Space-time analysis of meteorological parameters with geostatistical and machine learning methods

Vasiliki Agou

School of Mineral Resources Engineering

Technical University of Crete

Advisor: Prof. Dionissios T. Hristopulos

This dissertation is submitted for the degree of

Doctor of Philosophy

July 2023

“To my parents who did not get to see this adventure”

“A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Geotechnology & Environment, Department of Mineral Resources Engineering, Technical University of Crete”

Examination Committee:

Professor Dionissios T. Hristopulos (Technical University of Crete, School of Electrical and Computer Engineering): Supervisor and Chair of the examination committee

Professor Panagiotis Partsinevelos (Technical University of Crete, School of Mineral Resources Engineering): Member of the advisory committee

Professor George P. Karatzas (Technical University of Crete, School of Chemical and Environmental Engineering): Member of the advisory committee

Associate Professor Anastassia Baxevani (University of Cyprus, Department of Mathematics and Statistics): Member of the examination committee

Associate Professor Tryfonas Daras (Technical University of Crete, School of Chemical and Environmental Engineering): Member of the examination committee

Professor Nikolaos Nikolaidis (Technical University of Crete, School of Chemical and Environmental Engineering): Member of the examination committee

Assistant Professor Emmanouil Varouchakis (Technical University of Crete, School of Mineral Resources Engineering): Member of the examination committee

Acknowledgements

My most profound appreciation goes to my supervisor Prof. Dionissis Hristopulos for showing confidence in me by giving me the chance to work next to him all these years. Also for his dedicated support, guidance, feedback and overall insights throughout my doctoral studies. I could not have imagined having a better advisor and mentor for my Ph.D. studies. I also could not have undertaken this journey without my defense committee, who generously provided knowledge and expertise, insightful comments and encouragement. My committees' vast wisdom and wealth of experience have inspired me throughout my studies.

Special thanks to my colleagues at the Geostatistics Laboratory in the School of Mineral Resources Engineering at the Technical University of Crete, who have supported me and had to put up with my stresses and frustrations for the past years! In addition, I would like to thank Prof. Valérie Monbet (Professor at University Rennes 1, Department of Mathematics) for introducing me to the ERA5 data.

I would like to express my gratitude to all my friends, especially Ivi, Andreas, Michaela, Jaro and Růža for their friendship and support for all these years. Special thanks goes to my partner Sarantis for putting up with me being sat in the office for hours on end, and for providing guidance and a sounding board when required – I simply couldn't have done this without you. I would also like to thank my cats for all the entertainment and emotional support.

Last but not least, I would like to thank my family. Words cannot express how grateful I am for their sacrifices, support and encouragement. Without them, none of this would have been meaningful.

Financial Support

The research in this thesis was partially supported by the project "DES^{2}iRES: DESign of DESalination systems based on optimal usage of multiple Renewable Energy Sources" (ERANETMED_NEXUS-14-049) implemented within the framework of the ERA-NET Initiative "EURO-MEDITERRANEAN Cooperation through ERANET joint activities and beyond (ERANETMED)" under the topic ERANETMED Energy-Water nexus. The project DES^{2}iRES was funded by the Operational Program "Competitiveness, Entrepreneurship and Innovation 2014-2020" (co-funded by the European Regional Development Fund) and managed by the General Secretariat of Research and Technology, Ministry of Education, Research, and Religious Affairs under the project DES^{2}iRES (T3EPA-00017) of the ERAnet, `ERANETMED_NEXUS-14-049`.

Additional funding was provided by the project "Gaussian Anamorphosis with Kernel Estimators for Spatially Distributed Data and Time Series and Applications in the Analysis of Precipitation" (MIS 5052133), which was co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning 2014-2020».

Technical University of Crete, 2023

Vasiliki D. Agou

Abstract

Technological advancements have increased the availability of spatiotemporal data. However, meteorological data are usually non-Gaussian and correlated in space and time. In this dissertation, state-of-the-art geostatistical and machine-learning methodologies were utilized to analyze large-scale non-Gaussian meteorological space-time data. We carried out a series of numerical investigations utilizing 26 surface variables from the ERA5 reanalysis data sets collected for 65 grid locations on the island of Crete, Greece. The data sets correspond to multiple temporal scales (hourly to annually) and span the period from 1979 until 2019.

Four distinct approaches were implemented for the analysis of the meteorological parameters:

1. The ERA5 data set was used for the estimation of the standardized precipitation index (SPI) and the standardized precipitation evapotranspiration index (SPEI) to reveal the spatiotemporal patterns of drought in Crete.
2. Gaussian Anamorphosis with Hermite polynomials (GAH) was employed to transform non-Gaussian precipitation data into normally distributed variables. Ten processing scenarios were investigated and their performance with respect to spatial interpolation (based on Ordinary kriging) was evaluated. The scenarios include the application or exclusion of GAH with varying polynomial degrees, the utilization of either the exponential or Spartan variogram models, and the incorporation or omission of Monte Carlo simulations.

3. Twelve machine learning (ML) techniques were compared for the classification of precipitation data into eight classes. Twenty-six (26) numerical and categorical variables were used in a spatiotemporal predictive framework for precipitation. Due to pronounce class imbalance (dominance of “no rain” events), we first divided the data into two classes that represent the absence or occurrence of precipitation events. Then, the occurrence data set was split in five different classes to characterize the intensity of precipitation events.
4. Finally, we applied the Stochastic Local Interaction (SLI) model to perform temporal (precipitation, temperature and solar radiation) and spatiotemporal (precipitation and temperature) estimation of missing values (data gaps).

The most important conclusions derived in this dissertation are as follows:

1. The dry climate of Crete was confirmed by the estimation of the SPI and SPEI drought indices. It was found that the eastern part of the island is more prone to desertification than the northwestern part. Moreover, a temperature-inclusive drought index was shown to be more appropriate than a purely precipitation-based index for the study area.
2. Using higher-order (35 versus 20) polynomials in GAH has little effect on the cross-validation results for the monthly total precipitation data. In addition, the incorporation of Monte Carlo simulations does not universally improve the statistical measures.
3. With respect to the classification of hourly precipitation data, the method of Random Forests (Bagged Ensemble Trees) performs best for both the “Binary” (“rain” versus “no rain”) and the “Only Rain” classification cases.
4. SLI is a competitive method for interpolating large temporal and spatiotemporal data since it is fast and it performed very

well (compared to nearest-neighbor interpolation) across all the different hourly data sets (temperature, precipitation, and solar radiation).

The present study investigates a variety of methodological approaches for the analysis of non-Gaussian, large-scale meteorological variables. It provides an extensive analysis of precipitation, temperature, and solar radiation for the island of Crete using the ERA5 reanalysis data set. The meteorological data used involve multiple timescales. Two drought indices are evaluated and compared in order to assess the effect of warming trends on drought events. Various data processing scenarios that combine GAH, kriging interpolation and bootstrapping are studied and assessed. In addition, a comparison of twelve machine learning methods for the classification of precipitation data supported by 26 meteorological variables is conducted. Lastly, the computationally efficient SLI models are herein applied for the first time to spatiotemporal precipitation and solar radiation data.

Περίληψη

Οι τεχνολογικές εξελίξεις σε κλάδους όπως η τηλεπισκόπηση και το crowd-sourcing έχουν οδηγήσει σε αύξηση των διαθέσιμων χωροχρονικών δεδομένων. Τα πιο πρόσφατα διαθέσιμα δεδομένα τηλεπισκόπησης, όπως τα ERA5, έχουν αποδειχθεί πιο ακριβή από προηγούμενες συλλογές και επιτρέπουν την διερεύνηση σε περιοχές όπου το επίγειο δίκτυο σταθμών καταγραφής είναι περιορισμένο. Ωστόσο, η αποτελεσματική και ακριβής ανάλυση χωροχρονικών μετεωρολογικών δεδομένων εμπεριέχει διάφορες δυσκολίες. Κατά κανόνα τέτοιου είδους δεδομένα παρουσιάζουν εγγενείς συσχετίσεις μεταξύ χωρικών και χρονικών διαστάσεων και χαρακτηρίζονται από μη-Γκαουσιανές ιδιότητες.

Η ανάλυση χωροχρονικών δεδομένων για μια συγκεκριμένη περιοχή είναι ζωτικής σημασίας, διότι οι παρούσες τοπικές συνθήκες είναι ικανές να επηρεάσουν και να μεταβάλουν σημαντικά υποθέσεις οι οποίες προκύπτουν από μεγαλύτερης έκτασης μοντέλα. Συνεπώς, εμπειρικά μοντέλα που βασίζονται στην ανάλυση δεδομένων είναι απαραίτητα για ακριβείς προβλέψεις. Η ανάλυση τέτοιων μεταβλητών έχει διερευνηθεί εκτενώς μέσω γεωστατιστικών προσεγγίσεων, ωστόσο, πολλές φορές η εξάρτηση τους στην Γκαουσιανή υπόθεση δημιουργεί περιορισμούς. Ως εκ τούτου, είναι σημαντικό να αναπτυχθούν υβριδικές μεθοδολογίες οι οποίες θα αξιοποιούν τα πλεονεκτήματα διαφόρων διακριτών τεχνικών.

Ο πρωταρχικός στόχος αυτής της διατριβής είναι να διερευνήσει διαφορετικές μεθοδολογίες για χωροχρονική μοντελοποίηση εκτεταμένων και μη-Γκαουσιανών χωροχρονικών συνόλων δεδομένων. Χρησιμοποιήθηκαν προηγμένες γεωστατιστικές μεθοδολογίες και μεθοδολογίες μηχανικής μάθησης για την ανάλυση μετεωρολογικών δεδομένων που αποκλίνουν από τις συμβατικά χρησιμοποιούμενες παραμετρικές κατανομές.

Πραγματοποιήσαμε μια σειρά αριθμητικών πειραμάτων για το νησί της Κρήτης, χρησιμοποιώντας 26 επιφανειακές μεταβλητές από το ERA5 σύνολο δεδομένων οι οποίες συλλέχθηκαν για 65 τοποθεσίες σε κανονικό πλέγμα. Τα σύνολα δεδομένων αντιστοιχούν σε πολλαπλές χρονικές κλίμακες (ωριαία έως ετήσια) και καλύπτουν την περίοδο από το 1979 έως το 2019.

Τέσσερις διακριτές προσεγγίσεις εφαρμόστηκαν για την ανάλυση των μετεωρολογικών παραμέτρων:

1. Το σύνολο δεδομένων ERA5 χρησιμοποιήθηκε για τον προσδιορισμό των χαρακτηριστικών ξηρασίας για το νησί της Κρήτης. Η εκτίμηση πραγματοποιήθηκε χρησιμοποιώντας τους δείκτες ξηρασίας standardized precipitation index (SPI) και standardized precipitation evapotranspiration index (SPEI) για έξι χρονικές κλίμακες. Σκοπός της παρούσας μελέτης είναι να διαπιστωθεί η επίδραση της αύξησης της θερμοκρασίας στην συχνότητα των φαινομένων ξηρασίας.

2. Προκειμένου να καταστεί δυνατή η βέλτιστη απόδοση των κλασικών γεωστατιστικών μεθόδων όπως το kriging, χρησιμοποιήθηκε η Γκαουσιανή Αναμόρφωση με Ερμιτιανά πολυώνυμα (GAH) για τη μετατροπή των μη-Γκαουσιανών δεδομένων βροχόπτωσης σε κανονικά κατανεμημένα. Διερευνήθηκαν δέκα σενάρια επεξεργασίας και αξιολογήθηκε η απόδοσή τους σε σχέση με τη χωρική παρεμβολή (με βάση το κανονικό kriging). Τα σενάρια περιλαμβάνουν τη χρήση ή μη της GAH με ποικίλους πολυωνυμικούς βαθμούς, τη χρήση είτε του εκθετικού είτε του Σπαρτιάτικου μοντέλου βαριογράμματος και την ενσωμάτωση ή παράλειψη προσομοιώσεων Monte Carlo.

3. Χρησιμοποιήθηκαν δώδεκα τεχνικές μηχανικής μάθησης (ML) για την ταξινόμηση των δεδομένων βροχόπτωσης σε οκτώ τάξεις. Οι μέθοδοι περιλαμβάνουν τα fine, medium, και coarse classification trees, linear, quadratic, cubic, fine Gaussian, medium Gaussian, και coarse Gaussian Support Vector Machines, Boosted Ensem-

ble trees, Bagged Ensemble trees, και Ensemble RUSBoosted trees. Διερευνήθηκε η επίδραση είκοσι έξι (26) βοηθητικών μεταβλητών (ποιοτικές και ποσοτικές) στα πλαίσια της χωροχρονικής ταξινόμησης των βροχοπτώσεων. Λόγω της ανισοκατανομής των δεδομένων (κυριαρχία των γεγονότων «χωρίς βροχή»), το σύνολο χωρίστηκε περαιτέρω σε δύο ξεχωριστά σύνολα δεδομένων. Το πρώτο σύνολο περιέχει δύο κλάσεις, οι οποίες καθορίζονται από ένα ορισμένο κατώφλι και χαρακτηρίζεται ως το "Binary" σύνολο δεδομένων, το οποίο ταξινομεί την απουσία ή την ύπαρξη βροχόπτωσης. Το δεύτερο σύνολο αποτελείται αποκλειστικά από τις τάξεις που υπερβαίνουν το όριο κατωφλίου (πέντε τάξεις) και χαρακτηρίζεται ως το "Only Rain" σύνολο δεδομένων, το οποίο ταξινομεί την ένταση των συμβάντων βροχόπτωσης.

4. Τέλος, χρησιμοποιήσαμε τα Στοχαστικά μοντέλα Τοπικών Αλληλεπιδράσεων (Stochastic Local Interaction models, SLI) για την πλήρωση κενών σε δεδομένα βροχόπτωσης, θερμοκρασίας και ηλιακής ακτινοβολίας. Στη μελέτη μας χρησιμοποιήσαμε τη μέθοδο SLI για ανάλυση στο χρόνο και στο χωροχρόνο. Οι αναλύσεις στο χρόνο αποτελούνταν από 2 535 χρονοσειρές για τη βροχόπτωση, 2 600 χρονοσειρές για τη θερμοκρασία και επιπλέον 2 600 χρονοσειρές για δεδομένα ηλιακής ακτινοβολίας. Οι χωροχρονικές αναλύσεις αναφέρονται σε εκτιμήσεις βροχόπτωσης και θερμοκρασίας. Το χωροχρονικό σύνολο βροχόπτωσης περιλαμβάνει 10 920 ωριαίες τιμές που αντιστοιχούν σε επτά συνεχόμενες ημέρες και το σύνολο δεδομένων θερμοκρασίας περιέχει 10 920 ωριαίες τιμές για το ίδιο χρονικό διάστημα.

Τα σημαντικότερα συμπεράσματα που προκύπτουν σε αυτή τη διατριβή είναι:

1. Το ξηρό κλίμα της Κρήτης επιβεβαιώθηκε μέσω της εκτίμησης των δεικτών ξηρασίας SPI και SPEI. Το ανατολικό τμήμα του νησιού βρέθηκε να είναι πιο επιρρεπές στην ερημοποίηση από το βορειοδυτικό τμήμα. Επιπλέον, το Ηράκλειο πλήττεται σοβαρά από την

άνοδο της θερμοκρασίας και την αυξημένη εξατμισοδιαπνοή λόγω της κλιματικής αλλαγής. Οι αποκλίσεις μεταξύ των δεικτών που παρουσιάζονται στις τελευταίες δεκαετίες υποδηλώνουν ότι ο δείκτης ξηρασίας που περιλαμβάνει τη θερμοκρασία είναι πιο κατάλληλος για την εκτίμηση των φαινομένων ξηρασίας στην περιοχή μελέτης.

2. Η αύξηση της τάξης των πολυωνύμων στην εφαρμογή GAH έχει μικρή επίδραση στην ακρίβεια των αποτελεσμάτων για τα μηνιαία δεδομένα βροχόπτωσης. Επιπλέον, η ενσωμάτωση προσομοιώσεων δεν βελτιώνει απαραίτητα τα αποτελέσματα. Το Σπαρτιάτικο μοντέλο συνδιακύμανσης βρέθηκε να είναι πιο κατάλληλο για την αναμόρφωση χωρίς τις προσομοιώσεις, ενώ το εκθετικό μοντέλο βρέθηκε να είναι πιο κατάλληλο για τα σενάρια που ενσωματώνουν τις προσομοιώσεις.

3. Στην εφαρμογή των μοντέλων μηχανικής μάθησης για την ταξινόμηση της ωριαίας βροχόπτωσης, τα Τυχαία Δάση (Random Forests ή Bagged Ensemble Trees) έχουν την καλύτερη ακρίβεια και στα δύο σύνολα δεδομένων ("Binary" και "Only Rain"). Ένα υβριδικό μοντέλο μπορεί να είναι πιο κατάλληλο για το σύνολο δεδομένων "Only Rain" λόγω του πολύ μικρού δείγματος δεδομένων στις υψηλότερες τάξεις.

4. Το μοντέλο SLI είναι μια αποτελεσματική μέθοδος παρεμβολής μεγάλου συνόλου χρονικών και χωροχρονικών δεδομένων, καθώς δεν απαιτεί την αντιστροφή μεγάλων πινάκων ή εκτεταμένη προεργασία. Το SLI είναι πολύ αποτελεσματικό για όλα τα διαφορετικά σύνολα δεδομένων (θερμοκρασία, βροχόπτωση, ηλιακή ακτινοβολία) σε σύγκριση με την μέθοδο του κοντινότερου γείτονα (Nearest Neighbor interpolation).

Η παρούσα μελέτη διερευνά ένα εύρος μεθοδολογικών προσεγγίσεων για την ανάλυση μη-Γκαουσιανών, μεγάλης κλίμακας μετεωρολογικών μεταβλητών. Παρέχει μία εκτενή ανάλυση για δεδομένα βροχόπτωσης, θερμοκρασίας και ηλιακής ακτινοβολίας για το νησί της Κρήτης, χρησιμοποιώντας τα δεδομένα ERA5 σε πολλαπλές χρονικές κλίμακες. Επι-

πλέον, δύο δείκτες ξηρασίας συγκρίνονται με σκοπό να αξιολογηθεί η επίδραση της αύξησης της θερμοκρασίας λόγω κλιματικής αλλαγής στα φαινόμενα ξηρασίας στο νησί. Ποικίλα σενάρια επεξεργασίας τα οποία χρησιμοποιούν την GAH σε συνδυασμό με την εκτίμηση kriging και προσομοιώσεις Monte Carlo διερευνώνται και αξιολογούνται. Επιπρόσθετα, πραγματοποιείται σύγκριση δώδεκα μεθόδων μηχανικής μάθησης για την ταξινόμηση των δεδομένων βροχόπτωσης που ενισχύονται από πληροφορία από 26 μετεωρολογικές μεταβλητές. Τέλος, τα SLI μοντέλα εφαρμόζονται εδώ για πρώτη φορά σε δεδομένα χωροχρονικής βροχόπτωσης και ηλιακής ακτινοβολίας.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In recent years due to technological advancements, the availability of large spatiotemporal data sets has increased exponentially. Sources of these data sets include remote sensing and crowd-sourcing. Especially for the meteorological parameters from remote sensing sources, the data values have become more representative of the true ground values compared to the past collections and help vastly in cases where the ground station network is sparse. However, in order to understand the spatiotemporal meteorological values there are specific issues that need to be addressed. Usually, spatiotemporal data exhibit intrinsic correlations between space and time and they are non-Gaussian. There are many reasons why we need to analyze spatiotemporal data. Many times the variable that we need to understand is qualitative, other times a theoretical model able to describe the variable does not exist. Specifically for the analysis of meteorological parameters, the need lies in the significance of comprehending the patterns that determine the weather and climate. This is important because while general fundamental equations are established for the characterization of such phenomena, each area of interest experiences different patterns driven by local conditions. In this case, predictions are dependent on empirical models at the core of which is data analysis. For the analysis of variables that are correlated in space and time, such as meteorological parameters, classical geostatistical methodologies have been applied widely [Goovaerts, 2000; Kyriakidis and Journel, 1999; Wackernagel, 2003; Webster and Oliver, 2007]; however, they have several shortcomings such as their dependence on the Gaussian assumption. Additionally, because of the abundance of data available from remote sensing, and the number of correlated variables, their size has grown considerably. Machine learning methods that are able to handle big data can help in revealing the spatiotemporal patterns [Hastie et al., 2009; Huntingford et al., 2019] without the need for a deep understanding of the mathematical background, thus allowing non-experts to use them.

The stability of the climate in the lower layers of the atmosphere has allowed the viability of the human species as well as all forms of life on Earth. These conditions are created and directly depend on various meteorological parameters. Key parameters include precipitation, temperature, solar radiation, wind speed, and wind direction, humidity, and air pressure. Observations of these variables are used by experts to predict the weather and usually they do not follow the Gaussian distribution. Water is essential to life, household uses, industrial activities, agriculture, transportation, and virtually all processes that sustain life. Consequently, water resources are an integral part of life on Earth.

To understand and early acknowledge changes in the freshwater reserves either on a global or a local scale one must first estimate how precipitation is distributed in space and time. Improved models of spatiotemporal variations of precipitation are important for hydrological and climate studies as well as for water resources management [Agou et al., 2019; Kavetski et al., 2006; Varouchakis et al., 2018]. In particular, there is strong interest in the Mediterranean region where local economies depend on scarce water resources and climate change is expected to adversely affect water availability [Cannarozzo et al., 2006; Coscarelli and Caloiero, 2012; IPC, 2013].

Because of the uncertainties involved (spatial and temporal variability), probabilistic approaches are required to enable water resources managers to analyze risk under scenarios of climate change. Investigation of precipitation at fine temporal resolution, such as monthly [Hellström et al., 2001; Mendez et al., 2020] or even daily scales [Black, 2009; Chu et al., 2010; Zhang et al., 2011] is preferable, considering variations in seasonal patterns [Portmann et al., 2009; Vera et al., 2006] and the probability of extreme events [Kjellström et al., 2007; O'Gorman and Schneider, 2009; O'Gorman, 2015]. However, because of the computational load, the lower availability of data at finer temporal scales, and the mathematical challenges involved in treating fast-changing values, most of the studies from previous years involve lower resolution data sets [Kovats et al., 2014].

Geostatistical approaches have been successfully employed in different environmental and Earth sciences disciplines. One of the main advantages of geostatistical methods is their ability to handle sparse measurements [Agou et al., 2019; Varouchakis and Hristopulos, 2013]. Hence, they can provide space-time predictions for variables with environmental and socio-economic importance supplemented with estimates of the uncertainty of the results. Geostatistical methods are widely used nowadays, and they remain a big part of the proposed methodologies. However, in the last decades, due to the explosion in computing power, more demanding approaches are investigated such as Neural Networks [Moustris et al., 2011], simulations [Richardson, 1981], and hybrid approaches [Li et al., 2012] that incorporate more than one method.

Standard geostatistical techniques often rely on the normality assumption, meaning that the target variable must approximate the normal distribution. Nonetheless, this is not the case in environmental variables such as precipitation, which are oftentimes incomplete, and highly skewed [Agou et al., 2019]. Depending on the investigated timescale and the geographical location, the distribution of the data set could vary greatly. A few of the available options that address non-Gaussianity include the removal of a trend function or the application of a normalization technique such as the Box-Cox transform [Box and Cox, 1964]. The simplest normalization transforms, which are routinely used, are not always appropriate for the analyzed data. For example, in the case of precipitation height, because of the existence of zero values in the data set, the application of the Box-Cox or the logarithm transform without modifications is unsuited.

This thesis is motivated by the need for accurate interpolation methodologies that can help to determine the spatiotemporal variability of parameters which do not necessarily follow known probability distribution models [Pavlides et al., 2022]. ERA5[1] data became recently available to the public. They include val-

[1] ERA5 is a climate reanalysis data set (5th generation) from ECMWF (the European Centre for Medium-Range Weather Forecasts) with a spatial resolution of 0.25°(31 km), lower time

ues of meteorological parameters at fine temporal resolution across threatened, due to climate change geographical areas. The parameters available in ERA5 data include many variables relevant to precipitation estimation, such as temperature and cloud cover, they are also more accurate than their predecessor ERA-Interim [Hassler and Lauer, 2021] and they have many characteristics of the problems commonly found when analyzing spatiotemporal data such as the big data size, the non-Gaussianity and the hidden correlations in the data structures. Hence, ERA5 data are the focus of this research. Herein, we introduce spatiotemporal methodologies applied to reanalysis products, such as precipitation, temperature and solar radiation. It is quite demanding to capture the distribution of non-Gaussian meteorological data, e.g., precipitation, due to its complex spatiotemporal variability and physical mechanisms. Additional motivations of this thesis are the extreme events, the irregularities found in multiple meteorological variables, the changing climatic conditions, and the risk of desertification in various Mediterranean basins. All these reasons emphasize why a tool for accurate spatiotemporal modeling of parameters that vary depending on the latitude, the longitude and the timescale is crucial. Furthermore, the temperature increase over the last decades has affected the estimation of the drought conditions over an area. Localized estimations of drought events can be accomplished by the application of techniques that are able to fill in the missing gaps effectively which are usually present in ground station data.

In the following sections, several methodologies will be reviewed, however, every method has its limitations and drawbacks. There is yet to find the "perfect" algorithm for modeling data that do not follow parametric distributions, and we are still far from achieving that perfection. We aim to take steps towards that ultimate goal.

resolution at 1 hour, 137 vertical levels from the surface up to a height of 80 km into the atmosphere, and is spanning the period 1950 to present (available for use in 2020) [Copernicus Climate Change Service C3S, 2018]. Reanalysis is a systematic approach that employs data assimilation and numerical methods to generate weather and climate products over high-resolution grids [Dee et al., 2016].

1.2 Objectives

The main objective of this thesis is to provide various methodologies for space-time modeling of potentially large and non-Gaussian space-time data sets. In order to test progress towards this objective we use an extensive set of meteorological reanalysis data for the island of Crete, Greece. The specific goals pursued in this thesis are as follows:

1. We aim to show that geostatistical methods can be used to investigate the spatial and temporal variability of non-Gaussian data (e.g., precipitation) and provide a more user-friendly formulation, which does not require significant pre-processing. Most of the classical geostatistical methodologies such as kriging rely on the normality assumption, which is not met in precipitation data at fine temporal resolutions. According to the relevant bibliography [Andreou, 2022; Baxevani and Lennatsson, 2015; Gellens, 2002; Koutsoyiannis, 2004; Li et al., 2013; Papalexiou et al., 2018; Rho and Kim, 2019; Shoji and Kitaura, 2006; Ye et al., 2018], precipitation height can be approximately modeled by a variety of model distribution functions (pareto, GEV, gamma, lognormal etc.). The appropriate distribution is usually estimated from the data set that corresponds to a specific area and has a fixed temporal step. This introduces several steps prior to the actual estimation of the variable of interest, especially when the chosen approach relies on the Gaussian hypothesis, such as kriging methods. On that note, the objective was to incorporate a methodology that omits the fitting to a parametric distribution step but its application will result in a field of approximately normally distributed values, and at the same time it can be consistently applied regardless of the time step and the optimal fitted distribution for the specified data set [Agou et al., 2022; Pavlides et al., 2022]. We implement a normalization technique based on the Hermite polynomials (GAH) to drive high-resolution precipitation data to meet these requirements [Hristopulos, 2020; Wackernagel, 2003]. GAH uses the Hermite polynomials for the transformation of the data-based cumulative distribution function (CDF) to a Gaussian CDF. The same approach can be applied to other high resolution data that do not approximate the Gaussian distribution. Additional

to kriging methods, we utilized stochastic simulation in order to capture the entire variability of precipitation for the island of Crete.

2. Another goal was to use the Stochastic Local Interaction (SLI) methodology for the interpolation of meteorological variables in the space-time domain [Hristopulos and Agou, 2020]. Running large data sets with classical geostatistical methodologies such as kriging demands the covariance matrix inversion which is computationally expensive, in terms of computational power, memory and time. Consequently, this limits the number of case studies that qualify for such applications. Additionally, in methodologies such as the nearest neighbors interpolation, the number of neighbors should always be larger than the number of missing ones in a row, which is not needed in the case of the SLI method. Herein, we show that the SLI methodology proposed by Hristopulos [2015b], an alternative to kriging for large data sets, which uses local interactions of the data and the joint probability density function defined by energy functionals is sufficient for the interpolation of spatiotemporal meteorological data. We also show that the method gives improved predictions to those acquired by classic geostatistical methods.

3. We apply several machine learning methods (fine, medium, and coarse classification trees, linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian Support Vector Machines, Boosted Ensemble trees, Bagged Ensemble trees, and Ensemble RUSBoosted trees) for the classification of an imbalanced response variable based on various auxiliary meteorological variables in order to identify the most accurate among them [Breiman, 1996; Chase et al., 2022; Freund and Schapire, 1996; Opitz and Maclin, 1999; Rokach and Maimon, 2014; Rolnick et al., 2022; Seiffert et al., 2008; Wu et al., 2008]. As mentioned previously, the pre-processing steps for the interpolation of non-Gaussian variables are intricate and time-consuming. Furthermore, previous machine learning applications that include but not focus on the area of interest are scarce and contain information from a small number of parameters [Moustris et al., 2011; Papacharalampous et al., 2018]. We use machine learning methods because they

do not rely on parametric assumptions and are able to improve the estimation by incorporating supplementary information. For our case study, the response variable is precipitation intensity and occurrence and the auxiliary variables are twenty six high-resolution meteorological parameters. We identified the most relevant parameters to the amount of precipitation and then we used them to evaluate the performance of the models to binary and non-binary classification of precipitation for the island of Crete. The binary classification aims to identify if an event is without precipitation (zero) or with precipitation (one), meaning that in this case, we classify the occurrence of precipitation. The non-binary classification aims to classify the precipitation events into five separate distinctive classes that correspond to intensity ranges.

4. Lastly, we aim to investigate the drought characteristics of the area of interest based on the newly available ERA5 data and determine if the effects of climate change are noticeable by incorporating one index that accounts for temperature and one that does not. In previous studies, usually, one drought index is applied and most of the time it is calculated over one timescale [Koutroulis et al., 2011; Tsakiris et al., 2007b; Vrochidou and Tsanis, 2012]. Rarely, the use of multiple indices for various timescales is seen in practice. This prevents the comparison of widely used indices and the effects of precipitation and temperature trends in short and long-term drought conditions. In this research, we use two drought indices for the estimation of upcoming (hydrological or meteorological) droughts, namely the Standardized Precipitation Index (SPI) and the Standardized Evapotranspiration Index (SPEI). The estimation of the drought indices aims to identify the spatiotemporal character of drought events in Crete.

Since we performed temporal and spatial precipitation analysis based on the proposed methodologies, the comparison of the spatiotemporal results to the spatial and the temporal results was natural. This thesis may prove itself valuable to policymakers, further environmental studies, agricultural management and planning.

1.3 Innovation

The present study addresses the analysis of meteorological data on the island of Crete using state-of-the-art geostatistical and machine learning tools. The methodologies that we present include stochastic methods and machine learning classification methods for the spatial, temporal, and spatiotemporal analysis of non-Gaussian variables, while the data sets used for proof of concept are ERA5 reanalysis products for the island of Crete, analyzed at different time scales (refer to Section 1.5.8 for details on the variables and the timescales used for each methodology).

1. Stochastic Local Interaction models (SLI) models [Hristopulos, 2015a; Hristopulos and Agou, 2020] are herein applied for the first time to temporal and spatiotemporal precipitation and solar radiation data. The SLI models are inspired by Gaussian field theories and Gaussian Markov random fields. They are based on the creation of correlations generated by interactions between neighboring sites and times and can be used in scattered data compared to the Gaussian Markov random fields. The interactions between neighboring points are expressed in terms of suitably selected weighting functions, which are supplied by kernel functions. In SLI models the correlations are determined employing sparse precision matrices, in addition, there is no need for inversion of large covariance matrices, thus allowing their use even in standard computers for significantly bigger data sets than what would have been possible with classical geostatistical methods.

2. The ERA5 data are used for the first time for an extensive, localized analysis of precipitation, temperature, and solar radiation and for the estimation of the drought characteristics to the extent used herein for the island of Crete. To address the sparsity of the precipitation record commonly found in the ground data for Crete [Agou et al., 2019], we use ERA5 reanalysis products [Copernicus Climate Change Service C3S, 2018]. We focus on the total precipitation for timescales ranging between 1 hour to 1 year at the locations of the ERA5 grid. The use of an extensive set of ERA5 variables (26 parameters are used herein) for the island of Crete has not yet been

used outside the context of General Circulation Models (GCMs). Lastly, even with the ground station data, most of the studies that investigate the patterns of climatological data for the island of Crete usually focus only on a single timescale.

3. We use the Gaussian Anamorphosis with Hermite polynomials (GAH) to transform the non-Gaussian precipitation data to normally distributed in order to permit classic geostatistical methods such as kriging to perform optimally. GAH is not usually used in this context because of the complexities involved in the calculation of the relevant coefficients. We further generate several scenarios with different configurations and we evaluate their performance. The separate elements used for the scenarios include the use or not of the GAH with different polynomial orders, the variogram estimation based on the exponential or the Spartan variogram model, the Ordinary Kriging, and the use or not of Monte Carlo simulations. We generate precipitation estimates and their associated uncertainties across the island using geostatistical methods coupled with Monte Carlo simulation. We employ Monte Carlo simulations because precipitation data in every analyzed timescale does not obey a specific probability distribution, and the data in all cases are non-Gaussian.

4. We estimate and compare several drought indices (SPI, SPEI) based on the ERA5 data. We use these indices to generate finer-resolution maps which can help identify the most prone to desertification areas. To our knowledge, this is the first time that ERA5 data are used for the estimation of drought indices for the island of Crete. Furthermore, the estimation of SPEI for the area of interest has not been published before. Since the length of the data set is quite long (41 years), the presentation of all the results (SPI and SPEI for 1-, 3-, 6-, 9-, 12-, and 24-months) for all the locations (65 locations) is impossible. Therefore, we propose specific years that represent the range of the data, as well as specific locations that are of higher interest due to their geographic location. We suggest the use of the estimated characteristic year, based on the precipitation data, for the comparison of the indices results to the years that recorded the lowest and the highest annual precipitation.

Past research for the island of Crete has focused on a single drought index (usually SPI or a custom index) and a single time scale (3-, 12-, 24-, or 48-months) [Koutroulis et al., 2011; Tsakiris et al., 2007b; Vrochidou and Tsanis, 2012]. In addition, we take advantage of index calculations for multiple timescales (1-, 3-, 6-, 9-, 12-, and 24-months), which allows us to capture in more detail the area's drought characteristics regarding short and long-term patterns.

5. We use several machine learning (ML) methods to study and classify imbalanced precipitation ERA5 data (fine, medium, and coarse classification trees, linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian Support Vector Machines, Boosted Ensemble trees, Bagged Ensemble trees, and Ensemble RUSBoosted trees). We study the impact of several meteorological and hydrological variables in a spatiotemporal predictive framework for precipitation. While it is not the first time that classification methodologies have been applied to precipitation data (for a review paper see Tyralis et al. [2019]), the aforementioned machine learning methods are for the first time applied to an extensive set of 27 meteorological variables (the original 26 variables and one additional categorical variable that signifies the month) for the island of Crete.

1.4 Dissertation Outline

The remainder of this thesis is organized as follows:

- Section 1, Introduction, presents the basic concepts and variables that are commonly used to define the climate, including the definitions of precipitation and drought, temperature, solar radiation and wind. Also, it describes the most commonly used methodologies for the analysis of precipitation data, the problems encountered in reanalysis products, and the importance of the different spatial and temporal scales.
- Section 2, Basic Concepts of Geostatistics, focuses on the main background necessary for the application of geostatistical methodologies, as well as the

description of specific approaches. Specifically, it includes the definitions of random fields, and concepts used to characterize them such as probability distributions, statistical moments such as the covariance function, statistical homogeneity, and isotropy. Additionally, commonly used variogram models are presented, followed by a brief introduction to Kriging methods, simulation methods, Cross-validation techniques, and typically used validation metrics. Relevant bibliography accompanies the methodologies.

- Section 3, Exploratory Data Analysis, introduces the data sets that will be used for evaluating the performance of the different methods and their characteristics. This section also presents exploratory statistical analysis of the main variables and relevant information for the study area.

- Section 4, Estimation of Drought Indices for the Island of Crete, presents a brief description of the drought indices concept, discusses commonly used indices in areas with similar characteristics to the study area, and contains the definitions of the indices (SPI, SPEI, PET) used in this research. Furthermore, this section includes the calculation of the drought indices based on the previously presented precipitation and temperature data for numerous timescales (1-, 3-, 6-, 9-, 12-, and 24-months), finer-resolution maps and finally the results and comparison between the indices.

- Section 5, Gaussian Anamorphosis of Precipitation Data, introduces the formalism of normalization methods. Additionally, the application of Gaussian Anamorphosis with Hermite polynomials coupled with geostatistical simulation (for the estimation of monthly precipitation) is presented and compared in the framework of ten different scenarios.

- Section 6, Space-time Modeling with Machine Learning Methods, briefly introduces widely used classification methodologies focusing on the one known as Random Forests. It describes the problems occurring if the data sets are not uniformly distributed, and presents the application of the methods using 27 environmental variables for the classification of hourly precipitation values. Because our data set is extremely imbalanced, we studied various data splits. Finally, the classification methodologies applied and their results are compared. The Random Forests method was proven to be the most accurate method for precipitation classification for all our test cases.

- Section 7, Stochastic Local Interaction Models, presents the theory behind the Stochastic Local Interactions methodology, its application to temporal and spatiotemporal data sets and the results obtained. In particular, we apply the SLI methodology proposed by Hristopulos [2015b], an alternative to kriging for large data sets, which uses local interactions of the data and the joint probability density function defined by energy functionals, to temporal hourly precipitation, temperature, and solar radiation data, and spatiotemporal hourly precipitation and temperature data. The SLI method gives improved results compared to those acquired by simple interpolation methodologies (nearest-neighbor interpolation) for precipitation and significantly better for the temperature and solar radiation data sets.

- Finally, Section 8, Conclusions presents a general discussion of the results and concluding remarks.

- Appendix A presents the theoretical probability density functions (PDFs) and cumulative density functions (CDFs) of commonly used probability distributions for precipitation data modeling, as well as widely known methodologies used to test the proximity of a data set to the Gaussian (or another) distribution.

- Appendix B presents the summary statistics of the precipitation and temperature data in different timescales.

- Appendix C presents the summary statistics of the monthly precipitation for the dry period as well as the variogram fits with the exponential and the Spartan variogram models for each month for the entire period. Additionally, the validation results of the four scenarios (S1-S4) investigated in Section 5 are presented in detail.

1.5 Preliminaries

Weather is defined by means of the meteorological conditions at any given time. Climate is typically described as the average weather (or more precisely by means of the mean and the variability of pertinent parameters) over a period that can range from months to millions of years. According to the World Meteorological Organization (WMO), averaging meteorological variables over 30 years is considered adequate. The most appropriate variables are measured at the Earth's surface such as temperature, precipitation and wind [Hartmann et al., 2013].

Life, as we know it today, is strongly linked to weather and climate. The prosperity of human societies relies upon the relatively stable climate conditions that Earth has experienced since the ice age. Extreme events occurring with increasing frequency during the last century have raised awareness over climate change. Steps that individuals and societies can take to mitigate the effects of climate change have been proposed. However, the first step to planning any strategy is understanding and thoroughly evaluating the key factors that contribute to the changing climate.

As stated by the Sixth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) [IPCC, 2021], for the last four decades after 1850, every decade is characterized by higher global surface temperature compared to the preceding decade. Specifically, for 2001–2020 global surface temperature was 0.99 °Celsius higher than during the period 1850–1900 (Fig. 1.1). Based on 5 different emission scenarios, the global surface temperature is characterized by an increasing trend until at least 2050. Certain projections suggest that before the end of the 21st century, unless extensive reduction in CO_2 emissions is achieved, the annually-averaged global surface temperature is going to increase by 1.5 °Celsius – 2 °Celsius over the global surface temperature relative to 1850–1900 (Fig. 1.1) [IPCC, 2021].

The estimations for the globally averaged precipitation indicate an increase since 1950 with medium confidence. However, with high confidence it can be said that the frequency and the intensity of heavy precipitation events has increased for areas with adequate observations (Fig. 1.2b). In addition, it is clear that human influence likely contributed to the aforementioned changes with medium

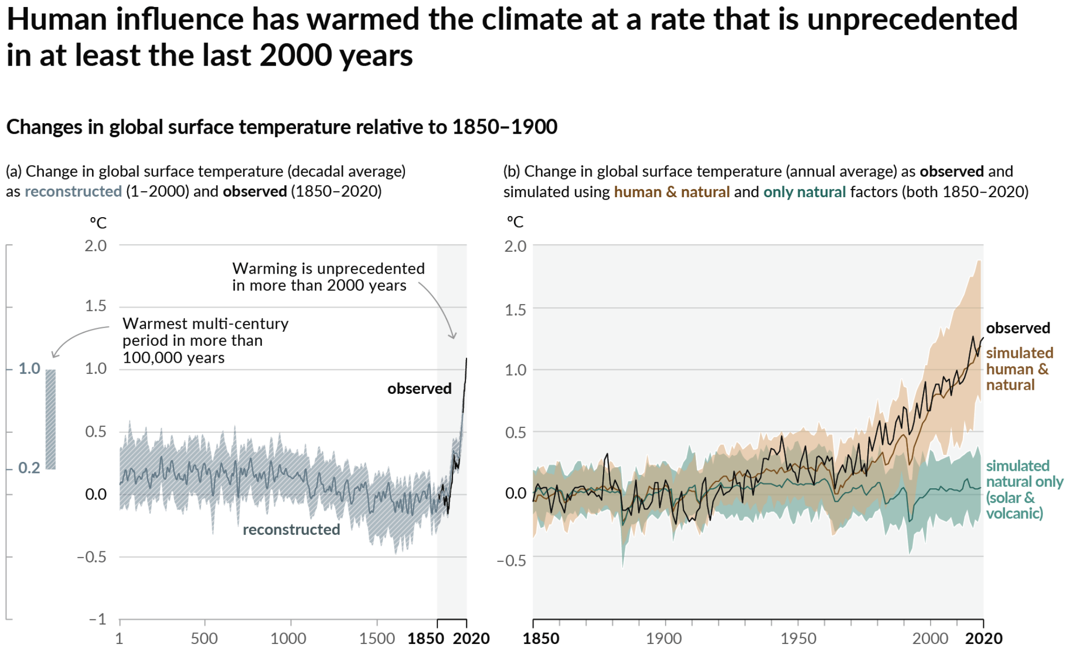
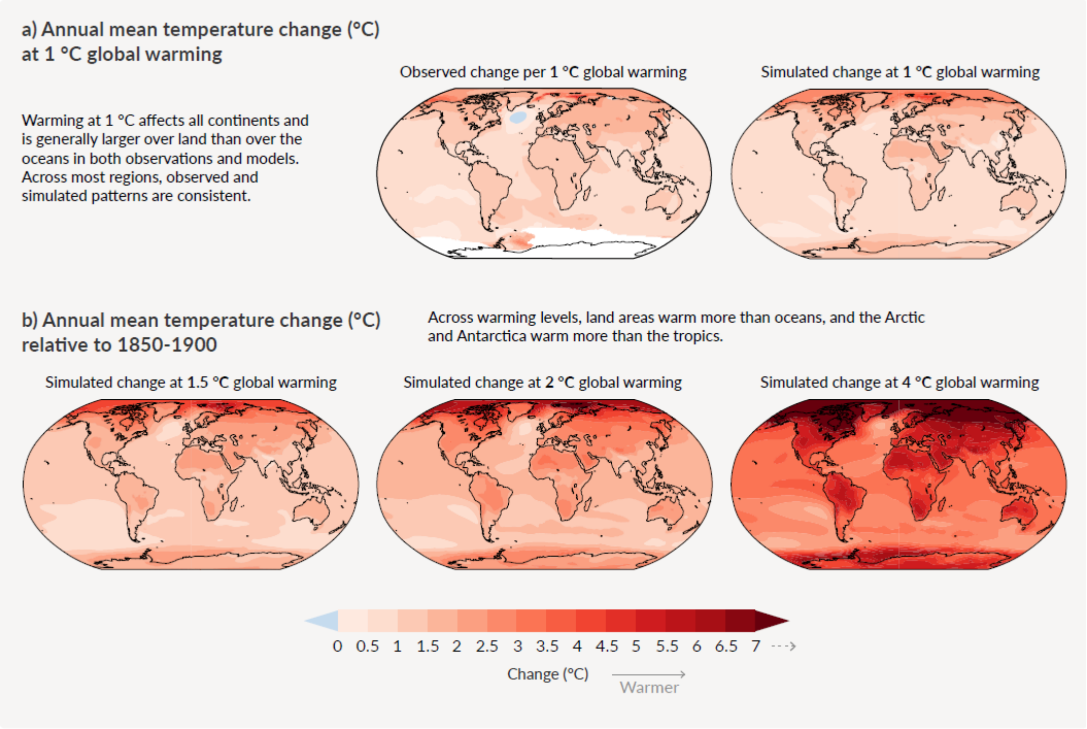


Figure 1.1: History of global temperature change and causes of recent warming. Left: Changes in global surface temperature reconstructed from paleoclimate archives (solid gray line, 1–2000) and direct observations (solid black line, 1850–2020), both relative to 1850–1900 and decadally averaged. The gray shading with white diagonal lines shows the very likely ranges for the temperature reconstructions. Right: Changes in global surface temperature over the past 170 years (black line) relative to 1850–1900 and annually-averaged, compared to climate models. Figure SPM.1 in IPCC [2021]. For a more detailed description, see the original published figure.

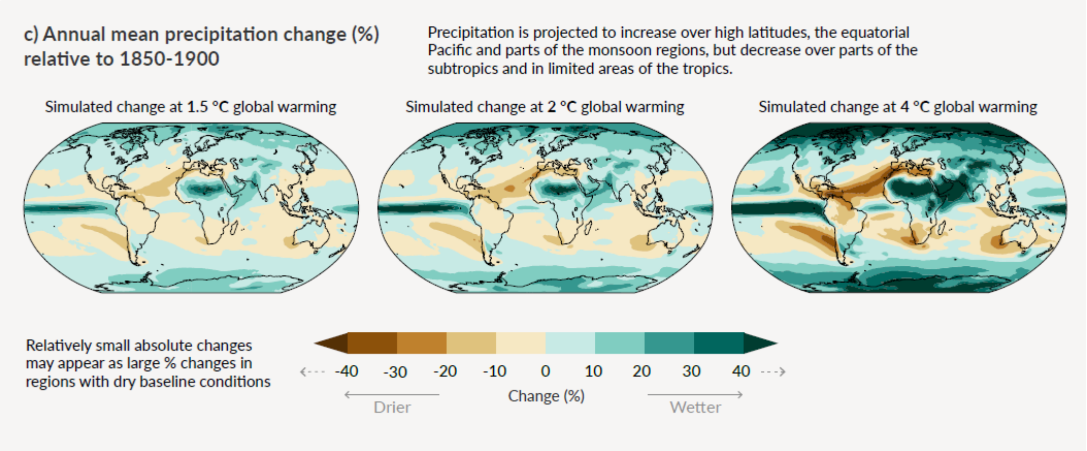
to high confidence [IPCC, 2021].

More specifically, for the Mediterranean region evidence indicates with high confidence that there will be an increase in summer temperature greater than in the global mean, while a reduction of summer precipitation is expected in southern Europe. Additionally, there are projections for expansion in the land area in danger of aridification, and an increase in the number of drought, aridity and fire weather events, which subsequently will affect the agriculture, forestry, and health sectors [IPCC, 2021].

Earlier reports [Kovats et al., 2014] have identified southern Europe as more sensitive than the rest of Europe and probably more severely affected by climate


a) Annual mean temperature change (°C)
at 1 °C global warming
Warming at 1 °C affects all continents and is generally larger over land than over the oceans in both observations and models. Across most regions, observed and simulated patterns are consistent.
Observed change per 1 °C global warming
Simulated change at 1 °C global warming
b) Annual mean temperature change (°C)
relative to 1850-1900
Across warming levels, land areas warm more than oceans, and the Arctic and Antarctica warm more than the tropics.
Simulated change at 1.5 °C global warming
Simulated change at 2 °C global warming
Simulated change at 4 °C global warming
0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5 5.5 6 6.5 7
Change (°C)
Warmer

(a)


c) Annual mean precipitation change (%)
relative to 1850-1900
Precipitation is projected to increase over high latitudes, the equatorial Pacific and parts of the monsoon regions, but decrease over parts of the subtropics and in limited areas of the tropics.
Simulated change at 1.5 °C global warming
Simulated change at 2 °C global warming
Simulated change at 4 °C global warming
Relatively small absolute changes may appear as large % changes in regions with dry baseline conditions
-40 -30 -20 -10 0 10 20 30 40
Change (%)
Drier
Wetter

(b)

Figure 1.2: Changes in annual mean surface temperature, and precipitation. Top: Comparison of observed and simulated annual mean surface temperature change. The left map shows the observed changes in annual mean surface temperature in the period of 1850–2020 per °C of global warming (°C). The local (i.e., grid point) observed annual mean surface temperature changes are linearly regressed against the global surface temperature in the period 1850–2020. Bottom of a): Simulated annual mean temperature change (°C), and b) precipitation change (%) at global warming levels of 1.5°C, 2°C, and 4°C (20-yr mean global surface temperature change relative to 1850–1900). Figure SPM.5 in IPCC [2021]. For a more detailed description, see the original published figure.

change. Recently, with many devastating events ravaging the Mediterranean, such as extensive fires and record temperatures, these effects are very pronounced. For example, in the middle of the summer of 2021 across the Mediterranean, many fires took place, while multiple were occurring synchronously at different locations in Greece, Italy, Spain, Turkey, North Macedonia (Skopje). Temperatures were recorded at 48.8 °Celsius in Sicily, on 11 August 2021, unofficially breaking the highest record in Europe, previously recorded at 48 °Celsius in Athens in 1977. These events resulted in many human and animal deaths, even more injuries, unprecedented loss of flora, constructions, and resources.

The situation is not different outside of Europe, with fires devastating California, Brazil, North and South Africa, Australia, India, and China [Aytekin, 2021]. The heatwave that hit North America on 28 June 2021, with extreme temperatures recorded at 49.6 °Celsius (new Canada record), destroyed the entire village of Lytton. The temperatures experienced could not be simulated, and according to the scientists it would have been "virtually impossible" without human influence [WMO, 2021].

Regarding precipitation, in the latest report of the IPCC, it is pointed out that precipitation trends over the Mediterranean depend on the time period as well as the study region and the season. While the projections for precipitation changes indicate an increase in high latitudes, over tropical regions and in parts of the monsoon region, over the subtropicals, including the Mediterranean, a decrease in precipitation is expected [IPCC, 2021].

Changes in temperature and precipitation due to climate change will profoundly affect multiple sectors of the society and the economy, starting with an extremely high number of deaths caused by fires, heatwaves, and other natural disasters. Grain harvest losses due to either flood, water deficiency, or fires will become more frequent, the pollution levels will get higher and we will experience great losses in terms of forest land, and flora in general, including several protected conservation sites (Natura, 2000) [Kovats et al., 2014].

1.5.1 Precipitation and Drought

1.5.1.1 Precipitation

In meteorology, any product of the condensation of atmospheric water that falls to Earth is called precipitation, and occurs when a portion of the atmosphere becomes saturated with water vapor so that the water condenses and "precipitates". The primary component of the water cycle, precipitation, is responsible for depositing fresh water on Earth and has a complicated spatiotemporal variability (Fig. 1.3). According to Chowdhury [2005], approximately 505 000 km^3 of water falls as precipitation each year, 398 000 km^3 of it over the oceans and 107 000 km^3 over land. The aforementioned suggests that the average annual precipitation across the entire Earth's surface is 990 mm, however only 715 mm of that precipitation falls over land.

Fresh water accounts for less than 4% of the world's total water supply, with more than 68 percent trapped in ice and glaciers and another 30 percent deposited underground. Fresh surface water sources, such as rivers and lakes, only constitute about 93 100 km^3, which is about 1/150th of one percent of the total water on the Earth. Nevertheless, rivers and lakes supply with fresh water the majority of the population [Shiklomanov, 1993].

The most commonly used data sources to obtain precipitation data fall into three categories. The first source involves the point measurements procured from rain gauges. The data acquired from this category are widely used to produce spatiotemporal estimates by applying numerous interpolation methodologies. However, they can be problematic in generating reliable precipitation estimates because the rain gauge network in some areas is too sparse in space or in time [Goovaerts, 2000]. For in-situ measurements of the precipitation height, the standard instrument is the standard rain gauge, consisting of a funnel emptying into a graduated cylinder, 2 cm in diameter, which fits inside a larger container which is 20 cm in diameter and 50 cm tall [Strangeways, 2006]. Other types of gauges include the wedge rain gauge (the most affordable and most fragile rain gauge), the tipping bucket rain gauge, and the weighing rain gauge.

The second source of precipitation data includes observations acquired from satellites or radars. These are available in specific areas, and the measurements

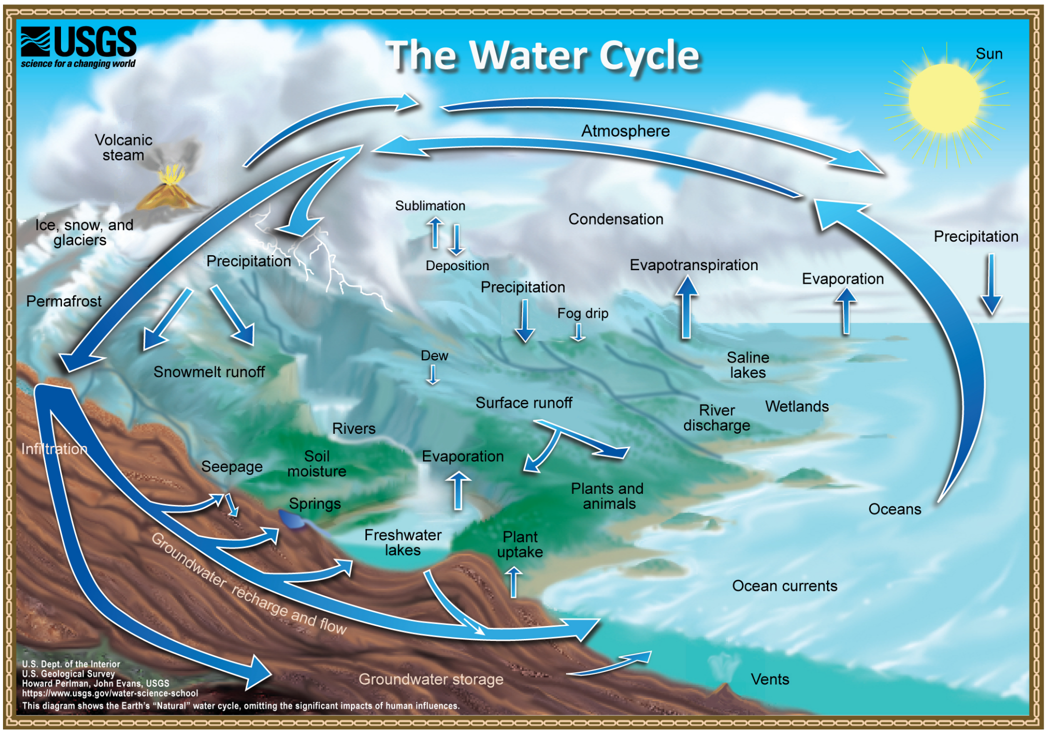


Figure 1.3: Water Cycle. Figure taken from Shiklomanov [1993].

are a direct interpretation of the water droplets' reflectivity [Atlas, 1990; Hong and Gourley, 2015]. As described in Seo et al. [1999] and Grzegorz et al. [2007], these measurements convey errors; thus, the appropriate correction must be applied before further use [Park et al., 2017]. In particular, precipitation measurements in vast expanses of the ocean and isolated land areas rely on satellite observations. Satellite sensors record electromagnetic spectrum, which is afterward translated via a mathematical formula to occurrence and intensity of precipitation. Sensors are classified based on the wavelength that they record into two categories. The thermal infrared (IR) sensor records a channel around 11-micron wavelength and primarily gives information about cloud tops (works best in the tropics). The second category includes sensors that record the microwave part of the electromagnetic spectrum (10 GHz to a few hundred GHz).

Finally, the last source for the acquisition of precipitation data involves the reanalysis estimates. Reanalysis estimates come from satellite missions that use their data to produce a more extensive data set. Such operations have been functioning since 1990. A few of them include the Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis (TMPA), the Global Precipitation Mapping (GPM), and the Global Change Observation Mission-Water (GCOM-W) [Hou et al., 2014; Imaoka et al., 2010; Kummerow et al., 1998].

In the last decades, remote sensing methods, and analysis methods that can handle bigger and longer data sets have evolved. This change brought progress in the quantitative mapping of various environmental variables, including precipitation. Consequently, the information that the meteorological variables enclose across different spatiotemporal scales became more understandable [Hu et al., 2019].

In our case, the data collections that we acquired and used for the applications in this dissertation are the ERA5 reanalysis products, from the ECMWF (the European Centre for Medium-Range Weather Forecasts). The introduction and the analysis of the data sets are presented in Section 3.

1.5.1.2 Drought

Drought is a difficult concept to define since several interpretations are possible. Drought is defined by the majority of people as a "prolonged absence or marked deficiency of precipitation," a "deficiency of precipitation that results in water shortage for some activity or for some group," or a "period of abnormally dry weather sufficiently prolonged for the lack of precipitation to cause a serious hydrological imbalance" [Heim, 2002].

Human activities, the underground surface, and the area's climatic conditions combined are the factors that can disturb the water budget equilibrium. The process of drought is much slower than other natural disasters such as hurricanes or floods. Since the consequences of drought are not that catastrophic until its full rise, it is often ignored throughout its development time [Dai, 2013; Mishra and Singh, 2010].

The types of drought are classified as meteorological/climatological, hydrological, and agricultural/ecological (Fig. 1.4). Meteorological is region-specific and links the deficiency of rainfall to the norm of the region. Hydrological drought suggests water deficiencies over a prolonged period that also affect the subsurface

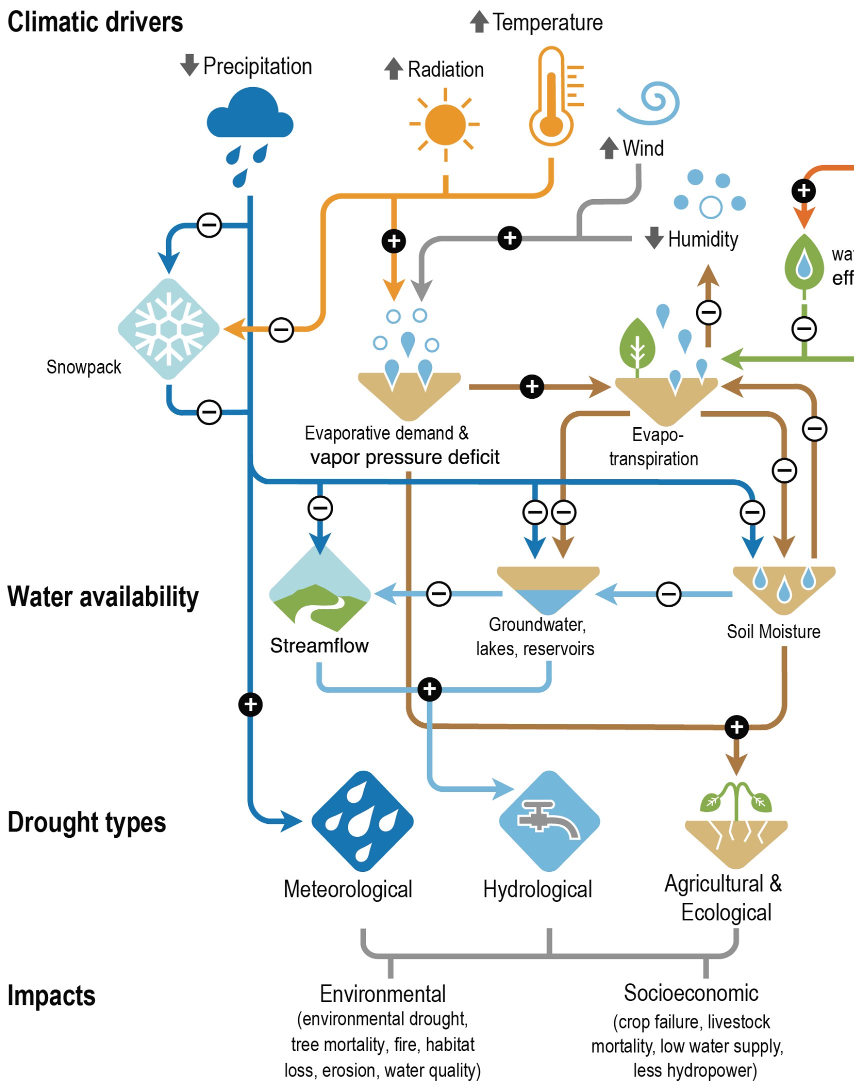

Climatic drivers
Precipitation
Radiation
Temperature
Wind
Humidity
CO_2
water use efficiency
plant growth
Snowpack
Evaporative demand & vapor pressure deficit
Evapo-transpiration
Water availability
Streamflow
Groundwater, lakes, reservoirs
Soil Moisture
Drought types
Meteorological
Hydrological
Agricultural & Ecological
Impacts
Environmental (environmental drought, tree mortality, fire, habitat loss, erosion, water quality)
Socioeconomic (crop failure, livestock mortality, low water supply, less hydropower)

Figure 1.4: Climatic drivers of drought, effects on water availability, and impacts. Plus and minus signs denote the direction of change that drivers have on factors such as snowpack, evapotranspiration, soil moisture, and water storage. The three main types of drought are listed, along with some possible environmental and socioeconomic impacts of drought (bottom). Figure 8.6 in IPCC [2021].

water supply. Finally, agricultural drought can be identified by its most prevalent characteristic, the lack of soil moisture. All types of droughts may have environmental and socioeconomic impacts. The environmental impacts include loss of habitat, fires, erosion and decline of the water quality. On the other hand, the socioeconomic impacts include food and water supply shortages, livestock mortality and reduction in the produced hydropower which have resulted from meteorological, agricultural, and hydrological drought elements.

Changes in the timing and the amount of soil moisture, evaporation, transpiration rates, and precipitation have an impact on the hydrology of a region, which in turn affects the region's susceptibility to drought. The effect of a drought varies according to vulnerability. For instance, subsistence farmers are more inclined to migrate during droughts because they do not have access to alternative food sources. Areas with populations that depend on water resources for food production are more vulnerable to famine.

1.5.2 Temperature

Temperature is a physical quantity that expresses hot and cold and can be measured with a thermometer. Various thermometers exist with numerous temperature scales. The most common temperature scales include the Celsius scale (°C), the Fahrenheit scale (°F), and the Kelvin scale (K).

Air temperature is another essential parameter that affects the climate and consequently the life on earth. The variability of the temperature measurements can yield devastating effects on human life and ecosystems. For instance, the rise of the air temperature leads to a higher probability of heat waves occurring, which endangers life and may induce more deaths, especially in the more vulnerable population [U.S. Environmental Protection Agency, 2021].

As mentioned earlier, the average surface temperature for each of the last four decades was warmer than any decade that preceded it since 1850 and is expected to continue rising [IPCC, 2021]. The consequences of an increase in the average global temperature can be catastrophic. The rise of the surface temperature results in more evaporation which consequently increases precipitation. However, an increase in the precipitation amount does not necessarily mean an increase in water resources. This is due to the increase of heavier precipitation events causing damage to crops, elevating the flood risk, and not enabling the replenishment of the underground aquifers [U.S. Environmental Protection Agency, 2021].

1.5.3 Solar Radiation

Solar irradiance represents the electromagnetic radiation measured by the appropriate instruments (pyranometers) per unit area, measured in watt per square meter (W/m^2). Most commonly used is the integration of solar irradiance over a given time period, which represents the solar irradiation, or the global horizontal irradiance (GHI) which is calculated taking into account the angle of the sun to the horizontal surface (the Earth's locations in measuring) and the diffuse horizontal irradiance [Stickler, 2015].

Research interest in solar irradiance, among other reasons, arises from the fact that it contributes the most to the prediction of energy production from solar power plants [Koutroulis et al., 2021], and plays a big part in climate modeling and forecasting. For example, the spatiotemporal distribution of solar radiation can greatly impact the timing and magnitude of snowmelt [Elder et al., 2015; Marks and Dozier, 1992].

1.5.4 Wind

In atmospheric sciences, the velocity of the air masses moving in the atmosphere from high to low-pressure areas is called wind speed and is highly affected by temperature changes. Wind speed is measured in meters per second (m/s in the SI) or another equivalent measure such as kilometers per hour (km/h) with anemometers. Another notable scale describing wind speed is the Beaufort scale. Nonetheless, it is an empirical scale based on visual observation of the sea, and every Beaufort number represents a range of wind speeds, thus, it is not used in scientific applications. Different wind speed measurements can have from assisting (pollination) to catastrophic effects (tornadoes) on the environment and the society.

1.5.5 Drought Indices

Drought indicators or indices have been developed to characterize and help monitor possible upcoming droughts in terms of their severity and duration. The consequences of droughts can vary greatly, from food supply and security to access to education. Depending on the relevant impacts of a drought, determining the proper drought indicator for early warning monitoring is crucial [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016].

Drought indices are quantitative measures created to characterize the drought levels, utilizing one or several variables (indicators). Some of these indicators are precipitation, evapotranspiration, temperature, and soil moisture. Depending on the target index, a large amount of data for one or multiple indicators are employed to calculate one single value. Since the possible combinations of variables are endless, the indices developed are over 150 according to Niemeyer [2008], and additional indices have recently been proposed [Karamouz et al., 2009; Rhee et al., 2010; Vasiliades et al., 2011; Vicente-Serrano et al., 2010; Zargar et al., 2011].

Drought indices have been implemented from the perspective of meteorology, hydrology and agriculture by scientists of different disciplines. Representative drought indices include the Standardized Precipitation Index (SPI, [McKee et al., 1993; Vrochidou, 2013; Wu et al., 2005]), the Standardized Precipitation Evapotranspiration Index (SPEI, [Vicente-Serrano et al., 2010]), the Palmer Drought Severity Index (PDSI, [Dai et al., 2004; Palmer, 1965]), the Drought Reconnaissance Index (DRI, [Tsakiris et al., 2007a]), the Rainfall deciles (RD Gibbs and Maher [1967]), the Reclamation Drought Index (RDI, [Weghorst, 1996]), the Crop Moisture Index (CMI, [Palmer, 1968]), the Surface Water Supply Index (SWSI, [Doesken and Garen, 2004; Doesken et al., 1991; Shafer and Dezman, 1982]), the Aggregate Drought Index (ADI, [Keyantash and Dracup, 2004]), and the Normalized Difference Vegetation Index (NDVI, [Kogan, 1995; Tarpley et al., 1984]).

In Section 4 we estimate the Standardized Precipitation Index (SPI) and the Standardized Evapotranspiration Index (SPEI) to monitor and extract conclusions for Crete —a drought-prone area— to assist the water resources management agencies with drought management policies and preparedness plans. We identified multiple drought occurrences across the investigation period and spatial “hot spots” more affected from the temperature increase during the last decades.

1.5.6 Bias Correction Methodologies

Methods employed to correct bias from model-derived data sets have different names in the literature. These include statistical downscaling, quantile mapping, histogram equalizing, and bias correction. Bias correction techniques usually involve a transfer function obtained through the observed and simulated cumulative density functions (CDF) (for CDF definition see Section 2.4) [Hansen et al., 2006; Piani et al., 2010]. In other words, bias correction is the process of calibrating climate model products to account for their systematic errors.

An abundance of satellite precipitation products followed the growth of the remote sensing field. However, as mentioned in Section 1.5.1.1, satellite data have some shortcomings despite the advancement of the field. First, downscaling techniques are inevitable for a local-scale analysis because of the coarse spatial resolution, especially in areas where rain gauges are sparse [Atkinson, 2013]. Numerous methodologies have been proposed for spatial downscaling, some of which use statistics/geostatistics while others use machine learning techniques [Ma et al., 2020; Park et al., 2017; Sharifi et al., 2019]. Additionally, some integrate auxiliary environmental variables. Apart from the estimation errors that every downscaling method produces, the model's performance is eventually subject to the accuracy of the input satellite precipitation product.

Rain gauge data have complementary characteristics in terms of availability and accuracy with satellite products; thus, they can be a valuable addition in precipitation mapping when appropriately integrated. Records retrieved from rain gauges are considered as true measurements, consequently, they can be used to correct biases in the remote sensing products, improving the model's prediction performance.

Integrating rain gauge data with satellite precipitation products poses some challenging issues, including the differences in scale, either spatial or temporal. For instance, data from a rain gauge corresponds to point measurements, whereas satellite precipitation products refer to the aggregated amount over a spatial grid cell. Inconsistencies between records from remote sensing and ground sources are common, exhibiting many wet days with low-intensity rainfall or extreme temperatures [Ines and Hansen, 2006], and under- or overestimation and incorrect seasonal variations of precipitation [Christensen et al., 2008; Terink et al., 2009; Teutschbein and Seibert, 2010, 2012], yielding unrealistic results in hydrological simulations [Soriano et al., 2019]. Hence, differences in scale or mismatching should be addressed appropriately during data integration from different supports [Porcù et al., 2014].

A variety of bias correction methodologies for merging rain gauge and satellite or radar precipitation products have been proposed and applied [Schmidli et al., 2006], some of which depend on geostatistics and specifically on different kriging methods [Berndt et al., 2014; Chappell et al., 2013; Erdin et al., 2012; Goudenhoofdt and Delobbe, 2009]. Other available methodologies include Bayesian techniques [Todini, 2001], double kernel smoothing [Li and Shao, 2010], conditional merging [Baik et al., 2016], and Quantile mapping (QM) [Li et al., 2010; Soriano et al., 2019; Teutschbein and Seibert, 2012]. Park et al. [2017] present a comparison between simple kriging with local means, kriging with an external drift, and conditional merging. Some of the methods are simple to apply, consisting of a linear model, or just an offset rainfall value, while others are more complex and require separate modeling.

In an attempt to map precipitation over South Korea Park et al. [2017] integrated coarse-resolution satellite monthly accumulated precipitation products (TRMM 0.25 degrees = 25km) and gauge data. They concluded that when the ground station network is sparsely distributed, incorporating satellite-derived estimates can significantly improve the precipitation mapping estimates. However, this is not the case when the ground station network is extensive. Additionally, the assimilation generates precipitation estimates that are more variant compared with the smoothed estimates that ordinary kriging (OK) generates, highlighting the benefit of the integration in areas with sparse rain gauge networks and the requirement of high-resolution mapping.

Soriano et al. [2019] applied different bias correction techniques to climate model projections of temperature and precipitation to estimate the influence on flood response for four catchments in Spain. They concluded that precipitation is the most crucial input on flood response, and quantile mapping was the best methodology for precipitation correction. To correct the daily projected temperature values, Soriano et al. [2019] measured the difference between the mean monthly values of the projected and the ground station values, and added the difference to the projected values, known as seasonal bias correction [Bergström et al., 2001]. Various studies proved that QM is the most appropriate method for precipitation bias correction, especially in cases with extreme values.

In this case study, we do not apply bias correction methodologies due to several reasons. Firstly, the ground stations network is very sparse in many regions around Crete which made the association of the ground measurements to the reanalysis products very difficult. Additionally, the first reason combined with the extreme variability of the ground measurements across the island [Agou, 2016] makes it impossible to estimate a universal model that can be applied to correct the data for the entire island. And finally, the ground measurements available correspond to different time period, and different time scale while also having many missing values. This means that before we could apply any bias correction methodology, we must fill in the missing values of the ground data set and downscale to a finer resolution, incorporating bias and uncertainty.

1.5.7 Statistical Modeling

A different approach for the characterization and the evaluation of the climatic conditions of a region besides the General Circulation Models (GCMs)[1], Regional Climate Models (RCMs)[2], and the drought indices is the geostatistical analysis. A wide range of methods comprises the geostatistical analysis, the foundation of which are mathematical functions. In this particular approach, mathematical expressions are utilized to model a variable of interest, such as precipitation, and to ascertain potential associations through space and time.

Studies that utilize geostatistical methodologies and incorporate topographical parameters in the algorithms are presented by Agou et al. [2019]; Goovaerts [2000]; Moral [2010]; Tushaus [2014], while studies such as those by Baxevani and Lennatsson [2015]; Li et al. [2013] focus on the generation of weather fields. Most

[1] GCMs are mathematical equations that reflect physical processes in the atmosphere, oceans, cryosphere, and land surface. They are the most prevalent used tools for climate simulation by dividing the earth, ocean and atmosphere into grid blocks [IPCC, 2015].

[2] RCMs is the most prevalent downscaling technique of GCMs with a horizontal grid resolution of around 25–50 km [Soriano et al., 2019].

of the studies in the field of geostatistics associated with precipitation use parametric distribution functions to approximate the distribution of the field [Gellens, 2002; Koutsoyiannis, 2004; Rho and Kim, 2019; Shoji and Kitaura, 2006], however studies that use non-parametric approximation are available [Harrold et al., 2003; Mosthaf and Bárdossy, 2017; Pavlides et al., 2021; Sharma and Lall, 1999].

Many meteorological variables display non-Gaussian characteristics, and variability thought space and time. Main step in most geostatistical methodologies is the identification of a probability distribution model that fits the variable of interest properly. For instance, precipitation displays significant spatial and temporal variability, therefore it can be modeled as a stochastic variable. Several probability distributions have been used to model precipitation. Commonly used parametric models include the exponential, gamma, lognormal, Weibull, generalized extreme value (GEV), pareto [Baxevani and Lennatsson, 2015], as well as hybrid mixtures of exponential with a Pareto tail [Li et al., 2013; Ye et al., 2018]. Recently, Andreou [2022] suggested the use of the compound Poisson-Gamma model which is a mixed type distribution that models the occurrence and the intensity of precipitation at the same time.

The gamma distribution has been extensively used in the analysis of precipitation for different time scales [Ye et al., 2018], including for modeling the daily rain rate [Cho et al., 2004], the monthly and seasonal anomalies [Wilks, 1990; Wilks and Eggleston, 1992], as well as to fit precipitation for the SPI development [McKee et al., 1993; Vrochidou, 2013]. Nevertheless, the gamma distribution is not defined for zero values that are present in precipitation measurements. Zero precipitation values are taken into account in the gamma model using Type I censoring of the distribution on the left [Wilks, 1990]. Left censoring means that the number of values that fall under a threshold is known but their values are considered unknown. This type of censoring can be applied to the distributions presented below. Nonetheless, recent studies have emphasized that the tails of the gamma distribution are not adequately heavy for heave rain events [Nerantzaki and Papalexiou, 2019; Wilson and Toumi, 2005]. This is because the exponential part of the gamma tail dominates the power-law term, and therefore in numerous cases the gamma tail behaves similarly to the exponential.

In other studies, the lognormal distribution is utilized to approximate rainrate [Biondini, 1976; Kedem and Chiu, 1987; Sauvageot, 1994], cumulus cloud populations [López, 1977], amount of precipitable water [Foster and Bevis, 2003; Foster et al., 2006], hourly precipitation [Shoji and Kitaura, 2006], and the average rain rate from satellite observations [Cho et al., 2004; Kedem et al., 1990]. In a recent study, Cho et al. [2004] concluded that the lognormal and gamma distributions fit adequately the TRMM data (obtained from the Tropical Rainfall Measuring Mission research satellite which operated from 1997 to 2015) of daily average rain-rate.

The GEV distribution, which is developed within extreme value theory, combines the Gumbel, Fréchet, and Weibull families also known as type I, II and III extreme value distributions respectively. It is applied in various studies to model precipitation and extremes for different time scales, including one-day maximum [Coles, 2001; Wang et al., 2017], k-day extreme precipitation (for k ranging between 1 and 30) [Gellens, 2002], daily precipitation forecasts [Scheuerer, 2014], and annual maximum precipitation [Koutsoyiannis, 2004]. In the earth system sciences, the GEV distribution finds oftentimes applications in hydrology for the study of extremes of several natural phenomena, including rainfall, floods, wind speeds and wave heights [Soriano et al., 2019]. Hybrid GEV models that allow more flexible tail behavior are considered in Papalexiou and Koutsoyiannis [2012]; Papalexiou and Serinaldi [2020]; Rho and Kim [2019].

In Appendix A we present the PDF (Eq. A.1) and CDF (Eq. A.2) equations as well as the plots of several relevant probability distribution functions including the Gaussian, gamma, GEV, lognormal, Weibull, Pearson's Type III, and Pareto Type II distribution (Figs. A1 and A2).

1.5.8 The Role of Spatial and Temporal Scales

Many studies are widely available that investigate the patterns of temperature and precipitation on a global scale. For example, while global scale models indicate an increase in mean temperature, downscaling to local and regional scales reveals different trends. This fact demonstrates that it is important to study different spatial scales to view the entire picture, and especially important to analyze local scales to yield conclusions for local regimes.

In the analysis of precipitation data, the selection of the time and spatial resolution is important. The time resolution can vary, for example, between daily, weekly, monthly, annual, seasonal, wet and dry periods. Annual precipitation is the sum of precipitation over the course of a year (365 days), wet periods correspond to October till March and dry periods to April till September. The seasonal time scale is calculated by summing the precipitation over the total days of each season. The time scale for the analysis must be determined according to the application. For instance, when investigating the precipitation for a region with arable crops, results based on the seasonal time-scale are more valuable [Rotter and Van De Geijn, 1999], while a daily temporal resolution is more appropriate when studying the impact of extreme precipitation.

According to McKee et al. [1993], at longer time scales drought becomes less frequent and lasts longer. In terms of precipitation this suggests that at longer time scales small amounts of precipitation become less frequent and have longer duration. In addition, the correlation between precipitation and topography increases with the length of the time interval. Finally, Bárdossy and Pergam [2013] demonstrated that interpolation quality correlates to aggregation time; longer aggregation times decrease the interpolation relative error.

Steps toward the understanding of precipitation distribution and change on global or large spatial scales have extensively been analyzed in the last decades. Although there is still an unknown territory on large scale estimation, the estimation on a smaller scale with finer resolution estimations is still in its first steps and needs to be addressed more thoroughly. This problem arises mostly because of the lack of an extensive ground stations network, or a higher resolution grid with remote sensing records. Nevertheless, the need for reliable estimations on a regional scale is necessary because of the imminent threats that the environment is projected to encounter in the next decade [IPCC, 2018].

The effects of climate change and the increase of 1.5 °C globally are very troubling [IPCC, 2018]. The value of mean temperature presents the outline on the effect of climate change on a global scale; however, the impacts will affect every region economically and socially at different levels [Marotzke et al., 2017]. Modeling the climate at large scales, e.g., global or continental, has been the field of study for many researchers in the last decades with significant results. Yet, we still need to break new ground in modeling climate change at local scales if we want to grasp the future. A better understanding at a regional level will allow each country's relevant authorities to take measures against climate change's adverse consequences [Marotzke et al., 2017].

The intermittent rainfall behavior is unveiled by the fluctuations in rainfall intensity and the alternation of wet and dry periods [Agnese et al., 2014; Molini et al., 2009; Schmitt et al., 1998]. The intermittence in rainfall rises with the decrease in the time-step; consequently, information noticeable in small time-scales is concealed at larger scales.

Onyutha and Willems [2017] displayed how the change in spatial and temporal resolution affected the variability in the rainfall amount and the computed climate indices in the Nile basin. Specifically, they noted that the variation in the indices explained the variability better at regional than location-specific scales. The aforementioned are indicators of the need to focus and extensively analyze areas at regional scales and not support all the planning and water management decisions on large-scale estimation results.

In this thesis we use multiple time scales, the finer temporal resolution in use is one hour in the temporal SLI application for the temperature, solar radiation, and precipitation data (Section 7.7, and in the spatiotemporal SLI application for the temperature and precipitation data (Section 7.8). Similarly, for the classification of precipitation presented in the Section 6.5 we use the hourly data from all the 26 variables included in the data set covering and surrounding the island of Crete. In the GAH applications we use the monthly ERA5 precipitation products for the wet period which includes 246 months (from January to March and October to December 1979 to 2019 - 41 years) for the entire grid (Section 5.6). Lastly, the investigation of the frequency and intensity of the drought occurrences is carried out with the monthly precipitation and temperature data based on the historical records during the period 1979–2019 (Section 4.5).

Chapter 2

Basic Concepts of Geostatistics

2.1 Summary

This chapter introduces several definitions and concepts that are helpful in spatial data modeling. They include definitions for random fields, statistical moments such as the mean value and the covariance function, statistical homogeneity and isotropy, different variogram models, modeling formulas for the kriging method, spatial model estimation via maximum likelihood estimation, a brief description of simulation methods, cross-validation types, and validation metrics.

2.2 Introduction

Earth science data exhibit vast variability in space and time. Geostatistics provides the tools to model and characterize the spatio-temporal attributes, including variability, based on the theory of random fields. The use of geostatistical methods to analyze data facilitates the estimation or prediction of the system's response and helps practitioners make informed decisions. The field of geostatistics originated from the mining and petroleum industries, which is apparent in the applications discussed in various geostatistical books (see Journel and Huijbregts [2003]; Olea [1999]) [Hristopulos, 2020]. However, since those initial applications, the field has expanded to many other areas, including hydrology, meteorology, forestry, and geochemistry. Some example applications using geostatistical methods involve random variables such as mineral grades [Hristopulos et al., 2021; Pavlides et al., 2015], porosities and pollutant concentrations, underground water [Varouchakis and Hristopulos, 2013] and meteorological variables such as temperature, precipitation [Agou et al., 2019, 2022; Varouchakis et al., 2018] and pressure [Chilès and Delfiner, 2012; Christakos, 1992; Goovaerts, 1997].

The motive behind the development of methods that can integrate the distribution of the data and their spatial correlations was profit. For the variable of interest (e.g. gold deposits), increasing the sample size will normally decrease the uncertainty of the estimation on neighboring positions, yet, it will also remarkably increase the cost. Geostatistics decreases the uncertainty by utilizing the current sample network, while simultaneously offering guidance if a network expansion is needed.

Measurements for a real-world variable distributed in space and/or time commonly appear to be correlated with each other. One of the most frequently followed approaches in geostatistics include the estimation of such correlations via the "structural analysis" also known as "variogram modeling", followed by the application of interpolation methods, such as kriging, for the estimation of the modeled process at unsampled locations, and finally the assessment of the uncertainty of the estimates.

2.3 Random Fields

A random variable is used to quantify outcomes from random occurrences, such as the result from a dice roll, the voltage of a random source, the cost of a random component, plume concentrations, temperature measurements, or any other numerical quantity. A random variable is a function whose domain is the set $\mathbf{s}$ of all experimental outcomes [Papoulis and Pillai, 2002]. A random variable x can be a discrete or continuous variable, if its sample domain is discrete (e.g. $x(\mathbf{s}) = 0, 1, 2, ...$) or continuous. Most environmental variables are continuous, for instance precipitation, wind speed, and solar irradiance. Because of the continuity, it is not possible to assign probabilities to all probable values of the random variable in a meaningful way [Coles, 2001].

A random field (RF) is a set of interdependent random variables that describe the spatial or spatiotemporal range of an attribute. They have unique mathematical properties that distinguish them from a set of independent random variables. Ω denotes a probability space, $\mathcal{F}$ is the σ–algebra on Ω, which is a collection of subsets of Ω that contains the full set and is closed under complementation and countable unions, and P is the family of probability measures. Let $(\Omega, \mathcal{F}, P)$ denote a probability space and $\mathbb{D} \subseteq \mathbb{R}^d$ the spatial domain of interest. Then an RF $X(\mathbf{s};\omega)$ is a collection of real-valued random variables distributed over $\mathbb{D}$. The RF is defined by a mapping from $\Omega \times \mathbb{D}$ into the set of real numbers $\mathbb{R}$. Hence, for any fixed $\mathbf{s} \in \mathbb{D}$, $X(\mathbf{s};\omega) \to X(\omega)$ is $\mathcal{F}$-measurable as a function of ω, and for a fixed ω, $X(\mathbf{s};\omega)|_{\omega=fixed} = x(\mathbf{s})$ is a deterministic function of $\mathbf{s}$ [Gikhman and Skorokhod, 1996].

Overall, we denote a field marked as $X(\mathbf{s})$ where the vector $\mathbf{s}$ corresponds to the position of a point in the study area, $x(\mathbf{s})$ denotes the values corresponding to a unique realization, and $X'(\mathbf{s})$ denotes the fluctuation of the field [Hristopulos, 2020]. Equivalently, a random field can be i) a *field of discrete values*, ii) a *field of continuous values*, iii) a *lattice field* if the locations where the field is defined are lying on a grid, and iv) a *continuum field* if the field extends over a continuous space.

It is important to point out that if the interdependence of the random variables is absent, the random field does not exhibit spatial continuity. Physical processes have intrinsic correlations, making it possible to model. For example, the distribution of precipitation over an area is governed by complex physical phenomena and attributes such as the movement of clouds in the atmosphere, the topography of the area, and the temperature. The regionalized aspect of fields is the single characteristic that differentiates geostatistics from pure statistics. Without spatial continuity, prediction of the field's value at an unobserved would not be possible.

In this case study, the main variables of interest are precipitation, temperature and solar radiation. We assume that they can be modeled as a random field defined in continuum space.

2.4 Probability Density Function

The probability density function (PDF) of the field is denoted with the symbol $f_X[x(\mathbf{s})]$, while the PDF $f_X(x)$ refers to a single point. In both cases, the subscript is the symbol indicating the field. This means that $f_X[x(\mathbf{s})]$ describes the joint PDF of the field values for any number (even infinite) of points. Therefore, the PDF in the case of the random field involves much more information than the PDF of a single variable.

The multidimensional PDF, is defined as $f_X(x_1, \dots, x_N; \mathbf{s}_1, \dots, \mathbf{s}_N)$ and describes the interdependence of possible states for a set of N points [Isaaks and Srivastava, 1989]. In the case of a single random variable, the PDF is the first derivative of the cumulative density function $F_X(x)$ (at every continuity point). It is normalized so that the total probability of all possible outcomes is equal to 1, i.e.,

$$\int_a^b f_X(x)\, \mathrm{d}x = 1. \tag{2.1}$$

The integral limits depend on the space where the field X is defined, with values ranging from $-\infty$ to ∞ or being limited to a specific interval [a,b].

In many cases, Gaussianity is a prerequisite for many applications; several tests may be used to determine how dissimilar the sample data are from Gaussian distributed data. For a brief introduction see Appendix A. Techniques to transform non-Gaussian to Gaussian distributed data are presented in Sections 5.3 and 5.4.

2.5 Statistical Moments

Statistical moments are deterministic functions that represent expectations over all possible states of the field. They are defined for various combinations of field values at one or more locations. In practice, the most commonly used moments are low order moments such as mean value, variance, covariance function, and semivariogram (or variogram) [Cressie, 1993].

The central value (*mean value*) of a distribution, also known as the first central moment, is the expectation $\mathbb{E}[X(\mathbf{s})]$ of the random field $X(\mathbf{s})$ at the position $\mathbf{s}$

of a point in the study area calculated over the ensemble of all states, i.e.

$$m_X(\mathbf{s}) = \mathbb{E}[X(\mathbf{s})] = \int_{-\infty}^{\infty} x \, f_X(x;\mathbf{s}) \, \mathrm{d}x, \tag{2.2}$$

where x are the values that correspond to a given state.

The central value of highly asymmetric distributions is more appropriately depicted by the median value, M_x, which is the value corresponding to a cumulative frequency of 0.5 [Goovaerts, 1997].

The *variance* of a distribution, also known as the second central moment, captures the dispersion of the values around the mean value. It is given by the mean value of the squared fluctuation, i.e.

$$\sigma_X^2(\mathbf{s}) \equiv \mathbb{E}\left[\{X(\mathbf{s}) - m_X(\mathbf{s})\}^2\right] = \mathbb{E}\left[X'^2(\mathbf{s})\right]. \tag{2.3}$$

Variance is sensitive to irregular high values due to the squared differences in the formulation [Hristopulos, 2020]. The square root of the variance, σ, is called standard deviation, and its ratio to the mean, σ/m, is the unit-free coefficient of variation for non-negative variables [Wackernagel, 2003].

Higher order moments include the *skewness* and the *kurtosis* of the distribution. The *skewness* is used to measure the asymmetry of the distribution about its mean value, however, its interpretation is complicated due to the fact that it cannot differentiate between fat and long tails. In other words, zero skewness indicates that the tails on either side of the mean balance out, yet, it does not necessarily indicate a symmetric distribution. The skewness is usually defined as

$$\text{coefficient of skewness} = \frac{\mathbb{E}\left[X'^3(\mathbf{s})\right]}{\sigma_X^3}. \tag{2.4}$$

The *kurtosis* k_X is a scaled fourth order moment measure that describes the shape of the distribution. It is used to assess how spread the tails of the distribution are around the mean value. Namely, high kurtosis implies that more of the variance is the outcome of extreme fluctuations, instead of recurring fairly sized fluctuations [Balanda and Macgillivray, 1988]. It is typical in practice to use the excess kurtosis, which is defined based on the kurtosis of a univariate normal distribution which is equal to 3, and is calculated by the equation

$$\text{excess kurtosis} = \frac{\mathbb{E}\left[X'^4(\mathbf{s})\right]}{\sigma_X^4} - 3 = k_X - 3. \tag{2.5}$$

In the rest of the text, for the sake of brevity, we will refer to the excess kurtosis as kurtosis. Both skewness and kurtosis can be unreliable estimators if the sample size is small, however, large values even for small sample sizes may merit attention, because they indicate that statistical approaches that lie on the Gaussian assumption may be improper [Williams, 2000].

2.6 Covariance Function

The centered covariance function (CCF), also known as covariance function for conciseness, is a mixed statistical moment that represents quantitatively the dependence of the fluctuations between two different points and is defined by the following equation:

$$\begin{aligned} c_X(\mathbf{s}_1, \mathbf{s}_2) &\equiv \mathbb{E}[X(\mathbf{s}_1)X(\mathbf{s}_2)] - \mathbb{E}[X(\mathbf{s}_1)]\,\mathbb{E}[X(\mathbf{s}_2)] \\ &\equiv \mathbb{E}[\{X(\mathbf{s}_1) - m_X(\mathbf{s}_1)\}\{X(\mathbf{s}_2) - m_X(\mathbf{s}_2)\}] \\ &\equiv \mathbb{E}[X'(\mathbf{s}_1)X'(\mathbf{s}_2)]. \end{aligned} \tag{2.6}$$

If the arguments of the covariance function coincide, its value becomes equal to the variance of the field at that point, that is

$$c_X(\mathbf{s}_1, \mathbf{s}_1) = \sigma_X^2(\mathbf{s}_1). \tag{2.7}$$

A function has to fulfill the permissibility conditions defined by Bochner's theorem to be accepted as a covariance function. This is expressed by means of the spectral density, which is given by the Fourier transformation of the covariance function $\tilde{c}_X(\mathbf{k})$ [Bochner et al., 1959]. According to the theorem, a function $c_X(\mathbf{r})$ is a permissible covariance function if the power spectral density $\tilde{c}_X(\mathbf{k})$ exists, is non-negative throughout the frequency domain, and the integral of $\tilde{c}_X(\mathbf{k})$ over the entire frequency domain is bounded [Agou, 2016].

2.7 Statistical Homogeneity & Isotropy

Statistical homogeneity and isotropy are two more concepts that are used to characterize a random field. A homogeneous field $X(\mathbf{s})$ in the weak sense has constant mean value, i.e., $m_x(\mathbf{s}) = m_x$. Also the covariance function of the field is defined and depends only on the distance vector $\mathbf{r} = \mathbf{s}_1 - \mathbf{s}_2$ between two points, in other words $c_X(\mathbf{s}_1, \mathbf{s}_2) = c_X(\mathbf{r})$, which implies that the variance of the field is constant. A random field is statistically homogeneous in the strong sense if the multidimensional PDF for N points, where N is any positive integer, remains unchanged by transformations that change the location of the points without changing the distances between them.

Therefore, statistical homogeneity indicates that the statistical characteristics of the field are independent of the spatial coordinates of the center of mass of the N points. Practically, statistical homogeneity insinuates the absence of spatial trends; thereby, fluctuations around a fixed level equal to the mean value can be used to explain the spatial variability of the field [Hristopulos, 2020].

Providing that a field is statistically homogeneous, even in the weak sense, it is also isotropic if the covariance function depends only on the distance r, and not on the direction of the distance vector $\mathbf{r}$. This addition is significant from a practical point of view because it facilitates identifying the spatial dependence through the omnidirectional variogram. Consequently, the field is by definition statistically homogeneous, if the covariance function is isotropic, but not the other way around [Olea, 1999].

2.8 Variogram Modeling

The semivariogram or variogram is a statistical moment that assesses how much the average similarity between two random variables declines as their distance increases. Stochastic interpolation algorithms, such as kriging-based methods, demand knowledge of the variogram or the covariance [Olea, 1999].

The variogram of a random field is defined by the following equation

$$\gamma_X(\mathbf{s},\mathbf{r}) = \frac{1}{2}\,\mathbb{V}\mathrm{ar}\left[X(\mathbf{s}+\mathbf{r}) - X(\mathbf{s})\right]. \tag{2.8}$$

The variogram is thus defined with respect to a pair of points, using the variance of the increment field (a.k.a. distance step $\mathbf{r}$), where the latter is defined as $\delta X(\mathbf{s};\mathbf{r}) \equiv X(\mathbf{s}+\mathbf{r}) - X(\mathbf{s})$.

If the field $X(\mathbf{s})$ is statistically homogeneous, the variogram is directly connected to the covariance function by means of the equation

$$\gamma_X(\mathbf{r}) = \sigma_X^2 - c_X(\mathbf{r}). \tag{2.9}$$

It follows from Eq. (2.9) that the variogram tends asymptotically to the variance $c_X(0)$, and if the covariance is known, the variogram is also known [Chilès and Delfiner, 2012]. However, the variogram can increase indefinitely if the variability of the process does not approach a limit at long distances. This is evidence that the random field is not statistically homogeneous. It is a commonly occurring practice to utilize the variogram than the covariance function to estimate the spatial dependence since it has a more general function and does not require knowledge of the mean value.

In practical applications, the variogram near the origin demonstrates two typical behaviors. On the one hand, it may show a discontinuity at distance zero; thus, $\gamma_X(\mathbf{r})$ does not seem to tend to zero as $r \to 0$ but to c_0, which is also called as nugget effect. The nugget effect might be present due to unresolvable fluctuations, a component of the phenomenon with a range shorter than the sampling support, measurement, or positioning errors [Chilès and Delfiner, 2012]. On the other hand, the experimental variogram might be a flat curve, indicating a pure nugget effect or white noise. This means that there is no correlation between any two points; it is an extreme case of a total absence of spatial structure.

The correlation length ξ specifies the distance over which the field values are statistically correlated, and the "speed" with which the variogram approaches the sill. An anisotropic dependence exists when correlation properties change in different directions in space. In practice, two types of anisotropy are observed: geometrical anisotropy and zonal anisotropy. Geometrical anisotropy refers to cases where the sill is independent of the direction, but the "speed" of approach to the sill depends on the direction, while in zonal anisotropy the sill depends on the spatial direction [Goovaerts, 1997].

Variogram models that are commonly used in practice include the Gaussian, the exponential, the spherical, the power-law, and the Matérn model [Hristopulos, 2020]. Their respective isotropic equations are listed below. For the following equations σ_x^2 is the variance, $\|\mathbf{r}\|$ is the Euclidean norm of the lag vector $\mathbf{r}$, and ξ is the characteristic length.

Exponential model:

$$\gamma_x(r) = \sigma_x^2 \left[1 - \exp\left(-\|\mathbf{r}\|/\xi\right)\right]. \tag{2.10}$$

Gaussian model:

$$\gamma_x(r) = \sigma_x^2 \left[1 - \exp\left(-\|\mathbf{r}\|^2/\xi^2\right)\right]. \tag{2.11}$$

Spherical model:

$$\gamma_x(r) = \begin{cases} \sigma_x^2 \left[1.5\left(\dfrac{\|\mathbf{r}\|}{\xi}\right) - 0.5\left(\dfrac{\|\mathbf{r}\|}{\xi}\right)^3\right], & \text{if } \|\mathbf{r}\| \leq \xi \\ \sigma_x^2, & \text{if } \|\mathbf{r}\| \geq \xi. \end{cases} \tag{2.12}$$

Power-law model:

$$\gamma_x\left(\|\mathbf{r}\|\right) = \alpha\|\mathbf{r}\|^{2H}, \qquad \text{where } 0 < H < 1, \tag{2.13}$$

where H is the Hurst exponent.

Matérn model:

$$\gamma_x\left(\|\mathbf{r}\|\right) = \sigma_x^2 \left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\sqrt{\nu}}{R}\|\mathbf{r}\|\right)^{\nu} K_\nu\left(\frac{2\sqrt{\nu}}{R}\|\mathbf{r}\|\right)\right], \tag{2.14}$$

where $\nu > 0$ is the smoothness parameter, $\Gamma(\cdot)$ is the gamma function and $K_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν [Stein, 1999].

Spartan model

The Spartan spatial random fields (SSRFs) are a relatively recently introduced family of geostatistical models [Hristopulos, 2003]. The SSRFs have been successfully applied in environmental risk assessment [Hristopulos and Elogne, 2007], atmospheric environment [Agou et al., 2019; Ẑukoviĉ and Hristopulos, 2008], hydrological data [Varouchakis, 2012; Varouchakis et al., 2012], and mining [Pavlides, 2016]. The SSRFs are generalized Gibbs random fields, equipped with a coarse-graining kernel that acts as a low-pass filter for the fluctuations [Hristopulos, 2003]. SSRFs are defined by means of physically motivated spatial interactions and a small set of free parameters.

Spartan model (d=1):

$$\gamma_0(r) = \begin{cases} \sigma_0^2 - \eta_0 e^{-|r|\beta_2/\xi} \left[\dfrac{\cos(|r|\beta_1/\xi)}{4\beta_2} + \dfrac{\sin(|r|\beta_1/\xi)}{4\beta_1} \right], & |\eta_1| < 2, \\[2ex] \sigma_0^2 - \eta_0 \dfrac{1 + |r|/\xi}{4}\, e^{-|r|/\xi}, & \eta_1 = 2, \\[2ex] \sigma_0^2 - \dfrac{\eta_0}{\sqrt{\eta_1^2 - 4}} \left[\dfrac{e^{-|r|\omega_1/\xi}}{2\omega_1} - \dfrac{e^{-|r|\omega_2/\xi}}{2\omega_2} \right], & \eta_1 > 2, \end{cases} \tag{2.15a}$$

Spartan model (d=2):

$$\gamma_0(\|\mathbf{r}\|) = \begin{cases} \sigma_0^2 - \dfrac{\eta_0}{\pi\sqrt{4 - \eta_1^2}}\, \mathfrak{I}\left[K_0\left(\dfrac{\|\mathbf{r}\|}{\xi}\omega_1 \right) \right], & |\eta_1| < 2, \\[2ex] \sigma_0^2 - \dfrac{\eta_0 \|\mathbf{r}\|}{4\pi\xi}\, K_{-1}\left(\dfrac{\|\mathbf{r}\|}{\xi} \right), & \eta_1 = 2, \\[2ex] \sigma_0^2 - \dfrac{\eta_0}{2\pi\sqrt{\eta_1^2 - 4}} \left[K_0\left(\dfrac{\|\mathbf{r}\|}{\xi}\omega_1 \right) - K_0\left(\dfrac{\|\mathbf{r}\|}{\xi}\omega_2 \right) \right], & \eta_1 > 2, \end{cases} \tag{2.15b}$$

Spartan model (d=3):

$$\gamma_0(\|\mathbf{r}\|) = \begin{cases} \sigma_0^2 - \dfrac{\eta_0\, e^{-\|\mathbf{r}\|\,\beta_2/\xi}}{2\,\pi\,\sqrt{4-\eta_1^2}} \left[\dfrac{\sin\left(\|\mathbf{r}\|\,\beta_1/\xi\right)}{\|\mathbf{r}\|/\xi}\right], & |\eta_1| < 2, \\ \sigma_0^2 - \dfrac{\eta_0\, e^{-\|\mathbf{r}\|/\xi}}{8\,\pi}, & \eta_1 = 2, \\ \sigma_0^2 - \dfrac{\eta_0}{4\,\pi\,\sqrt{\eta_1^2-4}} \left(\dfrac{e^{-\|\mathbf{r}\|\,\omega_1/\xi} - e^{-\|\mathbf{r}\|\,\omega_2/\xi}}{h}\right), & \eta_1 > 2. \end{cases} \tag{2.15c}$$

For the Spartan model, the variance σ_0^2 is determined from the hyperparameters η_0 (scale parameter), η_1 (rigidity coefficient) and ξ (characteristic length) as follows [Hristopulos, 2015a]:

$$(d=1) \quad \sigma_0^2 = \begin{cases} \frac{\eta_0}{2\sqrt{2+\eta_1}}, & |\eta_1| < 2, \\ \frac{\eta_0}{4}, & \eta_1 = 2, \\ \frac{\eta_0}{2\sqrt{\eta_1^2-4}} \left(\omega_1^{-1} - \omega_2^{-1}\right), & \eta_1 > 2\,. \end{cases} \tag{2.16a}$$

$$(d=2) \quad \sigma_0^2 = \begin{cases} \frac{\eta_0}{2\pi\sqrt{4-\eta_1^2}} \left[\frac{\pi}{2} - \arctan\left(\frac{\eta_1}{\sqrt{4-\eta_1^2}}\right)\right], & |\eta_1| < 2, \\ \frac{\eta_0}{4\pi}, & \eta_1 = 2, \\ \frac{\eta_0}{4\pi\sqrt{\eta_1^2-4}} \ln\left(\frac{\eta_1+\sqrt{\eta_1^2-4}}{\eta_1-\sqrt{\eta_1^2-4}}\right), & \eta_1 > 2\,. \end{cases} \tag{2.16b}$$

$$(d=3) \quad \sigma_0^2 = \begin{cases} \frac{\eta_0}{4\,\pi\sqrt{2+\eta_1}}, & |\eta_1| < 2, \\ \frac{\eta_0}{8\,\pi}, & \eta_1 = 2, \\ \frac{\eta_0}{4\,\pi\,\sqrt{\eta_1^2-4}} \left(\omega_2 - \omega_1\right), & \eta_1 > 2\,. \end{cases} \tag{2.16c}$$

In Eqs. (2.15) and (2.16), η_0 determines the total variance of the fluctuation, while $\eta_1 > -2$ is the rigidity hyperparameter (smaller η_1 allow oscillatory behavior of the covariance while $\eta_1 \geq 2$ lead to exponential decay). The hyperparameters $\omega_{1,2}$ and β_2 are dimensionless damping coefficients, β_1 is a dimensionless wave number. The coefficients $\beta_{1,2}$ are determined as $\beta_{1,2} = \frac{1}{2}|2 \mp \eta_1|^{1/2}$. The coefficients $\omega_{1,2}$ determine the decay of the slow and fast exponential functions (for $\eta_1 > 2$) and are given by means of $\omega_{1,2} = \left[\left(\eta_1 \mp \sqrt{\eta_1^2 - 4}\right)/2\right]^{1/2}$. The normalized lag vector is given by $\|\mathbf{h}\| = \|\mathbf{r}\|/\xi$, while $r = |\mathbf{r}|$ is the Euclidean norm and σ_X^2 is the variance. The exponential covariance function is obtained from Eq. (2.15c) for $\eta_1 = 2$ [Hristopulos, 2003, 2020].

In this case study, the exponential and the Spartan variogram models are used for the estimation of the spatial variability of the corresponding fields in Section 5.6.

2.9 Spatial Estimation (Kriging)

In geosciences, one of the most common problems to resolve is how to evaluate a variable in a location that does not coincide with the locations of the sampling network. Besides simple deterministic methods such as the inverse distance weighting or the natural neighbor interpolation [Mitas and Mitasova, 2005], stochastic methods are another family of approaches that can be employed for spatial prediction. Stochastic methods are in general more complex than deterministic methods, however, they provide more precise spatial predictions and additionally, a measure of their uncertainty.

Stochastic methods involve the use of multiple parameters that have to be estimated to optimally fit the spatial data. The estimate results from the optimization of a statistical measure, e.g. the maximization of likelihood [Fisher, 1922, 1925] or the minimization of the mean square estimation error [Chilès and Delfiner, 2012]. The stochastic spatial prediction methods are also known as kriging methods [Chilès and Delfiner, 2012; Christakos, 1992; Cressie, 1993; Olea, 1999; Wackernagel, 2003].

Kriging methods are named in honor of Danie Krige [Krige, 1951] who introduced stochastic linear interpolation in conjunction with the minimization of the mean square error of the estimate to estimate mineral deposits on unsampled locations [Matheron, 1963]. In most cases of practical interest, the ultimate goal is to estimate the field at a set of points. The estimated values at the new domain can be used for the construction of maps (e.g., precipitation maps) or the estimation of the concentration of the variable over that domain (e.g., mineral reserves). Kriging has been successfully used in environmental, meteorological and hydrological studies to generate spatial maps based on partial data [Agou et al., 2019; Boer et al., 2001; Guan et al., 2005; Moral, 2010; Verdin et al., 2016].

The problem of local estimation is usually expressed as follows: Based on a data set $x(\mathbf{s}_i)$, at $\mathbf{s}_i$ (where $i = 1, \ldots, N$) points located in a region Ω, calculate the value of the field at the estimation point $\mathbf{u} \in \Omega$, which does not coincide with any of the $\mathbf{s}_i$. The estimate at point $\mathbf{u}$ is denoted as $\hat{X}(\mathbf{u})$, while with $\hat{x}(\mathbf{u})$ we denote the specific value of the estimate derived from the available data, supplemented by an estimate of reliability, which determines the uncertainty of the estimation at each point [Agou, 2016].

Let $\omega(\mathbf{u})$ represent the correlation neighborhood of the point $\mathbf{u}$, which includes $n(\mathbf{u}) \leq N$ points than the size of $\mathbf{s}_i$. Due to computational difficulties encountered when processing large data sets, the estimation of the value $\hat{X}(\mathbf{u})$ is performed inside a neighborhood and not over the entire domain. The size of the neighborhood is defined in terms of the correlation length.

Kriging is a form of generalized linear regression that formulates the optimal estimator $\hat{X}(\mathbf{u})$ using linear weights that minimize the estimation error variance [Agou, 2016; Chilès and Delfiner, 2012]. The predictive equations used by kriging also appear in the framework of Gaussian process regression (GPR) [Hristopulos, 2020; Rasmussen and Williams, 2006]. Kriging and GPR methods assume that the data have an underlying joint normal (Gaussian) distribution, which simplifies the calculations and leads to explicit predictive expressions [Agou et al., 2022]. Several variations of kriging exist based on their underlying characteristics (e.g., ordinary kriging, regression kriging, co-kriging, etc.) [Goovaerts, 1997; Hristopulos, 2020; Journel, 1989]. Below, we present the most common formulation of kriging, the ordinary kriging (OK). Other formulations can derive from OK with the appropriate modifications to the mean $m_X(\mathbf{u})$ and the trend function. We present briefly the formulation of Simple and Regression Kriging.

Ordinary Kriging (OK)

Ordinary kriging (OK) is used when the mean value $m_X(\mathbf{u})$ is constant but unknown inside the local neighborhood $\omega(\mathbf{u})$ of the estimate point. The mean $m_X(\mathbf{s})$ may vary from neighborhood to neighborhood if the ordinary kriging is not applied over the entire domain.

The unknown local mean is filtered from the linear estimator by forcing the kriging weights λ_α to sum to one. This constraint enforces the zero bias condition. The ordinary kriging estimator $\hat{X}(\mathbf{u})$ is thus written as a linear combination of the $X(\mathbf{s}_\alpha)$ where $\alpha = 1, \ldots, n(\mathbf{u})$, as

$$\hat{X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha), \tag{2.17}$$

$$\text{with} \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1. \tag{2.18}$$

Equation (2.18) is the unbiasedness constraint. The estimator $\hat{X}(\mathbf{u})$ is a random variable, because it consists of a linear combination of random field values.

In the case of ordinary kriging, minimum mean square error should be calculated using the restriction imposed by the unbiasedness constraint. The minimization of the error variance under the non-bias condition $\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1$ uses the Lagrange multiplier method for constrained minimization. The error variance is calculated by means of the equation

$$\begin{aligned}\sigma^2_{E,OK}(\mathbf{u}) = \sigma^2_X(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\alpha \lambda_\beta \mathbb{E}[X'(\mathbf{s}_\alpha) X'(\mathbf{s}_\beta)] \\ -2 \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha \mathbb{E}[X'(\mathbf{s}_\alpha) X'(\mathbf{u})] + 2\mu \sum_{\beta=1}^{n(\mathbf{u})} (\lambda_\alpha - 1),\end{aligned} \tag{2.19}$$

where the constant 2μ is the Lagrange parameter. Using the covariance function, Eq. (2.19) is expressed as

$$\begin{aligned}\sigma^2_{E,OK}(\mathbf{u}) = \sigma^2_X(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\alpha \lambda_\beta \, c_X(\mathbf{s}_\alpha, \mathbf{s}_\beta) \\ -2 \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha c_X(\mathbf{s}_\alpha, \mathbf{u}) + 2\mu \sum_{\beta=1}^{n(\mathbf{u})} (\lambda_\alpha - 1).\end{aligned} \tag{2.20}$$

The optimal values of the linear weights and the parameter μ minimize the error variance. The weights are obtained by setting each of the $(n(\mathbf{u})+1)$ partial first derivatives equal to zero, i.e.,

$$\frac{\partial\sigma^2_{E,OK}(\mathbf{u})}{\partial\lambda_\alpha} = 0, \qquad \alpha = 1,\dots,n(\mathbf{u}), \tag{2.21}$$

$$\frac{\partial\sigma^2_{E,OK}(\mathbf{u})}{\partial\mu} = 0. \tag{2.22}$$

These conditions lead to the following linear system of equations for the linear weights,

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta \; c_X(\mathbf{s}_\alpha - \mathbf{s}_\beta) + \mu = c_X(\mathbf{s}_\alpha - \mathbf{u}), \qquad \alpha = 1,\dots,n(\mathbf{u}), \tag{2.23}$$

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1. \tag{2.24}$$

The above linear system of equations is written in the form of matrices as follows:

$$\begin{bmatrix} \sigma^2_X & c_X(\mathbf{s}_1 - \mathbf{s}_2) & \dots & c_X(\mathbf{s}_1 - \mathbf{s}_n) & 1 \\ c_X(\mathbf{s}_2 - \mathbf{s}_1) & \sigma^2_X & \dots & c_X(\mathbf{s}_2 - \mathbf{s}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_X(\mathbf{s}_n - \mathbf{s}_1) & c_X(\mathbf{s}_n - \mathbf{s}_2) & \dots & \sigma^2_X & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} c_X(\mathbf{s}_1 - \mathbf{u}) \\ c_X(\mathbf{s}_2 - \mathbf{u}) \\ \vdots \\ c_X(\mathbf{s}_n - \mathbf{u}) \\ 1 \end{bmatrix}. \tag{2.25}$$

The solution of the linear system is given by the following equation:

$$\lambda_\beta = C^{-1}_{\beta,\alpha} \; C_{\alpha,u}, \qquad \forall \quad \beta = 1,\dots,n(\mathbf{u}). \tag{2.26}$$

The optimal estimate of the kriging error variance is respectively given by the equation

$$\sigma^2_{E,OK}(\mathbf{u}) = \sigma^2_X - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha \; c_X(\mathbf{u}, \mathbf{s}_\alpha) - \mu, \tag{2.27}$$

with parameter $\mu < 0$ [Christakos, 1992; Goovaerts, 1997].

Simple Kriging (SK)

Simple kriging (SK) is a simple case of OK and is applied when the mean $m_X(\mathbf{u})$ is known and constant over the entire study area Ω, i.e. $\mathbb{E}[X(\mathbf{s})] = m_X$. In this case the kriging estimator is defined by the following equation:

$$\hat{X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha) - m_X \left[\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha - 1 \right]. \tag{2.28}$$

The Gaussian process regression (GPR) equations are similar to those of simple kriging and equivalent to those of universal kriging [Chilès and Delfiner, 2012]. In GPR the mean can comprise a superposition of basis functions with unknown coefficients (hyperparameters) which are estimated by maximizing the likelihood of the model [Rasmussen and Williams, 2006], while in simple kriging the mean is assumed constant and known [Agou et al., 2022].

Regression Kriging (RK)

With appropriate adaptations to the OK formulas, *Regression Kriging (RK)* is generated. RK combines a trend function with interpolation of the residuals. In RK the estimate is expressed as

$$\hat{X}(\mathbf{u}) = m_X(\mathbf{u}) + \hat{X}'(\mathbf{u}), \tag{2.29}$$

where $m_X(\mathbf{u})$ is the trend function, and $\hat{X}'(\mathbf{u})$ is the interpolated residual by means of OK [Rivoirard, 2002]. In the trend function a variety of auxiliary variables can be incorporated such as the topography of an area [Agou, 2016; Agou et al., 2019].

The method of regression kriging is used in applications, such as mapping of leaf area index (LAI) [Berterretche et al., 2005], mapping soil particle size fractions [Wang et al., 2020], mapping of precipitation height [Agou, 2016; Agou et al., 2019], mapping of groundwater levels [Varouchakis et al., 2012], and mineral resources [Hristopulos et al., 2021]. Generally, kriging methods are considered robust spatial prediction approaches that can be used to interpolate a variety of environmental variables. Kriging formulations that allow the incorporation of auxiliary variables (such as topographical characteristics) are particularly powerful, especially for the analysis of variables that are affected by various parameters, such as meteorological variables. Additionally, they allow the interpretation of the separate components. The complexity both of the implementation and the interpretation of RK makes its use limited.

In this thesis, we use OK to estimate the precipitation over the entire grid which covers the island of Crete based on the ERA5 precipitation data. The optimization of the parameters of the model is performed by maximizing the likelihood of the sample data [Fisher, 1922, 1925] and by minimizing the mean square estimation error [Chilès and Delfiner, 2012]. We use several cases of the initial precipitation values, including the original aggregated monthly precipitation, the normalized values calculated by means of the Gaussian Anamorphosis with Hermite polynomials, and the Monte Carlo simulations of the original or the normalized values (Section 5.6).

2.10 Spatial Model Estimation

Maximum likelihood estimation (MLE) is a method of estimating the optimal parameters of a predefined spatial model to fit the sample data [Fisher, 1922, 1925; Norden, 1972]. The estimation of the optimal parameters is a result of the maximization of the probability that the sample values will be generated for the given set of parameters [Chilès and Delfiner, 2012]. Applications of the MLE method for spatial data are widely available [Kitanidis and Lane, 1985; Mardia and Marshall, 1984; Mardia and Watkins, 1989].

For a sample set $\mathbf{x}$ with values $x_1, x_2, \ldots, x_n$ and a candidate parameter vector $\boldsymbol{\theta}$ for the determined spatial model, the optimal parameter vector $\hat{\boldsymbol{\theta}}$ can be estimated by maximizing the likelihood of the data set as follows

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \mathbf{x}), \tag{2.30}$$

where $\mathcal{L}(\boldsymbol{\theta}; \mathbf{x}) = f_x(\mathbf{x}|\boldsymbol{\theta})$, and f_x is the joint the probability of the data. To discourage the fast changes of the likelihood, the maximization of its logarithm (logarithm is a monotonically increasing function) is broadly calculated, or equivalently the minimization of the negative log-likelihood.

2.11 Simulation

In the earth sciences, oftentimes it is required to generate a realization of a domain based on sample values. Commonly used techniques include spatial and temporal interpolation such as kriging. A drawback of those methods is that they tend to over-smooth the estimation field. Simulation techniques are capable of estimating even the lower and higher probabilities, therefore they capture the true variability of the process. This is very important especially for the assessment of extreme events, and resource planning.

Simulation methods involve the generation of synthetic realizations based on the joint PDF of the field, respecting the field's basic characteristics such as the mean value, the variance, and the variogram. Simulations can be divided into two main categories, the conditional and the unconditional simulations. The former category additionally to the main statistical properties of the field also preserves the field values locally, while the latter does not. Most broadly encountered in field studies is the conditional simulation due to the constraints involved [Hristopulos, 2020].

Simulation approaches are complimentary to kriging because of their ability to provide a deeper understanding of the uncertainty of the estimated process due to the multiple calculated realizations. However, this advantage has a high computational cost, especially when the estimation grid is large.

Monte Carlo simulation approaches are very popular, and they can be divided into rejection sampling, importance sampling, and Markov Chain Monte Carlo. In Monte Carlo simulation a part of the simulated states is used and not all of them; this is the reason why there is a debate on whether they belong to the simulation methods or not. Other simulation techniques include the covariance matrix decomposition combined with kriging conditioning [Goovaerts, 1997; Pavlides, 2016], spectral methods such as the Fast-Fourier-Transform, sequential simulation, and simulated annealing [Hristopulos, 2020].

2.12 Cross Validation

Several methods are available for the evaluation of the predictive performance of the model. Those include the aforementioned MLE, and method of moments (MoM), as well as the Cross-validation (CV). CV is a very popular statistical approach often used for the selection of the model, the model's parameters, and the assessment of the model's performance. Essentially, to apply CV the data are split into two different sets, the training, and the validation set. The training set is used for the estimation of the model (or the model parameters) and the validation set is used for the comparison of the estimates at the locations of the validation set with the real values [Hristopulos, 2020].

The selection of the "best" model lies in the minimization of a statistical error measure (e.g., mean error) or a combination of multiple measures. In summary, CV combines average measures of fit (i.e., prediction error) to minimize the estimation error and obtain a more precise estimate of the model's predictive performance [Grossman , edit.].

Several statistical measures of predictive performance can be used in CV; a list is given below. In classification problems, CV is shown to be comparable with the bootstrap and the Akaike selection criterion in terms of model selection performance [Hastie et al., 2009; Hristopulos, 2020].

For the application of CV, three different strategies can be followed to split the data into training and validation sets: the k-fold CV, the leave-P-out CV, and the leave-one-out CV.

k-fold cross-validation In k-fold cross-validation, the sample is partitioned into k disjoint subsets (folds) of approximately equal size [Kohavi, 1995]. For k times, each of the k folds is selected as the validation set, while the remaining sets serve as the training set. The validation measures are obtained as the average over the k configurations. A standard number of splits is $k = 4$ or $k = 10$, but generally, k is an unfixed parameter [McLachlan et al., 2004]. We use the k-fold CV in the Random Forests application in Section 6.5.

Leave-p-out cross-validation The main idea in Leave-p-out cross-validation (LPO) is that the validation set contains p observations, and the remaining values specify the training set. The process of splitting the data and estimating-validating the model can be performed in an exhaustive way (all possible combinations of sets with p and $n-p$ points) or the number of splitting repetitions can be set at a specified low number to avoid the computational cost.

Leave-one-out cross-validation Leave-one-out cross-validation (LOOCV or LVO) is a special case of leave-p-out CV if $p = 1$, or a special case of the k-fold CV if $k = n$ (the number of observations). The advantage of the LOOCV is the low computational strain compared to the other approaches, and also the fact that the LOOCV error estimate is considered an almost unbiased estimate of the true error [Varma and Simon, 2006]. We use the LOOCV in the Stochastic Local Interaction models application as well as in the Gaussian Anamorphosis applications in Sections 7.8.1, and 5.6 respectively.

Cross-validation Metrics

To appraise the model's performance the following measures are commonly used. Those include: the mean error (bias) ($\varepsilon_{\text{bias}}$), the mean absolute error (ε_{MA}), the root mean square error (ε_{RMS}), the mean absolute relative error (ε_{MAR}), the root mean square relative error ($\varepsilon_{\text{RMSR}}$), the Pearson's linear correlation coefficient (ρ_{P}) the Spearman's (rank) correlation coefficient (ρ_{rank}), the minimum error (ε_{min}), and the maximum error (ε_{max}). Below we define these measures in the case of LOOCV. The values $\hat{x}_{-i}(\mathbf{s}_i)$ and $x(\mathbf{s}_i)$ are, respectively, the estimated (based on the $N-1$ data points excluding $\mathbf{s}_i$) and true value of the field at point $\mathbf{s}_i$, $\overline{x(\mathbf{s}_i)}$ denotes the spatial average of the data and $\overline{\hat{x}_{-i}(\mathbf{s}_i)}$ the spatial average of the predictions, while N is the number of observations [Agou, 2016].

Mean error (bias) (ME):

$$\varepsilon_{\text{bias}} = \frac{1}{N}\sum_{i=1}^{N}\left[\, x(\mathbf{s}_i) - \hat{x}_{-i}(\mathbf{s}_i)\,\right]. \tag{2.31}$$

The bias of the predictor is a crucial measure but needs to be interpreted carefully. For example, low mean error values do not necessarily indicate low bias but they may also signify that high positive errors are canceled out by large negative errors.

Mean absolute error (MAE)

$$\varepsilon_{\text{MA}} = \frac{1}{N}\sum_{i=1}^{N} \left| \hat{x}_{-i}(\mathbf{s}_i) - x(\mathbf{s}_i) \right| . \tag{2.32}$$

The absolute value in the formulation of the MAE helps to overcome the problem of ME to counterbalance large positive and negative errors. Thus, providing a measure of the true magnitude of the deviations between the estimations and the true values [Hristopulos, 2020].

Root mean square error (RMSE):

$$\varepsilon_{\text{RMS}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left[\hat{x}_{-i}(\mathbf{s}_i) - x(\mathbf{s}_i) \right]^2} . \tag{2.33}$$

Similar to the ε_{MA}, the ε_{RMS} measures the magnitude of the deviations between the estimations and the true values, however, it weighs more in favor of large errors.

Mean absolute relative error (MARE):

$$\varepsilon_{\text{MAR}} = \frac{1}{N}\sum_{i=1}^{N} \left| \frac{\hat{x}(\mathbf{s}_i) - x(\mathbf{s}_i)}{x(\mathbf{s}_i)} \right| . \tag{2.34}$$

The ε_{MAR} is a dimensionless measure that represents the difference between the true values and the approximations. It is important to note that the ε_{MAR} is undefined if the true value is zero, and it is a meaningful measure when the variable in question is measured on a ratio scale (i.e., values do not fall below zero). Precipitation is measured on a ratio scale, therefore the ε_{MAR} will not be sensitive to measurement units. However, in multiple time scales, true precipitation is zero, resulting in undefined MARE values.

Root mean square relative error (RMSRE):

$$\varepsilon_{\text{RMSR}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\hat{x}_{-i}(\mathbf{s}_i) - x(\mathbf{s}_i)}{x(\mathbf{s}_i)}\right]^2}. \tag{2.35}$$

The characteristics of the $\varepsilon_{\text{RMSR}}$ are similar to the ε_{MAR}, but it weighs heavier for large error values.

Pearson's Linear correlation coefficient (RP):

The correlation coefficient, ρ_{P}, measures the correlation between two variables. The formula for Pearson's linear correlation coefficient ρ_{P} is [Isaaks and Srivastava, 1989]

$$\rho_{\text{P}} = \frac{\mathbb{C}\mathbb{O}\mathbb{V}(x,\hat{x})}{\sigma_x\,\sigma_{\hat{x}}} = \frac{\sum_{i=1}^{N}\left[x(\mathbf{s}_i) - \overline{x(\mathbf{s}_i)}\right]\left[\hat{x}_{-i}(\mathbf{s}_i) - \overline{\hat{x}_{-i}(\mathbf{s}_i)}\right]}{\sqrt{\sum_{i=1}^{N}\left[x(\mathbf{s}_i) - \overline{x(\mathbf{s}_i)}\right]^2}\sqrt{\sum_{i=1}^{N}\left[\hat{x}_{-i}(\mathbf{s}_i) - \overline{\hat{x}_{-i}(\mathbf{s}_i)}\right]^2}}. \tag{2.36}$$

The Pearson correlation coefficient measures the dispersion of estimates with respect to the observed values and is a reliable measure when linear correlations are expected. The ρ_{P} takes values in the interval $-1 \le \rho_{\text{P}} \le 1$. The higher the $|\rho_{\text{P}}|$ value, the higher the linear correlations between the data sets.

Spearman (rank) correlation coefficient (RS)

$$\rho_{\text{rank}} = 1 - \frac{6\sum_{i=1}^{N}(R_{x_i} - R_{\hat{x}_i})^2}{N(N^2-1)}, \tag{2.37}$$

where R_{x_i} is the rank (order) of x_i among all the other x values. The lowest x value appears first on a sorted list and therefore receives a rank of 1; the highest x value appears last on the list and receives a rank of N.

Similar to Pearson's correlation coefficient, the Spearman correlation coefficient ρ_{rank} measures the correlation between two variables and ranges between -1 and 1, however, the ρ_{rank} can characterize non-linear relations. Large differences between ρ_{rank} and ρ_{P} are often quite revealing about the existence of extreme pairs on the scatterplot.

Error minimum

$$\varepsilon_{\min} = \min_{i=1,\dots,N} \{ x(\mathbf{s}_i) - \hat{x}_{-i}(\mathbf{s}_i) \} \tag{2.38}$$

The $\varepsilon_{\min}$ shows the minimum estimation error over all locations.

Error maximum

$$\varepsilon_{\max} = \max_{i=1,\dots,N}, \{ x(\mathbf{s}_i) - \hat{x}_{-i}(\mathbf{s}_i) \} \tag{2.39}$$

The $\varepsilon_{\max}$ shows the maximum estimation error over all locations.

Chapter 3

Exploratory Data Analysis

3.1 Summary

This chapter focuses on the study area and introduces the data sets (26 atmospheric reanalysis variables from the beginning of 1979 to February 2020) that will be used in the following chapters of this thesis. Initially, general information about the study area is presented, including topographic and demographic features. Then, the data are described followed by pre-processing operations on the main meteorological variables (precipitation and temperature). Finally, summary statistical properties for the monthly precipitation and temperature data for the wet season are illustrated by means of auxiliary tables and figures. The summary measures include the mean value, the median, the minimum and maximum values, the standard deviation, the coefficient of variation, the skewness and the kurtosis. This chapter is supplemented by Appendix B where the corresponding tables and figures for the daily, weekly and annual timescales are presented.

3.2 Information about the Study Area

The study area is the island of Crete (Greece) in the southeastern part of the Mediterranean basin. Crete is the largest island in Greece with an area of 8 336 km^2, length of 260 km, width ranging from 12 km to 57 km, and maximum elevation of 2 456 m [Hellenic Statistical Authority, 2014]. The island's climate exhibits a transition from Mediterranean to semi-arid as is common in Mediterranean regions [Agou et al., 2019; Watrous, 1982]. In spite of its rather small geographical size, temperature and precipitation exhibit significant local variations due to three mountain ranges which are among the highest in Europe. From west to east, the island is divided into four administrative regional units: Chania, Rethymno, Heraklion, and Lasithi.

The population of Crete approaches 624 408 people according to the 2021 census [Hellenic Statistical Authority, 2023]. Most of the population lives closer to the coastal areas where the most important agricultural areas are located (Fig. 3.1). Messara valley in the south of Heraklion prefecture is the island's largest and most productive plain. The valley of Ierapetra in the south-east also has significant agricultural activity. Agriculture is an important revenue contributor for the Cretan region, accounting for 13% of the local Gross Domestic Product (GDP) [Chartzoulakis et al., 2001].

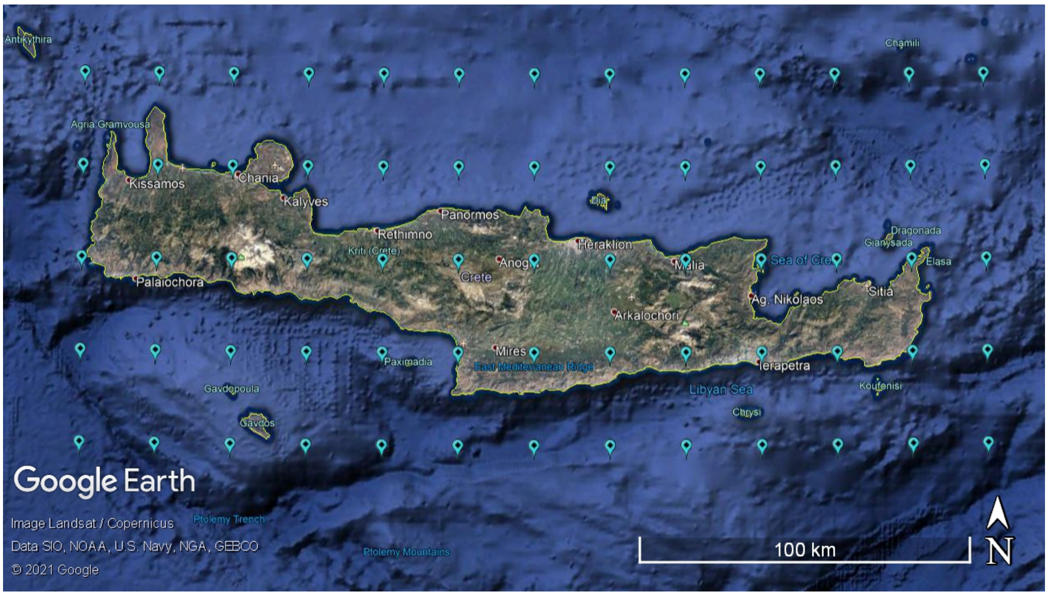


Figure 3.1: Geomorphological map of Crete showing the 65 ERA5 grid locations (blue markers) used in this study [Google Earth, 2015].

3.3 Data sets

Drought monitoring has become a prevalent research topic in the last century. The majority of the available studies from past decades involve data from ground stations. The low availability of older studies with applications on satellite images is due to the technical limitation that comes with big size data sets, as well as the low resolution that satellite and radar records used to have.

As remote sensing technology developed, and satellite or reanalysis data sets were widely and freely available for academic purposes, their use started to grow. Nowadays, satellite data consist of long-term and large-scale meteorological data that can be used to monitor environmental disasters [Tian et al., 2014; Xue et al., 2019]. Their spatial resolution commonly ranges between 0.25°(≈ 31 km) to 2.5°(≈ 310 km), while their time resolution ranges from 1 hour to one month, covering the globe.

Reanalysis is a systematic methodology that employs data assimilation and numerical methods to generate weather and climate products over high-resolution grids [Dee et al., 2016]. Data assimilation entails the application of mathematical tools to fuse data from many sources. Reanalysis products may contain bias due to errors and approximations in the observations and models used. This study does not employ bias correction approaches since the goal is to validate the proposed methodologies rather than to compare reanalysis-based interpolation to results obtained from ground measurements.

ERA5 is a climate reanalysis data set (5th generation) from ECMWF (the European Centre for Medium-Range Weather Forecasts) with a spatial resolution of 0.25°(31 km), lower time resolution at 1 hour, 137 vertical levels from the surface up to a height of 80 km into the atmosphere, and is spanning the period 1950 to present (available for use in 2020). Those high-resolution data sets can be utilized for weather and climate analysis and upcoming disaster prediction, but also for the assimilation of a sparse ground station network, like in our case. Recently, studies use reanalysis data for climate simulation prediction [Chen et al., 2019; Wang et al., 2016].

Several studies use the ERA-Interim data, the predecessor of the ERA5, to produce a global climatology of the stratosphere-troposphere exchange [Škerlak et al., 2014], to estimate the variability of the temperature globally [Simmons et al., 2014], to detect surface and upper-air temperature and humidity trends over Greece [Tzanis et al., 2019], or to estimate drought indices [Mavromatis and Voulanas, 2020]. According to Hassler and Lauer [2021], ERA5 outperforms ERA-Interim data. Although biases are still present in ERA5 data, issues such as the wet bias over Central Africa and the Indian Ocean and the dry bias over Northern Hemisphere continental regions are significantly decreased when compared to ERA-Interim. Furthermore, ERA5 show smaller biases in precipitation against other commonly used datasets such as JRA-55 and MERRA-2.

3.4 ERA5 Data

The ERA5 reanalysis data collection used to support this study was downloaded from the Copernicus Climate Change Service [Copernicus Climate Change Service C3S, 2018]. Based on the availability of the data, we selected twenty six (26) atmospheric reanalysis variables from 1979 to 2020, with the main variables being total precipitation and near-surface temperature. The entire collection of the atmospheric variables processed in this thesis are presented in Table 3.1 with their corresponding units. For a more detailed description of the variables, refer to the Climate's Data Store variable description available online in the Table: (MAIN VARIABLES).

3.4.1 Precipitation Data

The ERA5 reanalysis precipitation data correspond to hourly precipitation amount measured in m [Copernicus Climate Change Service C3S, 2018]. The data set includes 23 360 610 values of hourly total precipitation for a period of 41 years (from 01-Jan-1979 06:00:00 to 31-Dec-2019 23:00:00) at the nodes of a 5×13 spatial grid (see Fig. 3.1); the grid nodes cover the Greek island of Crete (see Fig. 3.1). The average spatial resolution is ≈ 0.28 degrees (grid cell size $\approx$ 31km). A total of 359 394 hourly precipitation values are available at each node.

Precipitation is higher at the west part of the island than it is in the rest of the island with the highest precipitation values for the weekly, monthly and annual

Table 3.1: ERA5 reanalysis atmospheric variables available from the Copernicus Climate Change Service [Copernicus Climate Change Service C3S, 2018]. All listed variables represent surface values.

Variable name	Unit	Type
10 metre U wind component	m/s	instantaneous
10 metre V wind component	m/s	instantaneous
2 metre dewpoint temperature	K	instantaneous
2 metre temperature	K	instantaneous
Evaporation	m of water equivalent	accumulations
Mean sea level pressure	Pa	instantaneous
Runoff	m	accumulations
Sea surface temperature	K	instantaneous
Snow density	kg/m^3	instantaneous
Snow depth	m of water equivalent	instantaneous
Snowfall	m of water equivalent	accumulations
Soil temperature level 1	K	instantaneous
Surface latent heat flux	J/m^2	accumulations
Surface sensible heat flux	J/m^2	accumulations
Surface net solar radiation	J/m^2	accumulations
Surface net thermal radiation	J/m^2	accumulations
Surface solar radiation downward clear-sky	J/m^2	accumulations
Surface solar radiation downwards	J/m^2	accumulations
Surface thermal radiation downward clear-sky	J/m^2	accumulations
Surface thermal radiation downwards	J/m^2	accumulations
Top net solar radiation	J/m^2	accumulations
Top net thermal radiation	J/m^2	accumulations
Total cloud cover	(0-1)	instantaneous
Total precipitation	m	accumulations

timescales occurring to the closest data node to the city center of Chania. Additionally, precipitation across the island, both geographically (east versus west) and physiographically (plains versus mountainous areas) varies greatly. Monthly precipitation peaks in December or January and attains a minimum in July and August which are almost dry months across the low-lying areas of Crete [Agou, 2016; Region of Crete Information Bull., 2002]. During the dry season months,

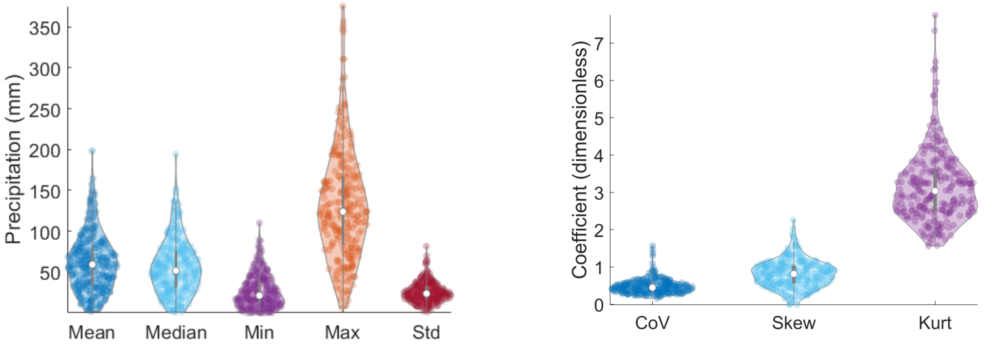
Table 3.2: Mean, median, minimum and maximum values (shown across rows) of monthly ERA5 precipitation statistics (shown across the columns) based on 246 monthly values. Each monthly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	61.25	55.69	26.19	132.70	25.53	0.48	0.82	3.16
Median	59.19	51.78	21.23	123.98	23.67	0.45	0.81	3.04
Minimum	1.75	1.05	0.05	6.10	1.16	0.16	−0.01	1.56
Maximum	198.27	194.15	110.03	375.32	81.54	1.57	2.26	7.75

that is from April to September, almost zero precipitation is present in the island. Therefore, for the monthly timesscale our analysis focuses on the wet period, that is from October to March [Nastos and Zerefos, 2009]. Daily, weekly, monthly and annual precipitation data sets were generated by aggregating the hourly values at each location over respective time windows.

Table 3.2 presents the summary statistics of the monthly precipitation data for the wet season. From the data of every time step (246 months) the mean value, the median, the minimum and maximum values, the standard deviation, the coefficient of variation, the skewness and the kurtosis are calculated. The mean, median, minimum, and maximum values of the resulted collection are calculated and presented in Table 3.2. The table is supplemented by Fig. 3.2 which shows the probability distribution of the monthly statistics (corresponding to different columns of Table 3.2) calculated over the 246 months. These plots exhibit asymmetric distribution of the statistics and considerable dispersion. The non-zero skewness, the deviation of the minimum and maximum kurtosis from the Gaussian value of three, and the unsuccessful fitting of the monthly histograms to the normal distribution (see Fig. 3.3 and Table B5), strongly suggest that monthly precipitation data follow non-Gaussian distributions. For the daily, weekly, and annual timescale see Appendix B.

To investigate the deviations from Gaussian behavior the data are first grouped



(a) Statistics measured in mm

(b) Dimensionless statistics

Figure 3.2: Violin plots for the mean, median, minimum and maximum values of monthly ERA5 precipitation statistics based on 246 monthly values. Each monthly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

by location and then by month. The following probability distribution models are tested: generalized Pareto, inverse Gaussian, lognormal, t-Scale location, Generalized Extreme Value, Weibull, Gaussian, Birnbaum-Saunders, exponential, extreme value, gamma, Nakagami, logistic, log-logistic, Rayleigh, and Rician. For the monthly precipitation data according to Akaike's Information Criterion (AIC), the Nakagami model is optimal at 45 of 65 nodes, the Weibull at 14, the gamma at 4 and the Rayleigh distribution at the remaining two locations. The results based on the Bayesian Information Criterion (BIC) are similar, with the Nakagami model being optimal at 40 of 65 nodes, the Weibull at 12, the gamma at 4, and the Rayleigh distribution at the remaining 9 nodes (Table B4). For the daily precipitation data according to Akaike's Information Criterion (AIC), the generalized Pareto model is optimal at all 65 nodes. The results based on the Bayesian Information Criterion (BIC) are identical. For the weekly scale, the generalzied Pareto distribution is optimal at 58 (or 50 based on the AIC) of 65 nodes, the gamma at one node, and the exponential at the remaining nodes. For the annual precipitation, AIC and BIC select the Nakagami model at 17 nodes, the gamma at 16, the inverse Gaussian at 8, the Rician at 8, the Birnbaum-Saunders at 5, the Weibull and the logistic at 4 each, the log-logistic at 2, and the lognormal at the last node.

For the monthly precipitation data grouped by month, the optimal distribution according to AIC is the Nakagami model for 8 out of the 246 wet months, the Weibull for 3, the gamma for 13, the GEV for 25, the Rayleigh for 1, the generalized Pareto for 126, the log-logistic for 2, the lognormal for 3, the Birnbaum-Saunders for 25, and the inverse Gaussian distribution for the remaining 40 wet months. Similar are optimal fits according to BIC, concluding to the generalized Pareto for most of the time steps (Table B5).

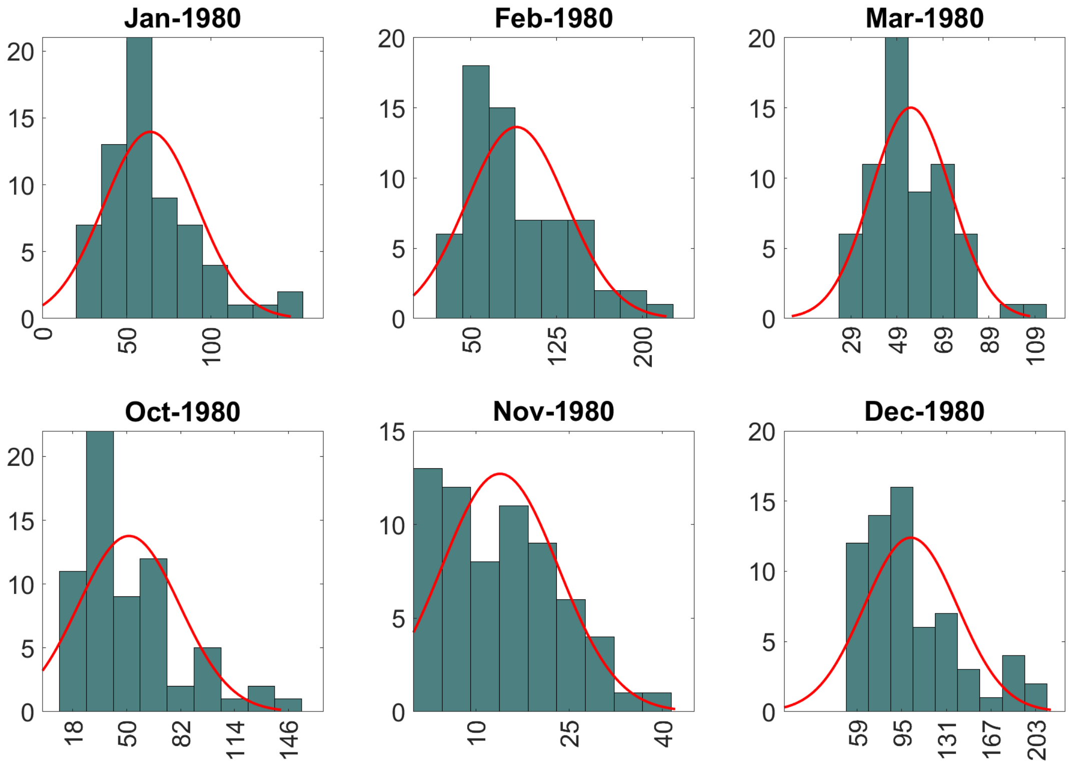


Figure 3.3: Distribution of monthly precipitation during the wet season of 1980. Histograms are based on ERA5 precipitation data at 65 grid locations over and around the island of Crete. Best fits to the optimal Gaussian PDF models (red line) are also shown. The vertical axis of the histograms represents frequency; the horizontal axis represents precipitation amount measured in mm.

According to calculations based on the hourly, daily, and monthly precipitation data, the characteristic year is 2006, 2006 and 1980 respectively. For

Table 3.3: Optimal probability distribution fits (based on BIC and AIC) for the monthly ERA5 precipitation data in the year 1980. The models studied include the following: "GP": Generalized Pareto, "InvGauss": Inverse Gaussian, "B-S": Birnbaum-Saunders, and "GEV": Generalized Extreme Value distribution.

	January	February	March	October	November	December
BIC	GP	InvGauss	B-S	GP	GP	GEV
AIC	GP	InvGauss	B-S	InvGauss	GP	InvGauss

illustration, the precipitation probability distributions for the year 1980 are investigated. Sixteen parametric probability distribution models (as listed above) were tested. The optimal probability model for the monthly data per each wet season month is presented in Table 3.3 (based on the BIC and AIC) (see also Fig. 3.3). The optimal distribution for most months is the generalized Pareto (GP) with the AIC and either the GP or the inverse Gaussian based on the BIC. The best model in these cases only mean that the GP (or the inverse Gaussian) achieves a better AIC (or BIC) value than the other models, but it does not ensure that the model is an accurate representation of the empirical distribution.

3.4.2 Temperature

The ERA5 reanalysis temperature data set includes 23 361 000 values of hourly 2 metre temperature (from now on referred to simply as temperature) for a period of 41 years (from 01-Jan-1979 00:00:00 to 31-Dec-2019 23:00:00) at the nodes of a 5×13 spatial grid that cover the Greek island of Crete (see Fig. 3.1). The average spatial resolution is ≈ 0.28 degrees (grid cell size $\approx$ 31km). A total of 359 400 hourly temperature values are available at each node. It must be noted that in the case of the instantaneous variables each location includes 6 more values. We have 390 more values in the entire temperature data set than in the precipitation (accumulations) data set because the first value of the set corresponds to the 01-Jan-1979 00:00:00 instead of the 01-Jan-1979 06:00:00.

Temperature is higher at the east part of the island than it is in the rest of the island with the highest temperature values for the daily, weekly, and monthly timescales for the ERA5 data set occurring in the south part of Heraklion prefecture while for the annual timescale the highest temperature is measured at the south-eastern part the entire grid. Our analysis focuses on the full period for all the timescales. Daily, weekly, monthly and annual temperature data sets were created by averaging the hourly values at each location over respective time windows.

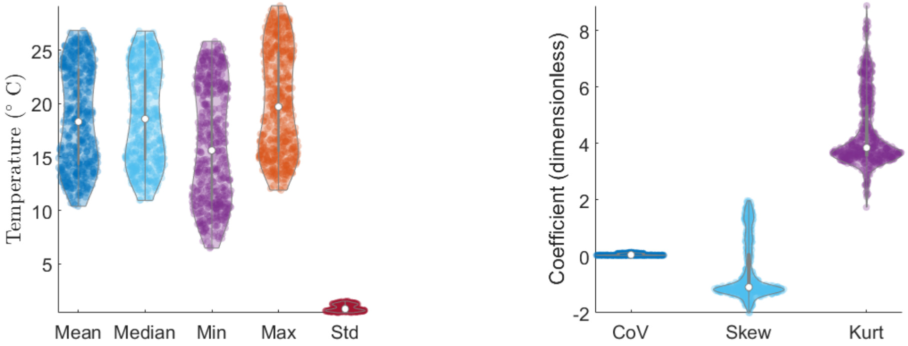
Table 3.4 presents the summary statistics of the monthly temperature data for the full period. They include the mean value, the median, the minimum and maximum values, the standard deviation, the coefficient of variation, the skewness and the kurtosis. The way to read this table is as follows: the second column corresponds to the minimum value (evaluated over all months) of the monthly statistic shown along a given row (evaluated for each month from the 65 sites). The table is supplemented by Fig. 3.4 which shows the probability distribution of the monthly statistics (corresponding to different columns of Table 3.4) calculated over the 492 months. Unlike the precipitation plots (Fig. 3.2) which exhibit highly asymmetrical distribution of the statistics, the plots of the temperature statistics as shown in Fig. 3.4 are distributed around a range of values without big dispersion almost presenting a bimodal distribution pattern. The non-zero skewness, the deviation of the minimum and maximum kurtosis from the Gaussian value of three, and the unsuccessful fitting of the monthly histograms to the normal distribution (see Fig. 3.5), strongly suggest that monthly temperature data follow non-Gaussian distributions. For the daily, weekly, and annual timescale see Section B2 in Appendix B.

Similarly to the precipitation data, we investigate the deviations of the temperature data from the Gaussian by location and then by month. The same seventeen probability distribution models tested for the precipitation data are tested for the temperature data. According to both criteria, AIC, and BIC, the GP model is optimal for the monthly temperature for all 65 nodes. The statistical criteria agree on the optimal model except for the annual scale (Table B9).

For the monthly temperature data grouped by month, the optimal distribution according to AIC is the GP model for 229 out of the 492 months, with second best the GEV for 114 months. Similar optimal fits gave the BIC, concluding to the generalized Pareto for most of the time steps (Table B10). For the data grouped

Table 3.4: Mean, median, minimum and maximum values (shown across rows) of monthly ERA5 mean hourly temperature statistics (shown across the columns) based on 492 monthly values (full period). Each monthly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	18.59	18.79	16.18	20.09	0.88	0.05	−0.57	4.41
Median	18.31	18.57	15.62	19.71	0.78	0.04	−1.10	3.84
Minimum	10.39	10.94	6.46	11.89	0.41	0.02	−2.01	1.72
Maximum	26.85	26.79	25.83	29.16	1.50	0.13	1.96	8.87



(a) Statistics measured in °C

(b) Dimensionless statistics

Figure 3.4: Violin plots for the mean, median, minimum and maximum values of monthly ERA5 temperature statistics based on 246 monthly values. Each monthly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

by the time-step for all the analyzed timescales, see Table B10 in Appendix B.

For illustration, the temperature probability distributions for the year 1980 are investigated. Seventeen parametric probability distribution models (as listed above) were tested. The optimal probability model per month is presented in Table 3.5 (based on BIC). The optimal distribution for most months is the gen-

Table 3.5: Optimal probability distribution fits (based on BIC and AIC) for the monthly ERA5 temperature data in the year 1980. The models studied include the following: "GP": Generalized Pareto, "GEV": Generalized Extreme Value, "t-scale": t-scale location, "Log": Logistic, and "EV": Extreme Value distribution.

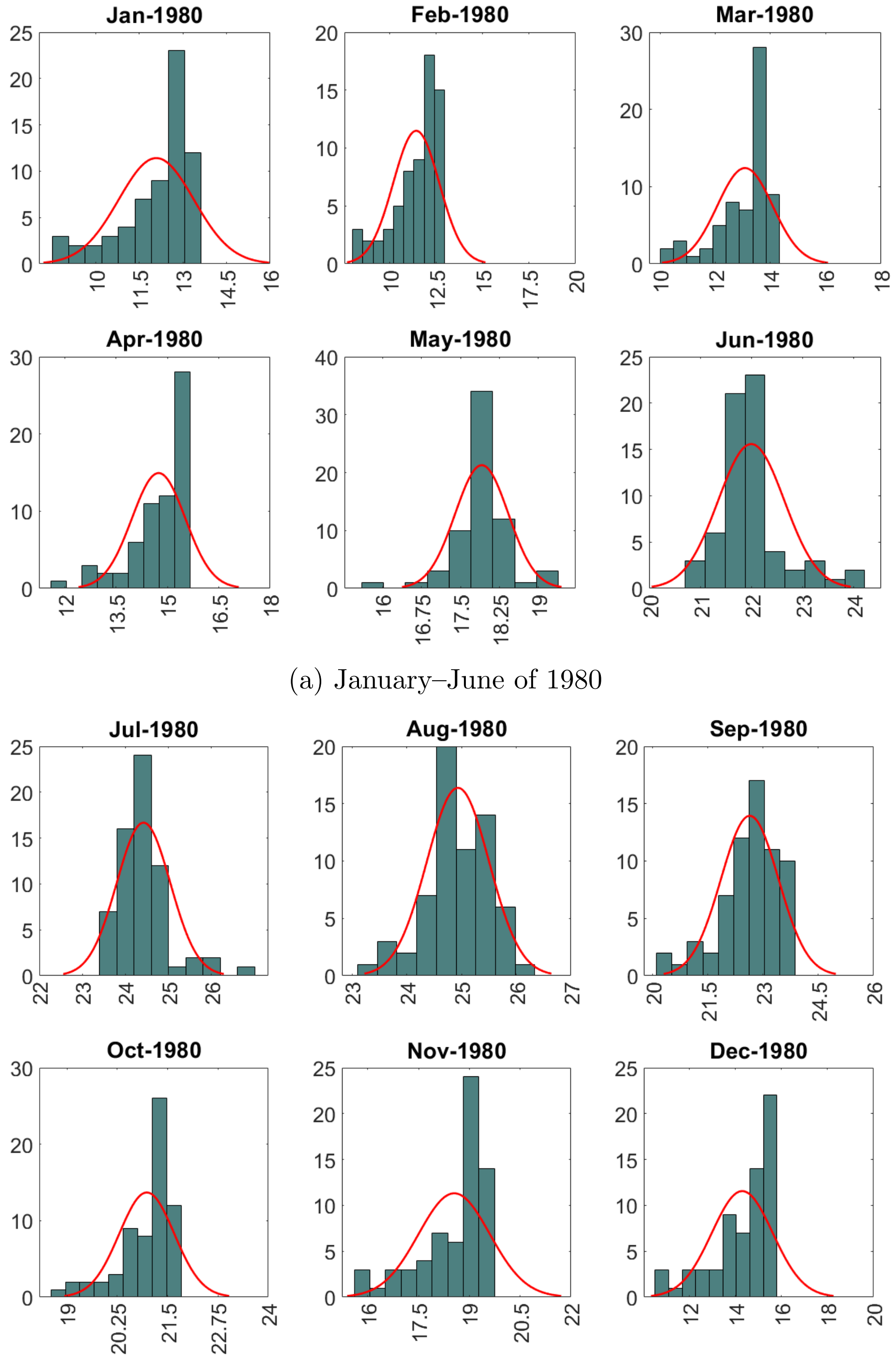
	January	February	March	April	May	June
BIC	GP	GP	GEV	GEV	t-scale	t-scale
AIC	GP	GP	GEV	GEV	t-scale	t-scale

	July	August	September	October	November	December
BIC	GEV	t-scale	Log	EV	GEV	GP
AIC	GEV	t-scale	Log	GEV	GEV	GP

eralized extreme value (GEV) (Table 3.5).

3.5 Computational environment

We implement the data analysis in the Matlab programming environment (Version 2018b) and in Python 3 (Version 3.8.2 64-bit) on a quad core Intel(R) Core(TM) i5-4570 CPU at 3.20GHz workstation with 24 GB installed RAM memory, running a 64-bit Windows 10 Education operating system.



(a) January–June of 1980

(b) July–December of 1980

Figure 3.5: Distribution of monthly temperature of January till June (top (a)) and July till December (bottom (b)) of 1980. Histograms are based on ERA5 temperature data at 65 grid locations over and around the island of Crete. Best fits to the optimal Gaussian PDF models (red line) are also shown. The vertical axis of the histograms represents frequency; the horizontal axis represents temperature measured in °C.

Chapter 4

Estimation of Drought Indices for the Island of Crete

4.1 Summary

In this chapter, we define drought, drought events and comment on their socio-economic consequences such as habitat loss and famine. Moreover, we provide details about the characteristics and the computational steps needed for the estimation of the Standardized Precipitation Index (SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI), which are used in this research. We use SPI and the SPEI drought indices to estimate and characterize the severity and intensity of droughts on the island of Crete. We calculate the indices based on the ERA5 monthly data. The spatial distribution of each index is visualized with interpolation maps for selected time steps, while their temporal distribution is presented in time-series plots that illustrate the temporal trends at different locations. For SPEI estimation, we first calculate the Potential Evapotranspiration (PET) from the temperature data.

4.2 Introduction

Drought is a natural disaster caused by water shortage over a prolonged period, resulting in disturbance of the water balance equilibrium. The consequences of a drought event for an area can be quite catastrophic, including deaths from water and food shortage. According to the World Meteorological Organization (WMO) [2014], for the period 1970–2012, there have been reported almost 690 000 deaths globally attributed to droughts. Although drought events occupy only 6% of all the natural disasters for that period, they result in 35% of the total deaths (1 944 653 deaths) and 8% of the total economic losses (2 390.7 billion USD). The correlation of the event occurrence and the death expectancy can characterize droughts as the highest fatality rate disaster. Drought monitoring and prediction are highly studied in the last century; however, most researchers used weather station data, which have their limitations as described in Section 1.5.1.1. The evolution of remote sensing has provided an abundance of new data products, making it easier to investigate regions that do not have established weather stations or their ground station networks are inadequate. Following these developments, monitoring environmental disasters using satellite and radar data became more widespread [Duan et al., 2014; Ezzine et al., 2014; Tian et al., 2014].

For the computation of a drought index, one or multiple indicators are required. The term indicators is used to denote the various variables and parameters, including precipitation, temperature, humidity, streamflow, evapotranspiration, groundwater levels, soil moisture and snowpack. As mentioned, indices are computed values that represent the drought severity, and they attempt to measure the qualitative state of droughts on the landscape for a distinct time period. Their main contribution is that they simplify a very complex event, helping a diverse audience understand and communicate the appropriate actions that need to be in place to resolve an upcoming adverse event.

The time in combination with the severity of a drought event's occurrence is crucial since it can be more catastrophic to have a low-severity event in a sensitive period (e.g. moisture-sensitive period of a stable crop) of the agricultural cycle than a more severe event in a less sensitive time. Hence, additional information on the exposed area and its vulnerable characteristics must be considered for the interpretation of an index result.

In this case study, we use the standardized precipitation index (SPI) and the standardized precipitation evapotranspiration index (SPEI) to monitor and extract conclusions for Crete —a drought-prone area— to assist the water resources management teams with drought management policies and preparedness plans. For the estimation of the SPEI we also calculate the Potential Evapotranspiration (PET).

Since the length of the data set is quite long (41 years), the presentation of all the results (SPI and SPEI for 1-, 3-, 6-, 9-, 12-, and 24-months) for all the locations (65 locations) is impossible. Therefore, we propose specific years that represent the range of the data, as well as specific locations that are of higher interest due to their geographic location. Based on the hourly, and the daily precipitation data, the estimated characteristic year is 2006, while for the monthly precipitation data, the estimated characteristic year is 1980. Since we are working on monthly data, we will consider 1980 as the characteristic year for precipitation on the island of Crete, and we will focus our results on that year. Additionally, we will compare the results for the year 1980 with the results for the years that had the lowest (1990) and the highest (2019) total annual precipitation.

To our knowledge, this is the first time where ERA5 data are used for the estimation of drought indices for the island of Crete. Furthermore, the estimation of SPEI for the area of interest has not been published before. Past research for the island of Crete has focused on a single drought index (usually SPI or a custom index) and a single time scale (3-, 12-, 24-, or 48-months). In addition, we take advantage of index calculations for multiple timescales (1-, 3-, 6-, 9-, 12-, and 24-months), which allows us to capture in more detail the area's drought characteristics regarding short and long-term patterns.

4.3 Drought Indices

Monitoring drought for early warning and risk assessment, utilizing drought indices, can be performed by estimating one or multiple indices or using/creating a composite or hybrid index. The most common approach until recently was to assess one or various indices. However, with the technological advancements and the need for more targeted results based on the specific area, scientists developed new indices and improved the visualization tools [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016].

With an abundance of indices available, determining which one is the most appropriate for the area of interest becomes almost impossible. Given that drought severity is best assessed based on multiple variables associated with water availability for a region, the composite or hybrid approach gained popularity in the last decade because it enables a collection of elements to be integrated into the estimation.

Determining which is the most suitable indicator, index, or combination of the above depends on a variety of reasons. It should be decided upon investigation and thorough analyses and after establishing the needs for a particular application, i.e., climate regimes, regions, basins and locations.

The most commonly used indices in Greece are the Standardized Precipitation Index, the Palmer Drought Severity Index, the Reclamation Drought Index and the Palfai Drought Index, while in Cyprus are the Standardized Precipitation Index and the Bhalme–Mooley Drought Intensity Index. Similar climatic conditions to Greece are observed in Spain, where the predominant indices are the Standardized Precipitation Index and the soil water content (available water calculated as percentage of soil water capacity from a soil water balance model). Likewise, in Turkey the Standardized Precipitation Index, the Percent of Normal Index and Palmer Drought Severity Index are the most widely used indices [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016].

The Palmer Drought Severity Index [Palmer, 1965], the self-calibrated Palmer Drought Severity Index [Wells et al., 2004], the Palmer Hydrological Drought Index, the Palmer moisture anomaly index (Z-Index), and the Palmer Modified Drought Index [Palmer, 1965] while available from the Python package used herein, will not be estimated due to lack of the available water content variable.

Below we will present the fundamental properties of the indices that we used in this study for the detection of drought in the study area. Those include: the Potential Evapotranspiration (PET), the Standardized Precipitation Index (SPI) and the Standardized Precipitation Evapotranspiration Index (SPEI). We opted out on estimating the Percentage of Normal Precipitation (PNP) because it is region-specific, making the comparison between regions or even seasons hard. Another reason for putting the index under scrutiny is that the precipitation data are not transformed, meaning that the mean and the median values can differ significantly, underestimating or overestimating the results and consequently its

Table 4.1: Indices used in this case study. P: precipitation and PET: potential evapotranspiration. Green: easy, Yellow: medium, Red: difficult, Part of the table taken from World Meteorological Organization (WMO) and Global Water Partnership (GWP) [2016].

Meteorology/Hydrology	Ease of use	Input parameters	Additional information
Standardized Precipitation Index (SPI)	Green	P	Highlighted by the World Meteorological Organization as a starting point for meteorological drought monitoring
Standardized Precipitation Evapotranspiration Index (SPEI)	Yellow	P, PET	Serially complete data required; output similar to SPI but with a temperature component

accuracy [Hayes, 2006; Zargar et al., 2011].

4.3.1 Standardized Precipitation Index

The Standardized Precipitation Index (SPI) [McKee et al., 1993] is probably the most popular meteorological drought index, mainly because it is easy to interpret and needs only precipitation data for its calculation. Also, it is effective in analyzing wet as well as dry periods [World Meteorological Organization (WMO), 2012]. For comparison purposes in drought severity between countries and regions, WMO recommended SPI in 2009 as the main meteorological drought index that countries should use [Hayes et al., 2011; World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016].

To calculate the SPI index, the user needs to have ideally at least 20–30 years of monthly precipitation values, while 50–60 years are considered optimal [Guttman, 1999]. The data set is allowed to have missing values, although the way they are distributed in the entire data set can dramatically affect the results. The ideal scenario would be to have a data set without missing data; yet, this scenario is unrealistic in our world, where most data sets have 90% or even 85% complete records. In some cases, where the records have even lower completion percentages, the user might need to apply first interpolation techniques to fill the missing values and afterward estimate the index. The confidence of the estimation index depends on the estimations' confidence and estimation error, meaning that the fewer the estimated data, the better the result [World Meteorological Organization (WMO), 2012].

The SPI was created to quantify the precipitation deficit over a variety of timescales. These timescales account for the impact of drought on the availability of various water resources. For example, short scale precipitation anomalies have effect on soil moisture, whereas longer scale anomalies reflect on groundwater, stream-flow and reservoir storage water availability. For these reasons, the SPI was initially estimated by McKee et al. [1993] for durations of 3, 6, 12, 24, and 48 months.

Nowadays, the index can be calculated at any number of timescales, from 1 month to 48 months or longer, typically it is applied for the 3, 6, 12, 24, and 48 month periods. The ability of SPI to be calculated at various timescales allows for a wide range of applications, depending on the drought impact. For instance, SPI values for 3 months or less might be useful for basic drought monitoring, values for 6 months or less for monitoring agricultural impacts and values for 12 months or longer for hydrological impacts [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016].

The initial step towards estimating the SPI index is the transformation of the input parameter to normally distributed. This is accomplished using an equal-probability transformation after fitting precipitation to a gamma or a Pearson Type III distribution. The mean value is set to zero so that the positive values correspond to wet periods, while the negative values indicate the dry periods. Specifically, the value of the index denotes how many standard deviations the cumulative precipitation differs from the normalized average [Zargar et al., 2011]. Positive SPI values indicate precipitation greater than the median precipitation, and negative values indicate precipitation less than the median precipitation [Tsakiris et al., 2007b]. Essentially, the SPI shows the actual precipitation compared to the probability of precipitation for different timescales.

For example, the estimation of the 3-month SPI for a specific 3-month period compares the precipitation over that specific period to the precipitation from the same 3-month period for all the years included in the historical record [Tsakiris and Vangelis, 2004]. In simple words, the SPI-3 at the end of November compares the September-October-November precipitation total in that specific year with the September-October-November precipitation totals of all the years on the historical record for that location [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016]. The SPI-3 at the end of May indicates the soil moisture conditions as the growing season begins. Because the 3-month SPI reflects the short term precipitation shortage, normal values of the index could be misleading in a region where it is normally dry during that 3-month period, and might mask drought events that will be obvious in longer timescales. A 6-month SPI can be highly useful for displaying seasonal precipitation and detecting medium-term trends in precipitation.

An event is considered as drought event when the index reaches values of −1 or less (see Table 4.2); the drought is considered in progress until the index value returns to zero [McKee et al., 1993]. Because SPI values are generated from the Gaussian distribution, it is natural that the probabilities are drawn based on that. This means that (for the selected time scale at a specific location) the probability of an event to be considered extremely wet (SPI ≥ 2) is 2.3%. The rest of the probabilities are shown in Table 4.2. The percentages based on this standardization indicate the rarity of the event. This means that we expect 1 extremely dry event per 50 years (2.3 per 100 years). Each drought event, has a "duration" defined by its beginning and end (in months), and an "intensity" for each month that the event continues. "Intensity" is the mean value of the SPI during the drought event. The absolute sum of the SPI for all the months within a drought event is called the drought's "magnitude" (DM) (a.k.a drought severity) [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016]. DM is defined as:

$$DM = -\left(\sum_{i=1}^{x} \mathrm{SPI}_{ij}\right)$$

where j starts with the first month of a drought event and continues to increase until the end of the drought (x) for any of the i time scales. The DM has units of months and would be numerically equivalent to drought duration if each month of the drought has SPI = -1.0. In fact, many droughts will have a DM very similar to the duration in months since most of the SPI values are between 0 and

-2.0 [McKee et al., 1993].

Table 4.2: Classification of the SPI and SPEI values.

Value	Characterization	Probability (%)
2.0+	Extremely wet	2.3
1.5 to 1.99	Very wet	4.4
1.0 to 1.49	Moderately wet	9.2
−0.99 to 0.99	Near normal	68.2
−1.49 to −1.0	Moderately dry	9.2
−1.99 to −1.5	Severely dry	4.4
< -2	Extremely dry	2.3

SPI does not come without shortcomings. By using precipitation records solely for its calculation, it disregards the impact of temperature (and of other parameters), which is an essential indicator of the overall water balance and water use, in the interpretation of the drought severity. This weakness can make the comparison between regimes with similar SPI values but with temperature differences more challenging. The flexibility of the index will help many people apply it. At the same time, without guaranteeing that the data's prior distribution agrees with the distribution that the index assumes, the results will be useless. Zargar et al. [2011] present an extensive review of various drought indices, including SPI, and identify their advantages and disadvantages.

The SPI has been investigated in numerous research papers. Utilizing the SPI for describing drought in the region of Crete, Tsakiris and Vangelis [2004] came to the conclusion that, the eastern side of the island experiences droughts more frequently. They also concluded that in the years 1973–74, 1976–77, 1985–86, and 1999–2000 distinct drought occurrences happened, whereas the years 1987–94 saw a relatively lasting drought event. Another study that was focused on the island of Crete was carried out by Koutroulis et al. [2011], where they introduced a variation of the SPI index in which the spatial patterns of precipitation are taken into account, resulting in an index that can be compared not only temporarily but also spatially. Their results (based on the SPI and their variant) showed that the southern and eastern part of the island suffers from drought occurrences.

Furthermore, according to the future scenarios they estimated that for the 48-month timescale, more than half of Crete will experience drought conditions during 28% of the 2010-2040 period.

4.3.2 Standardized Precipitation Evapotranspiration Index

Standardized Precipitation Evapotranspiration Index (SPEI) [Vicente-Serrano et al., 2010] is an extension of the SPI that also uses temperature and evapotranspiration data. It is sensitive to long-term temperature trends and becomes almost equivalent to SPI if no apparent trends are present. Indices that are solely dependent on precipitation, such as the SPI rely on the assumption that the variability of precipitation is higher than it is for other variables (e.g., temperature, PET), while these other variables are temporary stationary. This means that the driving parameter for droughts is precipitation. However, climate change and the evident rise in temperature (resulting in an increasing rate in evapotranspiration) has been shown by several studies that is not to be neglected in the estimation of drought severity [Vicente-Serrano et al., 2010]. The catastrophic 2003 central European heat wave made clear the significant impact of temperature on the severity of the drought. The extremely high temperatures substantially increased evapotranspiration and amplified summer drought stress [Martine et al., 2006]. Similar effects were observed throughout the summer of 2010, when a hot spell heightened forest drought stress and triggered extensive forest fires in eastern Europe and Russia [Barriopedro et al., 2011]. More recently, Rakovec et al. [2022] observed that the intensity of the 2018–2020 multi-year drought was record-breaking compared to the past 250 years. The temperature anomaly had great implications in the crops' yield. They estimated that the future events based on climate model simulations will have similar intensities but longer durations. Hence, empirical research has shown that rising temperatures amplify the effects of drought stress, and since the temperature increase during the last century has obvious side effects, it is safe to expect that these rising temperatures will have profound impacts on the drought conditions, with a reduced water availability as a result of evapotranspiration [Justin and Eric F., 2008].

The calculation of the SPEI is equivalent to the one used for the SPI. However, the SPEI takes the difference between precipitation and potential evapotranspiration (P - PET) as input rather than only precipitation (P). Essentially, the water balance compares the available water (precipitation) to the atmospheric evaporative demand (PET), providing a more reliable estimate of drought severity than explicitly considering precipitation. The water balance values are fitted to a probability distribution (in our case the gamma or the Pearson Type III distribution) and then the values are converted to standardized values that are comparable in space and time and at different time scales [Beguería et al., 2014]. To fit the water balance data different probability distributions have been utilized, those include the log-logistic [Vicente-Serrano et al., 2010], the generalized logistic, the generalized extreme value, the Gaussian, and the Pearson Type III distributions [Stagge et al., 2015]. The Thornthwaite equation was proposed in the original SPEI formulation for estimating PET [Thornthwaite, 1948]. Due to limited data availability, this equation only requires the mean daily temperature and latitude of the site. The FAO-56 Penman-Monteith equation [Allen et al., 1998] is considered more robust and is recommended if data is available (relative humidity, temperature, wind speed and solar radiation). If the required data are not available, the Hargreaves equation [Hargreaves and Samani, 1982] (first) or the Thornthwaite equation (second) are suggested.

Similar to SPI, SPEI takes positive and negative values identifying wet and dry events, respectively. It can be calculated for different time scales, starting from 1 month, creating an index that can identify slower developing droughts, but may miss recognizing the fast-developing. As stated previously, integrating temperature data into the computation of the SPEI, helps overcome the SPI's weakness to account for the impact of temperature changes. The requirement of a complete and extended monthly data set for both precipitation and temperature data might be a considerable drawback [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016].

From multiple studies across the Mediterranean area, based on the SPEI and the Palmer Drought Severity Index (PDSI), the probability and intensity of agricultural and ecological droughts has been increased. Increase of drought severity in South Europe was observed by Dai and Zhao [2017]; Spinoni et al. [2019];

Stagge et al. [2017], and in the Iberian Peninsula by González-Hidalgo et al. [2018]; Vicente-Serrano et al. [2014][IPCC, 2021, Table 11.18].

4.3.3 Potential Evapotranspiration

Potential evapotranspiration (PET) represents the summation of the water that transpires through the plants and the water that evaporates through the soil for a given area and during a constrained time period if a sufficient water source was available [Kirkham, 2014; Milly and Dunne, 2016]. It only occurs at the potential rate when the water available for this process is non-limiting. The rate of evaporation is influenced by climatic conditions, including the sun's solar radiation, wind, the air's vapor deficit, and temperature. Potential evaporation is often calculated from these measurements using the Penman Monteith equation [Penman, 1948]. It can also be estimated from readily available rainfall and temperature data using simple equations such as that of Thornthwaite [1948], and Hargreaves and Samani [1982]. This indicator describes the capacity of the prevailing climate to evaporate water from the soil, plants, open water, or other surfaces [Imeson, 2023]. Actual evapotranspiration is always less than or equal to PET because it is constrained by the amount of water available. In cases where the monthly precipitation is higher than the monthly PET, the residual water recharges the ground and runs off as steamflow. On the contrary, if the PET is higher than the precipitation for the month, the soil loses water. In dry (arid) climates annual potential evaporation exceeds annual precipitation.

The calculation of the SPEI relies on precipitation and PET values. In our study the calculation of the PET values from temperature data is performed based on the Thornthwaite equation. The potential evapotranspiration according to the Thornthwaite (1948) equation (mm/day) is calculated as follows

$$\mathrm{PET} = 16\left(\frac{L}{12}\right)\left(\frac{N}{30}\right)\left(\frac{10\,T}{I}\right)^{\alpha}, \tag{4.1}$$

where T is the mean daily air temperature (°C), N is the number of days in the month being calculated, L is the mean day length (hours) of the month being calculated, and I is a heat index which depends on the 12 monthly mean temperatures and is calculated as

$$I = \sum_{\text{Jan}}^{\text{Dec}} \left(\frac{\max\left(0, T_m\right)}{5} \right)^{1.514}, \tag{4.2}$$

where T_m is the mean air temperature for each month in the year (°C). The exponent α is calculated by the following regression formula

$$\alpha = 6.75 \times 10^{-7}\, I^3 - 7.71 \times 10^{-5}\, I^2 + 1.792 \times 10^{-2}\, I + 0.49239. \tag{4.3}$$

4.4 Case Study

The data used in this case study consists of the monthly precipitation and temperature for the period between 01.01.1979 to 31.12.2019 (41 years) from the ERA5 data set (dd.MM.yyyy). The data values correspond to the grid locations shown in Fig. 3.1, and they are complete without missing values. We will focus our analysis for this chapter in four main nodes, the two biggest cities of Crete, Heraklion and Chania, and the two main argicultural areas, the Messara plane and Ierapetra valley. The investigation nodes are those closer to those four locations. In Table 4.3 we present the selected locations, their coordinates and their mean precipitation, temperature and estimated PET. Furthermore, the variation of the monthly distribution of precipitation against the monthly variation of the estimated PET is shown in Fig. 4.1.

For the following analysis we use an open source Python package [Adams, 2019] downloaded from `https://github.com/monocongo/climate_indices`. It contains a variety of climate index algorithms that can offer information about the severity of temperature and precipitation anomalies in space and time, which can prove helpful for climate monitoring. The available indices are: SPI, SPEI, PET, PDSI, scPDSI, PHDI, Z-Index, PMDI, and PNP. Additional packages that are used include: numpy, pandas, scipy, matplotlib. The data set has to be formatted into a NetCDF file to be further processed.

Network Common Data Format (NetCDF) is a way to format a data set in a hierarchical structure that lets the users to access parts of the data set without loading the entire data set into the machine's memory. Thus, it is well suited to manage big numerical data sets. It is a commonly used format in the geoscience community. The NetCDF4 module is available from the Python Package Index (PyPi).

Table 4.3: Selected locations and their characteristics. The coordinates are in the World Geodetic System (WGS 84).

Name	Latitude (°)	Longitude (°)	Mean Precipitation (mm)	Mean Temperature (°C)	Mean Estimated PET (mm)
Chania	35.5	24.00	865	17.3	843
Heraklion	35.25	25.00	546	16.7	837
Messara	35.00	25.00	355	18.4	927
Ierapetra	35.00	25.75	299	18.1	884

As mentioned previously, when the mean PET for a location is higher than the mean precipitation for the same location, water has to be evaporated from the soil reserves. From Table 4.3 and Fig. 4.1 it is obvious that the climate of Crete has big variations from the west to the east as well as from the north to the south. In the north-eastern part (Chania) the mean monthly precipitation is almost equal to the mean monthly PET, while in the northern central part of the island (Heraklion) the precipitation is significantly less. Finally, in the south central Crete (Messara) and south west Crete (Ierapetra) the monthly precipitation is almost a third of the monthly PET, classifying the climate as dry.

4.5 Spatial and Temporal Analysis

In brief, the standardized values are calculated based on the following steps. A set of monthly precipitation values is prepared from the 41 years monthly data. For each timescale of i months (i.e., 1, 3, 6, 9, 12, 24 months) a set with averaging values is generated. The set's value for each month is created by averaging the previous i months. For every timescale the data set is fitted with the Pearson Type III probability distribution function. Then the probability for any observation is calculated and used to return the corresponding values from the standard normal distribution (Gaussian distribution with zero mean and one

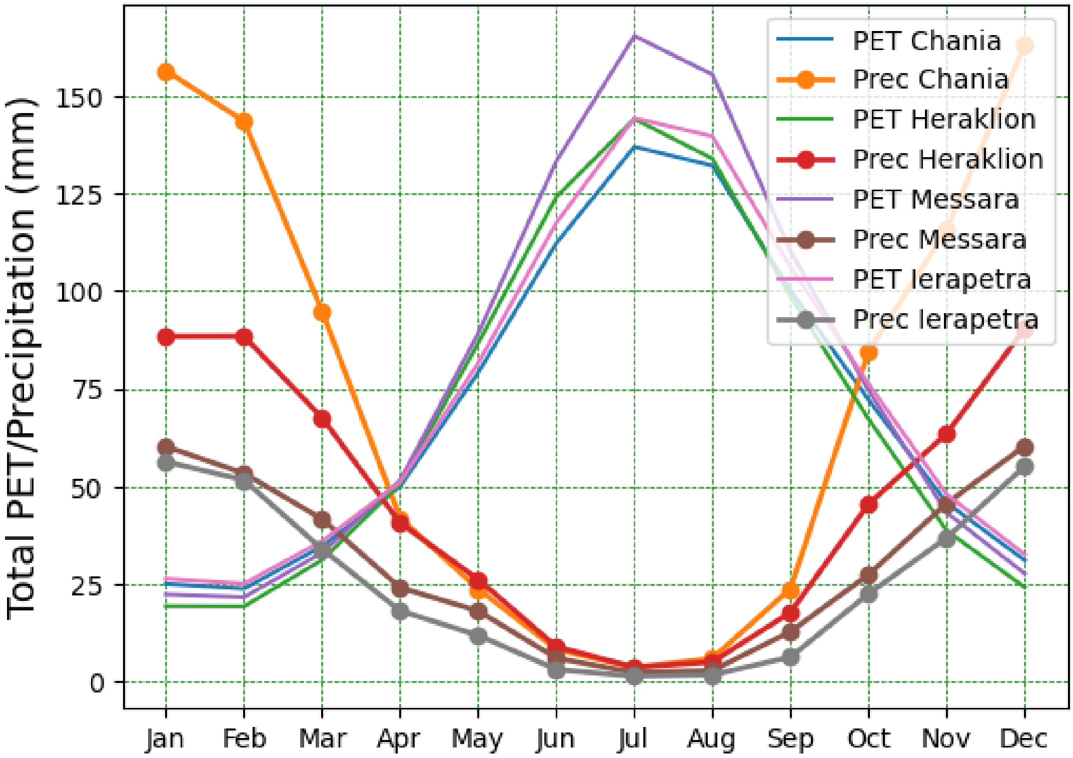


Figure 4.1: Mean precipitation versus mean PET calculated for the four locations presented in Table 4.3, historical data: 1979–2019.

standard deviation). The resulted value is the SPI value for the corresponding precipitation value. The same methodology is applied for the estimation of the SPEI values, however the initial set that needs to be fitted with the Pearson's Type III probability distribution function consists of the difference between the precipitation and the potential evapotranspiration.

An illustrative example for Crete is presented in Fig. 4.2 which shows the evolution of the SPI and the SPEI at different time scales from 1979 to 2019 for Chania located in the island of Crete (Greece). The most obvious characteristics of the drought events is that drought changes as the time scale changes. At longer time scales drought becomes less frequent but longer in duration (Fig. 4.2).

The SPI-1 is very closely related to the percentage of normal precipitation (PNP) for a period of 1-month, but because of the normalization process lying in the SPI calculation, the SPI-1 is a more accurate representation of monthly precipitation [World Meteorological Organization (WMO) and Global Water Partnership (GWP), 2016]. However, SPI-1 (and the PNP) may be ambiguous in regions where precipitation is usually low during a month. The same can be said for the SPI-3 for a Mediterranean climate because it is common to have dry conditions during a 3-month period (e.g., June–August). This may lead to high or low index values where precipitation does not deviate far from the mean. Because Crete has very low precipitation (many times equal to zero) during the dry season (April to September), the use of SPI-1 and the PNP is avoided. We present the SPI-1 in the temporal analysis figures (Figs. 4.2–4.5) but we do not focus on that timescale.

4.5.1 Drought Event Identification

Monthly values of SPI and SPEI were estimated for the four locations to provide an overview of drought events. Figures 4.2–4.5 present the estimated SPI and SPEI values for the analyzed timescales. Any specific evaluation of drought and its impacts requires a specification of time scale since drought initiation, intensity, duration, and magnitude (severity) are all dependent on time scale. The ordinate of the SPI/SPEI graphs has graduation lines at −2.5, −2, −1, 0, 1, 2, 2.5. Each of these values of SPI has a unique value of the probability that the SPI will be equal to or less than the stated value. These probability values are 0.006, 0.02, 0.16, 0.50, 0.84, 0.98, and 0.99 respectively.

Based on Figs. 4.2–4.5 we identify the following events:

- In 1982 a short dry event occurred for the locations in the north, however, in the south (Messara, Ierapetra) for that period the standardized values are near normal.

- In 1986–1987 we see a drought event for all four locations, with the lowest intensity estimated in Chania.

- In 1990–1992 a drought event occurred in Heraklion, followed by a near normal event for 2 years and another drought event in 1994–1996. Similar pattern is observed in Chania, while in Messara the same applies for the

SPEI values, however for the SPI-24 the drought was persistent and continued during 1992–1994. During this period we see the first big differences between the indices. Interestingly, in Ierapetra, which is the driest location of the four, the drought event had shorter duration and lasted from 1990–1992. We see differences between the indices during 1990–1994, especially for the 24-month timescale.

- Dry events occurred during 2000–2003 in Messara and Heraklion, while in Chania the duration of the event was one year shorter (2000–2002), and in Ierapetra half year longer.
- Between the years 2003 and 2016 in Heraklion and in Messara, no significant drought events are observed, however, moderately dry events were observed in Ierapetra during 2006–2011, and in Chania during 2009–2011. The most profound differences between the indices are present during the period 2013–2015 for all four locations.
- Lastly, dry events were identified during 2016–2019 in Heraklion with near normal values only between 2017–2018 for some of the timescales. Similar events were observed in Messara plane and in Ierapetra valley with higher intensities, and in Chania with near normal or even a wet spell for the 3-month and 6-month scales in 2017. For this period the events are almost entirely classified as severely dry (-1.99 to -1.5) and in some cases as extremely dry (≤ -2).

The differences between the indices and the events during the period 1990-1996 might indicate that the temperature started deviating from the historical normal values. The year that we estimated as the wettest (2019) can also be seen in Figs. 4.2–4.5, where the indices values start moving fast from negative values to positive values that are classified as severely and extremely wet events. Also, the frequency of the drought events has increased noticeably for all the locations after 2000.

Compared to the study by Tsakiris et al. [2007a] we observe drought occurrences during the period 2000–2003 while they resulted in wet events during the same period. However, it should be noted that they used meteorological data from ground stations that might be closer to the real values compared to the reanalysis products used herein or more unreliable due to the interpolation techniques used to fill in the missing values. Additionally, their analyzed period ends after the year 2005. Using data from a different period (especially if the climate change impacts have not yet been substantial) might affect the fitted parameters, and thus the standardized values. Nevertheless, during the years 1987–1994 where they observed a long drought event our results coincide and verify that during that period there were multiple drought occurrences based on multiple timescales from both SPI and SPEI results.

The frequency of droughts for the period 1979–2019 for the four areas is shown in Figs. 4.6–4.7. For the SPI, the events below minus one (≤ −1) add up to 13.7% for Chania, 14.8% for Heraklion, 16.5% for Messara, and 15% for Ierapetra. According to those values, Messara has experienced the most droughts during the investigated time period. Based on the SPEI results, the sum of the moderately, severely, and extremely dry events is 13.6% for Chania, 15.4% for Heraklion, 16.1% for Messara, and 14.2% for Ierapetra. The indices are very consistent for Chania, and Messara, however a considerable decrease (from 15% to 14.2%) in the amount of droughts is estimated based on the SPEI for Ierapetra and a considerable increase (from 14.8% to 15.4%) on the drought occurrences for the Heraklion area. Temperature has affected substantially the drought characterization for the eastern part of the island.

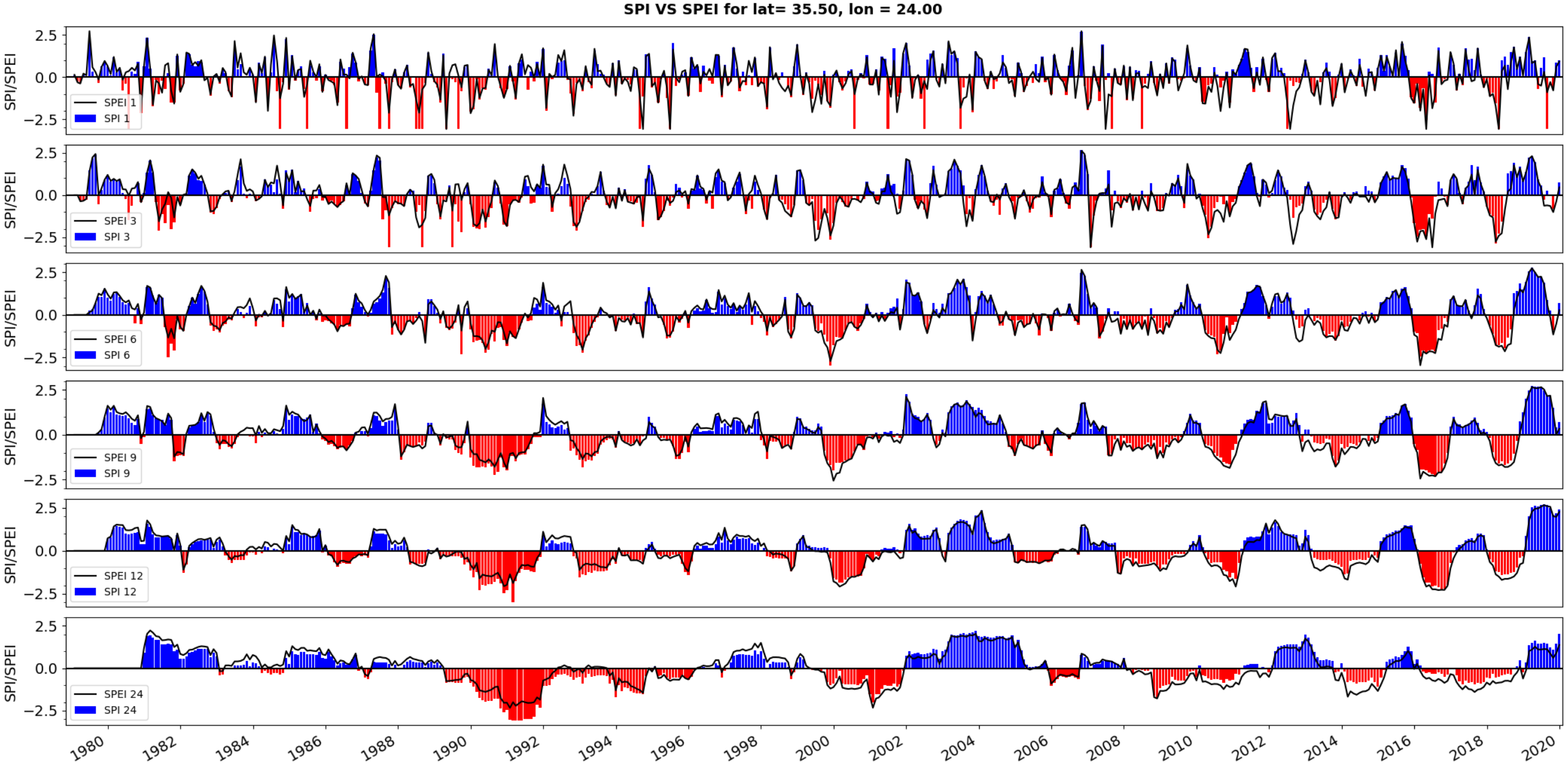


Figure 4.2: SPI vs SPEI timeseries calculated for Chania (the four locations are presented in Table 4.3), the historical data include values for the period 1979–2019 and the indices are calculated for time scales 1, 3, 6, 9, 12, and 24 months with the Pearson type III distribution.

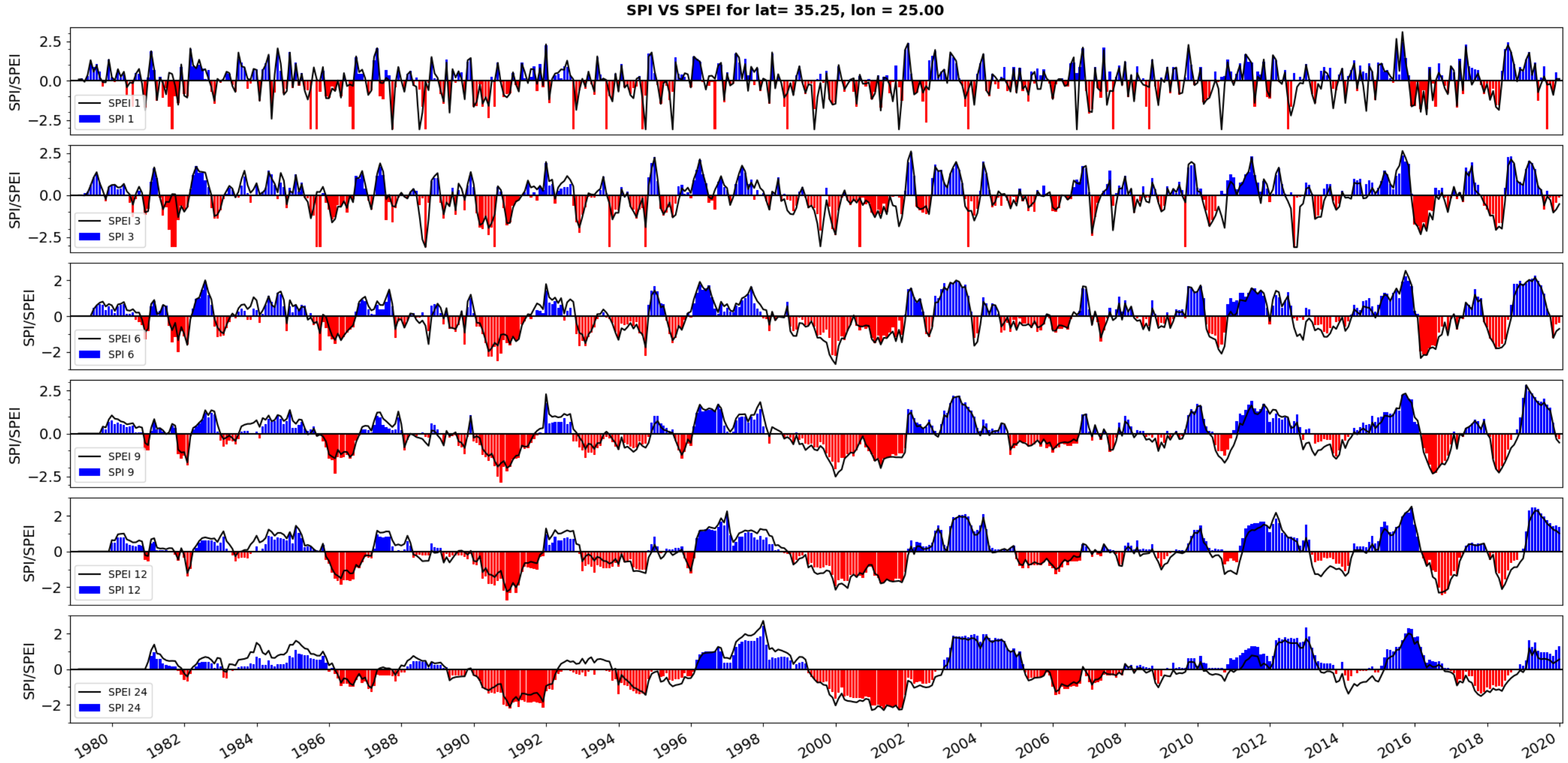


Figure 4.3: SPI vs SPEI timeseries calculated for Heraklion (the four locations are presented in Table 4.3), the historical data include values for the period 1979–2019 and the indices are calculated for time scales 1, 3, 6, 9, 12, and 24 months with the Pearson type III distribution.

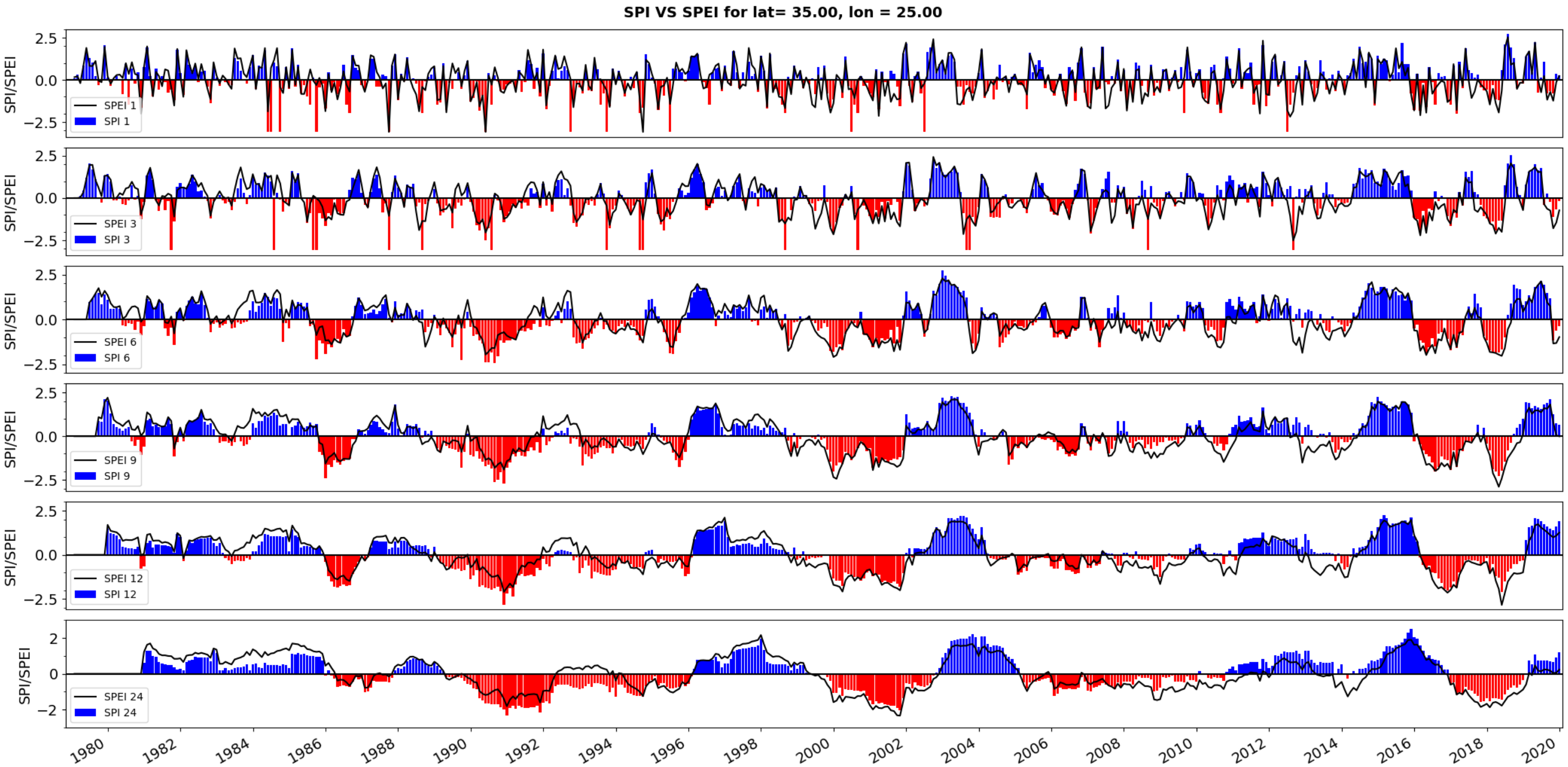


Figure 4.4: SPI vs SPEI timeseries calculated for Messara (the four locations are presented in Table 4.3), the historical data include values for the period 1979–2019 and the indices are calculated for time scales 1, 3, 6, 9, 12, and 24 months with the Pearson type III distribution.

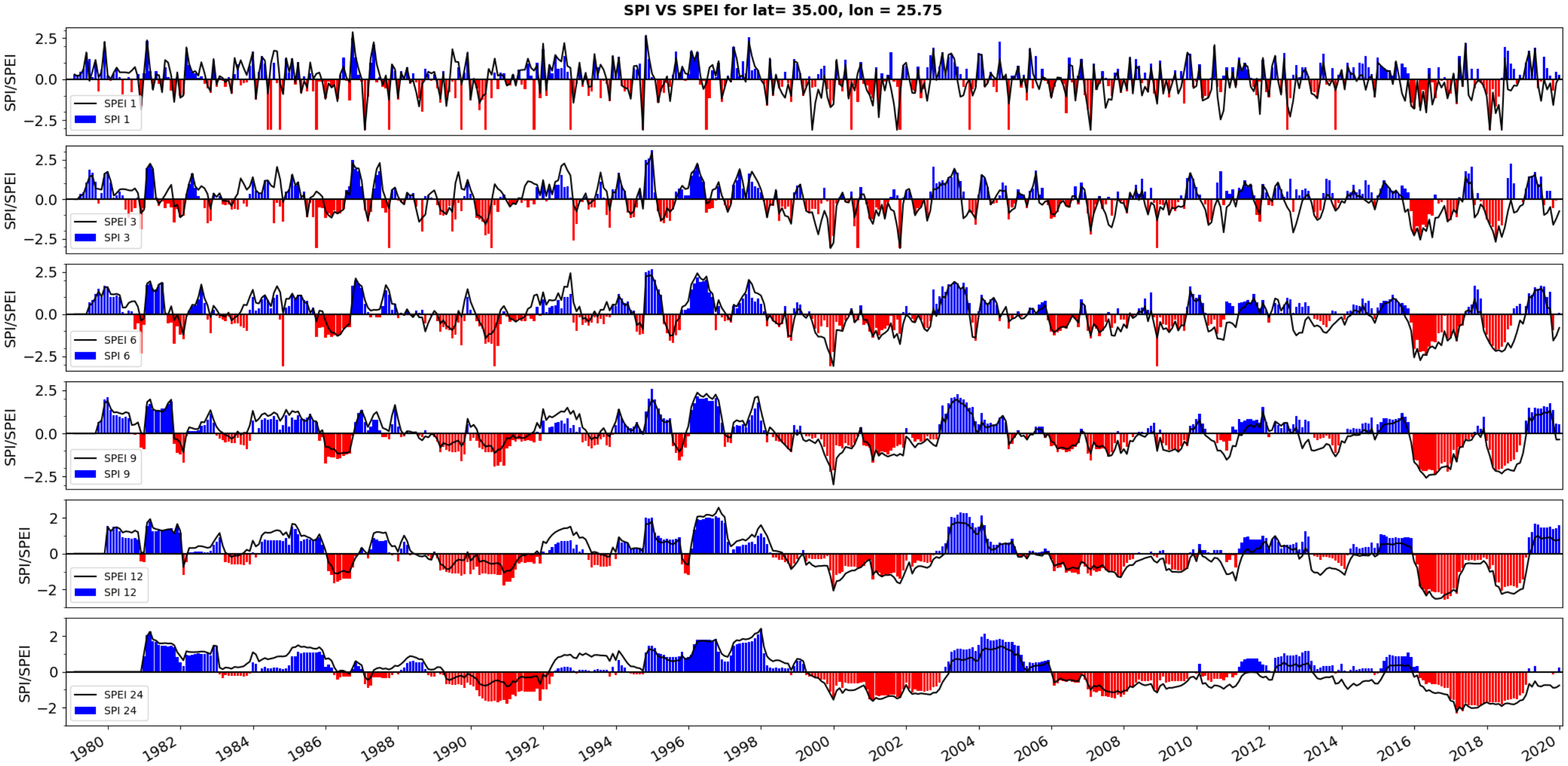
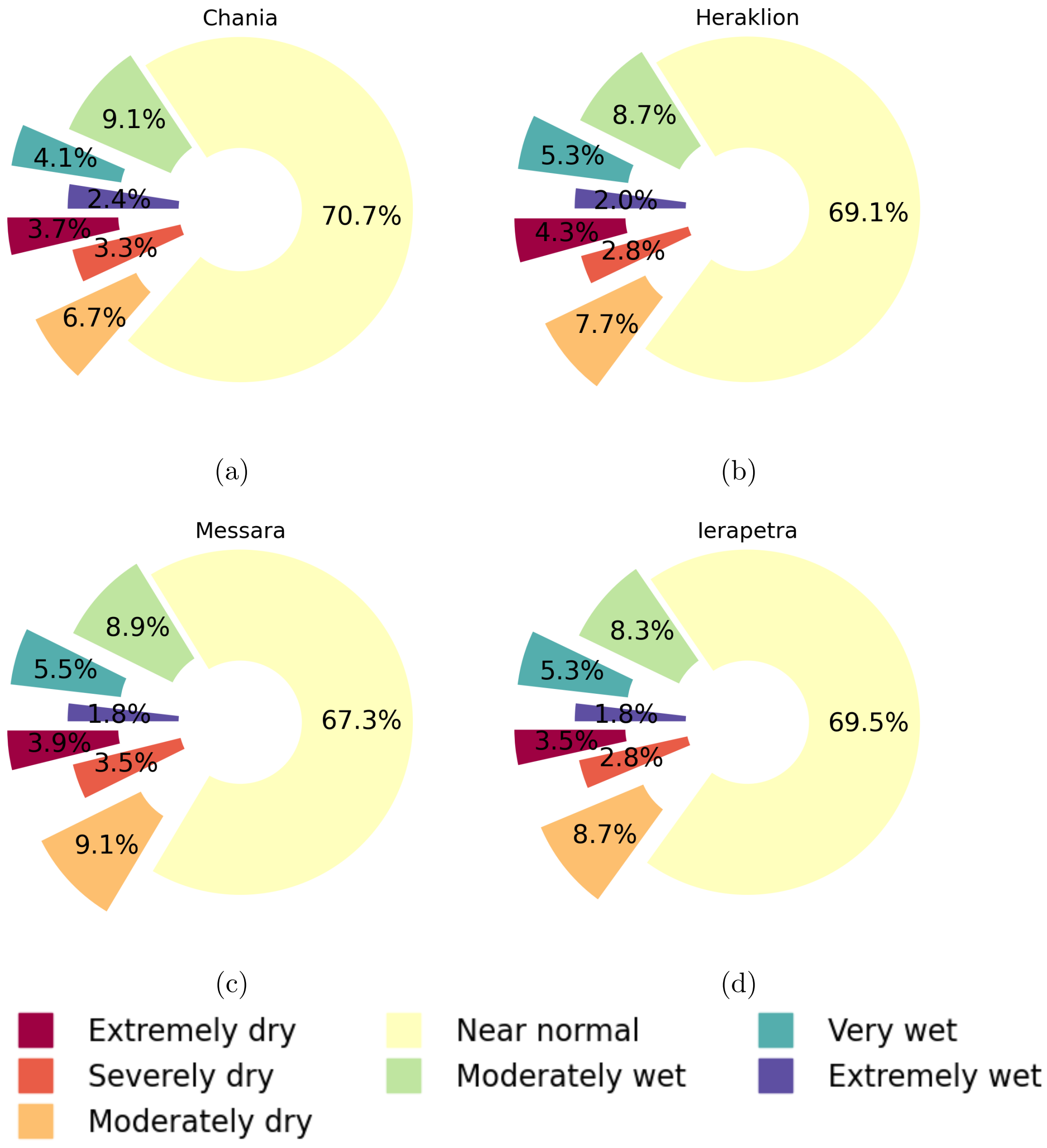


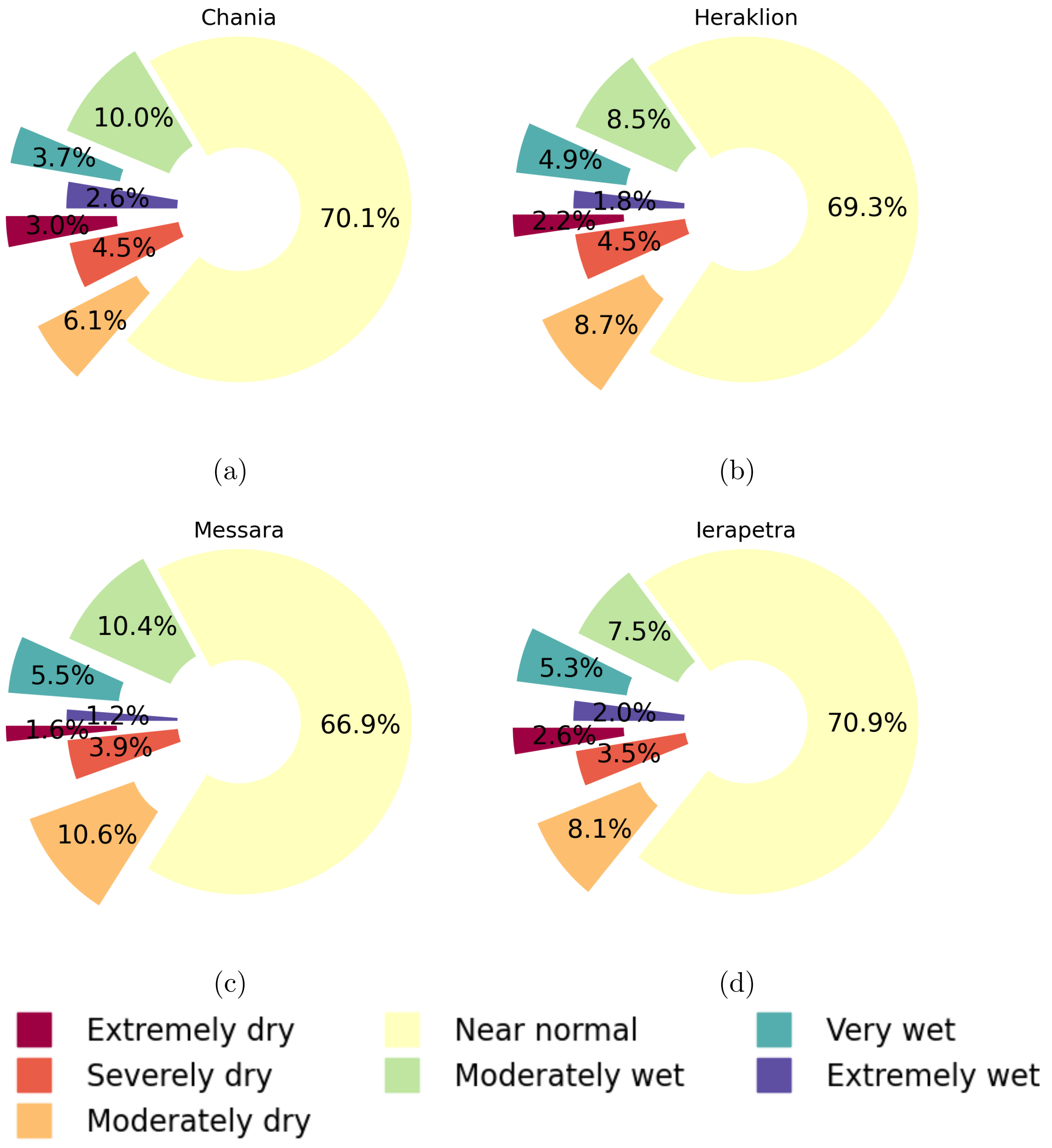
Figure 4.5: SPI vs SPEI timeseries calculated for Ierapetra (the four locations are presented in Table 4.3), the historical data include values for the period 1979–2019 and the indices are calculated for time scales 1, 3, 6, 9, 12, and 24 months with the Pearson type III distribution.



(a) (b) (c) (d)

(e)

Figure 4.6: Frequency of droughts based on the SPI-3 with the Pearson type III distribution for the four locations presented in Table 4.3.



(a)

(b)

(c)

(d)

(e)

Figure 4.7: Frequency of droughts based on the SPEI-3 with the Pearson type III distribution for the four locations presented in Table 4.3.

4.5.2 Drought Maps

A more understandable way to present the indices results compared to site specific data is by generating maps. The location data are placed on a map and interpolation techniques are used to fill the space in between them. Common interpolation methods include the kriging method (described in section 2.9, the spline method where the overall surface curvature is minimized, and the Inverse Distance Weighting especially in dense scattered data among others. In this case study we use the piecewise cubic interpolant to create smoothly interpolation maps [1]. For each point, the estimation is performed based on a cubic interpolation of the values at neighboring grid points.

Drought maps were generated using the estimated 3-month and 12-month values of SPI and SPEI. In Figs. 4.8–4.13 and Figs. 4.14–4.15 typical samples of these maps are presented. Specifically, for the 3-month timescale we focus on the year 1980 which is considered the characteristic year, the year 2019 which was the wettest year on record, and the year 1990 which was the driest year on record based on the monthly precipitation. Depending on the month of the year and the local conditions, the results are either equivalent or quite diverse. For the 12-month timescale we present results for the hydrological years 2008 to 2019 (hydrological year is a twelve month period starting from October and ending in September for the Northern Hemisphere).

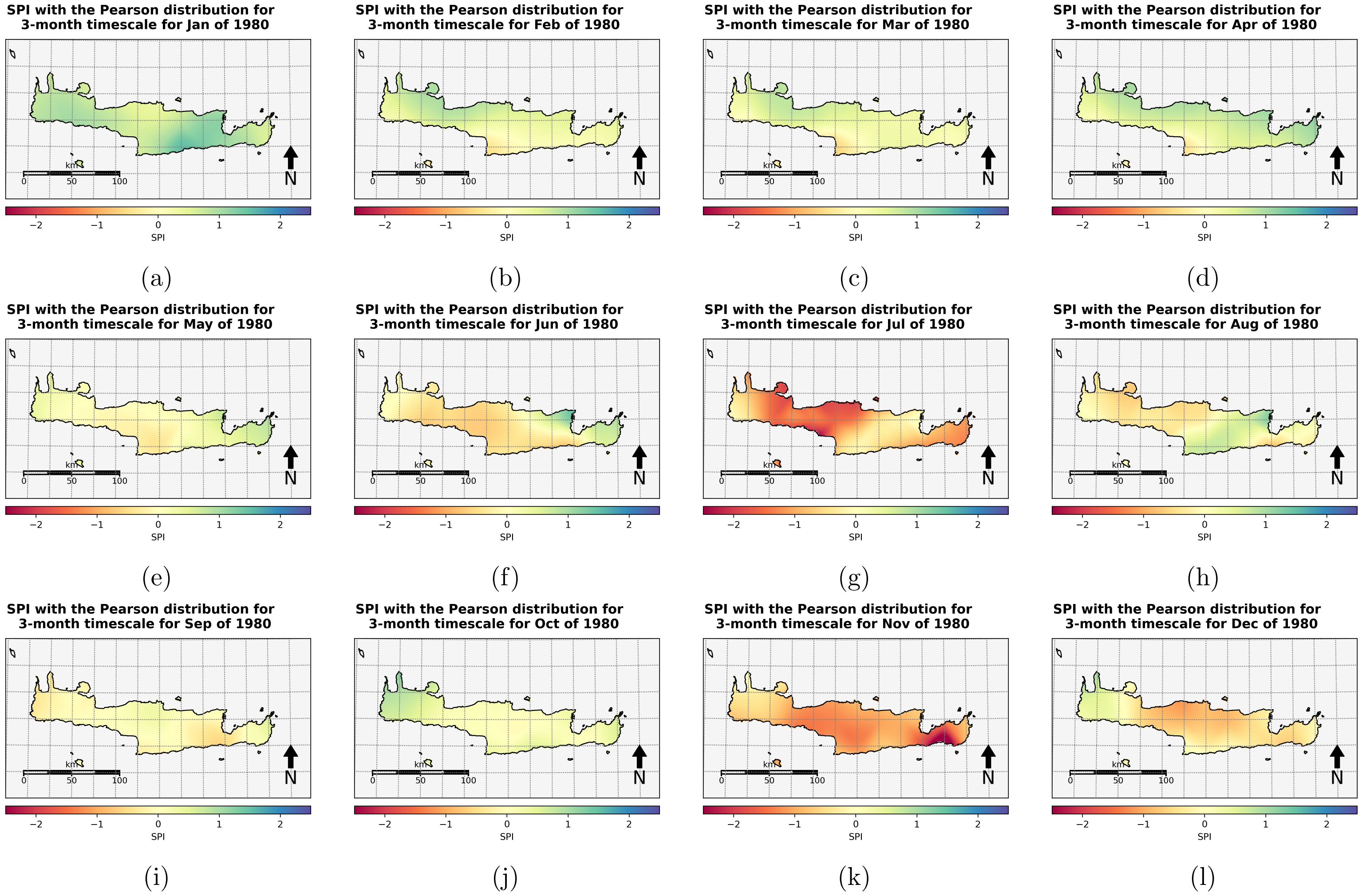
We observe that while comparing the SPI-3 to the SPEI-3 results for the characteristic year based on the precipitation data (1980 see Figs. 4.8, 4.9) the integration of temperature affects the results especially for the months of July (SPI-3 means May-July) and November (SPI-3 means September-November). The SPI-3 values for those months are considerably lower than for the SPEI, meaning that the temperature values were not high enough to affect the equilibrium. However, when we approach the more recent years (2019) where higher temperature values

[1] The piecewise cubic interpolation involves the triangulation of the input data for the generation of the convex hull by the Quickhull Algorithm [Barber et al., 1996], and the estimation of the polynomial coefficients based on the Bezier polynomials on each triangle of the convex [Alfeld, 1984; Farin, 1986]. The coefficients of the interpolant are selected so that the curvature of the interpolating surface is minimized. The estimation is performed using the global algorithm described in Nielson [1983] and Renka et al. [1984]. For more information see scipy.interpolate.CloughTocher2DInterpolator.

are present due to climate change, the differences between the indices are very pronounced, with SPI values being lower than the SPEI (see Figs. 4.10, 4.11), even though 2019 is the wettest year in our record. The differences between the indices are particularly prominent for the second half of the year. According to the SPEI, the increased evapotranspiration contributes to drier events compared to those where the temperature is not accounted for. In the case of the driest year (1990 see Figs. 4.12, 4.13), we observe the opposite behavior. Because temperatures were not that different from the previous century's mean temperatures, the dry conditions are more exaggerated by the SPI. Additionally, because the indices are calculated based on the historical data per location, for the first 3 months of the driest year the lower index values are higher in the eastern and northern part of the island which is also the part with the highest precipitation.

Because droughts usually take more than a season to develop, we present SPI-12 and SPEI-12 maps through September for twelve continuous years (2008–2019). Longer timescales such as the 12-month, normally approximate zero values because they consist of aggregated shorter periods, thus, values that deviate from zero suggest ongoing distinct dry or wet spells. Negative values for the 12-month timescale indicate that dryness has a considerable effect on water resources. This timescale reflects long-term precipitation patterns and is usually tied to streamflows, and reservoir levels, and might indicate an imbalance in the groundwater levels. The 12-month SPI is most highly correlated to the Palmer Index in certain locations, and the two indices can portray similar conditions.

In Figs. 4.14–4.15 the comparison between the indices indicate that the hydrological years 2010, 2013, 2016 and 2018 have significantly higher drought intensities according to the SPEI calculations versus the SPI results for the entire island. Those results are also visible in the timeseries plots (Figs. 4.2–4.5) especially for the locations of Heraklion, Messara and Ierapetra.



SPI with the Pearson distribution for 3-month timescale for Jan of 1980

(a)

SPI with the Pearson distribution for 3-month timescale for Feb of 1980

(b)

SPI with the Pearson distribution for 3-month timescale for Mar of 1980

(c)

SPI with the Pearson distribution for 3-month timescale for Apr of 1980

(d)

SPI with the Pearson distribution for 3-month timescale for May of 1980

(e)

SPI with the Pearson distribution for 3-month timescale for Jun of 1980

(f)

SPI with the Pearson distribution for 3-month timescale for Jul of 1980

(g)

SPI with the Pearson distribution for 3-month timescale for Aug of 1980

(h)

SPI with the Pearson distribution for 3-month timescale for Sep of 1980

(i)

SPI with the Pearson distribution for 3-month timescale for Oct of 1980

(j)

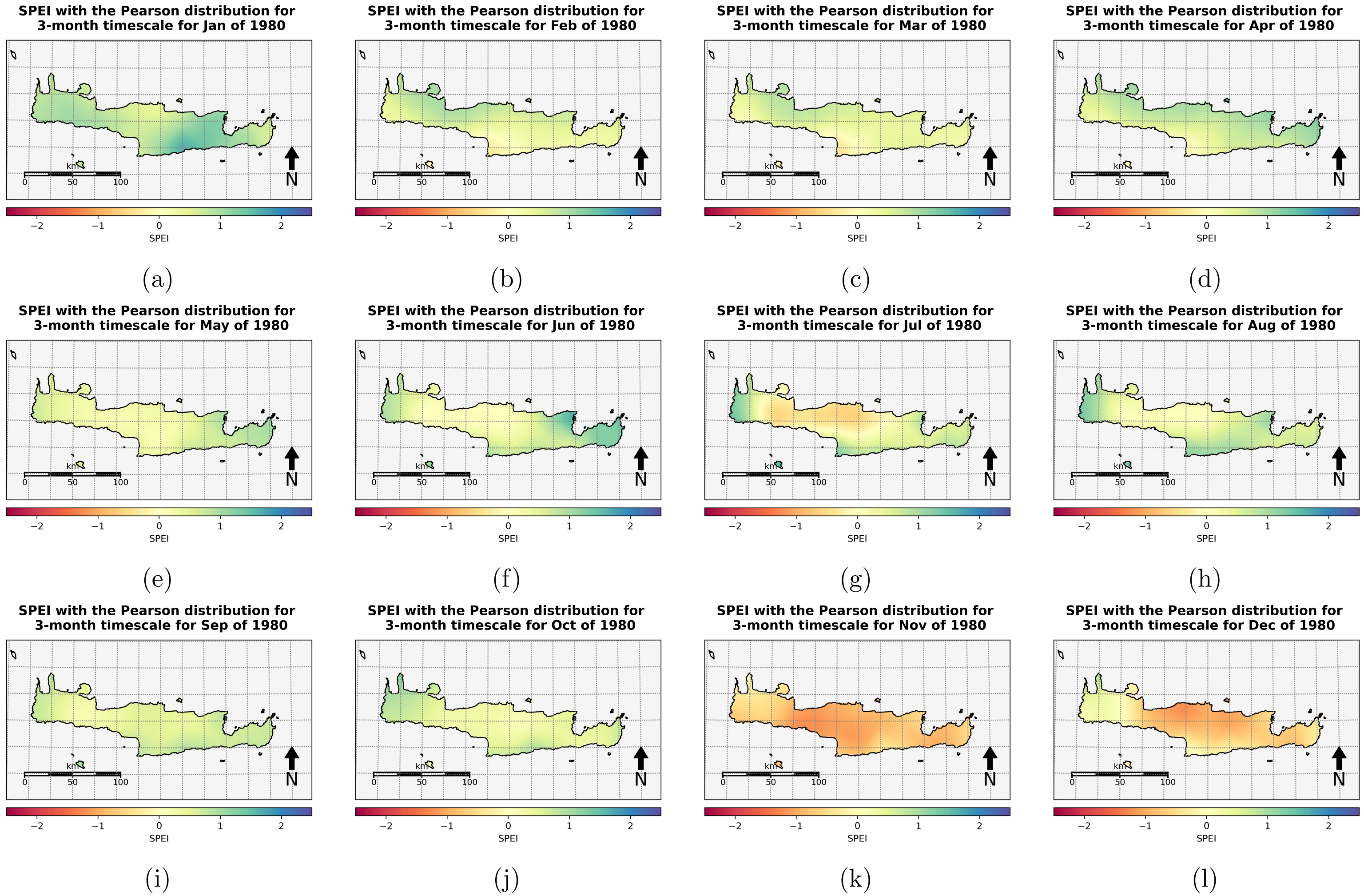
SPI with the Pearson distribution for 3-month timescale for Nov of 1980

(k)

SPI with the Pearson distribution for 3-month timescale for Dec of 1980

(l)

Figure 4.8: SPI spatial map for 1980, historical data: 1979–2019 using 3-month time scale with the Pearson type III distribution.



SPEI with the Pearson distribution for 3-month timescale for Jan of 1980

(a)

SPEI with the Pearson distribution for 3-month timescale for Feb of 1980

(b)

SPEI with the Pearson distribution for 3-month timescale for Mar of 1980

(c)

SPEI with the Pearson distribution for 3-month timescale for Apr of 1980

(d)

SPEI with the Pearson distribution for 3-month timescale for May of 1980

(e)

SPEI with the Pearson distribution for 3-month timescale for Jun of 1980

(f)

SPEI with the Pearson distribution for 3-month timescale for Jul of 1980

(g)

SPEI with the Pearson distribution for 3-month timescale for Aug of 1980

(h)

SPEI with the Pearson distribution for 3-month timescale for Sep of 1980

(i)

SPEI with the Pearson distribution for 3-month timescale for Oct of 1980

(j)

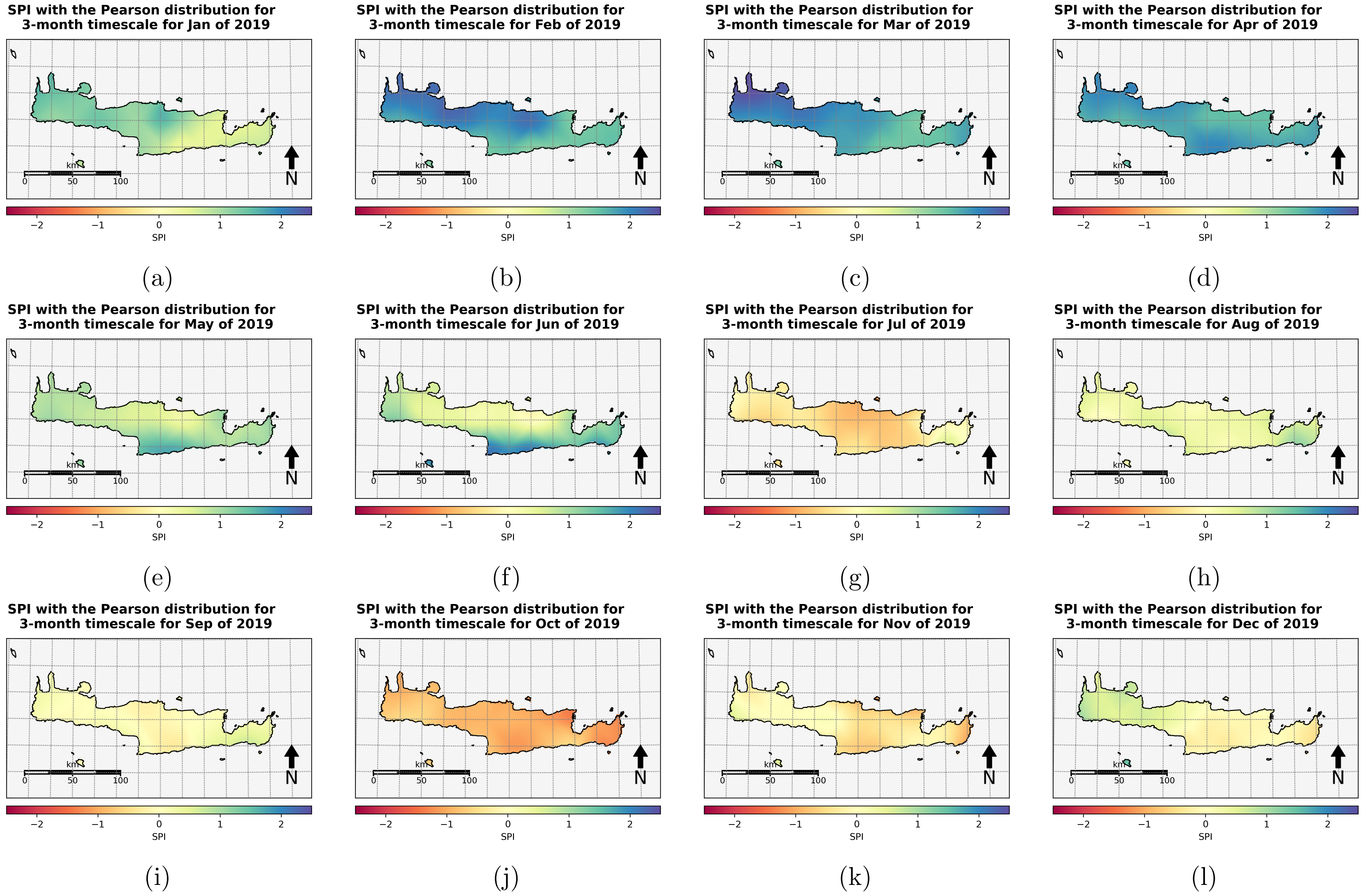
SPEI with the Pearson distribution for 3-month timescale for Nov of 1980

(k)

SPEI with the Pearson distribution for 3-month timescale for Dec of 1980

(l)

Figure 4.9: SPEI spatial map for 1980, historical data: 1979–2019 using 3-month time scale with the Pearson type III distribution.



SPI with the Pearson distribution for 3-month timescale for Jan of 2019

(a)

SPI with the Pearson distribution for 3-month timescale for Feb of 2019

(b)

SPI with the Pearson distribution for 3-month timescale for Mar of 2019

(c)

SPI with the Pearson distribution for 3-month timescale for Apr of 2019

(d)

SPI with the Pearson distribution for 3-month timescale for May of 2019

(e)

SPI with the Pearson distribution for 3-month timescale for Jun of 2019

(f)

SPI with the Pearson distribution for 3-month timescale for Jul of 2019

(g)

SPI with the Pearson distribution for 3-month timescale for Aug of 2019

(h)

SPI with the Pearson distribution for 3-month timescale for Sep of 2019

(i)

SPI with the Pearson distribution for 3-month timescale for Oct of 2019

(j)

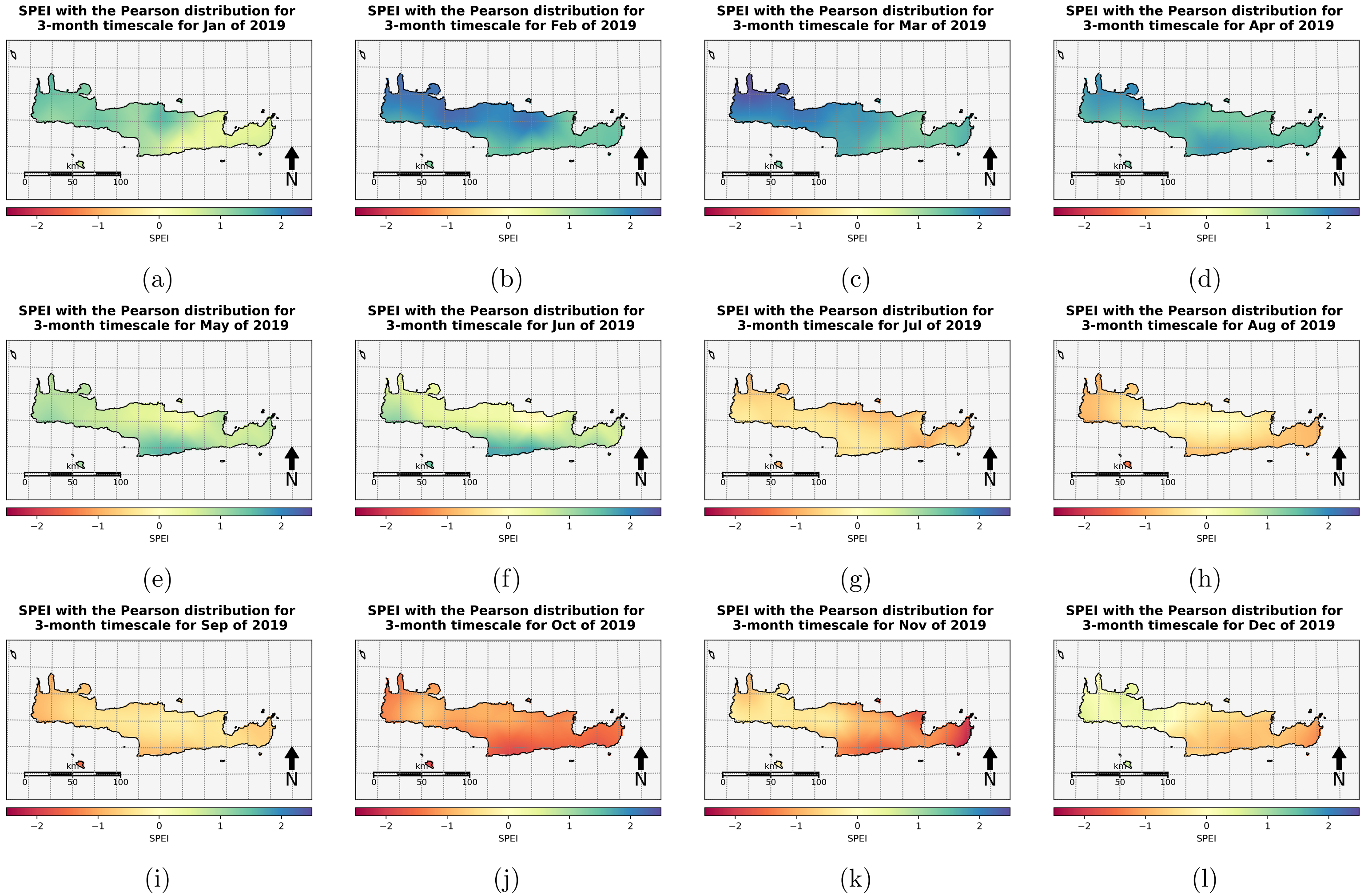
SPI with the Pearson distribution for 3-month timescale for Nov of 2019

(k)

SPI with the Pearson distribution for 3-month timescale for Dec of 2019

(l)

Figure 4.10: SPI spatial map for 2019, historical data: 1979–2019 using 3-month time scale with the Pearson type III distribution.



SPEI with the Pearson distribution for 3-month timescale for Jan of 2019

(a)

SPEI with the Pearson distribution for 3-month timescale for Feb of 2019

(b)

SPEI with the Pearson distribution for 3-month timescale for Mar of 2019

(c)

SPEI with the Pearson distribution for 3-month timescale for Apr of 2019

(d)

SPEI with the Pearson distribution for 3-month timescale for May of 2019

(e)

SPEI with the Pearson distribution for 3-month timescale for Jun of 2019

(f)

SPEI with the Pearson distribution for 3-month timescale for Jul of 2019

(g)

SPEI with the Pearson distribution for 3-month timescale for Aug of 2019

(h)

SPEI with the Pearson distribution for 3-month timescale for Sep of 2019

(i)

SPEI with the Pearson distribution for 3-month timescale for Oct of 2019

(j)

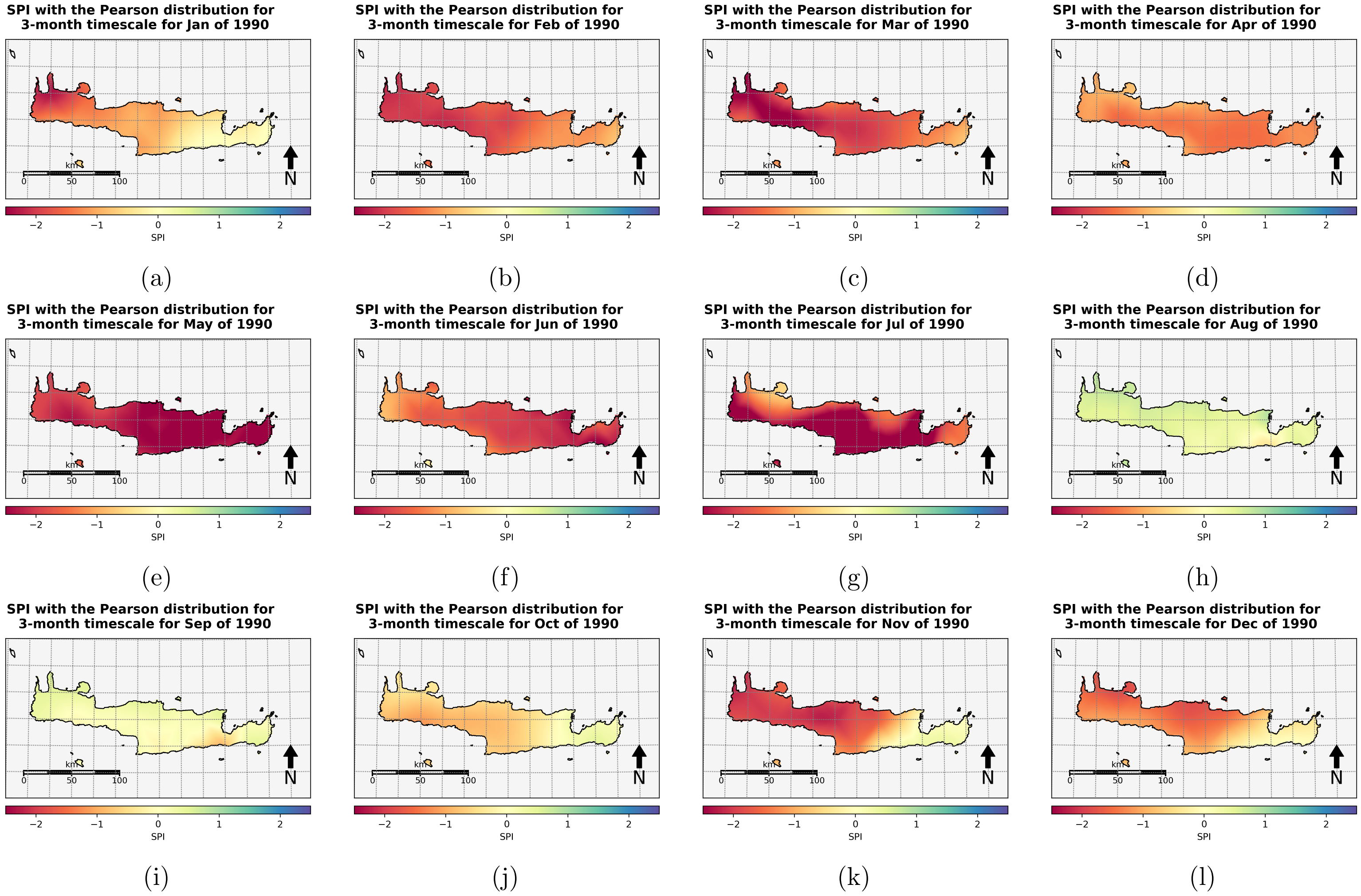
SPEI with the Pearson distribution for 3-month timescale for Nov of 2019

(k)

SPEI with the Pearson distribution for 3-month timescale for Dec of 2019

(l)

Figure 4.11: SPEI spatial map for 2019, historical data: 1979–2019 using 3-month time scale with the Pearson type III distribution.

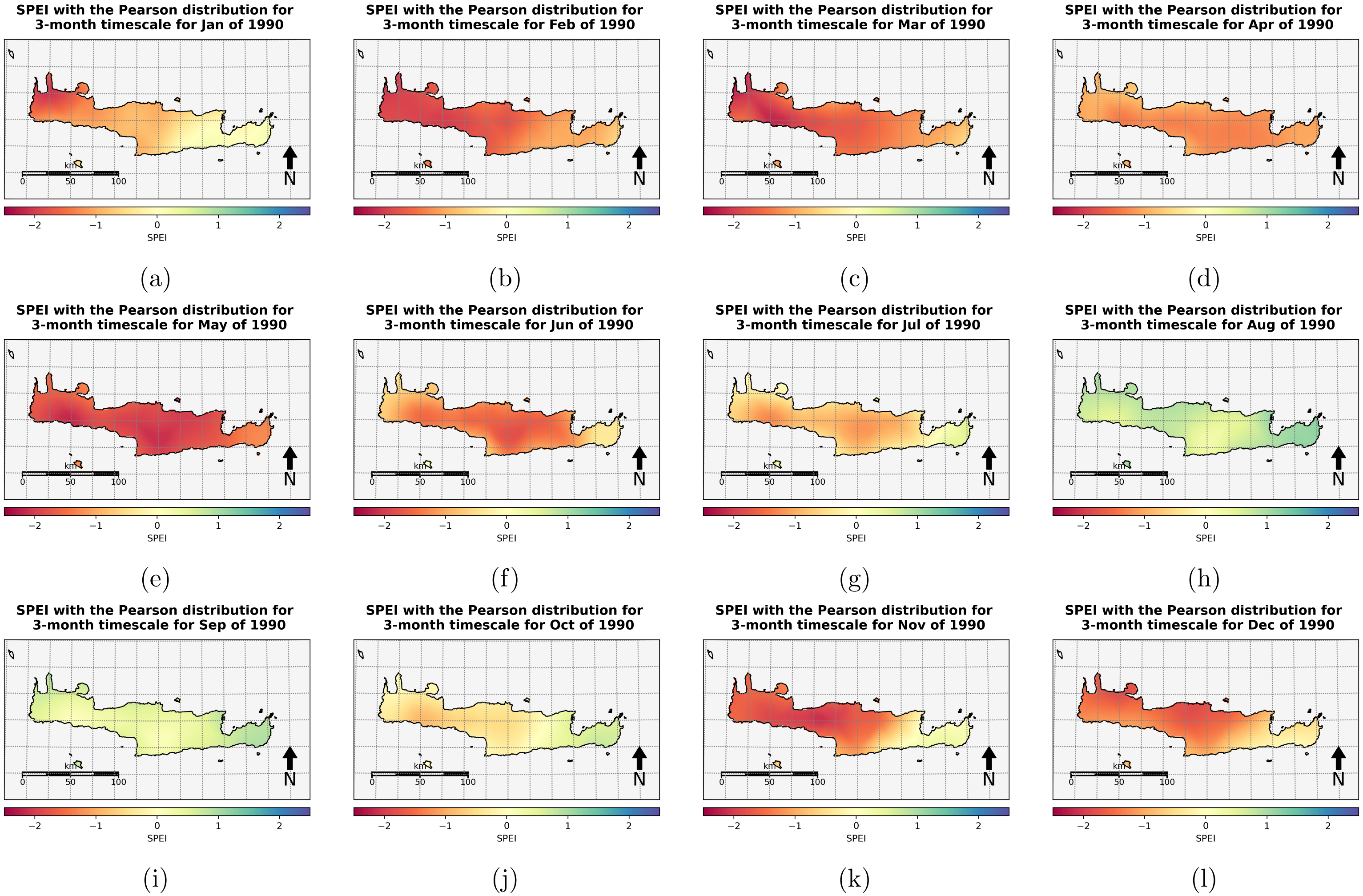

SPI with the Pearson distribution for 3-month timescale for Jan of 1990
SPI with the Pearson distribution for 3-month timescale for Feb of 1990
SPI with the Pearson distribution for 3-month timescale for Mar of 1990
SPI with the Pearson distribution for 3-month timescale for Apr of 1990
SPI with the Pearson distribution for 3-month timescale for May of 1990
SPI with the Pearson distribution for 3-month timescale for Jun of 1990
SPI with the Pearson distribution for 3-month timescale for Jul of 1990
SPI with the Pearson distribution for 3-month timescale for Aug of 1990
SPI with the Pearson distribution for 3-month timescale for Sep of 1990
SPI with the Pearson distribution for 3-month timescale for Oct of 1990
SPI with the Pearson distribution for 3-month timescale for Nov of 1990
SPI with the Pearson distribution for 3-month timescale for Dec of 1990
km
0
50
100
N
−2
−1
0
1
2
SPI

(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4.12: SPI spatial map for 1990, historical data: 1979–2019 using 3-month time scale with the Pearson type III distribution.


SPEI with the Pearson distribution for 3-month timescale for Jan of 1990

SPEI with the Pearson distribution for 3-month timescale for Feb of 1990

SPEI with the Pearson distribution for 3-month timescale for Mar of 1990

SPEI with the Pearson distribution for 3-month timescale for Apr of 1990

(a) (b) (c) (d)

SPEI with the Pearson distribution for 3-month timescale for May of 1990

SPEI with the Pearson distribution for 3-month timescale for Jun of 1990

SPEI with the Pearson distribution for 3-month timescale for Jul of 1990

SPEI with the Pearson distribution for 3-month timescale for Aug of 1990

(e) (f) (g) (h)

SPEI with the Pearson distribution for 3-month timescale for Sep of 1990

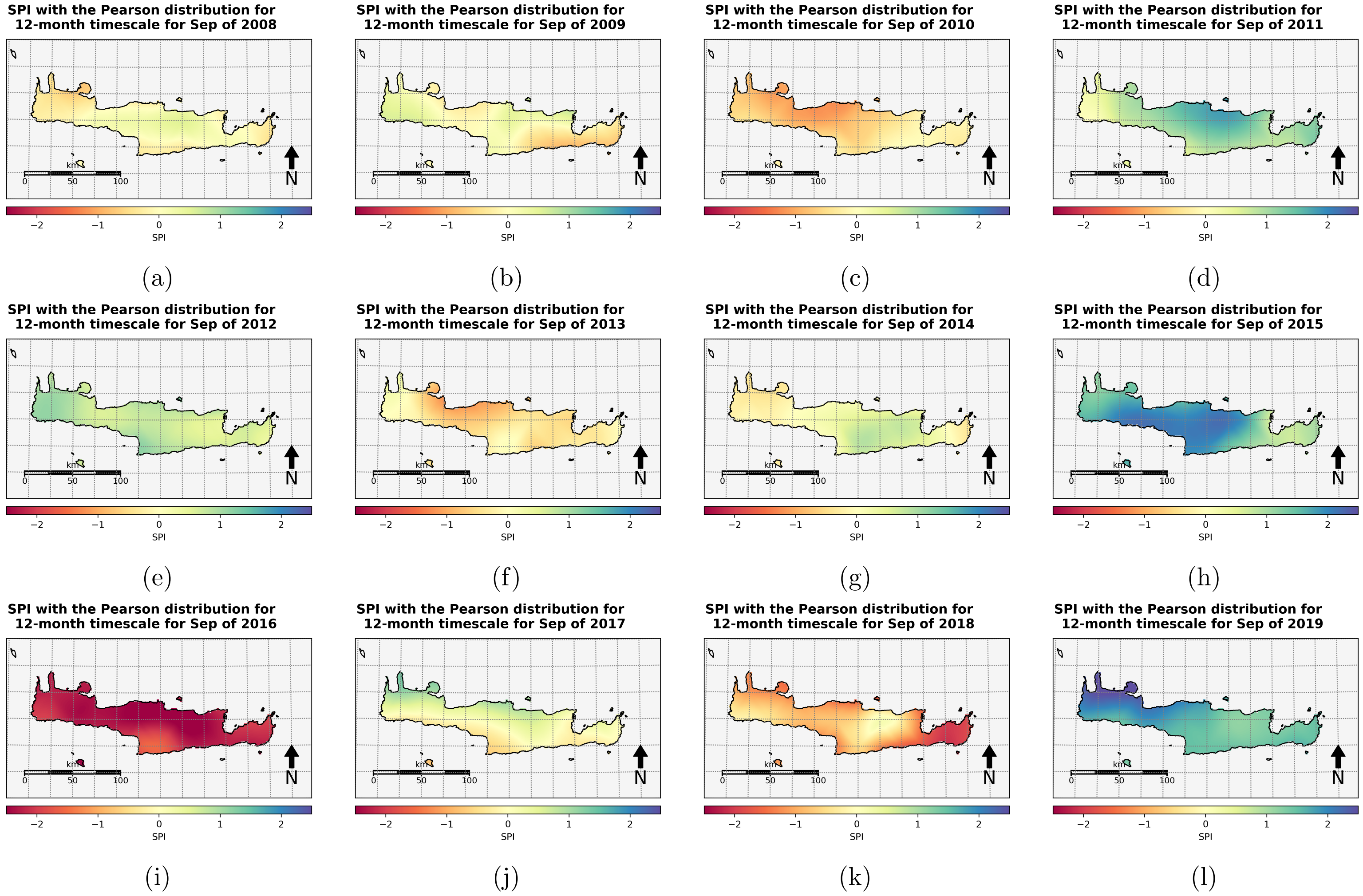
SPEI with the Pearson distribution for 3-month timescale for Oct of 1990

SPEI with the Pearson distribution for 3-month timescale for Nov of 1990

SPEI with the Pearson distribution for 3-month timescale for Dec of 1990

(i) (j) (k) (l)

Figure 4.13: SPEI spatial map for 1990, historical data: 1979–2019 using 3-month time scale with the Pearson type III distribution.

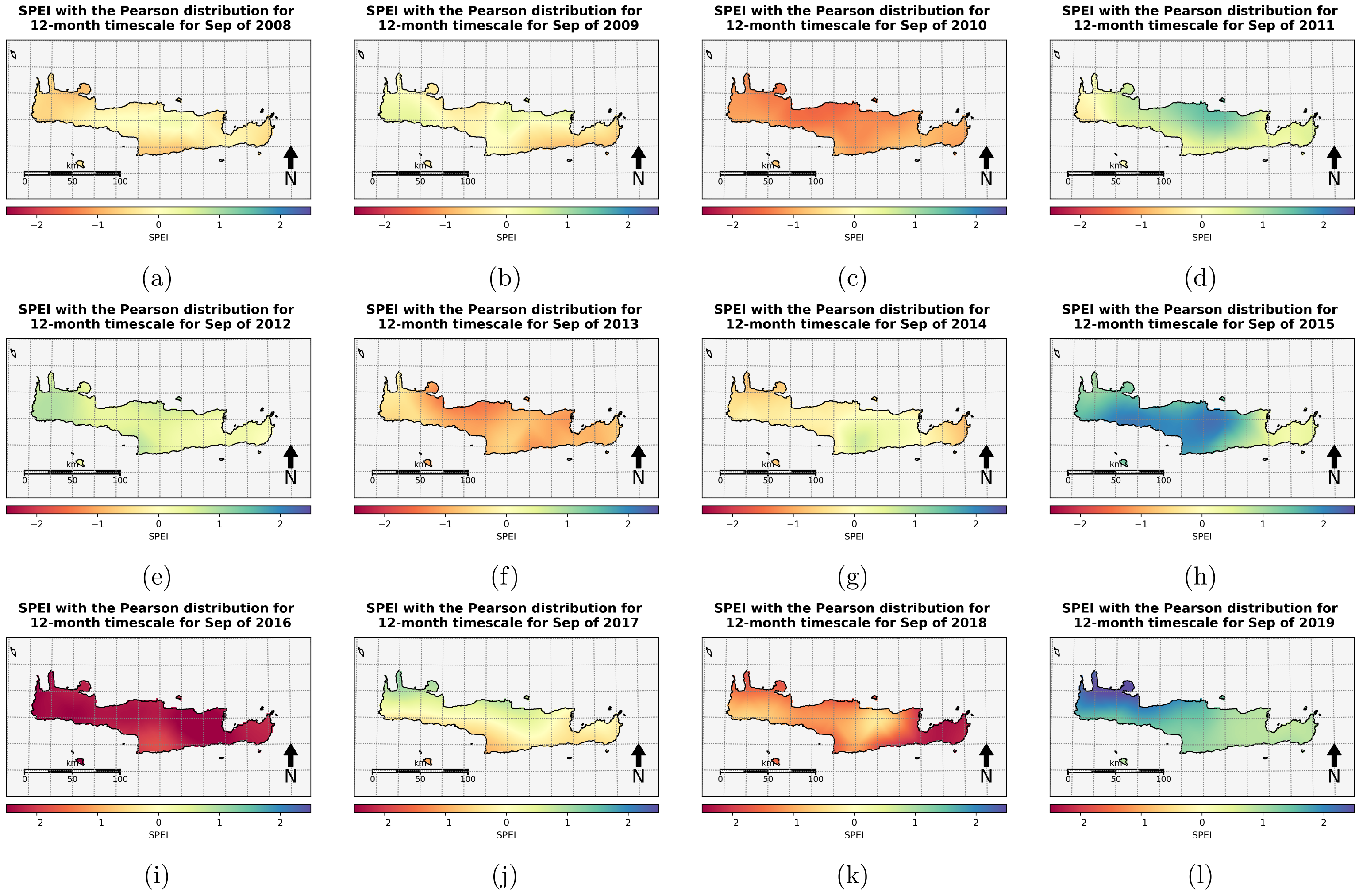

SPI with the Pearson distribution for 12-month timescale for Sep of 2008
SPI with the Pearson distribution for 12-month timescale for Sep of 2009
SPI with the Pearson distribution for 12-month timescale for Sep of 2010
SPI with the Pearson distribution for 12-month timescale for Sep of 2011
SPI with the Pearson distribution for 12-month timescale for Sep of 2012
SPI with the Pearson distribution for 12-month timescale for Sep of 2013
SPI with the Pearson distribution for 12-month timescale for Sep of 2014
SPI with the Pearson distribution for 12-month timescale for Sep of 2015
SPI with the Pearson distribution for 12-month timescale for Sep of 2016
SPI with the Pearson distribution for 12-month timescale for Sep of 2017
SPI with the Pearson distribution for 12-month timescale for Sep of 2018
SPI with the Pearson distribution for 12-month timescale for Sep of 2019
km
0
50
100
N
−2
−1
0
1
2
SPI

(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4.14: SPI-12 spatial map through September for the years 2008 to 2019, historical data: 1979–2019 with the Pearson type III distribution.


SPEI with the Pearson distribution for 12-month timescale for Sep of 2008
SPEI with the Pearson distribution for 12-month timescale for Sep of 2009
SPEI with the Pearson distribution for 12-month timescale for Sep of 2010
SPEI with the Pearson distribution for 12-month timescale for Sep of 2011
SPEI with the Pearson distribution for 12-month timescale for Sep of 2012
SPEI with the Pearson distribution for 12-month timescale for Sep of 2013
SPEI with the Pearson distribution for 12-month timescale for Sep of 2014
SPEI with the Pearson distribution for 12-month timescale for Sep of 2015
SPEI with the Pearson distribution for 12-month timescale for Sep of 2016
SPEI with the Pearson distribution for 12-month timescale for Sep of 2017
SPEI with the Pearson distribution for 12-month timescale for Sep of 2018
SPEI with the Pearson distribution for 12-month timescale for Sep of 2019
km
0
50
100
N
−2
−1
0
1
2
SPEI

(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

Figure 4.15: SPEI-12 spatial map through September for the years 2008 to 2019, historical data: 1979–2019 with the Pearson type III distribution.

4.6 Conclusions

The southernmost island of Greece, Crete is a drought-prone area with increased agricultural interest due to its advantageous location in the southeast Mediterranean sea. Reanalysis data covering the island are used to assess its drought conditions based on the SPI and the SPEI drought indices. Both SPI and SPEI are normalized, allowing the comparison between areas with different climates. Nevertheless, SPI does not account for the effects of climate change in the sense of the temperature rise, while SPEI does. This means that SPEI is more appropriate for comparison between areas that have diverse temperature profiles. The indices can be estimated for different timescales, thus letting for the assessment of short-term and long-term dry conditions. The historical record used in this analysis covers 41 continuous years starting in 1979, while the implementation was carried out through an open source Python package [Adams, 2019] downloaded from `https://github.com/monocongo/climate_indices`.

The temporal drought identification was focused on four important areas of the island, i.e., Chania, Heraklion, Messara, and Ierapetra —the first two are the biggest cities of Crete, while the rest are the main agricultural regions. The indices provide similar patterns for the four areas, however, some discrepancies are observed primarily at the longest timescales. Multiple drought events occurred during the analyzed period with the most severe events during the years 1990–1992, 2000–2002, and 2016–2019. Also, the frequency of the drought events has increased noticeably for all the locations after 2000.

Drought maps for the 3-month and the 12-month timescales were generated based on the piecewise cubic interpolation for both indices. The use of maps allows the visual inspection of spatial patterns.

Based on both the temporal and spatial analysis, it is evident that the eastern and southern parts of the island are more sensitive to drought events with longer duration. Additionally, it is revealed by the analysis that Heraklion, the biggest and the most urbanized city in Crete, had the most significant impact on drought characterization due to temperature change, resulting in an approximate 0.6% increase in drought events when the temperature (SPEI) is considered. Furthermore, significant changes in drought characterization we observe for Ierapetra, with 0.8% decrease in drought events below −1 when temperature is included in the calculation.

The indices provide similar results in the earlier years, however, as time evolves and the climate change impacts are more prominent the results start to differentiate. The good correlation between SPI and SPEI at different time scales for the early years implies that they were well adjusted to the study area, yet the deviations present in the recent decade suggest that an index that considers temperature is more appropriate.

Both indices are proven to be valuable tools for monitoring and assessment of the drought concerns in the study area. In general, indices are easy to implement, especially since the explosion in data availability from remote sensing sources, and with a basic understanding of the underlying climate of the region under inspection easy to comprehend and enable the appropriate agencies to take further action. Sub-seasonal forecasts for weather and climate conditions would allow water resource managers to prepare for changes in hydrologic regimes that can pose dangers to the management of water resources.

A future extension of this study, might include the estimation of the PET values from other formulations such as the Penman-Monteith equation [Allen et al., 1998], and the Hargreaves equation [Hargreaves and Samani, 1982]. Furthermore, the use of the log-logistic distribution for the SPEI calculation can be implemented, which is the recommended distribution by the creators of the SPEI [Beguería et al., 2014].

Chapter 5

Gaussian Anamorphosis of Precipitation Data

5.1 Summary

In this chapter, we begin by emphasizing the non-Gaussian character of the variables from the natural world, such as precipitation at different temporal resolutions. On the one hand, because Gaussianity is a prerequisite or an assumption for the optimal application of classical geostatistical methodologies we present strategies that are utilized to achieve that. We offer the definitions and formulations for the Box-Cox and Yeo-Johnson transformations as well as the Gaussian anamorphosis with Hermite polynomials (GAH) which is used in this research. On the other hand, fitting a parametric distribution to the input data is needed. The spatial and temporal scale of the data affects their variability, making it impossible to hypothesize about the ideal parametric distribution before investigation. This step introduces complexity and increases the pre-processing time. GAH combined with kriging has been used before in a similar way, however, to our knowledge never on precipitation data. The proposed approach overcomes both problems by estimating the cumulative density function of the data with Hermite polynomials (each time adapted to the data set) and further transforming it to Gaussian distributed values. Ten processing scenarios were investigated and their performance with respect to spatial interpolation (based on Ordinary kriging) was evaluated. The scenarios include the application or exclusion of GAH with varying polynomial degrees, the utilization of either the exponential or Spartan variogram models, and the incorporation or omission of Monte Carlo simulations.

5.2 Introduction

Gaussianity is a prerequisite for applying various statistical methodologies; however, in the real world, most of the documented phenomena, including precipitation (as well as other environmental variables) can be described better with non-Gaussian probability distribution functions [Papalexiou and Serinaldi, 2020; Papalexiou et al., 2021]. This means that spatial or spatiotemporal data (e.g., precipitation, wind speed, etc.) exhibit one or more of the following characteristics: asymmetric probability distributions, strictly positive values, long positive tails, and compact support [Hristopulos, 2020].

Various methodologies are available to address the problem that non-Gaussianity poses, yet it is still far from solved. The most widely known and used methods involve non-linear transformations of latent Gaussian random fields that generate either binary-valued (e.g., indicators) or continuous (e.g., lognormal) non-Gaussian random fields.

We will focus on non-Gaussian random fields generated by non-linear transformations such as normal scores, Box-Cox, and Hermite polynomials.

The terms normal scores transform and "Gaussian anamorphosis" refer to the bijective (one to one and invertible) mapping $g : X(\mathbf{s};\omega) \mapsto Y(\mathbf{s};\omega)$ from a random field $X(\mathbf{s};\omega)$ with non-Gaussian marginal distribution to a random field $Y(\mathbf{s};\omega) = g[X(\mathbf{s};\omega)]$ with a Gaussian marginal distribution [Chilès and Delfiner, 2012; Hristopulos, 2020; Wackernagel, 2003]. A similar "warping" approach has been applied to Gaussian process regression [Snelson et al., 2004]. The term "warping" herein refers to the nonlinear transformation of the Gaussian process.

For the transformation from a Non-Gaussian to Gaussian field the following expression is true: If $F_X(x)$ is the marginal CDF of $X(\mathbf{s};\omega)$, the function $g(\cdot)$ is defined by $g : x \mapsto y$ such that $y = \Phi^{-1}[F_X(x)]$. For the inverse transformation, i.e., from a Gaussian to Non-Gaussian field the following expression is true: Assuming that the inverse CDF, $F_X^{-1}(\cdot)$, exists, the anamorphosis function $\phi = g^{-1}$ defined by $\phi : y \mapsto x$ so that $x = F_X^{-1}[\Phi(y)]$, is a bijective mapping from the Gaussian variable y to the target variable x [Hristopulos, 2020].

The normal scores transform maps the sample values $\{x_n^*\}_{n=1}^N$ into a respective set of values $y_n^* = g(x_n^*)$ that follow the standard normal distribution. It is important to highlight that anamorphosis is not always attainable. For example, transforming a probability distribution with significant weight at a single value (e.g., if the observed process has a large number of zeros, a case commonly occurring in mineral reserves and precipitation data sets), transformation to a normal distribution does not work well [Armstrong and Matheron, 1986; Hristopulos, 2020].

Three of the most common and easily applicable normal scores transformations are the square root, the Johnson's Hyperbolic Sine and the Box-Cox transformations. The square root transformation has limited use because it can be applied to only non-negative values. Also, it is non-integrable at zero values, disqualifying it for applications on precipitation data. The parametric Johnson's Hyperbolic Sine transformation utilizes the hyperbolic sine function and is flexible enough to match any data set, due to its four parameters (a location, a scale, and two shape parameters) [Hristopulos, 2020]. In the following sections we will focus our applications on the ERA5 monthly precipitation data set in the island of Crete.

5.3 Normality Transformations

A widely adopted transformation that aims to approximate the Gaussian distribution is the Box-Cox transformation. The transformation is named after George Edward Pelham Box and Sir David Roxbee Cox [Box and Cox, 1964], the statisticians that developed the methodology.

The one-parameter Box–Cox transformation is defined as

$$Y(\mathbf{s};\omega) = \begin{cases} \dfrac{1}{\lambda}\left[X^{\lambda}(\mathbf{s};\omega) - 1\right], & \text{if } \lambda \in \mathbb{R} \text{ and } \lambda \neq 0 \\ \ln\left[X(\mathbf{s};\omega)\right], & \text{if } \lambda = 0, \end{cases} \tag{5.1}$$

and the two-parameter Box–Cox transformation as

$$Y(\mathbf{s};\omega) = \begin{cases} \dfrac{[X(\mathbf{s};\omega)+\lambda_2]^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda \in \mathbb{R} \text{ and } \lambda_1 \neq 0 \\ \ln[X(\mathbf{s};\omega)+\lambda_2], & \text{if } \lambda_1 = 0, \end{cases} \tag{5.2}$$

as described in the original article by Box and Cox [1964]. The transformation Eq. (5.1) holds for $X(\mathbf{s};\omega) > 0$, while Eq. (5.2) for $X(\mathbf{s};\omega) > -\lambda_2$. It is common to use for the parameter λ_2 the negative value of x_{min}, which guarantees that the quantity $[X(\mathbf{s};\omega)+\lambda_2]$ is always ≥ 0, thus viewing only λ_1 as the model parameter that needs optimization.

The parameter λ is estimated from the data so that the marginal probability distribution of $Y(\mathbf{s};\omega)$ approximates the Gaussian distribution. The selection can be settled via the maximum likelihood estimation or another model selection criterion (e.g., AIC, BIC).

The inverse one-parameter Box-Cox transformation is defined as follows

$$X(\mathbf{s};\omega) = \begin{cases} [\lambda Y(\mathbf{s};\omega)+1]^{1/\lambda}, & \text{if } \lambda \in \mathbb{R} \text{ and } \lambda \neq 0 \\ \exp[Y(\mathbf{s};\omega)], & \text{if } \lambda = 0. \end{cases} \tag{5.3}$$

The case $\lambda = 0$ implies that $X(\mathbf{s};\omega)$ follows the lognormal distribution if $Y(\mathbf{s};\omega)$ is a normally distributed random field.

The Yeo–Johnson transformation [Yeo and Johnson, 2000] is more flexible than the original Box-Cox transformation, since it allows also for zero and negative values of $X(\mathbf{s};\omega)$. λ can be any real number, where $\lambda = 1$ produces the identity transformation. The transformation is defined as

$$Y(\mathbf{s};\omega) = \begin{cases} \dfrac{1}{\lambda}\Big[(X(\mathbf{s};\omega)+1)^{\lambda}-1\Big], & \text{if } \lambda \neq 0 \text{ and } X(\mathbf{s};\omega) \geq 0 \\ \log[X(\mathbf{s};\omega)+1], & \text{if } \lambda = 0 \text{ and } X(\mathbf{s};\omega) \geq 0 \\ \dfrac{-\Big[(-X(\mathbf{s};\omega)+1)^{(2-\lambda)}-1\Big]}{2-\lambda}, & \text{if } \lambda \neq 2 \text{ and } X(\mathbf{s};\omega) < 0 \\ -\log[-X(\mathbf{s};\omega)+1], & \text{if } \lambda = 2 \text{ and } X(\mathbf{s};\omega) < 0. \end{cases} \tag{5.4}$$

If $X(\mathbf{s};\omega)$ is strictly positive, then the Yeo-Johnson transformation is the same as the Box-Cox power transformation of $[X(\mathbf{s};\omega)+1]$. If $X(\mathbf{s};\omega)$ is strictly negative, then the Yeo-Johnson transformation is the Box-Cox power transformation of $[-X(\mathbf{s};\omega)+1]$, but with power of $(2-\lambda)$. As in the case of the Box-Cox transformation, the parameter λ is estimated from the data so that the marginal probability distribution of $Y(\mathbf{s};\omega)$ approximates the Gaussian distribution. The selection can be settled via the maximum likelihood estimation or another model selection criterion (e.g., AIC, BIC). In the original study, Yeo and Johnson [2000] estimated the λ value by minimization of the Kullback-Leibler distance between the normal distribution and the transformed distribution.

One main difference between the Yeo-Johnson and the Box-Cox transformation is their behavior when $X(\mathbf{s};\omega)$ has values closer to zero. In that case, the Box-Cox creates a more significant difference between the original and the transformed values closer to zero than the Yeo-Johnson does.

The drawback when using a power transformation such as the aforementioned is that they work well only when the random field $X(\mathbf{s};\omega)$ differs slightly from the Gaussian distribution. The majority of precipitation data calculated in different timescales do not approximate the Gaussian distribution. The same is true for wind speed data, and solar radiation on horizontal plain data. Moreover, in semi-arid climates such as our case study, it is common to have zero precipitation even on a monthly timescale. However, aggregated precipitation for longer periods (e.g., annual) tends to approximate better the Gaussian distribution than for shorter timescales (e.g., daily).

5.4 Gaussian Anamorphosis with Hermite Polynomials

Gaussian Anamorphosis with Hermite Polynomials is a more sophisticated methodology than the preceding non-linear models. It aims to transform a given non-Gaussian random field $X(\mathbf{s};\omega)$ to a random field $Y(\mathbf{s};\omega)$ that approximates better the Gaussian distribution. In this case, the bijective transformation function $\phi = g^{-1}$ consists of several Hermite polynomials defined as derivatives of the Gaussian density function. In other words, the method employs a series of orthogonal Hermite polynomials to convert the original variable into a corresponding variable that follows the normal distribution. This transformation modifies the marginal distribution of the data to a normal marginal distribution [Hristopulos, 2020; Wackernagel, 2003].

The Hermite polynomials are defined as derivatives of the Gaussian density function. If $f(y)$ denotes the standard Gaussian density function ($f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, with $\mu = 0$ and $\sigma = 1$), then the probabilists' Hermite polynomials [Hristopulos, 2020] are given by

$$H_m(y) = (-1)^m \frac{\frac{d^m}{dy^m}(f(y))}{f(y)}, \qquad m \geq 0. \tag{5.5}$$

For example

$$\begin{aligned} H_0(y) &= \frac{f(y)}{f(y)} = 1, \\ H_1(y) &= (-1)\,\frac{\frac{d}{dy}(f(y))}{f(y)} = y, \\ H_2(y) &= (-1)^2\,\frac{\frac{d^2}{dy^2}(f(y))}{f(y)} = y^2 - 1. \end{aligned} \tag{5.6}$$

The m$^{\text{th}}$ degree Hermitian polynomials are calculated from the following recurrence expression:

$$H_{m+1}(y) = yH_m(y) - mH_{m-1}, \qquad m \geq 0. \tag{5.7}$$

For example, based on Eq. (5.7)

$$\begin{aligned} H_2(y) &= yH_1(y) - H_0(y) = y^2 - 1, \\ H_3(y) &= yH_2(y) - 2H_1(y) = y^3 - 3y, \\ H_4(y) &= yH_3(y) - 3H_2(y) = y^4 - 6y^2 + 3. \end{aligned} \tag{5.8}$$

For the application of the Gaussian Anamorphosis transformation, first, the empirical cumulative distribution $\hat{F}_X(x)$ of the non-Gaussian random variable X (e.g., rainfall height) is constructed from the data set (Fig. 5.1a). Next, the upper order M of the Hermite polynomials for the transformation $g : x \mapsto y = g\left[X\left(\mathbf{s};\omega\right)\right] = \Phi^{-1}\left[F_X(x)\right]$ is selected so that the transformed random variable follows the standard normal distribution (Fig. 5.1b), i.e., $Y \approx N(0,1)$, where $F_X^{-1}(x)$ is the inverse CDF of $Y(\mathbf{s};\omega)$. The inverse transformation $\phi = g^{-1} : y \mapsto x = g^{-1}[Y(\mathbf{s};\omega)] = F_X^{-1}[\Phi(y)]$ functions as a retrieval mechanism for the original variable $X(\mathbf{s};\omega)$ from the transformed one. Specifically, the random variable x can be expressed by means of the Hermite polynomials and the the normal values y as follows:

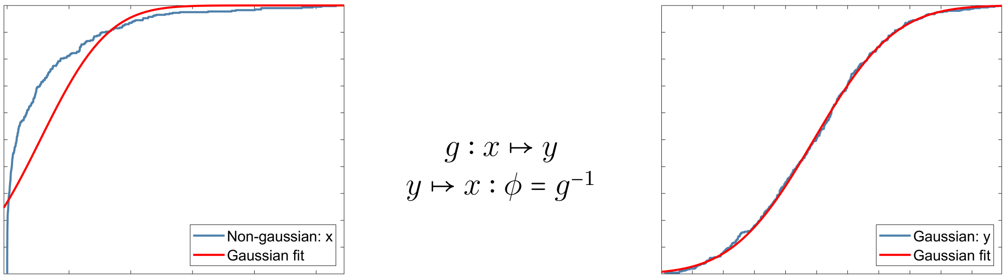
$$x = \phi(y) = \sum_{m=0}^{\infty} \frac{c_m}{m!} H_m(y). \tag{5.9}$$

The infinite series is usually truncated to an integer value M, which defines the order of the transformation. The coefficient c_m is equal to the expected value (expectation) of the product $x \times H_m(y)$. The calculation of the coefficients c_m from a sample with n elements $(x_1, x_2, \ldots, x_n)$ is performed based on the following equation [Wackernagel, 2003]:

$$\begin{aligned} c_m &= \int_{-\infty}^{\infty} \hat{\phi}(y)\, H_m(y)\, f(y)\, \mathrm{d}y \\ &= \sum_{i=1}^{m} x_{[i]} \int_{A_n} H_n(y)\, f(y)\, \mathrm{d}y \\ &= \sum_{i=1}^{m} x_{[i]} \left(H_{m-1}(y_i)\, f(y_i) - H_{m-1}(y_{i-1})\, f(y_{i-1})\right), \qquad m = 1, \ldots, M, \end{aligned} \tag{5.10}$$

where $\hat{\phi}(y)$ is the empirical Gaussian anamorphosis, $f(\cdot)$ is the probability density function of the standard normal distribution and $y_i = G^{-1}([F_X(x_i)])$, where $G^{-1}(\cdot)$ is the inverse cdf of the standard normal probability, and where $A_n, n = 1, \ldots, N$ are the following half-open intervals $A_n = \left(\Phi^{-1}\left(\frac{n-1}{N}\right), \Phi^{-1}\left(\frac{n}{N}\right)\right]$. The order M, in practice takes values from $M = 30$ to $M = 60$. The Gaussian anamorphosis transformation is then approached with a series of $M + 1$ Hermite polynomials of order from $m = 0$ to $m = M$ [Hristopulos, 2020]. For a more extensive analysis of the steps and the functions involved see Hristopulos [2020]; Wackernagel [2003]; Webster and Oliver [2007], and Chilès and Delfiner [2012].

The method of Gaussian anamorphosis with closed form equations is widely used for data transformations in signal analysis, time series, spatial data of environmental pollutants, ore concentration, and other applications [Alecu, 2006;



(a) non Gaussian cdf

(b) Gaussian cdf

Figure 5.1: Cumulative density function (blue lines) and Gaussian model fits (red lines) for non-Gaussian data (a) and their transformed (normalized) data (b) with the application of the Gaussian anamorphosis function.

Chilès and Delfiner, 2012; Lien et al., 2016; Wackernagel, 2003]. However, the use of the Gaussian anamorphosis with Hermitian polynomials is much more limited, mainly due to the increased complexity of the method and the discontinuity of the empirical probability distribution.

Although, the use of Hermite polynomials combined with the empirical cumulative distribution of the data improves the empirical Gaussian anamorphosis, it does not negate the fact that it uses the empirical cumulative density function $\hat{F}_X(x)$, which has a staircase form. To further improve the results, a non-parametric kernel-based density estimation as described in the work by Pavlides et al. [2022] can be implemented; such functions allow continuous representations of the probability density function.

5.5 Methodology

As mentioned before, precipitation does not follow the Gaussian distribution and is commonly fitted with the gamma, the lognormal, or the Generalized Extreme Value distributions, depending on the analyzed temporal and spatial scale. This creates additional complexity in the pre-processing step to apply methods dependent on the Gaussian assumption. Our proposed approach addresses this problem by transforming the data into Gaussian distributed values via the application of the Gaussian Anamorphosis with Hermite polynomials (GAH). After the transformation, the methodology consists of the following steps: (i) variogram estimation, (ii) bootstrap simulation, (iii) kriging prediction, (iv) back-transformation of the normal estimates to precipitation values, and (v) cross-validation measures. The steps are summarized in the flowchart of Fig. 5.2.

GAH in conjunction with kriging methods has been applied before. Methods that are similar but with minor differences with the approach followed here include the multi-Gaussian cokriging and factor kriging analysis [Buttafuoco et al., 2011], and Bi-Gaussian disjunctive kriging [Armstrong and Matheron, 1986; Guibal and Remacre, 1984; Matheron, 1984; Ortiz et al., 2005].

The spatial correlations of the precipitation field $X_t(\mathbf{s})$ (or the transformed gaussian field $Y_t(\mathbf{s})$) are determined by means of the variogram function γ_{X_t} [Chilès and Delfiner, 2012]. For a specific month t, the variogram $\hat{\gamma}_{X_t}(r)$ is defined as the half *variance of the increment field* $X_t(\boldsymbol{s}+\boldsymbol{r}) - X_t(\boldsymbol{s})$, where $\mathbf{r}$ is the *distance vector*. The variogram measures the average increase of the deviation between the field values $X_t(\mathbf{s})$ and $X_t(\mathbf{s}+\mathbf{r})$ as the distance $\|\mathbf{r}\|$ between them increases.

The variogram can be estimated even in cases that the covariance function is hard or impossible to estimate from the data. Thus, stochastic kriging methods are based on the variogram to predict the values of the random field at unsampled locations [Chilès and Delfiner, 2012; Cressie, 1993; Olea, 1999]. The estimation of the variogram from the data is performed by means of the *maximum likelihood estimation* [Aldrich, 1997; Fisher, 1922, 1925; Kitanidis and Lane, 1985; Mardia and Marshall, 1984; Norden, 1972] and by minimizing the mean square estimation error [Chilès and Delfiner, 2012].

To determine the spatial dependence at every possible lag distance r, we need to fit the empirical precipitation variogram with a theoretical variogram function [Chilès and Delfiner, 2012]. We use the *Spartan variogram family* [Hristopulos, 2003, 2015a; Hristopulos and Elogne, 2007], obtained from the Eq. (2.15b). The Spartan variogram includes three parameters, i.e., the scale coefficient η_0, the rigidity coefficient η_1, and the characteristic length ξ. For comparison reasons we also fit the well known Exponential model, given by the Eq. (2.8). The latitude and longitude of the data coordinates are expressed in degrees in the World Geodetic System (WGS 84). Before any calculations, they were transformed to the UTM (Universal Transverse Mercator) coordinate system.

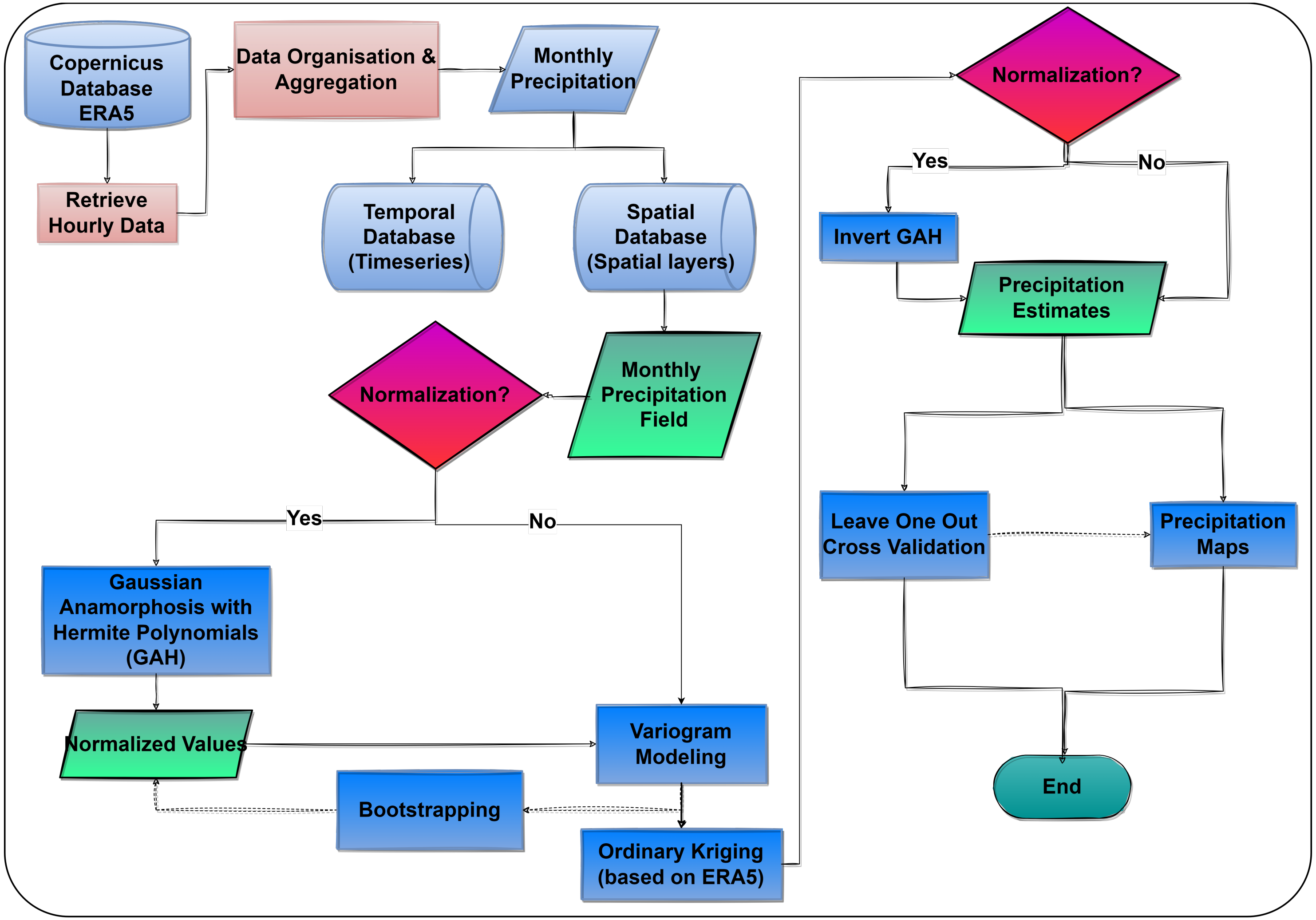

Copernicus Database ERA5
Data Organisation & Aggregation
Monthly Precipitation
Retrieve Hourly Data
Temporal Database (Timeseries)
Spatial Database (Spatial layers)
Monthly Precipitation Field
Normalization?
Yes
No
Gaussian Anamorphosis with Hermite Polynomials (GAH)
Normalized Values
Variogram Modeling
Bootstrapping
Ordinary Kriging (based on ERA5)
Normalization?
Yes
No
Invert GAH
Precipitation Estimates
Leave One Out Cross Validation
Precipitation Maps
End

Figure 5.2: Flowchart summarizing the main methodological steps for normalization of the data, spatial model estimation and mapping of the monthly precipitation field.

To assess the performance of the model on the precipitation reanalysis data, we use LOO-CV (see Section 2.12 for more details on the methodology). In LOO-CV the training set contains $N-1$ values and the validation set contains a single value. All N possible partitions of the data into training and validation sets are used.

The predictive performance of different models is assessed by means of statistical measures which include: the bias or mean error (ME), the mean absolute error (MAE), the root mean square error (RMSE), and the Pearson's linear correlation coefficient (RP) (see Section 2.12 for the definitions).

5.6 Results

In the following section we implement the newly described method that combines Gaussian anamorphosis with Hermite polynomials coupled with geostatistical simulation. The analysis is focused on the monthly ERA5 precipitation products for the wet period of 246 months (from January to March and October to December 1979 to 2019 - 41 years) for the entire grid. For brevity, we mostly discuss results from a single randomly chosen year, the year of 2008. We chose a year besides the characteristic (1980), the one with the lowest precipitation (1990), and the one with the highest precipitation (2019) that were previously discussed in Section 4 (Estimation of Drought Indices for the Island of Crete) in order to present results that cover a wider range of the analyzed time period. The scenarios tested include Ordinary Kriging, Gaussian Anamorphosis with Hermite polynomials, and Monte Carlo simulations. We use the Exponential or the Spartan model to capture the spatial variability.

Specifically, the first scenario uses the GAH transformation of order 20 to the monthly data, the exponential covariance model coupled with bootstrap simulations, and the OK method (S1). The second scenario uses the GAH transformation of order 20 to the monthly data, the Spartan covariance model coupled with bootstrap simulations, and the OK method (S2). The third scenario uses the GAH transformation of order 20 to the monthly data, the exponential covariance model, and the OK method (S3). The fourth scenario uses the monthly data, the exponential covariance model and OK method (S4). The rest of the scenarios follow the same structure with different polynomial order or different covariance model. All the scenarios are presented in Box 5.1.

Box 5.1 | Scenarios Configuration

1. GAH transformation of order 20 to the monthly data, exponential covariance model coupled with bootstrap simulations, and OK (S1)
2. GAH transformation of order 20 to the monthly data, Spartan covariance model coupled with bootstrap simulations, and OK (S2)
3. GAH transformation of order 20 to the monthly data, exponential covariance model, and OK (S3)
4. Exponential covariance model and OK (S4)
5. GAH transformation of order 35 to the monthly data, exponential covariance model coupled with bootstrap simulations, and OK (S5)
6. GAH transformation of order 35 to the monthly data, Spartan covariance model coupled with bootstrap simulations, and OK (S6)
7. Spartan covariance model and OK (S7)
8. GAH transformation of order 20 to the monthly data, Spartan covariance model, and OK (S8)
9. GAH transformation of order 35 to the monthly data, Spartan covariance model, and OK (S9)
10. GAH transformation of order 35 to the monthly data, exponential covariance, and OK (S10)

In Table 5.1, the summary statistics of the wet period monthly precipitation data are concentrated. They include the mean value, the median, the minimum and maximum values, the standard deviation, the coefficient of variation (ratio of

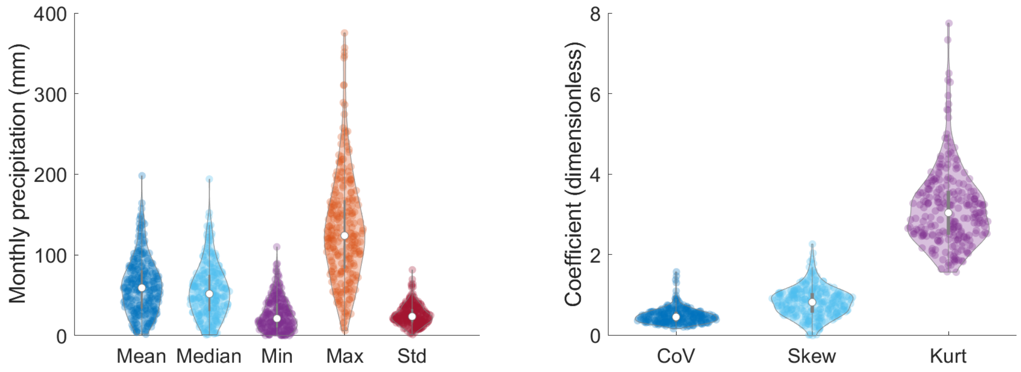
Table 5.1: Mean, median, minimum and maximum values (shown across different rows) of monthly ERA5 precipitation statistics (shown across the columns) based on 246 monthly values. Each monthly statistic is obtained from the 65 values in the respective spatial layer. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

	Mean	**Median**	**Min**	**Max**	**Std**	**CoV**	**Skew**	**Kurt**
Mean	61.25	55.69	26.19	132.70	25.53	0.48	0.82	3.16
Median	59.19	51.78	21.23	123.98	23.67	0.45	0.81	3.04
Minimum	1.75	1.05	0.05	6.10	1.16	0.16	−0.01	1.56
Maximum	198.27	194.15	110.03	375.32	81.54	1.57	2.26	7.75

the standard deviation over the mean), the skewness (coefficient of asymmetry) and the kurtosis. Note that Table 5.1 is the same as Table 3.2 but we additionally present it here for easy access and relevance to this Section. The table is supplemented by Fig. 5.3 which shows the probability distribution of the monthly statistics (corresponding to different columns of Table 5.1) calculated over the 246 months. As described in the Section 3.4.1, the monthly precipitation data for the wet period do not approximate well the Gaussian distribution. This fact can be supported by the statistical measures presented in Table 5.1 where the non-zero skewness and the deviation of kurtosis from three imply non-Gaussian distributions. The variability of the optimal fits is more obvious in Table B5. The same conclusion can be drawn for the dry period monthly precipitation (see Tables C1 and C2 in Appendix C).

For the data grouped by month, the optimal distribution according to AIC is the Nakagami model for 8 out of the 246 wet months, the Weibull for 3, the gamma for 13, the GEV for 25, the Rayleigh for 1, the generalized Pareto for 126, the log-logistic for 2, the lognormal for 3, the Birnbaum-Saunders for 25, and the inverse Gaussian distribution for the remaining 40 wet months.

Particularly for the year 2008, the optimal probability distribution fits for the monthly ERA5 precipitation data based on the AIC and the BIC criteria are



(a) Statistics measured in mm

(b) Dimensionless statistics

Figure 5.3: Violin plots for the mean, median, minimum and maximum values of monthly ERA5 precipitation statistics based on 246 monthly values from the wet period. Each monthly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

presented in Table 5.2. It is obvious that the distributions of the monthly precipitation data (Fig. 5.4) are not only non-Gaussian but they do not follow a specific distribution throughout the year. The optimal distribution to the majority of the optimal fits may be considered the generalized pareto (GP), however, this does not mean that the fit of the GP is representative for the data. This can only mean that from the 17 distributions tested, the GP gave the best fit based on the AICc and the BIC criteria. The distributions tested include the: generalized pareto, inverse Gaussian, lognormal, t-Scale location, Generalized Extreme Value, Weibull, Gaussian, Birnbaum-Saunders, exponential, extreme value, gamma, Nakagami, logistic, log-logistic, beta, Rayleigh, and Rician distribution. This fact makes even more urgent the need for a tool that can omit the distribution fitting step.

5.6.1 Data transformation with the Gaussian anamorphosis with Hermite polynomials

Initially, we estimate the anamorphosis function $\hat{\phi}(y)$, which is the empirical transformation calculated from the ordered sample values (see Section 5.4). In Fig. 5.4 the histogram of the initial precipitation values is plotted against the

Table 5.2: Optimal probability distribution fits for the monthly ERA5 precipitation data in the year 2008. The models studied include the following: "GP": Generalized Pareto, "InvGauss": Inverse Gaussian, "Logn": Lognormal, and "Wei" refers to the Weibull distribution.

January	**February**	**March**	**October**	**November**	**December**
GP	InvGauss	Logn	GP	GP	Wei

back-transformed values of the wet period monthly precipitation based on the GA with Hermite polynomials up to degree 20 for the year of 2008. The histogram of the monthly ERA5 precipitation appears to be in accord with the GAH inverted values. It can be expected that for a non-Gaussian field, the higher the polynomial order, the closer the GAH inverted values will be to the original field. However, as it will be obvious in the following sections, the computational strain from higher polynomial order did not improve the results dramatically. Additionally, we performed a preliminary sensitivity analysis which resulted in very small improvement of the estimations but with high computational cost when increasing the polynomial order above 20 (not presented).

Schematically, in Fig. 5.5 the anamorphosis function of orders 20 and 35 are illustrated against the empirical Gaussian anamorphosis for four different monthly data sets. Herein, we faced a well-known problem of Gaussian anamorphosis methodology, i.e., its difficulty to translate accurately the lowest and the highest values of the sample data to normalized values. This happens due to the oscillations created by the Hermite polynomials near the boundaries. This effect is especially apparent when there are several sample values near the minimum or maximum values in the sample.

5.6.2 Variogram estimation and modeling

The estimation of the theoretical model was carried out utilizing two covariance models, the Spartan variogram which is a three parameter model inspired from statistical physics [Hristopulos, 2003], and the extensively used exponential model. The empirical variograms of the monthly ERA5 precipitation for the

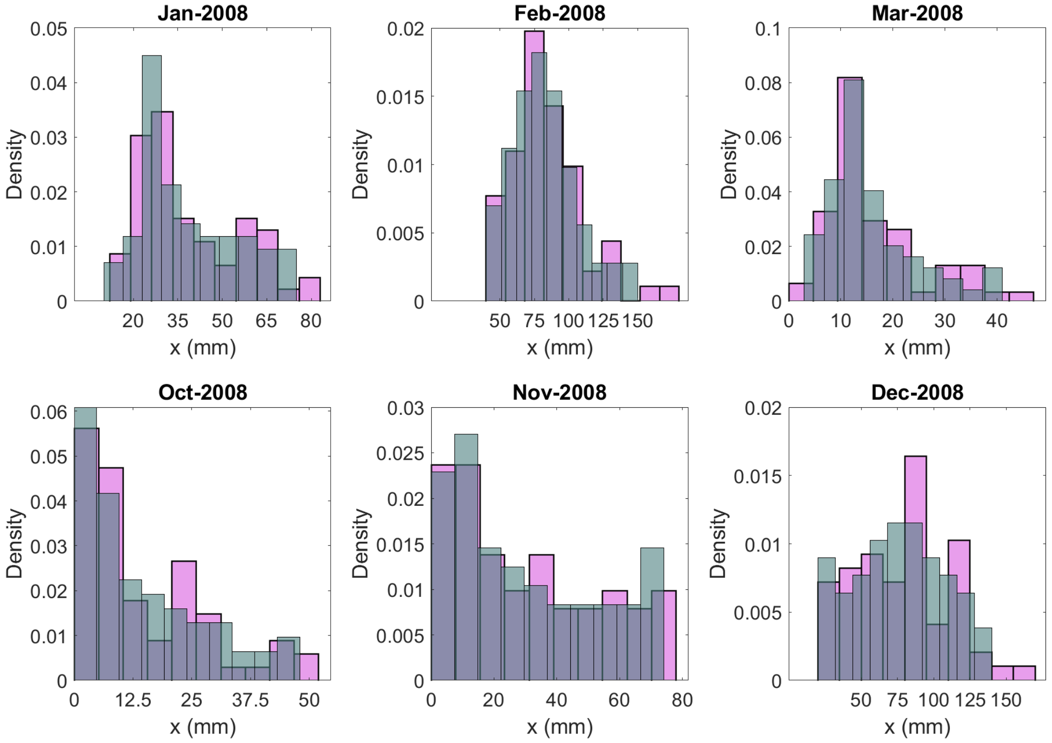
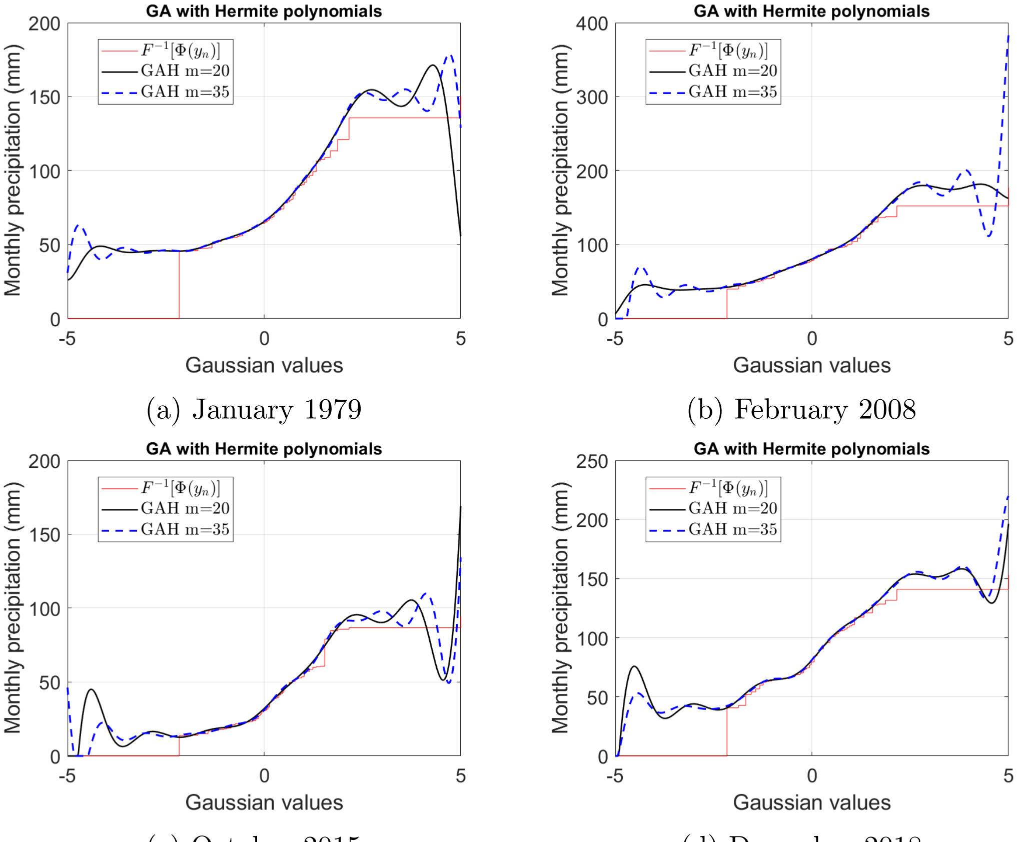


Figure 5.4: Empirical histogram for the monthly precipitation in the year 2008 (gray-green bins) and empirical histogram for the back-transformed data with GA using the first twenty Hermite polynomials (pink bins). The histograms are based on precipitation amounts at the ERA5 65 grid locations over and around Crete.

entire grid were estimated for all the time steps. We estimated the optimal parameters of the theoretical models using maximum likelihood optimization and weighted least-squares distance so that the theoretical model approximates the empirical model. We should point out that the data set for every timestep includes only 65 points. Note that this is a small data set for accurate estimation of the variogram model. In addition to that, the Spartan covariance model while being flexible due to its three parameters, it can vary a lot when there are not enough data, resulting in optimizing part of the variogram curve and diverting from the rest. This problem is more apparent with the MLE than it is with the minimization of the cost function which in our case is the weighted least-squares



(a) January 1979

(b) February 2008

(c) October 2015

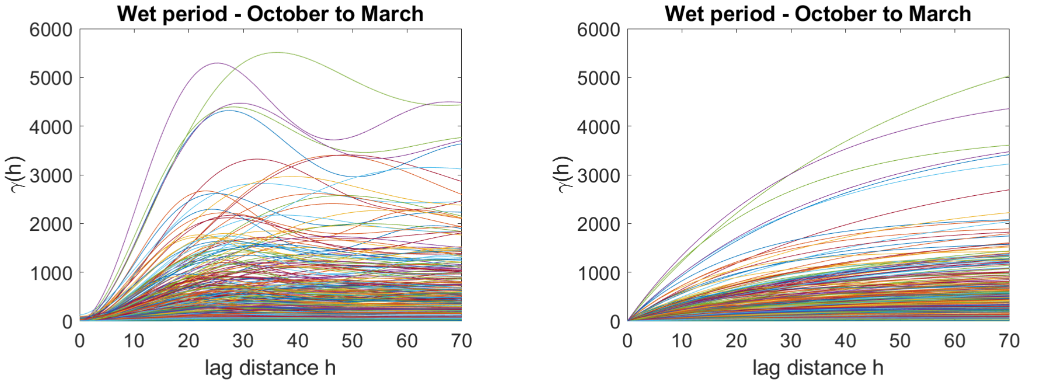
(d) December 2018

Figure 5.5: Precipitation versus standardized values y_n: empirical anamorphosis (staircase, red), GAH with 20 polynomials (continuous black line), and GAH with 35 polynomials (dash line, blue). The data correspond to precipitation amounts from (a) January 1979, (b) February 2008, (c) October 2015, and (d) December 2018 from the 65 ERA5 locations.

distance of the model fit to the empirical variogram. This commonly happens with the MLE method after arriving to a local minimum, hence optimizing a part of the variogram curve and not all of it.

Variogram estimation without transformation

The data sets representing the wet period, consists of the monthly precipitation from October till March, while the dry period data comprises the precipitation from April till September. The fitted variogram clouds for the wet period with the Spartan and the exponential variogram are presented in Fig. 5.6. The variograms for the dry period (Fig. C1) as well as a more detailed visual classification of the variogram clouds by month are presented in Appendix C (see Figs. C2 and C3). It can be noted that the variograms are very dispersed and their range vary greatly. The Spartan variograms estimated via the local MLE optimizer and the global MLE optimizer give the same results. The highest value of the variograms estimated with the WLS optimizer give almost 2000mm^2 and 5000mm^2 higher than the highest values estimated from the MLE method for the Spartan and the exponential variogram values respectively. Because the sill of the variogram is representative of the sample variance, the differences in the variograms' range indicate that the variance of the estimates will be higher with the WLS estimated parameters.



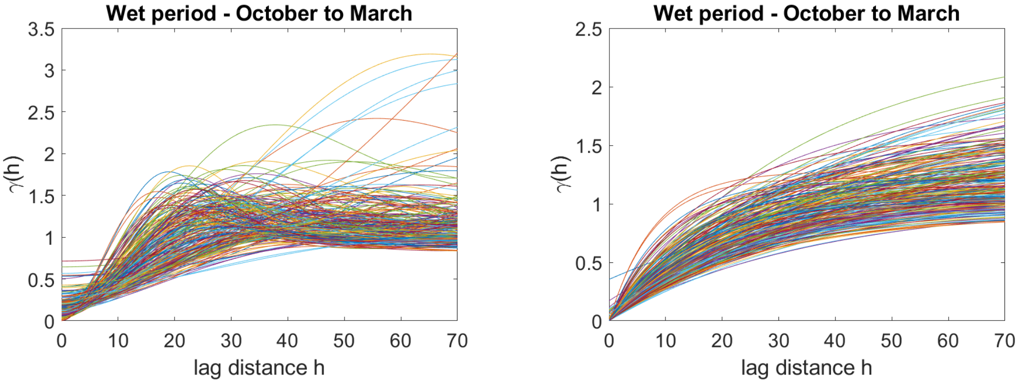
(a) Varios Wet GA false

(b) Varios Wet GA false Exponential

Figure 5.6: Spartan (left) and exponential (right) variogram fits of the monthly ERA5 precipitation obtained from different spatial layers for the wet period (from October till March). Each of the variograms is calculated based on 65 spatial locations for 246 months (corresponding to 246 spatial layers for the wet period). The precipitation values are used without the application of a normalizing transform.

Variogram estimation for the normalized data with GAH up to polynomial degree 20

Contrary to the variogram clouds estimated without the transformation, the variogram clouds estimated for the monthly precipitation after the normalization transform exhibit similar patterns, tend to concentrate more around the median variogram and have approximately the same range. Figure 5.7 presents the cloud variogram fits for the wet periods with the Spartan and the exponential variogram separately after the application of the GAH transformation of order 20. The corresponding cloud variograms for the dry period (Fig. C4) and the variogram clouds per month are presented in Appendix C (see Figs. C5 and C6). Even for the dry period where most of the values are close to zero we see that the theoretical variograms for both models are more concentrated and follow similar patterns when the GAH is applied.



(a) Varios Wet GA true

(b) Vario Wet GA true Exponential

Figure 5.7: Spartan (left) and exponential (right) variogram cloud fits of the monthly ERA5 precipitation obtained from different spatial layers for the wet period (from October till March). Each of the variograms is calculated based on 65 spatial locations for 246 months (corresponding to 246 spatial layers for the wet period). The normalized precipitation values are generated with the GAH normalizing transform up to degree 20.

5.6.3 Bootstrapping simulation

The bootstrapping simulation was performed for every time step (492 months) for four of the ten scenarios tested, namely (i) Gaussian anamorphosis of order 20 coupled with simulations and ordinary kriging with the exponential variogram model (S1), (ii) Gaussian anamorphosis of order 20 coupled with simulations and ordinary kriging with the Spartan model (S2), (iii) Gaussian anamorphosis of order 35 coupled with simulations and ordinary kriging with the exponential variogram model (S5), (iv) Gaussian anamorphosis of order 35 coupled with simulations and ordinary kriging with the Spartan model (S6).

For every time step in every LOOCV step, 100 simulated fields were generated based on the optimal parameters retrieved from the variogram estimation step. The simulation resulted in a total of 6500 fields for all 246 months covering the period under investigation. The spatial distribution of the estimated field for every timestep was determined by averaging the 100 bootstrapped simulated values obtained through the 65 LOOCV iterations. The resulting field contains 65 values, one for every location. While simulations are used mainly to capture the variability of the entire field, oftentimes they result in inaccurate estimations due to unreliable covariance function parameters.

5.6.4 Cross Validation

To assess the model's performance, we use the validation measures of Leave-one-out cross-validation. Table 5.3 presents the results for the scenarios tested. The description of the configuration for each scenario is presented in Box 5.1. For a detailed presentation of the validation measures for the first four scenarios see Table C3 in Appendix C.

The distributions of the main LOO-CV values for all the cases with the MLE method are illustrated in Figs. 5.8-5.11. Compared to the recent work by Agou et al. [2022] where they used the same data set we observe that our results provide some improvements. For the scenarios with anamorphosis we obtain lower ME with the GAH up to 20 or 35 degrees utilizing the Spartan variogram model (S8 and S9) that distribute around 0.5 mm (0.5 and 0.48), however with the exponential model (S3 and S10) in both cases we get higher ME values. In terms of the MAE, and RS metrics, we ended up with better measures by using the GAH method for both variogram models (S3, S8, S9, S10), ranging for the MAE from 6.49 to 6.89 mm compared to their 7.53 mm and for the RS from 0.91 to 0.94 versus their 0.90, while the RMSE metrics are better but quite similar. In the S4 and S7 where we used the exponential and the Spartan model without transformations, the average ME is −0.09 mm and −0.08 respectively, which is lower than their −0.15 mm average ME for both the exponential and the Mátern model.

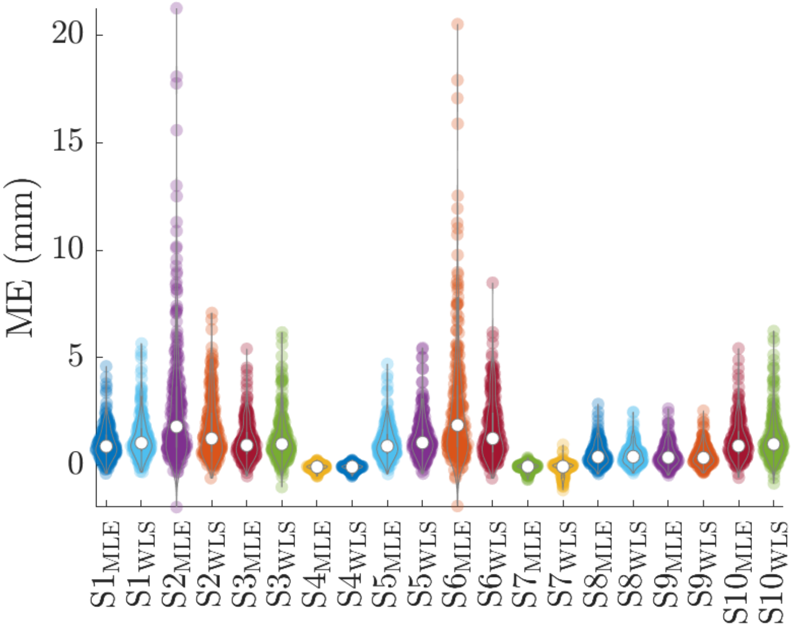


Figure 5.8: Violin plots of monthly precipitation cross-validation ME (CV) for the wet time period (246 months). The scenarios' configuration are shown in Box 5.1.

According to both optimization methodologies (for MLE see Table 5.3 and for WLS see Table 5.4) the addition of simulations for the scenario with the GAH of order 20, and the exponential covariance model (S1 versus S3) improved or maintained the same MARE, RMSRE, RS. However, for the scenarios with the SSRF covariance (S2 versus S8, S6 versus S9) all the validation metrics are worse with the incorporation of the simulations (S2, S6) than without them (S8, S9). For the scenario with the GAH of order 35 and the exponential covariance model (S5 versus S10) the results differ depending on the optimization method. With the MLE method, the metrics improve with the simulations while with the WLS most of the measures worsen or stay the same. Therefore, further investigation is needed to clarify why the simulations in the majority of the scenarios deteriorated the metrics contrary to what was expected. The RS results are comparable with better values with the exponential model and slightly better or worse for the Spartan depending on the optimization method (MLE gave slightly better RS and WLS gave slightly worse RS compared to the published work by Agou et al. [2022]).

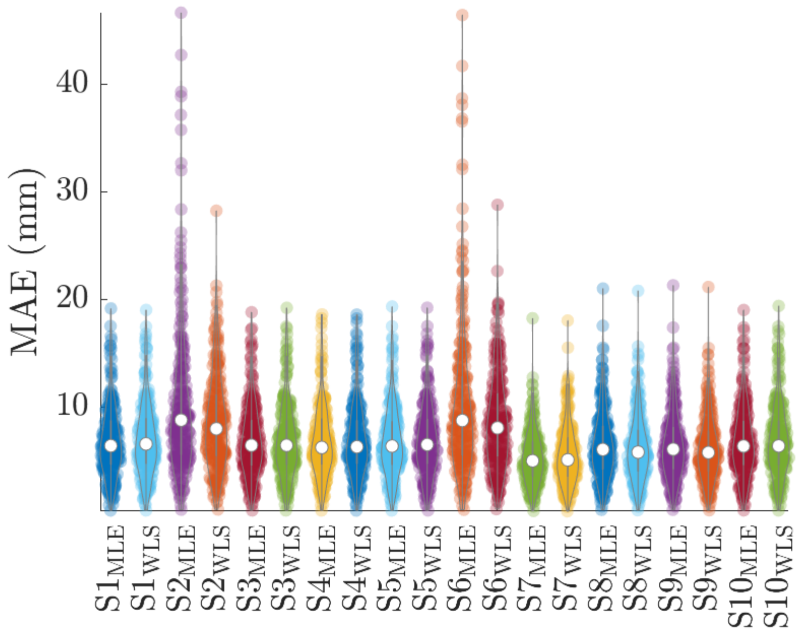


Figure 5.9: Violin plots of monthly precipitation cross-validation MAE (CV) for the wet time period (246 months).

The average values of the validation metrics over the 246 time slices are shown in Table 5.3 and Table 5.4. All the metrics have been rounded to the second decimal place. We provide a more analytical presentation of the first four scenarios in Table C3 in Appendix C, where we include the mean, median, minimum, maximum, and standard deviation values over the 246 time steps. According to the metrics (e.g. the median) and Figs. 5.8-5.11, there is practically no difference between results obtained with the two covariance kernels, despite the fact that the Spartan kernel which includes 3 hyperparameters allows for wider flexibility to data interpolation, except while coupled with simulations. Additionally, we do not observe significant differences between the application of the Gaussian anamorphosis with Hermite polynomials (S1 to S3, S5 to S6, S8 to S10) (a.k.a. warping Gaussian Process Regression wGPR) and without (S4 and S7). Also, the results based on the two optimization methods (MLE and WLS) are almost identical.

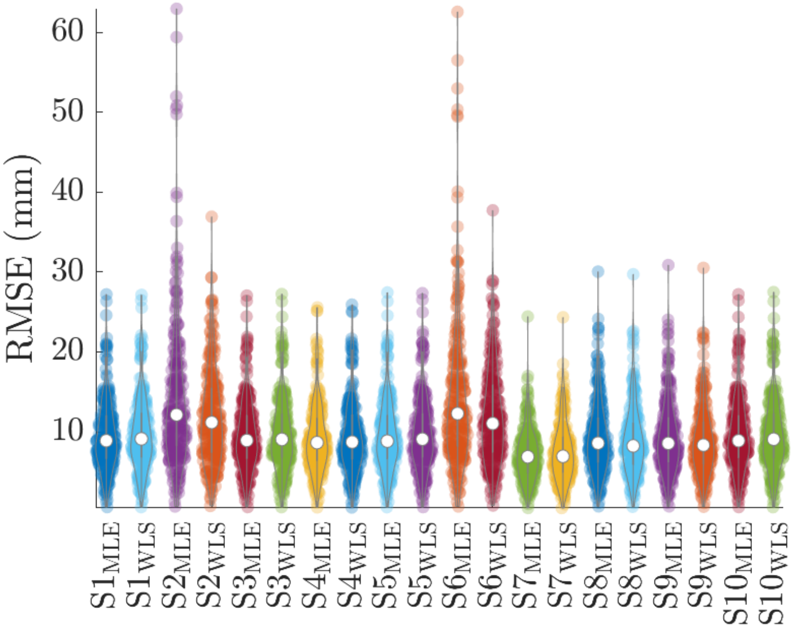


Figure 5.10: Violin plots of monthly precipitation cross-validation RMSE (CV) for the wet time period (246 months).

The main validation measure differences are in the mean error (bias), where in the S4 and S7 (only OK with the exponential or the Spartan model) the value is significantly less than the rest of the case studies. This is expected since the OK works by enforcing a zero-bias constrain, however, kriging variance is independent of the data values, thus, making it unreliable in skewed data sets. Nonetheless, the wGPR bias is still a small fraction of the average minimum of the data (cf. Table 5.1). The comparison (S3 vs S10 and S8 vs S9) of the same scenarios with higher Hermite polynomial order did not significantly improve the results (RMSE increases), but considerably increased the computational cost both in terms of memory and time. The simulation did not improve the validation measures compared with the measures resulted from the scenarios where simulations were not incorporated. Additionally, in terms of the mean value of the validation measures (MLE), the use of the three parametric Spartan model further improved the results in terms with the MAE, RMSE and RS.

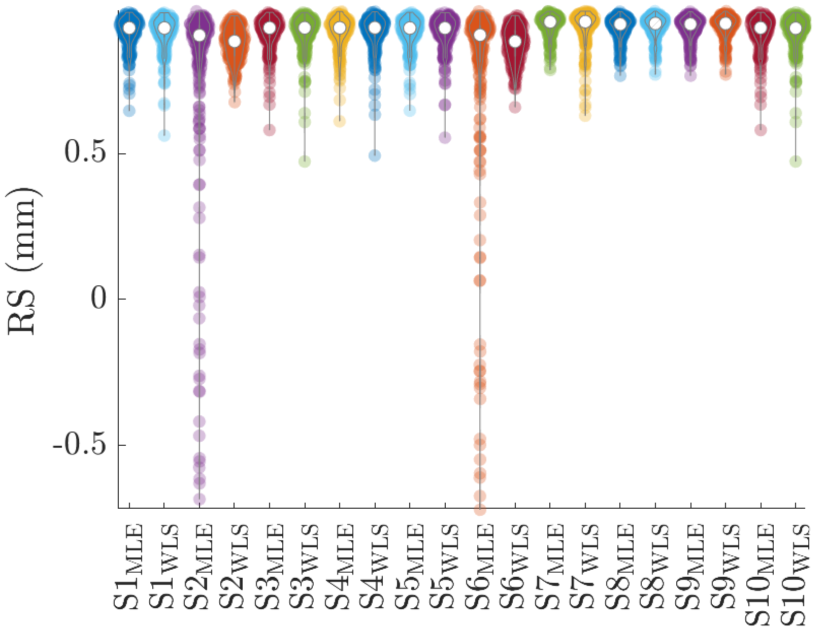


Figure 5.11: Violin plots of monthly precipitation cross-validation RS (CV) for the wet time period (246 months).

Table 5.3: The average value of the LOO-CV measures of monthly precipitation based on 10 different scenarios. The mean value of the following validation measures are shown: ME: mean error (bias); MAE: mean absolute error; MARE: mean absolute relative error; RMSE: root mean square error; RMSRE: root mean square relative error; *RS*: Pearson's correlation coefficient. The optimal values are shown with bold lettering. The optimization is performed using MLE.

Method	**ME (mm)**	**MAE (mm)**	**MARE**	**RMSE**	**RMSRE (mm)**	**RS**
$\mathrm{GAH20}_{\mathrm{S}_{\mathrm{Expo}}}$	1.04	6.83	0.13	9.62	0.18	0.92
$\mathrm{GAH20}_{\mathrm{S}_{\mathrm{SSRF}}}$	2.88	10.51	0.24	14.42	0.36	0.77
$\mathrm{GAH20}_{\mathrm{Expo}}$	1.15	6.89	0.14	9.69	0.19	0.91
$\mathrm{OK}_{\mathrm{Expo}}$	−0.09	6.84	0.15	9.32	0.22	0.92
$\mathrm{GAH35}_{\mathrm{S}_{\mathrm{Expo}}}$	1.05	6.78	0.13	9.63	0.18	0.92
$\mathrm{GAH35}_{\mathrm{S}_{\mathrm{SSRF}}}$	2.86	10.34	0.24	14.38	0.35	0.77
$\mathrm{OK}_{\mathrm{SSRF}}$	**−0.08**	**5.36**	0.14	**7.19**	0.22	**0.94**
$\mathrm{GAH20}_{\mathrm{SSRF}}$	0.50	6.52	0.13	9.39	0.18	0.93
$\mathrm{GAH35}_{\mathrm{SSRF}}$	0.48	6.49	**0.13**	9.44	**0.17**	0.93
$\mathrm{GAH35}_{\mathrm{Expo}}$	1.15	6.85	0.13	9.71	0.18	0.91

Table 5.4: The average value of the LOO-CV measures of monthly precipitation based on 10 different scenarios. The mean value of the following validation measures are shown: ME: mean error (bias); MAE: mean absolute error; MARE: mean absolute relative error; RMSE: root mean square error; RMSRE: root mean square relative error; *RS*: Pearson's correlation coefficient. The optimal values are shown with bold lettering. The optimization is performed using WLS.

Method	**ME (mm)**	**MAE (mm)**	**MARE**	**RMSE**	**RMSRE (mm)**	**RS**
GAH20$_{\mathrm{S_{Expo}}}$	1.29	7.00	0.14	9.87	0.19	0.91
GAH20$_{\mathrm{S_{SSRF}}}$	1.71	8.59	0.18	11.90	0.24	0.88
GAH20$_{\mathrm{Expo}}$	1.26	6.96	0.14	9.83	0.19	0.91
OK$_{\mathrm{Expo}}$	**−0.09**	6.98	0.16	9.49	0.23	0.92
GAH35$_{\mathrm{S_{Expo}}}$	1.28	6.96	0.14	9.88	0.19	0.91
GAH35$_{\mathrm{S_{SSRF}}}$	1.70	8.59	0.18	11.97	0.19	0.91
OK$_{\mathrm{SSRF}}$	−0.14	**5.58**	0.14	**7.52**	0.21	0.93
GAH20$_{\mathrm{SSRF}}$	0.48	6.21	0.12	9.01	0.17	0.94
GAH35$_{\mathrm{SSRF}}$	0.46	6.17	**0.12**	9.05	**0.17**	**0.94**
GAH35$_{\mathrm{Expo}}$	1.26	6.92	0.13	9.85	0.19	0.91

5.7 Discussion and Conclusions

Kriging methods or Gaussian processes (GP) are commonly used for the estimation of spatial and spatiotemporal data but they rely on the Gaussian assumption in order to give representative values for the prediction variance. In the cases of Gaussian distributed data the methodology can be applied as is to the observations. However, when the data exhibit non-Gaussianity nonlinear transformations should be applied prior to the application of the GP. In the field of geostatistics such techniques are known as Gaussian anamorphosis [Chilès and Delfiner, 2012; Wackernagel, 2003], while in the machine learning field they are known as Gaussian process warping [Agou et al., 2022; Peters et al., 2021; Snelson et al., 2004]. Non-linear transforms such as normal scores, logarithm, Box-Cox, hyperbolic tangent, or Hermite polynomials can be used to achieve the warping transformation [Chilès and Delfiner, 2012; Hristopulos, 2020; Peters et al., 2021; Snelson et al., 2004; Xu and Genton, 2017]. In this case study we use the Hermite polynomials to estimate the transform from the data-based CDF to a Gaussian CDF.

For non-parametric GP transformations the warping function adjusts to the characteristics of the data set at hand, providing higher flexibility rather than being determined by a closed-form expression. In our investigation we resulted in comparable but not improved approximation accuracy with the use of the GAH compared to the OK.

Our data correspond to monthly ERA5 precipitation products acquired for a grid covering the island of Crete located in the southeastern Mediterranean Sea, for a period of 492 consecutive months (from January 1979 to December 2019). We focus our investigation to the wet period months which includes 246 months (from October to March for all the years). Reanalysis data have been proven very valuable in areas where environmental monitoring systems are sparse.

We combine GAH transformation, and bootstrap simulations with Kriging prediction to estimate the prediction accuracy of monthly precipitation reanalysis (ERA5) data for the island of Crete. The covariance model parameters are estimated via variogram modeling and the optimization is carried out by two different ways. The hyperparameters are estimated by minimizing the weighted least squared distances (WLS) of the empirical to the model variogram or by maximizing the likelihood (MLE). In general, the WLS is less robust and less computationally intensive than the MLE. The $\mathcal{O}(N^3)$ computational cost of inverting the covariance for the MLE optimization especially for large data sets may be prohibitive. In our case studies the results based on the WLS and the MLE are very similar in terms of the LOOCV results.

The comparison of the cross-validation measures for all the implemented scenarios show that the OK method (S4, S7) has lower bias when applied solely than in combination with any other methodology (S1-S3, S5-S6, S8-S10). Regarding other measures, such as the MARE and the RMSRE, the use of GAH with up to 35 degrees combined with the Spartan variogram model gave the optimal results. Interestingly, the bootstrap simulations did not improve drastically the results. Comparing S1 to S3 (same configuration but S1 also includes simulations) with the MLE optimization method, we see that with the incorporation of simulations the validation measures improved imperceptibly, while with the WLS they deteriorated slightly. The same conclusion is drawn for S5 and S10. On the other hand, with the Spartan covariance model, the comparison of S2 and S8 (or S6 and S9) shows that the simulations significantly worsen the validation measures. The Spartan covariance kernel is found to be more appropriate for the cases with the anamorphosis but without the simulations, while the exponential kernel is found to be more suitable for the scenarios where we integrated the bootstrap simulations.

Several directions can be pursued for future research of the proposed methodology. Initially, incorporating a trend function can filter out the effect of the altitude on precipitation. Alternatively or in addition to the above, the omnidirectional variogram can be substituted by the anisotropic variogram. Agou et al. [2019] showed that the island of Crete is characterized by spatial precipitation patterns that differentiate from West to East and from North to South. Those extreme patterns are not that prominent in the reanalysis data. Another direction is to treat the entire data set in the space-time continuum. In that case the distances have to be readjusted to take into account the non-constant time step, and then the kernels have to be constructed in a way that they encapsulate the spatiotemporal correlations. To avoid the high computational cost of the covariance matrix inversion, the stochastic local interaction model (SLI) [Hristopulos, 2020; Hristopulos and Agou, 2020; Hristopulos et al., 2021] can be used instead of kriging. SLI employs sparse precision matrices to represent space-time correlations in such a way that results in highly sparse matrices, resulting to less computational stress. Another approach is to estimate the probability distribution of a continuously-valued variable, such as precipitation amount, with a kernel-based estimator (KCDE) like the one presented by Pavlides et al. [2022]. This technique avoids the disadvantages of a the step function (empirical CDF) by using the kernel-based estimator which is a continuous function. The KCDE method targets the CDF instead of the PDF [Harrold et al., 2003; Mosthaf and Bárdossy, 2017; Sharma and Lall, 1999] and is presented in their study by means of synthetic data sets and reanalysis precipitation data from the Mediterranean island of Crete (Greece). Lastly, GAH (with or without the KCDE) can be implemented as the first step of the data transformation to the Gaussian distribution in terms of the estimation of drought indices. As mentioned previously in Section 4, the first step for the estimation of the SPI and SPEI is to fit the precipitation values to a parametric distribution, usually the gamma or the lognormal distribution (in our case study the gamma and the Pearson type III distribution) and afterwards the transformation to index values is acquired. Oftentimes, the data do not follow the parametric distribution well, resulting in inaccurate index values. Thus, by fitting a model distribution that adapts to the data, and then transforming those to standardized normal variates (index values), those discrepancies are avoided.

To conclude, we provided an extensive analysis of multiple scenarios for the interpolation of monthly precipitation (based on Ordinary kriging) where some were equipped with Gaussian anamorphosis functions with Hermite polynomials, while others with Monte Carlo simulations. We showed that increasing the polynomials order improve the validation results but slightly, while the incorporation of the simulations gave improved results (compared to the cases without the simulations) only in some cases. The precipitation data sets used here do not follow the Gaussian distribution and based on our investigation they do not follow a specific parametric distribution across the months. We believe that GAH can improve the interpolation results in non-Gaussian data especially in bigger data sets, however further studies are needed for validation.

Chapter 6

Space-time Modeling with Machine Learning Methods

6.1 Summary

In this chapter we present the application of different machine learning classification methodologies for the estimation of hourly precipitation based on 27 meteorological and hydrological auxiliary variables. We use and briefly describe the following twelve classification methodologies: fine, medium, and coarse classification trees, linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian Support Vector Machines, Boosted Ensemble trees, Bagged Ensemble trees, and Ensemble RUSBoosted trees. Ensemble trees (a.k.a. Random Forests) are more extensively presented. The hourly precipitation data are imbalanced, meaning that their classes are very disanalogous to each other, specifically, they contain many low values and very few high values. Thus, we present ways by which classification methodologies are equipped to accurately predict imbalanced data, e.g., using different weights for each class, or applying undersampling and oversampling techniques. Also, we present the metrics that are commonly used to assess the model's performance. In terms of the application, we study the impact of auxiliary variables in a spatiotemporal predictive framework for hourly precipitation. We create eight classes and we split the data set into two different classification scenarios. The first one classifies the occurrence of precipitation while the other classifies the intensity of the precipitation events. We present the results for all the models tested in tabular form in order to compare their performance.

6.2 Introduction

Machine learning (ML) methods were developed in computer science to provide automated algorithms for pattern recognition. ML methods can be classified into three main categories: supervised learning, unsupervised learning, and reinforcement learning.

In Supervised Learning methods the training set has predetermined input features and assigned output labels. The objective of supervised learning involves acquiring a mapping function that can effectively forecast the output label for unobserved inputs. The model is generated through iteratively predicting and correcting until it attains a specified level of accuracy on the training data set. Subcategories of supervised learning include regression (e.g., temperature forecast) and classification methods (e.g., will it rain tomorrow or not). Various algorithms are available for employment, among which are the Linear and Logistic Regression, K-nearest neighbors, Decision trees, Support vector machines (SVMs), Random Forests (RFs), and Artificial Neural Networks [Chase et al., 2022; Rolnick et al., 2022].

On the contrary, the input data in Unsupervised Learning methods are unlabeled and the model is generated by searching structures and patterns present in the input data. Standard algorithms for unsupervised learning include the K-Means clustering and the Principal component analysis [Chase et al., 2022].

Finally, in Reinforcement Learning the training data is a combination of labeled and unlabeled inputs. Reinforcement learning methods are commonly applied when the environment is complex and changes dynamically (e.g., gaming, autonomous driving). Typical algorithms are the Deep Reinforcement learning and the Actor-Critic methods. In this chapter we focus on supervised learning methods.

The supervised ML methods used herein can be divided into three categories: decision trees, support vector machines, and ensemble methods. We used fine, medium, and coarse classification trees, linear, quadratic, and cubic SVMs, fine, medium, and coarse Gaussian SVMs, and Boosted, Bagged, and RUSBoosted trees.

In decision tree classification, a tree-like model is created to classify inputs into output classes based on feature values. The algorithm repeatedly divides a set of input data into smaller groups, relying on the characteristics of the input features to create the model. At every node of the tree structure, the algorithm identifies the attribute that gives the optimal partitioning of the data into distinct categories. The process stops when all the data in a subset are assigned under the same class or the algorithm achieves a fixed stopping criterion (e.g., accuracy, maximum tree depth). The root node symbolizes the original input data, the branches signify the different feature values that can lead to distinct subsets of the data, and the leaves are the output classes that the algorithm has been trained to estimate [Rokach and Maimon, 2014; Wu et al., 2008].

Support vector machines (SVMs) can be used both for classification and regression problems. The SVM maps a set of data into a high-dimension feature space. The simplest SVM model is the linear SVM which involves a straight line. In this case, a straight line is created in this feature space in a way that the width of the gap between the two groups is maximized [Cortes and Vapnik, 1995]. Then new data are mapped in the space between the groups and based on which side of the space they fall in a decision is made. If a cubic curve or a kernel function is used instead of a straight line (linear SVM) to create the hyperplane, then we result in more complex SVM models.

Ensemble learning methods are essentially a combination of other machine learning methods designed to increase the accuracy and decrease the bias of the resulting estimates [Opitz and Maclin, 1999]. The most popular ensemble methods are bagging (or bootstrap aggregation) [Breiman, 1996] and boosting [Freund and Schapire, 1996; Schapire, 1990]. Bagging creates subsamples of the training data to train the model with random data configurations. More about bagging in Section 6.3.2. Boosting creates a series of classifiers and tries to estimate and correct the predictions of the classifiers that resulted in incorrect predictions; thus, repeatedly tries to train the model to create more accurate classifiers [Opitz and Maclin, 1999]. RUSBoost combines data sampling and boosting, and is a good approach when the training data set is imbalanced [Seiffert et al., 2008].

ML methodologies have been applied to different meteorological data. For instance, they have been used for air quality modeling [Keller et al., 2017], heavy rainfall estimation [Moon et al., 2019], drought index estimation [Mokhtar et al., 2021], surface air temperature [Zhu et al., 2019], and data merging [Papacharalampous et al., 2023a,b].

Random Forests is a supervised machine learning algorithm. Despite the advantages of using RFs, the literature is limited in the water resources field. Recently the method has began to gain in popularity. For an extensive list of references refer to Tyralis et al. [2019]. In the following section we will introduce, and provide a brief description of the Random Forests methodology. Additionally, we will apply multiple classification methods to precipitation and we will discuss our findings.

RFs have been applied in various scientific fields, such as in remote sensing [Belgiu and Drăguţ, 2016; Maxwell et al., 2018], bioinformatics [Chen et al., 2011], agriculture [Liakos et al., 2018], biology [Goldstein et al., 2011], and mining [Rodriguez-Galiano et al., 2014]. Theoretical and practical aspects of RFs can be found in the review papers of Biau and Scornet [2016]; Boulesteix et al. [2012]; Criminisi et al. [2012] and Ziegler and König [2014]. The main advantages of RFs are their ability to handle big data, they have the ability to incorporate different types of information, and they require minimal parameter tuning since they can automatically search for identifiable patterns in the data based on simple and intuitive heuristics, making the method suitable for use by non-experts. Furthermore, the method is flexible enough to avoid assumptions that need to be fulfilled when geostatistical methods are applied, such as field stationarity and Gaussianity. In brief, RFs extend decision tree methods by introducing the idea of ensembles of trees (i.e., forests). A classification and regression tree (CART) is a predictive model that divides a target variable into homogeneous groups. However, these models have several weaknesses that prevent their application to real-world problems, mainly because of their lack of stability and tendency to overfitting [Legasa et al., 2022]. To predict the variable Y (in this case study hourly precipitation at a 65 node spatial grid) from the predictors X (in this study a set of 27 reanalysis large-scale variables) ensemble averaging improves the accuracy and uncertainty estimation of simple decision tree methods by combining an ensemble of N CARTs. In water resources, random forests are said to belong to the class of data-driven models (see e.g., Solomatine and Ostfeld [2008]).

6.3 Methodology

Multiple implementations of RFs exist in the literature, however, we will focus on the one introduced by Breiman [2001]. Essentially, what differentiates Breiman's RF-algorithm from other RF implementations is the use of classification and regression trees (CARTs, [Breiman et al., 1984]) as base learners [Biau and Scornet, 2016]. The RFs implementation is sufficiently adaptable to handle both supervised classification and regression tasks. In regression algorithms, the dependent variable is quantitative, whereas in classification algorithms the dependent variable is qualitative. In the latter case, the dependent variable can also be ordered; i.e., the values of the variable are ordered but no metric is defined/used to quantitatively assess the observed differences [Tyralis et al., 2019].

In the following, we briefly describe the way the algorithm works. Before building each tree (N trees in total), k observations are sampled randomly from the initial data. These k observations are considered for the construction of the tree. Then, at each cell of each tree, a partition is performed by maximizing the selected CART criterion. Finally, the construction of each tree stops when each cell contains fewer points than the node size. The final estimate depends only on pre-determined k data points [Biau and Scornet, 2016].

The Matlab platform provides several functions in the Statistics and Machine Learning toolbox which allow for rapid development of ML approaches [MATLAB, 2018].

6.3.1 Variable Importance Metrics & Selection

Random Forests can provide useful information about the importance of the variables. This means that the rank of the importance of each variable indicates the relative significance of the predictor variables in modeling the response variable. The measures of significance that can be used to assess the importance of the predictor variables are the Mean Decrease Impurity (MDI; see [Breiman, 2003]) and the Mean Decrease Accuracy (MDA; see [Breiman, 2001]). The former is based on the total decrease in node impurity from splitting on the variable, averaged over all trees, while the latter is on the idea that rearranging the values of the variable does not influence the prediction accuracy [Biau and Scornet, 2016].

For a review on variable selection refer to the work of Heinze et al. [2018]. In RFs, MDI and MDA are used to conclude variable selection by excluding the variables that resulted in approximately zero importance as non-significant. Díaz-Uriarte and Alvarez de Andrés [2006] offer a stepwise approach where a combination of predictor variables is tested and progressively removed until the lowest error is achieved.

6.3.2 Resampling or Bagging

As mentioned earlier, classification and regression trees (CARTs) are unstable and tend to overfitting [Legasa et al., 2022]. To overcome this problem a technique called bagging is utilized [Breiman, 2001]. Bagging is essentially a resampling mechanism, also known as bootstrap in the statistical literature [Efron, 1982; Politis et al., 1999]. Out-of-bag (OOB) errors are used to tune random forests' parameters. Out-of-bag samples (about one-third of the training set, see Biau and Scornet [2016]) are the samples that remain after bootstrapping the training set. Each tree is formed for the specific sub-sample of the selected data which is called "bagged". The remaining data, which are called "out-of-bag", can then be used to evaluate the tree performance. The preceding approach is similar to the well-known k-fold cross-validation [Hastie et al., 2009; Tyralis et al., 2019].

In Breiman's algorithm, in the resampling step, the tree estimates are calculated by choosing n times from n points with replacement. Bagging is the process of creating several bootstrap samples and averaging the predictions (bootstrap aggregation). When performing classification, the prediction is estimated by the majority class vote from each tree's class vote, while in regression, the prediction is obtained by the average of the predictions of each tree. The bagging of trees and the added randomization is used to reduce the correlation between the trees and, consequently, reduce the variance of the predictions [Tyralis et al., 2019].

According to Breiman [2001] bagging can provide improved accuracy as well as ongoing estimates of the generalized error of the combined ensemble of trees. This is achieved by leaving out one-third of the training set in each bootstrap sample.

By leaving out one-third of the training set in each bootstrap sample, the estimates are based on that subset of classifiers. Because the error decreases if the number of combinations increases, the out-of-bag estimates tend to overestimate the error rate. This means that in order to get unbiased estimates it is crucial to run past the point where the test set error converges [Breiman, 2001]. Yet, unlike cross-validation, where bias is present but the level of it is unclear, the out-of-bag estimations are unbiased.

6.3.3 Imbalanced data

The terminology imbalanced data is unclear, because a data set with data splits different than 100/(the number of classes) % for each class can be considered imbalanced. However, if we have a data set with a 90-10% split, or as it is in our case with a 70-11-6-7-4-2~0~0 split it seems obvious that this is an imbalanced set (this split considers the classes presented in Table 6.3 and the respective pie chart is presented in Fig. 6.1). Classification algorithms, which tend to be biased to the majority class, return inferior results when an imbalanced set is considered.

In real life, many classification problems are imbalanced. Most commonly the aim of those classification problems is to correctly classify the minority class. Such examples include fraud detection [Fawcett and Provost, 1997], rare disease diagnosing [Jabeen et al., 2022], and environmental disasters [Kubat et al., 1998].

A few techniques have been developed to deal with an imbalanced data set. They include two main approaches: the sampling method and the cost effective method. The cost effective approaches work by assigning a high cost to misclassification of the minority class, and trying to minimize the overall cost [Domingos, 1999; Pazzani et al., 1994]. The sampling methods include the undersampling and the oversampling techniques. For example, SHRINK is a system for the undersampling approach developed by Kubat et al. [1997] which labels a mixed region as minority class (regardless if it is) and then it searches for the minority class [Chen et al., 2004]. Undersampling has one main advantage, it reduces the size of the problem. Both methods aim to create a more balanced data set. In the oversampling approach, the sample is enriched with synthetic new data for the minority class or classes, while in the former (provided that the data set is big enough) values from the majority class are removed. Oversampling must be applied very carefully to avoid the risk of overfitting or incorporating noise in the data set, however, oversampling does not increase information but it increases the weight of the minority class. A combination of both oversampling and downsampling has also been used to improve classification performance. One oversampling algorithm that exists in the bibliography is the Synthetic Minority Oversampling Technique (SMOTE) [Chawla et al., 2002], which generates new synthetic points for the minority class between existing ones based on local densities and boundaries of other classes. The algorithm works as follows for each minority sample:

1. Find the k-nearest neighbors for the sample in the minority class.

2. Randomly select j neighbors (this number can be adapted and depends on the amount of oversampling required).

3. Randomly generate synthetic values along the line joining the original sample value and its j neighbors.

One main disadvantage of the SMOTE technique is overgeneralization because it generalizes the minority domain irrespective to the majority class and can return inconclusive and insufficient results. This is especially true when the distribution of the data is extremely skewed, where the minority class is sparse enough that presents high probability of mixing with the majority class. In our case the SMOTE technique cannot work sufficiently because the minority class is so rare that the probability that it is mixed into the majority class is extremely high.

Herein, we will apply the random Forest (RF) algorithm [Breiman, 2001] without balancing the data set and afterwards we will apply several other methods to improve and compare our results.

6.3.4 Performance Measurement

The classification performance can be assessed by a variety of performance measures, however, the accuracy of the model is the most prevalent. In the cases of imbalanced data classification, accuracy is frequently inappropriate measure for the model's success. For instance, even a trivial classifier can achieve very high accuracy if it predicts every case as the majority class. To evaluate the efficacy of the learning algorithms on imbalanced data, we utilize metrics such as precision, true positive rate (Acc+ or recall), true negative rate (Acc-), F-measure, and G-mean. These metrics have been widely used for comparison. All the metrics are functions of the confusion matrix (Tables 6.1 and 6.2) as shown in the following equations. The rows of the confusion matrix are actual classes, and the columns are the predicted classes. Based on the confusion matrix, the performance metrics are defined as follows:

$$\text{True Negative Rate } (\text{Acc}^{-}) = \frac{\text{TN}}{\text{TN} + \text{FP}} = 1 - \text{FPR},$$

$$\text{True Positive Rate } (\text{Acc}^{+}) = \frac{\text{TP}}{\text{TP} + \text{FN}} = 1 - \text{FNR} = \text{Recall},$$

$$\text{G-mean} = \left(\text{Acc}^{-} \times \text{Acc}^{+}\right)^{1/2},$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}},$$

$$\text{Accuracy of Single} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN}},$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = 1 - \text{FDR} = \text{Positive Predictive Value (PPV)},$$

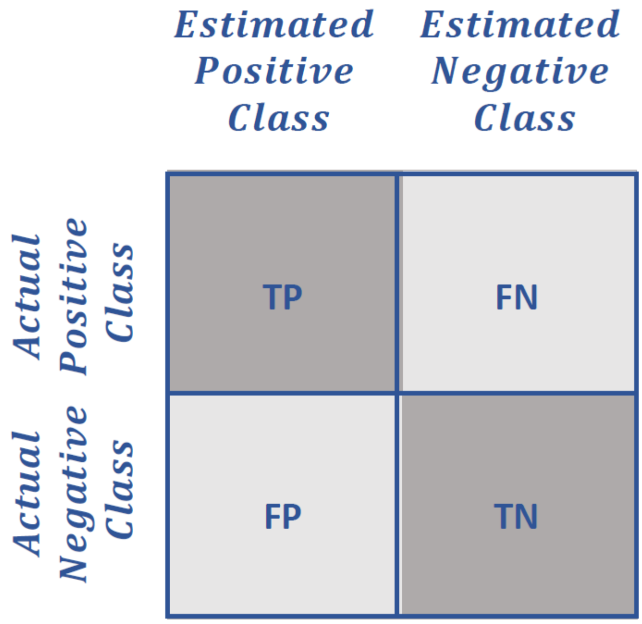
$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \text{F1-score},$$

$$\text{False Negative Rate (FNR)} = \frac{\text{FN}}{\text{FN} + \text{TP}} = 1 - \text{TPR},$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR},$$

$$\text{False Discovery Rate (FDR)} = \frac{\text{FP}}{\text{FP} + \text{TP}} = 1 - \text{PPV}.$$

Table 6.1: Confusion matrix for binary classification.



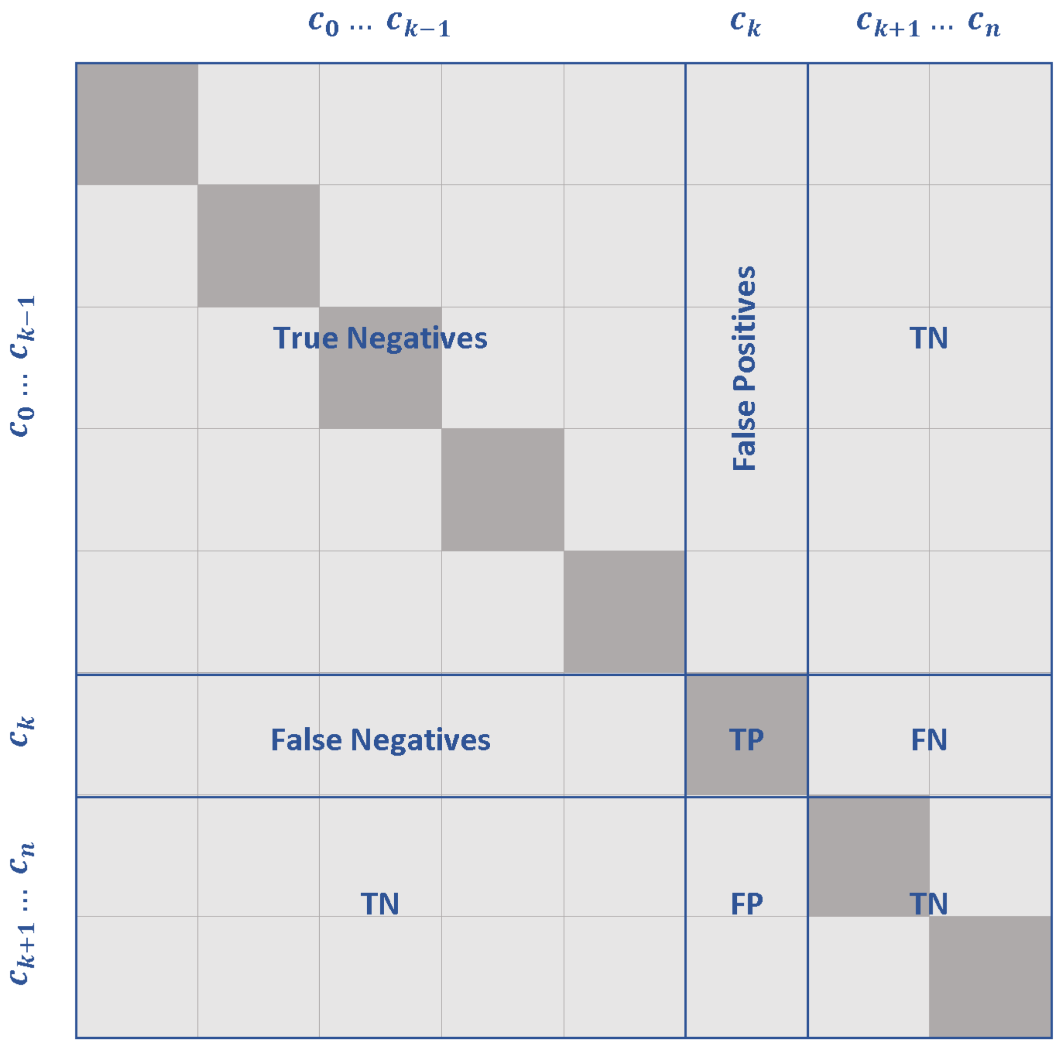
	Estimated Positive Class	Estimated Negative Class
Actual Positive Class	TP	FN
Actual Negative Class	FP	TN

6.4 Data Preparation and Pre-processing

The initial data set includes the variables presented in Table 3.1 for the entire grid shown in Fig. 3.1. According to multiple publications and organizations [American Meteorological society, 2022; Barthiban et al., 2012; United Kingdom Meteorological Office, 2022], precipitation intensity varies across the globe. For example in the US rainfall intensity between 2.6 and 7.6 mm/hr is classified as moderate rain, however, in the UK moderate rain is considered rainfall intensity of 0.5 to 4 mm/hr. Furthermore, we have noticed that in many cases, intensity of 0.5 mm/hr and lower is classified as no rain [Avanzato and Beritelli, 2020]. Another way to classify precipitation is by the percentile ranges [Kant, 2018]. In our case, this was not ideal since hourly precipitation data set for the area of interest contains a very high percentage of zero values. However, the application of the percentiles would have resulted in the following classification which is fairly similar to the one that we used and is presented in Table 6.3:

1. Very light spell: 0–50 % $\Longrightarrow$ (0, 0.00095367) mm,

Table 6.2: Confusion matrix for multi-class classification.



2. Light spell: 50–75 % $\Longrightarrow$ (0.00095367, 0.063896) mm,
3. Moderate spell: 75–90 % $\Longrightarrow$ (0.063896, 0.21648) mm,
4. Intense spell: 90–95 % $\Longrightarrow$ (0.21648, 0.52834) mm,
5. Very intense spell: 95–97.5 % $\Longrightarrow$ (0.52834, 3.022) mm,
6. Extremely intense spell: 97.5–99.9% $\Longrightarrow$ (3.022, 18.836) mm.

We classify the hourly precipitation in eight different classes based on various bibliographies [Kittredge, 1948; Lull, 1959; Meteoclub, 2013], making sure that the ranges are appropriate for the sample data. The classes are defined in Table 6.3.

Table 6.3: Definition of the hourly precipitation classes used. The numbers within parentheses in the first columns correspond to the class labels. LB and UB stand for the lower bound and upper bound of the respective class.

Precipitation class	LB (mm/hr)	UB (mm/hr)
No Rain (1)	0	0
Fog (2)	0	0.0127
Mist (3)	0.0127	0.0508
Drizzle (4)	0.0508	0.254
Light Rain (5)	0.254	1.016
Moderate Rain (6)	1.016	3.81
Heavy Rain (7)	3.81	15.24
Excessive Rain (8)	15.24	40.64

6.4.1 Prediction and Response Variables

In particular, our response variable is hourly precipitation, which is a semi-continuous variable with a probability distribution characterized by a spike at zero (dry days) followed by a continuous distribution with positive support (wet days). In our case study, we treat precipitation as a categorical ordinal variable according to Table 6.3. We opted for a categorical variable to cover for the big range in our data and the very asymmetrical distribution that the hourly precipitation data exhibit. Initially, all the variables in Table 3.1 besides precipitation are treated as the prediction variables.

We use the following twenty seven features as prediction variables: (i) Latitude of the target point expressed in degrees in the World Geodetic System (WGS 84), (ii) Longitude of the target point expressed in degrees in the World Geodetic System (WGS 84), (iii) timestamp in Matlab time (hours), (iv) 10 metre U wind component in m/s, (v) 10 metre V wind component in m/s, (vi) 2 metre dewpoint temperature in degrees Celsius, (vii) 2 metre temperature in degrees Celsius, (viii) evaporation in m of water equivalent, (ix) mean sea level pressure in Pa, (x) runoff in meters, (xi) sea surface temperature in degrees Celsius, (xii) snow density in kg/m^3, (xiii) snow depth in m of water equivalent, (xiv) snowfall in m of water equivalent, (xv) soil temperature level 1 in degrees Celsius, (xvi) surface latent heat flux in W/m^2, (xvii) surface sensible heat flux in W/m^2, (xviii) surface net solar radiation in W/m^2, (xix) surface net thermal radiation in W/m^2, (xx) surface solar radiation downward clear-sky in W/m^2, (xxi) surface solar radiation downwards in W/m^2, (xxii) surface thermal radiation downward clear-sky in W/m^2, (xxiii) surface thermal radiation downwards in W/m^2, (xxiv) top net solar radiation in W/m^2, (xxv) top net thermal radiation in W/m^2, (xxvi) total cloud cover in (0-1), and (xxvii) months in (1-12). The values of the latter variable are generated, they are not included in the initial data sets and they are treated as categorical.

6.4.2 Exploratory Data Analysis

We classify the observed precipitation classes as shown in Table 6.3. The distribution of the precipitation data per class is shown in the pie chart of Fig. 6.1. The spatio-temporal distribution of the precipitation classes is shown in Fig. 6.3. As mentioned previously and also shown in Fig. 6.1, the most common class (No Rain: 0 mm) is measured for ≈ 70%, while the most extreme class (above 15.24 mm) is observed for only ≈ 0.0003% (two events) of the available data.

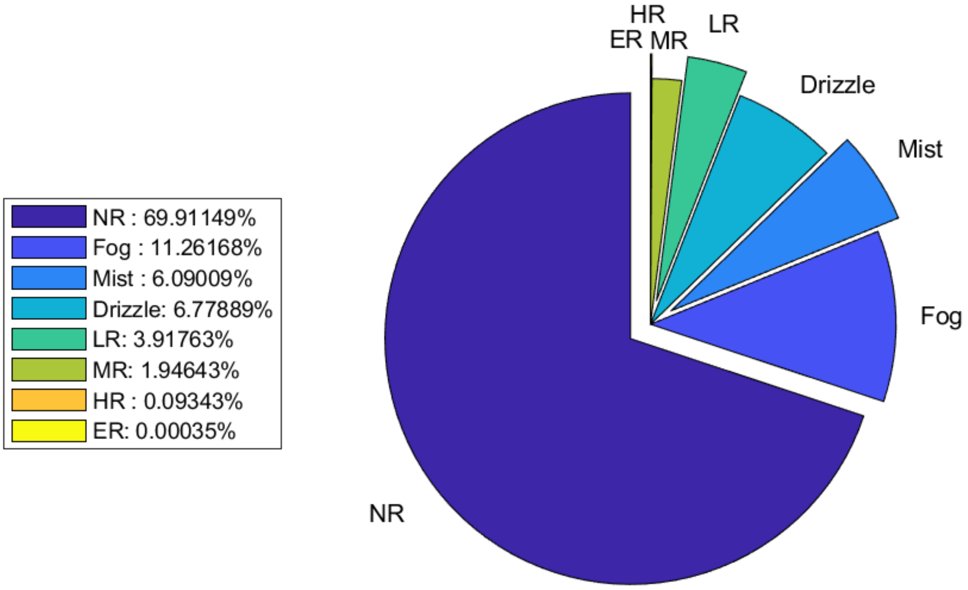


Figure 6.1: Pie chart of precipitation classes according to the thresholds defined in Table 6.3. The hourly values correspond to a time period of one year spanning from 01-Jan-2019 00:00:00 to 31-Dec-2019 23:00:00. NR: No Rain; F: Fog; M: Mist; D: Drizzle; LR: Light Rain; MR: Moderate Rain; HR: Heavy Rain; ER: Excessive Rain.

The spatial distribution of the precipitation classes depending on the month of the year is presented in Fig. 6.4. The percentages of each class in the entire data set are illustrated in Fig. 6.1. In particular, No Rain corresponds to ≈ 69.9% of the data, Fog to ≈ 11.3%, Mist to ≈ 6.1%, Drizzle to ≈ 6.8%, Light Rain to ≈ 3.9%, Moderate Rain to ≈ 1.9%, and Heavy Rain to ≈ 0.1% and Excessive Rain to ≈ 0%. Specifically, across the months of the year, No Rain presents a pick frequency in August, Fog in December, Mist, Drizzle, and Light Rain in January, Moderate Rain in February and Heavy Rain, and Excessive Rain in November. Excessive Rain has only two events and both are recorded in November. The lowest frequency for No Rain is observed in January, for Fog, Mist, Drizzle, Light Rain, and Moderate Rain in August, while Heavy rain does not have any events in the months between May to October. It follows that, the majority of the data with rain above zero are distributed with the highest frequencies in the winter months and the lowest during the summer months (see Fig. 6.4). Provided the predominance of the majority class (69.9%), the algorithm may default to predicting the majority class. The algorithm can maximize its accuracy by randomly predicting which majority class will occur each time. This is a trivial result and provides near-zero predictive power.

6.5 Data Analysis and Results

We build Random Forests using the predictor variables defined is Section 6.4.1. We consider the hourly precipitation class as the response variable. We use the Matlab `fitensemble` function to construct the Random Forest structure. We apply Random Forests (RFs) with 100 trees. The prediction variables are complete for the entire grid.

We randomly split the data into training and test sets of 70-30 percent size respectively. The classification error is the percentage of sites whose precipitation class is wrongly classified. Since we have eight classes, a completely random guess has a probability of success equal to 12.5%, implying an 87.5% classification error. We aim to improve on these odds using the Random Forest methodology. Indeed, the classification error of the Random Forest (i.e., the percentage of misclassified classes) is shown in Fig. 6.2 versus the number of trees in the forest. The classification error for the test set (blue line) is almost 14% if 100 are used, and it is very close to this figure even with as few as 50 trees. The classification error can also be estimated based on the samples that are omitted from each tree, averaged over all trees in the forest. This so-called out-of-bag error (red line) also converges to the same value when we incorporate more than 100 trees, albeit somewhat slower than the test-based error.

The classification error is global error measure that does not inform about the classification performance for the different classes individually. The confusion matrix $C_{i,j}$ is a useful measure of prediction accuracy which resolves such differences. The confusion matrix is formulated in terms of the values of the precipitation classes in the test set. If there are G different precipitation classes

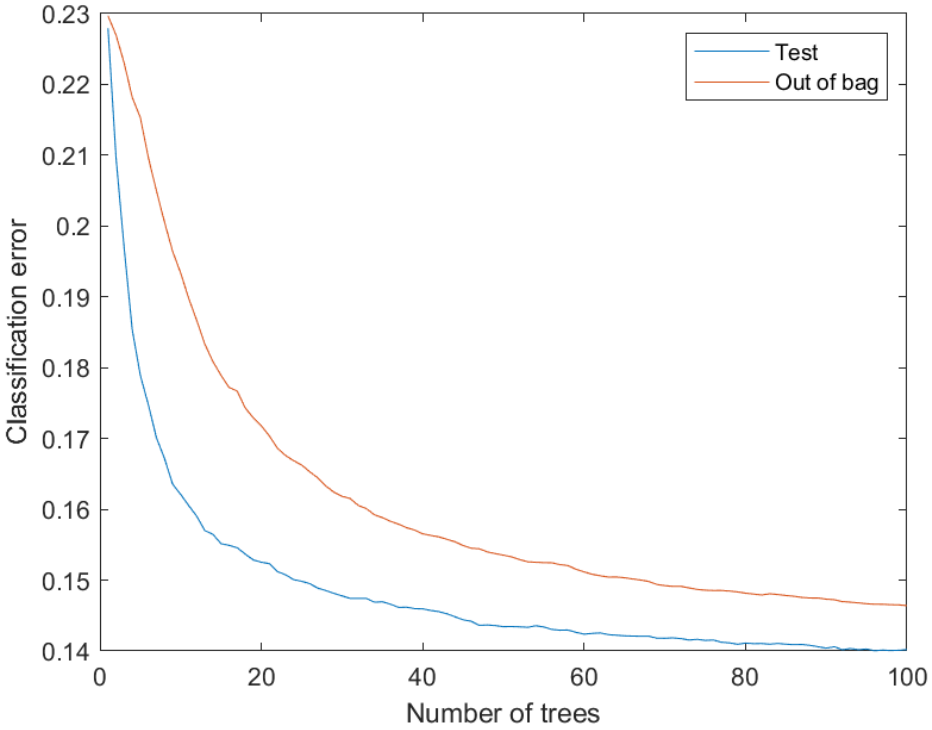
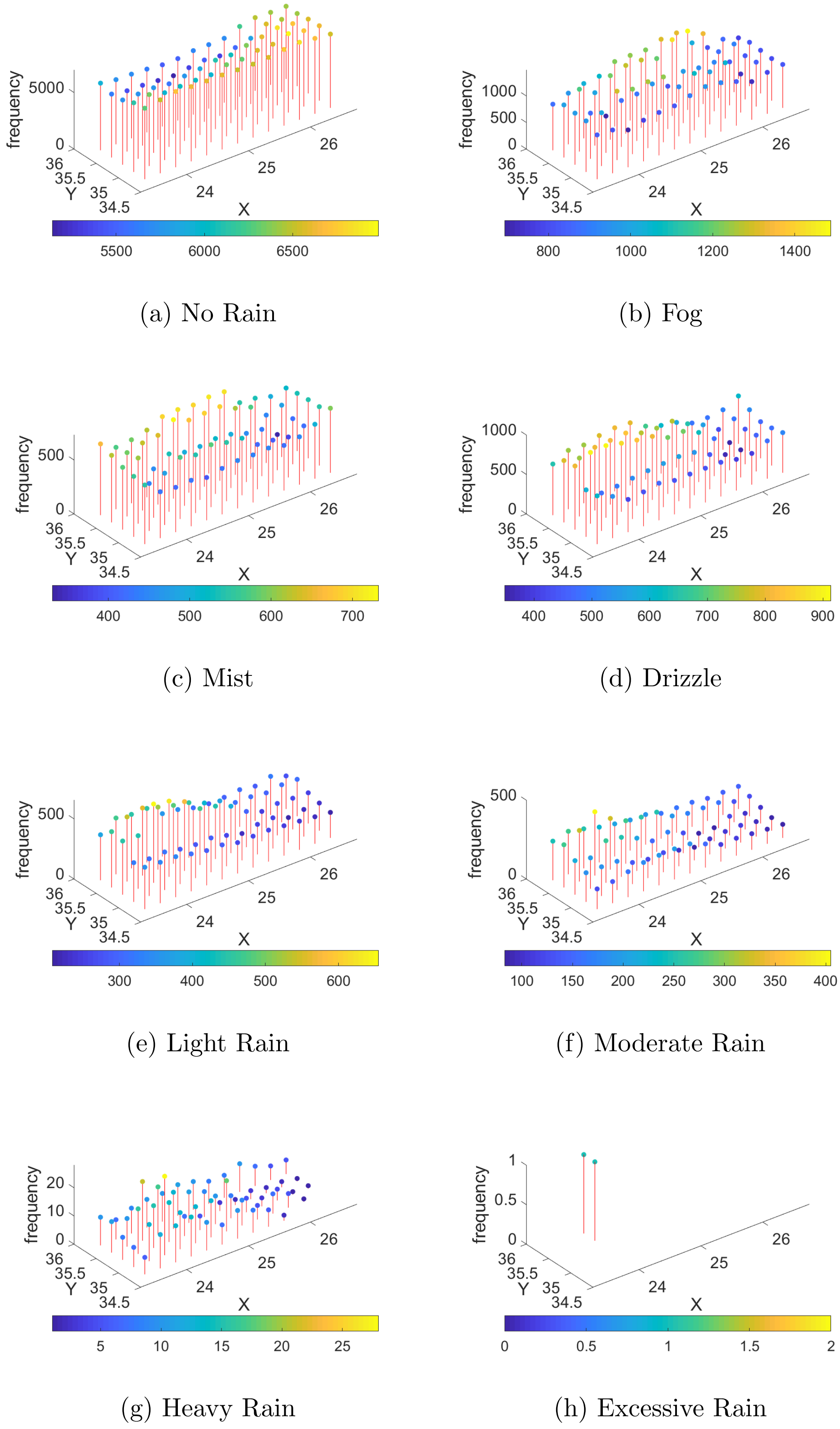


Figure 6.2: Classification error of precipitation classes as function of the number of Random Forest trees. The error is calculated using (i) the values in the test set and (ii) the out-of-bag samples.

(in our case $G = 8$), $[C_{i,j}]$ is a square $G \times G$ matrix, such that $C_{i,j}$ is equal to the number of times that precipitation with values belonging in class i is classified by the Random Forest as belonging to the class j. Ideally, the confusion matrix should be diagonal meaning that all the sites are correctly classified. This, however, does not happen in practice, leading to a leaking effect from the diagonal to off-diagonal entries which indicates a "confused" classifier. In the case of our data set, the confusion matrix is shown in Fig. 6.5.

The side sub-matrix, (columns) of Figs. 6.5, 6.8 and 6.10 display the number of correctly and incorrectly classified observations for each predicted class as percentages of the number of observations of the corresponding predicted class. The bottom sub-matrix (rows) represents the number of correctly and incorrectly classified observations for each true class as percentages of the number of observations



(a) No Rain

(b) Fog

(c) Mist

(d) Drizzle

(e) Light Rain

(f) Moderate Rain

(g) Heavy Rain

(h) Excessive Rain

Figure 6.3: Spatio-temporal distribution of the eight different classes of precipitation.

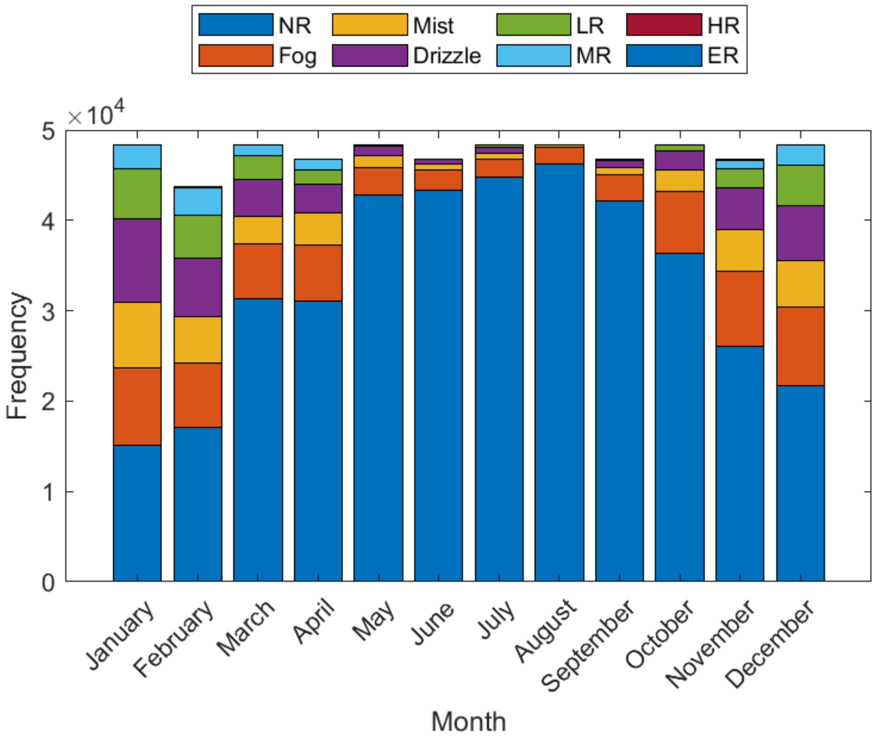


Figure 6.4: Bar graph showing the distribution of the eight different classes of precipitation classified by month.

of the corresponding true class [MATLAB, 2018].

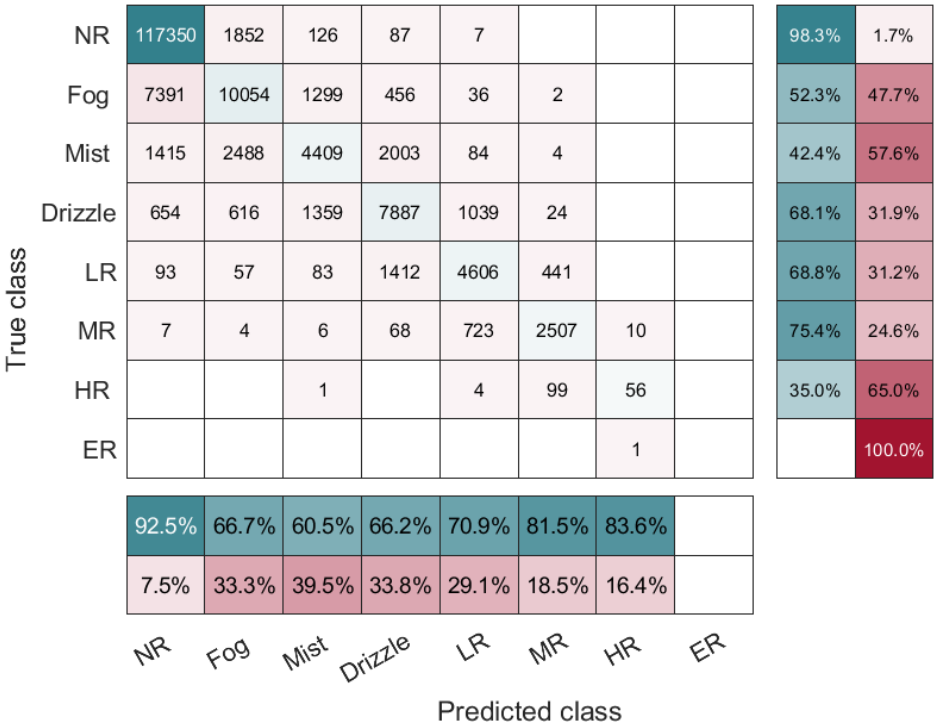
The classifier accurately predicts the lower class (98.3% accuracy), which is expected since the majority of the data falls under that class. Nevertheless, identifying the locations of the class with no precipitation provides insight for the dryness of the set. For the classes Fog and Mist the classification is poor but improved if it was performed by chance (87.5%), resulting in 52.3% and 42.4% accuracy accordingly. The classification error for the next three classes (Drizzle, Light Rain and Moderate Rain) are somewhat improved (≈ 70%) . For the highest precipitation class (above 15.24 mm), the classification works poorly misclassifying the values from the test set (only one event). For the Heavy Rain and the Excessive Rain class, the classification accuracy deteriorates significantly, with classification errors 65% and 100% respectively. This is not surprising, since the Heavy Rain and Excessive Rain classes occur very rarely (see the pie plot

Table 6.4: Classification performance comparison on the hourly precipitation data for the year 2019. The measures are described in Section 6.3.4. The predictor variable "month" is treated as categorical variable, as well as the response variable. Eight classification classes are used and their ranges are presented in Table 6.3. NA: not applicable; the respective measures involve division by zero or by a very small number.

Method	**Precision**	**Acc+(Recall)**	**Acc-**	**F-measure**	**G-mean**	**Accuracy**
Model performance						
Bagged Ensemble of 100 Trees	62.2	55.0	96.5	58.5	72.9	86.0
Performance metrics per class						
Class 1 - No Rain	92.5	98.3	81.4	95.3	89.4	98.3
Class 2 - Fog	66.7	52.3	96.7	58.6	71.1	52.3
Class 3 - Mist	60.5	42.4	98.2	49.9	64.5	42.4
Class 4 - Drizzle	66.2	68.1	97.5	67.1	81.5	68.1
Class 5 - Light Rain	70.9	68.8	98.8	69.8	82.5	68.8
Class 6 - Moderate Rain	81.5	75.4	99.7	78.3	86.7	75.4
Class 7 - Heavy Rain	83.6	35.0	100	49.3	59.2	35.0
Class 8 - Excessive Rain	NA	0	100	NA	0	0

in Fig. 6.1), thus making it very difficult to accurately train the classifier. If the percentage of the classes in the train set was higher, the classification would have worked better. This is one of the reasons that there is a separate field of studies that focuses on those percentages alone, and developing methods that can estimate those outliers better. It is important to point out that our data set is extremely imbalanced. Those data sets are very rarely modeled correctly and in the next step we will split the data set to two different case studies and we will further apply multiple classification methods to try recuperate this issue. The performance metrics for the model are also presented in Table 6.4 where we additionally present the metrics per class. The accuracy of the first (No Rain) class is very high, however the rest of the classes are poorly classified.

Random forests also allow us to estimate the relative importance of the different predictors. This is based on the idea of surrogate splits which is common in CART. The relative predictor importance for our data is shown in Fig. 6.6. The most important predictors are those associated with the mechanisms that water



	NR	Fog	Mist	Drizzle	LR	MR	HR	ER		
NR	117350	1852	126	87	7				98.3%	1.7%
Fog	7391	10054	1299	456	36	2			52.3%	47.7%
Mist	1415	2488	4409	2003	84	4			42.4%	57.6%
Drizzle	654	616	1359	7887	1039	24			68.1%	31.9%
LR	93	57	83	1412	4606	441			68.8%	31.2%
MR	7	4	6	68	723	2507	10		75.4%	24.6%
HR			1		4	99	56		35.0%	65.0%
ER							1			100.0%
	92.5%	66.7%	60.5%	66.2%	70.9%	81.5%	83.6%			
	7.5%	33.3%	39.5%	33.8%	29.1%	18.5%	16.4%			

True class
Predicted class

Figure 6.5: Confusion matrix for the Random Forest constructed from the ERA5 data over the island of Crete for the year 2019. The rows indicate reanalysis values from the test set while the columns correspond to predictions. The row entries show how the predictions of the respective class are distributed among different classes. The upper left table corresponds to the confusion matrix values. In the upper right sub-matrix, the first column corresponds to the sensitivity (Acc+) and the specificity (Acc-) values, while the second column to the FNR and FPR (the miss rate per class). In the lower sub-matrix, the first row corresponds to the Precision per class, while the second row to the FDR and FOR (the false discovery and omission rates).

evaporates from the earth into different layers in the atmosphere (total cloud, net thermal, temperature and runoff.) In addition to the relative importance, the use of surrogate splits allows the estimation of the association between the different predictor variables. Higher values of association between two predictors imply that including both variables in the predictor set may be redundant.

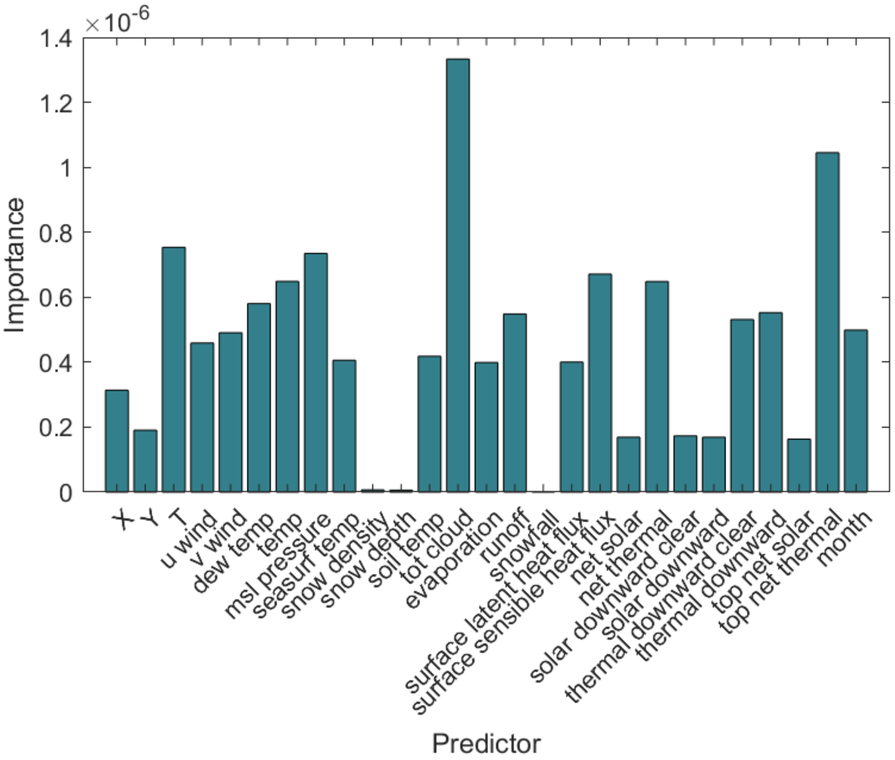
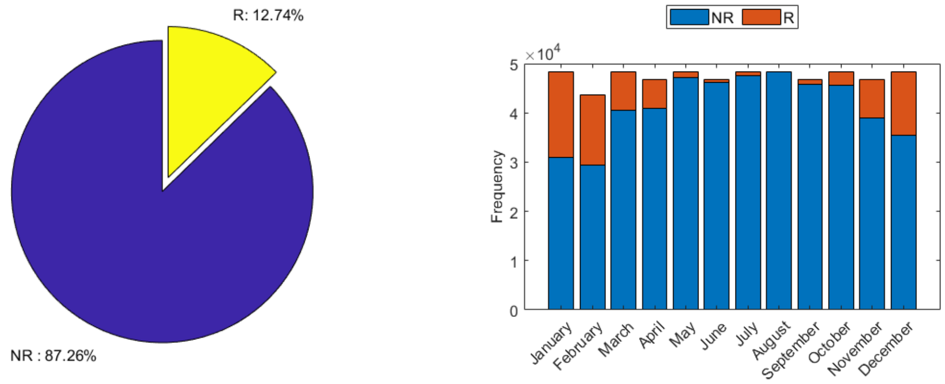


Figure 6.6: Graph of relative predictor importance for precipitation regression using the Random Forest method. The following abbreviations are used. X, Y: space coordinates, T: time. The rest as shown in the labels.

We split the entire data set to two different sets. We will call the first one the “Binary” data set and the second one the “Only Rain” data set. The “Binary” set characterizes all the values that fall below the set threshold (0.0508 mm/hr) as ’No Rain’, which corresponds to the values of the first 3 classes presented in Table 6.3, while the rest of the values are characterized as ’Rain’. The second data set includes only the values above the threshold, meaning that the “Only Rain” set includes only the values from the classes 4-8.

The distribution of the precipitation data per class for the two new sets is shown in the pie charts in Figs. 6.7a and 6.9a, while the histogram of the temporal distribution of the precipitation classes is shown in Figs. 6.7b and 6.9b. The "Binary" set includes 569 400 hourly values, while the "Only Rain" set includes 72 523 values. For both data sets we apply multiple classification methodologies and we present their classification measures in the Tables 6.5 and 6.6. We additionally present the confusion matrix for the model that resulted in the best performance according to the accuracy classification metric.



(a) Pie chart "Binary" 2019

(b) Histogram per month, "Binary" 2019

Figure 6.7: Pie chart and histogram of the temporal distribution of the two different classes of precipitation for the "Binary" data set. The hourly values correspond to a time period of one year spanning from 01-Jan-2019 00:00:00 to 31-Dec-2019 23:00:00. NR: No Rain corresponds to the values below the threshold (0.0508 mm/hr); R: Rain corresponds to the values above the threshold.

6.5.1 "Binary" data set

In the case of the "Binary" data set ($\approx 570\ \ 000$ values per variable), the methods used and their classification measures are presented in Table 6.5. Since we have two classes, a completely random guess has a probability of success equal to 50%, implying an 50% classification error. In order to decide which of the models is the best we will list the models that resulted in the best and the second best value for each metric. According to the precision metric, the best model is the Fine SVM Cross-Validation 5 folds (99.0%), followed by the Bagged Ensemble Trees Cross-

Validation 5 folds (98.9%). For the recall, the best model is the Bagged Ensemble Trees Cross-Validation Holdout 30% (98.8%), followed by the Linear SVM Holdout 30% (98.4%). For the specificity (Acc-), the best model is the My Ensemble RUSBoosted 500 Trees Holdout 50% (92.6%), followed by the same model with 1000 trees (92.5%). For this metric we see bigger differences between the models starting from 34.7% with the Cubic SVM Cross-Validation 5 folds. According to the F-measure, the best model is the Bagged Ensemble Trees Cross-Validation 5 folds and Holdout 30% (both at 98.0%), followed by the Bagged Ensemble Trees Cross-Validation Holdout 30% (98.0%). For the G-mean values, the best model is the Bagged Ensemble Trees Cross-Validation 5 folds (94.1%), followed by the My Ensemble RUSBoosted 500 Trees Holdout 50% (93.6%). Similarly to the Acc-, we observe big differences for the G-mean values between the models (minimum 56.3% for the Cubic SVM Cross-Validation 5 folds). Lastly, based on the accuracy metric, the best model is the Bagged Ensemble Trees Cross-Validation 5 folds (96.5%), followed by the Bagged Ensemble Trees Holdout 30% (96.4%).

We will consider the Bagged Ensemble Trees Cross-Validation 5 folds as the best model because it has the best classification measures according to 3 metrics (F-measure, G-mean and accuracy) and the second best according to another metric (Precision). As the second best model we will consider the model Bagged Ensemble Trees Holdout 30% because it has the best metrics for Acc- and the second best for F-measure and accuracy. From all the models tested, the best models are those that utilize the random forests methodology.

The confusion matrix of the best model (Bagged Ensemble Trees Cross-Validation 5 folds) according to the accuracy measure (see Table 6.5) is shown in Fig. 6.8. The accuracy of the model is 96.5%, which is definitely an improvement of the 50-50% classification accuracy by chance. In terms of the accuracy, all of the models performed well. To evaluate the overall performance of the model, accuracy is a good starting point but the rest of the metrics have to be considered. The sensitivity (TPR or Acc+ or Recall or sensitivity of single) value of 97.2% means that out of all of the values that were actually NR (No Rain) tested as NR. If we look at the specificity (TNR or Acc-) value of 91.1% we know that out of all the samples that were R (Rain) actually tested as R. In general, when it comes to sensitivity and specificity it is important to have a balance between the two values. In our model, we consider that the values are balanced. F-measure is commonly used to evaluate how balanced is precision and recall because it takes into account both in the calculation, however, in some cases if there is a need to prioritize one over the other, the F-measure should also be weighted. G-mean also provides insight into the classification performance. It is a measure that shows how balanced is the classification performance regarding both the majority and the minority class. Additionally, it is a indicator of the possibility of overfitting, meaning that low G-mean indicates that the model "prefers" one class over the other.

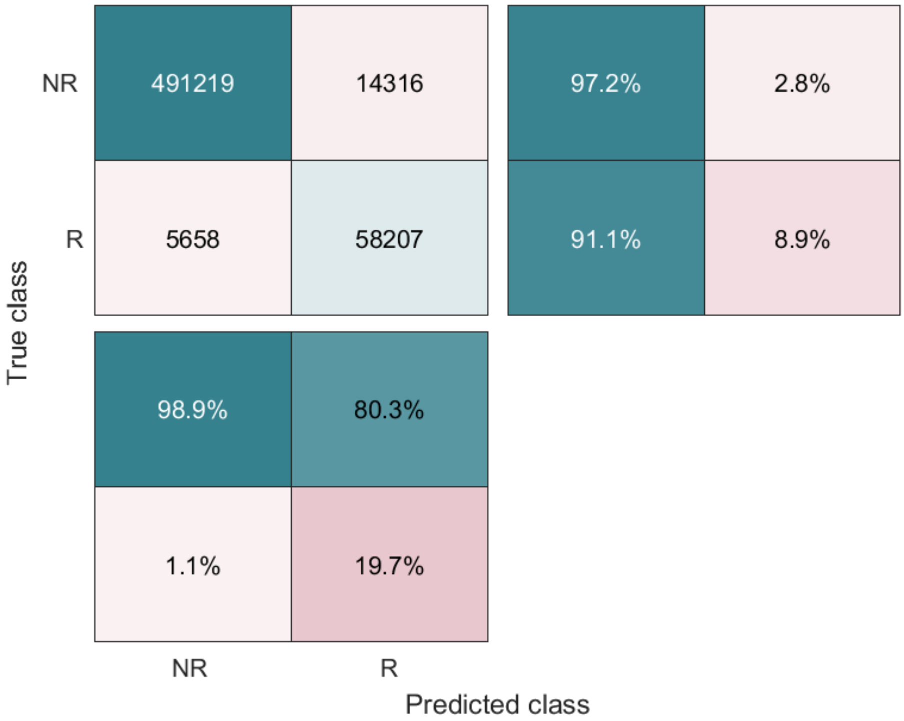
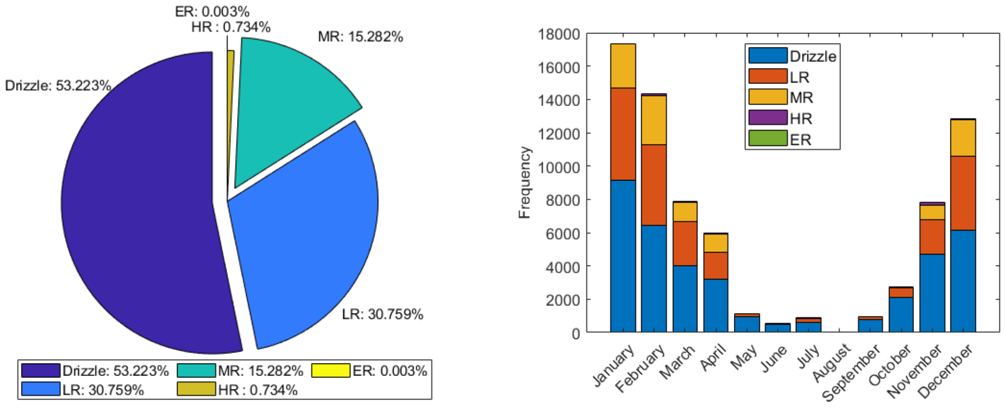


Figure 6.8: Confusion matrix for the Bagged Ensemble Trees Cross-Validation 5 folds model (Random Forest) constructed from the "Binary" data set for the year of 2019 from the ERA5 data over the island of Crete. For a description of the tables see caption in Fig. 6.5.



(a) Pie chart "Only Rain" 2019

(b) Histogram per month, "Only Rain" 2019

Figure 6.9: Pie chart and histogram of the temporal distribution of the five different classes of precipitation for the "Only Rain" data set. The hourly values correspond to a time period of one year spanning from 01-Jan-2019 00:00:00 to 31-Dec-2019 23:00:00. All of the values are above threshold (0.0508 mm/hr); D: Drizzle; LR: Light Rain; MR: Moderate Rain; HR: Heavy Rain; ER: Excessive Rain.

6.5.2 "Only Rain" data set

In the case of the "Only Rain" data set ($\approx 72\ 500$ values per variable), the methods used and their classification measures are presented in the Table 6.6. Since we have five classes, a completely random guess has a probability of success equal to 20%, implying an 80% classification error. In order to decide which of the models is the best we will list the models that resulted in the best and the second best value for each metric. The following metrics characterize the entire model and not each class specifically. According to the precision, the recall and the specificity metrics, the best model is the Bagged Ensemble Trees Cross-Validation 5 folds (62.7%, 52.2%, 93.1%), followed by the Fine SVM Cross-Validation 5 folds (60.3%, 44.8%, 90.2%). For the F-measure and the G-mean values, the best model is the Bagged Ensemble Trees Cross-Validation 5 folds (55.5%, 69.7%), followed by the Cubic SVM Cross-Validation 5 folds (50.1%, 64.6%). Lastly, based on the accuracy metric, the best model is the Bagged Ensemble Trees Cross-Validation 5 folds (80.0%), followed by the Fine SVM Cross-Validation 5 folds (73.8%).

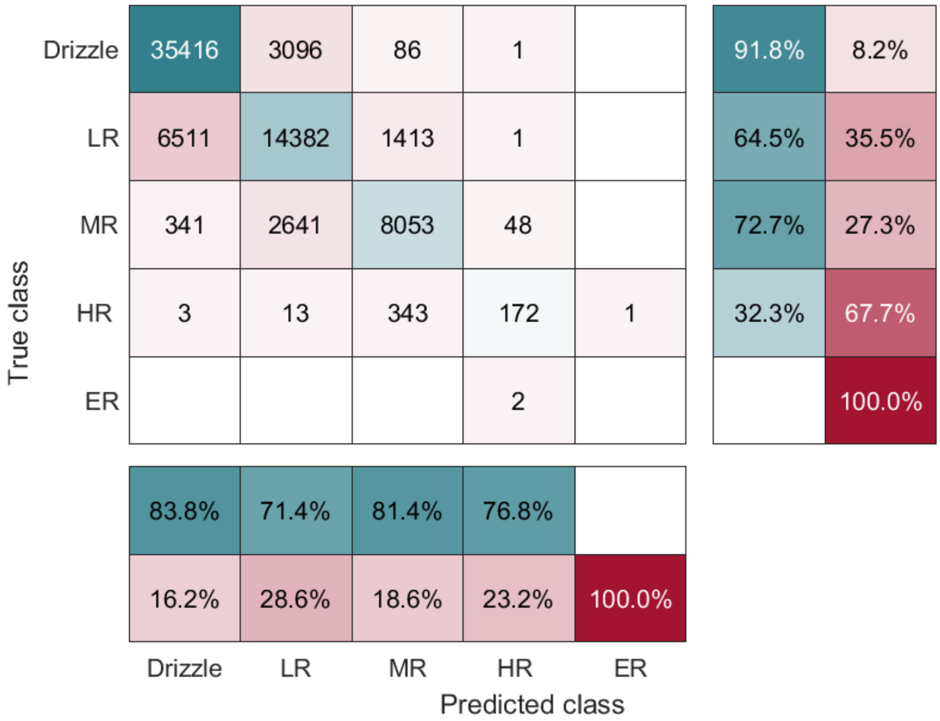
Based on the total model classification measures, the best model is the Bagged

Ensemble Trees Cross-Validation 5 folds because it has the best classification measures according to all of the metrics. As the second best model we will consider the model Fine SVM Cross-Validation 5 folds because it has the second best metrics for four out of the six measures (precision, recall, specificity and accuracy).

Specifically for the best model (Bagged Ensemble Trees Cross-Validation 5 folds), the performance metrics per class are shown in Table 6.6. We see that the metrics for the class Drizzle are excellent, for Light Rain poor, for Moderate Rain mediocre, while for Heavy and Excessive Rain the model greatly under-performs. However, it should be emphasized that the percentages of the classes in the data set are: Drizzle 53.2%, LR 30.8%, MR 15.3%, HR 0.7% and HR ≈ 0%. This means that from the 53.2% of the Drizzle data, the model correctly classified the 91.7% of that subset. Also for the 15.3% of the MR data, the model correctly classified the 72.7% of that subset. This is not trivial considering that the percentage in the entire data set is small enough that the model has difficulty to train. This is more apparent in the cases of the HR and ER classes where the percentages in the set are 0.7% and ≈0.

The confusion matrix of the best model (Bagged Ensemble Trees Cross-Validation 5 folds) according to the accuracy measure (see Table 6.6) is shown in Fig. 6.10. The accuracy of the model is 80.0%, which is definitely an improvement of the 20% by chance. In terms of the accuracy, most of the models performed above average.

What is interesting about the results in Table 6.6 is that while the performance of the RUSBoosted is poor, it is the only model that accurately predicted the Excessive Rain class (class 5). This is surprising since the rest of the classes were highly misclassified, yet, the fact that it could predict precisely the only two excessive rain occurrences is quite impressive and might imply that in cooperation with another method (e.g., the bagged trees ensemble) the classification metrics could improve even more.



	Drizzle	LR	MR	HR	ER		
Drizzle	35416	3096	86	1		91.8%	8.2%
LR	6511	14382	1413	1		64.5%	35.5%
MR	341	2641	8053	48		72.7%	27.3%
HR	3	13	343	172	1	32.3%	67.7%
ER				2			100.0%
	83.8%	71.4%	81.4%	76.8%			
	16.2%	28.6%	18.6%	23.2%	100.0%		

Figure 6.10: Confusion matrix for the Bagged Ensemble Trees Cross-Validation 5 folds model (Random Forest) constructed from the "Only Rain" data set for the year of 2019 from the ERA5 data over the island of Crete. For a description of the tables see caption in Fig. 6.5.

6.6 Conclusions

We analyzed the ERA5 hourly precipitation data using supervised machine learning methods. We investigated the correlation of the precipitation data with the auxiliary variables (Table 3.1) and then we constructed various classification models for the estimation of missing precipitation values by utilizing the information from the precipitation data as well as the predictor variables (auxiliary variables). Random Forests performed best in both the data sets analyzed here. RFs extends decision trees by introducing the idea of ensembles of trees (i.e., forests). This ML method improves the accuracy and uncertainty of other simple decision tree methods. RFs allow to incorporate auxiliary information into the spatial model without excessive tuning, which makes the method suitable for even inexperienced users.

We explored the importance of the auxiliary variables over the response variable. We show that some of the auxiliary variables can be eliminated due to low (almost zero) importance and a subset of the auxiliary variables can be further used as the predictor variables. This will result in a more compact, easier to interpret model, while also reducing the training and estimation time. We defined eight different precipitation classes. Our Random Forest model includes information from both numerical and categorical variables. In order to improve the classification results, we split the data into two different sets, the one containing two classes divided by a specified threshold (the "Binary" data set) and the second set containing the values above the threshold to the corresponding classes (the "Only Rain" data set). In essence, in the "Binary" data set we classify the occurrence of precipitation, while in the "Only Rain" data set we classify the intensity of the precipitation events.

For the "Binary" data set, the Bagged Ensemble Trees Cross-Validation 5 folds was evaluated as the best model. The accuracy of the model is 96.5%, which is definitely an improvement of the 50-50% classification accuracy by chance. In terms of the accuracy, all of the models performed well. The best RF model was trained with five folds or with a training set containing 50%, or with a training set containing 70% of the precipitation data and was validated with the remaining 50% and 30% of the data values for the "Binary" data set. Our findings are promising, with classification error less than 2.8% for the first class and 9% for the second class.

For the "Only Rain" data set, the Bagged Ensemble Trees Cross-Validation 5 folds was also evaluated as the best model. The accuracy of the model is 80.0%, which is definitely an improvement of the 20% by chance. Regarding the "Only Rain" data set the model presented a subpar performance, resulting to error for the first class less than 9%, however for the LR and MR classes the errors are approximately 35% and 27% respectively. The last two classes that represent a small portion of the entire data set return high classification errors, demonstrating unsatisfactory model performance. Nevertheless, our belief is that with more tuning or by combining multiple methodologies the results can be further improved. The results drawn herein can prove useful, especially as a first step into removing and refining the variables needed for resulting in a more accurate representation while modeling precipitation data. While combining several classes together would have resulted in a more balanced data set and probably in better classification metrics, this is not always applicable, especially in cases when the classes' effects are distinctive.

Lastly, it should be pointed out that despite the advantages and the ease of use of machine learning methods such as Random Forests, insight from specialists cannot be replaced. RFs can become a very powerful tool in the hands of experts.

Table 6.5: Classification performance comparison on the hourly precipitation data for the year of 2019 with the "Binary" data set. The measures are described in Section 6.3.4. The predictor variable "month" is treated as categorical variable, as well as the response variable. All of the models that do not specify the trees used, use 30 learners. The optimal values are shown with bold lettering and italics, while the second optimal values are shown with bold lettering. "failed:memory" means that the specific method was tested but failed to finish due to memory shortage.

Method	**Precision**	**Acc+(Recall)**	**Acc-**	**F-measure**	**G-mean**	**Accuracy**
Cross-Validation: 5 folds						
Fine Tree	97.6	94.3	78.1	95.9	85.8	92.7
Medium Tree	96.8	93.9	72.4	95.3	82.4	91.7
Coarse Tree	93.9	94.8	60.9	94.4	76.0	90.2
Logistic Regression	97.9	92.6	76.7	95.2	84.3	91.4
Linear SVM	98.4	92.7	81.1	95.5	86.7	91.8
Quadratic SVM	98.7	93.6	85.6	96.1	89.5	92.9
Cubic SVM	88.4	91.3	34.7	89.8	56.3	82.5
Fine Gaussian SVM	***99.0***	94.9	90.1	96.9	92.4	94.4
Medium Gaussian SVM	98.8	93.9	87.7	96.3	90.8	93.4
Coarse Gaussian SVM	98.8	92.8	85.2	95.7	88.9	92.2
Boosted Ensemble Trees	97.9	94.7	81.5	96.3	87.9	93.4
Bagged Ensemble Trees	**98.9**	97.2	91.1	***98.0***	***94.1***	***96.5***
Ensemble RUSBoosted Trees	86.4	98.1	48.8	91.9	69.2	86.7
Holdout Validation set: 50%						
My Ensemble RUSBoosted 1000 Trees	98.4	94.6	**92.5**	96.5	**93.6**	94.3
My Ensemble RUSBoosted 500 Trees	98.5	94.6	***92.6***	96.5	93.6	94.2
Holdout Validation set: 30%						
Fine Tree			failed:memory			
Medium Tree			failed:memory			
Coarse Tree	94.7	94.3	64.1	94.5	77.7	90.4
Logistic Regression			failed:memory			
Linear SVM	92.7	**98.4**	47.2	95.5	68.1	91.9
Quadratic SVM			failed:memory			
Cubic SVM			failed:memory			
Fine Gaussian SVM			failed:memory			
Medium Gaussian SVM			failed:memory			
Coarse Gaussian SVM			failed:memory			
Boosted Ensemble Trees	94.7	98.0	62.2	96.3	78.1	93.4
Bagged Ensemble Trees	97.1	***98.8***	80.0	**98.0**	88.9	**96.4**
Ensemble RUSBoosted Trees	98.1	86.7	88.7	92.0	87.6	86.9

Table 6.6: Classification performance comparison on the hourly precipitation data that are higher than 0.05 mm/hr for the year of 2019. The measures are described in Section 6.3.4. The predictor variable "month" is treated as categorical variable, as well as the response variable. Five classification classes are used and their ranges are presented in table 6.3 (Drizzle (4) to Excessive Rain (8)). All of the models use 30 learners. The optimal values are shown with bold lettering and italics, while the second optimal values are shown with bold lettering. "failed:memory" means that the specific method was tested but failed to finish due to memory shortage. NA: not applicable; the respective measures involve division by zero or by a very small number.

Method	**Precision**	**Acc+(Recall)**	**Acc-**	**F-measure**	**G-mean**	**Accuracy**
Cross-Validation: 5 folds						
Fine Tree	45.6	35.7	87.9	37.6	56.0	64.7
Medium Tree	35.0	32.3	87.2	33.0	53.1	62.2
Coarse Tree	30.8	27.1	85.5	26.4	48.2	56.9
Linear SVM	36.4	32.3	86.7	33.2	52.9	63.8
Quadratic SVM	55.4	41.0	88.3	44.6	60.1	68.4
Cubic SVM	57.5	46.5	89.7	50.1	64.6	72.3
Fine Gaussian SVM	60.3	44.8	90.2	48.8	63.5	73.8
Medium Gaussian SVM	55.9	37.9	88.7	40.2	58.0	69.5
Coarse Gaussian SVM	38.2	31.7	86.6	32.8	52.4	64.1
Boosted Ensemble Trees	36.9	32.7	87.3	33.6	53.4	64.6
Bagged Ensemble Trees	62.7	52.2	93.1	55.5	69.7	**80.0**
Ensemble RUSBoosted Trees	22.8	40.0	81.0	20.0	56.9	37.3
Best model: Bagged Ensemble Trees						
Class 1 - Drizzle	83.8	91.7	79.8	87.6	85.6	91.7
Class 2 - Light Rain	71.4	64.5	88.5	68.8	75.6	64.5
Class 3 - Moderate Rain	81.4	72.7	97.0	76.8	83.9	72.7
Class 4 - Heavy Rain	76.8	32.3	99.9	45.5	56.84	32.3
Class 5 - Excessive Rain	0	0	100	NA	0	0

Chapter 7

Stochastic Local Interaction Models

7.1 Summary

In this chapter, we discuss why more flexible and less demanding computational approaches than standard geostatistical methods are needed (e.g., remote sensing data availability). Then, we present a theoretical framework for the analysis of space-time data based on stochastic local interaction (SLI) models. We apply the SLI method to three temporal data sets which involve hourly reanalysis temperature, solar radiation on the horizontal plane, and precipitation data. Furthermore, we apply the SLI method to two spatiotemporal data sets which involve hourly reanalysis temperature and precipitation data. The nearest neighbor interpolation is also applied to these data. Finally, we compare the methodologies and we present our conclusions.

7.2 Introduction

In the current era, the availability of data is becoming more and more staggering. The ways that the data are collected are numerous, including remote sensing, extensions of the ground-based networks, sensors of unmanned aerial vehicles as well as crowd-sourcing [Council et al., 2013]. This explosion in data availability has affected science and engineering, providing more information in many cases than what is possible right now to be processed. Many theoretical and technical challenges arise in the processing and modeling of such immense data sets. Due to that, flexible and computationally powerful solutions are in need. Most of the past developed methodologies are not functional for extremely big and hyper-dimensional data. For instance, classical geostatistical and machine learning methods [Chilès and Delfiner, 2012; Rasmussen and Williams, 2006] are limited by the cubic dependence of the computational time on data size, which is prohibitive even for large purely spatial data [Hristopulos and Agou, 2020].

State-of-the-art methods and better computational resources are integral for processing and modeling, especially but not limited to space-time (ST) data. For instance, correlations in the ST domain are often overlooked by methods that extend the spatial statistics by merely adding a separable time dimension [Christakos, 1992; Cressie and Wikle, 2011]. Non-separable covariance models have been developed for that reason [De Iaco et al., 2002; Kolovos et al., 2004; Varouchakis and Hristopulos, 2019]. Additionally, many methods suffer from scalability issues, regardless if they are derived from geostatistics [Chilès and Delfiner, 2012; Cressie and Wikle, 2011; Gneiting et al., 2006] or machine learning [Rasmussen and Williams, 2006]. The main origin of the problem is the computational cost of the inversion of large covariance (Gram) matrices [Rasmussen and Williams, 2006; Sun et al., 2012]. Therefore, classical methods executed on standard desktop computers are limited to data sets with size $N \sim \mathcal{O}(10^3) - \mathcal{O}(10^4)$.

Gaussian field theories and Gaussian Markov random fields (GMRFs) are both characterized by local structures which are derived from the derivatives of the field [Mussardo, 2010] or the interactions created by local neighborhoods [Rue and Held, 2005] accordingly. Similarly, the SLI models, which are inspired by the aforementioned methods, are based on the creation of correlations generated by interactions between neighboring sites and times.

In the following Sections 7.3-7.6 we present a theoretical framework for the analysis of space-time (ST) data that is based on *stochastic local interaction* (SLI) models [Hristopulos, 2015a; Hristopulos and Tsantili, 2017]. This formulation can assist in filling missing values by interpolation in environmental ST data sets. For example, gaps in records of meteorological variables need to be reconstructed to evaluate renewable energy potential at candidate sites [Koutroulis and Kolokotsa, 2010], while ground-based rainfall gauge networks often have missing data [Bárdossy and Pegram, 2014]. The main idea in SLI is that the ST correlations are determined employing *sparse precision matrices* that only involve couplings between near neighbors (in the ST domain). The advantage of SLI against GMRF models is that it is suitable for direct application to scattered data and stochastic graph processes, versus the need for regular lattice data defined on continuum spaces [Hristopulos, 2015b; Hristopulos et al., 2021]. However, it is also applicable to data on regular space-time lattices [Hristopulos and Agou, 2020].

The SLI models configuration captures the local spatial dependence thought ideas from *kernel regression* [Nadaraya, 1964; Watson, 1964]. The correlations between neighboring points are expressed in terms of suitably selected weighting functions, supplied by *kernel* functions. In Section 7.3.2, additional information about kernel functions can be found, as well as their corresponding equations (Table 7.1).

Herein, we apply the SLI method to three temporal data sets and two ST data sets. The temporal data sets involve reanalysis temperature, solar radiation on horizontal plane, and precipitation data, while the ST data sets involve reanalysis temperature data, and reanalysis precipitation data. Finally, we present our conclusions and a brief discussion in Section 7.9.

7.3 ST Model based on Stochastic Local Interactions

A *space-time scalar random field (STRF)* $X(\mathbf{s}, t; \omega) \in \mathbb{R}$ where $\mathbf{s}, t \in \mathbb{R}^d \times \mathbb{R}$ and $\omega \in \Omega$ is defined as a mapping from the probability space (Ω, A, P) into the space of real numbers $\mathbb{R}$. For each ST coordinate $(\mathbf{s}, t)$, $X(\mathbf{s}, t; \omega)$ is a measurable function of ω, where ω is the state index [Christakos, 1992]. The states (realizations) of the random field $X(\mathbf{s}, t; \omega)$ are real-valued functions $x(\mathbf{s}, t)$ obtained for a specific ω. In the following, the state index ω is dropped to simplify notation.

We focus on partially sampled realizations $\mathbf{x} = (x_1, \ldots, x_N)^\top$ of the random field, where $N \in \mathbb{N}$ is the sample size. The vector $\mathbf{x}$ comprises the field values at the ST point set $\mathbb{S} = \{(\mathbf{s}_1, t_1), \ldots, (\mathbf{s}_N, t_N)\}$. The point set is assumed to be general; it may represent a time sequence of lattice sites, randomly scattered points in space and time, or a collection of time series at random locations in space.

7.3.1 Energy of the exponential joint density

In the following we will present a simplified version of the initial proposed SLI model in [Hristopulos, 2015b] that involved squared fluctuation and gradient terms.

The SLI model is defined in terms of a *Boltzmann-Gibbs* exponential joint PDF and the energy function $\mathcal{H}(\cdot;\cdot)$ that represents the "cost" of a specific configuration

$$f_{\mathrm{x}}(\mathbf{x};\boldsymbol{\theta}) = \frac{\mathrm{e}^{-\mathcal{H}(\mathbf{x};\boldsymbol{\theta})}}{Z(\boldsymbol{\theta})}, \tag{7.1}$$

where $\boldsymbol{\theta}$ is a vector of model parameters, and the denominator $Z(\boldsymbol{\theta})$ –known in physics as the *partition function*– represents a normalization constant [Mussardo, 2010].

The energy-based approach is commonly used in statistical physics [Kardar, 2007; Mussardo, 2010]. Its main advantage is that it expresses statistical dependence in terms of interactions between space locations and time instants which can be local, without recourse to the concept of the covariance matrix. Depending on the form of the interactions involved in the energy, both Gaussian and non-Gaussian probability density functions can be obtained. The most famous example of non-Gaussian dependence is the magnetic Ising model [Ising, 1925] which was introduced in spatial statistics by Besag [1974]. While non-Gaussian models are definitely interesting, their Gaussian counterparts lead to explicit predictive expressions and uncertainty estimates based on the conditional variance. Hence, herein we focus on a Gaussian SLI model [Hristopulos, 2020].

We assume that $\mathcal{H}(\mathbf{x};\boldsymbol{\theta})$ satisfies the following properties for any vector $\mathbf{x} \in \mathbb{R}^N$ and $N \in \mathbb{N}$:

1. *Gaussianity:* $\mathcal{H}(\mathbf{x};\boldsymbol{\theta})$ is a quadratic function of the data vector $\mathbf{x}$ that can

be expressed as

$$\mathcal{H}(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{2}(\mathbf{x}-\mathbf{m}_{\mathrm{x}})^{\top}\mathbf{J}(\boldsymbol{\theta}')(\mathbf{x}-\mathbf{m}_{\mathrm{x}}), \tag{7.2}$$

where $\mathbf{m}_{\mathrm{x}} = \left(m_{\mathrm{x};1},\ldots,m_{\mathrm{x};N}\right)^{\top}$ is a vector of mean (trend) values such that $m_{\mathrm{x};,i} = \mathcal{E}\left[\,X(\mathbf{s}_i,t_i)\,\right]$, where $\mathcal{E}\left[\,\cdot\,\right]$ is the expectation operator. On the other hand, $\mathbf{J}(\boldsymbol{\theta}')$ is the $N\times N$ precision matrix[1]. The latter depends on the parameter vector $\boldsymbol{\theta}' = \boldsymbol{\theta}\smallsetminus\{b_1,\ldots,b_K\}$ which excludes the trend coefficients. The vector $\mathbf{m}_{\mathrm{x}}$ incorporates both periodic and aperiodic trend components.

2. *Positive-definiteness:* $\mathcal{H}(\mathbf{x};\boldsymbol{\theta}) > 0$ for all $\mathbf{x}$ that are not identically equal to zero. This is equivalent to the *precision matrix* $\mathbf{J}(\boldsymbol{\theta})$ being a positive-definite matrix.

3. *Sparseness:* $\mathbf{J}(\boldsymbol{\theta}')$ is a *sparse matrix*[2] that incorporates the local interactions.

More specifically, we focus on the following SLI energy function for a ST field with N data points which satisfies the properties of Gaussianity, positive-definiteness and sparseness:

$$\mathcal{H}(\mathbf{x};\boldsymbol{\theta}) = \frac{1}{2\lambda}\left[\sum_{n=1}^{N}\frac{1}{N}(x_n - m_{\mathrm{x};n})^2 + c_1\left\langle\,(x'_n - x'_k)^2\,\right\rangle\right]. \tag{7.3}$$

We assume that the mean is modeled by means of a trend function which can be expressed as $m_{\mathrm{x}}(\mathbf{s},t) = \sum_{k=1}^{K} b_k f_k(\mathbf{s},t)$ in terms of a suitable ST function basis $\{f_k(\mathbf{s},t)\}_{k=1}^{K}$, where $\{b_k\}_{k=1}^{K}$ is a set of real-valued trend coefficients and $f_k : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, for $k = 1,\ldots,K$.

The variables x_n, x_k stand for $x(\mathbf{s}_n,t_n)$ and $x(\mathbf{s}_k,t_k)$ respectively, where $n,k = 1,\ldots N$ while x'_n, x'_k represent the residuals after the trend values are removed. The term $\left\langle\,(x'_n - x'_k)^2\,\right\rangle$ represents a kernel weighted average of the squared increments. However, instead of focusing on all $\mathcal{O}(N^2)$ pairs, the average defined

[1] The precision matrix $\mathbf{J}(\boldsymbol{\theta}')$ is the inverse covariance.

[2] A sparse matrix is a specific type of matrix in which the proportion of zero entries to non-zero entries is significantly larger.

below selects only pairs within a local neighborhood around each point $\mathbf{s}_n$ ($\mathbf{s}_N, t_N$ in the ST domain).

The *SLI parameter vector* $\boldsymbol{\theta}$ includes the trend coefficients $\{b_k\}_{k=1}^{K}$, the overall scale parameter λ (which is proportional to the variance), and the increment coefficient c_1 (a dimensionless factor that multiplies the contribution from the squares of the increments). The vector $\boldsymbol{\theta}$ includes additional parameters that determine the local ST neighborhoods used in the average of the squared increments $\langle \cdot \rangle$. The average is defined in Eq. (7.4) below.

7.3.2 Kernel-based averaging

The weights in the average of the squared increments are defined by means of the *Nadaraya-Watson* equation [Nadaraya, 1964; Watson, 1964], i.e.,

$$\langle (x'_n - x'_k)^2 \rangle = \frac{\sum_{n=1}^{N} \sum_{k=1}^{N} w_{n,k} \, (x'_n - x'_k)^2}{\sum_{n=1}^{N} \sum_{k=1}^{N} w_{n,k}}. \tag{7.4}$$

The coefficients $w_{n,k}$ are defined in terms of *compactly supported ST kernel functions* $K(\cdot,\cdot) : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R}$, where $q = d + 1$ for an ST kernel, $q = d$ for a spatial kernel, and $q = 1$ for a temporal kernel. Kernel functions are symmetric, real-valued functions; herein they are assumed to take values in the interval $[0, 1]$ without loss of generality. Moreover, we will assume spatially homogeneous and temporally stationary kernel functions, i.e., $K(\mathbf{s}_1, \mathbf{s}_2) = K(\mathbf{s}_1 - \mathbf{s}_2)$, $K(t_1, t_2) = K(t_1 - t_2)$, and $K(\mathbf{s}_1, t_1; \mathbf{s}_2, t_2) = K(\mathbf{s}_1 - \mathbf{s}_2, t_1 - t_2)$. Furthermore, it will be assumed for simplicity that the kernel function depends only on the magnitude of the ST distance [Hristopulos and Agou, 2020].

In the case of the spatial model we omit the terms that corresponds to the time dimension, while in the one dimensional case the variables x_n, x_k stand for $x(t_n)$ and $x(t_k)$ respectively, where $n, k = 1, \dots N$ correspond only to different time steps for a specific location. Application of the method in two dimensional data can be found in the work of Hristopulos et al. [2021].

Kernel Functions

A kernel (weighting) function $K(x, x_n^*)$ is a non-negative function that assigns a real value to any pair of points x and x_n^*, where $\{x_n^*\}_{n=1}^N$ are the sample values. For $x, x_n^* \in \mathbb{R}^d$ let $u = \|x - x_n^*\|/h$ represent the normalized distance, and $h > 0$. A kernel function is (i) non-negative, i.e., $K(u) \geq 0$ for all u; (ii) symmetric, i.e., $K(-u) = K(u)$ for all $u \in \mathbb{R}_{0,+}$; (iii) maximized at $u = 0$; and (iv) a continuous function of u [Hristopulos, 2020]. In addition, kernel functions can be normalized so that their integral over the entire space equals one [Pavlides et al., 2022], this property is relevant to our study, since we want to use it to replace the empirical cumulative distribution function.

A list of common kernel functions is given in Table 7.1. The first seven functions of the table (uniform, triangular, Epanechnikov, quartic, tricube, spherical, Cauchy) are compactly supported, while the last two functions (exponential, Gaussian) are infinitely extended but integrable. The use of compactly supported kernel functions leads to local predictors, while infinitely extended kernels allow averaging over a larger part of the domain at the expense of increased computational cost [Hristopulos et al., 2021].

The standard Parzen–Rosenblatt kernel density estimator of the marginal PDF of $X(\mathbf{s};\omega)$ based on the sample $\{x_n^*\}_{n=1}^N$ is given by

$$\hat{f}_X(x) = \sum_{n=1}^{N} \frac{1}{N\,h} K\left(\frac{x - x_n^*}{h}\right), \tag{7.5}$$

where $K(\cdot)$ is a kernel function (Table 7.1) and h is the bandwidth parameter [Ghosh, 2017]. The range of the kernel, i.e., the bandwidth parameter, can be estimated employing several techniques, such as the minimization of the mean integrated square error (MISE) of the estimator (minimizing Eq. (7.5)), which additionally provides an explicit equation for the PDF at every $x \in \mathbb{R}$ [Hristopulos, 2020]. An estimate of the empirical CDF is then obtained as follows

$$\hat{F}_X(x) = \sum_{n=1}^{N} \frac{1}{N\,h} \tilde{K}\left(\frac{x - x_n^*}{h}\right), \tag{7.6}$$

Table 7.1: List of common kernel functions. The normalized distance u is given by $u = \|\mathbf{r}\|/h$ where $\|\mathbf{r}\| = \|x - x_n^*\| \geq 0$ denotes the Euclidean norm. $\vartheta(x) = 1$ if $x \geq 0$ and $\vartheta(x) = 0$ if $x < 0$ is the unit step function.

Name	Equation ($u \geq 0$)
Uniform	$K(u) = \vartheta(1-u)$
Triangular	$K(u) = (1-u)\,\vartheta(1-u)$
Epanechnikov	$K(u) = (1-u)^2\,\vartheta(1-u)$
Quartic (biweight)	$K(u) = (1-u^2)^2\,\vartheta(1-u)$
Tricube	$K(u) = (1-u^3)^3\,\vartheta(1-u)$
Spherical	$K(u) = (1-1.5u+0.5u^3)\,\vartheta(1-u)$
Truncated Cauchy	$K(u) = 1/(1+u^2)\,\vartheta(1-u)$
Exponential	$K(u) = \exp(-u)$
Gaussian	$K(u) = \exp(-u^2)$

where $\tilde{K}(\cdot)$ represents the kernel's integral, i.e.,

$$\tilde{K}\left(\frac{x - x_n^*}{h}\right) = \int_{-\infty}^{x} \mathrm{d}x'\, K\left(\frac{x' - x_n^*}{h}\right). \tag{7.7}$$

7.4 Definition of space-time distance

The space-time distance used in the kernel weights determines the structure of correlations that we impose in the space-time domain. Both separable and non-separable space-time metric distances are possible as discussed below [Hristopulos and Agou, 2020].

Composite space-time distance: In this case the spatial and temporal coordinates are intertwined in the distance metric. For example, the differential of the space-time distance between two points using the Riemannian metric is

$$dq = \sqrt{\sum_{i=1}^{d+1}\sum_{j=1}^{d+1} g_{i,j}\, dz^{(i)} dz^{(j)}}, \tag{7.8}$$

where $\{g_{i,j}\}_{i,j=1}^{d+1}$ are the elements of the metric tensor $\mathbf{g}$, $\{dz^{(i)}\}_{i=1}^{d}$, are the differentials of the spatial distance in the d orthogonal directions, and $dz^{(d+1)}$ is the time differential [Christakos, 2017; Christakos et al., 2000].

In the Euclidean case the metric tensor $\mathbf{g}$, is given by

$$g_{i,j} = \delta_{i,j}\left[1 + (\alpha - 1)\,\delta_{i,d+1}\right],\; i,j = 1,\ldots,d+1, \tag{7.9}$$

where α is a parameter that controls the contribution of the time lag in the composite distance. The kernel coefficient based on the composite Euclidean metric can be expressed as

$$w_{n,k} = K\left(\frac{\sqrt{\mathbf{r}_{n,k}^2 + \alpha^2\tau_{n,k}^2}}{h_{s,n}}\right). \tag{7.10}$$

In Eq. (7.10), $\mathbf{r}_{n,k} = (\mathbf{s}_n - \mathbf{s}_k)$ is the spatial lag between the initial point $\mathbf{s}_n$ and the target point $\mathbf{s}_k$, and $h_{s,n}$ is the *local spatial bandwidth* at $\mathbf{s}_n$. In addition, $\tau_{n,k} = t_n - t_k$ is the temporal lag between the initial and target times. The space-time distance for the composite metric leads to ellipsoidal neighborhoods as shown in the schematic of Fig. 7.1a. The temporal bandwidth in this case is $h_{t,n} = h_{s,n}/\alpha$ [Hristopulos and Agou, 2020].

Separable space-time distance: The coefficients $w_{n,k}$ for a separable space-time neighborhood are defined as

$$w_{n,k} = K\left(\frac{\|\mathbf{r}_{n,k}\|}{h_{s,n}}\right) K\left(\frac{|\tau_{n,k}|}{h_{t,n}}\right),\; n,k = 1,\ldots,N. \tag{7.11}$$

In the weight equation (Eq. (7.11)), $h_{s,n}$ is the *local spatial bandwidth* at $\mathbf{s}_n$ and $h_{t,n}$ is the *temporal bandwidth.* The space-time distance for the separable space-time metric leads to cylindrical neighborhoods as shown in Fig. 7.1b.

7.4.1 Definition of bandwidths

For each ST point $\{(\mathbf{s}_n, t_n)\}_{n=1}^{N}$, the *spatial bandwidth* $h_{s,n}$ is determined from the geometry of the sampling network around the spatial point $\mathbf{s}_n$, while the temporal bandwidth $h_{t,n}$ is based on the time neighborhood around t_n. In general, this means that the number of bandwidth parameters scales linearly with the sampling size, leading to an under-determined estimation problem when the additional parameters are accounted for.

To simplify the bandwidth estimation we use a trick that reduces the dimensionality of the problem. We assign to each point a bandwidth which is proportional to the spatial distance $D_{n,[K_s]}(\mathbb{S})$ between this point and its K_s-nearest neighbor in the point set $\mathbb{S}$. Thus, it holds that $h_{s,n} = \mu_s \, D_{n,[K_s]}(\mathbb{S})$, where typically $K_s = 2, 3, 4$, and $\mu_s > 0$ is a dimensionless spatial bandwidth parameter to be estimated from the data.

In the case of the *composite space-time* distance the temporal bandwidths $h_{t,n}$ are determined from the $h_{s,n}$ and the additional parameter α. For a *separable ST* distance metric, the temporal bandwidths are determined by means of $h_{t,n} = \mu_t \, \tilde{D}_{n,[K_t]}(\mathbb{S})$, where K_t is the order of the temporal neighbor and $\mu_t > 0$ is a dimensionless temporal bandwidth parameter. This definition of the temporal bandwidth in the case of uniform time step implies uniform bandwidths for all except the initial and final times, where the bandwidth is automatically increased to account for the missing left and right neighbors respectively.

7.4.2 Properties of kernel weights

The kernel-average of the squared increments (Eq. (7.4)) can be expressed in terms of normalized weights $u_{n,k}$ as follows

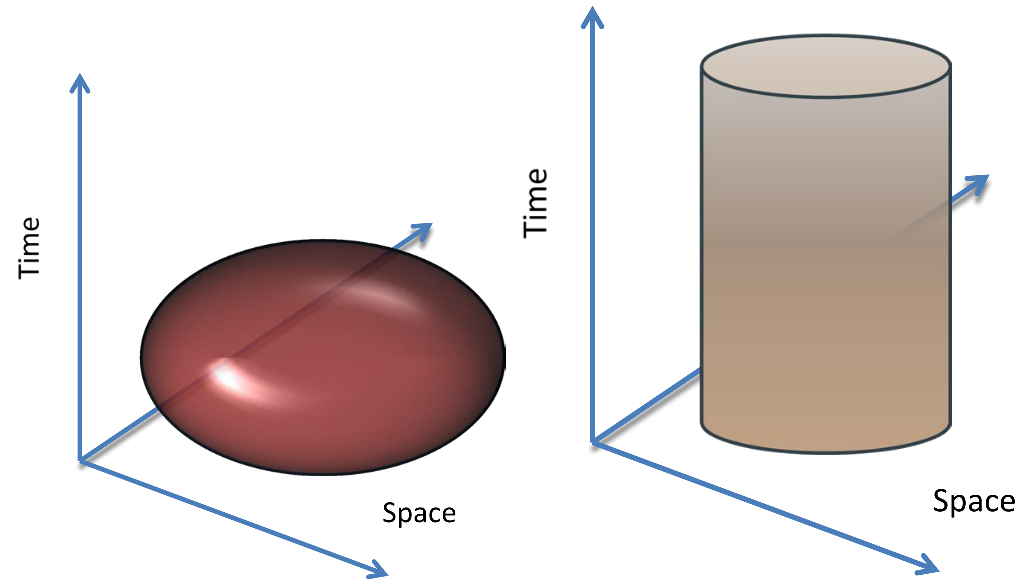
$$\langle (x'_n - x'_k)^2 \rangle = \sum_{n=1}^{N} \sum_{k=1}^{N} u_{n,k} \, (x'_n - x'_k)^2 \, , \tag{7.12a}$$

$$u_{n,k} = \frac{w_{n,k}}{\sum_{n=1}^{N} \sum_{k=1}^{N} w_{n,k}} . \tag{7.12b}$$

Normalization: The Eq. (7.12b) of the kernel weights implies that

$$\sum_{n=1}^{N} \sum_{k=1}^{N} u_{n,k} = 1. \tag{7.13}$$

Asymmetry: The Eq. (7.12b) of the bandwidths is based on the local ST neighborhood. This implies that the *spatial weights* are in general asymmetric,



(a) Composite

(b) Separable

Figure 7.1: Schematics of kernel-based neighborhoods for composite (left) and separable (right) space-time structures. Figure taken from Hristopulos and Agou [2020].

i.e., $w_{n,k} \neq w_{k,n}$ if $\mathbf{s}_n \neq \mathbf{s}_k$, since the sampling density around the point $\mathbf{s}_n$ can be quite different than around the point $\mathbf{s}_k$.

Non-separability: The kernel weights $u_{n,k}$ are non-separable for both the composite and the separable ST distance metrics. In the first case this is obvious from the Eq. (7.10). In the second case, even though the $w_{n,k}$ are separable, the normalized weights $u_{n,k}$ are non-separable functions of space and time due to the kernel summation in the denominator of Eq. (7.4).

Robustness with respect to general distance metrics: Regardless of the distance metric used, the kernel-based weights $u_{n,k}$ are non-negative. This implies that the SLI energy function (Eq. (7.3)) is positive, and consequently the precision matrix is positive definite. Hence, general distance metrics, e.g., Manhattan (also known as city block and taxicab) distance, can be used in the SLI model.

In the following we develop the SLI formalism for a separable space-time metric structure.

7.4.3 Squared increments for separable space-time metric

In this section we formulate the average squared increments for separable space-time kernel functions using matrix operations.

First, we define the square kernel matrices $\mathbf{K}_s$ of dimension $N_s \times N_s$ and $\mathbf{K}_t$ of dimension $N_t \times N_t$ as follows

$$\mathbf{K}_s = \begin{bmatrix} K\left(\frac{\|\mathbf{r}_{1,1}\|}{h_{s,1}}\right) & \dots & K\left(\frac{\|\mathbf{r}_{1,N_s}\|}{h_{s,1}}\right) \\ \vdots & \vdots & \vdots \\ K\left(\frac{\|\mathbf{r}_{N_s,1}\|}{h_{s,N_s}}\right) & \dots & K\left(\frac{\|\mathbf{r}_{N_s,N_s}\|}{h_{s,N_s}}\right) \end{bmatrix}, \tag{7.14a}$$

$$\mathbf{K}_t = \begin{bmatrix} K\left(\frac{\|\tau_{1,1}\|}{h_{t,1}}\right) & \dots & K\left(\frac{\|\tau_{1,N_t}\|}{h_{t,1}}\right) \\ \vdots & \vdots & \vdots \\ K\left(\frac{\|\tau_{N_t,1}\|}{h_{t,N_t}}\right) & \dots & K\left(\frac{\|\tau_{N_t,N_t}\|}{h_{t,N_t}}\right) \end{bmatrix}. \tag{7.14b}$$

Then, the $N \times N$ matrix $\mathbf{W}$ of ST kernel weights is given by the following Kronecker product (denoted by $\otimes$):

$$\mathbf{W} = \mathbf{K}_s \otimes \mathbf{K}_t. \tag{7.14c}$$

For compactly supported kernel functions the matrix $\mathbf{W}$ given by Eq. (7.14c) is sparse (for data size bigger than the neighborhood size).

The matrix $\mathbf{U}$ of the *normalized kernel weights* is then defined by means of

$$\mathbf{U} = \frac{\mathbf{W}}{\|\mathbf{W}\|_1}, \tag{7.15a}$$

where the denominator $\|\mathbf{W}\|_1$ represents the entry-wise L_1 norm of the matrix $\mathbf{W}$ and is given by

$$\|\mathbf{W}\|_1 = \sum_{k=1}^{N} \sum_{l=1}^{N} |W_{k,l}|. \tag{7.15b}$$

In terms of the above matrices, the average squared increment of Eq. (7.12) is expressed as follows

$$\langle (x'_n - x'_k)^2 \rangle = \left\| \left[(\mathbf{x}' \otimes \mathbf{1}) - (\mathbf{x}' \otimes \mathbf{1})^\top \right] \circ \mathbf{U} \circ \left[(\mathbf{x}' \otimes \mathbf{1}) - (\mathbf{x}' \otimes \mathbf{1})^\top \right] \right\|_1 \tag{7.16}$$

where $\mathbf{1} = (1, \ldots, 1)^\top$ is the $N \times 1$ vector of ones, and $\circ$ denotes the Hadamard product, i.e., $[\mathbf{A} \circ \mathbf{B}]_{i,j} = A_{i,j} B_{i,j}$.

The computational complexity of the operations in Eq. (7.16) is $\mathcal{O}(N^2)$, if the sparsity of the matrix $\mathbf{W}$ is not taken into account. However, the numerical complexity can be improved using sparse-matrix operations [Hristopulos and Agou, 2020]. We have implemented all the calculations which involve the precision matrix using sparse matrix functionality.

7.4.4 Precision matrix formulation

In light of the above definitions, the SLI energy function (Eq. (7.3)) involves the following parameter vector

$$\boldsymbol{\theta} = (b_1, \ldots, b_K, \lambda, c_1, \mu_s, \mu_t, K_s, K_t)^\top, \tag{7.17}$$

where $\{b_k\}_{k=1}^K$ are the coefficients of the trend model, λ is the SLI scaling factor, c_1 is the square increment coefficient, μ_s, μ_t the dimensionless scaling factors used to determine the bandwidths, and K_s, K_t are the orders of spatial and temporal near neighbors respectively.

The SLI energy function (Eq. (7.3)) can be transformed into a quadratic energy functional, i.e., of the form of Eq. (7.2), by defining the precision matrix $\mathbf{J}(\boldsymbol{\theta}')$ as follows

$$\mathbf{J}(\boldsymbol{\theta}') = \frac{1}{\lambda} \left\{ \frac{\mathbf{I}_N}{N} + c_1 \, \mathbf{J}_1(\mathbf{h}; \boldsymbol{\theta}'') \right\}, \tag{7.18}$$

where $\mathbf{I}_N$ is the $N \times N$ identity matrix: $[\mathbf{I}_N]_{i,j} = 1$ if $i = j$ and $[\mathbf{I}_N]_{i,j} = 0$ otherwise. The precision matrix $\mathbf{J}(\boldsymbol{\theta}')$ involves the parameter vector $\boldsymbol{\theta}' = (\lambda, c_1, \mu_s, \mu_t, K_s, K_t)^\top$. The matrix $\mathbf{J}_1(\mathbf{h}; \boldsymbol{\theta}'')$ is derived from the average squared increments (Eq. (7.12)), and $\boldsymbol{\theta}'' = (\mu_s, \mu_t, K_s, K_t)^\top$ is the parameter vector which determines the kernel bandwidths. The matrix $\mathbf{J}_1(\mathbf{h}; \boldsymbol{\theta}'')$ is expressed in terms of the normalized weights $u_{n,k}$ according to

$$[\mathbf{J}_1(\mathbf{h};\boldsymbol{\theta}'')]_{n,k} = -u_{n,k} - u_{k,n} + [\mathbf{I}_N]_{n,k} \sum_{l=1}^{N} (u_{n,l} + u_{l,n}), \tag{7.19}$$

where the normalized weights $u_{n,k}$ are given by Eq. (7.12b). Hence, the precision matrix (Eq. (7.18)) is determined by the sampling pattern, the kernel functions, and the bandwidths.

7.5 ST Prediction

In this section we consider ST prediction by means of the SLI model at the set of space time points $\mathbb{G} = \{\tilde{\mathbf{s}}_p\}_{p=1}^{P}$, where $\tilde{\mathbf{s}}_p = (\tilde{\mathbf{s}}_p, \tilde{t}_p)$, assuming that the model parameters are known. It is further assumed that the sets $\mathbb{S}$ and $\mathbb{G}$ are disjoint. For example, the set $\mathbb{G}$ could comprise all the nodes of a regular map grid at a time instant t_p for which measurements are not available. Alternatively, $\mathbb{G}$ could comprise all the nodes of an irregular spatial sampling network at a time instant with no measurements.

7.5.1 SLI energy function including prediction set

The SLI energy function that incorporates the prediction sites is given by straightforward extension of Eq. (7.3). Thus, the following expression that involves block vectors of sampling and prediction sites and respective precision block matrices is obtained

$$\mathcal{H}(\mathbf{x}, \mathbf{x}_{\mathbb{G}}; \boldsymbol{\theta}^{*}) = \frac{1}{2} \begin{bmatrix} \mathbf{x}'^{\top} & \mathbf{x}'_{\mathbb{G}} \end{bmatrix} \begin{bmatrix} \mathbf{J}_{\mathbb{S},\mathbb{S}} & \mathbf{J}_{\mathbb{S},\mathbb{G}} \\ \mathbf{J}_{\mathbb{G},\mathbb{S}} & \mathbf{J}_{\mathbb{G},\mathbb{G}} \end{bmatrix} \begin{bmatrix} \mathbf{x}' \\ \mathbf{x}'_{\mathbb{G}} \end{bmatrix}, \tag{7.20}$$

where $\mathbf{x}' = \mathbf{x} - \mathbf{m}_{\mathrm{x}}$ is the detrended data vector, $\mathbf{x}'_{\mathbb{G}} = \mathbf{x}_{\mathbb{G}} - \mathbf{m}_{\mathrm{x}}$ is the fluctuation vector at the prediction points, and $\boldsymbol{\theta}^{*}$ is the estimate of the parameter vector based on the data. Let the sets A, B denote either of the disjoint sets $\mathbb{S}$ or $\mathbb{G}$. Then, the block precision matrices $\mathbf{J}_{A,B}$ are expressed as

$$\mathbf{J}_{A,B}(\boldsymbol{\theta}'^{*}) = \frac{1}{\lambda} \left[c_0 \mathbf{I} + c_1 \mathbf{J}^{(1)}_{A,B}(\boldsymbol{\theta}''^{*}) \right]. \tag{7.21}$$

The block sub-matrices $\mathbf{J}^{(1)}_{A,B}$ are defined as follows:

$$\left[\mathbf{J}^{(1)}_{\mathbb{S},\mathbb{S}}\right]_{n,k} = -\,u_{n,k} - u_{k,n}, \qquad n,k = 1,\ldots,N,\ n \neq k \tag{7.22a}$$

$$\left[\mathbf{J}^{(1)}_{\mathbb{S},\mathbb{S}}\right]_{n,n} = \sum_{l=1\neq n}^{N} \left(u_{n,l} + u_{l,n}\right) + \sum_{p=1}^{P} \left(u_{n,p} + u_{p,n}\right), \qquad n = 1,\ldots,N, \tag{7.22b}$$

$$\left[\mathbf{J}^{(1)}_{\mathbb{S},\mathbb{G}}\right]_{n,p} = -\,u_{n,p} - u_{p,n}, \qquad n = 1,\ldots,N,\ p = 1,\ldots P, \tag{7.22c}$$

$$\left[\mathbf{J}^{(1)}_{\mathbb{G},\mathbb{S}}\right] = \mathbf{J}^{(1)\top}_{\mathbb{S},\mathbb{G}}, \tag{7.22d}$$

$$\left[\mathbf{J}^{(1)}_{\mathbb{G},\mathbb{G}}\right]_{p,q} = -\,u_{p,q} - u_{q,p}, \qquad p \neq q = 1,\ldots,P, \tag{7.22e}$$

$$\left[\mathbf{J}^{(1)}_{\mathbb{G},\mathbb{G}}\right]_{p,p} = \sum_{l=1}^{N} \left(u_{p,l} + u_{l,p}\right) + \sum_{q\neq p=1}^{P} \left(u_{p,q} + u_{q,p}\right), \qquad p = 1,\ldots,P. \tag{7.22f}$$

7.5.2 Prediction based on stationary point of the energy

The Boltzmann-Gibbs PDF of the field at the prediction sites conditional on the data is given by $\exp\left[-\mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*})\right]/Z(\boldsymbol{\theta}^{*})$. The prediction $\hat{\mathbf{x}}_{\mathbb{G}}$ maximizes the PDF, which is equivalent to minimizing the energy, i.e.,

$$\hat{\mathbf{x}}_{\mathbb{G}} = \arg\min_{\mathbf{x}_{\mathbb{G}}} \mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*}). \tag{7.23}$$

The SLI energy (Eq. (7.20)) can be further expressed in terms of the precision matrix as follows

$$\mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*}) = \mathcal{H}_s(\mathbf{x};\boldsymbol{\theta}^{*}) + \frac{1}{2}\left(\mathbf{x}'^{\top}_{\mathbb{G}}\mathbf{J}_{\mathbb{G},\mathbb{S}}\mathbf{x}' + \mathbf{x}'^{\top}\mathbf{J}_{\mathbb{S},\mathbb{G}}\mathbf{x}'_{\mathbb{G}} + \mathbf{x}'^{\top}_{\mathbb{G}}\mathbf{J}_{\mathbb{G},\mathbb{G}}\mathbf{x}'_{\mathbb{G}}\right),$$

where $\mathcal{H}_s(\mathbf{x};\boldsymbol{\theta}^{*}) = \mathbf{x}'^{\top}\mathbf{J}_{\mathbb{S},\mathbb{S}}\,\mathbf{x}'/2$ depends only on the data and is thus irrelevant for the prediction. The condition for a stationary point of the energy function is

$$\frac{\partial \mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*})}{\partial x'_p} = 0, \text{ for all } \tilde{\mathbf{s}}_p \in \mathbb{G}. \tag{7.24}$$

The Hessian of the energy is $\nabla'\nabla'\mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*})$, where the prime denotes differentiation with respect to $\mathbf{x}'$. For the stationary point to represent a minimum of the energy (and thus a maximum of the Boltzmann-Gibbs PDF), $\nabla'\nabla'\mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*})$ must be positive definite. From Eq. (7.24) it follows that $\nabla'\nabla'\mathcal{H}(\mathbf{x},\mathbf{x}_{\mathbb{G}};\boldsymbol{\theta}^{*}) = \mathbf{J}_{\mathbb{G},\mathbb{G}}$.

Since the SLI precision matrix is positive definite by construction, so is the Hessian as well.

Finally, the SLI prediction is given by the following equation

$$\hat{\mathbf{x}}_{\mathbb{G}}(\boldsymbol{\theta}^{*}|\mathbf{x}) = \mathbf{m}_{\mathrm{x}} - \mathbf{J}^{-1}_{\mathbb{G},\mathbb{G}}(\boldsymbol{\theta}'^{*})\,\mathbf{J}_{\mathbb{G},\mathbb{S}}(\boldsymbol{\theta}'^{*})\,\mathbf{x}', \tag{7.25}$$

where $\mathbf{m}_{\mathrm{x}}$ is the $P \times P$ diagonal trend matrix, i.e., $[\mathbf{m}_{\mathrm{x}}]_{p,q} = \delta_{p,q} m_{\mathrm{x}}(\mathbf{s}_p, t_p)$ and the precision matrices $\mathbf{J}_{\mathbb{G},\mathbb{G}}$ and $\mathbf{J}_{\mathbb{G},\mathbb{S}}$ are defined by means of Eq. (7.21) and Eqs. (7.22c)–(7.22f).

Note that due to the matrix product $\mathbf{J}^{-1}_{\mathbb{G},\mathbb{G}}\,\mathbf{J}_{\mathbb{G},\mathbb{S}}$ and in light of Eq. (7.21) the SLI prediction is independent of the scale parameter λ. This property is analogous to the independence of the kriging prediction from the variance, since the latter is proportional to λ.

7.5.3 Prediction intervals

Since the precision matrix of the SLI model is known, it is straightforward to obtain the *conditional variance* at the prediction sites using the result known in Markov random field theory [Rue and Held, 2005]. Hence,

$$\sigma^2_{\mathrm{SLI}}(\tilde{\mathbf{s}}_p) = \frac{1}{J_{p,p}(\boldsymbol{\theta}^{*})}, \; \tilde{\mathbf{s}}_p \in \mathbb{G}, \tag{7.26}$$

where $J_{p,p}(\boldsymbol{\theta}^{*})$ is the p-th diagonal entry of the precision matrix $\mathbf{J}_{\mathbb{G},\mathbb{G}}$ which is determined from Eqs. (7.21) and (7.22f).

Based on the above, *prediction intervals* at the site $\tilde{\mathbf{s}}_p \in \mathbb{G}$ can be constructed as follows

$$[\hat{x}_p - z_q \sigma_{\mathrm{SLI}}(\tilde{\mathbf{s}}_p),\; \hat{x}_p + z_q \sigma_{\mathrm{SLI}}(\tilde{\mathbf{s}}_p)],$$

where $0 \le q \le 1$ is a specified level (e.g., $q = 0.95$), and z_q is the respective quantile of the standard normal distribution.

7.6 Parameter Estimation

We use Leave-one-out cross validation to estimate the SLI model parameter vector Eq. (7.17). The orders of the spatial and temporal neighbors K_s and K_t are set in advance to low integer values larger than one. This does not have a significant impact on the results, since the bandwidth parameters μ_s, μ_t compensate for the choice of the neighbor order [Hristopulos and Agou, 2020; Hristopulos et al., 2021].

The *cost function* optimized with respect to the parameters is the root mean square error (RMSE) of the predictions. The latter are based on the SLI prediction equation (7.25); for each sampling point the prediction is based on the remaining $N - 1$ points. Hence, the RMSE is given by

$$\text{RMSE}(\tilde{\boldsymbol{\theta}}) = \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left[\hat{x}_n(\tilde{\boldsymbol{\theta}} \mid \mathbf{x}_{-n}) - x_n\right]^2}, \tag{7.27}$$

where $\mathbf{x}_{-n}$ denotes the data vector $\mathbf{x}$ from which the point x_n is removed. Note that the dependence of the cost function on the bandwidth controlling parameters μ_s and μ_t is highly nonlinear, due to the presence of the latter in kernel function sums both in the numerator and the denominator of the normalized kernel weights (7.15a).

The optimal parameter vector $\tilde{\boldsymbol{\theta}}^*$ is then determined by means of

$$\tilde{\boldsymbol{\theta}}^* = \arg\min_{\tilde{\boldsymbol{\theta}}} \text{RMSE}(\tilde{\boldsymbol{\theta}}).$$

Finally, the value of the scale parameter is selected to set equal to the value that maximizes the likelihood of the SLI model given the other parameters, i.e., [Hristopulos, 2015a]

$$\lambda^* = \frac{2\mathcal{H}(\mathbf{x}; \boldsymbol{\theta}^*_{-\lambda})}{N}, \tag{7.28}$$

where

$$\boldsymbol{\theta}^*_{-\lambda} = (m_{\mathrm{x}}{}^*, 1, K_s, K_t, \mu_s^*, \mu_t^*, c_1^*)^\top .$$

7.7 Application to Timeseries

The following section presents results based on the SLI algorithm to three temporal data sets which involve temperature, solar radiation on horizontal plane, and precipitation data from the ERA5 data set. We investigate SLI interpolation and nearest neighbor (NN) interpolation for the reanalysis data (temperature in degrees Celsius, solar irradiance in Watt per square meter, and precipitation in mm) for every location of the grid on the island of Crete. For brevity we focus on results from a representative node (see Fig. 3.1) located near the Messara valley, which is the major agricultural area of Crete [Varouchakis and Hristopulos, 2013]. The coordinates (LAT, LON) of the test node in the World Geodetic System (WGS 84) are 35°0'0.00"N and 25°0'0.00"E. Additionally, we present the cross validation results of all the timeseries together in violin plots in order to capture the variability for the entire investigation. Each one of the timeseries corresponds to the hourly values across a year for one location.

In order to evaluate the models, we use partial timeseries from the ERA5 data on the island of Crete. The scenario corresponds to a timeseries set where we randomly remove 10% of the data with the restriction that after the removal the missing values do not correspond to more than six continuous time steps. We use 2 535 timeseries for the wet season hourly values for the precipitation data (39 years (1980–2018) for 65 locations), with each of the series containing 3 931 for the training set and 437 values for the validation set. We use 2 600 timeseries for the entire season hourly values for the temperature data (40 years (1980–2019) for 65 locations), with each of the series containing 7 884 for the training set and 876 values for the validation set. And lastly, we use 2 600 timeseries for the entire season hourly values for the solar irradiance data (40 years (1980–2019) for 65 locations), with each of the series containing 7 884 for the training set and 876 values for the validation set.

7.7.1 Scenario analysis for the Test location

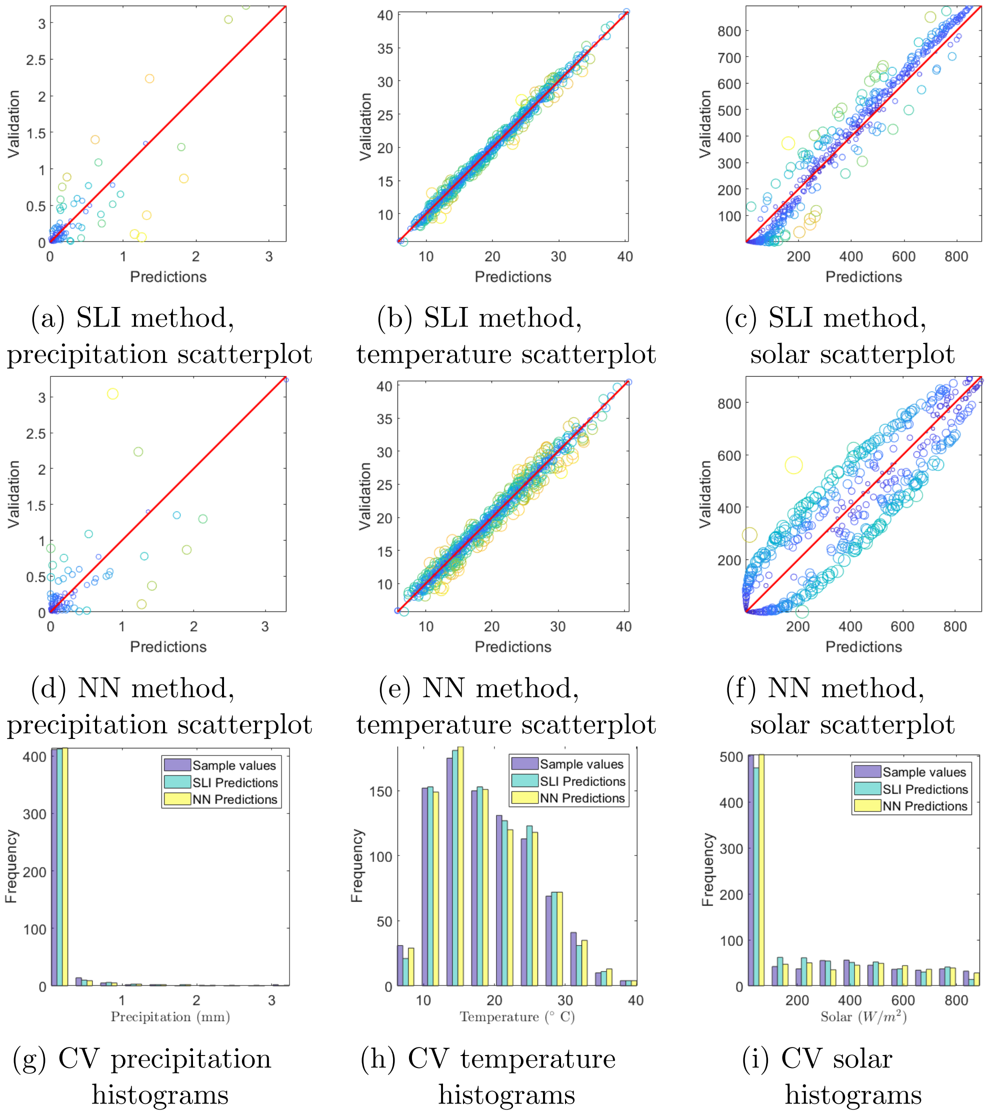
The entire data set used for the Scenario includes 4 368 hourly precipitation values distributed through the wet period of the year 2008 (01-10-2008 00:00:00 to 31-03-2009 23:00:00) at the test location, with missing values that add up to 437 values. Respectively, for the temperature and the solar radiation data, the data set includes 8 760 hourly values for the year 2008 (01-01-2008 00:00:00 to 31-12-2008 23:00:00), with 876 missing values for each variable.

The scatter plot of the SLI predictions for the Scenario is shown in Fig. 7.2 and exhibits good agreement between the predictions and the data, especially for the temperature and solar radiation variables. The histogram plots of the predicted versus the sample values appear to be quite similar for both the SLI and the NN, for all the variables. The cross validation measures are shown in Table 7.2 and confirm that the interpolation performance for the SLI model is better in most cases.

According to the validation results, SLI is a better method for filling the gaps particularly for the temperature and the solar radiation data sets. Nevertheless, for the nearest neighbor interpolation cases, the mean error for all the variables is lower versus those calculated with the SLI, while for the precipitation data Spearman's coefficient is also higher for the nearest neighbor interpolation case. It should be noted that the ME in all cases is a very small fraction of the variable's mean value, meaning that both predictors seem to be unbiased. Additionally, the SLI method is superior to the NN interpolation because it gives an estimation of the error variance and it allows variability on the kernel distances, overcoming the problem of estimating consecutive missing values. This problem was encountered during the application of the NN interpolation, leading to estimation of some and not all of the missing values. Thus, to avoid the estimation of NaN correlation coefficients, their estimation is carried out from the values that had finite estimations, resulting in potentially biased and higher values.

7.7.2 Results for the temporal data

In the following section we present the results for the entire investigation of the SLI methodology for the precipitation, temperature and solar radiation data for the complete grid. As mentioned previously, the collection of the results corresponds to 2 535 timeseries for the precipitation data, and to 2 600 timeseries for the temperature as well as the solar radiation data. The investigation was performed with a training set that includes 90% of the data and the remaining



(a) SLI method, precipitation scatterplot

(b) SLI method, temperature scatterplot

(c) SLI method, solar scatterplot

(d) NN method, precipitation scatterplot

(e) NN method, temperature scatterplot

(f) NN method, solar scatterplot

(g) CV precipitation histograms

(h) CV temperature histograms

(i) CV solar histograms

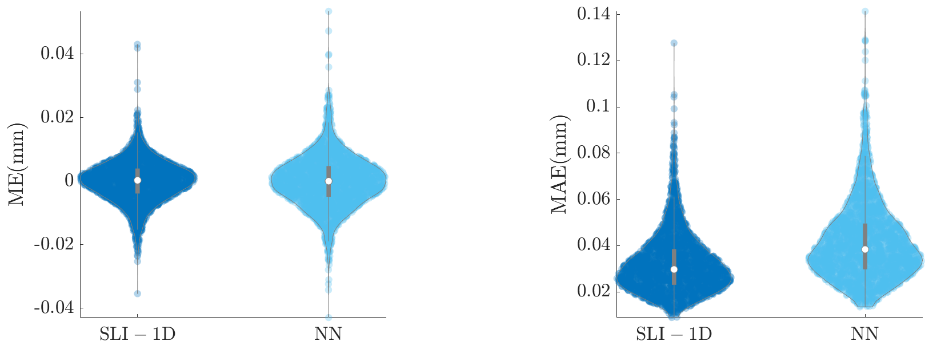
Figure 7.2: Top: scatterplots for the SLI predictions versus the real values for precipitation (left), temperature (middle) and solar (right) data for the test location in Crete. Middle: scatterplots for the nearest neighbor interpolation predictions versus the real values for precipitation (left), temperature (middle) and solar (right) data for the same location. Bottom: histograms for the real precipitation (left), temperature (middle) and solar (right) values and both the SLI and nearest neighbor interpolation predictions. All the data correspond to the test location for the year 2008. For more information on the data see Section 7.7.1.

LAT, LONG: 35°0’0.00”N, 25°0’0.00”E						
	SLI interpolation			Nearest neighbor interpolation		
Measure	Precipitation	Temperature	Solar Radiation	Precipitation	Temperature	Solar Radiation
ME	0.0031	0.0408	−0.9098	−0.0019	0.0222	−0.4323
MAE	0.0411	0.3975	27.2329	0.0469	0.6653	55.6723
MARE	NA	0.0239	NA	NA	0.0389	NA
RMSE	0.1468	0.5470	42.5869	0.1801	0.9174	86.5416
RMSRE	NA	0.0352	NA	NA	0.0569	NA
RP	0.8663	0.9967	0.9899	0.8029	0.9905	0.9497
RS	0.8899	0.9965	0.9482	0.9276	0.9903	0.9419
ErrMin	−1.1892	−2.7662	−176.1218	−1.1654	−3.6342	−214.0267
ErrMax	0.8728	2.7159	212.1669	2.1720	3.0869	377.0311

Table 7.2: Cross validation (CV) interpolation performance for the ERA5 time-series (precipitation, temperature and solar radiation data). The CV measures are calculated by comparing the true variable values of each missing hour and either the SLI predictions (2nd, 3rd and 4th column), or the nearest neighbor interpolation predictions (5th, 6th and 7th column). NA: not applicable; the respective measures involve division by zero or by a very small number. Note: For the NN interpolation, the RP and RS coefficients are calculated after the removal of a set of NaN values obtained at locations with insufficient number of neighbors.

10% is used as the validation set as described in the previous section.

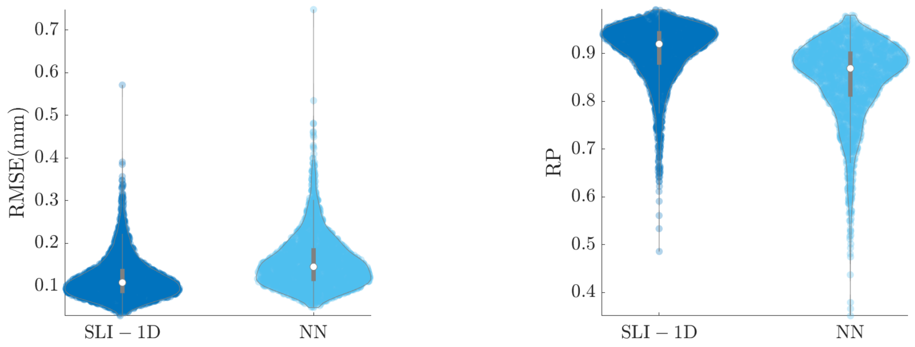
The results are presented in the form of violin plots per variable for the two methods. The ME for the wet-season hourly precipitation (Figs. 7.3, and 7.4) is comparable between the methods with smaller dispersion for the SLI method. Additionally, MAE, RMSE and RS coefficient indicate better performance for the SLI methodology both in terms of the mean value and their dispersion. Similar results were obtained for the hourly temperature (Figs. 7.5 and 7.6). For the hourly solar radiation (Figs. 7.7, and 7.8) the ME has greater dispersion with higher median value in the case of NN method, while the rest of the measures (MAE, RMSE, RP) have significant differences between the methodologies, indicating superior performance with the SLI approach.



(a) ME

(b) MAE

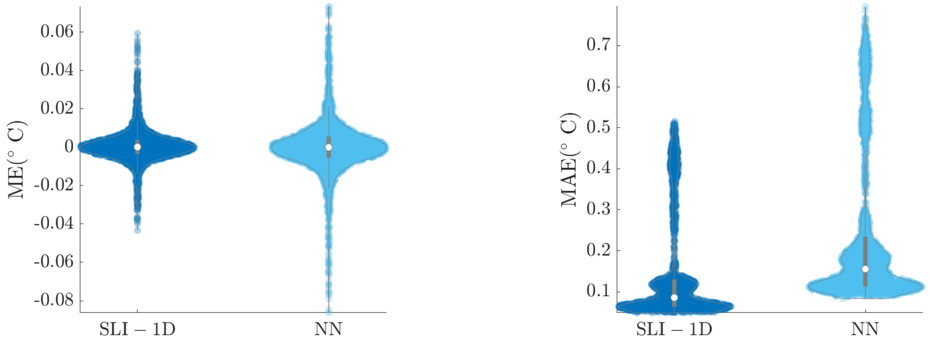
Figure 7.3: SLI-1D vs NN LOO-CV mean error (ME) and mean absolute error (MAE) for the wet-season ERA5 hourly precipitation data (1 October 00:00:00 to 31 March 23:00:00) for all the timeseries.



(a) RMSE

(b) RP

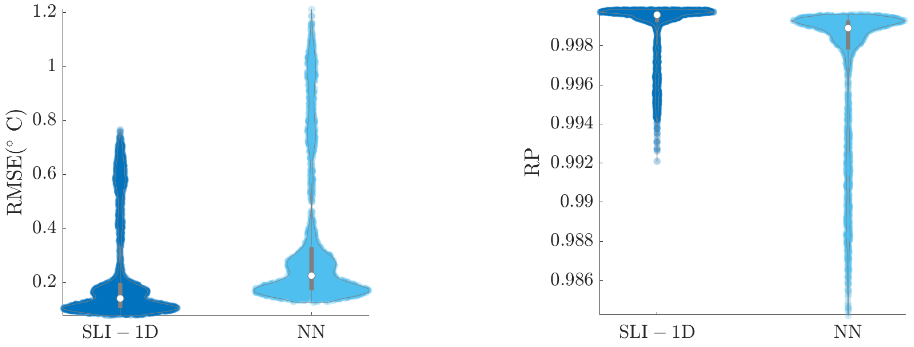
Figure 7.4: SLI-1D vs NN LOO-CV root mean error error (RMSE) and the Pearson correlation coefficient (RP) between the true and predicted values for the wet-season ERA5 hourly precipitation data (1 October 00:00:00 to 31 March 23:00:00) for all the timeseries.



(a) ME

(b) MAE

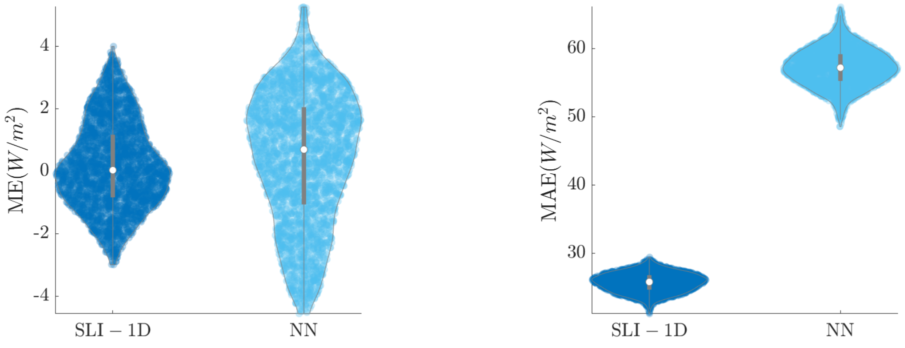
Figure 7.5: SLI-1D vs NN LOO-CV mean error (ME) and mean absolute error (MAE) for the entire-season ERA5 hourly temperature data (1 January 00:00:00 to 31 December 23:00:00) for all the timeseries.



(a) RMSE

(b) RP

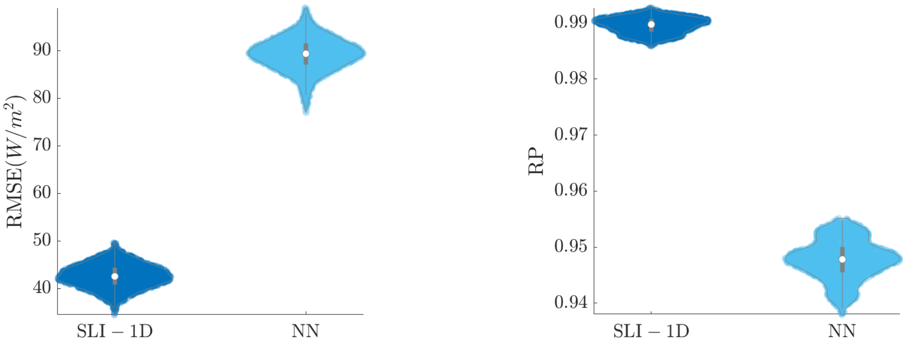
Figure 7.6: SLI-1D vs NN LOO-CV root mean error error (RMSE) and the Pearson correlation coefficient (RP) between the true and predicted values for the entire-season ERA5 hourly temperature data (1 January 00:00:00 to 31 December 23:00:00) for all the timeseries.



(a) ME

(b) MAE

Figure 7.7: SLI-1D vs Nearest Neighbor LOO-CV mean error (ME) and mean absolute error (MAE) for the entire-season ERA5 hourly solar radiation on horizontal plane data (1 January 00:00:00 to 31 December 23:00:00) for all the timeseries.



(a) RMSE

(b) RP

Figure 7.8: SLI-1D vs NN LOO-CV root mean error error (RMSE) and the Pearson correlation coefficient (RP) between the true and predicted values for the entire-season ERA5 hourly solar radiation on horizontal plane data (1 January 00:00:00 to 31 December 23:00:00) for all the timeseries.

7.8 Application to ST Data

7.8.1 Hourly precipitation reanalysis data

We investigate SLI-based interpolation for reanalysis ST hourly precipitation data (mm) in Crete. The data set includes 10 920 points that correspond to hourly values for seven consecutive days (December 25-31, 2008) at the 65 nodes of the spatial grid around the island of Crete (Greece) as shown in Fig. 3.1. The data are displayed as time series in Fig. 7.9. Hristopulos and Agou [2020] offer a more extensive analysis of the ST-SLI methodology with its application to synthetic (simulated) data, reanalysis ST temperature data, and ozone measurements over France.

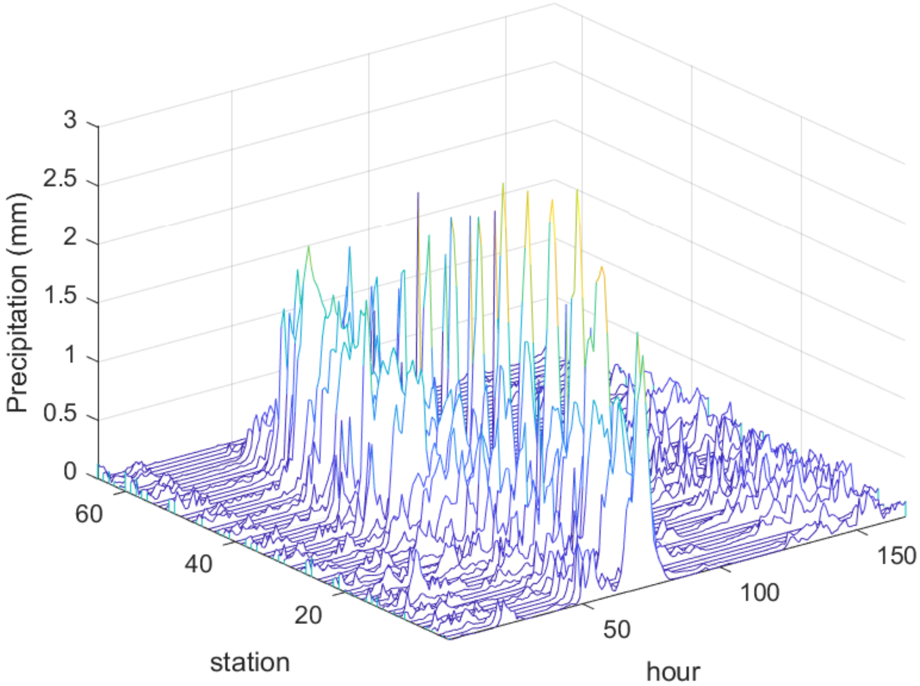


Figure 7.9: Time series of precipitation (in mm) at the ERA5 grid sites shown in Fig. 3.1. The hourly values correspond to a time period of seven days spanning from 25 to 31 December of 2008.

SLI parameter estimation using LOOCV

The optimal model parameters are estimated with LOOCV by minimizing the RMSE between the estimations and the real values. The orders of the spatial and temporal neighbors are set to $K_s = K_t = 3$. The initial guesses for the SLI parameters and the parameter bounds are given in Table 7.3. The value of the cost function for the optimal SLI parameters is $\approx 8.0793 \times 10^{-2}$. The values of the optimal SLI parameters are listed in Table 7.3. The precision matrix has a sparsity index $\approx 0.027\%$, corresponding to 32 890 non-zero entries out of 119 246 400 entries.

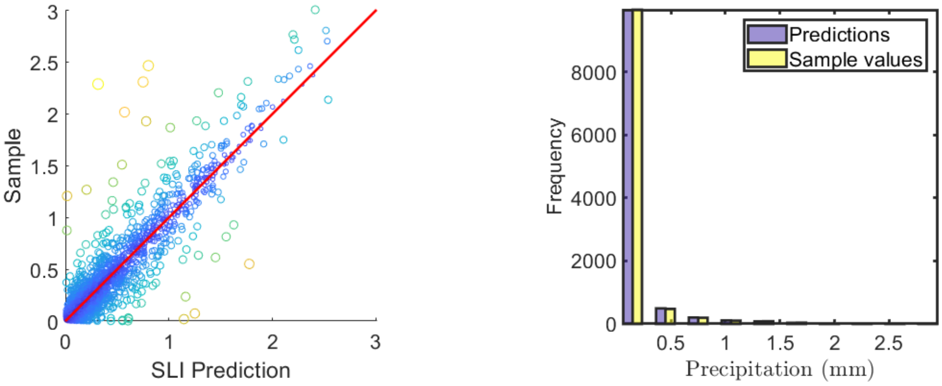
	m_{x}	λ	μ_t	μ_s
Initial values	0.1029	300	3	2.5
Lower bounds	0	1	2.2e−16	2.2e−16
Upper bounds	0.6134	1000	10	10
Based on LOOCV	0.1013	300.0007	0.9533	0.0806

Table 7.3: SLI parameters for the precipitation ST data (25–31 Dec, 2008) based on LOOCV. The lower and upper bounds on the mean are based on $\overline{x} \mp 2\sigma_x$, where $\overline{x}$ is the sample mean and σ_x is the sample standard deviation. We restrict the LB value to zero since the is not possible to have negative precipitation.

SLI model performance

To test the performance of the estimated SLI model we use one-slice-out cross validation: we remove and subsequently predict all the values for one time slice using the sample values at the $N_t - 1$ remaining time slices. We repeat this experiment by removing sequentially all the time slices, one at a time. The scatter plot of the predictions (for all N points) versus the sample values (Fig. 7.10a) as well as the histogram of the predictions versus the sample values (Fig. 7.10b) demonstrate overall very good agreement between the two sets.

The validation measures presented in Table 7.4 indicate overall very good performance of the SLI model with small bias $\approx -1.9579 \times 10^{-5}$ and very good correlation ≈ 0.95. The RMSE is ≈ 0.08.



(a) Scatter plot

(b) Histogram

Figure 7.10: (a): Scatter plot of the predictions versus the sample values for the precipitation space-time data (25–31 Dec, 2008). (b): Histograms of the sample (yellow) and SLI-predicted (purple) values.

ME	MAE	MARE	RMSE	RMSRE	RP	RS
−0.00002	0.0286	NA	0.0808	NA	0.9486	0.9412

Table 7.4: One-slice-out cross validation (CV) test of the SLI interpolation performance for the ERA5 precipitation data (25–31 Dec, 2008). The CV measures are calculated by comparing the true precipitation values of each hourly time slice (from 1 to 168) and the SLI predictions that are based on the SLI model with the LOOCV parameters reported in Table 7.3. The predictions are based on $N_t - 1$ time slices excluding the predicted slice.

7.8.2 Hourly temperature reanalysis data

The ST temperature data set includes 10 920 points that correspond to hourly values for seven consecutive days (December 25–31, 2008) at the 65 nodes of the spatial grid around the island of Crete (Greece) as shown in Fig. 3.1. The data are displayed as time series in Fig. 7.11. The temperature data does not exhibit temporal or spatial trend for the studied time period.

The SLI parameter estimation and the performance assessment are carried out as in the precipitation data case study (Section 7.8.1). The parameter estimates for the temperature data are shown in Table 7.5. The precision matrix has a sparsity index $\approx 0.034\%$, corresponding to 40 662 non-zero entries out of

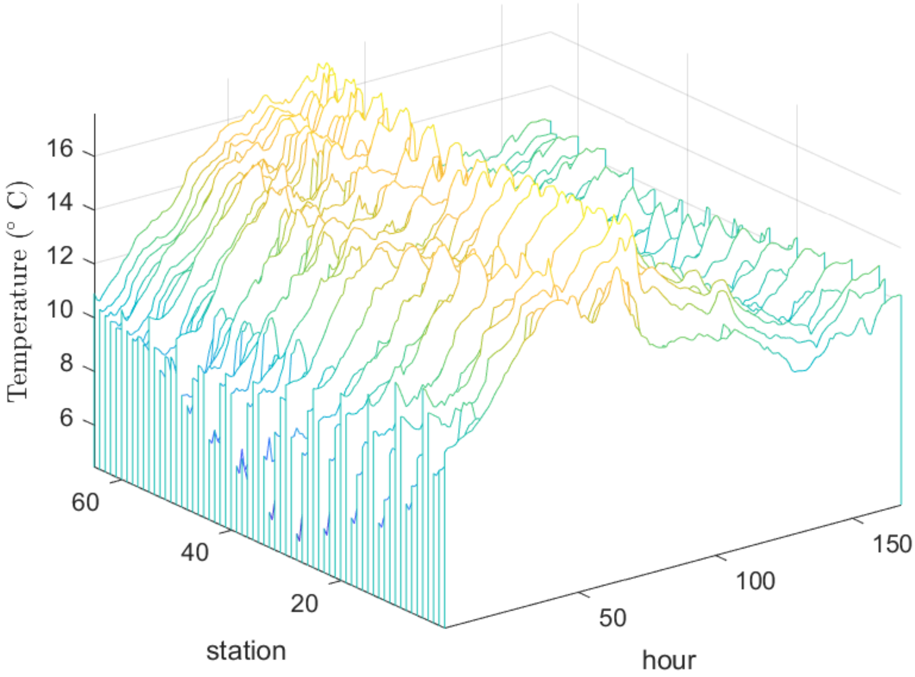


Figure 7.11: Time series of temperature (in degrees Celsius) at the ERA5 grid sites shown in Fig. 3.1. Time period between the 25th and 31st of December of 2008.

119 246 400 entries.

The scatter plot of the predictions (for all N points) is shown in Fig. 7.12a and exhibits good agreement between the predictions and the data. The histogram plots of the predicted versus the sample values (Fig. 7.12b) also show very good performance of the ST-SLI method. The cross validation measures (obtained by sequentially removing each of the 168 hourly time slices) are shown in Table 7.6 and confirm the interpolation performance for the SLI model. The correlation coefficients for the temperature data ($\approx$ 100% correlation) are even higher than the already high values for the precipitation case study ($\approx$ 95% correlation).

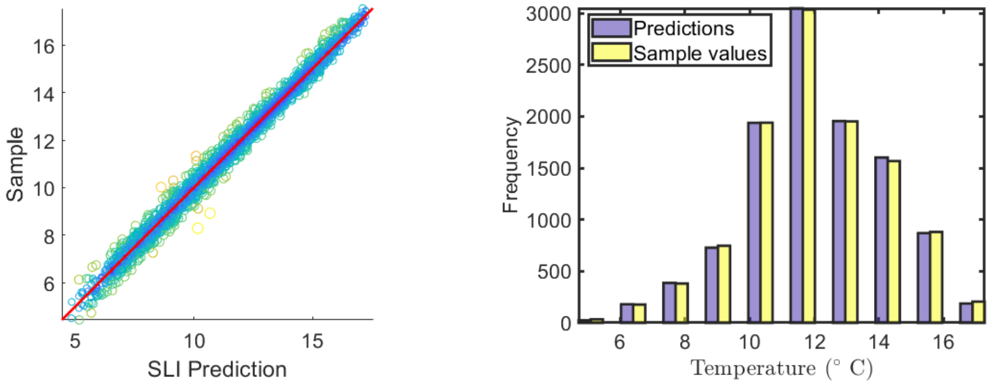
Note that $\mu_s = 0.7099$, which implies that the spatial bandwidth is small but it is higher than the $\mu_s = 0.0806$ for the precipitation data. Similar results we see for the temporal correlation for the temperature data, with $\mu_t \approx 0.7$. On the

	m_{x}	λ	μ_t	μ_s
Initial values	12.016	300	3	2.5
Lower bounds	7.6647	1	2.2e−16	2.2e−16
Upper bounds	16.3670	1000	10	10
Based on LOOCV	12.0300	300.2732	0.6665	0.7099

Table 7.5: SLI model parameters for the ERA5 ST temperature data (25–31 Dec, 2008) based on LOOCV. The lower and upper bounds on the mean are based on $\overline{x} \mp 2\sigma_x$, where $\overline{x}$ is the sample mean and σ_x is the sample standard deviation.

ME	MAE	MARE	RMSE	RMSRE	RP	RS
−0.0015	0.0984	0.0092	0.1647	0.0171	0.9972	0.9976

Table 7.6: One-slice-out cross validation (CV) test of the SLI interpolation performance for the ERA5 temperature data (25–31 Dec, 2008). The CV measures are calculated by comparing the true temperature values of each hourly time slice (from 1 to 168) and the SLI predictions that are based on the SLI model with the LOOCV parameters reported in Table 7.5. The predictions are based on $N_t - 1$ time slices excluding the predicted slice.



(a) Scatter plot

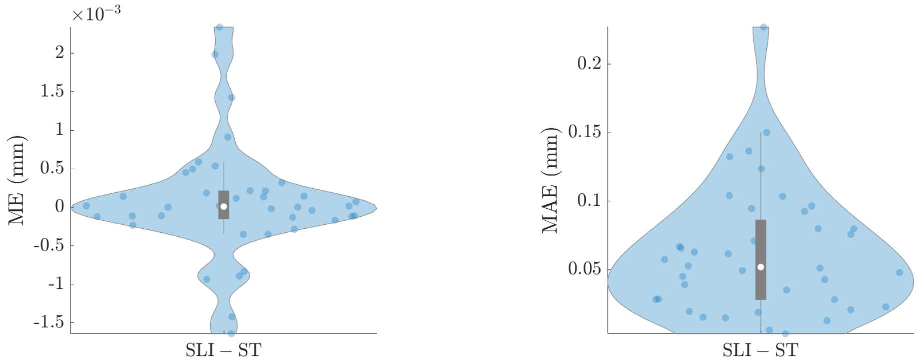
(b) Histogram

Figure 7.12: (a): Scatter plot of the predictions versus the sample values for the temperature space-time data (25–31 Dec, 2008). (b): Histograms of the sample (yellow) and SLI-predicted (purple) values.

other hand, $\mu_t \approx 1$ for the precipitation data set. This implies that the precipitation data for the case study display equal correlation in space and time, and higher correlation in time than the temperature data. In essence, $\mu_t \approx 0.7$ and $K_t = 3$ indicate that the temporal bandwidth is $\mu_t(K_t - 1)\delta t \approx 1.4$ hr ($\delta t = 1$ hour). This result implies that the SLI predictions are at most locations based on the two temporal nearest neighbors (one forward and one backward). Respectively, for $\mu_s \approx 0.7$ and $K_s = 3$ indicate that the temporal bandwidth is $\mu_s(K_s - 1)\delta s \approx 11.66$(dimensionless) ($\delta s = 8.33$ normalized distance which corresponds to 0.25°WGS). This result indicates that the SLI predictions are at most locations based on the 11 nearest neighbors in space. We see this behavior because for the specific area the differences in temperature are not happening extremely fast resulting to values that oscillate smoothly during the day.

7.8.3 ST-SLI results for all the years

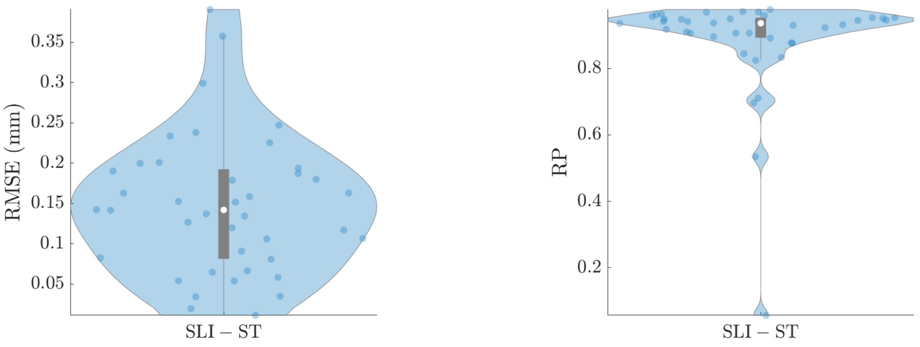
In the following section we present in violin plots the LOOCV measures for all the years for the precipitation (Figs. 7.13, 7.14) and temperature (Figs. 7.15, 7.16) data. The results were produced by performing the ST-SLI methodology for each variable forty times (40 years in total: 1980–2019) for the time period between the 25th and 31st of December of 2008.



(a) ME

(b) MAE

Figure 7.13: LOO-CV mean error (ME) and mean absolute error (MAE) for the ERA5 hourly precipitation data (25–31 Dec) for all the years (1980–2019).

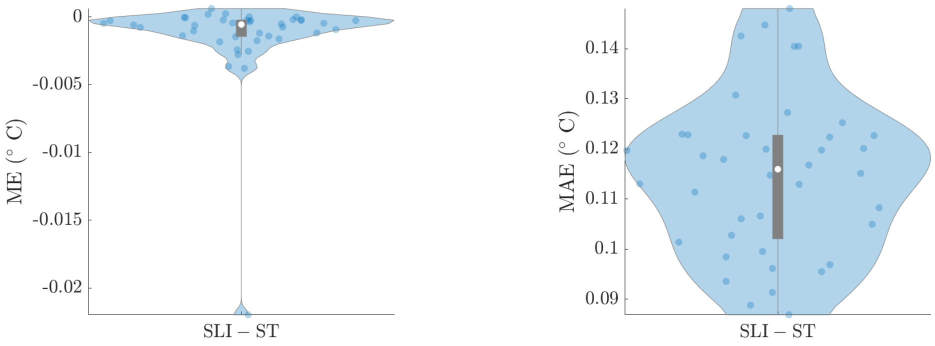


(a) RMSE

(b) RP

Figure 7.14: LOO-CV root mean error error (RMSE) and the Pearson's correlation coefficient (RP) between the true and predicted values for the ERA5 hourly precipitation data (25–31 Dec) for all the years (1980–2019).

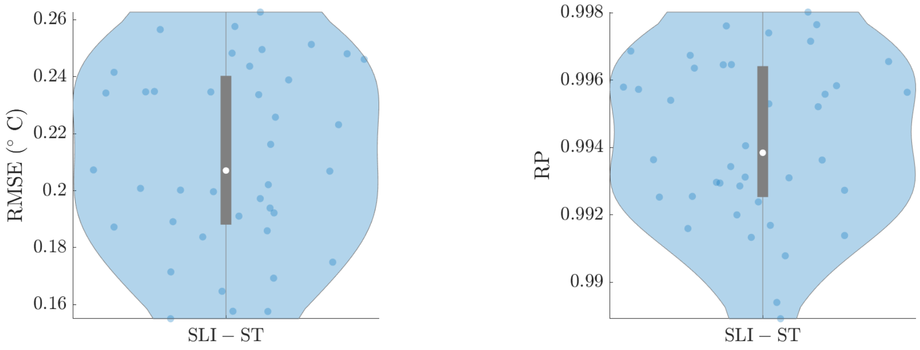
The resulted validation measures from the ST-SLI are better in absolute values than the SLI application in 1D. However, direct comparison between the 1-D and the ST cases is not possible since the range of the values from the 1D cases are greater than in the spatiotemporal cases (values from December only). This applies for both the precipitation and the temperature data. It is impossible to run in the testing computational environment (see configuration in Section 3.5) a data set greater than 15 600 values, which would correspond to hourly values for 10 days for the 65 node grid of Crete.



(a) ME

(b) MAE

Figure 7.15: LOO-CV mean error (ME) and mean absolute error (MAE) for the ERA5 hourly temperature data (25–31 Dec) for all the years (1980–2019).



(a) RMSE

(b) RP

Figure 7.16: LOO-CV root mean error error (RMSE) and the Pearson's correlation coefficient (RP) between the true and predicted values for the ERA5 hourly temperature data (25–31 Dec) for all the years (1980–2019).

7.9 Discussion and Conclusions

Herein, we provide the application of the theoretical framework presented in [Hristopulos and Agou, 2020], which describes ST model construction based on the exponential Boltzmann-Gibbs joint probability density functions. The ST-SLI model takes advantage of an energy function with local interactions, resulting in a sparse structure on the precision matrix. Local interactions are implemented via compactly supported kernel functions, which compensate for the lack of a structured lattice. However, the concept is also applicable to normal lattice data. In this scenario, the SLI model is equivalent to a Gauss Markov random field with a specified precision matrix structure. ST-SLI model extends the purely spatial SLI model [Hristopulos, 2015a] into the spatio-temporal domain. To our knowledge, this structure has never been used before in precipitation ST data. For environmental ST data, the graph topology is determined from the data by LOOCV and is not given a priori.

The sparse precision matrix enables the computationally efficient implementation of parameter estimation and prediction techniques. This is especially true when the domain is irregular and the goal is to create a finer regular estimation map. The computational savings arises from the fact that big and dense covariance matrices are not required to be stored and inverted in SLI.

The optimization of the cost function was based on an interior point algorithm that stops at local minima. The optimization algorithm for the cost function is also capable of searching for the global minimum. From numerous previous experiments [Hristopulos, 2015a; Hristopulos and Agou, 2020; Hristopulos et al., 2021], the estimated parameters from the local minima of the cost function are sufficient for interpolation purposes and many times identical to those derived if the algorithm was run for global minimization.

In our investigation, for the temporal case studies which involve 2 535 timeseries for the precipitation and 2 600 timeseries for the temperature data, the cross validation results (Table 7.2) for the SLI and the NN (Figs. 7.3-7.6) methods are comparable in terms of the ME. However, the rest of the measures (MAE, RMSE, RS) demonstrate better performance for the SLI model. While for the hourly solar radiation in 1D for the test location (Table 7.2) the ME is smaller with the NN than with the SLI method, according to Figs. 7.7-7.8 that represent the entire investigation period, the ME has greater dispersion in the case of NN method. The rest of the measures (MAE, RMSE, RP) have significant differences between the methodologies, indicating superior performance with the SLI approach.

In terms of prediction performance for the ST precipitation data, we have shown (see Table 7.4) that the cross validation statistics indicate overall very good performance of the SLI model with small bias $\approx -1.9579e{-5}$ and very good correlation ≈ 0.95. The cross-validation measures for the ST temperature data (obtained by sequentially removing each of the 168 hourly time slices) are shown in Table 7.6 and confirm the high interpolation performance for the SLI model. The correlation coefficients for the temperature data (≈ 100% correlation) are even higher than the already high values for the precipitation case study (≈ 95% correlation).

The performance comparisons of different methods concerning prediction performance may change depending on the particular data set. In our opinion, the results presented here indicate that SLI is a competitive method for interpolating temporal and spatiotemporal data. Further research can elaborate on the performance of SLI relative to other methods.

The formulation presented here can be extended to multivariate random fields by choosing the energy function appropriately. Additionally, there is a possibility of incorporating different spatial distance metrics in the kernel functions, anisotropy, and periodicity (in space and in time). Furthermore, incorporation of a trend model into the estimation with auxiliary correlated data may provide important information. Extensions such as these may improve the performance of the model but they will additionally increase the computational cost.

Chapter 8

Conclusions

The main objective of this thesis is to provide various methodologies for space-time modeling of potentially large and non-Gaussian space-time data sets. Cutting-edge geostatistical and machine-learning techniques were employed to examine meteorological data that do not follow commonly used parametric distributions. The investigated approaches involve utilizing probabilistic techniques and algorithms based on machine learning to examine non-Gaussian variables in terms of their spatial, temporal, and spatiotemporal characteristics. To demonstrate the validity of our methods, we have conducted experiments at various time scales using 26 surface variables from the ERA5 reanalysis data sets for the island of Crete, Greece.

The ERA5 data reanalysis data set is used for the first time for an extensive, localized analysis of precipitation, temperature, and solar radiation for the island of Crete, Greece in multiple time scales. Two drought indices are used for the estimation of the drought characteristics for the study area. Specifically, we use the ERA5 data for the estimation of the SPI and SPEI drought indices, at six different time scales, and we compare the results in order to assess the effect of warming trends on drought events. We additionally apply classical geostatistical methodologies such as kriging and Nearest Neighbor interpolation, and we compare their results with more novel approaches such as Gaussian Anamorphosis with Hermite polynomials coupled with Monte Carlo simulations and the Stochastic Local Interaction models. In addition, we implement multiple classification problems using a set of 26 meteorological variables for the classification of precipitation occurrence and intensity. In the following paragraphs we briefly summarize the conclusions from each methodological approach.

The dry climate of Crete is verified by the estimation of the Standardized Precipitation Index (SPI) and the Standardized Potential Evapotranspiration Index (SPEI) for various timescales. We analyze the indices for 1-, 3-, 6-, 9-, 12-, and 24-months, but we focus on the 3-months and 12-months timescales that indicate short-term and long-term drought conditions respectively. We opted to use both indices, because the SPI is recommended by the World Meteorological Organization (WMO), it is widely used in Greece, and it does not need any other input variable besides precipitation; on the other hand, the SPEI includes the potential evapotranspiration (which is estimated based on the temperature data), and thus it accounts for the effects of temperature changes.

Initially, we demonstrate that the eastern part of the island is more prone to desertification than the north-western part. Based on the discrepancies between the indices, it is concluded that Heraklion is severely affected by the rising temperature and the consequently increased evapotranspiration due to climate change. This results approximately in 5% increase of drought events when the temperature (SPEI) is considered. We identify multiple drought events in the entire record (1979-2019) with the most severe events during the years 1990–1992, 2000–2002, and 2016–2019, however, the events both wet and dry tend to become more severe and the indices start to differentiate in recent years for the whole island. Additionally, we observe a noticeable increase in the frequency of drought occurrences for all the locations after 2000. The good correlation between SPI and SPEI for the early years of the study at different time scales imply that both indices are well adjusted to the study area. However, the deviations present in the recent decade suggest that a drought index which considers temperature is more appropriate. Both indices provide valuable tools for monitoring and assessment of the drought risks in the study area, they are easy to implement and interpret. Combined with a basic understanding of the underlying climate of the region under inspection they can enable the responsible agencies to formulate suitable management plans.

Additionally, we model precipitation with geostatistical methodologies. We propose the transformation of non-Gaussian variables, in our case monthly precipitation, into Gaussian-distributed values by applying Gaussian Anamorphosis with Hermite polynomials (GAH), to overcome the common problem of Gaussian methods (e.g., kriging) to accurately represent the prediction variance. We provide an extensive analysis of ten scenarios for the spatial interpolation (based on Ordinary kriging) of monthly precipitation that use different methodological configurations. The scenarios include the application or exclusion of GAH with varying polynomial degrees, the utilization of either the exponential or Spartan variogram models, and the incorporation or omission of Monte Carlo simulations.

We show that increasing the polynomial order improves the validation results only slightly, while the incorporation of simulations leads to improved results (compared to the cases without the simulations) only in some cases. The precipitation data sets used here do not follow the Gaussian distribution, and based on our investigation they do not follow consistently any of the commonly used parametric distributions across different months. For non-parametric Gaussian Processes transformations the anamorphosis function adjusts to the characteristics of the data set at hand, providing higher flexibility than closed-form expressions. In our investigations, we obtain comparable—but not improved–approximation accuracy with GAH compared to Ordinary Kriging. The Spartan covariance kernels are found to be more appropriate for the anamorphosis (without using simulations), while the exponential kernel is found to be more suitable for the scenarios that integrate the bootstrap simulations. We believe that GAH can improve the interpolation results in non-Gaussian data, but further investigation to other non-Gaussian data is needed.

We use several machine learning (ML) methods (fine, medium, and coarse classification trees, linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian Support Vector Machines, Boosted Ensemble trees, Bagged Ensemble trees, and Ensemble RUSBoosted trees) to classify imbalanced precipitation ERA5 data. We use 26 variables (predictor variables) to help us classify hourly precipitation data. We investigate the correlation of the precipitation data with the auxiliary variables, and then we construct various classification models for the estimation of missing precipitation values by utilizing the information from the precipitation data as well as the predictor variables. We show that some of the auxiliary variables can be eliminated due to low (almost zero) importance, and that a subset of the auxiliary variables is adequate for prediction. This results in a more compact, easier to interpret model while also reducing the training and estimation times.

We define eight different precipitation classes. However, because of their imbalanced nature, we split the data into two different sets: one containing two classes divided by a specified threshold ("Binary" data set), and another one containing values above the threshold to the corresponding classes ("Only Rain" data set). The models include information from both numerical and categorical variables. Random Forests (RFs) perform best for both the data sets analyzed here ("Binary" and "Only Rain"). RFs extend decision trees by introducing the idea of ensembles of trees and improves the accuracy and uncertainty of other simple decision tree methods. For the "Binary" data set, the Bagged Ensemble Trees Cross-Validation 5 folds was evaluated as the best model. The accuracy of the model is 96.5%, which is definitely an improvement over the 50-50% classification accuracy of random classification. In terms of the accuracy, all of the models performed well. For the "Only Rain" data set, the Bagged Ensemble Trees Cross-Validation 5 folds was also evaluated as the best model. The accuracy of the model is 80.0%, which is definitely an improvement over the 20% expected if the data were classified by chance. Our findings are promising especially for the "Binary" data set. For the "Only Rain" data set a hybrid model might prove more appropriate, since it is characterized by classes that occupy a very small fraction of the entire set. The results drawn herein can prove useful, especially as a first step towards removing and refining the variables needed for a more accurate representation for modeling precipitation data while simultaneously reducing the computational cost of the estimation.

Finally, we apply another ML method, i.e., the Stochastic Local Interaction (SLI) model for the estimation of missing values in precipitation, temperature and solar radiation data. The application of the SLI models avoid the inversion of the covariance matrix needed in kriging methods, because they use local interactions between neighboring sites (and times) to capture the correlations in the data with the help of kernel functions. Additionally, the current implementation takes advantage of sparse precision matrices that only involve couplings between near neighbors, thus allowing computationally efficient parameter estimation and prediction procedures. In our investigations, the temporal case studies involve 2 535 time series for the precipitation and 2 600 time series for the temperature data. The cross validation results for the SLI and the Nearest Neighbors (NN) methods are comparable in terms of the ME. However, the rest of the validation measures (MAE, RMSE, RS) show better performance for the SLI model. For the hourly solar radiation data (2 600 timeseries) in one dimension, the ME shows more dispersion in the case of NN method, while the rest of the measures (MAE, RMSE, RP) exhibit significant differences between the methodologies, indicating superior performance of the SLI approach overall. In terms of prediction performance for the spatiotemporal precipitation and temperature data, the cross validation statistics indicate overall very good performance of the SLI model with small bias and excellent correlation (≈ 95% for precipitation and ≈ 100% for temperature). In our opinion, the results indicate that SLI is a competitive method for interpolating temporal and spatiotemporal data. Further research can elaborate on the performance of SLI relative to other methods.

8.1 Future work

1. A future extension of the indices study, might include the estimation of the PET values from other formulations such as the Penman-Monteith equation [Allen et al., 1998], and the Hargreaves equation [Hargreaves and Samani, 1982]. Furthermore, the use of the log-logistic distribution for the SPEI calculation can be implemented, which is the recommended distribution by the creators of the SPEI [Beguería et al., 2014]. Lastly, GAH (with or without the KCDE [Pavlides et al., 2022]) can be implemented as the first step of the data transformation to the Gaussian distribution in terms of the estimation of drought indices. That way, the fitting to the gamma or the Pearson distribution – which sometimes is inappropriate – can be avoided.

2. For further research of the proposed methodology (GAH combined with simulations and kriging methods) different directions can be followed. Initially, incorporating a trend function can filter out the effect of the altitude

on precipitation. Alternatively or supplementary to that the omnidirectional variogram can be substituted by the anisotropic variogram. Agou et al. [2019] showed that the island of Crete is characterized by spatial precipitation patterns that differentiate from West to East and from North to South. Those extreme patterns are not that prominent in the reanalysis data. Another direction is to treat the entire data set in the space-time continuum. In that case the distances have to be readjusted to take into account the non-constant time step, and then the kernels have to be constructed in a way that they embody the spatiotemporal correlations. To avoid the high computational cost of the covariance matrix inversion, the stochastic local interaction model (SLI) [Hristopulos, 2020; Hristopulos and Agou, 2020; Hristopulos et al., 2021] can be used instead of kriging. SLI employs sparse precision matrices to represent space-time correlations in such a way that results in highly sparse matrices, resulting to less computational stress. Another approach is to estimate the probability distribution of a continuously-valued variable, such as precipitation amount, with a kernel-based estimator (KCDE) like the one presented by Pavlides et al. [2021]. This technique avoids the disadvantages of a the step function (empirical CDF) by using the kernel-based estimator which is a continuous function. The KCDE method targets the CDF instead of the PDF [Harrold et al., 2003; Mosthaf and Bárdossy, 2017; Sharma and Lall, 1999] and is presented in their study by means of synthetic data sets and reanalysis precipitation data from the Mediterranean island of Crete (Greece).

3. The construction of a hybrid model for the classification of the precipitation data is a natural extension of the machine learning methods investigation.

4. The formulation presented here for the SLI models can be extended to multivariate random fields by choosing the energy function appropriately. Additionally, there is a possibility of incorporating different spacial distance metrics in the kernel functions, anisotropy, and periodicity (in space and in time), as well as trend model for the incorporation of information from auxiliary variables. Extensions such as these may improve the performance of the model but they will additionally increase the computational cost.

Appendix A

A1 Probability distribution models

The theoretical PDFs of commonly used probability distributions for precipitation data modeling are given by the following Eq. (A.1), and they include the Gaussian (for comparison), the gamma (GAM), the Generalized Extreme Value (GEV), the lognormal (LGN), the Weibull (WBL), the Pearson Type-III (P-III), and the Pareto Type-II (P-II) models.

$$\text{Gaussian: } f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{A.1a}$$

$$\text{gamma: } f(x;\xi,\sigma) = \frac{x^{\xi-1} e^{-\frac{x}{\sigma}}}{\sigma^{\xi}\Gamma(\xi)}, \qquad x > 0 \text{ and } \xi,\sigma > 0, \tag{A.1b}$$

$$\begin{aligned} &\text{GEV: } f(x;\mu,\sigma,\xi) = \frac{1}{\sigma} y(x)^{\xi+1} \exp\left[-y(x)\right], \quad x \in \mathbb{S}(\xi), \\ &\text{where } y(x) = \begin{cases} \left[1 + \xi\left(\dfrac{x-\mu}{\sigma}\right)\right]^{-1/\xi} & \xi \neq 0, \\ \exp\left(-\dfrac{x-\mu}{\sigma}\right) & \xi = 0, \end{cases} \end{aligned} \tag{A.1c}$$

$$\text{lognormal: } f(x;\mu,\sigma) = \frac{1}{x\,\sigma\sqrt{2\pi}} \exp\left\{\frac{-(\log x - \mu)^2}{2\sigma^2}\right\}, \quad x > 0,\ \sigma > 0\,, \tag{A.1d}$$

$$\text{Weibull: } f(x;\sigma,\xi) = \begin{cases} \dfrac{\xi}{\sigma}\left(\dfrac{x}{\sigma}\right)^{\xi-1} e^{-(x/\sigma)^{\xi}}, & x \geq 0, \\ 0, & x < 0, \end{cases} \tag{A.1e}$$

$$\text{Pearson Type-III: } f(x;\mu,\sigma,\xi) = \begin{cases} \dfrac{(x-c)^{\alpha-1}\, e^{-(x-c)/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)}, & \xi > 0,\ c \leq x < \infty, \\ \dfrac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, & \xi = 0,\ -\infty < x < \infty, \\ \dfrac{(c-x)^{\alpha-1}\, e^{-(c-x)/\beta}}{\beta^{\alpha}\,\Gamma(\alpha)}, & \xi < 0,\ -\infty < x \leq c, \end{cases} \tag{A.1f}$$

$$\text{Pareto Type-II: } f(x;\mu,\sigma,\xi) = \begin{cases} \dfrac{\xi}{\sigma}\left[1 + \dfrac{x-\mu}{\sigma}\right]^{-(\xi+1)}, & x \geq \mu, \\ 0, & x < \mu\,. \end{cases} \tag{A.1g}$$

In the above expressions, $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ is the scale parameter that controls the dispersion of the distribution, and ξ is a shape parameter. $\Gamma(\cdot)$ is the gamma function, while the Pareto Type-II becomes the Lomax distribution for $\mu = 0$ [Lomax, 1954]. Note that for $\xi = 1$, the gamma distribution defaults to the exponential model.

For the GEV model, ξ determines the type (I, II, or III) of the distribution. The Gumbel distribution (Type-I) is obtained for $\xi = 0$, the Fréchet distribution (Type-II) for $\xi < 0$, and the Reverse-Weibull distribution (Type-III) for $\xi > 0$ [de Haan and Ferreira, 2010]. The support $\mathbb{S}(\xi)$ of the GEV distribution is $\mathbb{S}(\xi) = [\mu - \sigma/\xi, \infty)$ for $\xi > 0$, $\mathbb{S}(\xi) = (-\infty, \infty)$ for $\xi = 0$, and $\mathbb{S}(\xi) = (-\infty, \mu - \sigma/\xi]$ for $\xi < 0$. For $\xi \in (-0.278, 1)$ the GEV distribution is positively skewed; it has a finite mean given by $m = \mu + \sigma\frac{\Gamma(1-\xi)-1}{\xi}$ for $\xi \neq 0$, and $m = \mu + \sigma\gamma_E$ for $\xi = 0$, where $\gamma_E \approx 0.5772$ is the Euler-Mascheroni constant [Scheuerer, 2014].

For the Pearson Type III distribution, $\alpha = 4/\xi^2$, $\beta = \frac{1}{2}\sigma|\xi|$, and $c = \mu - 2\sigma/\xi$. Note that the P-III becomes the normal distribution for $\xi = 0$, the exponential when $\xi = 2$, and the reverse exponential when $\xi = -2$. Usually the case of $\xi > 0$ is considered as the Pearson type III distribution [Hosking and Wallis, 1997].

The respective CDFs of the Gaussian, gamma, GEV, lognormal, Weibull, Pearson type-III, and Pareto type-II models are given below:

$$\text{Gaussian: } F(x;\mu,\sigma^2) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{A.2a}$$

$$\text{gamma: } F(x;\xi,\sigma) = \int_0^x f(u;\xi,\sigma)\,\mathrm{d}u = \frac{\gamma(\xi,\frac{x}{\sigma})}{\Gamma(\xi)}, \tag{A.2b}$$

$$\text{GEV: } F(x;\mu,\sigma,\xi) = \exp\left[-y(x)\right], \quad x \in \mathbb{S}(\xi), \tag{A.2c}$$

$$\text{lognormal: } F(x;\mu,\sigma) = \frac{1}{2}\left\{1 + \operatorname{erf}\left(\frac{\log x - \mu}{\sigma\sqrt{2}}\right)\right\}, \quad \text{for } x > 0, \tag{A.2d}$$

$$\text{Weibull: } F(x;\sigma,\xi) = 1 - e^{-(x/\sigma)^{\xi}}, \quad \text{for } x > 0\,, \tag{A.2e}$$

$$\text{Pearson Type-III: } F(x;\mu,\sigma,\xi) = \int_0^x f(u;\mu,\sigma,\xi)\,\mathrm{d}u =$$
$$\begin{cases} \dfrac{\gamma(\alpha,\frac{x-c}{\beta})}{\Gamma(\alpha)}, & \xi > 0,\ c \le x < \infty, \\ \dfrac{1}{2}\left[1 + \operatorname{erf}\left(\dfrac{x-\mu}{\sigma\sqrt{2}}\right)\right], & \xi = 0,\ -\infty < x < \infty, \\ 1 - \dfrac{\gamma(\alpha,\frac{c-x}{\beta})}{\Gamma(\alpha)}, & \xi < 0,\ -\infty < x \le c, \end{cases} \tag{A.2f}$$

$$\text{Pareto Type-II: } F(x;\mu,\sigma,\xi) = 1 - \left[1 + \left(\frac{x-\mu}{\sigma}\right)\right]^{-\xi}, \quad \text{for } x \ge \mu. \tag{A.2g}$$

The function $y(x)$ used in Eq. (A.2c), is defined in Eq. (A.1c), $\gamma(\xi,\cdot)$ represents the lower incomplete gamma function, and $\mathrm{erf}(\cdot)$ is the error function [Abramowitz et al., 1988]. For the gamma and lognormal distributions $F(0;\cdot,\cdot) = 0$, while for the GEV finite CDF values are possible for $x < 0$. Since such values are not acceptable for precipitation amounts, one sets $F(x < 0) = 0$ [Scheuerer, 2014]. The respective PDFs and CDFs are shown in Figs. A1 and A2 respectively.

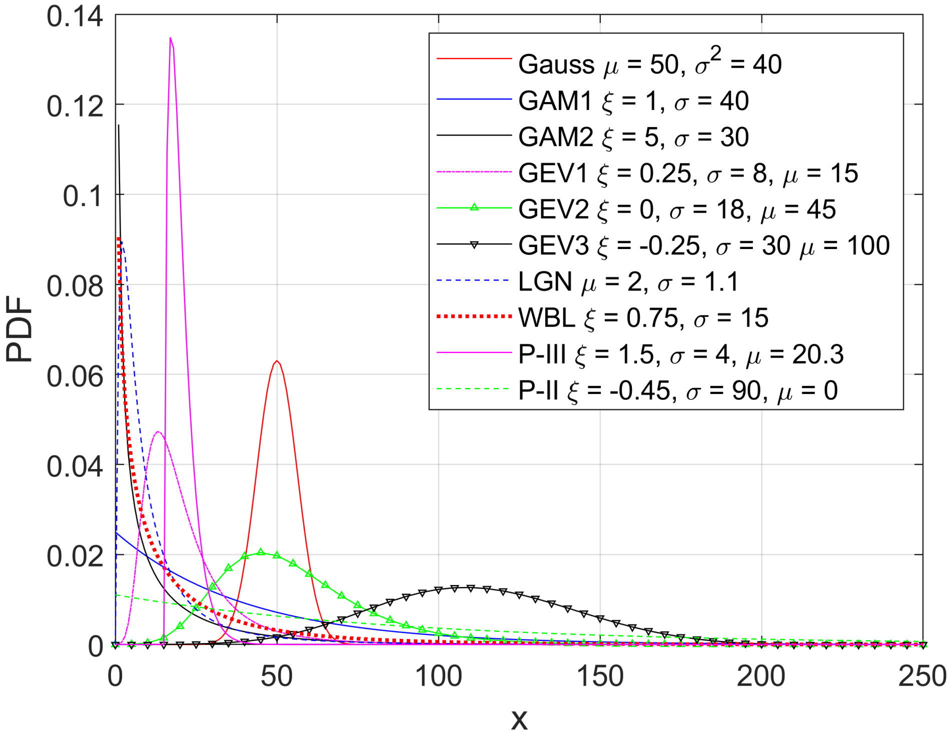


Figure A1: PDF plots for the seven probability models [see Eq. (A.1)]. The horizontal axis represents synthetic monthly precipitation amount (mm). The parameters ξ, σ, μ for each model are defined with reference to Eq. (A.1).

A2 Normality tests

In order to define how different the under investigation data are from Gaussian distributed data, various tests can be used. They are sorted into two main cat-

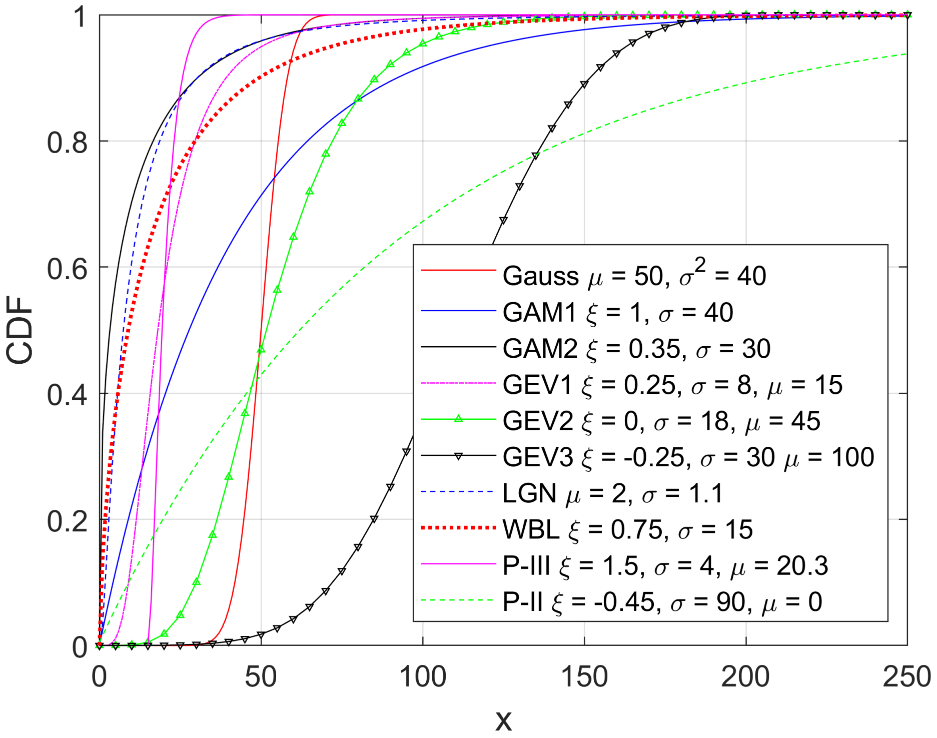


Figure A2: CDF plots for the seven probability models [see Eq. (A.2)]. The horizontal axis represents synthetic monthly precipitation amount (mm). The parameters ξ, σ, μ for each model are defined with reference to Eq. (A.2).

egories: the visualization techniques, and the statistical inference (Hypothesis Testing).

A2.1 Visualization techniques

Visualization techniques that are used to interpret the normality of a data set include the histogram, the BoxPlot and the QQ plot.

- The histogram shows the frequency of a specific value to occur in the data set. In a histogram the center (i.e., location) of the data, the spread (i.e., scale), the skewness, possible extreme values, and the mode values are shown. In the case of a gaussian distribution, the mean, the median and the mode values are the same. A supplementary line with the fitted Gaussian distribution with the sample's mean and standard deviation values can be overlaid on the same plot.

- The BoxPlot is more difficult to interpretation than the histogram, but includes 5 different values in one graph and can be used for multiple variables. The values that a BoxPlot carries are the median (the central mark), the 25th and 75th percentiles (bottom and top edges of the box), the most extreme values (where the horizontal lines (whiskers) extend), and the outliers that fall outside the whiskers (the furthest marks '+').
- Finally, the QQ plot is a graphical technique that carries information about the quantiles of the sample data versus the theoretical quantile values from a distribution (for normality tests, the Gaussian distribution). It also marks with a different line type the first through the third quantiles of the data, and the ends of the data. This visualization technique is more difficult to understand totally, but the main idea is that the closer the quantiles of the sample are to the straight line, the more confident we can be that the data follows the Gaussian distribution. If the quantile values deviate in the tails, a hypothesis test must be carried out.

Visualization techniques are easy and quick ways to detect if there is a need for more complicated tests to be implemented, since in some cases the deviation of a data set from the Normal distribution is so obvious to be detected.

A2.2 Statistical Testing

Several hypothesis testing approaches have been developed for testing whether a sample comes from a specified distribution function. Some commonly used tests include the Kolmogorov-Smirnov, Jarque-Bera, Lilliefors, Anderson-Darling, Cramer Von-Mises, and Shapiro–Wilk goodness-of-fit tests.

- The Kolmogorov-Smirnov test (K-S) measures the difference between the empirical cdf of the sample data and a specified theoretical distribution function. The test statistic is defined as the least upper bound of the set of those distances. One significant limitation of the K-S test is that the distribution must be properly defined; alternatively, if the location, scale, and shape parameters are estimated from the data, the critical region of the K-S test should be assessed by Monte Carlo simulation. [Clauset et al., 2009].

- The Jarque-Bera test is a function of the measures of skewness and kurtosis computed from the sample. It is very popular among econometricians and performs well in comparison with some other tests for normality discussed in the literature if the alternatives to normal distribution belong to the Pearson family [Jarque and Bera, 1980, 1987; Thadewald and Büning, 2007].

- The Lilliefors test is closely related to the K-S test with the difference that the mean and the variance of the distribution can be estimated from the population and does not need to be pre-specified by the user [Lilliefors, 1967, 1969].

- The Anderson-Darling test and the Cramer Von-Mises test are also closely related to the K-S test with a few refinements, and this is one of the reasons that many practitioners opt for those over the original K-S test [Anderson and Darling, 1952].

- When testing against the normal distribution, the Shapiro-Wilk test is the most powerful test, however it is not appropriate for samples with many identical values. It was originally designed for testing against the normal distribution and cannot be used to test against other distributions, unlike the KS test [Shapiro and Wilk, 1965].

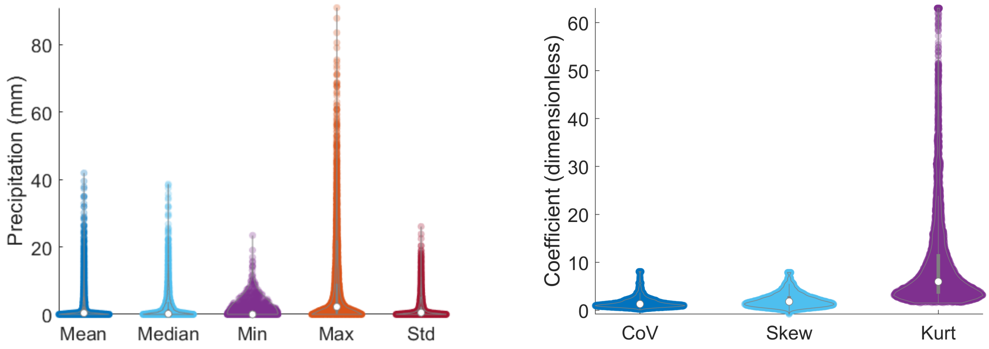
The best hypothesis test depends on the specific case study. For example, for a quick visual identification of the proximity of a sample or multiple samples to the normal distribution, the use of a QQ plot or a BoxPlot respectively are more applicable. For a deeper investigation the Shapiro-Wilk test is the more powerful but can be used only against the normal distribution. For other distributions the Anderson-Darling or the K-S test are more appropriate.

Appendix B

B1 Summary Statistics of Precipitation

Table B1: Mean, median, minimum and maximum values (shown across rows) of daily ERA5 precipitation statistics (shown across the columns) based on 7 472 daily values. Each daily statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	2.02	1.54	0.23	6.84	1.63	1.47	2.08	9.26
Median	0.40	0.15	5.93 0	2.26	0.50	1.24	1.75	5.76
Minimum	0	0	0	0	0	0.10	−0.80	1.53
Maximum	41.94	38.65	23.42	90.99	26.06	8.06	7.87	63.01



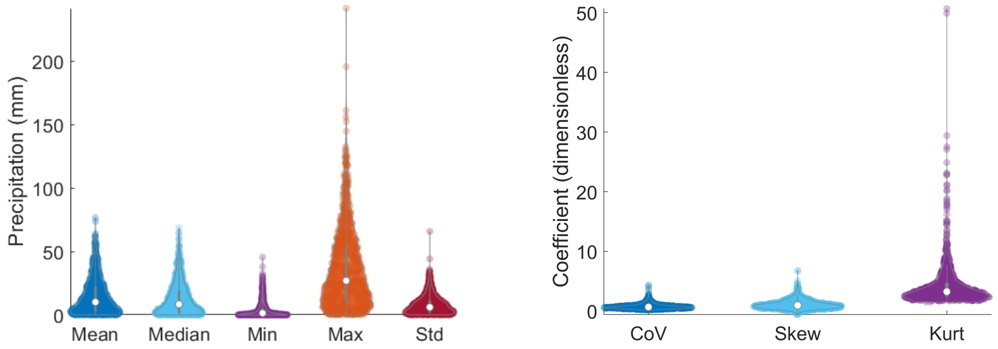
(a) Statistics measured in mm

(b) Dimensionless statistics

Figure B1: Violin plots for the mean, median, minimum and maximum values of daily ERA5 precipitation statistics based on 7 472 daily values. Each daily statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

Table B2: Mean, median, minimum and maximum values (shown across rows) of weekly ERA5 precipitation statistics (shown across the columns) based on 1 068 weekly values. Each weekly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	14.19	12.41	3.88	35.66	7.90	0.78	1.09	4.32
Median	10.13	8.36	1.40	26.93	6.06	0.67	0.95	3.25
Minimum	$8.51\ 10^{-4}$	0	0	0.01	$2.23\ 10^{-3}$	0.17	−0.60	1.51
Maximum	76.71	68.76	45.49	241.82	65.77	4.35	6.77	50.67



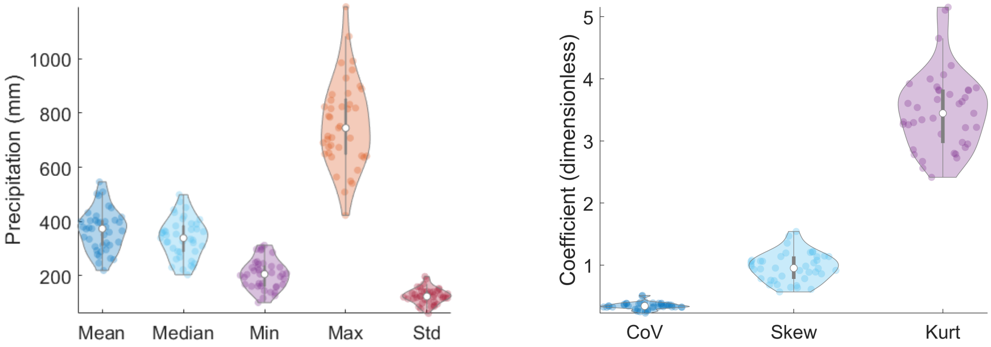
(a) Statistics measured in mm

(b) Dimensionless statistics

Figure B2: Violin plots for the mean, median, minimum and maximum values of weekly ERA5 precipitation statistics based on 1 068 weekly values. Each weekly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

Table B3: Mean, median, minimum and maximum values (shown across rows) of annual ERA5 precipitation statistics (shown across the columns) based on 41 annual values. Each annual statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	367.49	338.83	205.53	758.60	125.42	0.34	0.97	3.48
Median	373.64	339.20	205.89	745.01	123.62	0.34	0.95	3.44
Minimum	219.38	203.23	100.81	422.57	61.24	0.23	0.57	2.41
Maximum	546.20	499.00	312.60	1192.15	197.37	0.51	1.54	5.15



(a) Statistics measured in mm

(b) Dimensionless statistics

Figure B3: Violin plots for the mean, median, minimum and maximum values of annual ERA5 precipitation statistics based on 41 annual values. Each annual statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

Table B4: Number of times $\in \{0, 1, 2 \ldots, 65\}$ each parametric model is selected as the optimal distribution for the daily, weekly, monthly and annual precipitation amounts for the wet period (October to March) for the 65 ERA5 nodes around Crete. Measures of fit: Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), AIC/BIC. Measures of fit for the optimal models are boldfaced.

Timescale Model	Daily	Weekly	Monthly	Annual
beta	0/0	0/0	0/0	0/0
Nakagami	0/0	0/0	**45/40**	**17/17**
Weibull	0/0	0/0	14/12	4/4
gamma	0/0	1/1	4/4	16/16
GEV	0/0	0/0	0/0	0/0
Rayleigh	0/0	0/0	2/9	0/0
Rician	0/0	0/0	0/0	8/8
GP	**65/65**	**58/50**	0/0	0/0
logistic	0/0	0/0	0/0	4/4
t-scale	0/0	0/0	0/0	0/0
normal	0/0	0/0	0/0	0/0
log-logistic	0/0	0/0	0/0	2/2
lognornal	0/0	0/0	0/0	1/1
exponential	0/0	6/14	0/0	0/0
Birnbaum-Saunders	0/0	0/0	0/0	5/5
EV	0/0	0/0	0/0	0/0
inverse normal	0/0	0/0	0/0	8/8

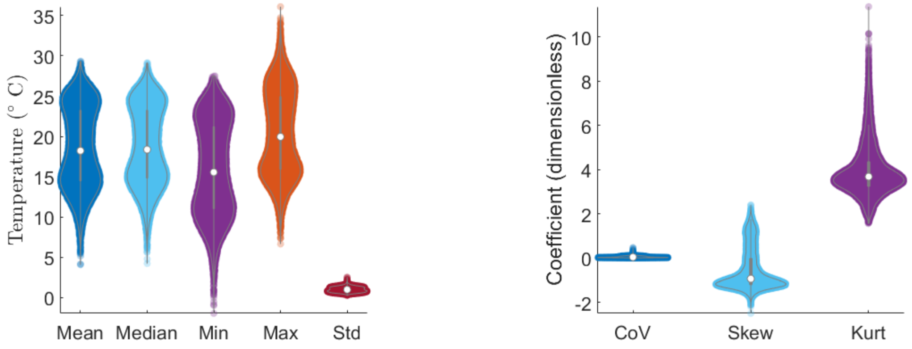
Table B5: Number of times each parametric model is selected as the optimal distribution for the daily ($\in \{0, 1, 2 \ldots, 7\,472\}$ time steps), weekly ($\in \{0, 1, 2 \ldots, 1\,068\}$ time steps), monthly ($\in \{0, 1, 2 \ldots, 246\}$ time steps) and annual ($\in \{0, 1, 2 \ldots, 41\}$ time steps) precipitation amounts for the wet period (October to March) for the 65 ERA5 nodes around Crete depending on the timestamp of each timescale. Measures of fit: Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), AIC/BIC. Measures of fit for the optimal models are boldfaced.

Timescale Model	Daily	Weekly	Monthly	Annual
beta	18/12	1/0	0/0	0/0
Nakagami	645/637	63/59	8/8	0/0
Weibull	187/191	25/25	3/5	0/0
gamma	183/164	44/44	13/12	0/0
GEV	440/422	31/23	25/15	**20/11**
Rayleigh	45/135	20/64	1/5	0/0
Rician	7/6	4/5	0/0	0/0
GP	**4602/3924**	**673/517**	**126/106**	0/0
logistic	3/3	1/1	0/0	0/0
t-scale	202/202	1/0	0/0	0/0
normal	1/0	0/0	0/0	0/0
log-logistic	63/70	22/22	2/2	0/0
lognornal	69/79	12/14	3/3	0/0
exponential	593/1108	29/96	0/1	0/0
Birnbaum-Saunders	365/434	93/124	25/30	2/2
EV	0/2	1/2	0/0	0/0
inverse normal	59/83	48/72	40/59	**19/28**

B2 Summary Statistics of Temperature

Table B6: Mean, median, minimum and maximum values (shown across rows) of daily ERA5 temperature statistics (shown across the columns) based on 14 975 daily values. Each daily statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	18.62	18.79	15.93	20.40	1.01	0.06	−0.53	3.92
Median	18.22	18.37	15.56	19.96	0.99	0.05	−0.92	3.68
Minimum	4.09	4.24	−1.95	6.66	0.27	0.01	−2.49	1.57
Maximum	29.32	29.14	27.51	36.10	2.56	0.48	2.41	11.36



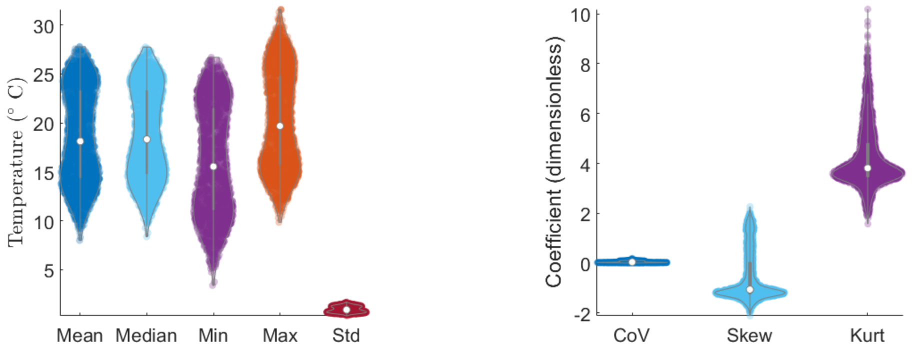
(a) Statistics measured in ° C

(b) Dimensionless statistics

Figure B4: Violin plots for the mean, median, minimum and maximum values of daily ERA5 temperature statistics based on 14 975 daily values. Each daily statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

Table B7: Mean, median, minimum and maximum values (shown across rows) of weekly ERA5 temperature statistics (shown across the columns) based on 2 140 weekly values. Each weekly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	18.62	18.81	16.11	20.22	0.93	0.06	−0.56	4.23
Median	18.11	18.32	15.53	19.68	0.89	0.05	−1.05	3.81
Minimum	7.97	8.35	3.37	9.83	0.26	0.01	−2.11	1.57
Maximum	27.81	27.76	26.70	31.62	1.68	0.20	2.27	10.19



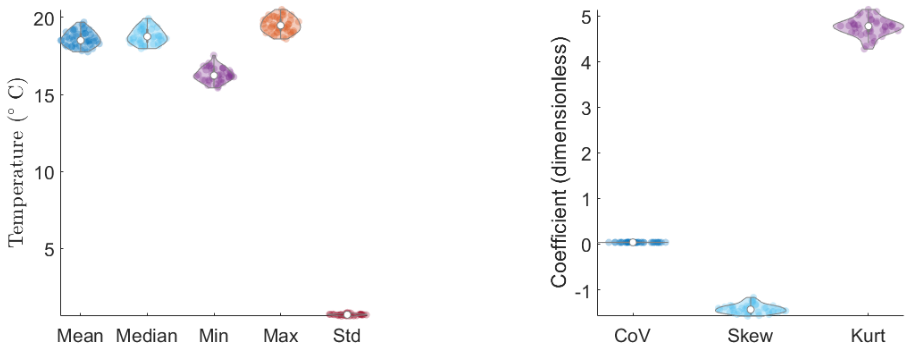
(a) Statistics measured in ° C

(b) Dimensionless statistics

Figure B5: Violin plots for the mean, median, minimum and maximum values of weekly ERA5 temperature statistics based on 2 140 weekly values. Each weekly statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

Table B8: Mean, median, minimum and maximum values (shown across rows) of annual ERA5 temperature statistics (shown across the columns) based on 41 annual values. Each annual statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

	Mean	Median	Min	Max	Std	CoV	Skew	Kurt
Mean	18.62	18.81	16.28	19.52	0.71	0.04	−1.42	4.75
Median	18.1	18.76	16.23	19.48	0.71	0.04	−1.43	4.77
Minimum	17.77	19.98	15.42	18.60	0.61	0.03	−1.58	4.27
Maximum	19.69	19.93	17.54	20.52	0.86	0.05	−1.17	5.14



(a) Statistics measured in ° C

(b) Dimensionless statistics

Figure B6: Violin plots for the mean, median, minimum and maximum values of annual ERA5 temperature statistics based on 41 annual values. Each annual statistic is based on the data at the 65 ERA5 grid nodes. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in °C.

Table B9: Number of times $\in \{0, 1, 2 \ldots, 65\}$ each parametric model is selected as the optimal distribution for the daily, weekly, monthly and annual temperature amounts for the full period (October to March) for the 65 ERA5 nodes around Crete. Measures of fit: Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), AIC/BIC. Measures of fit for the optimal models are boldfaced.

Timescale Model	Daily	Weekly	Monthly	Annual
beta	0/0	0/0	0/0	0/0
Nakagami	0/0	0/0	0/0	0/0
Weibull	0/0	0/0	0/0	0/0
gamma	0/0	0/0	0/0	0/0
GEV	**65/65**	**62/62**	0/0	0/0
Rayleigh	0/0	0/0	0/0	0/0
Rician	0/0	0/0	0/0	0/0
GP	0/0	3/3	**65/65**	**49/44**
logistic	0/0	0/0	0/0	0/0
t-scale	0/0	0/0	0/0	0/0
normal	0/0	0/0	0/0	0/0
log-logistic	0/0	0/0	0/0	0/0
lognornal	0/0	0/0	0/0	0/0
exponential	0/0	0/0	0/0	0/0
Birnbaum-Saunders	0/0	0/0	0/0	0/0
EV	0/0	0/0	0/0	0/0
inverse Gaussian	0/0	0/0	0/0	16/21

Table B10: Number of times each parametric model is selected as the optimal distribution for the daily ($\in \{0, 1, 2 \ldots, 14\,975\}$ time steps), weekly ($\in \{0, 1, 2 \ldots, 2\,140\}$ time steps), monthly ($\in \{0, 1, 2 \ldots, 492\}$ time steps) and annual ($\in \{0, 1, 2 \ldots, 41\}$ time steps) temperature amounts for the full period (January to December) for the 65 ERA5 nodes around Crete depending on the timestamp of each timescale. Measures of fit: Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), AIC/BIC. Measures of fit for the optimal models are boldfaced.

Timescale Model	Daily	Weekly	Monthly	Annual
beta	0/0	0/0	0/0	0/0
Nakagami	35/36	2/3	1/1	0/0
Weibull	800/1086	68/106	13/17	0/0
gamma	20/21	4/4	1/1	0/0
GEV	3950/2717	543/396	114/88	**34/26**
Rayleigh	0/0	0/0	0/0	0/0
Rician	220/334	31/40	6/7	0/0
GP	**6219/5958**	**957/917**	**229/226**	2/2
logistic	505/635	60/86	11/15	0/0
t-scale	912/607	220/165	74/63	0/0
normal	2/2	0/0	0/0	0/0
log-logistic	486/809	62/107	11/22	0/0
lognornal	0/0	0/0	0/0	0/0
exponential	0/0	0/0	0/0	0/0
Birnbaum-Saunders	21/22	4/4	0/0	0/0
EV	1519/2282	149/251	22/38	5/13
inverse Gaussian	286/466	40/61	10/14	0/0

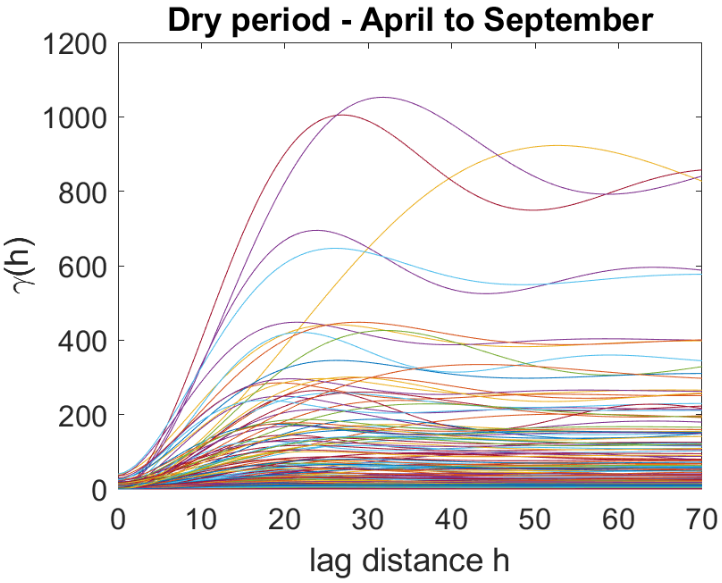
Appendix C

Table C1: Mean, median, minimum and maximum values (shown across different rows) of monthly ERA5 precipitation statistics (shown across the columns) based on 246 monthly values for the dry months. Each monthly statistic is obtained from the 65 values in the respective spatial layer. The values for CoV (coefficient of variation), Skew (skewness) and Kurt (kurtosis) are dimensionless. All other values are measured in mm.

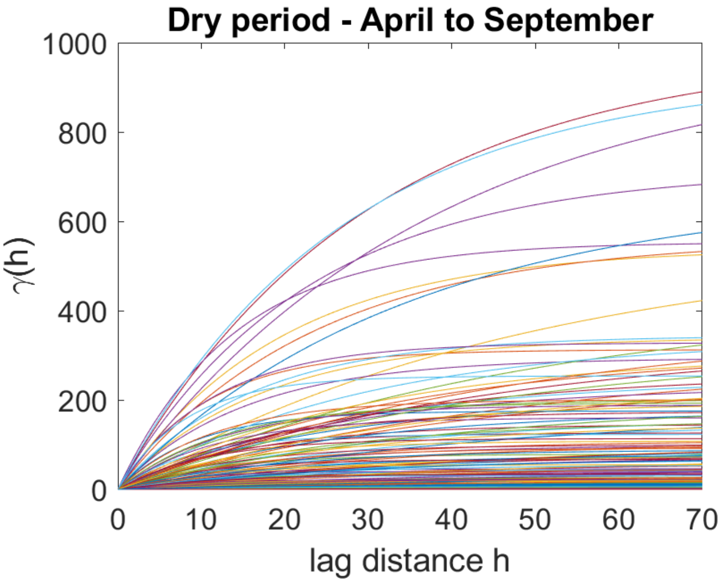
	Mean	**Median**	**Min**	**Max**	**Std**	**CoV**	**Skew**	**Kurt**
Mean	8.59	6.92	1.89	26.61	5.84	1.10	1.62	6.18
Median	3.51	2.16	0.08	15.48	3.35	1.05	1.55	5.19
Minimum	0.02	0.00	0.00	0.14	0.03	0.22	−0.21	1.57
Maximum	78.23	73.49	35.65	179.92	31.33	2.74	5.15	33.80

Table C2: Optimal probability distribution fits for the monthly ERA5 precipitation data. The models studied include the following: "GP": Generalized Pareto, "t-scale": t-Scale location, and "GEV" refers to the Generalized Extreme Value distribution.

April	**May**	**June**	**July**	**August**	**September**
t-scale	GP	GP	GEV	GP	GP

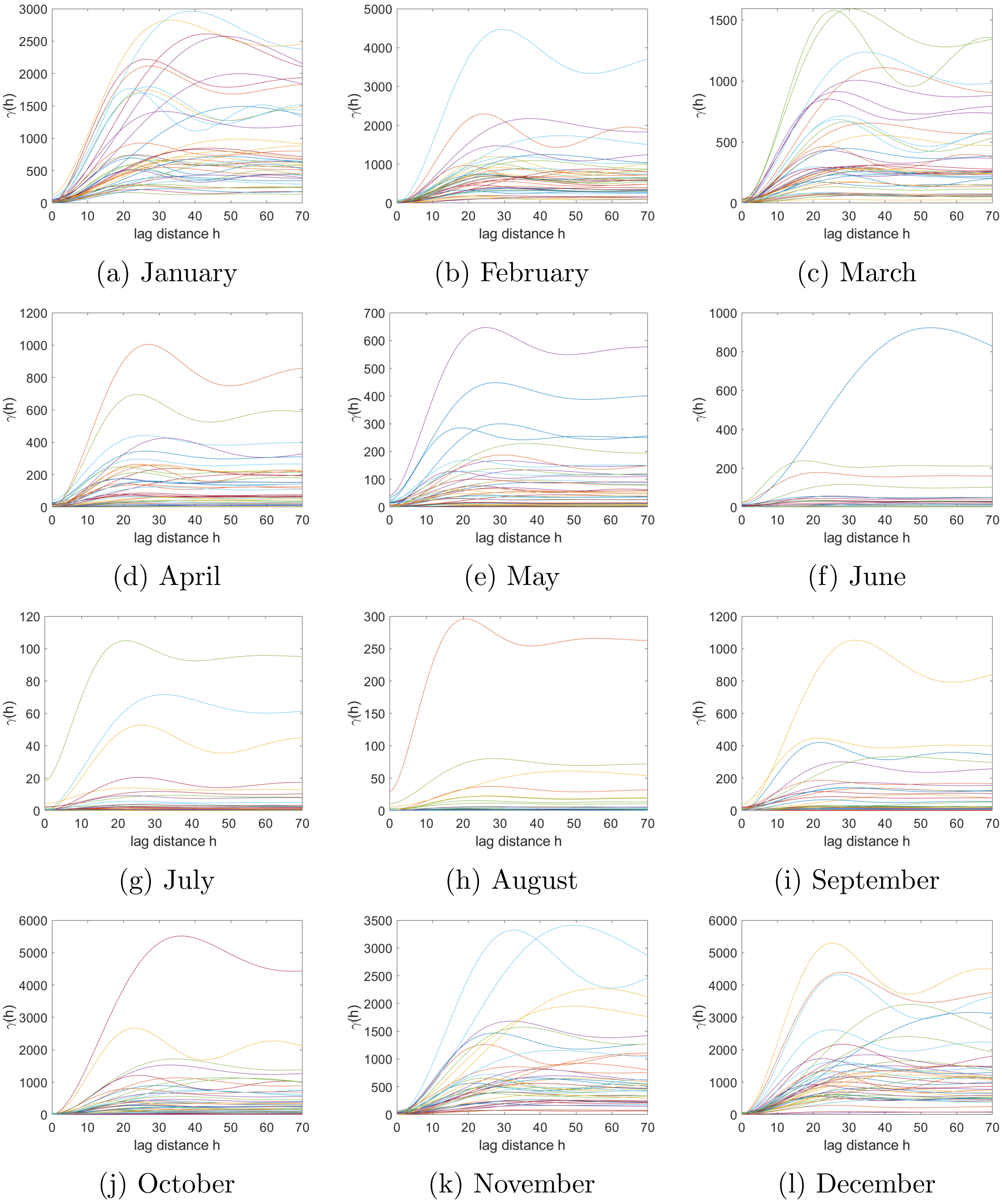


(a) Varios Dry GA false SSRF



(b) Varios Dry GA false Exponential

Figure C1: Spartan (top) and exponential (bottom) variogram fits of the monthly ERA5 precipitation obtained from different spatial layers for the dry period (from April till September). Each of the variograms is calculated based on 65 spatial locations for 246 months (corresponding to 246 spatial layers for the dry period). The precipitation values are used without the application of a normalizing transform.



(a) January

(b) February

(c) March

(d) April

(e) May

(f) June

(g) July

(h) August

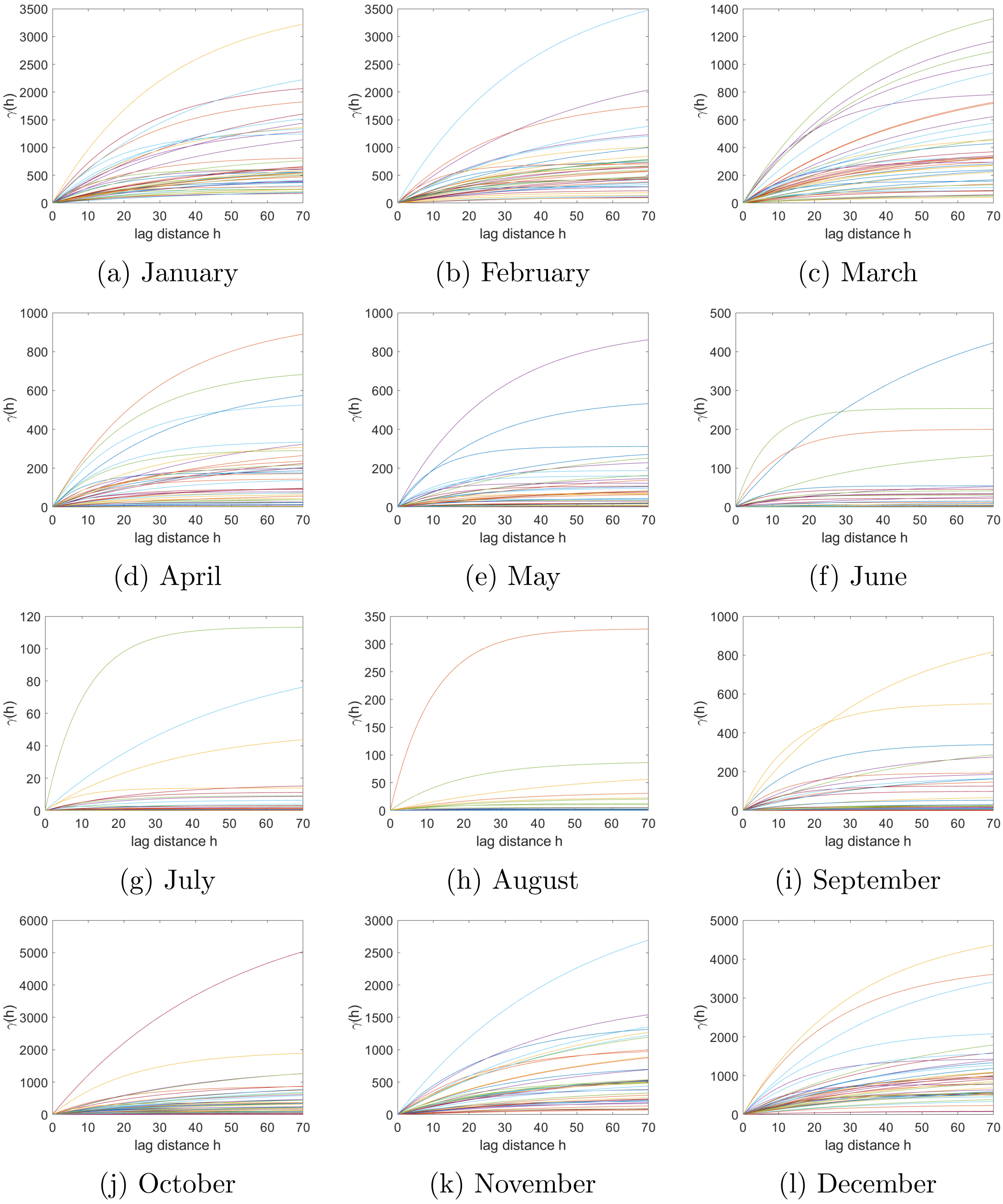
(i) September

(j) October

(k) November

(l) December

Figure C2: Spartan variogram fit for the monthly ERA5 precipitation data set, without any transformation imposed, presented separately for every month in a variogram cloud plot. Detailed view of Fig. 5.6a and C1a.



(a) January

(b) February

(c) March

(d) April

(e) May

(f) June

(g) July

(h) August

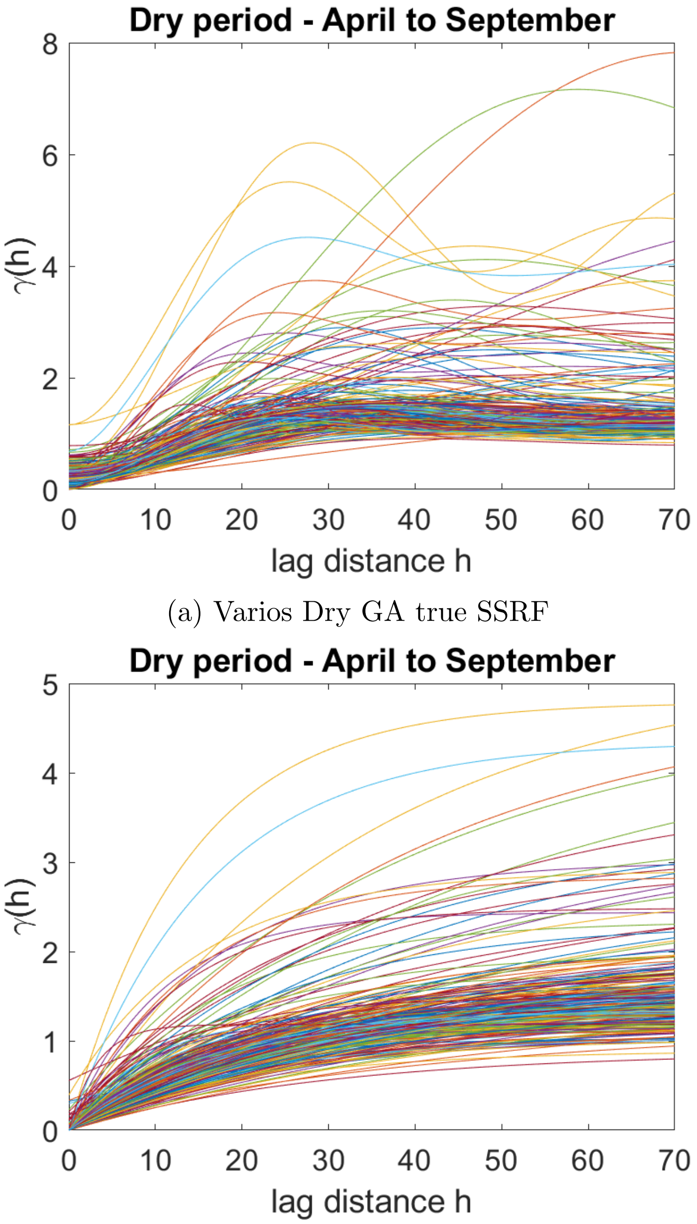
(i) September

(j) October

(k) November

(l) December

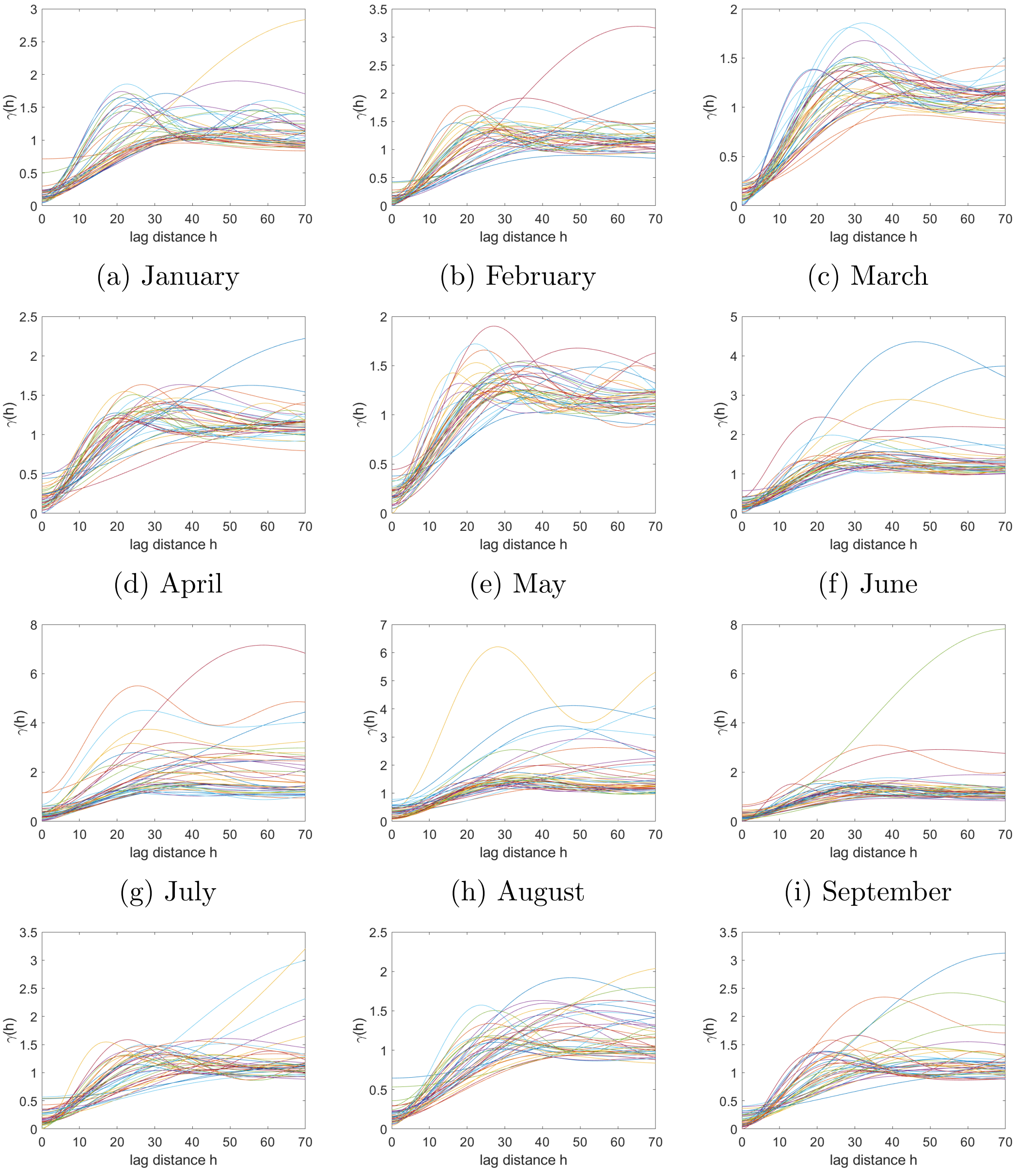
Figure C3: Exponential variogram fit for the monthly ERA5 precipitation data set, without any transformation imposed, presented separately for every month in a variogram cloud plot. Detailed view of Figs. 5.6b and C1b.



(a) Varios Dry GA true SSRF

(b) Vario Dry GA true Exponential

Figure C4: Spartan (top) and exponential (bottom) variogram cloud fits of the monthly ERA5 precipitation obtained from different spatial layers for the dry period (from April till September). Each of the variograms is calculated based on 65 spatial locations for 246 months (corresponding to 246 spatial layers for the dry period). The normalized precipitation values are generated with the GAH normalizing transform up to degree 20.



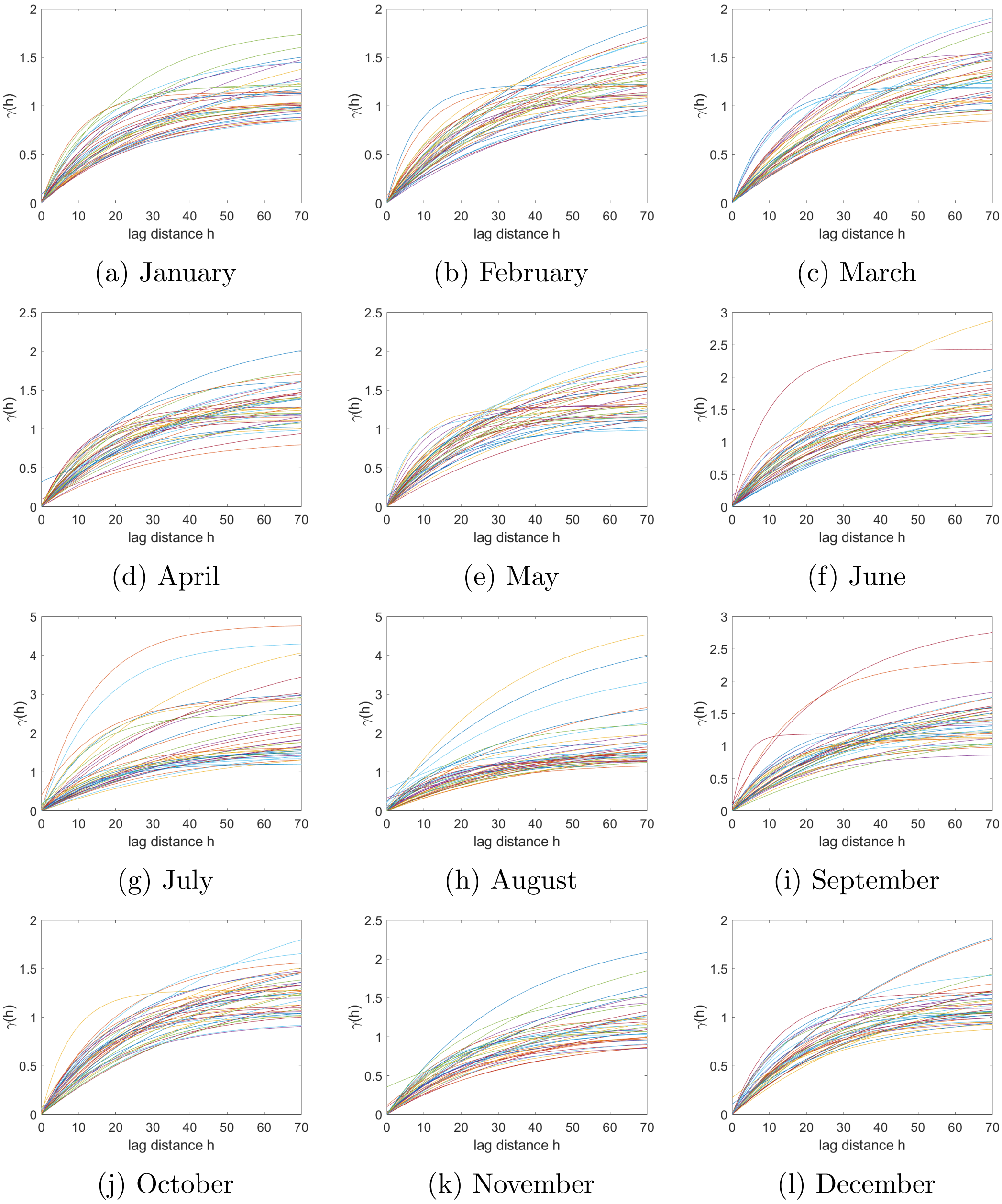
(a) January (b) February (c) March

(d) April (e) May (f) June

(g) July (h) August (i) September

(j) October (k) November (l) December

Figure C5: Spartan variogram fit for the monthly ERA5 precipitation data set, with GAH transformation with 20 hermite polynomials imposed, presented separately for every month in a variogram cloud plot. Detailed view of Fig. 5.7a and C4a.



(a) January (b) February (c) March

(d) April (e) May (f) June

(g) July (h) August (i) September

(j) October (k) November (l) December

Figure C6: Exponential variogram fit for the monthly ERA5 precipitation data set, with GAH transformation with 20 hermite polynomials imposed, presented separately for every month in a variogram cloud plot. Detailed view of Fig. 5.7b and C4b.

Table C3: This table presents the mean, median, minimum, maximum, and standard deviation values (shown across different rows) of the LOO-CV measures of the OK with the exponential model on the monthly ERA5 precipitation (shown across the columns) based on 246 wet monthly values coupled with 100 bootstrap simulations. The precipitation values used are the monthly ERA5 precipitation values and their summary measures are presented in Figs. 5.8-5.11 and correspond to S1 to S4 with the WLS optimization method. All values, except Pearson's (RP) and Spearman's (RS) correlation coefficients are in mm. Correlation coefficients are dimensionless. Analytically the scenarios can be found in Table 5.1.

	ME	**MAE**	**RMSE**	**RP**	**RS**	**ErrMin**	**ErrMax**
				S1			
Mean	1.04	6.82	9.62	0.91	0.92	−25.56	33.34
Median	0.89	6.41	8.90	0.93	0.93	−23.22	28.94
Minimum	−0.37	0.41	0.58	0.71	0.65	−80.22	2.24
Maximum	4.60	19.15	27.20	0.98	0.98	−1.55	103.60
Std	0.77	3.33	4.57	0.05	0.05	13.30	17.36
				S2			
Mean	2.88	10.51	14.42	0.76	0.77	−24.17	45.17
Median	1.80	8.79	12.13	0.88	0.90	−25.06	36.48
Minimum	−1.94	0.48	0.66	−0.63	−0.68	−95.30	2.07
Maximum	21.23	46.60	62.93	0.98	0.99	−1.89	231.59
Std	3.21	7.40	9.63	0.34	0.35	16.40	30.96
				S3			
Mean	1.15	6.89	9.70	0.91	0.91	−25.03	32.69
Median	0.93	6.45	8.93	0.93	0.93	−22.67	29.20
Minimum	−0.50	0.41	0.59	0.66	0.58	−76.83	2.43
Maximum	5.41	18.81	27.05	0.98	0.98	−1.35	106.50
Std	0.90	3.37	4.62	0.05	0.06	12.86	17.72
				S4			
Mean	−0.09	6.84	9.32	0.92	0.92	−21.78	30.52
Median	−0.07	6.25	8.65	0.93	0.93	−20.07	26.86
Minimum	−0.50	0.41	0.55	0.68	0.61	−62.70	2.12
Maximum	0.24	18.60	25.56	0.99	0.99	−1.10	95.72
Std	0.14	3.34	4.438	0.05	0.05	10.35	16.32

Journal publications linked to PhD research

V. D. Agou, A. Pavlides, and D. T. Hristopulos. Spatial modeling of precipitation based on data-driven warping of Gaussian Processes. *Entropy*, 24(3), 2022. ISSN 1099-4300. doi: 10.3390/e24030321.

D. T. Hristopulos and V. D. Agou. Stochastic local interaction model with sparse precision matrix for space-time interpolation. *Spatial Statistics*, 40:100403, 2020. doi: 10.1016/j.spasta.2019.100403.

D. T. Hristopulos, A. Pavlides, V. D. Agou, and P. Gkafa. Stochastic local interaction model: An alternative to kriging for massive datasets. *Mathematical Geosciences*, 2021. doi: 10.1007/s11004-021-09957-7.

E. Koutroulis, G. Petrakis, V. Agou, A. Malisovas, D. Hristopulos, P. Partsinevelos, A. Tripolitsiotis, N. Halouani, P. Ailliot, M. Boutigny, V. Monbet, D. Allard, A. Cuzol, D. Kolokotsa, E. Varouchakis, K. Kokolakis, and S. Mertikas. Site selection and system sizing of desalination plants powered with renewable energy sources based on a web-gis platform. *International Journal of Energy Sector Management*, 2021. doi: 10.1108/IJESM-04-2021-0018.

A. Pavlides, V. D. Agou, and D. T. Hristopulos. Non-parametric kernel-based estimation and simulation of precipitation amount. *Journal of Hydrology*, 612: 127988, 2022. doi: 10.1016/j.jhydrol.2022.127988.

Conference presentations linked to PhD research

V. D. Agou, A. Pavlides, and D. T. Hristopulos. Space-time analysis of precipitation reanalysis data for the island of Crete using Gaussian anamorphosis with Hermite polynomials. In *EGU General Assembly 2021*, Vienna, Austria, 2021. doi: 10.5194/egusphere-egu21-3088. abstract no. EGU2021-3088.

D. T. Hristopulos and V. D. Agou. Stochastic local interaction model for spatial and space-time data. In *9th Workshop on Spatio-temporal modeling - METMA IX*, Montpellier, France, 2018.

D. T. Hristopulos, A. Pavlides, V. D. Agou, and P. Gkafa. Introduction to a stochastic local interaction model and applications. In *19th Annual Conference of the International Association for Mathematical Geosciences (IAMG2018)*, Olomouc, Czech Republic, 2018.

D. T. Hristopulos, V. D. Agou, A. Pavlides, and P. Gkafa. Stochastic local interaction models for processing spatiotemporal datasets. In *EGU General Assembly 2020*, Vienna, Austria, 2020. doi: 10.5194/egusphere-egu2020-2954. abstract no. EGU2020-2954.

References

M. Abramowitz, I. A. Stegun, and R. H. Romer. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* American Association of Physics Teachers, 1988.

J. Adams. climate_indices, an open source python library providing reference implementations of commonly used climate indices, 2019. Retrieved on October 7, 2021 from: `https://github.com/monocongo/climate_indices`.

C. Agnese, G. Baiamonte, and C. Cammalleri. Modelling the occurrence of rainy days under a typical mediterranean climate. *Advances in Water Resources*, 64: 62–76, 2014. ISSN 0309-1708. doi: 10.1016/j.advwatres.2013.12.005.

V. Agou. Geostatistical analysis of precipitation on the island of Crete. Master's thesis, Technical University of Crete, Chania, Crete, 73100 Greece, 2016.

V. D. Agou, E. A. Varouchakis, and D. T. Hristopulos. Geostatistical analysis of precipitation in the island of Crete (Greece) based on a sparse monitoring network. *Environmental Monitoring and Assessment*, 191(353):1573–2959, 2019. doi: 10.1007/s10661-019-7462-8.

V. D. Agou, A. Pavlides, and D. T. Hristopulos. Spatial modeling of precipitation based on data-driven warping of Gaussian Processes. *Entropy*, 24(3), 2022. ISSN 1099-4300. doi: 10.3390/e24030321.

J. Aldrich. R.A. Fisher and the making of maximum likelihood 1912-1922. *Statistical Science*, 12(3):162–176, 1997. doi: 10.1214/ss/1030037906.

S. Alecu. The Gaussian Transform of Distributions: Definition, Computation and Application. *IEEE Transactions on Signal Processing*, 54(8):2976–2985, 2006. doi: 10.1109/TSP.2006.877657.

P. Alfeld. A trivariate clough—tocher scheme for tetrahedral data. *Computer Aided Geometric Design*, 1(2):169–181, 1984. ISSN 0167-8396. doi: https://doi.org/10.1016/0167-8396(84)90029-3.

R. G. Allen, L. S. Pereira, D. Raes, and M. Smith. *Crop Evapotranspiration: Guidelines for Computing Crop Requirements, Irrigation and Drainage*. FAO: Roma, Italia, 1998.

American Meteorological society. Glossary of Meteorology, October 2022. Online: `https://glossary.ametsoc.org/wiki/Rain`.

T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193–212, 1952. doi: 10.1214/aoms/1177729437.

C. M. Andreou. *Construction and Fitting of Random Field Models to Precipitation Data*. PhD thesis, University of Cyprus, Faculty of Pure and Applied Sciences, Nicosia, Cyprus, 2022. URL `http://gnosis.library.ucy.ac.cy/handle/7/65162`.

M. Armstrong and G. Matheron. Disjunctive kriging revisited: Part I. *Mathematical Geology*, 18(8):711–728, 1986.

P. M. Atkinson. Downscaling in remote sensing. *International Journal of Applied Earth Observation and Geoinformation*, 22:106–114, 2013. ISSN 0303-2434. doi: 10.1016/j.jag.2012.04.012. Spatial Statistics for Mapping the Environment.

D. Atlas. *Radar in Meteorology*. Springer: New York, NY, USA, 1990.

R. Avanzato and F. Beritelli. An innovative acoustic rain gauge based on convolutional neural networks. *Information*, 11(4), 2020. ISSN 2078-2489. doi: 10.3390/info11040183.

E. Aytekin. Wildfires ravaging forestlands in many parts of globe, 2021. URL `https://www.aa.com.tr/en/world/wildfires-ravaging-forestlands-in-many-parts-of-globe/2322512`. Retrieved 03.08.2021.

J. Baik, J. Park, D. Ryu, and M. Choi. Geospatial blending to improve spatial mapping of precipitation with high spatial resolution by merging satellite-based and ground-based data. *Hydrological Processes*, 30(16):2789–2803, 2016. doi: 10.1002/hyp.10786.

K. Balanda and H. Macgillivray. Kurtosis: A critical review. *The American Statistician*, 42(2):111–119, 1988. doi: 10.1080/00031305.1988.10475539.

C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4): 469–483, 1996.

A. Bárdossy and G. Pegram. Infilling missing precipitation records–a comparison of a new copula-based method with other techniques. *Journal of Hydrology*, 519, Part A:1162–1170, 2014.

A. Bárdossy and G. Pergam. Interpolation of precipitation under topographic influence at different time scales. *Water Resources Research*, 49:4545–4565, 2013. doi: 10.1002/wrcr.20307.

D. Barriopedro, E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera. The hot summer of 2010: redrawing the temperature record map of Europe. *Science*, 332(6026):110–224, 2011.

S. Barthiban, B. Lloyd, and M. Maier. Sanitary hazards and microbial quality of open dug wells in the Maldives Islands. *Journal of Water Resource and Protection*, 4(7):474–486, 2012. doi: 10.4236/jwarp.2012.47055.

A. Baxevani and J. Lennatsson. A spatiotemporal precipitation generator based on a censored latent Gaussian field. *Water Resources Research*, 51(6), 2015. doi: 10.1002/2014WR016455.

S. Beguería, S. M. Vicente-Serrano, F. Reig, and B. Latorre. Standardized precipitation evapotranspiration index (spei) revisited: parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. *International Journal of Climatology*, 34(10):3001–3023, 2014. doi: 10.1002/joc.3887.

M. Belgiu and L. Drăguţ. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114:24–31, 2016. ISSN 0924-2716. doi: 10.1016/j.isprsjprs.2016.01.011.

S. Bergström, B. Carlsson, M. Gardelin, G. Lindström, A. Pettersson, and M. Rummukainen. Climate change impacts on runoff in Sweden assessments by global climate models, dynamical downscaling and hydrological modelling. *Climate research*, 16(2):101–112, 2001.

C. Berndt, E. Rabiei, and U. Haberlandt. Geostatistical merging of rain gauge and radar data for high temporal resolutions and various station density scenarios. *Journal of Hydrology*, 508:88–101, 2014. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2013.10.028.

M. Berterretche, A. Hudak, W. Cohen, T. Maiersperger, S. Gower, and J. Dungan. Comparison of regression and geostatistical methods for mapping leaf area index (LAI) with landsat ETM+ data over a boreal forest. *Remote Sensing of Environment*, 96(3):49–61, 2005.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

G. Biau and E. Scornet. A random forest guided tour. *TEST*, 25(2):197–227, 2016. doi: 10.1007/s11749-016-0481-7.

R. Biondini. Cloud motion and rainfall statistics. *Journal of Applied Meteorology and Climatology*, 15(3):205–224, 1976.

E. Black. The impact of climate change on daily precipitation statistics in Jordan and Israel. *Atmospheric Science Letters*, 10(3):192–200, 2009.

S. Bochner, M. Tenenbaum, and H. Pollard. *Lectures on Fourier Integrals*. Annals of mathematics studies. Princeton University Press, 1959. ISBN 9780691079943.

E. P. J. Boer, K. M. de Beurs, and A. D. Hartkamp. Kriging and thin plate splines for mapping climate variables. *International Journal of Applied Earth Observation and Geoinformation*, 3(2):146–154, 2001.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining and Knowledge Discovery*, 2 (6):493–507, 2012. doi: 10.1002/widm.1072.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society*, 26(2), 1964. doi: 10.1002/2014WR016455.

L. Breiman. Stacked regressions. *Machine learning*, 24:49–64, 1996. doi: 10.1007/BF00117832.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. doi: 1573-0565.

L. Breiman. Manual-setting up, using, and understanding random forests v4.0. Technical report, University of California, Berkeley. Department of Statistics, Berkeley, CA, 2003. 33 pp.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series, Belmont, C.A., 1984. Wadsworth Advanced Books and Software.

G. Buttafuoco, T. Caloiero, and R. Coscarelli. Spatial and temporal patterns of the mean annual precipitation at decadal time scale in southern Italy (Calabria region). *Theoretical and Applied Climatology*, 105:431–444, 2011.

M. Cannarozzo, L. V. Noto, and F. Viola. Spatial distribution of rainfall trends in Sicily (1921–2000). *Physics and Chemistry of the Earth, Parts A/B/C*, 31 (18):1201–1211, 2006.

A. Chappell, L. J. Renzullo, T. H. Raupach, and M. Haylock. Evaluating geostatistical methods of blending satellite and gauge data to estimate near real-time daily rainfall for Australia. *Journal of Hydrology*, 493:105–114, 2013. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2013.04.024.

K. Chartzoulakis, N. Paranychianakis, and A. Angelakis. Water resources management in the island of crete, greece, with emphasis on the agricultural use. *Water Policy*, 3:193–205, 2001.

R. J. Chase, D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern. A machine learning tutorial for operational meteorology. part I: Traditional machine learning. *Weather and Forecasting*, 37(8):1509 – 1529, 2022. doi: 10.1175/WAF-D-22-0070.1.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, 2002. ISSN 1076-9757.

C. Chen, A. Liaw, and L. Breiman. Using random forest to learn imbalanced data. *University of California, Berkeley*, 110(1–12):24, 2004.

S. Chen, T. Y. Gan, X. Tan, D. Shao, and J. Zhu. Assessment of CFSR, ERA-Interim, JRA-55, MERRA-2, NCEP-2 reanalysis data for drought analysis over China. *Climate Dynamics*, 53(1–2):737–757, 2019. doi: 10.1007/s00382-018-04611-1.

X. Chen, M. Wang, and H. Zhang. The use of classification trees for bioinformatics. *WIREs Data Mining and Knowledge Discovery*, 1(1):55–63, 2011. doi: 10.1002/widm.14.

J. Chilès and P. Delfiner. *Geostatistics: modeling spatial uncertainty*. Wiley series in probability and statistics. Wiley, 2012. ISBN 9780470183151.

H.-K. Cho, K. P. Bowman, and G. R. North. A comparison of gamma and lognormal distributions for characterizing satellite rain rates from the tropical rainfall measuring mission. *Journal of Applied Meteorology*, 43(11):1586–1597, 2004. doi: 10.1175/JAM2165.1.

Chowdhury. *Dr. Chowdhury's Guide to Planet Earth: "The Water Cycle"*. West ed. retrieved 2006-10-24 edition, 2005.

G. Christakos. *Random Field Models in Earth Sciences*. Academic Press, Chapel Hill, North Carolina, 1992. ISBN 0-12-174230-X.

G. Christakos. *Spatiotemporal Random Fields: Theory and Applications*. Elsevier, Amsterdam, Netherlands, 2017.

G. Christakos, D. T. Hristopulos, and P. Bogaert. On the physical geometry concept at the basis of space/time geostatistical hydrology. *Advances in Water Resources*, 23(8):799–810, 2000.

J. H. Christensen, F. Boberg, O. B. Christensen, and P. Lucas-Picher. On the need for bias correction of regional climate change projections of temperature and precipitation. *Geophysical Research Letters*, 35(20), 2008. doi: 10.1029/2008GL035694.

J. Chu, J. Xia, C.-Y. Xu, and V. Singh. Statistical downscaling of daily mean temperature, pan evaporation and precipitation for climate change scenarios in Haihe River, China. *Theoretical and Applied Climatology*, 99:149–161, 2010.

A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009. ISSN 00361445, 10957200. URL http://www.jstor.org/stable/25662336.

S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001. ISBN 1-85233-459-2.

Copernicus Climate Change Service C3S. ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate, 2018. Data retrieved from: https://cds.climate.copernicus.eu/cdsapp#!/home.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20:273–297, 1995.

R. Coscarelli and T. Caloiero. Analysis of daily and monthly rainfall concentration in Southern Italy (Calabria region). *Journal of Hydrology*, 416-417:145–156, 2012.

N. R. Council et al. *Frontiers in Massive Data Analysis*. National Academies Press, Washington, DC, 2013.

N. Cressie and C. L. Wikle. *Statistics for Spatio-temporal Data*. John Wiley and Sons, New York, 2011.

N. A. C. Cressie. *Statistics for spatial data*. Wiley series in probability and statistics. Wiley, revised edition edition, 1993. ISBN 9780471002550.

A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends ®in Computer Graphics and Vision*, 7(2–3):81–227, 2012. ISSN 1572-2740. doi: 10.1561/0600000035.

A. Dai. Increasing drought under global warming in observations and models. *Nature Climate Change*, 3(1):52–58, 2013. doi: 10.1038/nclimate1633.

A. Dai and T. Zhao. Uncertainties in historical changes and future projections of drought. Part I: estimates of historical drought changes. *Climatic Change*, 144:519–533, 2017.

A. Dai, K. E. Trenberth, and T. Qian. A global dataset of palmer drought severity index for 1870–2002: relationship with soil moisture and effects of surface warming. *Journal of Hydrometeorology*, 5:1117–1130, 2004.

L. de Haan and A. Ferreira. *Extreme Value Theory: An Introduction (Springer Series in Operations Research and Financial Engineering)*. Springer, New York, NY 10013, U.S.A., 1st edition. edition, 2010. ISBN 144192020X. doi: 10.1007/0-387-34471-3.

S. De Iaco, D. E. Myers, and D. Posa. Nonseparable space-time covariance models: some parametric families. *Mathematical Geology*, 34(1):23–42, 2002.

D. Dee, J. Fasullo, D. Shea, and J. Walsh. The climate data guide: Atmospheric reanalysis: Overview & comparison tables. last modified 12 dec 2016, 2016. Retrieved on October 7, 2021 from: `https://climatedataguide.ucar.edu/climate-data/atmospheric-reanalysis-overview-comparison-tables`.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:1471–2105, 2006.

N. J. Doesken and D. Garen. Drought monitoring in the western united states using a surface water supply index. *Seventh Conference on Applied Climatology*, pages 266–269, 2004. American Meteorological Society.

N. J. Doesken, T. B. McKee, and J. Kleist. Development of a surface water supply index for the western United States: final report, 1991. URL `https://climate.colostate.edu/pdfs/climo_rpt_91-3.pdf`. Climatology Report 91-3, Colorado Climate Center.

P. Domingos. Metacost: A general method for making classifiers cost-sensitive. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 155—164, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581131437. doi: 10.1145/312129.312220.

S.-B. Duan, Z.-L. Li, B.-H. Tang, H. Wu, and R. Tang. Evaluation of real-time satellite precipitation data for global drought monitoring. *Remote Sensing of Environment*, 140:339–349, 2014.

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38. Society for Industrial and Applied Mathematics, Philadelphia, USA, 1982. doi: 10.1137/1.9781611970319.

K. Elder, J. Dozier, and J. Michaelsen. 1991 accumulation and distribution in an alpine watershed. *Water Resources Research*, 27:1541–1552, 2015.

R. Erdin, C. Frei, and H. R. Künsch. Data transformation and uncertainty in geostatistical combination of radar and rain gauges. *Journal of Hydrometeorology*, 13(4):1332–1346, 2012. doi: 10.1175/JHM-D-11-096.1.

H. Ezzine, A. Bouziane, and D. Ouazar. Seasonal comparisons of meteorological and agricultural drought indices in morocco using open short time-series data. *International Journal of Applied Earth Observations & Geoinformation*, 26(1): 36–48, 2014.

G. Farin. Triangular Bernstein-Bézier patches. *Computer Aided Geometric Design*, 3(2):83–127, 1986. ISSN 0167-8396. doi: https://doi.org/10.1016/0167-8396(86)90016-6.

T. Fawcett and F. Provost. Adaptive fraud detection. *Data Mining and Knowledge Discovery*, 1(3):291–316, 1997. doi: 10.1023/A:1009700419189.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A*, 222(594-604): 309–368, 1922.

R. A. Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, volume 22, pages 700–725. Cambridge University Press, 1925.

J. Foster and M. Bevis. Lognormal distribution of precipitable water in Hawaii. *Geochemistry, Geophysics, Geosystems*, 4(7):1065, 2003. doi: 10.1029/2002GC000478.

J. Foster, M. Bevis, and W. Raymond. Precipitable water and the lognormal distribution. *Journal of Geophysical Research: Atmospheres*, 111(D15), 2006. doi: 10.1029/2005JD006731.

Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, ICML'96, page 148–156, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1558604197.

D. Gellens. Combining regional approach and data extension procedure for assessing GEV distribution of extreme precipitation in Belgium. *Journal of Hydrology*, 268(1-4):113–126, 2002. doi: 10.1016/S0022-1694(02)00160-9.

S. Ghosh. *Kernel Smoothing - Principles, Methods and Applications.* John Wiley & Sons, Ltd, New Jersey, USA, 2017. doi: 10.1002/9781118890370.

W. J. Gibbs and J. V. Maher. Rainfall deciles as drought indicators. *Bureau of Meteorology Bulletin*, 1967. Commonwealth of Australia, Melbourne.

I. Gikhman and A. Skorokhod. *Introduction to the theory of random processes.* Dover Publications, INC, Mineola, New York, 1996. ISBN 0-486-69387-2.

T. Gneiting, M. G. Genton, and P. Guttorp. Geostatistical space-time models, stationarity, separability, and full symmetry. In B. Finkelstádt, L. Held, and V. Isham, editors, *Statistical Methods for Spatio-Temporal Systems*, volume 107, pages 151–175. Chapman & Hall, 2006.

B. A. Goldstein, E. C. Polley, and F. B. S. Briggs. Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, 10(1), 2011. doi: 10.2202/1544-6115.1691.

J. C. González-Hidalgo, S. M. Vicente-Serrano, D. Peña-Angulo, C. Salinas, M. Tomas-Burguera, and S. Beguería. High-resolution spatio-temporal analyses of drought episodes in the western mediterranean basin (spanish mainland, iberian peninsula). *Acta Geophysica*, 66:381–392, 2018.

Google Earth. Crete island, Greece, 2015. SIO, NOAA, U.S. Navy, NGA, GEBCO. DigitalGlobe 2015. Data retrieved from: `http://www.earth.google.com`.

P. Goovaerts. *Geostatistics for Natural Resources Evaluation.* Applied geostatistics series. Oxford University Press, 1997. ISBN 9780195115383.

P. Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129, 2000.

E. Goudenhoofdt and L. Delobbe. Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrology and Earth System Sciences*, 13(2):195–203, 2009. doi: 10.5194/hess-13-195-2009.

R. Grossman (edit.), G. Seni, and J. Elder. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool, California, United States, 2010. doi: 10.2200/S00240ED1V01Y200912DMK002.

J. C. Grzegorz, F. K. Witold, and V. Gabriele. Product-error-driven uncertainty model for probabilistic quantitative precipitation estimation with nexrad data. *Journal of Hydrometeorology*, 8(6):1325–1347, 2007. doi: 10.1175/2007JHM814.1.

H. Guan, J. L. Wilson, and O. Makhnin. Geostatistical mapping of mountain precipitation incorporating autosearched effects of terrain and climatic characteristics. *Journal of Hydrometeorology*, 6(6):1018–1031, 2005.

D. Guibal and A. Remacre. *Local Estimation of the Recoverable Reserves: Comparing Various Methods with the Reality on a Porphyry Copper Deposit*, pages 435–448. Springer Netherlands, Dordrecht, 1984. ISBN 978-94-009-3699-7. doi: 10.1007/978-94-009-3699-7_25.

N. B. Guttman. Accepting the standardized precipitation index: a calculation algorithm 1. *JAWRA Journal of the American Water Resources Association*, 35(2):311–322, 1999. doi: 10.1111/j.1752-1688.1999.tb03592.x.

J. W. Hansen, A. Challinor, A. Ines, T. Wheeler, and V. Moron. Translating climate forecasts into agricultural terms: advances and challenges. *Climate Research*, 33:27–41, 2006. doi: 10.3354/cr033027.

G. H. Hargreaves and Z. A. Samani. Estimating potential evapotranspiration. *Journal of the irrigation and Drainage Division*, 108(3):225–230, 1982.

T. I. Harrold, A. Sharma, and S. J. Sheather. A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research*, 39(12), 2003. doi: 10.1029/2003WR002570.

D. L. Hartmann, A. M. G. Klein Tank, M. Rusticucci, L. V. Alexander, S. Bronnimann, Y. Charabi, F. J. Dentener, E. J. Dlugokencky, D. R. Easterling,

A. Kaplan, B. J. Soden, P. W. Thorne, M. Wild, and P. M. Zhai. *Observations: Atmosphere and Surface*, chapter 2, pages 159–254. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. [Stocker, T. F. and Qin, D. and Plattner, G.-K. and Tignor, M. and Allen, S. K. and Boschung, J. and Nauels, A. and Xia, Y. and Bex, V. and Midgley, P. M. (eds.)].

B. Hassler and A. Lauer. Comparison of reanalysis and observational precipitation datasets including ERA5 and WFDE5. *Atmosphere*, 12(11), 2021. ISSN 2073-4433. doi: 10.3390/atmos12111462.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag New York, 2 edition, 2009. doi: 10.1007/978-0-387-84858-7.

M. Hayes, M. Svoboda, N. Wall, and M. Widhalm. The lincoln declaration on drought indices: Universal meteorological drought index recommended. *Bulletin of the American Meteorological Society*, 92(4):485–488, 2011. doi: 10.1175/2010BAMS3103.1.

M. J. Hayes. *Drought Indices*. John Wiley & Sons, Ltd, 2006. ISBN 9780471743989. doi: 10.1002/0471743984.vse8593.

R. R. Heim. A review of twentieth century drought indices used in the United States. *Bull. American Meteorological Society*, 83:1149–1165, 2002.

G. Heinze, C. Wallisch, and D. Dunkler. Variable selection – a review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3):431–449, 2018. doi: 10.1002/bimj.201700067.

Hellenic Statistical Authority. Demographic and social characteristics of the Resident Population of Greece according to the 2011 Population - Housing Census revision of 20/3/2014, September 2014. Online: http://www.statistics.gr/en/2011-census-pop-hous.

Hellenic Statistical Authority. Census Results of Population and Housing 2021 - ELSTAT 2021, March 2023. Online: https://www.statistics.gr/en/2021-census-res-pop-results.

C. Hellström, D. Chen, C. Achberger, and J. Räisänen. Comparison of climate change scenarios for sweden based on statistical and dynamical downscaling of monthly precipitation. *Climate Research*, 19(1):45–55, 2001.

Y. Hong and J. Gourley. *Radar Hydrology: Principles, Models, and Applications*. CRC Press, 2015. doi: 10.1201/b17921.

J. R. M. Hosking and J. R. Wallis. *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, 1997. doi: 10.1017/CBO9780511529443. Appendix A.9.

A. Y. Hou, R. K. Kakar, S. Neeck, A. A. Azarbarzin, C. D. Kummerow, M. Kojima, R. Oki, K. Nakamura, and T. Iguchi. The global precipitation measurement mission. *Bulletin of the American Meteorological Society*, 95(5):701–722, 2014. doi: 10.1175/BAMS-D-13-00164.1.

D. T. Hristopulos. Spartan Gibbs random field models for geostatistical applications. *SIAM Journal on Scientific Computing*, 24(6):2125–2162, 2003.

D. T. Hristopulos. Covariance functions motivated by spatial random field models with local interactions. *Stochastic Environmental Research and Risk Assessment*, 29(3):739—754, 2015a.

D. T. Hristopulos. Stochastic local interaction (SLI) model: Bridging machine learning and geostatistics. *Computers and Geosciences*, 85(Part B):26–37, 2015b.

D. T. Hristopulos. *Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers*. Springer Netherlands, 2020. doi: 10.1007/978-94-024-1918-4.

D. T. Hristopulos and V. D. Agou. Stochastic local interaction model with sparse precision matrix for space-time interpolation. *Spatial Statistics*, 40:100403, 2020. doi: 10.1016/j.spasta.2019.100403.

D. T. Hristopulos and S. N. Elogne. Analytical properties and covariance functions for a new class of generalized gibbs random fields. *IEEE Transactions on Information Theory*, 53(12):4667–4679, 2007.

D. T. Hristopulos and I. C. Tsantili. Space–time covariance functions based on linear response theory and the turning bands method. *Spatial Statistics*, 22, Part 2:321–337, 2017.

D. T. Hristopulos, A. Pavlides, V. D. Agou, and P. Gkafa. Stochastic local interaction model: An alternative to kriging for massive datasets. *Mathematical Geosciences*, 53:1907–1949, 2021. doi: 10.1007/s11004-021-09957-7.

Q. Hu, Z. Li, L. Wang, Y. Huang, Y. Wang, and L. Li. Rainfall spatial estimations: A review from spatial interpolation to multi-source data merging. *Water*, 11 (579), 2019. ISSN 2073-4441. doi: 10.3390/w11030579.

C. Huntingford, E. S. Jeffers, M. B. Bonsall, H. M. Christensen, T. Lees, and H. Yang. Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, 14(12):124007, nov 2019. doi: 10.1088/1748-9326/ab4e55.

K. Imaoka, M. Kachi, H. Fujii, H. Murakami, M. Hori, A. Ono, T. Igarashi, K. Nakagawa, T. Oki, Y. Honda, and H. Shimoda. Global change observation mission (gcom) for monitoring carbon, water cycles, and climate change. *Proceedings of the IEEE*, 98(5):717–734, 2010. doi: 10.1109/JPROC.2009.2036869.

A. C. Imeson. Desertification Indicator System for Mediterranean Europe, 2023. URL `https://esdac.jrc.ec.europa.eu/public_path/shared_folder/projects/DIS4ME/indicator_descriptions/potential_evapotranspiration.htm`. Foundation for Sustainable Development (3D-EC).

A. V. Ines and J. W. Hansen. Bias correction of daily GCM rainfall for crop simulation studies. *Agricultural and Forest Meteorology*, 138(1):44–53, 2006. ISSN 0168-1923. doi: 10.1016/j.agrformet.2006.03.009.

Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on

Climate Change, Cambridge, United Kingdom and New York, NY, USA, 2013. IPCC, Cambridge University Press. doi: 10.1017/CBO9781107415324. [Stocker, T. F. and Qin, D. and Plattner, G.-K. and Tignor, M. and Allen, S. K. and Boschung, J. and Nauels, A. and Xia, Y. and Bex, V. and Midgley, P. M. (eds.)].

IPCC. Guidance on the use of data: What is a GCM?, 2015. URL `http://www.ipcc-data.org`.

IPCC. *Global Warming of 1.5°C.An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2018. doi: 10.1017/9781009157940. [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)].

IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021. doi: 10.1017/9781009157896. [Masson-Delmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou (eds.)].

E. H. Isaaks and R. M. Srivastava. *Applied Geostatistics.* Oxford University Press, New York, USA, 1989. ISBN 9780195050134.

E. Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1): 253–258, 1925.

K. Jabeen, M. A. Khan, M. Alhaisoni, U. Tariq, Y.-D. Zhang, A. Hamza, A. u. Mickus, and R. Damaševičius. Breast cancer classification from ultrasound images using probability-based optimal deep learning feature fusion. *Sensors*, 22(3), 2022. doi: 10.3390/s22030807.

C. M. Jarque and A. K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3):255–259, 1980. ISSN 0165-1765. doi: 10.1016/0165-1765(80)90024-5.

C. M. Jarque and A. K. Bera. A test for normality of observations and regression residuals. *International Statistical Review / Revue Internationale de Statistique*, 55(2):163–172, 1987. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1403192.

A. G. Journel. Fundamentals of geostatistics in five lessons. *American Geophysical Union*, page 45, 1989.

A. G. Journel and C. J. Huijbregts. *Mining Geostatistics*. Blackburn Press, 2003. ISBN 9781930665910.

S. Justin and W. Eric F. Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, IPCC AR4 simulations. *Climate Dynamics*, 31:79–105, 2008.

S. Kant. Trend and variability of hourly intensity of rainfall over eastern and northern part of uttar pradesh during 1969-2014. *MAUSAM*, 69(4):577—588, 2018. doi: 10.54302/mausam.v69i4.422.

M. Karamouz, K. Rasouli, and S. Nazif. Development of a hybrid index for drought prediction: Case study. *Journal of Hydrologic Engineering*, 14(6): 617–627, 2009. doi: 10.1061/(ASCE)HE.1943-5584.0000022.

M. Kardar. *Statistical Physics of Fields*. Cambridge University Press, 2007.

D. Kavetski, G. Kuczera, and S. W. Franks. Bayesian analysis of input uncertainty in hydrological modeling: 1. theory. *Water Resources Research*, 42(3), 2006. doi: 10.1029/2005WR004368.

B. Kedem and L. S. Chiu. On the lognormality of rain rate. *Proceedings of the National Academy of Sciences*, 84(4):901–905, 1987. ISSN 0027-8424. doi: 10.1073/pnas.84.4.901.

B. Kedem, L. S. Chiu, and G. R. North. Estimation of mean rain rate: Application to satellite observations. *Journal of Geophysical Research: Atmospheres*, 95 (D2):1965–1972, 1990. doi: 10.1029/JD095iD02p01965.

C. A. Keller, M. J. Evans, J. N. Kutz, and S. Pawson. Machine learning and air quality modeling. In *2017 IEEE International Conference on Big Data (Big Data)*, pages 4570–4576, 2017. doi: 10.1109/BigData.2017.8258500.

J. A. Keyantash and J. A. Dracup. An Aggregate Drought Index: Assessing drought severity based on fluctuations in the hydrologic cycle and surface water storage. *Water Resources*, 40, 2004. W09304.

M. Kirkham. Chapter 28 - potential evapotranspiration. In M. Kirkham, editor, *Principles of Soil and Plant Water Relations (Second Edition)*, pages 501–514. Academic Press, Boston, second edition edition, 2014. ISBN 978-0-12-420022-7. doi: 10.1016/B978-0-12-420022-7.00028-8.

P. K. Kitanidis and R. W. Lane. Maximum likelihood parameter estimation of hydrologic spatial processes by the Gauss-Newton method. *Journal of Hydrology*, 79(1):53–71, 1985. ISSN 0022-1694. doi: 10.1016/0022-1694(85)90181-7.

J. Kittredge. *Forest Influences: The Effects of Woody Vegetation on Climate, Water, and Soil, with Applications to the Conservation of Water and the Control of Floods and Erosion.* American forestry series. McGraw-Hill Book Company, 1948. ISBN 9780486209425.

E. Kjellström, L. Bärring, D. Jacob, R. Jones, G. Lenderink, and C. Schär. Modelling daily temperature extremes: recent climate and future changes over europe. *Climatic Change*, 81(Suppl 1):249–265, 2007.

F. N. Kogan. Droughts of the late 1980s in the united states as derived from noaa polar-orbiting satellite data. *Bulletin of the American Meteorology Society*, 76:655–668, 1995. URL https://journals.ametsoc.org/view/journals/bams/76/5/1520-0477_1995_076_0655_dotlit_2_0_co_2.xml.

R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–43, San Mateo, CA, 1995.

A. Kolovos, G. Christakos, D. Hristopulos, and M. L. Serre. Methods for generating non-separable spatiotemporal covariance models with potential environmental applications. *Advances in Water Resources*, 27(8):815–830, 2004.

A. G. Koutroulis, A.-E. K. Vrohidou, and I. K. Tsanis. Spatiotemporal characteristics of meteorological drought for the island of crete. *Journal of Hydrometeorology*, 12(2):206–226, 2011. doi: 10.1175/2010JHM1252.1.

E. Koutroulis and D. Kolokotsa. Design optimization of desalination systems power-supplied by PV and W/G energy sources. *Desalination*, 258(1-3):171–181, 2010.

E. Koutroulis, G. Petrakis, V. Agou, A. Malisovas, D. Hristopulos, P. Partsinevelos, A. Tripolitsiotis, N. Halouani, P. Ailliot, M. Boutigny, V. Monbet, D. Allard, A. Cuzol, D. Kolokotsa, E. Varouchakis, K. Kokolakis, and S. Mertikas. Site selection and system sizing of desalination plants powered with renewable energy sources based on a web-gis platform. *International Journal of Energy Sector Management*, 2021. doi: 10.1108/IJESM-04-2021-0018.

D. Koutsoyiannis. Statistics of extremes and estimation of extreme rainfall: II. empirical investigation of long rainfall records. *Hydrological Sciences Journal*, 49(4):591–610, 2004. doi: 10.1623/hysj.49.4.591.54424.

R. S. Kovats, R. Valentini, L. M. Bouwer, E. Georgopoulou, D. Jacob, E. Martin, M. Rounsevell, and J.-F. Soussana. Europe. In *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter 23, pages 1267–1326. Cambridge University

Press, Cambridge, United Kingdom and New York, NY, USA, 2014. [Barros, V. R., Field, C. B. and Dokken, D. J. and Mastrandrea, M. D. and Mach, K. J. and Bilir, T. E. and Chatterjee, M. and Ebi, K. L. and Estrada, Y. O. and Genova, R. C. and Girma, B. and Kissel, E. S. and Levy, A. N. and MacCracken, S. and Mastrandrea, P. R. and White, L. L. (eds.)].

D. G. Krige. A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52(6):119–139, 1951. doi: 10.2307/3006914.

M. Kubat, R. Holte, and S. Matwin. Learning when negative examples abound. In M. van Someren and G. Widmer, editors, *Machine Learning: ECML-97*, pages 146–153, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. ISBN 978-3-540-68708-5.

M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite Radar images. *Machine Learning*, 30(2):195–215, 1998. doi: 10.1023/A:1007452223027.

C. Kummerow, W. Barnes, T. Kozu, J. Shiue, and J. Simpson. The tropical rainfall measuring mission (trmm) sensor package. *Journal of Atmospheric and Oceanic Technology*, 15(3):809–817, 1998. URL `https://journals.ametsoc.org/view/journals/atot/15/3/1520-0426_1998_015_0809_ttrmmt_2_0_co_2.xml`.

P. C. Kyriakidis and A. G. Journel. Geostatistical space–time models: a review. *Mathematical geology*, 31:651–684, 1999.

M. N. Legasa, R. Manzanas, A. Calviño, and J. M. Gutiérrez. A posteriori random forests for stochastic downscaling of precipitation by predicting probability distributions. *Water Resources Research*, 58(4):e2021WR030272, 2022. doi: 10.1029/2021WR030272.

C. Li, V. P. Singh, and A. K. Mishra. Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water resources research*, 48(3), 2012.

H. Li, J. Sheffield, and E. F. Wood. Bias correction of monthly precipitation and temperature fields from intergovernmental panel on climate change ar4 models using equidistant quantile matching. *Journal of Geophysical Research: Atmospheres*, 115(D10), 2010. doi: 10.1029/2009JD012882.

M. Li and Q. Shao. An improved statistical approach to merge satellite rainfall estimates and raingauge data. *Journal of Hydrology*, 385(1):51–64, 2010. ISSN 0022–1694. doi: 10.1016/j.jhydrol.2010.01.023.

Z. Li, F. Brissette, and J. Chen. Finding the most appropriate precipitation probability distribution for stochastic weather generation and hydrological modelling in nordic watersheds. *Hydrological Processes*, 27(25):3718–3729, 2013. doi: 10.1002/hyp.9499.

K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis. Machine learning in agriculture: A review. *Sensors*, 18(8), 2018. ISSN 1424-8220. doi: 10.3390/s18082674.

G.-Y. Lien, E. Kalnay, T. Miyoshi, and G. J. Huffman. Statistical Properties of Global Precipitation in the NCEP GFS Model and TMPA Observations for Data Assimilation. *Monthly Weather Review*, 144(2):663–679, 2016. doi: 10.1175/MWR-D-15-0150.1.

H. W. Lilliefors. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318): 399–402, 1967. doi: 10.1080/01621459.1967.10482916.

H. W. Lilliefors. On the Kolmogorov-Smirnov test for the exponential distribution with mean unknown. *Journal of the American Statistical Association*, 64(325): 387–389, 1969. doi: 10.1080/01621459.1969.10500983.

K. S. Lomax. Business failures: Another example of the analysis of failure data. *Journal of the American Statistical Association*, 49(268):847–852, 1954. doi: 10.1080/01621459.1954.10501239.

R. E. López. The lognormal distribution and cumulus cloud populations. *Monthly Weather Review*, 105(7):865–872, 1977. URL https://journals.ametsoc.org/view/journals/mwre/105/7/1520-0493_1977_105_0865_tldacc_2_0_co_2.xml.

H. Lull. *Soil Compaction on Forest and Range Lands*. Miscellaneous publication (United States. Department of Agriculture). Forest Service, U.S. Department of Agriculture, 1959.

Z. Ma, J. Xu, K. He, X. Han, Q. Ji, T. Wang, W. Xiong, and Y. Hong. An updated moving window algorithm for hourly-scale satellite precipitation downscaling: A case study in the southeast coast of china. *Journal of Hydrology*, 581:124378, 2020. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2019.124378.

K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984. ISSN 00063444. URL http://www.jstor.org/stable/2336405.

K. V. Mardia and A. J. Watkins. On multimodality of the likelihood in the spatial linear model. *Biometrika*, 76(2):289–295, 1989. ISSN 00063444. URL http://www.jstor.org/stable/2336662.

D. Marks and J. Dozier. Climate and energy exchange at the snow surface in the alpine region of the sierra nevada. *Water Resources Research*, 28(11):3043–3054, 1992.

J. Marotzke, C. Jakob, S. Bony, P. A. Dirmeyer, P. A. O'Gorman, E. Hawkins, S. Perkins-Kirkpatrick, C. Le Quere, S. Nowicki, K. Paulavets, S. I. Seneviratne, B. Stevens, and M. Tuma. Climate research must sharpen its view. *Nature climate change*, 7(2):89–91, 2017.

R. Martine, M. Helmut, D. Olivier, S. Dirk, G. Karl, K. Jürgen P., and M. Anette. Heat and drought 2003 in Europe: a climate synthesis. *Ann. For. Sci.*, 63(6): 569–577, 9 2006. doi: 10.1051/forest:2006043.

G. Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963. doi: 10.2113/gsecongeo.58.8.1246.

G. Matheron. *Isofactorial Models and Change of Support*, pages 449–467. Springer Netherlands, Dordrecht, 1984. ISBN 978-94-009-3699-7. doi: 10.1007/978-94-009-3699-7_26.

MATLAB. *version 9.5.0 (R2018b)*. The MathWorks Inc., Natick, Massachusetts, 2018.

T. Mavromatis and D. Voulanas. Evaluating ERA-Interim, Agri4Cast, and E-OBS gridded products in reproducing spatiotemporal characteristics of precipitation and drought over a data poor region: The Case of Greece. *International Journal of Climatology*, 41(3):2118–2136, 2020. doi: 10.1002/joc.6950.

A. E. Maxwell, T. A. Warner, and F. Fang. Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, 39(9):2784–2817, 2018. doi: 10.1080/01431161.2018.1433343.

T. McKee, N. Doesken, and J. Kleist. The relationship of drought frequency and duration to time scales. *Proceedings of the 8th Conference on Applied Climatology*, pages 179—-184, 1993. Boston.

G. J. McLachlan, K.-A. Do, and C. Ambroise. *Analyzing microarray gene expression data*. Wiley & Sons, New Jersey, USA, 2004.

M. Mendez, B. Maathuis, D. Hein-Griggs, and L.-F. Alvarado-Gamboa. Performance evaluation of bias correction methods for climate change monthly precipitation projections over Costa Rica. *Water*, 12(2):482, 2020.

Meteoclub. Spatial distribution of the mean annual index of the hourly rainrate (mm/hr) in greece, 2013. URL `https://www.meteoclub.gr`. (in Greek).

P. C. D. Milly and K. A. Dunne. Potential evapotranspiration and continental drying. *Nature Climate Change*, 6(10):946–949, 2016.

A. K. Mishra and V. P. Singh. A review of drought concepts. *Journal of Hydrology*, 391(1–2):202–216, 2010. doi: 10.1016/j.jhydrol.2010.07.012.

L. Mitas and H. Mitasova. *Spatial Interpolation*, chapter 34, pages 481–492. Wiley, New York, USA, 2005. ISBN 9780471735458. URL `http://www.geos.`

ed.ac.uk/~gisteac/gis_book_abridged/. [Longley, P. A. and Goodchild, M. F. and Maguire, D. J. and Rhind, D. W. (eds.)].

A. Mokhtar, M. Jalali, H. He, N. Al-Ansari, A. Elbeltagi, K. Alsafadi, H. G. Abdo, S. S. Sammen, Y. Gyasi-Agyei, and J. Rodrigo-Comino. Estimation of spei meteorological drought using machine learning algorithms. *IEEE Access*, 9:65503–65523, 2021. doi: 10.1109/ACCESS.2021.3074305.

A. Molini, G. G. Katul, and A. Porporato. Revisiting rainfall clustering and intermittency across different climatic regimes. *Water Resources Research*, 45 (11), 2009. doi: 10.1029/2008WR007352. W11403.

S.-H. Moon, Y.-H. Kim, Y. H. Lee, and B.-R. Moon. Application of machine learning to an early warning system for very short-term heavy rainfall. *Journal of Hydrology*, 568:1042–1054, 2019. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2018.11.060.

F. J. Moral. Comparison of different geostatistical approaches to map climate variables: application to precipitation. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 30(4):620–631, 2010. doi: 10.1002/joc.1913.

T. Mosthaf and A. Bárdossy. Regionalizing nonparametric models of precipitation amounts on different temporal scales. *Hydrology and Earth System Sciences*, 21(5):2463–2481, 2017. doi: 10.5194/hess-21-2463-2017.

K. P. Moustris, I. K. Larissi, P. T. Nastos, and A. G. Paliatsos. Precipitation forecast using artificial neural networks in specific regions of Greece. *Water resources management*, 25:1979–1993, 2011.

G. Mussardo. *Statistical Field Theory*. Oxford University Press, 2010.

E. A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9(1):141–142, 1964.

P. T. Nastos and C. S. Zerefos. Spatial and temporal variability of consecutive dry and wet days in Greece. *Atmospheric Research*, 94(4):616–628, 2009. doi: 10.1016/j.atmosres.2009.03.009.

S. D. Nerantzaki and S. M. Papalexiou. Tails of extremes: Advancing a graphical method and harnessing big data to assess precipitation extremes. *Advances in Water Resources*, 134:103448, 2019. ISSN 0309-1708. doi: 10.1016/j.advwatres.2019.103448.

G. M. Nielson. A method for interpolating scattered data based upon a minimum norm network. *Mathematics of Computation*, 40(161):253–271, 1983.

S. Niemeyer. New drought indices. *Options Méditerranéennes. Série A: Séminaires Méditerranéens*, 80:267–274, 2008.

R. H. Norden. A survey of maximum likelihood estimation. *International Statistical Review / Revue Internationale de Statistique*, 40(3):329–354, 1972. ISSN 03067734, 17515823. URL `http://www.jstor.org/stable/1402471`.

P. A. O'Gorman and T. Schneider. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *Proceedings of the National Academy of Sciences*, 106(35):14773–14777, 2009.

R. A. Olea. *Geostatistics for Engineers and Earth Scientists*. Springer US, 1999. ISBN 9780792385233.

C. Onyutha and P. Willems. Influence of spatial and temporal scales on statistical analyses of rainfall variability in the river nile basin. *Dynamics of Atmospheres and Oceans*, 77:26–42, 2017. ISSN 0377-0265. doi: 10.1016/j.dynatmoce.2016.10.008.

D. Opitz and R. Maclin. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999. doi: 10.1613/jair.614.

J. M. Ortiz, B. Oz, and C. V. Deutsch. *A Step by Step Guide to Bi-Gaussian Disjunctive Kriging*, pages 1097–1102. Springer Netherlands, Dordrecht, 2005. ISBN 978-1-4020-3610-1. doi: 10.1007/978-1-4020-3610-1_114.

P. A. O'Gorman. Precipitation extremes under climate change. *Current climate change reports*, 1:49–59, 2015.

W. C. Palmer. Meteorological drought. *Research Paper No 45*, 1965. US Weather Bureau, Washington, DC.

W. C. Palmer. Keeping track of crop moisture conditions, nationwide: The new Crop Moisture Index. *Weatherwise 21*, pages 156–161, 1968.

G. Papacharalampous, H. Tyralis, and D. Koutsoyiannis. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. *Water resources management*, 32(15):5207–5239, 2018.

G. Papacharalampous, H. Tyralis, A. Doulamis, and N. Doulamis. Comparison of tree-based ensemble algorithms for merging satellite and earth-observed precipitation data at the daily time scale. *Hydrology*, 10(2), 2023a. ISSN 2306-5338. doi: 10.3390/hydrology10020050.

G. Papacharalampous, H. Tyralis, A. Doulamis, and N. Doulamis. Comparison of machine learning algorithms for merging gridded satellite and earth-observed precipitation data. *Water*, 15(4), 2023b. ISSN 2073-4441. doi: 10.3390/w15040634.

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

S. M. Papalexiou and F. Serinaldi. Random fields simplified: Preserving marginal distributions, correlations, and intermittency, with applications from rainfall to humidity. *Water Resources Research*, 56(2):e2019WR026331, 2020. doi: 10.1029/2019WR026331.

S. M. Papalexiou, A. AghaKouchak, and E. Foufoula-Georgiou. A diagnostic framework for understanding climatology of tails of hourly precipitation extremes in the United States. *Water Resources Research*, 54(9):6725–6738, 2018. doi: 10.1029/2018WR022732.

S. M. Papalexiou, F. Serinaldi, and E. Porcu. Advancing space-time simulation of random fields: From storms to cyclones and beyond. *Water Resources Research*, 57(8):e2020WR029466, 2021. doi: 10.1029/2020WR029466.

A. Papoulis and S. U. Pillai. *Probability, Random Variables and Stochastic Process*. McGraw-Hill Inc., New York, fourth edition, 2002.

N.-W. Park, P. C. Kyriakidis, and S. Hong. Geostatistical integration of coarse resolution satellite precipitation products and rain gauge data to map precipitation at fine spatial resolutions. *Remote Sensing*, 9(3), 2017. ISSN 2072-4292. doi: 10.3390/rs9030255.

A. Pavlides, D. T. Hristopulos, C. Roumpos, and Z. Agioutantis. Spatial modeling of lignite energy reserves for exploitation planning and quality control. *Energy*, 93:1906–1917, 2015. doi: 10.1016/j.energy.2015.10.049.

A. Pavlides, V. D. Agou, and D. T. Hristopulos. Non-parametric kernel-based estimation of probability distributions for precipitation modeling, 2021.

A. Pavlides, V. D. Agou, and D. T. Hristopulos. Non-parametric kernel-based estimation and simulation of precipitation amount. *Journal of Hydrology*, 612: 127988, 2022. doi: 10.1016/j.jhydrol.2022.127988.

A. G. Pavlides. *Development of new geostatistical methods for spatial analysis and applications in reserves estimation and quality characteristics of coal deposits*. PhD thesis, Technical University of Crete, School of Mineral Resources Engineering, Chania, Greece, 2016.

M. J. Pazzani, C. J. Merz, P. M. Murphy, K. M. Ali, T. Hume, and C. Brunk. Reducing misclassification costs. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning*, pages 217—225, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

H. L. Penman. Natural evaporation from open water, bare soil and grass. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 193(1032):120–145, 1948.

G. W. Peters, I. Nevat, S. G. Nagarajan, and T. Matsui. Spatial warped gaussian processes: Estimation and efficient field reconstruction. *Entropy*, 23(10), 2021. doi: 10.3390/e23101323.

C. Piani, J. O. Haerter, and E. Coppola. Statistical bias correction for daily precipitation in regional climate models over europe. *Theoretical and Applied Climatology*, 99:187–192, 2010. doi: 10.1007/s00704-009-0134-9.

D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer New York, New York, NY, 1999. ISBN 978-1-4612-1554-7. doi: 10.1007/978-1-4612-1554-7_2.

F. Porcù, L. Milani, and M. Petracca. On the uncertainties in validating satellite instantaneous rainfall estimates with raingauge operational network. *Atmospheric Research*, 144:73–81, 2014. ISSN 0169-8095. doi: 10.1016/j.atmosres.2013.12.007. Perspectives of Precipitation Science - Part II.

R. W. Portmann, S. Solomon, and G. C. Hegerl. Spatial and seasonal patterns in climate change, temperatures, and precipitation across the United States. *Proceedings of the National Academy of Sciences*, 106(18):7324–7329, 2009.

O. Rakovec, L. Samaniego, V. Hari, Y. Markonis, V. Moravec, S. Thober, M. Hanel, and R. Kumar. The 2018–2020 multi-year drought sets a new benchmark in europe. *Earth's Future*, 10(3):e2021EF002394, 2022. doi: 10.1029/2021EF002394.

C. E. Rasmussen and C. K. I. Williams. Gaussian processes for machine learning. Technical report, MA: MIT Press, Cambridge, MA, 2006. URL `www.GaussianProcess.org/gpml`.

Region of Crete Information Bull. Region of crete (2002) sustainable management of water resources in crete., November 2002. pp. 24 (in Greek).

R. J. Renka, R. L. Renka, and A. K. Cline. A triangle-based c^1 interpolation method. *The Rocky Mountain journal of mathematics*, pages 223–237, 1984.

J. Rhee, J. Im, and G. J. Carbone. Monitoring agricultural drought for arid and humid regions using multi-sensor remote sensing data. *Remote Sensing of Environment*, 114(12):2875–2887, 2010. ISSN 0034-4257. doi: 10.1016/j.rse.2010.07.005.

H. Rho and J. H. Kim. Modelling the entire range of daily precipitation using phase-type distributions. *Advances in Water Resources*, 123:210–224, 2019. ISSN 0309-1708. doi: 10.1016/j.advwatres.2018.11.014.

C. W. Richardson. Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water resources research*, 17(1):182–190, 1981.

J. Rivoirard. On the structural link between variables in kriging with external drift. *Mathematical Geology*, 34(7):797–808, 2002.

V. Rodriguez-Galiano, M. Chica-Olmo, and M. Chica-Rivas. Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the rodalquilar area, Southern Spain. *International Journal of Geographical Information Science*, 28(7):1336–1354, 2014. doi: 10.1080/13658816.2014.885527.

L. Rokach and O. Maimon. *Data Mining with Decision Trees*. WORLD SCIENTIFIC, Singapore, 2nd edition, 2014. ISBN 978-9814590075. doi: 10.1142/9097.

D. Rolnick, P. L. Donti, L. H. Kaack, K. Kochanski, A. Lacoste, K. Sankaran, A. S. Ross, N. Milojevic-Dupont, N. Jaques, A. Waldman-Brown, A. S. Luccioni, T. Maharaj, E. D. Sherwin, S. K. Mukkavilli, K. P. Kording, C. P. Gomes, A. Y. Ng, D. Hassabis, J. C. Platt, F. Creutzig, J. Chayes, and Y. Bengio. Tackling climate change with machine learning. *ACM Comput. Surv.*, 55(2), feb 2022. ISSN 0360-0300. doi: 10.1145/3485128.

R. Rotter and S. C. Van De Geijn. Climate change effects on plant growth, crop yield and live stock. *Clim. Change*, 43:651–681, 1999.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, Boca Raton, FL, 2005.

H. Sauvageot. The probability density function of rain rate and the estimation of rainfall by area integrals. *Journal of Applied Meteorology and Climatology*, 33(11):1255–1262, 1994. URL `https://journals.ametsoc.org/view/journals/apme/33/11/1520-0450_1994_033_1255_tpdfor_2_0_co_2.xml`.

R. E. Schapire. The strength of weak learnability. *Machine learning*, 5:197–227, 1990. doi: 10.1007/BF00116037.

M. Scheuerer. Probabilistic quantitative precipitation forecasting using Ensemble Model Output Statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1086–1096, 2014. doi: 10.1002/qj.2183.

J. Schmidli, C. Frei, and P. L. Vidale. Downscaling from gcm precipitation: a benchmark for dynamical and statistical downscaling methods. *International Journal of Climatology*, 26(5):679–689, 2006. doi: 10.1002/joc.1287.

F. Schmitt, S. Vannitsem, and A. Barbosa. Modeling of rainfall time series using two-state renewal processes and multifractals. *Journal of Geophysical Research: Atmospheres*, 103(D18):23181–23193, 1998.

C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Rusboost: Improving classification performance when training data is skewed. In *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008. doi: 10.1109/ICPR.2008.4761297.

D.-J. Seo, J. Breidenbach, and E. Johnson. Real-time estimation of mean field bias in radar rainfall data. *Journal of Hydrology*, 223(3):131–147, 1999. ISSN 0022-1694. doi: 10.1016/S0022-1694(99)00106-7.

B. A. Shafer and L. E. Dezman. Development of a surface water supply index (swsi) to assess the severity of drought conditions in snowpack runoff areas. In *Proceedings of the Western Snow Conference*, pages 164–175, Colorado State University, Fort Collins, CO, 1982.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3–4):591–611, 12 1965. ISSN 0006-3444. doi: 10.1093/biomet/52.3-4.591.

E. Sharifi, B. Saghafian, and R. Steinacker. Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *Journal of Geophysical Research: Atmospheres*, 124 (2):789–805, 2019. doi: 10.1029/2018JD028795.

A. Sharma and U. Lall. A nonparametric approach for daily rainfall simulation. *Mathematics and Computers in Simulation*, 48(4):361–371, 1999. ISSN 0378-4754. doi: 10.1016/S0378-4754(99)00016-6.

I. Shiklomanov. *Water in Crisis: A Guide to the World's Fresh Water Resources*. Oxford University Press, New York, 1993.

T. Shoji and H. Kitaura. Statistical and geostatistical analysis of rainfall in central Japan. *Computers & Geosciences*, 32(8):1007–1024, 2006. doi: 10.1016/j.cageo.2004.12.012.

A. J. Simmons, P. Poli, D. P. Dee, P. Berrisford, H. Hersbach, S. Kobayashi, and C. Peubey. Estimating low-frequency variability and trends in atmospheric temperature using era-interim. *Quarterly Journal of the Royal Meteorological Society*, 140(679):329–353, 2014. doi: 10.1002/qj.2317.

E. Snelson, C. E. Rasmussen, and Z. Ghahramani. Warped Gaussian processes. *Advances in Neural Information Processing Systems*, 16:337–344, 2004.

D. P. Solomatine and A. Ostfeld. Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics*, 10(1):3–22, 2008. ISSN 1464-7141. doi: 10.2166/hydro.2008.015.

E. Soriano, L. Mediero, and C. Garijo. Selection of bias correction methods to assess the impact of climate change on flood frequency curves. *Water*, 11(2266), 2019. doi: 10.3390/w11112266.

J. Spinoni, P. Barbosa, A. De Jager, N. McCormick, G. Naumann, J. V. Vogt, D. Magni, D. Masante, and M. Mazzeschi. A new global database of meteorological drought events from 1951 to 2016. *Journal of Hydrology: Regional Studies*, 22:100593, 2019.

J. Stagge, D. Kingston, L. Tallaksen, and D. Hannah. Observed drought indices show divergence across europe. *Scientific Reports*, 7(1), Dec. 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-14283-2.

J. H. Stagge, L. M. Tallaksen, L. Gudmundsson, A. F. Van Loon, and K. Stahl. Candidate distributions for climatological drought indices (SPI and SPEI). *International Journal of Climatology*, 35(13):4027–4040, 2015. doi: 10.1002/joc.4267.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, USA, 1999.

G. Stickler. Educational Brief - Solar Radiation and the Earth System. National Aeronautics and Space Administration, 2015. URL `https://web.archive.org/web/20160425164312/http://education.gsfc.nasa.gov/experimental/July61999siteupdate/inv99Project.Site/Pages/science-briefs/ed-stickler/ed-irradiance.html`. Archived from the original on 25 April 2016. Retrieved 14 April 2022.

I. Strangeways. *Precipitation: Theory, Measurement and Distribution*. Cambridge University Press, 2006. ISBN 9780511535772. doi: 10.1017/CBO9780511535772. Cambridge Books Online.

Y. Sun, B. Li, and M. G. Genton. Geostatistics for large datasets. In E. Porcu, J. Montero, and M. Schlather, editors, *Advances and Challenges in Space-time Modelling of Natural Events*, Lecture Notes in Statistics, pages 55–77. Springer Berlin Heidelberg, 2012.

J. D. Tarpley, S. R. Schneider, and R. L. Money. Global vegetation indices from the noaa-7 meteorological satellite. *Journal of Climate and Applied Meteorology*, 23:491–494, 1984. URL `https://journals.ametsoc.org/view/journals/apme/23/3/1520-0450_1984_023_0491_gviftn_2_0_co_2.xml`.

W. Terink, R. T. W. L. Hurkmans, P. J. J. F. Torfs, and R. Uijlenhoet. Bias correction of temperature and precipitation data for regional climate model application to the rhine basin. *Hydrology and Earth System Sciences Discussions*, 6:5377–5413, 2009. doi: 10.5194/hessd-6-5377-2009.

C. Teutschbein and J. Seibert. Regional climate models for hydrological impact studies at the catchment scale: A review of recent modeling strategies. *Geography Compass*, 4(7):834–860, 2010. doi: 10.1111/j.1749-8198.2010.00357.x.

C. Teutschbein and J. Seibert. Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, 456-457:12–29, 2012. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2012.05.052.

T. Thadewald and H. Büning. Jarque–Bera Test and its Competitors for Testing Normality – A Power Comparison. *Journal of Applied Statistics*, 34(1):87–105, 2007. doi: 10.1080/02664760600994539.

C. W. Thornthwaite. An approach toward a rational classification of climate. *Geographical Review*, 38(1):55–94, 1948. ISSN 00167428. URL `http://www.jstor.org/stable/210739`.

Z. Tian, B. Nijssen, G. J. Huffman, and D. P. Lettenmaier. Evaluation of real-time satellite precipitation data for global drought monitoring. *Journal of Hydrometeorology*, 15(4):1651–1660, 2014. doi: 10.1175/JHM-D-13-0128.1.

E. Todini. A bayesian technique for conditioning radar precipitation estimates to rain-gauge measurements. *Hydrology and Earth System Sciences*, 5(2):187–199, 2001. doi: 10.5194/hess-5-187-2001.

G. Tsakiris and H. Vangelis. Towards a drought watch system based on spatial SPI. *Water Resources Management*, 18 (1):1573–1650, 2004.

G. Tsakiris, D. Pangalou, and H. Vangelis. Regional drought assessment based on the Reconnaissance Drought Index (RDI). *Water Resources Management*, 21 (5):821–833, 2007a.

G. Tsakiris, D. Tigkas, H. Vangelis, and D. Pangalou. *Regional Drought Identification and Assessment. Case Study in Crete*, volume 62, pages 169–191. Springer Netherlands, Dordrecht, Netherlands, 2007b. ISBN 978-1-4020-5924-7. doi: 10.1007/978-1-4020-5924-7_9.

S. Tushaus. Bayesian statistics yields new insights into topographically influenced precipitation. physicstoday, 2014.

H. Tyralis, G. Papacharalampous, and A. Langousis. A brief review of random forests for water scientists and practitioners and their recent history in water resources. *Water*, 11(5), 2019. ISSN 2073-4441. doi: 10.3390/w11050910.

C. G. Tzanis, I. Koutsogiannis, K. Philippopoulos, and D. Deligiorgi. Recent climate trends over greece. *Atmospheric Research*, 230:104623, 2019. ISSN 0169-8095. doi: 10.1016/j.atmosres.2019.104623.

United Kingdom Meteorological Office. , October 2022. Online: http://www.metoffice.gov.uk/.

U.S. Environmental Protection Agency. Climate Change Indicators in the United States. U.S. and Global Mean Temperature and Precipitation. Report, 2021. [Last accessed 14 April 2022 from `https://cfpub.epa.gov/roe/indicator.cfm?i=89`].

B. Škerlak, M. Sprenger, and H. Wernli. A global climatology of stratosphere–troposphere exchange using the ERA-Interim data set from 1979 to 2011. *Atmospheric and Physics*, 14(2):913–937, 2014. doi: 10.5194/acp-14-913-2014.

S. Varma and R. Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(91):1471–2105, 2006. doi: 10.1186/1471-2105-7-91.

E. A. Varouchakis. *Geostatistical Analysis and Space-Time Models of Aquifer Levels: Application to Mires Hydrological Basin in the Prefecture of Crete.* PhD thesis, Technical University of Crete, 2012.

E. A. Varouchakis and D. T. Hristopulos. Improvement of groundwater level prediction in sparsely gauged basins using physical laws and local geographic features as auxiliary variables. *Advances in Water Resources*, 52:34–49, 2013. doi: 10.1016/j.advwatres.2012.08.002.

E. A. Varouchakis and D. T. Hristopulos. Comparison of spatiotemporal variogram functions based on a sparse dataset of groundwater level variations. *Spatial Statistics*, 34:100245, 2019. doi: 10.1016/j.spasta.2017.07.003.

E. A. Varouchakis, D. T. Hristopulos, and G. P. Karatzas. Improving kriging of groundwater level data using nonlinear normalizing transformations – a field application. *Hydrological Sciences Journal*, 57(7):1–16, 2012.

E. A. Varouchakis, G. A. Corzo, G. P. Karatzas, and A. Kotsopoulou. Spatio-temporal analysis of annual rainfall in Crete, Greece. *Acta Geophysica*, 66(3): 319–328, 2018. doi: 10.1007/s11600-018-0128-z.

L. Vasiliades, A. Loukas, and N. Liberis. A water balance derived drought index for pinios river basin, greece. *Water Resources Management*, 25(4):1573–1650, 2011. doi: 10.1007/s11269-010-9665-1.

C. Vera, G. Silvestri, B. Liebmann, and P. González. Climate change scenarios for seasonal precipitation in South America from IPCC-AR4 models. *Geophysical research letters*, 33(13), 2006.

A. Verdin, C. Funk, B. Rajagopalan, and W. Kleiber. Kriging and local polynomial methods for blending satellite-derived and gauge precipitation estimates to support hydrologic early warning systems. *IEEE Transactions on Geoscience and Remote Sensing*, 54(5):2552–2562, 2016.

S. M. Vicente-Serrano, S. Beguería, and J. I. López-Moreno. A multi-scalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of Climate*, 23:1696–1718, 2010. doi: 10.1175/2009JCLI2909.1.

S. M. Vicente-Serrano, J.-I. Lopez-Moreno, S. Beguería, J. Lorenzo-Lacruz, A. Sanchez-Lorenzo, J. M. García-Ruiz, C. Azorin-Molina, E. Morán-Tejeda, J. Revuelto, R. Trigo, et al. Evidence of increasing drought severity caused by temperature rise in southern europe. *Environmental Research Letters*, 9(4): 044001, 2014.

A.-E. Vrochidou. *Spatiotemporal drought analysis and climate change impact on hydrometeorological variables for the island of Crete.* PhD thesis, Environmental Engineering of Technical University of Crete, Chania, Greece, 2013.

A.-E. K. Vrochidou and I. K. Tsanis. Assessing precipitation distribution impacts on droughts on the island of Crete. *Hazards Earth Syst. Sci.*, pages 1159—1171, 2012. doi: 10.5194/nhess-12-1159-2012.

H. Wackernagel. *Multivariate Geostatistics.* Springer, Berlin, Germany, 3rd edition, 2003. ISBN 978-3-642-0791-5. doi: 10.1007/978-3-662-05294-5.

Z. Wang, K. Guan, J. Sheffield, and E. F. Wood. Depiction of drought over sub-Saharan Africa using reanalyses precipitation datasets: depiction of drought using reanalyses. *Journal of Geophysical Research Atmospheres*, 121(18):10555–10574, 2016. doi: 10.1002/2016JD024858.

Z. Wang, Z. Zeng, C. Lai, W. Lin, X. Wu, and X. Chen. A regional frequency analysis of precipitation extremes in mainland china with fuzzy c-means and L-moments approaches. *International Journal of Climatology*, 37(S1):429–444, 2017. doi: 10.1002/joc.5013.

Z. Wang, W. Shi, W. Zhou, X. Li, and T. Yue. Comparison of additive and isometric log-ratio transformations combined with machine learning and regression kriging models for mapping soil particle size fractions. *Geoderma*, 365: 114214, 2020. ISSN 0016-7061. doi: 10.1016/j.geoderma.2020.114214.

L. V. Watrous. *Lasithi: A History of Settlement on a Highland Plain in Crete.* Princeton, NJ: American School of Classical Studies in Athens., 1982.

G. S. Watson. Smooth regression analysis. *Sankhya Ser. A*, 26(1):359–372, 1964.

R. Webster and M. A. Oliver. *Geostatistics for Environmental Scientists.* John Wiley & Sons, Ltd, 2007. ISBN 9780470517277.

K. Weghorst. The reclamation drought index: Guidelines and practical applications. *Bureau of Reclamation, Denver (CO)*, 1996.

N. Wells, S. Goddard, and M. J. Hayes. A self-calibrating palmer drought severity index. *Journal of Climate*, 17(12):2335–2351, 2004. URL https://journals.ametsoc.org/view/journals/clim/17/12/1520-0442_2004_017_2335_aspdsi_2.0.co_2.xml.

D. S. Wilks. Maximum likelihood estimation for the gamma distribution using data containing zeros. *Journal of Climate*, 3(12):1495–1501, 1990. URL https://journals.ametsoc.org/view/journals/clim/3/12/1520-0442_1990_003_1495_mleftg_2_0_co_2.xml.

D. S. Wilks and K. L. Eggleston. Estimating monthly and seasonal precipitation distributions using the 30-and 90-day outlooks. *Journal of Climate*, 5(3):252–259, 1992. URL https://www.jstor.org/stable/26197095.

C. L. Williams. Data analysis and introductory statistical inference with statistical formulae and tables with sas implementations, 2000.

P. S. Wilson and R. Toumi. A fundamental probability distribution for heavy rainfall. *Geophysical Research Letters*, 32(14), 2005. doi: 10.1029/2005GL022465.

WMO. North america heatwave almost impossible without climate change, 2021. URL https://public.wmo.int/en/media/news/north-america-heatwave-almost-impossible-without-climate-change. Retrieved 08.07.2021.

World Meteorological Organization (WMO). Standardized Precipitation Index User Guide (M. Svoboda, M. Hayes and D. Wood), 2012. URL https://library.wmo.int/doc_num.php?explnum_id=7768. (WMO-No. 1090).

World Meteorological Organization (WMO). Atlas of mortality and economic losses from weather, climate and water extremes (1970–2012), 2014. URL https://library.wmo.int/index.php?lvl=notice_display&id=16279.

World Meteorological Organization (WMO) and Global Water Partnership (GWP). Handbook of drought indicators and indices (m. svoboda and b.a. fuchs). integrated drought management programme (idmp), integrated drought management tools and guidelines series

2, 2016. URL https://public.wmo.int/en/resources/library/handbook-of-drought-indicators-and-indices.

H. Wu, M. J. Hayes, D. A. Wilhite, and M. D. Svoboda. The effect of the length of record on the standardized precipitation index calculation. *International Journal of Climatology*, 25:505–520, 2005. doi: 10.1002/joc.1142.

X. Wu, V. Kumar, J. Ross Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, et al. Top 10 algorithms in data mining. *Knowledge and information systems*, 14:1–37, 2008. doi: 10.1007/s10115-007-0114-2.

G. Xu and M. G. Genton. Tukey g-and-h random fields. *Journal of the American Statistical Association*, 112(519):1236–1249, 2017. doi: 10.1080/01621459.2016.1205501.

C. Xue, H. Wu, and X. Jiang. Temporal and spatial change monitoring of drought grade based on ERA5 analysis data and BFAST method in the Belt and Road area during 1989–2017. *Advances in Meteorology*, 2019:1–10, 2019. doi: 10.1155/2019/4053718.

L. Ye, L. S. Hanson, P. Ding, D. Wang, and R. M. Vogel. The probability distribution of daily precipitation at the point and catchment scales in the United States. *Hydrology and Earth System Sciences*, 22(12):6519–6531, 2018. doi: 10.5194/hess-22-6519-2018.

I. Yeo and R. A. Johnson. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959, 12 2000. ISSN 0006-3444. doi: 10.1093/biomet/87.4.954.

A. Zargar, R. Sadiq, B. Naser, and F. I. Khan. A review of drought indices. *Environmental Reviews*, 19(NA):333–349, 2011. doi: 10.1139/a11-013.

X. Zhang, L. Alexander, G. C. Hegerl, P. Jones, A. K. Tank, T. C. Peterson, B. Trewin, and F. W. Zwiers. Indices for monitoring changes in extremes based on daily temperature and precipitation data. *Wiley Interdisciplinary Reviews: Climate Change*, 2(6):851–870, 2011.

X. Zhu, Q. Zhang, C.-Y. Xu, P. Sun, and P. Hu. Reconstruction of high spatial resolution surface air temperature data across China: A new geo-intelligent multisource data-based machine learning technique. *Science of the Total Environment*, 665:300–313, 2019. doi: 10.1016/j.scitotenv.2019.02.077.

A. Ziegler and I. R. König. Mining data with random forests: current options for real-world applications. *WIREs Data Mining and Knowledge Discovery*, 4(1): 55–63, 2014. doi: 10.1002/widm.1114.

M. Ẑukoviĉ and D. T. Hristopulos. Spartan random processes in time series modeling. *Physica A-statistical Mechanics and Its Applications*, 387:3995–4001, 2008. doi: 10.1016/j.physa.2008.01.051.