NATIONAL TECHNICAL UNIVERSITY OF ATHENS
INTERDEPARTMENTAL PROGRAM
“WATER RESOURCES SCIENCE AND TECHNOLOGY”

MASTER’S THESIS:
“IMPACT OF TIME’S ARROW ON STREAMFLOW AND ITS STOCHASTIC MODELLING”

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ:
“Η ΕΠΙΡΡΟΗ ΤΟΥ ΒΕΛΟΥΣ ΤΟΥ ΧΡΟΝΟΥ ΣΤΗΝ ΑΠΟΡΡΟΗ ΚΑΙ ΣΤΗΝ ΣΤΟΧΑΣΤΙΚΗ ΤΗΣ ΜΟΝΤΕΛΟΠΟΙΗΣΗ”

VAVOULOGIANNIS STYLIANOS

ΒΑΒΟΥΛΟΓΙΑΝΝΗΣ ΣΤΥΛΙΑΝΟΣ

SUPERVISOR: PROF. KOUTSOYIANNIS DEMETRIS

ΕΠΙΒΛΕΠΩΝ: ΚΑΘ. ΚΟΥΤΣΟΓΙΑΝΝΗΣ ΔΗΜΗΤΡΗΣ

ATHENS, FEBRUARY, 2020

ΕΥΧΑΡΙΣΤΙΕΣ

Η μεταπτυχιακή αυτή εργασία εκπονήθηκε στα πλαίσια της ολοκλήρωσης των μεταπτυχιακών μου σπουδών στο διατμηματικό μεταπτυχιακό πρόγραμμα: "Επιστήμη και Τεχνολογία Υδατικών Πόρων".

Θα ήθελα να ευχαριστήσω θερμά τον Δ. Κουτσογιάννη για την ευκαιρία που μου έδωσε να αναλάβω ένα τόσο ενδιαφέρον θέμα αλλά και για τη διαρκή του υποστήριξη καθ'όλη τη διάρκεια της εκπόνησης της εργασίας. Το μάθημα του αλλά και η μεταπτυχιακή εργασία με έκαναν να αλλάξω τελείως τον τρόπο που βλέπω την επιστήμη της Υδρολογίας.

Παράλληλα θα ήθελα να ευχαριστήσω την Θ. Ηλιοπούλου που παρείχε διαρκή υποστήριξη και καθοδήγηση σε όλα τα στάδια της εργασίας. Με τις συμβουλές της είχε πολύ καθοριστική συμβολή στην ολοκλήρωση της εργασίας.

Επιπλέον θα ήθελα να ευχαριστήσω τον Π. Δημητριάδη για τις διαλέξεις του μου προκάλεσαν το ενδιαφέρον για το αντικείμενο των στοχαστικών και επίσης για τις επισημάνσεις του σε καίρια ζητήματα της μεταπτυχιακής εργασίας.

Τέλος θα ήθελα να ευχαριστήσω το Νίκο και την Άννα που συνέβαλαν με τον τρόπο τους στην ολοκλήρωση της εργασίας αλλά και όλο το διδακτικό προσωπικό του μεταπτυχιακού για τις πολύτιμες γνώσεις που μου μετέδωσαν.

ABSTRACT

We investigate the impact of time's arrow on hydrological processes. The role of stochastic simulation and uncertainty is first investigated. Uncertainty is a major factor in physical sciences and engineering. The probabilistic behavior of an engineering system is essential considering that uncertainty issues are important and must be managed. The true distribution for the system response is subject to parameter uncertainty and is in most of the times difficult or even impossible to calculate. This is due to the complexity of the hydrosystems. In such cases, stochastic simulation else known as Monte Carlo simulation is a viable tool to provide numerical estimations of the stochastic features of the system response. Long range dependence is a feature connected to uncertainty and is very important in hydrology. It is being discussed through relevant literature and models. Time's arrow or temporal asymmetry is also related to uncertainty and randomness and has an important role in science. It has been implemented in stochastics for some time but it has recently attracted attention in relevant publications in hydrology. Studies have shown that the temporal asymmetry of the streamflow process is marked for scales up to several days and this highlights the need to reproduce it in flood simulations. After a review of the relevant literature, an analytical method based on an asymmetric moving average (AMA) scheme is being used to simulate time series with temporal asymmetry. The temporal asymmetry of real world streamflow time series is being investigated at hourly scale from the large USGS database. Finally a modification of the method that can simulate time asymmetry at two time scales simultaneously is proposed. The method is successfully tested in the physical world through a case study.

ΠΕΡΙΛΗΨΗ

Διερευνούμε την επιρροή του βέλους του χρόνου στις υδρολογικές διεργασίες. Στην αρχή παρουσιάζεται ο ρόλος της στοχαστικής προσομοίωσης και της αβεβαιότητας. Η αβεβαιότητα είναι ένας σημαντικός παράγοντας στις φυσικές επιστήμες και στην επιστήμη του μηχανικού. Η μελέτη της πιθανοτικής συμπεριφοράς ενός συστήματος είναι απαραίτητη λαμβάνοντας υπόψιν ότι το θέμα της αβεβαιότητας είναι σημαντικό και πρέπει το τα διαχειριστούμε. Σε τέτοιες περιπτώσεις, η στοχαστική προσομοίωση ή αλλιώς προσομοίωση Monte Carlo είναι ένα χρήσιμο εργαλείο για την παροχή αριθμητικών εκτιμήσεων των στοχαστικών χαρακτηριστικών της απόκρισης του συστήματος. Η μακροπρόθεσμη χρονική εξάρτηση ή εμμονή είναι στενά συνδεδεμένη με την αβεβαιότητα και είναι σημαντική για την υδρολογία. Μελετάται μέσω σχετικής βιβλιογραφίας και μοντέλων. Το βέλος του χρόνου συνδέεται επίσης στενά με την τυχαιότητα και την αβεβαιότητα και έχει σημαντικό ρόλο στην επιστήμη. Έχει εφαρμοστεί στις στοχαστικές μεθόδους για κάποιο χρονικό διάστημα, αλλά έχει προσελκύσει πρόσφατα την προσοχή σε δημοσιεύσεις σχετικές με την υδρολογία. Έρευνες κατέληξαν στο συμπέρασμα ότι η χρονική ασυμμετρία της φυσικής διεργασίας της απορροής υπάρχει για την κλίμακα μερικών ημερών και αυτό υπογραμμίζει την ανάγκη αναπαραγωγής της σε προσομοιώσεις πλημμυρών. Μετά από μία ανασκόπηση της σχετικής βιβλιογραφίας χρησιμοποιείται μια αναλυτική μέθοδος βασισμένη σε ένα σχήμα ασύμμετρου κινητού μέσου (AMA) για την παραγωγή συνθετικών χρονοσειρών με χρονική ασυμμετρία. Αυτή η εργασία προτείνει μια τροποποίηση της μεθόδου για την προσομοίωση της ασυμμετρίας του χρόνου σε δύο κλίμακες ταυτόχρονα. Για να δοκιμαστεί η αποτελεσματικότητα αυτής της μεθόδου στον φυσικό κόσμο, γίνεται μια μελέτη με πραγματικά δεδομένα. Διερευνάται η χρονική ασυμμετρία των χρονοσειρών απορροής, από τη μεγάλη βάση δεδομένων του USGS, σε ωριαία κλίμακα.

ΕΚΤΕΝΗΣ ΠΕΡΙΛΗΨΗ

Εισαγωγή

Αντικείμενο αυτής της μεταπτυχιακής διπλωματικής εργασίας είναι η μελέτη της επίδρασης του βέλους του χρόνου στα στοχαστικά μοντέλα μικρής χρονικής κλίμακας.

Η εργασία αυτή έχει επηρεαστεί απο την πρόσφατη μελέτη του Δ. Κουτσογιάννη (2019): «Το βέλος του χρόνου στο στοχαστικό χαρακτηρισμό και την προσομοίωση των ατμοσφαιρικών και υδρολογικών διεργασιών». Στην αρχή της διπλωματικής γίνεται μια προσπάθεια παρουσίασης μέρους του επιστημονικού πλαισίου που αποτελεί τη βάση της πρόσφατης μελέτης. Συγχρόνως συζητείται και άλλη σχετική βιβλιογραφία. Αυτό γίνεται για να επισημανθεί η σημασία και η χρησιμότητα των μετέπειτα αποτελεσμάτων

Πραγματικές χρονοσειρές από μεγάλη βάση δεδομένων χρησιμοποιούνται για τη διερεύνηση της μη αντιστρεψιμότητας σε ωριαία κλίμακα. Χρησιμοποιούνται χρονοσειρές απορροής μέχρι και την εκατοστή συναθροισμένη κλίμακα.

Σε αυτή τη μελέτη γίνεται μια τροποποίηση της υπάρχουσας μεθόδου, η οποία διατηρεί τη μη αντιστρεψιμότητα μόνο στην πρώτη κλίμακα και την καθιστά ικανή να διατηρεί την μη αντιστρεψιμότητα στην πρώτη και στη δεύτερη κλίμακα ταυτόχρονα. Στο τέλος, η βασική μέθοδος και η τροποποιημένη επαληθεύονται από φυσικά δεδομένα.

Αβεβαιότητα, αξιοπιστία και μέθοδοι Monte Carlo στα υδροσυστήματα

Στο πλαίσιο της παραγωγής υδρολογικών μοντέλων, η αβεβαιότητα είναι ένας τεράστιος παράγοντας. Γύρω από αυτόν τον όρο υπάρχουν πολλές παρανοήσεις. Επίσης πολλές φορές δεν λαμβάνεται υπόψιν και το αποτέλεσμα είναι δαπανηρό.

Στην πραγματικότητα οι ντετερμινιστικοί νόμοι και η τυχαιότητα συνυπάρχουν και πρέπει να αντιμετωπίζονται και να μοντελοποιούνται με ολιστικό τρόπο. Το αν μια διεργασία είναι πιο στοχαστική ή ντετερμινιστική είναι μόνο ζήτημα χρονικού ορίζοντα.

Η πηγές της αβεβαιότητας στα υδροσυστήματα μπορεί να είναι (Μακρόπουλος και Ευστρατιάδης, 2018):

1. Απλοϊκές παραδοχές μοντέλου για κρίσιμες διεργασίες του συστήματος (σφάλματα δομικών μοντέλων - δομική αβεβαιότητα). Αυτό συμβαίνει όταν λαμβάνουμε υπόψη λιγότερους κανόνες από ότι στην πραγματικότητα υπάρχουν.

2. Ευαισθησία στις αρχικές και οριακές συνθήκες (χαοτικά συστήματα). Όπως και στο προηγούμενο σχήμα.

3. Ανεπαρκής γνώση των κρίσιμων παραμέτρων του συστήματος.

4. Στοχαστική φύση και χωροχρονική μεταβλητότητα των υδρομετεωρολογικών διαδικασιών (π.χ. βροχή, εξάτμιση, απορροή, άνεμος).

5. Σφάλματα μέτρησης και ανακρίβειες.

6. Αλλαγή συστήματος με την πάροδο του χρόνου (λόγω εξωτερικών παραγόντων).

7. Αλλαγές στις αποφάσεις / πολιτικές και ως εκ τούτου στα μέτρα απόδοσης.

Ο βασικός στόχος της ανάλυσης αβεβαιότητας είναι να προσδιοριστούν τα χαρακτηριστικά αβεβαιότητας της εξόδου του συστήματος ως συνάρτηση των αβεβαιοτήτων που σχετίζονται με το ίδιο το μοντέλο συστήματος και τις στοχαστικές παραμέτρους του. Η ανάλυση αβεβαιότητας παρέχει μια επίσημη και μεθοδική δομή για τη μέτρηση της αβεβαιότητας του συστήματος. Επιπλέον, παρέχει πληροφορίες για τη συμβολή κάθε στοχαστικής βασικής παραμέτρου στη συνολική αβεβαιότητα των εξόδων του συστήματος. Οι πληροφορίες αυτές είναι απαραίτητες και μπορούν να οδηγήσουν στον εντοπισμό των παραμέτρων που παίζουν σημαντικότερο ρόλο στην αβεβαιότητα. Η εκτίμησή τους θα οδηγήσει στη μείωση της συνολικής αβεβαιότητας του συστήματος (Tung and Yen, 2005).

Στη συνέχεια γίνεται μία εισαγωγή στις πιθανότητες και τη στατιστική. Δίνονται τα βασικά αξιώματα και ορισμοί.

Ο ορισμός μιας στοχαστικής ανέλιξης $\{X(t), t \in T\}$ είναι ότι είναι "μια οικογένεια τυχαίων μεταβλητών" (Κουτσογιάννης, 1997). Δηλαδή, για κάθε $t \in T, X(t)$ είναι μια τυχαία μεταβλητή. Αναφερόμαστε στο $X(t)$ ως την κατάσταση της ανέλιξης κατά το χρόνο t εάν ο δείκτης t παριστάνει το χρόνο. Οι στοχαστικές ανελίξεις μπορούν να χρησιμοποιηθούν για να περιγράψουν τη χρονική εξέλιξη ή τις χωρικές σχέσεις τυχαίων μεταβλητών

Οι υδρολογικές μεταβλητές μπορούν να θεωρηθούν στοχαστικές ανελίξεις (Κουτσογιάννης, 1997). Το γεγονός ότι μια φυσική διεργασία θεωρείται στοχαστική ανέλιξη δεν σημαίνει ότι δεν έχει καθοριστικό ρόλο. Είναι γνωστό ότι πολλές υδρολογικές διεργασίες παρουσιάζουν ετήσια ντετερμινιστική μεταβλητότητα π.χ. απορροής. Αυτή η μεταβλητότητα μπορεί να θεωρηθεί ως τυχαία μεταβλητότητα που συμβαίνει σε διάφορες χρονικές κλίμακες. Το στοχαστικό μέρος της διαδικασίας δεν είναι εντελώς τυχαίο, έχει στοχαστική δομή ή στοχαστική μνήμη.

Στη συνεχεια αναπτύσσεται η έννοια των υδροσυστημάτων καθώς και θέματα διαχείρησης τους και ποσοτικοποίησης της αξιοπιστίας.

Ένα υδροσύστημα είναι ένα σύστημα που αποτελείται από φυσικά υδάτινα σώματα και τεχνικά έργα που συνεργάζονται για την εξυπηρέτηση ενός ή περισσοτέρων σκοπών, τα οποία αναφέρονται τόσο στην εκμετάλλευση του νερού ως φυσικού πόρου όσο και στην προστασία από την καταστροφική του δράση ως φυσικό κίνδυνο (Κουτσογιάννης και Ξανθόπουλος, 2014).

Η διακινδύνευση ορίζεται ως η πιθανότητα αδυναμίας επίτευξης του στόχου. Η αξιοπιστία ορίζεται μαθηματικά ως συμπλήρωμα της διακινδύνευσης. Η μαθηματική ανάλυση της διακινδύνευσης και της αξιοπιστίας ορίζεται ως ανάλυση αξιοπιστίας.

Εάν το X αντιπροσωπεύει τη μέγιστη τιμή της φυσικής διεργασίας σε ετήσια βάση (π.χ. μέγιστη ετήσια πλημμύρα) και n είναι η διάρκεια ζωής του έργου, τότε το γεγονός $\{L \leq C\}$ ισούται με τις διαδοχικές εμφανίσεις του συμβάντος $\{X \leq C\}$. Προκειμένου να μην υπερβεί η τιμή c καθ 'όλη τη διάρκεια του έργου, δεν θα πρέπει να υπάρχει υπέρβαση σε όλα τα έτη της διάρκειας αυτής. Θεωρώντας ότι οι πλημμύρες διαδοχικών ετών είναι στοχαστικά ανεξάρτητες, η διακινδύνευση δίνεται από (Κουτσογιάννης, 1997):

$$R = 1 - [P(X \leq C)]^n = 1 - [F_X(C)]^n \qquad \text{(ΕΠ.1)}$$

Η στοχαστική συμπεριφορά ενός μηχανικού συστήματος είναι απαραίτητη εάν αποδεχτούμε ότι τα θέματα αβεβαιότητας είναι σημαντικά και πρέπει να τα διαχειριστούμε. Η πραγματική κατανομή της απόκρισης του συστήματος που υπόκειται στην αβεβαιότητα των παραμέτρων είναι πολλές φορές δύσκολη ή και αδύνατη να υπολογιστεί. Αυτό οφείλεται στην πολυπλοκότητα των υδροσυστημάτων. Σε τέτοιες περιπτώσεις, η προσομοίωση Monte Carlo είναι ένα χρήσιμο εργαλείο για την παροχή αριθμητικών εκτιμήσεων των στοχαστικών χαρακτηριστικών της απόκρισης του συστήματος (Tung and Yen, 2005).

Στοχαστικά εργαλεία και μακροπρόθεσμη χρονική εξάρτηση

Εδώ εισάγονται κάποια εργαλεία που βοηθούν στη διερεύνηση της στοχαστικής δομής πραγματικών η συνθετικών χρονοσειρών.

Το κλιμακόγραμμα ορίζεται ως η διασπορά του συναθροισμένου μέσου της διεργασίας $\underline{x}$(t) σε κλιμακα συνάθροισης k και συμβολίζεται $\gamma(k)$. Ορίζουμε τη διεργασία $\underline{x}^{(k)}$ σε κάθε κλίμακα $k \geq 1$ ως:

$$\underline{x_i}^{(k)} := \frac{1}{k} \sum_{l=(i-1)k+1}^{ik} \underline{x_j} \qquad \text{(ΕΠ.2)}$$

Το φάσμα ισχύος $s(\omega)$ της στοχαστικής ανέλιξης σε διακριτο χρόνο $t = 0,1, \ldots,$ με συνάρτηση αυτοσυνδιασποράς $\gamma_m = Cov[x_t, x_{t+m}], m = 0, \pm 1, \ldots,$ είναι ο αντιστροφος μετασχηματισμός Fourier της συνάρτησης αυτοσυνδιασποράς με ω στο διάστημα $[0,1/2]$. Η επόμενη σχέση ισχύει (Κουτσογιάννης, 2013):

$$s(\omega) = 2\gamma_0 + 4 \sum_{m=1}^{\infty} \gamma_m cos(2\pi m\omega) \qquad \text{(ΕΠ.3)}$$

Το κλιμακοφάσμα είναι ένα νεοεισαχθέν στοχαστικό εργαλείο. Ορίζεται απο τον Κουτσογιάννη (2017):

$$\zeta(k) := \frac{k(\gamma(k) - \gamma(2k))}{ln2} \qquad \text{(ΕΠ.4)}$$

Ένας άλλος σημαντικός παράγοντας που πρέπει να λαμβάνεται υπόψη κατά την προσπάθεια κατανόησης των υδρολογικών διεργασιών είναι η μακροχρόνια χρονική εξάρτηση ή η εμμονή. Στην υδρολογία ο όρος αυτός είναι ισοδύναμος με τον όρο "φαινόμενο Hurst". Η μακροπρόθεσμη χρονική εξάρτηση ορίζεται από τους Everitt και Skrondal (2010) ως: "Μικρές αλλά αργά φθίνουσες συσχετίσεις σε μια στοχαστική ανελιξη. Τέτοιες συσχετίσεις συχνά δεν ανιχνεύονται με τυποποιημένα στατιστικά τεστ, αλλά το αποτέλεσμά τους μπορεί να είναι αρκετά σημαντικό. " Στην υδρολογία αυτό παρατηρήθηκε αρχικά από τον Hurst. Είναι ουσιαστικά η τάση των υγρών ετών να συσσωρεύονται και να σχηματίζουν μεγαλύτερες υγρές περιόδους και τα ξηρά χρόνια σχηματίζοντας παρομοίως περιόδους ξηρασίας.

Το μοντέλο Filtered Hurst Kolmogorov (Κουτσογιάννης, 2015) είναι ένα μοντέλο που μπορεί

να προσομοιώσει φυσικά φαινόμενα με μακροχρόνια εμμονή.

Το κλιμακόγραμμα της ανέλιξης δίνεται απο τον παρακάτω τύπο:

$$\gamma(\Delta) = \lambda\left(1 + \left(\frac{\Delta}{\alpha}\right)^{2M}\right)^{\frac{H-1}{M}} \quad \text{(ΕΠ.5)}$$

Το α και λ είναι παράμετροι κλίμακας με μονάδες $[t]$ and $[x]^2$, αντιστοίχως, H είναι η παράμετρος Hurst, στο διάστημα (0,1), και M μιά δεύτερη παράμετρος στο διάστημα (0,1). Το H καθορίζει τις καθολικές ιδιότητες τις ανέλιξης καθώς ($t \to \infty$) και το M καθορίζει τις τοπικές ιδιότητες καθώς ($t \to 0$).

Βέλος του χρόνου και στοχαστικές ανελίξεις

Ο όρος "βέλος του χρόνου" αναπτύχθηκε αρχικά από τον Eddington (1928) για να περιγράψει την κατεύθυνση χρόνου, η οποία μπορεί να προσδιοριστεί με τη μελέτη της οργάνωσης ατόμων, μορίων και σωμάτων. Η διαισθητική αντίληψή μας για τον χρόνο οτι κυλάει μόνο προς τα εμπρός χρόνο μπορεί να απορριφθεί ως απλώς υποκειμενική.

Η κατεύθυνση του χρόνου μπορεί να οριστεί από μια κατηγορία διαδικασιών που καταστρέφουν πληροφορίες και δημιουργούν χάος. Οι μη αναστρέψιμες διεργασίες που καταστρέφουν την μακροσκοπική πληροφορία είναι εκδηλώσεις του δεύτερου νόμου της θερμοδυναμικής. Από την άλλη πλευρά, υπάρχουν πολλές διεργασίες που είναι μη αναστρέψιμες και είναι αντιδιαμετρικά αντίθετες. Όλες αυτές οι διεργασίες έχουν κάτι κοινό παράγουν τάξη ή πληροφορία. Εκτρέπουν ένα σύστημα από μία απλή κατάσταση σε πιο σύνθετη.

Ο Weiss (1975) ορίζει μία στοχαστική ανέλιξη $\underline{x}(t)$, σε συνεχή χρόνο t, με nιοστής τάξης συνάρτηση κατανομής.

$$F(x_1, x_1, \dots, x_n; t_1, t_2, \dots, t_n) \coloneqq P\{\underline{x}(t_1) \leq x_1, \underline{x}(t_2) \leq x_2, \dots \underline{x}(t_n) \leq x_n\} \quad \text{(ΕΠ.6)}$$

ως συμμετρική στο χρόνο αν η κοινή κατανομή δεν αλλάζει μετά απο αντιστροφή του χρόνου γύρω απο την αρχή των αξόνων, δηλαδή αν για κάθε $n, t_1;\ t_2;\ \dots;\ t_{n-1};\ t_n$,

$$F(x_1, x_1, \dots, x_n; t_1, t_2, \dots, t_n) = F(x_1, x_1, \dots, x_n; -t_1, -t_2, \dots, -t_n) \quad \text{(ΕΠ.7)}$$

Η πρόσφατη μελέτη από τον Κουτσογιάννη (2019) παρέχει μία μεθοδολογία αναπαραγωγής της αντιστρεψιμότητας σε συνθετικές χρονοσειρές. Ο δείκτης αντιστρεψιμότητας ορίζεται ως ο λόγος της ασυμμετρίας της διαφορικής διεργασίας προς την ασυμμετρία της αρχικής διεργασίας.

Η μελέτη της μη αναστρεψιμότητας απο τον Κουτσογιάννη (2019) δηλώνει ότι η ασυμμετρία απαιτεί τη μελέτη της τρίτης ροπής και του συντελεστή ασυμμετρίας της διεργασίας, της αρχικής αλλά και της διαφοροποιημένης. Η πρώτη ροπής (μέσος όρος) της διαφορικής διεργασίας είναι πάντα μηδενική ενώ η δεύτερη (διακύμανση) είναι πάντα θετική και έτσι καταλήγει στο συμπέρασμα ότι δεν παρέχουν ενδείξεις σχετικά με την ασυμμετρία του

χρόνου. Ως εκ τούτου, η ροπή ελάχιστης τάξης που μπορεί να χρησιμοποιηθεί για την ανίχνευση της αντιστρεψιμότητας είναι η τρίτη.

Η ασυμμετρία στις στοχαστικές ανελίξεις είναι συνώνυμη με απότομους ανοδικούς κλάδους και ομαλότερους καθοδικούς κλάδους στις δειγματοσυναρτήσεις. Το ίδιο συμβαίνει και σε ένα υδρογράφημα όταν εξετάζουμε παροχές μικρής χρονικής κλίμακας. Αυτή η συμπεριφορά γίνεται προσπάθεια να αναπαραχθεί με την έννοια της χρονικής ασυμμετρίας.

Το μοντέλο (MA) (Κουτσογιάννης, 2000),

$$\underline{x}_\tau = \sum_{i=-\infty}^{\infty} a_i \underline{v}_{\tau-i} \tag{ΕΠ.8}$$

έχει την παρακάτω λύση

$$a_\eta = \int_{-1/2}^{1/2} e^{2\pi i(\theta(\omega)-\eta\omega)} A^R(\omega) d\omega \tag{ΕΠ.9}$$

Όπου i είναι η φανταστική μονάδα, $\theta(\omega)$ είναι οποιαδήποτε περιττή πραγματική συνάρτηση (σημαίνει πως $\theta(-\omega) = -\theta(\omega)$) και

$$A^R(\omega) \coloneqq \sqrt{2 s_d(\omega)} \tag{ΕΠ.10}$$

Τροποποίηση αλγορίθμου για περαιτέρω διατήρηση χρονικής ασυμμετρίας

Στην παρούσα μελέτη γίνεται προσπάθεια τροποποίησης του αλγορίθμου που πρότεινε ο Κουτσογιάννης (2019). Ο στόχος είναι να προσομοιωθούν οι χρονοσειρές που διατηρούν τη χρονική ασυμμετρία σε μεγαλύτερες κλίμακες συνάθροισης της διεργασίας.

Για μήκος προσομοίωσης i, το AMA μοντέλο θα μπορούσε να γραφτεί επίσης ως:

$$X_i = \sum_{j=1}^{2q+1} a_{2q+2-j} V_{i+j-1} = a_{2q+1} V_i + \cdots + a_1 V_{i+2q+1} \tag{ΕΠ.11}$$

Η δεύτερη ροπή της αρχικής ακολουθίας σε δεύτερη κλίμακα υπολογίζεται ως:

$$M_{orig.}^{(2)\ (\kappa=2)} = \sum_{j=1}^{2q+1} \frac{\left(a_{2q+1-j} + a_{2q+2-j}\right)^2}{4} + \frac{{a_{2q+1}}^2}{4} \tag{ΕΠ.12}$$

Η δεύτερη ροπή της διαφορικής ακολουθίας σε δεύτερη κλίμακα υπολογίζεται ως:

$$M_{differ.}^{(2)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1}\left[+\frac{\left(a_{2q-1-j}+a_{2q-j}-a_{2q+1-j}-a_{2q+2-j}\right)^2}{4}\right] + \frac{\left(a_{2q}+a_{2q-1}-a_{2q+1}\right)^2}{4} + \frac{\left(a_{2q}+a_{2q+1}\right)^2}{4} + \frac{a_{2q+1}{}^2}{4} \quad \text{(ΕΠ.13)}$$

Η τρίτη ροπή της αρχικής ακολουθίας σε δεύτερη κλίμακα υπολογίζεται ως:

$$M_{orig.}^{(3)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1}\frac{\left(a_{2q+1-j}+a_{2q+2-j}\right)^3}{8} + \frac{a_{2q+1}{}^3}{8} \quad \text{(ΕΠ.14)}$$

Η τρίτη ροπή της διαφορικής ακολουθίας σε δεύτερη κλίμακα υπολογίζεται ως:

$$M_{differ.}^{(3)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1}\left[-\frac{\left(a_{2q-1-j}+a_{2q-j}-a_{2q+1-j}-a_{2q+2-j}\right)^3}{8}\right] - \frac{\left(a_{2q}+a_{2q-1}-a_{2q+1}\right)^3}{8} - \frac{\left(a_{2q}+a_{2q+1}\right)^3}{8} - \frac{a_{2q+1}{}^3}{8} \quad \text{(ΕΠ.15)}$$

Μετά τον υπολογισμό των ροπών δειγματος, τα υπολογιστικά εργαλεία πρέπει να εξισώσουν τις ροπες του δείγματος (εμπειρικές) και τις ακολουθίας (θεωρητικές) έτσι ώστε να βρεθούν οι παράμετροι του μοντέλου. Τα εργαλεία βελτιστοποίησης χρησιμοποιούνται για την εύρεση των απαραίτητων παραμέτρων. Η παραμετροποίηση ακολουθεί την ίδια μεθοδολογία όπως στον Κουτσογιάννη (2019): ένας ορισμός του $\theta(\omega)$ ως το ομαλό ελάχιστο των δύο υπερβολικών συναρτήσεων της συχνότητας, δηλαδή :

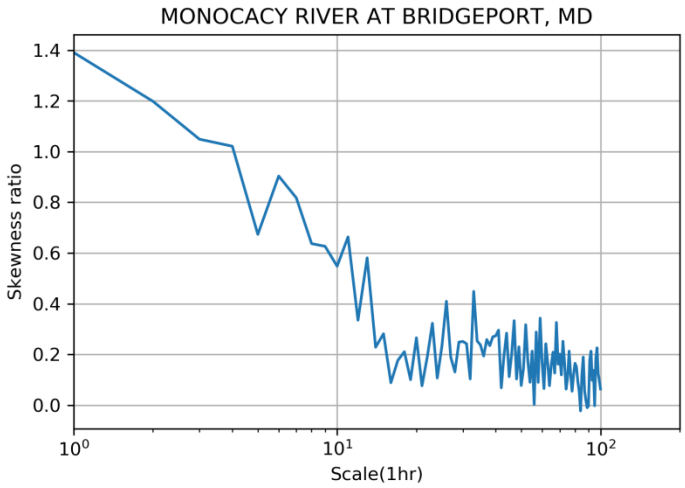
$$\theta(\omega) = \frac{1}{\zeta}\ln\left(e^{\zeta\theta_1(\omega)} + e^{\zeta\theta_2\left(1/2-\omega\right)}\right), \qquad \theta_i(\omega) \coloneqq \frac{C_{1,i}\omega}{C_{2,i}+\omega} + C_{0,i} \quad \text{(ΕΠ.16)}$$

Αποτελέσματα για τη μελέτη της περίπτωσης του ποταμού Monacacy

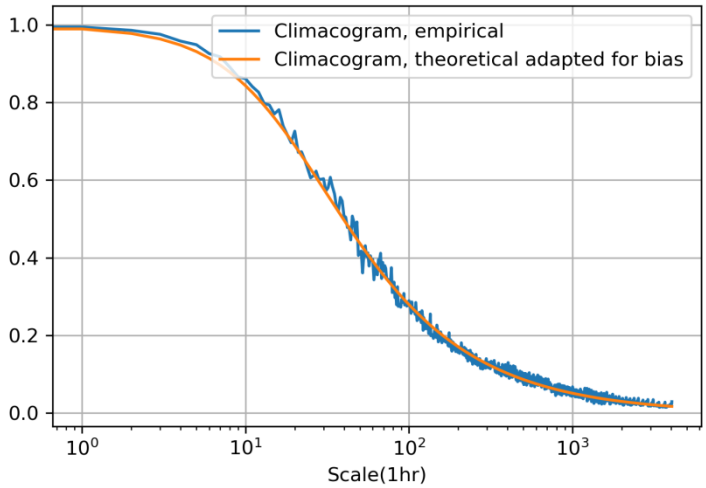
Σε αυτό το κεφάλαιο εφαρμόζεται ένας τροποποιημένος αλγορίθμος που περιγράφεται στο προηγούμενο κεφάλαιο με πραγματικά δεδομένα. Αναλύεται η περίπτωση του ποταμού Monocacy απο τη βάση δεδομένων USGS. Στην πρώτη περίπτωση γίνονται προσομοιώσεις με διατήρηση της χρονικής ασυμμετρίας στην πρώτη κλίμακα (Σχήμα 8). Στη δεύτερη περίπτωση γίνονται προσομοιώσεις με διατήρηση στις δύο πρώτες κλίμακες (Σχήμα 9).

Τα αρχικά δεδομένα προέρχονταν από 15λεπτες μετρήσεις αλλά συναθροίστηκαν σε ωριαία κλίμακα. Μετά από αυτό, θεωρήθηκε σημαντικό να πραγματοποιηθεί η στασιμοποίηση της χονοσειράς και έγινε το τεστ αντιστρεψιμότητας (Σχήμα 1). Για την προσαρμογή του μοντέλου Filtered Hurst Kolmogorov ταυτόχρονα χρησιμοποιήθηκαν και το κλιμακόφασμα (Σχήμα 3) με έμφαση μικρές κλίμακες και στο κλιμακόγραμμα (Σχήμα 2) στις μεγαλύτερες κλίμακες. Για το σκοπό αυτον έγινε πρώτα μονιμοποίηση. Οι παράμετροι υπολογίστηκαν: $a =$

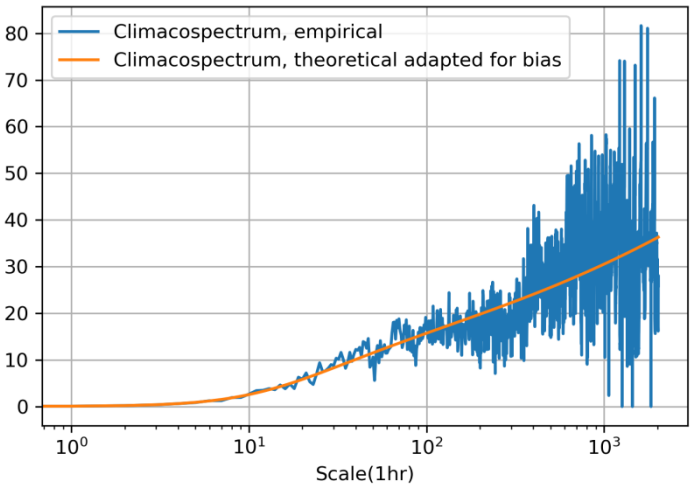
19.399, $H = 0.628$, $M = 0.724$. Στην συνέχεια μέσω υπολογισμών που περιέχουν το φάσμα ισχύος (Σχήμα 4) υπολογίζονται οι συντελεστές α_η του AMA για την πρώτη και δεύτερη περίπτωση (Σχήμα 5 και 7). Για τη δεύτερη περίπτωση υπολογίζονται επίσης και οι συντελεστές $\theta(\omega)$ (Σχήμα 6).



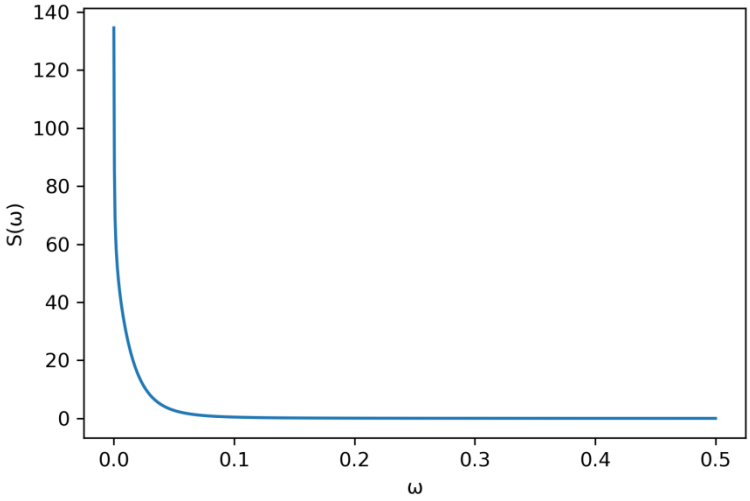
Σχήμα 1 Τεστ αντιστρεψιμότητας.



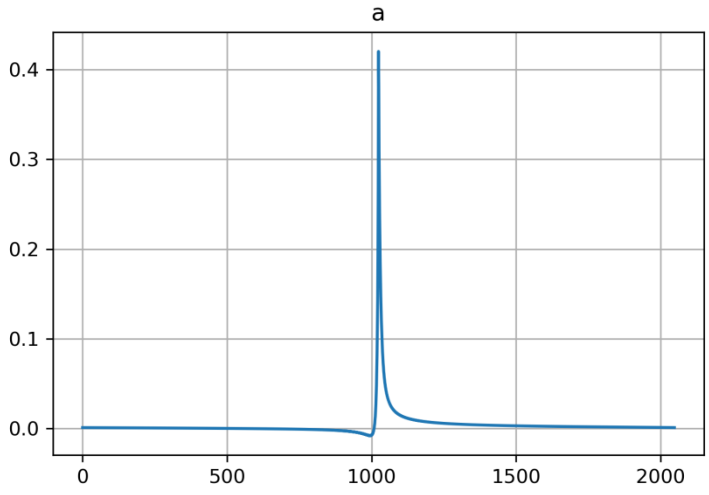
Σχήμα 2 Προσαρμογή δεδομένων με κλιμακόγραμμα .



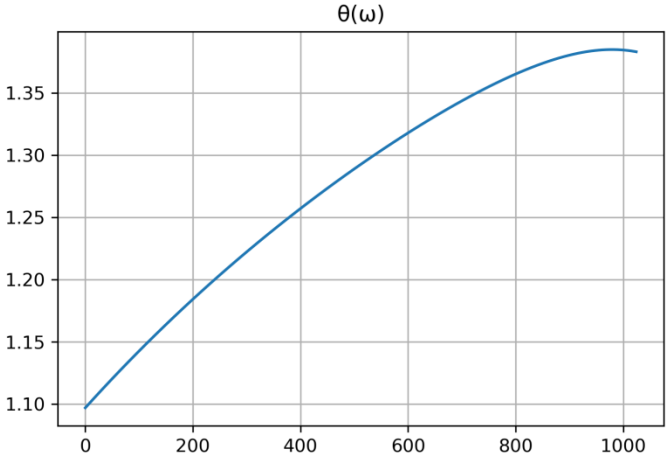
Σχήμα 3 Προσαρμογή δεδομένων με κλιμακόφασμα.



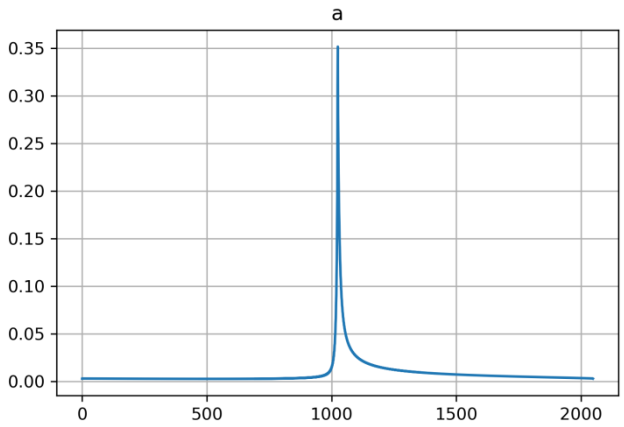
Σχήμα 4 Διακριτό φάσμα ισχύος.



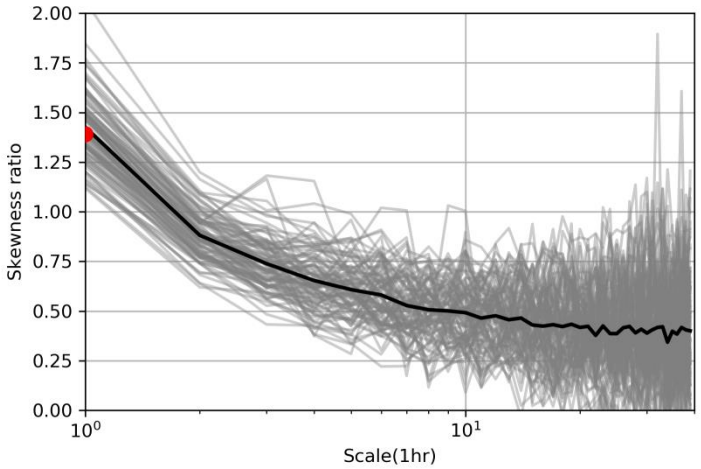
Σχήμα 5 a_η ακολουθία απο την πρώτη περίπτωση με σταθερό θ.



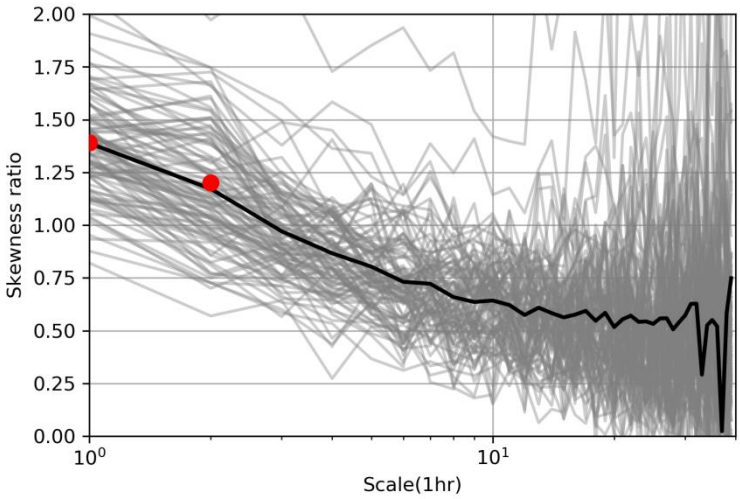
Σχήμα 6 $\boldsymbol{\theta}$ ακολουθία για τη δεύτερη περίπωση διατήρησης της ασυμμετρίας και στις δύο κλίμακες



Σχήμα 7 $\boldsymbol{a_\eta}$ ακολουθία απο την δεύτερη περίπτωση με μεταβλητό $\boldsymbol{\theta}$.



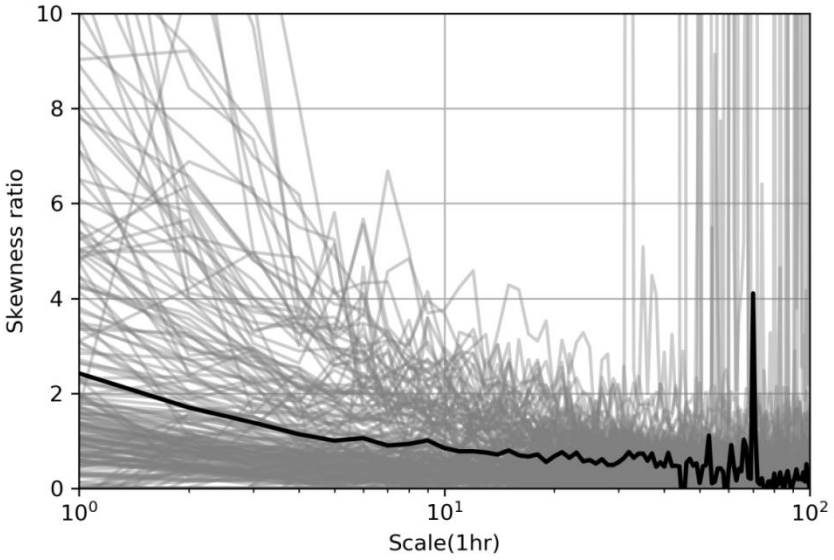
Σχήμα 8 100 προσομοιώσεις με 10000 μήκος, διατηρώντας τη χρονική αντιστρεψιμότητα μόνο στη πρώτη κλίμακα.



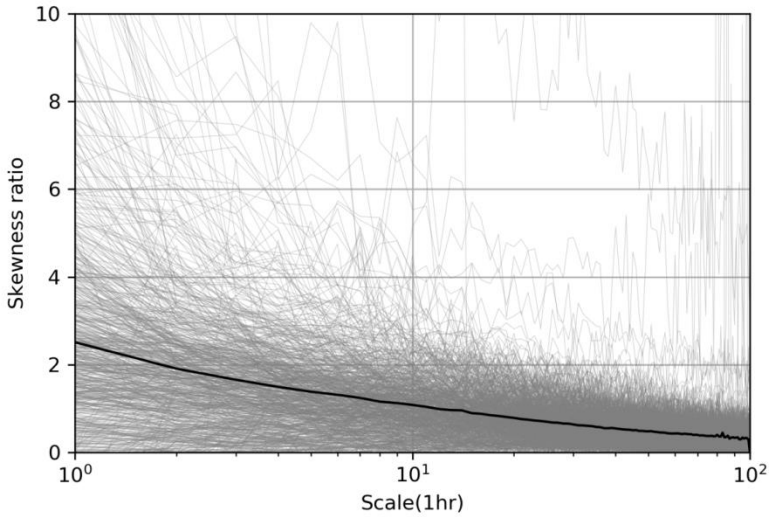
Σχήμα 9 100 προσομοιώσεις με 10000 μήκος, διατηρώντας τη χρονική αντιστρεψιμότητα ταυτόχρονα και στις δύο κλίμακες (πρώτη και δεύτερη)

Αποτελέσματα απο τη διερεύνυση της αντιστρεψιμότητας στη βάση δεδομένων του USGS

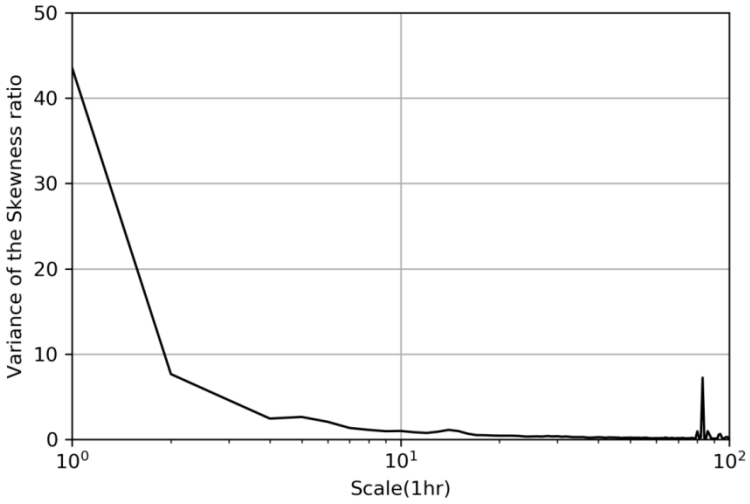
Σε αυτή την ενότητα επιχειρείται η ποσοτικοποίηση της μη αναστρέψιμότητας στις πρώτες 100 κλίμακες από μιά μεγάλη βάση δεδομένων απορροής. Ο στόχος είναι να μελετηθούν πολλοί σταθμοί και να βρεθεί η μέση τιμή του δείκτη ασυμμετρίας για κάθε κλίμακα. Η πρώτη περίπτωση είναι η πολιτεία του Maryland που αποτελείται από 222 σταθμούς (Σχήμα 10). Η δεύτερη περίπτωση είναι ένα ακόμη μεγαλύτερο σύνολο δεδομένων που αποτελείται από 762 σταθμούς στις ΗΠΑ (Σχήμα 11). Για τη συλλογή και επεξεργασία των πρωτογενών δεδομένων χρησιμοποιήθηκαν κωδικές python και συγκεκριμένα το πακέτο climata (www.pypi.org/project/climata/) και ενω ήταν σε 15λεπτη κλίμακα μετατράπηκαν σε ωριαία δεδομένα. Όλες οι χρονοσειρές έχουν στασιμοποιηθεί με τον ίδιο τρόπο όπως στη μελέτη περίπτωσης. Στο τέλος υπολογίστηκε η διασπορά σε κάθε κλίμακα για τη δεύτερη περίπτωση (Σχήμα 12).



Σχήμα 10 Αντιστρεψιμότητα για ολο το Μέριλαντ, Η.Π.Α., 222 σταθμοί



Σχήμα 11 Αντιστρεψιμότητα για 762 σταθμούς στις Η.Π.Α



Σχήμα 12 Διασπορά του κριτηρίου αντιστρεψιμότητας για τη δεύτερη περίπτωση.

Συμπεράσματα

Η αβεβαιότητα είναι ένας σημαντικός παράγοντας στις φυσικές επιστήμες και στην επιστήμη του μηχανικού. Η πιθανοτική συμπεριφορά ενός μηχανικού συστήματος είναι απαραίτητο να μελετάται, δεδομένου ότι τα ζητήματα αβεβαιότητας είναι σημαντικά και πρέπει να αντιμετωπιστούν. Η στοχαστική προσομοίωση είναι ένα χρήσιμο εργαλείο για την παροχή αριθμητικών εκτιμήσεων των στοχαστικών χαρακτηριστικών της απόκρισης του συστήματος.

Το βέλος του χρόνου έχει σημαντικό ρόλο στην επιστήμη και σχετίζεται στενά με την τυχαιότητα και την αβεβαιότητα. Η χρονική ασυμμετρία της απορροής σημειώνεται για κλίμακες αρκετών ημερών και αυτό υπογραμμίζει την ανάγκη αναπαραγωγής σε προσομοιώσεις πλημμύρας.

Η ασυμμετρία στις στοχαστικές ανελίξεις είναι συνώνυμη με απότομους ανοδικούς κλάδους και ομαλότερους καθοδικούς κλάδους στις δειγματοσυναρτήσεις. Το ίδιο συμβαίνει και σε ένα υδρογράφημα όταν εξετάζουμε παροχές μικρής χρονικής κλίμακας. Αυτή η συμπεριφορά γίνεται προσπάθεια να αναπαραχθεί με την έννοια της χρονικής ασυμμετρίας.

Πραγματικές χρονοσειρές από μεγάλη βάση δεδομένων χρησιμοποιούνται για τη διερεύνηση της μη αντιστρεψιμότητας σε ωριαία κλίμακα. Χρησιμοποιούνται χρονοσειρές απορροής μέχρι και την εκατοστή συναθροισμένη κλίμακα. Η χρονική ασυμμετρία της απορροής στις ΗΠΑ τουλάχιστον, έχει αναμενόμενη τιμή γιά το κριτήριο αντιστρεψιμότητας στην πρώτη κλίμακα γύρω στο 2,5 και στη δεύτερη κλίμακα γύρω στο 1,9. Ωστόσο, αυτό το αποτέλεσμα έχει μια πολύ μεγάλη διακύμανση στην πρώτη κλίμακα, η οποία τείνει να μειώνεται όσο αυξάνονται οι κλίμακες.

Η μελέτη αυτή προτείνει μια τροποποίηση της υπάρχουσας μεθόδου από τον Κουτσογιάννη (2019) που διατηρεί την μη αντιστρεψιμότητα μόνο στην πρώτη κλίμακα και την καθιστά ικανή να διατηρεί την μη αντιστρεψιμότητα ταυτόχρονα στην πρώτη και στη δεύτερη κλίμακα. Για να ελέγξουμε τη μέθοδο, χρησιμοποιούμε πραγματικά δεδομένα. Τα αποτελέσματα επαληθεύουν τη μέθοδο με επιτυχία

TABLE OF CONTENTS

TABLE OF FIGURES

1. Introduction

1.1 Study scope

The aim of this master thesis is to study the influence of time's arrow on small scale stochastic models. This work is inspired and heavily influenced by the recent study by Koutsoyiannis (2019): "Time's arrow in stochastic characterization and simulation of atmospheric and hydrological processes" that is part of a broader stochastics framework. At the beginning of the thesis there is an attempt to cover part of the framework that is the basis for the recent study. At the same time other relevant literature is discussed. The reason behind that is to highlight the importance and the usefulness of the results.

Later, real world data are being used to investigate the irreversibility of hourly scale stremaflow time series at scales up to one hundred. The aim is to find out the importance of the irreversibility in small scale streamflow data and at which degree it should affect its modeling.

The last aim of this study is to modify the existing method by Koutsoyiannis (2019) that conserves irreversibility at the first scale only and make it capable of preserving the irreversibility simultaneously at the first and second scale. For example if there is hourly streamflow data, the irreversibility quantification method that is used should give the desired result after aggregation at the 2-hour scale. In the end the basic method and the modified are to be verified by real world data.

1.2 Work structure

In the first chapter an introduction to the thesis is made. The aims of the thesis are being presented. In the second chapter, the sometimes controversial concept of uncertainty is being discussed because it is the reason behind stochastic simulation. Also relevant terms such as uncertainty analysis are being defined and presented. In the third chapter there is an introduction to probability and statistics as it is the basis behind stochastics. Some basic principles and axioms are given in order set the foundation behind the concept of a stochastic process. Later, at chapter four the subject of reliability in Hydrosystems is being discussed and relevant literature is provided. Proper definitions are being given for all terms and the important subject of Monte Carlo simulation is being theoretically backed up. At chapter five some stochastic tools are being presented with their advantages/disadvantages and the methods of using them. They are all relevant to the later study and most of them are used. The chapter six presents information on the subject of long range dependence in hydrology. It essentially provides a basis for the modeling framework and in the end describes the model used in the study: the Filtered Hurst Kolmogorov process. In chapter six the concept of time's arrow is being discussed and the final method by Koutsoyiannis (2019) is presented. In the same chapter the model is modified. In the eighth chapter the case study is being shown as an application of the method modification. In the ninth chapter the streamflow reversibility at small time scales is being investigated through a large database. In the end at chapter ten the

conclusions are presented.

2. Dealing with Uncertainty in Hydrosystems

2.1 Introduction

In the frame of hydrological modeling, uncertainty is a huge factor. Around this term there is a lot of misconceptions. A lot of times it is not taken into account and the result is costly.

Uncertainty can be defined as the manifestation of events that are beyond one's control (Mays and Tung 1992). Natural phenomena are commonly separated into two divisions regarding uncertainty and randomness, random (or stochastic) and deterministic. Koutsoyiannis (2010) argues that this view should be reconsidered.

Deterministic laws and randomness coexist and should be modeled and represented in a holistic way. Whether a process is more random or deterministic is only a matter of time horizon. Uncertainty does not only exist in nature (Figure 2.1) but also in deterministic models. One can easily observe this by using a deterministic non-linear model that is complex enough and realize that a small change in the initial values can cause high uncertainty at the end of the simulation. This can be seen at Figure 2.2.

Papoulis (1991) writing about causality and randomness states: "We conclude with a brief comment on the apparent controversy between causality and randomness. There is no conflict between causality and randomness or between determinism and probability if we agree, as we must, that scientific theories are not discoveries of the laws of nature but rather inventions of the human mind. Their consequences are presented in deterministic form if we examine the results of a single trial; they are presented as probabilistic statements if we are interested in averages of many trials. In both cases, all statements are qualified. In the first case, uncertainties are of the form "with certain errors and in certain ranges of the relevant parameters": in the second "with a high degree of uncertainty if the number of trials is large enough"."

The causes of uncertainty can be due to either the fundamental stochastic nature of the process or to limited knowledge or resources to model it perfectly.

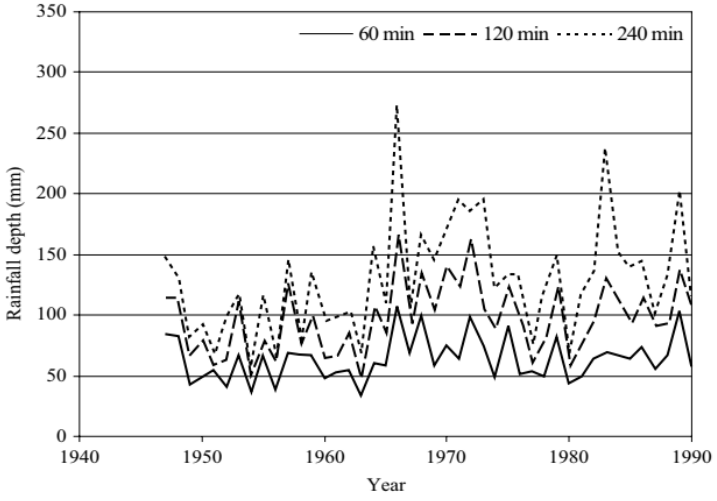


Figure 2.1 Annual maximum rainfall series of different durations (1947–1990) at Hong Kong Observatory, Hong Kong (Tung and Yen, 2005)

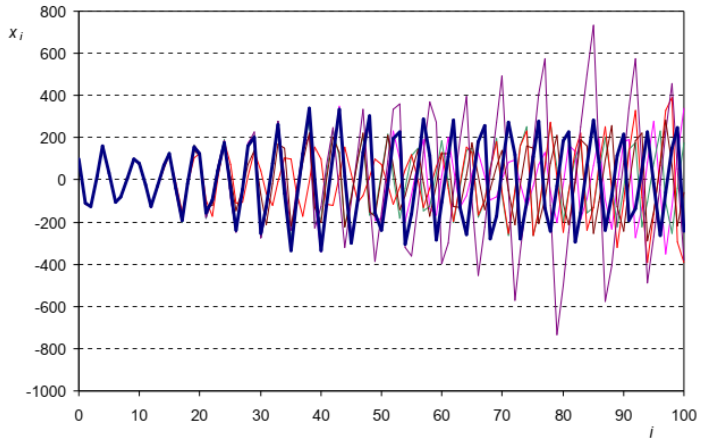


Figure 2.2 A simulation with 1% uncertainty in initial conditions Koutsoyiannis (2010)

2.2 Sources of uncertainty

The source of uncertainty in hydrosystems can be (Makropoulos and Efstratiadis, 2018):

1. Simplistic model assumptions for critical system processes (structural model errors - structural uncertainty). It happens when we take into consideration fewer rules than there actually are.
2. Sensitivity to initial and boundary conditions (chaotic systems). As in the previous figure.
3. Inadequate knowledge of critical system parameters.
4. Stochastic nature and spatio-temporal variability of hydrometeorology processes (e.g. rain, evaporation, runoff, wind).
5. Measurement errors and inaccuracies.
6. System change over time (due to external factors).
7. Changes in decisions / policies and hence in performance measures.

2.3 Purpose of uncertainty analysis

The main purpose of uncertainty analysis is to quantify the uncertainty by estimating statistical properties of the system outputs that are affected by the stochastic nature of the natural process or sensitivity in the initial conditions. Design quantities and system outputs are dependent on several system parameters that cannot always be accurately evaluated.

The task of uncertainty analysis is to determine the uncertainty features of the system output as a function of uncertainties associated with the system model itself and its stochastic parameters. Uncertainty analysis provides a formal and methodical structure to measure the uncertainty of the system. In addition, it offers information into the contribution of each stochastic basic parameter to the overall uncertainty of system outputs. This information is

essential and can lead to the identification of the parameters that play a more important role in the uncertainty. Their assessment will result to the reduction of the overall uncertainty of the system output (Tung and Yen, 2005).

2.4 Measures of uncertainty

There are some expressions used to show the measure of uncertainty. In general, the uncertainty linked with a parameter, a function, a model, or a system, results from the combined effects of the uncertainties of these contributing parameters.

The statistical moments, associated with a quantity subject to uncertainty, are a simple way to assess uncertainty. The second-order moment called *variance* is a measure of the dispersion of a random variable and can be used. At the instance of comparing or combining uncertainties of different variables, the coefficient of variation can be used. It is the ratio of standard deviation to the mean, offers a normalized measure of uncertainty.

The most complete and ideal description of the uncertainty features of a quantity can be given by the probability density function (PDF). However it is more difficult to find (Tung and Yen, 2005).

Informational entropy can be a measure of uncertainty as discussed in the chapter 6.

3. Basic Theory: Probability and Statistics

3.1 Fundamentals of probability

In 1933 Kolmogorov published the axiomatic basis of modern probability theory. The modern approach to probability theory is based on set theory. It is built on three fundamental concepts and three axioms. The fundamental concepts are the following (Koutsoyiannis, 1997):

- The sample space is defined as the set Ω, the elements of which correspond to the possible outcomes of an experiment.

For example for the throw of a dice the sample space is: $\Omega = \{1,2,3,4,5,6\}$

- The subsets of F subsets are called events: We say that event A happens when the outcome ω of the experiment is an element of A.
- The probability measure is a function P on F. In each event A we assign a number $P(A)$ that says the probability of event A.

The three elements (Ω, F, P) define what is called a probability space. The function P must satisfy the following axioms of probability theory:

1. $P(A) \geq 0$ (3.1)
2. $P(\Omega) = 1$ (3.2)
3. $P(A \cup B) = P(A) + P(B)$, if also $A \cap B = \emptyset$ (3.3)
4. $P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$, if also $A_i \cap A_j = \emptyset, i \neq j$ (3.4)

3.2 Random variable and probability distributions

Each event ω is associated with a number $X(\omega)$ according to some predefined rule through a function X. This function is defined on a sample space Ω and is called "random variable". The outcome ω may be a number and the predefined rule a mathematical function (Koutsoyiannis, 1997). Usually we omit the element ω and simply write X. For the random variable itself we use capital letters while for the value of the random variable we use small letters. For instance we write $\{X \leq x\}$ meaning to show the event that is composed of all events ω such that the values $X(\omega)$ are less than or equal to the number x. The probability of this event is expressed as $P(\{X(\omega) \leq x\})$ or for simplicity $P(X \leq x)$ (Koutsoyiannis, 1997).

3.3 Distribution function

The distribution function $F(x)$ is a function of x defined by the following equation (Koutsoyiannis, 1997):

$$F_X(x) = P(X \leq x), x \in R, F_X \in [0,1] \tag{3.5}$$

It must be stated that F_X is not a function of the random variable X but just connected to it. It is obviously a function of x. Also the domain of F is not identical to the range of X (ω) but is always $(-\infty, +\infty)$. F follows the inequality:

$$0 = F_X(-\infty) \leq F_X(x) \leq F_X(+\infty) = 1 \tag{3.6}$$

F_X is also called cumulative distribution function or non-exceedance probability. If F_X (x) is continuous for all x, then the random variable X (ω) is also continuous. In this case the sample space Ω is an infinite and uncountable set. On the other hand if F_X (x) is a step function, then the random variable X(ω) is called discrete. In this case the sample space Ω is a finite set or an infinite and countable set. It is important to note however that even for discrete random variables, the cumulative distribution function is always defined for all x ∈ R. For continuous random variables, the derivative of the cumulative distribution function is called probability density function:

$$f_x(x) := \frac{dF(x)}{dx} \tag{3.7}$$

The distribution function can be calculated through the inverse of the above equation:

$$F_X(x) = \int_{-\infty}^{x} f_x(\xi) d\xi \tag{3.8}$$

3.4 Non-disjoint events

For two non-disjoint events A and B it is shown that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3.9}$$

3.5 Conditional probabilities

The conditional probability of an event is the probability of its occurrence given that another event has occurred. The conditional probability is denoted and defined as

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{3.10}$$

3.6 Independent events

According to Ross (2004) two events are said to be independent if the knowledge that one

event has occurred has no effect on the probability of occurrence of the other. In mathematical terms:

$$P(A\,B) \;=\; P(A)\,P(B) \tag{3.11}$$

This also implies:

$$P(A|B) \;=\; P(A) \tag{3.12}$$

Two events *A* and *B* that are not independent are said to be *dependent*.

3.7 Expected value

The expected value of a random variable X, generally denoted as $E(X)$. If the variable is discrete with probability distribution, $P(X = x)$, then $E(X) = \sum_x xP(X = x)$. If the variable is continuous the summation is replaced by an integral. The expected value of a function of a random variable, $f(x)$, is defined :

$$E\big(f(x)\big) = \int_u f(u)g(u)du \tag{3.13}$$

where $g(x)$ is the probability distribution of x

3.8 Stochastic processes

The definition of a stochastic process $\{X(t), t \in T\}$ is that it is "a family of random variables" (Koutsoyiannis, 1997). That is, for each $t \in T, X(t)$ is a random variable. We refer to $X(t)$ as the state of the process at time t if the index t is represents time. Stochastic processes can be used to describe the temporal evolution or the spatial relations of random variables. The set T is called the index set of the process. The stochastic process is said to be a discrete-time process. If T is an interval of the real line, the stochastic process is a continuous-time process. For example, $\{Xn, n = 0, 1, \dots\}$ is a discrete-time stochastic process indexed by the nonnegative integers; while $\{X(t), t \geq 0\}$ is a continuous-time stochastic process indexed by the nonnegative real numbers. The state space of a stochastic process is defined as the set of all possible values that the random variables $X(t)$ can assume. Thus, a stochastic process is a family of random variables that describes the evolution through time of some (physical) process.

A stochastic process $\{X(t), t\ 0\}$ is said to be a *stationary process* if for all $n, s, t_1, \dots, t_n$ the random vectors $X(t_1), \dots, X(t_n)$ and $X(t_1 + s), \dots, X(t_n + s)$ have the same joint distribution. In other words, a process is stationary if, in choosing any fixed point s as the origin, the ensuing process has the same probability law (Ross, 2004).

Hydrologic variables can be considered stochastic processes (Koutsoyiannis, 1997). The fact that a physical process is considered a stochastic process does not mean that it has no deterministic part. It is well known that a lot of hydrologic processes show annual deterministic variability e.g. streamflow. This variability is seen sometimes as a trend or a jump by some authors. Koutsoyiannis (1997) prefers to envisage it as random variability that happens at various time scales. More specifically: The stochastic part of the process is not completely random, it has a *stochastic structure* or *stochastic memory.* That means that there is stochastic dependence at contiguous time moments and a larger factor of autocovariance of the process.

3.9 Fundamentals of statistics

Statistics is the applied branch of probability theory that deals with samples and populations. The most important objective of statistics is to estimate and to forecast. When the sample, that is represented by a random variable is known, the calculation of parameters is needed e.g. of a distribution, an estimation is made. Contrariwise when the parameters are known, and the random variable is needed, a forecast is made.

The definition of a statistical function and an estimator are provided by (Koutsoyiannis, 1997):

"By statistical function we mean any function of random variables of the sample in the form $\theta = g(X_1, \dots, X_n)$. From the sample observations we can directly calculate the value $\theta = g(x_1, \dots, x_n)$ of the statistical function"

"Statistical functions are used for estimating parameters of population. For every parameter η the population one or more statistical functions can be found, of the form $\theta = g(X_1, \dots, X_n)$, suitable for estimating this parameter. In in this case we say that $\theta = g(X_1, \dots, X_n)$ is the estimator of parameter η and that the arithmetic value of $\theta = g(x_1, \dots, x_n)$ is an assessment of η."

Probably the most usual statistical function is the *sample mean*, which is an average value estimator and is defined by the relationship.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.13}$$

The unbiased (and consistent) estimator of dispersion is the following, known as *sample dispersion*:

$$S_X{}^{*2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})}{n-1} \tag{3.14}$$

The estimator of the *sample's standard deviation* is the square root of the above equation which is biased.

The unbiased estimator of the third central moment is given by the following equation:

$$M_X^{*(3)} = \frac{n\sum_{i=1}^{n}(X_i - \bar{X})}{(n-1)(n-2)} \tag{3.15}$$

For estimating the skewness coefficient C_{sx} of the sample the following biased estimator is used:

$$C_{sx} = \frac{M_X{}^{(3)}}{S_X^3} \tag{3.16}$$

There are some equations that limit the bias but not an unbiased one.

The unbiased (and consistent) estimator of covariance is the following known as the *sample covariance*:

$$S_{XY}^{*} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)} \tag{3.17}$$

The estimator of the coefficient of correlation ρ_{XY} is known as *sample correlation coefficient* is considered approximately unbiased and is given by the following formula:

$$R_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})}} \tag{3.18}$$

4. Hydrosystems' Management and Reliability

4.1 Definitions

As a system is defined a set of independent elements, which are characterized by (Mays & Tung, 1992):

1. A boundary that determines whether the element belongs to the system or the environment
2. Interactions with environment (entry and exit)
3. Relationships between elements of and inputs and outputs

The first mention of a hydrosystem was given by Chow (1988). He considered hydrologic phenomena extremely complex and believed that they may never be completely understood thereat the "systems" concept was introduced. A hydrosystem is a system consisting of natural water bodies and technical projects that work together to serve one or more purposes, both of which refer to the exploitation of water as a natural resource, and the protection against its destructive action as a natural hazard (Koutsoyiannis and Xanthopoulos, 2014).

4.2 Reliability, Risk and Failure

The following definitions are provided by Mays and Tung (1992). *Risk* is defined as the probability of failure to reach the goal. *Reliability* is defined mathematically as the complement of the risk. The mathematical analysis of risk and reliability is defined as *reliability analysis*.

If X represents the maximum value of the physical process on a yearly basis (e.g. maximum annual flood) and n is n lifetime of the project, then the event $\{L \leq C\}$ equals n successively occurrences of the event $\{X \leq C\}$. In order not to exceed the value c throughout the lifetime of the project there should not be an exceeding in all n years of this duration. Considering that the floods of successive years are stochastically independent, risk is given by (Koutsoyiannis, 1997):

$$R = 1 - [P(X \leq C)]^n = 1 - [F_X(C)]^n \tag{4.1}$$

Failure of a hydrosystem can be defined as a situation in which the load L (demands) on the system exceeds the resistance C (capacity, or supply) of the system. L depends on the variability of the physical process as well as the time the project will be exposed to physical danger. This is called *project life time*. The *reliability* p_s of an hydrosystem is defined as the probability of nonfailure in which the resistance of the system exceeds the load; that is,

$$p_s = P(L \leq C) \tag{4.2}$$

In the above equation P (·) denotes probability. Conversely, the risk is the probability of failure when the load is larger than the resistance. Thus the failure probability (risk) p_f can be expressed mathematically as:

$$p_f = P(L > C) = 1 - p_s \tag{4.3}$$

There are two types of failure: structural failure and functional failure. *Structural failure* involves damage or change of the structure or facility, resulting in inability to function as desired. Contrarily, *performance failure* does not always involve structural damage. Nevertheless, manifestation of undesirable results in performance and elsewhere, occur. Generally, the two types of failure are related. Some structures, such as dams, levees, and, are designed on the concept of structural failure, whereas others, such as sewers, water supply systems are designed on the basis of performance failure (Mays and Tung, 1992).

In Figure 4.1 is shown the effect of hydraulic uncertainty on the overall failure probability. The assumption is that both random load and resistance are independent log-normal random variables. COV stands for coefficient of variation, a measure of uncertainty. Considering that there is an uncertainty not only in load but also in resistance, it is shown that the annual failure probability is significantly underestimated. It must be stated that the inherent natural randomness of hydrologic processes is not enough to estimate the total uncertainty. This figure clearly demonstrates a reason why the conventional frequency-analysis approach in reliability assessment of hydrosystems is deficient.

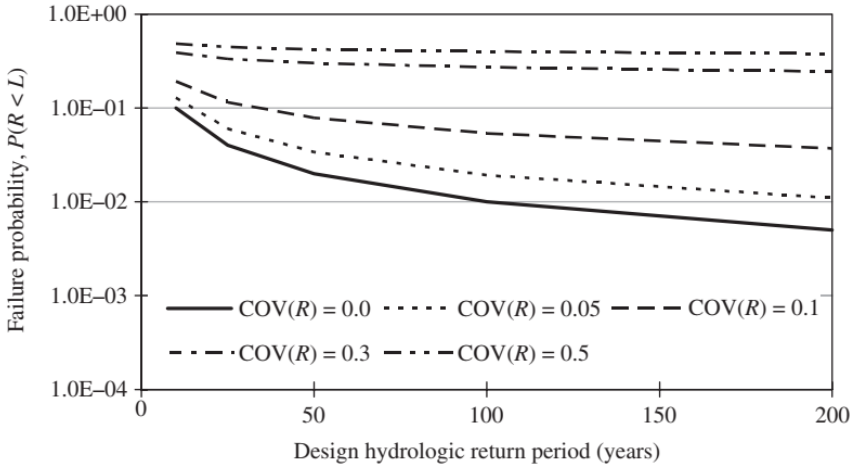


Figure 4.1 Effect of resistance uncertainty on failure probability under COV(L) = 0.1 (COV stands for coefficient of variation that can be used as a measure of uncertainty as stated before). (Mays and Tung, 1992).

The return period, T, of a given value of x variable X (which actually represents one stochastic process) is defined as the average number of time intervals (in these case, hydrological years) between two successive years of occurrence of the random variable of a value greater than or equal to that the given value x. Each occurrence must be stochastically independent of the previous and the random variable must be continuous. The Return Period can be given by the following equations, where R represents the risk, (Koutsoyiannis, 1997):

$$T = \frac{1}{P(X \geq x)} \tag{4.4}$$

$$T = \frac{1}{1 - (1 - R)^{1/n}} \tag{4.5}$$

Another reason conventional frequency-analysis approach in reliability assessment of hydrosystems is deficient is that there is a sometimes a long-range dependence, raising uncertainty. In other words the stochastic nature of the system is partially ignored.

4.3 Monte Carlo Methods

Monte Carlo methods are used nowadays used in hydrosystems for the many advantages and are a great way to deal with uncertainty. Monte Carlo methods are defined as: "Methods for finding solutions to mathematical and statistical problems by simulation. Used when the analytic solution of the problem is either intractable or time consuming." (Everitt and Skrondal, 2010).

The probabilistic behavior of an engineering system is essential if we accept that uncertainty issues are important and must be managed. The true distribution for the system response subject to parameter uncertainty is a lot of times difficult or even impossible sometimes to calculate. This is due to the complexity of the hydrosystems. In such cases, Monte Carlo simulation is a viable tool to provide numerical estimations of the stochastic features of the system response **(**Tung and Yen, 2005).

Practicing the method, random sampling is used from certain probability distributions to provide the random numbers used for generating the objects. The idea of the Monte Carlo techniques is to repeat the experiment many times to obtain many quantities of interest using the Law of Large Numbers and other methods of statistical inference (Kroese et al., 2014). Instead of repeating many times it is also possible to generate one long simulation. The choice depends on whether the initials values stop influencing sooner or later the generated values.

The law of large numbers formalizes the intuitive notion of probability which assumes that if in n identical trials an event A occurs nA times, and if n is very large, then nA/n should be near the probability of A. The formalization involves translating 'identical trials' as Bernoulli trials with probability p of a success. The law then states that as n increases, the probability that the average number of successes deviates from p by more then any preassigned value ε where $\varepsilon > 0$ is arbitrarily small but fixed, tends to zero (Everitt and Skrondal, 2010).

In Monte Carlo simulation the system performance measure is repeatedly measured under various system parameter sets that are generated from assumed probabilistic laws. It offers a practical approach to the uncertainty analysis because the stochastic behavior of the system response can be probabilistically duplicated **(**Tung and Yen, 2005).

Some uses of the Monte Carlo method are provided by Kroese et al. (2014).The first is sampling. In this case information is gathered about a random object by generating many realizations of it. That could be a hydrological model that represents a real physical system with rainfall, runoff, evapotranspiration etc. Another example can be a stochastic model with long range dependence or a hydrosystem.

The second is estimation. Here certain numerical quantities are determined related to a

simulation model. An example is the evaluation of multi-dimensional integrals via Monte Carlo techniques.

Optimization is a key factor in hydrosystems' management and engineering. The Monte Carlo method is used very effectively for the optimization of complicated objective functions, this is called stochastic optimization. In many applications these functions are deterministic but in order for the optimization to be more successful, randomness is integrated.

4.4 The parameterization-simulation-optimization method approach

One approach to the hydrosystems control problem is the parameterization-simulation-optimization method (Koutsoyiannis and Economou, 2003).

In contrast to most common methods that require a lot of control variables, the less widespread parameterization-simulation-optimization (PSO) method is a low-dimensional method. Few control variables are used, which are parameters of a simple rule that is exists through the entire control period. Through this rule the releases from different reservoirs are calculated. Specifically, the set of control variables consists of a ''target variable'' depending on the objective of the problem examined and a few parameters that determine a simple expression for distributing the degrees of freedom of the reservoir system operation. The parameterization of the rule is associated with the simulation of the reservoir system, which enables the calculation of a performance measure of the system for given parameter values, and nonlinear optimization, which enables determination of the optimal parameter values.

PSO does not only reach solutions that are not inferior to those of the benchmark methods but also has several advantages in some domains. In the theoretical level, the PSO method exhibits the following advantages over a high-dimensional method.

First of all, contrary to typical methods that use hundreds or thousands control variables PSO needs just a few for the same simulation period. As a result it is very effective and efficient in locating its optimal solution.

Second is the fact that the required computing time in PSO increases only linearly with the number of simulation steps n whereas in other methods computation time increase faster with n. Because of this, the performance measure in PSO can be based on a large simulation period, thus making it possible for a long year basis that leads to taking into account future impacts on the system.

Third is the fact that PSO avoids simplification of the system by describing its dynamics with a simulation model of the system, incorporating stochastic and deterministic components.

Fourth, the parametric method is compatible with the stochastic nature of the reservoir problems and very easily incorporates concepts like probability, reliability, expected value, etc., also assigning values to such quantities.

Fifth, the optimal values of the control variables do not depend on any any quantity that has a stochastic behavior and that means they do not have to be changed unless the system characteristics, the inflow statistics, or the operational objectives and constraints changed.

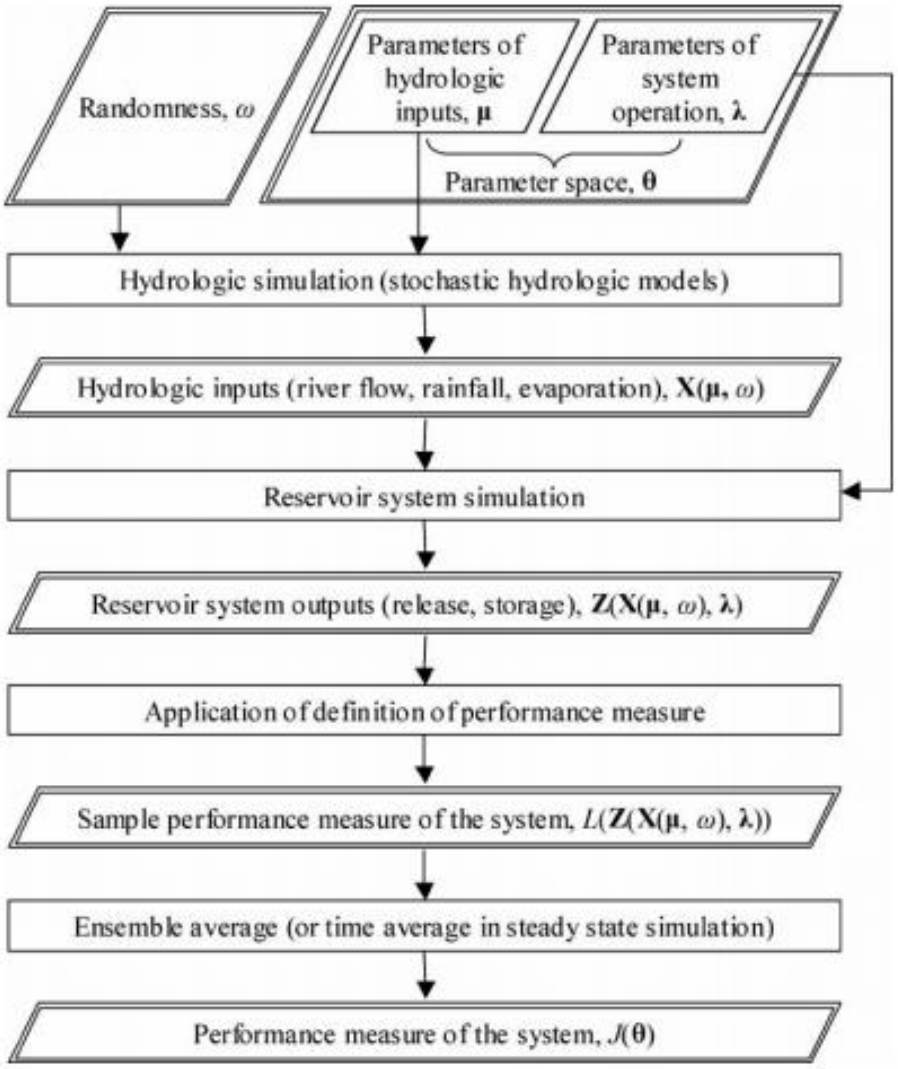


Figure 4.2 PSO method (Koutsoyiannis and Economou, 2003)

Figure 4.2 PSO method (Koutsoyiannis and Economou, 2003)

Sixth, the system can be very easily operated applying the parametric reservoir rule without model runs at all, once it is optimized with the PSO method. Similar to this, the model parameters and do not depend on forecasted values of inflows, high in uncertainty. In this way the operation policy is also not affected by this uncertainty.

In figure 4.2 we can see a representation of the method.

5. Stochastic Tools

Here some tools are imported for examining statistical properties in times series. Also, information about computational methods and their advantages/disadvantages and are being provided. Source of the latter is the study by Dimitriadis and Koutsoyiannis (2015) where the uncertainty and bias for three of the stochastic (autocovariance, power spectrum, climacogram) tools were calculated and this way the three were compared.

5.1 Autocorrelation/Autocovariance

The internal correlation of the observations in a time series, usually expressed as a function of the time lag between observations. The autocorrelation at lag k, $\gamma(k)$, is defined mathematically as:

$$\gamma(k) = \frac{E(X_t - \mu)(X_{t+k} - \mu)}{E(X_t - \mu)^2} \tag{5.1}$$

Where $X_t, t = 0, \pm 1, \pm 2, ...$, represent the values of the series and μ is the mean of the series. E denotes expected value. The sample statistic is given by the equation below:

$$\hat{\gamma}(k) = \frac{\sum_{i=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{i=1}^{n}(x_t - \bar{x})^2} \tag{5.2}$$

Where $\bar{x}$ is the mean of the series of observed values, $x_1; x_2; \ldots; x_n$. A plot of the sample values of the autocorrelation against the lag is known as the autocorrelation function or correlogram. The numerator of $\gamma(k)$ is called autocovariance. Autocorrelation is autocovariance standardized and is related to the discrete-time power spectrum by:

$$c_\eta = \int_0^{1/2} s_d(\omega) \cos(2\pi\omega\eta)\, d\omega \tag{5.3}$$

Autocovariance is intuitive in its definition and it is one of the tools most commonly used in time series analyzing and at the process of model selecting. It is well-defined and its bias can be easily estimated. However it has estimation errors larger than those of the climacogram (a tool discussed later). Besides its large bias, it is also prone to discretization errors as its value can never be equal with the true value in continuous time, even for an infinite sample size. Additional disadvantages are its negative values in the high lag tail (Dimitriadis and Koutsoyiannis, 2015).

5.2 Power spectrum

Power spectrum is a function (ω) , defined on $-\pi < \omega < \pi$ for a stationary time series, which has the following properties (Everitt and Skrondal, 2010):

1. The function defines the contribution to the total variance of the time series made by the frequencies in the band $[\omega, \omega\delta\omega]$
2. Harmonic components with finite power produce spikes $s(\omega)$
3. For real series the spectrum is symmetric, $s(\omega) = s(-\omega)$

The function is related to the autocovariance function of the series by:

$$s(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \gamma(k) \cos k\omega \tag{5.4}$$

The power spectrum of a stochastic process is discrete time $t = 0,1, \ldots$, with autocovariance function $\gamma_m = Cov[x_t, x_{t+m}], m = 0, \pm 1, \ldots$, the inverse discrete Fourier transformation of the autocovariance function is called power spectrum $s(\omega)$ with ω in the band $[0,1/2]$ .The following formula exists (Koutsoyiannis 2013):

$$s(\omega) = 2\gamma_0 + 4 \sum_{m=1}^{\infty} \gamma_m cos(2\pi m\omega) \tag{5.5}$$

For quick computation calculation the use of Fast Fourier Transform (FFT) is appropriate. For the calculation of the Fourier transform of a time series $x_0, x_1, \ldots x_{n-1}$, $d(\omega_p)$ given by:

$$d(\omega_p) = \sum_{t=0}^{n-1} x_t e^{i\omega_p t}, \qquad p = 0, 1, 2, \ldots n-1; \qquad \omega_p = \frac{2\pi p}{n} \tag{5.6}$$

The number of computations required to calculate $\{d(\omega_p)\}$ is n^2 which can be very large. The FFT reduces the number of computations to $O(nlog_2\, n)$ and operates in the following way:

Let $n = rs$ where r and s are integers. Let $= rt_1 + t_0$, $t = 0,1,2, \ldots r-1$. Further let $p_1 = 0,1,2, \ldots, r-1, p_0 = 0,1, \ldots, s-1$. The FFT can now be written:

$$d(\omega_p) = \sum_{t=0}^{n-1} x_t e^{i\omega_p t} = \sum_{t_0=0}^{r-1} \sum_{t_1=0}^{s-1} x_{rt_1+t_0} e^{\frac{2\pi ip}{n}(rt_1+t_0)} \tag{5.7}$$

$$= \sum_{t_0=0}^{r-1} e^{\frac{2\pi ip}{n}t_0} a(p_0, t_0)$$

where

$$a(p_0, t_0) = \sum_{t_1=0}^{s-1} x_{rt_1+t_0} e^{\frac{2\pi i t_1 p_0}{S}} \tag{5.8}$$

Estimating the amount of $a(p_0, t_0)$ needs only s^2 operations, and $d(\omega_p)$ only rs^2. The evaluation of $d(\omega_p)$ reduces to the evaluation of $a(p_0, t_0)$ which is itself a Fourier transform. Repeating the above task, the computation of $a(p_0, t_0)$ can be reduced in a similar way. The scheme can be repeated until a single term is reached (Everitt and Skrondal, 2010).

At the study by Dimitriadis and Koutsoyiannis (2015) the following conclusions were made about the power spectrum. It has the largest values of estimation error (between autocovariance, climacogram and power spectrum). It has a discretization error as its value even for an infinite sample size, can never be equal to the true value in continuous time. Additionally, in theory it is always positive, practical applications can result in negative values. Finally, it often has the highest value of skewness for its regular values and the smallest one for its NLD (negative logarithmic derivative) ones. The latter advantage of the power spectrum means that its mode should be close to the expected one.

5.3 Climacogram

Climacogram is defined as the variance of the averaged process $\underline{x}$(t) (assuming stationary) versus averaging time scale m and is symbolized by $\gamma(m)$. The climacogram is useful for detecting the long term change (or else dependence, persistence, clustering) of a process or multi-scale stochastic representation. Based on the process x_i at scale 1, we define a process $\underline{x}^{(k)}$ at any scale $k \geq 1$ as:

$$\underline{x}_i{}^{(k)} := \frac{1}{k} \sum_{l=(i-1)k+1}^{ik} \underline{x}_i \tag{5.9}$$

It is related to the autocorrelogram by the following transition:

$$\sigma^{(k)} = \frac{\sigma}{\sqrt{k}} \sqrt{a_k} \tag{5.10}$$

Where

$$a_k = a + 2 \sum_{j=1}^{k-1} (1 - \frac{j}{k}) \rho_j \tag{5.11}$$

And the term ρ_j is given by:

$$\rho_j = \frac{j+1}{2} a_{j+1} - j a_j + \frac{j-1}{2} a_{j-1} \tag{5.12}$$

Also the following classic statistical law exists:

$$\sigma^{(k)} = \frac{\sigma}{\sqrt{k}} \tag{5.13}$$

The bias in the climacogram can be calculated as follows. As shown in Koutsoyiannis (2011), assuming that we have $n = T/\Delta$observations of the averaged process $\underline{x}_i^{(\Delta)}$, where the observation period T is an integer multiple of Δ, the expected value of the empirical (sample) climacogram $\hat{\underline{\gamma}}(\Delta)$:

$$E\left[\hat{\underline{\gamma}}(\Delta)\right] = \frac{\gamma(\Delta) - \gamma(T)}{1 - \Delta/T} \tag{5.14}$$

The climacogram is also related with the power spectrum and the climacospectrum (presented below).

At the same study by Dimitriadis and Koutsoyiannis (2015) the following advantages/disadvantages were discovered. The climacogram had the smallest estimation error, between the three tools, in estimating the true values but also the true logarithmic derivatives. Its bias can be computed simply and analytically. Additionally the fact that its values are always positive is an advantage in stochastic modeling. Moreover it is well-defined with an intuitive definition and mostly monotonic. Finally, it has (for all the examined processes) values of sample skewness close to 0, for the small scale tail, while in the large scale tail; its skewness is increasing up to 3.

5.4 Climacospectrum

The climacospectrum is a newly introduced stochastic tool. It is defined by Koutsoyiannis (2017):

$$\zeta(k) := \frac{k(\gamma(k) - \gamma(2k))}{ln2} \tag{5.15}$$

It is written alternatively in terms of frequency $\omega = \frac{1}{k}$

$$\overline{\zeta}(\omega) \coloneqq \zeta\left(\frac{1}{\omega}\right) = \frac{\gamma\left(\frac{1}{\omega}\right) - \gamma(\frac{2}{\omega})}{(ln2)\omega} \tag{5.16}$$

Its name: “climacospectrum” comes from the fact that it has characteristics similar to these of the power spectrum. The entire area under the curve $\overline{\zeta}(\omega)$ is precisely equal to the variance $\gamma(0)$ of the instantaneous process, an attribute that is also evident in power spectrum $s(w)$. This is not the only connection with the power spectrum. The climacospectrum has also the same asymptotic behaviour with it:

$$\hat{\zeta}^{\#}(0) = -\zeta^{\#}(\infty) = s^{\#}(0), \qquad \hat{\zeta}^{\#}(\infty) = -\zeta^{\#}(0) = s^{\#}(\infty) \tag{5.17}$$

Specifically, the asymptotic behaviour of the second-order characteristics of a process for

$k \to 0$ and $k \to \infty$ is characterized by two parameters, M and H, which are given by:

$$M := \frac{\nu^{\#}(0)}{2} = \frac{\zeta^{\#}(0) - 1}{2}, \qquad H := 1 + \frac{\gamma^{\#}(\infty)}{2} = \frac{\zeta^{\#}(\infty) + 1}{2} \tag{5.18}$$

The climacospectrum has the following advantages (Koutsoyiannis, 2017): In comparison with the power spectrum it is superior in respect to the connection with conditional entropy production. Specifically it is more precise and without exceptions at all. Additionally, the variance, on which the definition of the climacospectrum is based, is more closely related to uncertainty, and as a result to the entropy of the process, than the power spectrum and the autocovariance. It is also very easy to calculate in contrast to the power spectrum that needs Fourier transformation. Furthermore, like the climacogram, it is not affected by discretization (while autocovariance and power spectrum are) and has a very small bias because of its definition as a difference of two variances, in which the biases tend to cancel out. In the end the empirical climacogram and climacospectrum are easily determined from data using nothing more than the standard statistical estimator of variance and they have a smooth shape, much smoother than those of the empirical autocovariance and power spectrum, thus enabling better model identification and fitting.

In Figure 5.1 a climacospectrum of of simulated time series is shown.

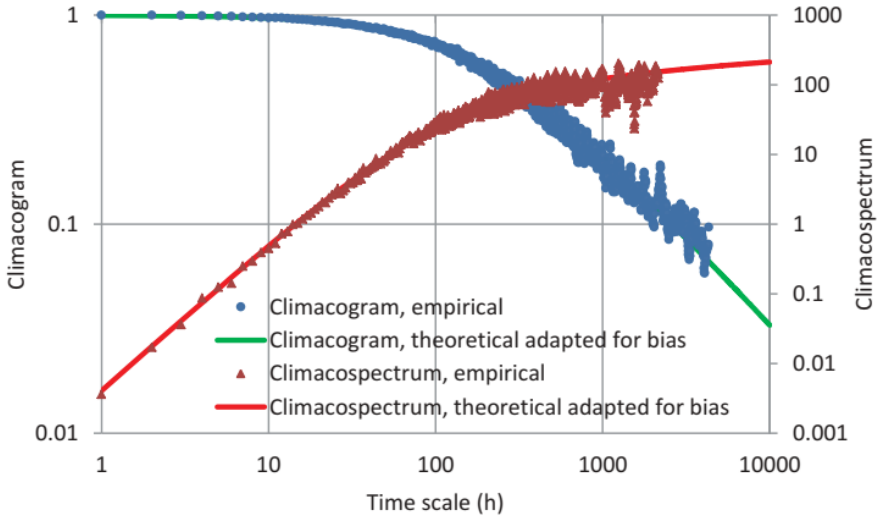


Figure 5.1 Comparison of the climacogram and climacospectrum of generated series with the FHK-C model (Koutsoyiannis, 2019).

5.5 Cross Climacogram

The climacogram can be further expanded to describe the dependence of different processes, replacing the concept of cross-correlogram of two stationary processes $\underline{x}(t)$ and $\underline{y}(t)$ and by the standardized crossclimacogram (SCC) for scale k and lag h:

$$\rho_{xy}(k,h) := var\left[\frac{\underline{X}(k)}{2\sqrt{\Gamma_x(k)}} + \frac{\underline{Y}(k+h) - \underline{Y}(k)}{2\sqrt{\Gamma_y(k)}}\right] \tag{5.19}$$

$$var\left[\frac{\underline{X}(k)/k}{2\sqrt{\gamma_x(k)}}+\frac{(\underline{Y}(k+h)-\underline{Y}(k))/k}{2\sqrt{\gamma_y(k)}}\right] \tag{5.20}$$

The cross-covariance can be replaced by the cross-climacogram (CC) and the cumulative crossclimacogram (CCC):

$$\gamma_{xy}(k,h) \coloneqq \rho_{xy}(k,h)\sqrt{\gamma_x(k)\gamma_y(k)} \tag{5.21}$$

$$\Gamma_{xy}(k,h) \coloneqq \rho_{xy}(k,h)\sqrt{\Gamma_x(k)\Gamma_y(k)} \tag{5.22}$$

This tool is used to detect time irreversibility in bivariate processes.

6. Long- Range Dependence in Hydrological Processes

6.1 Definition and Introduction

One important factor that should be taken into account when trying to understand hydrologic processes is long-range dependence or persistence. In hydrology this term is equivalent to the term “Hurst phenomenon”. Long range-dependence is defined by Everitt and Skrondal (2010) as: “Small but slowly decaying correlations in a stochastic process. Such correlations are often not detected by standard tests, but their effect can be quite strong.”

In hydrology this was first observed by Hurst while investigating the discharge of the Nile River. It is essentially the tendency of wet years to cluster and form bigger wet periods and the dry years similarly forming periods of draught.

6.2 Stochastic representation by a Markov process

A Markov process is a memoryless stochastic process that implies that to make predictions about the future behaviour of the system it suffices to consider only its present state and not its past history.

The Markov process is the most easy to use (simple expression of second order characteristics), the most parsimonious and it is commonly used due to its advantages. The disadvantages are: “its neutrality in terms of smoothness and persistence, and more specifically the low entropy production for large time scales” (Koutsoyiannis, 2017). In the end, these reasons do not make it always a good candidate to model natural behaviors. Nevertheless the Markov process can be approximated as a case of the Filtered Hurst Kolmogorov process discussed later on.

6.3 Stochastic representation by a Hurst-Kolmogorov process

The representation of the above physical phenomenon can be accomplished through stochastics, assuming a stochastic process. A sufficient model is the Hurst-Kolmogorov process. It is also called Fractional Gaussian Noise (FGN) in continuous time, introduced by Mandelbrot (1965). Alternative names for this are stationary increments of self-similar process and simple scaling process.

The advantage of the FGN model is apart from its very good fit to hydrologic time series is that it is parsimonious having only one parameter. A definition of parsimony is given by Everitt and Skrondal (2010): “The general principle that among competing models, all of which provide an adequate fit for a set of data, the one with the fewest parameters is to be preferred.”

Let k be a positive integer that represents a larger than the basic timescale of the process X_i .

The aggregated stochastic process on that timescale is denoted as:

$$Z_i^{(k)} \coloneqq \sum_{l=(i-1)k+1}^{ik} X_l \tag{6.1}$$

The statistical properties of Z_i for any time scale k are the following. The mean is given the equation below:

$$E\left[Z_i^{(k)}\right] = k\mu \tag{6.2}$$

The variance for a stochastic process that follows the Hurst phenomenon $\gamma_0^{(k)}$ is related to variance at the first scale γ_0. The bottom index is represents the lag for the autocovariance of the stochastic process.

$$\gamma_0^{(k)} \coloneqq Var[Z_i^{(k)}]k^{2H}\gamma_0 \tag{6.3}$$

H is called "Hurst coefficient" a measure of persistence or antipersistence of a process and its values can be in the interval [0, 1]. Values in the interval [0, 0.5] show no interest in hydrology. The value 0.5 corresponds to random noise. The standard deviation is given below:

$$\sigma^{(k)} \coloneqq (\gamma_0^{(k)})^{1/2}k^H\sigma \tag{6.4}$$

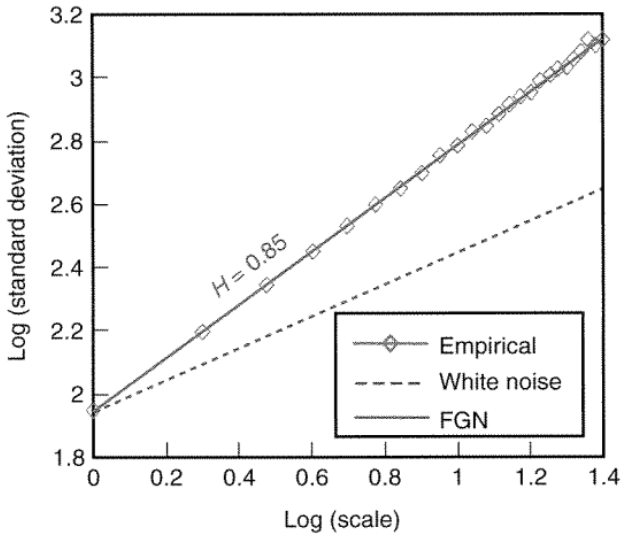


Figure 6.1 Aggregated standard deviation plot of the Nile timeseries (Koutsoyiannis, 2004)

The Figure 6.1 is produced using the formula 6.4. For H=0.85 the suggested stochastic representation seems to be in reality accurate. If there was not any long range dependence the white noise and empirical graph should be identical. However it is clearly evident that they are not.

It can be shown that the autocovariance at any scale and any lag is given by:

$$\rho_j^{(k)} = \rho_j = \frac{1}{2}(|j+1|^{2H} + |j-1|^{2H} - j^{2H} \tag{6.5}$$

As $j \to \infty$ the above approximates to:

$$\rho_j^{(k)} \approx H(2H-1)|j|^{2H-1} \tag{6.6}$$

It can also be shown that the power spectrum of the process is approximated by the following formula:

$$s_\gamma^{(k)}(\omega) = 2\sum_{j=-\infty}^{\infty} \gamma_j^{(k)} \cos(2\pi j\omega) \tag{6.7}$$

If we assume the process demonstrates scale invariant properties the following generalization is possible.

$$\left(Z_i^{(k)} - k\mu\right) = \left(\frac{k}{l}\right)^H (Z_j^{(l)} - l\mu) \tag{6.8}$$

This equation is valid for any integer i and j (with the process being stationary as a prerequisite) and any time scale k and l (Koutsoyiannis, 2002)

6.4 Stochastic Representation by a Filtered Hurst-Kolmogorov process

The Filtered Hurst Kolmogorov process is another model that can represent physical phenomena with long-range dependence. It is also called Hybrid Hurst-Kolmogorov process in its introduction by Koutsoyiannis (2015).

This process has some advantages over the HK process, while maintaining the persistence or antipersistence properties. These are: the variance of the instantaneous process is always finite ($\gamma_0 = \gamma(0) = \lambda$), while even for $0 < H < 0.5$ the initial part of the autocovariance function for small lags is positive for all variants of the process. A further important feature of this process is that it allows explicit control of the asymptotic behaviour of all properties at both ends, which are different at each end, opposite to the HK process, which implies simple scaling laws. The asymptotic properties are also easy to calculate.

The climacogram of the process is given below:

$$\gamma(\Delta) = \lambda\left(1 + \left(\frac{\Delta}{\alpha}\right)^{2M}\right)^{\frac{H-1}{M}} \tag{6.9}$$

Here α and λ are scale parameters with dimensions $[t]$ and $[x]^2$, respectively, H is the Hurst coefficient as in the HK process, a scaling parameter in the interval (0,1), and M is a second scaling parameter and in the interval (0,1), which will be called the fractal parameter. Both parameters are dimensionless. Parameter H determines the global properties of the process (as $t \to \infty$) and M determines the local properties (as $t \to 0$).

The process incorporates both the Markov and the HK processes. In the occasion when $H = M = 0.5$, Filtred Hurst-Kolmogorov is practically indistinguishable from a Markov

process. Furthermore, as $\alpha \rightarrow 0$, the process tends to a pure HK process with the same Hurst coefficient H. Additionally, under any parametric scenario, FHK exhibits Markov behavior for small time scales (if $M = 0.5$, or similar to Markov if $M \neq 0.5$) and Hurst behavior for large time scales.

6.5 Reproduction algorithm: Symmetric Moving Average scheme (SMA)

In 2000 a generalized framework for single-variate and multivariate simulation in stochastic hydrology was proposed by Koutsoyiannis. It is appropriate for short-term or long-term dependency processes and preserves the Hurst coefficient. Simultaneously, it explicitly preserves the coefficients of skewness of the processes.

The proposed framework incorporates short memory (autoregressive moving average) and long-memory (fractional Gaussian noise) models, considering them as special instances of a parametrically defined generalized autocovariance function. The generalized autocovariance function is then implemented in a generalized moving average generating scheme that yields a new time-symmetric (backward-forward) representation. It is called symmetric moving average (SMA).It can be used to generate any kind of stochastic process with any autocorrelation structure or power spectrum. The SMA scheme is given below:

$$X_\iota = \sum_{j=-q}^{q} a_{|j|} V_{i+j} = a_q V_{i-q} + \cdots + a_1 V_{i-1} + a_0 V_i + a_1 V_{i+1} + \cdots + a_q V_{i+q} \tag{6.10}$$

Where q theoretically is infinity but in for practical applications can be restricted to a finite number with a small error, as the sequence of weights a_j tends to zero for increasing j. The discrete Fourier transform $s_a(\omega)$of the a_j sequence is related to the power spectrum of the process $s_\gamma(\omega)$ by:

$$s_a(\omega) = \sqrt{2 s_\gamma(\omega)} \tag{6.11}$$

In addition with the formula 4.7 the discrete Fourier transform can be calculated:

$$s_a(\omega) = 2\sqrt{(2-2H)\gamma_0}(2\omega)^{0.5-H} \tag{6.12}$$

The scheme is accompanied by the following formulas:

$$a_0 = \frac{\sqrt{(2-2H)\gamma_0}}{1.5-H} \tag{6.13}$$

$$a_j \approx \frac{a_0}{2}[(j+1)^{H+0.5} + (j-1)^{H+0.5} - 2j^{H+0.5} \tag{6.14}$$

This method can also preserve the process skewness ξ_X by appropriately choosing the skewness of the white noise ξ_V. μ corresponds to the mean. The effectiveness is shown in Figure 6.2. The formulas are the following:

$$\left(a_0 + 2\sum_{j=1}^{s} a_j\right) E[V_i] = \mu \tag{6.15}$$

$$\left(a_0^3 + 2\sum_{j=1}^{q} a_j^3\right) \xi_V = \xi_X \gamma_0^{3/2} \tag{6.16}$$

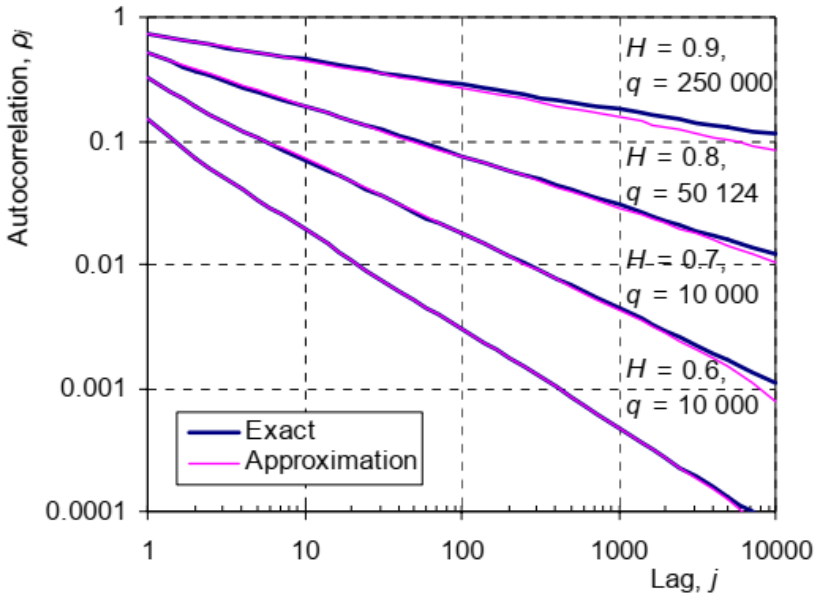


Figure 6.2 Approximate autocorrelation functions based on HK *vs* the exact autocorrelation functions of FGN for various values of the Hurst exponent *H* and the number of weights *q* (Koutsoyiannis, 2002)

7. Time's Arrow

7.1 Definition

The term "time's arrow" was developed at first by Eddington (1928) to describe time directionality, which can be determined by studying the organization of atoms, molecules and bodies. He states "Let us draw an arrow arbitrarily. If as we follow the arrow we find more and more of the random element in the state of the world, then the arrow is pointing towards the future; if the random element decreases the arrow points towards the past. That is the only distinction known to physics. This follows at once if our fundamental contention is admitted that the introduction of randomness is the only thing which cannot be undone. I shall use the phrase 'time's arrow' to express this one-way property of time which has no analogue in space. It is a singularly interesting property from a philosophical standpoint. We must note that:

- It is vividly recognized by consciousness.
- It is equally insisted on by our reasoning faculty, which tells us that a reversal of the arrow would render the external world nonsensical.
- It makes no appearance in physical science except in the study of organization of a number of individuals. Here the arrow indicates the direction of progressive increase of the random element."

7.2 General information and intuitive examples

Time's arrow can be a difficult term to grasp. This is due to the fact that the usual physical processes that are comprehended in the everyday life are usually irreversible in time. However that is not always the case. Microscopic physics, for example, gives no special status to any moment, and it distinguishes only weakly between the direction of the past and that of the future. Our intuitive perception of the world as unfolding in time therefore cannot be dismissed as being merely subjective. Here are given some information and examples for the better understanding of the scientific term.

This subjective intuitive perspective that we have, that times only moves forward, can help us understand an aspect of the problem. Let's imagine that we videotape an experiment of a dynamic system and reversing the video. For example the physical experiment of a ball falling from the sky. The reversed video does not seem very weird for our perception, as it shows a ball going up into the air (perhaps it is thrown). In contrast, the ball bouncing on the ground until it stays still, after it falls from the sky, is a process that if reversed seems peculiar. The reason is that the second process is irreversible in time. A second example is putting some milk in a cup of coffee. The reversed video of this process, milk parting the coffee after they have been mixed, seems peculiar since it never happens. In reality if we have one molecule of coffee and another one molecule of milk it is 100% probable that the milk will part from the coffee. It is easily understood that adding more molecules to the experiment just makes the probability of this smaller.

The direction of time can be also defined by a class of processes that destroy information and generate disorder. The irreversible processes that destroy macroscopic information (in the coffee-milk example molecular diffusion) are manifestations of the second law of thermodynamics. This law states that all natural processes generate entropy, a measure of disorder. The irreversible destruction of macroscopic order defines what can be called the "thermodynamic" arrow of time.

On the other hand there are a lot of processes that are irreversible and are diametrically opposite. All these processes have a quality in common: they generate order, or information; they transform a simpler state into a more complex one. In the phrase of Sir Arthur Eddington, they indicate which way "time's arrow" is pointing; they define what can be called the "historical" arrow of time (Layzer, 1975).

7.3 Informational entropy and uncertainty

The processes that define the historical and the thermodynamic arrows of time generate information and entropy mutually. Shannon showed in 1946, that information is part of the statistical description of a physical system. It is measured in bits, one bit is the quantity of information needed to decide between two equally probable cases. Additionally information can also be viewed as a measure of how highly organized the physical systems are. Shannon shows that the information content of a system is the minimum number of bits needed to encode a completely describe a system statistically.

The concept of entropy is directly linked to the concept of information. Entropy was first defined (by Clausius) in the context of thermodynamics and connected with time irreversibility. It measures the displacement of a system from thermodynamic equilibrium; at equilibrium the entropy is maximized.

Using a formula first produced by Ludwig Boltzmann and J. Willard Gibbs, Shannon defined the entropy of information theory. This entropy which measures the uncertainty associated the system in statistical terms. The thermodynamic entropy and the statistical entropy of have the same mathematical properties and they seem to represent different views of the same subject.

Entropy and information are related by a simple conservation law, which states that the sum of the information and the entropy is constant and equal to the system's maximum attainable information or entropy under given conditions. As a result a gain of information is always compensated for by an equal loss of entropy.

The following example and explanation is given by Layzer (1975). Let’s assume some physical system has eight possible states; in binary numbering they could be represented by: $000, 001, 010, 011, 100, 101, 110$ and 111. To specify a particular state, for example the one labeled 001, requires three binary digits, which is the amount of information needed to have the information of the exact state. The uncertainty or entropy associated with this description is clearly zero. Let’s now assume we had no information about the state of the system, we would assign equal probabilities to each of the eight possible states. In this case the information is evidently zero. Since the sum of the entropy and the information in the system is constant, the entropy must now be three bits. In general, if a system has 2^r possible states, where r an integer, the maximum quantity information or entropy is equal to the logarithm to the base 2 of 2^r, or r.

7.4 Irreversibility in stochastic processes

Many processes occurring in the natural sciences, engineering, finance, and economics exhibit some form of nonlinear behavior. This means that some of the characteristics cannot be modeled using Gaussian linear processes because time reversibility is not evident. Weiss (1975) showed that if the process $\underline{x}(t)$ is Gaussian then it is time reversible. As a result a directional process cannot be Gaussian. He also showed that a discrete-time autoregressive moving-average (ARMA) process is reversible if and only if it is Gaussian. This conclusion is very important because is shows that stationary series which show evidence of directionality cannot be modeled by Gaussian ARMA models. So other models should be used to model accurately this behavior (Koutsoyiannis, 2019).

As it is previously defined, a stochastic process $\underline{x}(t)$ is a collection of (usually infinitely many) random variables $\underline{x}$ indexed by t, typically representing time. In turn, a random variable, $\underline{x}$, is an abstract mathematical entity, associated with a probability distribution function $F(x) \coloneqq P\{\underline{x} \leq x\}$ where x is any numerical value numerical value (a regular variable).. The stochastic process x(t) represents the evolution of the system over time, while a trajectory $x(t)$ is a realization of $\underline{x}(t)$; if it is known at certain points t_i, it is a time series.

Weiss (1975) defines a stochastic process $\underline{x}(t)$, at (continuous) time t, with nth order distribution function:

$$F(x_1, x_1, \dots, x_n; t_1, t_2, \dots, t_n) \coloneqq P\{\underline{x}(t_1) \leq x_1, \underline{x}(t_2) \leq x_2, \dots \underline{x}(t_n) \leq x_n\} \tag{7.1}$$

as time reversible or time symmetric if its joint distribution does not change after reflection of time about the origin, i.e., if for any $n, t_1;\ t_2;\ \dots;\ t_{n-1};\ t_n$,

$$F(x_1, x_1, \dots, x_n; t_1, t_2, \dots, t_n) = F(x_1, x_1, \dots, x_n; -t_1, -t_2, \dots, -t_n) \tag{7.2}$$

The next important step studying irreversibility is the method for its detection and quantification. A method proposed by Psaradakis (2008) is to measure the probability of the differenced process. As the probability is positive, there is a deviation of the median of the differenced process from zero. In a more recent and study, Müller et al. (2017) they propose a class of new — easy to calculate — tests for time reversibility and suggest different ways to implement it. They used as an indicator of asymmetry the third moment of differences, but of the empirical copulas rather than of the time series. Further, they performed simulations of combined sewer systems with original and time-reversed time series and found “significant deviations of more than 10%”. In a study by Serinaldi and Kilsby (2016) directed horizontal visibilitygraphs (DHVGs) were used to perform an analysis of the dynamics of streamflow fluctuations with focus on time irreversibility and long range dependence. The study of irreversibility in Koutsoyiannis (2019) states that time asymmetry requires the study of third moment μ_3 and the coefficient of skewness C_s of the process, original and differenced. The first moment (mean) of the differenced process is always zero while the second one (variance) is always positive and thus he concludes that they don’t provide indications on time asymmetry. Hence, the least-order moment that can be used to detect reversibility is the third. The cumulative process enables representation of the process in discrete time τ:

$$\underline{x}_\tau = \frac{1}{D}\int_{(t-1)D}^{tD} \underline{x}(u)du = \frac{\underline{X}(\tau D) - \underline{X}((t-1)D)}{D} \tag{7.3}$$

D is the time unit. The symbol for continuous time is t and for discrete time is τ.This can be expanded to define a discrete time process averaged at scale $k = \kappa D$.

$$\underline{x}_\tau{}^{(\kappa)} := \frac{\underline{X}(\tau\kappa D) - \underline{X}((t-1)\kappa D)}{\kappa D}$$

$$= \frac{\underline{x}_{(\tau-1)\kappa+1}, \underline{x}_{(\tau-1)\kappa+2} + \cdots + \underline{x}_{\tau\kappa}}{\kappa} \tag{7.4}$$

To study the time asymmetry of processes we define the differenced process in discrete and continuous time, respectively, as:

$$\tilde{\underline{x}}_\tau := \underline{x}_\tau - \underline{x}_{\tau-1}, \qquad \tilde{\underline{x}}_{\tau,\eta} := \underline{x}_\tau - \underline{x}_{\tau-\eta}$$

$$\tilde{\underline{x}}(t, D) := \underline{x}(t) - \underline{x}(t-D) \tag{7.5}$$

The cumulative process for discrete time is:

$$\tilde{\underline{X}}_\kappa = \tilde{\underline{x}}_1 + \tilde{\underline{x}}_2 + \cdots + \tilde{\underline{x}}_\kappa$$

$$= \underline{x}_1 - \underline{x}_0 + \underline{x}_2 - \underline{x}_1 + \cdots + \underline{x}_\kappa - \underline{x}_{\kappa-1} \tag{7.6}$$

And for continuous time:

$$\tilde{\underline{X}}(k, D) := \int_0^k \tilde{\underline{x}}(t, D)dt = \int_0^k \Big(\underline{x}(t) - \underline{x}(t-D)\Big)\,dt \tag{7.7}$$

$$= \int_0^k \underline{x}(t)dt - \int_{-D}^{k-D} \underline{x}(t)dt = \underline{X}(k) - \underline{X}(k-D) + \underline{X}(-D) \tag{7.8}$$

For $k = \kappa D$:

$$\tilde{\underline{X}}(k, D) = D\big(\underline{x}_\kappa - \underline{x}_0\big) = D\tilde{\underline{X}}_\kappa \tag{7.9}$$

For the averaged differenced process at discrete time scale κ we have:

$$\tilde{\underline{x}}_\tau{}^{(\kappa)} := \frac{\tilde{\underline{X}}(\tau\kappa D) - \tilde{\underline{X}}((\tau-1)\kappa D)}{\kappa D} = \frac{\tilde{\underline{X}}_{\tau\kappa} - \tilde{\underline{X}}_{(\tau-1)\kappa}}{\kappa}$$

$$= \frac{\underline{x}_{\tau\kappa} - \underline{x}_{(\tau-1)\kappa}}{\kappa} \tag{7.10}$$

For the original process, averaged at the integer time scale κ, the marginal third moment

characteristics are:

$$\mu_3(\kappa) \coloneqq E\left[\left(\underline{x}_\tau^{(\kappa)} - \mu\right)^3\right], \qquad C_s(\kappa) \coloneqq \frac{\mu_3(\kappa)}{(\gamma(\kappa))^{3/2}} \tag{7.11}$$

the second and third moments of the averaged-differenced process:

$$\tilde{\gamma}(\kappa) \coloneqq var\left[\underline{x}_\tau^{(\kappa)}\right], \qquad var\left[\underline{\tilde{X}}_\kappa\right] = \kappa^2\tilde{\gamma}(\kappa) \tag{7.12}$$

$$\tilde{\mu}_3(\kappa) \coloneqq E\left[\left(\underline{\tilde{x}}_\tau^{(\kappa)}\right)^3\right], \qquad C_s(\kappa) = \frac{\tilde{\mu}_3(\kappa)}{(\tilde{\gamma}(\kappa))^{3/2}} \tag{7.13}$$

7.5 Irreversibility in streamflow at small scales

Koutsoyiannis (2019) stated that the irreversibility of streamflow is marked for scales of several days and this highlights the need to reproduce it in flood simulations. This is offcourse neglected by the models used today.

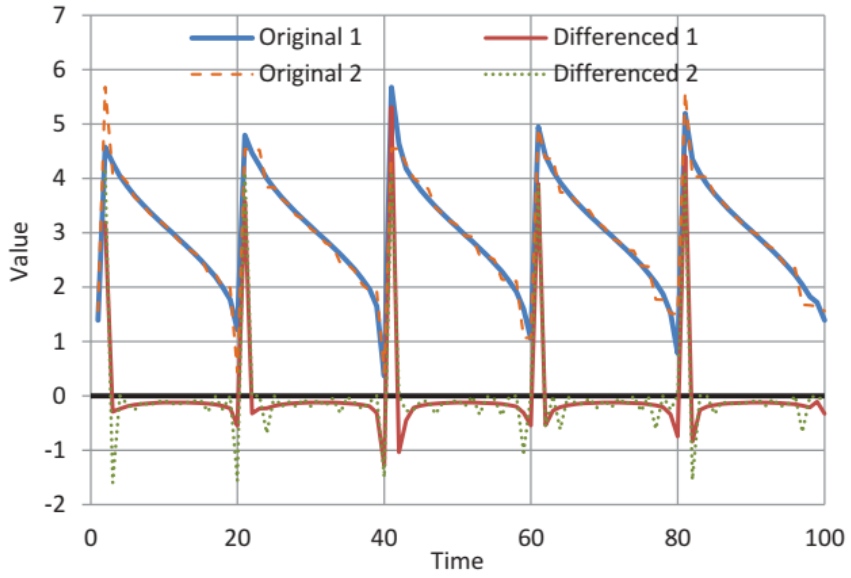


Figure 7.1 Plot of two synthetic time series generated by maximizing time irreversibility properties of a process restricted to be marginally Gaussian (N(3, 1)) with lag one autocorrelation 0.5, so that the variance of the differenced process is also 1 (equal to that of the original process). Solution 1 maximizes the skewness of the differenced process. (Koutsoyiannis, 2019)

In Figure 7.1 we can see how maximized irreversibility affects time series. It can be understood that the hydrograph partly mimics the above behavior. More specifically the ascending part of the hydrograph is steeper than the descending one. This property is in fact a result of time's arrow and can be modeled as a statistical parameter. Koutsoyiannis (2019) proposes a model called AMA for this purpose. Mathai and Mujumdar (2019) in recent study have also built a model to simulate time irreversible streamflow at multiple sites. Multisite correlated streamflow states were generated and then flow sequences that are constructed

considering the ascension and recession limbs of the hydrograph at individual sites independently.

In the following figures (7.2, 7.3) real world data is shown to highlight the resemblance with the previous figure (7.1). In the figure 7.3 we can see that this behavior is evident at different scales. In the next two figures (7.4, 7.5) synthetic time series are plotted to show the difference in a time symmetric model and an asymmetric one.

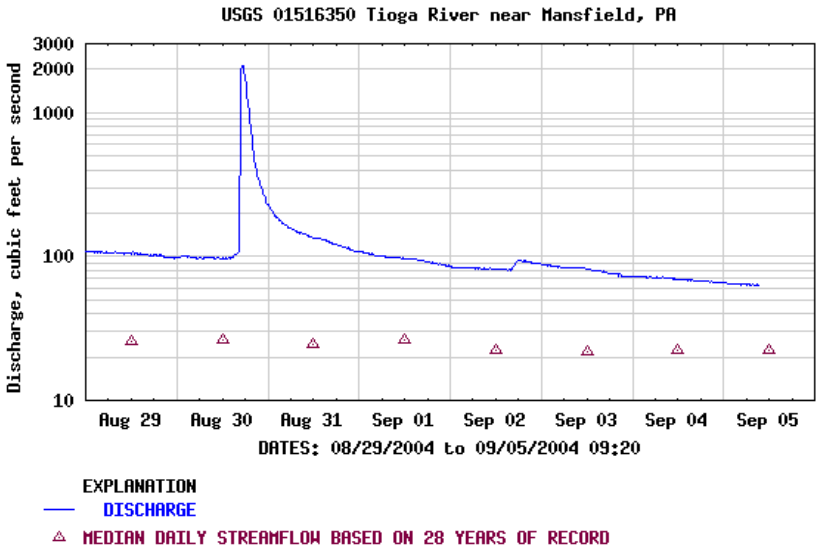


Figure 7.2 Real world data hydrograph

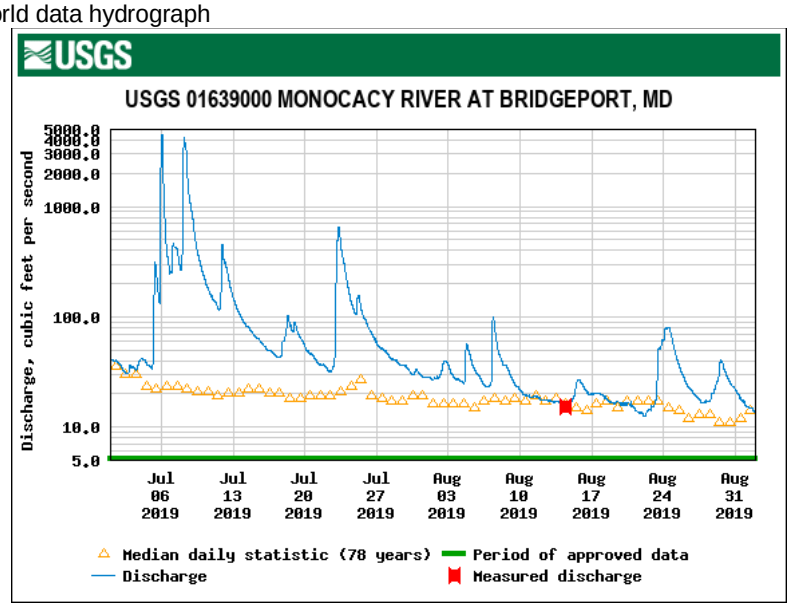


Figure 7.3 Real world data showing irreversibility at multiple scales

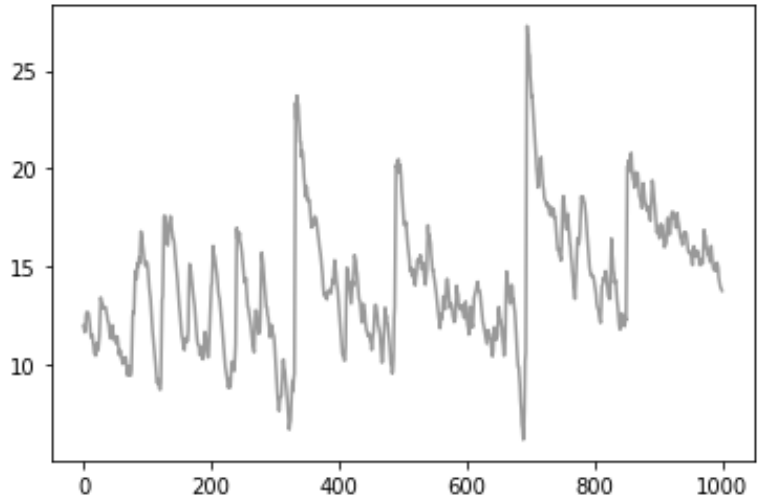


Figure 7.4 Synthetic time series with reversibility parameter equal to13.Produced using the the original methodology with the AMA scheme by Koutsoyiannis (2019)

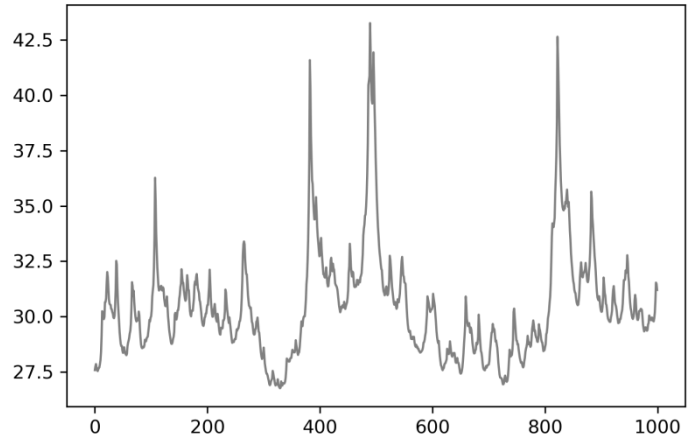


Figure 7.5 Synthetic time series with zero reversibility. Produced using the time symmetric SMA scheme by Koutsoyiannis (2000).

In the figure 7.6 we can see the irreversibility in a river from the USGS database at various scales.

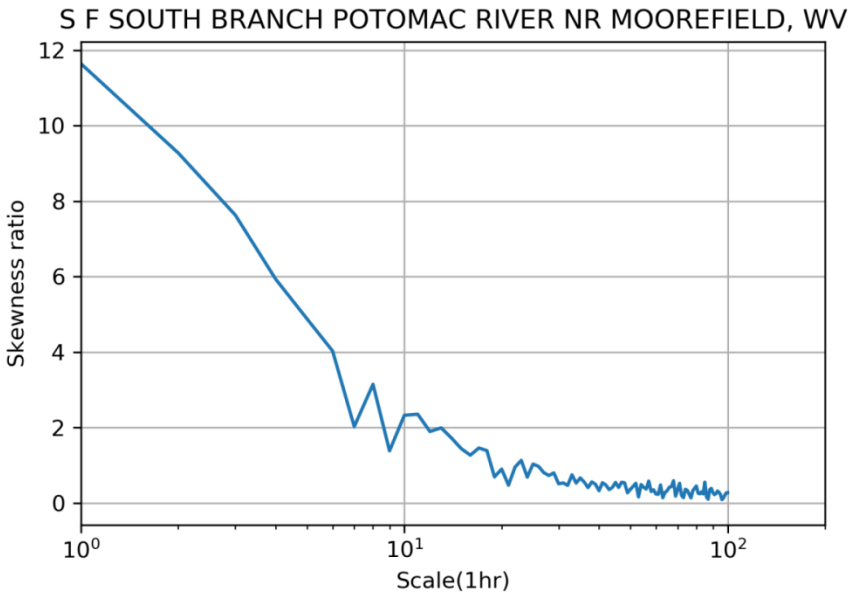


Figure 7.6 Reversibility test for a station from USGS database. The term "Skewness ratio" refers to the ratio of the skewness of the differenced process and the original process. The code used for its production is given in Apendix: C.

7.6 Reproduction algorithm for irreversible processes of one variable

Koutsoyiannis (2019) has proposed a mathematical framework to simulate irreversible processes. At first the moving average scheme (MA) is used (Koutsoyiannis, 2000),

$$\underline{x}_\tau = \sum_{i=-\infty}^{\infty} a_i \underline{v}_{\tau-i} \tag{7.14}$$

where a_i are coefficients to be calculated from the autocovariance function and v_i is white noise averaged in discrete-time. Writing the above equation for $\underline{x}_{\tau+\eta}$, multiplying it with the MA equation and taking expected values we find the convolution expression:

$$c_\eta = \sum_{l=-\infty}^{\infty} a_l a_{\eta+l} \tag{7.15}$$

The key task is to find the sequence of $a_\eta, \eta = \cdots, -1,0,1, \ldots$, so that above equation holds true. A known solution (Koutsoyiannis 2000) is the symmetric moving average (SMA) scheme in which $a_{-\eta} = a_\eta$.

He also provides a generic solution of an asymmetric moving average scheme in which the coefficients a_η are given by:

$$a_\eta = \int_{-1/2}^{1/2} e^{2\pi i(\theta(\omega)-\eta\omega)} A^R(\omega) d\omega \tag{7.16}$$

Where i is the imaginary unit, $\theta(\omega)$ is any (arbitrary) odd real function (meaning $\theta(-\omega) = -\theta(\omega)$) and

$$A^R(\omega) \coloneqq \sqrt{2 s_d(\omega)} \tag{7.17}$$

The equation can be written also as:

$$a_\eta = 2\int_0^{1/2} \cos\big(2\pi(\theta(\omega) - \eta\omega)\big) A^R(\omega) d\omega \tag{7.18}$$

To calculate the sequence of a_η we must first know the frequency functions $A^R(\omega)$ (from the power spectrum) and $\theta(\omega)$. For an array of frequencies $\omega_j = j\, w_0, j = 0,1, \dots, q, w_0 \coloneqq 1 \,/\, qD$, we form data arrays (vectors) A^R and A^I, with the superscripts R and I standing for a real and an imaginary vector, respectively. On these we perform FFT methodology discussed earlier. Both vectors are of size $2q$ indexed as $0, \dots, 2q - 1$. FFT works only if we have chosen q as a power of 2. When q is not a power of 2 DFT (Direct Fourier Transform). The real vector has elements:

$$[\boldsymbol{A}^R]_j = \frac{A^R(\omega_j)\cos\left(2\pi\theta(\omega_j)\right)}{4q}, \qquad j = 0, \dots, q \tag{7.19}$$

$$[\boldsymbol{A}^R]_j = [\boldsymbol{A}^R]_{2q-j}, \qquad j = q+1, \dots, 2q-1 \tag{7.20}$$

And the imaginary vector:

$$[\boldsymbol{A}^I]_j = \frac{A^R(\omega_j) sin\left(2\pi\theta(\omega_j)\right)}{4q}, \qquad j = 0, \dots, q-1 \tag{7.21}$$

$$[\boldsymbol{A}^I]_j = 0, \qquad j = q \tag{7.22}$$

$$[\boldsymbol{A}^I]_j = -[\boldsymbol{A}^I]_{2q-j}, \qquad j = q+1, \dots, 2q-1 \tag{7.23}$$

Performing the Fourier transform and getting the real part of the result for $j = 0, \dots, q$, we will have the sequence of a_η

7.7 Algorithm modification for further irreversibility conservation

In the current study there is an attempt to modify the algorithm proposed by Koutsoyiannis (2019). The goal is to simulate time series that conserve the irreversibility at larger scales without disturbing the balance of the existing method and keeping all of its necessary features. The existing algorithm simulates time series preserving irreversibility at only first scale therefore the attempt is to preserve irreversibility at both scales: first and second. At first it is important to calculate the theoretical moments of the AMA model at the first and second scale.

For simulation length i, the AMA model can be written as:

$$X_i = \sum_{j=1}^{2q+1} a_{2q+2-j} V_{i+j-1} = a_{2q+1} V_i + \cdots + a_1 V_{i+2q+1} \tag{7.24}$$

Where V_i is lognormal white noise. For the first scale we must calculate the second and third moment of the differenced and the original sequence respectively.

The second moment of the original sequence is:

$$M_{orig.}^{(2)} = \sum_{j=1}^{2q+1} \left(a_{2q+2-j}\right)^2 \tag{7.25}$$

Also the second moment of the differenced sequence is:

$$M_{differ.}^{(2)} = \sum_{j=1}^{2q+1} \left(a_{2q+1-j} - a_{2q+2-j}\right)^2 + {a_{2q+1}}^2 \tag{7.26}$$

The third moment of the original sequence is:

$$M_{orig.}^{(3)} = \sum_{j=1}^{2q+1} \left(a_{2q+2-j}\right)^3 \tag{7.27}$$

Finally the third moment of the differenced sequence is:

$$M_{differ.}^{(3)} = \sum_{j=1}^{2q+1} \left(a_{2q+1-j} - a_{2q+2-j}\right)^3 + {a_{2q+1}}^3 \tag{7.28}$$

At last, for the second scale we must calculate the second and third moment of the differenced and the original sequence respectively. We highlight that at first the process is averaged at the second scale and afterwards differenced and not the other way around.

The second moment of the original sequence at second scale:

$$M_{orig.}^{(2)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1} \frac{\left(a_{2q+1-j} + a_{2q+2-j}\right)^2}{4} + \frac{a_{2q+1}{}^2}{4} \tag{7.29}$$

Also the second moment of the differenced sequence at second scale is:

$$\begin{aligned} M_{differ.}^{(2)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1} \left[+ \frac{\left(a_{2q-1-j} + a_{2q-j} - a_{2q+1-j} - a_{2q+2-j}\right)^2}{4} \right] \\ + \frac{\left(a_{2q} + a_{2q-1} - a_{2q+1}\right)^2}{4} + \frac{\left(a_{2q} + a_{2q+1}\right)^2}{4} + \frac{a_{2q+1}{}^2}{4} \end{aligned} \tag{7.30}$$

The third moment of the original sequence at the second scale is:

$$M_{orig.}^{(3)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1} \frac{\left(a_{2q+1-j} + a_{2q+2-j}\right)^3}{8} + \frac{a_{2q+1}{}^3}{8} \tag{7.31}$$

Finally the third moment of the differenced sequence at the second scale is:

$$\begin{aligned} M_{differ.}^{(3)}{}^{(\kappa=2)} = \sum_{j=1}^{2q+1} \left[- \frac{\left(a_{2q-1-j} + a_{2q-j} - a_{2q+1-j} - a_{2q+2-j}\right)^3}{8} \right] \\ - \frac{\left(a_{2q} + a_{2q-1} - a_{2q+1}\right)^3}{8} - \frac{\left(a_{2q} + a_{2q+1}\right)^3}{8} - \frac{a_{2q+1}{}^3}{8} \end{aligned} \tag{7.32}$$

After calculating the sample moments, computational tools have to equalize between the sample (empirical) and the sequence (theoretical) moments so that the parameters are satisfied. Optimization tools are used to find the parameters needed. The parameterization follows the same methodology as in Koutsoyiannis (2019): A definition of $\theta(\omega)$ as the smooth minimum of two hyperbolic functions of frequency, i.e.:

$$\theta(\omega) = \frac{1}{\zeta} \ln\left(e^{\zeta\theta_1(\omega)} + e^{\zeta\theta_2\left(1/2-\omega\right)}\right), \qquad \theta_i(\omega) \coloneqq \frac{C_{1,i}\omega}{C_{2,i} + \omega} + C_{0,i} \tag{7.33}$$

After parameter estimation we also use the function of $\theta(\omega)$ that was found to perform the Fourier transform and get the real part of the result for $j = 0, \dots, q$, At last we have the sequence of a_η. At the end we have simulations that conserve reversibility at two scales. For example if we choose as the first scale as one hour the second scale to be conserved is two hours.

8. Case Study: Monocacy River

8.1 Introduction

In this chapter there is a case study to test the results from using the modified algorithm described in the previous chapter with real data. The case of Monocacy River is investigated. The station name is: “MONOCACY RIVER AT BRIDGEPORT” and is located in the state Maryland in the U.S.A. The reversibility of the time series is being quantified and conserved at both scales (hourly and two hours). In the figure 8.1 daily mean discharge is given for the last five years.

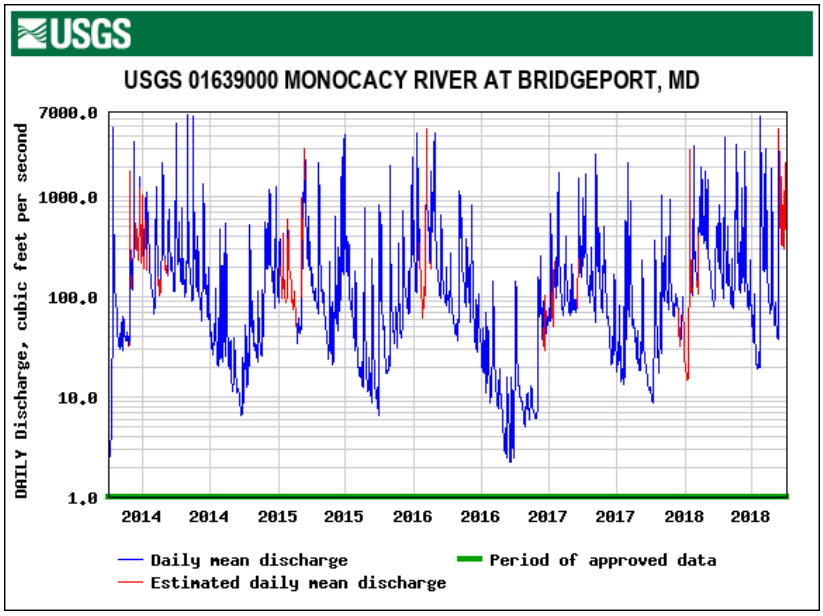


Figure 8.1 Daily mean discharge of Monocacy River at Bridgeport, Maryland.

Figure 8.2 Monocacy River.

8.2 Computational tools

For the computational part of the study Spyder was used. Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language. In Spyder's website it is stated: "Spyder is a powerful scientific environment written in Python, for Python, and designed by and for scientists, engineers and data analysts." Python is a general purpose language created by Guido van Rossum and first released in 1991. Python's design philosophy emphasizes code readability . Its language constructs and object oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

The following packages were used with Python:

- Numpy: Base N-dimensional array package
- Math: for mathematical functions
- Pandas: data structures and data analysis tools
- Matplotlib: for plotting
- Scipy: package for scientific computing
- Climata: is a pythonic interface for loading and processing time series data from climate and flow monitoring stations and observers. climata leverages a number of web services as listed below. Climata is powered by wq.io, and shares its goal of maximizing the reusability of data parsing code, by smoothing over some of the differences between various data formats.

  For optimization the scipy.optimize.minimize function was used. It minimizes a scalar function of one or more variables using Sequential Least SQuares Programming (SLSQP).

8.3 Database

Time series were downloaded from the water department of the United States Geological Survey (USGS). USGS is the largest provider of in situ water data in the world, and the Water Resources Mission is committed to observe, understand, predict, and deliver water data and information.

The USGS works with partners to monitor, assess, conduct targeted research, and deliver information on a wide range of water resources and conditions including streamflow, groundwater, water quality, and water use and availability. It has collected water-resources data at approximately 1.5 million sites in all 50 States.

There is a variety of types of data, but generally fit into the broad categories of surface water and groundwater. Surface-water data, such as gage height (stage) and streamflow (discharge), are collected at major rivers, lakes, and reservoirs. Groundwater data, such as water level, are collected at wells and springs. Water-quality data are available for both surface water and

groundwater. Examples of water-quality data collected are temperature, specific conductance, pH (www.usgs.gov).

8.4 Station Information

Figure 8.3 Monocacy River at Bridgeport.

Figure 8.4 Monocacy River at Bridgeport Station.

The exact location of the station is: (39°40'44.6, 77°14'04.3). It is in Frederick County, Maryland and the Hydrologic Unit's number is: 02070009. It is located on right bank at downstream side of bridge on State Highway 140 at Bridgeport, 0.9 mi upstream from Cattail Branch, 3.4 miles northwest of Taneytown, 4.8 miles downstream from confluence of Rock and Marsh Creeks at Pennsylvania-Maryland State line, and 52 miles upstream from mouth.

The drainage area is 173 miles^2 and the period of record is from May 1942 to now.

Some remarks are that there is an occasional regulation at low flow from Lake Heritage and other unknown sources upstream from station.

Extremes for the period of record: maximum discharge, 24,400 ft3/s, June 19, 1996, gage height, 25.42 ft; minimum discharge, 0.0 ft3/s, July 24-29, 1966.

Extremes outside the period of record Flood of Aug. 24, 1933, reached a stage of about 25 ft, present site and gage datum, from floodmarks, discharge, about 23,000 ft3/s. Stage exceeded that of June 1889, from information by local residents (www.usgs.gov).

8.5 Methodology

The methodology implemented is discussed in this section.

First of all the station was picked randomly. The time series was downloaded with the help of climata.usgs package for a period of 5 years (from November 2014 until November 2019). The data was at first 15 minute measurements but was aggregated to 1 hour scale – relevant scale for hydrosystem management.

After that, stationarization of the time series was considered important to take place. Specifically the effect of the annual cycle was "removed" by multiplying the discharge values by 12 different coefficients, one per month, summing up to 1. These coefficients were found by minimizing the total variance of the transformed time. For optimization the scipy.optimize.minimize function was used. It minimizes a scalar function of one or more variables using Sequential Least SQuares Programming (SLSQP). A python function was developed to perform the above task. The function uses as an input the station "code number" and gives back the time series aggregated at hourly scale and stationarized.

Additionally the data had to be fitted into the Filtered Hurst Kolmogorov model to estimate the parameters H, M and a. The time series were normalized at first. For these tasks another function was built. The function used both the climacogam and the climacospectrum (empirical and theoretical) to fit the data giving emphasis to the climacospectrum at the finer scales and to the climacogram in the greater scales. The reason behind that is in the theoretical context of these stochastic tools and has been discussed earlier. Also the fitting was not done with the pure theoretical climacogram but with the one adapted for bias, in a way also discussed earlier. The same happened with the climacospectrum.

More Python functions were built to calculate the discrete power spectrum through FFT and the AMA coefficient. These were translated to Python from the VBA file accompanying the study by Koutsoyiannis (2019).

The next step was to develop a Python function to detect reversibility scalewise. It aggregates the data until scale 100 is reached and calculates the sample skewness of the differenced and the original process. The ratio is the reversibility estimator and the output is a plot. Also the reversibility of scale 1 and 2 was important to be calculated because it was to be conserved later by the algorithm.

In the first case, when reversibility is conserved at only scale one, optimization tools are used to find the parameters needed to find the constant θ. In the second case for sequence $\theta(\omega)$ which is defined as the smooth minimum of two hyperbolic functions of frequency, again optimization was used. The concept is that after building functions to calculate the sample and theoretical moments, computational tools have to minimize difference between the sample(real) and the sequence(theoretical) moments. In the second case the difference is that this happens for two scales and the square error is minimized. The output is the $\theta(\omega)$ sequence.

With the knowledge of the power spectrum and the θ or $\theta(\omega)$ sequence we are able to calculate the AMA coeffiecients, the a_η sequence through another Python function. After that synthetic time series can be simulated.

In the end we make 100 simulations with 10000 length for each case and test them for reversibility using the same criterion as before (skewness ratio). The average of the

simulations at each scale is being highlighted in black and each simulation is shown in grey, a denotation of variability.

8.6 Results

In figure 8.5 the reversibility of Monocacy River at Bridgeport scalewise. The ratio at scale one was calculated as: $r_1 = 1.390$ and at scale two: $r_2 = 1.197$. It seems that it safe to say that the physical process is reversible at scale 100 (approximately 4 days).

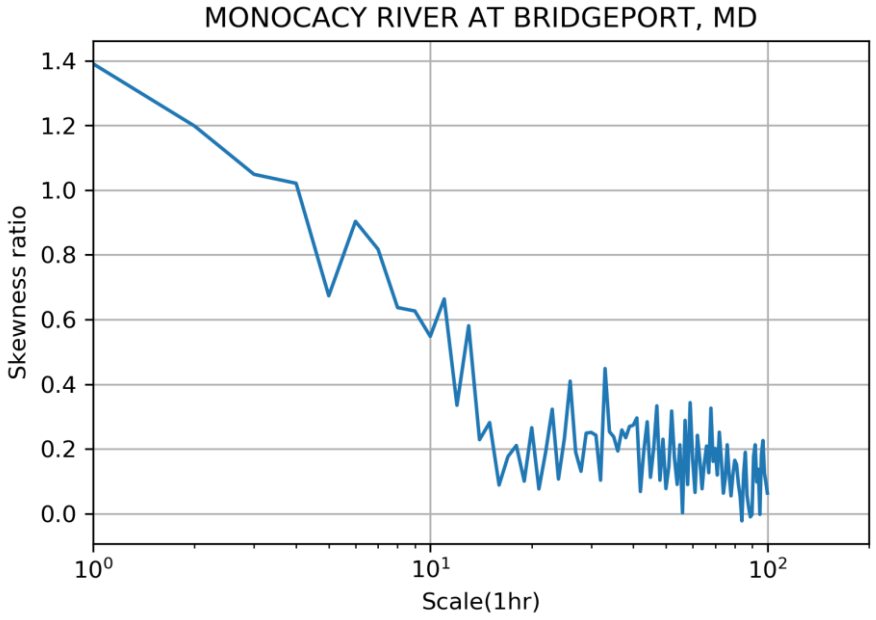


Figure 8.5 Reversibility test.

As described before the processed time series were fitted into the Filtered Hurst Kolmogorov model through both the climacogram (8.6) and the climacospectrum (8.7). The results were assumed satisfactory. The parameters calculated were: $a = 19.399, H = 0.628, M = 0.724$

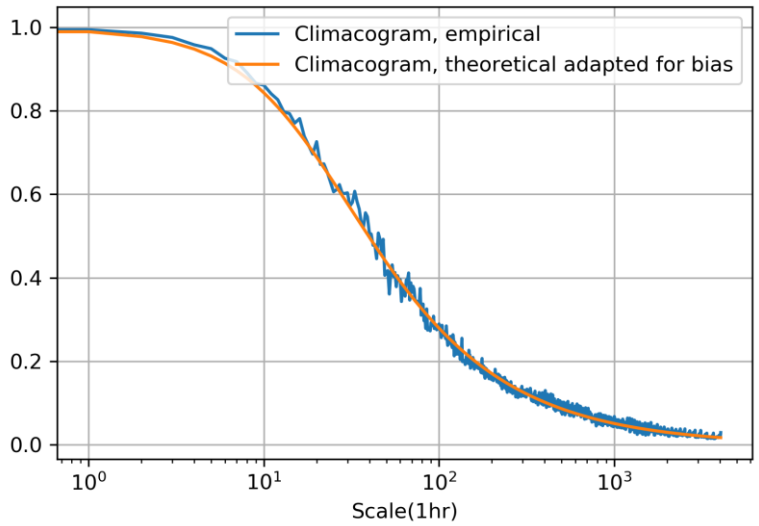


Figure 8.6 FHK Climacogram data fit.

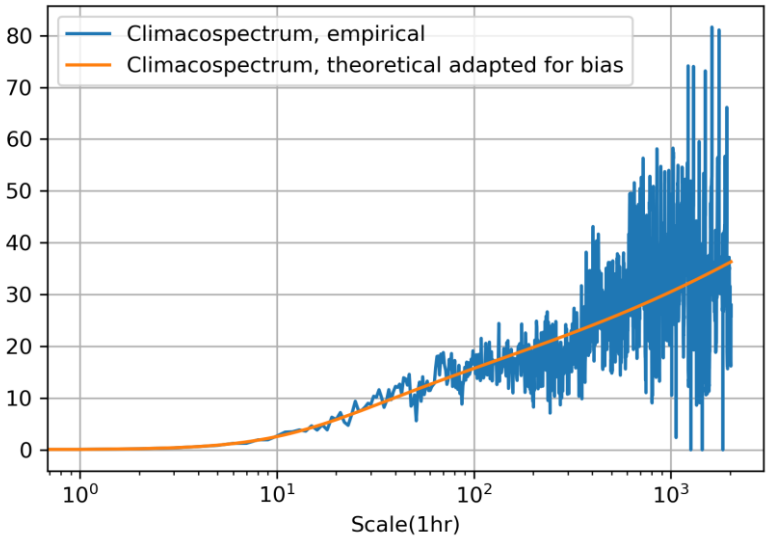


Figure 8.7 FHK Climacospectrum data fit.

In the following figure the discrete power spectrum is shown (8.8). However it is drawn in a continuous line.

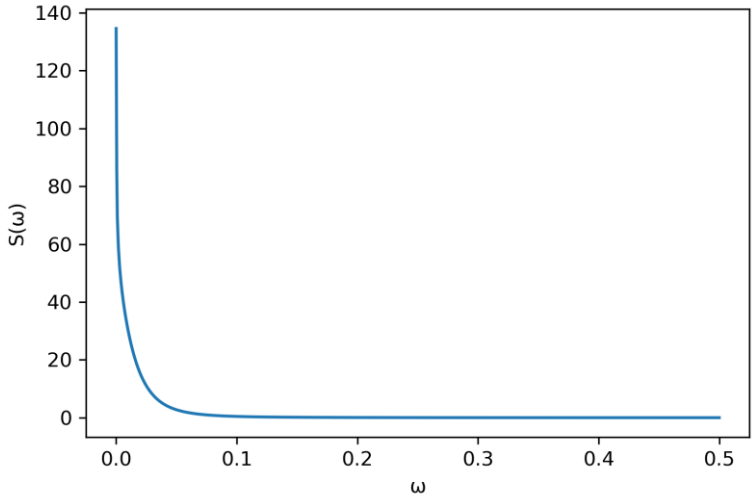


Figure 8.8 Descrite Power Spectrum

In figure 8.9 we can see the a_η sequence for the first case of conserving the reversibility at only one scale.

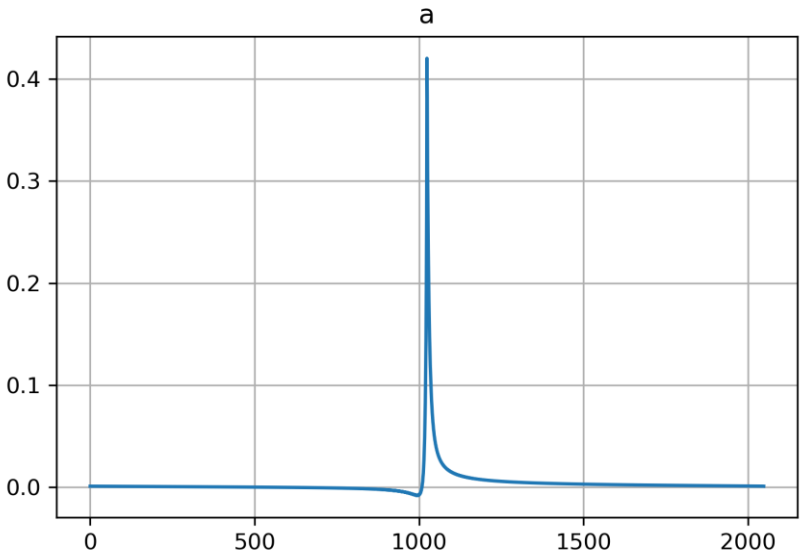


Figure 8.9 a_η sequence results for the first case with constant θ.

A stated before $\theta(\omega)$ which is defined as the smooth minimum of two hyperbolic functions of frequency, after optimization was used the following sequence was found. In the plot (8.10) it seen as a line but in reality there are 1024 $\theta(\omega)$parameters.

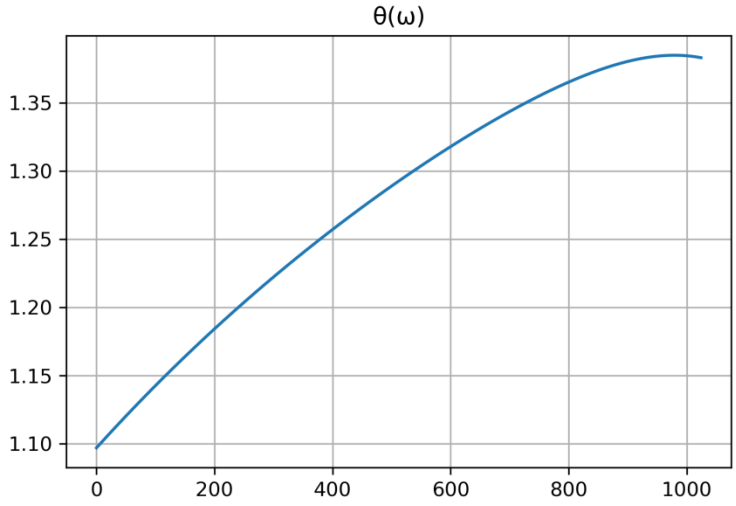


Figure 8.10 $\boldsymbol{\theta}$ sequence for the second case of conserving reversibility at both scales.

After using the $\theta(\omega)$ parameters, the the a_η sequence for the second case is found (8.11). In this case for conserving the reversibility at both scales.

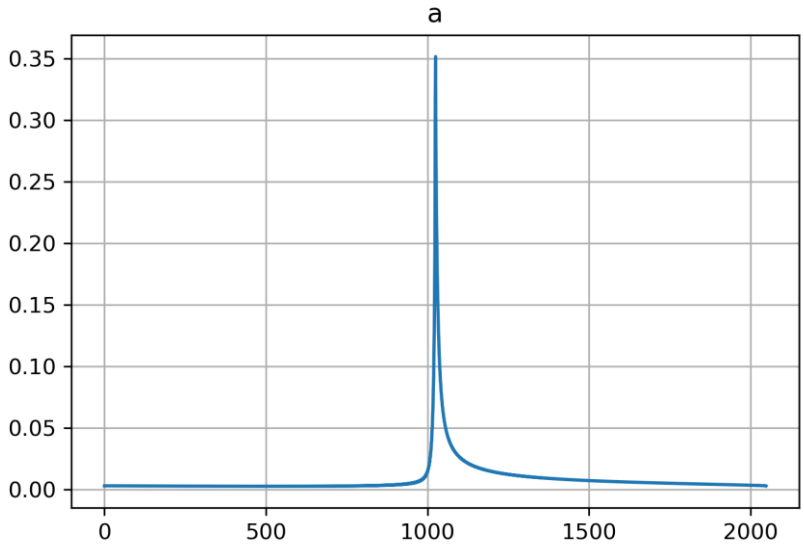


Figure 8.11 $\boldsymbol{a_\eta}$ sequence results for the second case with varying $\boldsymbol{\theta}$.

In the end we have the results from 100 simulations of 10000 in length for the first (8.12) and the second case (8.13) respectively. The red dots indicate the reversibility that was aimed to be conserved. It is shown that the irreversibility targets are achieved. It is also observed that the first method cannot achieve the second scale target efficiently as it was expected. However it is significantly close to it.

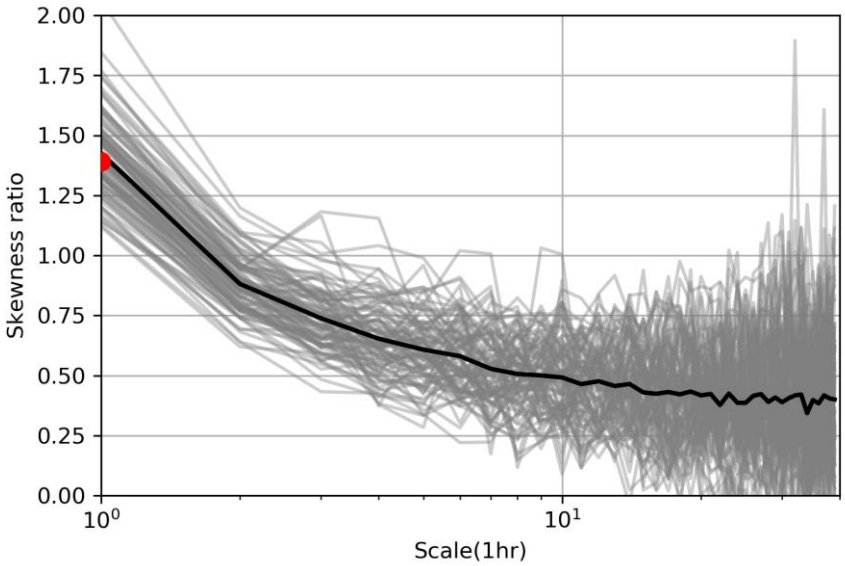


Figure 8.12 100 simulations with 10000 length, conserving the reversibility only at the first scale.

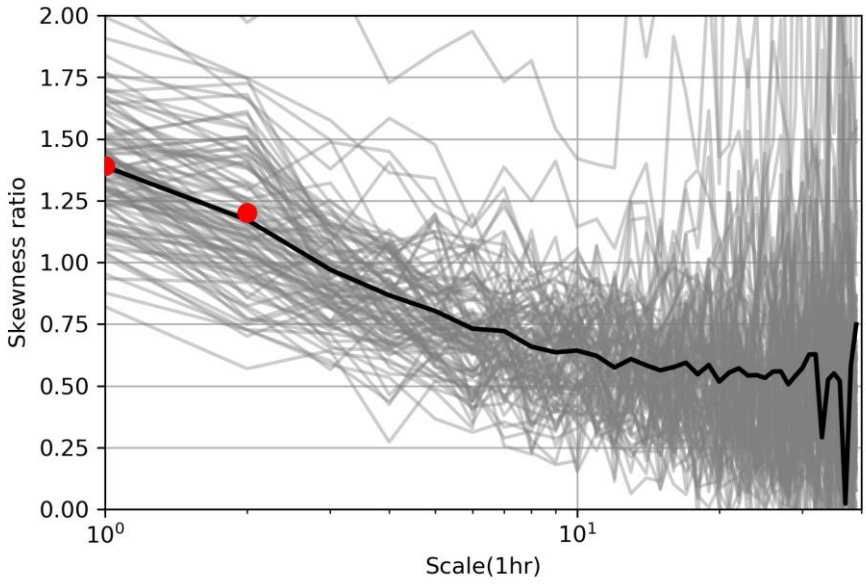


Figure 8.13 100 simulations with 10000 length, conserving the reversibility at both first and second scale.

8.7 Conclusions

The results seem to be generally satisfying. The optimization was successful for both the first and the second case finding adequate θ parameters.

More specifically for the first case, where the reversibility has to be conserved only for the first scale, the average of the simulations is very close to the sample skewness ratio. This case study further verifies the effectiveness of the original algorithm and model. The original algorithm actually tends to conserve also the second scale (through the power spectrum of the process). That means that the new modification acts as a small adjustment.

In the second case again there is high precision with only 100 simulations. The average of the simulations is very close to the sample skewness ratio at both scales this time.

A higher skewness ratio means that more simulations are needed to achieve the same result.

At last the modification affects the irreversibility at even greater scales implicitly. Furthermore the same method could be used to conserve the reversibility at even greater scales with high precision.

9. Irreversibility investigation from the USGS Database

9.1 Introduction

In this section there is an attempt to quantify the irreversibility at the first 100 scales from a large sample of streamflow data. The aim is to study a lot of stations and find the average value of skewness ratio for each scale. The first case is the state of Maryland consisting of 222 stations (9.1). The second case is an even bigger dataset consisting 762 stations around the USA (9.2). At first the station number was higher (802) but some stations were excluded due to data management criteria. The criterion for missing values was 10%. In the first case the data recording period was 2013-2018 or less (mostly 5 year in length). In the second case the data recording period was 1900-2019 or less (mostly 20-30 years in length).

9.2 Methodology

The methodology is quite similar to the one for the Monocacy River reversibility test. The dataset was downloaded with the exact wanted timeframe using the climate package. The data was at first 15 minute measurements but was aggregated to 1 hour scale – relevant scale for hydrosystem management. After that, stationarization of the time series took place. Specifically the effect of the annual cycle was “removed” by multiplying the discharge values by 12 different coefficients, one per month, summing up to 1. These coefficients were found by minimizing the total variance of the transformed time. For optimization the scipy.optimize.minimize function was used. A python function was developed to perform the above task. The function uses as an input the station “code number” and gives back the time series aggregated at hourly scale and stationarized.

The next step was to develop a Python function to detect reversibility scalewise. It aggregates the data until scale 100 is reached and calculates the sample skewness of the differenced and the original process. For each station the data were saved in a matrix and after the average and the variance for each scale were calculated, the plots were produced.

9.3 Results

The results are presented in the form of plots. The skewness ratio of the first scale for the two cases is 2.42 and 2.51 respectively. The skewness ratios of the second scale for the two cases are 1.7 and 1.9 respectively. The variance at the first scale of the second case is 43.58 and for the second 7.65 (9.3).

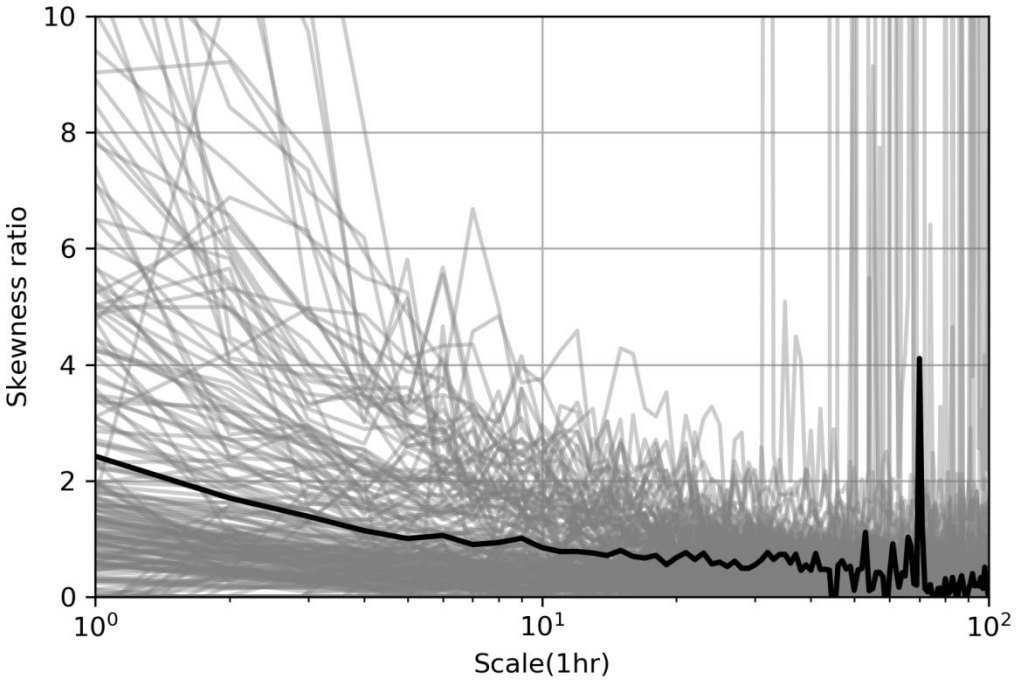


Figure 9.1 Maryland, 222 stations skewness ratio.

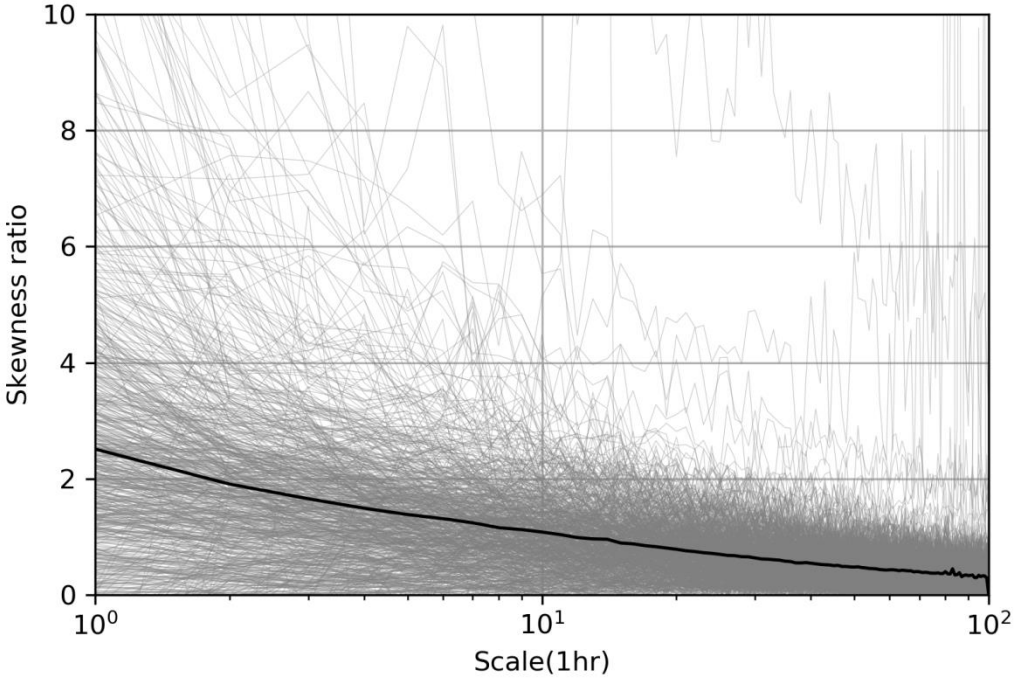


Figure 9.2 Skewness ratio from 762 stations around USA.

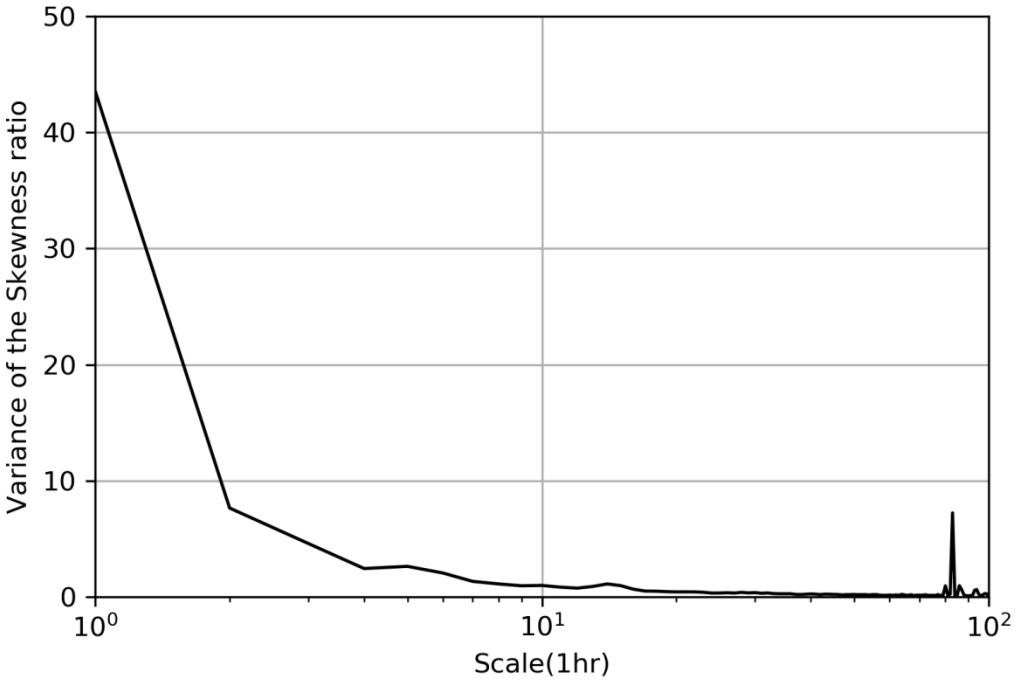


Figure 9.3 Variance of the skewness ratio of the second case.

9.4 Conclusions

The conclusion from the above is that the reversibility of streamflow in the USA at least, has expected value of skewness ratio at the first scale around 2.5 and at the second scale around 1.9. However this result has a very high variance at the first scale that tends to get smaller at as the scales get higher.

10. Conclusions and Future Research

Uncertainty is a major factor in physical sciences and engineering. The main purpose of uncertainty analysis is to quantify the uncertainty by estimating statistical properties of the system outputs that are affected by the stochastic nature of the natural process or sensitivity in the initial conditions.

It is essential to study he probabilistic behavior of an engineering system is essential considering that uncertainty issues are important and must be managed. The true distribution for the system response subject to parameter uncertainty is a lot of times difficult or even impossible sometimes to infer. This is due to the complexity of the hydrosystems. In such cases, Monte Carlo simulation is a very useful tool to provide numerical estimations of the stochastic features of the system response.

Time's arrow has an important role in science and is related to randomness and uncertainty. It has been implemented in stochastics for some time and it has recently attracted attention in hydrological relevant publications. Time asymmetry in stochastic processes is synonymous with steeper ascending parts and gradual descending parts in a realization. The same happens in a hydrograph where irreversibility is manifested by the steeper rise of the climbing limb in contrast to the falling limb. We try to reproduce this behavior through the concept of time asymmetry. Some studies have found that the irreversibility of streamflow is marked for scales of several days and this highlights the need to reproduce it in flood simulations.

In this study real world streamflow data from a large database i.e. the USGS database are used to investigate the irreversibility of hourly scale streamflow time series at scales up to one hundred hours. The aim is estimate the temporal asymmetry of streamflow data at fine timescales in order to assess the importance of taking it into account in its modelling. The reversibility of streamflow in the USA, has expected value of skewness ratio at the first scale around 2.5 and at the second scale around 1.9. However this result has a very high variance at the first scale that tends to get smaller at as the scales get higher.

This study proposes a modification to the existing method by Koutsoyiannis (2019) that conserves irreversibility only at the first scale and makes it capable of preserving the irreversibility simultaneously at the first and second scale. To test the method, we use real world data. The results verify the method successfully.

The modification affects the irreversibility at even greater scales implicitly. Furthermore the same method could be used to conserve the reversibility at even greater scales.

Temporal asymmetry investigation from large databases around the world should be a topic for future research. Other research topics could be the connection of temporal asymmetry with conceptual characteristics of the basin e.g. the surface area. Downstream stations are suspected to have higher irreversibility than the upstream ones and this could also be studied. Parametric equations could be made for further irreversibility conservation. At last other odd functions could be used as the $\theta(\omega)$ function to conserve irreversibility at greater scales implicitly.

REFERENCES

- Everitt, BS. And Skrondal, A., 2010 The Cambridge Dictionary of Statistics (4th edn). Cambridge University Press: Cambridge, U.K.
- Box, G. E. P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2016, Time Series Analysis: Forecasting and Control, John Wiley & Sons, Inc., Hoboken, New Jersey.
- Chow, Ven Te, Maidment, 1988 D.R., and Mays, L.W. Applied Hydrology. New York: McGraw-Hill.
- Dimitriadis, P. and Koutsoyiannis, D., 2015. Climacogram versus autocovariance and power spectrum in stochastic modelling for Markovian and Hurst–Kolmogorov processes. Stochastic Environmental Research and Risk Assessment, 29 (6), 1649–1669. doi:10.1007/s00477-015- 1023-7.
- Eddington, A., 1928. The nature of the physical world. Cambridge, UK: Cambridge University Press.
- Kolmogorov, A. N., 1933, Grundbegriffe der Wahrscheinlichkeitsrechnung, translation, 1950, Foundations of the Theory of Probability, 2nd edition, 84 pages, Chelsea Publishing Company, New York
- Koutsoyiannis, D., 2005, Hydrologic persistence and the Hurst phenomenon, Water Encyclopedia, Vol. 4, Surface and Agricultural Water, edited by J. H. Lehr and J. Keeley, 210–221, doi:10.1002/047147844X.sw434, Wiley, New York.
- Koutsoyiannis, D., 2017 Entropy production in stochastics. Entropy 19(11):581
- Koutsoyiannis, D., 2019 Simple stochastic simulation of time irreversible and reversible processes, Hydrological Sciences Journal, doi: 10.1080/02626667.2019.1705302
- Koutsoyiannis, D., 2016, Generic and parsimonious stochastic modelling for hydrology and beyond, Hydrological Sciences Journal, 61:2, 225-244, doi: 10.1080/02626667.2015.1016950
- Koutsoyiannis, D. and Xanthopoulos, Th., 2014 Engineering Hydrology (in Greek), National Technical University, Athens.
- Koutsoyiannis, D., 1997, Statistical Hydrology, 4th edition, in Greek, 312 pages, National Technical University of Athens, Athens
- Koutsoyiannis, D., 2000. A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. Water Resources Research, 36 (6), 1519–1533. doi:10.1029/2000WR900044
- Koutsoyiannis, D., 2002. The Hurst phenomenon and fractional Gaussian noise made easy. Hydrological Sciences Journal, 47 (4), 573–595. doi:10.1080/02626660209492961
- Koutsoyiannis, D., 2010. HESS Opinions: “a random walk on water”. Hydrology and Earth System Sciences, 14, 585–601. doi:10.5194/hess-14-585-2010

- Koutsoyiannis, D., 2011. Hurst-Kolmogorov dynamics as a result of extremal entropy production. Physica A: Statistical Mechanics and its Applications, 390 (8), 1424–1432. doi:10.1016/j.physa.2010.12.035
- Koutsoyiannis, D., 2019, Time's arrow in stochastic characterization and simulation of atmospheric and hydrological processes, Hydrolog. Sci. J., 64, 1013–1037, https://doi.org/10.1080/02626667.2019.1600700.
- Koutsoyiannis, D., 2019. Knowable moments for high-order stochastic characterization and modelling of hydrological processes. Hydrological Sciences Journal, 64 (1), 19–33
- Koutsoyiannis, D., and Economou A., 2003, Evaluation of the parameterization-simulation-optimization approach for the control of reservoir systems, Water Resour. Res., 39(6), 1170, doi:10.1029/2003WR002148.
- Kroese, D. P., Brereton T. , Taimre T. , and Botev Z. I., 2014 "Why the Monte Carlo method is so important today," Wiley Interdisciplinary Rev.: Computat. Statist., vol. 6, no. 6, pp. 386–392
- "Kuhlman, D., 2009 A Python Book: Beginning Python, Advanced Python, and Python Exercises. Dave Kuhlman."
- Layzer, D., 1975, The arrow of Time, Scientific. American., 233(6), 56-69.
- Mandelbrot, B. B. and van Ness J. W., 1968, Fractional Brownian Motion, fractional noises and applications, SIAM Review, 10, 422–437.
- Mathai, J., Mujumdar, PP., 2019. Multisite Daily Streamflow Simulation with Time Irreversibility. Water Resources Research, 55,9334-9350, https://doi.org/10.1029/2019WR025058
- Mays, L., and Tung, Y.-K., 1992. Hydrosystems engineering and management, McGraw-Hill, New York.
- Müller, T., Schütze, M., and Bárdossy, A., 2017. Temporal asymmetry in precipitation time series and its influence on flow simulations in combined sewer systems. Advances in Water Resources, 107, 56–64. doi:10.1016/j.advwatres.2017.06.010
- N. Metropolis., 1987, The beginning of the Monte Carlo method. Los Alamos Science, 15:125–130
- Papoulis, A., 1991, Probability, random variables and stochastic processes. 3rd ed. New York: McGraw-Hill.
- Psaradakis, Z., 2008. Assessing time-reversibility under minimal assumptions. Journal of Time Series Analysis, 29 (5), 881–905. doi:10.1111/ j.1467-9892.2008.00587.x
- Ross, S., 2014, Introduction to probability models. Academic Press.
- Serinaldi, F.; Kilsby, C.G., 2016 Irreversibility and complex network behavior of stream flow fluctuations. Phys. A Stat. Mech. Appl., 450, 585–600
- Shannon, C. E., 1948, A Mathematical Theory of Communication, Mobile Computing and Communications Review, 5(1), 3-55
- Tung, Y.-K. and Yen, B.-C., 2005 Hydrosystems Engineering Uncertainty Analysis.

New York: McGraw-Hill.

- Weiss, G., 1975. Time-reversibility of linear stochastic processes. Journal of Applied Probability, 12 (4), 831–836. doi:10.2307/3212735
- Κουτσογιάννης, Δ., 2013, *Σημειώσεις Στοχαστικών Μεθόδων στους Υδατικούς Πόρους*, Εκδοση 4, 100 pages, Εθνικό Μετσόβιο Πολυτεχνείο, Αθήνα.
- Μακρόπουλος, Χ. και Ευστρατιάδης, Α., 2018, Σημειώσεις μαθήματος "Βελτιστοποίηση Συστημάτων Υδατικών Πόρων – Υδροπληροφορική", 151 pages, Τομέας Υδατικών Πόρων και Περιβάλλοντος – Εθνικό Μετσόβιο Πολυτεχνείο.
- www.usgs.gov
- www.pypi.org/project/climata/

APENDIX A

Models for AMA (models_for_AMA)

```
def FHK_Cauchy_(n, M, H, a, la):
 import numpy as np
gamma_k=np.zeros((n+2))
 c_h=np.zeros((n+2))
for i in range(0, n+2):
   gamma_k[i]=la * (1 +(i/ a) ** (2 * M)) ** ((H - 1) / M)
 for i in range(0, n+1):
   c_h[i]=((i-1 ) ** 2 * gamma_k[ i-1 ] + (i+1 ) ** 2 * gamma_k[...
...i+1 ]) / 2 - (i) ** 2 *gamma_k[i]
 c_h[0]=gamma_k[1]
 c_h=c_h[:-1]
return c_h

def Markov_(n, a, la):
 import numpy as np
 import math
 #import matplotlib.pyplot as plt
#import pandas as pd
 n=1024
 la=1
 a=20
 gamma_k=np.zeros((n+1))
 c_h=np.zeros(n+1)
#la einai to λ =γ(0)
 gamma_k[0]=la
 for i in range(1, n+1):
   gamma_k[i]=(2 * la) * (a / i) * (1 - (1 -...
... math.exp((-i) / a)) * (a / i))
 c_h[0]=gamma_k[1]
 for i in range(1, n):
   c_h[i]=((i - 1) ** 2 * gamma_k[ i - 1] + (i + 1) ** 2...
... * gamma_k[ i + 1]) / 2 - (i) ** 2 *gamma_k[i]
return c_h, gamma_k
```

APENDIX B

Spectra for AMA (last_try)

```
def DFT_(f,t):
    import numpy as np
    import math
    n=f.shape[0]+1
    df=np.zeros((n,2),dtype=np.single)
    w =np.float16( 2 * math.pi / n)
    for k  in range (0,n - 1):
          for M in range(0, n - 1):
            c = math.cos(k * w * M)
            S = math.sin(k * w * M) * t
            df[k, 0] = df[k, 0] + f[M, 0] * c + f[M, 1] * S
            df[k, 1] = df[k, 1] + f[M, 1] * c - f[M, 0] * S
          if t == 1 :
            df[k, 0] = df[k, 0] / n
            df[k, 1] = df[k, 1] / n
            #print('dft mpike2')
    for k in range( 0, n - 1):
         f[k, 0] = df[k, 0]
         f[k, 1] = df[k, 1]
    return f

def FFT__(f, t):
 #print(t)
 import math
 n=f.shape[0]
 M=math.log(n)/math.log(2)
 for j in range(1 , int(M)+1):
  l = int(n / 2 ** j)
  for k in range( 0 , n ,2 * l):
     w =( math.pi / l)
     for i in range( 0 , l ):
      c =( math.cos(i * w))
      S =(math.sin(i * w) * t)
```

```
      x = k + i
      y = k + i + l
      f[x, 0] = f[x, 0] + f[y, 0]
      #if (i==0) and (j==1):
       #   print (f[0,0])
      f[x, 1] = f[x, 1] + f[y, 1]
      tr = (f[x, 0] - f[y, 0] - f[y, 0])
      ti = (f[x, 1] - f[y, 1] - f[y, 1])
      f[y, 0] = tr * c + ti * S
      f[y, 1] = ti * c - tr * S

 f=Bit_Reverse (f, n)

 if abs(t-1)<0.1 :
    for i in range (0, (n )):
        f[i, 0] = f[i, 0] / n
        f[i, 1] = f[i, 1] / n


 return f
def Bit_Reverse(f,n):
    j = 0
    for i in range (0, n - 1):
          if i < j :
             t=f[i, 0] #swaping
             f[i, 0] =f[j, 0]
             f[j, 0]=t

             t=f[i, 1] #swaping
             f[i, 1] =f[j, 1]
             f[j, 1]=t

          k = n / 2
          while k<(j+1) :
              j = j - k
              k = k / 2
          j = j + k
```

```
            j=int(j)

    return f
def FFTPowSpec_(A):
    import numpy as np
    # n=2**n+1
    #n=1025 KANONIKA
    n=1025

    f=np.zeros((2*n-3+1,2))
    df=np.zeros((n-1+1,2))
    for i in range( 0 , n ):
      f[i,0] = 4 * (n - 1) * A[i ]

    for i in range( n , 2 * n - 2):
      f[i,0 ] = f[2 * n - i - 2,0]

    f=FFT__ (f, 1)

    for i in range(0 , n ):
      df[i, 0] = (i) / 2 / (n-1)
      df[i, 1] = f[i,0]
      #if df[i,1]<0:
          #print(df[i,1])
          #import sys
          #sys.exit("Error message")

    return df
        df[i + n - 1-1, 1] = f[2 * (n - 2) + 2 - i-1, 0]#to allaksa

    def FFT_AMACoef_(spectrum, theta):
    import last_try
  #  import pandas as pd
    import numpy as np
    #theta=0.0625
    #spectrum=pd.read_excel...
...(r'C:\Users\USER\Desktop\gia_python_paradeigma9.xlsx')
```

```
    #spectrum=pd.DataFrame.to_numpy(spectrum)
    #spectrum=spectrum[:,1]
    n=spectrum.shape[0]
    th=np.zeros((n))
    th[1:n]=theta
    th[0]=0
    #import numpy as np
    import math

    #n=spectrum.shape[0]
    ss=np.zeros((n))
    for i in range( 0 , n ):
        ss[i] = (2 * spectrum[i ]) ** 0.5 * …
…math.cos(2 * math.pi * th[i]) / (4 * (n-1 ))
    f=np.zeros((2*n-2,2))
    for i in range( 0 , n ):
        f[i, 0] = ss[i]
        f[i, 1] = ss[i] * math.tan(2 * math.pi * th[i])
    f[n - 1, 1] = 0
    for i in range ( n, 2 * n - 2):
        f[i, 0] = f[2 * n - i - 2, 0]
        f[i, 1] = -f[2 * n - i - 2, 1]
    f=last_try.FFT__(f,-1)
    df=np.zeros((2 * (n - 1) + 1, 2))
    for i in range(0 , n ):
      #print(i + n - 1, 2 * (n - 2) + 2 - i)
      df[i, 0] = i - n + 1
      df[i + n - 1, 0] = i
      df[i, 1] = f[n - 1 - i, 0]
      if i == 0 :
        df[i + n - 1, 1] = f[n - 1, 0]#to allaksa
      else:
        df[i + n - 1, 1] = f[2 * (n - 2) + 2 - i, 0]#to allaksa
    return df
def FFT_AMACoef_var_theta(spectrum, theta):
    import last_try
  #  import pandas as pd
```

```
    import numpy as np
    #theta=0.0625
    #spectrum=                                          pd.read_excel
(r'C:\Users\USER\Desktop\gia_python_paradeigma9.xlsx')
    #spectrum=pd.DataFrame.to_numpy(spectrum)
    n=spectrum.shape[0]
    th=theta
    #import numpy as np
    import math
    #n=spectrum.shape[0]
    ss=np.zeros((n))
    for i in range( 0 , n ):
        ss[i] = (2 * spectrum[i ]) ** 0.5 * math.cos(2 * math.pi *
th[i]) / (4 * (n-1 ))
    f=np.zeros((2*n-2,2))
    for i in range( 0 , n ):
        f[i, 0] = ss[i]
        f[i, 1] = ss[i] * math.tan(2 * math.pi * th[i])
    f[n - 1, 1] = 0
    for i in range ( n, 2 * n - 2):
        f[i, 0] = f[2 * n - i - 2, 0]
        f[i, 1] = -f[2 * n - i - 2, 1]
    f=last_try.FFT__(f,-1)
    df=np.zeros((2 * (n - 1) + 1, 2))
    for i in range(0 , n ):
      #print(i + n - 1, 2 * (n - 2) + 2 - i)
      df[i, 0] = i - n + 1
      df[i + n - 1, 0] = i
      df[i, 1] = f[n - 1 - i, 0]
      if i == 0 :
        df[i + n - 1, 1] = f[n - 1, 0]#to allaksa
      else:
        df[i + n - 1, 1] = f[2 * (n - 2) + 2 - i, 0]#to allaksa
    return df
```

APENDIX C

Scripts for USGS Database study

Script: Reversibility_test_2

```
import numpy as np
import math
import pandas as pd
import scipy.stats
import matplotlib.pyplot as plt
def aggr(n,A):

    if A.shape[0]>n:
        B=np.zeros(np.int(A.shape[0]/n))
        B[0]=np.average(A[0:n])
        for i in range(1,B.shape[0]):
          B[i]=np.average(A[i*n:(i+1)*n])
        return B

def dif_proc_(A):
        dif_proc= np.empty((A.shape[0]-1,1))
        for i in range(0,A.shape[0]-1):
         dif_proc[i]=A[i+1]-A[i]
        return dif_proc

def reversibility_test(A,name):

    max_=100
    intervals=100
    k=np.linspace(1,max_,intervals,dtype=int)
    l=np.zeros((k.shape[0]+1,1))
    for i in range(0,(k.shape[0])):
     q1=scipy.stats.skew(dif_proc_(aggr(k[i],A)),bias=False,
nan_policy='omit')
     q2=scipy.stats.skew(aggr(k[i],A),bias=False, nan_policy='omit')
     l[i+1]=q1/q2
    l[0]=0
```

```
    plt.plot(l)
    plt.xscale('log')
    plt.title(name)
    plt.xlabel('Scale(1hr)')
    plt.xlim(1, 200)
    plt.ylabel('Skewness ratio')
    plt.grid(True)
    #plt.show()
    plt.savefig(name,dpi=300)
    print(l[1],l[2])
    plt.close()
    import winsound
    duration = 1000  # milliseconds
    freq = 440  # Hz
    winsound.Beep(freq, duration)
```

def reversibility_test_no_plot(A,name):

```
    max_=100
    intervals=100
    k=np.linspace(1,max_,intervals,dtype=int)
    l=np.zeros((k.shape[0]+1))
    for i in range(0,(k.shape[0])):
     q1=scipy.stats.skew(dif_proc_(aggr(k[i],A)),bias=False,
nan_policy='omit')
     q2=scipy.stats.skew(aggr(k[i],A),bias=False, nan_policy='omit')
     l[i+1]=q1/q2
    l[0]=0
    return l
```

Script:Rev_test_generator

```
#import sys
#sys.path.append(r'C:\Users\guest13\Desktop\python_path')
import pandas as pd
import numpy as np
from stationarize_new import import_and_stationarise
#import matplotlib.pyplot as plt
```

```
from climata.usgs import InstantValueIO,DailyValueIO
from scipy.optimize import minimize



#from reversibility_test_2 import reversibility_test
from reversibility_test_2 import reversibility_test_no_plot
#xronoseira                          =                      pd.read_excel
(r'C:\Users\USER\Desktop\gia_python_paradeigma4.xlsx')
#xronoseira['values']    =    xronoseira['values'].apply(lambda    x:
'{0:0>8}'.format(x))
#xronoseira=pd.read_excel()
#xronoseira=np.loadtxt('C:\\Users\\USER\\Desktop\\zeros.txt',dtype=s
tr)
df = pd.read_csv('station_ids_final.txt', dtype = str,header=None)
A=pd.DataFrame.to_numpy(df ,dtype=str)
#matr=np.zeros((7951,101))
matr=np.load(r'C:\Users\USER\Desktop\python_path\matr.npy')
k=np.load(r'C:\Users\USER\Desktop\python_path\k.npy')
for i in range(k,0,-1):
    try:
        station_id = A[i,0]
        #station_id="03076500"
        a=import_and_stationarise(station_id)
        #reversibility_test(a[0],a[1])
        matr[i,:]=reversibility_test_no_plot(a[0],a[1])
        print(i)
        k=i
        np.save('k',k)
        np.save('matr',matr)
    except UnboundLocalError:
        pass
    except ConnectionError:
        np.save('matr',matr)

    except TypeError:
        pass
```

Script:stationarize_new

```
import pandas as pd
def import_and_stationarise(station_id):

    from climata.usgs import InstantValueIO,DailyValueIO
    import numpy as np
    from scipy.optimize import minimize

    #print(station_id)
    # set parameters
    #station_id = "01496200"
    #station_id = "01484719"
    param_id = "00060"

    datelist   =...   pd.date_range(start='1900-10-01',end='2019-12-
12',freq='15T').tolist()
    data = InstantValueIO(start_date=datelist[0],end_date=datelist[-
1],station=station_id,parameter=param_id,)
    # create lists of date-flow values
    flow_=np.empty((1000000,1))
    for series in data:
        flow_ = [r[1] for r in series.data]
        dates = [r[0] for r in series.data]
    flow_=np.asarray(flow_)
    #plt.plot(dates, flow)
    #plt.xlabel('Date')
    #plt.ylabel('Streamflow')
    #plt.title(series.site_name)
    #plt.xticks(rotation='vertical')
    #plt.show()
    name=series.site_name
    n=len(flow_)

    cn=0
    for i in range (0,flow_.shape[0]):
        if flow_[i]<0.5:
            cn=cn+1
        if flow_[i]<0.0:
```

```
            flow_[i]=np.NaN
#
np.savetxt(str(station_id)+'_'+str(name),(flow_,dates),delimiter="
",header=str(station_id)+'  '+str(name)+'  '+'Missing  Values'+'  '+
str(cn/flow_.shape[0])+'%')
    flow_=pd.Series(flow_)
    flow_.index=dates
    flow_.to_csv(str(station_id)+'_'+str(name)+'.txt')
#
np.savetxt(str(station_id)+'_'+str(name),(flow_,dates),delimiter="
")


    if (cn/flow_.shape[0])>0.1:
        print('too few')
        return
    month_=np.zeros(n,dtype=np.int8)
    for i in range (0, n):
        month_[i]=np.int(dates[i].month)
    del data,dates,datelist
    flow1_=np.empty((n,1))
    flow1_[:,0]=0.0283168466*flow_

    const=np.empty(12)
    const[:]=0.08
    flow2=np.empty((n,1))

    def constraints1(const) :
      return sum(const)-1
```

Script:usgs_stationarise

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from climata.usgs import InstantValueIO,DailyValueIO
from scipy.optimize import minimize
```

def import_and_save(station_id):

```
    param_id = "00060"


    datelist     =     pd.date_range(start='2013-10-01',end='2018-10-
1',freq='H').tolist()
    data = InstantValueIO(start_date=datelist[0],end_date=datelist[-
1],station=station_id,parameter=param_id,)
    # create lists of date-flow values
    flow_=np.empty((360000,1))
    for series in data:
        flow_ = [r[1] for r in series.data]
        dates = [r[0] for r in series.data]
    flow_=np.asarray(flow_)


    name=series.site_name



    np.savetxt(str(station_id)+'_'+str(name),flow_,delimiter="
",header=str(station_id)+' '+str(name))
    cn=0
    for i in range (0,flow_.shape[0]):
        if flow_[i]<0.5:
            cn=cn+1
    if (cn/flow_shape[0])>0.1:
        return
    return flow_
def stationarise(station_id):


    param_id = "00060"


    datelist     =     pd.date_range(start='2013-10-01',end='2018-10-
1',freq='H').tolist()
    data = InstantValueIO(start_date=datelist[0],end_date=datelist[-
1],station=station_id,parameter=param_id,)
    # create lists of date-flow values
    flow_=np.empty((360000,1))
    for series in data:
        flow_ = [r[1] for r in series.data]
        dates = [r[0] for r in series.data]
```

```
    flow_=np.asarray(flow_)

    name=series.site_name
    n=len(flow_)
    month_=np.zeros(n,dtype=np.int8)
    for i in range (0, n):
        month_[i]=np.int(dates[i].month)
    del data,dates,datelist
    flow1_=np.empty((n,1))
    flow1_[:,0]=0.0283168466*flow_

    const=np.empty(12)
    const[:]=0.08
    flow2=np.empty((n,1))

    def constraints1(const) :
      return sum(const)-1

    cons = [{'type':'eq', 'fun': constraints1}]
    bnds = ((0, None), (0, None),(0, None), (0, None),(0, None), (0, None),(0, None), (0, None),(0, None), (0, None))

    def var_(const):

     flow2[:,0]=flow1_[:,0]*const[month_[:]-1]
     var_=np.std(flow2)
     ret=abs(var_)
     print (ret)
     return ret

    rez=minimize(var_,const,constraints=cons,bounds=bnds)
    const=rez['x']
    flow2[:,0]=flow1_[:,0]*const[month_[:]-1]

    def aggr_sum(n,A):
     if A.shape[0]>n:
```

```
        B=np.zeros(np.int(A.shape[0]/n))
        B[0]=np.sum(A[0:n])
        for i in range(1,B.shape[0]):
          B[i]=np.sum(A[i*n:(i+1)*n])
        return B
    flow_1hr=aggr_sum(4,flow2)
    return flow_1hr,name
```

Script: maryland_skewness_ratio

```
import numpy as np

import matplotlib.pyplot as plt
fig=np.zeros((101))
for i in range(0,100):
 fig[i]=np.mean(matr[:,i])
for i in range (0,222):
 plt.plot(matr[i,0:101],'grey',linewidth=1.5,alpha=0.4)

plt.plot(fig[0:101],'black',linewidth=2)
plt.ylim(0, 10)
plt.xlim(1, 100)
plt.xscale('log')
#plt.title('Simulations using AMA at scale 1hr and 2hr')
plt.xlabel('Scale(1hr)')
plt.ylabel('Skewness ratio')
plt.grid(True)
#plt.show()
plt.savefig('Maryland_sk',dpi=300)
plt.close()
```

Script: import_usgs_and_normalise

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from climata.usgs import InstantValueIO,DailyValueIO
from scipy.optimize import minimize
```

```
# set parameters
station_id = "01639000"
param_id = "00060"

datelist = pd.date_range(start='2013-10-01',end='2018-10-1',freq='15T').tolist()
data = InstantValueIO(start_date=datelist[0],end_date=datelist[-1],station=station_id,parameter=param_id,)
# create lists of date-flow values
flow_=np.empty((360000,1))
for series in data:
    flow_ = [r[1] for r in series.data]
    dates = [r[0] for r in series.data]
flow_=np.asarray(flow_)
#plt.plot(dates, flow)
#plt.xlabel('Date')
#plt.ylabel('Streamflow')
#plt.title(series.site_name)
#plt.xticks(rotation='vertical')
#plt.show()
n=len(flow_)
n1=4
n2=np.int(n/n1)
month_=np.zeros(n2,dtype=np.int8)
for i in range (0, n2):
    month_[i]=np.int(dates[i*n1].month)
del data,dates,datelist
flow1_=np.empty(n)
flow1_[:]=0.0283168466*flow_

def aggr_sum(n,A):
 if A.shape[0]>n:
    B=np.zeros(np.int(A.shape[0]/n))
    B[0]=np.sum(A[0:n])
    for i in range(1,B.shape[0]):
      B[i]=np.sum(A[i*n:(i+1)*n])
    return B
```

```
flow3=np.empty((n2,1))
flow3=aggr_sum(4,flow1_)

d = {'col1': (flow3), 'col2': (month_)}
df_=pd.DataFrame(index=None,data=d,dtype=float)

mean_=df_.groupby(['col2']).mean()
mean_1=pd.DataFrame.to_numpy(mean_)
std_=df_.groupby(['col2']).std()
std_1=pd.DataFrame.to_numpy(std_)
flow2_=np.empty((n2,1))
flow4=np.empty((n2,1))
flow2_[:,0]=flow3[:]

flow4[:,0]=(flow2_[:,0]-mean_1[month_[:]-1,0])/std_1[month_[:]-1,0]
var_epal=np.var(flow4)
mean_epal=np.mean(flow4)
```

APENDIX D

Scripts about parameter estimation

Script:theta_parameter

```
def theta_parameter(spectrum,A):#spectrum momo h mia sthlh, A h xronoseira pou tha vrethei to skewness ratio
    X1=spectrum

    def rev_parameter(A):

        sh=A.shape[0]
        skewA=(scipy.stats.skew(A[0:sh-1],bias=False, nan_policy='omit')) # unbiased, omit agnoei ta NaN
        skewAdif=(scipy.stats.skew((A[0:sh-1]-A[1:sh]),bias=False, nan_policy='omit'))
        rev_param=skewAdif/skewA
        return rev_param

    rev_param=rev_parameter(A)
    def theta(theta):

      X=AMA_coeffiecient.FFT_AMACoef_(X1,theta)
      X=X[:,1]
      X_dif=X[1:(X.shape[0])]-X[0:X.shape[0]-1]
      orig_var=sum(X**2)
      dif_var=sum(X_dif**2)+X[0]**2
      orig_m3=sum(X**3)
      dif_m3=sum(X_dif**3)+X[0]**3
      orig_skew=orig_m3/orig_var**1.5
      dif_skew=dif_m3/dif_var**1.5
      sk_ratio=dif_skew/orig_skew

      print(theta,(sk_ratio-rev_param))
      return abs(sk_ratio-rev_param)
    #    M=HHK_parameter...
```

```
    #    dif=AMA(theta,H,M)
    #    return dif

    res = minimize(theta, 0.1,bounds=((0,0.251),))
    return res['x']
```

Script:scale_1k_reversibility

```
import numpy as np
import math
import pandas as pd
import AMA_coeffiecient
import matplotlib.pyplot as plt
import AMA_coeffiecient
import scipy.stats
from scipy.optimize import minimize
import last_try

Xw=pd.DataFrame.to_numpy(pd.read_excel
(r'C:\Users\USER\Desktop\gia_python_paradeigma8.xlsx') )
Xp=pd.DataFrame.to_numpy(pd.read_excel
(r'C:\Users\USER\Desktop\gia_python_paradeigma7.xlsx'))

#Xw=np.zeros((g.shape[0],1))
#Xp=np.zeros((g.shape[0],1))
Xw[:,0]=p[:,0]
Xp[:,0]=p[:,1]

def theoretic_sk_ratio_(X):
  X_dif=X[1:(X.shape[0])]-X[0:X.shape[0]-1]
  orig_var=np.nansum(X**2)
  dif_var=np.nansum(X_dif**2)+X[0]**2
  orig_m3=np.nansum(X**3)
  dif_m3=np.nansum(X_dif**3)+X[0]**3
  orig_skew=orig_m3/orig_var**1.5
  dif_skew=dif_m3/dif_var**1.5
  theoretic_sk_ratio=dif_skew/orig_skew
  return theoretic_sk_ratio
```

```
bnds = ((None))
#theta_m=np.ones((3,3))

theta_s=0.1

def theta(theta_s):

      Xa=AMA_coeffiecient.FFT_AMACoef_(Xp, theta_s)
      Xa_=np.zeros((Xa.shape[0],1))
      Xa_[:,0]=Xa[:,1]
      #print(theta,(sk_ratio-rev_param))
      rev_1=1.39

      rev_th_1=theoretic_sk_ratio_(Xa_)

      error=(rev_1-rev_th_1)**2
      print(rev_th_1,error)
      return (error)
    #    M=HHK_parameter...
    #    dif=AMA(theta,H,M)
    #    return dif

res = minimize(theta, theta_s)
print(res['x'])
theta_s=res['x']

Xa=AMA_coeffiecient.FFT_AMACoef_(Xp, theta_s)
plt.plot(Xa[:,1])
plt.title('a')
plt.grid(True)
plt.savefig('af',dpi=300)
#plt.show()

#rez=minimize(var_,const,constraints=cons,bounds=bnds)
#const=rez['x']
```

Script:scale_2k_reversibility

```
import numpy as np
import math
import pandas as pd
import AMA_coeffiecient
import matplotlib.pyplot as plt
import AMA_coeffiecient
import scipy.stats
from scipy.optimize import minimize
import last_try


Xw=pd.DataFrame.to_numpy(pd.read_excel
(r'C:\Users\USER\Desktop\gia_python_paradeigma8.xlsx') )
Xp=pd.DataFrame.to_numpy(pd.read_excel
(r'C:\Users\USER\Desktop\gia_python_paradeigma7.xlsx'))


#Xw=np.zeros((g.shape[0],1))
#Xp=np.zeros((g.shape[0],1))
#Xw[:,0]=g[:,0]
#Xp[:,0]=g[:,1]


def theoretic_sk_ratio_(X):
  X_dif=X[1:(X.shape[0])]-X[0:X.shape[0]-1]
  orig_var=np.nansum(X**2)
  dif_var=np.nansum(X_dif**2)+X[0]**2
  orig_m3=np.nansum(X**3)
  dif_m3=np.nansum(X_dif**3)+X[0]**3
  orig_skew=orig_m3/orig_var**1.5
  dif_skew=dif_m3/dif_var**1.5
  theoretic_sk_ratio=dif_skew/orig_skew
  return theoretic_sk_ratio


def theoretic_sk_ratio_scale_2(X):



  end=X.shape[0]
```

```
  orig_var=X[0]**2/4+np.sum((X[0:(end-1)]+X[1:(end)])**2)/4
  dif_var=X[0]**2/4+(X[0]+X[1])**2/4+(X[1]+X[2]-
X[0])**2/4+np.sum((X[2:end-1]+X[3:end]-X[1:end-2]-X[0:end-3])**2/4)
  orig_m3=X[0]**3/8+np.sum((X[0:(end-1)]+X[1:end])**3/8)
  dif_m3=-X[0]**3/8-(X[0]+X[1])**3/8-(X[1]+X[2]-X[0])**3/8-
np.sum((X[2:end-1]+X[3:end]-X[1:end-2]-X[0:end-3])**3/8)
  dif_skew=dif_m3/dif_var**1.5
  orig_skew=orig_m3/orig_var**1.5
  theoretic_sk_ratio_scale_2=dif_skew/orig_skew
  return theoretic_sk_ratio_scale_2


bnds = ((None, None), (None, None),(None, None), (None, None),(None,
None), (None, None),(None, None), (None, None),(None, None))
#theta_m=np.ones((3,3))
theta_s=np.ones(9)
theta_s[:]=1
#print(theta_m[2,1])
def theta(theta_s):
      end1=Xw.shape[0]
      U1=np.zeros((end1,1))
      U2=np.zeros((end1,1))
     # U3=np.zeros((end1,1))
      #print(theta_m)

U1[:,0]=theta_s[1]*Xw[0:end1,0]/(theta_s[2]+Xw[0:end1,0])+theta_s[0]
      U2[:,0]=theta_s[4]*(0.5-Xw[:,0])/(theta_s[5]+0.5-
Xw[:,0])+theta_s[3]
     # U3[:,0]=theta_s[7]*(Xw[:,0])/(theta_s[8]+Xw[:,0])+theta_s[6]


      th=np.zeros(Xw.shape[0])
      th=np.log(np.exp(-10*U1[:])+np.exp(-10*U2[:]))/(-10)
      Xa=AMA_coeffiecient.FFT_AMACoef_var_theta(Xp, th)
      Xa_=np.zeros((Xa.shape[0],1))
      Xa_[:,0]=Xa[:,1]
      #print(theta,(sk_ratio-rev_param))
      rev_1=1.39
      rev_2=-1.2
      rev_th_1=theoretic_sk_ratio_(Xa_)
```

```
      rev_th_2=theoretic_sk_ratio_scale_2(Xa_)
      error=(rev_1-rev_th_1)**2+(rev_2-rev_th_2)**2
      print(rev_th_1,rev_th_2)
      return (error)
    #    M=HHK_parameter...
    #    dif=AMA(theta,H,M)
    #    return dif


res = minimize(theta, theta_s)
print(res['x'])
theta_s=res['x']
end1=Xw.shape[0]
U1=np.zeros((end1,1))
U2=np.zeros((end1,1))
  #print(theta_m)
U1[:,0]=theta_s[1]*Xw[0:end1,0]/(theta_s[2]+Xw[0:end1,0])+theta_s[0]
U2[:,0]=theta_s[4]*(0.5-Xw[:,0])/(theta_s[5]+0.5-Xw[:,0])+theta_s[3]

th=np.zeros(Xw.shape[0])
th=np.log(np.exp(-10*U1[:])+np.exp(-10*U2[:]))/(-10)

plt.plot(th)
plt.title('θ')
plt.grid(True)
plt.savefig('theta',dpi=300)
#plt.show()
plt.close()

Xa=AMA_coeffiecient.FFT_AMACoef_var_theta(Xp, th)
plt.plot(Xa[:,1])
plt.title('a')
plt.grid(True)
plt.savefig('a, sc=22',dpi=300)
#plt.show()

#rez=minimize(var_,const,constraints=cons,bounds=bnds)
```

```
#const=rez['x']
```

Script:Parameters_final

```
import numpy as np
from scipy.optimize import curve_fit
import matplotlib.pyplot as plt
from scipy.optimize import minimize
import pandas as pd

def HHK(l0,a,M,H):
    #  for i in range(0,Nj):
        #l0=1
        n=A.shape[0]
        gamma_=np.zeros(np.int(A.shape[0]/10))
        for k in range(0,(int(A.shape[0]/10))):
         gamma_[k]=l0*(1+(k/a)**(2*M))**((H-1)/M)
        g_n=l0*(1+(int(A.shape[0])/a)**(2*M))**((H-1)/M)

        return gamma_,g_n,n

def aggr(n,A):

    if A.shape[0]>n:
        B=np.zeros(np.int(A.shape[0]/n))
        B[0]=np.average(A[0:n])
        for i in range(1,B.shape[0]):
          B[i]=np.average(A[i*n:(i+1)*n])
        return B

def Climacospectrum(A):#to A einai o pinakas me to HHK climacogram
       # A=l
        cli_spec=np.zeros((int(A.shape[0]/2),1))
        for i in range(0,int(A.shape[0]/2)):
           cli_spec[i]=(i)*(A[i]-A[2*(i)])/np.log(2)
           if cli_spec[i]<0 :
               cli_spec[i]=0.00001
        cli_spec=cli_spec[1:(int(A.shape[0]/2)+1)]
        return cli_spec
```

```
def Climacogram(A,ints):

        max_=int(A.shape[0]/10)
        intervals=int(A.shape[0]/10)

        k=np.linspace(1,int(max_),intervals,dtype=int)
        l=np.zeros((k.shape[0],1))
        for i in range(0,(int(k.shape[0]))):
         q1=np.nanvar(aggr(k[i],A))
         l[i]=q1

         #print(q1)
        return l

def Climacogram_biased(A,g_n,n): #A is for theoretical climacogram

        max_=int(A.shape[0])
        intervals=int(A.shape[0])
        k=np.linspace(1,int(max_),intervals,dtype=int)
#        l1=np.zeros((k.shape[0],1))
        l2=np.zeros((k.shape[0],1))
        l1=A
        for i in range(0,(int(k.shape[0]-1))):
         q1=(l1[i]-g_n)/(1-(i+1)/n)
         l2[i]=q1
        l2[-1]=l2[-2]
        return l2

param=np.array([31.704337789505427,              0.7458597738634244,
0.7922178198926836])
def parameter_estimator(param):

    a=param[0]
    M=param[1]
    H=param[2]
```

```
    clgram_real=Climacogram(A,0)
    clspec_real=Climacospectrum(clgram_real)

    clgram_theory=np.zeros((int(A.shape[0]/10),1))
    res1=HHK(clgram_real[0],a,M,H)
    clgram_theory=Climacogram_biased(res1[0],res1[1],res1[2])
    clspec_theory=Climacospectrum(clgram_theory)

    numb=np.arange(1,clgram_real.real.shape[0]+1)

b2=np.sum(np.log(clgram_theory[:]/clgram_real[:])**2+numb[:]**0.5)

a1=np.sum(np.log(clspec_theory[:]/clspec_real[:])**2+numb[:]**0.5)
    b1=(clgram_theory[0]-clgram_real[0])**2
    a2=b1+b2*10**5
    suma=a1+a2
    print(suma,param)
    return suma
bnds = ((0, 100), (0.1, 0.9999999),(0.1, 0.9999999))
strval=np.array([52.52,0.6421,0.7294])
res = minimize(parameter_estimator,strval,bounds=bnds)

import winsound
duration = 1000  # milliseconds
freq = 440  # Hz
winsound.Beep(freq, duration)

#plt.plot(c,label='Climacogram, theoretical')
plt.plot(clgram_real,label='Climacogram, empirical')
plt.plot(clgram_theory,label='Climacogram, theoretical adapted for
bias')
#plt.plot(clspec_real,label='Climacospectrum, empirical')
#plt.plot(clspec_theory,label='Climacospectrum, theoretical adapted
for bias')

plt.xscale('log')
plt.grid(True)
#plt.title('Climacogram')
```

```
#plt.title('Climacospectrum')
plt.xlabel('Scale(1hr)')
plt.legend()
plt.grid(True)
#plt.show()
plt.savefig('Climac31,704337789505427,           0,7458597738634244,
0,7922178198926836,',dpi=300)
#plt.savefig('Climacospectrum',dpi=2000)
plt.close()
```

APENDIX E

Scripts for simulation

```
def AMA_simulation(C1,sim_length):
  #   import models_for_AMA
 #    import last_try
  #   import AMA_coeffiecient
     import time
     import numpy as np

     arr = np.array(C1)
     C1 = arr[::-1]

     #start = time.time()

     import numpy as np
     import math
     from scipy.stats import norm
    # from scipy.stats import pearsonr

     n=C1.shape[0]
     #sim_length=1000

   #  logparam=1
     V=np.zeros(sim_length+n)
     rnd=np.random.rand(sim_length+n)
     V=norm.ppf(rnd[0:sim_length+n])
     for i in range(0,sim_length+n):
        V[i]=math.exp(V[i])
     X=np.zeros((sim_length,1),dtype=float)

     for i in range(0,sim_length):

      X[i]=(sum(np.multiply(C1[0:n],V[i:i+n])))
     #end = time.time()
    # print ((end-start),'seconds simulation')
```

```
     return X
```

Script: 1000simulations

```
import AMA_simulation
import numpy as np
import math
import pandas as pd
import scipy.stats
import matplotlib.pyplot as plt

def aggr(n,A):

    if A.shape[0]>n:
        B=np.zeros(np.int(A.shape[0]/n))
        B[0]=np.average(A[0:n])
        for i in range(1,B.shape[0]):
          B[i]=np.average(A[i*n:(i+1)*n])
        return B

def dif_proc_(A):
        dif_proc= np.empty((A.shape[0]-1,1))
        for i in range(0,A.shape[0]-1):
         dif_proc[i]=A[i+1]-A[i]
        return dif_proc

def reversibility_matrix(A):

    #dif_proc= np.empty((A.shape[0]-1,1))
    #for i in range(0,A.shape[0]-1):
    # dif_proc[i]=A[i+1]-A[i]

    max_=100
    intervals=100
    k=np.linspace(1,max_,intervals,dtype=int)
    l=np.zeros((k.shape[0]+1))
    for i in range(0,(k.shape[0])):
     q1=scipy.stats.skew(dif_proc_(aggr(k[i],A)),bias=False,
nan_policy='omit')
```

```
     q2=scipy.stats.skew(aggr(k[i],A),bias=False, nan_policy='omit')
     l[i+1]=q1/q2
    l[0]=0
    return l

count=0
nn=100
rev_matr=np.zeros((101))
var_prev=np.zeros((101))
rev_matr_f=np.zeros((101))

for j in range(0,nn):
    po=AMA_simulation.AMA_simulation(Xa[:,1],10000)
    var=reversibility_matrix(po)
    rev_matr[:]=var_prev[:]+var[:]
    count=count+1
    print(count)
    var_prev=rev_matr
    plt.plot(var[0:40],'grey',linewidth=1.5,alpha=0.4)
rev_matr_f[:]=rev_matr[:]/count
x=[1,2]
y=[1.39,1.2]
plt.scatter(x,y,color='black',s=60)
plt.plot(rev_matr_f[0:40],'black',linewidth=2)
plt.ylim(0, 2)
plt.xlim(1, 40)
plt.xscale('log')
#plt.title('Simulations using AMA at scale 1hr and 2hr')
plt.xlabel('Scale(1hr)')
plt.ylabel('Skewness ratio')
plt.grid(True)
plt.scatter(x,y,color='red',s=60,zorder=nn+1)
#plt.show()
plt.savefig('Simulations dfsd',dpi=300)
plt.close()

import winsound
```

```
duration = 1000  # milliseconds
freq = 440  # Hz
winsound.Beep(freq, duration)
```