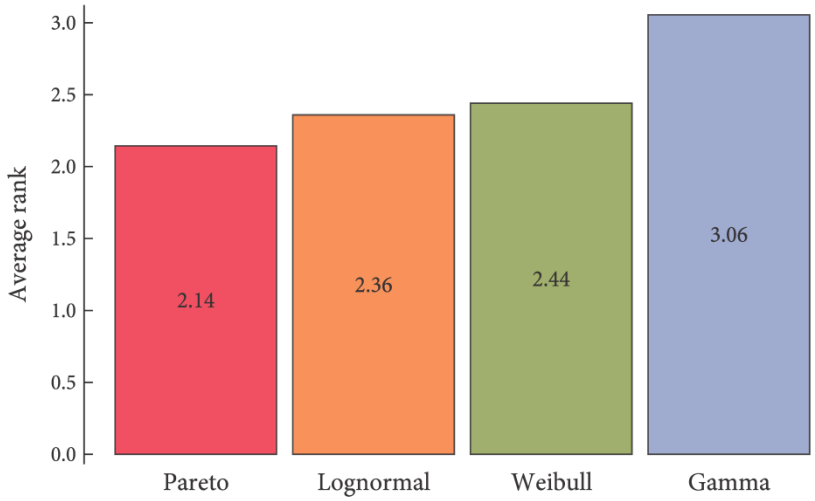



**Figure 4.5.** Mean ranks of the tails for all records. The best-fitted tail is ranked as 1 while the worst-fitted as 4. A lower average rank indicates a better performance.

Figure 4.6 and Figure 4.7 depict, respectively, the empirical distributions of the shape and of the scale parameters of the fitted tails. It is well-known that the most probable values are the ones around the mode, which for the Pareto shape parameter is 0.134. Interestingly, this value is close to the one determined in a different context by *Koutsoyiannis* [1999] using *Hershfield's* [1961] dataset. This implies that power-type distributions, which asymptotically behave like the Pareto, will not have finite power moments of order greater than $1 / 0.134 \approx 7.5$. Moreover, as the empirical distribution of the Pareto shape parameter in Figure 4.6 attests, values around 0.2 are also common, implying the non-existence of moments greater than the fifth order. This entails that sample moments of that or higher order (sometimes appearing in research papers) may not exist. Regarding the Weibull tail, the estimated mode of its shape parameter is 0.661, implying a much heavier tail compared to the exponential one. Finally, it is worth noting that the estimated mode of the Gamma shape parameter is as low as 0.092. The shape parameter of the Gamma distribution controls mainly the behaviour of the left tail, resulting in J- or bell-shaped densities (loosely speaking, the right tail is dominated by the exponential function and thus behaves like an exponential tail). A value that low corresponds to an extraordinarily J-shaped density which would be unrealistic for describing the whole distribution body of daily rainfall. In other words, a Gamma distribution fitted to the whole set of points would most probably underestimate the behaviour of extremes.

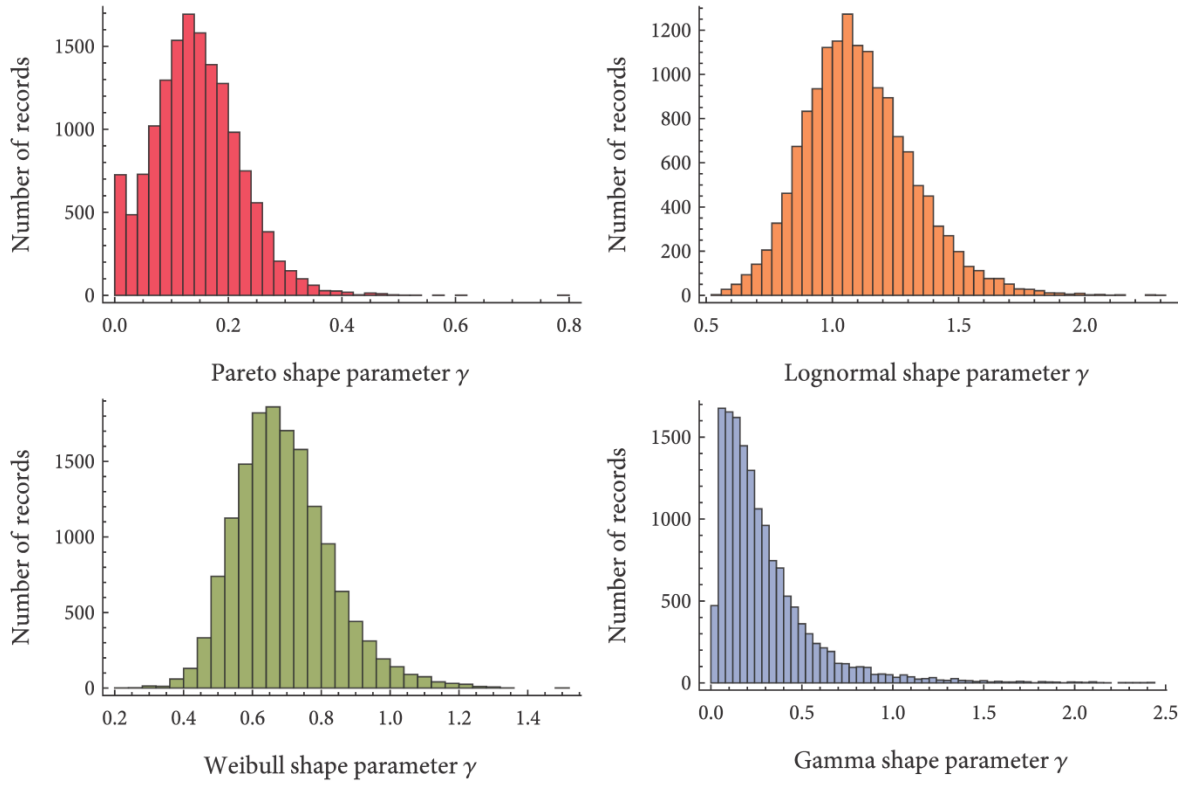



**Figure 4.6.** Histograms of the shape parameters of the fitted tails.

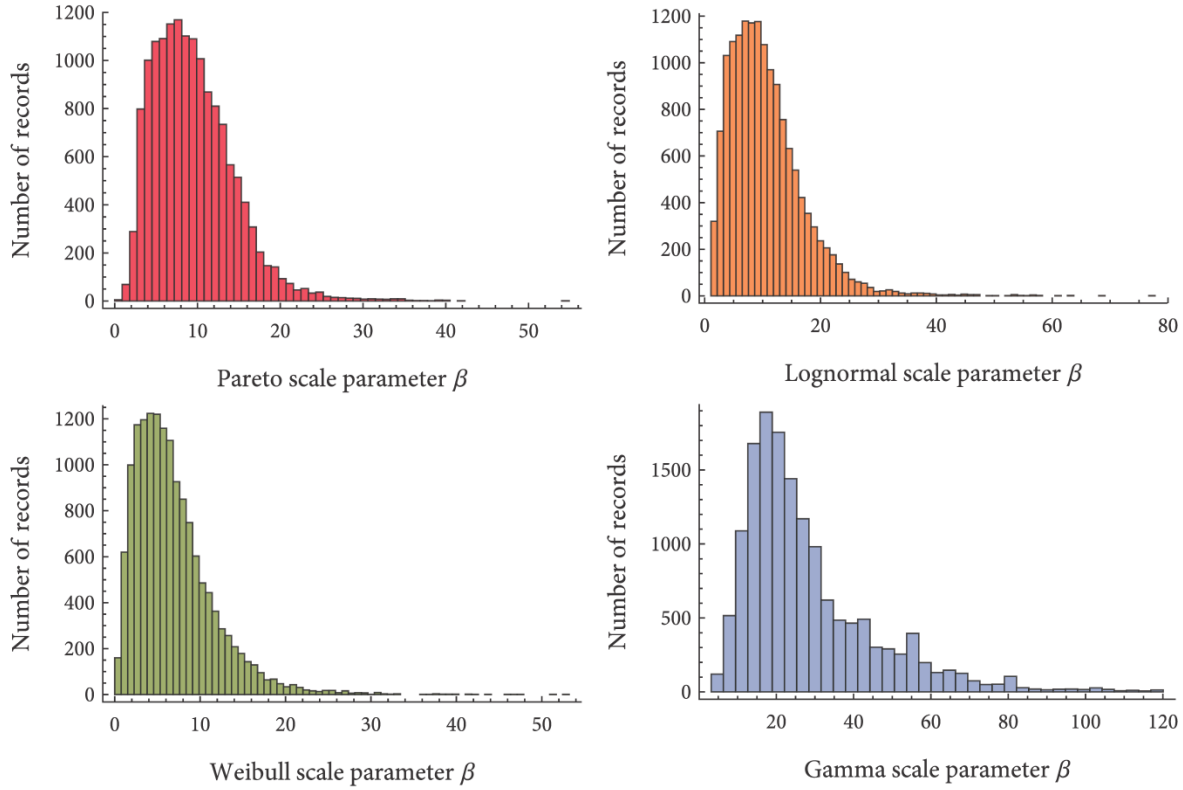



**Figure 4.7.** Histograms of the scale parameters of the fitted tails.

The existence of geographical patterns that potentially define climatic zones, in the best fitted tails, was also investigated, i.e., the existence of zones in the world where the majority of the records were better described by one of the studied distribution tails. The maps in Figure 4.8, which depict the locations of the stations where each distribution tail was best fitted, did not unveil any regular patterns in terms of the best fitted distribution but rather seem to follow a random variation.

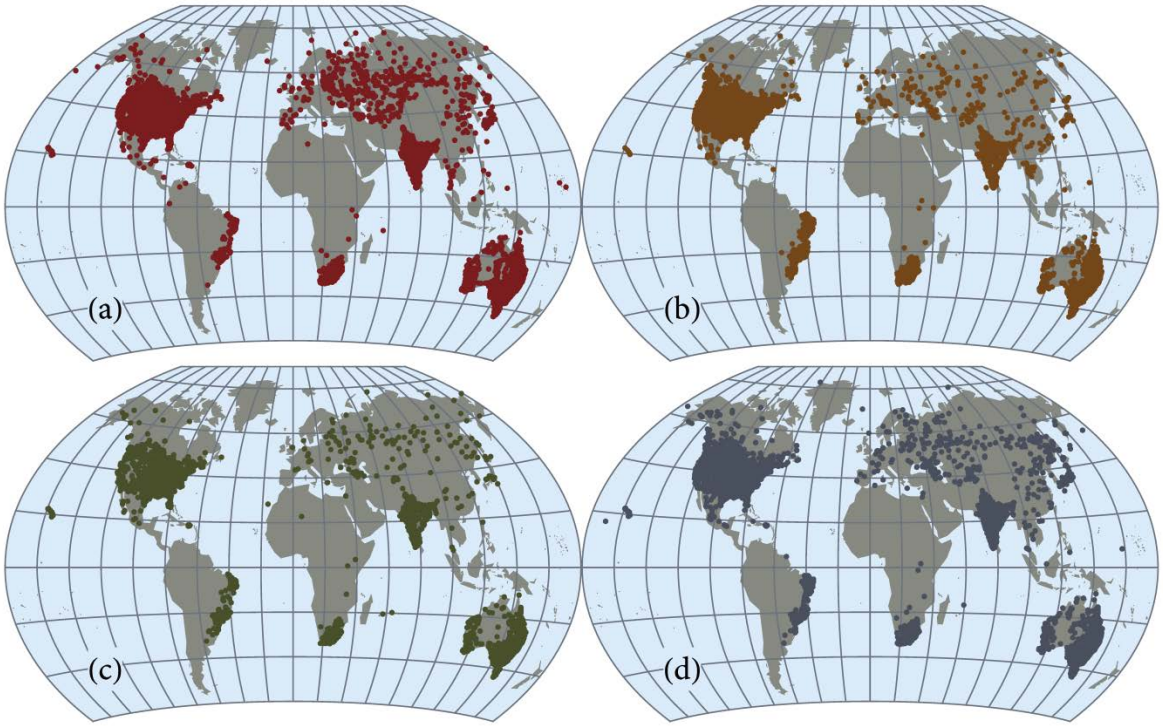



**Figure 4.8.** Geographical depiction of the 15 029 stations where the best fitted tail is (a) Pareto in 4 621, (b) Lognormal in 4 486, (c) Weibull in 2 051, and (d) Gamma in 3 871.

Another way to investigate for geographical patterns, as the previous map did not reveal any useful information, is to study the fitted tails grouped into two coarser groups:, the subexponential group and the exponential-hyperexponential group. The former includes the Pareto, the Lognormal and the Weibull with $\gamma < 1$ tails, while the latter includes the Gamma and the Weibull with $\gamma \geq 1$ tails. Among the 15 029 records, subexponential tails were best fitted in 10 911 cases or in 72.6% while exponential-hyperexponential tails were best fitted in 4 118 or in 27.4%. Further, in order to get a clearer picture instead of constructing maps with the locations where the first-group or the second-group tails were best fitted, the study focused on the percentage of subexponential tails that were best fitted in large regions. Specifically, a grid covering the entire earth was constructed using a latitude difference $\Delta\varphi = 2.5°$ and longitude difference $\Delta\lambda = 5°$. The

percentage of the best fitted subexponential tails in each grid cell is simply estimated by counting the number of the best fitted subexponential tails divided by the total number of records within the cell. These percentages are presented in the form of a map in Figure 4.9, using a colour scale as shown in the map's legend. The cells plotted in the map are those containing at least two records, so that the calculation of percentages has some meaning.

The map of Figure 4.9 clearly shows that in the vast majority of cells subexponential tails dominate (percentage > 60%). Particularly, out of 532 cells having at least two records, 255 and 163 have percentages of subexponential tails between 60-80%, and >80%, respectively. In contrast, in only 35 and 79 cells are the percentage values in the ranges 0-40% and 40-60%, respectively.

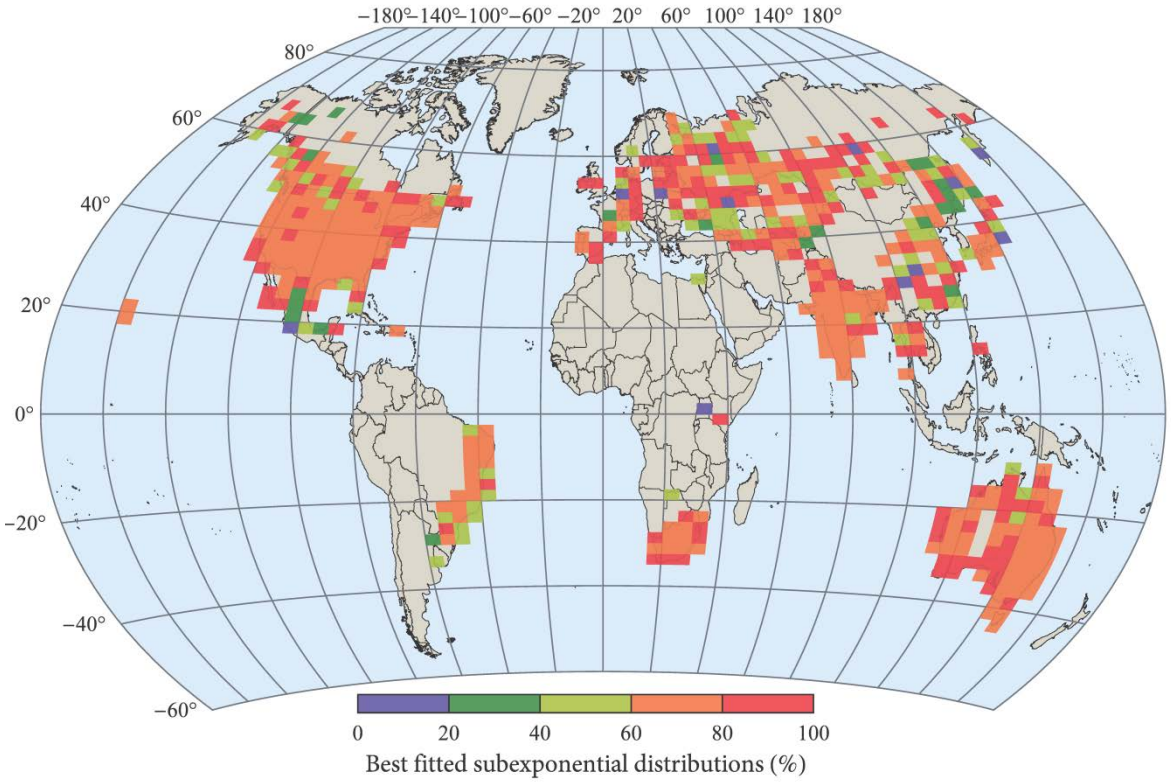



**Figure 4.9.** Geographical variation of the percentage of best fitted subexponential tails in cells defined by latitude difference $\Delta\varphi = 2.5°$ and longitude difference $\Delta\lambda = 5°$. In total, in 72.6% of the 15 029 records analysed, the subexponential tails were the best fitted.

## 4.6 Verification of the fitting method

The use of a different norm for fitting the tail into the empirical data could potentially modify the conclusions drawn. Nevertheless, this argument is pointless in the sense that the main concern should be the efficiency of the norm used, i.e., if it possesses desired properties, e.g., if it is unbiased and has lower variance in comparison to other candidates.

Usually, the error is expressed in terms of random variable values, e.g., rainfall values, and not in terms of probability. However, a literature search did not reveal or verify that the commonly used norms, e.g., the classical MSE norm, are better than the norm N1 used here (see Eq. (4.5)).

For this reason, a Monte Carlo scheme was implemented, which actually replicates the method followed, i.e., the performance of the norm N1 was evaluated and also compared with the more common norms N2 and N3 defined as

$$\mathrm{N2} = \frac{1}{N} \sum_{i=n-N+1}^{n} \left( \frac{x_u}{x_{(i)}} - 1 \right)^2 \tag{4.15}$$

$$\mathrm{N3} = \frac{1}{N} \sum_{i=n-N+1}^{n} \left( x_u - x_{(i)} \right)^2 \tag{4.16}$$

Here $x_u = Q(u)$ is the value predicted by the quantile function $Q$ of the distribution under study for $u$ equal to the empirical probability of $x_{(i)}$ (the $i$th element the sample ranked in ascending order) according to the Weibull plotting position. The norm N2 has the same rationale as the one used but the error is estimated in terms of rainfall values, rather than in terms of probability, while the norm N3 is the classical and most commonly used MSE norm.

The Monte Carlo scheme performed can be summarized in the following steps: (a) 1000 random samples are generated from each one of the four distributions studied with sample size equal to 6600 values which is approximately the average number of nonzero daily rainfall values per record; (b) selection of the scale and the shape parameter values to be approximately equal with the median values resulted from the analysis of the real world dataset (see Table 4.3) in order for the generated random samples to be representative of the real data; and (c) each distribution is fitted to its corresponding random sample and estimated the parameters by applying our method for each one of the three norms, while $N$ is set equal to 80 years, which is approximately the average record length.

The results are presented in Figure 4.10. The whiskers of the box plots express the 95% Monte Carlo confidence interval of the parameters while the dashed lines show the true parameter values. It is clear that the norm N1 used in this study results in almost unbiased estimation of the parameters while especially for the Pareto and the Lognormal distributions results in markedly smaller variance compared to the classical norm N3. The norm N2 seems to perform very well for the Pareto, Lognormal and Weibull distributions

(although somewhat biased) but the results are poor for the Gamma distribution. The classical and the most commonly used norm N3 is by far the worst in term of bias excepting the Gamma distribution, for which it performs equally well as N1. In particular, for the subexponential distributions of this simulation, i.e., the Pareto, the Lognormal and the Weibull, the classical norm N3 fails to provide good results. This may point to a more general conclusion, i.e., that the classical MSE, which is inspired based on properties of the normal distribution, is not a good choice for subexponential distributions. This needs to be further investigated; however, it is reasonable to assume that there is a rationale supporting this conclusion: subexponential distributions can generate "extremely" extreme values compared to the main "body" of values, and thus, in the classical norm these values will contribute "extremely" to the total error heavily affecting the fitting results.

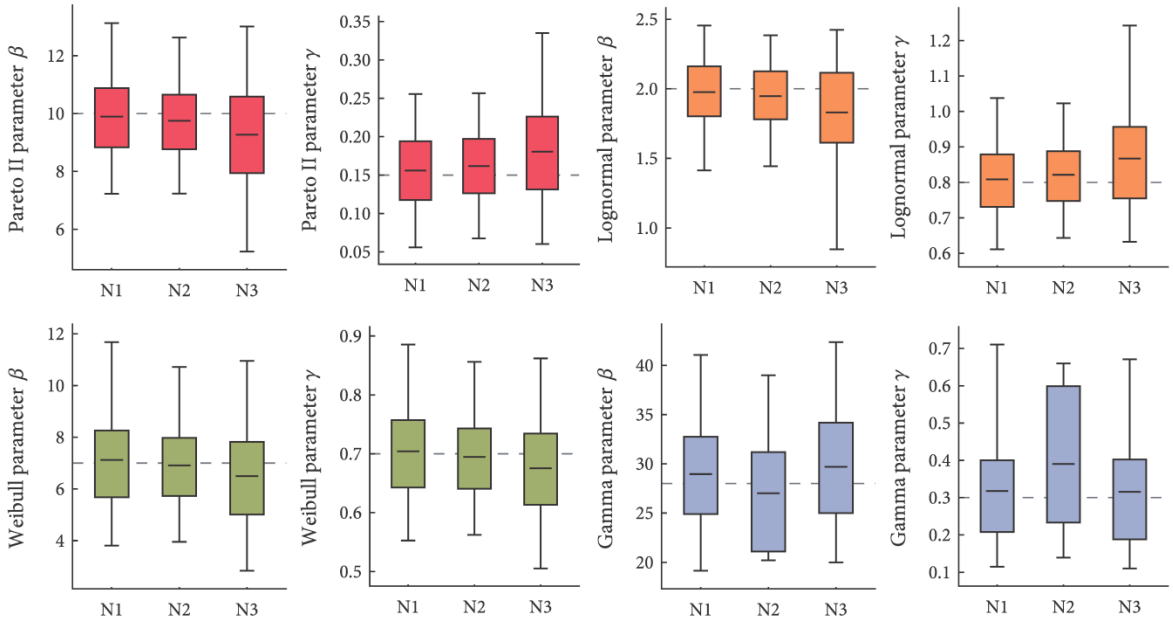



**Figure 4.10.** Results of a Monte Carlo scheme implemented to evaluate the performance of the norm N1 used in fitting of tails in this study, in comparison to commonly used ones (N2, N3).

Another issue of potential concern for the validity of the conclusions drawn is the impact of the sample size, i.e., the number of the $N$ largest events, for which the four distribution tails are fitted. As mentioned before, the annual exceedance series used here is a standard method in hydrology in which $N$ equals the number of the record's years. Obviously, $N$ can be defined in many different ways, either with reference to record length or as a fixed number for every record studied.

In order to assess the impact of the number of events in the performance of the four fitted distribution tails 2 000 records were randomly selected among the 15 029 analysed

and the four distribution tails were fitted using six different methods for defining $N$. The first method (M1) is the one used for all above analyses, in which $N$ equals the number of the record's years. In the second (M2) and third (M3) methods the threshold $x_L$ is defined as the 90th- and the 95th-percentiles, respectively, so that $N$ equals the number of events included in the upper 10% and 5%, respectively, of the nonzero values. Obviously, in these two methods $N$ varies from record to record depending on the total number of nonzero values and on the average it equals 667 and 333 values for M2 and M3, respectively. In the rest three methods (M4, M5 and M6) $N$ is defined as a fixed number for every record, i.e., 50, 100 and 200 values, respectively.

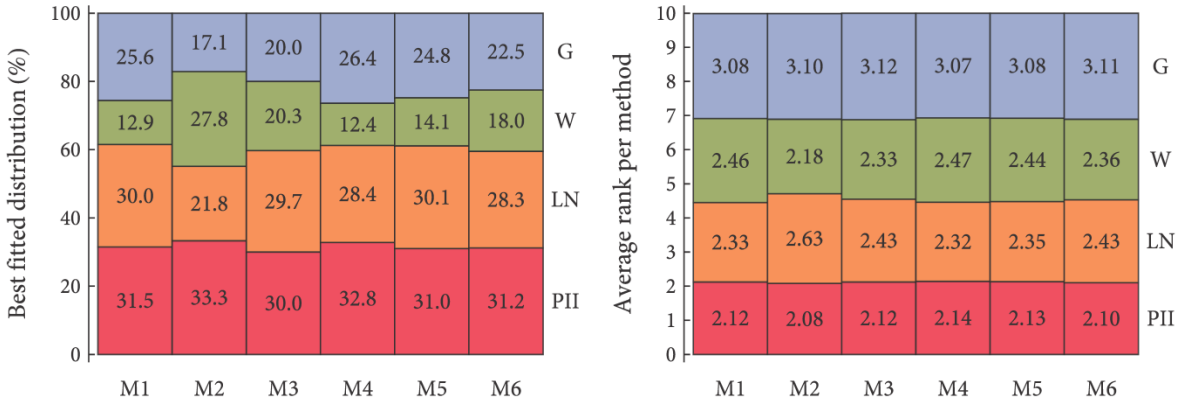



**Figure 4.11.** Performance results of the four fitted tails in 2000 randomly selected records using six different methods for selecting the sample size: (a) Percentage of records that each distribution tail was best fitted; (b) Average ranks of the fitted tails (lower rank indicates better performance).

The performance results comparing the six methods are given in Figure 4.11 which presents (a) the estimated percentages that each distribution was best fitted and (b) the average rank of each distribution tail. Again the Pareto II tail was best fitted in a higher percentage of records in all cases (M1-M6) with the percentage values varying in a narrow range. The results are essentially the same with those obtained from the analysis of the whole database. The only noticeable difference regards the method M2, in which the Weibull tail seems to "gain ground" over the Gamma and the Lognormal tails. In general it seems that the Weibull tail increases its performance as $N$ increases. Thus, in M4 where $N$ has the lowest value, i.e., 50 values, it performs the worst, while in M2 where $N$ is maximum (667 values on the average), it performs the best. The average rank, which is a better measure of the overall performance of the distribution tails, remains essentially the same for each distribution in all methods. An exception is observed again in M2 where the Weibull tail performs better than the Lognormal tail. Apart from this exception the general

conclusion is again that the Pareto II performs the best, followed by the Lognormal and the Weibull tails, while the Gamma tail performs the worst in all cases.

## 4.7 Summary and conclusions

Daily rainfall records from 15 029 stations are used to investigate the performance of four common tails that correspond to the Pareto, the Weibull, the Lognormal and the Gamma distributions. These theoretical tails were fitted to the empirical tails of the records and their ability to capture adequately the behaviour of extreme events was quantified by comparing the resulting MSE. The ranking from best to worst in terms of their performance is: (a) the Pareto, (b) the Lognormal, (c) the Weibull, and (d) the Gamma distributions. The analysis suggests that heavier-tailed distributions in general performed better than their lighter-tailed counterparts. Particularly, in 72.6% of the records subexponential tails were better fitted while the exponential-hyperexponential tails were better fitted is only 27.4%. It is instructive that the most popular model used in practice, the Gamma distribution, performed the worst, revealing that the use of this distribution underestimates in general the frequency and the magnitude of extreme events. Nevertheless, is should not be neglected that the Gamma distribution was the best fitted in 25.8% of the records.

Additionally, it is noted that heavy tails tend to be hidden [*Koutsoyiannis*, 2004a, 2004b; *Papalexiou and Koutsoyiannis*, 2013] especially when the sample size is small. Thus, it could be argued that even in the cases where the Gamma tail performed well, the true underlying distribution tail may be heavier. This leads to the recommendation that heavy-tailed distributions are preferable as a means to model extreme rainfall events worldwide. It is also noted, that the tails studied here are as simple as possible, i.e., only one shape parameter controls their asymptotic behaviour. Yet there are many distributions with more than one shape parameters which may affect their tail behaviour. Particularly, the Generalized Gamma [*Stacy*, 1962] and the Burr type XII distributions were compared as candidates for the daily rainfall (based on L-moments) in an earlier study, using thousands of empirical daily records and the former performed better [*Papalexiou and Koutsoyiannis*, 2012].

The key implication of this analysis is that the frequency and the magnitude of extreme events have generally been underestimated in the past. Engineering practice needs to acknowledge that extreme events are not as rare previously thought and to shift toward the heavy-tailed probability distributions.

# CHAPTER 5

*"Φύσις κρύπτεσθαι φιλεῖ"*

*HERACLITUS OF EPHESUS*

# ON THE DISTRIBUTION OF ANNUAL MAXIMUM DAILY RAINFALL

**ABSTRACT**

Theoretically, if the distribution of daily rainfall is known or justifiably assumed, then one could argue, based on extreme value theory, that the distribution of the annual maxima of daily rainfall would resemble one of the three limiting types: (a) type I, known as Gumbel, type II, known as Fréchet and, type III, known as reversed Weibull. Yet, the parent distribution usually is not known and often only records of annual maxima are available. Thus, the question that naturally arises is which one of the three types better describes the annual maxima of daily rainfall. The question is of great importance as the naïve adoption of a particular type may lead to serious underestimation or overestimation of the return period assigned to specific rainfall amounts. To answer this question, the annual maximum daily rainfall of 15 137 records from all over the world is analysed, with lengths varying from 40 to 163 years. The Generalized Extreme Value (GEV) distribution, which comprises the three limiting types as special cases for specific values of its shape parameter, is fitted and the fitting results are examined focusing on the behaviour of the shape parameter. The analysis reveals that: (a) the record length strongly affects the estimate of the GEV shape parameter and long records are needed for reliable estimates, (b) when the effect of the record length is corrected the shape parameter varies in a narrow range, (c) the geographical location of the globe may affect the value of the shape parameter, and (d) the winner of this battle is the Fréchet law.

## 5.1 Introduction

Arguably, the statistical behaviour of the annual maximum daily rainfall has been the cornerstone of statistical hydrology, as it is directly related to the design of hydraulic infrastructures and to extreme floods. In hydrology, the study of rainfall or flood extremes has been an active research field and a matter of debate for more than half a century dating back to the works of E. J. Gumbel in 1940s; however, the field of extreme value theory seems to have originated more than three centuries ago in the works of Nicolaus Bernoulli [see e.g. *Gumbel*, 1958]. Yet, it was during the 20th century when the theory was rapidly evolved and found applications in astronomy, hydrology and engineering in general.

A detailed historical survey on the subject would be out of the scope of this study. Nevertheless, here are mentioned some of the milestones of this fascinating field [for a more complete historical note see e.g. *Kotz and Nadarajah*, 2000]. It seems that the first methodical approach was due to von *Bortkiewicz* [1922] regarding the range of random samples. In the sequel, *Fréchet* [1922]identified one of the asymptotic distributions of maxima, and, soon after, Fisher and *Tippett* [1928] showed that there are only three possible limiting distributions for extremes. These findings were strengthened by *von Mises* [1936] who identified some sufficient conditions for convergence to the three limiting laws. Yet, it was *Gnedenko* [1943] who set the solid foundations of the asymptotic theory of extremes providing the precise conditions for the weak convergence to the limiting laws. All these initial theoretical results were refined and generalized later in the works of *Juncosa* [1949], *Smirnov* [1949], *Watson* [1954], *Jenkinson* [1955], *Barndorff-Nielsen* [1963], *Berman* [1964], *de Haan* [1971], *Balkema and de Haan* [1972], *Galambos* [1972] and *Pickands III* [1975] to mention some of them. Numerous real-world applications followed this theoretical progress not only in flood and rainfall analysis. It is worth noting in this respect *Gumbel's* [1958] celebrated book who was one of the pioneers promoting and applying the formal theory into engineering practice.

Accordingly, the central question in extreme rainfall analysis is: which one of the three extreme value distributions, i.e., the Gumbel, the Fréchet or the reversed Weibull, should be chosen to describe extreme rainfall? Its answer is not only of academic interest, but mainly constitutes a practical matter of eminent significance as the wrong choice may severely underestimate the design rainfall of hydraulic infrastructures leading thus to infrastructure failures and other negative consequences. Overestimation can also be a possibility, which again has negative consequences in terms of the infrastructure cost. During the last decades, accumulation of observations and advances in computers

facilitated the analysis of extreme rainfall and literally thousands of studies or technical reports have been published using, or arguing for or against, a particular extreme value distribution. Yet, most of these studies are of "local" character, e.g., case studies analysing extreme rainfall in particular areas. As an exception, the study by *Koutsoyiannis* [2004a,b] used records from several sites in the globe but the number of records was small (169 rainfall records worldwide each having 100-154 years of data). Here, the aim is to investigate the behaviour of the annual maximum daily rainfall at a global scale, using more than 15 000 rainfall records distributed across the globe, and to provide a better answer to the question addressed.

## 5.2 Theoretical issues of extreme analysis

### 5.2.1 The three limiting laws

It is well known that if a random variable (RV) $X$ follows the distribution $F_X(x)$ then according to the classical extreme value theory the distribution function of the maximum of $n$ independent and identically distributed (iid) RV's, i.e., $Y_n = \max(X_1,\ldots,X_n)$ is given by

$$G_{Y_n}(x) = \left(F_X(x)\right)^n \tag{5.1}$$

Now, loosely speaking, if $n \to \infty$ three limiting laws can emerge from Eq. (5.1). Actually, as $\lim_{n\to\infty}\left(F(x)\right)^n$ results in a degenerate distribution, the limiting laws are obtained from $\lim_{n\to\infty}\left(F(a_n x + b_n)\right)^n$ for appropriate constants $a_n > 0$ and $b_n$ [*Fisher and Tippett*, 1928]. In addition, these limiting laws emerge not only for iid RV's as *Juncosa* [1949] extended these results to the case of non-iid random variables and *Leadbetter* [1974] proved that the limiting distributions hold also for dependent random variables, given that there is no long range dependence of high level exceedances.

The three limiting laws are the type I or Gumbel (G), the type II or Fréchet (F) and the type III or reversed Weibull (RW) with distribution functions respectively given by

$$G_{\mathrm{G}}(x) = \exp\left(-\exp\left(-\frac{x-\alpha}{\beta}\right)\right), \qquad x \in \mathbb{R} \tag{5.2}$$

$$G_{\mathrm{F}}(x) = \exp\left(-\left(\frac{x-\alpha}{\beta}\right)^{-1/\gamma}\right), \qquad x \geq \alpha \tag{5.3}$$

$$G_{\mathrm{RW}}(x) = \exp\left( -\left( -\frac{x-\alpha}{\beta} \right)^{1/\gamma} \right), \qquad x \le \alpha \tag{5.4}$$

All three distributions comprise a location parameter $\alpha \in \mathbb{R}$ and a scale parameter $\beta > 0$, with the Fréchet and the reversed Weibull distributions having the additional shape parameter $\gamma > 0$. Although the expressions of the Fréchet and the reversed Weibull distributions look very similar, i.e., they differ in a couple of signs, the distributions behave completely differently as the first is bounded from below while the second is bounded from above. Noteworthy, the exponential form of the Fréchet distribution does not imply an exponential right tail, i.e., the Fréchet distribution behaves like a power-type distribution as it can be easily proved that for $\gamma > 0$ the function $1 - \exp(-x^{-1/\gamma})$ is asymptotically equivalent to $x^{-1/\gamma}$ (it is reminded that two functions $f(x)$ and $g(x)$ are asymptotically equivalent if $\lim_{x\to\infty} f(x)/g(x) = 1$). Likewise, the double exponential form of the Gumbel distribution does not imply a double exponential tail, as its right tail is asymptotically equivalent with the exponential tail, i.e., $\exp(-x)$.

Now, any specific parent distribution $F_X(x)$ belongs to the domain of attraction of one the aforementioned limiting laws. To which one depends mainly on the form of its right tail. Several formal mathematical conditions determine the distribution's domain of attraction (formed originally by *von Mises* [1936] and *Gnedenko* [1943] and extended by several other authors ; [for a complete account see e.g. *Embrechts et al.*, 1997; *Reiss and Thomas*, 2007]). Generally speaking, distributions with right tail regularly varying in infinity or, equivalently, not having all of their moments finite, belong to the domain of attraction of the Fréchet law. These include power-type distributions like the Pareto, the Burr type XII and III, the Log-Gamma, the Cauchy and others. In contrast, in the domain of attraction of the Gumbel law belong all distributions with right tail tending to zero faster than any power-type tail, or equivalently distributions having all of their moments finite, e.g., Normal, Lognormal, Gamma, Weibull and others. Finally, in the domain of attraction of the reversed Weibull law belong distributions bounded from above [see e.g. *Kotz and Nadarajah*, 2000].

The afore mentioned three limiting distribution laws can be unified into a single expression known as the Generalized Extreme Value (GEV) distribution (also known as the Fisher-Tippet) with probability distribution function given by

$$G_{\mathrm{GEV}}(x)=\exp\left(-\left(1+\gamma\frac{x-\alpha}{\beta}\right)^{-1/\gamma}\right),\qquad 1+\gamma\frac{x-\alpha}{\beta}\geq 0 \tag{5.5}$$

This parameterization was proposed by *von Mises* [1936], although it is commonly attributed to *Jenkinson* [1955]. The distribution comprises the location parameter $\alpha\in\mathbb{R}$ the scale parameter $\beta>0$ and the shape parameter $\gamma\in\mathbb{R}$. It can be easily seen that for $\gamma>0$ it is bounded from below, ($x\geq\alpha-\beta/\gamma$) while for $\gamma<0$ it is bounded from above ($x\leq\alpha-\beta/\gamma$) (notice that here positive $\gamma$ means a GEV bounded from below, while some texts use opposite sign convention). Essentially, the GEV distribution formula can be seen as a simple reparameterization of the Fréchet formula as the Fréchet parameters (indexed with F in Eq. (5.3)) are related with the GEV parameters, i.e., $\alpha_{\mathrm{F}}=\alpha-\beta/\gamma$, $\beta_{\mathrm{F}}=\beta/\gamma$ and $\gamma_{\mathrm{F}}=\gamma$. This simple reparameterization exploits the limiting definition of the exponential function, i.e., $\lim_{\gamma\to 0}(1+\gamma x)^{-1/\gamma}=\exp(-x)$ so that the Gumbel distribution emerges for $\gamma\to 0$.

### 5.2.2 Convergence to the limiting laws

The distribution of the maximum value, given in Eq. (5.1), converges to one of the three liming laws (depending on the parent distribution) given that the maximum value is selected from a number of variables which tends to infinity. In real world, convergence practically holds if this number is very large. However, in daily rainfall it seems that this number is not even large as in the best case it would equal the number of the year's days, i.e., 365 or 366 values. Actually, the number of rainy days $N_{\mathrm{R}}$ that depends on the probability dry is always smaller than the number of year's days and varies from year to year. Thus, whether or not the annual maximum can actually be modelled by one the three limiting laws should not be taken for granted [see also *Koutsoyiannis*, 2004a].

To demonstrate this issue, results from a previous study are used [*Papalexiou and Koutsoyiannis*, 2012] where more than ten thousand daily rainfall records were analysed and was found that the Burr type XII distribution (BrXII) and the Generalized Gamma distribution (GG), are both very good models for describing the non-zero daily rainfall. Their probability density functions are given, respectively, by

$$f_{\mathrm{BrXII}}(x)=\frac{1}{\beta}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\left(1+\gamma_2\left(\frac{x}{\beta}\right)^{\gamma_1}\right)^{-\frac{1}{\gamma_1\gamma_2}-1}\qquad x\geq 0 \tag{5.6}$$

$$f_{\mathrm{GG}}(x)=\frac{\gamma_2}{\beta\Gamma(\gamma_1/\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma_2}\right),\qquad x\geq 0 \tag{5.7}$$

Hence, assuming that both of these distributions can serve as parent distributions, and assuming a constant number of rainy days $N_{\mathrm{R}}$, the exact distribution of the annual maximum then would respectively be $G_{\mathrm{BrXII}}(x)=\left(F_{\mathrm{BrXII}}(x)\right)^{N_{\mathrm{R}}}$ and $G_{\mathrm{GG}}(x)=\left(F_{\mathrm{GG}}(x)\right)^{N_{\mathrm{R}}}$. It is noted that the BrXII distribution as a power type distribution belongs to the domain of attraction of the Fréchet law; in contrast, the GG distribution is of exponential type, having all of its moments finite and thus belonging to the domain of attraction of the Gumbel law. So, theoretically speaking the first is expected to converge to the Fréchet law and the second to the Gumbel law.

The different daily rainfall records analysed in the aforementioned study had different statistical characteristics, yet in order to illustrate the convergence rate based on real world evidence the next procedure was followed. First, the medians (closer to the mode than the mean value) of the sample estimates of the first L-moment $\lambda_1$ (mean), of L-variation $\tau_2$ and of L-skewness $\tau_3$ are considered as representative statistics of the nonzero daily rainfall; their numerical estimates are $\lambda_1 = 9.86$, $\tau_2 = 0.58$, $\tau_3 = 0.45$ (all parameters with dimensions, e.g., $\lambda_1$ or scale parameters, are expressed in mm). Additionally, the median of probability dry was 76.3% corresponding approximately to $N_{\mathrm{R}} = 87$ rainy days. These statistics can be reproduced by a BrXII distribution with parameters $\beta = 8.47$, $\gamma_1 = 0.91$, $\gamma_2 = 0.18$, and a GG distribution with parameters $\beta = 1.83$, $\gamma_1 = 1.16$, $\gamma_2 = 0.54$. The parameters of the exact distribution of the annual maximum, for these parent distributions and for $N_{\mathrm{R}} = 87$, were numerically calculated. Namely, the $G_{\mathrm{BrXII}}$ would have $\lambda_1 = 77.62$, $\tau_2 = 0.23$, $\tau_3 = 0.30$ and the $G_{\mathrm{GG}}$ would have $\lambda_1 = 73.71$, $\tau_2 = 0.20$, $\tau_3 = 0.24$. Next the GEV and Gumbel distributions, corresponding to these parameters, were determined, i.e., for the $G_{\mathrm{BrXII}}$ parameters the GEV will have $\alpha = 60.71$, $\beta = 20.85$, $\gamma = 0.19$, and the Gumbel will have $\alpha = 62.72$, $\beta = 25.80$. Likewise, for the $G_{\mathrm{GG}}$ parameters the GEV will have $\alpha = 60.48$, $\beta = 19.15$, $\gamma = 0.10$, and the Gumbel will have $\alpha = 61.43$, $\beta = 21.28$.

This analysis is graphically depicted in Figure 5.1 where the fitted distributions are formed in a Rainfall *vs.* Return period plot. It can be easily shown that the exact annual maximum laws, i.e., the $G_{\mathrm{BrXII}}$ and the $G_{\mathrm{GG}}$ are given by the relationship $x(T)=Q_{X|X>0}\left(\left(1-1/T\right)^{1/N_{\mathrm{R}}}\right)$, where $T$ denotes the return period in years and $Q_{X|X>0}$ the quantile function of the representative BrXII or GG distribution describing the nonzero

daily rainfall. The graph reveals that the exact annual maximum law, assuming as a parent distribution the BrXII, quickly converges to the anticipated Fréchet law or GEV with positive $\gamma$. Noteworthy, the tail index of the representative BrXII, expressed by the shape parameter $\gamma_2$, and the shape parameter $\gamma$ of the GEV distribution, theoretically should be the same. In reality, while they are not exactly the same, they are very close, i.e., $\gamma_2 = 0.19$ and $\gamma = 0.18$, verifying thus a satisfactory convergence. On the other hand, assuming the GG as a parent distribution, it is observed that not only does the exact law $G_{GG}$ not converge to the Gumbel law as theoretically expected, but it is better described by the Fréchet law. In this case the GEV overestimates the rainfall for large return periods, yet, it is on the safe side, whereas it is clear that the Gumbel distribution severely underestimates it.

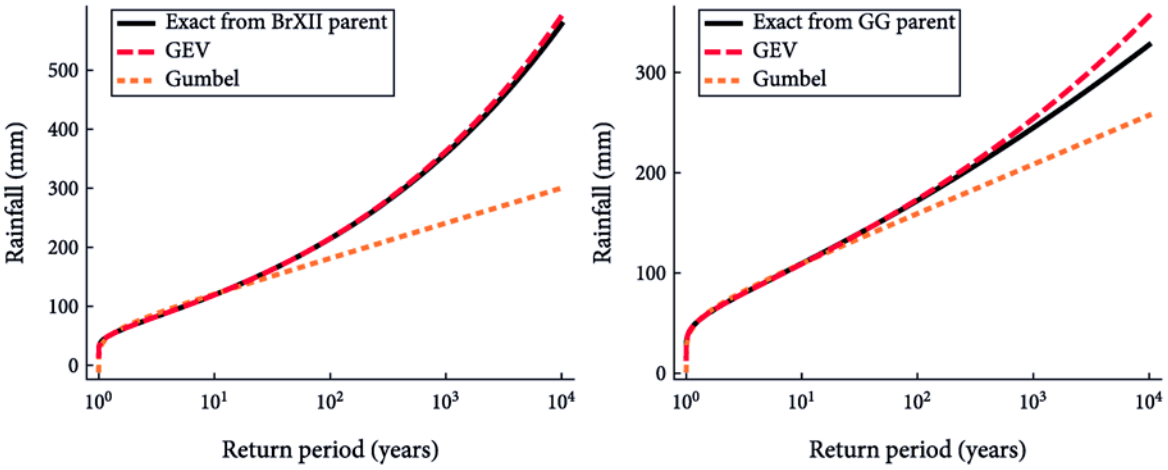



**Figure 5.1.** Demonstration of the convergence of the true distribution of maxima to the limiting laws.

This analysis indicates that even if the parent distribution of daily rainfall is of exponential type, belonging thus theoretically to the domain of attraction of the Gumbel law, the annual maximum is better described by the Fréchet law [see also *Koutsoyiannis*, 2004a]. Is this a paradox? The answer is no. The reason is that the convergence to the Gumbel law is very slow; actually, it does not converge satisfactorily even for $n = 10^7$ as our tests showed. On the contrary, the additional shape parameter of the Fréchet law or of the GEV distribution, adds the required flexibility to this distribution to "imitate" the shape characteristics annual maxima even if the parent distribution does not belong to its domain of attraction. Thus, although the Fréchet law has a power type tail, its flexibility enables it to better describe, compared to Gumbel law, other heavy-type tails like the stretched exponential or the lognormal. Noteworthy, a recent study [*Papalexiou et al.*, 2013] where

more than 15 000 daily records were analysed focusing on the tail behaviour of the parent distribution, revealed that the daily rainfall tail is better described by heavy tails. This offers a theoretical argument favouring the use of the Fréchet law in any case instead of Gumbel.

## 5.3 The original dataset

This study uses more than 15 000 rainfall records distributed across the globe. The original data were daily rainfall records obtained from the Global Historical Climatology Network-Daily database (version 2.60, www.ncdc.noaa.gov/oa/climate/ghcn-daily) which includes thousands of records worldwide. It is mentioned though, that many records of this database have a large percentage of missing values, are short in length, e.g., just a few years, or, contain suspicious values in terms of quality (for the quality flags used refer to the aforementioned website).

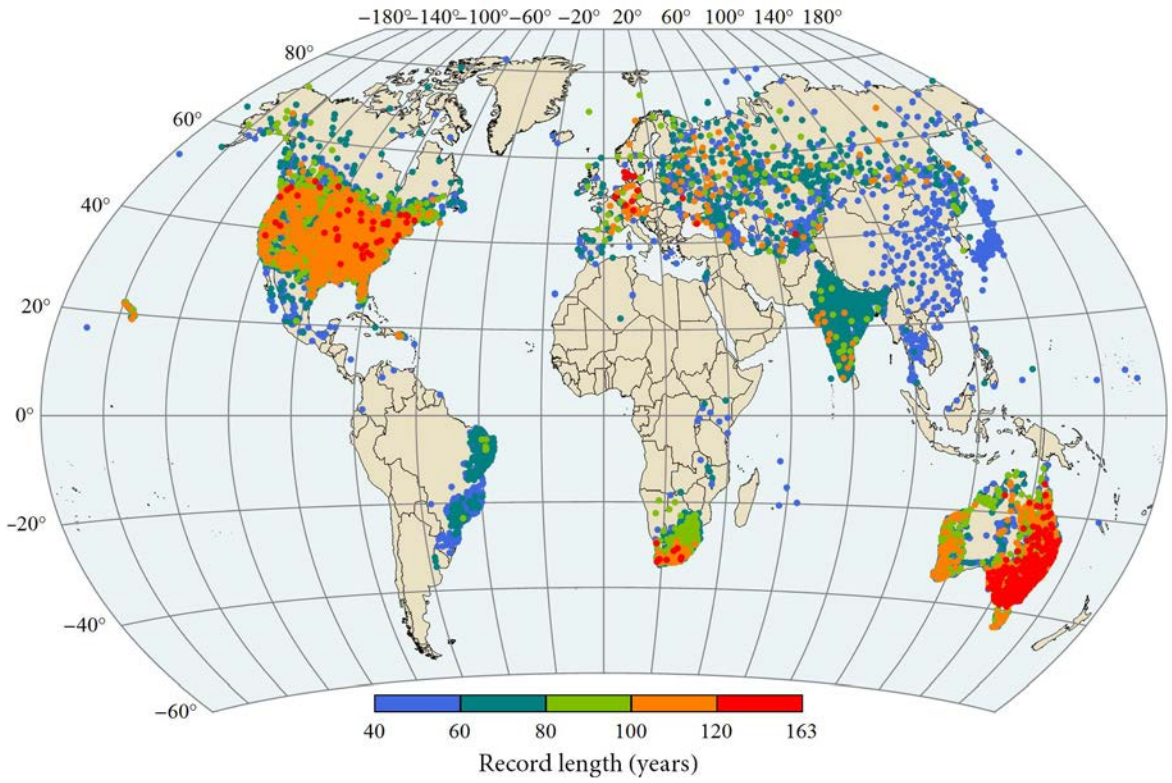



**Figure 5.2.** Locations of the 15 137 stations with annual maximum records of daily rainfall analysed with number of values ranging from 40 to 163 years. Note that there are overlaps with points corresponding to high record lengths shadowing (being plotted in front of) points of lower record lengths.

Thus, among the several thousands of records studied only those satisfying the following criteria: (a) record length greater or equal than 50 years, (b) percentage of missing values per record less than 20%, and (c) percentage of values assigned with "quality

flags" per record less than 0.1%. Special attention was given to values assigned with quality flags "G" (failed gap check) or "X" (failed bounds check) as these values are suspiciously large, e.g., could be orders of magnitude larger compared to the record's second larger value. These extremely large values (probably resulting from recording or registering errors), could alter the record's statistics, and thus they had to be identified and deleted (yet, only 594 records contained such values and typically one or two values at each record had to be deleted). The resulted number of records after screening with these criteria is 15 137 (for further details on the dataset please Appendix B). The locations of those records are depicted in the map given in Figure 5.2.

## 5.4 A method for extracting the maxima

### 5.4.1 Selection procedure

The original dataset comprises daily rainfall records, thus, in order to study the annual maximum daily rainfall the time series of annual maxima had to be formed. If the original records did not contain any missing-values then forming the annual maximum time series would be trivial. Yet, missing-values occur commonly, and specifically, in the dataset analysed here records may contain up to 20% of missing-values. Usually, within a record only some years are incomplete, (contain missing-values); hence, the problem is how is it possible to extract the maximum value of incomplete years. Evidently, the recorded maximum value of an incomplete year may not be the real one, as it is likely for a larger value to have occurred in days of missing data. Moreover, as the percentage of missing values gets higher the more probable it becomes that the real maximum has been recorded. Thus, years with missing values, if not treated appropriately, could result in significant errors that may affect the conclusions drawn from the data analysis.

Basically, one could think of three different methods to extract the annual maxima from a daily time series containing missing values: (a) in the first method (M1), specific criteria are used to assess the validity of the annual maxima, e.g., the annual maximum value could be considered valid only if the missing-values percentage is small, (b) in the second method (M2), only the maxima of complete years are accepted as valid while those of incomplete years are assumed unknown, and (c), in the third method (M3), the annual maxima are extracted irrespective of the years' missing-values percentage. Clearly, the method M3 is not safe because, if the missing-values percentage is high, it will result in underestimated maxima. Method M2 is safe and assures that the extracted maxima are the real ones, yet it does not fully utilize the available information. For example, a record may contain many years with just a few missing values per year; according to method M2 all

these years would be excluded, thus leading to an unjustifiably small sample. So, it is clear that the most reasonable choice is to set some criteria that need to be fulfilled in order to accept an extracted annual maximum as valid.

It is reasonable to assume that it is safe to extract the annual maximum of those years with small missing-values percentage. Nevertheless, two problems arise. First, the definition of "small" would be subjective, e.g., 1% or 10% could be considered small, and second and most important, maxima of incomplete years may be much greater compared to those of complete years. For example, a year with 90% of missing values may contain the record's maximum; would it be rational to exclude this value? Of course, larger values may have occurred within an incomplete year but this would be unlikely. For these reasons it is rational to assume that the acceptance or not of a value extracted from an incomplete year, as the annual maximum, should be based on two criteria; first, on the missing-values percentage, and second, on the value's rank, i.e., its relative position in the extracted sample of maxima after it has been sorted in ascending order (the smallest rank is given to the smallest value).

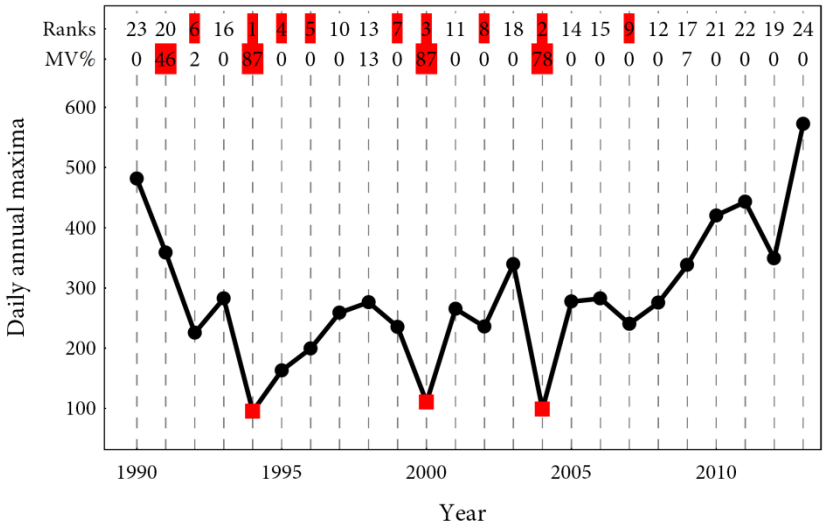



**Figure 5.3.** Explanatory plot of the maxima extraction method. The annual maximum daily rainfall is considered unknown (red rectangles) if its rank is in the smaller 40% of ranks (red shaded ranks) and the missing-value percentage (MV%) of the year it belongs is larger than 1/3 (red shaded percentages).

Accordingly, the annual maxima time series are formed in two steps: (a) the maximum of each year is extracted irrespective of the year's missing-values percentage and, (b) the values of this initial series are tested according to the criteria set and those not fulfilling them are deleted from the time series, i.e., they are assumed unknown. Namely,

two criteria, whose validity is justified in section 4.3, were set to justify deletion of a value whenever both hold: (a) the rank is smaller or equal than 40% × $N$ (where $N$ is the sample size) which means that the particular value belongs to the 40% of the lowest values, and (b) the missing-values percentage within a year is larger than or equal to 1/3 which means that in the particular year approximately the values of more than four months are missing. The method is graphically explained in Figure 5.3 which depicts along with the annual maxima time series the corresponding percentages and ranks of missing values. Essentially, the method's rationale is simple; if an incomplete year has a high percentage of missing values and its maximum is small compared to the maxima of the other years, then there is a high probability for larger values to have occurred within this year and thus this value should not be accepted as the real annual maximum.

### 5.4.2 Validation of the method

One could argue that the criteria defined previously are subjective and different values could be set as thresholds both for the rank and percentage of the missing values. Yet, these thresholds where not selected unjustifiably, but rather emerged after extended Monte Carlo simulations. Particularly, a Monte Carlo scheme was planned and performed in order to validate the method performance and specify the appropriate criteria values. The Monte Carlo scheme could be summarized in four basic steps: (a) a subset of complete daily records is selected and the annual maxima series are created, (b) this daily-records subset is modified to contain missing values, (c) annual maxima series are extracted from the modified daily-records subset by utilizing the maxima extraction method for various criteria values, and (d) the real maxima series created in step (a) are compared with those created in step (c). In other words, the basic idea is to find, if possible, those threshold values resulting in maxima series with statistical characteristics similar to the real ones.

Obviously, to validate the method complete daily time series are needed. Yet only few records of the dataset are totally complete, hence, for start only those with very small missing-values percentage were selected, i.e., less than 0.1%, while the few incomplete years per record, if existed, were deleted in order to be absolutely certain for the resulting annual maxima series. The result was 1 003 daily rainfall records with lengths varying from 38 to 155 years.

Now, the records of the dataset analysed here contain missing-values up to 20%, and these values are distributed among some of the record's years, i.e., only a percentage of the record's years are incomplete. To identify how the percentage of incomplete years per record is distributed all 15 137 records were studied. The empirical distribution is

presented in Figure 5.4, as well as a fitted Beta($\alpha$,$\beta$) distribution, that will be valuable in the sequel, with estimated parameters $\alpha = 1.32$ and $\beta = 2.41$.

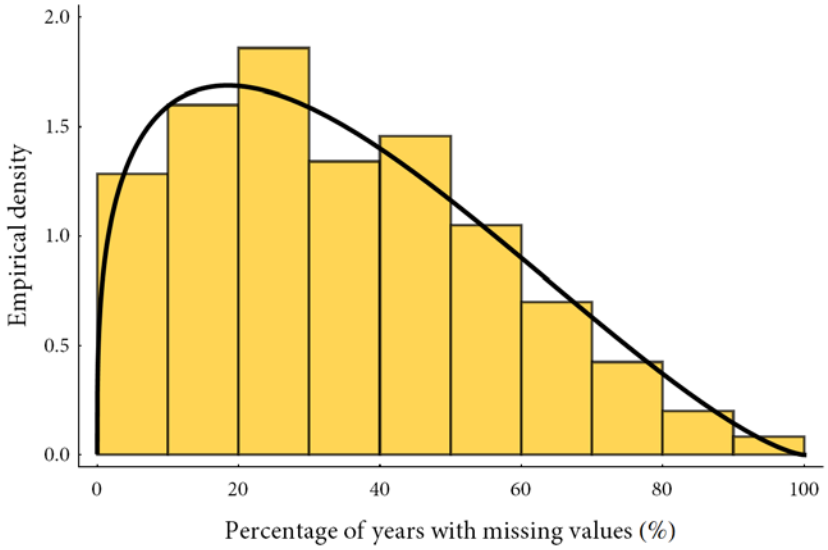



**Figure 5.4.** Empirical distribution of the year's percentage per record having missing values as resulted from the analysis of the 15 137 records; the solid line depicts a fitted Beta distribution.

In order to construct time series with missing values distributed similar to the real ones each one of the aforementioned daily records was modified by the following procedure: (a) a random number $p_{MV}$ less than 20% that represents the missing-values percentage of the record was generated, (b) the record's total missing-values number is then defined as $n_{MV} = p_{MV} \times 365 \times N$, where $N$ is the record's length in years, (c) the $n_{MV}$ missing values is distributed to $N_{MV} = p_Y \times N \geq n_{MV} / 365$ years, where $p_Y$ is the percentage of incomplete years which was randomly generated from the fitted Beta distribution depicted in Figure 5.4, (d) the number $n_{MV}$ is randomly split into $N_{MV}$ parts in order to define the number of missing values for each incomplete year, and (e) $N_{MV}$ years were randomly selected from the record and the number of values previously defined were randomly deleted from each year.

Finally, the annual maxima series extracted by the modified records were compared to the corresponding real ones based on four basic statistics, i.e., the mean as a measure of central tendency, the L-variation as a measure of dispersion, and the L-skewness and L-kurtosis as measures of shape characteristics. The maxima extraction method (M1) was repeatedly applied by altering the criteria values until the resulting series were statistically similar to the real ones; this led to the aforementioned threshold values. The maxima series extracted by methods M2 and M3 were also compared to the real ones. Figure 5.5 presents

the box plots formed by the 1 003 differences between the statistics of the real annual maxima series and the ones extracted from the daily series modified to contain missing values.

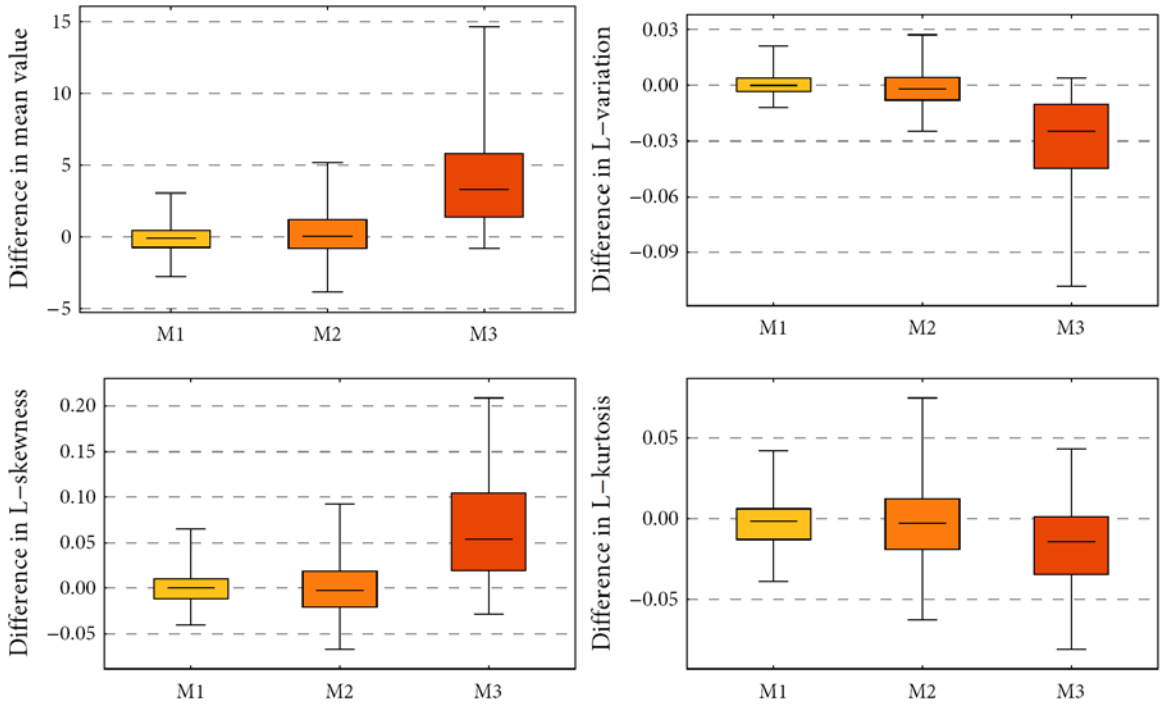



**Figure 5.5.** Box plots depicting the resulting sample differences of various statistics between the real annual maxima series and the ones created from the incomplete daily series. The advantage of the first method compared to the others is clearly seen by the smaller range of the box plots. The lower and upper fences of the box plots represent the sample quantiles $Q_1$ and $Q_{99}$, respectively.

As expected, method M3 (the one in which maxima are extracted irrespective of the percentage of missing-values) is inappropriate because it significantly alters the statistical character of the extracted maxima series while method M2 does not. Interestingly, not only does method M1 preserve the statistical characteristics (the median is zero and approximately equals the mean as the box plots are almost symmetric) but performs better than method M2. The explanation is that method M1 generates time series with larger length, compared to those of method M2, as fewer values are deleted. Apparently, larger time series means more information and thus more accurate sample estimates. Finally, it is worth noting that the overall range of the differences, taking into account that sample estimates of shape characteristics are usually very uncertain, is very small.

## 5.5 Analysis and results

### 5.5.1 Fitting results

The application of the maxima extraction method (it is noted that the annual maximum value is determined per calendar year, which is a more appropriate time basis for a study of global rainfall) produced 15 137 annual maximum daily rainfall time series with length varying from 40 to 163 years. To obtain a general idea of the statistical behaviour of the annual maximum daily rainfall the basic summary statistics for all records of maxima were calculated. The results are given in Table 5.1. Noteworthy, all statistical characteristics (mean, standard deviation, skewness, L-skewness, L-kurtosis) vary significantly; for example, the mean ranges, from 9.1 mm to 863.7 mm and the standard deviation from 3.9 mm to 430.7 mm. In particular, the large variation of shape characteristics indicates that any distribution with fixed shape will be inadequate for describing the annual maximum daily rainfall. Consequently, this portends the Gumbel distribution's inability as a universal model as its shape characteristics are fixed.

**Table 5.1.** Basic summary statistics of the 15 137 records; *Q* indicates the empirical quantile.

| | Record Length | Median | Mean | SD | Skew | L-scale $\lambda_2$ | L-skew $\tau_3$ | L-kurtosis $\tau_4$ |
|---|---|---|---|---|---|---|---|---|
| min | 40 | 7.40 | 9.10 | 3.94 | -0.71 | 2.15 | -0.16 | -0.06 |
| $Q_5$ | 49 | 25.60 | 28.51 | 11.00 | 0.53 | 5.80 | 0.10 | 0.09 |
| $Q_{25}$ | 58 | 39.20 | 43.13 | 17.41 | 0.98 | 9.06 | 0.18 | 0.14 |
| $Q_{50}$ | 68 | 57.20 | 62.24 | 23.73 | 1.35 | 12.35 | 0.23 | 0.18 |
| $Q_{75}$ | 91 | 77.50 | 83.96 | 33.84 | 1.84 | 17.43 | 0.28 | 0.22 |
| $Q_{95}$ | 117 | 114.80 | 126.23 | 57.81 | 3.03 | 29.86 | 0.37 | 0.30 |
| max | 163 | 864.50 | 863.69 | 430.69 | 9.87 | 244.66 | 0.76 | 0.73 |
| Mean | 74.85 | 61.97 | 67.73 | 27.72 | 1.51 | 14.40 | 0.23 | 0.18 |
| SD | 21.84 | 30.71 | 33.16 | 15.38 | 0.85 | 7.98 | 0.08 | 0.06 |
| Skew | 0.80 | 2.68 | 2.37 | 2.72 | 2.06 | 3.16 | 0.15 | 0.85 |
| L-scale $\lambda_2$ | 12.07 | 15.97 | 17.35 | 7.80 | 0.43 | 4.01 | 0.04 | 0.03 |
| L-skew $\tau_3$ | 0.22 | 0.19 | 0.20 | 0.27 | 0.23 | 0.28 | 0.02 | 0.10 |

It could be expected that in some cases the Gumbel distribution suits better, while in other cases the Fréchet, or, even the reversed Weibull are more appropriate; in fact all three distributions have been used in the literature. Theoretically, the estimated shape parameter of a fitted GEV distribution reveals which one of the three distributions performs better, as all of them emerge for specific values of $\gamma$. Yet, the Gumbel distribution arises for $\gamma \to 0$, and thus, even if the sample is indeed drawn from a Gumbel distribution the estimated GEV shape parameter (irrespective of the fitting method used) will never be exactly zero. In the literature more than thirteen tests can be found for testing whether the estimated

GEV shape parameter can be assumed zero [*Hosking*, 1984]. Nevertheless, all these tests examine whether the null hypothesis $H_0$: $\gamma = 0$ can be rejected or not. Clearly, a sample not rejecting the null hypothesis does not imply that $\gamma = 0$, or equally, that the underlying distribution is the Gumbel. It is highly probable for a null hypothesis with small values of $\gamma$, e.g., $H_0$: $\gamma = -0.01$, or, $H_0$: $\gamma = 0.01$, not to be rejected. Hence, it is reasonable to assume that it is not possible to conclude with certainty based on statistical tests whether the underlying distribution is Gumbel or GEV with $\gamma$ close to zero.

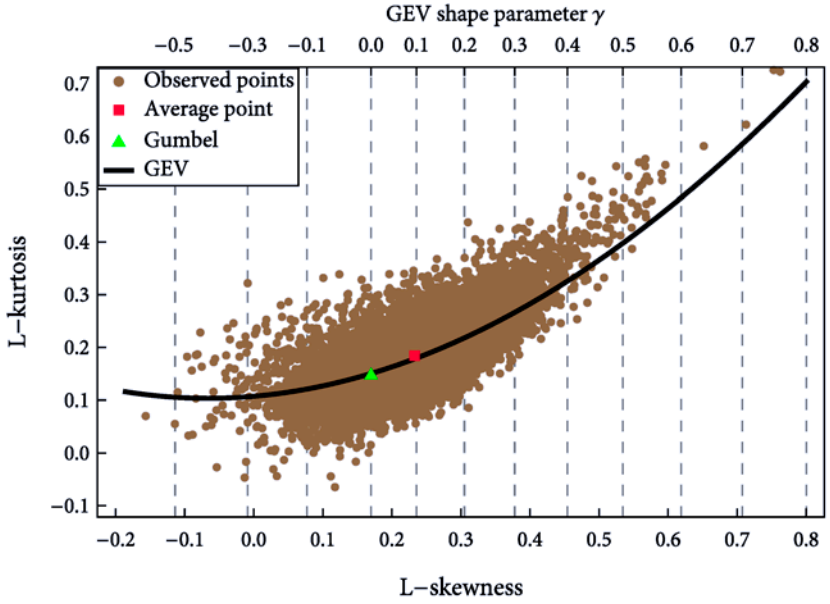



**Figure 5.6.** Observed L-kurtosis *vs*. L-skewness points of the 15 137 annual maximum daily rainfall records and the theoretical point and line of the Gumbel and GEV distribution, respectively.

Nevertheless, apart from the aforementioned tests, graphical tools exist that are especially useful when dealing with a large number of records, which can help to make inference about the underlying distribution. A graphical tool that has gained popularity over the last decade, introduced by Hosking [1990], is provided by the L-moments ratio diagrams. L-ratio plots have superseded classical moments ratio plots as they are superior in many aspects [see e.g., *Hosking and Wallis*, 1993; *Hosking*, 1992; *Peel et al.*, 2001; *Vogel and Fennessey*, 1993]. Essentially, this tool provides a graphical comparison between observed L-ratio values and points or lines or even areas formed by the theoretical formulas of parametric distributions. Figure 5.6 depicts in an L-kurtosis *vs*. L-skewness plot the 15 137 observed points as well as the theoretical point and line corresponding to the Gumbel and the GEV distributions, respectively. Interestingly, only 20% of points lie on the left of the Gumbel distribution (corresponding to a GEV distribution with $\gamma < 0$;

reversed Weibull law), while 80% of points lie on the right (corresponding to a GEV distribution with $\gamma > 0$; Fréchet law). Also it is worth noting that the average point lies almost exactly on the GEV line and corresponds to $\gamma \approx 0.1$. Figure 5.6 may not reveal the percentage of points that could be described by a Gumbel distribution, yet, it offers a clear indication that the Fréchet law prevails.

**Table 5.2.** Summary statistics of the estimated parameter of the fitted Gumbel and GEV distributions to the 15 137 annual maximum daily rainfall records; the fitting was done by the method of L-moments.

| | Gumbel parameters | | GEV parameters | | |
|---|---|---|---|---|---|
| | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\gamma$ |
| min | 6.81 | 3.10 | 6.00 | 2.66 | -0.587 |
| $Q_5$ | 23.21 | 8.37 | 22.59 | 7.36 | -0.107 |
| $Q_{25}$ | 35.26 | 13.07 | 34.67 | 11.71 | 0.020 |
| $Q_{50}$ | 51.54 | 17.82 | 50.82 | 16.16 | 0.093 |
| $Q_{75}$ | 70.07 | 25.15 | 69.24 | 22.69 | 0.169 |
| $Q_{95}$ | 102.54 | 43.09 | 101.14 | 38.53 | 0.283 |
| max | 659.96 | 352.97 | 688.17 | 401.68 | 0.760 |
| Mean | 55.74 | 20.77 | 54.95 | 18.71 | 0.092 |
| SD | 27.21 | 11.51 | 27.08 | 10.68 | 0.120 |
| Skew | 2.23 | 3.16 | 2.38 | 4.67 | -0.130 |
| L-scale $\lambda_2$ | 14.30 | 5.78 | 14.17 | 5.25 | 0.067 |
| L-skewness $\tau_3$ | 0.18 | 0.28 | 0.18 | 0.27 | -0.017 |
| L-kurtosis $\tau_4$ | 0.13 | 0.18 | 0.14 | 0.18 | 0.158 |

As mentioned before, the GEV shape parameter value indicates the type of the limiting law, a fact that emphasizes the importance to study in depth the behaviour of this parameter. To this aim, the GEV distribution was fitted to all available records, and for the completeness of the analysis the Gumbel distribution was also fitted. Both distributions were fitted using the method of L-moments [see e.g., *Hosking*, 1990], as especially for the GEV distribution it has been shown [*Hosking et al.*, 1985] that L-moments estimators are even better than maximum likelihood estimators in terms of bias and variance for samples up to 100 values. The fitting results are shown in Table 5.2 where various summary statistics of the estimated parameters are given. The table shows the large variation of the estimated GEV shape parameter, which ranges from −0.59 to 0.76 with mean value 0.093; the 90% empirical confidence interval is evidently much smaller, i.e., form −0.11 to 0.28 . The empirical distribution of the GEV shape parameter is depicted on Figure 5.7 along with a fitted normal distribution with mean 0.093 and standard deviation 0.12.

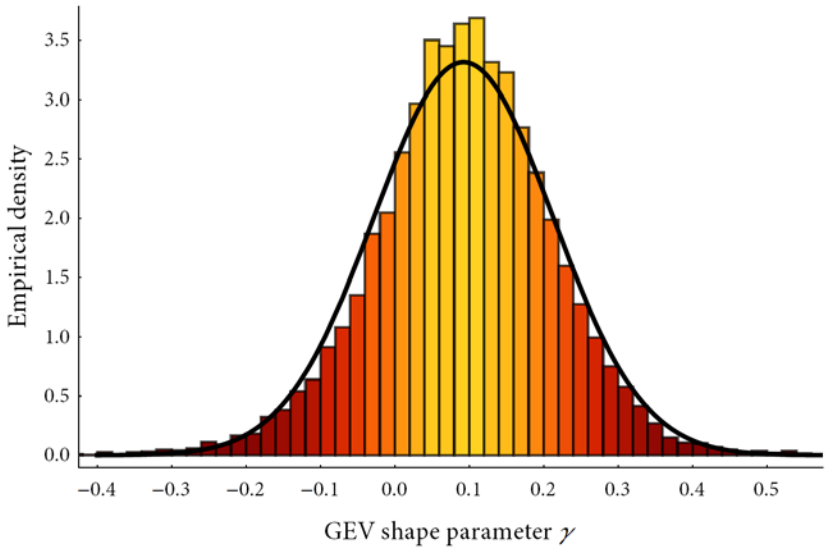



**Figure 5.7.** Empirical distribution of the GEV shape parameter as resulted by fitting the GEV distribution to the 15 137 annual maximum daily rainfall records. The solid line depicts a fitted normal distribution.

### 5.5.2 GEV shape parameter *vs.* record length

Larger samples offer more accurate estimates because, obviously, the variance of an estimator decreases as the sample size gets larger. Unambiguously thus, the estimate of the GEV shape parameter is expected to be more accurate if based for example on a 100-year record rather than on a ten-year record. In this respect, the estimated GEV shape parameter was studied in relationship with the record length as records vary in length from 40 to 163 years. First, the 15 137 estimated shape parameter values were gathered into nine groups based on the length of the record that were estimated; and second, various statistics were estimated for each group. The summary statistics of each group are given in Table 5.3, while the mean value and the percentage of records with positive shape parameter in each group are depicted in Figure 5.8. Clearly, Figure 5.8 indicates an upward "trend" in the mean shape parameter value over record length, e.g., for the 40-50 years group the mean value of $\gamma$ is 0.077 while for the last group (with ≥ 121 years) it is markedly larger, i.e., 0.116. Additionally, as the values of Table 5.3 attest, the standard deviation, as expected, decreases over the record length, e.g., for the 40-50 years group it is 0.141 while for the one with ≥ 121 years it is 0.088. Obviously the smaller the standard deviation the smaller the parameter range, yet a drastic decrease is observed, e.g., in the 90% empirical confidence interval (ECI) of $\gamma$, which for the 40-50 years group is $[-0.152, 0.312]$ while for the one with ≥ 121 years it is $[-0.029, 0.263]$. Another key issue to emphasize is the upward "trend" of the percentage of positive $\gamma$ over record length. This percentage is large (71.8%) even in

the 40-50 years and for the group with ≥ 121 years it gets as high as 91.0%, providing a clear indication that the Fréchet law prevails.

**Table 5.3.** Summary statistics of the estimated GEV shape parameter for various record length categories.

| Record length (years) | 40 - 50 | 51-60 | 61-70 | 71-80 | 81-90 | 91-100 | 101-110 | 110-120 | ≥ 121 |
|---|---|---|---|---|---|---|---|---|---|
| Records No. | 1 161 | 3 610 | 3 972 | 1 467 | 1 134 | 1 164 | 1 132 | 1 017 | 480 |
| Records % ($\gamma > 0$) | 71.8 | 72.9 | 77.8 | 83.6 | 85.0 | 86.8 | 88.1 | 91.1 | 91.0 |
| Records % ($\gamma \leq 0$) | 28.2 | 27.1 | 22.2 | 16.4 | 15.0 | 13.2 | 11.9 | 8.9 | 9.0 |
| | GEV shape parameter $\gamma$ | | | | | | | | |
| min | -0.461 | -0.587 | -0.493 | -0.307 | -0.287 | -0.283 | -0.188 | -0.193 | -0.204 |
| $Q_5$ | -0.152 | -0.156 | -0.112 | -0.086 | -0.068 | -0.048 | -0.046 | -0.035 | -0.029 |
| $Q_{25}$ | -0.014 | -0.009 | 0.011 | 0.030 | 0.036 | 0.042 | 0.049 | 0.047 | 0.060 |
| $Q_{50}$ | 0.079 | 0.082 | 0.086 | 0.102 | 0.100 | 0.106 | 0.108 | 0.102 | 0.118 |
| $Q_{75}$ | 0.172 | 0.166 | 0.166 | 0.176 | 0.169 | 0.175 | 0.169 | 0.158 | 0.170 |
| $Q_{95}$ | 0.312 | 0.290 | 0.291 | 0.285 | 0.268 | 0.271 | 0.271 | 0.247 | 0.263 |
| max | 0.541 | 0.706 | 0.760 | 0.567 | 0.539 | 0.573 | 0.750 | 0.471 | 0.345 |
| Mean | 0.077 | 0.077 | 0.089 | 0.103 | 0.101 | 0.108 | 0.110 | 0.105 | 0.116 |
| SD | 0.141 | 0.138 | 0.124 | 0.112 | 0.102 | 0.100 | 0.096 | 0.088 | 0.088 |
| Skew | -0.135 | -0.253 | 0.120 | 0.096 | -0.029 | 0.171 | 0.367 | 0.220 | -0.137 |
| L-scale $\lambda_2$ | 0.079 | 0.077 | 0.069 | 0.063 | 0.057 | 0.056 | 0.053 | 0.048 | 0.049 |
| L-skewness $\tau_3$ | -0.012 | -0.034 | 0.015 | 0.006 | 0.002 | 0.014 | 0.023 | 0.024 | -0.011 |
| L-kurtosis $\tau_4$ | 0.142 | 0.149 | 0.153 | 0.134 | 0.135 | 0.137 | 0.144 | 0.166 | 0.156 |

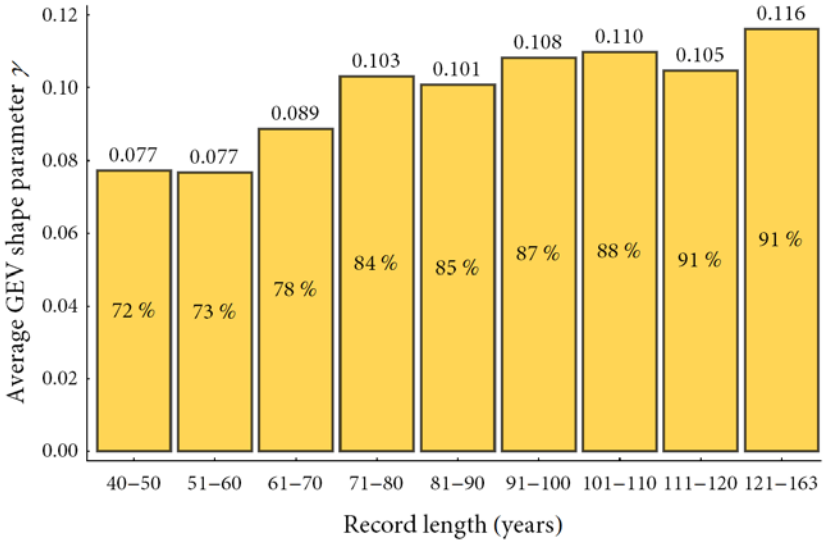



**Figure 5.8.** Mean value of the GEV shape parameter for various categories of record length. The numbers in the boxes indicates the percentage of records with positive shape parameter value.

The previous analysis gave a clear indication that a relationship between the estimated GEV shape parameter and the record length exists, yet, this relationship is not

exactly revealed as the variation in the mean value, as shown in Figure 5.8, does not suggest a precise law. Nevertheless, if such a law exists, it is clear that the previous grouping technique fails to reveal its exact form because the record length is not uniformly distributed within the groups (e.g., the 51-60 years group contains 3610 records but this does not imply that there are 361 records of 51 years, 361 records of 52 years, etc.). Thus, in order to create records with exactly the same length, the existing ones were modified by partitioning or cutting off a number of values. Specifically, only records with length greater or equal than 80 years were selected (5 049 records; it would be extremely laborious to use all records), and each one was partitioned into lengths ranging from ten to 115 years increased by a step of five years. The 115-year "upper limit" emerged by demanding at least 1000 records at each record length, a number that could be reasonably assumed large enough to offer a robust analysis (there are 1046 records with length ≥ 115 years and only 540 with length ≥ 120 years). For instance, applying this technique, a 112-year record is partitioned into eleven 10-year records or yields only one 90-year record and obviously none 115-year record. In total the 5 049 selected records generated, for example, 49 270 ten-year records and 1046 115-year records. For all these records and for each record length the GEV shape parameter was estimated using the L-moments method.

Figure 5.9a depicts the observed mean and the 95% confidence interval (CI) values of the GEV shape parameter for the various record lengths as well as the corresponding fitted theoretical functions. The fitted curves have the form $g(L) = a + bL^{-c}$, with $c > 0$, $L$ denoting the record length and $a$, $b$, $c$ parameters estimated here with a least square error fitting. This formula was figured out so as to have two desiderata: The first stems from the fact that the observed values indicate clearly that the mean and the CI values do not increase or decrease linearly over the record length. Rather, it is reasonable to assume that they tend asymptotically to a fixed value. Clearly, as $L \to \infty$ the function $g(x) \to a$ with $a$ thus expressing the limiting value. The second desideratum is this function to be simple and flexible. Indeed, for $b < 0$ it is concave and for $b > 0$ it is convex, thus being suitable to describe both upward and downward "trends" that converge to a liming value. The estimated parameters for the fitted curves are as follows: (a) for the lower CI curve, $a = 0.021$, $b = -3.90$, $c = 0.80$, (b) for the mean value curve, $a = 0.114$, $b = -0.69$, $c = 0.98$, and (c) for the upper CI curve, $a = 0.195$, $b = 1.29$, $c = 0.55$. Undoubtedly, Figure 5.9a indicates a perfect match of the fitted functions to the observed values, unveiling thus the underlying laws. Noteworthy, the 95% limiting CI is very narrow $(0.021, 0.195)$ with the lower bound positive, while the mean value of $\gamma$ converges to $\mu_\gamma \simeq 0.114$.

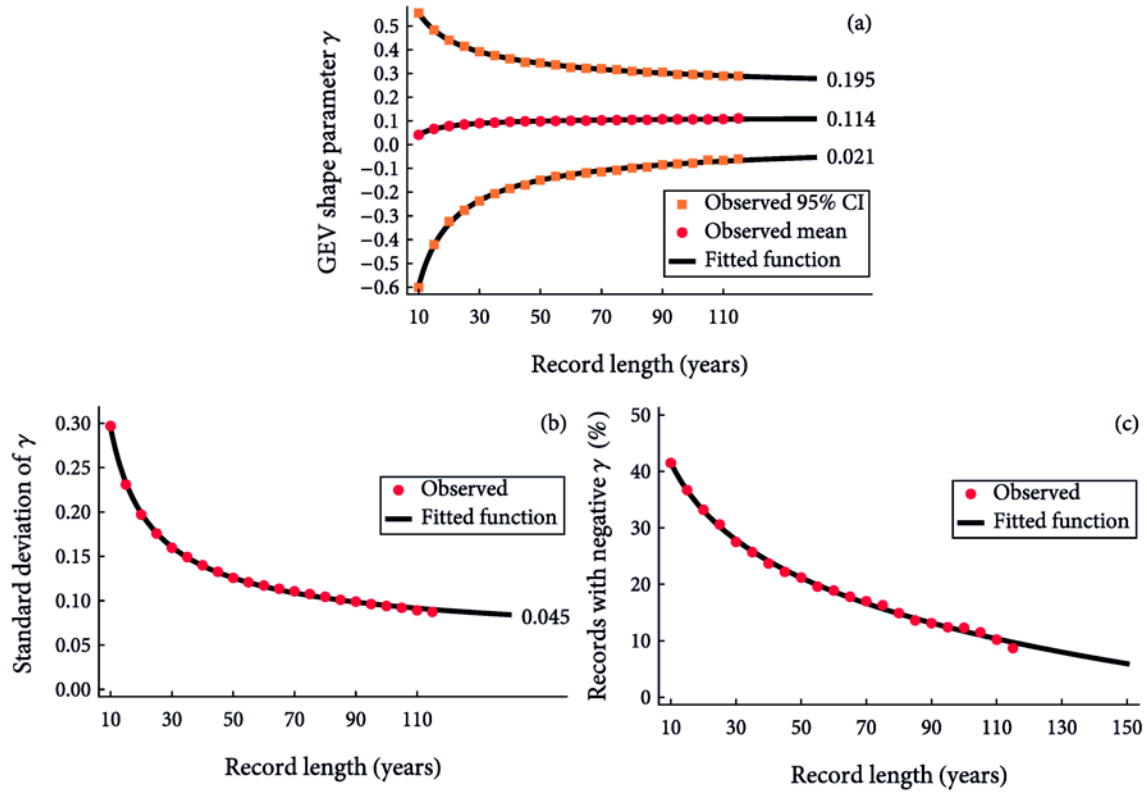



**Figure 5.9.** (a) Mean, quantiles $Q_5$ and $Q_{95}$ as estimated for various records lengths and their fitted asymptotic values; (b) standard deviation; (c) percentage of records with negative shape parameter.

In order to identify the true underlying distribution of the GEV shape parameter (assuming it is well approximated by a normal distribution), apart from the limiting mean value estimated before, estimation of the limiting value of the standard deviation is also necessary. Figure 5.9b depicts the estimated standard deviation values versus record length and a fitted curve of the same form used for the mean. The estimated parameters of the fitted curve are $a = 0.045$, $b = 1.27$ and $c = 0.70$, indicating thus that the true standard deviation of $\gamma$ is $\sigma_\gamma \simeq 0.045$, a value significantly smaller than the smallest observed. Interestingly, assuming that the shape parameter follows the estimated normal distribution, i.e., $\gamma \sim N(\mu_\gamma, \sigma_\gamma^2)$, the 95% CI of $\gamma$ would be (0.03, 0.21) which is very close to the limiting CI estimated and depicted in Figure 5.9a. Furthermore the 99% CI (rounded at the second decimal digit) is estimated at (0, 0.23), and apparently the probability for a negative shape parameter to occur is only 0.005.

Additionally, Figure 5.9c depicts the percentage of records with negative $\gamma$ over record length. Evidently, the observed points suggest a quickly non-linear decreasing "trend". The fitted curve has the same simple form as above but with $c < 0$. With estimated

parameters $a = 221.3$, $b = -154.1$, $c = -0.067$ it crosses the horizontal axis at $L = (-a/b)^{-1/c} \approx 226$ years, implying that for record length greater than 226 years the percentage of records with negative $\gamma$ would be zero. Indeed, none of the 16 records available with length greater than 140 years resulted in negative $\gamma$. This indicates a deviation from the fitted curve; yet, the number of stations for this record length is very small to take it into account but this is additional evidence that the Fréchet law prevails.

Finally, based on the previous findings, it is possible to create an "unbiased" or record-length-free estimator for the GEV shape parameter that incorporates its relation with the record length. Given that the true distribution of $\gamma$ is the $N(\mu_\gamma, \sigma_\gamma^2)$ while for specific record length $n$ is the $N(\mu_\gamma(n), \sigma_\gamma^2(n))$, with $\mu_\gamma(n) = \mu_\gamma - 0.69\, n^{-0.98}$ and $\sigma_\gamma(n) = \sigma_\gamma + 1.27\, n^{-0.70}$ being the functions fitted previously for the mean and the standard deviation, it can be easily proved that an "unbiased" estimator $\tilde{\gamma}(n)$ is the

$$\tilde{\gamma}(n) = \frac{\sigma_\gamma}{\sigma_\gamma(n)}\left(\hat{\gamma} - \mu_\gamma(n)\right) + \mu_\gamma \tag{5.8}$$

where $n$ is sample size (number of years), $\hat{\gamma}$ is the L-moments estimate of $\gamma$, whereas $\mu_\gamma \simeq 0.114$ and $\sigma_\gamma \simeq 0.045$ are the limiting mean and standard deviation values estimated previously.

### 5.5.3 Monte Carlo validation of the results

In order to validate our results regarding the underlying distribution of the GEV shape parameter a Monte Carlo simulation was performed. Specifically, 15 137 random samples were generated, with sizes precisely equal with the original records lengths, from a GEV distribution with the shape parameter being randomly generated from the anticipated normal distribution, i.e., the $N(\mu_\gamma, \sigma_\gamma^2)$, and with the location and scale parameter fixed to their mean values given in Table 5.2 as they do not affect the shape parameter estimates. In sequel, the shape parameter values of those samples were estimated and the empirical distribution shown in Figure 5.10 was formed. It is observed that while the prior distribution of $\gamma$ was the $N(\mu_\gamma, \sigma_\gamma^2)$ the estimated posterior is almost identical with the empirical distribution emerged from the real records given in Figure 5.7. The comparison of the two distributions reveals a very close match, i.e., the empirical distribution emerged from the real records has mean and the standard deviation, respectively, equal to 0.092 and

0.12 while the corresponding values for the empirical distribution emerged from the synthetic records are, respectively, 0.104 and 0.11.

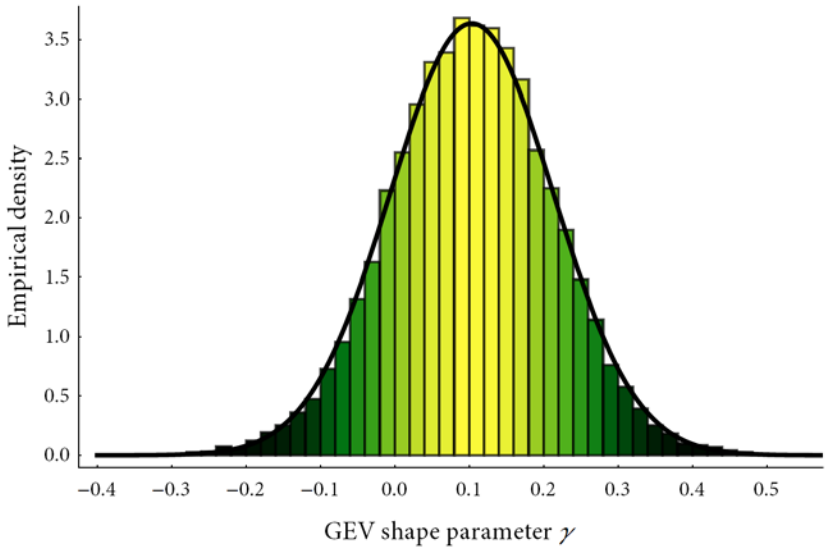



**Figure 5.10.** Empirical distribution of the GEV shape parameter as resulted from the Monte Carlo simulation where 15 137 synthetic records generated with the shape parameter being randomly sampled from the N($\mu_\gamma$, $\sigma_\gamma^2$). The solid line depicts the fitted normal distribution.

This minor deviation is probably justified by the fact that the L-skewness and the L-kurtosis of the empirical distribution of $\gamma$, which are −0.017 and 0.158, respectively, deviate slightly from the theoretical values of a normal distribution which are 0 and 0.123. The small negative skewness may have caused the slight decrease in the mean value while the higher L-kurtosis implies more extremes $\gamma$ values, both negative and positive, and this obviously leads to higher variance. The fact is that both the empirical evidence and the Monte Carlo simulation suggest that the distribution of the GEV shape parameter is very well approximated by the normal distribution N($\mu_\gamma$, $\sigma_\gamma^2$). Even if the shape characteristics between the empirical and the Monte Carlo distributions do not match exactly (mainly the L-kurtosis) this is something anticipated; when a set of 15 137 real-world records is analysed it is expected that some records may either contain incorrectly recorded values or some extraordinary events occurred, leading thus to unrealistically small or large shape parameter estimates. For example a couple or even one "extremely" extreme event in a relatively small sample, e.g., 40-60 years may alter significantly the value of L-skewness and consequently the estimate of the shape parameter $\gamma$ resulting thus in a distribution that may not describe realistically the behaviour of the rainfall in general. "Errors" of this kind are unavoidable as it is possible for a small sample to contain, e.g., the 1000-year event.

The previous analysis also indicated that the true mean value of the underlying distribution of the GEV shape parameter is $\mu_\gamma = 0.114$, markedly larger than zero, i.e. the value specifying the Gumbel distribution. This consequently leads us to assume that the Gumbel distribution is not a good model in general for annual maximum daily rainfall. Nevertheless, it does not reveal how bad or good the Gumbel model is if compared to the GEV model or more specifically to the Fréchet law. Obviously the GEV and the Gumbel distributions cannot be compared directly in the sense that the first one is a three-parameter model while the second one is a two-parameter model and a special case of the first one. For this reason it is valuable to compare the Gumbel distribution with a representative fixed-shape-parameter GEV distribution, i.e., a GEV with shape parameter equal to $\mu_\gamma = 0.114$.

Specifically, 15 137 random samples were generated, with sizes equal to those of the original records using: (a) a Gumbel distribution, and (b) a GEV distribution with $\gamma = 0.114$ (the location and scale parameters were fixed in both distributions as their values do not affect the shape characteristics). Next, the Monte Carlo (MC) L-kurtosis *vs.* L-skewness points were estimated and depicted them in comparison with the observed ones already presented in Figure 5.6. The idea is to compare the extent of the area formed by the MC points with the area formed by the points of the real records.

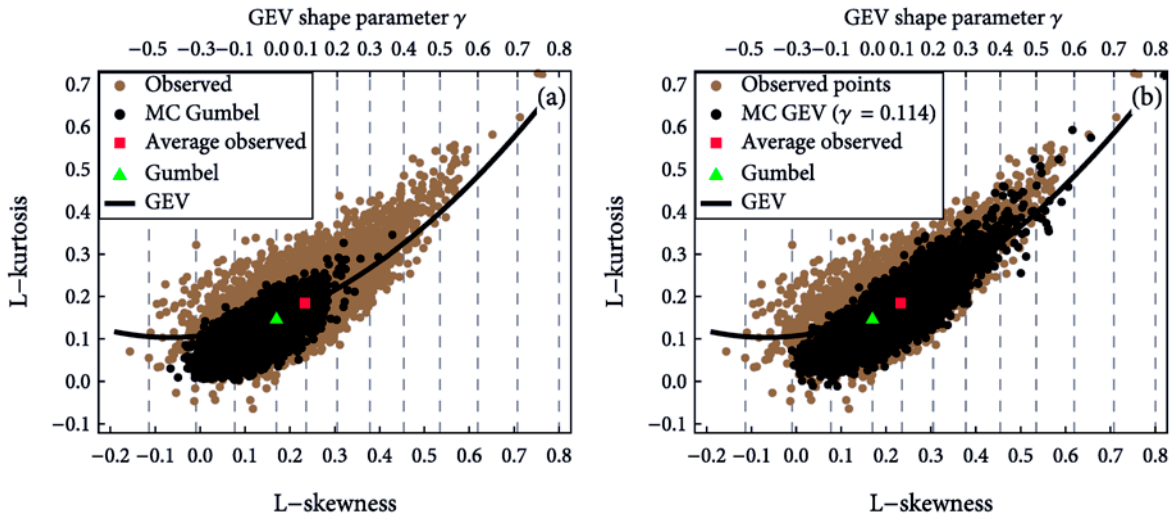



**Figure 5.11.** Monte Carlo points estimated (a) for the Gumbel distribution, and (b) for the GEV distribution with fixed shape parameter $\gamma = 0.114$, depicted in comparison to the observed ones.

The results of this Monte Carlo simulation are presented in Figure 5.11. For the Gumbel case (left graph) it is observed that indeed there is a spread around the theoretical Gumbel point, yet, the area covered by the MC points is significantly smaller than the one

formed by the observed points and the cloud of points are placed toward the left. Clearly, the Gumbel distribution fails to generate points with high values of L-skewness. In the GEV case with fixed $\gamma$ (right graph) it is observed that not only the expected shift of the cloud of the MC points toward the right, but also the expansion of this cloud, so that the area formed is much larger compared to that of the Gumbel case. In addition, the MC area better fits the one formed by the empirical points. This reveals that the GEV distribution with fixed $\gamma$ performs in general much better compared with the Gumbel distribution.

### 5.5.4 Geographical variation of the GEV shape parameter

The previous analysis reveals that the GEV shape parameter estimates depend on the record length and that essentially the parameter varies in the interval $(0, 0.23)$. Thus, the question that naturally arises is how the parameter varies over geographical location, as it is reasonable to expect that different areas of the world exhibit different behaviour not only in the mean annual rainfall but also the in the shape of distribution of the annual extremes. Yet it is stressed that that even if the behaviour of extreme rainfall is the same in a big area, in practice the estimated GEV shape parameters in different locations within the area will differ due to sampling effects. As a consequence, the different estimates may lead to false conclusions.

Thus, in order to reduce the sampling effect and to investigate the geographical distribution of the GEV shape parameter seeking to reveal any kind of geographical pattern, the earth's surface was divided into cells and the mean value of the GEV shape parameter within the cells was studied; obviously the mean value offers a simple and rational smoothing. Each cell is defined by a latitude difference of $\Delta\varphi = 2.5°$ and longitude difference of $\Delta\lambda = 5°$; as latitude $\varphi$ ranges from $-90°$ to $90°$ and longitude $\lambda$ from $-180°$ to $180°$, a total of 5 184 cells emerged. The mean value of the GEV shape parameter of each cell is simply estimated as the average of those shape parameter estimates that correspond to stations lying within the cell, given that the cell contains at least two records, Clearly, the number of stations within each cell is not constant, and most of the cells (notably those in the oceans) do not contain any stations while there are 258 cells containing only one record. Specifically, from the 5184 cells formed, only 792 cells had available records and only 534 had at least two records, while there are 46 cells with more than 100 records each.

The results using the typical (record-length dependent) estimates of the GEV shape parameter are depicted in the world map given in Figure 5.12 where the cell's mean value is expressed by colouring the cell according to the map's legend. It is noted that the values defining the bins in the map's legend are defined by the minimum value, the $Q_{10}$, $Q_{25}$, $Q_{50}$,

$Q_{75}$, and $Q_{90}$ empirical quantile (or percentile) points and the maximum value of the 534 mean shape parameter values after rounding off to the second decimal, e.g., the central 50% of values or the interquartile range is approximately form 0.06 to 0.14. The numbers of cells with mean values at each successive bin (from low to high values) are: 57, 76, 146, 115, 89 and 51, while the number of cells with negative mean values is 52. Clearly, the map reveals that large and discrete areas exist with the same behaviour in extreme rainfall manifested by the approximately equal GEV shape parameter values.

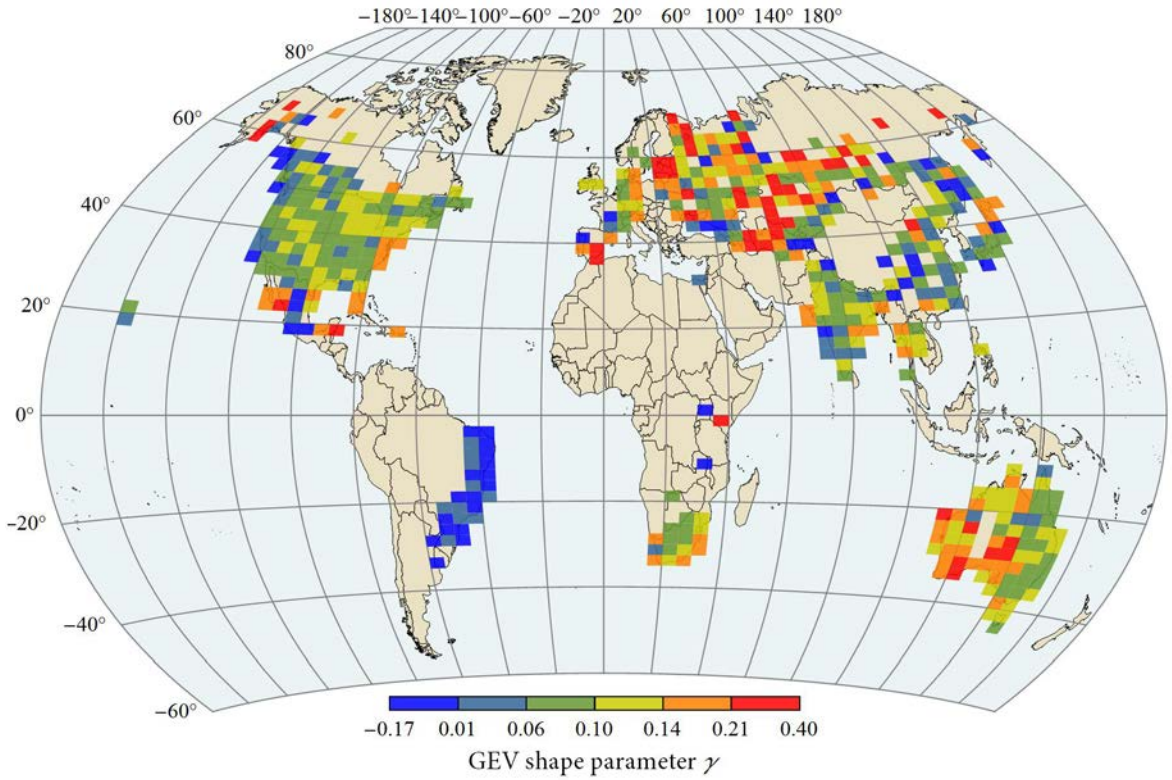



**Figure 5.12.** Geographical distribution of the mean value of the GEV shape parameter (estimated by the standard L-moment estimator) in regions of latitude difference $\Delta\varphi = 2.5°$ and longitude difference $\Delta\lambda = 5°$.

Nevertheless, the analysis of the previous section unveiled the clear relationship of the estimated GEV shape parameters with the record length. Consequently, a more accurate map should incorporate these findings as a region contains records of variable length leading thus to a record-length depended estimate of the mean value. Additionally, it was shown that the GEV shape parameter estimates can be corrected by Eq. (5.8) to be record-length free and follow the normal distribution $N(\mu_\gamma, \sigma_\gamma^2)$ which constitutes a very good approximation of the true distribution of the GEV shape parameter. For these reasons, a reconstructed map was formed by using the unbiased (free of record-length

dependence) estimate of the shape parameter values according to Eq. (5.8). The results are presented in Figure 5.13. As in the previous map, the bins are defined the same way but obviously the values differ as the range of variation is much smaller. The numbers of cells with values spotted in each successive bin are different from the previous map, i.e., 59, 88, 105, 143, 93 and 46 (due to rounding of the quantile values), while the number of points representing negative values is now zero. Comparing the two maps it is observed that they look almost the same but in fact they differ. Finally, it is notable that large areas or zones are formed by points representing shape parameter values belonging in a very narrow range. For example, in the US there are two large zones where the shape parameter ranges from 0.10 to 0.11 in the one (green colour) and from 0.11 to 0.13 in the other (yellow-green colour); additionally, in the entire Atlantic coasts of South America a zone of low values is formed while a large area of high values can be spotted in South-West Australia.

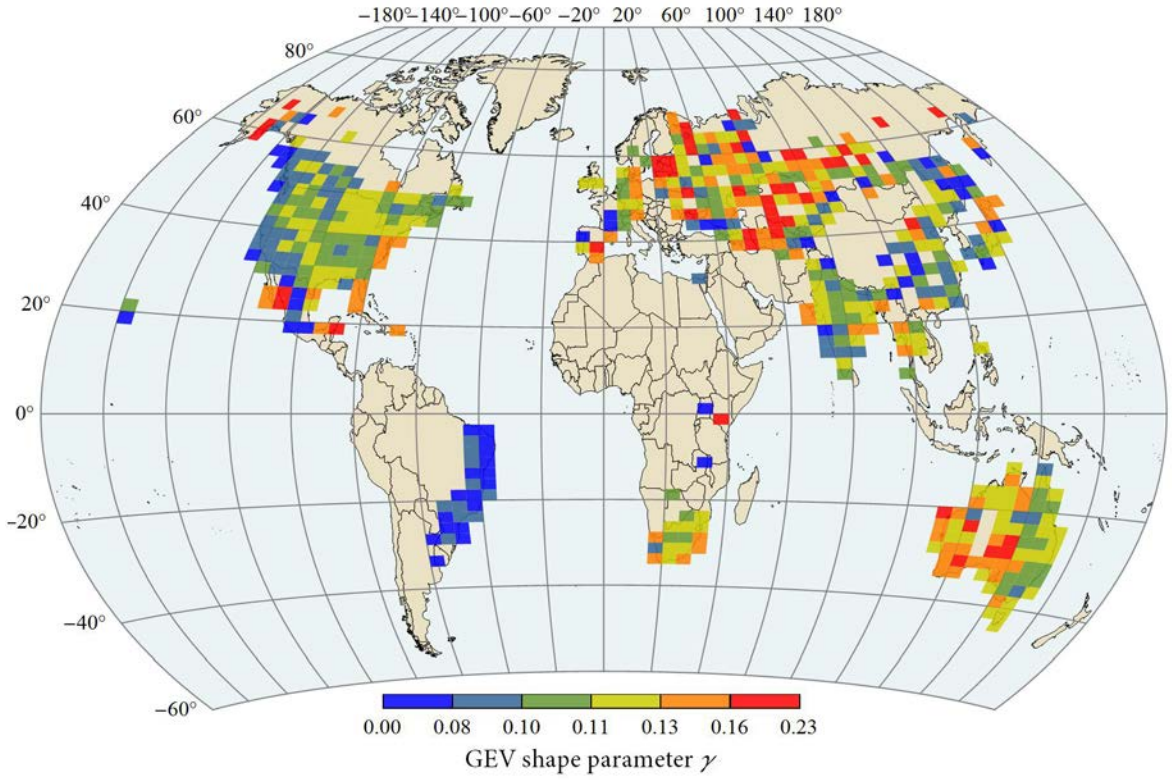



**Figure 5.13.** Geographical distribution of the mean value of the GEV shape parameters estimated by the unbiased estimator of Eq. (5.8) that corrects the sample-size effect; notice the difference in the values of the legend with the legend of Figure 5.12.

Obviously, the accuracy in the estimation of the shape parameter mean values is not the same for every cell as the number of records per cell is not constant. Thus, in order to provide a measure of uncertainty or a measure of estimation error, the map given in Figure

5.14 was constructed that presents each cell's standard error (SE) values with respect to the mean values given in the map Figure 5.13 (unbiased estimates). The SE is defined as $\mathrm{SE} = \sigma / \sqrt{n}$ and in this case $\sigma$ is the sample standard deviation of the shape parameter values of the cell and $n$ the number of those values. In order for the estimates of SE to be relatively accurate, only those cells that contain at least six records (a total of 281 cells) were selected, as it is well-known that the estimation of the standard deviation is markedly biased for very small samples. A cell's SE expresses the standard deviation of the cell's shape parameter mean value, and can be used directly to calculate the 95% CI of this estimate as it is well-known that the 95% CI is given by $\overline{\gamma} \pm 1.96\,\mathrm{SE}$, where $\overline{\gamma}$ is the cell's shape parameter mean value. The values defining the bins of SE in the map's legend (Figure 5.14) are defined by the minimum value, the $Q_{25}$, $Q_{50}$, $Q_{75}$ empirical quantile (or percentile) points and the maximum value of the 281 SE values after rounding off to the third decimal, e.g., the 50% of SE values are less than 0.008. The numbers of cells with SE values at each successive bin (from lower to higher values) are: 67, 75, 68, and 71. As expected, areas with high density of stations and large records have very low values of SE.

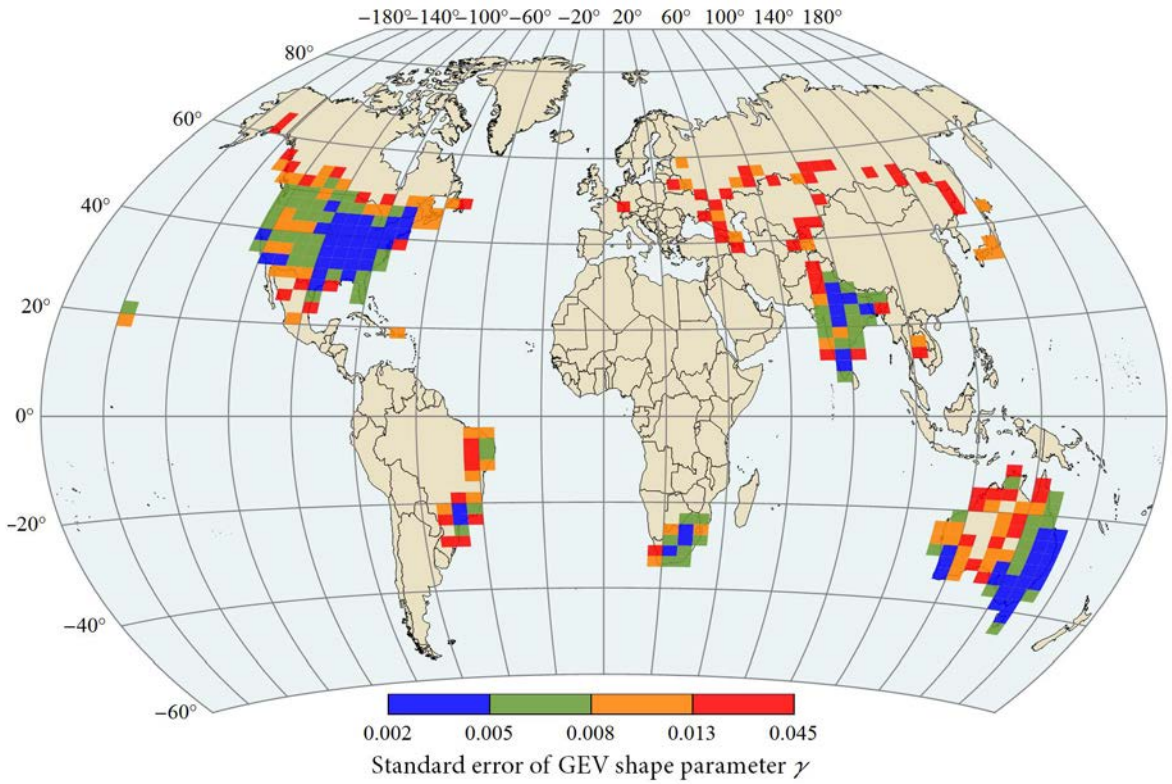



**Figure 5.14.** Standard error values of the GEV shape parameter mean values that are given in the map of Figure 5.13.

## 5.6 Summary and conclusions

Extreme value distributions have been extensively used in hydrology for more than half a century as a basic tool for estimating the design rainfall of infrastructures or assessing flood risks; however, selecting the appropriate law is usually based on small samples without guaranteeing the correct choice or the accurate estimate of the law's parameters. In this study, 15 137 rainfall records are analysed from all over the world aiming to assess which one of the three limiting distributions better describes the annual maximum daily rainfall. Initially, a method was formed comprising two simple criteria, in order to treat the very common problem of extracting annual maxima of daily rainfall from records containing missing values. The method was successfully validated and applied to form the annual maximum daily rainfall records.

The question, which of the three limiting extreme value distributions to use, is the focus of this study. Starting from the reversed Weibull distribution, it is noted that it implies a parent distribution for daily rainfall with an upper bound; a fact that seems to be physically inconsistent and moreover distributions bounded from above have never been used for daily rainfall in competent studies. With reference to the Fréchet *vs*. Gumbel "battle", it was shown that, as strange it may seem, annual maxima extracted from a parent distribution that belongs to the domain of attraction of the Gumbel law, are better described by the Fréchet law. This occurs for two reasons: first, the convergence rate to the Gumbel law is extremely slow, and second, the shape parameter of the Fréchet law enables the distribution to approximate quite well not only distributions with power-type tails but also other heavy-tailed distributions.

The empirical investigation using 15 137 records started with an L-moments ratio plot which reveals that 80% of observed points are located on the right of the "Gumbel point" providing clear evidence that the Fréchet law prevails. Additionally, the analysis of the estimated GEV shape parameters unveils a clear relationship between the shape parameter value over the record length, implying that only very large samples can reveal its true distribution or the true behaviour of the extreme rainfall. The "asymptotic" analysis performed, based on the fitted functions to the mean and standard deviation of the GEV shape parameter over record length, suggests that the distribution of the GEV shape parameter that would emerge if extremely large samples were available is approximately normal with mean value 0.114 and standard deviation 0.045. The meaning of this finding is that the GEV shape parameter is expected to belong in a narrow range, approximately from 0 to 0.23 with confidence 99%. Essentially, the analysis shows that data cannot be trusted

blindly, as small samples may distort the true picture. In this direction, the use of Eq. (5.8) is proposed that corrects the L-moments estimate of the GEV shape parameter removing the bias due to limited sample size.

While originally a small percentage of records have negative shape parameter (reversed Weibull law), the analysis reveals that this percentage rapidly decreases over sample size, while the fitted function indicates that for record length greater than 226 years this percentage would be zero. Interestingly, none of the 16 records available with length greater than 140 years resulted in negative $\gamma$. Moreover, the probability for a negative shape parameter to occur, according to the distribution fitted, is only 0.005, and combined with the previous findings suggests that a GEV distribution with negative shape parameter (bounded from above) is completely inappropriate for rainfall. Concerning the geographical distribution of the GEV shape parameter, the constructed maps show that large areas of the world share approximately the same GEV shape parameter, yet different areas of the world exhibit different behaviour in extremes.

It seems that the "verdict" is clear: the Fréchet law, or else the GEV law with positive shape parameter, should prevail over the Gumbel law and a fortiori over the reversed Weibull law, with latter suggesting a dangerous choice. If a rule of thumb had to be formed, then it would be this: even in the case where the data suggest a GEV distribution with negative shape parameter, it should not be used; instead it is more reasonable to use a Gumbel or, for additional safety, a GEV distribution with a shape parameter value equal to 0.114. The prevailing practice of the past that favoured the use of the Gumbel distribution does not suggest a proof of its outperformance over the Fréchet law, as it seems it takes a long time to reveal Nature's "secrets" and its true behaviour. As Heraclitus of Ephesus stated more than 2500 years ago in the aphorism given in the introduction (loosely translated) "Nature loves to hide".

# CHAPTER 6

*"Simplicity is the ultimate sophistication."*

LEONARDO DA VINCI

# CAN A SIMPLE RAINFALL MODEL MEET THE COMPLEX REALITY?

**ABSTRACT**

Several of the existing rainfall models involve diverse assumptions, a variety of uncertain parameters, complicated mechanistic structures, use of different model schemes for different time scales, and possibly classifications of rainfall patterns into different types. However, the parsimony of a model is recognized as an important desideratum as it improves its comprehensiveness, its applicability and possibly its predictive capacity. To investigate the question if a single and simple stochastic model can generate a plethora of temporal rainfall patterns, as well as to detect the major characteristics of such a model (if it exists), a dataset with very fine timescale rainfall is used. This is the well-known dataset of the University of Iowa comprising measurements of seven storm events at a temporal resolution of 5-10 seconds. Even though only seven such events have been observed, their diversity can help investigate these issues. An evident characteristic resulting from the stochastic analysis of the events is the scaling behaviours both in state and in time. Utilizing these behaviours, a stochastic model is constructed which can represent all rainfall events and all rich patterns, thus suggesting a positive reply to the above question. In addition, it seems that the most important characteristics of such a model are a power-type distribution tail and an asymptotic power-type autocorrelation function. Both power-type distribution tails and autocorrelation functions can be viewed as properties enhancing randomness and uncertainty, or entropy.

## 6.1 Introduction and motivation

Rainfall has been traditionally regarded as a random process with several peculiarities, mostly related to intermittency and non Gaussian behaviour. However, many have been not satisfied with the idea of a pure probabilistic or stochastic description of rainfall and favoured a deterministic modelling option. For example, *Eagleson* [1970] states "The spacing and sizing of individual events in the sequence is probabilistic, while the internal structure of a given storm may be largely deterministic". Such a perception of rainfall is also reflected in common engineering practices, such as the construction of design storms, in which the total depth may be determined by probabilistic considerations but the arrangement of rainfall depth increments follows a deterministic procedure, e.g. a pre-specified dimensionless hyetograph.

More recently, developments of nonlinear dynamical systems and chaos allowed many to apply algorithms from these disciplines in rainfall and claim for having discovered low dimensional deterministic dynamics in rainfall [*Puente and Sivakumar*, 2007; see e.g., *Sivakumar*, 2000]. However, such results have been disputed by others [*Koutsoyiannis*, 2006b; e.g., *Schertzer et al.*, 2002]. In the latter study, among other datasets, a high temporal resolution data record was used, in which the application of chaos detection algorithms did not give any indication of low dimensional chaos.

This high resolution record is one of seven storms that were measured by the Hydrometeorology Laboratory at the University of Iowa using devices that are capable of high sampling rates, once every 5 or 10 seconds [*Georgakakos et al.*, 1994]. This unique dataset allows inspection of the rainfall process at very fine time scales and was the subject of several extensive analyses including multifractal analysis and multiplicative cascades [*Cârsteanu and Foufoula-Georgiou*, 1996] and wavelet analysis [*Kumar and Foufoula-Georgiou*, 1997]. However, apart from such more technical analyses, this unique dataset offers a basis for simpler yet more fundamental investigations that could provide insights for the characterization and mathematical modelling of the rainfall process; this will be attempted in the next sections. In this respect, the Iowa dataset allows revisiting and acquiring better insight on the questions whether a single model can or cannot generate different types of events with enormous differences among them and, if yes, how such a model would look like. First, will it be deterministic or stochastic? A deterministic perception of the rainfall process may seem in accord to the high temporal dependence (autocorrelation) of the rainfall process at small lag times. However, this may indicate a misconception because au fond high autocorrelation without a specified underlying reason

(an a priori known deterministic control) may increase rather than reduce uncertainty [*Tyralis and Koutsoyiannis*, 2010] and thus may require a stochastic description. In the latter case, fundamental behaviours to be explored are (a) the long (e.g., power-law) or short (e.g., exponential) tails in probability distribution function and (b) the long or short tails of the autocorrelation function. In both cases, long tails imply high uncertainty and may comply with the maximum entropy principle applied with certain constraints [*Koutsoyiannis*, 2005a, 2005b].

It should be emphasized from the beginning that this paper is more explanatory than descriptive. In this respect, some general properties of a candidate rainfall modelling approach, rather than the construction of a complete and accurate model, are sought. Besides, as the empirical basis of this study is the Iowa dataset which comprises only seven uninterrupted single storms, it is impossible to study all aspects of the rainfall process and generalize the validity of our findings for other seasons or other locations. For example intermittency, a very important peculiarity of the rainfall process is left out of this study. For the latter, and especially its relationship to the maximum entropy principle, the interested reader is referred to a study by *Koutsoyiannis* [2006a].

## 6.2 General properties of rainfall dataset

### 6.2.1 The data

Seven storm events of high temporal resolution, recorded by the Hydrometeorology Laboratory at the Iowa University [*Georgakakos et al.*, 1994], are the dataset of this study. The original measurements were taken every 5 or 10 seconds; however, for uniformity here the 10-second resolution is used for all events. Figure 6.1 illustrates the patterns of the seven storms.

**Table 6.1.** Summary statistics of the seven storm events.

| Event No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | All |
|---|---|---|---|---|---|---|---|---|
| Sample size | 9 697 | 4 379 | 4 211 | 3 539 | 3 345 | 3 331 | 1 034 | 29 536 |
| Mean (mm/h) | 3.89 | 0.50 | 0.38 | 1.14 | 3.03 | 2.74 | 2.70 | 2.29 |
| Standard deviation (mm/h) | 6.16 | 0.97 | 0.55 | 1.19 | 3.39 | 2.20 | 2.00 | 4.11 |
| Skewness | 4.84 | 9.23 | 5.01 | 2.07 | 3.95 | 1.47 | 0.52 | 6.54 |
| Kurtosis | 47.12 | 110.24 | 37.38 | 5.52 | 27.34 | 2.91 | -0.59 | 91.00 |
| Hurst Exponent | 0.94 | 0.79 | 0.89 | 0.94 | 0.89 | 0.87 | 0.97 | 0.89 |

The events are characterized by a variable duration and also exhibit large statistical differences among them. Specifically, summary statistics like the mean, the standard deviation, the skewness and the kurtosis, shown in Table 6.1, differ notoriously among the

events, up to two orders of magnitude (e.g., the kurtosis coefficient). In the following analyses, the different events are analysed either separately or jointly. For the latter type of analysis, which is consistent with the scope of the paper to seek whether a single model can or cannot generate all different types, a merged sample of all events is used.

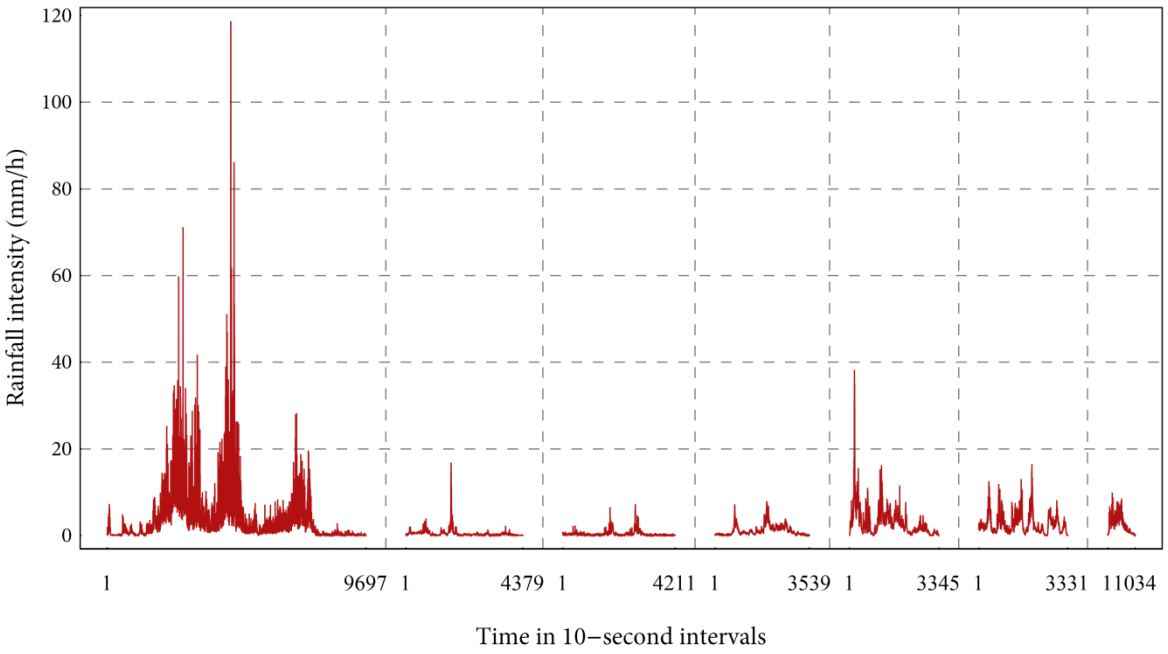



**Figure 6.1.** The seven storm events recorded by the Hydrometeorology Laboratory at the Iowa University.

### 6.2.2 Scaling in state

The term scaling in state [see e.g., *Koutsoyiannis*, 2005a] refers to the power-law behaviour of the probability distribution of a process. Whether or not a natural process is characterized by a power-law distribution is of great importance, as a power-law process implies that extreme events are not only more frequent in comparison to an exponential-law process, but also more severe. Clearly, the frequency and the magnitude of extreme events in natural processes like rainfall, have many practical applications, e.g., in the design of hydraulic works.

In practice, the identification and the characterization of a natural process as a power-law process is a difficult task. Natural processes that are considered to be power-law, do not exhibit a single power law distribution over the entire domain. Thus, the range over which the power-law holds, i.e. the distribution tail, must be identified and this is not trivial. Actually, inferences related to distribution tail that are based on sample data are uncertain. Therefore, in the best case, the validity of a power law might be conjectured, if

the empirical data are consistent with the hypothesized power law and do not falsify the power-law hypothesis.

Generally, there are several methods for identifying power-law behaviour in empirical data, e.g., methods based on least-square fitting or maximum likelihood, but none of them seems to be universally accepted [see e.g., *Clauset et al.*, 2009 and references therein]. Nevertheless, one of the most common procedures used for discerning power-law behaviour in empirical data, which dates back to the end of 19th century in the works of Pareto, is based on least-square fitting.

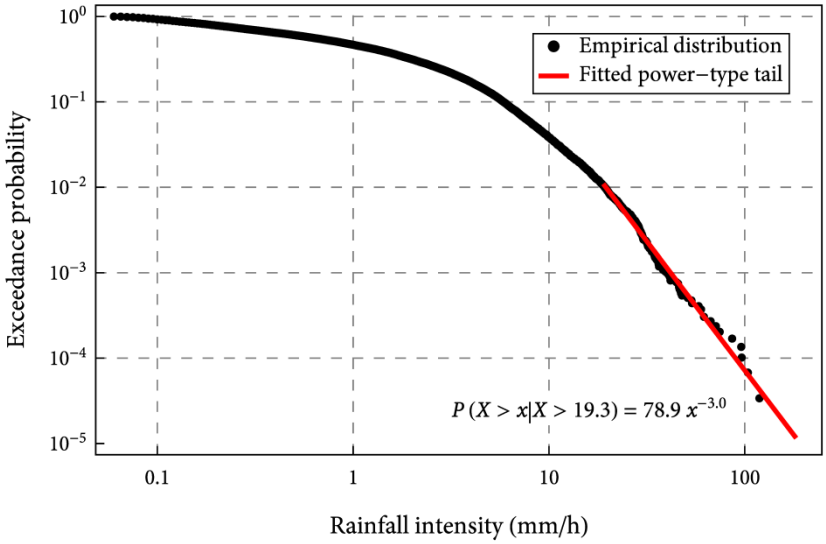



**Figure 6.2.** Empirical probability distribution (Weibull plotting position) of the merged Iowa dataset and the least-square-fitted line to the empirical tail.

Mathematically, a random variable $X$ follows a power-law distribution, if its probability density function is of the form

$$f_X(x) \sim L(x)x^{-\gamma-1} \tag{6.1}$$

where $\gamma > 0$ is a constant known as the scaling exponent or the tail index, and $L(x)$ is a slowly varying function, that is a function satisfying $\lim_{x\to\infty} L(cx)/L(x) = 1$, where $c$ is a constant. The essence of a slowly varying function is that asymptotically it does not affect the power-law behaviour of the distribution, thus controlling the shape of the distribution only over a finite domain of values. Straightforwardly from Eq. (6.1), the $q$th moment of a power-law distribution, defined as $m_q := \int_{-\infty}^{\infty} x^q f_X(x)\mathrm{d}x$, diverges if $q > \gamma$.

It is also apparent from Eq. (6.1) that in a double-logarithmic plot, a power-law distribution (for both the probability density function and the probability distribution function) would be depicted as a straight line—at least in the range of values where the power law holds, i.e. the distribution tail. Thus, the slope of the least-square-fitted line to the tail of the empirical distribution (which, by virtue of Eq. (6.1) is proportional $x^{-\gamma}$) is an estimate of the state-scaling exponent. Using the aforementioned Pareto's method, Figure 6.2 depicts the empirical probability distribution (constructed by using the Weibull plotting position) of the merged Iowa dataset and the least-square-fitted line to the empirical tail. A power law with $\gamma \approx 3$ seems to describe the tail (at probability of exceedance smaller than 1%).

### 6.2.3 Scaling in time

Since *Hurst* [1951] empirically discovered scaling in time, else known as long-term persistence (LTP), this same behaviour has been identified in many other natural processes, as well as time series from many other scientific disciplines, e.g., in economy and in network traffic [e.g., *Baillie*, 1996; *Leland et al.*, 2002]. Ever since, LTP has been an active research field, as its importance necessitated not only theoretical accounts, but also, practical approaches concerning primarily the estimation of its strength and the development of models capable of generating synthetic time series with LTP behaviour.

Basically, scaling in time can be defined in terms of the averaged process on several time scales $k$, i.e.

$$X^{(k)}(\tau) := \frac{1}{k} \sum_{t=(\tau-1)k+1}^{k\tau} X(t) \tag{6.2}$$

In a scaling process the following expression holds, i.e.,

$$\left(X^{(k)}(\tau) - \mu_X\right) \overset{\mathrm{d}}{=} k^{H-1}\left(X(t) - \mu_X\right) \tag{6.3}$$

for any $t$ and $\tau$, where $H$ is the scaling exponent or the so-called Hurst coefficient, and $\overset{\mathrm{d}}{=}$ stands for equality in probability distribution. This process has recently been termed the Hurst-Kolmogorov process (HK; to give credit to Kolmogorov, 1940, who was the first to propose it). If $X$ is Gaussian the process is also called fractional Gaussian noise (fGn), due to *Mandelbrot and Van Ness* [1968]. As can be easily derived by Eq. (6.3), $\sigma_{X^{(k)}} = k^{H-1}\sigma_X$, that is, the aggregated process's standard deviation is proportional to $k^{H-1}$ and not to $k^{-0.5}$

as is in the case of independent processes. In addition, the autocorrelation function $\rho(\tau) \sim \tau^{2H-2}$ as $\tau \to \infty$ and the spectral density $S(\omega) \sim \omega^{1-2H}$. While in the HK process the property in Eq. (6.3) holds for all time scales, in other processes it may hold only asymptotically, as scale tends to infinity. Again the Hurst coefficient $H$ is an important characteristic of the asymptotic behaviour. For example, in a Markovian process, $H = 0.5$ (as in independent processes).

With reference to LTP identification and parameter estimation—a non-trivial issue—many methods have been developed (e.g. based on maximum likelihood, the periodogram, the variance, the rescaled range and others concepts [e.g., *Taqqu and Teverovsky*, 1998; *Taqqu et al.*, 1995; *Tyralis and Koutsoyiannis*, 2010], each having its advantages and drawbacks.

In this study, the Hurst coefficient $H$ is estimated for each of the seven storm events, and additionally for the merged dataset, by using a method that is based on the scaling property of the standard deviation, i.e., $\sigma_{X^{(k)}} = k^{H-1}\sigma_X$. Taking the logarithms, it follows that $\ln\sigma_{X^{(k)}} = (H-1)\ln k + \ln\sigma_X$, and consequently, the aggregated sample standard deviation $\sigma_{X^{(k)}}$ versus the timescale $k$ in a double-logarithmic plot, would be depicted as a straight line (at least in the timescale range where the scaling holds) and the estimated Hurst coefficient is $H = 1 + \eta$, where $\eta$ is the slope of the fitted linear regression line.

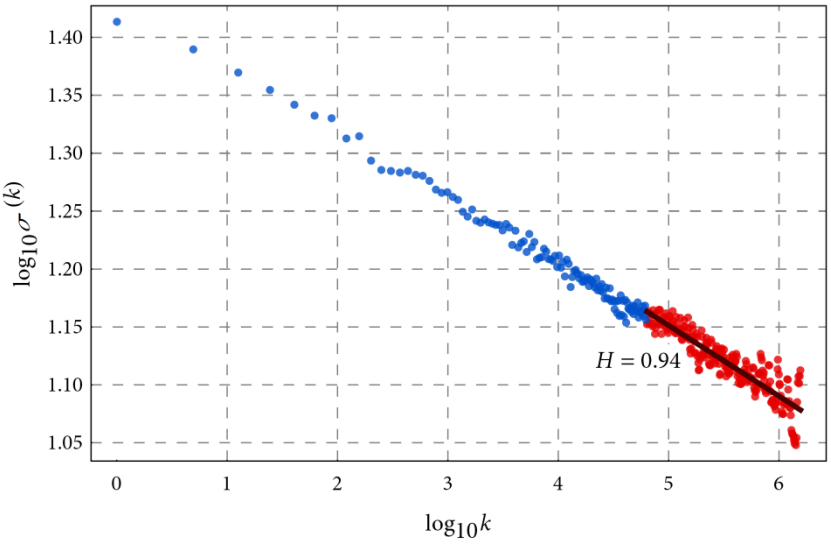



**Figure 6.3.** Double logarithmic plot of sample standard deviation versus scale of averaging for the normalized merged event.

The estimated Hurst coefficients of the seven storm events are presented in Figure 6.3; the variation among the estimated coefficients is high, from 0.77 to 0.97 with a mean

value 0.88. Nevertheless, under the assumption that the seven storm events can be considered as the realizations of a single process, a better estimate of the Hurst coefficient would result if the estimation is carried out on the merged dataset, taking care in the aggregation procedure that individual storm events do not interfere with each other. As Figure 6.3 reveals, the scaling in the merged event seems to hold over the whole range of timescales, while the estimated Hurst coefficient is 0.94.

## 6.3 Stochastic analysis of the rainfall dataset

### 6.3.1 The simulation scheme

As previously mentioned, the major target of this study is first to explore if the seven storm events could be considered as the outcome of a sole and simple stochastic process, and second, to identify the basic characteristics of this process. To this aim the approach followed is heuristic; that is, a stochastic simulation scheme was formed in order to generate synthetic rainfall series whose statistics are subsequently compared with those of the observed records. The aim is to check whether one cannot reject the hypothesis that the statistics themselves are coincident and therefore the observed and synthetic records could be regarded as realizations of the same stochastic process. Different stochastic processes are considered; in principle, candidate processes should include power law as well exponential solutions for both the marginal probability distribution and the autocorrelation. As mentioned above, there are two major questions that need to be answered: the first, concerns the scaling in state, i.e., whether or not, the stochastic process's marginal probability distribution is power type or exponential type. The second, concerns the scaling in time, i.e., whether or not, the autocorrelation structure is power type or exponential.

Regarding the marginal distribution, it is straightforward that realizations from a stochastic process with a power-type marginal distribution would exhibit large differences from an exponential marginal distribution, mainly because a power-type distribution assigns large probabilities to the extreme events, which signifies high variability and uncertainty. Clearly, this behaviour is in agreement with the large variability observed in the seven recorded storm events, and in addition, the whole dataset does not falsify the power-law hypothesis of the marginal distribution (see section 6.2.2). Consequently, the power-law hypothesis of the marginal distribution is accepted as rational and valid and only stochastic processes with power law marginal distribution were considered for the simulation.

In contrast to the choice of the marginal distribution, the a priori decision of a particular autocorrelation structure for the stochastic process is not simple. Short term

persistence (STP) models have been a frequent choice in simulating natural phenomena, but they are often unjustifiably adopted [see e.g., *Koutsoyiannis and Montanari*, 2007]. In fact, an LTP autocorrelation structure, in many cases, may be more appropriate [see, for instance, *Mandelbrot and Van Ness*, 1968, 1968]. Additionally, it is not clear, how intensively the autocorrelation structure of a stochastic process—taking into account that the marginal distribution remains the same—affects the variability of the sample statistics among different realizations, e.g., the statistics among the simulated storm events addressed in this study. Thus, even in the case when the empirical evidence supports the adoption of a certain autocorrelation structure, and in view of the intrinsic uncertainty of this choice, it is valuable to perform a comparison of different scenarios, i.e., a comparison between STP and LTP autocorrelation structures. Therefore, this rationale suggests a side-by-side comparison between an STP model and an LTP model in view of the behaviours of the observed data.

The following sections present the simulation scheme which consists of the following seven steps: (1) application of an appropriate normalizing transformation to the original dataset (section 6.3.2); (2) analysis of the empirical ACF (section 6.3.4); (3) identification and calibration of an STP model and an LTP model (sections 6.3.5 and 6.3.6) to the normalized dataset; (4) correction of the model standard deviation bias (section 6.3.7); (5) simulation of normal synthetic time series (section 6.3.8); (6) generation of the synthetic rainfall time series by applying the inverse transformation (see section 6.3.2) to the normal synthetic time series; and (7) statistical analysis of the synthetic time series (section 6.4).

### 6.3.2 Normalizing the original data

The Gaussian or the Normal distribution is probably the most known and the most widely used distribution in statistics, with applications also in natural sciences. There are two theoretical reasons that justify the ubiquity of the Normal distribution in statistics and its application in other scientific fields. The first relates to the central limit theorem (CLT) that—loosely speaking—states that the sum of independently and identically distributed (i.i.d) random variables tends to the Normal distribution as the number of summands tends to infinity. The second is the principle of maximum entropy [*E. T. Jaynes*, 1957b], which states that, among all possible distributions with known mean and variance, the normal distribution is the one that maximizes the Boltzmann-Gibbs-Shannon information entropy [see also *Shannon Claude and Weaver*, 1948].

Nevertheless, it seems that geophysical data are seldom normal. Empirical data show that many geophysical processes, like rainfall and river discharge, may depart mildly or

severely from normality, especially at small time scales. A relevant example is the dataset addressed in this study. Specifically, departures from normality may be identified in skewness, e.g., positively or negatively skewed empirical data, in the asymptotic behaviour of the distribution tail, e.g., a stretched exponential tail or a power-type tail, and of course, in the variable's domain. Thus, as there exist theoretical reasons that favour normality in many cases, theoretical reasons also exist that do not support it [see e.g., *Koutsoyiannis*, 2005a; *Papalexiou and Koutsoyiannis*, 2012].

For instance, it is well known that a normal variable ranges over the whole real axis, while many natural processes are positively defined, that is, have a lower limit at zero, while a solid reason to fix an upper limit very rarely exists. While the previous reasons explain why departures from normality are so common in nature, a formal and generalized method for simulating non-normal data with a certain autocorrelation structure does not exist, although heuristic solutions were frequently proposed [for a hydrological example, see *Montanari et al.*, 1997]. In contrast, several methods exist addressing the simulation of normal data with STP or LTP autocorrelation structures [e.g., *Box et al.*, 1994; *Brockwell and Davis*, 2009; *Koutsoyiannis*, 2000]. A common technique for simulating non-normal data consists of transforming the non-normal dataset to normal, by applying a normalizing transformation, next, simulating normal data by implementing a standard model, and finally, de-normalizing the normal data by applying the inverse transformation. Basically, this is the methodology followed also in this study, which presents the inconvenience that finding an appropriate normalizing transformation is not always a trivial task, and clearly, a general method for normalizing all types of data does not exist. It is well known that there are some general and commonly used families of transformations, like the Box-Cox family of transformations [*Box and Cox*, 1964], that in many cases give satisfactory results. Unfortunately, such general and simple transformations were not effective for the case of the Iowa dataset. In particular, while the application of the Box-Cox transformation resulted in approximately normal data for the upper empirical tail, it failed to normalize the lower tail, namely the values near zero. A frequently used solution to solve this problem is the normal quantile transform [also called normal quantile score; *Kelly and Krzysztofowicz*, 1997] which, however, is an empirical transformation that is defined over the range of the observed data only and cannot be extrapolated.

Therefore, in order to normalize the Iowa dataset a five-parameter normalizing transformation is introduced here heuristically [by extending a transformation by *Koutsoyiannis et al.*, 2008] given by

$$z(t) = g\left(x(t)\right) = \left(\alpha x(t)^{-\zeta} + \beta\right)\left(\gamma + \left(\left(1+1/\delta\right)\ln\left(1+\delta\left(x(t)-\gamma\right)^2\right)\right)^{1/2}\right) \tag{6.4}$$

where $z(t)$ and $x(t)$ are the transformed and original values of the rainfall intensity, which are realisations of the stochastic processes $Z(t)$ and $X(t)$, respectively, and $\alpha$, $\beta$, $\gamma$, $\delta$, $\zeta$ are the parameters to be estimated. The two factors of the product in the right hand side are introduced to normalize the lower and the larger values, respectively.

While this transformation was identified heuristically, its construction was based on two theoretical aspects. First, Eq. (6.4) ensures that the random variable $Z \sim \mathrm{N}(0,1)$ ranges from $-\infty$ to $\infty$. Obviously, inspection of (6.4) reveals that for $\{\alpha,\beta,\delta,\zeta\} \in (0,1)$ and $\gamma \in (-\infty,0)$, the random variable $Z \in (-\infty,\infty)$, as $\lim_{x\to 0^+} g(x) = -\infty$ and $\lim_{x\to\infty} g(x) = \infty$. Second, the probability density function (pdf) of the random variable $X$ should be long tailed as the empirical evidence supports this assumption (see section 6.2.2). Again, inspection of Eq. (6.4) reveals that for large values of $x$, $g(x) \sim \left(2\beta^2(1+1/\delta)\ln x\right)^{1/2}$, and taking into account that $f_Z(z) \sim \exp(z^{-2}/2)$ and combining the two equations, we get $f_X(x) \sim f_Z\left(g(x)\right) \sim x^{-\beta^2(1+1/\delta)}$ and thus the pdf of the variable $X$ is long tailed.

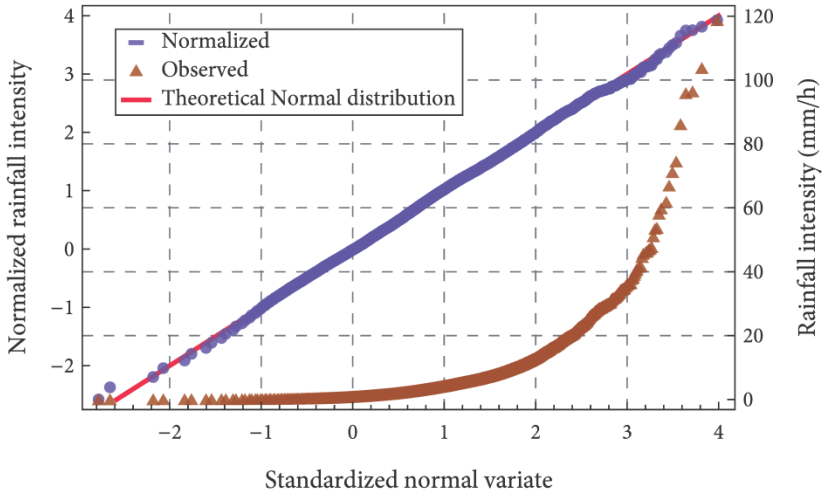



**Figure 6.4.** Probability plot of the natural (recorded) and the normalized rainfall intensity data.

Finally, the parameters of Eq. (6.4) were estimated for the transformed merged Iowa dataset by using the method of least-squares, and particularly, by numerically minimizing the sum of squared errors between values of the standardized normal variate that correspond to the values of the empirical normal distribution (obtained by applying the normal quantile transformation) and the respective values result from the application of

Eq. (6.4) to the original rainfall values. The resulted estimates were $\alpha = 0.41$, $\beta = 2.49$, $\gamma = -2.13$, $\delta = 4.09$ and $\zeta = 1.18$. The transformed data in comparison with the original data are presented in Figure 6.4. Clearly, as the Figure 6.4 demonstrates, the transformed data are satisfactorily normalized.

### 6.3.3 Identification and calibration of the stochastic models

A Gaussian (normal) stochastic process is completely characterized when its second-order distribution, i.e., $F_X(x_i, x_j; t_i, t_j) = P\left(X(t_i) \le x_i, X(t_j) \le x_j\right)$ for any $i \neq j$, is known. Normalizing the marginal distribution of a stochastic process by a transformation, does not necessarily result in jointly normal distribution [*Feller*, 1971, p.70]. However, it is important to check if a particular, marginally normalized, data has also become Gaussian in terms of the multivariate joint distribution or not. A rough indication of joint normality is provided by the linear relation of conditional expectation of a variable $X_i$ given $X_j$ for $i \neq j$. Figure 6.5 depicts the normalized rainfall intensity versus the 1-time-step and 10-time-step shifted normalized rainfall intensity. It can be seen that the empirical points are spread around a straight line, which is an indication of joint normality. This linearity should not be regarded as a surprise, given that it is consistent with the principle of maximum entropy applied on a multivariate setting with constraints of known mean, variance and lag-1 autocorrelation.

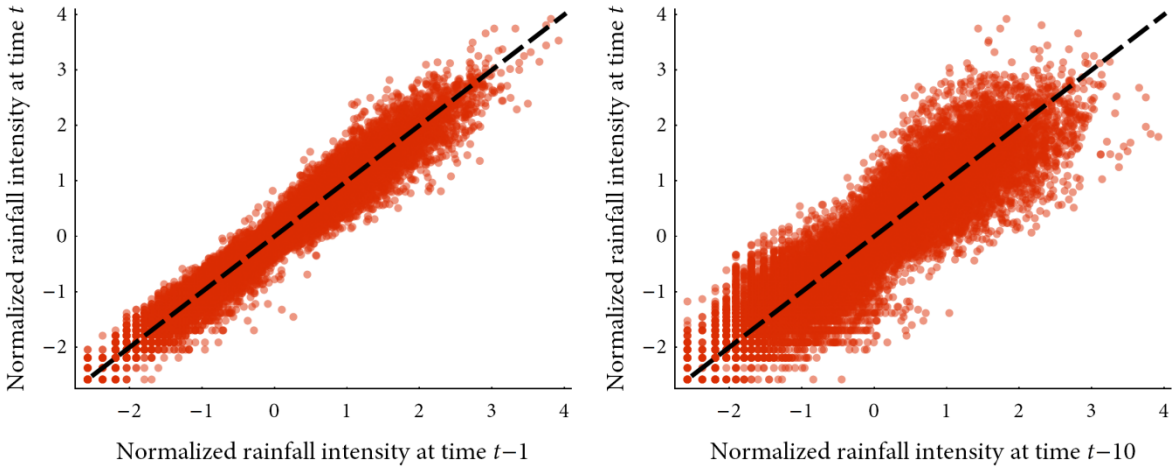



**Figure 6.5.** Scatter plot of normalized rainfall intensity for time lags 1 and 10.

As discussed in section 6.2.3, scaling in time exists and is quantified by an estimated Hurst coefficient $H = 0.94$. Similarly, analysis of the transformed dataset, using the same methods as in section 6.2.3, reveals also a high value of the Hurst coefficient, i.e., $H = 0.92$. Thus, accepting the assumption of scaling in time, a serious issue arises; that is, almost all

classical estimators of statistics (exception is the mean value) are highly biased [e.g., *Koutsoyiannis*, 2003]. So in order to set up an accurate and consistent stochastic model to simulate a normal process—that is, a model that sufficiently reproduces the mean, the standard deviation and the autocorrelation structure of the observed sample—unbiased and accurate estimates of the aforementioned statistics are necessary.

### 6.3.4 Empirical autocorrelation function (ACF)

It is well known, that for finite samples the typical estimate $\hat{\rho}_l$ of the lag-$l$ autocorrelation is a biased estimator of the true autocorrelation $\rho_l$ and the more intense the autocorrelation structure is the more biased the estimator becomes. In particular, in the presence of scaling in time the bias can be corrected by the following formula [see *Koutsoyiannis*, 2003 and the references therein],

$$\tilde{\rho}_l = \hat{\rho}_l\left(1-\frac{1}{n^{2-2H}}\right)+\frac{1}{n^{2-2H}} \tag{6.5}$$

where $\tilde{\rho}_l$ stands for the unbiased estimator and $H$ is the Hurst coefficient.

In this study, the unbiased estimator given in Eq. (6.5) is used to estimate the empirical autocorrelation coefficients. It is clarified, that to estimate $\hat{\rho}_l$ and consequently to estimate the unbiased estimator given in Eq. (6.5), the transformed merged sample was used that comprises the seven transformed storm events. This is a reasonable choice if the seven events are considered as the outcome of a single process; and thus, while the empirical ACF may differ among events, all events share the same theoretical ACF. Furthermore, it is noted that special care was taken in the estimation of the covariance in order to avoid overlapping among the events; specifically, all products of the form $\left(x_t-\hat{\mu}_X\right)\left(x_{t-l}-\hat{\mu}_X\right)$ were eliminated when $x_t$ and $x_{t-l}$ do not belong in the same storm event, and adjusted accordingly the number $n$ of the sample size. The estimated unbiased empirical ACF—given a Hurst coefficient equal to $H = 0.92$, and for lags approximately up to 1000—is depicted in Figure 6.6. Clearly, as Figure 6.6 attests, the empirical autocorrelation structure is very intense, and particularly, the values of the small-lag autocorrelation coefficients are near to 1, while for lags near to 1000 the values are as high as 0.85.

### 6.3.5 The short-term persistence model

Probably, the most common STP stochastic model is the lag-one autoregressive model AR(1). This model belongs to the general family of stochastic models known as

autoregressive moving-average models ARMA($p$,$q$)—comprehensively presented in *Box et al.* [1994]. It is important to note that the ARMA($p$,$q$) family, and especially the AR(1) model are not able to reproduce the scaling behaviour in time or to preserve the Hurst coefficient [e.g., *Box et al.*, 1994]. Consequently, they may be inappropriate for simulating natural phenomena exhibiting LTP.

Nevertheless, while from a theoretical viewpoint ARMA($p$,$q$) models are considered STP models, for increasing values of the autoregressive and moving average order $p$ and $q$ they can provide very good approximations of the LTP structure and thus manage to reproduce, from a practical point of view, the scaling in time or to preserve the Hurst coefficient at least for small sample sizes [*Papalexiou*, 2007]. It is clear, though, that high order ARMA($p$,$q$) models are not parsimonious, i.e., many parameters need to be estimated therefore increasing the estimation variance.

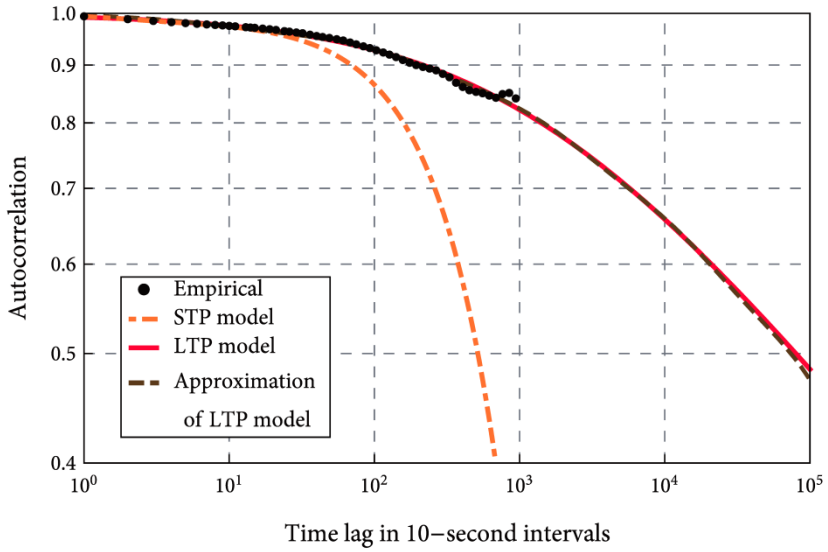



**Figure 6.6.** Empirical ACF of the normalized merged event (corrected for bias), theoretical ACF of the fitted STP model, fitted power-type ACF given in Eq. (6.9), and approximation of the latter by the sum of five AR(1) processes.

Here, the ARMA(2,2) model was chosen for the simulation of the normalized rainfall intensity. It is a model frequently used in hydrology that is able to generate time series that preserve the mean value $\mu_X$, the variance $\sigma_X^2$ and the first four autocorrelation coefficients $\rho_1$, $\rho_2$, $\rho_3$, $\rho_4$. The stochastic process $\{X(t),\ t \in T\}$ that results from an ARMA(2,2) model is defined by

$$X(t) = \alpha_1 X(t-1) + \alpha_2 X(t-2) + \beta_1 \varepsilon(t-1) + \beta_2 \varepsilon(t-2) + \varepsilon(t) \tag{6.6}$$

where $\alpha_1$, $\alpha_2$, $\beta_1$, $\beta_2$ are parameters, and $\varepsilon(t)$ is a normal white-noise process, i.e. consisting of independently, identically and normally-distributed random variables with mean $\mu_\varepsilon = 0$ and variance $\sigma_\varepsilon^2$. Using typical estimation methods [*Box et al.*, 1994], the resulting parameters for the transformed merged Iowa dataset are $a_1 = 1.51$, $\alpha_2 = 0.51$, $\beta_1 = -0.57$, $\beta_2 = -0.19$, $\sigma_\varepsilon = 0.11$.

Once the model parameters are estimated, the theoretical ACF of the ARMA(2,2) for lags $\tau \geq 3$ degenerates to the ACF of an AR(2), i.e., $\rho(\tau) = \alpha_1 \rho(\tau - 1) + \alpha_2 \rho(\tau - 2)$ and thus can be calculated recursively. Figure 6.6 depicts the theoretical ACF of the fitted ARMA(2,2) model in comparison with the empirical ACF. Clearly, it preserves the first four autocorrelation coefficients, as expected, and also performs well for lags up to 50. Nevertheless, for higher lags, it clearly deviates from the empirical ACF as the exponential character of the theoretical ACF unfolds.

### 6.3.6 The long-term persistence model

Since the time when *Hurst* [1951] discovered the LTP behaviour, the necessity to consistently simulate natural phenomena that exhibit LPT has led to the development of several stochastic processes and algorithmic procedures that reproduce the LTP behaviour. Among the most common models are several algorithmic approximations of the HK (or fGn) process by *Mandelbrot and Wallis* [1969], *Mandelbrot* [1971], *O'Connell* [1974], *Koutsoyiannis* [2002], and the FARIMA($p$,$d$,$q$) models introduced by *Granger and Joyeux* [1980] and *Hosking* [1981], that have gained popularity mainly in the last decade [for an application to hydrology see *Montanari et al.*, 1997].

The theoretical ACFs of the HK and FARIMA(0,$d$,0) processes are

$$\rho_{\mathrm{FGN}}(\tau) = \frac{1}{2}\left( |\tau - 1|^{2H} - 2|\tau|^{2H} + |\tau + 1|^{2H} \right) \sim \tau^{2H-2} \tag{6.7}$$

$$\rho_{\mathrm{FARIMA}}(\tau) = \frac{\Gamma(1-d)\Gamma(\tau + d)}{\Gamma(d)\Gamma(\tau + 1 - d)} \sim \tau^{2d-1} \tag{6.8}$$

respectively. Clearly, the ACFs of those two models are asymptotically coincident, with $d = H - 1/2$, as Eq. (6.7) and Eq. (6.8) attest, whereas, time series generated by both of them preserve the scaling exponent $H$. Moreover, while the HK process model is a very simple model—essentially is one-parameter model, the FARIMA($p$,$d$,$q$) models are much more flexible as the orders of $p$ and $q$ controls the STP behaviour of the model.

Here a simple yet general approach was used to simulate LTP, obtained by approximating the real process with the sum of five independent AR(1) processes (note that *Koutsoyiannis* [2002] has shown that good approximations can be obtained even with summing three independent AR(1) processes). The implementation comprises two steps: first, fitting a generalized power-type (GP) ACF to the empirical ACF (see 6.3.4) and second approximating the fitted ACF by the ACF obtained as the sum of five independent AR(1).

Regarding the first step of this approach, a theoretical ACF was fitted (consistent with the empirical evidence) to the empirical ACF in order to be able to extrapolate the correlation coefficients for lags as high as desired, instead of being confined in the lag-range provided by the estimated empirical ACF.

Here, a theoretical three-parameter power-type ACF was used that has the form

$$\rho_{\mathrm{GP}}(\tau) := \left(1 + c\left(\frac{\tau}{a}\right)^{b}\right)^{-1/c} \tag{6.9}$$

where $a > 0$, $b > 0$ and $c > 0$ are parameters. The form of (6.9) can be considered as a natural generalization of an exponential ACF as the $\lim_{c\to 0} \rho_{\mathrm{GP}}(\tau) = \exp(-\tau / a)^{b}$. Asymptotically Eq. (6.9) behaves as $\rho_{\mathrm{GP}}(\tau) \sim \tau^{-b/c}$ and therefore, Eq. (6.9) and Eq. (6.7) possesses the same asymptotic behaviour if $b/c = 2(1-H)$. As a result, the fitted $\rho_{\mathrm{GP}}(\tau)$ would be consistent with the estimated $H = 0.92$ if $b/c = 0.16$. Thus, the $\rho_{\mathrm{GP}}(\tau)$ is fitted by minimizing the square error between the $\rho_{\mathrm{GP}}(\tau)$ and the empirical ACF and by setting as a constraint $b/c = 0.16$. The estimated parameters are $a = 12\,881$, $b = 0.51$, $c = 3.18$. The fitted $\rho_{\mathrm{GP}}(\tau)$ is depicted in Figure 6.6, which shows that the fit is satisfactory.

Turning to the second step of the LTP simulation procedure mentioned above, an LTP model was used made up by the sum of five independent AR(1) process by following the idea that was first introduced by *Mandelbrot* [1971], to approximate the HK process. The same method was used by *Koutsoyiannis* [1994], for the same purposes, while *Mudelsee* [2007] proved empirically that the sum of $n$ inflows generated by an AR(1) model in a river network, with $n$ sufficiently large, ends up with a collective river discharge that exhibits LTP behaviours.

Therefore the LTP model that was used herein to simulate the normalized rainfall intensity is given by

$$Y(t) = \sum_{i=1}^{5} Y_i(t) \tag{6.10}$$

where $Y_i(t) = a_i Y_i(t-1) + \varepsilon_i(t)$ is the *i*-th AR(1) process with mean $\mu_{Y_i} = 0$, variance $\sigma_{Y_i}$, lag-*one* autocorrelation coefficient $a_i$ and and $\varepsilon_i(t)$ is a normal white-noise process, with mean $\mu_{\varepsilon_i} = (1-a_i)\mu_{Y_i} = 0$ and variance $\sigma^2_{\varepsilon_i} = (1-a_i^2)\sigma^2_{Y_i}$. Under the assumption of independence of the five AR(1) processes it can be easily proven that the theoretical ACF of Eq. (6.10) is given by

$$\rho_{\mathrm{LTP}}(\tau) = \sum_{i=1}^{5} a_i^{\tau} \sigma^2_{Y_i}, \quad \text{with} \sum_{i=1}^{5} \sigma^2_{Y_i} = 1 \tag{6.11}$$

The parameters of the five independent AR(1) processes were estimated by minimizing the square error between the Eq. (6.9) and Eq. (6.11) for lags as high as $10^4$. The resulting estimates are $a_1 = 0.9943$, $\sigma^2_{Y_1} = 0.075$, $a_2 = 0.8719$, $\sigma^2_{Y_2} = 0.029$, $a_3 = 0.9999$, $\sigma^2_{Y_3} = 0.179$, $a_4 = 0.9994$, $\sigma^2_{Y_4} = 0.138$ and $a_5 = 0.9999$, $\sigma^2_{Y_5} = 0.578$. As the Figure 6.6 reveals, the fitted $\rho_{\mathrm{LTP}}(\tau)$, up to the lag-$10^4$, is satisfactory.

### 6.3.7 The standard deviation bias

One issue in stochastic modelling that may have serious consequences on the validity and accuracy of the simulation, and is often neglected, concerns the differences in statistics that may occur between the theoretical process and its realizations. While the estimate of the mean is unbiased regardless of the dependence structure, this does not hold for the standard deviation. In fact, it is well known that the standard estimator *S* of the standard deviation is slightly biased even in the case of normally distributed and independent data [e.g., *Bolch*, 1968]. However, the bias may become very large in a time dependent process as it increases monotonically with the increase of the autocorrelation. For certain known ACFs, like the one of the HK process, unbiased estimators have been developed [see *Koutsoyiannis*, 2003 and references therein].

In order to assess the standard deviation bias in random samples generated by the STP and the LTP models described in section 6.3.5 and 6.3.6, and for several different sample sizes, a Monte Carlo simulation was performed. Specifically, at first, 5000 independent samples were generated by each model and for several sample sizes, and in turn, a standard deviation correction factor was calculated, defined by $c_{\mathrm{SD}} := \sigma / E(S)$ where

$\sigma$ is the true standard deviation of the STP and the LTP models, chosen as 1 in our simulations and $E(S)$ is calculated by Monte Carlo simulation.

The results that are depicted in Figure 6.7, are remarkable, especially in the case of the LTP model. In fact, the bias correction factors, for a sample size of 1 000, are as high as 1.9 and 3.1 for the STP and the LTP models, respectively, while even for a very large sample size equal to 50 000, in the LTP case, the correction factor sustains a value of 1.7. Given that the correction factor depends of the sample size, the choice of the appropriate correction factor should be carried out by considering the number of the data generated with the simulation. Given that the normalizing transformation was applied to a sample of 29 536 values that comprised the seven storm events, it follows that a unit standard deviation was imposed to that complete sample. Consequently, all samples generated in this study, irrespective of their size, were multiplied by the correction factor that corresponds to a size of 29 536 size, that is, $c_{SD} = 1.04$ for the STP model and $c_{SD} = 1.81$ for the LTP model. In this way the correction to the standard deviation was imposed depending on the sample size that was used to constrain the standard deviation itself during the normalizing transformation.

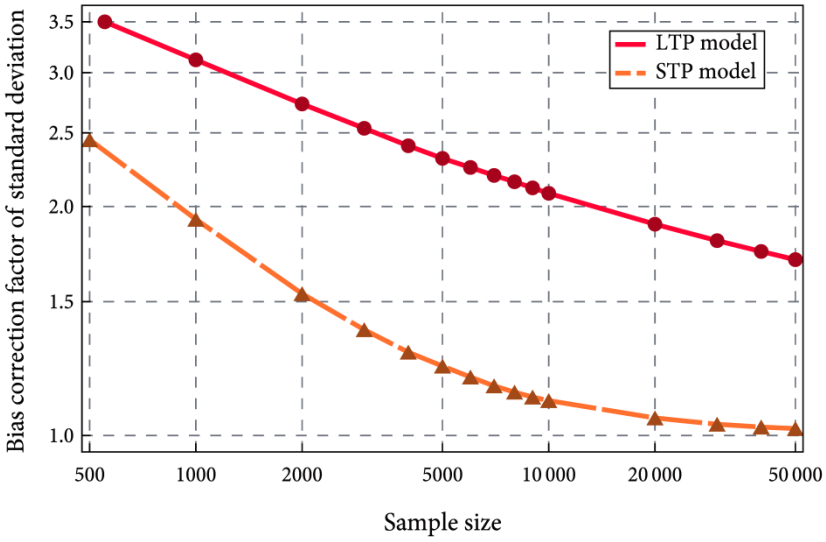



**Figure 6.7.** Standard deviation bias correction factors for the STP and the LTP models for various sample sizes (dots and triangles) calculated by Monte Carlo simulation.

### 6.3.8 Sample size and number of samples

As shown in Figure 6.1, the seven recorded storm events have all different sample lengths varying from 1 034 to 9 697 values. In order to compare the observed statistics with those as the synthetic series, it would be appropriate that the simulations have the same length of the observed records. For practicality only 3 sample sizes were used namely: 1 000 (L1),

which is very close to the size of event 7; 4 000 (L2), close to the size of events from 2 to 6; and 10 000 (L3) representing event 1.

Finally, 10 000 synthetic series were generated for each sample length and for each model. In sequel, the mean, the standard deviation, the skewness, the kurtosis and the autocorrelations, were calculated for every synthetic series and were compared with the respective statistics of the observed records.

## 6.4 Results of the stochastic simulation

Figure 6.8 reports an example of visualized simulated events for the three different sample sizes (L1, L2 and L3) considered here and the two different models (three events generated by the LTP model on the left and three by the STP model on the right). Some differences in the patterns generated by the models are visible. For example, the pattern of the LTP model is characterized by a higher variability (although the marginal distributions are the same). By comparing the patterns with those of the observed records, which are shown in Figure 6.1, one may notice that the variability of the observed record looks better reproduced by the LTP model.

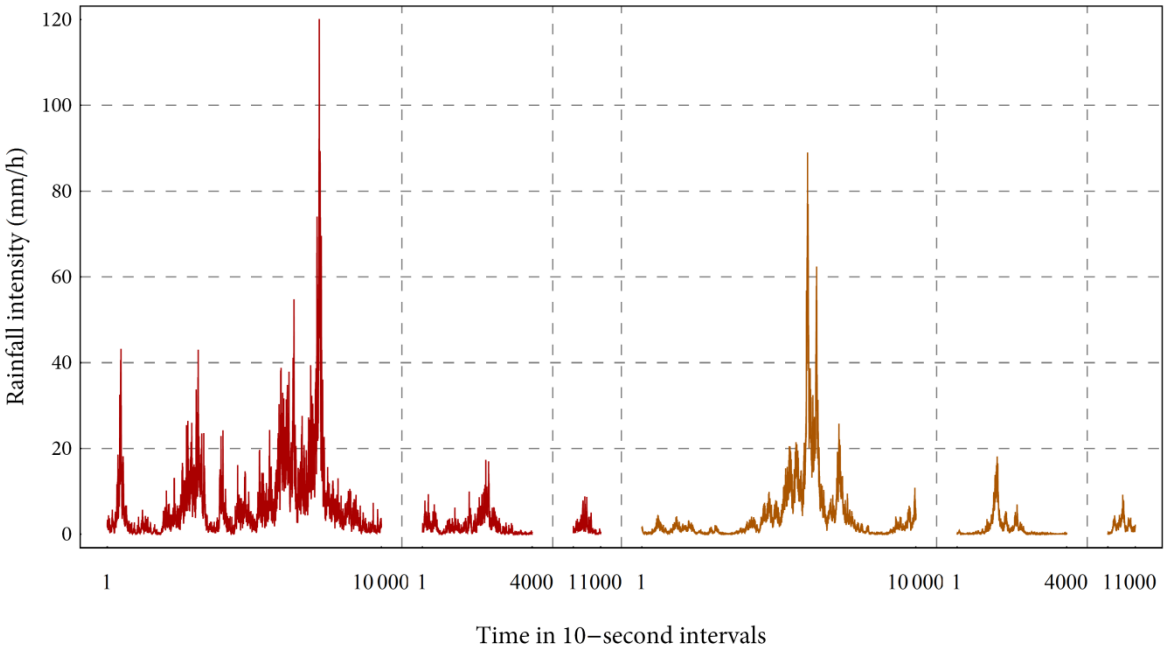



**Figure 6.8.** Synthetic rainfall events generated by the LTP (the first three) and the STP (the last three) models for three characteristic samples sizes.

Figure 6.9 shows box plots of selected statistics computed on the simulated data (in this case also by referring to the original probability distribution), namely, mean, standard deviation, skewness and kurtosis. The observed statistics are also shown with dots. The box

and the whiskers encompass 50% and 99%, respectively, of the computed statistics, while the median is indicated by a horizontal straight line. The box plots clearly show the different behaviours of the two models. Looking at the mean value, one should note that, not surprisingly, the two models are characterized by nearly the same median of the mean, but the variability in the LTP model is higher. Also expected is the higher variability of the standard deviation, skewness and kurtosis that is depicted in the other box plots. One may note that the LTP model is more skewed than the STP one. This result is explained by the higher variability of a process (rainfall) that is bounded at zero. In general one may note that the higher uncertainty of the LTP model makes the fit more satisfactory, even though the observed points are very few and therefore do not allow more than a qualitative assessment.

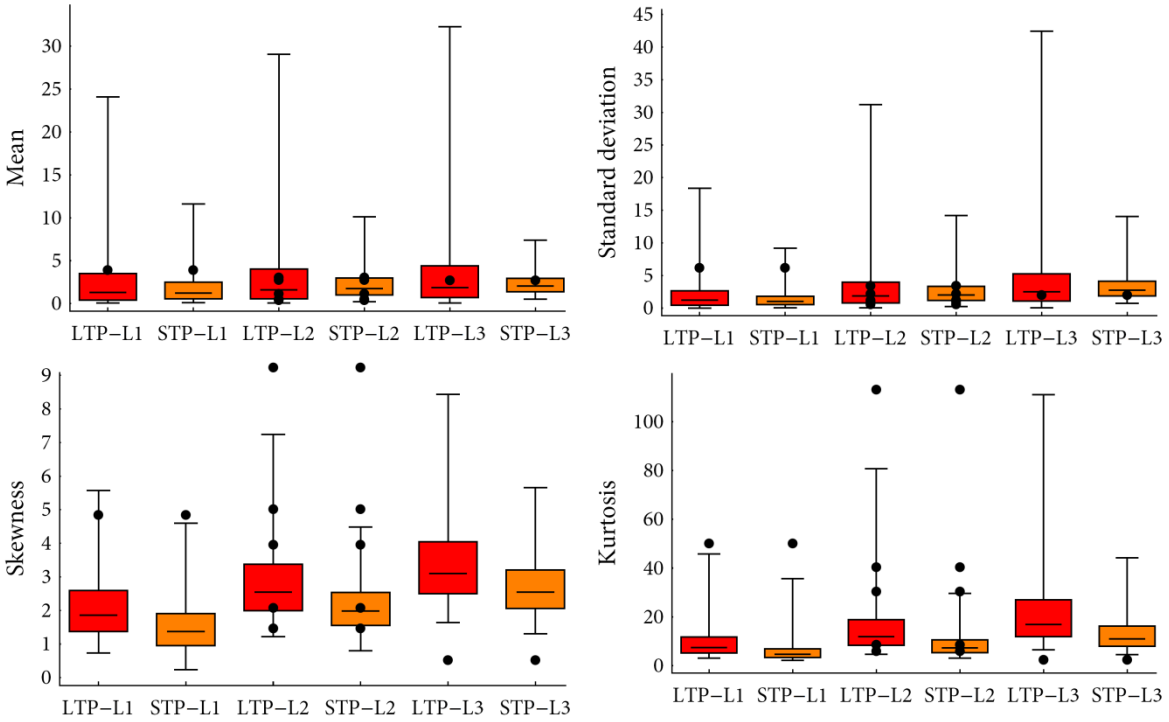



**Figure 6.9.** Box plots of sample statistics estimated from the synthetic rainfall events generated by the LTP and STP models for the three characteristic sample sizes L1, L2 and L3. The dots represent the empirical points of the seven rainfall events.

Figure 6.10 shows a comparison between the observed autocorrelation functions with those simulated by the models. First of all, one notes that the autocorrelation coefficients of the LTP model are higher in the tail of the autocorrelation function. This result is expected. Another relevant feature is the higher correlations shown by the STP model for low lags. This result, which is not intuitive, is due to the fact that the STP model, in order to reach a

better fit of the tail of the ACF, reacts by increasing also the autocorrelation coefficients for low lags. Conversely, the power law behaviour of the LTP model allows one to reach a better fit of the tail of the ACF without increasing much the correlation for low lags. Even in this case, the autocorrelation function of the LTP appears to be more convincing in view of the observed pattern. One should note that this assessment is again qualitative in view of the small number of observed events.

However, apart from the comparison between the two models, one relevant conclusion is that both models look able to provide, within a relatively simple framework, a satisfactory fit of the observed behaviours.

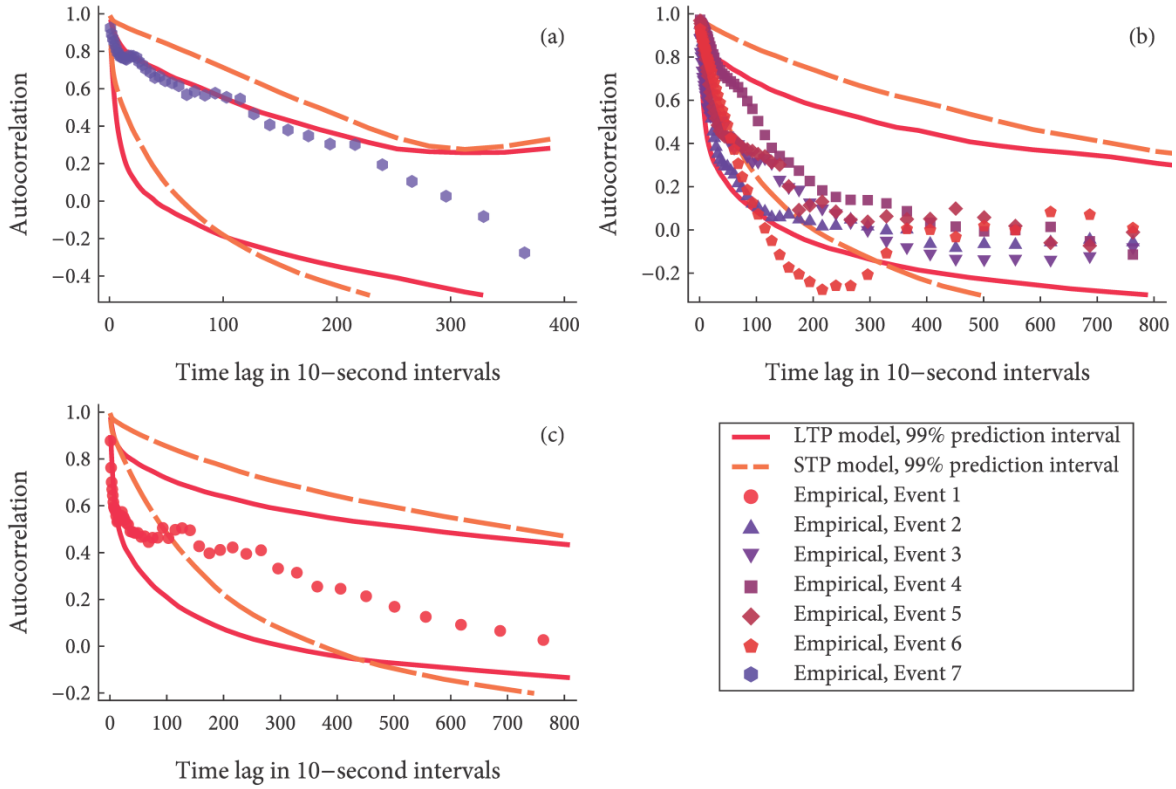



**Figure 6.10.** Empirical autocorrelation functions of the seven rainfall events and 99% prediction intervals of ACF for the LTP and STP models.

## 6.5 Conclusions and discussion

Summarizing the above investigations, it can be said that a single and rather simple stochastic model can represent all rainfall events and all rich patterns appearing in each of the separate events making them look very different from one another. From a practical view point, such a model is characterized by high autocorrelation at fine scales, slowly decreasing with lag, as well as by distribution tails slowly decreasing with rainfall intensity. Such an autocorrelation form can indeed produce huge differences among different events

and such a distributional form can produce enormously high rainfall intensities at times. Both these behaviours are just opposite to the more familiar processes resembling Gaussian white noise, which would produce very "stable" events with infrequent high intensities. In this respect, both high autocorrelations and distribution tails can be viewed as properties enhancing randomness and uncertainty (or entropy).

Whether the tails of both the marginal distribution and autocorrelation functions are long (meaning that are described by power-law functions) is difficult to conclude based merely on the dataset of this study. Both these power-law functions are by definition asymptotic properties, and the exponents of power laws are theoretically defined for state or lag tending to infinity. In this respect, it seems impossible to verify such asymptotic laws by empirical studies, which necessarily imply finite sample sizes. But it is important that the empirical evidence presented in the current study does not falsify the hypothesis that both tails are long. Other empirical studies published recently [*Papalexiou et al.*, 2013] do not falsify this hypothesis as well.

If the hypothesis of a long tail of the distribution function is accepted, it seems that this can be quantified by an exponent $\alpha$ of about 3, which implies that only the first three moments of the distribution exist whereas all others are infinite. If the hypothesis of a long tail of the autocorrelation function is accepted, it seems that this can be quantified by a Hurst coefficient $H$ as high as 0.94. Based on these findings, the construction of a stochastic model admitting asymptotically long tails from the outset seems a reasonable choice. After all, in a dynamical systems context, even the randomness is an asymptotic property per se, in the sense that it implies an infinite number of degrees of freedom. The fact that an infinite number of degrees of freedom cannot be verified (and perhaps neither falsified) empirically, does not preclude us from successfully using probabilistic descriptions and stochastic models of several processes including rainfall.

As mentioned earlier, long tails can be viewed as an enhancement of randomness and uncertainty in these processes. In the framework of this enhanced randomness, it seems to be useless to analyse each rainfall event separately as an attempt to infer dynamics of rainfall. Such an attempt, even using sophisticated methods such as wavelets, can perhaps be paralleled with one's attempt to explain the dynamics of the tossing of a coin by observing a series of "heads" and "tails". In both cases, it may be misleading to seek substantial information in extremely random occurrences. A more useful target for such cases would be to elevate from the obscurity the underlying randomness and seek its own laws.

# CHAPTER 7

*"Beauty will save the world."*

*FYODOR DOSTOEVSKY*

# CONCLUSIONS

## 7.1 Summary

This research has been focused on three main topics: (a) the use of a theoretical principle, i.e., the Principle of Maximum Entropy as a theoretical basis to derive probability distributions suitable for geophysical processes, (b) the statistical analysis, at a global scale, of daily rainfall and of daily rainfall extremes, and (c) the stochastic analysis of rainfall at fine temporal scales. The major objectives of this research were to formulate some simple yet fundamental and of wide interest questions and try to give answers not only of theoretical value but mainly of practical one.

With respect to the Principle of Maximum Entropy, the study focused on the possibility to use the classical definition of entropy, i.e., the Boltzmann-Gibbs-Shannon entropy, avoiding thus the use of generalized entropy measures, to derive suitable probability distributions for rainfall, or more generally, for positively defined geophysical random variables. The emphasis was on formulating and theoretically or logically justifying specific constraints, with the premise to be as simple and general as possible that would lead into flexible and simple distributions.

Regarding the statistical analysis of daily rainfall, which constitutes the largest part of this research, three different aspects of daily rainfall were examined. First, the seasonal variation of daily rainfall was investigated focusing on the properties of its marginal distribution. A massive empirical analysis of more than 170 000 monthly daily rainfall records was performed from more than 14 000 stations from all over the globe aiming to answer two major questions: (a) which statistical characteristics of daily rainfall vary the most over the months and how much, and (b) whether or not there is a relatively simple

probability model that can describe the nonzero daily rainfall at every month and every area of the world. Second, the focus was on the distribution tail of daily rainfall, i.e., the distribution's part that describes the extremes events. Data from more than 15 000 stations were used to test the performance of four common tails that correspond to the Pareto, the Weibull, the Lognormal and the Gamma distributions aiming to find out which type of tail better describes the behaviour of extreme events. Third, annual maxima of daily rainfall from thousands of stations from all over the world were extracted and analysed trying to answer one of the most basic questions in statistical hydrology, i.e., which one of the three Extreme Value distributions better describes the annual maximum daily rainfall.

Finally, rainfall was examined at fine temporal scales by studying a dataset comprising measurements of seven storm events at a temporal resolution of 5-10 seconds and tried to answer the question if a single and simple stochastic model can generate a plethora of temporal rainfall patterns, as well as to detect the major characteristics of such a model.

## 7.2 Conclusions

### 7.2.1 On the Principle of Maximum Entropy

- ***Why and how could the Principle of Maximum Entropy help derive or choose suitable probability distributions for a random variable?***

  The number of well-known distributions may be less than a hundred while from a mathematical point of view this number is literally infinite as an infinite number of functions can be formed with the properties of a probability distribution. The common technique to choose a distribution is usually based on trial-and-error methods, i.e., fitting the commonly used distributions to the data and selecting the best fitted according to a fitting measure. Moreover, this procedure, at least theoretically, could be endless if one decides to form new distributions to test. On the contrary entropy maximization offers a solid theoretical basis for identifying a probabilistic law based on the available information. Yet the key issue in using successfully this principle is to incorporate all available information in the form of constraints.

- ***What form should these constraints have for geophysical variables, e.g., like rainfall?***

  The rationale formed here is based on the premises that the constraints should be as few and simple as possible and incorporate prior information on the process of interest. This prior information for example may concern the general shape properties of the density function of the variable under study and could be obtained by an intensive empirical

analysis. Three particular constraints were studied and conceptually justified that are related to the logarithmic and power functions, which can be suitable for positive, highly varying and asymmetric RVs, characteristics that are usually in geophysical processes, e.g., like rainfall. Namely, the constraints are the expected values of (a) the $\ln x$; (b) the $x^q$; and (c) the $\ln(1+px^q)/p$, with the last constraint, named $p$-moments, offering a generalization of the classical moments.

- ***What types of distributions are derived using these constraints?***
  The BGS entropy maximization under two simple combinations of these constraints leads into two flexible distributions, i.e., a three-parameter exponential type, known as the Generalized Gamma (GG), and, a four-parameter power type, known as the Generalized Beta of the second kind (GB2) with the former being a particular limiting case of the latter. For practical purposes the use of a three-parameter power type distribution is proposed, known as the Burr type XII, which is easily derived as simplification of the GB2 distribution. Both the GG and Burr type XII distributions are very flexible as, apart from a scale parameter, comprise two shape parameters giving control over both tails (left and right).

- ***Are generalized entropy measures necessary to obtain heavy-tailed distributions?***
  Maximization of the BGS entropy has been "traditionally" used by imposing constraints that led to exponential type distributions having light right tails, e.g., like the Exponential or the Normal distributions. The empirical analysis of various phenomena, however, indicated that these distributions in many cases are inadequate to describe reality since heavy-tailed distributions are also common. This led in the introduction of generalized entropy measures which however raised doubts regarding their validity compared to the classical and well justified BGS entropy. Instead of using these kinds of generalized measures, the constraints formed here and used with the BGS entropy, especially $p$-moments, naturally lead to power-type distributions adhering to the classical entropy definition.

### 7.2.2 On the seasonal variation of rainfall

- ***Which characteristics of the marginal distribution of daily rainfall exhibit seasonal variation?***
  The empirical analysis of the monthly variation of probability dry, of the mean value, and of two measures of shape of nonzero daily rainfall, i.e., the L-variation and the L-skewness, revealed, in general, sinusoidal-like patterns for all statistics indicating thus

seasonal variation. According to the seasonal variation test that was formed and applied, it was observed a clear monthly variation in probability dry and in the mean value of nonzero daily rainfall in 95.1% and in 91.7%, respectively, of the stations studied while the corresponding percentages of the shape characteristics, i.e., of L-variation and L-skewness, were 66.1% and 54.2%, respectively. These results if combined with the general picture obtained by the analysis in the hemispheres indicate that the shape of the marginal distribution varies too, in addition to the probability dry and the mean value.

- ***Which statistics have higher seasonal variation?***
  The monthly variation of those statistics at each station was quantified by various deviation measures with respect to the average of all months. The analysis showed that the highest monthly variation is observed in the mean value of nonzero rainfall followed by probability dry, L-skewness and finally by L-variation, implying that, although the shape characteristics vary, their variability is not very high.

- ***What is the general shape of the nonzero daily rainfall distribution?***
  The variations of statistical measures studied, as well as the fitted distributions, indicate that the density function of nonzero rainfall may significantly differ from station to station not only in its general shape, i.e., J-shaped or Bell-shaped, but also in its tail behaviour implying different behaviour of the extremes.

- ***Are the commonly used two-parameter models adequate models for daily rainfall?***
  The seasonal and the spatial variability observed in the shape characteristics point out that the commonly used two-parameter models, e.g., the Gamma, the Weibull, the Lognormal, the Pareto, etc. cannot serve as adequate or "universal" models for the daily rainfall as their flexibility is limited and thus they cannot describe sufficiently both the main body and left and the right tails of the distribution.

- ***Is there a "universal" model capable of describing daily rainfall at all seasons and at every area of the world?***
  This analysis suggests that a "universal" probability model for daily rainfall must have at least two shape parameters, one to control the left tail and one to control the right tail. Two distributions with the above characteristics which were derived using the Principle of Maximum Entropy are the Burr type XII distribution and the Generalized Gamma distribution. Both distributions performed very well with the latter performing even better than the former providing thus an excellent model choice. These two distributions have some of their characteristics complementary to each other, thus the

the GB2 distribution, which includes both of them as special cases, can be used to model the entire dataset for all months and all stations.

- ***What do the parameter values of the best fitted distribution reveal?***
  The shape parameter $\gamma_2$ of the Generalized Gamma distribution, which controls the right tail and thus the extreme values, for the vast majority of records analysed is $\gamma_2 < 1$, with $\gamma_2 = 1$ corresponding to the Gamma distribution; this implies that some of the most commonly used exponential-tail distributions like the Exponential, the Gamma or mixed Exponentials may constitute a dangerous choice and should not be used unjustifiably in practice as they can severely underestimate the magnitude and the frequency of the extreme daily rainfall.

### 7.2.3 On the rainfall extremes

- ***What type of distribution tail better describes daily rainfall extremes above threshold?***
  The analysis suggests that heavier-tailed, or else, subexponential distributions in general performed better than their lighter-tailed counterparts. Particularly, in 72.6% of the records studied subexponential tails were better fitted while the exponential-hyperexponential tails were better fitted is only 27.4%. The ranking from best to worst in terms of their performance is: (a) the Pareto, (b) the Lognormal, (c) the Weibull, and (d) the Gamma distributions.

- ***Are the most commonly used models for rainfall adequate and reliable to model the extreme events above threshold?***
  The analysis revealed that the most popular model used in practice, the Gamma distribution, performed the worst, implying that the use of this distribution underestimates in general the frequency and the magnitude of extreme events. This leads to the recommendation that subexponential distributions are preferable to model extreme rainfall events worldwide.

- ***What are the implications of subexponential distribution tails in practice?***
  The key implication of this analysis is that the frequency and the magnitude of extreme events have generally been underestimated in the past given that the most commonly used distributions for daily rainfall are light-tailed. This implies that the hydrological design based on these distributions might be a dangerous choice and thus, engineering practice needs to recognize that extreme events are not as rare as it is believed and to shift toward the heavy-tailed probability distributions.

- ***Which one of the three extreme value distributions can better describe annual maxima?***

  Starting with some theoretically based arguments it is noted that the reversed Weibull distribution implies a parent distribution for daily rainfall with an upper bound which appears physically inconsistent, while distributions bounded from above have not been used for daily rainfall in competent studies. With reference to the Fréchet *vs*. Gumbel "battle", it was shown that, as strange it may seem, annual maxima extracted from a parent distribution that belongs to the domain of attraction of the Gumbel law, are better described by the Fréchet law. This occurs for two reasons: first, the convergence rate of subexponential parent distributions to the Gumbel law is extremely slow, and second, the shape parameter of the Fréchet law enables the distribution to approximate quite well not only distributions with power-type tails but also other heavy-tailed distributions. In terms of empirical evidence the investigation of more than 15 000 records provided a clear "verdict", i.e., the Fréchet law prevails.

- ***Is there any relationship between the estimated value of the GEV shape parameter and the record length?***

  The analysis unveils a clear relationship between the shape parameter value over the record length, implying that only very large samples can reveal its true distribution or the true behaviour of the extreme rainfall.

- ***What is the true distribution of the GEV shape parameter?***

  The "asymptotic" analysis performed, based on the fitted functions to the mean and standard deviation of the GEV shape parameter over record length, suggests that the distribution of the GEV shape parameter that would emerge if extremely large samples were available is approximately normal with mean value 0.114 and standard deviation 0.045.

- ***In which interval the GEV shape parameter is expected to vary and can we trust the usual estimators?***

  According to the analysis the GEV shape parameter is expected to belong in a narrow range, approximately from 0 to 0.23 with confidence 99%. Essentially, the analysis shows that data cannot be trusted blindly, as small samples may distort the true picture. In this direction, an equation (Eq. (5.8)) was developed that corrects the L-moments estimates of the GEV shape parameter removing the bias due to limited sample size.

- ***Is it valid to use the GEV distribution with negative shape parameter which implies a bounded from above distribution?***

  In small percentage of the records initially studied the estimated GEV shape parameter value was negative (reversed Weibull law), yet the analysis reveals that this percentage rapidly decreases over sample size, while the fitted function expressing the relationship with sample size indicates that for record length greater than 226 years this percentage would be zero. Additionally, none of the 16 records available with length greater than 140 years resulted in negative shape parameter. Moreover, the probability for a negative shape parameter to occur, according to the distribution fitted, is only 0.005, and combined with the previous findings suggests that a GEV distribution with negative shape parameter (bounded from above) is completely inappropriate for rainfall.

- ***Does the GEV shape parameter vary in different areas of the world?***

  The study of the average GEV shape parameter value within regions defined by latitude difference $\Delta\varphi = 2.5°$ and longitude difference $\Delta\lambda = 5°$ and the constructed maps show that large areas of the world share approximately the same GEV shape parameter, yet different areas of the world exhibit different behaviour in extremes.

- ***What is the importance of these findings and what can be suggested as a rule of thumb?***

  The analysis revealed that the Fréchet law, or else the GEV law with positive shape parameter, prevails over the Gumbel law and a fortiori over the reversed Weibull law, with the latter being a dangerous choice in hydrological design. As a rule of thumb it is proposed that even in the case where data suggest a GEV distribution with negative shape parameter, it should not be used. Instead it is more reasonable to use a Gumbel or, for additional safety, a GEV distribution with a shape parameter value equal to 0.114.

### 7.2.4 On the stochastic properties of rainfall at fine temporal scales

- ***Can a simple stochastic model generate rainfall events that differ significantly with each other?***

  The analysis showed that it is feasible for a single and rather simple stochastic model to generate rainfall events at fine temporal scales with sample statistics varying enormously making them "look" very different to each other.

- ***What are the characteristics of such a model and how do they relate to uncertainty?***

  Such a model is characterized by an intense autocorrelation structure, slowly decreasing with lag, as well as by distribution tail slowly decreasing with rainfall intensity. Such an autocorrelation form can produce huge differences among different events and such a

distribution tail can produce enormously high rainfall intensities at times. Both these behaviours are just opposite to the more familiar processes resembling Gaussian white noise, which would produce very "stable" events with infrequent high intensities. In this respect, both high autocorrelations and distribution tails can be viewed as properties enhancing randomness and uncertainty (or entropy).

# BIBLIOGRAPHY


Ahmad, M. I., C. D. Sinclair, and A. Werritty (1988), Log-logistic flood frequency analysis, *Journal of Hydrology*, *98*(3-4), 205–224, doi:10.1016/0022-1694(88)90015-7.

Alsmeyer, G., and M. Sgibnev (1998), On the tail behaviour of the supremum of a random walk defined on a Markov chain, [online] Available from: http://kamome.lib.ynu.ac.jp/dspace/handle/10131/5689 (Accessed 10 November 2012)

Baillie, R. T. (1996), Long memory processes and fractional integration in econometrics, *Journal of Econometrics*, *73*(1), 5–59, doi:10.1016/0304-4076(95)01732-1.

Balkema, A. A., and L. D. Haan (1972), On R. Von Mises' Condition for the Domain of Attraction of exp(-e-x)1, *The Annals of Mathematical Statistics*, *43*(4), 1352–1354.

Balkema, A. A., and L. de Haan (1974), Residual Life Time at Great Age, *Ann. Probab.*, *2*(5), 792–804, doi:10.1214/aop/1176996548.

Barndorff-Nielsen, O. (1963), On the Limit Behaviour of Extreme Order Statistics, *The Annals of Mathematical Statistics*, *34*(3), 992–1002.

Ben-Zvi, A. (2009), Rainfall intensity–duration–frequency relationships derived from large partial duration series, *Journal of Hydrology*, 367(1–2), 104–114, doi:10.1016/j.jhydrol.2009.01.007.

Beguería, S., S. M. Vicente-Serrano, J. I. López-Moreno, and J. M. García-Ruiz (2009), Annual and seasonal mapping of peak intensity, magnitude and duration of extreme precipitation events across a climatic gradient, northeast Spain, *International Journal of Climatology*, *29*(12), 1759–1779.

Berman, S. M. (1964), Limit Theorems for the Maximum Term in Stationary Sequences, *The Annals of Mathematical Statistics*, *35*(2), 502–516.

Bhattarai, K. P. (2004), Partial L-moments for the analysis of censored flood samples, *Hydrological sciences journal*, *49*(5), 855–868.

Biondini, R. (1976), Cloud Motion and Rainfall Statistics, *Journal of Applied Meteorology*, *15*(3), 205–224, doi:10.1175/1520-0450(1976)015<0205:CMARS>2.0.CO;2.

Bolch, B. W. (1968), More on unbiased estimation of the standard deviation, *American Statistician*, *22*(3), 27–27.

Von Bortkiewicz, L. (1922), Variationsbreite und mittlerer Fehler, *Sitzungsber. Berli. Math. Ges.*, (21), 3–11.

Box, G. E. ., and D. R. Cox (1964), An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 211–252.

Box, G. E. ., G. M. Jenkins, and G. C. Reinsel (1994), *Time series analysis: forecasting and control*, 3rd ed., Prentice Hall, New Jersey.

Brockwell, P. J., and R. A. Davis (2009), *Time series: theory and methods*, Springer Verlag.

Bruhn, J. A., W. E. Fry, and G. W. Fick (1980), Simulation of Daily Weather Data Using Theoretical Probability Distributions, *Journal of Applied Meteorology*, *19*(9), 1029–1036, doi:10.1175/1520-0450(1980)019<1029:SODWDU>2.0.CO;2.

Buishand, T. A. (1978a), Some remarks on the use of daily rainfall models, *Journal of Hydrology*, *36*(3–4), 295–308, doi:10.1016/0022-1694(78)90150-6.

Buishand, T. A. (1978b), Some remarks on the use of daily rainfall models, *Journal of Hydrology*, *36*(3-4), 295–308, doi:10.1016/0022-1694(78)90150-6.

Burr, I. W. (1942), Cumulative Frequency Functions, *The Annals of Mathematical Statistics*, *13*(2), 215–232.

Cârsteanu, A., and E. Foufoula-Georgiou (1996), Assessing dependence among weights in a multiplicative cascade model of temporal rainfall, *Journal of Geophysical Research*, *101*(D21), 26363.

Chow, V. T. (1964), *Handbook of applied hydrology: a compendium of water-resources technology*, McGraw-Hill.

Clauset, A., C. R. Shalizi, and M. E. . Newman (2009), Power-law distributions in empirical data, *SIAM review*, *51*(4), 661–703.

Cunnane, C. (1973), A particular comparison of annual maxima and partial duration series methods of flood frequency prediction, *Journal of Hydrology*, *18*(3–4), 257–271, doi:10.1016/0022-1694(73)90051-6.

Eagleson, P. S. (1970), *Dynamic Hydrology*, McGraw-Hill Inc., New York.

El Adlouni, S., B. Bobée, and T. B. M. J. Ouarda (2008), On the tails of extreme event distributions in hydrology, *Journal of Hydrology*, *355*(1-4), 16–33, doi:10.1016/j.jhydrol.2008.02.011.

Embrechts, P., and C. M. Goldie (1982), On convolution tails, *Stochastic Processes and their Applications*, *13*(3), 263–278, doi:10.1016/0304-4149(82)90013-8.

Embrechts, P., C. Klüppelberg, and T. Mikosch (1997), *Modelling extremal events for insurance and finance*, Springer Verlag, Berlin Heidelberg.

Esteban, M. D., and D. Morales (1995), A summary on entropy statistics, *Kybernetika*, *31*(4), 337–346.

European Commission (2007), Directive 2007/60/EC of the European Parliament and of the Council of 23 October 2007 on the assessment and management of flood risks, *Official Journal of the European Communities, L*, *288*(6.11), 27–34.

Feller, W. (1971), *An introduction to probability theory and its applications*, 2nd ed., John Wiley & Sons Inc, New York.

Fisher, R. A., and L. H. C. Tippett (1928), Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample, *Mathematical Proceedings of the Cambridge Philosophical Society*, *24*(02), 180–190, doi:10.1017/S0305004100015681.

Fitzgerald, D. L. (1989), Single station and regional analysis of daily rainfall extremes, *Stochastic Hydrol Hydraul*, *3*(4), 281–292, doi:10.1007/BF01543461.

Foufoula-Georgiou, E., and D. P. Lettenmaier (1987), A Markov Renewal Model for rainfall occurrences, *Water Resources Research*, *23*(5), 875–884, doi:10.1029/WR023i005p00875.

Galambos, J. (1972), On the Distribution of the Maximum of Random Variables, *The Annals of Mathematical Statistics*, *43*(2), 516–521.

Geng, S., F. W. T. Penning de Vries, and I. Supit (1986), A simple method for generating daily rainfall data, *Agricultural and Forest Meteorology*, *36*(4), 363–376, doi:10.1016/0168-1923(86)90014-6.

Georgakakos, K. P., A. A. Carsteanu, P. L. Sturdevant, and J. A. Cramer (1994), Observation and analysis of Midwestern rain rates, *Journal of applied meteorology*, *33*(12), 1433–1444.

Gnedenko, B. (1943), Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire, *The Annals of Mathematics*, *44*(3), 423–453, doi:10.2307/1968974.

Goldie, C. M., and C. Klüppelberg (1998), Subexponential distributions, in *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, edited by R. Adler, R. Feldman, and M. s. Taggu, pp. 435–459, Birkhäuser Boston.

Granger, C. W. J., and R. Joyeux (1980), An introduction to long-memory time series and fractional differencing, *J Time Series Analysis*, *1*(1), 15–29, doi:10.1111/j.1467-9892.1980.tb00297.x.

Gumbel, E. J. (1958), *Statistics of Extremes*, Columbia University Press.

Gupta, S. K. (2011), *Modern Hydrology and Sustainable Water Development*, John Wiley & Sons.

Haan, C. T., D. M. Allen, and J. O. Street (1976), A Markov Chain Model of daily rainfall, *Water Resources Research*, *12*(3), 443–449, doi:10.1029/WR012i003p00443.

De Haan, L. (1971), A form of regular variation and its application to the domain of attraction of the double exponential, *Z. Wahrsch. Geb.*, (17), 241–258.

Havrda, J., and F. Charvát (1967), Concept of structural a-entropy, *Kybernetika*, *3*, 30–35.

Heo, J. H., D. C. Boes, and J. D. Salas (2001a), Regional flood frequency analysis based on a Weibull model: Part 1. Estimation and asymptotic variances, *Journal of Hydrology*, *242*(3-4), 157–170.

Heo, J. H., J. D. Salas, and D. C. Boes (2001b), Regional flood frequency analysis based on a Weibull model: Part 2. Simulations and applications, *Journal of Hydrology*, *242*(3-4), 171–182.

Hershfield, D. M. (1961), Estimating the probable maximum precipitation, in *Proc. ASCE, J. Hydraul. Div*, vol. 87, p. 106.

Hosking, J. R. . (1981), Fractional differencing, *Biometrika*, *68*(1), 165–176.

Hosking, J. R. M. (1984), Testing Whether the Shape Parameter is Zero in the Generalized Extreme- Value Distribution, *Biometrika*, *71*(2), 367–374, doi:10.2307/2336254.

Hosking, J. R. M. (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), 105–124.

Hosking, J. R. M. (1992), Moments or L Moments? An Example Comparing Two Measures of Distributional Shape, *The American Statistician*, *46*(3), 186–189, doi:10.2307/2685210.

Hosking, J. R. M. (1994), The four-parameter kappa distribution, *IBM Journal of Research and Development*, *38*(3), 251–258.

Hosking, J. R. M., and J. R. Wallis (1993), Some statistics useful in regional frequency analysis, *Water Resour. Res.*, *29*(2), PP. 271–281, doi:199310.1029/92WR01980.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood (1985), Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments, *Technometrics*, *27*(3), 251–261, doi:10.2307/1269706.

Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Transactions of the American Society of Civil Engineers*, *116*, 770–808.

Jaynes, E. T. (1957a), Information Theory and Statistical Mechanics, *Physical review*, *106*(4), 620, doi:10.1103/PhysRev.106.620.

Jaynes, E. T. (1957b), Information theory and statistical mechanics. II, *Physical review*, *108*(2), 171–190.

Jaynes, E. T. (1957c), Information Theory and Statistical Mechanics. II, *Physical review*, *108*(2), 171, doi:10.1103/PhysRev.108.171.

Jaynes, E. T. (2003), *Probability: The logic of science*, Cambridge University Press.

Jenkinson, A. F. (1955), The frequency distribution of the annual maximum (or minimum) values of meteorological elements, *Quarterly Journal of the Royal Meteorological Society*, *81*(348), 158–171, doi:10.1002/qj.49708134804.

Juncosa, M. L. (1949), The asymptotic behavior of the minimum in a sequence of random variables, *Duke Mathematical Journal*, *16*(4), 609–618, doi:10.1215/S0012-7094-49-01658-0.

Kapur, J. N. (1989), *Maximum-entropy models in science and engineering*, John Wiley & Sons.

Kelly, K. S., and R. Krzysztofowicz (1997), A bivariate meta-Gaussian density for use in hydrology, *Stochastic Hydrology and Hydraulics*, *11*(1), 17–31.

Kleiber, C., and S. Kotz (2003), *Statistical size distributions in economics and actuarial sciences*, Wiley-Interscience.

Klüppelberg, C. (1988), Subexponential Distributions and Integrated Tails, *Journal of Applied Probability*, *25*(1), 132–141, doi:10.2307/3214240.

Klüppelberg, C. (1989), Subexponential distributions and characterizations of related classes, *Probability Theory and Related Fields*, *82*(2), 259–269, doi:10.1007/BF00354763.

Kottek, M., J. Grieser, C. Beck, B. Rudolf, and F. Rubel (2006), World Map of the Koppen-Geiger climate classification updated, *Meteorologische Zeitschrift*, *15*(3), 259–263, doi:10.1127/0941-2948/2006/0130.

Kotz, S., and S. Nadarajah (2000), *Extreme value distributions: theory and applications*, Imperial College Press.

Koutsoyiannis, D. (1999), A probabilistic view of Hershfield's method for estimating probable maximum precipitation, *Water resources research*, *35*(4), 1313–1322.

Koutsoyiannis, D. (2000), A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series, *Water Resour. Res.*, *36*(6), 1519–1533.

Koutsoyiannis, D. (2002), The Hurst phenomenon and fractional Gaussian noise made easy, *Hydrological Sciences Journal*, *47*(4), 573–596.

Koutsoyiannis, D. (2003), Climate change, the Hurst phenomenon, and hydrological statistics, *Hydrological Sciences Journal*, *48*(1), 3–24.

Koutsoyiannis, D. (2004a), Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation, *Hydrological Sciences Journal*, *49*(4), 575–590.

Koutsoyiannis, D. (2004b), Statistics of extremes and estimation of extreme rainfall, 1, Theoretical investigation, *Hydrological Sciences Journal*, *49*(4), 575–590.

Koutsoyiannis, D. (2004c), Statistics of extremes and estimation of extreme rainfall, 2, Empirical investigation of long rainfall records, *Hydrological Sciences Journal*, *49*(4), 591–610.

Koutsoyiannis, D. (2005a), Uncertainty, entropy, scaling and hydrological stochastics, 1, Marginal distributional properties of hydrological processes and state scaling, *Hydrological Sciences Journal*, *50*(3), 381–404.

Koutsoyiannis, D. (2005b), Uncertainty, entropy, scaling and hydrological stochastics. 2. Time dependence of hydrological processes and time scaling/Incertitude, entropie, effet d'échelle et propriétés stochastiques hydrologiques. 2. Dépendance temporelle des processus hydrologiques et échelle temporelle, *Hydrological Sciences Journal*, *50*(3), 1–426.

Koutsoyiannis, D. (2006a), An entropic-stochastic representation of rainfall intermittency: The origin of clustering and persistence, *Water Resources Research*, *42*(1), doi:10.1029/2005WR004175. [online] Available from: http://www.itia.ntua.gr/en/docinfo/675/ (Accessed 3 November 2010)

Koutsoyiannis, D. (2006b), On the quest for chaotic attractors in hydrological processes/Sur la recherche d'attracteurs chaotiques dans des processus hydrologiques, *Hydrological Sciences Journal*, *51*(6), 1065–1091.

Koutsoyiannis, D., and A. Montanari (2007), Statistical analysis of hydroclimatic time series: Uncertainty and insights, *Water Resources Research*, *43*(5), W05429.

Koutsoyiannis, D., H. Yao, and A. Georgakakos (2008), Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods/Prévision du débit du Nil à moyen terme: une comparaison de méthodes stochastiques et déterministes, *Hydrological Sciences Journal*, *53*(1), 142–164.

Kroll, C. N., and J. R. Stedinger (1996), Estimation of moments and quantiles using censored data, *Water Resources Research*, *32*(4), 1005–1012.

Kumar, P., and E. Foufoula-Georgiou (1997), Wavelet analysis for geophysical applications, *Reviews of Geophysics*, *35*(4), 385–412.

Leadbetter, M. R. (1974), On extreme values in stationary sequences, *Probability theory and related fields*, *28*(4), 289–303.

Leland, W. E., M. S. Taqqu, W. Willinger, and D. V. Wilson (2002), On the self-similar nature of Ethernet traffic (extended version), *Networking, IEEE/ACM Transactions on*, *2*(1), 1–15.

Mandelbrot, B. B. (1971), A Fast Fractional Gaussian Noise Generator, *Water Resour. Res.*, *7*(3), 543–553.

Mandelbrot, B. B., and J. W. Van Ness (1968), Fractional Brownian motions, fractional noises and applications, *SIAM review*, *10*(4), 422–437.

Mandelbrot, B. B., and J. R. Wallis (1969), Computer Experiments With Fractional Gaussian Noises: Part 1, Averages and Variances, *Water Resour. Res.*, *5*(1), 228, doi:10.1029/WR005i001p00228.

McDonald, J. B. (1984), Some generalized functions for the size distribution of income, *Econometrica: Journal of the Econometric Society*, 647–663.

Mielke Jr, P. W. (1973), Another Family of Distributions for Describing and Analyzing Precipitation Data, *Journal of Applied Meteorology*, *12*(2), 275–280.

Mielke Jr, P. W., and E. S. Johnson (1973), Three-Parameter Kappa Distribution Maximum Likelihood Estimates and Likelihood Ratio Tests, *Monthly Weather Review*, *101*(9), 701–707.

Mielke Jr, P. W., and E. S. Johnson (1974), Some generalized beta distributions of the second kind having desirable application features in hydrology and meteorology, *Water Resources Research*, *10*(2), 223–226.

Mimikou, M. (1983), Daily precipitation occurrences modelling with Markov chain of seasonal order, *Hydrological Sciences Journal*, *28*(2), 221–232, doi:10.1080/02626668309491962.

Mimikou, M. (1984), A study for improving precipitation occurrences modelling with a Markov chain, *Journal of Hydrology*, *70*(1–4), 25–33, doi:10.1016/0022-1694(84)90111-2.

Von Mises, R. (1936), La distribution de la plus grande de n valeurs, *Rev. math. Union interbalcanique*, *1*(1).

Mitzenmacher, M. (2004), A brief history of generative models for power law and lognormal distributions, *Internet mathematics*, *1*(2), 226–251.

Moisello, U. (2007), On the use of partial probability weighted moments in the analysis of hydrological extremes, *Hydrological processes*, *21*(10), 1265–1279.

Montanari, A., R. Rosso, and M. Taggu (1997), Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation, *Water Resources Research*, *33*(5), 1035–1044.

Mudelsee, M. (2007), Long memory of rivers from spatial aggregation, *Water Resour. Res.*, *43*(1), W01202.

O'Connell, P. E. (1974), A simple stochastic modelling of Hurst's law, *Mathematical models in hydrology*, *1*. [online] Available from: http://www.csa.com/partners/viewrecord.php?requester=gs&collection=ENV&recid=7707666 (Accessed 20 January 2013)

Papalexiou, S. M. (2007), Stochastic modelling of skewed data exhibiting long-range dependence. [online] Available from: http://www.itia.ntua.gr/en/docinfo/835/ (Accessed 3 November 2010)

Papalexiou, S. M., and D. Koutsoyiannis (2008a), Ombrian curves in a maximum entropy framework, p. 00702. [online] Available from: http://www.itia.ntua.gr/en/docinfo/851/ (Accessed 6 October 2010)

Papalexiou, S. M., and D. Koutsoyiannis (2008b), Probabilistic description of rainfall intensity at multiple time scales. [online] Available from: http://www.itia.ntua.gr/en/docinfo/884/ (Accessed 6 October 2010)

Papalexiou, S. M., and D. Koutsoyiannis (2012), Entropy based derivation of probability distributions: A case study to daily rainfall, *Advances in Water Resources*, *45*, 51–57, doi:10.1016/j.advwatres.2011.11.007.

Papalexiou, S. M., and D. Koutsoyiannis (2013), Battle of extreme value distributions: A global survey on extreme daily rainfall, *Water Resources Research*, *49*(1), 187–201, doi:10.1029/2012WR012557.

Papalexiou, S. M., D. Koutsoyiannis, and C. Makropoulos (2013), How extreme is extreme? An assessment of daily rainfall distribution tails, *Hydrol. Earth Syst. Sci.*, *17*(2), 851–862, doi:10.5194/hess-17-851-2013.

Park, J.-S., S.-C. Seo, and T. Y. Kim (2009), A kappa distribution with a hydrological application, *Stoch Environ Res Risk Assess*, *23*(5), 579–586, doi:10.1007/s00477-008-0243-5.

Peel, M. C., Q. Wang, R. M. Vogel, and T. A. McMahon (2001), The utility of L-moment ratio diagrams for selecting a regional probability distribution, *Hydrological sciences journal*, *46*(1), 147–156.

Peel, M. C., B. L. Finlayson, and T. A. McMahon (2007), Updated world map of the Köppen-Geiger climate classification, *Hydrol. Earth Syst. Sci.*, *11*(5), 1633–1644, doi:10.5194/hess-11-1633-2007.

Pickands III, J. (1975), Statistical Inference Using Extreme Order Statistics, *The Annals of Statistics*, *3*(1), 119–131.

Puente, C. E., and B. Sivakumar (2007), Modeling geophysical complexity: a case for geometric determinism, *Hydrology and Earth System Sciences*, *11*(2), 721–724.

Reiss, R. D., and M. Thomas (2007), *Statistical analysis of extreme values: with applications to insurance, finance, hydrology and other fields*, Birkhauser.

Schertzer, D., I. Tchguirinskaia, S. Lovejoy, P. Hubert, H. Bendjoudi, and M. Larchvêque (2002), Which chaos in the rainfall–runoff process, *Hydrol. Sci. J*, *47*(1), 139–148.

Schoof, J. T., and S. C. Pryor (2008), On the Proper Order of Markov Chain Model for Daily Precipitation Occurrence in the Contiguous United States, *Journal of Applied Meteorology and Climatology*, *47*(9), 2477–2486, doi:10.1175/2008JAMC1840.1.

Shannon, C. E. (1948), The mathematical theory of communication, *Bell System Technical Journal*, *27*, 379–423.

Shannon Claude, E., and W. Weaver (1948), The mathematical theory of communication, *Bell System Technical Journal*, *27*, 379–423.

Shimizu, K. (1993), A bivariate mixed lognormal distribution with an analysis of rainfall data, *Journal of Applied Meteorology;(United States)*, *32*(2). [online] Available from: http://www.osti.gov/energycitations/product.biblio.jsp?osti_id=6883866 (Accessed 27 November 2012)

Sivakumar, B. (2000), Chaos theory in hydrology: important issues and interpretations, *Journal of hydrology*, *227*(1-4), 1–20.

Smirnov, N. V. (1949), Limit distributions for the terms of a variational series, *Trudy Matematicheskogo Instituta im. VA Steklova*, *25*, 3–60.

Smith, R. L., and H. A. Schreiber (1974), Point Processes of Seasonal Thunderstorm Rainfal 2. Rainfall Depth Probabilities, *Water Resources Research*. [online] Available from: http://ddr.nal.usda.gov/bitstream/10113/6178/1/CAIN749068818.pdf (Accessed 27 November 2012)

Stacy, E. W. (1962), A Generalization of the Gamma Distribution, *The Annals of Mathematical Statistics*, *33*(3), 1187–1192.

Swift, L. W., and H. T. Schreuder (1981), Fitting Daily Precipitation Amounts Using the SB Distribution, *Monthly Weather Review*, *109*(12), 2535–2540, doi:10.1175/1520-0493(1981)109<2535:FDPAUT>2.0.CO;2.

Tadikamalla, P. R. (1980), A Look at the Burr and Related Distributions, *International Statistical Review / Revue Internationale de Statistique*, *48*(3), 337–344.

Taqqu, M. S., and V. Teverovsky (1998), On estimating the intensity of long-range dependence in finite and infinite variance time series, *A practical guide to heavy tails: statistical techniques and applications*, 177–218.

Taqqu, M. S., V. Teverovsky, and W. Willinger (1995), Estimators for long range dependence, *Fractals*, *3*(4), 785–798.

Tavares, L. V., and J. E. Da Silva (1983), Partial duration series method revisited, *Journal of Hydrology*, *64*(1–4), 1–14, doi:10.1016/0022-1694(83)90056-2.

Teugels, J. (1975), Class of subexponential distributions, *Ann. Probab.*, *3*(6), 1000–1011, doi:10.1214/aop/1176996225.

Todorovic, P., and D. A. Woolhiser (1975), A stochastic model of n-day precipitation., *Journal of Applied Meteorology*, *14*, 17–24.

Tsallis, C. (1988), Possible generalization of Boltzmann-Gibbs statistics, *Journal of Statistical Physics*, *52*(1), 479–487, doi:10.1007/BF01016429.

Tyralis, H., and D. Koutsoyiannis (2010), Simultaneous estimation of the parameters of the Hurst-Kolmogorov stochastic process, [online] Available from: http://www.scopus.com/inward/record.url?eid=2-s2.0-77954344030&partnerID=40&md5=fcdcb53db3e84a17d2b0567736c8860a (Accessed 3 November 2010)

Vogel, R. M., and N. M. Fennessey (1993), L moment diagrams should replace product moment diagrams, *Water Resources Research*, *29*(6), 1745–1752.

Wang, Q. J. (1996), Using partial probability weighted moments to fit the extreme value distributions to censored samples, *Water resources research*, *32*(6), 1767–1771.

Watson, G. S. (1954), Extreme Values in Samples from m-Dependent Stationary Stochastic Processes, *The Annals of Mathematical Statistics*, *25*(4), 798–800.

Waymire, E., and V. K. Gupta (1981), The mathematical structure of rainfall representations: 1. A review of the stochastic rainfall models, *Water Resources Research*, *17*(5), 1261–1272, doi:10.1029/WR017i005p01261.

Werner, T., and C. Upper (2004), Time variation in the tail behavior of Bund future returns, *Journal of Futures Markets*, *24*(4), 387–398.

Wilks, D. S. (1998), Multisite generalization of a daily stochastic precipitation generation model, *Journal of Hydrology*, *210*(1-4), 178–191, doi:10.1016/S0022-1694(98)00186-3.

Wilks, D. S. (1999), Simultaneous stochastic simulation of daily precipitation, temperature and solar radiation at multiple sites in complex terrain, *Agricultural and Forest Meteorology*, *96*(1–3), 85–101, doi:10.1016/S0168-1923(99)00037-4.

Wilson, P. S., and R. Toumi (2005), A fundamental probability distribution for heavy rainfall, *Geophysical Research Letters*, *32*(14), L14812.

Woolhiser, D., and J. Roldán (1982), Stochastic daily precipitation models: 2. a comparison of distributions of amounts., *Water resources research*, *18*(5), 1461–1468.

# APPENDIX A

*"Science commits suicide when it adopts a creed."*

*THOMAS HENRY HUXLEY*

# DERIVATION OF THE ENTROPIC DISTRIBUTIONS

The maximum entropy distributions given in Chapter 1 and sequentially used in the statistical analysis of daily rainfall in Chapter 2, emerged by maximizing the classical definition of entropy, i.e., the BGS entropy given in Eq. (2.1). These distributions can easily arise by using the general solution of the maximum entropy distributions given in Eq. (2.5) and by replacing the arbitrary constraints with specific ones. Particularly, the Generalized Gamma distribution emerged by using the constraints given in Eq. (2.8) and Eq. (2.10) as follows:

$$\begin{aligned} f_X(x) &= \exp\left(-\lambda_0 - \sum_{j=1}^{n}\lambda_j g_j(x)\right) = \exp\left(-\lambda_0 - \lambda_1 \ln x - \lambda_2 x^q\right) \\ &= \exp\left(-\lambda_0 + \ln x^{-\lambda_1} - \lambda_2 x^q\right) = \exp\left(-\lambda_0\right)\exp\left(\ln x^{-\lambda_1}\right)\exp\left(-\lambda_2 x^q\right) \\ &= \exp\left(-\lambda_0\right) x^{-\lambda_1} \exp\left(-\lambda_2 x^q\right) \end{aligned} \tag{6.12}$$

If we set $\lambda_0 = -\ln\dfrac{\gamma_2}{\beta^{\gamma_1}\Gamma(\gamma_1/\gamma_2)}$, $\lambda_1 = 1-\gamma_1$, $\lambda_2 = \beta^{-\gamma_2}$ and $q = \gamma_2$ we find that

$$f_X(x) = \frac{\gamma_2}{\beta\Gamma(\gamma_1/\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1-1}\exp\left(-\left(\frac{x}{\beta}\right)^{\gamma_2}\right) \tag{6.13}$$

which is the familiar form the GG distribution.

The Generalized Beta distribution of Second Kind emerged by imposing the constraints given in Eq. (2.8) and Eq. (2.12) as follows:

$$\begin{aligned} f_X(x) &= \exp\left(-\lambda_0 - \sum_{j=1}^{n} \lambda_j g_j(x)\right) = \exp(-\lambda_0 - \lambda_1 \ln x - \lambda_2 \ln(1+px^q)/p) \\ &= \exp\left(-\lambda_0 + \ln x^{-\lambda_1} + \ln(1+px^q)^{-\lambda_2/p}\right) \\ &= \exp\left(-\lambda_0\right)\exp\left(\ln x^{-\lambda_1}\right)\exp\left(\ln(1+px^q)^{-\lambda_2/p}\right) \\ &= \exp\left(-\lambda_0\right)x^{-\lambda_1}(1+px^q)^{-\lambda_2/p} \end{aligned} \tag{6.14}$$

If we set $\lambda_0 = -\ln\frac{\gamma_2}{\beta^{\gamma_1}\Gamma(\gamma_1/\gamma_2)}$, $\lambda_1 = 1-\gamma_1$, $\lambda_2 = \beta^{-\gamma_2}$ and $q = \gamma_2$ we find that

$$f_X(x) = \frac{\gamma_3}{\beta \mathrm{B}(\gamma_1,\gamma_2)}\left(\frac{x}{\beta}\right)^{\gamma_1\gamma_3-1}\left(1+\left(\frac{x}{\beta}\right)^{\gamma_3}\right)^{-(\gamma_1+\gamma_2)} \tag{6.15}$$

which is the familiar form the GB2 distribution.

# APPENDIX B

*"It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts."*

ARTHUR CONAN DOYLE

## THE DATASET

The original database used in this thesis, i.e., the Global Historical Climatology Network-Daily (GHCND) database (version 2.60, www.ncdc.noaa.gov/oa/climate/ghcn-daily) comprises more than 80 000 daily precipitation records from stations all over the world. The spatial distribution of those stations is given in Figure B.1 which presents the number of stations in geographical cells defined by latitude and longitude differences $\Delta\varphi = 2.5°$ and $\Delta\lambda = 5°$, respectively.

Nevertheless, for the purposes of the analyses contacted here among those thousands of stations only those satisfying the following criteria were selected: (a) record length greater or equal than 50 years, (b) percentage of missing values per record less than 20%, and (c) percentage of values assigned with "quality flags" per record less than 0.1%. These criteria resulted in a total of 15 137 stations. The spatial distribution of those stations in depicted in the map of Figure B.2, while the map given in Figure B.3 presents the average record length of those stations per cell. Obviously, many stations have the same record length, yet the period they cover might differ. The graphs of Figure B.4 present the number of stations *vs*. the starting (Figure B.4a) and ending (Figure B.4b) recording year.

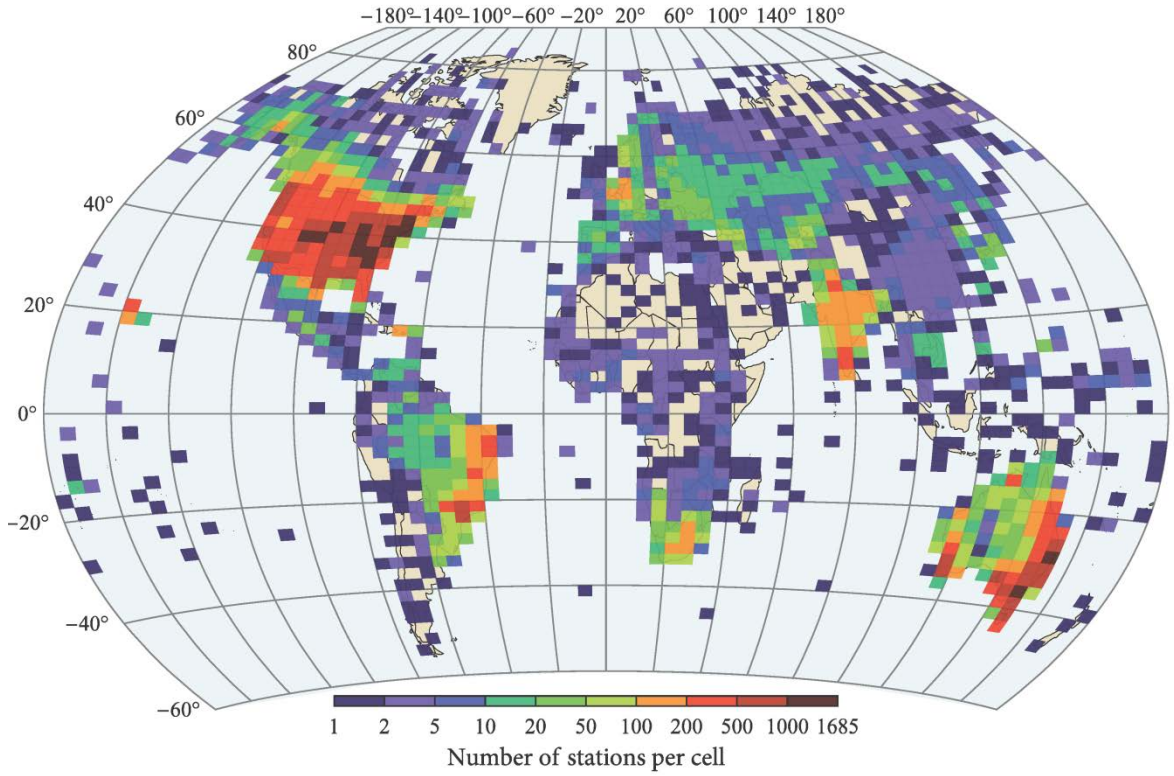



**Figure B.1.** Spatial distribution of the stations comprised in the original database which contains more than 80 000 stations.

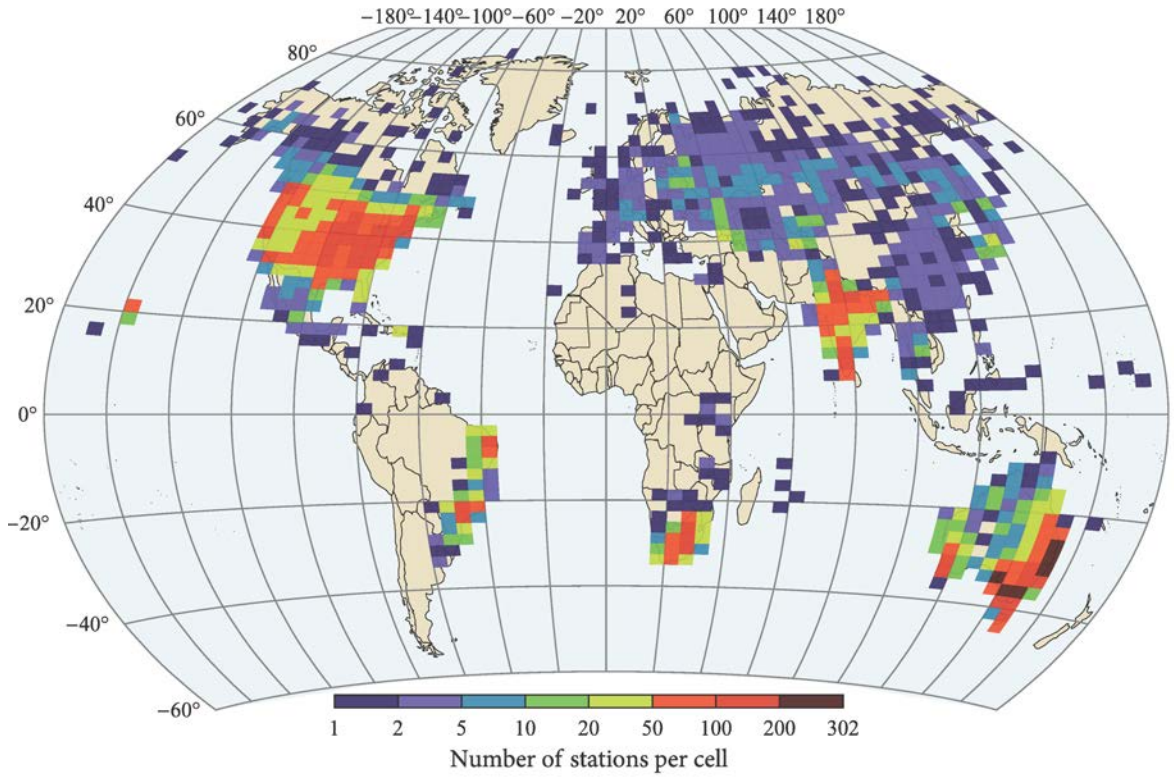



**Figure B.2.** Spatial distribution of the 15 137 stations selected.

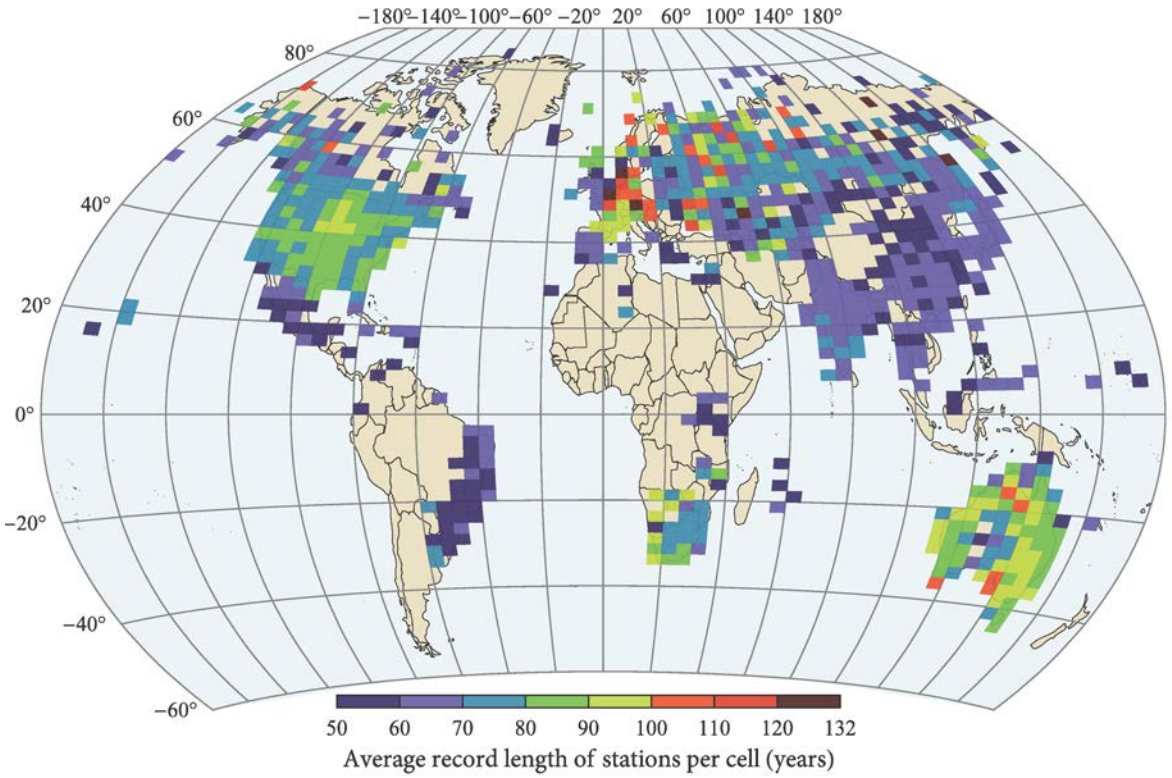



**Figure B.3.** Average record length per cell of the 15 137 stations selected.

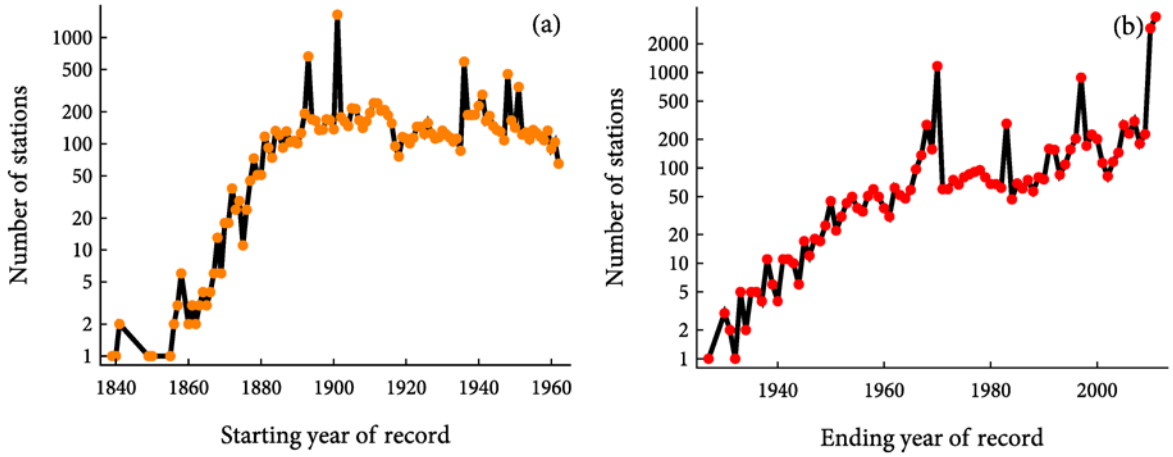



**Figure B.4.** Number of the 15 137 stations *vs.*: (a) starting record year, and (b) ending record year.

Additional information is given in the graphs of Figure B.5 that present the empirical distributions or the histograms of: (a) the record length, (b) the probability dry, (c) the percentage of missing values, and (d) the number of quality flags. It is noted that the

majority of records studied have record length less than 100 years, yet there are a few thousand larger records. The most common probability dry values lie between 70% and 90%. The Figure B.5c indicates that most stations have missing values less than 10% while the most common value lies in between 0% and 2%. Regarding the quality flags, it is apparent from the Figure B.5d that the vast majority of stations have only up to two daily values assigned with quality flags. Moreover, Figure B.6 present the empirical distributions or the histograms of: (a) the total number of daily values, (b) the number of nonzero daily values, (c) the number of zero values, and (d) the number of missing daily values of the 15 137 stations studied.

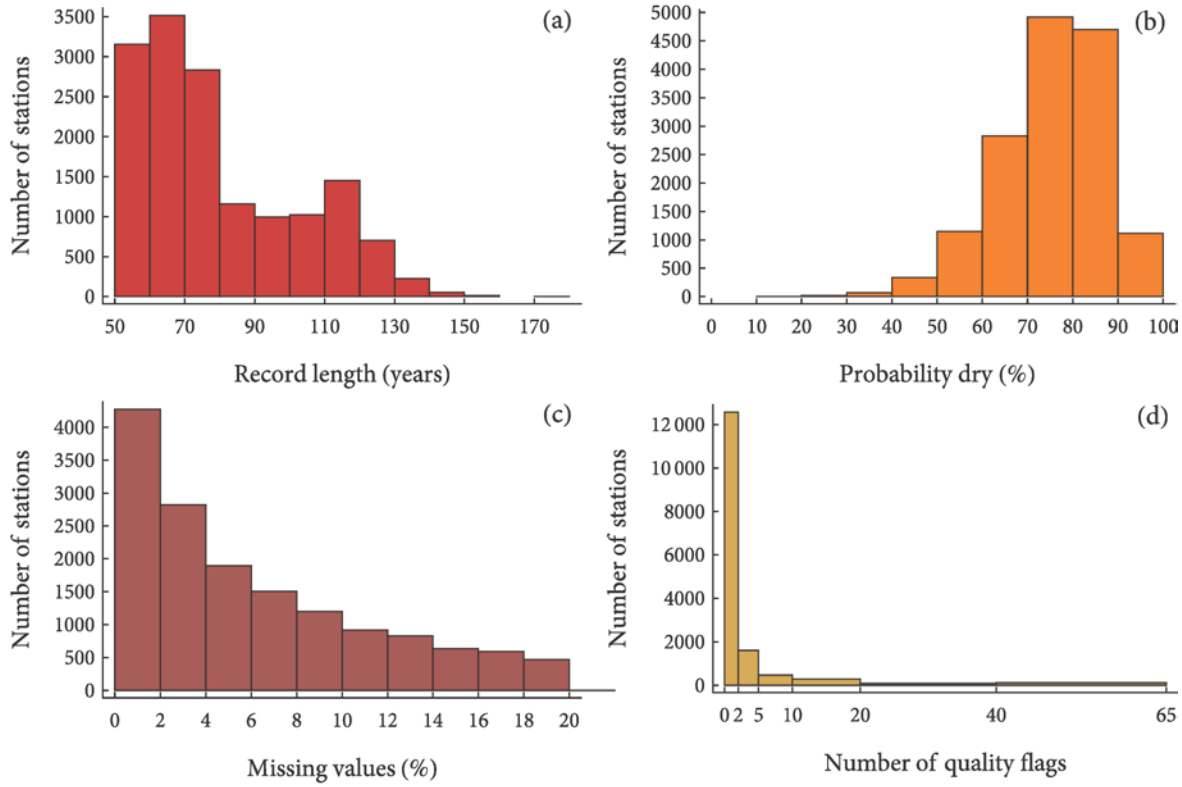



**Figure B.5.** Empirical distributions of the 15 137 stations for: (a) the record length, (b) the probability dry, (c) the missing values, and (d) the number of quality flags.

Finally, Table B.1 provides a summary of the data used in the chapters of this thesis. The data used in Chapters 2, 3 and 4 were extracted from the 15 137 records of daily precipitation resulted from the aforementioned criteria applied to the original GHCND database.

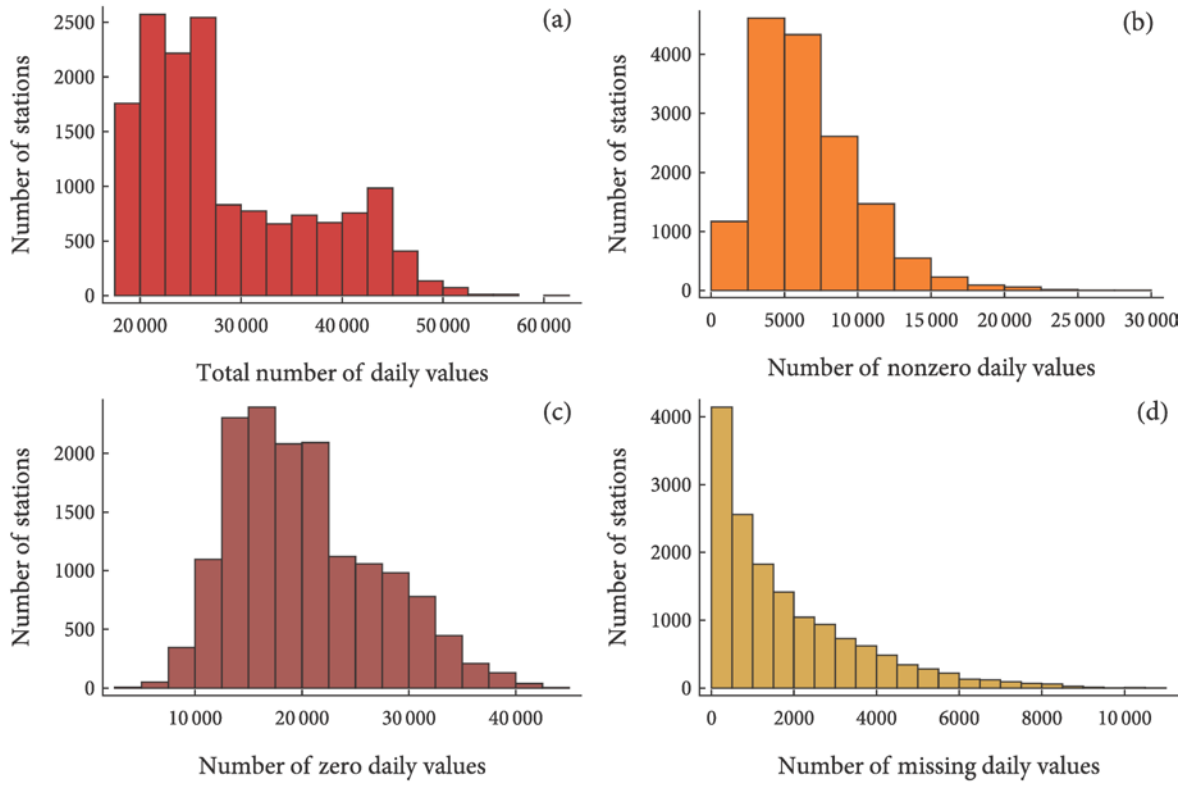



**Figure B.6.** Empirical distributions of the 15 137 stations for: (a) the total number of daily values, (b) the number of nonzero daily values, (c) the number of zero values, and (d) the number of missing daily values.

**Table B.1.** A summary of the data used in this thesis.

| Chapter | Data type | Records No. | Comments |
|---|---|---|---|
| Ch. 2 | i. Daily precipitation<br>ii. Monthly daily precipitation | 14 157<br>169 884 | The monthly daily records were constrained to have at least 20 nonzero values to assure the reliability of the analysis. This additional criterion excluded 980 stations. |
| Ch. 3 | Annual exceedance precipitation | 15 029 | The fit of the theoretical tails failed in 108 records due to algorithmic convergence issues. |
| Ch. 4 | Annual maxima of daily precipitation | 15 137 | Annual maxima records were successfully extracted from all available daily records. |
| Ch. 5 | Precipitation events with temporal resolution 10 s | 7 | The original resolution of some records was 5 s; these records, for uniformity, were transformed also to the 10 s resolution. |

# APPENDIX C

*"Somewhere, something incredible is waiting to be known."*

*CARL SAGAN*

# L-RATIO PLOTS OF DAILY RAINFALL

Figures C.1-C.4 present the observed L-points of the nonzero daily rainfall for individual months while Figure C.5 present the observed L-points of the nonzero daily rainfall of all months. The observed L-points are superimposed over the theoretical L-areas formed by the GG and Burr type XII distributions. At each plot empirical points are colored in three ways; the red-colored points lie outside the area; the dark-colored indicate a Bell-shaped distribution; the light-colored indicate a J-shaped distribution. Additionally, Table C.1 presents some basic summary statistics of the estimated shape parameters of the fitted GG and BrXII distributions.

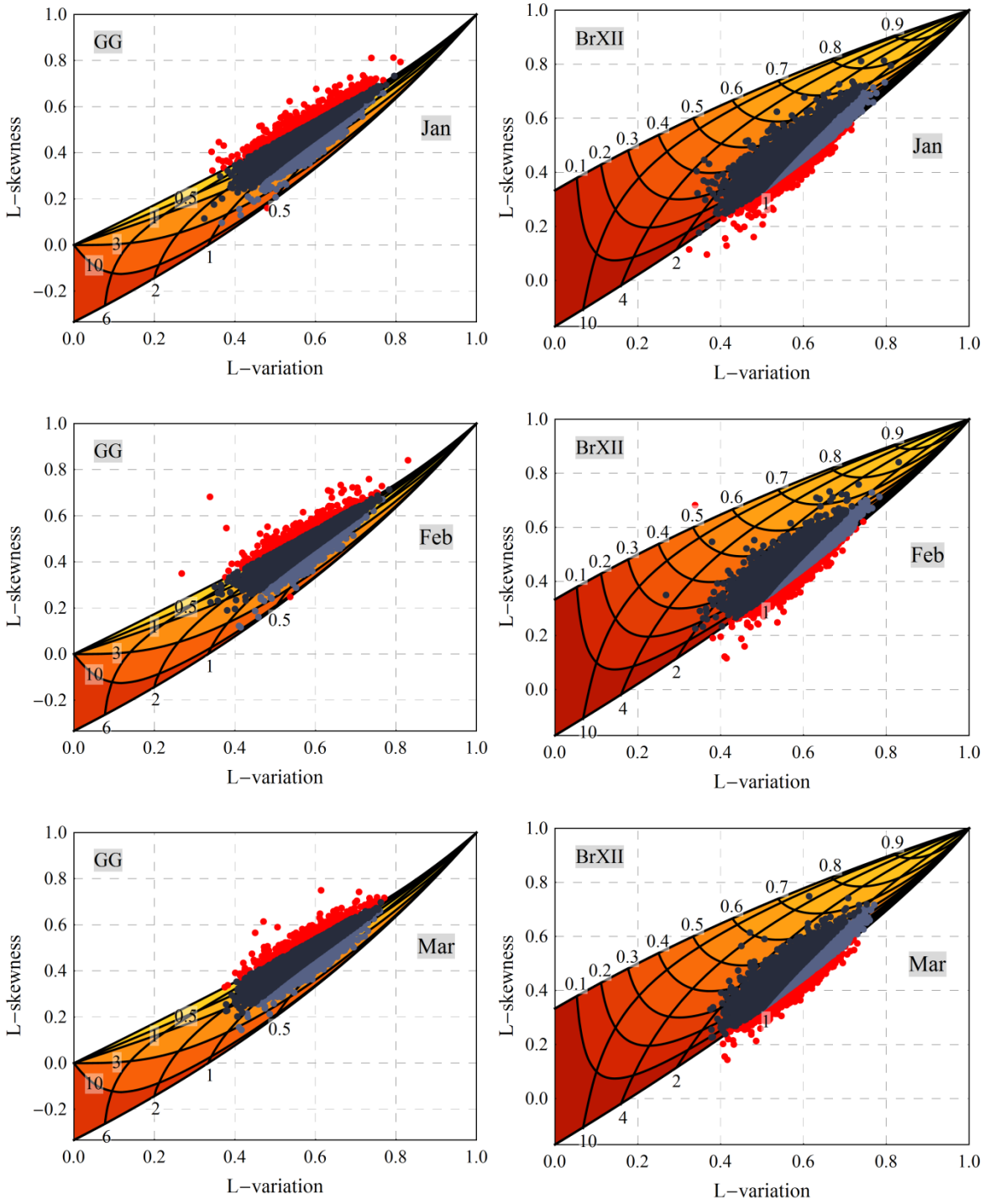



**Figure C.1.** Observed L-points of the 14 157 stations studied for the months January to March.

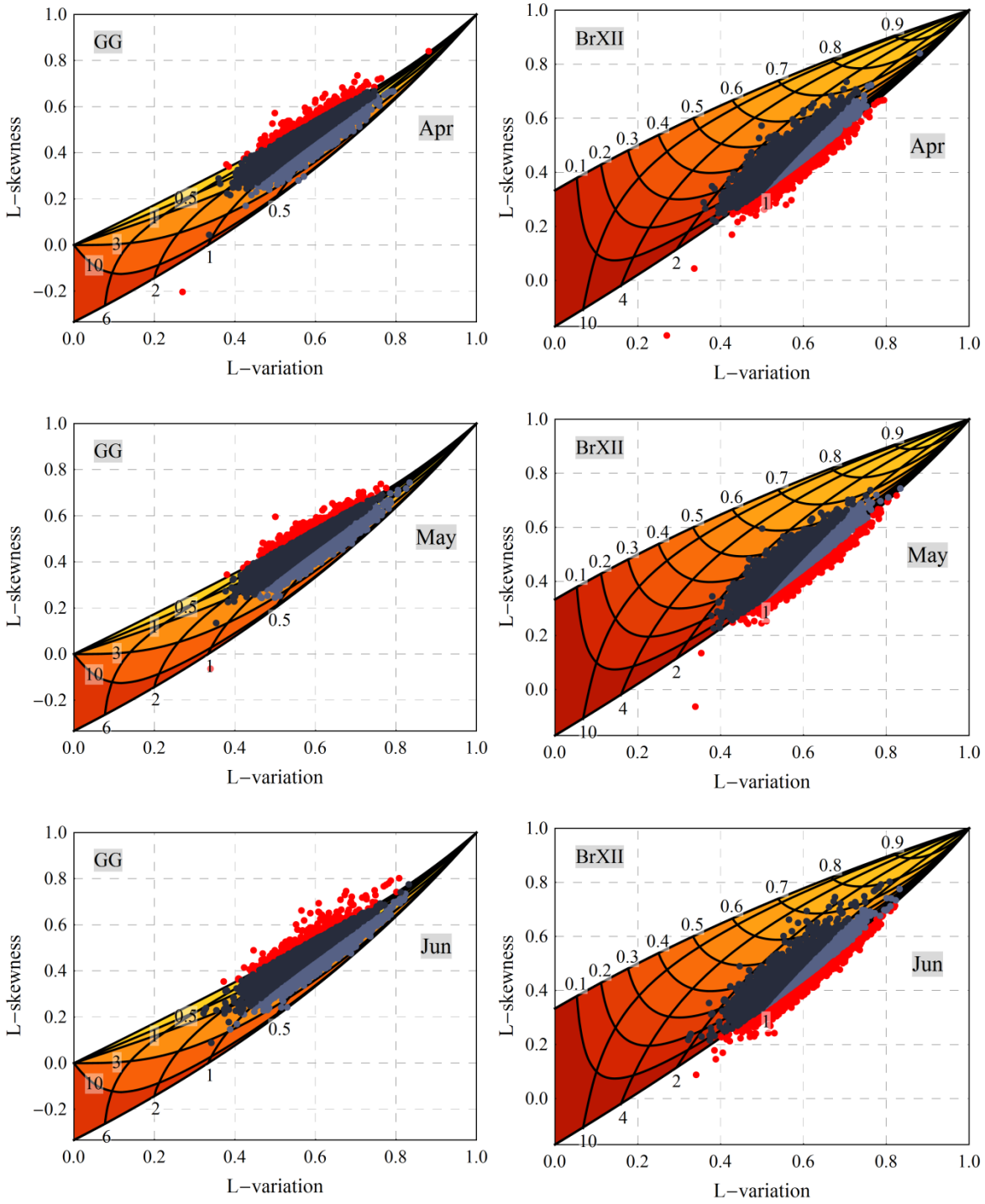



**Figure C.2.** Observed L-points of the 14 157 stations studied for the months April to June

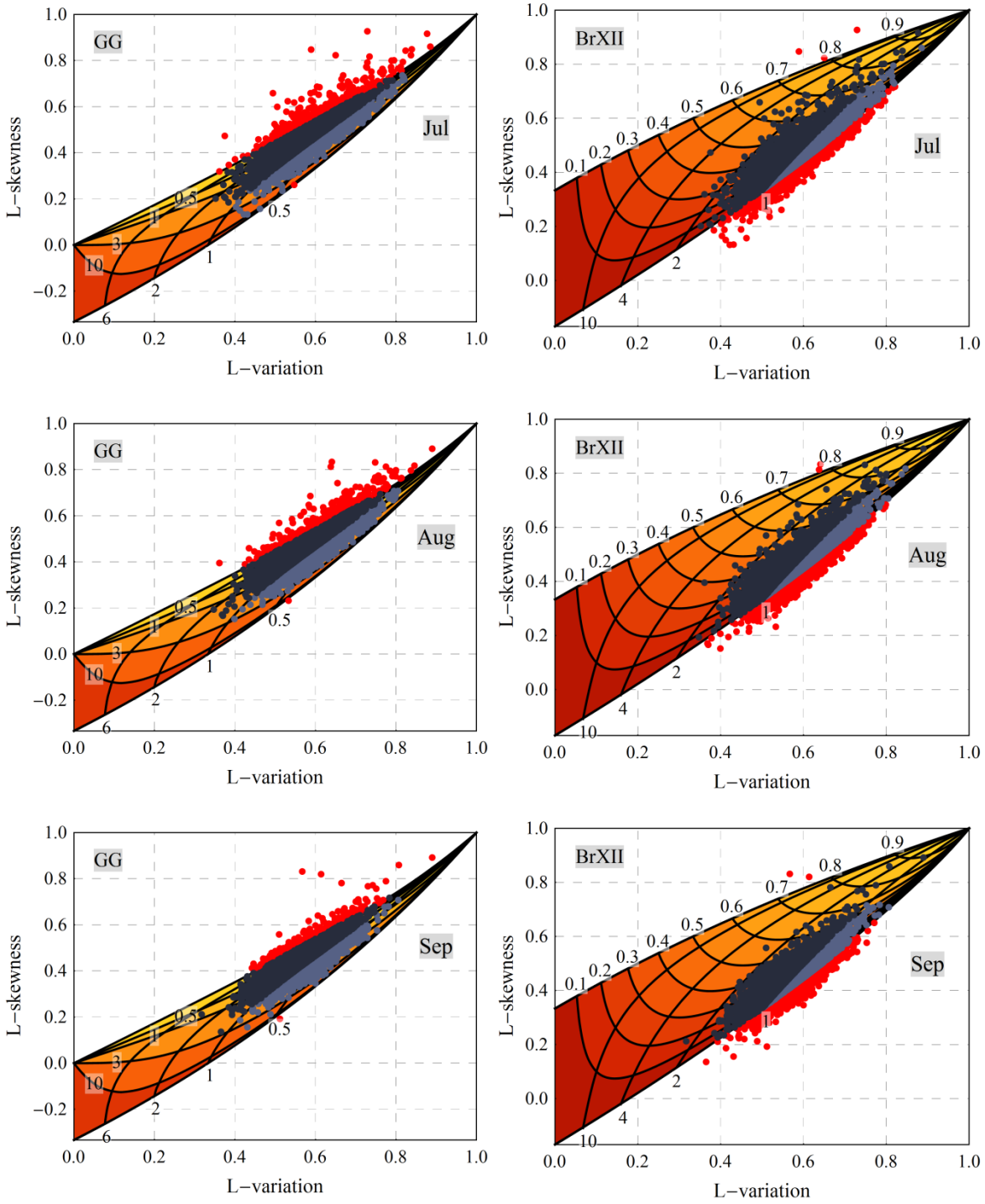



**Figure C.3.** Observed L-points of the 14 157 stations studied for the months July to September.

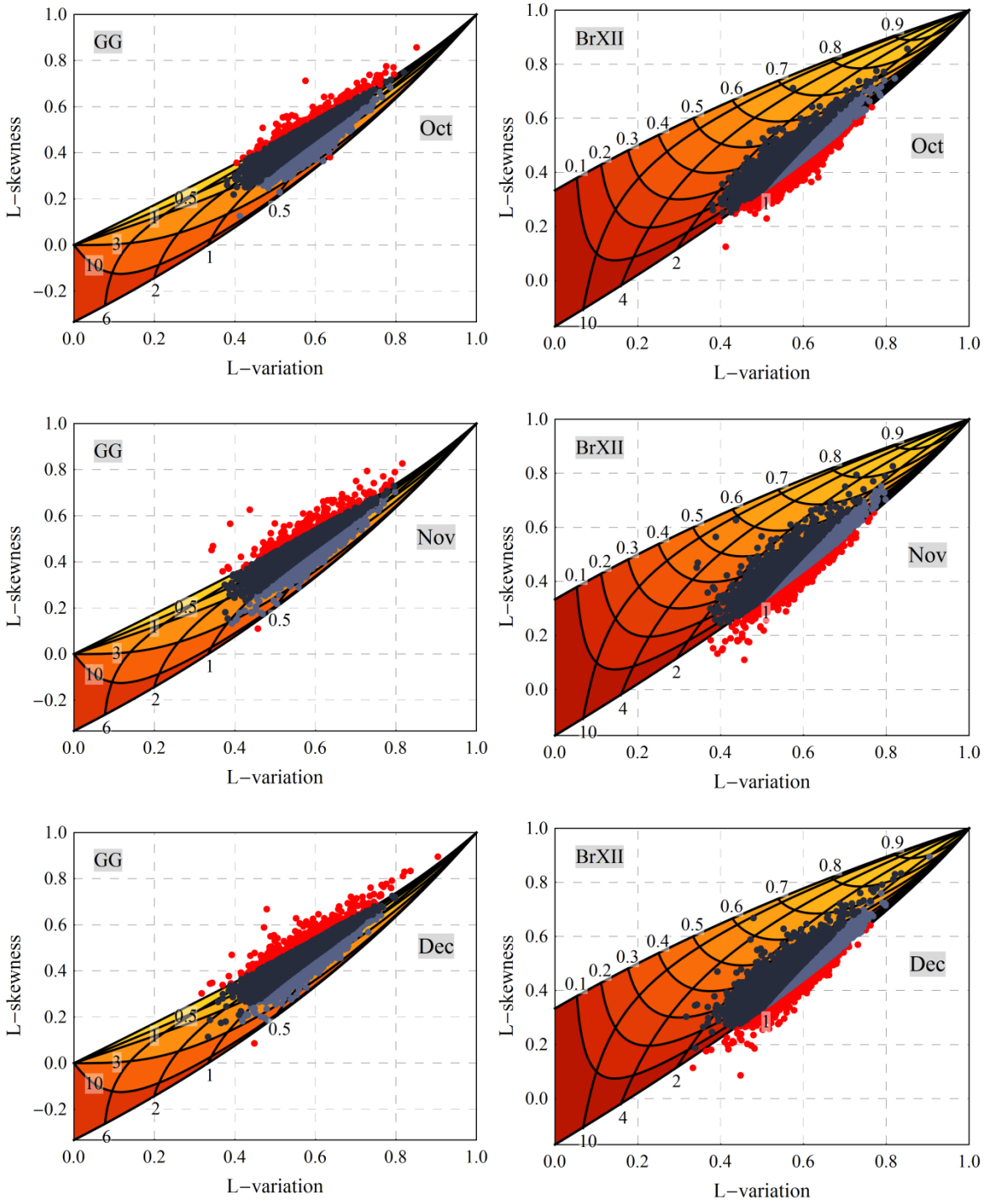



**Figure C.4.** Observed L-points of the 14 157 stations studied for the months October to December.

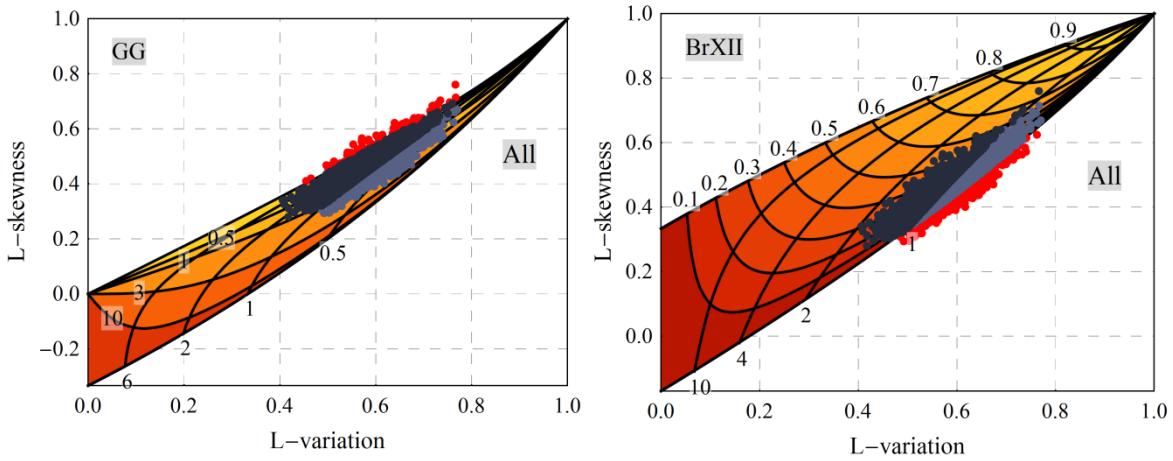



**Figure C.5.** Observed L-points of the 14 157 stations studied for all months.

**Table C.1.** Basic summary statistics of the estimated shape parameters of the GG and BrXII distributions.

| | All | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GG distribution | | | | | | | | | | | | | |
| Fit No. | 13826 | 12729 | 13012 | 13116 | 13353 | 13445 | 13491 | 13292 | 13317 | 13509 | 13620 | 13410 | 13000 |
| Parameter $\gamma_1$ | | | | | | | | | | | | | |
| $Q_{50}$ | 1.20 | 1.23 | 1.22 | 1.17 | 1.13 | 1.09 | 1.08 | 1.09 | 1.10 | 1.09 | 1.10 | 1.13 | 1.21 |
| $\mu$ | 1.50 | 1.63 | 1.59 | 1.53 | 1.45 | 1.39 | 1.36 | 1.41 | 1.43 | 1.41 | 1.42 | 1.49 | 1.61 |
| $\sigma$ | 0.94 | 1.22 | 1.15 | 1.07 | 1.00 | 0.97 | 0.94 | 1.01 | 1.04 | 1.02 | 1.02 | 1.11 | 1.20 |
| $\tau_2$ | 0.29 | 0.34 | 0.33 | 0.32 | 0.31 | 0.30 | 0.30 | 0.31 | 0.32 | 0.31 | 0.31 | 0.33 | 0.34 |
| $\tau_3$ | 0.38 | 0.43 | 0.42 | 0.42 | 0.42 | 0.43 | 0.43 | 0.43 | 0.44 | 0.44 | 0.43 | 0.43 | 0.42 |
| Parameter $\gamma_2$ | | | | | | | | | | | | | |
| $Q_{50}$ | 0.52 | 0.54 | 0.54 | 0.58 | 0.61 | 0.62 | 0.61 | 0.60 | 0.59 | 0.59 | 0.60 | 0.60 | 0.56 |
| $\mu$ | 0.53 | 0.58 | 0.58 | 0.59 | 0.62 | 0.62 | 0.62 | 0.61 | 0.60 | 0.60 | 0.61 | 0.63 | 0.60 |
| $\sigma$ | 0.22 | 0.30 | 0.31 | 0.28 | 0.28 | 0.26 | 0.27 | 0.28 | 0.27 | 0.27 | 0.28 | 0.32 | 0.31 |
| $\tau_2$ | 0.23 | 0.28 | 0.28 | 0.26 | 0.25 | 0.23 | 0.23 | 0.24 | 0.24 | 0.23 | 0.24 | 0.26 | 0.28 |
| $\tau_3$ | 0.06 | 0.14 | 0.14 | 0.08 | 0.06 | 0.04 | 0.08 | 0.09 | 0.09 | 0.10 | 0.10 | 0.12 | 0.13 |
| Burr XII distribution | | | | | | | | | | | | | |
| Fit No. | 12744 | 11900 | 11827 | 11810 | 11555 | 11460 | 11544 | 11737 | 11878 | 11768 | 11503 | 11203 | 11551 |
| Parameter $\gamma_1$ | | | | | | | | | | | | | |
| $Q_{50}$ | 0.94 | 1.00 | 0.98 | 0.98 | 0.97 | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 | 0.96 | 0.99 | 1.01 |
| $\mu$ | 0.96 | 1.05 | 1.03 | 1.01 | 1.00 | 0.99 | 0.98 | 0.99 | 0.99 | 0.98 | 0.99 | 1.02 | 1.05 |
| $\sigma$ | 0.16 | 0.24 | 0.23 | 0.21 | 0.18 | 0.18 | 0.19 | 0.21 | 0.20 | 0.19 | 0.19 | 0.23 | 0.24 |
| $\tau_2$ | 0.09 | 0.12 | 0.12 | 0.11 | 0.10 | 0.10 | 0.10 | 0.11 | 0.11 | 0.10 | 0.10 | 0.11 | 0.11 |
| $\tau_3$ | 0.14 | 0.21 | 0.21 | 0.20 | 0.18 | 0.19 | 0.21 | 0.22 | 0.19 | 0.19 | 0.16 | 0.18 | 0.19 |
| Parameter $\gamma_2$ | | | | | | | | | | | | | |
| $Q_{50}$ | 0.21 | 0.25 | 0.24 | 0.22 | 0.20 | 0.19 | 0.19 | 0.20 | 0.20 | 0.19 | 0.20 | 0.21 | 0.24 |
| $\mu$ | 0.22 | 0.25 | 0.24 | 0.23 | 0.22 | 0.21 | 0.20 | 0.21 | 0.21 | 0.21 | 0.21 | 0.22 | 0.24 |
| $\sigma$ | 0.11 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.13 | 0.12 | 0.12 | 0.13 | 0.13 |
| $\tau_2$ | 0.30 | 0.30 | 0.30 | 0.33 | 0.35 | 0.36 | 0.35 | 0.35 | 0.34 | 0.33 | 0.33 | 0.32 | 0.31 |
| $\tau_3$ | 0.02 | 0.05 | 0.04 | 0.07 | 0.09 | 0.11 | 0.12 | 0.12 | 0.12 | 0.10 | 0.09 | 0.07 | 0.04 |

# LIST OF PUBLICATIONS

- ***Peer-reviewed articles***


[1]. Lombardo, F., E. Volpi, D. Koutsoyiannis, and S. M. Papalexiou (2013), Just two moments! A cautionary note against use of high-order moments in multifractal models in hydrology, Hydrology and Earth System Sciences Discussions, 10(4), 4627–4654, doi:10.5194/hessd-10-4627-2013.

[2]. Papalexiou, S. M., and D. Koutsoyiannis (2013a), A global survey on the seasonal variation of the marginal distribution of daily rainfall, Journal of Hydrology, submitted.

[3]. Papalexiou, S. M., and D. Koutsoyiannis (2013b), Battle of extreme value distributions: A global survey on extreme daily rainfall, Water Resources Research, 49(1), 187–201, doi:10.1029/2012WR012557.

[4]. Papalexiou, S. M., D. Koutsoyiannis, and C. Makropoulos (2013), How extreme is extreme? An assessment of daily rainfall distribution tails, Hydrol. Earth Syst. Sci., 17(2), 851–862, doi:10.5194/hess-17-851-2013.

[5]. Papalexiou, S. M., and D. Koutsoyiannis (2012), Entropy based derivation of probability distributions: A case study to daily rainfall, Advances in Water Resources, 45, 51–57, doi:10.1016/j.advwatres.2011.11.007.

[6]. Papalexiou, S. M., D. Koutsoyiannis, and C. Makropoulos (2012), How extreme is extreme? An assessment of daily rainfall distribution tails, Hydrology and Earth System Sciences Discussions, 9(5), 5757–5778, doi:10.5194/hessd-9-5757-2012.

[7]. Papalexiou, S.-M., D. Koutsoyiannis, and A. Montanari (2011), Can a simple stochastic model generate rich patterns of rainfall events?, Journal of Hydrology, 411(3–4), 279–289, doi:10.1016/j.jhydrol.2011.10.008.

[8]. Papalexiou, S. M., and D. Koutsoyiannis (2006), A probabilistic approach to the concept of Probable Maximum Precipitation, Adv. Geosci., 7, 51–54, doi:10.5194/adgeo-7-51-2006.

- ***Book chapters***

[9]. Grimaldi, S., S.-C. Kao, A. Castellarin, S.-M. Papalexiou, A. Viglione, F. Laio, H. Aksoy, and A. Gedikli (2011), 2.18 - Statistical Hydrology, in Treatise on Water Science, edited by Peter Wilderer, pp. 479–517, Elsevier, Oxford. [online] Available from: http://www.sciencedirect.com/science/article/pii/B9780444531995000464 (Accessed 12 June 2013)

- ***Conference publications and presentations with evaluation of abstract***

[10]. Anagnostopoulou, E. et al. (2013), Record breaking properties for typical autocorrelation structures, in *European Geosciences Union General Assembly 2013*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1331/ (Accessed 12 June 2013)

[11]. Mystegniotis, A., V. Vasilaki, I. Pappa, S. Curceac, D. Saltouridou, N. Efthimiou, G. Papatsoutsos, S. M. Papalexiou, and D. Koutsoyiannis (2013), Clustering of extreme events in typical stochastic models, in *European Geosciences Union General Assembly 2013*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1332/ (Accessed 12 June 2013)

[12]. Nerantzaki, S., S. M. Papalexiou, and D. Koutsoyiannis (2013), Extreme rainfall distribution tails: Exponential, subexponential or hyperexponential?, in *European Geosciences Union General Assembly 2013*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1333/ (Accessed 12 June 2013)

[13]. Venediki, A., S. Giannoulis, C. Ioannou, L. Malatesta, G. Theodoropoulos, G. Tsekouras, Y. Dialynas, S. M. Papalexiou, A. Efstratiadis, and D. Koutsoyiannis (2013), The Castalia stochastic generator and its applications to multivariate disaggregation of hydro-meteorological processes, in *European Geosciences Union General Assembly 2013*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1325/ (Accessed 12 June 2013)

[14]. Iliopoulou, T., S. M. Papalexiou, and D. Koutsoyiannis (2013), Assessment of the dependence structure of the annual rainfall using a large dataset, in *European Geosciences Union General Assembly 2013*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1335/ (Accessed 12 June 2013)

[15]. Markonis, Y., S. M. Papalexiou, and D. Koutsoyiannis (2013), The role of teleconnections in extreme (high and low) precipitation events: The case of the Mediterranean region, in *European Geosciences Union General Assembly 2013*.

[online] Available from: http://www.itia.ntua.gr/en/docinfo/1324/ (Accessed 12 June 2013)

[16]. Giannoulis, S., C. Ioannou, E. Karantinos, L. Malatesta, G. Theodoropoulos, G. Tsekouras, A. Venediki, P. Dimitriadis, S. M. Papalexiou, and D. Koutsoyiannis (2012), Long term properties of monthly atmospheric pressure fields, in *European Geosciences Union General Assembly 2012*, p. 4680. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1208/ (Accessed 12 June 2013)

[17]. Houdalaki, E. et al. (2012), On statistical biases and their common neglect, in *European Geosciences Union General Assembly 2012*, p. 4388. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1205/ (Accessed 12 June 2013)

[18]. Papalexiou, S. M., and D. Koutsoyiannis (2012), A global survey on the distribution of annual maxima of daily rainfall: Gumbel or Fréchet?, in *European Geosciences Union General Assembly 2012*, p. 10563. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1207/ (Accessed 12 June 2013)

[19]. Koutsoyiannis, D., and S. M. Papalexiou (2011), Scaling as enhanced uncertainty, in *European Geosciences Union General Assembly 2011*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1115/ (Accessed 12 June 2013)

[20]. Papalexiou, S. M., and D. Koutsoyiannis (2011a), A worldwide probabilistic analysis of rainfall at multiple timescales based on entropy maximization, in *European Geosciences Union General Assembly 2011*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1129/ (Accessed 12 June 2013)

[21]. Papalexiou, S. M., and D. Koutsoyiannis (2011b), Entropy maximization, p-moments and power-type distributions in nature, in *European Geosciences Union General Assembly 2011*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1127/ (Accessed 12 June 2013)

[22]. Papalexiou, S. M., E. Kallitsi, E. Steirou, M. Xirouchakis, A. Drosou, V. Mathios, H. Adraktas-Rentis, I. Kyprianou, M.-A. Vasilaki, and D. Koutsoyiannis (2011), Long-term properties of annual maximum daily rainfall worldwide, in *European Geosciences Union General Assembly 2011*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1124/ (Accessed 12 June 2013)

[23]. Dialynas, Y., P. Kossieris, K. Kyriakidis, A. Lykou, Y. Markonis, C. Pappas, S. M. Papalexiou, and D. Koutsoyiannis (2010), Optimal infilling of missing values in hydrometeorological time series, in *European Geosciences Union General Assembly*

*2010*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/981/ (Accessed 12 June 2013)

[24]. Efstratiadis, A., and S. M. Papalexiou (2010), The quest for consistent representation of rainfall and realistic simulation of process interactions in flood risk assessment, in *European Geosciences Union General Assembly 2010*, p. 11101. [online] Available from: http://www.itia.ntua.gr/en/docinfo/961/ (Accessed 12 June 2013)

[25]. Papalexiou, S. M., and D. Koutsoyiannis (2010a), A world-wide investigation of the probability distribution of daily rainfall, in *International Precipitation Conference (IPC10)*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1003/ (Accessed 12 June 2013)

[26]. Papalexiou, S. M., and D. Koutsoyiannis (2010b), On the tail of the daily rainfall probability distribution: Exponential-type, power-type or something else?, in *European Geosciences Union General Assembly 2010*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/977/ (Accessed 12 June 2013)

[27]. Papalexiou, S. M., D. Koutsoyiannis, and A. Montanari (2010), Mind the bias!, in *STAHY Official Workshop: Advances in statistical hydrology*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/985/ (Accessed 12 June 2013)

[28]. Bouziotas, D., G. Deskos, N. Mastrantonas, D. Tsaknias, G. Vangelidis, S. M. Papalexiou, and D. Koutsoyiannis (2011), Long-term properties of annual maximum daily river discharge worldwide, in *European Geosciences Union General Assembly 2011*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/1128/ (Accessed 12 June 2013)

[29]. Katerinopoulou, A., K. Kagia, M. Karapiperi, A. Kassela, A. Paschalis, G.-M. Tsarouchi, Y. Markonis, S. M. Papalexiou, and D. Koutsoyiannis (2009), Reservoir yield-reliability relationship and frequency of multi-year droughts for scaling and non-scaling reservoir inflows, in *European Geosciences Union General Assembly 2009*, p. 8063. [online] Available from: http://www.itia.ntua.gr/en/docinfo/899/ (Accessed 12 June 2013)

[30]. Papalexiou, S. M., and D. Koutsoyiannis (2009a), An all-timescales rainfall probability distribution, in *European Geosciences Union General Assembly 2009*, p. 13469. [online] Available from: http://www.itia.ntua.gr/en/docinfo/902/ (Accessed 12 June 2013)

[31]. Papalexiou, S. M., and D. Koutsoyiannis (2009b), Ombrian curves: from theoretical consistency to engineering practice, in *8th IAHS Scientific Assembly / 37th IAH*

*Congress*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/926/ (Accessed 12 June 2013)

[32]. Papalexiou, S. M., and N. Zarkadoulas (2009), The trendy trends: a fashion or a science story?, in *European Geosciences Union General Assembly 2009*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/946/ (Accessed 12 June 2013)

[33]. Koutsoyiannis, D., N. Mamassis, A. Christofides, A. Efstratiadis, and S. M. Papalexiou (2008), Assessment of the reliability of climate predictions based on comparisons with historical time series, in *European Geosciences Union General Assembly 2008*, p. 09074. [online] Available from: http://www.itia.ntua.gr/en/docinfo/850/ (Accessed 12 June 2013)

[34]. Papalexiou, S. M., and D. Koutsoyiannis (2008a), Ombrian curves in a maximum entropy framework, in *European Geosciences Union General Assembly 2008*, p. 00702. [online] Available from: http://www.itia.ntua.gr/en/docinfo/851/ (Accessed 12 June 2013)

[35]. Papalexiou, S. M., and D. Koutsoyiannis (2008b), Probabilistic description of rainfall intensity at multiple time scales, in *IHP 2008 Capri Symposium: "The Role of Hydrology in Water Resources Management."* [online] Available from: http://www.itia.ntua.gr/en/docinfo/884/ (Accessed 12 June 2013)

[36]. Zarkadoulas, N., D. Koutsoyiannis, N. Mamassis, and S. M. Papalexiou (2008), Climate, water and health in ancient Greece, in *European Geosciences Union General Assembly 2008*, p. 12006. [online] Available from: http://www.itia.ntua.gr/en/docinfo/848/ (Accessed 12 June 2013)

[37]. Koutsoyiannis, D., S. M. Papalexiou, and A. Montanari (2007), Can a simple stochastic model generate a plethora of rainfall patterns? (invited), in *The Ultimate Rainmap: Rainmap Achievements and the Future in Broad-Scale Rain Modelling*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/789/ (Accessed 12 June 2013)

[38]. Mackey, R., and S. M. Papalexiou (2007), Sources of the stochastic regulation of climate, in *European Geosciences Union General Assembly 2007*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/836/ (Accessed 12 June 2013)

[39]. Montanari, A., D. Koutsoyiannis, and S. M. Papalexiou (2007), The omnipresence of scaling behaviour in hydrometeorological time series and its implications in climatic change assessments, in *XXIV General Assembly of the International Union of Geodesy*

*and Geophysics*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/786/ (Accessed 12 June 2013)

[40]. Papalexiou, S. M., A. Montanari, and D. Koutsoyiannis (2007), Scaling properties of fine resolution point rainfall and inferences for its stochastic modelling, in *European Geosciences Union General Assembly 2007*, p. 11253. [online] Available from: http://www.itia.ntua.gr/en/docinfo/751/ (Accessed 12 June 2013)

[41]. Papalexiou, S. M. (2007), Stochastic modelling of skewed data exhibiting long-range dependence, in *XXIV General Assembly of the International Union of Geodesy and Geophysics*. [online] Available from: http://www.itia.ntua.gr/en/docinfo/835/ (Accessed 12 June 2013)

[42]. Efstratiadis, A., A. Tegos, I. Nalbantis, E. Rozos, A. Koukouvinos, N. Mamassis, S. M. Papalexiou, and D. Koutsoyiannis (2005), Hydrogeios, an integrated model for simulating complex hydrographic networks - A case study to West Thessaly region, in *7th Plinius Conference on Mediterranean Storms*, EGU. [online] Available from: http://www.itia.ntua.gr/en/docinfo/693/ (Accessed 12 June 2013)

[43]. Papalexiou, S. M., and D. Koutsoyiannis (2005), A probabilistic approach to the concept of Probable Maximum Precipitation, in *7th Plinius Conference on Mediterranean Storms*, EGU. [online] Available from: http://www.itia.ntua.gr/en/docinfo/694/ (Accessed 12 June 2013)