

**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ
ΤΟΜΕΑΣ ΥΔΑΤΙΚΩΝ ΠΟΡΩΝ ΚΑΙ
ΠΕΡΙΒΑΛΛΟΝΤΟΣ**



**ΔΙΕΡΕΥΝΗΣΗ ΤΟΥ ΠΡΟΒΛΗΜΑΤΟΣ ΜΕΣΟΠΡΟΘΕΣΜΗΣ
ΠΡΟΓΝΩΣΗΣ ΤΩΝ ΕΙΣΡΟΩΝ ΤΑΜΙΕΥΤΗΡΩΝ ΜΕ ΧΡΗΣΗ
ΣΤΟΧΑΣΤΙΚΩΝ ΤΕΧΝΙΚΩΝ ΚΑΙ ΕΡΓΑΛΕΙΩΝ ΜΗΧΑΝΙΚΗΣ
ΜΑΘΗΣΗΣ**

Παναγιώτης Παγωτέλης

Επιβλέπων: Ανδρέας Ευστρατιάδης

Αθήνα, Νοέμβριος 2024

Υδροδοτικό Σύστημα της Αθήνας

Η Αθήνα και ένα μεγάλο μέρος της Αττικής υδροδοτούνται από ένα σύστημα αποτελούμενο στο μεγαλύτερο μέρος του από 3 ταμιευτήρες, εκ των οποίων οι δύο είναι κύριοι και ο ένας χρησιμοποιείται σαν εφεδρικός.

Οι κύριοι ταμιευτήρες είναι αυτοί του Ευήνου και του Μόρνου, οι οποίοι είναι τεχνητοί και έχουν χωρητικότητα και μέγιστο ωφέλιμο όγκο $138 / 113 \text{ hm}^3$ και $764 / 630 \text{ hm}^3$. Ο εφεδρικός ταμιευτήρας του συστήματος είναι η φυσική λίμνη της Υλίκης, με μέγιστη χωρητικότητα 553 hm^3 και μέγιστο ωφέλιμο όγκο 543 hm^3 (ΕΥΔΑΠ).

Με την υψηλή κατανάλωση νερού που παρατηρείται, συνδυασμένη με εκτενείς περιόδους ανομβρίας, διαφαίνεται η σημασία μίας ολοκληρωμένης και αξιόπιστης στρατηγικής διαχείρισης των συγκεκριμένων υδατικών πόρων, προκειμένου να εξασφαλίζεται η αδιάλειπτη παροχή νερού για τις ανάγκες ύδρευσης της Αττικής, αλλά και άρδευσης του κάμπου της Κοπαΐδας.



Αντικείμενο και σκοπός της εργασίας

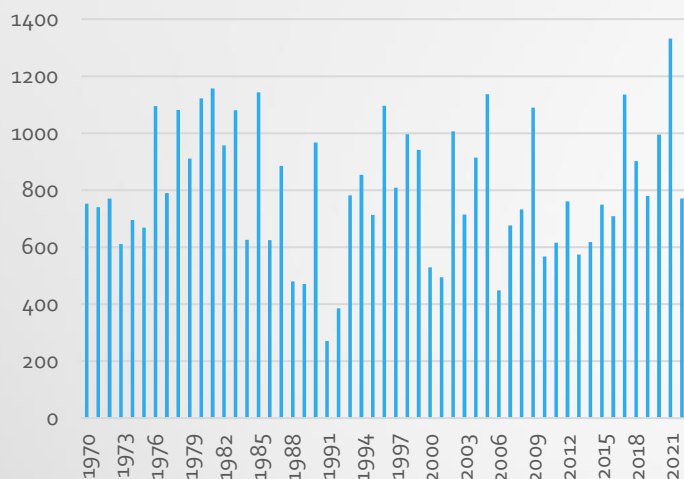
- Ο σκοπός της εργασίας είναι η αξιολόγηση εργαλείων μηχανικής μάθησης για την πρόβλεψη εισροών σε ταμιευτήρες που αποσκοπούν στην υδροδότηση της Αθήνας, χρησιμοποιώντας δεδομένα βροχόπτωσης και απορροής από προηγούμενους μήνες. Το πρόβλημα παρουσιάζει αρκετό ενδιαφέρον, καθώς ουσιαστικά εμπεριέχει την αναγνώριση σύνθετων κλιματικών διαδικασιών, οι οποίες παρουσιάζουν μεγάλη τυχαιότητα.
- Συγκεκριμένα, για τον σκοπό αυτό επιλέχθηκε η χρήση ενός μοντέλου LSTM, ενός αρκετά σύγχρονου και σύνθετου μοντέλου που ανήκει στην κατηγορία της επιβλεπόμενης μηχανικής μάθησης και χρησιμοποιείται κυρίως για την πρόβλεψη σειριακών δεδομένων, όπως είναι και οι χρονοσειρές.
- Περισσότερο ενδιαφέρον προσθέτει το γεγονός ότι η εκπαίδευση του μοντέλου θα γίνει καθαρά χρησιμοποιώντας συνθετικά δεδομένα, προκειμένου να ελαχιστοποιηθούν οι πιθανότητες να ενταχθούν μεροληψίες, οι οποίες θα επηρεάσουν αρνητικά τις επιδόσεις του μοντέλου όταν πια αυτό έχει να αντιμετωπίσει πραγματικά δεδομένα.
- Ταυτόχρονα, γίνεται χρήση και ενός απλού γραμμικού μοντέλου, το οποίο χρησιμοποιείται σαν benchmark προκειμένου να μπορούν να συγκριθούν οι επιδόσεις του LSTM για την πρόβλεψη στις 3 λεκάνες.

Το πρόβλημα της πρόβλεψης

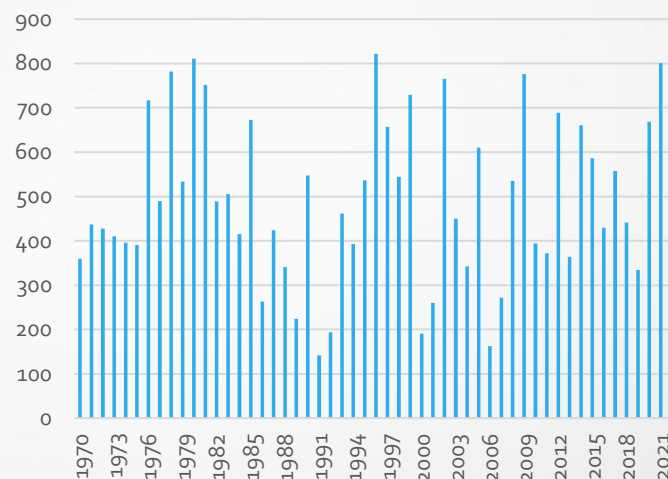
- Ίσως και η σημαντικότερη απόφαση που πρέπει να ληφθεί κατά την ανάπτυξη ενός τέτοιου συστήματος είναι η επιλογή του χρονικού ορίζοντα της πρόβλεψης, ώστε να μπορεί να προσαρμόζεται ταυτόχρονα και στις ανάγκες του διαχειριστικού συστήματος του οποίου αποτελεί μέρος, αλλά και στα πραγματικά δεδομένα που υπάρχει δυνατότητα να συλλεχθούν. Διερευνήθηκαν οι επιλογές της χρήσης (δεδομένα / προβλέψεις) ενός διαχωρισμού 3 / 9, 4 / 8 και 5 / 7 μηνών, όμως τελικά επιλέχθηκε ο διαχωρισμός 5 / 7, δηλαδή να γίνεται πρόβλεψη της απορροής των αντίστοιχων λεκανών για τους τελευταίους 7 μήνες του υδρολογικού έτους με την χρήση δεδομένων από τους 5 πρώτους μήνες.
- Άλλο ένα σημαντικό πρόβλημα που εμφανίζεται στη διαδικασία πρόβλεψης είναι το γεγονός πως τα ιστορικά δεδομένα που είναι διαθέσιμα για τη βροχόπτωση και την απορροή στις λεκάνες που μελετούνται είναι πολύ περιορισμένα (~ 50 έτη). Αυτό είναι ένα πρόβλημα που συναντάται παγκοσμίως, με τη μοναδική λεκάνη για την οποία υπάρχουν δεδομένα χιλιετιών (μεταβαλλόμενης, βέβαια, ποιότητας και αξιοπιστίας) να είναι αυτή του Νείλου στην Αίγυπτο. Έτσι, γεννήθηκε η ιδέα της χρήσης συνθετικών δεδομένων, τα οποία έχουν παραχθεί από τα πραγματικά, ώστε να υπάρχει αρκετός όγκος για να καταστεί δυνατή η εκπαίδευση των μοντέλων. Κύριος στόχος κατά τη σύνθεση των δεδομένων ήταν η διατήρηση της εμμορφής των ιστορικών χρονοσειρών, της μακροπρόθεσμης εξάρτησης, δηλαδή, που έχουν οι διαδοχικές τιμές μεταξύ τους.

Ιστορικά Δεδομένα – Λεκάνη Ευήνου

Αθροιστική βροχόπτωση για το πεντάμηνο
Οκτωβρίου – Φεβρουαρίου (mm)



Αθροιστική απορροή για το πεντάμηνο
Οκτωβρίου – Φεβρουαρίου (mm)

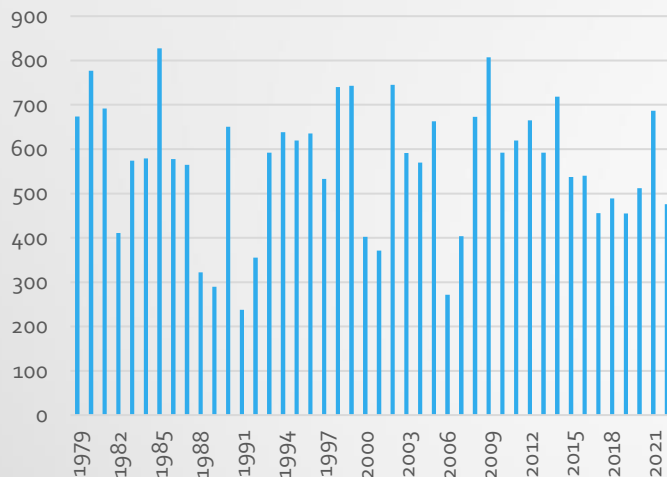


Στατιστικά μεγέθη χρονοσειράς απορροής λεκάνης Ευήνου για τα υδρολογικά έτη 1970-71 έως 2022-23, για διάφορες χρονικές κλίμακες (mm).

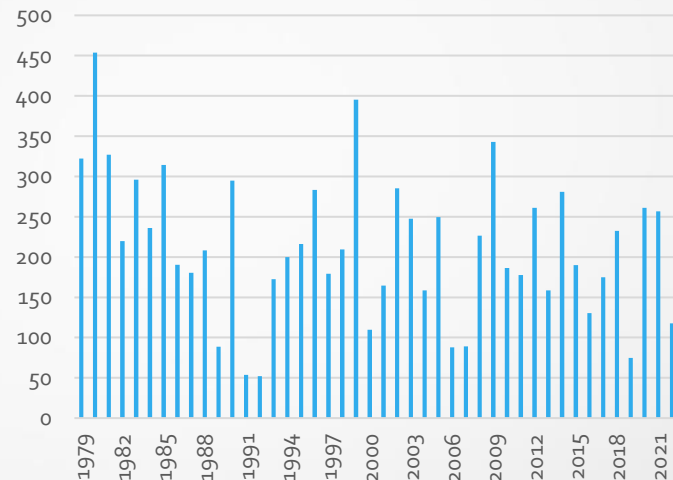
	Οκτ.	Νοε.	Δεκ.	Ιαν.	Φεβ.	Οκτ.- Φεβ.	Μαρ.- Σεπ.	Έτος
Μέση τιμή	24.1	72.0	139.0	124.7	128.6	488.4	318.0	806.4
Τυπική Απόκλιση	18.5	48.6	99.6	77.1	68.4	182.3	93.6	219.6
Ελάχιστη τιμή	5.4	9.6	21.0	16.2	12.3	141.6	97.9	321.7
Μέγιστη τιμή	77.0	220.7	475.0	344.0	284.9	821.8	497.6	1182.6

Ιστορικά Δεδομένα – Λεκάνη Μόρνου

Αθροιστική βροχόπτωση για το πεντάμηνο
Οκτωβρίου – Φεβρουαρίου (mm)



Αθροιστική απορροή για το πεντάμηνο
Οκτωβρίου – Φεβρουαρίου (mm)

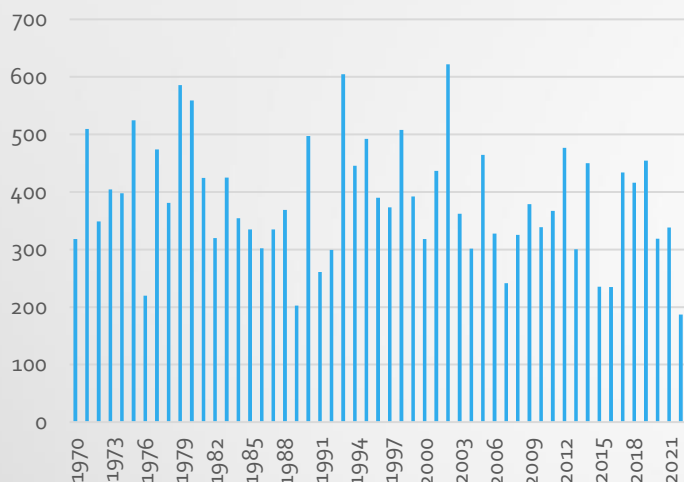


Στατιστικά μεγέθη χρονοσειράς απορροής λεκάνης Μόρνου για τα υδρολογικά έτη 1979-80 έως 2022-23, για διάφορες χρονικές κλίμακες (mm).

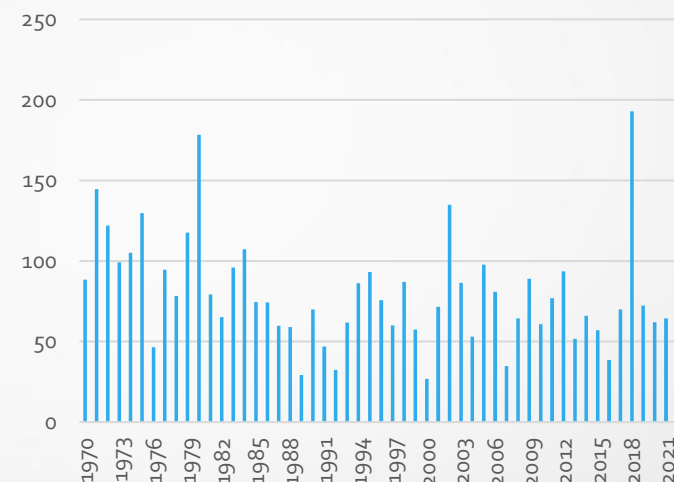
	Οκτ.	Νοε.	Δεκ.	Ιαν.	Φεβ.	Οκτ.- Φεβ.	Μαρ.- Σεπ.	Έτος
Μέση τιμή	14.4	34.6	53.3	54.2	56.2	212.6	197.2	409.9
Τυπική Απόκλιση	9.1	24.4	36.2	34.7	31.1	88.2	56.0	125.6
Ελάχιστη τιμή	1.1	5.0	7.0	5.4	7.1	52.0	66.7	155.3
Μέγιστη τιμή	42.4	108.0	167.6	142.8	141.7	453.6	339.4	710.8

Ιστορικά Δεδομένα – Λεκάνη Υλίκης

Αθροιστική βροχόπτωση για το πεντάμηνο
Οκτωβρίου – Φεβρουαρίου (mm)



Αθροιστική απορροή για το πεντάμηνο
Οκτωβρίου – Φεβρουαρίου (mm)



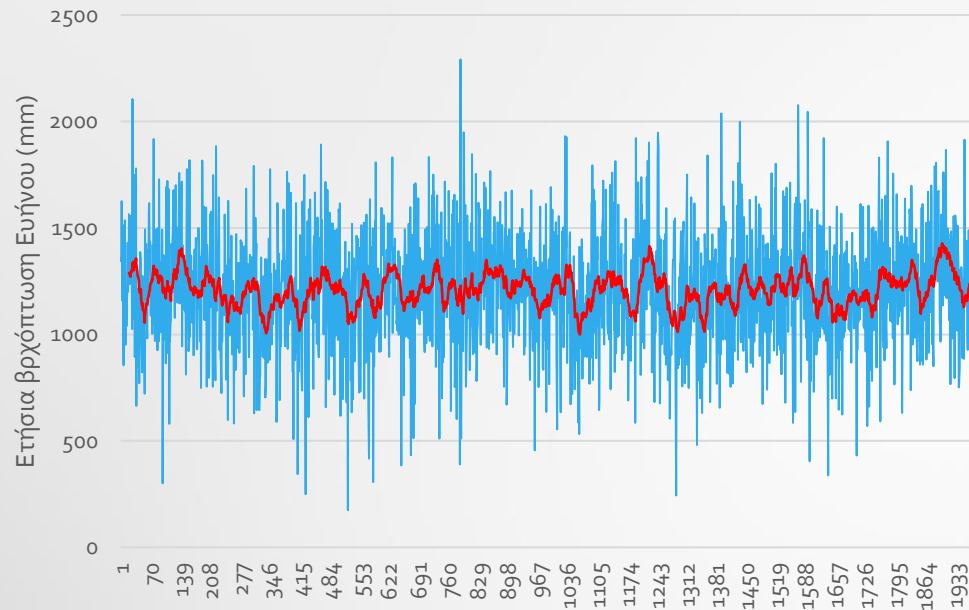
Στατιστικά μεγέθη χρονοσειράς απορροής λεκάνης Υλίκης για τα υδρολογικά έτη 1970-71 έως 2022-23, για διάφορες χρονικές κλίμακες (mm).

	Οκτ.	Νοε.	Δεκ.	Ιαν.	Φεβ.	Οκτ.- Φεβ.	Μαρ.- Σεπ.	Έτος
Μέση τιμή	7.5	10.2	15.2	22.1	24.2	79.2	62.7	141.9
Τυπική Απόκλιση	5.0	4.7	6.4	13.4	13.5	33.8	36.1	63.4
Ελάχιστη τιμή	0.0	3.4	4.7	7.4	1.8	26.9	5.4	33.5
Μέγιστη τιμή	21.8	25.8	31.5	72.5	64.8	192.9	148.0	340.9

Συνθετικά Δεδομένα

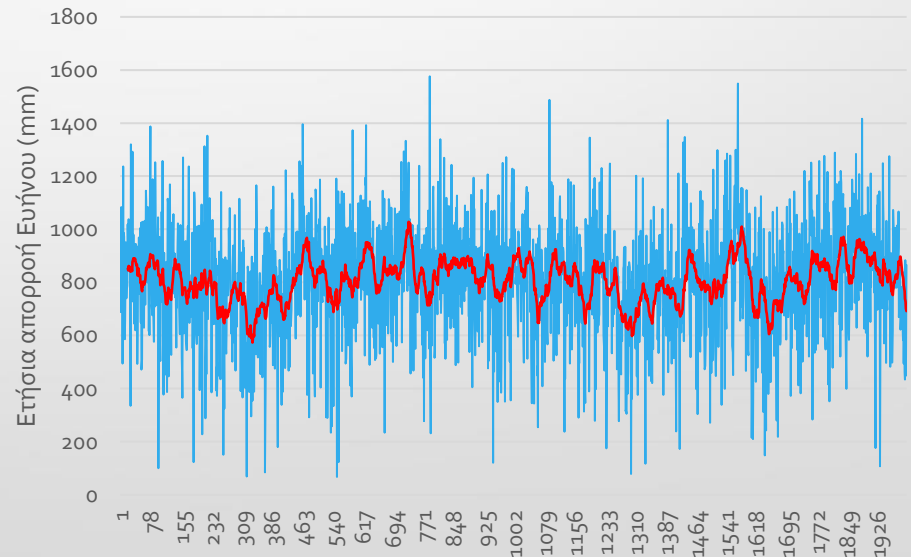
- Τα συνθετικά δεδομένα που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων στο πλαίσιο αυτής της εργασίας έχουν παραχθεί στο πλαίσιο της ερευνητικής εργασίας (Α. Ευστρατιάδης, και Γ.-Κ. Σακκή, Αναλύσεις υδατικών ισοζυγίων και απολογιστική έκθεση υδρολογικού έτους 2022-23, Εκσυγχρονισμός της διαχείρισης του συστήματος των υδατικών πόρων ύδρευσης της Αθήνας – Αναθεώρηση, 30 pages, Τομέας Υδατικών Πόρων και Περιβάλλοντος – Εθνικό Μετσόβιο Πολυτεχνείο, Ιανουάριος 2024). Η παραγωγή των συνθετικών δεδομένων δεν αποτελεί σκοπό της παρούσας εργασίας, αλλά μόνο η αξιοποίησή τους.
- Η γέννηση των συνθετικών χρονοσειρών έγινε μέσω πολυμεταβλητής στοχαστικής προσομοίωσης δύο χρονικών κλιμάκων, προκειμένου να αναπαράγονται σε αυτές τα στατιστικά χαρακτηριστικά και χρονικές και χωρικές συσχετίσεις των ιστορικών δειγμάτων. Σημαντικό ζητούμενο ήταν η αναπαραγωγή της μακροπρόθεσμης εμμονής, που αναφέρεται και ως δυναμική Hurst – Kolmogorov (Koutsoyiannis, 2011· Dimitriadis and Koutsoyiannis, 2015), προκειμένου να περιγραφούν οι μεγάλης κλίμακας υδροκλιματικές μεταβολές, με συσσώρευση μακρών περιόδων ξηρών και υγρών ετών. Η απαίτηση αυτή επιτεύχθηκε με την εισαγωγή μιας θεωρητικής συνάρτησης αυτοσυνδιασποράς, με κατάλληλη προσαρμογή των παραμέτρων της, για την περιγραφή της δομής αυτοσυσχέτισης των διεργασιών στην ετήσια κλίμακα (Koutsoyiannis, 2000· Efstratiadis *et al.*, 2014· Tsoukalas *et al.*, 2018). Με τη διαδικασία αυτή παρήχθησαν χρονοσειρές με έντονη εμμονή, με συντελεστές Hurst της τάξης του 0.75 έως 0.80 για τις απορροές (πιο υψηλοί στην καρστική λεκάνη του Βοιωτικού Κηφισού, όπως υποδεικνύουν και τα ιστορικά δεδομένα), και 0.65 έως 0.70 για τις βροχοπτώσεις. Οι χρονοσειρές που παρήχθησαν έχουν μήκος 2000 ετών για όλες τις λεκάνες.

Συνθετικά Δεδομένα – Λεκάνη Ευήνου

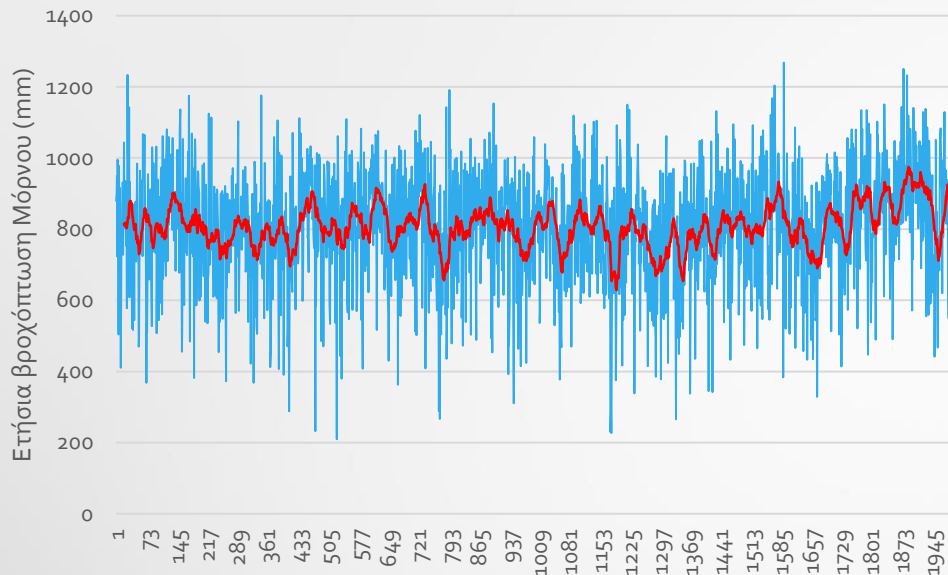


Συνθετική χρονοσειρά βροχόπτωσης λεκάνης Ευήνου, μήκους 2000 ετών, και κυλιόμενος μέσος όρος 20 ετών (mm).

Συνθετική χρονοσειρά απορροής λεκάνης Ευήνου, μήκους 2000 ετών, και κυλιόμενος μέσος όρος 20 ετών (mm).

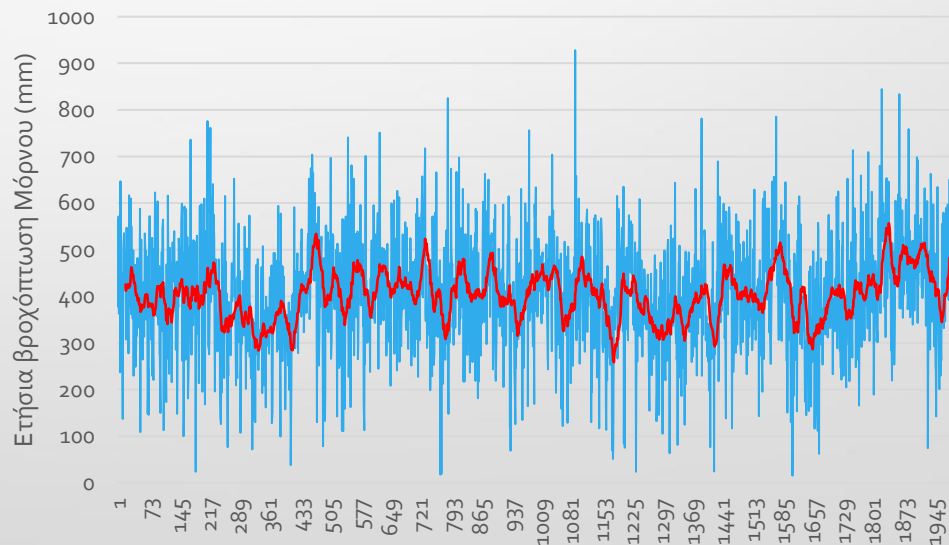


Συνθετικά Δεδομένα – Λεκάνη Μόρνου

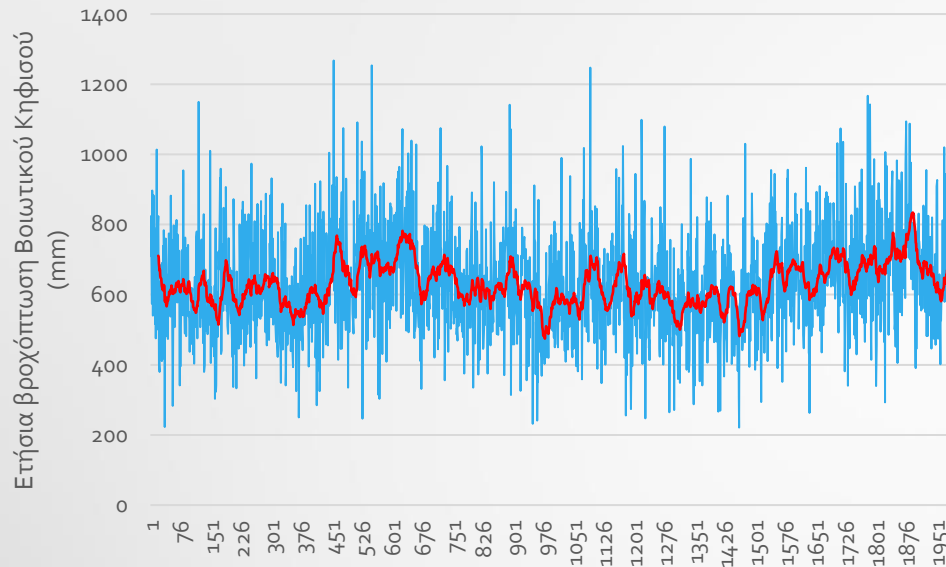


Συνθετική χρονοσειρά βροχόπτωσης λεκάνης Μόρνου, μήκους 2000 ετών, και κυλιόμενος μέσος όρος 20 ετών (mm).

Συνθετική χρονοσειρά απορροής λεκάνης Μόρνου, μήκους 2000 ετών, και κυλιόμενος μέσος όρος 20 ετών (mm).

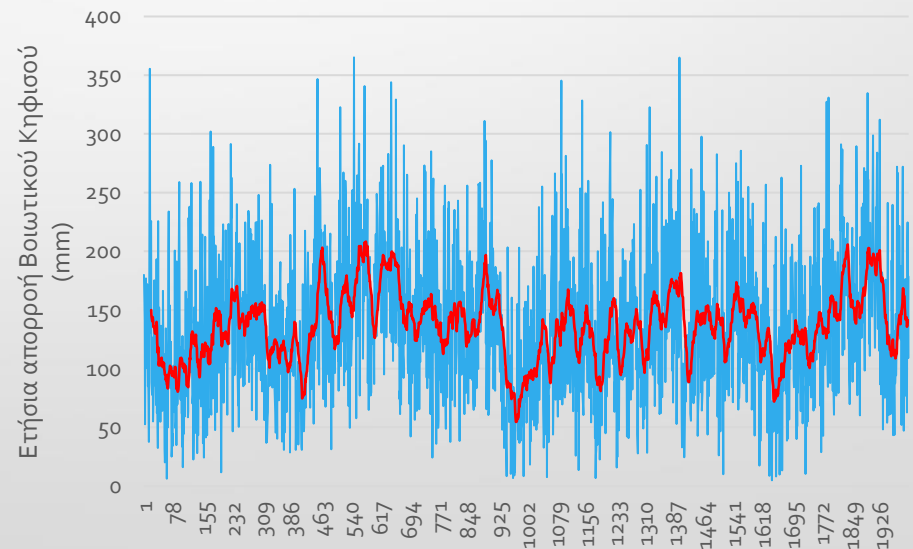


Συνθετικά Δεδομένα – Λεκάνη Υλίκης



Συνθετική χρονοσειρά βροχόπτωσης λεκάνης Υλίκης, μήκους 2000 ετών, και κυλιόμενος μέσος όρος 20 ετών (mm).

Συνθετική χρονοσειρά απορροής λεκάνης Υλίκης, μήκους 2000 ετών, και κυλιόμενος μέσος όρος 20 ετών (mm).

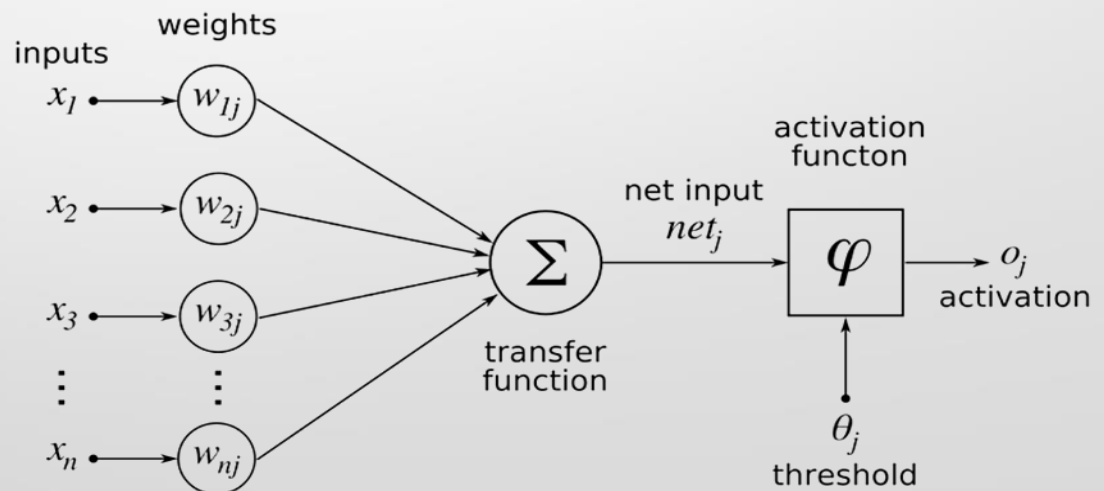


Μηχανική Μάθηση

- Η μηχανική μάθηση (MM), ή Machine Learning, αποτελεί ένα πεδίο της τεχνητής νοημοσύνης που έχει ως σκοπό την ανάπτυξη αλγορίθμων και μοντέλων για την αυτοματοποιημένη βελτίωση της απόδοσης συστημάτων μέσω εμπειρίας. Στηρίζεται στη στατιστική και την αλγοριθμική ανάλυση και έχει τη δυνατότητα να μαθαίνει από δεδομένα για να επιλύει προβλήματα όπως η κατηγοριοποίηση, η πρόβλεψη και η αναγνώριση μοτίβων.
- Οι αλγόριθμοι της μηχανικής μάθησης διακρίνονται κυρίως σε τρεις κατηγορίες:
 1. **Επιβλεπόμενη Μάθηση (Supervised Learning):** Περιλαμβάνει εκπαίδευση με δεδομένα που περιέχουν ετικέτες ή κατηγορίες, π.χ., για την κατηγοριοποίηση (classification) ή την παλινδρόμηση (regression). Σε αυτήν την κατηγορία ανήκει το μοντέλο LSTM που χρησιμοποιείται στη συγκεκριμένη περίπτωση.
 2. **Μη Επιβλεπόμενη Μάθηση (Unsupervised Learning):** Αναζητά κρυφά μοτίβα και σχέσεις στα δεδομένα χωρίς να απαιτείται σήμανση ή κατηγοριοποίηση των δεδομένων.
 3. **Ενισχυτική Μάθηση (Reinforcement Learning):** Το μοντέλο μαθαίνει μέσω αλληλεπίδρασης με το περιβάλλον του, βελτιστοποιώντας τις αποφάσεις του μέσω ανταμοιβών ή ποινών.

Τεχνητά Νευρωνικά Δίκτυα (1 / 2)

- Τα τεχνητά νευρωνικά δίκτυα, κοινώς νευρωνικά δίκτυα (neural networks) αποτελούν ένα υποπεδίο των αλγορίθμων μηχανικής μάθησης που έχουν ως στόχο την επίλυση των προβλημάτων μηχανικής μάθησης χρησιμοποιώντας αλγορίθμους των οποίων η δομή ομοιάζει με τους ανθρώπινους βιολογικούς νευρώνες.
- Τα περισσότερα τεχνητά νευρωνικά δίκτυα έχουν:
 1. Ένα στρώμα εισόδου, μέσα από το οποίο το μοντέλο «διαβάζει» τα δεδομένα,
 2. Ενδιάμεσα κρυφά στρώματα τα οποία αποσκοπούν στην αναγνώριση πιο σύνθετων και έντονα μη γραμμικών σχέσεων στα δεδομένα,
 3. Μία συνάρτηση ενεργοποίησης που υπάρχει στο στρώμα εξόδου, όπου το μοντέλο εξάγει τα αποτελέσματά του.



Τεχνητά Νευρωνικά Δίκτυα (2 / 2)

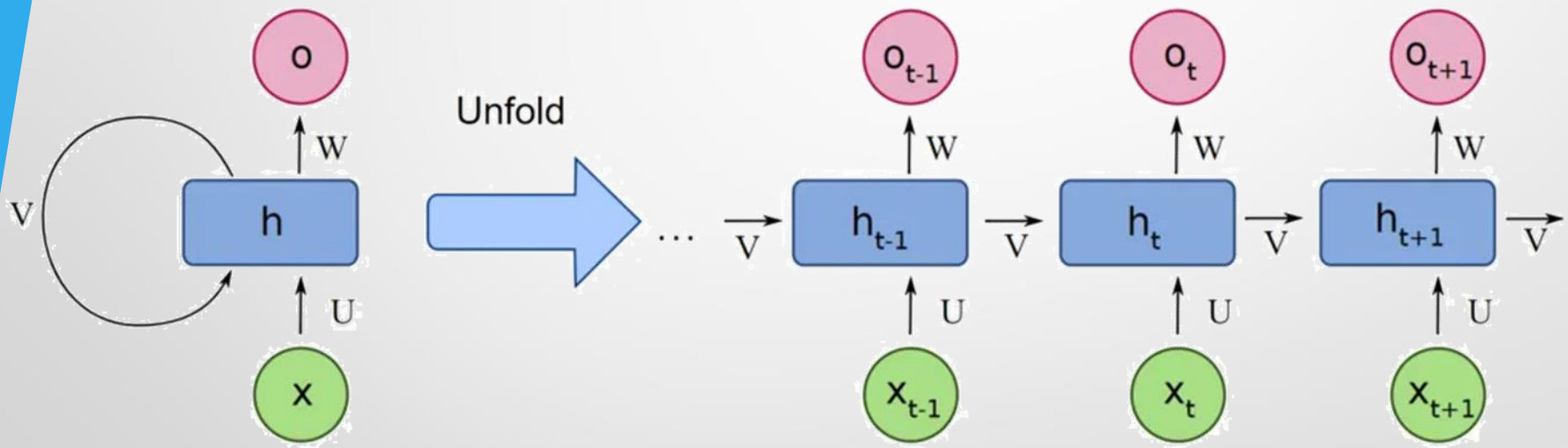
- Στην περίπτωση της παλινδρόμησης στη Μηχανική Μάθηση, όπως είναι και η συγκεκριμένη, η απόδοση ενός νευρωνικού δικτύου στις προβλέψεις του πρέπει να ποσοτικοποιείται με κάποιες μετρικές. Σκοπός δηλαδή είναι η ελαχιστοποίηση ενός όρου σφάλματος, ο οποίος συχνά, όπως και εδώ, επιλέγεται να είναι το μέσο τετραγωνικό σφάλμα MSE, που υπολογίζεται με τον συγκεκριμένο τύπο:

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- Η επιλογή του συγκεκριμένου όρου σφάλματος έναντι άλλων, όπως για παράδειγμα είναι το μέσο απόλυτο σφάλμα MAE, έγινε σκόπιμα. Ήταν ζητούμενο το μοντέλο να είναι αρκετά ευαίσθητο σε ακραίες τιμές (outliers), καθώς η πρόβλεψη συνεχόμενων περιόδων χαμηλής παροχής και ανομβρίας είναι ένας από τους κύριους σκοπούς της εργασίας. Ταυτόχρονα, με την ύψωση του απόλυτου σφάλματος σε δύναμη, τιμωρούνται εκθετικά τα μεγαλύτερα σφάλματα, κάτι που επίσης ταιριάζει στη συγκεκριμένη περίπτωση.
- Για την εκπαίδευση του μοντέλου και την ελαχιστοποίηση του όρου σφάλματος έγινε χρήση του αλγορίθμου Adam, ο οποίος αποτελεί μετεξέλιξη της μεθόδου Στοχαστικής Κατάβασης Κλίσης (Stochastic Gradient Descent, SGD). Χρησιμοποιείται ευρέως καθώς προσφέρει σημαντικά βελτιωμένη ικανότητα σύγκλισης για αρκετά μικρότερο υπολογιστικό κόστος σε σχέση με άλλους αλγορίθμους, όπως ο κλασικός SGD.

Αναδρομικά Δίκτυα RNN

- Τα Αναδρομικά Νευρωνικά Δίκτυα ή Recursive Neural Networks, αποτελούν υποκατηγορία νευρωνικών δικτύων τα οποία έχουν ενισχυμένες εσωτερικές συνδέσεις, οι οποίες είναι σχεδιασμένες ειδικά για να επεξεργάζονται και να προβλέπουν ακολουθιακά δεδομένα, όπως οι χρονοσειρές. Σε κάθε βήμα πρόβλεψης, μπορούν να εκμεταλλεύονται τόσο τα δεδομένα που τους δίνονται, όσο και τις σχέσεις αυτών των με δεδομένα προηγούμενων χρονικών βημάτων, και αυτό να συμβαίνει για πολλά χαρακτηριστικά ταυτόχρονα.



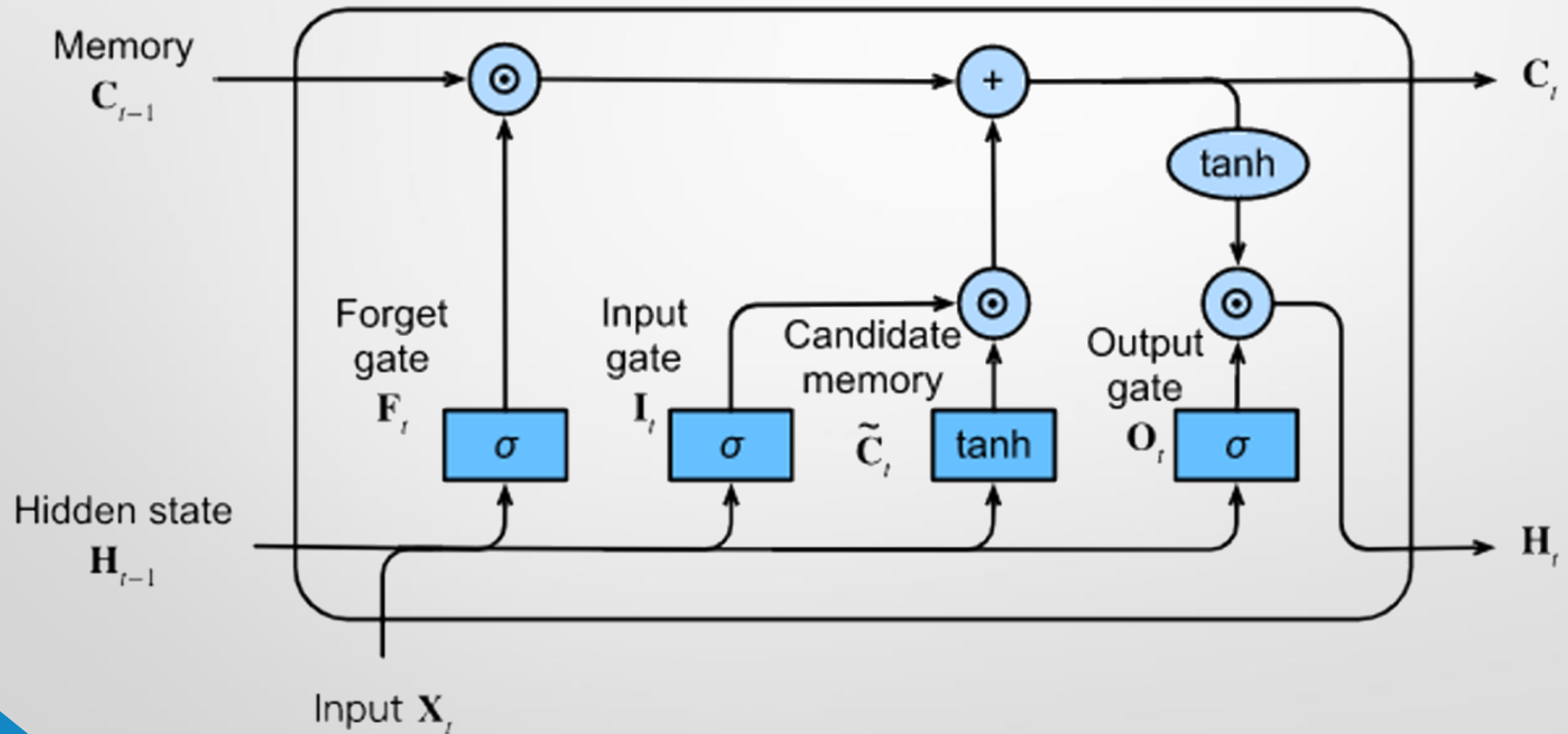
Σε ένα χρονικό βήμα t ισχύει: $h_t = f(W_x x_t + W_h h_{t-1} + b)$ και $y_t = g(W_y h_t + b_y)$, όπου h_t είναι η μνήμη του μοντέλου την χρονική στιγμή t , W_x τα «βάρη» των νευρώνων, b ένας όρος μεροληψίας και f η συνάρτηση ενεργοποίησης για την έξοδο.

Δίκτυα LSTM (1 / 2)

- Τέλος, υπάρχει μία εξειδίκευση των δικτύων RNN, τα δίκτυα Long Short – Term Memory, ή αλλιώς δίκτυα μακράς βραχύχρονης μνήμης. Δημιουργήθηκαν για να καλύψουν κάποιες εγγενείς αδυναμίες των RNN και χρησιμοποιούν στην αρχιτεκτονική τους πύλες για να επιτύχουν ορθότερη διαχείριση της πληροφορίας, κυρίως έχοντας τη δυνατότητα να διαχειριστούν ακολουθίες που χαρακτηρίζονται από πολύ μακρά μνήμη.
- Συγκεκριμένα, περιλαμβάνουν ένα κρυφό επίπεδο που περιέχει μία δομή που λέγεται κύτταρο (cell) και χρησιμοποιεί:
 1. Πύλες εισόδου (input gates), οι οποίες είναι υπεύθυνες για το ποια δεδομένα θα εισαχθούν και θα ληφθούν υπόψη από το μοντέλο,
 2. Πύλες λυσμόνησης (forget gates), πύλες που βρίσκονται εσωτερικά στην αρχιτεκτονική του μοντέλου και είναι υπεύθυνες για το ποια δεδομένα θα κρατηθούν στη μακρά μνήμη και ποια όχι,
 3. Και πύλες εξόδου (output gates), που είναι υπεύθυνες για την έξοδο του μοντέλου.
- Στα δίκτυα LSTM έρχεται να προστεθεί και μία τεχνική που ονομάζεται Bidirectional LSTM (BiLSTM), η οποία χρησιμοποιεί δύο μοντέλα LSTM ταυτόχρονα για να επεξεργαστεί τα δεδομένα και προς τις δύο χρονικές κατευθύνσεις και πιθανόν να εντοπίσει ακόμα πιο σύνθετες εξαρτήσεις που μπορεί να υπάρχουν σε αυτά.

Δίκτυα LSTM (2 / 2)

- Παρατίθεται ένα διάγραμμα που απεικονίζει τον τρόπο λειτουργίας της μνήμης των LSTM. Φαίνονται οι διαφορετικές πύλες και η συνάρτηση ενεργοποίησης.



Υλοποίηση του LSTM στην Python (1 / 2)

- Η υλοποίηση των δικτύων LSTM στο πλαίσιο αυτής της εργασίας έγινε με την χρήση της βιβλιοθήκης Keras που παρέχεται για την γλώσσα Python, ενώ υποστηρικτικά έγινε χρήση και της βιβλιοθήκης Μηχανικής Μάθησης scikit – learn. Επιλέχθηκε ένα δίκτυο με δύο στρώματα LSTM 128 και 64 νευρώνων, το οποίο σε κάθε βήμα εκπαίδευσης λάμβανε 2 χαρακτηριστικά (βροχόπτωση και απορροή) με 5 χρονικά βήματα το καθένα και επέστρεφε 7 εξόδους.
- Για τα μοντέλα έγινε μία αναζήτηση πλέγματος, με σκοπό την εύρεση των υπερπαραμέτρων τους οι οποίες θα οδηγούσαν στην καλύτερη απόδοση. Συγκεκριμένα, δοκιμάστηκαν:
 1. Δύο διαφορετικοί τρόποι τυποποίησης / μετασχηματισμού των δεδομένων για επεξεργασία από τα μοντέλα, η κανονικοποίηση $\frac{x-\mu}{\sigma}$, καθώς και η λογαριθμική τυποποίηση,
 2. Διαφορετικά batch sizes, δηλαδή μεγέθη των «πακέτων» στα οποία χωρίζει το δίκτυο τα δεδομένα κατά την εκπαίδευσή του,
 3. Διαφορετικά learning rates ή ρυθμοί μάθησης, που καθορίζουν την τάξη μεγέθους της ενημέρωσης που γίνεται στα βάρη μετά από κάθε βήμα μάθησης,
 4. Και τέλος διαφορετικό dropout, που είναι μία παράμετρος που χρησιμοποιείται για την αποφυγή της υπερπροσαρμογής (overfitting), που συμβαίνει όταν το μοντέλο ακολουθεί υπερβολικά πιστά μοτίβα που διδάχθηκε κατά την εκπαίδευσή του και δυσκολεύεται να γενικεύσει σε νέα δεδομένα.
- Αυτή η αναζήτηση πλέγματος έγινε με σκοπό να βρεθούν οι υπερπαραμέτροι που οδηγούν στην καλύτερη απόδοση, με κριτήριο τα συνθετικά δεδομένα, χωρίς να είναι γνωστό αν αυτό μεταφράζεται άμεσα στα πραγματικά, αφού δε συμμετείχαν σε κανένα στάδιο της διαδικασίας μάθησης.

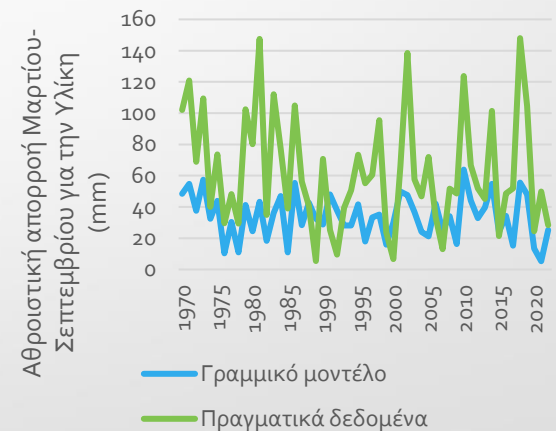
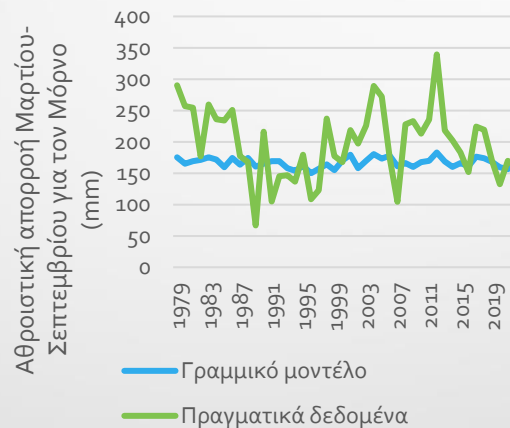
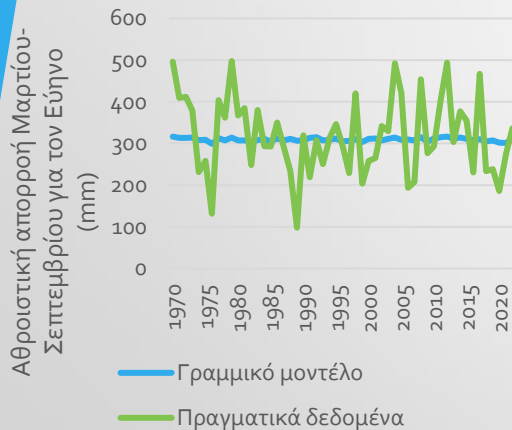
Υλοποίηση του LSTM στην Python (2 / 2)

- Αξίζει εδώ να σημειωθεί πως το πρώτο 80% των συνθετικών δεδομένων χρησιμοποιήθηκε για την εκπαίδευση των μοντέλων και το τελευταίο 20% για την επαλήθευσή τους, για την εύρεση δηλαδή των βέλτιστων υπερπαραμέτρων και για μία αρχική αξιολόγηση πάνω σε δεδομένα αντίστοιχα με αυτά της εκπαίδευσης. Όπως αναφέρθηκε και νωρίτερα, λόγω της εκπαίδευσης με καθαρά συνθετικά δεδομένα, ένας από τους κύριους στόχους της διαδικασίας ήταν η αποφυγή της υπερπροσαρμογής των μοντέλων στα συνθετικά δεδομένα. Αυτό επιχειρήθηκε να επιτευχθεί με δύο τρόπους, οι οποίοι εφαρμόστηκαν ταυτόχρονα:
 1. Την παράμετρο dropout του πρώτου στρώματος LSTM, η οποία καθορίζει το ποσοστό των νευρώνων που «απενεργοποιούνται» σε διαφορετικές στιγμές της εκπαιδευτικής διαδικασίας με σκοπό να μειώσει την ευαισθησία τους και να εντάξει περισσότερη αβεβαιότητα στη διαδικασία,
 2. Με χρήση Early Stopping. Συγκεκριμένα, ορίστηκε αρχικά μία σταθερή παράμετρος εκπαίδευσης 100 «εποχών» για τα μοντέλα, 100 «περασμάτων» δηλαδή από τα δεδομένα. Μετά το πέρας κάθε εποχής, το μοντέλο παρακολουθεί τις επιδόσεις του τόσο στα δεδομένα εκπαίδευσης (80%), όσο και στα δεδομένα επικύρωσης (20%). Μετά από έναν συγκεκριμένο αριθμό εποχών, η επίδοση του μοντέλου στα δεδομένα επικύρωσης αρχίζει να μειώνεται, δείγμα ότι αρχίζει να υπάρχει υπερπροσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης. Έτσι, η εκπαίδευση μπορεί και σταματάει αυτόματα όταν αρχίζει να συμβαίνει αυτό.

Γραμμικό Μοντέλο

- Σαν benchmark / κριτήριο για την επίδοση του LSTM, έγινε και εκπαίδευση ενός απλού γραμμικού μοντέλου, το οποίο λάμβανε ως μεταβλητές εισόδου τις αθροιστικές βροχοπτώσεις και απορροές για τους πρώτους 5 μήνες του υδρολογικού έτους και επέστρεφε την αθροιστική απορροή της λεκάνης για τους τελευταίους 7 μήνες. Ο τύπος του είναι:

$$\hat{Y} = w_1x_1 + w_2x_2 + b.$$



Προβλέψεις - Υπερπαράμετροι

- Για κάθε λεκάνη κρατήθηκαν τελικά 4 δίκτυα LSTM, δύο με κανονική τυποποίηση και δύο με λογαριθμική, στα οποία έγινε διαφορετική αναζήτηση πλέγματος για τη βελτιστοποίηση των υπερπαραμέτρων τους. Οι 12 συνδυασμοί για τις 3 λεκάνες που προέκυψαν είναι οι εξής:

Για τον Εύηνο

Τυποποίηση	Learning Rate	Batch Size	Dropout
Κανονική	0.001	4	0
Κατανομή	0.001	2	0
Log	0.001	8	0.1
	0.001	16	0.1

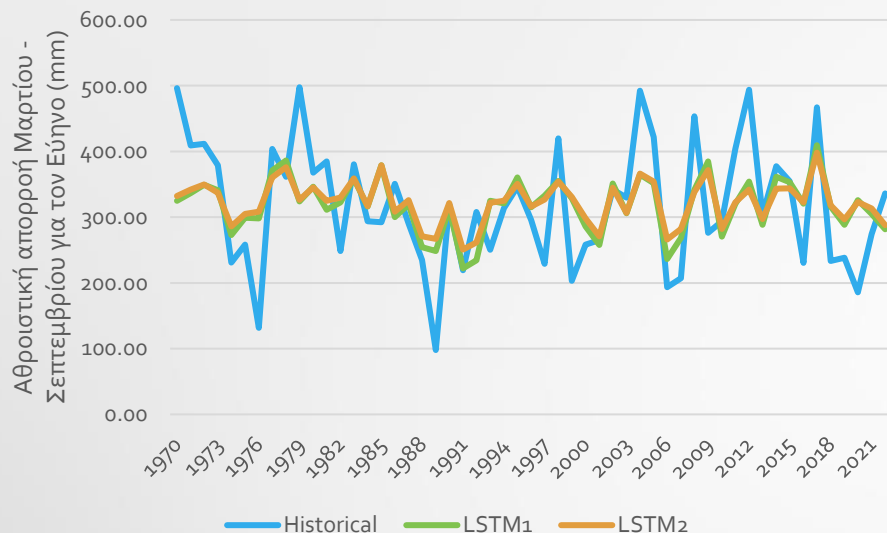
Για την Υλίκη

Τυποποίηση	Learning Rate	Batch Size	Dropout
Κανονική Κατανομή	0.01	32	0
	0.005	32	0
Log	0.001	16	0
	0.0005	8	0

Για τον Μόρνο

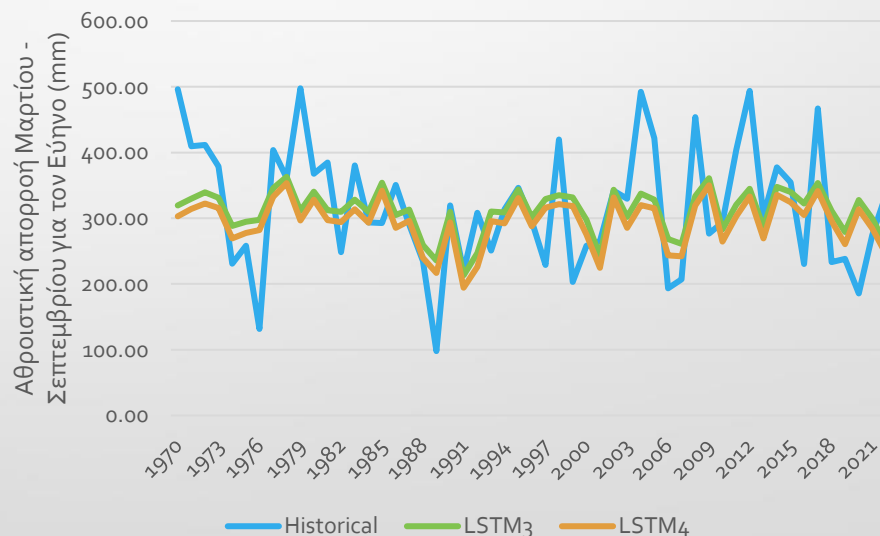
Τυποποίηση	Learning Rate	Batch Size	Dropout
Κανονική Κατανομή	0.01	32	0.5
	0.005	32	0.4
Log	0.001	16	0
	0.0005	8	0

Προβλέψεις - Εύηνος

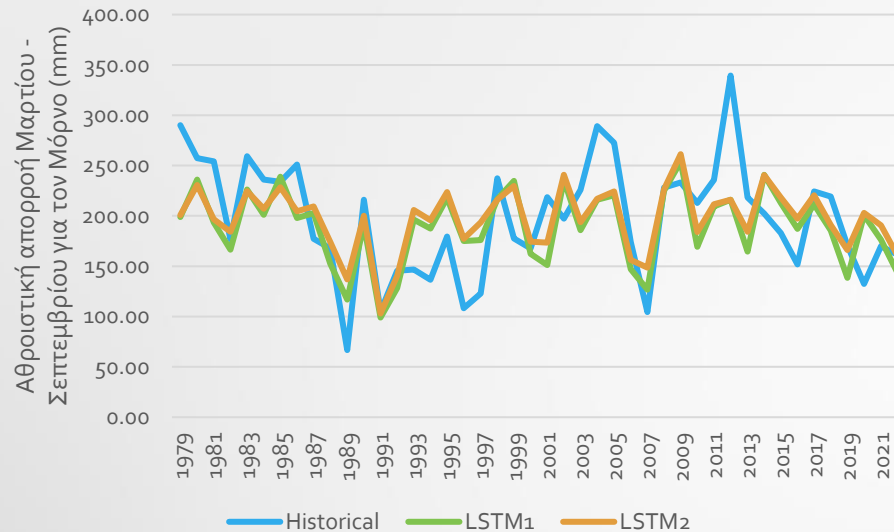


Για τα μοντέλα LSTM1 και LSTM2, παρατηρείται πως μπορούν να εντοπίζουν ικανοποιητικά τις τάσεις της πραγματικής χρονοσειράς, με κάποιες παραφωνίες, αλλά είναι ξεκάθαρο πως αντιμετωπίζουν πρόβλημα στην πρόβλεψη της έντασης των πιο ακραίων φαινομένων, είτε αυτά είναι περίοδοι ξηρασίας, είτε περίοδοι πολυομβρίας.

Για τα μοντέλα LSTM3 και LSTM4, οι επιδόσεις είναι αντίστοιχες, με βελτίωση σε κάποια σημεία, αλλά και επιδείνωση της πρόβλεψης σε κάποια άλλα. Στη συγκεκριμένη περίπτωση η λογαριθμική τυποποίηση των δεδομένων δεν αποδίδει για τα πιο ακραία φαινόμενα, γεγονός που κρίνεται κάπως αναπάντεχο.

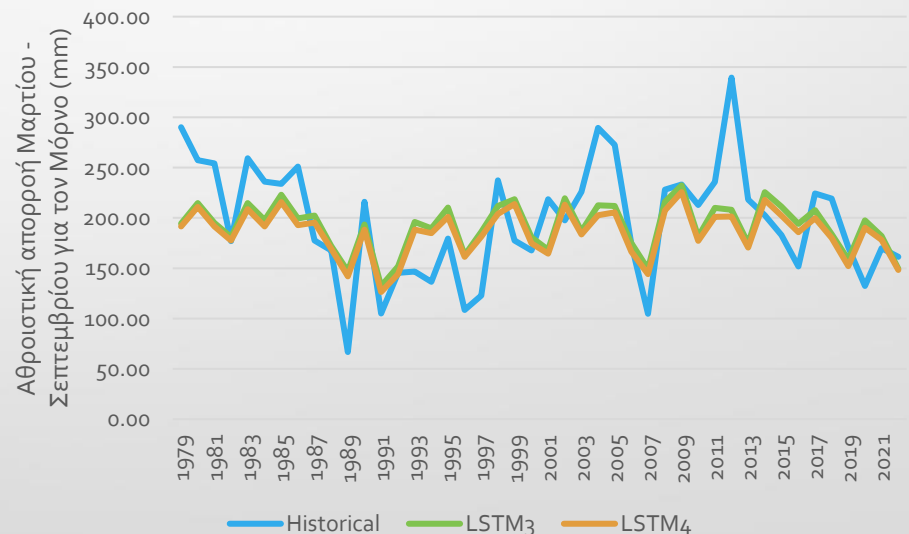


Προβλέψεις - Μόρνος

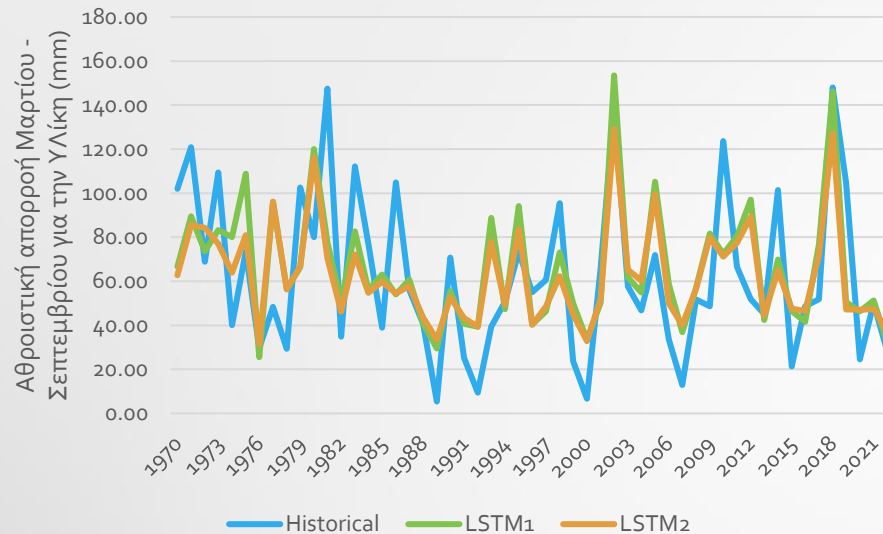


Για τα μοντέλα LSTM1 και LSTM2, παρατηρείται πως μπορούν να εντοπίζουν τις τάσεις της πραγματικής χρονοσειράς με αρκετή ακρίβεια, εμφανώς καλύτερα από τη λεκάνη του Ευήνου. Και πάλι όμως φαίνεται μία αδυναμία στον εντοπισμό της έντασης των πιο ακραίων φαινομένων και περιόδων, με βελτιωμένη βέβαια επίδοση στις χαμηλές ροές.

Στα μοντέλα LSTM3 και LSTM4, οι επιδόσεις είναι αρκετά χειρότερες, με το μοντέλο να παρουσιάζει μειωμένη δυνατότητα και στις υψηλές και στις χαμηλές ροές, με γενικότερη δυσκολία στον εντοπισμό της διακύμανσης των τιμών, γεγονός και πάλι παράξενο, όπως στη λεκάνη του Ευήνου.

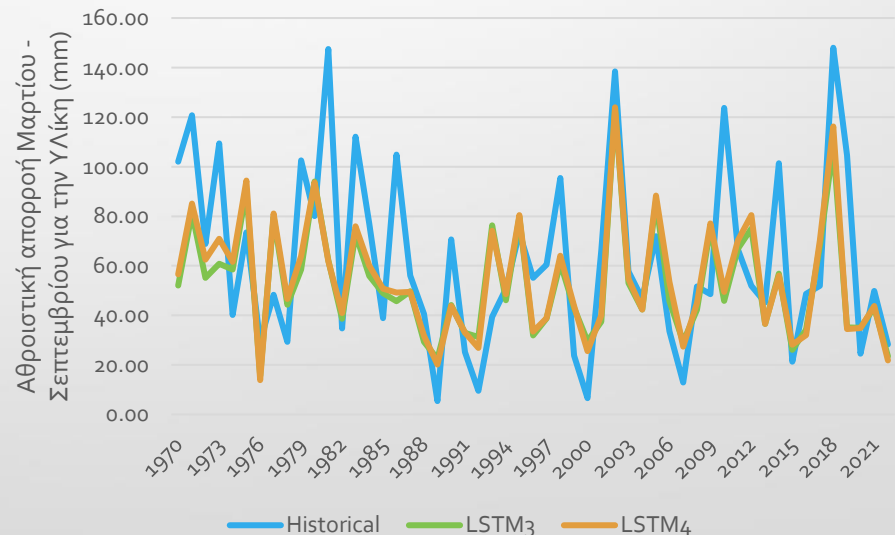


Προβλέψεις - Υλίκη



Για τα μοντέλα LSTM1 και LSTM2, παρατηρείται πως μπορούν να εντοπίζουν πολύ καλά τις τάσεις της πραγματικής χρονοσειράς, ειδικότερα στις μέσες και υψηλές ροές, με την απόδοση στις χαμηλές ροές να είναι επίσης ικανοποιητική, καλύτερη από τις άλλες λεκάνες. Βέβαια, η λεκάνη του Βοιωτικού Κηφισού έχει γενικότερα διαφορετική συμπεριφορά, όπως φαίνεται και από το γραμμικό μοντέλο.

Για τα μοντέλα LSTM3 και LSTM4, οι επιδόσεις είναι και πάλι αρκετά καλές γενικότερα, αυτή τη φορά όμως περισσότερο στις χαμηλές ροές, τις οποίες φαίνεται το μοντέλο να μπορεί να εντοπίζει με μεγάλη πιστότητα, τόσο στο κομμάτι της τάσης, όσο και στην ένταση των περιόδων ανομβρίας.



Ποσοτικοποίηση των Επιδόσεων των Μοντέλων

- Πέρα από την οπτική αξιολόγηση των επιδόσεων των μοντέλων, πρέπει να γίνει και η ποσοτικοποίησή τους με τις κατάλληλες μετρικές για να εντοπιστεί η επίδοσή τους σε διάφορα ζητούμενα υποπροβλήματα που πρέπει να λυθούν. Για αυτόν τον λόγο, επιλέγονται 3 μετρικές που αποτελούν θεμέλιους λίθους στον τομέα της αξιολόγησης υδρολογικών μοντέλων.

1. Αρχικά, επιλέγεται η Nash – Sutcliffe Efficiency, μία πολύ κλασική μετρική. Ο τύπος της είναι:

$$NSE = 1 - \frac{\sum_{i=1}^n (Q_{o,i} - Q_{p,i})^2}{\sum_{i=1}^n (Q_{o,i} - \bar{Q}_o)^2}$$

2. Έπειτα, χρησιμοποιείται η LNSE, η οποία είναι η NSE με τους λογαρίθμους των τιμών των προβλέψεων και των παρατηρήσεων. Η συγκεκριμένη μετρική είναι ιδιαίτερα χρήσιμη για την ποσοτικοποίηση της επίδοσης των μοντέλων στις χαμηλές ροές.
3. Τέλος, χρησιμοποιείται η επίσης κλασική μετρική στην Υδρολογία, Kling – Gupta Efficiency ή KGE, η οποία είναι σημαντική καθώς συνδυάζει τρεις διαφορετικές παραμέτρους: τη συσχέτιση, τη μεροληψία και τη μεταβλητότητα μεταξύ των τιμών των προβλέψεων και των παρατηρήσεων. Ο τύπος της είναι:

$$KGE = 1 - \sqrt{(r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$$

- Όπου r είναι ο συντελεστής συσχέτισης Pearson, α ο λόγος της τυπικής απόκλισης των προβλέψεων προς αυτής των παρατηρήσεων και β ο λόγος των μέσων τιμών τους.

Επιδόσεις - Εύηνος

	Historical	LSTM ₁	LSTM ₂	LSTM ₃	LSTM ₄	Linear
Μέσος όρος	318.05	318.99	322.57	310.67	294.92	309.44
Τυπική απόκλιση	93.60	41.02	32.26	33.43	35.64	3.97
Συσχέτιση	1.00	0.58	0.60	0.54	0.54	0.66
MSE	0	5959	6219	6544	6978	8363
NSE	1.00	0.32	0.29	0.25	0.20	0.05
LNSE	1.00	0.29	0.24	0.26	0.26	0.04
KGE	1.00	0.30	0.23	0.21	0.22	-0.02

- Όπως ήταν εμφανές και από τα διαγράμματα σύγκρισης των πραγματικών μετρήσεων με τις προβλέψεις των μοντέλων, το μοντέλο που παρουσιάζει τις καλύτερες επιδόσεις με βάση όλες τις μετρικές είναι το LSTM₁, το οποίο χρησιμοποιεί τυποποίηση των δεδομένων με βάση την κανονική κατανομή.
- Επίσης, υπάρχει επιβεβαίωση των συμπερασμάτων των διαγραμμάτων στο γεγονός ότι όλα τα μοντέλα παρουσιάζουν αρκετά μέτριες επιδόσεις, ιδιαίτερα στην πρόβλεψη της έντασης των φαινομένων, όπως γίνεται φανερό από τον χαμηλό δείκτη LNSE, αλλά και από το γεγονός ότι η τυπική απόκλιση των προβλέψεων του καλύτερου μοντέλου είναι λιγότερη από τη μισή της τυπικής απόκλισης της πραγματικής χρονοσειράς.
- Συμπερασματικά, η επίδοση των μοντέλων φαίνεται μέτρια και είναι προφανές πως απαιτείται περισσότερη διερεύνηση για να βρεθεί ένα μοντέλο που ανταποκρίνεται ορθά στη λεκάνη του Ευήνου.

Επιδόσεις - Μόρνος

	Historical	LSTM ₁	LSTM ₂	LSTM ₃	LSTM ₄	Linear
Μέσος όρος	197.22	189.39	197.31	191.61	185.74	166.67
Τυπική απόκλιση	55.96	36.17	30.74	23.70	22.81	7.60
Συσχέτιση	1.00	0.61	0.60	0.62	0.63	0.61
MSE	0	2013	1675	1721	1796	2994
NSE	1.00	0.36	0.36	0.34	0.31	-0.15
LNSE	1.00	0.44	0.40	0.38	0.38	-0.01
KGE	1.00	0.48	0.40	0.31	0.30	0.04

- Και πάλι, προς επιβεβαίωση των αρχικών συμπερασμάτων, φαίνεται πως το μοντέλο LSTM₁ υπερέχει σημαντικά σε σχέση με τα άλλα σε όλες τις μετρικές, κάτι που φαίνεται να ενισχύει την υπόθεση πως μοντέλα που θεωρητικά είναι λιγότερο βελτιστοποιημένα στα συνθετικά δεδομένα παρουσιάζουν καλύτερη συμπεριφορά στην πρόβλεψη των πραγματικών, πιθανόν λόγω καλύτερης δυνατότητας γενίκευσης.
- Όλα τα μοντέλα παρουσιάζουν και πάλι μέτριες επιδόσεις, αν και εμφανώς καλύτερες από τον Εύηνο, ιδιαίτερα στην πρόβλεψη της έντασης των φαινομένων, όπως ήταν φανερό και από τα διαγράμματα. Όσον αφορά τα στατιστικά μεγέθη, υπάρχει σημαντικά βελτιωμένη εκτίμηση της τυπικής απόκλισης της πραγματικής χρονοσειράς, ιδίως από το LSTM₁.
- Συμπερασματικά, η επίδοση των μοντέλων φαίνεται μέτρια προς καλή, αλλά και πάλι είναι προφανές πως απαιτείται περισσότερη διερεύνηση για να βρεθεί ένα μοντέλο που ανταποκρίνεται ορθά στη λεκάνη του Μόρνου, αν και το LSTM₁ φαίνεται υποσχόμενο.

Επιδόσεις - Υλίκη

	Historical	LSTM1	LSTM2	LSTM3	LSTM4	Linear
Μέσος όρος	62.75	66.12	63.04	53.22	54.89	34.04
Τυπική απόκλιση	36.09	27.11	22.34	22.47	23.71	13.87
Συσχέτιση	1.00	0.63	0.62	0.61	0.63	0.64
MSE	0	812	798	908	840	1680
NSE	1.00	0.38	0.39	0.30	0.36	-0.29
LNSE	1.00	0.37	0.36	0.41	0.47	-0.43
KGE	1.00	0.55	0.46	0.44	0.48	0.15

- Στη λεκάνη απορροής του Βοιωτικού Κηφισού, στη λίμνη Υλίκη, φαίνεται να υπάρχει σημαντική βελτίωση των επιδόσεων των μοντέλων. Όπως και στις προηγούμενες λεκάνες, και πάλι υπερέχει το μοντέλο LSTM1 στις πιο γενικές μετρικές, όμως το μοντέλο LSTM4 παρουσιάζει μία πολύ καλή επίδοση στη μετρική LNSE, κάτι που είναι θεωρητικά αναμενόμενο για ένα μοντέλο με δεδομένα που έχουν τυποποιηθεί λογαριθμικά.
- Όλα τα μοντέλα παρουσιάζουν συνολικά καλύτερες επιδόσεις, κάτι που πιθανότατα οφείλεται στη διαφορετική συμπεριφορά της λεκάνης που μελετάται. Όσον αφορά τα στατιστικά μεγέθη, υπάρχει πολύ βελτιωμένη εκτίμηση της τυπικής απόκλισης της πραγματικής χρονοσειράς από το LSTM1. Ειδική μνεία αξίζει και το LSTM4, καθώς επιβεβαιώνεται και αριθμητικά αυτό που ήταν φανερό οπτικά, ότι δηλαδή έχει μία πολύ καλή επίδοση στην πρόβλεψη των χαμηλών ροών.
- Συμπερασματικά, η επίδοση των μοντέλων φαίνεται και πάλι μέτρια προς καλή, όμως για τη συγκεκριμένη λεκάνη φαίνεται πολλά υποσχόμενος ένας συνδυασμός των LSTM1 και LSTM4, ο οποίος μπορεί εύκολα να πραγματοποιηθεί.

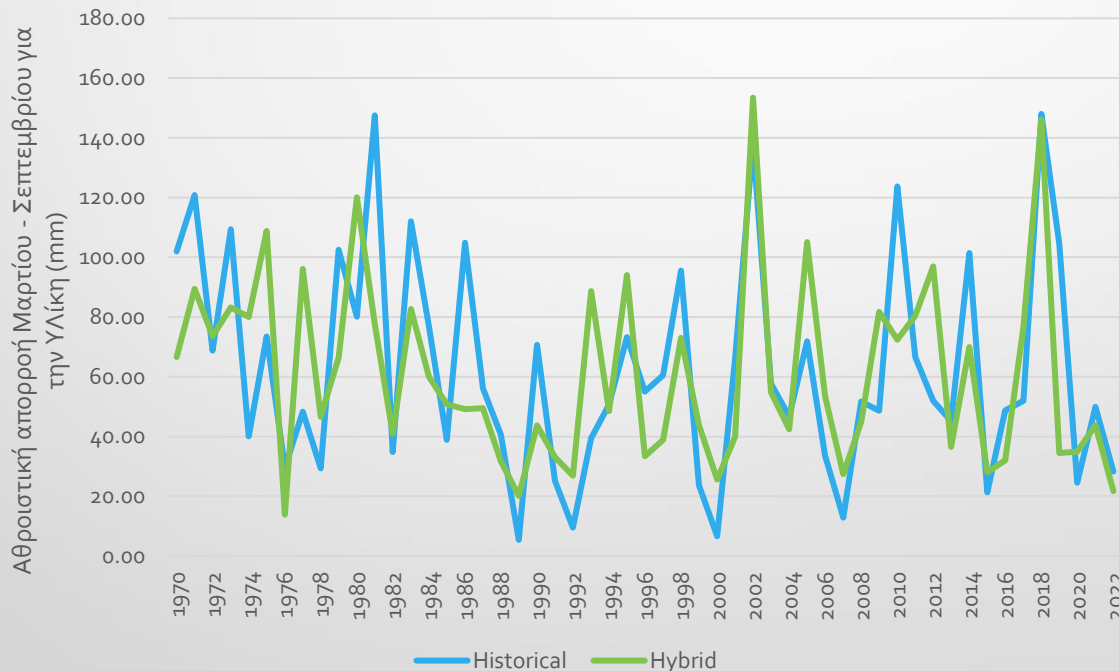
Συνδυασμός LSTM1 και LSTM4 – Υλίκη (1 / 2)

- Για τον συνδυασμό των δύο μοντέλων με στόχο τη δημιουργία ενός υβριδικού μοντέλου, επιλέγεται η εξής προσέγγιση: Σε κάθε βήμα γίνεται σύγκριση μεταξύ της τιμής του LSTM1 και του LSTM4 και αν $LSTM1 \geq \mu(LSTM1)$, τότε επιλέγεται η τιμή LSTM1, αλλιώς επιλέγεται η τιμή του LSTM4.
- Προκύπτουν τα εξής αποτελέσματα, επιβεβαιώνοντας την υπόθεση πως ένα απλό υβριδικό μοντέλο μπορεί να συνδυάσει τις καλές επιδόσεις των δύο μοντέλων:

	Historical	LSTM ₁	LSTM ₂	LSTM ₃	LSTM ₄	Linear	Hybrid
Μέσος όρος	62.75	66.12	63.04	53.22	54.89	34.04	61.03
Τυπική απόκλιση	36.09	27.11	22.34	22.47	23.71	13.87	30.99
Συσχέτιση	1.00	0.63	0.62	0.61	0.63	0.64	0.65
MSE	0	812	798	908	840	1680	813
NSE	1.00	0.38	0.39	0.30	0.36	-0.29	0.38
LNSE	1.00	0.37	0.36	0.41	0.47	-0.43	0.47
KGE	1.00	0.55	0.46	0.44	0.48	0.15	0.62

Συνδυασμός LSTM1 και LSTM4 – Υλίκη (2 / 2)

- Ο αρκετά καλός LNSE και ο εξαιρετικός KGE επιβεβαιώνονται γραφικά, καθώς είναι εμφανές ότι το υβριδικό μοντέλο παρουσιάζει αρκετά καλές επιδόσεις, ιδιαίτερα στις χαμηλές ροές, με κάποιες απώλειες σε περιόδους πολυομβρίας. Αποδεικνύεται λοιπόν πως ο συνδυασμός LSTM με διαφορετικούς τρόπους τυποποίησης των δεδομένων είναι μία προσέγγιση που μπορεί να αποδώσει καρπούς.



Συμπεράσματα

- **Ικανοποιητική απόδοση του LSTM και στις τρεις υδρολογικές λεκάνες:** Το μοντέλο LSTM κατάφερε να αποδώσει ικανοποιητικά στην πρόβλεψη απορροής σε τρεις διαφορετικές υδρολογικές λεκάνες, παρά την πολυπλοκότητα του προβλήματος και τη χρήση συνθετικών δεδομένων. Αυτή η επιτυχία δείχνει την ικανότητα των νευρωνικών δικτύων, και ιδιαίτερα των LSTM, να χειρίζονται ικανά τις μη γραμμικές και σύνθετες σχέσεις που χαρακτηρίζουν το συγκεκριμένο υδρολογικό πρόβλημα.
- **Ακριβείς προβλέψεις απορροής στους Εύηνο και Μόρνο:** Στους ποταμούς Εύηνο και Μόρνο, το LSTM κατάφερε να εκτιμήσει με ικανοποιητική ακρίβεια τη συμπεριφορά της απορροής και τη μεταβλητότητά της, υπερβαίνοντας κατά πολύ τις επιδόσεις του απλού γραμμικού μοντέλου. Ειδικότερα, το LSTM ήταν σε θέση να μοντελοποιήσει την έντονα μη γραμμική φύση του υδρολογικού κύκλου και να εντοπίσει την γενικότερη τάση της χρονοσειράς, έχοντας όμως δυσκολία στην πρόβλεψη των πιο ακραίων φαινομένων.
- **Ανώτερη απόδοση στη λεκάνη του Βοιωτικού Κηφισού / Υλίκης με υβριδικό μοντέλο:** Στη λεκάνη του Βοιωτικού Κηφισού / Υλίκης, το LSTM απέδωσε πολύ καλά, ειδικά όταν έγινε συνδυασμός 2 LSTM σε ένα υβριδικό μοντέλο και αξιοποιήθηκαν δύο διαφορετικές τεχνικές τυποποίησης των δεδομένων εισόδου και εξόδου. Ο συνδυασμός προηγμένων μεθόδων όπως τα LSTM με κλασικές τεχνικές δημιουργίας υβριδικών μοντέλων δύναται να βελτιώσει σημαντικά τις προβλέψεις, ενισχύοντας τις επιδόσεις του LSTM και καθιστώντας το ακόμα πιο αξιόπιστο για επίλυση προβλημάτων στον χώρο της Υδρολογίας.

Προτάσεις

- **Πιο διαδεδομένη χρήση των LSTM για την πρόβλεψη απορροής και συνδυασμός με μεθόδους επαύξησης δεδομένων:** Τα μοντέλα LSTM έχουν σημαντικό δυναμικό για την πρόβλεψη χρονοσειρών απορροής, ιδιαίτερα όταν συνδυάζονται με μεθόδους επαύξησης των δεδομένων, όπως η παραγωγή συνθετικών χρονοσειρών, και με άλλες στατιστικές τεχνικές όπως οι συζεύξεις (corulas). Αυτός ο συνδυασμός επιτρέπει την καλύτερη δυνατή αξιοποίηση των διαθέσιμων δεδομένων και βελτιώνει την απόδοση των μοντέλων στην πρόβλεψη, κάτι που είναι ιδιαίτερα χρήσιμο για τη διαχείριση των υδάτινων πόρων σε περιοχές όπως η Αττική.
- **Περαιτέρω βελτιστοποίηση των LSTM και δημιουργία υβριδικών μοντέλων:** Η περαιτέρω βελτιστοποίηση των μοντέλων LSTM μπορεί να γίνει με την προσαρμογή της ανατομίας τους (π.χ. αριθμός στρωμάτων και νευρώνων, αλλαγή υπερπαραμέτρων) και την εφαρμογή διαφορετικών τεχνικών τυποποίησης και προεπεξεργασίας δεδομένων. Επιπλέον, η δημιουργία υβριδικών μοντέλων που συνδυάζουν LSTM με κλασικά μοντέλα πρόβλεψης, άλλα νευρωνικά δίκτυα ή και σύνθετα μοντέλα βροχής – απορροής πιθανόν να καταφέρει ακόμα καλύτερες επιδόσεις, προσφέροντας ακριβή αποτελέσματα χωρίς σημαντική αύξηση του υπολογιστικού κόστους, το οποίο μπορεί να αυξηθεί εκθετικά.



Ευχαριστώ για την προσοχή σας!