

Geostatistical Analysis of Precipitation on the Island of Crete



Agou Vasiliki

School of Mineral Resources Engineering

Technical University of Crete

Advisor : Prof. Dionissios T. Hristopulos

A thesis submitted for the degree of

Master of Science

January 2016

I would like to dedicate this thesis to my loving parents ...

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dionissis Hristopulos for showing confidence in me by giving me the chance to work in his lab. Also for the continuous support of my Master's study and related research, for his patience, caring, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Master's research.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Panagiotis Partsinevelos and Prof. George P. Karatzas, for their insightful comments and encouragement, but also for the hard questions which motivated me to widen my research horizons in various perspectives.

My sincere thanks also goes to the members of the Geostatistics Laboratory in the School of Mineral Resources Engineering at the Technical University of Crete, Dr. Ivi Tsantili, Andreas Pavlides, Manolis Petrakis and Konstantinos Giannis for the stimulating conversations throughout my research. I specially would like to thank Ivi and Andreas for their friendship and support for all these years. I would also like to thank Emmanouil Varouchakis for sharing with me the data set of precipitation from the Region of Crete. NASA SRTM data sets were used. These data are distributed by the Land Processes Distributed Active Archive Center (LP DAAC). This work was supported by the project SPARTA 1591 implemented under the ARISTEIA Action of the operational programme Education and Lifelong Learning co-funded by the European Social Fund (ESF) and National Resources.

Besides my lab-mates, I would like to express my gratitude to all my friends for their support. Special thanks goes to Sarantis for his help, support, quiet patience and encouragement.

Last but not least, I would like to thank my family. Words cannot express how grateful I am to my mother, my father, my sister and my brother for supporting me spiritually, encouraging me with their best wishes and for all of the sacrifices that they have made on my behalf. Without them, none of this would have been meaningful.

Technical University of Crete, 2016

Vasiliki D. Agou

Abstract

Life on Earth is strongly linked to weather and climate; consequently knowing how to evaluate and interpret factors that affect such phenomena is vital for environmental and economic sustainability. The sensitivity of Europe to climate change has a distinct north–south gradient, with many studies indicating that Southern Europe will be more severely affected leading to warmer and drier climate, which will threaten its waterways, hydropower, agricultural production and timber harvests.

This thesis focuses on the investigation of precipitation trends on the island of Crete. In order to provide an efficient methodology to predict precipitation and to produce reliable precipitation maps, geostatistical approaches are required. The data available for this study are monthly measurements of precipitation from 54 rain gauges around Crete, with records from 1948 until 2012. The data for the time period from 1948 until 1964 present a lot of missing records and have been used only for a preliminary analysis.

For the analysis of precipitation measurements for the time period 1965 until 2012 a geostatistical methodology has been developed. At first, original monthly measurements of precipitation were analysed over different time and spatial scales, in order to determine their statistical properties. Then, an investigation on the topographic parameters that contribute to precipitation amount and distribution through space were conducted. Multilinear regression was used to determine a trend model for every year. The precipitation residuals were shown to be suitable for the application of regression kriging. In order to apply kriging, the spatial variability of the data sets was investigated

using the Spartan variogram family. To determine the trend at every point in space, the Digital Elevation Model of Crete, created through Geographical Information Systems, was used. The model was validated using leave-one-out cross-validation and measures of uncertainty were estimated. The analysis of the validation measures shows that the differences between the true values and the estimated values are small. Finally, precipitation maps were generated using the geostatistical model constructed.

The most important conclusions that are derived in this thesis are:

1. Spatial correlations of the precipitation were identified both locally and between the Eastern and Western parts of the island.
2. The statistical analysis of precipitation data is not consistent with a significant change in annual precipitation for the period 1965 to 2012.
3. The elevation is the most important topographic parameter for the estimation of annual precipitation levels.
4. There is increased uncertainty in the precipitation estimates in the coastal areas of the island.
5. The network of the meteorological stations is very sparse, especially in the western part of the island.

The above findings improve our understanding of the space-time precipitation patterns and its statistical distribution over Crete. The proposed model provides a flexible and mathematical advanced tool with potential applications in effective water resources management.

Περίληψη

Η ζωή στη Γη είναι άρρηκτα συνδεδεμένη με τον καιρό και το κλίμα. Κατά συνέπεια, πληροφορίες οι οποίες βοηθούν στην ερμηνεία των παραγόντων που επηρεάζουν τέτοιου είδους φαινόμενα είναι ζωτικής σημασίας για το περιβάλλον και την οικονομική βιωσιμότητα. Η ευαισθησία της Ευρώπης στην κλιματική αλλαγή αυξάνεται με κατεύθυνση από το βορρά προς το νότο, με πολλές μελέτες να δείχνουν ότι η νότια Ευρώπη θα πληγεί πιο σοβαρά. Πιο συγκεκριμένα, η νότια Ευρώπη αναμένεται να γίνει πιο θερμή και ξηρή, επιφέροντας συνέπειες στους υδατικούς πόρους, την υδροηλεκτρική ενέργεια, την γεωργική παραγωγή και την ξυλεία της. Εξαιτίας της ευαισθησίας της Μεσογείου σε τέτοια φαινόμενα, πολλές περιοχές της έχουν μελετηθεί εκτενώς. Οι μελέτες αυτές υποδεικνύουν μία τάση μείωσης της ατμοσφαιρικής κατακρήμνισης, κυρίως στις νοτιοανατολικές περιοχές.

Η παρούσα εργασία επικεντρώνεται στη διερεύνηση τάσεων που σχετίζονται με τις βροχοπτώσεις στο νησί της Κρήτης. Προκειμένου να δημιουργηθεί μια αποτελεσματική μεθοδολογία για την πρόβλεψη χωρικών και χρονικών τάσεων της βροχόπτωσης και την παραγωγή αντίστοιχων χαρτών, απαιτείται η χρήση της Γεωστατιστικής. Η παρούσα διατριβή χρησιμοποιεί δεδομένα βροχόπτωσης από βροχομετρικούς σταθμούς στην Κρήτη. Τα δεδομένα αυτά είναι μηνιαία και αναφέρονται στην χρονική περίοδο από το 1948 έως το 2012. Χρησιμοποιήθηκαν δεδομένα από 54 σταθμούς και για την χρονική

περίοδο από το 1948 έως και το 1964 τα δεδομένα παρουσιάζουν πολλές ελλείψεις και δεν χρησιμοποιήθηκαν πέρα από μια διερευνητική ανάλυση.

Για την μελέτη των δεδομένων βροχόπτωσης στο χρονικό διάστημα από 1965 έως 2012 αναπτύχθηκε γεωστατιστική μεθοδολογία ανάλυσης. Αρχικά, οι πρωτογενείς μηνιαίες μετρήσεις της βροχόπτωσης αναλύθηκαν σε διαφορετικές χρονικές και χωρικές κλίμακες, ώστε να εκτιμηθούν οι στατιστικές τους ιδιότητες.

Στη συνέχεια, διερευνήθηκε η συμβολή των τοπογραφικών παραμέτρων στο ύψος και την χωρική κατανομή της βροχόπτωσης. Με εφαρμογή της πολλαπλής γραμμικής παλινδρόμησης (multilinear regression) δημιουργήθηκε ένα μοντέλο χωρικής τάσης της βροχόπτωσης ανά έτος. Κατόπιν αφαίρεσης της τάσης από τη συνολική βροχόπτωση προκύπτει το χωρικό πεδίο υπολοίπων της βροχόπτωσης, το οποίο χρησιμοποιείται για γεωστατιστική ανάλυση με τη μέθοδο regression kriging. Η χωρική μεταβλητότητα των δεδομένων ποσοτικοποιείται με τη βοήθεια της συνάρτησης του βαριογράμματος, για την οποία χρησιμοποιείται το Σπαρτιάτικο μοντέλο βαριογράμματος. Για την κατασκευή χαρτών βροχόπτωσης απαιτείται και η εκτίμηση των χωρικών τάσεων σε κάθε κόμβο του πλέγματος χαρτογράφησης. Για να επιτευχθεί αυτό, χρησιμοποιήθηκε το ψηφιακό μοντέλο εδάφους της Κρήτης, το οποίο δημιουργήθηκε μέσω της πλατφόρμας των Γεωγραφικών Πληροφοριακών Συστημάτων.

Για την επαλήθευση του γεωστατιστικού μοντέλου, χρησιμοποιήθηκε η μέθοδος της διασταυρωτικής επιβεβαίωσης (leave-one-out cross-validation) και υπολογίσθηκαν στατιστικά μέτρα επίδοσης. Η ανάλυση

των μέτρων αυτών έδειξε ότι η απόκλιση μεταξύ των εκτιμώμενων και των πραγματικών τιμών της βροχόπτωσης είναι μικρή. Τέλος, δημιουργήθηκαν χάρτες βροχόπτωσης χρησιμοποιώντας το παραπάνω γεωστατιστικό μοντέλο.

Τα πιο σημαντικά συμπεράσματα που προκύπτουν από την παρούσα διατριβή είναι τα εξής:

1. Ανιχνεύθηκε συστηματική διαφορά της βροχόπτωσης ανάμεσα στην Ανατολική και την Δυτική Κρήτη.
2. Η στατιστική ανάλυση των δεδομένων βροχόπτωσης δεν υποστηρίζει σημαντική μεταβολή της ετήσιας βροχόπτωσης για το χρονικό διάστημα 1965 έως 2012.
3. Το υψόμετρο αποτελεί την σημαντικότερη από τις τοπογραφικές παραμέτρους για τον προσδιορισμό της ετήσιας βροχόπτωσης.
4. Παρατηρείται αυξημένη αβεβαιότητα στις εκτιμήσεις της βροχόπτωσης στις παράλιες περιοχές του νησιού.
5. Το βροχομετρικό δίκτυο είναι ιδιαίτερα αραιό, ιδιαίτερα στο δυτικό τμήμα του νησιού.

Τα παραπάνω ευρήματα βελτιώνουν την κατανόηση μας, όσον αφορά τις χωροχρονικές μεταβολές της βροχόπτωσης, καθώς και την στατιστική κατανομή της βροχόπτωσης σε όλη την έκταση της Κρήτης. Το προτεινόμενο μοντέλο παρέχει ένα ευέλικτο και μαθηματικά εξελιγμένο εργαλείο το οποίο μπορεί να βρει εφαρμογές στην πιο αποτελεσματική διαχείριση των υδάτινων πόρων.

Contents

Contents	ix
List of Figures	xiii
List of Tables	xxiii
1 Introduction	1
1.1 Objectives	2
1.2 Innovation	2
1.3 Thesis Outline	3
1.4 Basic Concepts and Literature Review	3
1.4.1 Precipitation and Drought	5
1.4.1.1 Precipitation	5
1.4.1.2 Drought	9
1.4.2 Drought Indices	10
1.4.3 Climate Models	11
1.4.4 Statistical Modeling	13
1.4.5 Probabilistic Modeling of Precipitation	13
1.4.5.1 The Normal Distribution	14
1.4.5.2 The Gamma Distribution	16
1.4.5.3 Generalized Extreme Value Distribution	19
1.4.5.4 Student's t Distribution	23
1.4.6 The Role of Spatial and Temporal Scales	24

2	Geostatistical Methods	27
2.1	Randomness	28
2.2	Random Fields	28
2.3	Basic Concepts of Random Fields	30
2.4	Probability Density Function	30
2.5	Statistical Homogeneity	31
2.6	Statistical Isotropy	32
2.7	Moments	32
2.7.1	Expectation	33
2.7.2	Variance	33
2.7.3	Skewness	34
2.7.4	Kurtosis	35
2.8	Covariance Function	36
2.9	Variogram Function	37
2.10	Variogram Models	41
2.10.1	Spartan Variogram Model	43
2.11	Variogram Estimation	45
2.11.1	Method of Moments	45
2.12	Spatial Estimation	48
2.12.1	Kriging Methods	50
2.12.1.1	Simple Kriging (SK)	51
2.12.1.2	Ordinary Kriging (OK)	54
2.12.1.3	Kriging with a Trend Model – Universal Kriging(UK)	56
2.12.1.4	Regression Kriging – Residual Kriging (RK)	57
2.13	Spatial Model Validation	59
2.13.1	Cross-validation Approach	60
2.13.2	Cross-validation Error Measures	63
3	Exploratory Data Analysis	67
3.1	Information about the Study Area	67
3.1.1	Temperature	69
3.1.2	Humidity	71

3.1.3	Topography	71
3.1.4	Precipitation	72
3.2	Precipitation Records	73
3.3	Statistical Analysis of Monthly Data	75
3.4	Precipitation Fields and Series	76
3.5	Precipitation Comparison in the Eastern and Western Regions . .	87
4	Results of Geostatistical Analysis	95
4.1	Model of Annual Precipitation Field	95
4.1.1	Topographic Trend	96
4.1.2	Variogram Modeling	99
4.1.3	Regression Kriging	101
4.1.4	Leave-One-Out Cross Validation	103
4.2	Model of Average Monthly Precipitation	105
4.2.1	Topographic Trend	105
4.2.2	Variogram Modeling	109
4.2.3	Regression Kriging	111
4.2.4	Leave one Out Cross Validation	111
4.3	Spatiotemporal Patterns of Precipitation	115
5	Conclusions	121
	Appendix A	125
	References	173

List of Figures

1.1	(a) Global mean surface temperature (GMST) anomalies relative to a 1961-1990 climatology based on HadCRUT4 annual data. The straight black lines are least squares trends for 1901-2012, 1901-1950 and 1951-2012. (b) Same data as in (a), with smoothing spline (solid curve) and the 90% confidence interval on the smooth curve (dashed lines). Note that the (strongly overlapping) 90% confidence intervals for the least squares lines in (a) are omitted for clarity. Figure taken from Hartmann et al. [2013]	4
1.2	Projected changes in heavy precipitation (in %) in winter and summer from 1971–2000 to 2071–2100 for the RCP8.5 scenario based on the ensemble mean of different regional climate models (RCMs) nested in different general circulation models (GCMs). Figure taken from EEA [2015]	7
1.3	Water Cycle. Figure taken from Shiklomanov [1993]	8
1.4	Schematic illustrating the relation between General Circulation Models and Regional Circulation Models. Figure taken from WMO [2015]	12
1.5	Plots of the probability density function and the cumulative distribution function for the Normal or Gaussian Distribution.	15
1.6	Plots of the probability density function and the cumulative distribution function for the Gamma Distribution.	18
1.7	Plots of the probability density function and the cumulative distribution function for the Generalized Extreme Value Distribution.	20

1.8	Plots of the probability density function and the cumulative distribution function for the Student's t Distribution.	24
2.1	Schematic figure of the region $B(\mathbf{r}_k)$ around the distance vector [Hristopulos, 2012]	46
3.1	Morphological map of the island of Crete, Greece [Google Earth, 2015].	68
3.2	Digital Elevation Model of Crete	72
3.3	Geomorphological map of Crete showing the locations of the 54 stations (yellow circles) with rain gauges used in this study [Google Earth, 2015].	73
3.4	Matrix showing the missing data problem related to the precipitation records used. Rows correspond to stations while columns correspond to years. Circles (blue) mark locations with complete yearly records while with stars mark years that miss one or more monthly measurements.	74
3.5	Left: Monthly precipitation series at the station C1 which exhibits the maximum mean monthly precipitation. Right: Monthly precipitation series at the station L10 which has recorded the minimum mean monthly precipitation.	76
3.6	Empirical and model cumulative probability functions for the two precipitation time series shown in Fig. 3.5 C1 and L10 correspond to the empirical cdfs (blue line). The model fits include the following: "GEV" refers to the generalized extreme value distribution (red line), "Normal" refers to the the Gaussian distribution (green line), "Exponential" refers to the exponential distribution (cyan line), and "Gamma" refers to the gamma distribution (brown line).	77

- 3.7 Empirical and model cumulative probability functions for the wet period data (consists of monthly observations from October till March) of the two precipitation time series shown in Fig. 3.5 C1 and L10 correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line). 78
- 3.8 Empirical and model cumulative probability functions for the dry period data (consists of monthly observations from April till September) of the two precipitation time series shown in Fig. 3.5 C1 and L10 correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line). 79
- 3.9 Three-dimensional stem plots representing the average monthly precipitation field (left) and the annual precipitation field (right) for the year 2002. Each stem refers to the average monthly precipitation at every station based on the Eq. (3.1) (left) and to the cumulative monthly precipitation based on the Eq. (3.3) (right) for the year 2002. 79
- 3.10 Empirical cumulative distribution function (cdf) and fits to model cdfs for the average monthly precipitation. The empirical cdf is given by the staircase (blue line), the Generalized Extreme Value distribution is given by the red line, the normal distribution is given by the dark green line, the exponential distribution by the cyan line, and the Gamma distribution by the brown line. 80

3.11	Empirical cumulative distribution function (cdf) and fits to model cdfs for the annual precipitation. The empirical cdf is given by the staircase (blue line), the Generalized Extreme Value distribution is given by the red line, the normal distribution is given by the dark green line, the exponential distribution by the cyan line, and the Gamma distribution by the brown line.	82
3.12	Empirical and model cumulative probability functions for the average monthly precipitation time series of station C1 (left) and L10 (right) correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).	88
3.13	Empirical and model cumulative probability functions for the annual precipitation time series of station C1 (left) and L10 (right) correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).	88
3.14	Flow chart of the preliminary data processing and geostatistical methodology applied to the spatial database. Data Organisation & Transformations refers to the preliminary investigation of the needed transformations in order to keep the physical meaning of the data, but also to find the appropriate data set . “GEV” refers to the generalized extreme value distribution	89
3.15	Time series of average monthly precipitation (left) and annual precipitation (right) for the Western (blue line) and the Eastern (red line) regions of Crete based on Eqs. (3.7) and (3.9) respectively. The sample mean of the precipitation in the West (East) is indicated by the green (magenta) straight line in the diagrams.	91

3.17	Correlation plots between Western and Eastern regions of Crete for the annual precipitation. On the left diagonal, the histograms show the probability densities of the Western and Eastern annual precipitation time series (Eq. (3.10)). On the right diagonal the scatterplots of the data are presented. The Pearson's correlation coefficient calculated by Eq. (2.77) equals to 0.71 (left) and Spearman's correlation coefficient based on Eq. (2.78) equals to 0.71 (right).	92
3.16	Correlation plots between Western and Eastern regions of Crete for the average monthly precipitation. On the left diagonal, the histograms show the probability densities of the Western and Eastern average monthly precipitation time series (Eq. (3.8)). On the right diagonal the scatterplots of the data are presented. The Pearson's correlation coefficient calculated by Eq. (2.77) equals to 0.69 (left) and Spearman's correlation coefficient based on Eq. (2.78) equals to 0.76 (right).	93
4.1	Multilinear regression of the precipitation field on the predictors which comprise longitude, latitude, and elevation, for the annual precipitation field in the year 1971. The trend equation for this year is given by $f = -4.1919 \cdot 10^4 - 0.0019x + 0.0012y + 0.5688z$. The coordinates x , y , z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y representing the easting and the northing coordinates measured in meters, and z being the elevation measured in meters above sea level.	98
4.2	Scatter plot of the annual precipitation field values and the precipitation residuals obtained after the application of multilinear regression according to the trend equation shown in Fig. 4.1, for the year 1971. The blue circles correspond to the annual precipitation field values, while the red asterisks correspond to the precipitation residuals. The square boxes mark the highest values for the respective data set.	99

- 4.3 Empirical and model probability density function (left) and normal probability plot (right) for the year 1971 of the precipitation residuals provided in Fig. 4.2. On the left, bars correspond to the empirical histogram of the data (blue line), while the model fits include the following: “t Scale–Location” refers to the Student’s t distribution (red line), “Logistic” refers to the Logistic distribution (blue dotted line), “GEV” refers to the generalized extreme value distribution (brown line), and “Normal” refers to the Gaussian distribution (gray line). On the right, cross markers are used for the sample data, while the solid line corresponds to the model normal probability (red line). 100
- 4.4 Empirical variogram (crosses) and theoretical Spartan variogram model (continuous blue line) based on Eq. (2.30) for the precipitation residuals of the year 1971 (shown in Fig. 4.2). The horizontal axis is calibrated using normalized distances; one unit corresponds to a lag distance of approximately 2.8 km in x direction, 1.2 km in y direction, and 2.4 m in z direction. Bars count the available pairs of sample points for every lag distance. The estimated parameters for the theoretical model are as follows: nugget variance $\sigma_n^2 = 173.9258 \text{ mm}^2$, $\eta_0 = 1.9764e + 05 \text{ mm}^2$, $\eta_1 = -0.8643$, $\xi = 9.9620$. 101
- 4.5 Map of estimated annual precipitation residuals for 1971, based on the Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.2. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA ’87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters. 102
- 4.6 Map of estimated annual precipitation for 1971, based on Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.2. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA ’87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters. 103

4.7	Map of estimated coefficient of variation of annual precipitation for 1971, based on Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.2. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.	104
4.8	Kriging-based leave-one-out cross validation predictions versus sample values for 1971 precipitation (based on the data set shown in Fig. 4.2).	104
4.9	Histogram of leave-one-out cross validation errors for 1971 precipitation (based on the data set shown in Fig. 4.2).	105
4.10	Comparison between the actual precipitation values of the year 1971 at each station and the predictions derived using leave-one-out cross validation.	106
4.11	Multilinear regression of the precipitation field on the predictors in which comprise longitude, latitude, and elevation, for the average monthly precipitation field in the year 1971. The trend equation for this year is given by $f = -3.4932 \cdot 10^3 - 0.0002x + 0.0010y + 0.0474z$. The coordinates x, y, z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y representing the easting and northing coordinates measured in meters, and z being the elevation measured in meters above sea level.	107
4.12	Scatter plot of the average monthly precipitation field values and the precipitation residuals obtained after the application of multilinear regression according to the trend equation shown in Fig. 4.11, for the year 1971. The blue circles correspond to the monthly precipitation field values, while the red asterisks correspond to the precipitation residuals. The square boxes mark the highest values for the respective data set.	108

- 4.13 Empirical and model probability density function (left) and normal probability plot (right) for the year 1971 of the precipitation residuals provided in Fig. 4.12. On the left, bars correspond to the empirical histogram of the data (blue line), while the model fits include the following: “t Scale–Location” refers to the Student’s t distribution (red line), “Logistic” refers to the Logistic distribution (blue dotted line), “GEV” refers to the generalized extreme value distribution (brown line), and “Normal” refers to the Gaussian distribution (gray line). On the right, cross markers are used for the sample data, while the solid line corresponds to the model normal probability (red line). 109
- 4.14 Empirical variogram (crosses) and theoretical Spartan variogram model (continuous blue line) based on Eq. (2.30) for the precipitation residuals of the year 1971 (shown in Fig. 4.12). The horizontal axis is calibrated using normalized distance; one unit corresponds to a lag distance of approximately 2.8 km in x direction, 1.2 km in y direction, and 2.4 m in z direction. Bars count the available pairs of sample points for every lag distance. The estimated parameters for the theoretical model are as follows: nugget variance $\sigma_n^2 = 2.2204e-16 \text{ mm}^2$, $\eta_0 = 2.6979e+03 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 15.8604$. 110
- 4.15 Map of estimated average monthly precipitation for 1971, based on Spartan variogram model (Eq. 2.30) and the data set shown in Fig. 4.12. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA ’87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters. 111
- 4.16 Map of estimated coefficient of variation of average monthly precipitation for 1971, based on Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.12. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA ’87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters. 112

4.17	Kriging-based leave-one-out cross validation predictions versus sample values for 1971 precipitation (based on the data set shown in Fig. 4.12).	113
4.18	Histogram of leave-one-out cross validation errors for the 1971 precipitation (based on the data set shown in Fig. 4.12).	113
4.19	Comparison between the actual precipitation values of the year 1971 at each station and the predictions derived using leave-one-out cross validation.	114
4.20	Overall trend calculated from the recorded data sets. The averaged annual precipitation values are calculated with the calculation of the mean values of annual precipitation of every station in Crete for every year (left), while the averaged average monthly precipitation values are calculated with the calculation of the mean values of average monthly precipitation of every station in Crete for every year (right). The solid red line shows the trend equation, while the red dashed line shows the 95% coefficient bounds.	115
4.21	Overall trend calculated from the estimations. The averaged annual precipitation values are calculated with the calculation of the mean values of the estimated annual precipitation of every station in Crete for every year (left), while the averaged average monthly precipitation values are calculated with the calculation of the mean values of estimated average monthly precipitation of every station in Crete for every year (right). The solid red line shows the trend equation, while the red dashed line shows the 95% coefficient bounds.	116
4.22	Kriging maps of precipitation from 2004 until 2009, created based on the Spartan variogram model (Eq. (2.30)). The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.	118

List of Tables

1.1	Estimates of the mean change in global (spatial) mean surface temperature between 1901 and 2012, 1901 and 1950, and 1951 and 2012, obtained from the linear (least squares) and nonlinear (smoothing spline) trend models. Half-widths of the 90% confidence intervals are also provided for the estimated changes based on the two trend methods. Table taken from Hartmann et al. [2013] .	5
1.2	Impacts of climate extremes in the last decade in Europe. Table taken from Kovats et al. [2014] .	6
3.1	Crops cultivated in Crete and their respective farming area. The units are in $(\times 10^3 \text{ha})^a$. Table taken from Chartzoulakis et al. [2001] .	69
3.2	Monthly temperature minima and maxima recorded on Crete between minimum and maximum recorded temperatures ($^{\circ}\text{C}$).	70
3.3	Location coordinates, optimal GEV distribution type and parameters for all the stations on the island, calculated with the average monthly precipitation time series data set (based on Eq. (3.5)). The coordinates x, y, z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y being northing and easting measured in meters, and z being the elevation measured in meters above sea level. The GEV distribution and its parameters are defined in Eq. (1.17). “Weibull” refers to the Reversed Weibull distribution, that belongs to the Generalized Extreme Value distribution family.	83

3.4	Location coordinates, optimal GEV distribution type and parameters for all the stations on the island, calculated with the annual precipitation time series data set (based on Eq. (3.6)). The coordinates x , y , z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y being northing and easting measured in meters, and z being the elevation measured in meters above sea level. The GEV distribution and its parameters are defined in Eq. (1.17). “Weibull” refers to the Reversed Weibull distribution, that belongs to the Generalized Extreme Value distribution family. . . .	85
4.1	Statistics of multilinear regression for all the possible combinations of longitude (x), latitude (y) and elevation (z), for the annual precipitation field in the year 1971 (Eq. (3.4)). The coordinates x , y , z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y representing the longitude and latitude measured in meters, and z being the elevation measured in meters above sea level. The value of the coefficient of determination R^2 shows how much of the variability in the response variable (precipitation) is explained by the model (e.g. for x , y , z : 74%). The F–statistic gives the relationship between the response variable and the predictor variables. The p–value is used in the context of null hypothesis testing of zero correlation in order to quantify the idea of statistical significance of evidence. The p–value refers to the F–statistic test, lower value are preferable.	97
4.2	Trend model parameters, for annual precipitation field in the year 1971 (Eq. (3.4)).	97
4.3	Summary statistics for the precipitation data set depicted in Fig. 4.2. The parameter n is the number of available stations of the year. .	97

4.4	Cross validation performance measures calculated through leave-one-out cross validation for the precipitation data of year 1971 for annual recorded precipitation data set. ME: Mean error (bias) (Eq. (2.72)), MAE: Mean absolute error (Eq. (2.73)), MARE: Mean absolute relative error (INF if z contains zeros) (Eq. (2.74)), RMSE: Root mean square error (Eq. (2.75)), RMSRE: Root mean square relative error (INF if the actual precipitation value is zero) (Eq. (2.76)), RP: Pearson's linear correlation coefficient (Eq. (2.77)), RS: Spearman (rank) correlation coefficient (Eq. (2.78)), ErrMin: Difference between minimum prediction and minimum sample value, ErrMax: Difference between maximum prediction and maximum sample value.	106
4.5	Trend model parameters, for average monthly precipitation field in the year 1971 (Eq. (3.2)).	107
4.6	Summary statistics for the precipitation data set depicted in Fig. 4.12. The parameter n is the number of available stations of the year. .	109
4.7	Cross validation performance measures calculated through leave-one-out cross validation for the precipitation data of year 1971 for the average monthly precipitation data set. ME: Mean error (bias) (Eq. (2.72)), MAE: Mean absolute error (Eq. (2.73)), MARE: Mean absolute relative error (INF if z contains zeros) (Eq. (2.74)), RMSE: Root mean square error (Eq. (2.75)), RMSRE: Root mean square relative error (INF if the actual precipitation value is zero) (Eq. (2.76)), RP: Pearson's linear correlation coefficient (Eq. (2.77)), RS: Spearman (rank) correlation coefficient (Eq. (2.78)), ErrMin: Difference between minimum prediction and minimum sample value, ErrMax: Difference between maximum prediction and maximum sample value	114

Chapter 1

Introduction

During the last decades, geostatistical approaches have been successfully applied in different environmental and earth sciences disciplines. Geostatistics has methods that can be applied in difficult situations, such as sparse measurements, is able to provide space–time predictions for variables with environmental and socioeconomic importance and provides estimates of the uncertainty of the results. Nowadays, geostatistical methods are at the core of new research methodologies that are being proposed.

Because of the uncertainties involved, probabilistic approaches are required to enable water resources managers to undertake analyses of risk under scenarios of climate change. Mathematical techniques are being developed to construct probability distributions for specified outcomes. Most studies of climate change impact on water stress in countries assess demand and supply on an annual basis. Analysis at monthly or higher temporal resolution scale is desirable, since changes in seasonal patterns and the probability of extreme events may offset the positive effect of increased availability of water resources [Kovats et al., 2014].

This thesis is motivated by the need for accurate interpolation methodologies that can help to determine the spatiotemporal variability of precipitation. Thus, below we introduce spatiotemporal methodologies applied on field data from rain gauges. The methods that we propose combine spatial trend models based on features of the local topography with fluctuations the correlations of which are modeled by means of the Spartan variogram family.

In addition, this thesis is motivated by the dramatic decrease of available water

resources and predictions that identify climate change as a potential trigger for upcoming socioeconomic disaster. More specifically, many Mediterranean basins with significant agricultural activity are threatened by desertification. Therefore, accurate spatiotemporal modeling of precipitation is significant and needed for the identification of future climatic conditions, in order to enable policy makers to design and implement more efficient water resources management.

1.1 Objectives

The main objective of this thesis is to develop and improve the results of standard interpolation methodologies leading to more precise mapping of precipitation on the island of Crete (Greece). The broader aim is to better understand the distributions of precipitation around Crete at different space and time scale. A secondary goal is to study the impact of the island's topography on the rainfall variations and its contribution on precipitation measurements. Another objective is to introduce the recently established Spartan variogram family in climatology. The final goal of this research is to use stochastic methods for the spatiotemporal simulation of precipitation in one of the most mountainous islands of Europe which has a terrain marked by big plains between mountains and climate that belongs to two different climatic zones.

1.2 Innovation

The present research addresses the geostatistical analysis of precipitation data on the island of Crete. The methodology that we present includes stochastic methods for the spatial and temporal analysis of precipitation patterns in Crete, analyzed at different scales. The Spartan variogram family is herein applied for the first time to precipitation data. We generate maps of precipitation and its uncertainty across the island using geostatistical methods.

1.3 Thesis Outline

The remainder of this thesis is organized as follows. In Section 1.4 background information and theory pertaining to meteorological parameters are presented as well as some of the most commonly used methodologies for the analysis of precipitation data. In Section 2 geostatistical methodologies are reviewed. Section 3 presents relevant information for the study area (Crete) and an exploratory statistical analysis of the available rainfall data set. In Section 4 we present the geostatistical methodology applied to the data set and the results obtained. Section 5 contains a general discussion of the results and concluding remarks. Finally, in Appendix A all the resulted plots for every year are gathered.

1.4 Basic Concepts and Literature Review

Climate in a narrow sense is usually defined as the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities over a period of time ranging from months to thousands or million of years. The classical period for averaging these variables is 30 years, as defined by the World Meteorological Organization. The relevant quantities are most often surface variables such as temperature, precipitation and wind [Hartmann et al., 2013].

Life on Earth is strongly linked to weather and climate; consequently knowing how to evaluate and interpret key factors for weather and climate is vital. Human societies have developed and prospered under relatively stable climate since the ice age which ended several thousand years ago. In the last century, however, climate trends have been observed that have caused international concern.

More specifically, according to the Fifth Assessment Report [Hartmann et al., 2013] the mean surface temperature seems to follow an increasing trend of 0.075°C per decade from 1901 to 2012 (90% confidence interval: 0.062 to 0.088°C) as shown in Fig. 1.1. The HadCRUT4 dataset that is used for the Fig. 1.1 has been developed by the Climatic Research Unit at the University of East Anglia in conjunction with the Met Office Hadley Centre [Jones et al., 2015]. HadCRUT4 is a gridded dataset of global historical surface temperature anomalies relative to

a 1961-1990 reference period. Data are available for each month since January 1850, on a 5 degree grid [Morice, 2015].

The observed trend on a shorter time scale is presented in Table 1.1. Climate change is caused by non-anthropogenic factors such as biotic processes, variations in the solar radiation received by the Earth, plate tectonics, and volcanic eruptions. However, most scientists believe that there is a definite anthropogenic component of climate change [IPC, 2013].

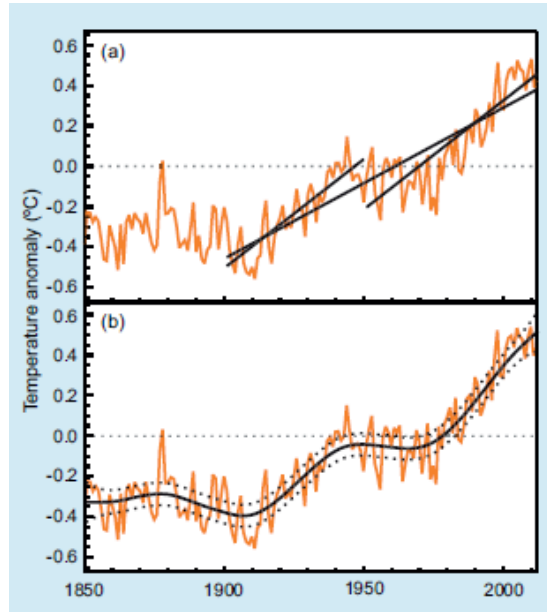


Figure 1.1: (a) Global mean surface temperature (GMST) anomalies relative to a 1961-1990 climatology based on HadCRUT4 annual data. The straight black lines are least squares trends for 1901-2012, 1901-1950 and 1951-2012. (b) Same data as in (a), with smoothing spline (solid curve) and the 90% confidence interval on the smooth curve (dashed lines). Note that the (strongly overlapping) 90% confidence intervals for the least squares lines in (a) are omitted for clarity. Figure taken from Hartmann et al. [2013].

The climate of Europe is temperate and is divided into five types: maritime, transitional, continental, polar and Mediterranean. The sensitivity of Europe to climate change has a distinct north-south gradient, with many studies indicating that Southern Europe will be more severely affected by climate change. Southern Europe has a hot semi-arid climate which is expected to become warmer and drier, threatening the continent's waterways, hydropower, agricultural production and

Table 1.1: Estimates of the mean change in global (spatial) mean surface temperature between 1901 and 2012, 1901 and 1950, and 1951 and 2012, obtained from the linear (least squares) and nonlinear (smoothing spline) trend models. Half-widths of the 90% confidence intervals are also provided for the estimated changes based on the two trend methods. Table taken from [Hartmann et al. \[2013\]](#).

Method	Trend in °C per decade			
	1901-2012	1901-1950	1951-2012	
Least squares	0.075± 0.013	0.107± 0.026	0.106± 0.027	
Smoothing spline	0.081± 0.010	0.070± 0.016	0.090± 0.018	

timber harvests.

Climate trends will lead to changes in temperature and precipitation, affecting multiple sectors of the society and the economy (Table 1.2). For example, as extreme weather events (meteorological droughts, heavy precipitation events, heat waves, floods and fires) become more frequent, their impacts on economic, social and health sectors will become more visible. In Southern Europe projections call for an increase in the frequency and the intensity of heat waves and a decrease in precipitation, contrary to the projected precipitation increase for Northern Europe [[Kovats et al., 2014](#)]. Nevertheless, the impact of this change over the Mediterranean region is not clear due to the prevalence of strong convective rainfall and its great spatial variability. Climate change could lead to increased evaporative losses, leading to significant reduction of water availability in rivers and groundwater resources [[Kovats et al., 2014](#)].

1.4.1 Precipitation and Drought

1.4.1.1 Precipitation

In meteorology, any product of the condensation of atmospheric water that falls under gravity to Earth is called precipitation. Precipitation can be divided into three categories, based on whether it falls as liquid water (rain, drizzle, sleet), liquid water that freezes on contact with the surface, or ice (snow, hail). Precipitation occurs when a portion of the atmosphere becomes saturated with water

Table 1.2: Impacts of climate extremes in the last decade in Europe. Table taken from Kovats et al. [2014].

Year	Region	Meteorological characteristics	Production systems and physical infrastructure, settlements	Agriculture, fisheries, forestry, bioenergy	Health and social welfare	Environmental quality and biological conservation	Mega-fire
2003	Western and central Europe	Hottest summer in at least 500 years [Luterber et al., 2004]	Damage to road and rail transport systems. Reduced/interrupted operation of nuclear power plants (mostly in France). High transport prices on the Rhine due to low water levels	Grain harvest losses of 20% [Ciais et al., 2005]	35 000 deaths in August in central and western Europe [Robine et al., 2008]	Decline in water quality [Daufresne et al., 2007]. High outdoor pollution	Yes
2004/2005	Iberian Peninsula	Hydrological drought		Grain harvest losses of 40			
2007	Southern Europe	Hottest summer on record in Greece since 1891 [Founda and Giannakopoulos, 2009]	1710 buildings burned down or rendered uninhabitable in Greece [JRC, 2008]	~ 575 500 hectares burnt area [JRC, 2008]	6 deaths Portugal, 80 deaths in Greece [JRC, 2008]	Several protected conservation sites (Natura, 2000) were destroyed [JRC, 2008]	Yes, Greece
2007	England and Wales	May–July wettest since records began in 1766	Estimated total losses 4 billion (£3 billion insured losses) [Chatterton et al., 2010]	78 farms flooded. Impacts on agriculture £50 million [Chatterton et al., 2010]	13 deaths and 48 000 flooded homes [Pitt, 2008]. Damage costs for health effects, including loss of access to education, £287 million [Chatterton et al., 2010]		
2010	Western Russia	Hottest summer since 1500 [Barriopedro et al., 2011]		Fire damage to forests [Shvidenko et al., 2011]. Reduction in crop yields [Barriopedro et al., 2011] [Coumou and Rahmstorf, 2012]	Estimated 10 000 excess deaths due to heat wave in Moscow in July and August [Revich and Shaposhnikov, 2012]	High outdoor pollution levels in Moscow [Bondur, 2011], [Revich and Shaposhnikov, 2012]	Yes
2011	France	Hottest and driest spring in France since 1880	Reduction in snow cover for skiing	8% decline in wheat yield [AGRESTE, 2011]			

vapour, so that the water condenses and “precipitates”.

Over 96 percent of the world’s total water supply is saline. Over 68 percent of the total freshwater is locked up in ice and glaciers, while another 30 percent of freshwater is in the ground. Fresh surface–water sources, such as rivers and lakes, only constitute about 93 100 km³, which is about 1/150th of one percent of the total water on the Earth; yet, rivers and lakes are the sources of most of the water people use everyday [Shiklomanov, 1993].

Precipitation is a major component of the water cycle, and it is responsible for depositing the fresh water on the planet (Fig. 1.3). Approximately 505 000 km³ of water falls as precipitation each year, 398 000 km³ of it over the oceans and 107 000 km³ over land [Chowdhury, 2005]. Given the Earth’s surface area, the above implies that the globally averaged annual precipitation is 990 mm, but only 715 mm of that precipitation falls over land. Climate classification systems such as the Köppen classification use average annual rainfall to help differentiate between climate regimes.

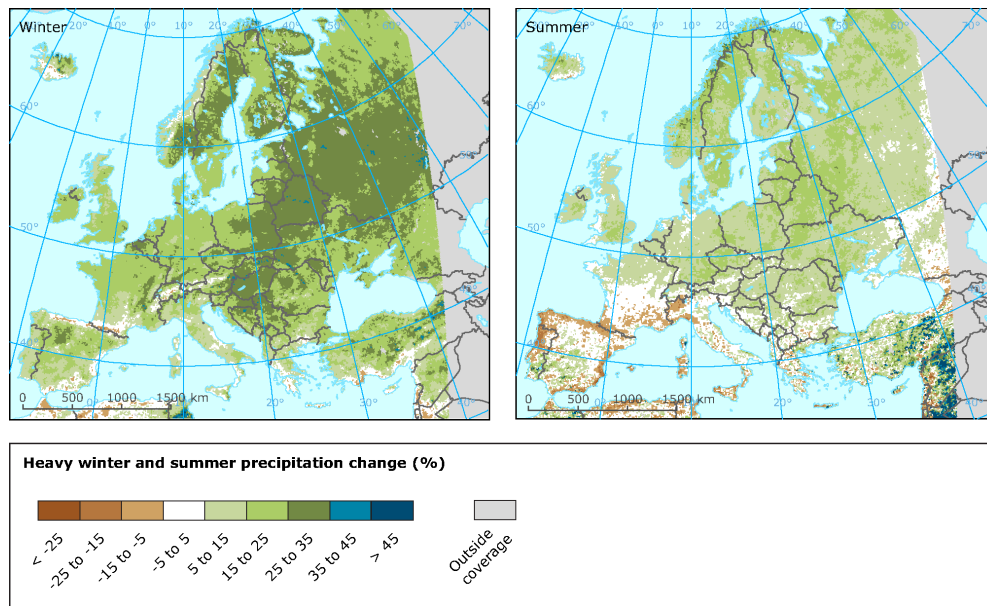


Figure 1.2: Projected changes in heavy precipitation (in %) in winter and summer from 1971–2000 to 2071–2100 for the RCP8.5 scenario based on the ensemble mean of different regional climate models (RCMs) nested in different general circulation models (GCMs). Figure taken from EEA [2015].

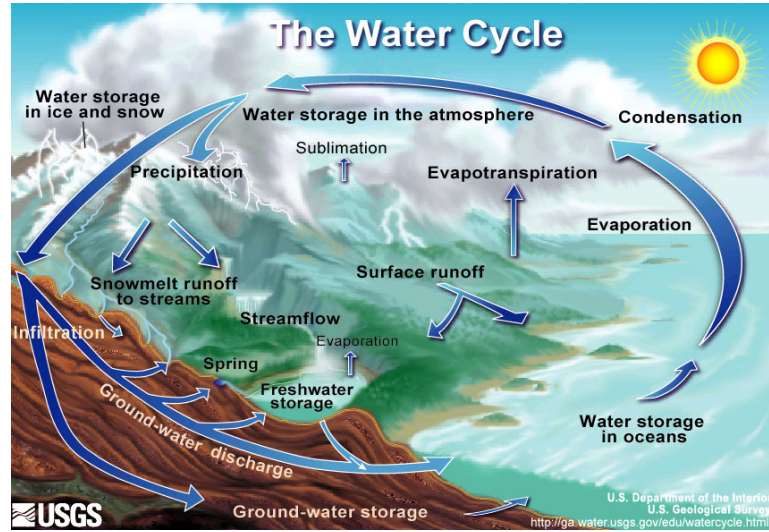


Figure 1.3: Water Cycle. Figure taken from Shiklomanov [1993].

The standard instrument for measuring rainfall or snowfall is the standard rain gauge, developed at the start of the 20th century. The gauge consists 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. If the rainwater overflows the graduated inner cylinder, the larger outer container will catch it. When measurements are taken, the height of the water in the small graduated cylinder is measured, and the excess overflow in the large container is carefully poured into another graduated cylinder and measured to give the total rainfall. Sometimes a cone meter is used to prevent leakage that can corrupt the data. The cylinder is usually marked in mm and will measure up to 250 mm of rainfall with 0.5 mm markings that define the resolution [Strangeways, 2006].

Other types of gauges include the wedge gauge (the cheapest and most fragile rain gauge), the tipping bucket rain gauge, and the weighing rain gauge. Contrary to in-situ methods, radars and satellites are used for measuring rainfall with remote sensing methods. In particular, precipitation measurements in vast expanses of the ocean and remote land areas depend on satellite observations. There is a relationship between the occurrence and the intensity of precipitation with the recorded electromagnetic spectrum. Satellite sensors that are used for precipitation, fall into two categories. Thermal infrared (IR) sensors record a

channel around 11 micron wavelength and primary give information about cloud tops. These information combined with mathematical schemes, algorithms and atmospheric properties, create the precipitation estimation from IR data. The IR sensors work best in cases of deep, vigorous convection —such as the tropics— and becomes progressively less useful in areas where stratiform precipitation dominates. The second category of sensor channels is in microwave part of the electromagnetic spectrum, with frequencies ranging between 10 GHz to a few hundred GHz. Satellites such as the Tropical Rainfall Measuring Mission (TRMM) and the Global Precipitation Measurements (GPM) mission employ microwave sensors to estimate precipitation.

1.4.1.2 Drought

The definition of drought is complex because different interpretations are possible. The majority of people may consider drought 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].

Drought can be divided into four categories:

1. **Meteorological or Climatological.** Meteorological drought is defined usually on the basis of the degree of dryness—in comparison to some “normal” or average amount—and the duration of the dry period. Definitions of meteorological drought must be considered as region specific since the atmospheric conditions that result in deficiencies of precipitation are highly variable from region to region [Wilhite and Glantz, 1985].
2. **Agricultural.** The onset of an agricultural drought may lag that of a meteorological drought, depending on the prior moisture status of the surface soil layers.
3. **Hydrological.** Precipitation deficits over a prolonged period that affect surface or subsurface water supply, thus reducing streamflow, groundwater,

reservoir, and lake levels, will result in a hydrological drought, which will persist long after a meteorological drought has ended.

4. **Socioeconomic.** Socioeconomic drought associates the supply and demand of certain economic goods with elements of meteorological, agricultural and hydrological drought.

The relationship between the different types of drought is complex. For example, streamflow is the key variable in the analysis of droughts for many water supply activities such as hydropower generation, recreation, and irrigated agriculture where crop growth and yield are largely dependent on water availability in the streams [Condra, 1944].

The hydrology of a region is affected by changes in the timing and amount of precipitation, evaporation, transpiration rates, and soil moisture, parameters which in turn also affect the drought characteristics of a region. The effect of a drought varies according to vulnerability. For example, subsistence farmers are more likely to migrate during droughts because they do not have alternative food sources. Areas with populations that depend on water resources to grow their food sources are more vulnerable to famine.

1.4.2 Drought Indices

Drought indices are quantitative measures that characterize drought levels by assimilating data from one or several variables (indicators) such as precipitation and evapotranspiration into a single numerical value. The nature of drought indices reflects different events and conditions; they can reflect climate dryness anomalies (mainly based on precipitation) or correspond to delayed agricultural and hydrological impacts such as soil moisture loss or lowered reservoir levels [Zargar et al., 2011].

Some of the drought indices incorporate a large amount of data on rainfall, streamflow, snow and other hydrometeorological indicators, and they transform these huge datasets into a single number. Since the development of a drought index can be based on multiple factors (e.g. drought's nature and characteristics and the impacts considered), multiple drought indices have been developed

(more than 150, according to Niemeyer [2008]). They include the Standardized Precipitation Index (SPI, McKee et al. [1993]), the Palmer Drought Severity Index (PDSI, Palmer [1965]), the Reconnaissance Drought Index (RDI, Tsakiris et al. [2007]), deciles (Gibbs and Maher [1967]), the Reclamation Drought Index (RDI, Weghorst [1996]), the Crop Moisture Index (CMI, Palmer [1968]), the Surface Water Supply Index (SWSI, Shafer and Dezman [1982]), and the Aggregate Drought Index (ADI, Keyantash and Dracup [2004]) [Vrochidou, 2013].

1.4.3 Climate Models

Numerical models of climate, i.e., General Circulation Models or GCMs, represent physical processes in the atmosphere, oceans, cryosphere and land surface, in terms of mathematical equations that are solved numerically. They are the most commonly used tools for simulating the climate [IPCC, 2015]. The models divide the earth, ocean and atmosphere into grid blocks. The examined variables, such as precipitation, are calculated at each grid point as a function of time. The time step is a function of the grid size: the finer the resolution, the shorter the interval between each computation [WMO, 2015]. Typically, a three-dimensional grid has horizontal resolution of between 250 and 600 km, 10 to 20 vertical layers in the atmosphere, and sometimes as many as 30 layers in the oceans (Fig. 1.4). Thus, the resolution of GCMs is quite coarse and many physical processes that occur at smaller scales, such as those related to clouds, cannot be properly modeled.

There are different types of GCMs based on the area studied, i.e., studies of the oceans can concentrate on three-dimensional properties of the oceans and are generally known as ocean GCMs. When it comes to simulating the general behaviour of the climate system over lengthy periods, however, it is essential to use models that represent, and where necessary conserve, the important properties of the atmosphere, land surface and the oceans in three dimensions. At the interfaces, the atmosphere is coupled to the land and oceans through exchanges of heat, moisture and momentum. These models of the climate system are usually known as coupled GCMs.

A technique known as parameterization refers to averaging known properties over the larger scale, which is one source of uncertainty in GCM-based simula-

tions of future climate. The use of all the available GCM experiments does not guarantee a representative range for the simulated climate variables, especially for estimates of future atmospheric composition, due to uncertainties such as the parameterization technique used (averaging a known property over the larger scale) [IPCC, 2015].

To overcome the coarse spatial resolution of a coupled GCM, the technique of nested modeling is used (Fig. 1.4). This technique is applied after the general analysis using the GCM output as initial and boundary conditions for appropriate Regional Climate Models (RCM). This enables the RCM to enhance the detailed regional model climatology. This downscaling procedure can be extended to finer grids in terms of local models [WMO, 2015].

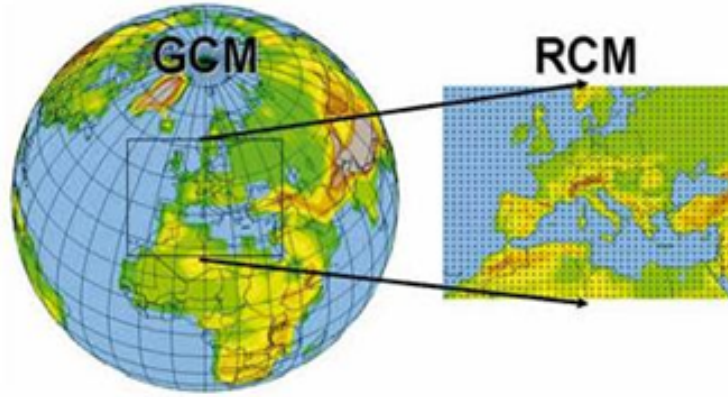


Figure 1.4: Schematic illustrating the relation between General Circulation Models and Regional Climate Models. Figure taken from WMO [2015].

The first GCM was created by Phillips [1956] at Princeton and was quickly hailed as “a classic experiment”. Later, Neumann and Charney developed a general circulation model of the entire three-dimensional global atmosphere built directly from the primitive equations. Many researchers responded to the challenge of creating new GCMs. Dickinson et al. [1989] conducted an innovative study in which they introduced the RCM, which plays a key role in the scientific investigations during the last three decades. Recent research on the connections between GCMs and RCMs include the following papers: Christensen et al. [2007]; Giorgi and Mearns [1999]; Grotch and MacCracken [1991]; Hulme [1992]; McCabe

and Dettinger [1995]; Wilby et al. [1998] and Fowler et al. [2007].

1.4.4 Statistical Modeling

In addition to climate models and drought indices, that have been used widely to characterize a region, another approach to evaluate the climatic conditions observed in an area is the geostatistical analysis. In this approach, mathematical functions are used to model a variable of interest, e.g., precipitation, and to determine potential correlations through space and time.

Such studies have been presented by Goovaerts [2000] who has used three multivariate geostatistical algorithms that incorporate a digital elevation model into the spatial prediction of rainfall: simple kriging with varying local means, kriging with an external drift, and colocated cokriging. Another study that uses geostatistical tools is provided by Tushaus [2014] and focuses in the topographical parameters that affect mountain rain. Moral [2010] gives a comparison of different geostatistical approaches for mapping climate variables incorporating geographical information systems (GIS) and carrying out cross validation to obtain the prediction errors created by the algorithms. Lately, Baxevas and Lennatsson [2015] developed a daily stochastic spatio-temporal precipitation generator based on a censored latent Gaussian field. The distributions used for the description of precipitation intensity are a gamma distribution for observations below some threshold and a generalized Pareto distribution to model the excesses above the threshold.

1.4.5 Probabilistic Modeling of Precipitation

Precipitation is best modelled as a stochastic variable due to its significant spatial and temporal variability. There are certain typical features of these data that must be taken into account to produce useful results, including the non-Gaussian distribution of precipitation values, with the presence of many zero values, the low density of observations, and the temporal and spatial variability of spatial correlation patterns.

Various probability distributions have been used to model precipitation. The most commonly used probability distribution for precipitation data analysis based

on recorded measurements is the gamma distribution [Baxevasi and Lennatsson, 2015; McKee et al., 1993; Vrochidou, 2013]. We present plots of several relevant probability distribution functions including the normal (Gaussian), the gamma, the generalized extreme value and Student's t distribution.

1.4.5.1 The Normal Distribution

The normal distribution also known as the Gauss distribution, is the most important distribution in statistics. It is often encountered in natural phenomena as a result of the Central Limit Theorem. The latter —loosely expressed— states that the normal distribution is an attractor for averages of identically distributed random variables so long as their probability distributions do not have heavy tails. Its crucial property is that any affine combination of independent normal random variables is also normal [Kroese et al., 2011]. An affine combination of vectors x_1, \dots, x_n is a vector $\sum_{i=1}^n a_i x_i = a_1 x_1 + a_2 x_2 + \dots + a_n x_n$, called a linear combination of x_1, \dots, x_n , in which the sum of the coefficients is equal to 1, thus, $\sum_{i=1}^n a_i = 1$. The probability density function of a normal distribution is given as follows:

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}, \quad (1.1)$$

where μ is a location parameter, equal to the mean, and σ is the standard deviation. For $\mu = 0$ and $\sigma = 1$ we refer to this distribution as the standard normal distribution. In many connections it is sufficient to use this simpler form since μ and σ simply may be regarded as a shift and scale parameter, respectively [Walck, 1996]. In Fig. 1.5 different types of the normal distribution are presented. In precipitation data analysis, the Gaussian distribution is used mostly after the transformation of the data sets into a normalized index, i.e., the transformation of precipitation data into the Standardized precipitation index, resulting into a new data set of normalized values [Karavitis et al., 2011; McKee et al., 1993]. In this particular study, normal probability distribution is used to fit the precipitation residuals, that are used for the application of the residual kriging.

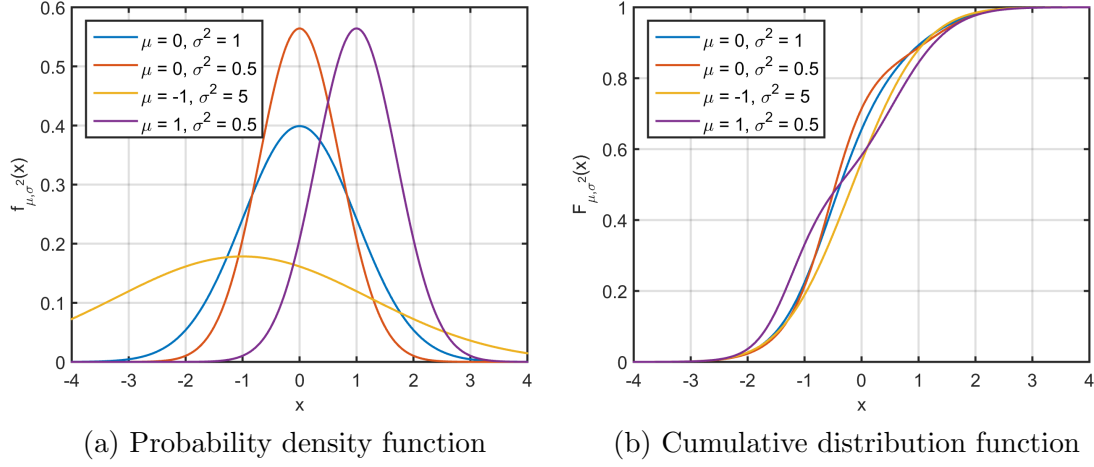


Figure 1.5: Plots of the probability density function and the cumulative distribution function for the Normal or Gaussian Distribution.

Location parameters

The normal distribution $f(x)$, with any mean μ and any positive deviation σ , is symmetric around the point $x = \mu$, which is at the same time the mode, the median and the mean of the distribution. The following results hold for $N(\mu, \sigma^2)$ variables

$$\text{Mean} = \text{Mode} = \mu = \mu_1, \quad \text{and} \quad \text{Variance} = \sigma^2 = \mu_2, \quad (1.2)$$

$$\mu_3 = 0, \quad \text{and} \quad \mu_4 = 3\sigma^4, \quad (1.3)$$

where $\mu_1, \mu_2, \mu_3, \mu_4$ are the first, the second, the third and the fourth moments [Jagdish and Campbell, 1982](see section 2.7 on page 32).

Skewness

The skewness is defined in terms of the centered third-order moment of the distribution. The skewness is a measure of the symmetry or asymmetry of the probability density function. The skewness of the normal distribution and all other symmetric distributions is equal to zero. Any data with a symmetric empirical distribution should also have a skewness near zero.

The skewness of a distribution is defined as

$$s = \frac{\mu_3^2}{\mu_2^3} = 0. \quad (1.4)$$

Kurtosis

The kurtosis is defined in terms of the centered fourth-order moment of the distribution. The kurtosis is a measure of whether the data are more or less concentrated near the peak relative to a normal distribution. The kurtosis for the standard normal distribution is equal to three, calculated by

$$\text{kurtosis} = \frac{\mu_4}{\mu_2^2} = 3. \quad (1.5)$$

1.4.5.2 The Gamma Distribution

The gamma distribution is a two-parameter family of continuous probability distributions. Gamma distribution is the most commonly used distribution in the analysis of precipitation data, since it is used to fit precipitation for the SPI development [McKee et al., 1993; Vrochidou, 2013] but it is also used to fit precipitation using geostatistical tools such as the study conducted by Baxevani and Lennatsson [2015]. The parameters used in this case include the shape parameter a and the scale parameter b . Both parameters are positive real numbers.

If a is an integer, the distribution represents an Erlang distribution; i.e., the sum of a independent exponentially distributed random variables, each of which has a mean equal to b [Devroye, 1986; Kroese et al., 2011].

A random variable X that is gamma-distributed with shape a and scale b is denoted as:

$$X \sim \Gamma(a, b) \equiv \text{Gamma}(a, b). \quad (1.6)$$

The probability density function of the gamma distribution using the shape-scale parametrization is:

$$f(x; a, b) = \frac{x^{a-1} e^{-\frac{x}{b}}}{b^a \Gamma(a)} \quad \text{for } x > 0 \text{ and } a, b > 0. \quad (1.7)$$

Here $\Gamma(a)$ is the gamma function evaluated at a . The gamma function is defined

by means of the following integral

$$\Gamma(a) = \int_0^{\infty} e^{-x} x^{a-1} dx, \quad a > 0. \quad (1.8)$$

The cumulative distribution function is the regularized gamma function

$$F(x; a, b) = \int_0^x f(u; a, b) du = \frac{\gamma(a, \frac{x}{b})}{\Gamma(a)}, \quad (1.9)$$

where $\gamma(a, x/b)$ is the lower incomplete gamma function [Kroese et al., 2011], defined by means of the following integral

$$\gamma(a, x) = \frac{1}{\Gamma(a)} \int_0^x e^{-t} t^{a-1} dt, \quad a > 0, x > 0. \quad (1.10)$$

For positive integer values of $a = n$, we have

$$\gamma(n, x) = 1 - e^{-x} \sum_{k=0}^{n-1} \frac{x^k}{k!}. \quad (1.11)$$

In Fig. 1.6, plots of the gamma distributions are presented with different parametrization. [McKee et al., 1993]

Location parameters

The gamma distribution $f(x)$, with any mean μ , any shape parameter $\alpha > 0$ and any scale parameter $b > 0$ has mean value

$$\mu = ab, \quad (1.12)$$

and variance

$$\sigma^2 = ab^2. \quad (1.13)$$

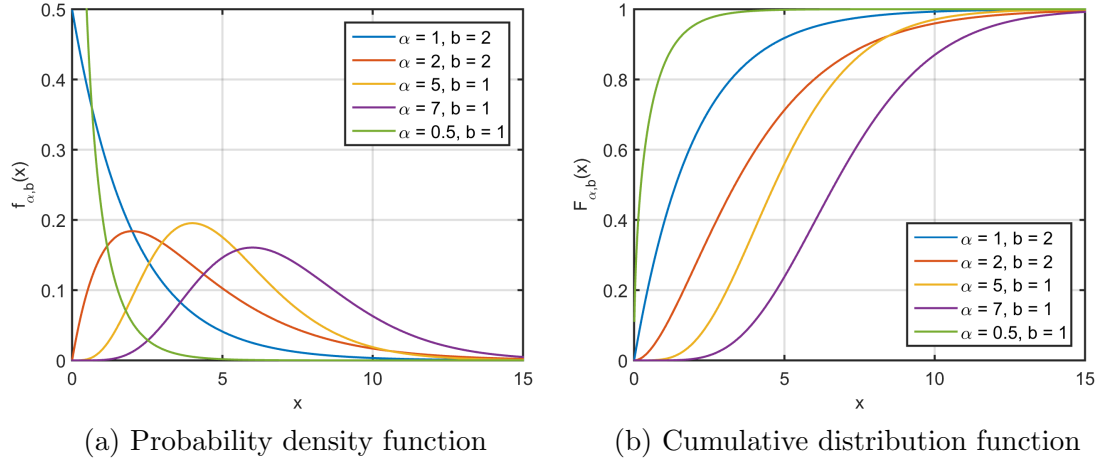


Figure 1.6: Plots of the probability density function and the cumulative distribution function for the Gamma Distribution.

Skewness

The skewness of the gamma distribution is equal to $2/\sqrt{a}$. The skewness depends only on the shape parameter a . The gamma distribution approaches a normal distribution when a is large (approximately when $a > 10$).

Median

Unlike the mode and the mean which have readily calculable formulas based on the parameters of the gamma distribution, the median does not obey a closed form equation. The median for this distribution is defined as the value ν such that

$$\frac{1}{\Gamma(a)b^a} \int_0^\nu x^{a-1} e^{-\frac{x}{b}} dx = \frac{1}{2}. \quad (1.14)$$

A formula for approximating the median for any gamma distribution, when the mean is known, has been derived based on the fact that the ratio $\mu/(\mu - \nu)$ is approximately a linear function of a when $a \geq 1$ [Devroye, 1986]. The approximation formula is:

$$\nu \approx \mu \frac{3a - 0.8}{3a + 0.2}. \quad (1.15)$$

Superposition

If X_i follows a $\text{Gamma}(a_i, b)$ distribution for $i = 1, 2, \dots, N$ (i.e., all distributions have the same scale parameter b), then

$$\sum_{i=1}^N X_i \sim \text{Gamma}\left(\sum_{i=1}^N a_i, b\right), \quad (1.16)$$

where \sim denotes equality in distribution, provided all X_i are independent.

1.4.5.3 Generalized Extreme Value Distribution

The generalized extreme value (GEV) distribution is a family of continuous probability distributions developed within extreme value theory to combine the Gumbel, Fréchet and Weibull families also known as type I, II and III extreme value distributions respectively. In hydrology, the Gumbel distribution is used to analyze variables such as monthly and annual maximum values of daily rainfall and river discharge volumes [Oosterbaan, 1994], and also to describe droughts [Burke et al., 2010]. Fréchet distribution is also applied to hydrology to extreme events such as annually maximum one-day rainfalls and river discharges [Vivekanandan, 2013]. The Weibull distribution has been used to describe wind speed distributions [Odo et al., 2012; Yanping et al., 2010]. In general, the GEV distribution has applications in meteorological analysis and widely to the economical domain [Ghosh, 2010; Gilli and Kellezi, 2006; Sharma and Singh, 2010]. This study uses the GEV distribution to fit precipitation data of a dry region, such as Crete, locally, at the position of every station over the recorded years.

By virtue of the extreme value theorem (also known as the FisherTippettGnedenko theorem) the GEV distribution is the limit distribution of properly normalized maxima of a sequence of independent and identically distributed random variables. More specifically, let X_1, X_2, \dots, X_n be a sequence of independent and identically distributed random variables, and $M_n = \max\{X_1, \dots, X_n\}$. If a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and $\lim_{n \rightarrow \infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, where F is a non degenerate distribution function, then the limit distribution F belongs to either the Gumbel, the Fréchet or the Weibull family [wikipedia, 2015b]. Because of this feature, the GEV distribu-

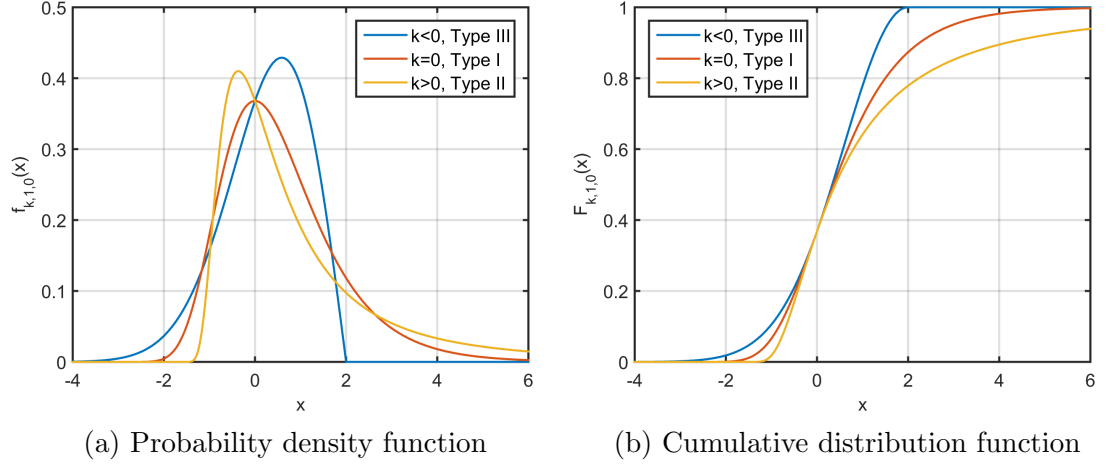


Figure 1.7: Plots of the probability density function and the cumulative distribution function for the Generalized Extreme Value Distribution.

tion is used as an approximation to model the maxima of long (finite) sequences of random variables [Coles, 2001]. In Fig. 1.7, generalized extreme value distributions with different parametrization are presented.

Distribution and Probability Density Functions

The generalized extreme value distribution has the following cumulative distribution function

$$F(x; \mu, \sigma, k) = \exp \left\{ - \left[1 + k \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/k} \right\}, \quad (1.17)$$

for $1 + k(x - \mu)/\sigma > 0$, where $\mu \in \mathbb{R}$ is the location parameter, $\sigma > 0$ the scale parameter, and $k \in \mathbb{R}$ the shape parameter. For $k = 0$ the expression is formally undefined and is understood as a limiting case.

The density function is, consequently

$$f(x; \mu, \sigma, k) = \frac{1}{\sigma} \left[1 + k \left(\frac{x - \mu}{\sigma} \right) \right]^{(-1/k)-1} \exp \left\{ - \left[1 + k \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/k} \right\}, \quad (1.18)$$

again, for $1 + k(x - \mu)/\sigma > 0$.

Statistics

Some simple statistics of the GEV distribution are:

$$E(X) = \mu - \frac{\sigma}{k} + \frac{\sigma}{k} g_1, \quad (1.19)$$

$$Var(X) = \frac{\sigma^2}{k^2} (g_2 - g_1^2), \quad (1.20)$$

$$Mode(X) = \mu + \frac{\sigma}{k} [(1+k)^{-k} - 1]. \quad (1.21)$$

Skewness

$$\text{skewness}(X) = \frac{g_3 - 3g_1g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}} \times \text{sign}(k). \quad (1.22)$$

Kurtosis

$$\text{kurtosis}(X) = \frac{g_4 - 4g_1g_3 + 6g_2g_1^2 - 3g_1^4}{(g_2 - g_1^2)^2} - 3, \quad (1.23)$$

where $g_i = \Gamma(1 - ik)$, $i = 1, 2, 3, 4$ and $\Gamma(t)$ is the gamma function.

Sub-families

The shape parameter k governs the tail behaviour of the distribution. The sub-families defined by $k = 0$, $k > 0$ and $k < 0$ correspond, respectively, to the Gumbel, Fréchet and Weibull families, whose cumulative distribution functions are displayed below [[Kroese et al., 2011](#)].

- Gumbel or type I extreme value distribution ($k = 0$)

$$F(x; \mu, \sigma, 0) = e^{-e^{(x-\mu)/\sigma}} \quad \forall x \in \mathbb{R}. \quad (1.24)$$

- Fréchet or type II extreme value distribution, if $k = \alpha^{-1} > 0$

$$F(x; \mu, \sigma, k) = \begin{cases} 0 & x \leq \mu, \\ e^{-((x-\mu)/\sigma)^{-\alpha}} & x > \mu. \end{cases} \quad (1.25)$$

- Reversed Weibull or type III extreme value distribution, if $k = -\alpha^{-1} < 0$,

$$F(x; \mu, \sigma, k) = \begin{cases} e^{-(-(x-\mu)/\sigma)^\alpha} & x < \mu, \\ 1 & x \geq \mu, \end{cases} \quad (1.26)$$

where $\sigma > 0$.

The equations above relate to maxima, and the distribution being discussed is an extreme value distribution for maxima. A generalised extreme value distribution for minima can be obtained, for example by substituting $(-x)$ for x in the distribution function, and subtracting from one: this yields a separate family of distributions [Coles, 2001].

The ordinary Weibull distribution arises in reliability applications and is obtained from Reversed Weibull distribution using the variable transformation $t = \mu - x$, which gives a strictly positive support. In term of an equation this means

$$F(t; \sigma, k) = \begin{cases} \frac{k}{\sigma} (t/\sigma)^{k-1} e^{-(t/\sigma)^k} & t \geq 0, \\ 0 & t < 0. \end{cases} \quad (1.27)$$

This arises because the Weibull distribution is used to characterize minima rather than maxima. Importantly, in applications of the GEV, the upper bound is unknown and must be estimated from the data, while in applications of the Weibull distribution the lower bound is known to be zero. Note the differences in the ranges of interest for the three extreme value distributions: Gumbel is unlimited, Fréchet has a lower limit, while the reversed Weibull has an upper limit.

One can link the type I to types II and III in the following way: if the cumulative distribution function of some random variable X is of type II and has a positive support, i.e. $F(x; 0, \sigma, \alpha)$, then the cumulative distribution function of $\ln X$ is of type I, namely $F(x; \ln \sigma, 1/\alpha, 0)$. Similarly, if the cumulative distribution function of X is of type III, and with negative support, i.e. $F(x; 0, \sigma, -\alpha)$, then the cumulative distribution function of $\ln(-X)$ is of type I, namely $F(x; -\ln \sigma, 1/\alpha, 0)$ [Coles, 2001].

1.4.5.4 Student's t Distribution

The probability density function of Student's t distribution, or simply the t distribution, is given by:

$$f(x; \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad (1.28)$$

where $\nu > 0$. If the parameter ν takes integer values, then it is referred to as the *degrees of freedom* of the t distribution. The distribution arises in statistics in the estimation of the mean of a normally distributed population when the population variance is unknown (and thus estimated from the data) and the sample size is small. Whereas the normal distribution describes the full population, t-distributions describe samples drawn from the population; accordingly, the t-distribution for each sample size is different; larger samples converge to the normal distribution [Devroye, 1986].

The cumulative Student distribution function is expressed as follows:

$$F(t; \nu) = \int_{-\infty}^t \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi} \Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} dt. \quad (1.29)$$

In Fig. 1.8, the probability density function and the cumulative distribution function of Student's t distribution with different degrees of freedom are presented.

Statistics

The Student's t distribution $f(x)$, with ν degrees of freedom, has mean value

$$\mu = \begin{cases} 0 & \forall \nu > 1, \\ \text{undefined} & \text{otherwise,} \end{cases} \quad (1.30)$$

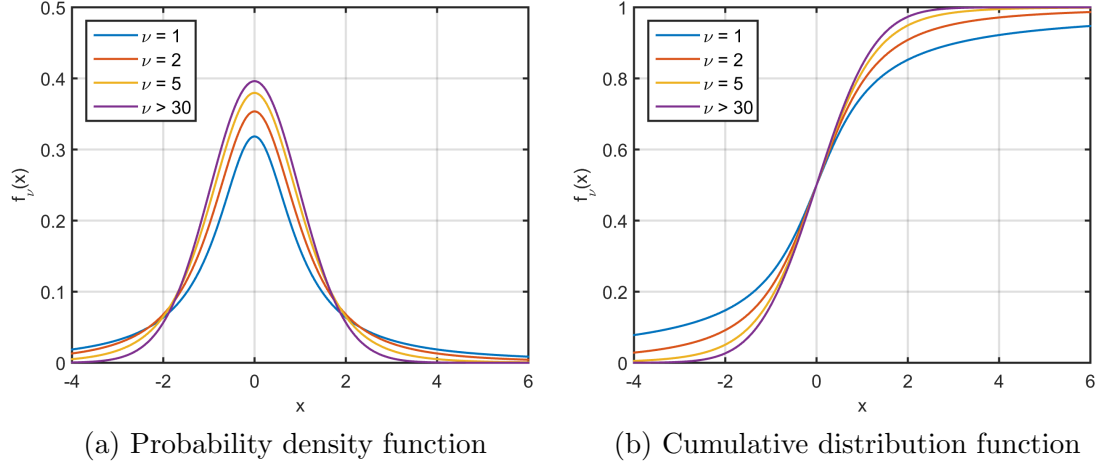


Figure 1.8: Plots of the probability density function and the cumulative distribution function for the Student's t Distribution.

variance

$$\sigma^2 = \begin{cases} \frac{\nu}{\nu-2} & \forall \nu > 2, \\ \infty & \text{for } 1 < \nu \leq 2, \\ \text{undefined} & \text{otherwise,} \end{cases} \quad (1.31)$$

skewness

$$\mu = \begin{cases} 0 & \forall \nu > 3, \\ \text{undefined} & \text{otherwise.} \end{cases} \quad (1.32)$$

1.4.6 The Role of Spatial and Temporal Scales

One of the most important issues in the analysis of precipitation data is the time scale of the observations, i.e., whether they correspond to days, months, seasons, wet and dry periods or years. Annual means of precipitation correspond to the sum of precipitation over 365 days of each year, wet periods correspond to October till March and dry periods to April till September. Seasonal time scale is established in practice of all the four seasons by summing the precipitation over the total days of each seasons. Changes in seasonal precipitation distribution and intensities would matter more for arable crops than changes in annual precipitation [Rotter and Van De Geijn, 1999]. Accordingly, monthly time scale refers to the summation of daily precipitation for each of the calendar months of a year

[Kumar et al., 2013].

As proven by McKee et al. [1993], at longer time scales drought becomes less frequent and of longer duration. In terms of precipitation this means that at longer time scales small amounts of precipitation become less frequent and of longer duration. Also correlation between precipitation and topography increases with the length of time interval. Finally, as proven by Bárdossy and Pergam [2013], interpolation quality depends on the aggregation time; longer aggregation times, reduce the relative error of the interpolation.

Chapter 2

Geostatistical Methods

Earth science data are distributed through space and/or time. The analysis and prediction of spatial phenomena, such as mineral grades, porosities and pollutant concentrations are performed using geostatistical methods [Chilès and Delfiner, 2012; Christakos, 1992; Goovaerts, 1997]. Geostatistics comprises a collection of numerical methods that can be used to model and characterize spatial attributes, based on the theory of random fields. Random fields are the proper mathematical framework for analyzing spatial data, in the same way that time series analysis is used for temporal data. Because geographical location varies continuously, the number of variables required to represent a spatial process is infinite even for areas of interest of finite size. Such a collection of random variates is called a random function or a random field [Olea, 1999].

The values of a variable that is distributed in space and/or time are often correlated with each other. The study of such correlations is usually conducted using a method known as “structural analysis” or “variogram modeling”. After structural analysis, predictions of the modeled process at unsampled locations are derived using geostatistical methods, such as kriging interpolation. The data processing steps involved in a geostatistical study involve: (i) exploratory data analysis (ii) structural analysis (iii) generating predictions and (iv) estimating the uncertainty.

2.1 Randomness

Randomness characterizes phenomena in which the knowledge of a situation with absolute precision is impossible due to various restrictions that are intrinsic, i.e. strong spatial and temporal variability of the observed process, or imposed by the experimental procedure, e.g. limited resolution, random errors, or caused by fluctuations of environmental factors, e.g. temperature and humidity. In these cases, the measured properties at different points in space are defined by means of respective probability distribution functions, which determine the probabilities of occurrence for each possible result [Papoulis and Pillai, 2002].

2.2 Random Fields

Ω denotes a probability space, \mathcal{F} is the Borel's σ -algebra on Ω , which is a collection of subsets of Ω that contains the null set and is closed under complementation and countable unions, and P is the probability measure. Let (Ω, \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 $\mathbf{s} \in \mathbb{D}$, $X(\mathbf{s}; \omega) \rightarrow X(\omega)$ 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]. A random variable could be the gain in a game of chance, the voltage of a random source, the cost of a random component, pollutant concentrations, wind direction or any other numerical quantity that is of interest. 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 is said to be a discrete random variable if its sample space is discrete: $x(\mathbf{s}) = 0, 1, 2, \dots$, for example. This class includes the indicator random variable which has a binary sample space, meaning there are only two possible outcomes: $x(\mathbf{s}) = 0, 1$. Most of the random variables used to represent environmental variables are continuous random variables: they have a sample space, $x(\mathbf{s})$, that is continuous. For example, precipitation, sea-levels, wind speeds, and breaking strengths all take values on continuous scales. 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].

Mathematical properties characterizing the random variables are extended for the case of variables distributed in space. A random field consists of a set of random variables that describe the spatial (or space–time) changes of an attribute. Therefore, a random field may be viewed as a multidimensional random variable. Because of the interdependence of physical quantities at different locations in space, random fields have unique mathematical properties that distinguish them from a set of independent random variables. 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 [Hristopulos, 2012].

Random fields are divided into the following categories based on their values and their support:

1. If the field takes values from a finite set of numbers (e.g., from the set of integers), it is called a *field of discrete values*.
2. If the field values are drawn from a continuum of real numbers, then it is called a *field of continuous values*.
3. If the locations where the field is defined coincide with the positions of a grid, we refer to a *lattice field*.
4. If the field extends over a continuous space, we have a *continuum field*.

Random variables and random fields are described by an ensemble of states. Each state (realization) is also a sample of the field with a corresponding probability determined by the multidimensional probability density function of the field.

In this case study, the variable of interest is precipitation. We assume that precipitation can be modeled as a random field defined on a continuum space. The interpolation is performed over a rectangular grid leading to a discretised lattice field.

2.3 Basic Concepts of Random Fields

Random fields describe different types of spatial dependence with different properties. The probability distribution of a random field and its parameters are defined as extensions of those applicable to independent random variables. Henceforth, the following symbols are going to be used:

- In capital letters, $X(\mathbf{s})$ denotes the random field.
- In lowercase letters, $x(\mathbf{s})$, denotes the values corresponding to a unique realization.
- The fluctuation of the field is denoted with a prime next to the symbol of the field. For example $X'(\mathbf{s})$ indicates the fluctuation of the field $X(\mathbf{s})$.

The integral limits depend on the space where the field X is defined. If the field takes positive and negative values, the integral ranges from $-\infty$ to ∞ . If the field takes only positive values, the integral ranges from 0 to ∞ . If it is known that the field values are limited to a specific interval $[a,b]$, then the integral is calculated over this interval.

2.4 Probability Density Function

The probability density function (pdf) of the field is denoted with the symbol $f_x[x(\mathbf{s})]$. For the pdf the subscript is the symbol indicating the field, while the argument of the function are the values of the state $x(\mathbf{s})$.

In the case of a single random variable, the pdf $f_x(x)$ is a function that refers to a single point. In contrast, in the case of a random field, $f_x[x(\mathbf{s})]$ contains the values for the entire area where the field is defined. 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 one-dimensional (scalar) pdf of the field at the point \mathbf{s}_1 is defined as $f_x(x_1, \mathbf{s}_1)$ and describes the possible states of the field at that point. In the case

of non-homogeneity it is possible for the one-dimensional pdf to vary from point to point.

Accordingly, the two-dimensional pdf of the field is defined as $f_X(x_1, x_2; \mathbf{s}_1, \mathbf{s}_2)$, and expresses the interdependence of possible states at two locations. The multi-dimensional pdf, $f_X(x_1, \dots, x_N; \mathbf{s}_1, \dots, \mathbf{s}_N)$, which describes the interdependence of possible states for a set of N points, is similarly defined [Isaaks and Srivastava, 1989].

The pdf is the first derivative of the cumulative density function. It is normalized so that the total probability of all possible outcomes is equal to 1, i.e.,

$$\int_{-\infty}^{\infty} f_X(x) dx = 1. \quad (2.1)$$

2.5 Statistical Homogeneity

The concept of statistical homogeneity extends the classical definition of homogeneity (a property is homogeneous if the corresponding variable is constant in space). Thus, a random field $X(\mathbf{s})$ is statistically homogeneous if the following conditions are fulfilled:

1. The mean value (expectation) is constant, meaning $m_X(\mathbf{s}) = m_X$.
2. The covariance function is defined and depends only on the distance vector $\mathbf{r} = \mathbf{s}_1 - \mathbf{s}_2$ between two points, meaning $c_X(\mathbf{s}_1, \mathbf{s}_2) = c_X(\mathbf{r})$.
3. The second condition implies that the variance of a statistically homogeneous field is constant.

These conditions define statistical homogeneity in a weak sense. 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.

Accordingly, the concept of statistical homogeneity implies that the statistical properties of the field do not depend on the spatial coordinates of the center of

mass of the N points. Practically, statistical homogeneity implies that there are no spatial trends, so that the spatial variability of the field can be attributed to fluctuations around a constant level equal to the mean value [Hristopulos, 2012].

2.6 Statistical Isotropy

Statistical isotropy is a property that assumes and further constrains statistical homogeneity. A field is statistically isotropic if it is statistically homogeneous and its covariance function depends only on the distance r , but not on the direction of the distance vector \mathbf{r} . This is important from a practical point of view, because it facilitates the identification of the spatial dependence by means of the omnidirectional variogram. Therefore, if the covariance function is isotropic, the field is by definition statistically homogeneous, but not vice versa [Olea, 1999].

2.7 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. The expectation of a quantity $A(X)$ which depends on the field is denoted by $\mathbb{E}[A(X)]$. For example, the general form of a multidimensional moment $\mathbb{E}[X^{k_1}(\mathbf{s}_1) \dots X^{k_N}(\mathbf{s}_N)]$, where $k_1 + \dots + k_N = K$, is given by the following K -dimensional integral

$$\mathbb{E}[X^{k_1}(\mathbf{s}_1) \dots X^{k_N}(\mathbf{s}_N)] = \int dx_1 \dots \int dx_N f_x(x_1, \dots, x_N; \mathbf{s}_1, \dots, \mathbf{s}_N) x_1^{k_1} \dots x_N^{k_N}. \quad (2.2)$$

The limit of integration in the above integral depend on the spatial support of the joint pdf. In general, we can extend the limits of integration from minus infinity to infinity taking into account that the pdf vanishes outside the support.

In practice most commonly used moments are low order moments such as mean value, variance, covariance function and semivariogram [Cressie, 1993].

2.7.1 Expectation

An important feature of a distribution is its central value (mean value). Roughly, this is similar to the arithmetic average of the data values in a sample. The mean value of a random field is defined by

$$m_x(\mathbf{s}) = \mathbb{E}[X(\mathbf{s})]. \quad (2.3)$$

$\mathbb{E}[X(\mathbf{s})]$ denotes the expectation, calculated over the ensemble of all states of the field, i.e.

$$\mathbb{E}[X(\mathbf{s})] = \int dx f_x(x; \mathbf{s}) x, \quad (2.4)$$

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

In Eq. (2.4) it can be noted that the mean value may depend on the position, \mathbf{s} , which comes from a possible position dependence of the one-dimensional pdf. Since the pdf is not always known in advance, the expectation is estimated from the sample using statistical methods [Cressie, 1993].

For highly asymmetric distributions, a more appropriate central value is the median, M_x , which is the value corresponding to a cumulative frequency of 0.5. Hence, the median splits the distribution into two equal halves [Goovaerts, 1997]. The median, M_x of the sample is the midpoint of the observed values if they are arranged in increasing order and is calculated by means of [Isaaks and Srivastava, 1989]

$$M_x = \begin{cases} x \left(r_{\frac{N+1}{2}} \right) & \text{if } N \text{ is odd,} \\ \frac{1}{2} \left(x \left(r_{\frac{N}{2}} \right) + x \left(r_{\frac{N}{2}+1} \right) \right) & \text{if } N \text{ is even,} \end{cases} \quad (2.5)$$

where r refers to rank of the values $x(\mathbf{s})$ arranged in ascending order.

2.7.2 Variance

The variance is a measure of the distribution's dispersion around the mean value. It is given by the mean value of the squared fluctuation according to the following equation:

$$\sigma_x^2(\mathbf{s}) \equiv \mathbb{E} [\{X(\mathbf{s}) - m_x(\mathbf{s})\}^2] = \mathbb{E} [X'^2(\mathbf{s})]. \quad (2.6)$$

Since the variance involves squared differences, it is sensitive to erratic high values. It is possible for the variance to vary from point to point. If the field is statistically homogeneous, however, the variance is the same at every point [Hristopulos, 2012].

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 [Goovaerts, 1997].

2.7.3 Skewness

In probability theory and statistics, the skewness of a probability distribution is a measure of the distribution's asymmetry about its mean. The skewness can be positive or negative, or even undefined. The skewness is usually defined as

$$\text{coefficient of skewness} = \frac{\mathbb{E} \left[X'^3(\mathbf{s}) \right]}{\sigma_x^3}. \quad (2.7)$$

The qualitative interpretation of the skewness is complicated. For a unimodal distribution, negative skewness indicates that the tail on the left side of the probability density function is longer or fatter than the right side —skewness does not distinguish between these two possibilities. Conversely, positive skewness indicates that the tail of the right side is longer or fatter than the left side. In cases where one tail is long but the other tail is fat, skewness does not obey a simple rule. For example, a zero value indicates that the tails on both sides of the mean balance out. This occurs both for a symmetric distribution and for asymmetric distributions the asymmetries of which even out, such as one tail being long but thin and the other being short but fat. In case of a unimodal symmetric distribution, the mean is equal to the median and the mode. Further, in multimodal and discrete distributions, skewness is also difficult to interpret. Importantly, the skewness does not determine the relationship of the mean and the median [Dean and Illowsky, 2012].

2.7.4 Kurtosis

One common measure that describes the shape of a probability distribution is kurtosis. Kurtosis measures the heaviness of the tails of distributions, originated with Pearson [Shohat, 1929] and is based on a scaled version of the fourth moment of the data or population. The coefficient of kurtosis is usually defined as

$$k_x = \frac{\mathbb{E} \left[X'^4(\mathbf{s}) \right]}{\sigma_x^4}. \quad (2.8)$$

There are various interpretations of kurtosis based on peakedness (width of peak), tail weight, and lack of shoulders (distribution primarily peak and tails, not in between). However, heavy tails have much more influence on kurtosis than the shape of the distribution near the mean [Ali, 1974; Kaplansky, 1945]. Also it has been argued that kurtosis really measures heavy tails, and not peakedness [Balanda and Macgillivray, 1988]. For this measure, higher kurtosis means more of the variance is the result of infrequent extreme deviations, as opposed to frequent modestly sized deviations. It is common practice to use an adjusted version of Pearson's kurtosis, the excess kurtosis, to provide a comparison of the shape of a given distribution to that of the normal distribution. The coefficient of excess kurtosis is defined based on the kurtosis of a univariate normal distribution which is equal to 3. Thus, the excess kurtosis is calculated by the equation

$$\text{excess kurtosis} = \frac{\mathbb{E} \left[X'^4(\mathbf{s}) \right]}{\sigma_x^4} - 3 = k_x - 3. \quad (2.9)$$

Sample skewness and kurtosis are rather unreliable estimators of the corresponding population parameters if the sample is small; they become better estimators as the sample size increases. However, large values of skewness or kurtosis may merit attention even in small samples, because such values indicate that statistical methods based on normality assumptions may be inappropriate [Williams, 2000].

2.8 Covariance Function

Another moment which characterizes a random field is the centered covariance function (CCF), also known as covariance function for short, is defined by the following equation:

$$c_X(\mathbf{s}_1, \mathbf{s}_2) \equiv \mathbb{E}[X(\mathbf{s}_1)X(\mathbf{s}_2)] - \mathbb{E}[X(\mathbf{s}_1)]\mathbb{E}[X(\mathbf{s}_2)]. \quad (2.10)$$

Equation (2.10) is equivalent to the following

$$c_X(\mathbf{s}_1, \mathbf{s}_2) \equiv \mathbb{E}[\{X(\mathbf{s}_1) - m_X(\mathbf{s}_1)\}\{X(\mathbf{s}_2) - m_X(\mathbf{s}_2)\}]. \quad (2.11)$$

The random field $X'(\mathbf{s}_1) \equiv X(\mathbf{s}_1) - m_X(\mathbf{s}_1)$ corresponds to the fluctuation of the field $X(\mathbf{s}_1)$ around the mean value at point \mathbf{s}_1 . The mean value of the fluctuation field equals to zero, meaning

$$\mathbb{E}[X'(\mathbf{s}_1)] = 0. \quad (2.12)$$

Based on the previous equations it holds that

$$c_X(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{E}[X'(\mathbf{s}_1)X'(\mathbf{s}_2)]. \quad (2.13)$$

Consequently, the CCF represents quantitatively the dependence of the fluctuations between two different points. [Hristopulos, 2012].

In statistically homogeneous and isotropic fields the two most important parameters of the covariance are (i) the variance $\sigma_X^2 = c_X(0)$ and (ii) the correlation length ξ . The variance is a measure of the magnitude of the field fluctuations. The correlation length is the parameter which normalizes the distance (in covariance function, distance is shown as the ratio r/ξ). The correlation length defines the distance over which the field values are statistically correlated. In case of anisotropic dependence, different correlation lengths are observed along the principal directions of anisotropy [Varouchakis, 2012].

If the arguments of the covariance function coincide, its value becomes equal

to the variance of the field at that point, that is

$$c_x(\mathbf{s}_1, \mathbf{s}_1) = \sigma_x^2(\mathbf{s}_1). \quad (2.14)$$

Spectral density

Not every function is acceptable as a covariance function. Permissibility conditions for the covariance function are defined by Bochner's theorem. This is expressed by means of the spectral density, which is given by the Fourier transformation of the covariance function [Bochner et al., 1959]. The Fourier transform is defined by the following integral

$$\tilde{c}_x(\mathbf{k}) = \int d\mathbf{r} e^{-i \mathbf{k} \cdot \mathbf{r}} c_x(\mathbf{r}), \quad (2.15)$$

where \mathbf{r} is the distance vector between two points and \mathbf{k} is the vector of spatial frequency (wave-vector).

The inverse Fourier transform is given occurs by the following integral:

$$c_x(\mathbf{r}) = \frac{1}{(2\pi)^d} \int d\mathbf{k} e^{i \mathbf{k} \cdot \mathbf{r}} \tilde{c}_x(\mathbf{k}). \quad (2.16)$$

Bochner Theorem: A function $c_x(\mathbf{r})$ is a permissible covariance function if the following conditions apply:

1. The power spectral density $\tilde{c}_x(\mathbf{k})$ exists,
2. it is non-negative throughout the frequency domain, and
3. the integral of $\tilde{c}_x(\mathbf{k})$ over the entire frequency domain is bounded.

2.9 Variogram Function

The semivariogram (or more commonly known as variogram for simplicity) is a statistical moment that assesses the average decrease in similarity between two random variables as the distance between them increases. Stochastic interpolation

algorithms, such as kriging-based methods, require knowledge of the variogram or the covariance [Olea, 1999].

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

$$\gamma_x(\mathbf{s}, \mathbf{r}) = \frac{1}{2} \mathbb{E} \{ [X(\mathbf{s} + \mathbf{r}) - X(\mathbf{s})]^2 \}. \quad (2.17)$$

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

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

$$\gamma_x(\mathbf{r}) = \sigma_x^2 - c_x(\mathbf{r}). \quad (2.18)$$

Thus the variogram of an SRF is bounded by the variance, i.e., by $c_x(0)$. From the above equation it follows that (i) the variogram tends asymptotically to the variance and (ii) if the covariance is known, the variogram is also known. For statistically homogeneous fields, the variogram contains the same information as the covariance function [Chilès and Delfiner, 2012].

There are two reasons to favor the variogram over the covariance function for an SRF with unknown mean. First, it is a more general function than the covariance and second the variogram does not require knowledge of the mean to compute the covariance. The mean has to be estimated from the data; this introduces bias that cannot be corrected unless the covariance function, or at least the correlation function, is already known. The variogram is not affected by these problems since it automatically filters out the mean.

If the increment $\delta X(\mathbf{s}; \mathbf{r})$ is statistically homogeneous, the random field $X(\mathbf{s})$ is called homogeneous field with statistically homogeneous differences. In this case the variogram $\gamma_x(\mathbf{r})$ depends solely on the distance \mathbf{r} between the points. This is the result of statistical homogeneity of the field increment. If the field $X(\mathbf{s})$ is statistically homogeneous, the same applies to the increment $\delta X(\mathbf{s}; \mathbf{r})$. However the opposite is not necessarily true, i.e. if the field of increment is statistically homogeneous, the initial field $X(\mathbf{s})$ is not necessarily homogeneous [Hristopulos,

2012].

For a field $X(\mathbf{s})$ with homogeneous increments, the covariance function may depend on both the distance \mathbf{r} and the positions of the points. Instead, the variogram function depends only on the distance between the points.

The variogram parameters determine the spatial dependence of the field values at two neighboring points. From the definition of the variogram, using the mean squared differences, it is shown that the variogram is a non-negative function, i.e., $\gamma_x(\mathbf{r}) \geq 0$. However, the reverse is not always true, i.e. a non-negative function is not necessarily an admissible variogram [Hristopulos, 2012].

Statistically homogeneous field

For homogeneous fields with isotropic spatial dependence, the variogram is determined by two parameters: the upper bound (sill) and the correlation length. The value of the variogram for long distances \mathbf{r} tends asymptotically to the upper bound (sill) which in theory is equal to the variance σ_x^2 of the random field. This property is based on the relation $\gamma_x(\mathbf{r}) = \sigma_x^2 - c_x(\mathbf{r})$ and the fact that at long distance the value of the covariance function tends towards zero. The variogram can increase indefinitely if the variability of the process does not approach a limit at long distances. In this case, however, the random field is not statistically homogeneous.

The correlation length determines the "speed" with which the variogram approaches the sill and also the range within which two points are correlated. It gives a precise meaning to the conventional notion of the area of influence for the sample.

Variogram behavior near the origin has two typical behaviors:

1. A discontinuity at 0 known as the nugget effect. $\gamma_x(\mathbf{r})$ does not seem to tend to zero as $r \rightarrow 0$. This means that the random field is generally not continuous and thus very irregular. The term nugget effect is due to discontinuities of the variogram of mineral grades at the origin. By extension, the term "nugget effect" is applied to all discontinuities at the origin, even if their cause is different. In general, the nugget effect is due to [Chilès and Delfiner, 2012]:

- i. a microstructure or “noise”, namely a component of the phenomenon with a range shorter than the sampling support (true nugget effect),
 - ii. a structure with a range shorter than the smallest interpoint distance,
 - iii. measurement or positioning errors.
2. A flat curve which indicates pure nugget effect or white noise. There is no correlation between any two points, however close they may be. This is the extreme case of total absence of spatial structure.

If correlation characteristics vary in different directions in space, the dependence is anisotropic. There are two main types of anisotropy mainly encountered in practice.

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. In this case the variogram is expressed as a function $\gamma_X \left(\frac{r_1}{\xi_1}, \dots, \frac{r_d}{\xi_d} \right)$ of dimensionless distances $\frac{r_1}{\xi_1}, \dots, \frac{r_d}{\xi_d}$, where ξ_1, \dots, ξ_d are the correlation lengths in the corresponding directions and d is the number of spatial dimensions ($d=1$ for drillholes, $d=2$ for data on a plane, and $d=3$ for data in 3D space).

Zone anisotropy, also called stratified anisotropy, refers to the case where the sill depends on the spatial direction. Then the variogram can be expressed as the sum of two components such that

$$\gamma_X(\mathbf{r}) = \gamma_{X,1}(r) + \gamma_{X,2}(\hat{\mathbf{r}}). \quad (2.19)$$

In the above equation the function $\gamma_{X,1}(r)$, where $r = \|\mathbf{r}\|$ is the magnitude of the distance vector, represents an isotropic dependence, while the function $\gamma_{X,2}(\hat{\mathbf{r}})$ represents the anisotropic dependence of the sill on the direction of the unit vector $\hat{\mathbf{r}}$.

In the case of geometric anisotropy d correlation lengths are required ξ_1, \dots, ξ_d , which are not usually equal to each other. Accordingly, additional parameters are required for the determination of the anisotropy. For two dimensional systems with ξ_x and ξ_y , corresponding to the correlation lengths along the main axis, the anisotropy parameters are:

1. the ratio of correlation lengths $\rho_{y/x} \equiv \xi_y/\xi_x$, which is called anisotropy ratio,
2. the orientation angle θ , which defines the orientation of the main anisotropy axis in relation to the Cartesian coordinate system.

Any isotropic model can be interpreted as a particular case of the geometric anisotropic model with anisotropic ratio $\rho_{y/x}$ equal to one. If $\rho_{y/x} = 1$, it does not matter what the orientation angle θ is [Goovaerts, 1997; Hristopulos, 2012].

The variogram generally increases, but not necessarily linearly, with the distance between the points, while in contrast the correlation function decreases. This is due to the fact that the correlation function describes the dependence between the field values at two different points in space, and their dependence decreases at larger distances. In contrast, the variogram measures the difference between field values as a function of their distance. Therefore, variogram values increase when the distance increases [Journel, 1989].

2.10 Variogram Models

The most commonly used theoretical variogram models include the exponential, gaussian, spherical, power-law, linear and the Matérn functions. Their respective equations are listed below. For the following equations which define the isotropic versions of the models, 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

$$\gamma_x(r) = \sigma_x^2 [1 - \exp(-\|\mathbf{r}\|/\xi)]. \quad (2.20)$$

The exponential model approaches the sill asymptotically. A practical definition of the range is the distance 3ξ where the variogram is equal to $0.95\sigma_x^2$. Geometrically, a tangent at the origin intersects the asymptote σ_x^2 at lag ξ .

Gaussian

$$\gamma_x(r) = \sigma_x^2 [1 - \exp(-\|\mathbf{r}\|^2/\xi^2)]. \quad (2.21)$$

The sill is approached asymptotically. A practical rule is to consider the range as the distance $\sqrt{3\xi}$ where the variogram becomes equal to $0.95\sigma_x^2$. A graph of the model has a parabolic form near the origin. Despite its theoretical permissibility, normal equations involving Gaussian models may be numerically singular [Olea, 1999].

Spherical

$$\gamma_x(r) = \begin{cases} \sigma_x^2 \left[1.5 \left(\frac{\|\mathbf{r}\|}{\xi} \right) - 0.5 \left(\frac{\|\mathbf{r}\|}{\xi} \right)^3 \right] & \text{if } \|\mathbf{r}\| \leq \xi, \\ \sigma_x^2 & \text{if } \|\mathbf{r}\| \geq \xi. \end{cases} \quad (2.22)$$

Near the origin this variogram behaves as a linear function. The model is said to be transitive, because it reaches a finite sill at a finite range ξ . The tangent to this variogram at the origin intersects the sill at two-thirds of the range. The model is not permissible in spaces with dimension larger than 3 [Walck, 1996].

Power-law

$$\gamma_x(\|\mathbf{r}\|) = \alpha \|\mathbf{r}\|^{2H} \quad \text{where } 0 < H < 1, \quad (2.23)$$

where H is the Hurst exponent.

This is a non-transitive model. The linear variogram is a special case for $H = 1/2$, and the pure nugget effect is a special case for $H = 0$. The main advantage of the power model is its simplicity. The power variogram may be used to model an experimental variogram that is not transitive or to model transitive variograms if the maximum lag in the normal equations is smaller than the range, taking the dependence of the variogram beyond the range immaterial [Olea, 1999].

Linear

$$\gamma_x(\|\mathbf{r}\|) = \alpha \|\mathbf{r}\|. \quad (2.24)$$

Matérn

$$\gamma_x(\|\mathbf{r}\|) = \sigma_x^2 \left[1 - \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\sqrt{\nu}}{R} \|\mathbf{r}\| \right)^{\nu} K_{\nu} \left(\frac{2\sqrt{\nu}}{R} \|\mathbf{r}\| \right) \right], \quad (2.25)$$

where ν 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 . For $\nu = 0.5$, the exponential model is recovered, while the Gaussian model is recovered for $\nu \rightarrow \infty$. For all cases $\nu > 0$ [Stein, 1999]. Some special cases are recovered for different values of ν , e.g., for $\nu = 1/3, 1, 3/2, 5/2$ the recovered models are von Kármán, Whittle, Second-order autoregressive and Third-order autoregressive models, respectively [Guttorp and Gneiting, 2006].

The Matérn family has been applied extensively to environmental data as to the analysis of temperature data in the northern United States by Handcock and Wallis [1994], to the statistical assessment of deterministic air quality models by Fuentes [2002] and to the analysis of gravitational fields to diagnose geodetic networks by Meier [1981]. Also Mejía and Rodríguez-Iturbe [1974] established the use of stochastic process models in hydrology using the Matérn model.

Lately, the Matérn correlation family has attracted attention in the machine learning community [Rasmussen and Williams, 2006; Seeger, 2004].

2.10.1 Spartan Variogram Model

Spartan spatial random fields (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. SSRFs have been applied in environmental risk assessment [Hristopulos and Elogne, 2007], atmospheric environment [Žuković and Hristopulos, 2008] and in hydrological data [Varouchakis et al., 2012].

The multivariate probability density function of Gibbs random fields is expressed as:

$$f_x [X(\mathbf{s})] = Z^{-1} \exp\{-H [X(\mathbf{s})]\}. \quad (2.26)$$

The exponent $H [X(\mathbf{s})]$ represents the value of the energy functional $H [X]$ for the state $X(\mathbf{s})$, and the normalization constant $Z = \sum_{X(\mathbf{s})} \exp\{-H [X(\mathbf{s})]\}$ is the partition function.

The isotropic Fluctuation–gradient–curvature (FGC) SSRF functional is given

by the following equation:

$$H[X] = \frac{1}{2\eta_0\xi^d} \int \mathbf{ds} [\{X(\mathbf{s}) - \mathbb{E}[X(\mathbf{s})]\}^2 + \eta_1\xi^2\{\nabla X(\mathbf{s})\}^2 + \xi^4\{\nabla^2 X(\mathbf{s})\}^2]. \quad (2.27)$$

The scale parameter η_0 determines the total variance of the fluctuations, the coefficient η_1 the shape of the covariance spectral density, and ξ is the characteristic length [Hristopulos, 2003].

The spectral density is expressed in terms of η_0 , η_1 , ξ by the equation

$$\tilde{C}_x(\mathbf{k}) = \frac{|\tilde{G}(\mathbf{k})|^2 \eta_0 \xi^d}{1 + \eta_1 (k\xi)^2 + (k\xi)^4}. \quad (2.28)$$

The permissibility conditions for the new parameters require that either the set of relations (i) or the set of relations (ii) below is valid:

$$(i) \quad \eta_0 > 0, \quad \xi > 0, \text{ and } \eta_1 \geq 0,$$

$$(ii) \quad \eta_0 > 0, \quad \xi > 0, \text{ and } \eta_1 < 0, \eta_1^2 < 4.$$

The maximum of the spectral density for $\eta_1 > 0$ is at $k_m = 0$. For $\eta_1 < 0$, the spectral density curve is nonmonotonic, suggesting oscillatory behaviour of the covariance. The maximum of the spectral density occurs at the finite frequency $k_m = \xi^{-1} \sqrt{|\eta_1|/2}$. The height of the spectral peak is $\tilde{C}_x(k_m) = 4\eta_0 \xi^d / (4 - \eta_1^2)$. As $\eta_1 \rightarrow -2$, the width tends to zero while the height tends to infinity.

The spectral representation of the covariance function is given by means of the following one-dimensional integral, where $J_{d/2-1}(r)$ is the Bessel function of the first kind of order $d/2 - 1$ [Hristopulos and Elogne, 2007]

$$C_x(\mathbf{r}) = \frac{\eta_0 r \xi^d}{(2\pi r)^{d/2}} \int_0^{k_c} dk \frac{k^{d/2} J_{d/2-1}(kr)}{1 + \eta_1 (k\xi)^2 + (k\xi)^4}. \quad (2.29)$$

The Spartan covariance function in 3-dimensions is expressed as follows:

$$C_x(\mathbf{h}) = \begin{cases} \frac{\eta_0 e^{-h\beta_2}}{2\pi\Delta} \left[\frac{\sin(h\beta_1)}{h\beta_1} \right] & \text{for } |\eta_1| < 2, \sigma_x^2 = \frac{\eta_0}{2\pi\Delta}, \\ \frac{\eta_0 e^{-h}}{8\pi} & \text{for } \eta_1 = 2, \sigma_x^2 = \frac{\eta_0}{8\pi}, \\ \frac{\eta_0}{4\pi\Delta} \left(\frac{e^{-h\omega_1} - e^{-h\omega_2}}{h} \right) & \text{for } \eta_1 > 2, \sigma_x^2 = \frac{\eta_0}{4\pi\Delta}, \end{cases} \quad (2.30)$$

where, $\omega_{1,2} = (|\eta_1 \mp \Delta|/2)^{1/2}$, $\Delta = |\eta_1^2 - 4|^{1/2}$, and $\beta_{1,2} = |2 \mp \eta_1|^{1/2}/2$.

In Eq. (2.30), $\omega_{1,2}$ and β_2 are dimensionless damping coefficients, β_1 is a dimensionless wave number, ξ is the characteristic length, $\|\mathbf{h}\| = \|\mathbf{r}\|/\xi$ is the normalized lag vector, $h = |\mathbf{h}|$ is the Euclidean norm and σ_x^2 is the variance. The exponential covariance function is obtained from Eq. (2.30) for $\eta_1 = 2$.

In this case study, the Spartan variogram model is used for the estimation of the spatial variability.

2.11 Variogram Estimation

The estimation of the variogram is performed using the sample values and is based on the ergodic hypothesis, which allows us to replace the stochastic mean [calculated with regard to all situations which correspond to the pair of points $X(\mathbf{s}), X(\mathbf{s} + \mathbf{r})$], with the average of all pairs of points that are “approximately” apart at a distance equal to \mathbf{r} . There are various methods for estimating the variogram. We discuss the most commonly used in applied studies below.

2.11.1 Method of Moments

The most commonly used method for estimating the variogram is the method of moments, in which the estimator is expressed by means of the equation

$$\hat{\gamma}_x(\mathbf{r}_k) = \frac{1}{2n(\mathbf{r}_k)} \sum_{i,j=1}^N \{ [X(\mathbf{s}_i) - X(\mathbf{s}_j)]^2 \} \vartheta_{ij}(\mathbf{r}_k), \quad (k = 1, \dots, N_c), \quad (2.31)$$

where,

$$\vartheta_{ij}(\mathbf{r}_k) = \begin{cases} 1, & \mathbf{s}_i - \mathbf{s}_j \in B(\mathbf{r}_k), \\ 0, & \text{otherwise.} \end{cases} \quad (2.32)$$

- The class function $\vartheta_{ij}(\mathbf{r}_k)$ selects those vectors that corresponds to a closed area $B(\mathbf{r}_k)$ centered around the vector \mathbf{r}_k , as shown in Fig. 2.1.
- The variable $n(\mathbf{r}_k)$ is equal to the number of point pairs contained in the class $B(\mathbf{r}_k)$.
- The sample variogram is defined for a finite set of discrete distances \mathbf{r}_k ($k = 1, \dots, N_c$), where N_c is the total number of classes.

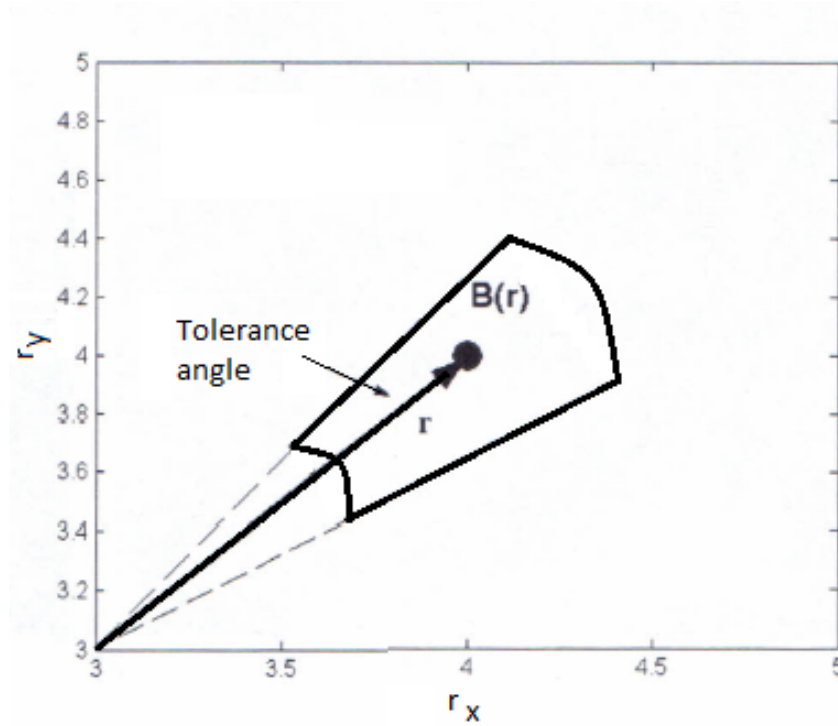


Figure 2.1: Schematic figure of the region $B(\mathbf{r}_k)$ around the distance vector [Hristopulos, 2012]

Therefore, this calculation determines a value for the sample variogram for each \mathbf{r}_k , according to the average value of the squared differences $[X(\mathbf{s}_i) - X(\mathbf{s}_j)]^2$

over all pairs of points, whose distance vector belongs in the $B(\mathbf{r}_k)$ region. The estimator $\hat{\gamma}_X(\mathbf{r}_k)$ is a good approximation of the true $\gamma_X(\mathbf{r}_k)$ if the average of the squared differences in the \mathbf{r}_k class is close to the expectation $\mathbb{E}\{X(\mathbf{s}) - X(\mathbf{s} + \mathbf{r}_k)\}$ [Hristopulos, 2012].

The estimation of the variogram implies an ergodic assumption, which allows to toggle between the stochastic expectation and the sample mean. The ergodic property can be considered as valid if the following conditions hold

1. The increment field $X(\mathbf{s}_i) - X(\mathbf{s}_j)$ is statistically homogeneous. This implies that the process to be investigated shows no significant changes in its statistical properties within the study area.
2. The number of pairs in each class is large enough so the sample average of the squared increments can be calculated with good statistical accuracy.
3. The number of classes is large enough to allow a sufficiently dense mapping of the variogram as a function of distance.

After the experimental variogram is calculated, it is fit to a theoretical model which allows the calculation of the variogram for every possible distance. The theoretical model is needed for the estimation (prediction) of the field at points where measurements are not available. A theoretical variogram is admissible if it is a conditionally negative definite function. This means that for any set of linear coefficients λ_α , where $\alpha = 1, \dots, n$, that satisfies the equation

$$\sum_{\alpha=1}^n \lambda_\alpha = 0, \quad (2.33)$$

the following inequality must apply

$$-\sum_{\alpha=1}^n \sum_{\beta=1}^n \lambda_\alpha \lambda_\beta \gamma_X(\mathbf{s}_\alpha - \mathbf{s}_\beta) \geq 0. \quad (2.34)$$

The variogram permissibility conditions are expressed more concisely with the help of Bochner's theorem as follows: The function $\gamma_X(\mathbf{r})$ is a valid variogram in d dimensions if

1. $\gamma_X(0) = 0$,
2. the generalized Fourier transformation $\tilde{\gamma}_X(\mathbf{k})$ exists,
3. $\tilde{\gamma}_X(\mathbf{k})$ satisfies the inequality $-k^2\tilde{\gamma}_X(\mathbf{k}) \geq 0$, and
4. $\lim \gamma_X(\mathbf{r})/r^2 = 0, \quad r \rightarrow \infty$.

If the random field $X(\mathbf{s})$ is statistically homogenous, it is easier to check the acceptance of a variogram model using the covariance $\sigma_X^2 - \gamma_X(\mathbf{r})$. If the function $\gamma_X(\mathbf{r})$ is a permissible variogram, then the function $c_X(\mathbf{r}) = \sigma_X^2 - \gamma_X(\mathbf{r})$ is an acceptable covariance function and vice versa [Christakos, 1984].

2.12 Spatial Estimation

An important problem in the geosciences is the estimation of a variable over an entire spatial domain based on measurements carried out at a limited number of points. From the mathematical viewpoint this is an interpolation problem. The variable of interest is approximated by a parametric function whose form is assumed in advance, either explicitly or implicitly. If this function is parametric, its parameters are selected so as to optimize some fitting criterion—with respect to the data. Once the approximation function is determined, it is a simple matter to evaluate it wherever needed [Chilès and Delfiner, 2012].

The term estimate includes all the mathematical procedures needed to calculate the field values at points where no measurements of the property are available. The estimate is local, if it refers to a specific point, or global, if it aims to calculate a characteristic (representative) value over an entire region. Based on the parametric function which contains the spatial dependence, it is possible to obtain local estimates in the field where there are no measurements, using the neighboring measured field values. However, in most cases of practical interest, the final objective is to estimate the field at a set of points instead of a single point. For example, in order to construct maps of environmental properties, the field must be estimated at all the nodes of a suitable interpolation grid.

Various models of spatial estimation (interpolation) exist in the literature. The main idea is that the value at the estimation point is given by a linear or

nonlinear combination of the neighboring values. The estimate results from the optimization of a statistical measure, e.g. the maximization of likelihood or the minimization of the mean square estimation error. Interpolation methods are divided in three categories:

- (i) the local neighbourhood approach, which includes methods such as inverse distance weighted that assumes that the value at an unsampled point can be approximated as a weighted average of values at points within a certain cut-off distance; usually the weights are inversely proportional to a power of distance. Also, natural neighbour interpolation falls into the same category, but the weighted averages are dependent on areas or volumes [Mitas and Mitasova, 2005].
- (ii) The variational approach which assumes that the interpolation function should pass through (or close to) the data points and, at the same time, should be as smooth as possible. These two requirements are combined into a single condition of minimising the sum of the deviations from the measured points and the smoothness seminorm of the spline function.
- (iii) The third category includes the geostatistical approaches, interpolation using kriging, the surface or volume is assumed to be one realisation of a random function with a certain spatial covariance [Journel and Huijbregts, 2003].

The most popular methods are based on stochastic linear interpolation in conjunction with the minimization of the mean square error of the estimate. This set of methods is known as “kriging”, a term coined by Matheron [1963] in honor of the South African engineer who first used this method in mineral resources exploration, Danie Krige [1951].

The problem of local estimation is usually expressed as follows: Based on a data set $x(\mathbf{s}_i)$, at \mathbf{s}_i (where $i = 1, \dots, N$) points located within a region Ω , determine 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.

The estimation process is usually repeated at every node of a grid suitably defined for the particular application. This allows the creation of maps representing the isolevel contours of the random field. These maps should be accompanied by an estimate of reliability, which determines the uncertainty of the estimation at each point.

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 . The size of the neighborhood is determined in terms of the correlation length. The neighborhood could ideally be extended over the entire domain, but for large data sets this may lead to computational difficulties. Next, we briefly outline how the kriging methods are formulated by focusing on the simple kriging method. The fluctuation of the field at the estimation point is expressed according to the following linear combination:

$$\hat{X}(\mathbf{u}) - m_X(\mathbf{u}) = \sum_{\mathbf{s}_\alpha \in \omega(\mathbf{u})} \lambda_\alpha [X(\mathbf{s}_\alpha) - m_X(\mathbf{s}_\alpha)] = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha [X(\mathbf{s}_\alpha) - m_X(\mathbf{s}_\alpha)]. \quad (2.35)$$

The coefficients λ_α represent the linear weights. Therefore, the above equation expresses the fluctuation at the estimation point as a function of the fluctuation of the sampling points inside the correlation neighborhood. The estimate of the field is respectively given by the equation

$$\hat{X}(\mathbf{u}) = m_X(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha [X(\mathbf{s}_\alpha) - m_X(\mathbf{s}_\alpha)]. \quad (2.36)$$

The estimator $\hat{X}(\mathbf{u})$ is a random variable, because it consists of a linear combination of random field values. Accordingly, the estimation error $\varepsilon(\mathbf{u}) = \hat{X}(\mathbf{u}) - X(\mathbf{u})$, is also a random variable. Kriging is a form of generalized linear regression formulates the optimal estimator $\hat{X}(\mathbf{u})$ using weights that minimize the estimation error variance [Chilès and Delfiner, 2012].

2.12.1 Kriging Methods

There exist various formulations of kriging which aim to adapt to different types of natural variability. Below, we present some formulations of kriging which differ

between them in how they treat the trend function $m_X(\mathbf{u})$.

2.12.1.1 Simple Kriging (SK)

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

$$\hat{X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} X(\mathbf{s}_{\alpha}) - m_X \left[\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} - 1 \right]. \quad (2.37)$$

The linear weights λ_{α} are then determined so as to minimize the error variance given by the equation

$$\sigma_{E,SK}^2(\mathbf{u}) = \text{Var} \left[X(\mathbf{u}) - \hat{X}(\mathbf{u}) \right] = \text{Var} \left[\hat{X}(\mathbf{u}) - m_X - X'(\mathbf{u}) \right]. \quad (2.38)$$

The equation of the estimator $\hat{X}(\mathbf{u})$ leads to the following relation for the fluctuation of the random variable $\hat{X}(\mathbf{u})$

$$\hat{X}(\mathbf{u}) - m_X = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} [X(\mathbf{s}_{\alpha}) - m_X] = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} X'(\mathbf{s}_{\alpha}). \quad (2.39)$$

The error variance can thus be expressed as a quadratic function of the weights with coefficients that are determined by the linear weights λ_{α} , $\alpha = 1, \dots, n(\mathbf{u})$.

$$\begin{aligned} \sigma_{E,SK}^2(\mathbf{u}) &= \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_{\alpha} \lambda_{\beta} \mathbb{E} \left[X'(\mathbf{s}_{\alpha}) X'(\mathbf{s}_{\beta}) \right] + \mathbb{E} \left[X'^2(\mathbf{u}) \right] \\ &\quad - 2 \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} \mathbb{E} \left[X'(\mathbf{s}_{\alpha}) X'(\mathbf{u}) \right] \\ &= \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_{\alpha} \lambda_{\beta} c_X(\mathbf{s}_{\alpha} - \mathbf{s}_{\beta}) + \sigma_X^2 - 2 \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} c_X(\mathbf{s}_{\alpha} - \mathbf{u}). \end{aligned} \quad (2.40)$$

The optimal values of the linear weights minimize the error variance. The weights are obtained by setting to zero each of the $n(\mathbf{u})$ partial first derivatives

of the error variance with respect to the weights, i.e.,

$$\frac{\partial \sigma_{E,SK}^2(\mathbf{u})}{\partial \lambda_\alpha} = 0, \quad \alpha = 1, \dots, n(\mathbf{u}). \quad (2.41)$$

The system of $n(\mathbf{u})$ linear equations is known as the system of normal equations [Luenberger, 1969]. After calculating the derivatives, the system is expressed in the following form and is expressed by the equation

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta c_X(\mathbf{s}_\alpha - \mathbf{s}_\beta) = c_X(\mathbf{s}_\alpha - \mathbf{u}), \quad \alpha = 1, \dots, n(\mathbf{u}). \quad (2.42)$$

The above system of linear equations may also be expressed as the matrix equation

$$\mathbf{C}_{\alpha,\beta} \lambda_\beta = \mathbf{C}_{\alpha,u}. \quad (2.43)$$

The matrix $\mathbf{C}_{\alpha,\beta}$, represents the covariance matrix with elements $\mathbf{C}_{\alpha,\beta} = c_X(\mathbf{s}_\alpha - \mathbf{s}_\beta)$. The vector $\mathbf{C}_{\alpha,u}$ represents the values of the covariance function between the sample points and the estimation point, i.e., $\mathbf{C}_{\alpha,u} = c_X(\mathbf{s}_\alpha - \mathbf{u})$.

Using the equation $c_X(0) = \sigma_X^2$, the linear system is written in more detail in the form of matrices as follows:

$$\begin{bmatrix} \sigma_X^2 & \dots & \dots & c_X(\mathbf{s}_1 - \mathbf{s}_n) \\ c_X(\mathbf{s}_2 - \mathbf{s}_1) & \dots & \dots & c_X(\mathbf{s}_2 - \mathbf{s}_n) \\ \vdots & \vdots & \vdots & \vdots \\ c_X(\mathbf{s}_n - \mathbf{s}_1) & \dots & \dots & \sigma_X^2 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} c_X(\mathbf{s}_1 - \mathbf{u}) \\ c_X(\mathbf{s}_2 - \mathbf{u}) \\ \vdots \\ c_X(\mathbf{s}_n - \mathbf{u}) \end{bmatrix} \quad (2.44)$$

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

$$\lambda_\beta = C_{\beta,\alpha}^{-1} C_{\alpha,u}, \quad \forall \beta = 1, \dots, n(\mathbf{u}). \quad (2.45)$$

The covariance matrix elements are calculated from the optimum variogram by means of the equation

$$\sigma_X^2 - \gamma_X(\mathbf{s}_\alpha, \mathbf{s}_\beta) = c_X(\mathbf{s}_\alpha, \mathbf{s}_\beta). \quad (2.46)$$

The linear system (Eq. (2.45)) has a solution for the coefficients λ_β , where $\beta = 1, \dots, n(\mathbf{u})$, if the covariance function is permissible (semi positive definite), and if each point has a unique value. The values of the linear coefficients are independent of the sill of the variogram, but they depend on the covariance model, i.e., on the form of the spatial dependence. Kriging is considered as an exact interpolator, in the sense that at every point where a measurement is available, the kriging estimate coincides with the sample value. This exactitude property is not true if the variogram (or the covariance, equivalently) involves a nugget effect. Then, the nugget variance is added to the diagonal elements of the covariance matrix on the left side of Eq. (2.44). As a result of this additional variance, the exactitude property is lost.

The reliability (uncertainty) of the estimate is determined by the square root of the variance of the estimation error. The variance $\sigma_{E,SK}^2(\mathbf{u})$ is determined from the following equation based on Eqs. (2.40), (2.45)

$$\sigma_{E,SK}^2(\mathbf{u}) = \sigma_x^2 - \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} C_{u,\alpha} C_{\alpha,\beta}^{-1} C_{\beta,u}. \quad (2.47)$$

This is equivalently expressed as

$$\sigma_{E,SK}^2(\mathbf{u}) = \sigma_x^2 \left[1 - \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \rho_{\alpha,u} \rho_{\alpha,\beta}^{-1} \rho_{\beta,u} \right], \quad (2.48)$$

where $\rho_{\alpha,\beta}$ are the elements of the correlation matrix defined by means of $C_{\alpha,\beta} = \sigma_x^2 \rho_{\alpha,\beta}$.

Equations (2.47), (2.48) show that the error variance increases proportionally to the random field variance σ_x^2 . Also, assuming positive values of the weights λ_α , the error increases as the distance $|\mathbf{u} - \mathbf{s}_\alpha|$ between the estimation points and the sample point decreases, because then the correlation $\rho_{u,\alpha}$ tends to one [Chilès and Delfiner, 2012; Goovaerts, 1997].

2.12.1.2 Ordinary Kriging (OK)

Ordinary kriging (OK) is applied when the mean $m_x(\mathbf{u})$ 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 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 only of the $X(\mathbf{s}_\alpha)$ where $\alpha = 1, \dots, n(\mathbf{u})$, as

$$\hat{X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha), \quad (2.49)$$

$$\text{with } \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1. \quad (2.50)$$

Equation (2.50) is the unbiasedness constraint.

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

$$\begin{aligned} \sigma_{E,OK}^2(\mathbf{u}) = & \sigma_X^2(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\alpha \lambda_\beta \mathbb{E} \left[X'(\mathbf{s}_\alpha) X'(\mathbf{s}_\beta) \right] \\ & - 2 \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha \mathbb{E} \left[X'(\mathbf{s}_\alpha) X'(\mathbf{u}) \right] + 2\mu \sum_{\beta=1}^{n(\mathbf{u})} (\lambda_\beta - 1), \end{aligned} \quad (2.51)$$

where the constant 2μ is the Lagrange parameter. Using the covariance function,

Eq. (2.51) is expressed as

$$\begin{aligned} \sigma_{E,OK}^2(\mathbf{u}) = & \sigma_X^2(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_{\alpha} \lambda_{\beta} c_X(\mathbf{s}_{\alpha}, \mathbf{s}_{\beta}) \\ & - 2 \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} c_X(\mathbf{s}_{\alpha}, \mathbf{u}) + 2\mu \sum_{\beta=1}^{n(\mathbf{u})} (\lambda_{\beta} - 1). \end{aligned} \quad (2.52)$$

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.,

$$\frac{\partial \sigma_{E,OK}^2(\mathbf{u})}{\partial \lambda_{\alpha}} = 0, \quad \alpha = 1, \dots, n(\mathbf{u}), \quad (2.53)$$

$$\frac{\partial \sigma_{E,OK}^2(\mathbf{u})}{\partial \mu} = 0. \quad (2.54)$$

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

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_{\beta} c_X(\mathbf{s}_{\alpha} - \mathbf{s}_{\beta}) + \mu = c_X(\mathbf{s}_{\alpha} - \mathbf{u}), \quad \alpha = 1, \dots, n(\mathbf{u}), \quad (2.55)$$

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} = 1. \quad (2.56)$$

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

$$\begin{bmatrix} \sigma_X^2 & c_X(\mathbf{s}_1 - \mathbf{s}_2) & \dots & c_X(\mathbf{s}_1 - \mathbf{s}_n) & 1 \\ c_X(\mathbf{s}_2 - \mathbf{s}_1) & \sigma_X^2 & \dots & c_X(\mathbf{s}_2 - \mathbf{s}_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ c_X(\mathbf{s}_n - \mathbf{s}_1) & c_X(\mathbf{s}_n - \mathbf{s}_2) & \dots & \sigma_X^2 & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} c_X(\mathbf{s}_1 - \mathbf{u}) \\ c_X(\mathbf{s}_2 - \mathbf{u}) \\ \vdots \\ c_X(\mathbf{s}_n - \mathbf{u}) \\ 1 \end{bmatrix} \quad (2.57)$$

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

$$\lambda_\beta = C_{\beta,\alpha}^{-1} C_{\alpha,u}, \quad \forall \quad \beta = 1, \dots, n(\mathbf{u}). \quad (2.58)$$

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

$$\sigma_{E,OK}^2(\mathbf{u}) = \sigma_X^2 - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha c_X(\mathbf{u}, \mathbf{s}_\alpha) - \mu, \quad (2.59)$$

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

2.12.1.3 Kriging with a Trend Model – Universal Kriging(UK)

Universal kriging(UK) considers that the unknown local mean varies smoothly within each local neighborhood $\omega(\mathbf{u})$. In this case, the trend component is modeled as a linear combination of known functions. Hence, the trend in each correlation neighborhood is given as

$$m_X(\mathbf{s}') = \sum_{k=0}^K \alpha_k(\mathbf{s}') \psi_k(\mathbf{s}') \approx \sum_{k=0}^K \alpha_k(\mathbf{u}) \psi_k(\mathbf{s}'), \quad \mathbf{s}' \in \omega(\mathbf{u}). \quad (2.60)$$

The functions $\psi_k(\mathbf{s}')$ are known, whereas the coefficients $\alpha_k(\mathbf{s}')$ are unknown and deemed constant within each local neighborhood $\omega(\mathbf{u})$. The linear estimator is thus written

$$\begin{aligned} \hat{X}(\mathbf{u}) &= \sum_{k=0}^K \alpha_k(\mathbf{u}) \psi_k(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha \left[X(\mathbf{s}_\alpha) - \sum_{k=0}^K \psi_k(\mathbf{s}_\alpha) \alpha_k(\mathbf{u}) \right] \\ &= \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha) + \sum_{k=0}^K \alpha_k(\mathbf{u}) \left[\psi_k(\mathbf{u}) - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha \psi_k(\mathbf{s}_\alpha) \right]. \end{aligned} \quad (2.61)$$

Consequently, the minimized error variance is calculated as

$$\sigma_{E,UK}^2(\mathbf{u}) = c_{X,R}(0) - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha c_{X,R}(\mathbf{s}_\alpha - \mathbf{u}) - \sum_{k=0}^K \mu \psi_k(\mathbf{u}), \quad (2.62)$$

where μ are the Lagrange parameters and $c_{x,r}$ is the covariance of the residual component [Goovaerts, 1997].

2.12.1.4 Regression Kriging – Residual Kriging (RK)

Regression Kriging (RK) combines a trend function with interpolation of the residuals. In RK the estimate is expressed as

$$\hat{X}(\mathbf{u}) = m_x(\mathbf{u}) + \hat{X}'(\mathbf{u}), \quad (2.63)$$

where $m_x(\mathbf{u})$ is the trend function, and $\hat{X}'(\mathbf{u})$ is the interpolated residual by means of OK [Rivoirard, 2002]. Typically, the trend function is modeled as

$$m_x(\mathbf{u}) = \sum_{k=0}^p \beta_k q_k(\mathbf{u}) \quad q_0(\mathbf{u}) = 1, \quad (2.64)$$

where $q_k(\mathbf{u})$ are the values of auxiliary variables at \mathbf{u} , β_k are the estimated regression coefficients and p is the number of auxiliary variables [Draper and Smith, 1981; Hengl, 2007; Hengl et al., 2007]. Auxiliary variables could include polynomials of the data coordinates (x,y). The regression coefficients, β_k , are estimated from the sample using ordinary least squares (OLS) or optimally, generalized least squares (GLS) as

$$\beta_{GLS} = (\mathbf{q}^T \cdot \mathbf{c}^{-1} \cdot \mathbf{q})^{-1} \cdot \mathbf{q}^T \cdot \mathbf{c}^{-1} \cdot \mathbf{X}, \quad (2.65)$$

where β_{GLS} is the vector of estimated regression coefficients, \mathbf{c} is the covariance matrix of the residuals, \mathbf{q} is a matrix of predictors at the sampling locations, and \mathbf{X} is the vector of measured values of the variable.

The variance of the estimates follows from the equations [Hengl et al., 2003, 2007]:

$$\sigma_{E,RK}^2(\mathbf{u}) = \sigma_d^2(\mathbf{u}) + \sigma_f^2(\mathbf{u}), \quad (2.66)$$

$$\sigma_d^2(\mathbf{u}) = \mathbf{q}_u^T \left(\mathbf{q}^T \mathbf{c}_{x'}^{-1} \mathbf{q} \right)^{-1} \mathbf{q}(u), \quad (2.67)$$

$$\sigma_f^2(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha} c_{x'}(s_{\alpha} - \mathbf{u}) + \mu, \quad (2.68)$$

where $\sigma_d^2(\mathbf{u})$ is the drift prediction variance, \mathbf{q}_u is the vector of $(p+1) \times 1$ predictors at the unvisited location, \mathbf{q} is the matrix of $(n_u + 1) \times (p + 1)$ predictors at the sampling points in the search neighborhood, $c_{x'}$ is the variogram matrix of the $(n_u + 1) \times (n_u + 1)$ residuals at the measured locations (neighborhood) and $\sigma_f^2(\mathbf{u})$ is the kriging (OK) variance of residuals. The matrices and vectors involved in the RK variance [Varouchakis, 2012]:

$$\mathbf{q}_u = \begin{bmatrix} q_1(\mathbf{u}) \\ q_2(\mathbf{u}) \\ \vdots \\ q_p(\mathbf{u}) \\ 1 \end{bmatrix} \quad (2.69)$$

$$\mathbf{q} = \begin{bmatrix} q_1(\mathbf{u}_1) & \dots & q_p(\mathbf{u}_1) & 1 \\ q_1(\mathbf{u}_2) & \dots & q_p(\mathbf{u}_2) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ q_1(\mathbf{u}_{n_u}) & \dots & q_p(\mathbf{u}_{n_u}) & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix} \quad (2.70)$$

$$\mathbf{c}_{X'} = \begin{bmatrix} \sigma_{x'}^2 & c_{x'}(\mathbf{s}_1 - \mathbf{s}_2) & \dots & c_{x'}(\mathbf{s}_1 - \mathbf{s}_n) & 1 \\ c_{x'}(\mathbf{s}_2 - \mathbf{s}_1) & \sigma_{x'}^2 & \dots & c_{x'}(\mathbf{s}_2 - \mathbf{s}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{x'}(\mathbf{s}_n - \mathbf{s}_1) & c_{x'}(\mathbf{s}_n - \mathbf{s}_2) & \dots & \sigma_{x'}^2 & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \quad (2.71)$$

The method of regression kriging is used in applications, such as the modeling of spatial variability in tropical rainforest soils [Yemefack et al., 2005], mapping of leaf area index (LAI) [Berterretche et al., 2005], modeling spatial distribution of human diseases [Pleydell et al., 2004], and mapping of groundwater levels [Varouchakis et al., 2012]. It is a powerful spatial prediction technique that can be used to interpolate environmental variables (both continuous and categorical) from large point sets. The asset of the method relies on its ability to combine a trend model between the dependent variable and auxiliary variables (such as land

surface parameters), and allow separate interpretation of the two interpolated components [Hengl et al., 2007]. The use of regression kriging is limited because the analyst must carry out various steps in different software environments, both statistical and GIS, it is computationally demanding and it is sometimes difficult to find the auxiliary variables that affects the data sets.

In this study, we use regression kriging interpolation to estimate the precipitation over the entire island of Crete based on available measurements from a limited number of stations. We opt for this method because we can use physical motivation to construct a trend model in terms of auxiliary variables and the residuals obtained after removing the trend are close to the Gaussian distribution.

2.13 Spatial Model Validation

Validation methods provide the means for assessing the performance of different spatial models or of the same model with different parameters, in terms of statistical measures. There are methods such as likelihood maximization, least squares and empirical contrast minimization, to quantify the fit of the data to specific spatial models. These methods provide a first account of the model’s ability to represent the data. Validation typically involves methods that also measure the predictive performance of the model to predict based on the available data. Let us call statistical algorithm any function that returns an estimator from data, such as maximum likelihood. Then, model selection can be seen as the selection of a particular (statistical) algorithm [Sylvain and Alain, 2010].

Cross-validation (CV) —a popular strategy for algorithm selection— is a model validation technique for assessing the predictive performance of a statistical spatial model. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. The main idea behind CV is to split data, once or several times, in order to estimate the accuracy and reliability of each algorithm. Part of the data (the training sample) is used for training each algorithm, while the remaining data (the validation sample) is used to evaluate the predictive performance of the algorithm. CV selects the algorithm with the smallest estimated “risk” (where the notion of risk may involve a single statistical measure or a combination of

statistical measures).

The goal of cross-validation is to estimate the expected level of fit of a model to a data set that is independent of the data (training sample) that were used to train the model. It can be based on any quantitative measure of fit that is appropriate for the data and the model. Most forms of cross-validation are straightforward to implement so long as an implementation of the prediction method is available. In particular, the prediction method needs only to be available as a “black box”; there is no need to have access to the internals of its implementation [Grossman et al., 2010].

Furthermore, one of the main reasons for using cross-validation instead of using the conventional validation (e.g. partitioning the data set into two sets of 70% for training and 30% for test) is that the error (e.g. Root Mean Square Error) on the training set in the conventional validation is not a useful estimator of model performance and thus the error on the test data set does not properly represent the assessment of model performance. This may be because there is not enough data available or there is not a good distribution and spread of data to partition it into separate training and test sets in the conventional validation method. In these cases, a fair way to properly estimate model prediction performance is to use cross-validation as a powerful general technique.

In summary, cross-validation combines average measures of fit (i.e., measures of prediction error) to minimize the estimation error and derive a more accurate estimate of model prediction performance [Grossman et al., 2010].

Below we present the most commonly used methods for model validation.

2.13.1 Cross-validation Approach

Most classical CV estimators split the data of the training set and are distinguished into two main categories: exhaustive data splitting and partial data splitting.

Exhaustive data splitting

Exhaustive cross-validation methods are CV methods which learn and test on all possible ways to divide the original sample into a training and a validation set.

1. **Leave-p-out cross-validation** (LPO), also called delete-p CV or delete-p multifold [Zhang, 1993] involves using p observations as the validation set and the remaining observations as the training set. This is repeated on all ways to cut the original sample with n observations on a validation set of p observations and a training set. LPO cross-validation requires to learn and validate C_p^n times, where the possible combinations of the set C are calculated by the binomial coefficient $\binom{n}{p}$. The binomial coefficient is defined by the relation $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, where n! denotes the factorial of n. Hence, if $n \gg 1$, it becomes impossible to calculate.
2. **Leave-one-out cross-validation** (LOOCV or LVO), also called delete-one CV [Li, 1987], ordinary CV [Burman, 1989; Stone, 1974] or simply CV [Efron, 1983; Li, 1987], is a particular case of leave-p-out cross-validation with $p = 1$. It can be shown that the CV error estimate is an almost unbiased estimate of the true error expected on an independent test set [Sudhir and Richard, 2006]. LVO cross-validation does not face the computational constraints of general LPO cross-validation because $C_1^n = n$.

In this case study, the LVO method of cross-validation is applied to evaluate the fit between the theoretical model and the data.

Partial data splitting

Non-exhaustive cross validation methods do not compute all possible ways of splitting the original sample. The methods are approximations of leave-p-out cross-validation.

1. In **k-fold cross-validation**, sometimes called rotation estimation, the original sample is randomly partitioned into k mutually exclusive subsets (the folds) of approximately equal size [Kohavi, 1995]. Of the k subsets, a single subset is retained as the validation data for testing the model, and the remaining k-1 subsets are used as training data. The cross-validation process is then repeated k times, with each of the k subsets used exactly once as the validation data. The k results from the folds are then averaged (or otherwise combined) to produce a single estimate. The advantage of

this method over repeated random sub-sampling is that all observations are used for both training and validation, and each observation is used for validation exactly once. 10-fold cross-validation is commonly used, but in general k remains an unfixed parameter [McLachlan et al., 2004].

When $k = n$ (the number of observations), the k -fold cross-validation becomes equivalent to LVO CV.

2. **Repeated learning-testing** (RLT) randomly splits the dataset into training and validation data. For each such split, the model is fit to the training data, and predictive accuracy is assessed using the validation data. The results are then averaged over the splits. The advantage of this method (over k -fold cross-validation) is that the proportion of the training/validation split is not dependent on the number of iterations (folds). The disadvantage of this method is that some observations may never be selected in the validation subsets, whereas others may be selected more than once. In other words, validation subsets may overlap. This method also exhibits Monte Carlo variation, meaning that the results vary if the analysis is repeated with different random splits [Breiman et al., 1984; Burman, 1989; Zhang, 1993].

As the number of random splits goes to infinity, the repeated random sub-sampling validation is arbitrary close to the leave- p -out cross-validation. In a stratified variant of this approach, the random samples are generated in such a way that the mean response value (i.e. the dependent variable in the regression) is equal in the training and testing sets. This is particularly useful if the responses are dichotomous with an unbalanced representation of the two response values in the data.

3. **Monte-Carlo cross validation** (MCCV) [Picard and Cook, 1984] is very similar to RLT. Unlike RLT, MCCV allows the same split to be chosen several times [Sylvain and Alain, 2010].

2.13.2 Cross-validation Error Measures

In order to assess the model performance certain statistical measures need to be evaluated. These measures include: the mean error (bias) (ME), the mean absolute error (MAE), the mean absolute relative error (MARE), the root mean square error (RMSE), the root mean square relative error (RMSRE), Pearson's linear correlation coefficient (RP) and Spearman (rank) correlation coefficient (RS). Below we define these measures in the case of leave-one-out cross validation. For the following measures, $x^*(\mathbf{s}_i)$ and $x(\mathbf{s}_i)$ are, respectively, the estimated (based on the $N - 1$ data that do not include point \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 $\overline{x^*(\mathbf{s}_i)}$ the spatial average of the estimates, while N is the number of observations.

Mean error (bias) (ME)

The mean error is calculated as follows:

$$\varepsilon_{\text{bias}} = \frac{1}{N} \sum_{i=1}^N x^*(\mathbf{s}_i) - x(\mathbf{s}_i). \quad (2.72)$$

Mean absolute error (MAE)

The mean absolute error is calculated as follows:

$$\varepsilon_{\text{MA}} = \frac{1}{N} \sum_{i=1}^N |x^*(\mathbf{s}_i) - x(\mathbf{s}_i)|. \quad (2.73)$$

Mean absolute relative error (MARE)

The mean absolute relative error is calculated as follows:

$$\varepsilon_{\text{MAR}} = \frac{1}{N} \sum_{i=1}^N \left| \frac{x^*(\mathbf{s}_i) - x(\mathbf{s}_i)}{x(\mathbf{s}_i)} \right|. \quad (2.74)$$

Root mean square error (RMSE)

The root mean square error is calculated as follows:

$$\varepsilon_{\text{RMS}} = \sqrt{\frac{1}{N} \sum_{i=1}^N [x^*(\mathbf{s}_i) - x(\mathbf{s}_i)]^2}. \quad (2.75)$$

Root mean square relative error (RMSRE)

The root mean square relative error is calculated as follows:

$$\varepsilon_{\text{RMSR}} = \sqrt{\frac{1}{N} \sum_{i=1}^N \left[\frac{x^*(\mathbf{s}_i) - x(\mathbf{s}_i)}{x(\mathbf{s}_i)} \right]^2}. \quad (2.76)$$

Pearson's Linear correlation coefficient (RP)

The correlation coefficient, ρ , is the statistic that is most commonly used to summarize the relationship between two variables. The formula for Pearson's linear correlation coefficient ρ is [Isaaks and Srivastava, 1989]

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}.$$

In case of cross-validation for a sample of size N , ρ is defined as

$$\bar{\rho}_{X,X^*} = \frac{\sum_{i=1}^N [x(\mathbf{s}_i) - \overline{x(\mathbf{s}_i)}] [x^*(\mathbf{s}_i) - \overline{x^*(\mathbf{s}_i)}]}{\sqrt{\sum_{i=1}^N [x(\mathbf{s}_i) - \overline{x(\mathbf{s}_i)}]^2} \sqrt{\sum_{i=1}^N [x^*(\mathbf{s}_i) - \overline{x^*(\mathbf{s}_i)}]^2}}. \quad (2.77)$$

The correlation coefficient measures the dispersion of estimates with respect to the observed values. This relation can best be illustrated in terms of a scatterplot. If $\rho = +1$, the scatterplot is a straight line with a positive slope; if $\rho = -1$, the scatterplot is a straight line with a negative slope. For $|\rho| < 1$ the scatterplot appears as a cloud of points that becomes more diffuse as $|\rho|$ decreases from 1 to 0 [Isaaks and Srivastava, 1989].

Spearman (rank) correlation coefficient (RS)

It is important to note that ρ provides a measure of the linear relationship between two variables. If the relationship between two variables is not linear, the correlation coefficient may be a very poor summary statistic. It is often useful to supplement the linear correlation coefficient with another measure of the strength of the relationship, the rank correlation coefficient. To calculate the rank correlation coefficient, one applies Eq. (2.77) to the ranks of the data values rather than to the original sample values:

$$\rho_{rank} = 1 - \frac{6 \sum_{i=1}^N (R_{x_i} - R_{y_i})^2}{N(N^2 - 1)}, \quad (2.78)$$

where R_{x_i} is the rank of x_i among all the other x values. The rank is usually calculated by sorting the x values in ascending order; the rank of a given value is equal to its order of appearance in the sorted list. 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 .

Large differences between ρ_{rank} and ρ are often quite revealing about the existence of extreme pairs on the scatterplot. Unlike the traditional correlation coefficient, the rank correlation coefficient is not strongly influenced by extreme pairs. Large differences between the two may be due to the location of extreme pairs on the scatterplot. A high value of ρ_{rank} and a low value of ρ may be due to the fact that a few erratic pairs have adversely affected an otherwise good correlation. If, on the other hand, it is ρ that is quite high while ρ_{rank} is quite low, then it is likely that the high value of ρ is due largely to the influence of a few extreme pairs.

Differences between ρ and ρ_{rank} may also reveal important features of the relationship between two variables. If the rank correlation coefficient is $+1$, then the ranks of the two variables are identical: the largest value of x corresponds to the largest value of y , and the smallest value of x corresponds to the smallest value of y .

The value of ρ is often a good indicator of how successful we might be when trying to predict the value of one variable from the other with a linear equation. If $|\rho|$ is large, then for a given value of one variable, the other variable is restricted

to only a small range of possible values. On the other hand, if $|\rho|$ is small, then knowing the value of one variable does not help us very much in predicting the value of the other [Isaaks and Srivastava, 1989].

Chapter 3

Exploratory Data Analysis

3.1 Information about the Study Area

The study area is the island of Crete (Greece), in the southeastern part of the Mediterranean region. Crete is the largest island in Greece and one of its thirteen (13) geographical regions. The island is divided into four prefectures, Chania, Rethimno, Heraklion and Lasithi, with Heraklion as the capital and the largest city.

The island of Crete covers an area of 8 336 km² in the Southern Aegean. It has a coastline of 1 046 km and its population reaches 623 065 people. The Sea of Crete lies to the north, the Lybian Sea to the south, the Myrtoan Sea to the west and the Karpathanian Sea to the East of the island. Crete has an elongated shape, spanning 260 km from east to west; it is 60 km at its widest point and narrows to only 12 km at the narrowest point [[Hellenic Statistical Authority, 2014](#)]. Comparing Crete with other major European islands, that cover an area of more than 200 km², it ranks at the 12th place, with United Kingdom being the biggest. The size of Crete is similar to the size of Cyprus and Corsica [[wikipedia, 2015a](#)].

Crete is one of the most mountainous islands of Europe with high mountain ranges crossing the island from West to East forming three large mountain complexes. In between lie lower mountains and semi-mountainous areas. In Western Crete, the White Mountains or “Lefka Ori” with Pachnes (2 456m) as its highest

peak, is also the highest point on Crete. The middle of the island is dominated by Mount Ida or “Psiloritis”, a single elongated mountain mass whose highest peak is Timios Stavros (2 453m) while in eastern Crete Dikti Mountains rise with their highest peak at 2 148m (Fig. 3.1) [Wolffenbuttel and van der Meijden, 2015].



Figure 3.1: Morphological map of the island of Crete, Greece [Google Earth, 2015].

The prefecture of Lasithi is categorized mostly as a semiarid bioclimatic zone with mild or warm winter. The sub-mountainous areas belong to the semi-dry bioclimatic zone with cold winter, while the mountainous areas (i.e. Lasithi Plateau) belong to dry bioclimatic zone with cold winter [Watrous, 1982].

Only the northern part of Heraklion prefecture belongs to the semiarid bioclimatic zone with warm winter. The rest belongs to the semi-arid with mild to warm winter. The plains and sub-mountainous areas of both Chania and Rethimno prefectures belong to semi-arid bioclimatic zone with mild or warm winter. The mountainous areas belong to dry bioclimatic zone with mild to cold winter. A very small part of the very high mountainous areas belong to a humid bioclimatic zone with harsh winter [Vrochidou, 2009].

The few plains are limited to the coastal area, where most of the population lives and the most important agricultural areas are located (Fig. 3.1). The largest and most productive plain of the island is Messara valley in the south of Heraklion prefecture. Ierapetra's valley in the south-east also has significant agricultural

Table 3.1: Crops cultivated in Crete and their respective farming area. The units are in $(\times 10^3 \text{ha})^a$. Table taken from [Chartzoulakis et al. \[2001\]](#).

Category	Cultivated area	Percentage of the total
Row crops	32.1	9.96
Vegetable crops	8.9	2.76
Vineyards	30.6	9.50
Fruit crops	182.6	56.70
Forage crops	14.2	4.50
Fallow fields	53.6	16.58
Arable land	322.0	100.00

activity.

Agriculture is an important source of income for the region of Crete, contributing to local Gross Domestic Product (GDP) by 13%. Fruit crops cover 60.8% of the total cultivated area, from which about 89% represents olive orchards [[Chartzoulakis et al., 2001](#)]. Grapevines and greenhouse vegetable crops are also important sources of income for farmers. About 42.3% of the cultivated land is irrigated; the percentage of irrigated land is higher for vegetable crops and lower for tree crops and grapevines (Table 3.1).

During the past thirty years, groundwater level has decreased due to over-exploitation [[Varouchakis et al., 2012](#)]. Potential future climatic changes in the Mediterranean region, population increase, and extensive agricultural activity generate concern over the sustainability of water resources. The accurate estimation of the spatial and temporal variability of precipitation is important for integrated water resources management plans that will help reduce the risk of water deficits and desertification. It is also important in helping to understand the potential impact of global climate change on the island.

3.1.1 Temperature

In order to describe the climate in Crete, temperature records were used for the time period of 2006 till 2013. Evaluation of the data showed that:

1. The minimum temperature of -6.9°C was observed in Samaria at 2012.

2. The maximum temperature of 45°C was observed in Paleochora at 2007.

These data are presented on Table 3.2 separated into monthly maxima and minima.

Table 3.2: Monthly temperature minima and maxima recorded on Crete between minimum and maximum recorded temperatures (°C).

Month	January		February		March		April	
Min	−5.6	Samaria 2012	−6.9	Samaria 2012	−6.4	Samaria 2011	0.3	Samaria 2012
Max	30.4	Vrysses 2010	32.1	Vrysses 2010	30.1	Vrysses 2010	36.7	Chania 2008
Month	May		June		July		August	
Min	3.4	Samaria 2011	9.0	Samaria 2013	10.8	Samaria 2013	11.9	Anogeia 2009
Max	38.8	Chania 2010	45	Paleochora 2007	42.4	Paleochora 2007	40.4	Moires 2012
Month	September		October		November		December	
Min	7.5	Samaria 2008	2.2	Samaria 2011	−0.1	Samaria 2011	−6.1	Samaria 2010
Max	37.7	Chania 2007	34.2	Chania 2010	30.8	HeraclionPort 2012	29.6	Vrysses 2010

Evaluation of the data shows that the minimum temperature is lower in the west side of Crete and increases from west to east, from −6.9°C in the west to 2.6°C in the east.

Based on the available temperature data of Crete, the following conclusions can be drawn:

- The lowland area of Crete exhibits a climatic transition from the Mediterranean to the semi-desert climate. It is characterized by little rainfall, mild winters and long dry seasons. Summer is relatively cool and lasts from June to September, because of the sea breeze and the annual winds. The hottest months are July and August. In this region, frost is never observed and the temperature seldom falls below 0°C. The temperature differences along the lowland over the year are moderate.

- The mountainous zone of Crete shows greater variation and an annual average temperature 2 – 3°C lower than the lowland. The mean temperature of the warmest month (July) is comparable to the average temperature of the warmest month of the lowland stations, but the average temperature of the coldest month (February) is approximately 3°C lower [Region of Crete Information Bull., 2002; Vrochidou, 2009].

3.1.2 Humidity

The mean relative atmospheric humidity is minimum in June and maximum in December across the north of the island. In southern Crete, the minimum mean monthly relative humidity occurs in July (Ierapetra and Timbaki stations). Among the northern stations, the minimum mean relative humidity decreases from east to west (Sitia 59.88%, Heraklion 55.4% and Souda 48.90%) while the maximum relative humidity is similar in Sitia and Souda (about 72%) and significantly lower in Heraklion (67%). The greatest variation of the relative humidity occurs at the southern stations (27%) while the lowest occurs at Heraklion station (only 12%).

During the summer months, when the temperatures are particularly high (especially in southern Crete), the low mean relative humidity makes summer pleasant which is especially important in tourist areas [Region of Crete Information Bull., 2002].

3.1.3 Topography

In order to incorporate topographic parameters into the spatial model that we will construct in Section 4, the Digital Elevation Model (DEM) of Crete is necessary. The DEM was created in Geographical Information Systems (GIS) [esri, 2015], and more specifically in ArcGIS, release version 10.0, and the mostly used feature was the 3D Analyst. Firstly, seven separate SRTM data sets comprising Crete were downloaded from Processes Distributed Active Archive Center [LP DAAC]. The NASA SRTM data sets result from a collaborative effort by the National Aeronautics and Space Administration (NASA) and the National Geospatial-Intelligence Agency (NGA), as well as the participation of the Ger-

man and Italian space agencies, to generate a near-globe digital elevation model (DEM) of Earth using radar interferometry. The data set used in this case is SRTM Non-Void Filled elevation data, which was processed from raw C-band radar signals spaced at intervals of 1 arc-second (approximately 30 meters) at NASA's Jet Propulsion Laboratory (JPL). This version was then edited or finished by the NGA to delineate and flatten water bodies, better define coastlines, remove spikes and wells, and fill small voids.

Using tools from 3D Analyst feature in ArcGIS, one single raster image was created from the seven separate SRTMs. The spatial resolution is 3601×3601 , the referenced horizontal datum is the World Geodetic System 84 (WGS84), and the vertical datum is the Earth Gravitational Model 1996 (EGM96) ellipsoid. The resulting image is then processed in Matlab and the resulting DEM of Crete is shown in Fig. 3.2.

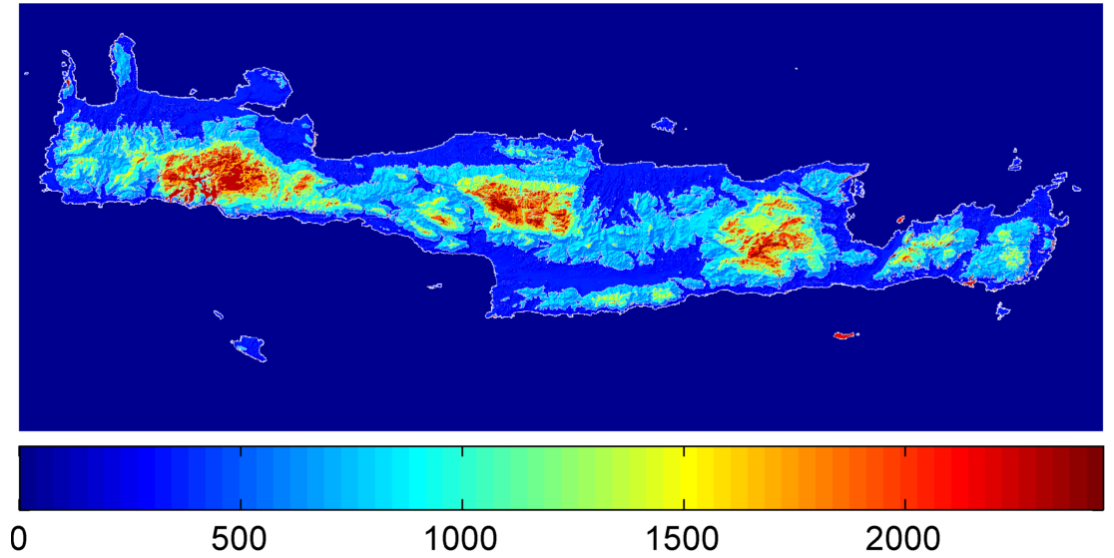


Figure 3.2: Digital Elevation Model of Crete

3.1.4 Precipitation

The average annual rainfall increases from the east to the west and from the south to the north. The mean precipitation based on all stations in each prefecture over the study period is 723 mm for Lasithi, 700 mm for Heraklion, 990 mm for

Rethimno and 1 279 mm for Chania.

The lowest annual precipitation is recorded at Demati station (196 mm) in Heraklion, while in stations placed on mountainous areas show significant variations. At Askufou station the mean annual precipitation is 3 146 mm, the highest recorded for the study period. Crete presents significantly variability of annual precipitation geographically (east versus west) and physiographically (lowland versus mountainous areas), displaying increasing precipitation with altitude (i.e., positive lapse rate) which ranks among the largest in Greece. Monthly precipitation peaks in December or January and attains a minimum in July and August which are almost drought months across the low-lying areas of Crete [Region of Crete Information Bull., 2002].

3.2 Precipitation Records

The data available for this study are monthly measurements of precipitation from fifty four (54) rain gauges distributed around Crete shown in Fig. 3.3. More specifically, six (6) of the stations are located in the prefecture of Chania, nine (9) in Rethimno, twenty eight (28) in Heraklion and eleven (11) in Lasithi. The study spans the time period from 1948 to 2012.

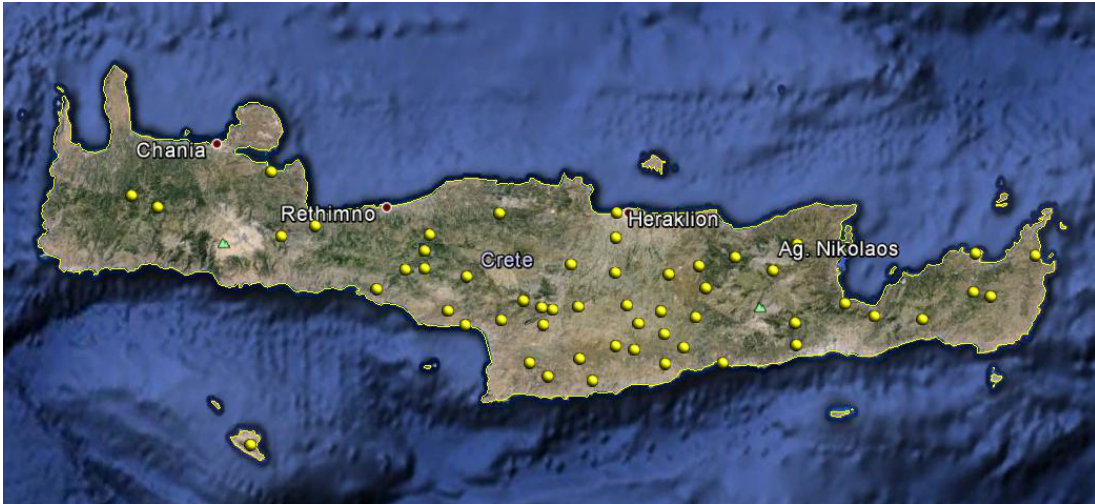


Figure 3.3: Geomorphological map of Crete showing the locations of the 54 stations (yellow circles) with rain gauges used in this study [Google Earth, 2015].

Before exploring the data, it should be noted that the records are incomplete. Figure 3.4 shows for which years and stations there are missing data. As evidenced in Fig. 3.4 there is a significant number of stations with incomplete records. The presence of missing data emphasizes the need to use geostatistical methods in order to understand the variability of precipitation.

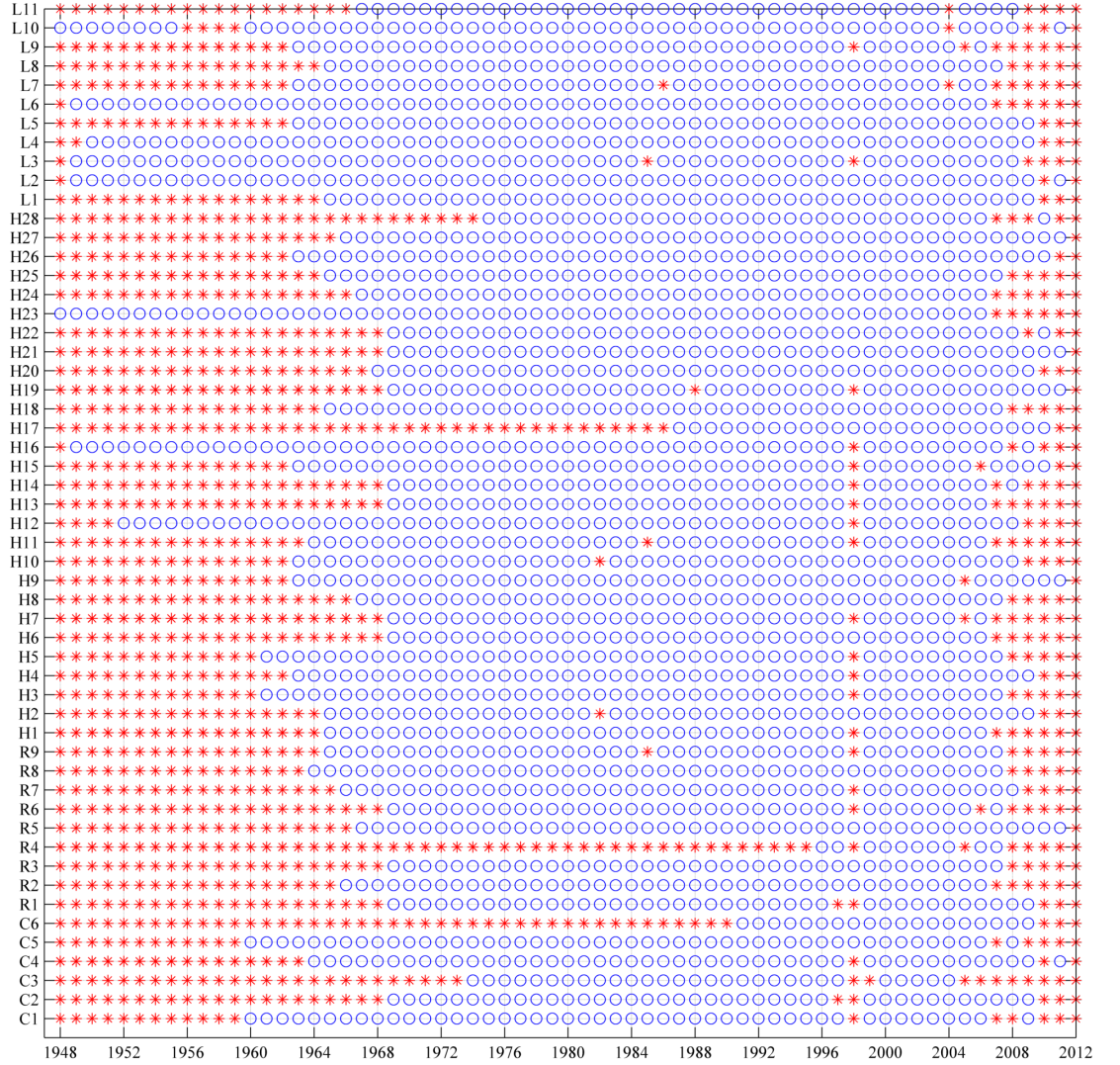


Figure 3.4: Matrix showing the missing data problem related to the precipitation records used. Rows correspond to stations while columns correspond to years. Circles (blue) mark locations with complete yearly records while with stars mark years that miss one or more monthly measurements.

3.3 Statistical Analysis of Monthly Data

The recorded precipitation as shown in Fig. 3.3 are sparse distributed over the domain, and also exhibit extreme dry condition during the summer period, with a lot of zero precipitation. This fact adds more complexity to the data analysis, making it difficult to conclude to a set of data appropriate for further investigation. Therefore, the recorded data are analyzed extensively and transformed into different data sets. The transformations used are chosen carefully, firstly to preserve a physical meaning, i.e., data sets of dry and wet periods, but also able to give a good fit to a model probability distribution function.

In order to fit the data to specific probability distribution models, distributions, such as the gamma distribution, empty records are removed in preliminary data processing. In addition, since several probability models do not allow zero values, such records are replaced by the machine resolution which is a very small real number (e.g., 2.2204^{-16}).

Based on a review of the literature, the gamma distribution is widely used to fit precipitation data [Baxevasi and Lennatsson, 2015; McKee et al., 1993; Vrochidou, 2013]. Hence, in the exploratory data analysis we fit monthly data with known distributions at every station, including the gamma distribution. Fits of the empirical cumulative probability distributions (cdfs) to respective model cdfs using the method of maximum likelihood result in unsatisfactory agreement between the empirical and the model cdf as shown in Fig. 3.6. This is mostly due to the presence of many months (especially in the summer) with zero precipitation values (Fig. 3.5).

To overcome this problem, the rainfall data sets are accordingly separated into wet and dry periods. The series representing the wet periods, consists of the monthly observations from October till March for all the years in the study period. Similarly, the dry period series comprises the measurements from April till September. Fitting wet and dry periods with model distributions gives good results for some of the stations (Fig. 3.7) but not for all of them (Fig. 3.8). More specifically, the wet period data set was fitted reasonably well by the Generalized Extreme Value distribution. On the other hand, the dry period data do not lead to good fits, because of the drought conditions observed in Crete during these

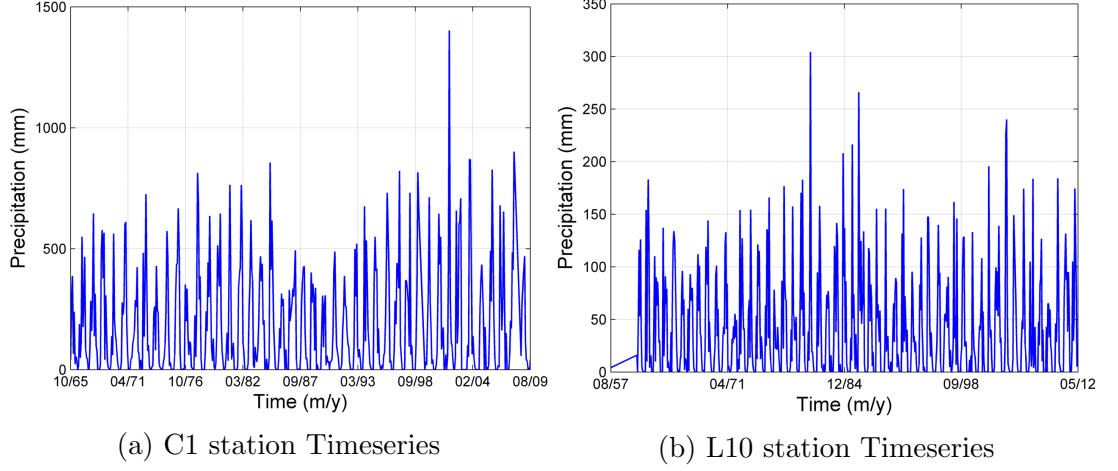


Figure 3.5: Left: Monthly precipitation series at the station C1 which exhibits the maximum mean monthly precipitation. Right: Monthly precipitation series at the station L10 which has recorded the minimum mean monthly precipitation.

months.

3.4 Precipitation Fields and Series

Average monthly precipitation fields from 1948 until 2012 are generated by aggregating the average values of the monthly precipitation measurements of each station for every year. This results in 65 average monthly precipitation fields (65 years) consisting of maximum 54 values (54 stations) for the entire island. Every field consists of the average monthly precipitation value for the particular year at every station available, and it is referred to the **average monthly precipitation**. We can also consider the distribution of the average monthly precipitation field to estimate the statistical fluctuations over the study area for specific years. For example, we could focus to the precipitation mean values for the year 1950 (which is the year with the minimum average monthly precipitation) and for the year 2002 (which is the year with the maximum average monthly precipitation). This data allows us to fit model distributions for every year. Representative results are presented in Fig. 3.10.

The average monthly precipitation for a given location S_j and a specific year

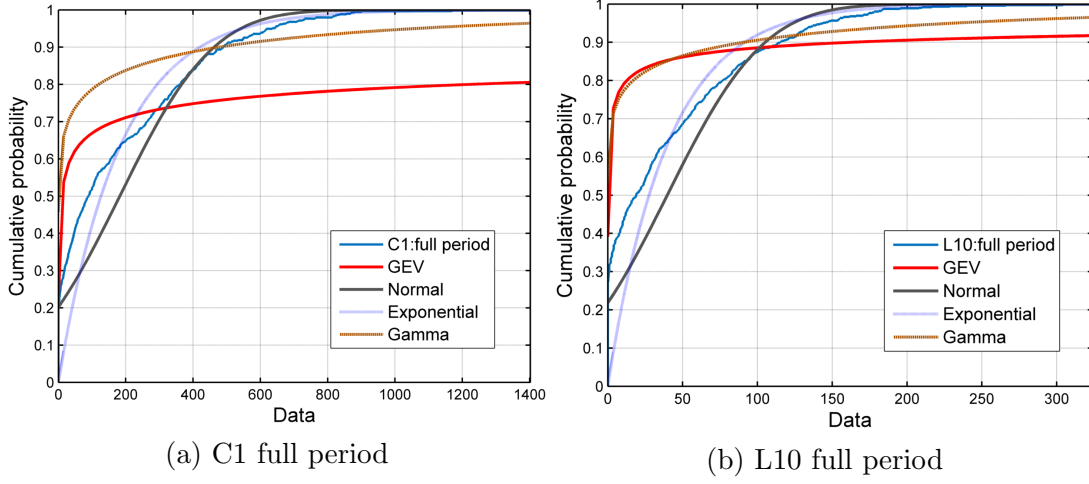


Figure 3.6: Empirical and model cumulative probability functions for the two precipitation time series shown in Fig. 3.5 C1 and L10 correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).

T is defined as

$$\bar{P}(S_j, T) = \frac{1}{m} \sum_{i=1}^{12} (I(S_j, t_i) \cdot P(S_j, t_i)) , \quad (3.1)$$

where T is the specific year, S_j is the station name, $P(S_j, t_i)$ are monthly precipitation levels, $I(S_j, t_i)$ is an indicator with value 1 for available monthly record and 0 otherwise, t_i is the specific month with values ranging between 0 and 12, and m is the number of the recorded months for the particular year ¹. If this value is calculated for every station for 2002, then the field of the average monthly precipitation of the year 2002 is produced. More specifically, in Fig. 3.9a the values of average monthly precipitation for every available station for the year 2002 are shown with the height of the stem.

The average monthly precipitation field for the year T consists of 0 to 54

¹The calculation is conducted with the matlab nanmean command which ignores the missing months

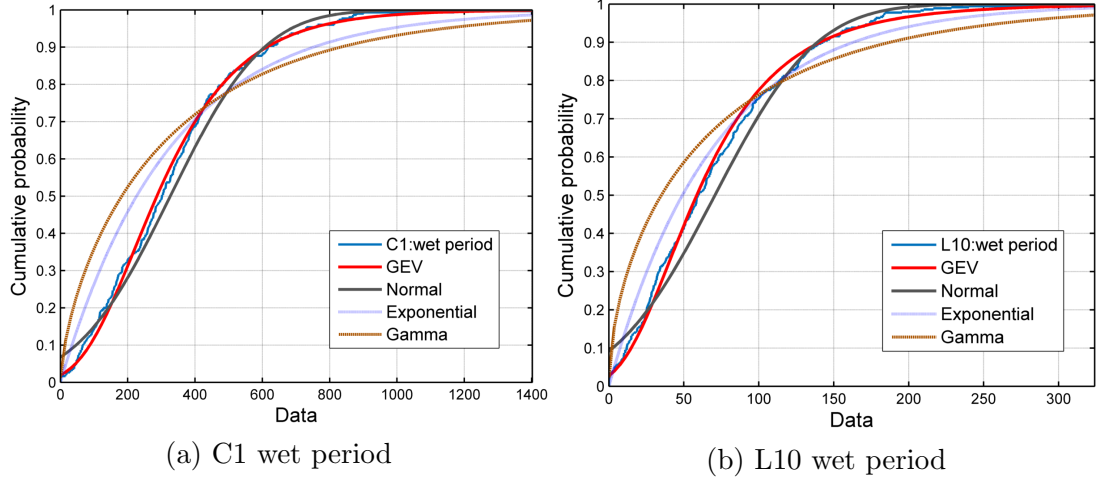


Figure 3.7: Empirical and model cumulative probability functions for the wet period data (consists of monthly observations from October till March) of the two precipitation time series shown in Fig. 3.5 C1 and L10 correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).

values, depending on how many stations are available. For example, for year T :

$$AMP(T) = \left(\bar{P}(S_{i_1}, T), \bar{P}(S_{i_2}, T), \dots, \bar{P}(S_{i_T}, T) \right), \quad (3.2)$$

where $AMP(T)$ is the average monthly precipitation field in constant time, T is a specific year, $\bar{P}(S_i, T)$ are the results of Eq. (3.1). S_1, S_2, \dots, S_N correspond to stations, N is the size of the stations network ($N=54$), and $1 \leq i_1 \leq 54 \dots 1 \leq i_T \leq 54$. Note that $AMP(T)$ vector’s length is not necessarily equal to 54, because there are not always available records for all the stations. If the values of the stems of Fig. 3.9a are gathered together, the average monthly precipitation field for year 2002 is created. Examples of the cdf fits of average monthly precipitation fields for years 1950 and 2002 are shown in Fig. 3.10. Based on the maximum likelihood, the best probability distribution model is the GEV distribution, shown with the red line. It is obvious from the figure that for year 1950 the model cumulative distribution function has higher variability and not so good fit as for year 2002.

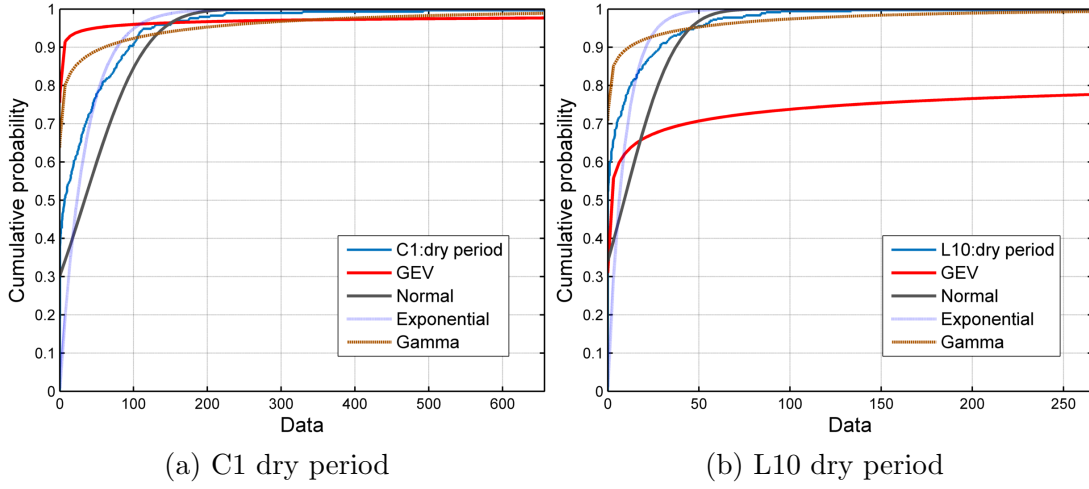


Figure 3.8: Empirical and model cumulative probability functions for the dry period data (consists of monthly observations from April till September) of the two precipitation time series shown in Fig. 3.5 C1 and L10 correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).

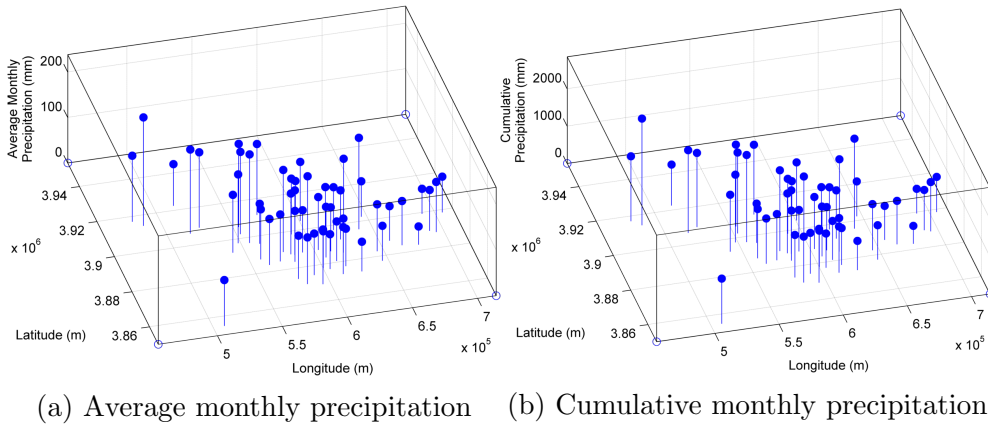


Figure 3.9: Three-dimensional stem plots representing the average monthly precipitation field (left) and the annual precipitation field (right) for the year 2002. Each stem refers to the average monthly precipitation at every station based on the Eq. (3.1) (left) and to the cumulative monthly precipitation based on the Eq. (3.3) (right) for the year 2002.

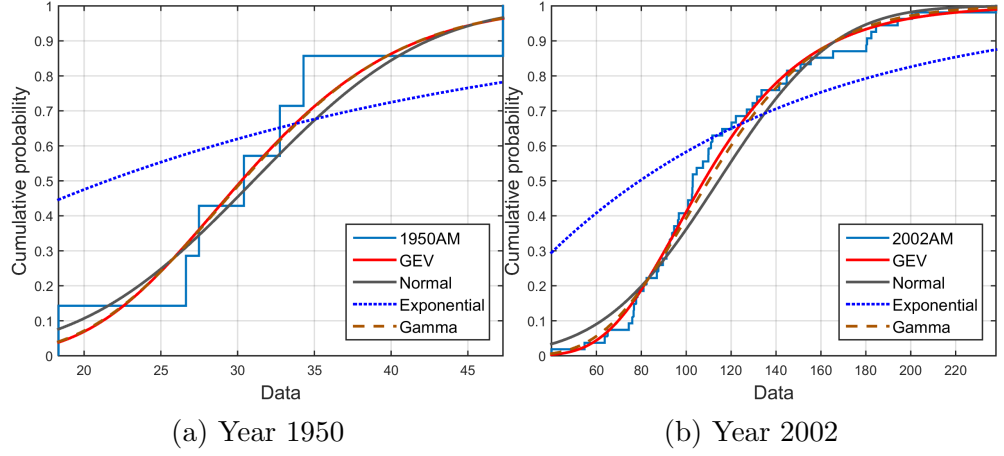


Figure 3.10: Empirical cumulative distribution function (cdf) and fits to model cdfs for the average monthly precipitation. The empirical cdf is given by the staircase (blue line), the Generalized Extreme Value distribution is given by the red line, the normal distribution is given by the dark green line, the exponential distribution by the cyan line, and the Gamma distribution by the brown line.

This is mostly because of the small data set available for this year. Even though the data set is small, the GEV distribution gives the best fit.

In addition, annual precipitation fields from 1948 until 2012 are generated by aggregating the cumulative monthly values. This results in 65 annual precipitation fields, one for every year, with maximum 54 values for the entire island. Every field consists of the cumulative monthly precipitation for the particular year at every station available, and it is referred to the **annual precipitation**. We can also consider the distribution of the annual precipitation field for a specific year to estimate the statistical fluctuations over the study area. For example, we could focus to the precipitation mean values for the year 1950 and for the year 2002 for the reasons explained above. This data allows us to fit model distributions for every year. Representative results are presented in Fig. 3.11.

The cumulative monthly precipitation at station S_j and for year T is defined as

$$CP(S_j, T) = \sum_{i=1}^{12} (I(S_j, t_i) \cdot P(S_j, t_i)), \quad (3.3)$$

where T , S_j , t_i and $P(S_j, t_i)$ are as above, and $I(S_j, t_i)$ is an indicator with value 1

if all $P(S_j, t_i)$ monthly records are available for this year and 0 otherwise ¹. Figure 3.9b shows the values of cumulative monthly precipitation for every available station with the height of the stem.

The data sets of cumulative monthly precipitation field, henceforth referred as **annual precipitation**, for the year T comprises between zero and 54 values, depending on how many stations have available data in this year. For example, for year T :

$$AP(T) = (CP(S_{i_1}, T), CP(S_{i_2}, T), \dots, CP(S_{i_N}, T)), \quad (3.4)$$

where $AP(T)$ is the annual precipitation field in constant time, T is the specific year, $CP(S_j, T)$ are the results of Eq. (3.3). S_1, S_2, \dots, S_N correspond to stations, N is the size of the stations network ($N=54$), and $1 \leq i_1 \leq 54, \dots, 1 \leq i_N \leq 54$. Note that $AP(T)$ vector's length is not necessarily equal to 54, because there are not always available records for all the stations. If we focus in Fig. 3.4, vertically, for every circle (blue mark) we have one value of $CP(T)$. The summation of the circles, gives the length of the vector for the estimated year. For example, for year 2000, we have 54 values to generate the annual precipitation field. Focusing in Fig. 3.9b, gathering the values of the stems together, the annual precipitation field for year 2002 is created. Examples of the cdf fits of those data sets are shown in Fig. 3.11. Based on the maximum likelihood, the best probability distribution model is the GEV distribution, shown with the red line.

Next we construct time series of precipitation at specific locations. The average monthly precipitation time series for the station S consists of at least 1 to 65 values, depending on how many years are available. Specifically, for station S :

$$AMP(S) = (\bar{P}(S, T_{j_1}), \bar{P}(S, T_{j_2}), \dots, \bar{P}(S, T_{j_S})), \quad (3.5)$$

where $AMP(S)$ is the average monthly precipitation time series at specified location, S is the specific station, $\bar{P}(S, T_i)$ are the results of Eq. (3.1) for the year S which have measurements at this station, and $1 \leq j_1 < j_2 < \dots < j_S \leq 65$. $AMP(S)$ vector's length is not necessarily equal to 65, because there are not

¹The calculation is conducted with the sum command which leaves out the years that have missing months

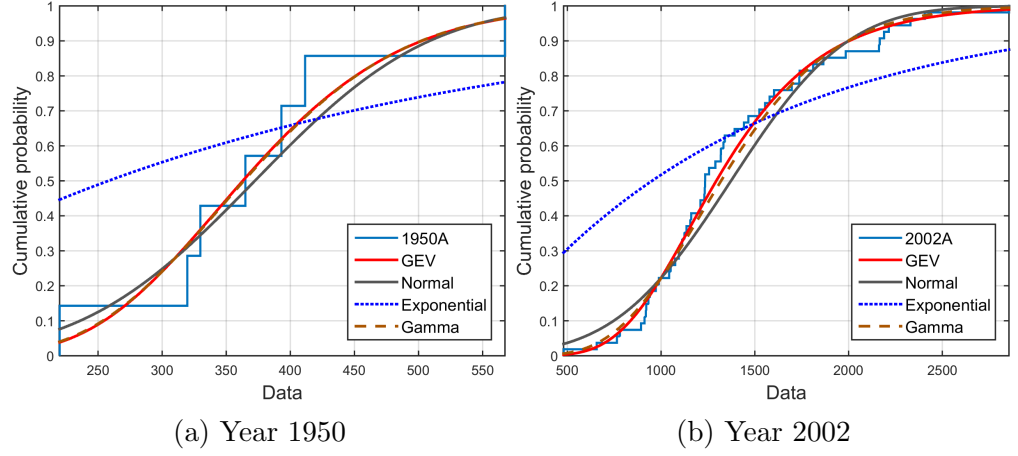


Figure 3.11: Empirical cumulative distribution function (cdf) and fits to model cdfs for the annual precipitation. The empirical cdf is given by the staircase (blue line), the Generalized Extreme Value distribution is given by the red line, the normal distribution is given by the dark green line, the exponential distribution by the cyan line, and the Gamma distribution by the brown line.

always available records for all the previous years. If we focus again in Fig. 3.4, along the rows, every circle (blue mark) corresponds to one value of $CP(S)$. The summation of the circles, gives the length of the vector for the estimated year. For example, for station H23, we have 59 values to generate the annual precipitation time series. For example, for station S :

$$AP(S) = (CP(S, T_{j_1}), CP(S, T_{j_2}), \dots, CP(S, T_{j_S})) , \quad (3.6)$$

where $AP(S)$ is the annual precipitation time series at specified location, S is the specific station, $CP(S, T_i)$ are the results of Eq. (3.3) at year S which have measurements at this station, and $1 \leq j_1 < j_2 < \dots < j_S \leq 65$. $AP(S)$ vector's length is not necessarily equal to 65, because there are not always available records for all the previous years.

The analysis of the time series, gives similar results with the field data sets: the model cumulative distribution function that fits best the time series is the GEV distribution. The optimal distribution, GEV, and its parameters are identified by applying the maximum likelihood estimation method, and they are presented in Table 3.3 for the average monthly precipitation time series data sets (based

on Eq. (3.5)) and in Table 3.4 for the annual precipitation time series data sets (based on Eq. (3.6)).

Table 3.3: Location coordinates, optimal GEV distribution type and parameters for all the stations on the island, calculated with the average monthly precipitation time series data set (based on Eq. (3.5)). The coordinates x, y, z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y being northing and easting measured in meters, and z being the elevation measured in meters above sea level. The GEV distribution and its parameters are defined in Eq. (1.17). “Weibull” refers to the Reversed Weibull distribution, that belongs to the Generalized Extreme Value distribution family.

Latitude	Longitude	Elevation	k	σ	μ	Distr. Type
CHANIA						
517156	3906520	740	−0.06	42.24	164.18	Weibull
508459	3854769	10	−0.01	11.55	44.48	Weibull
514927	3922549	20	−0.10	12.90	54.24	Weibull
525676	3908980	50	−0.14	21.44	83.42	Weibull
479746	3917260	316	−0.13	23.87	98.52	Weibull
486286	3914250	520	0.14	25.98	128.49	Fréchet
RETHIMNO						
562696	3883780	20	−0.01	11.55	44.48	Weibull
552861	3902359	260	−0.01	20.40	89.11	Weibull
563266	3895810	310	−0.11	14.33	56.59	Weibull
571668	3911363	260	−0.60	40.89	92.56	Weibull
556336	3897010	660	−0.16	24.35	99.21	Weibull
554179	3906323	580	0.01	19.24	72.80	Fréchet
540774	3893055	90	0.15	16.71	59.28	Fréchet
558436	3887380	560	0.20	12.10	56.24	Fréchet
547996	3897760	390	−0.36	31.01	83.40	Weibull
HRAKLEIO						
630233	3899448	230	−0.05	17.46	63.35	Weibull
590910	3887874	570	−0.27	15.32	71.86	Weibull
583036	3870700	450	−0.03	10.96	39.20	Weibull
622750	3891949	450	−0.14	16.66	59.97	Weibull

600049	3877943	200	−0.03	10.74	44.07	Weibull
612352	3873325	680	−0.16	13.56	52.05	Weibull
591040	3874936	190	0.07	10.94	38.02	Fréchet
613595	3895526	330	0.02	14.95	56.03	Fréchet
577228	3889597	520	−0.02	21.80	82.87	Weibull
584536	3887230	450	0.11	16.78	64.42	Fréchet
617026	3877270	210	0.04	15.35	35.19	Fréchet
581799	3887788	500	0.02	15.68	62.12	Fréchet
612261	3880719	200	0.10	10.44	42.21	Fréchet
594138	3869421	800	0.04	14.13	50.42	Fréchet
620116	3884740	320	−0.22	11.99	43.30	Weibull
621158	3897473	350	−0.07	14.57	55.78	Weibull
626716	3873280	10.	0.01	12.24	34.61	Fréchet
589276	3898390	500	−0.14	17.40	72.29	Weibull
571666	3884860	140	−0.13	13.81	41.26	Weibull
603058	3888005	430	0.00	11.12	54.76	Fréchet
582134	3883486	400	−0.08	13.75	51.40	Weibull
611476	3886360	400	−0.29	16.64	48.53	Weibull
578536	3874150	150	0.09	11.23	37.00	Fréchet
604705	3876931	225	0.06	9.65	40.77	Fréchet
600316	3896140	380	0.13	17.00	59.65	Fréchet
605866	3883420	360	−0.15	15.43	51.56	Weibull
600526	3904810	40	−0.17	19.09	54.63	Weibull
600886	3910990	15	−0.20	8.38	37.80	Weibull
LASITHI						
694096	3888550	480	0.14	14.43	62.44	Fréchet
645796	3902380	240	−0.15	13.90	61.90	Weibull
639586	3895960	840	−0.22	30.36	106.86	Weibull
657492	3887454	20	−0.07	8.64	38.19	Weibull
644896	3882850	590	0.13	15.22	54.91	Fréchet
689776	3889690	150	−0.04	11.06	48.76	Weibull

645166	3877390	200	−0.37	15.79	40.97	Weibull
705496	3898540	25	−0.11	11.41	38.04	Weibull
664846	3884170	50	0.10	13.86	42.14	Fréchet
690436	3899320	114	−0.08	9.91	34.67	Weibull
676876	3883150	325	−0.05	20.16	54.92	Weibull

Table 3.4: Location coordinates, optimal GEV distribution type and parameters for all the stations on the island, calculated with the annual precipitation time series data set (based on Eq. (3.6)). The coordinates x , y , z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y being northing and easting measured in meters, and z being the elevation measured in meters above sea level. The GEV distribution and its parameters are defined in Eq. (1.17). “Weibull” refers to the Reversed Weibull distribution, that belongs to the Generalized Extreme Value distribution family.

Latitude	Longitude	Elevation	k	σ	μ	Distr. Type
CHANIA						
517156	3906520	740	−0.37	479.69	2003.56	Weibull
508459	3854769	10	0.04	128.43	518.52	Fréchet
514927	3922549	20	0.09	135.25	617.16	Fréchet
525676	3908980	50	−0.12	247.90	998.53	Weibull
479746	3917260	316	−0.13	289.29	1181.73	Weibull
486286	3914250	520	0.08	327.65	1575.42	Fréchet
RETHIMNO						
562696	3883780	20	0.04	128.43	518.52	Fréchet
552861	3902359	260	−0.05	257.20	1085.55	Weibull
563266	3895810	310	−0.11	174.37	680.48	Weibull
571668	3911363	260	−0.38	316.23	1225.18	Weibull
556336	3897010	660	−0.16	292.18	1190.56	Weibull
554179	3906323	580	−0.01	221.97	859.72	Weibull
540774	3893055	90	−0.05	183.86	709.37	Weibull
558436	3887380	560	0.10	142.20	674.99	Fréchet

547996	3897760	390	−0.09	220.53	1030.68	Weibull
HRAKLEIO						
630233	3899448	230	−0.02	206.82	742.58	Weibull
590910	3887874	570	−0.28	188.52	862.30	Weibull
583036	3870700	450	−0.02	129.75	462.50	Weibull
622750	3891949	450	−0.07	169.69	724.75	Weibull
600049	3877943	200	0.00	129.39	522.43	Fréchet
612352	3873325	680	−0.12	145.83	638.97	Weibull
591040	3874936	190	0.11	127.79	444.80	Fréchet
613595	3895526	330	−0.08	171.41	677.94	Weibull
577228	3889597	520	−0.02	260.10	998.18	Weibull
584536	3887230	450	−0.04	184.44	775.45	Weibull
617026	3877270	210	−0.09	142.60	420.80	Weibull
581799	3887788	500	−0.04	177.18	736.91	Weibull
612261	3880719	200	−0.03	126.78	518.20	Weibull
594138	3869421	800	0.00	164.07	600.35	Fréchet
620116	3884740	320	−0.17	132.43	530.57	Weibull
621158	3897473	350	−0.20	165.08	665.85	Weibull
626716	3873280	10	−0.13	138.43	413.79	Weibull
589276	3898390	500	−0.10	198.62	878.17	Weibull
571666	3884860	140	0.10	117.51	502.35	Fréchet
603058	3888005	430	0.00	136.38	656.62	Fréchet
582134	3883486	400	−0.07	162.48	611.59	Weibull
611476	3886360	400	−0.06	134.79	589.88	Weibull
578536	3874150	150	−0.05	121.72	439.90	Weibull
604705	3876931	225	0.02	107.23	480.16	Fréchet
600316	3896140	380	−0.06	176.62	700.59	Weibull
605866	3883420	360	−0.09	160.01	618.70	Weibull
600526	3904810	40	−0.07	185.50	669.27	Weibull
600886	3910990	15	−0.20	100.52	453.63	Weibull
LASITHI						

694096	3888550	480	0.09	164.19	742.30	Fréchet
645796	3902380	240	−0.14	167.15	740.95	Weibull
639586	3895960	840	−0.19	350.46	1286.88	Weibull
657492	3887454	20	−0.08	106.02	460.25	Weibull
644896	3882850	590	−0.02	177.46	663.75	Weibull
689776	3889690	150	−0.12	131.17	586.13	Weibull
645166	3877390	200	−0.24	150.42	515.18	Weibull
705496	3898540	25	−0.11	135.43	455.56	Weibull
664846	3884170	50	−0.08	150.40	498.48	Weibull
690436	3899320	114	−0.08	119.66	418.62	Weibull
676876	3883150	325	−0.14	171.06	699.29	Weibull

By the above investigation and taking into account the statistical properties, in both cases, average monthly precipitation time series and annual precipitation time series, Generalized Extreme Value distribution (GEV) is the more representative distribution. For visual evidence of the goodness of the GEV fit, we present below the cdf plots of two stations for both average monthly and annual precipitation time series (Figs. 3.12, 3.13). Although, Gamma distribution function seems to have a good fit here, the GEV distribution fits better the time series, with higher differences calculated with the maximum likelihood between them in cases of smaller data sets. Through the above investigation, we can say that the best model distribution function to fit the data sets, for both spatial and temporal scale is the GEV distribution function.

3.5 Precipitation Comparison in the Eastern and Western Regions

To further study the spatial patterns of precipitation and to investigate the role of longitude on recorded precipitation values on Crete, we divide the island into two different geographical regions, i.e., into Western and Eastern Crete. As mentioned before, the available precipitation data are measurements from 54 meteorologi-

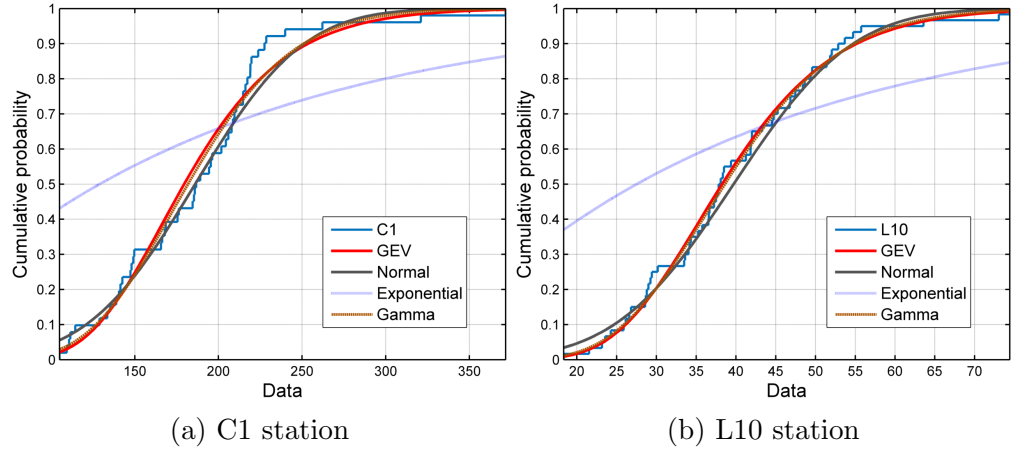


Figure 3.12: Empirical and model cumulative probability functions for the average monthly precipitation time series of station C1 (left) and L10 (right) correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).

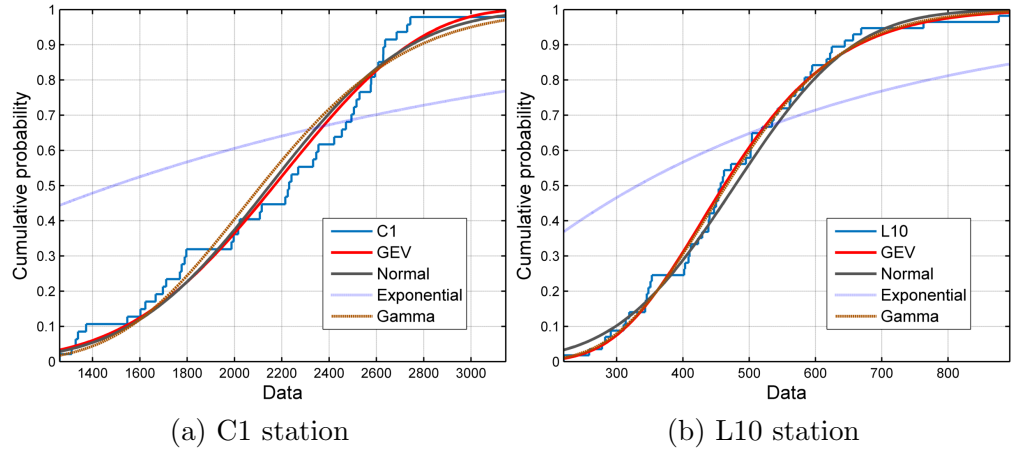


Figure 3.13: Empirical and model cumulative probability functions for the annual precipitation time series of station C1 (left) and L10 (right) correspond to the empirical cdfs (blue line). The model fits include the following: “GEV” refers to the generalized extreme value distribution (red line), “Normal” refers to the the Gaussian distribution (green line), “Exponential” refers to the exponential distribution (cyan line), and “Gamma” refers to the gamma distribution (brown line).

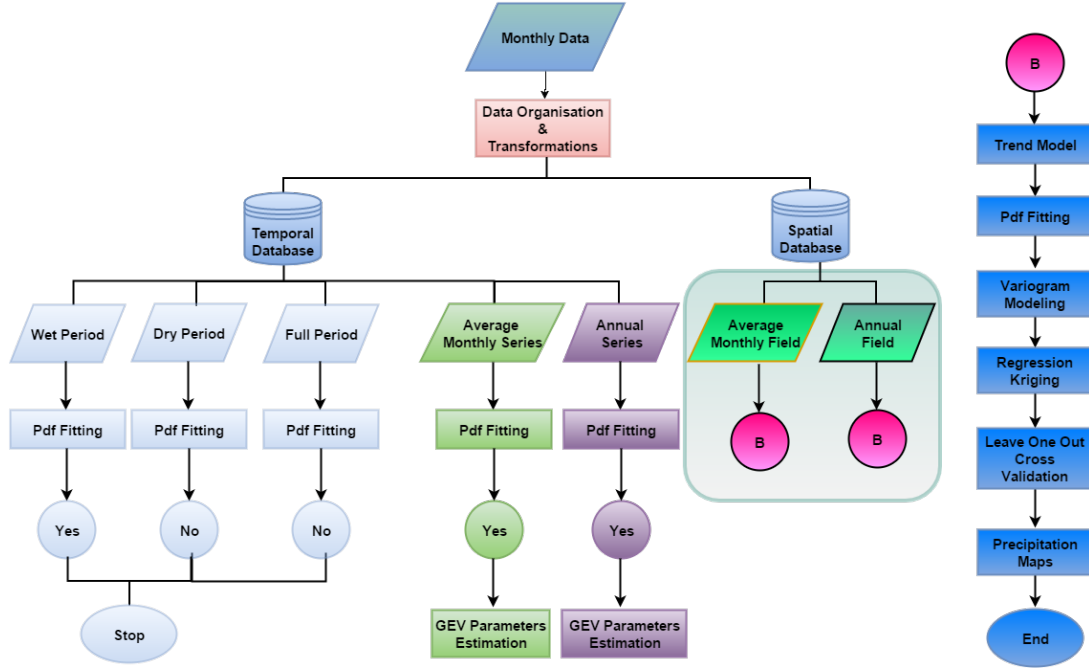


Figure 3.14: Flow chart of the preliminary data processing and geostatistical methodology applied to the spatial database. Data Organisation & Transformations refers to the preliminary investigation of the needed transformations in order to keep the physical meaning of the data, but also to find the appropriate data set . “GEV” refers to the generalized extreme value distribution

cal stations around Crete, 6 of the meteorological stations are located in the prefecture of Chania, 9 in Rethimno, 28 in Heraklion and 11 in Lasithi.

More specifically, West Crete consists of fifteen (15) stations and for sixty five (65) years there should be eleven thousand and seven hundred records (11 700). Due to lack of recorded measurements during the first twelve years there exist recorded data only for fifty three (53) years in total that include at most nine thousand five hundred and forty measurements for the 15 stations (9 540). From those 9 540 measurements the 2 400 are missing (25%). While, East Crete consists of thirty nine (39) stations and for 65 years there should be thirty thousand and four hundred and twenty measurements (30 420), from those eight thousand and three hundred eighty the 8 380 are missing (27%).

To compare the Eastern and Western regions of Crete, we need a representative data set for each. We create the average monthly precipitation time series

for West and for East by gathering one representative value for every region for every year. Below we present the methodology used to create the time series for the West region. The same methodology can be applied to the East.

The averaged monthly precipitation time series of West Crete are generated by aggregating the mean values of the average monthly precipitation of the 15 stations located in the prefectures of Chania and Rethimno for every year (for the East we take into account the 39 stations located in the prefectures of Heraklion and Lasithi). The average monthly precipitation for West Crete is calculated as follows:

$$\overline{PW}(T) = \left(\frac{1}{m} \sum_{i=1}^{15} \left(I(S_i, T) \cdot \overline{P}(S_i, T) \right) \right), \quad (3.7)$$

where T is the specific year, S_i is the station name, $\overline{P}(S_i, T)$ are the average monthly precipitation calculated from Eq. (3.1), $I(S_i, T)$ is an indicator with value 1 if there is a record of average monthly precipitation for this year at this station, and $m \leq 15$ is the sum of the available stations in the West for the particular year.

If this value is calculated for every year, then the time series of average monthly precipitation for West Crete is produced. The average monthly precipitation time series for West Crete begins in 1960, because there are no available measurements before this year. The formula that gives the average monthly precipitation time series for West Crete is

$$AMPW = \left(\overline{PW}(T_{j_1}), \overline{PW}(T_{j_2}), \dots, \overline{PW}(T_{j_S}) \right), \quad (3.8)$$

where $\overline{PW}(T_j)$ are the results of Eq. (3.7) for year T_j , where $1 \leq j_1 < j_2 < \dots < j_S \leq 65$.

For the generation of the annual precipitation time series for West Crete we firstly calculate the mean cumulative monthly precipitation for West Crete for every year as

$$\overline{CPW}(T) = \left(\frac{1}{m} \sum_{i=1}^{15} \left(I(S_i, T) \cdot CP(S_i, T) \right) \right), \quad (3.9)$$

where T , S_i , m , and $I(S_i, T)$ as above, and $CP(S_i, T)$ are the cumulative monthly

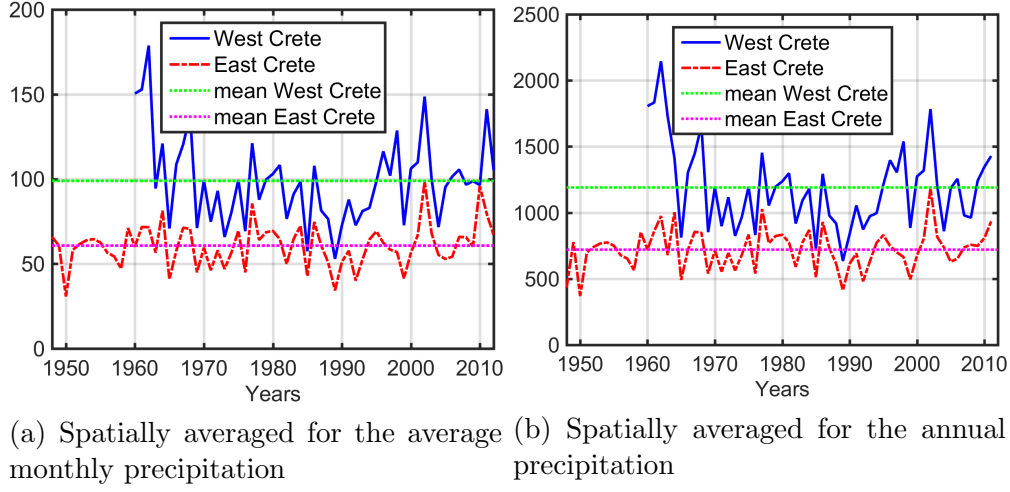


Figure 3.15: Time series of average monthly precipitation (left) and annual precipitation (right) for the Western (blue line) and the Eastern (red line) regions of Crete based on Eqs. (3.7) and (3.9) respectively. The sample mean of the precipitation in the West (East) is indicated by the green (magenta) straight line in the diagrams.

precipitation calculated from Eq. (3.3).

If this value is calculated for every year, then the time series of average monthly precipitation for West Crete is produced. The average monthly precipitation time series for West Crete begins from 1960, because there were not available measurements before that year. The formula that gives the average monthly precipitation time series for West Crete is

$$APW = \left(\overline{CPW}(T_{j_1}), \overline{CPW}(T_{j_2}), \dots, \overline{CPW}(T_{j_s}) \right), \quad (3.10)$$

where $\overline{CPW}(T_j)$ are the results of Eq. (3.9) for year T_j .

Figure 3.15a and Fig. 3.15b reveal strong correlations in precipitation between the western and eastern parts of the island. A spatial trend is observed with the spatially averaged data calculated with the average monthly precipitation time series in the West measured at about 40 mm than in the East (Eq. (3.7) for the West, accordingly we proceed with the calculation of the East). For the spatially averaged data with the annual precipitation time series, the spatial correlation is at about 450 mm higher in the West than in the East (Eq. (3.9) for the West,

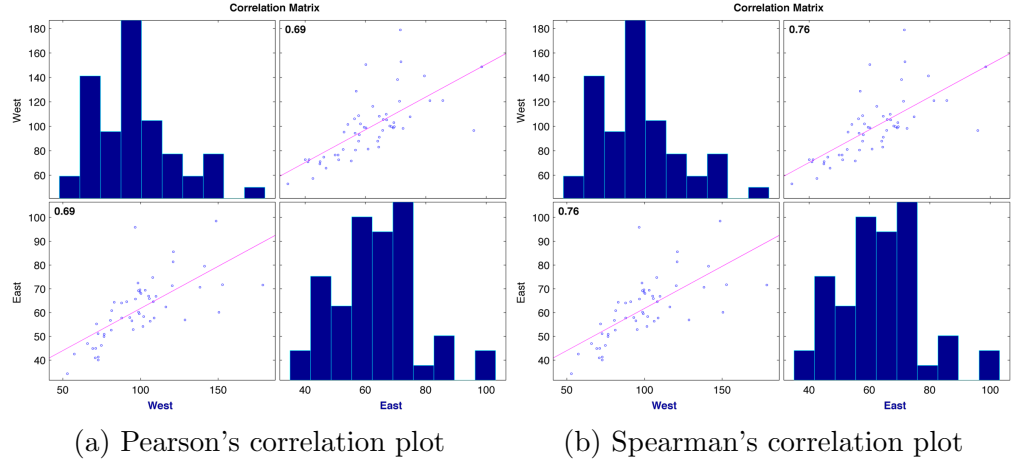


Figure 3.16: Correlation plots between Western and Eastern regions of Crete for the average monthly precipitation. On the left diagonal, the histograms show the probability densities of the Western and Eastern average monthly precipitation time series (Eq. (3.8)). On the right diagonal the scatterplots of the data are presented. The Pearson's correlation coefficient calculated by Eq. (2.77) equals to 0.69 (left) and Spearman's correlation coefficient based on Eq. (2.78) equals to 0.76 (right).

accordingly we proceed with the calculation for the East). In other words, if we have a mean value of precipitation for the one region, i.e., for West Crete, we are able to make a first assumption for East Crete, the value of which would be 40 mm lower in the case of the average monthly precipitation and lower at about 450 mm in the case of the annual precipitation.

We quantify the correlations shown in Fig. 3.15 by means of the correlation plots shown in the plots of Figs. 3.16 and 3.17. The estimated Pearson's correlation coefficient for the average monthly precipitation is 0.69 and the Spearman's is equal to 0.76. On the other hand, the correlation coefficients between West and East Crete for the annual precipitation is 0.71 for both Pearson's and Spearman's coefficients. We have lower dependence with the average monthly data because the data sets have bias from the beginning, as the average values are calculated by the available months. This mean that we could have stations with missing measurements in the summer, while for stations in the opposite region we could have missing values in the winter. If we focus to the annual precipitation, longitude seems to affects precipitation in Crete.

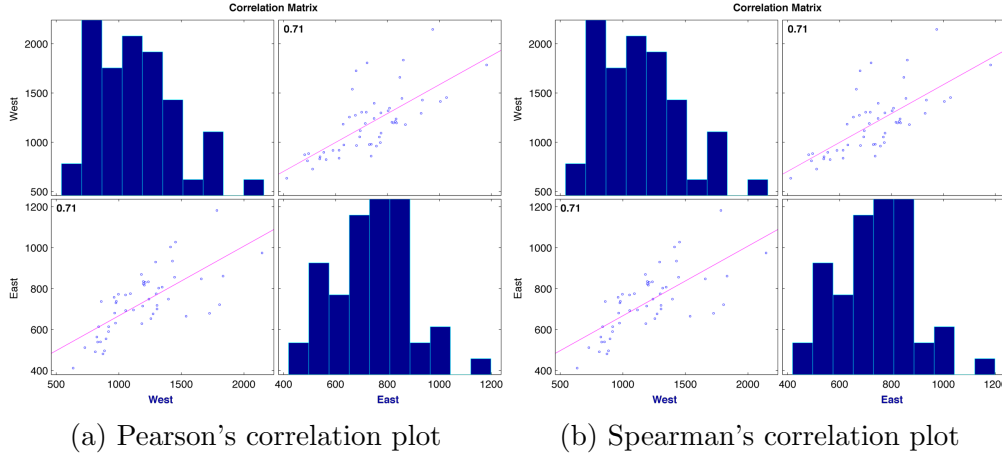


Figure 3.17: Correlation plots between Western and Eastern regions of Crete for the annual precipitation. On the left diagonal, the histograms show the probability densities of the Western and Eastern annual precipitation time series (Eq. (3.10)). On the right diagonal the scatterplots of the data are presented. The Pearson's correlation coefficient calculated by Eq. (2.77) equals to 0.71 (left) and Spearman's correlation coefficient based on Eq. (2.78) equals to 0.71 (right).

Chapter 4

Results of Geostatistical Analysis

In a study area that lacks homogeneity and has mountainous areas across, topography is an important factor for precipitation generating a noticeable trend. The analysis on both the average monthly and the annual precipitation fields gave similar results. Below, we will present in detail the results of the analysis for the annual precipitation field of the year 1971. Representative figures for the remaining years, from 1965 until 2011, are presented in [Appendix A](#). For comparison reasons, results for the average monthly precipitation field in 1971 are presented in [Section 4.2](#). In [Sections 4.1](#) and [4.2](#) we perform a local analysis of the precipitation field, while in [Section 4.3](#) we focus on a global analysis that refers to the entire island.

4.1 Model of Annual Precipitation Field

First, we remove the trend from the precipitation data. We accomplish this with the application of multilinear regression of the response (e.g., the precipitation) on the predictors, which contain longitude, latitude and elevation.

The residuals remaining after the application of the multilinear regression are shown in [Fig. 4.2](#) (in red). The plots in [Fig. 4.3](#) show how known probability density functions, including the normal distribution, fit the precipitation residuals. As shown in [Fig. 4.3b](#) the residuals can be considered as approximately Gaussian; hence, Ordinary Kriging can be used for interpolation.

4.1.1 Topographic Trend

In order to decide which of the topographic parameters contribute to precipitation, multilinear regression was performed using all the possible combinations of longitude, latitude and elevation as predictor variables. The results of the calculations are presented in Table 4.1. The coefficient of determination R^2 is low if only one or two of the parameters are used. Hence, we include all three variables (longitude, latitude, and elevation) in the trend function. The coefficient of determination is defined as

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \quad (4.1)$$

where SS_{res} is the sum of squares of residuals, also called the residual sum of squares, and SS_{tot} is the total sum of squares (proportional to the variance of the data), and has values among 0 and 1. More specifically, the estimates of the coefficients for the regression model of the annual precipitation field are presented in Table 4.2. The p-value is very low, thus rejecting the null hypothesis of zero correlation.

The most important topographic factor seems to be the elevation, which contributes most to the trend model. This conclusion is based on the value of the coefficient that multiplies the elevation in the trend function. Figure 4.1 shows a plot of the trend function and the data versus easting and elevation.

In Fig. 4.2 the annual precipitation field values are marked with blue circles, while the residual precipitation resulting after the removal of the trend is marked with red asterisks. The plots of the model probability density functions and the normal probability plot of the precipitation residuals are presented in Fig. 4.3. These plots show that the residuals are approximately normally distributed. As mentioned above, the Student's t distribution describes the average of samples from a full normal population. Hence, because of the small sample size (only 49 available stations), it is expected that the residual precipitation fit best the Student's t distribution, in agreement with the plot shown in Fig. 4.3a. Nonetheless, in the following it is assumed that the residuals are treated as normally distributed.

Summary statistics of the annual data set for the year 1971 are presented in Table 4.3. They include the mean value, the minimum, the maximum value, the

Table 4.1: Statistics of multilinear regression for all the possible combinations of longitude (x), latitude (y) and elevation (z), for the annual precipitation field in the year 1971 (Eq. (3.4)). The coordinates x, y, z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y representing the longitude and latitude measured in meters, and z being the elevation measured in meters above sea level. The value of the coefficient of determination R^2 shows how much of the variability in the response variable (precipitation) is explained by the model (e.g. for x, y, z : 74%). The F-statistic gives the relationship between the response variable and the predictor variables. The p-value is used in the context of null hypothesis testing of zero correlation in order to quantify the idea of statistical significance of evidence. The p-value refers to the F-statistic test, lower value are preferable.

Regression with	R^2	F-statistic	p-value	Error variance
x	0.2434	15.1196	3.1593e-04	5.1638e+04
y	0.3383	24.0259	1.1723e-05	4.5163e+04
z	0.3066	20.7799	3.6774e-05	4.7326e+04
x, y	0.5179	24.7092	5.1510e-08	3.3618e+04
x, z	0.4869	21.8277	2.1586e-07	3.5779e+04
y, z	0.6108	36.0980	3.7461e-10	2.7139e+04
x, y, z	0.7415	43.0295	2.8437e-13	1.8426e+04

Table 4.2: Trend model parameters, for annual precipitation field in the year 1971 (Eq. (3.4)).

b	$\times 1$	$\times x$	$\times y$	$\times z$
$b_i, i=0, \dots, 3$	-4.1919e+04	-0.0019	0.0112	0.5688
lower bound	-5.5151e+04	-0.0027	0.0078	0.3852
upper bound	-2.8686e+04	-0.0011	0.0146	0.7524

Table 4.3: Summary statistics for the precipitation data set depicted in Fig. 4.2. The parameter n is the number of available stations of the year.

1971	n	Mean	Median	Minimum	Maximum	Std.dev.	Coef.of var.	Skewness
Precipitation	49	634.8980	581.1000	291.2000	1602.4000	258.5111	0.4072	1.5316
Residuals	49	2.2065e-12	-21.9881	-339.2030	381.6012	131.4318	5.9567e+13	0.4557

standard deviation of the annual precipitation as well as the respective values for precipitation residuals. The zero mean of the residuals, the approximately zero

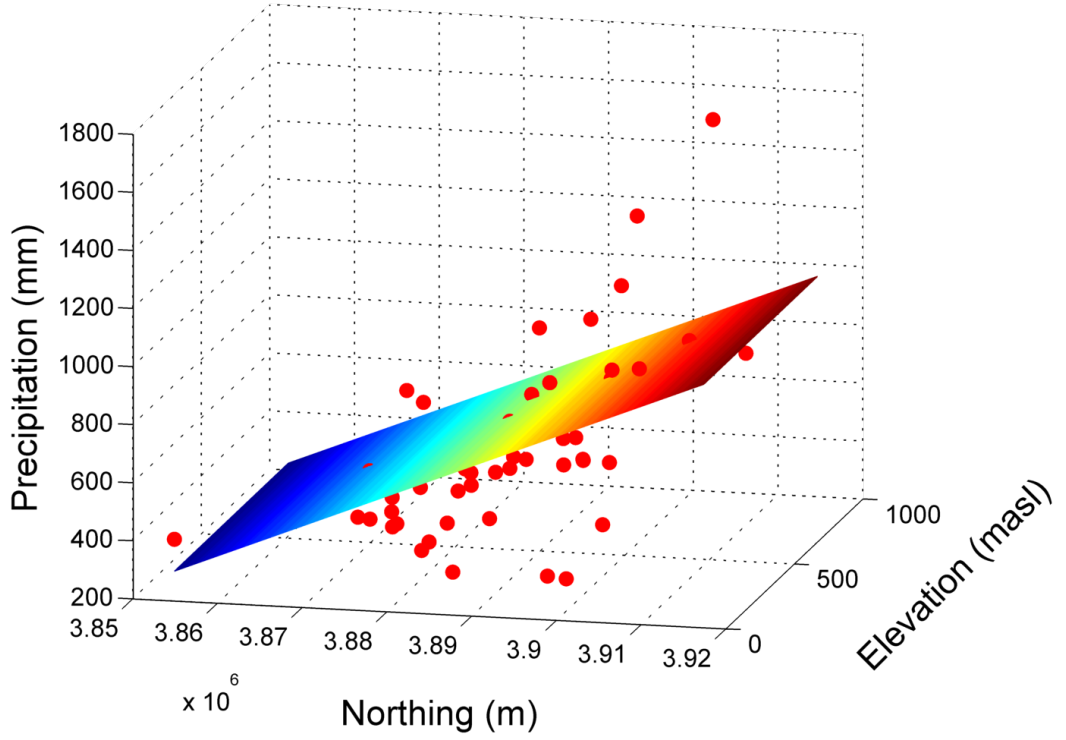


Figure 4.1: Multilinear regression of the precipitation field on the predictors which comprise longitude, latitude, and elevation, for the annual precipitation field in the year 1971. The trend equation for this year is given by $f = -4.1919 \cdot 10^4 - 0.0019x + 0.0012y + 0.5688z$. The coordinates x , y , z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y representing the easting and the northing coordinates measured in meters, and z being the elevation measured in meters above sea level.

skewness, and the proximity of the residuals to the straight line on the normal probability plot as shown in Fig. 4.3b prove that the residuals are close to the Gaussian distribution.

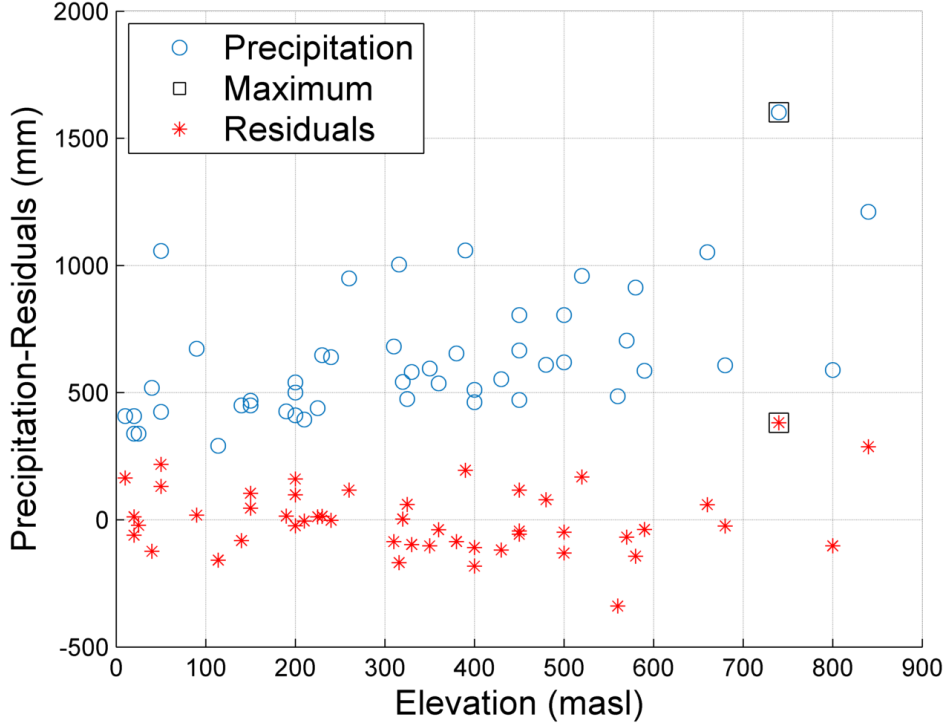


Figure 4.2: Scatter plot of the annual precipitation field values and the precipitation residuals obtained after the application of multilinear regression according to the trend equation shown in Fig. 4.1, for the year 1971. The blue circles correspond to the annual precipitation field values, while the red asterisks correspond to the precipitation residuals. The square boxes mark the highest values for the respective data set.

4.1.2 Variogram Modeling

To construct a spatial model of the residual precipitation, a suitable variogram model needs to be determined. After the calculation of the experimental variogram based on the method of moments, the Spartan variogram is fitted as a theoretical model. This allows estimating the parameters η_0 , η_1 , and the characteristic length ξ of the Spartan model (see Eq. (2.30)). Figure 4.4 shows the graphical representation of the variogram of the above data set. The scale parameter η_0 that determines the total variance of the fluctuations is equal to $1.9764e+05 \text{ mm}^2$, the shape parameter $\eta_1 = -0.8643$, and the correlation length ξ that shows the range of the spatial dependence equals 9.9620. The shape parameter η_1 is dimensionless. Particularly for this case, the correlation length ξ is

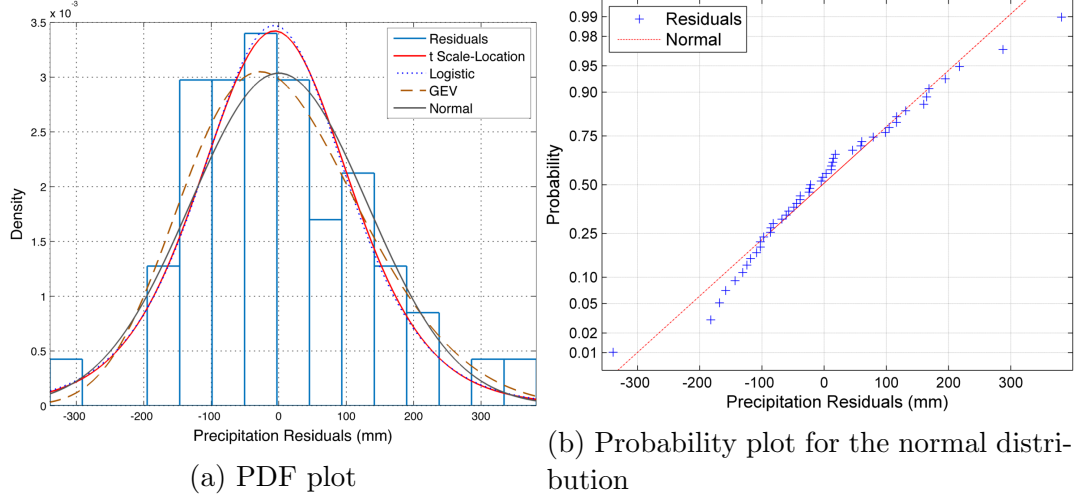


Figure 4.3: Empirical and model probability density function (left) and normal probability plot (right) for the year 1971 of the precipitation residuals provided in Fig. 4.2. On the left, bars correspond to the empirical histogram of the data (blue line), while the model fits include the following: “t Scale–Location” refers to the Student’s t distribution (red line), “Logistic” refers to the Logistic distribution (blue dotted line), “GEV” refers to the generalized extreme value distribution (brown line), and “Normal” refers to the Gaussian distribution (gray line). On the right, cross markers are used for the sample data, while the solid line corresponds to the model normal probability (red line).

also in dimensionless units, since the distances were normalized for the estimation of the variogram parameters. The lengths in the three orthogonal directions are scaled in such a way that the entire 3D map area is contained within the rectangular domain with size 100 x 100 x 1000. This implies that the lengths in all three directions are scaled by appropriate factors according to the following normalizing equations

$$\xi_x = \frac{\xi}{3.5326 \cdot 10^{-4}} \quad (4.2)$$

$$\xi_y = \frac{\xi}{8.3615 \cdot 10^{-4}} \quad (4.3)$$

$$\xi_z = \frac{\xi}{0.4103} \quad (4.4)$$

Inverting normalization, we obtain the anisotropic correlation lengths $\xi_x = 28\text{km}$, $\xi_y = 12\text{km}$, $\xi_z = 24\text{m}$.

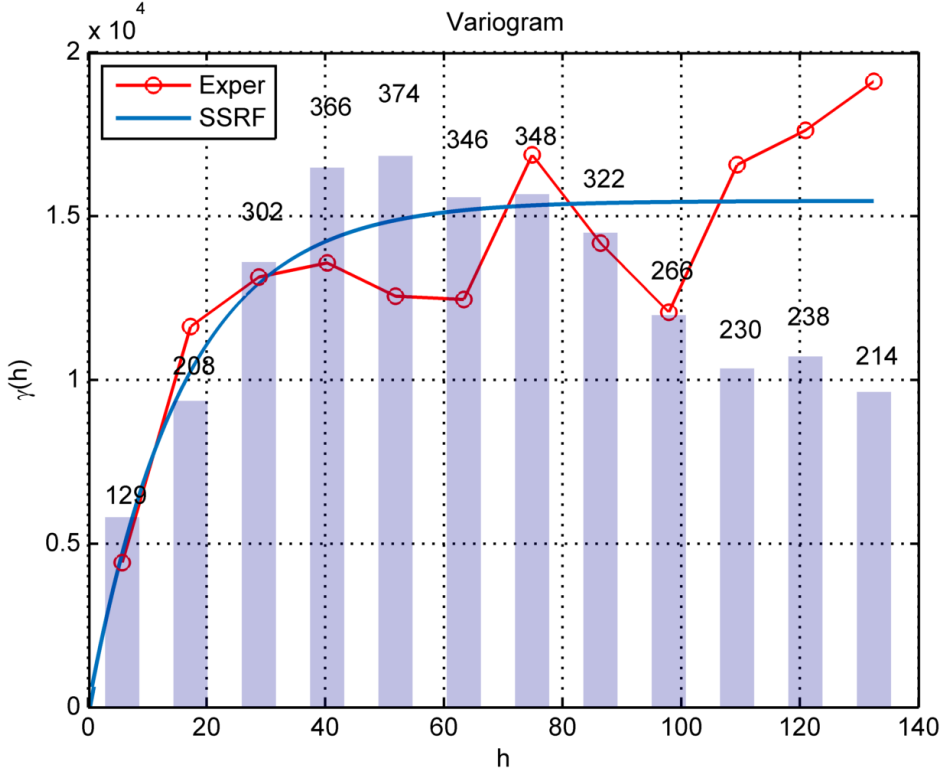


Figure 4.4: Empirical variogram (crosses) and theoretical Spartan variogram model (continuous blue line) based on Eq. (2.30) for the precipitation residuals of the year 1971 (shown in Fig. 4.2). The horizontal axis is calibrated using normalized distances; one unit corresponds to a lag distance of approximately 2.8 km in x direction, 1.2 km in y direction, and 2.4 m in z direction. Bars count the available pairs of sample points for every lag distance. The estimated parameters for the theoretical model are as follows: nugget variance $\sigma_n^2 = 173.9258 \text{ mm}^2$, $\eta_0 = 1.9764e + 05 \text{ mm}^2$, $\eta_1 = -0.8643$, $\xi = 9.9620$.

4.1.3 Regression Kriging

As shown above, the residual precipitation approximately follows a Gaussian distribution. Hence, kriging can be used to create precipitation maps, based on the theoretical model of the variogram (see Fig. 4.5). The necessary equations for the application of the residual kriging are presented in detail in Section 2.12.1.4. Kriging is used to generate a map of the residuals which then has to be transformed into a map of precipitation values. This is accomplished by adding the trend values at each map grid point. The trend at the grid points is estimated

using the elevation values from the Digital Elevation Model which was developed using Geographical Information Systems (see Fig. 3.2), combined with the trend function calculated —the trend function’s coefficients are shown in Table 4.2. The precipitation map thus generated is shown in Fig. 4.6. We observe three areas with higher prediction values than the rest of the island, which coincide with the three mountainous areas shown in the digital elevation model. This reflects the significant influence of the topographic parameters on the precipitation. More specifically, at the highest peak of Lefka Ori, which is located in Chania, the predicted precipitation value equals 2326 mm, at mountain Psiloritis in Rethimno the predicted precipitation is equal to 2055 mm, while on mountain Dicti in the regional unit of Lasithi the precipitation equals 1775 mm. These values are also in agreement with the overall trend of a negative west–east precipitation gradient.

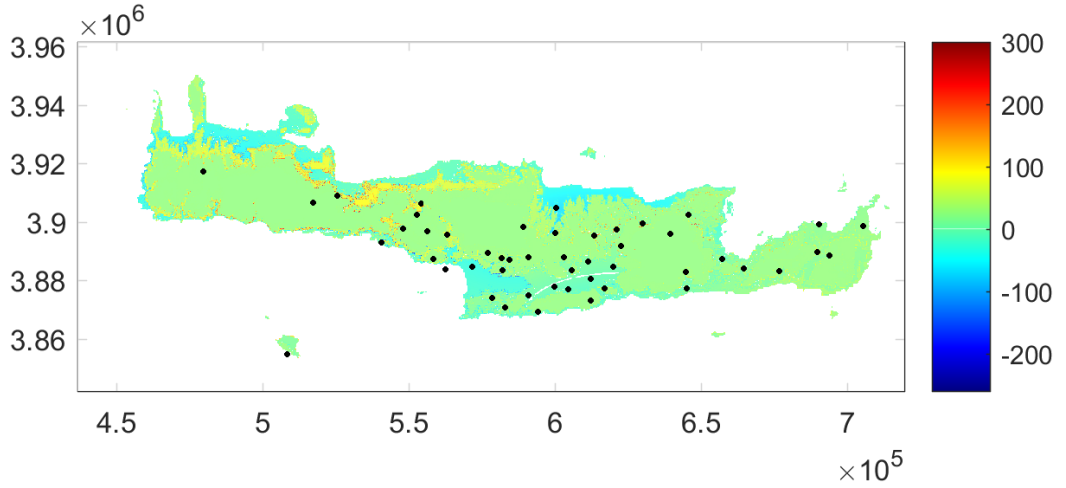


Figure 4.5: Map of estimated annual precipitation residuals for 1971, based on the Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.2. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

The ratio of the kriging standard deviation divided by the kriging estimate at the grid points, generates a map showing the kriging coefficient of variation. The map of the coefficient of variation is presented in Fig. 4.7. The biggest uncertainties are marked in red and they occur in the south–eastern part of Crete. The areas with high coefficient of variation are low–land areas with lower

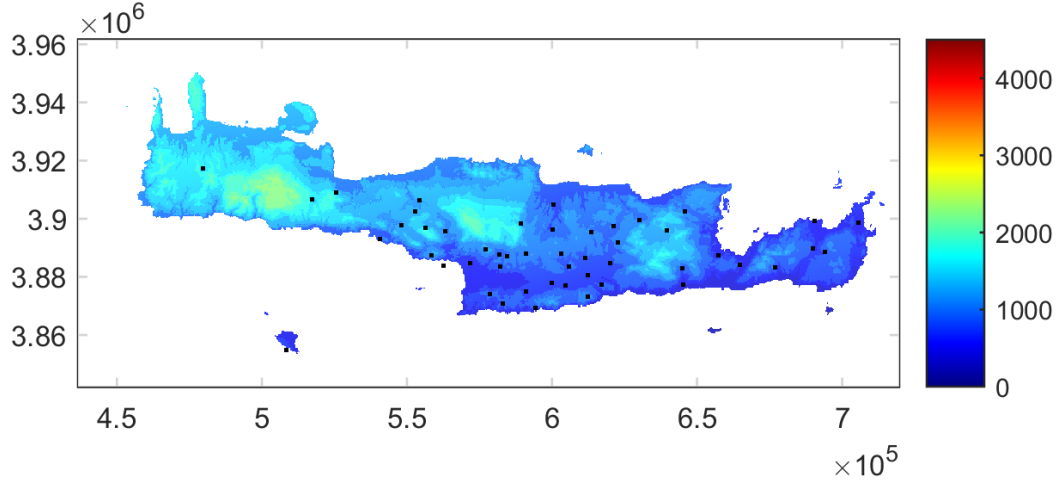


Figure 4.6: Map of estimated annual precipitation for 1971, based on Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.2. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

precipitation. Even though the standard deviation in these areas is lower than in the rest of Crete, the precipitation estimate is not analogously lower, resulting to higher sensitivity to changes. These errors are also partially due to the lack of measuring stations along the coastline.

4.1.4 Leave-One-Out Cross Validation

For the assessment of the spatial precipitation model's performance we use the method of leave-one-out cross validation. This method is described in detail in Section 2.13.1. The results of the cross validation analysis are presented in Figs. 4.8 and 4.9 and in Table 4.4. The latter summarizes the cross validation measures first for the estimated residuals and secondly for the estimated precipitation. The low mean error value means low bias, while the transformation from residuals to precipitation decreases the mean absolute relative error and the root mean square relative error. We notice the high correlation, based on both Pearson's and Spearman's correlation coefficients, between the estimated precipitation values and the actual precipitation values. This is achieved after the incorporation of the precipitation trend to the residuals.

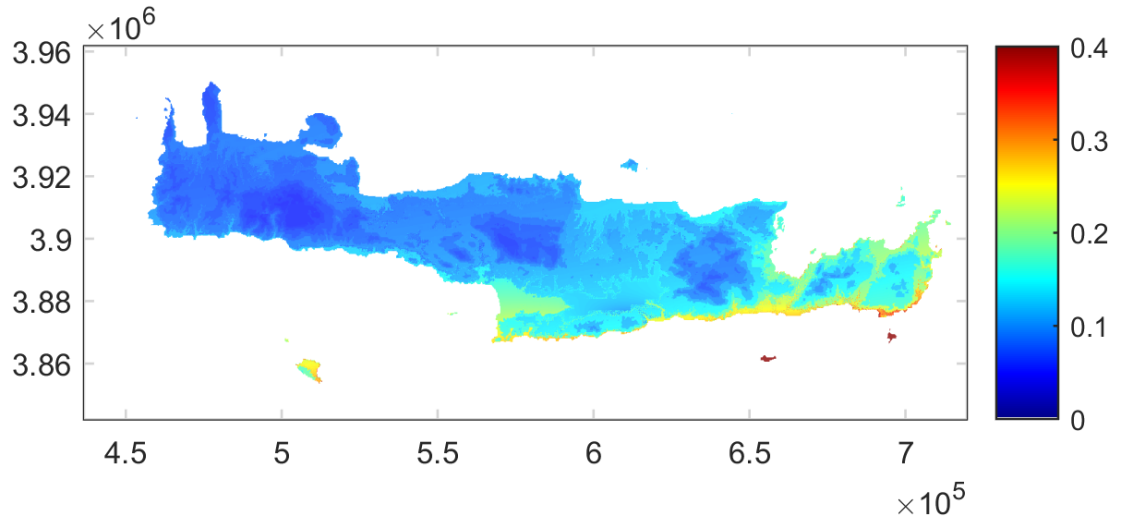


Figure 4.7: Map of estimated coefficient of variation of annual precipitation for 1971, based on Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.2. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

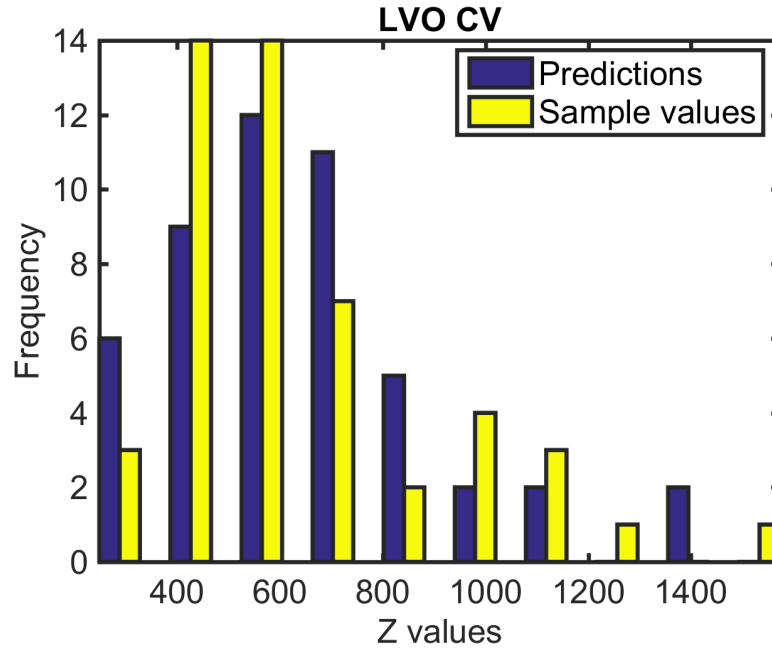


Figure 4.8: Kriging-based leave-one-out cross validation predictions versus sample values for 1971 precipitation (based on the data set shown in Fig. 4.2).

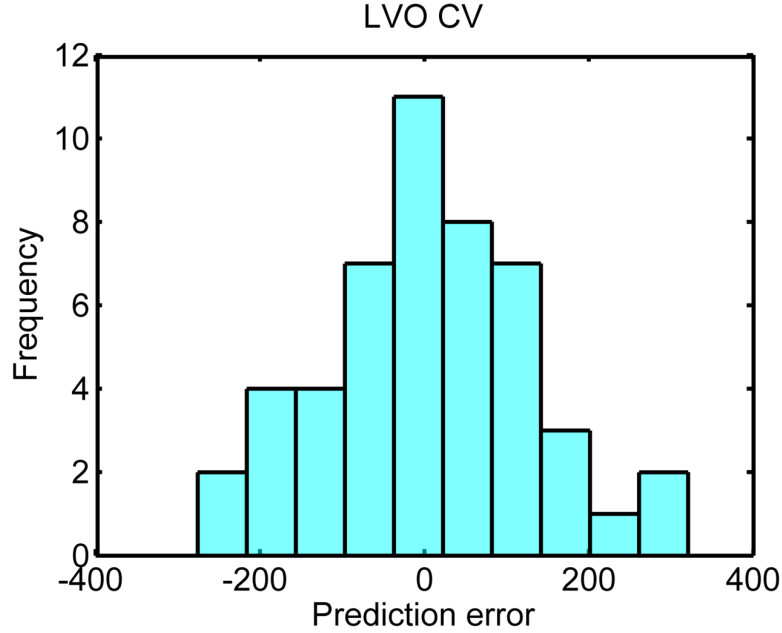


Figure 4.9: Histogram of leave-one-out cross validation errors for 1971 precipitation (based on the data set shown in Fig. 4.2).

A comparison between the actual precipitation at each station and the respective value obtained by means of leave-one-out cross validation is shown in the bar plot of Fig. 4.10.

4.2 Model of Average Monthly Precipitation

For the average monthly precipitation the results obtained by the application of the spatial analysis are very similar to those extracted from the analysis of the annual precipitation data set. Below, we repeat the analysis conducted in Section 4.1 for the annual precipitation field, and we present the respective results for the average monthly precipitation field in the year 1971.

4.2.1 Topographic Trend

In order to remove the trend from the precipitation data, we performed multilinear regression of the response (e.g., the precipitation) on the predictors, which contain the longitude, latitude and elevation. The residuals remaining after the

Table 4.4: Cross validation performance measures calculated through leave-one-out cross validation for the precipitation data of year 1971 for annual recorded precipitation data set. ME: Mean error (bias) (Eq. (2.72)), MAE: Mean absolute error (Eq. (2.73)), MARE: Mean absolute relative error (INF if z contains zeros) (Eq. (2.74)), RMSE: Root mean square error (Eq. (2.75)), RMSRE: Root mean square relative error (INF if the actual precipitation value is zero) (Eq. (2.76)), RP: Pearson's linear correlation coefficient (Eq. (2.77)), RS: Spearman (rank) correlation coefficient (Eq. (2.78)), ErrMin: Difference between minimum prediction and minimum sample value, ErrMax: Difference between maximum prediction and maximum sample value.

1971	ME(mm)	MAE(mm)	MARE	RMSE(mm)	RMSRE	RP	RS	ErrMin(mm)	ErrMax(mm)
Residuals	2.9540	96.9050	2.3637	129.3083	6.5976	0.3132	0.3212	157.5841	-145.7528
Precipitation	2.9540	96.9050	0.1623	129.3083	0.2171	0.8703	0.8742	-72.6871	-238.2975

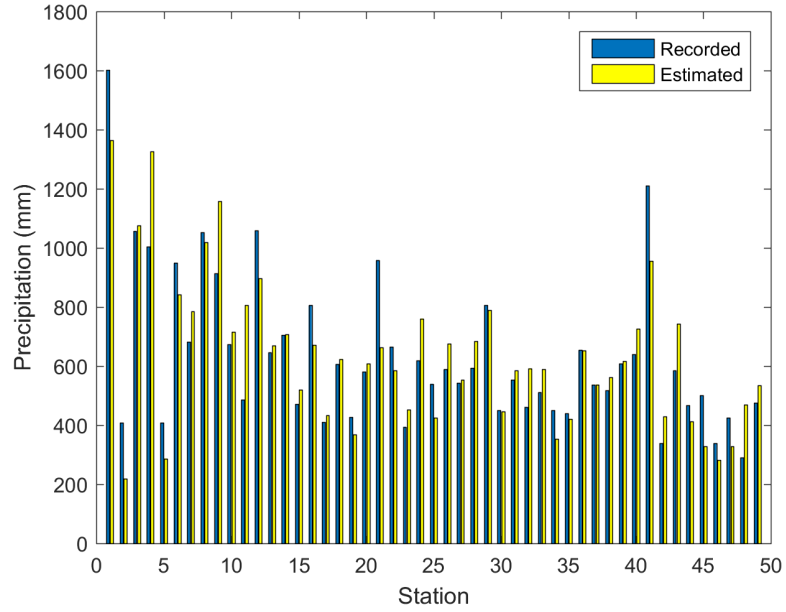


Figure 4.10: Comparison between the actual precipitation values of the year 1971 at each station and the predictions derived using leave-one-out cross validation.

application of the multilinear regression method are shown in Fig. 4.12 (in red). The estimates of the coefficients for the regression model of the average monthly precipitation field in the year 1971 are presented in Table 4.5.

Table 4.5: Trend model parameters, for average monthly precipitation field in the year 1971 (Eq. (3.2)).

b	$\times 1$	$\times x$	$\times y$	$\times z$
$b_i, i=0, \dots, 3$	-3.4932e+03	-0.0002	0.0010	0.0474
lower bound	-4.5959e+03	-2.2887e-04	0.0007	0.0321
upper bound	-2.3905e+03	-9.2975e-05	0.0012	0.0627

As mentioned above for the annual precipitation, elevation contributes most to the trend model, the same applies for the average monthly precipitation field. Figure 4.11 shows a plot of the trend function and the data versus northing and elevation.

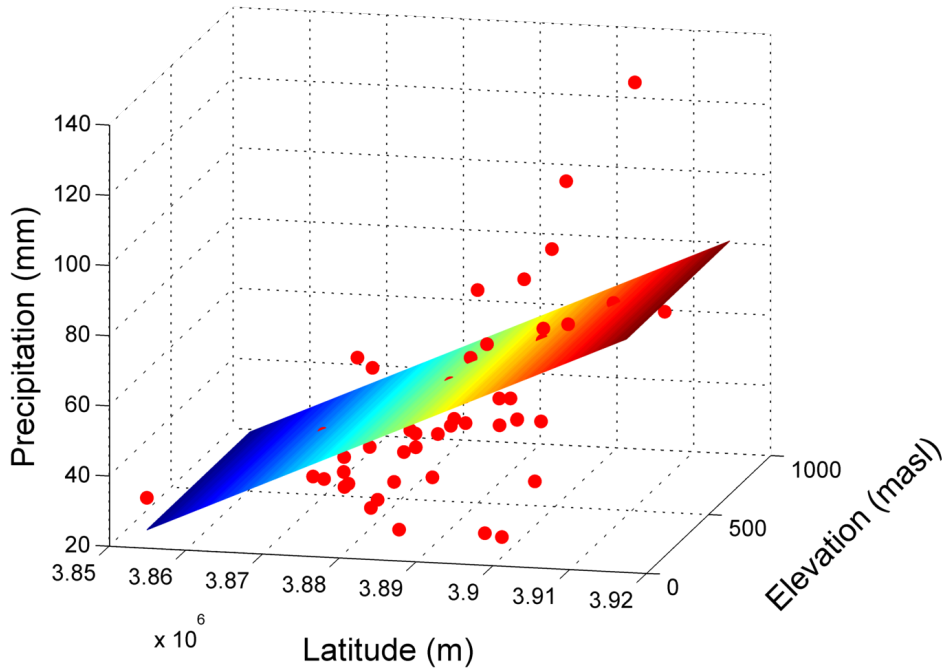


Figure 4.11: Multilinear regression of the precipitation field on the predictors in which comprise longitude, latitude, and elevation, for the average monthly precipitation field in the year 1971. The trend equation for this year is given by $f = -3.4932 \cdot 10^3 - 0.0002x + 0.0010y + 0.0474z$. The coordinates x, y, z are measured in the Greek Geodetic Reference System (EGSA '87), with x and y representing the easting and northing coordinates measured in meters, and z being the elevation measured in meters above sea level.

In Fig. 4.12 the average monthly precipitation field values are marked with blue circles, while the residual precipitation resulting after the removal of the trend is marked with red asterisks. The plots of the model probability density functions and the normal probability plot of the precipitation residuals are presented in Fig. 4.13. These plots show that the residuals are approximately normally distributed.

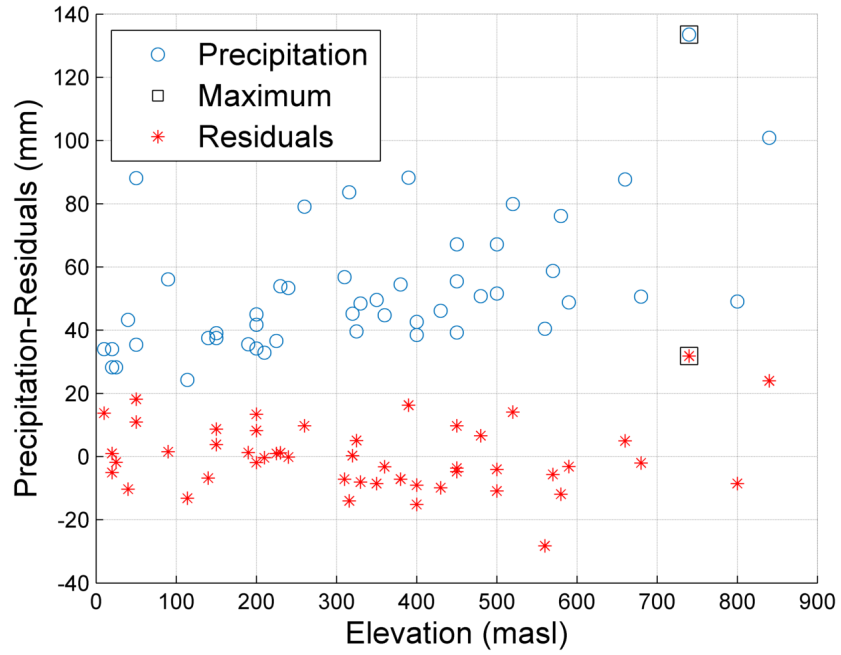


Figure 4.12: Scatter plot of the average monthly precipitation field values and the precipitation residuals obtained after the application of multilinear regression according to the trend equation shown in Fig. 4.11, for the year 1971. The blue circles correspond to the monthly precipitation field values, while the red asterisks correspond to the precipitation residuals. The square boxes mark the highest values for the respective data set.

Summary statistics of the average monthly precipitation data set for the year 1971 are presented in Table 4.6. They include the mean value, the minimum, the maximum value, the standard deviation of the average monthly precipitation as well as the respective values for precipitation residuals. The zero mean of the residuals, the approximately zero skewness (see Table 4.6), and the proximity of the residuals to the straight line on the normal probability plot as shown in Fig. 4.13b prove that the residuals are close to the Gaussian distribution.

Table 4.6: Summary statistics for the precipitation data set depicted in Fig. 4.12. The parameter n is the number of available stations of the year.

1971	n	Mean	Median	Minimum	Maximum	Std.dev.	Coef.of var.	Skewness
Precipitation	49	52.9082	48.4250	24.2667	133.5333	21.5426	0.4072	1.5316
Residuals	49	8.7977e-13	-1.8323	-28.2669	31.8001	10.9527	1.2449e+13	0.4557

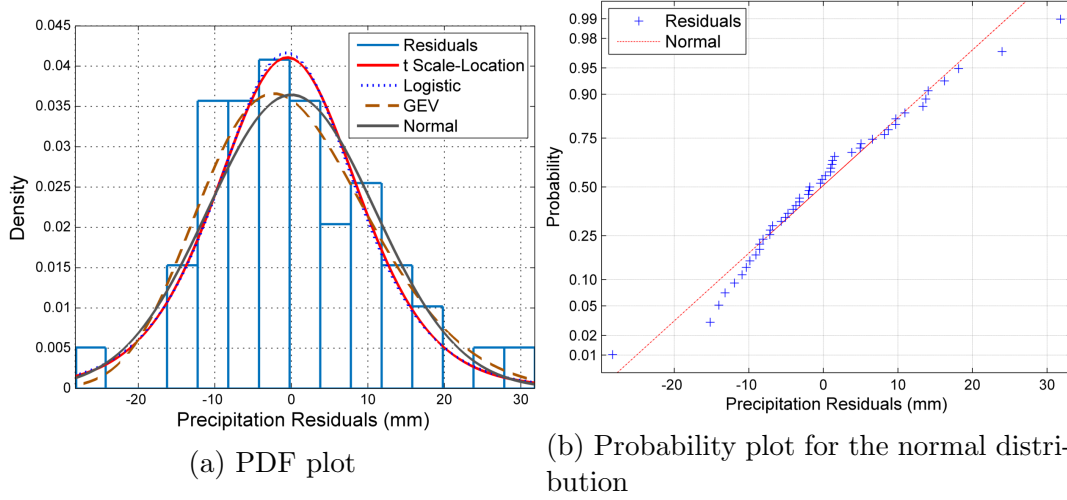


Figure 4.13: Empirical and model probability density function (left) and normal probability plot (right) for the year 1971 of the precipitation residuals provided in Fig. 4.12. On the left, bars correspond to the empirical histogram of the data (blue line), while the model fits include the following: “t Scale-Location” refers to the Student’s t distribution (red line), “Logistic” refers to the Logistic distribution (blue dotted line), “GEV” refers to the generalized extreme value distribution (brown line), and “Normal” refers to the Gaussian distribution (gray line). On the right, cross markers are used for the sample data, while the solid line corresponds to the model normal probability (red line).

4.2.2 Variogram Modeling

We fitted the theoretical Spartan variogram model to the experimental variogram, to construct the spatial model. This allows estimating the parameters η_0 , η_1 , and the characteristic length ξ of the Spartan model (see Eq. (2.30)). Figure 4.14 shows the graphical representation of the variogram of the above data set. The scale parameter η_0 that determines the total variance of the fluctuations is equal to $2.6979e + 03 \text{ mm}^2$, the shape parameter $\eta_1 = 1.9999$, and the correlation length ξ that shows the range of the spatial dependence equals 15.8604. The

shape parameter η_1 is in dimensionless units. Particularly for this case, the correlation length ξ is also dimensionless, since the distances were normalized for the estimation of the variogram parameters. The same normalization that was applied for the annual precipitation is applied here too. The biggest contribution to the variogram estimation has the vertical direction, which implies anisotropy focused on the elevation. Inverting normalization, we obtain the anisotropic correlation lengths $\xi_x = 45\text{km}$, $\xi_y = 19\text{km}$, $\xi_z = 39\text{m}$.

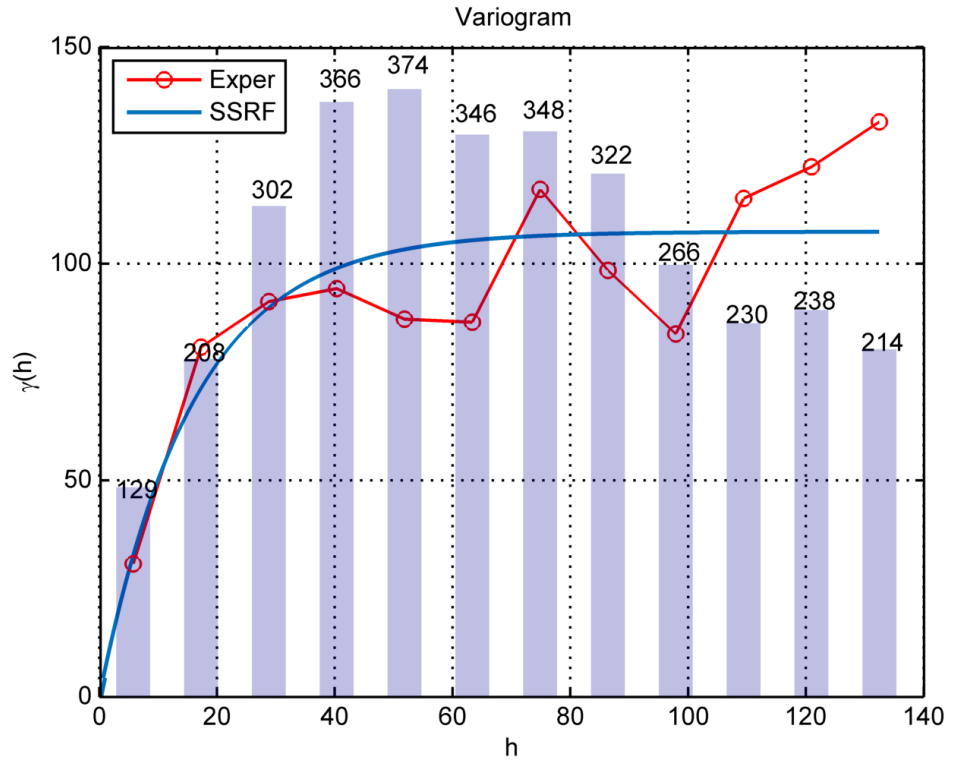


Figure 4.14: Empirical variogram (crosses) and theoretical Spartan variogram model (continuous blue line) based on Eq. (2.30) for the precipitation residuals of the year 1971 (shown in Fig. 4.12). The horizontal axis is calibrated using normalized distance; one unit corresponds to a lag distance of approximately 2.8 km in x direction, 1.2 km in y direction, and 2.4 m in z direction. Bars count the available pairs of sample points for every lag distance. The estimated parameters for the theoretical model are as follows: nugget variance $\sigma_n^2 = 2.2204e - 16 \text{ mm}^2$, $\eta_0 = 2.6979e + 03 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 15.8604$.

4.2.3 Regression Kriging

The residual precipitation follows a Gaussian distribution as shown above, thus kriging can be used to create precipitation maps, based on the theoretical model of the variogram (see Fig. 4.15). The necessary equations for the application of the residual kriging are presented in detail in Section 2.12.1.4.

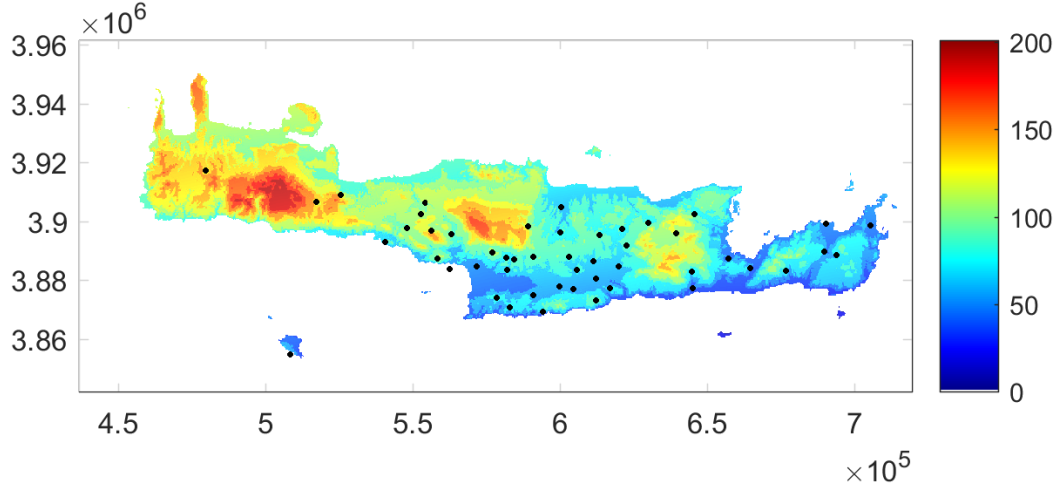


Figure 4.15: Map of estimated average monthly precipitation for 1971, based on Spartan variogram model (Eq. 2.30) and the data set shown in Fig. 4.12. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

The ratio of the kriging standard deviation divided by the kriging estimate at the grid points, generates a map showing the kriging coefficient of variation. The map of the coefficient of variation is presented in Fig. 4.16. The coefficient of variation map of average monthly precipitation is very similar to the one created for the annual precipitation data set, with the biggest estimation errors in the south-eastern part of Crete.

4.2.4 Leave one Out Cross Validation

The results of leave-one-out cross validation analysis —the method used for the assessment of the model's performance— are presented in Figs. 4.17 and 4.18 and in Table 4.7. The latter summarizes the cross validation measures firstly for the

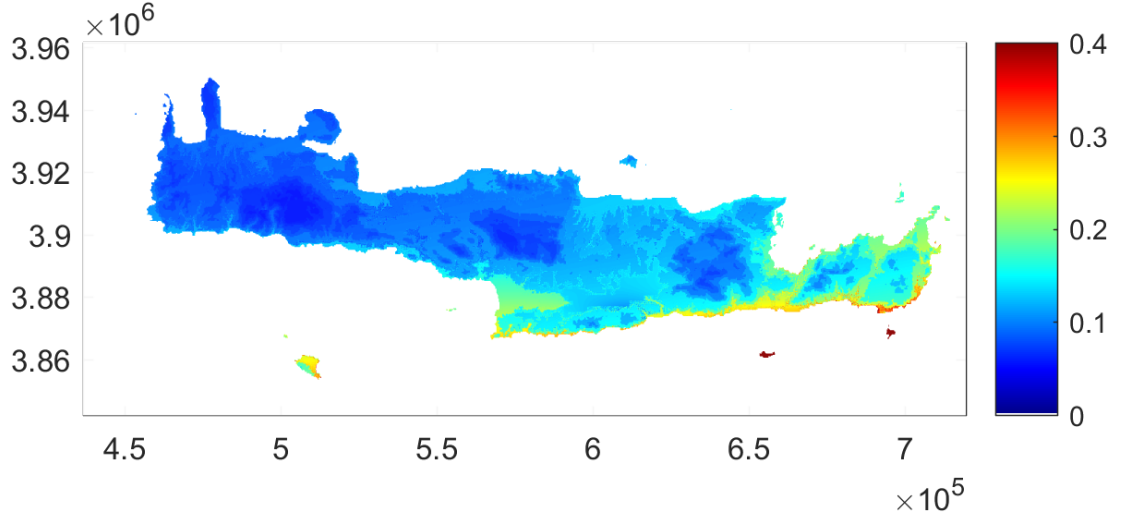


Figure 4.16: Map of estimated coefficient of variation of average monthly precipitation for 1971, based on Spartan variogram model (Eq. (2.30)) and the data set shown in Fig. 4.12. The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

estimated residuals and secondly for the estimated precipitation. As observed for the annual precipitation data set, we extracted similar conclusion from the cross validation measures for the average monthly precipitation. We observe low mean error (low bias), and decrease to the mean absolute relative error and the root mean square relative error after the incorporation of the precipitation trend to the residuals. Also after that transformation, we observe high correlation coefficient ranks, based on both Pearson's and Spearman's correlation coefficients, between the estimated precipitation values and the actual precipitation values.

A comparison between the actual precipitation at each station and the respective value obtained by means of leave-one-out cross validation is shown in the bar plot of Fig. 4.19.

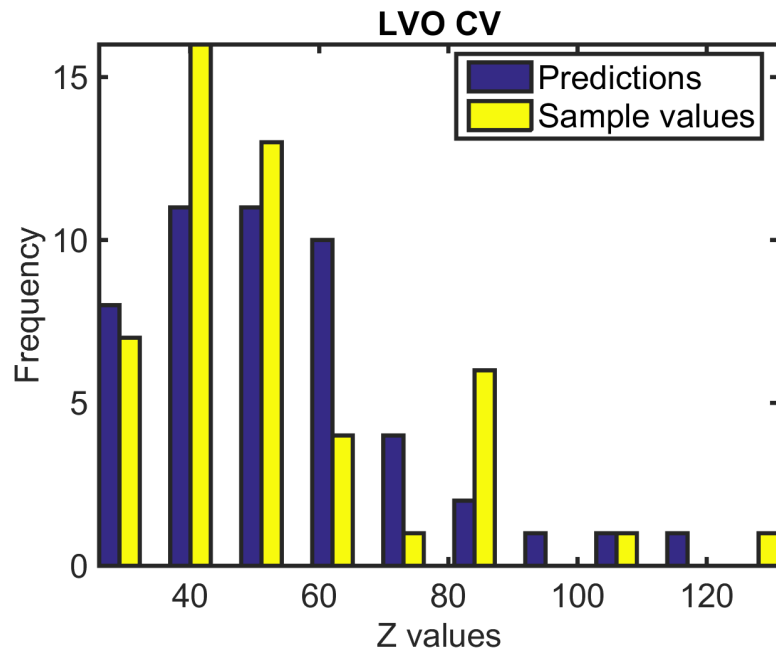


Figure 4.17: Kriging-based leave-one-out cross validation predictions versus sample values for 1971 precipitation (based on the data set shown in Fig. 4.12).

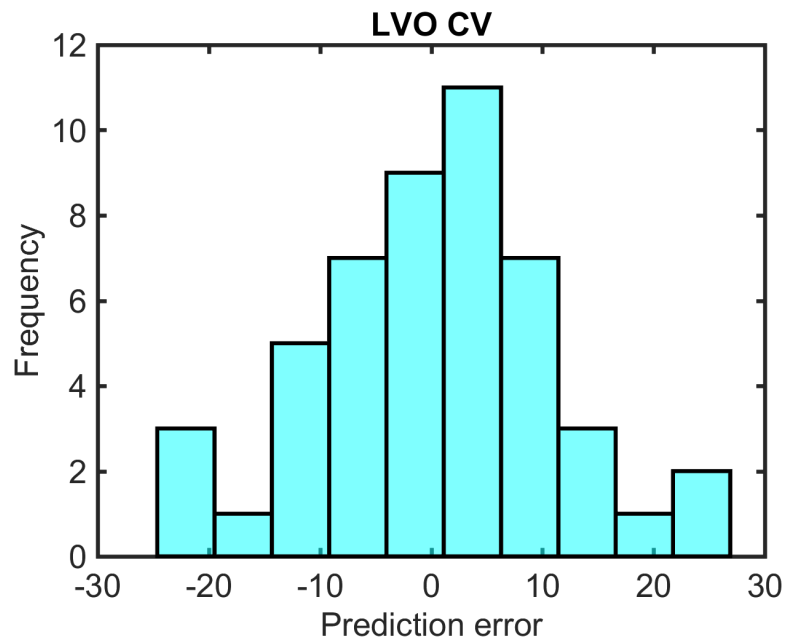


Figure 4.18: Histogram of leave-one-out cross validation errors for the 1971 precipitation (based on the data set shown in Fig. 4.12).

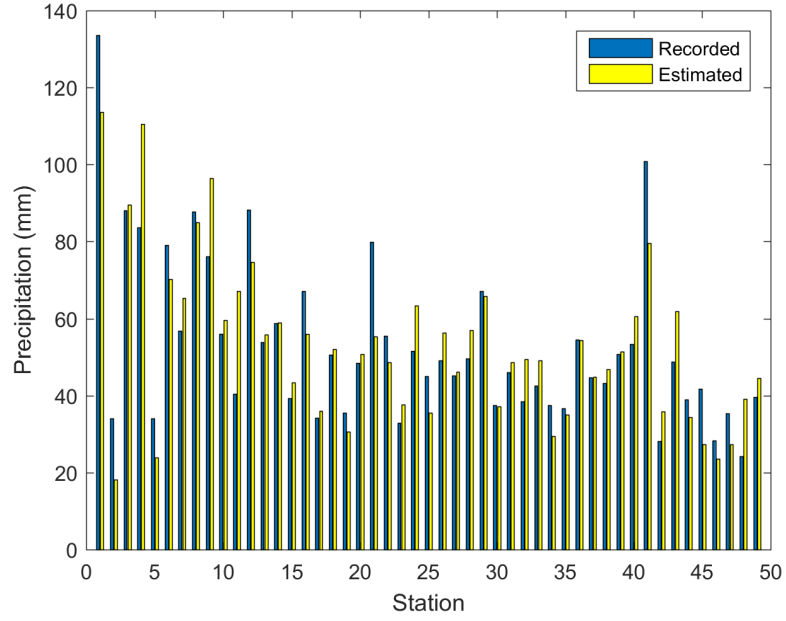


Figure 4.19: Comparison between the actual precipitation values of the year 1971 at each station and the predictions derived using leave-one-out cross validation.

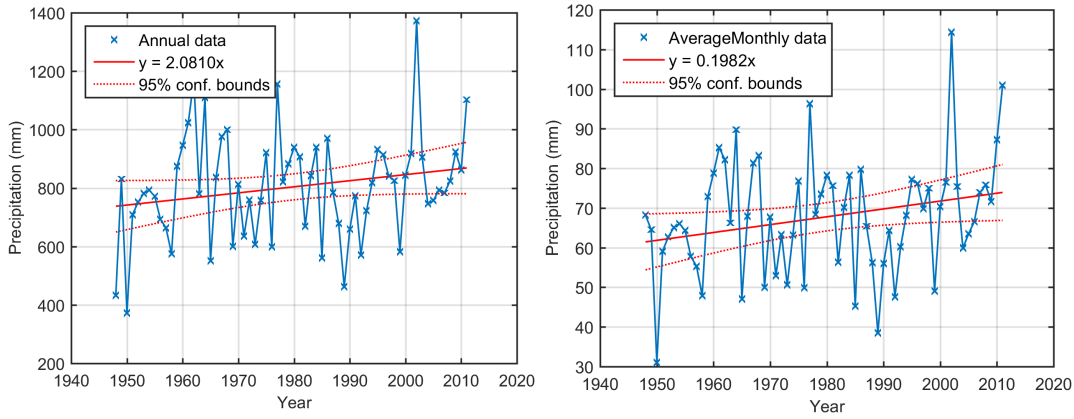
Table 4.7: Cross validation performance measures calculated through leave-one-out cross validation for the precipitation data of year 1971 for the average monthly precipitation data set. ME: Mean error (bias) (Eq. (2.72)), MAE: Mean absolute error (Eq. (2.73)), MARE: Mean absolute relative error (INF if z contains zeros) (Eq. (2.74)), RMSE: Root mean square error (Eq. (2.75)), RMSRE: Root mean square relative error (INF if the actual precipitation value is zero) (Eq. (2.76)), RP: Pearson's linear correlation coefficient (Eq. (2.77)), RS: Spearman (rank) correlation coefficient (Eq. (2.78)), ErrMin: Difference between minimum prediction and minimum sample value, ErrMax: Difference between maximum prediction and maximum sample value

1971	ME(mm)	MAE(mm)	MARE	RMSE(mm)	RMSRE	RP	RS	ErrMin(mm)	ErrMax(mm)
Residuals	0.2462	8.0754	2.3637	10.7757	6.5976	0.3132	0.3212	13.1320	-12.1461
Precipitation	0.2462	8.0754	0.1623	10.7757	0.2171	0.8703	0.8742	-6.0573	-19.8573

In both cases, the estimation smoothen local fluctuations. At stations with low precipitation values the kriging method tends to force the estimate to higher values compared to the original record and closer to the mean value. On the other hand, at stations with high values, the estimates are lower than the true values. This smoothing effect is a well-known property of kriging interpolators.

4.3 Spatiotemporal Patterns of Precipitation

The calculation of the overall trend of precipitation in Crete, based on the recorded values, seems to have an increasing pattern. More specifically, the trend calculated with mean values of every year's records of annual precipitation refers to an increase of 2.1 mm of precipitation per year, this increase corresponds to 0.28% increase per year (Fig. 4.20b). For the average monthly precipitation, the trend calculated with the data set mean values, is an increase of 0.2 mm of precipitation, that corresponds to an increase of 0.32% per year for average monthly precipitation values.



(a) Annual precipitation, approximately 2.1 mm increase or 0.28% increase per year. (b) Average monthly precipitation, approximately 0.2 mm increase or 0.32% increase per year.

Figure 4.20: Overall trend calculated from the recorded data sets. The averaged annual precipitation values are calculated with the calculation of the mean values of annual precipitation of every station in Crete for every year (left), while the averaged average monthly precipitation values are calculated with the calculation of the mean values of average monthly precipitation of every station in Crete for every year (right). The solid red line shows the trend equation, while the red dashed line shows the 95% coefficient bounds.

The results derived from the application of multilinear regression for the estimated precipitation values, gave the similar patterns as the results taken from the recorded values. Particularly, for the annual estimated precipitation, there is a 0.24% increase per year (Fig. 4.21a), while for the average monthly estimated

precipitation, the increase equals to 0.36% per year (Fig. 4.21b).

These are pretty interesting results, since they contradict the results taken from various GCMs according to the latest report of the Intergovernmental Panel on Climate Change [Hartmann et al., 2013], that monitors climate change on behalf of the UN Programme for the Environment (UNEP). The IPCC results refers to a decrease by 20% in South and Eastern Europe. In Southern Europe precipitation will show a decrease of 1% per decade, while in the summer season this reduction will reach 5%.

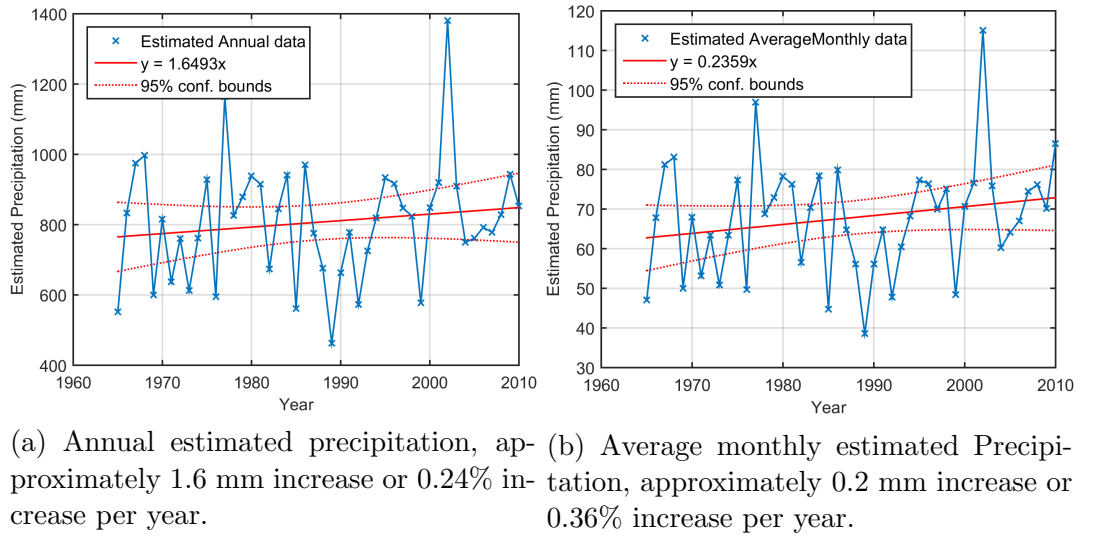


Figure 4.21: Overall trend calculated from the estimations. The averaged annual precipitation values are calculated with the calculation of the mean values of the estimated annual precipitation of every station in Crete for every year (left), while the averaged average monthly precipitation values are calculated with the calculation of the mean values of estimated average monthly precipitation of every station in Crete for every year (right). The solid red line shows the trend equation, while the red dashed line shows the 95% coefficient bounds.

Also according to Vrochidou et al. [2013] in the results presented for annual precipitation over Crete, the calculated trend follows a negative ratio, especially when applying the worst case scenarios. The estimation trends are decrease in precipitation amount by 6%, 17% and 26% concerning three different scenarios (best:RCP2.6, moderate:4.5 and worst:8.5) respectively.

Below we present a figure of a six years period results of precipitation maps

for visible proof that our results did not seem to present decrease in precipitation amount. We do not present the years after 2009, because we have only a few stations available to extract results.

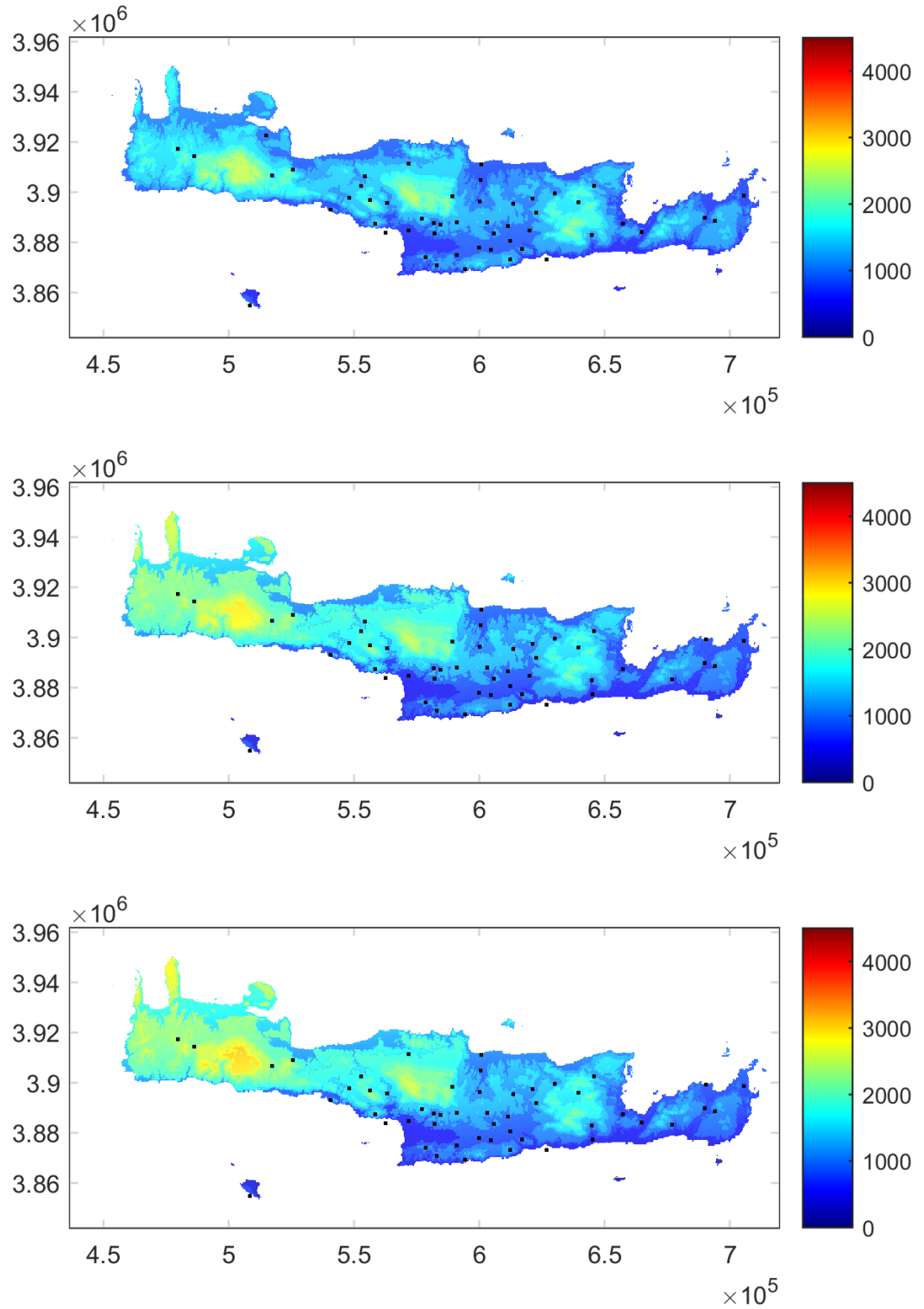


Figure 4.22: Kriging maps of precipitation from 2004 until 2009, created based on the Spartan variogram model (Eq. (2.30)). The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

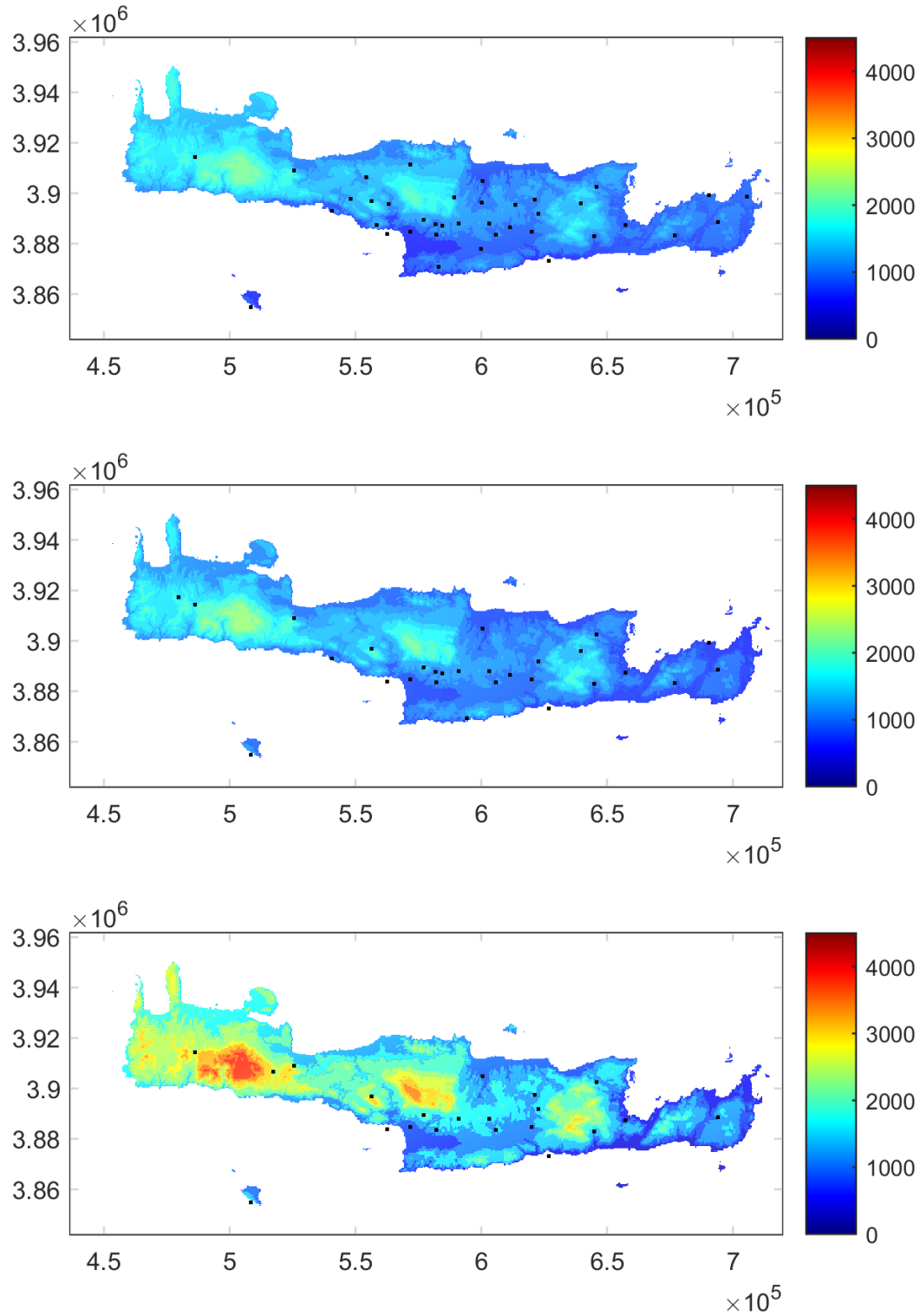


Figure 4.22: Kriging maps of precipitation from 2007 until 2009 (cont.).

Chapter 5

Conclusions

This study attempts for the first time a comprehensive analysis of the spatial variability of monthly rainfall over the island of Crete using geostatistical methods. Our goal is to better understand the spatial and temporal patterns that underlie the considerable variability of the rainfall process over the island.

The spatio-temporal analysis of precipitation data for the period 1948–2011 for the island of Crete yields useful information about climate change and precipitation trends on the island. We estimate the precipitation probability distributions for the dry and wet periods, separately for the west and east regions of the island, both at fixed time scales (fields), but also at specific spatial scales (time series). The original monthly data sets for the wet periods are fitted accurately with the Generalized Extreme Value distribution, but the data for the dry period do not give good fits because of many precipitation records with zero values. A spatial trend is observed, with the average monthly precipitation in the West measuring about 40 mm higher than in the East. Accordingly, a spatial trend is observed with the annual precipitation in the West measuring 450 mm higher than in the East.

In addition to the geographical East–West pattern of precipitation on the island, we investigate the topographic dependence of precipitation. The analysis reveals positive correlation between elevation and precipitation, and between latitude and precipitation, whereas the longitude is negatively correlated with precipitation. A linear regression model using all three explanatory variables is used, because it gives the lowest coefficient of determination R^2 . It is deduced

that topographic parameters are strongly correlated with precipitation and contribute both to the intensity of precipitation as well as its spatial distribution. The multilinear regression model is used to remove topographic trends and to obtain precipitation residuals. Trend removal helps to transform the initially non-Gaussian precipitation data into Gaussian distributed residuals.

The spatial variability of the residuals is separately investigated for every year through the calculation of the empirical variogram and the subsequent fitting of Spartan variograms as theoretical models. This study is the first application of the Spartan variogram to precipitation data. We apply regression kriging to the precipitation residuals in order to generate precipitation maps. Regression kriging incorporates the trend function at every node of the map grid based on the elevation obtained from the Digital Elevation Model of Crete. Performance measures estimated by means of cross validation prove that the geostatistical approach leads to a representative spatial model, which takes into account the significant spatial variability of the rainfall process on the island. The proposed approach can be integrated as a modeling tool in a comprehensive water resources management plan.

According to the latest report of the Intergovernmental Panel on Climate Change [Hartmann et al., 2013] that monitors climate change on behalf of the UN Programme for the Environment (UNEP), an increase of 10% to 40% of rainfall in northern Europe and a respective decrease of 20% in South and Eastern Europe from 1900 to 2005 is expected. In Southern Europe precipitation is projected to decrease by 1% per decade, while in the summer season this reduction will reach 5%. Also according to Vrochidou et al. [2013] the analysis of annual precipitation over Crete based on climate models shows a negative trend (reduction of precipitation), especially if the worst case scenarios are applied. In particular, a decrease in the amount of precipitation by 6%, 17% and 26% is projected based on three different scenarios (best:RCP2.6, moderate:RCP4.5 and worst:RCP8.5) respectively.

Studying the evolution of the precipitation over time based on the geostatistical analysis, we observe that the annual precipitation has a cyclical behavior with a variable period. It is interesting that the overall rainfall trend does not seem to support the scenario of reduced rainfall due to climate change at least over

the length of this study. In particular, the overall trend represents an increase by approximately 0.28% from 1948 to 2011 for the annual precipitation and 0.32% from 1948 to 2011 for the average monthly precipitation.

A natural evolution of the present study is to extend the methodology presented herein to a more rigorous analysis of anisotropy as well as to a joint spatio-temporal model. To improve the knowledge of the precipitation distribution and water availability, a possible approach is to incorporate additional variables in the model, such as groundwater levels, evapotranspiration and runoff. Such an integrated model could provide a valuable tool for water resources management on the island of Crete. A study involving longer time series is needed to more accurately evaluate the potential impact of climate change.

Our current model cannot handle zero precipitation values. A possible improvement could involve the Generalized Pareto distribution model proposed by [Baxevani and Lennatsson \[2015\]](#). This will also allow extension of the geostatistical model to finer time scales (i.e., monthly or daily scale). Finally, it is necessary to compare our results with other types of statistical analysis of extreme events (e.g., [Davison et al., 2013](#)).

Finally it is important to establish more monitoring stations on the island, particularly in the municipality of Chania, where data from only six stations are available. This will give a denser network, that will improve the results of geostatistical analysis.

Appendix A

In [Appendix A](#) we present figures from 1965 until 2010 for the annual precipitation fields. The figures include:

- a. Empirical variogram (crosses) and theoretical Spartan variogram model (continuous blue line) based on Eq. (2.30) for the precipitation residuals. The horizontal axis is calibrated using normalized distances; one unit corresponds to a lag distance of approximately 2.8 km in x direction, 1.2 km in y direction, and 2.4 m in z direction. Bars count the available pairs of sample points for every lag distance. The estimated parameters for the theoretical model are presented separately on every figure's caption.
- b. Kriging-based leave-one-out cross validation predictions versus sample values for precipitation.
- c. Map of estimated annual precipitation, based on Spartan variogram model (Eq. (2.30)). The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.
- d. Map of estimated coefficient of variation of annual precipitation, based on Spartan variogram model (Eq. (2.30)). The Cartesian coordinates are measured in the Greek Geodetic System (EGSA '87), with the horizontal axis representing the easting and the vertical axis representing the northing. Both axes are measured in meters.

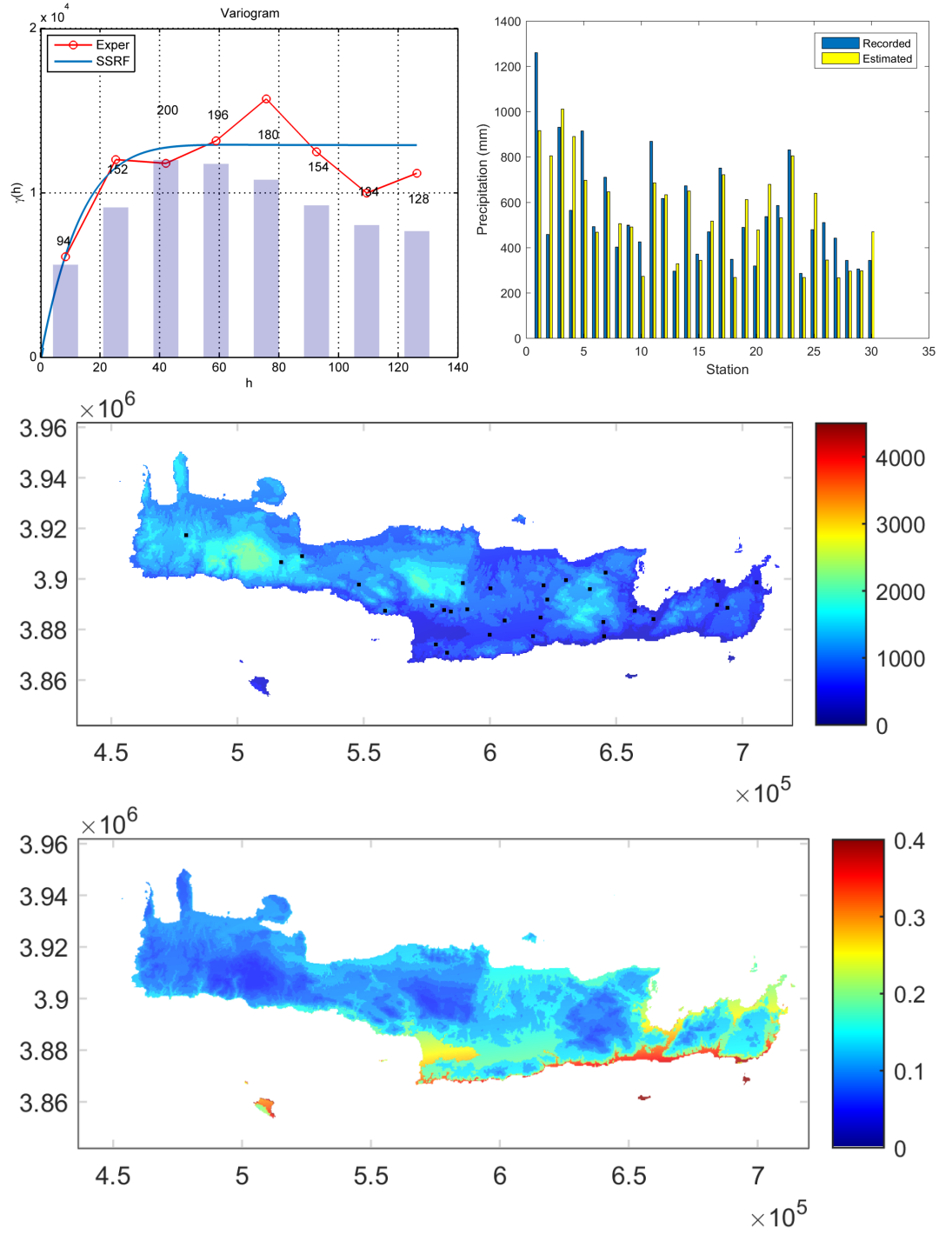


Figure 1: Year 1965 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 0.0041 \text{ mm}^2$, $\eta_0 = 2.5294e + 05 \text{ mm}^2$, $\eta_1 = 0.4333$, $\xi = 10.6993$.

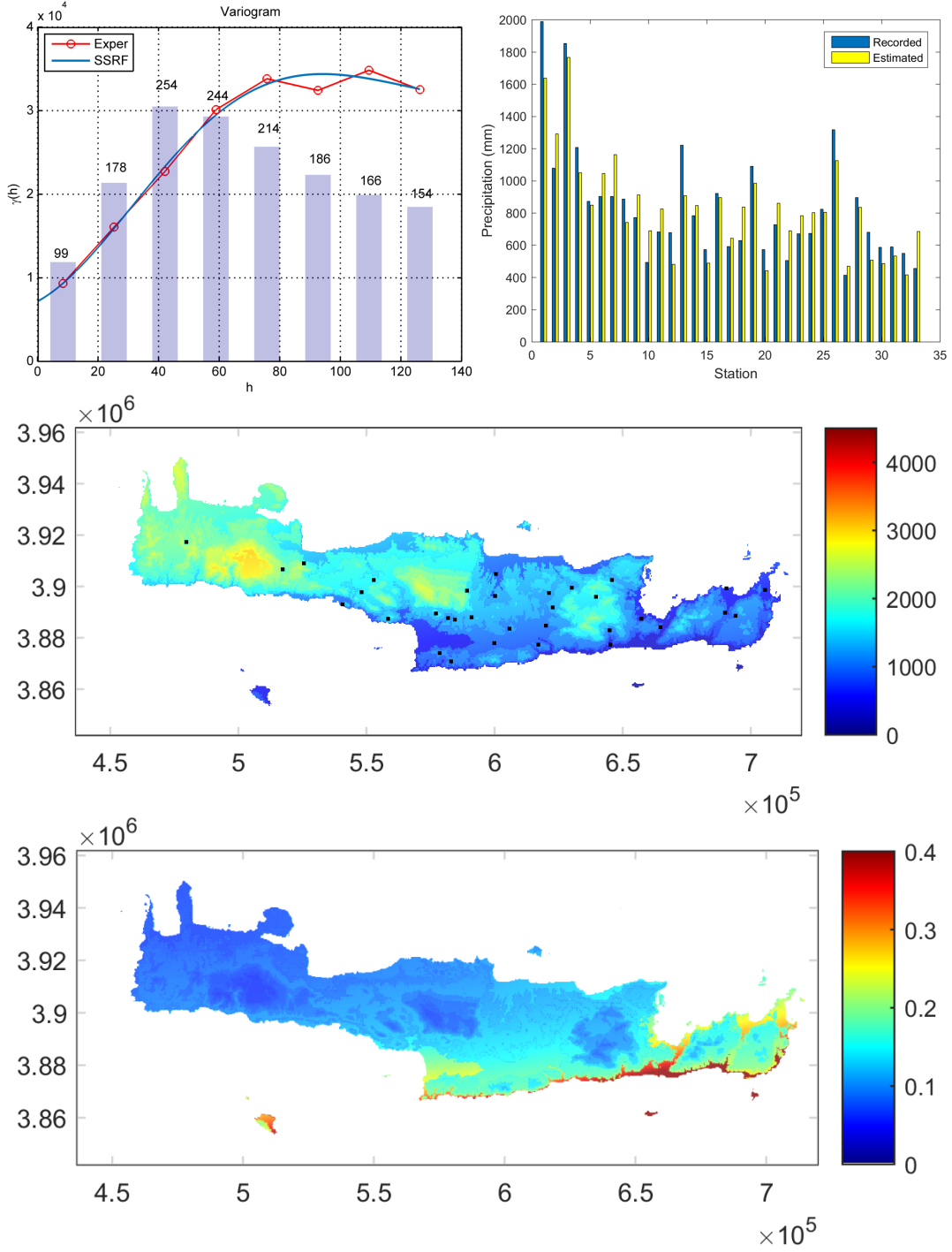


Figure 2: Year 1966 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 7.1825e + 03 \text{ mm}^2$, $\eta_0 = 1.0504e + 05 \text{ mm}^2$, $\eta_1 = -1.8852$, $\xi = 21.4235$.

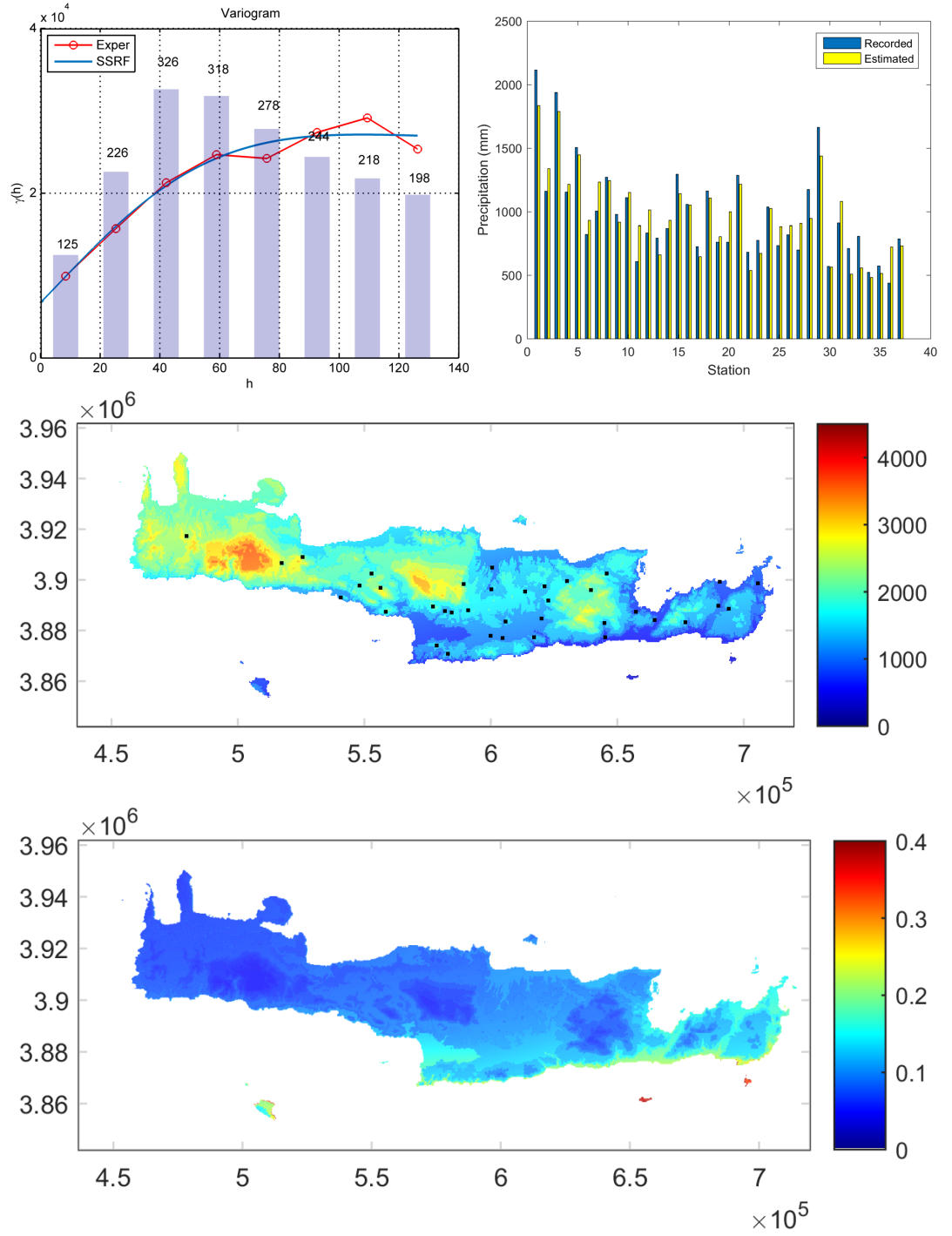


Figure 3: Year 1967 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 6.7205e + 03 \text{ mm}^2$, $\eta_0 = 2.2078e + 05 \text{ mm}^2$, $\eta_1 = -1.2175$, $\xi = 23.7127$.

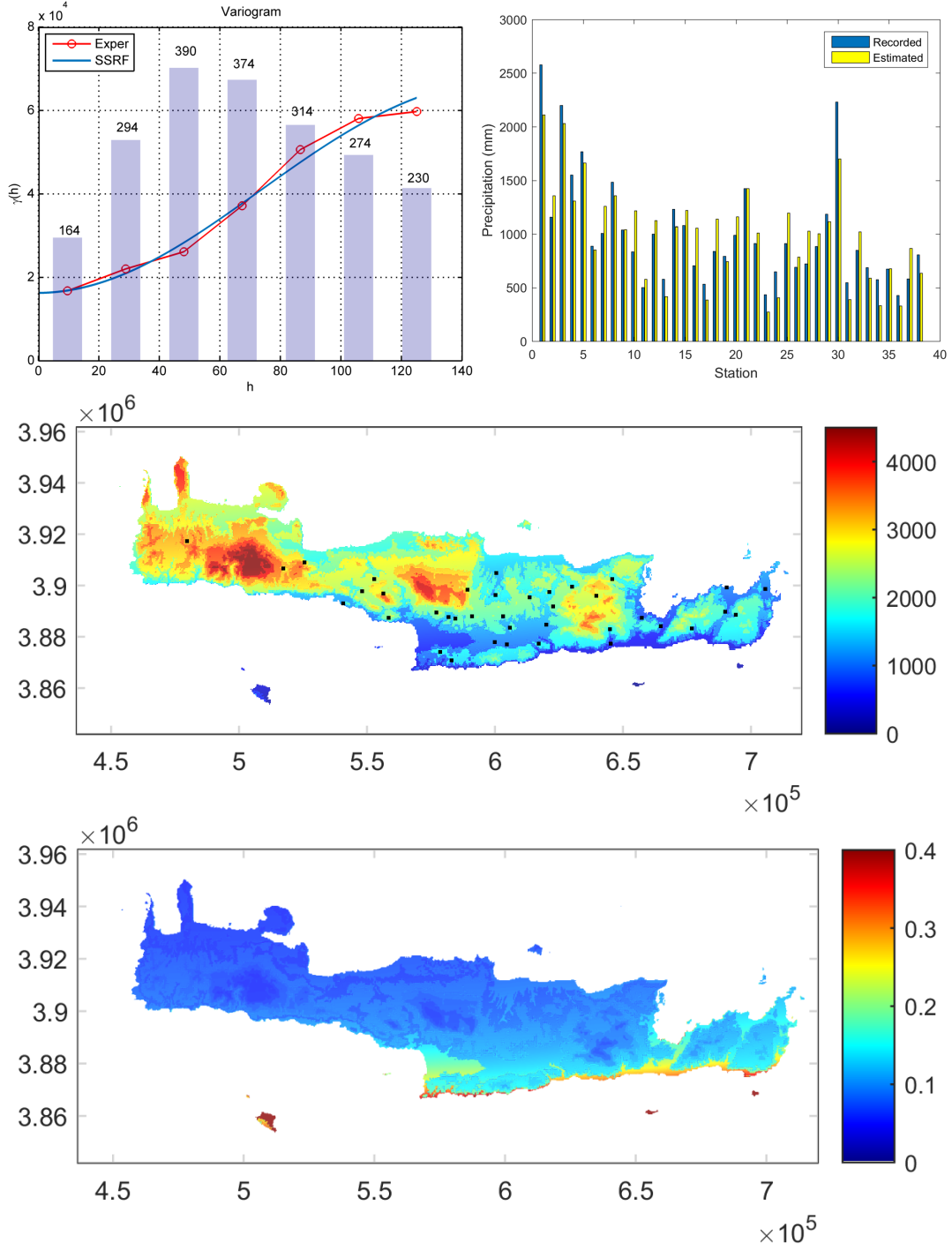


Figure 4: Year 1968 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.6271e + 04 \text{ mm}^2$, $\eta_0 = 5.6444e + 03 \text{ mm}^2$, $\eta_1 = -1.9999$, $\xi = 35.4525$.

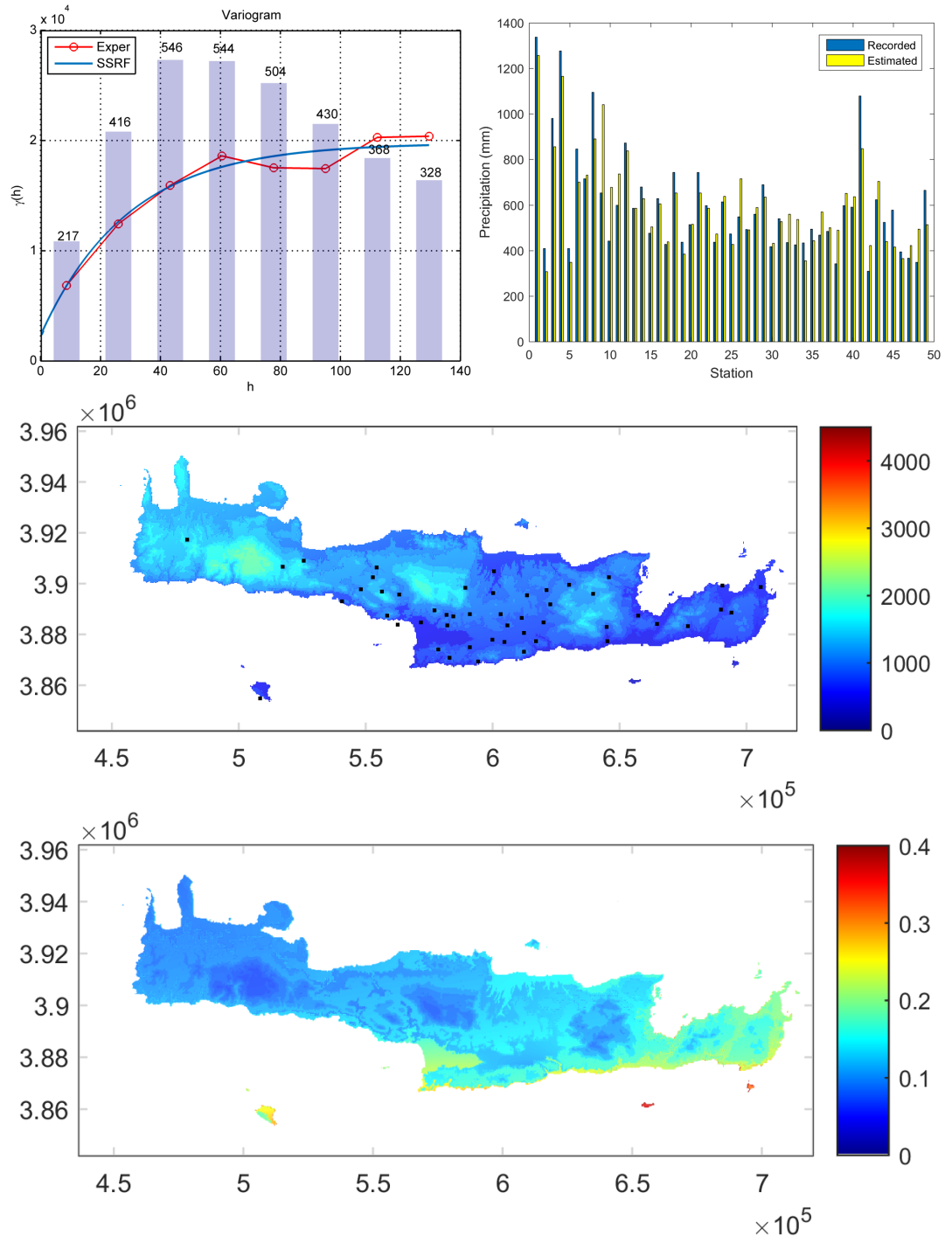


Figure 5: Year 1969 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.3427e + 03 \text{ mm}^2$, $\eta_0 = 4.3808e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 29.0385$.

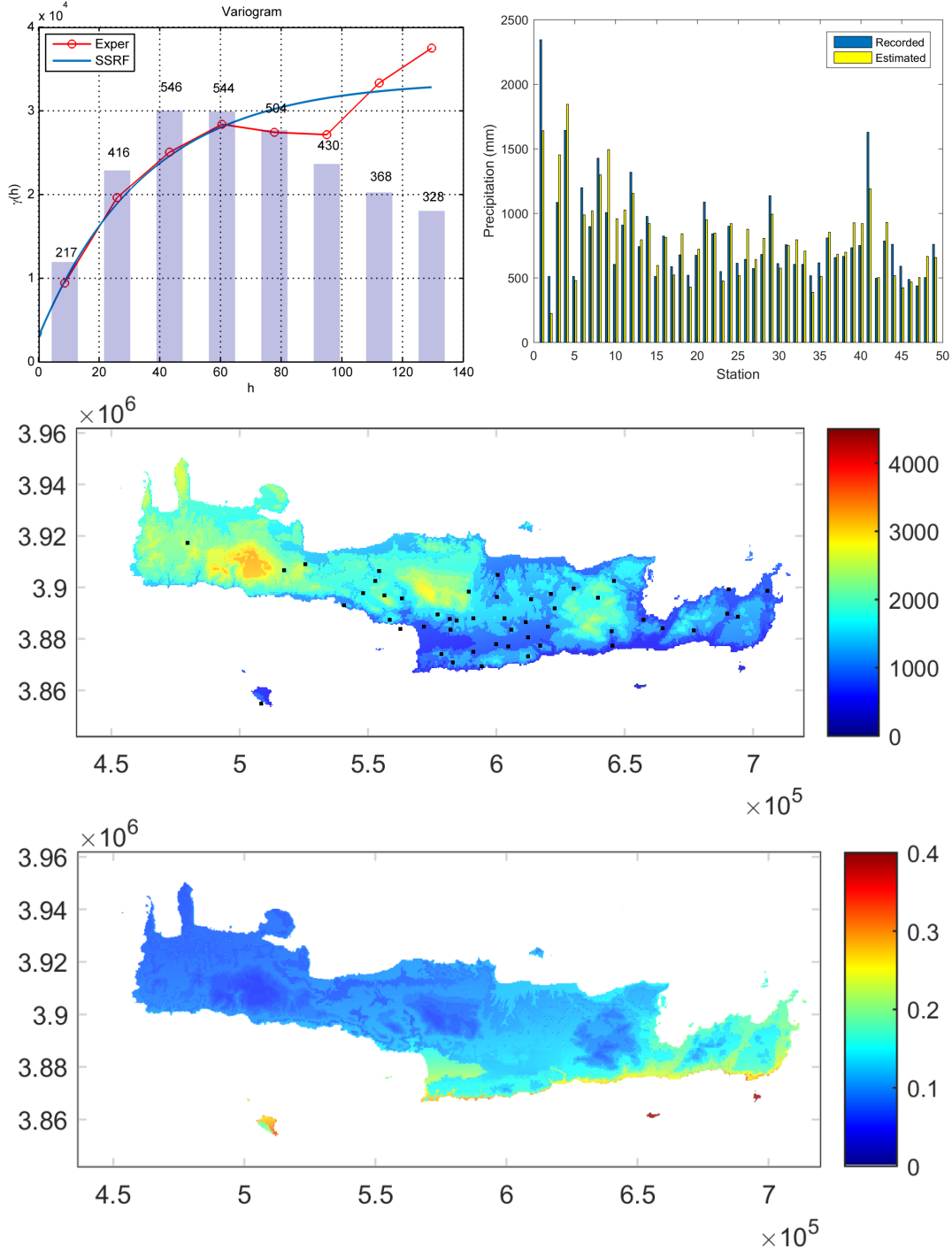


Figure 6: Year 1970 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.9939e + 03 \text{ mm}^2$, $\eta_0 = 7.6891e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 35.2855$.

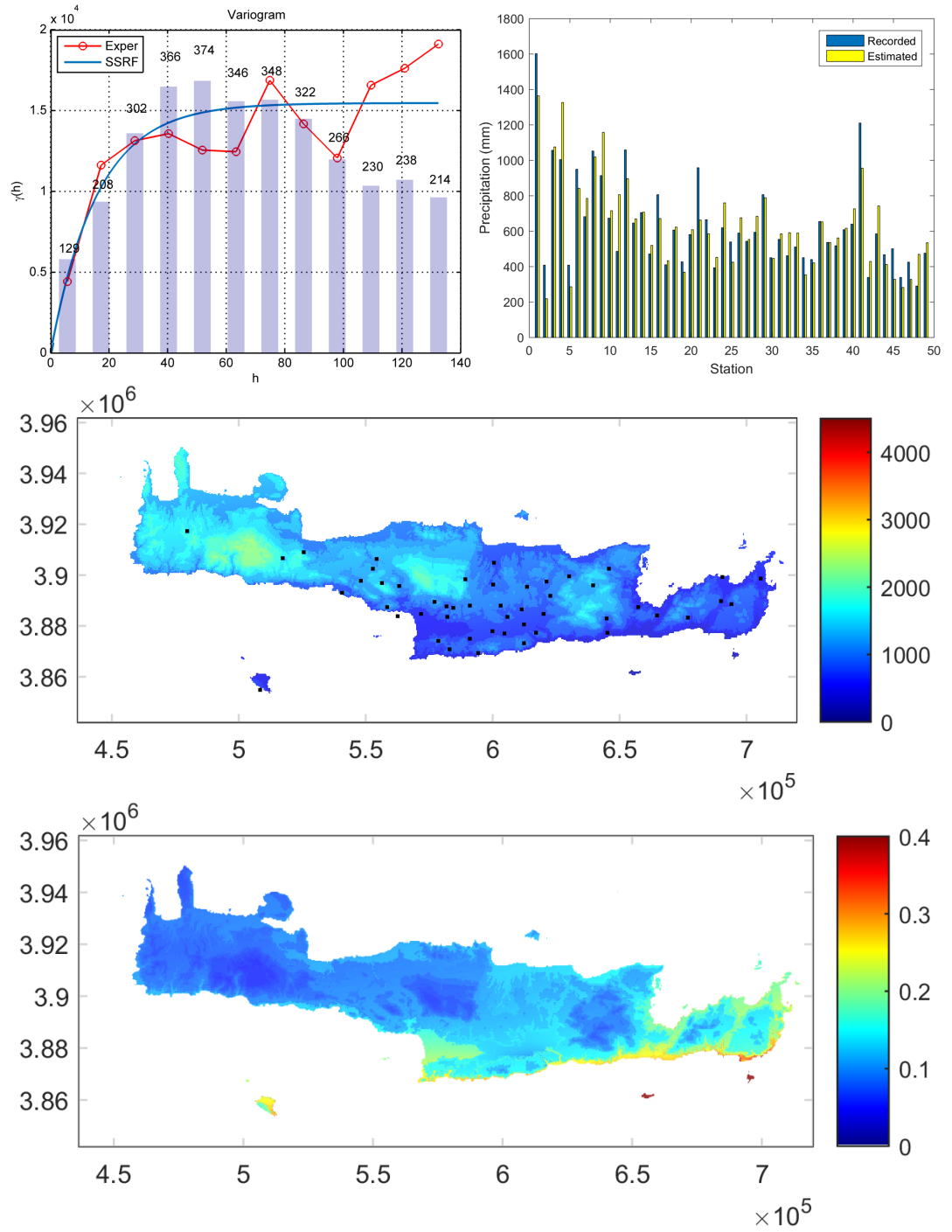


Figure 7: Year 1971 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 173.9258 \text{ mm}^2$, $\eta_0 = 1.9764e+05 \text{ mm}^2$, $\eta_1 = -0.8643$, $\xi = 9.9620$.

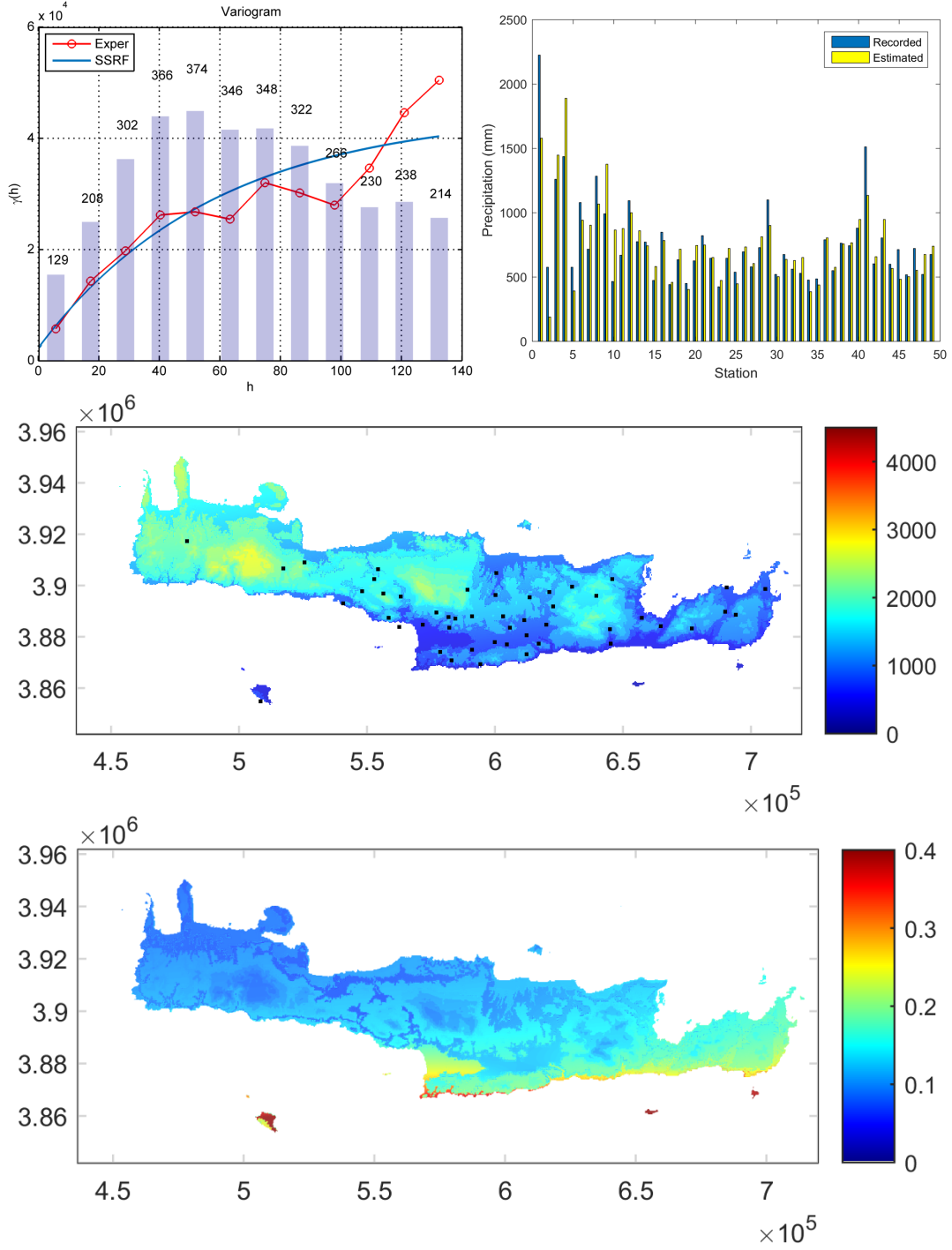


Figure 8: Year 1972 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.3323e + 03 \text{ mm}^2$, $\eta_0 = 1.0643e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 58.0892$.

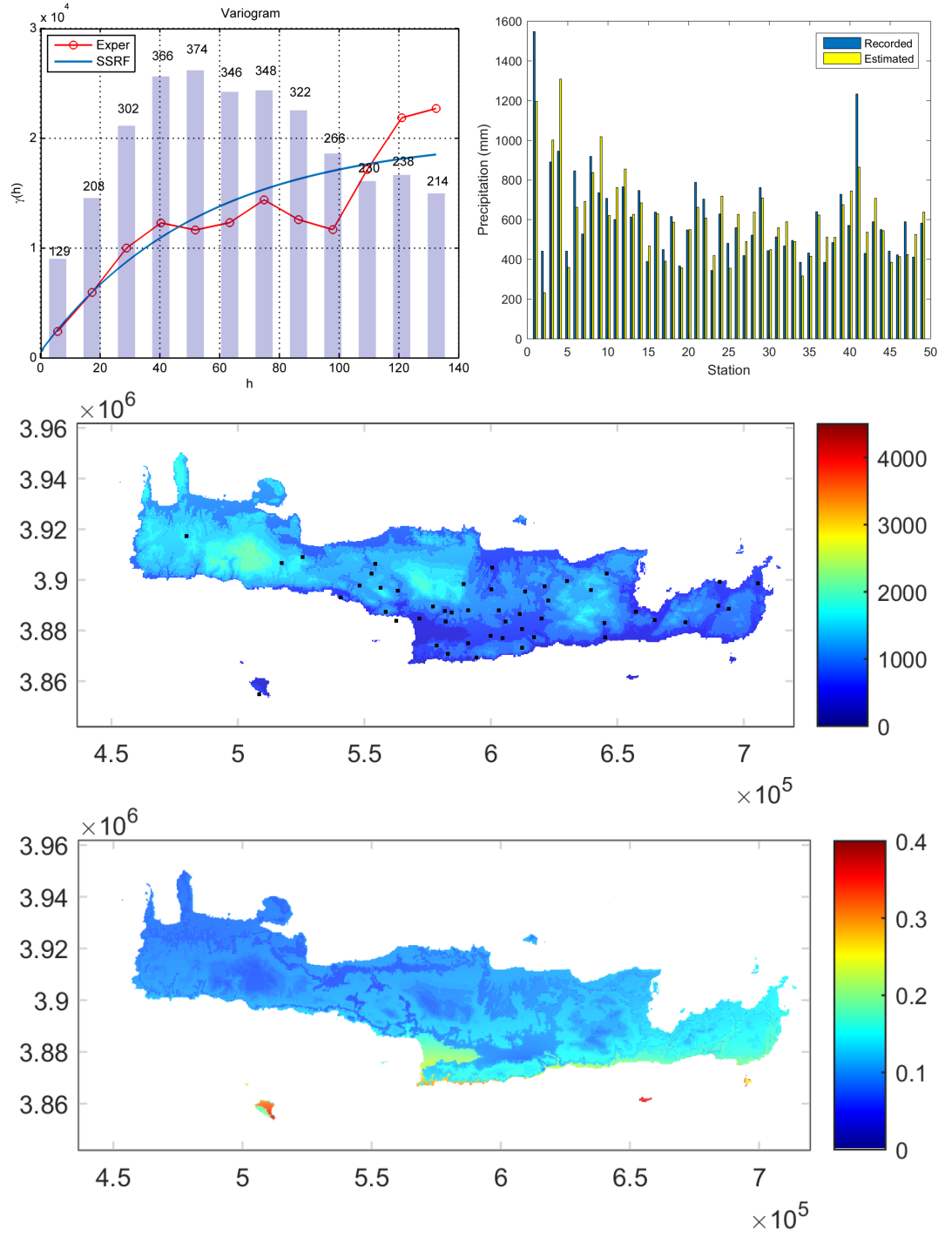


Figure 9: Year 1973 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 574.7301 \text{ mm}^2$, $\eta_0 = 4.9180e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 53.2182$.

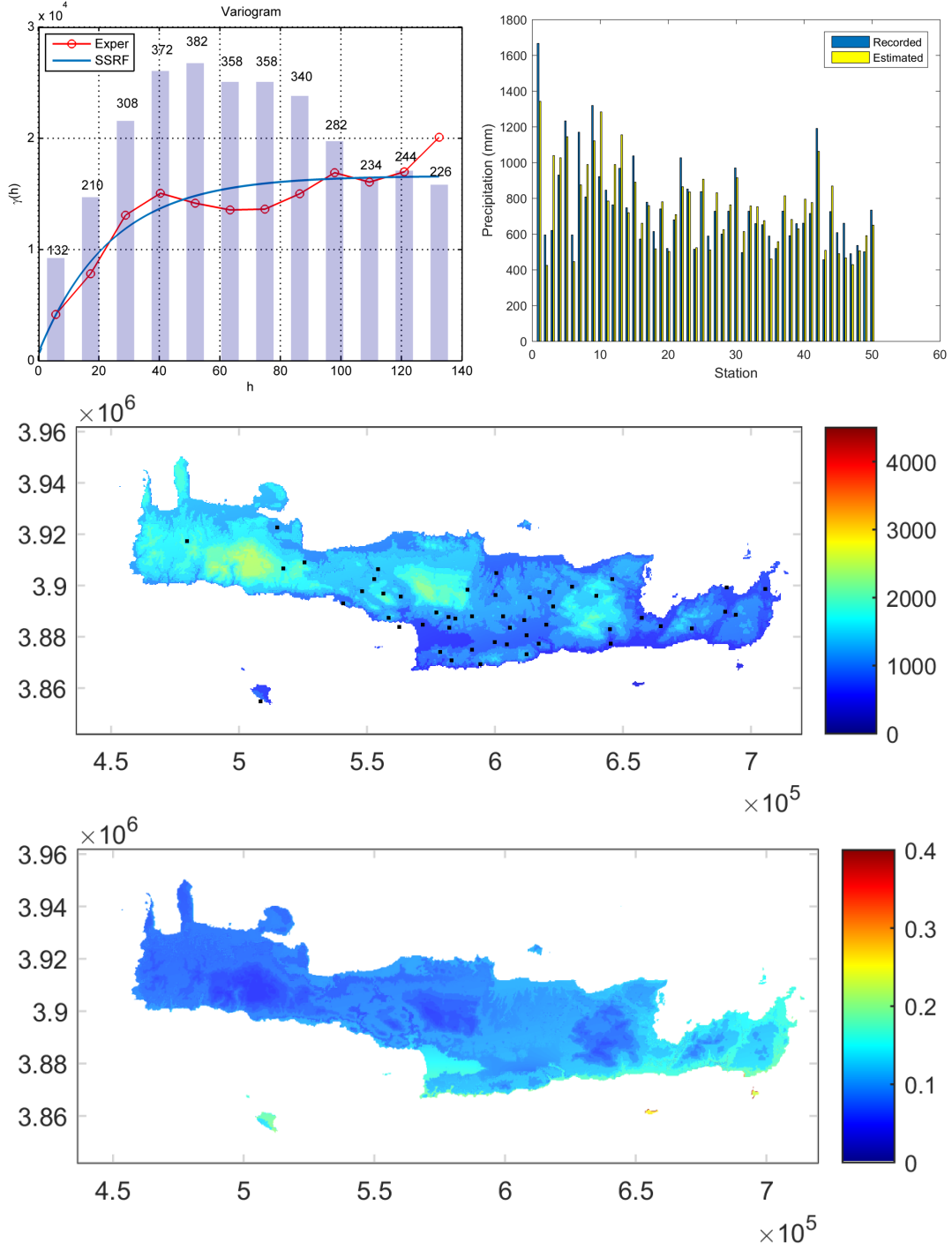


Figure 10: Year 1974 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 705.9062 \text{ mm}^2$, $\eta_0 = 4.0002e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 23.7951$.

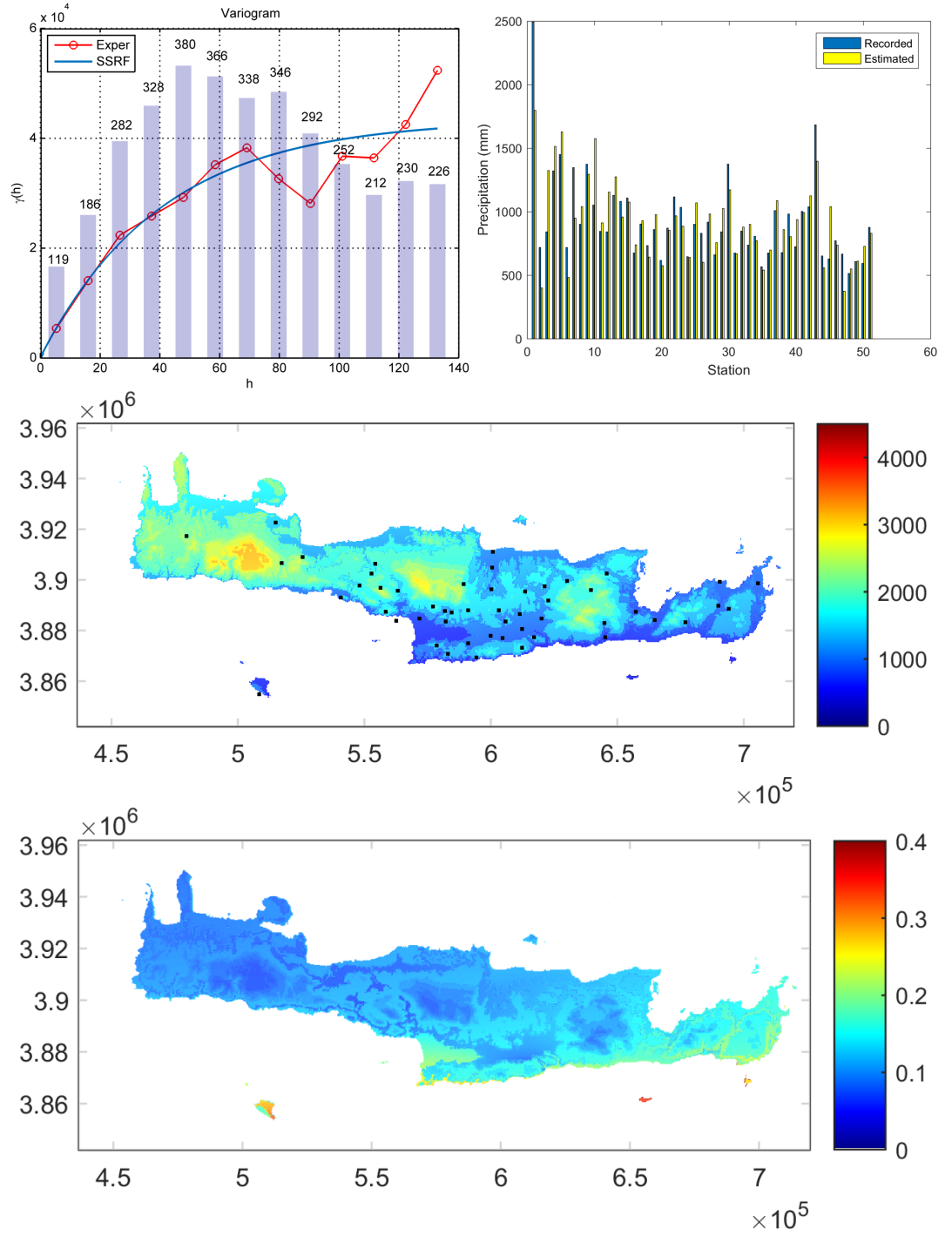


Figure 11: Year 1975 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 171.9771 \text{ mm}^2$, $\eta_0 = 1.0865e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 40.6173$.

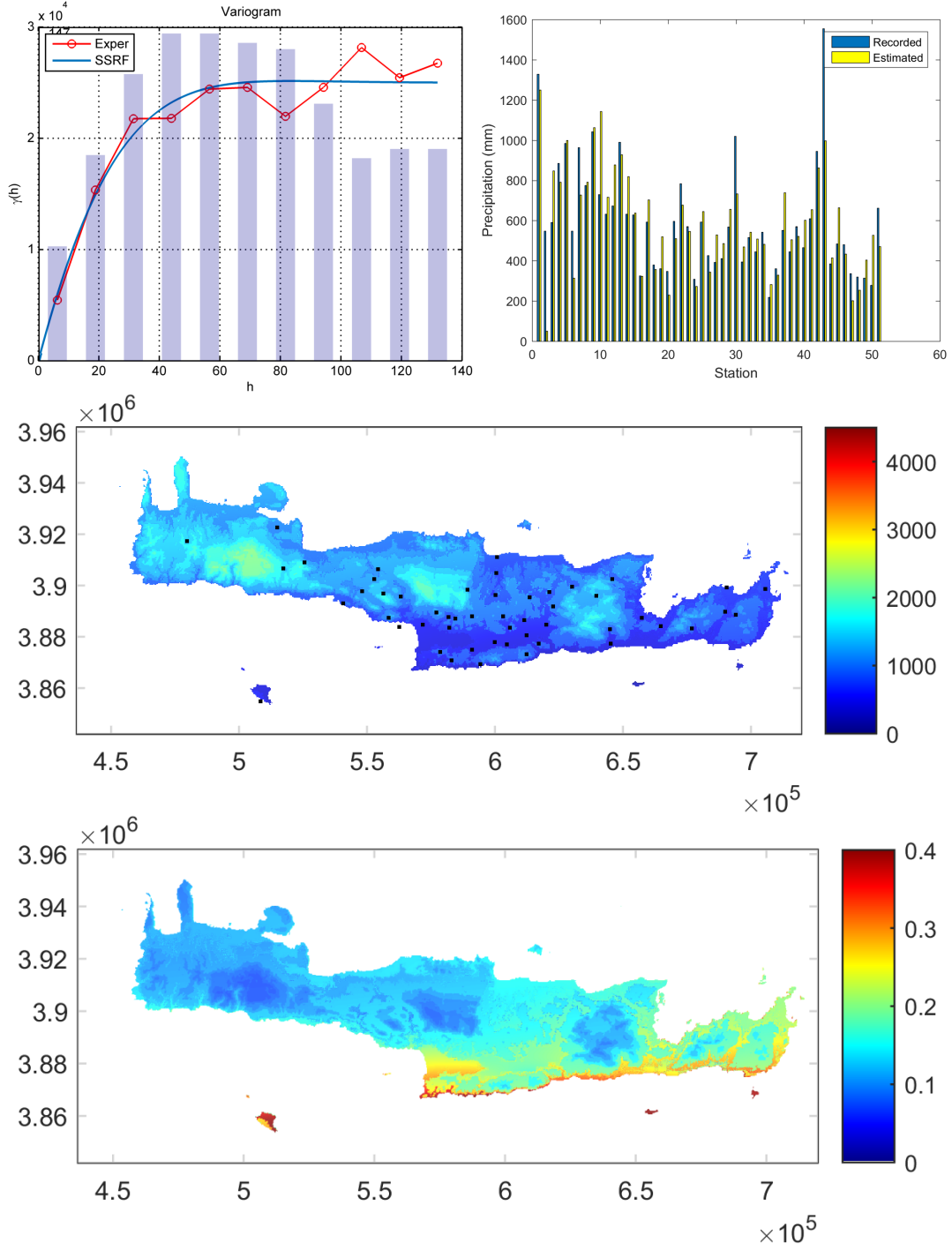


Figure 12: Year 1976 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.4903e + 04 \text{ mm}^2$, $\eta_0 = 4.1825e + 05 \text{ mm}^2$, $\eta_1 = -0.2280$, $\xi = 16.1434$.

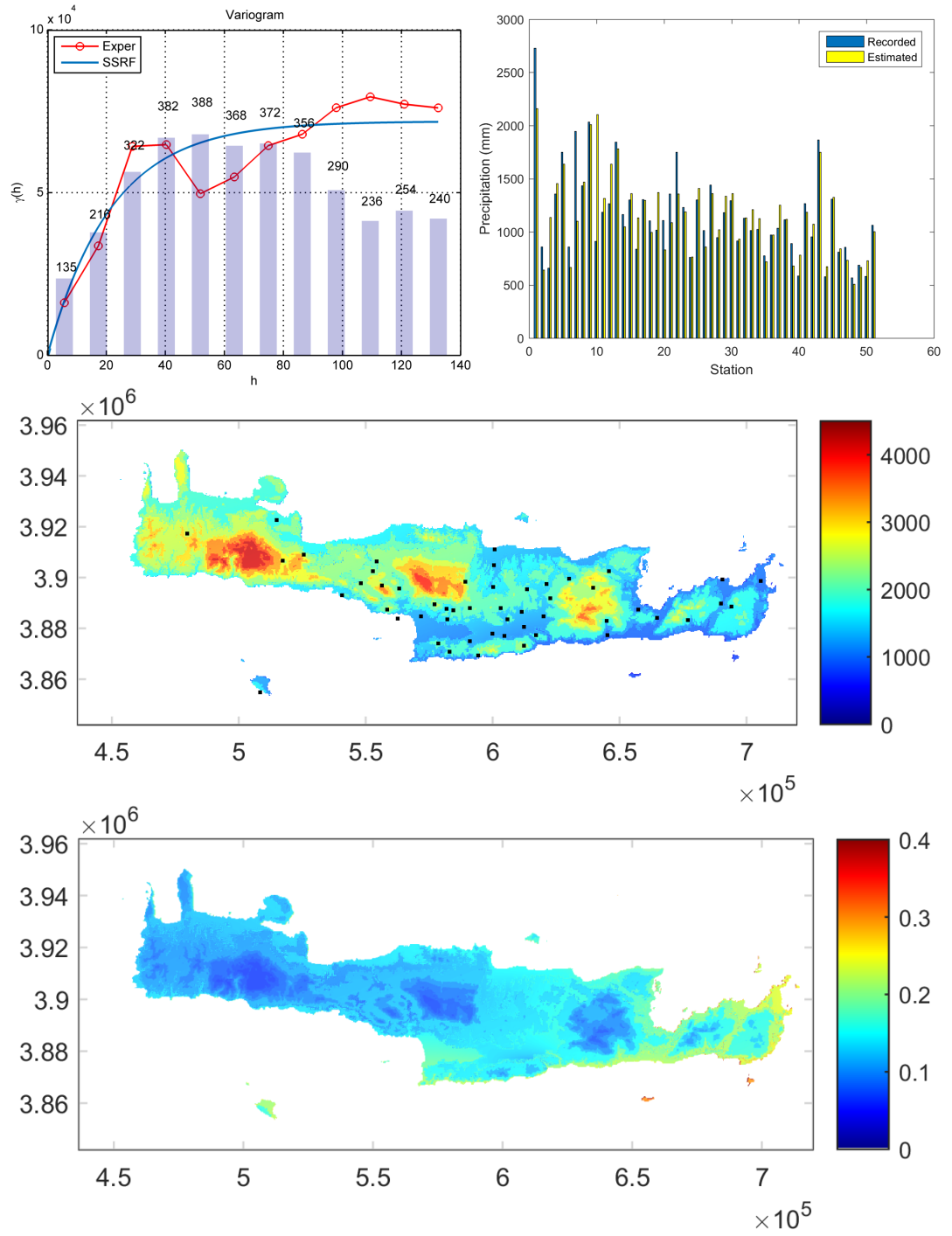


Figure 13: Year 1977 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 8.9749e - 04 \text{ mm}^2$, $\eta_0 = 1.8066e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 21.6769$.

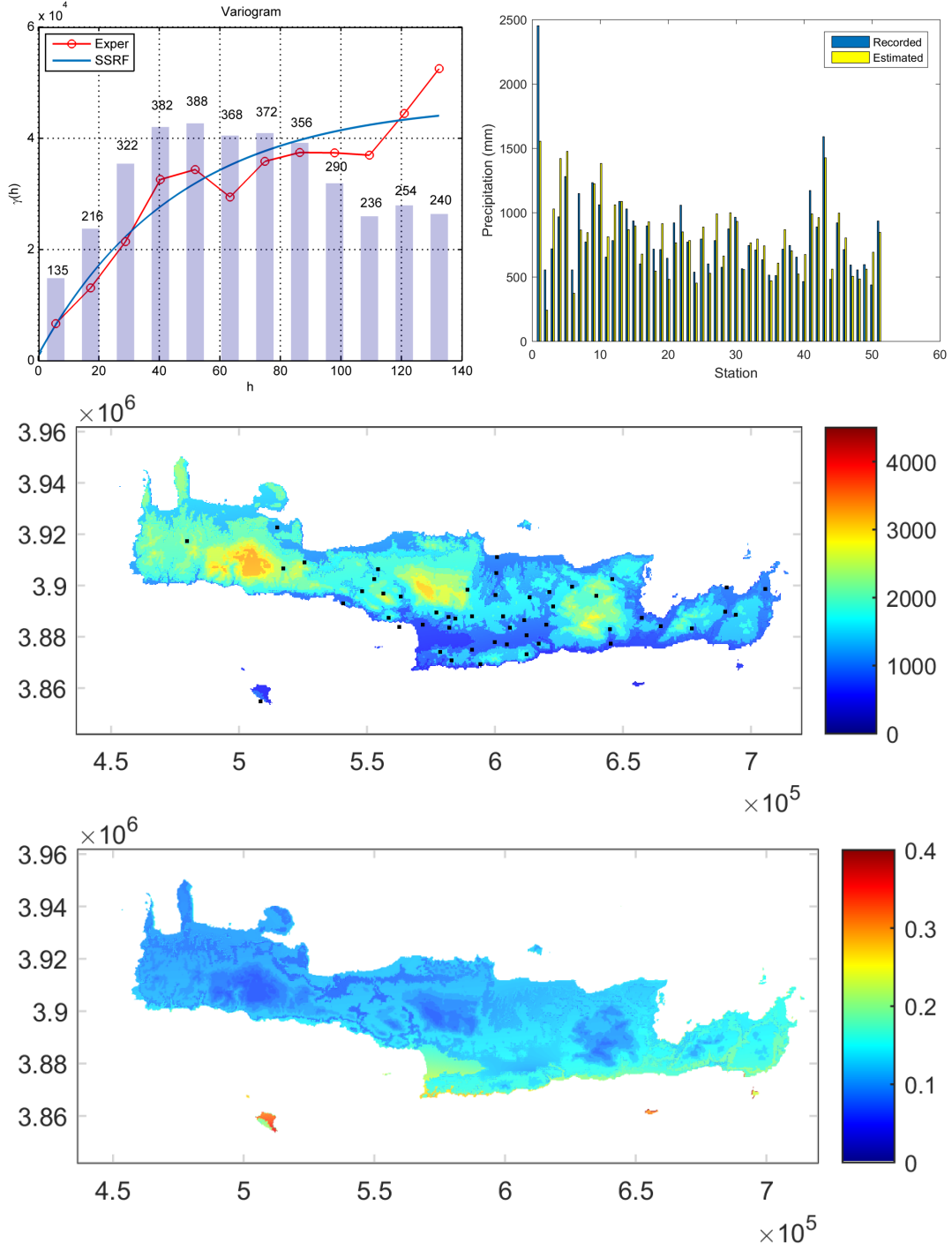


Figure 14: Year 1978 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.1239e + 03 \text{ mm}^2$, $\eta_0 = 1.1422e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 45.8321$.

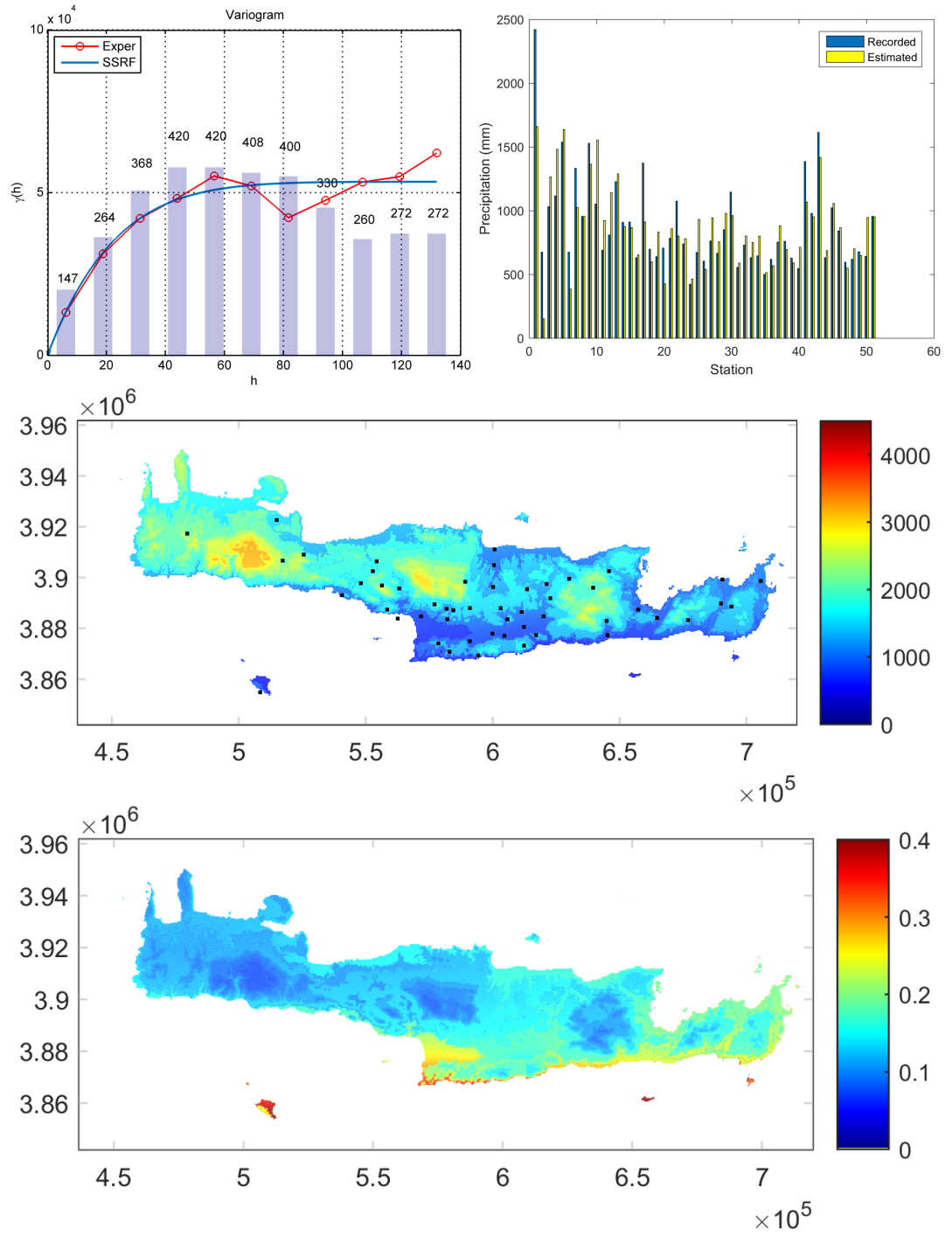


Figure 15: Year 1979 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 0.0024 \text{ mm}^2$, $\eta_0 = 1.1146e+06 \text{ mm}^2$, $\eta_1 = 0.7711$, $\xi = 18.5273$.

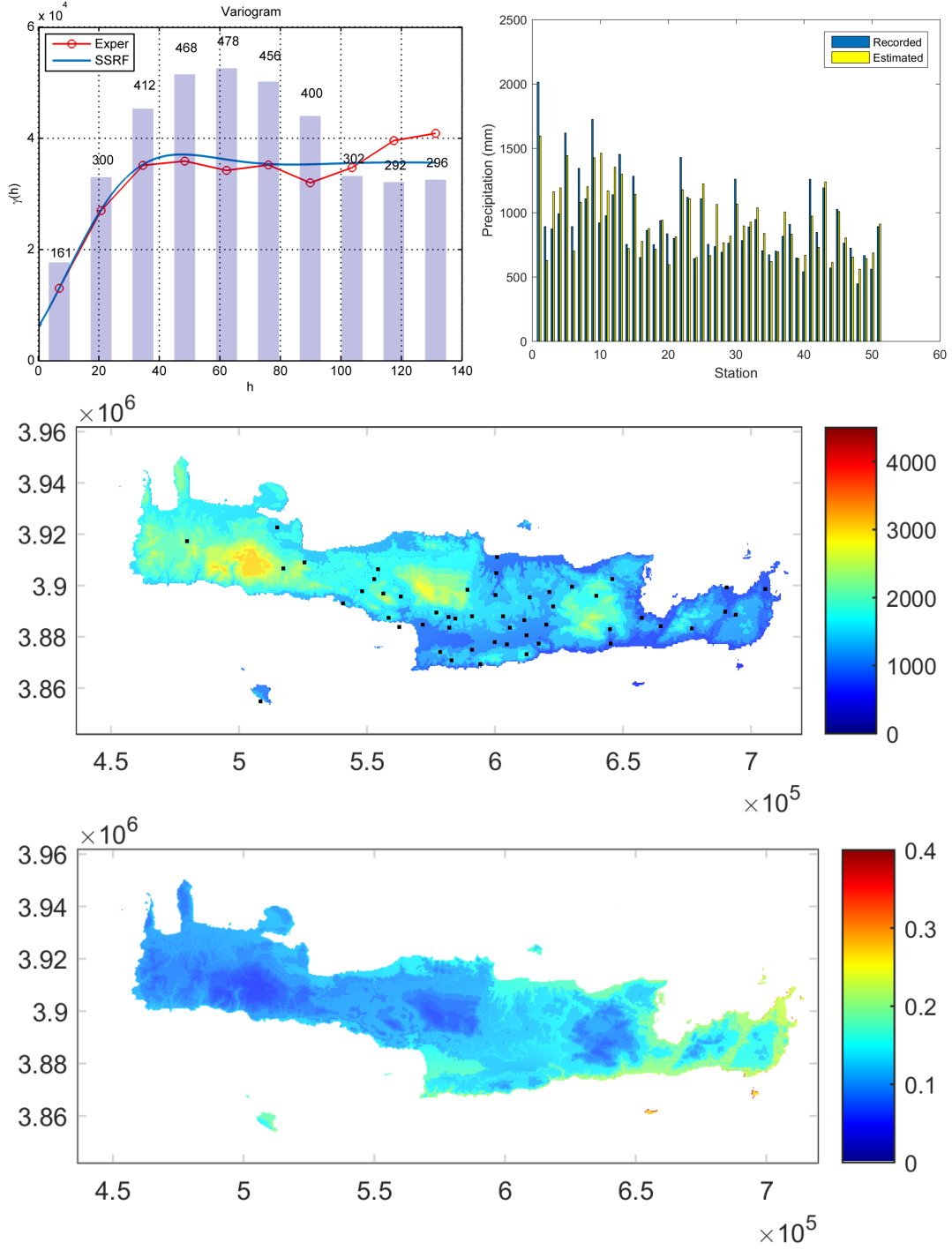


Figure 16: Year 1980 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 6.1494e + 03 \text{ mm}^2$, $\eta_0 = 2.3375e + 05 \text{ mm}^2$, $\eta_1 = -1.6003$, $\xi = 10.7215$.

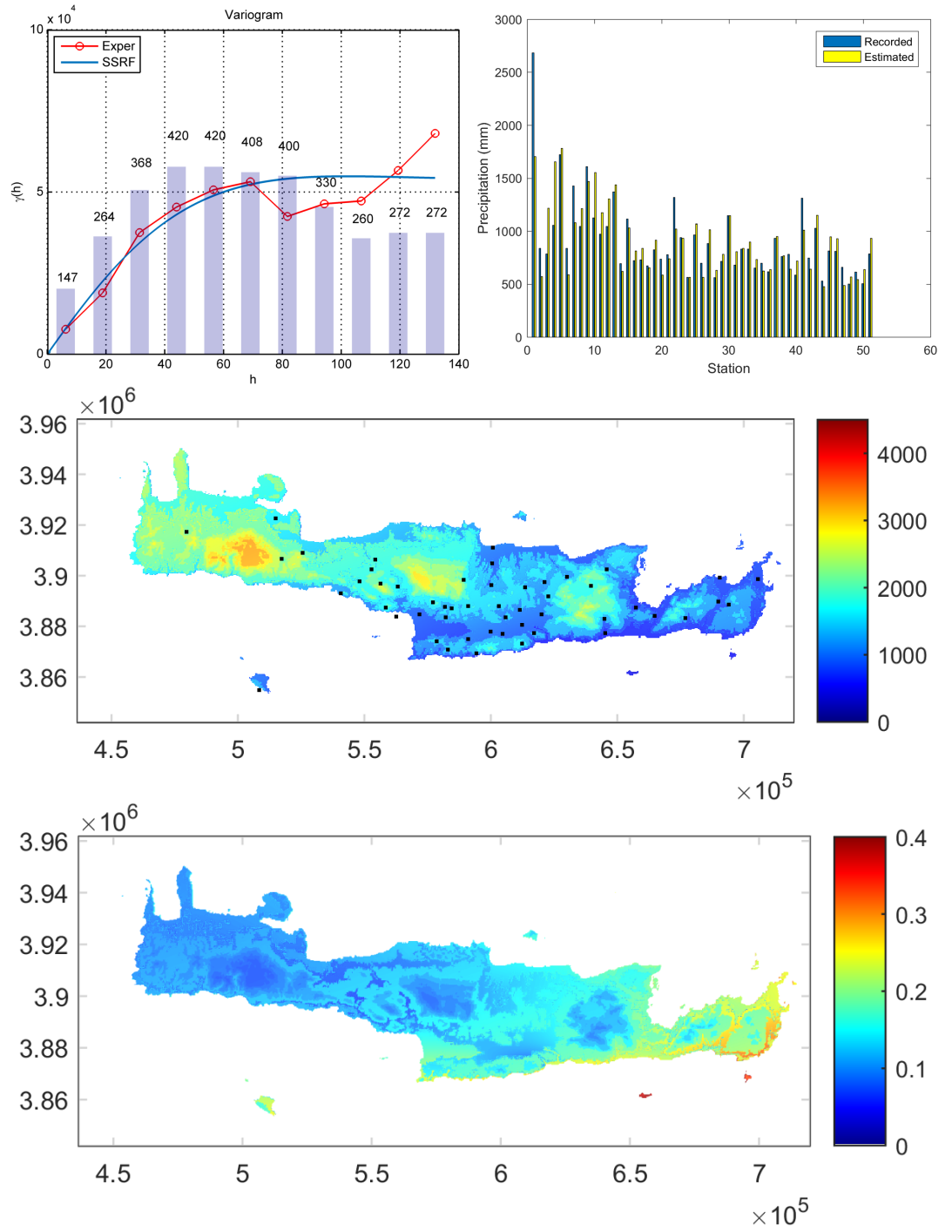


Figure 17: Year 1981 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.0419e - 04 \text{ mm}^2$, $\eta_0 = 6.9971e + 05 \text{ mm}^2$, $\eta_1 = -0.9333$, $\xi = 21.7992$.

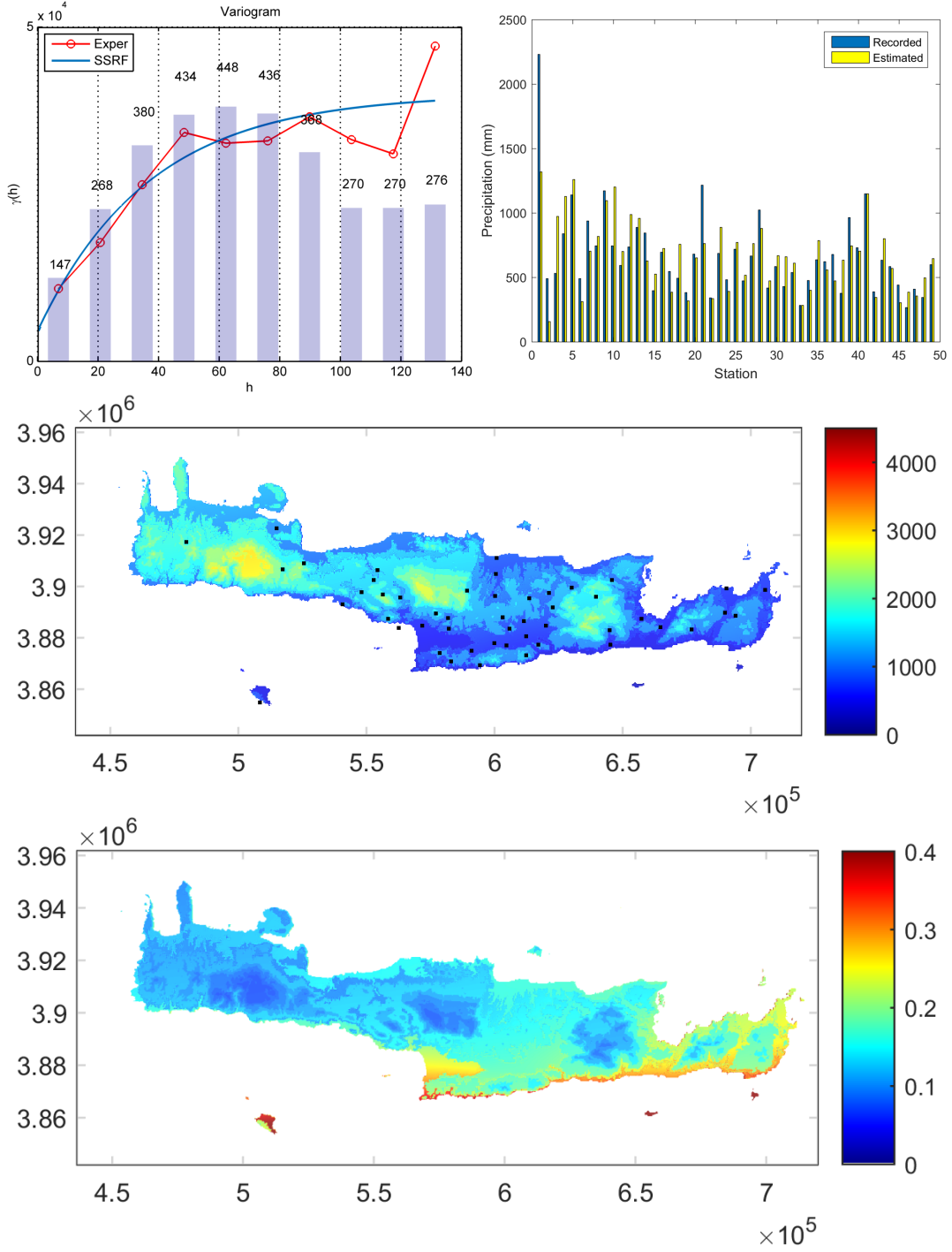


Figure 18: Year 1982 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 4.5221e + 03 \text{ mm}^2$, $\eta_0 = 8.9089e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 36.7636$.

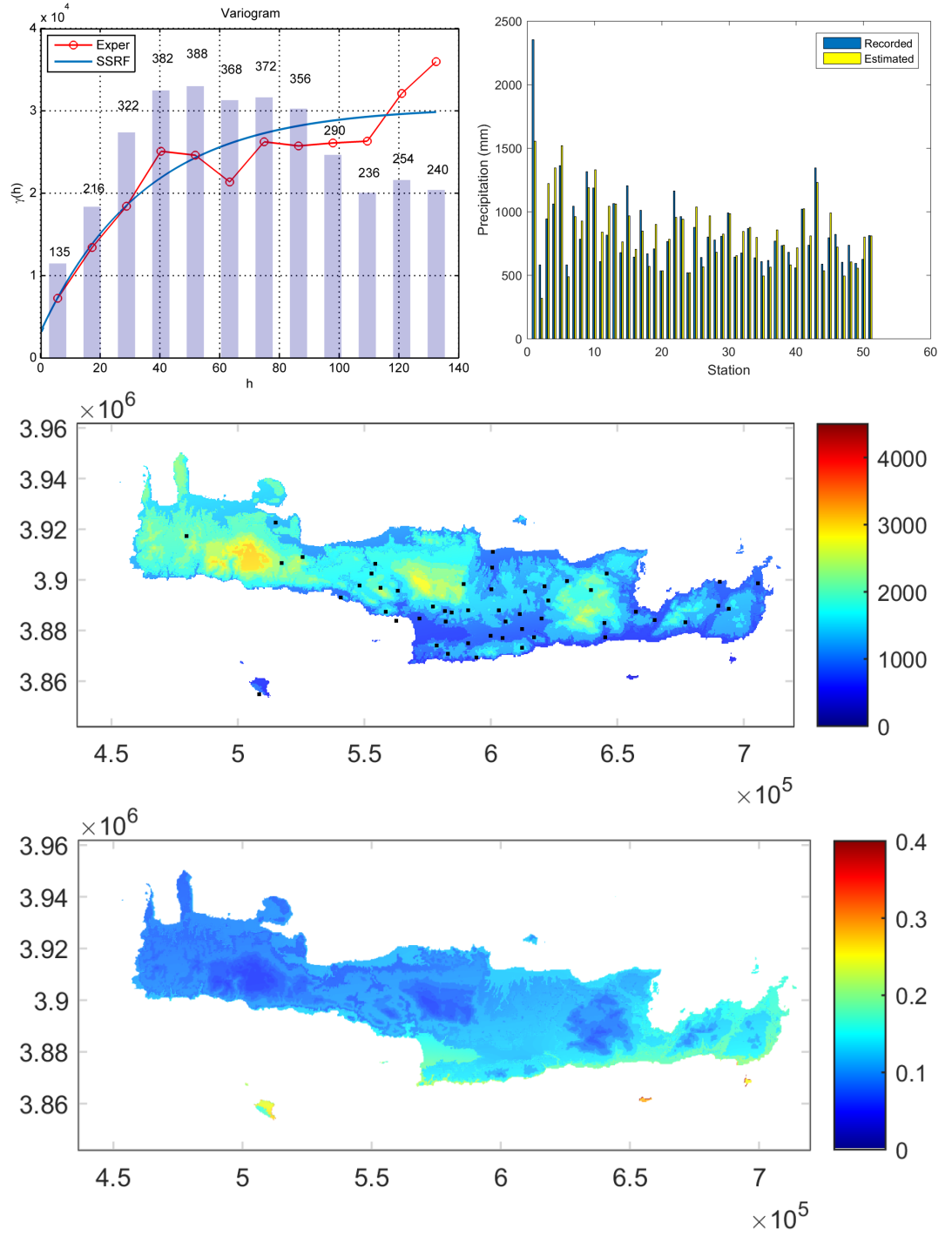


Figure 19: Year 1983 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 3.1588e + 03 \text{ mm}^2$, $\eta_0 = 6.8585e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 34.7984$.

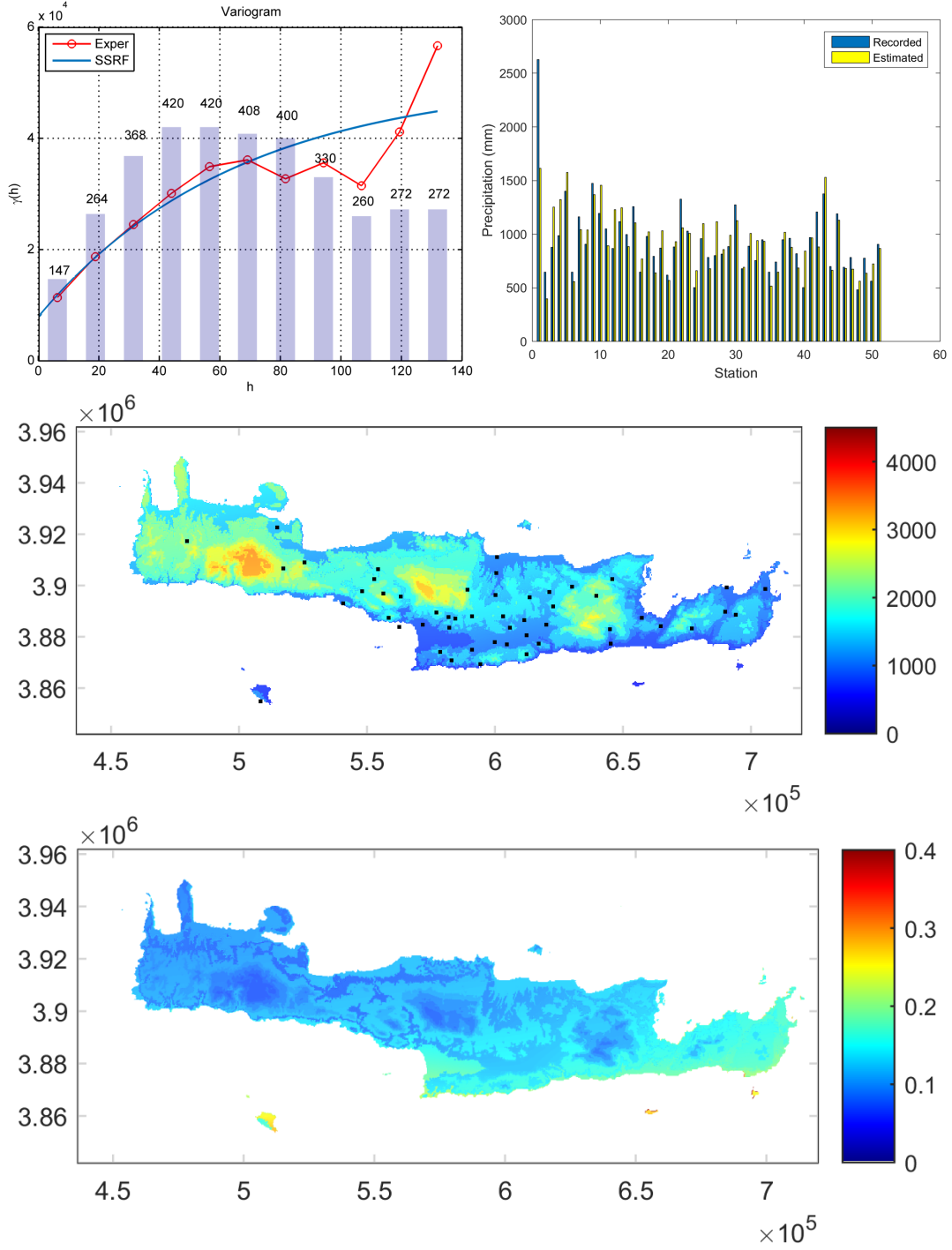


Figure 20: Year 1984 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 7.9446e + 03 \text{ mm}^2$, $\eta_0 = 1.0700e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 65.4678$.

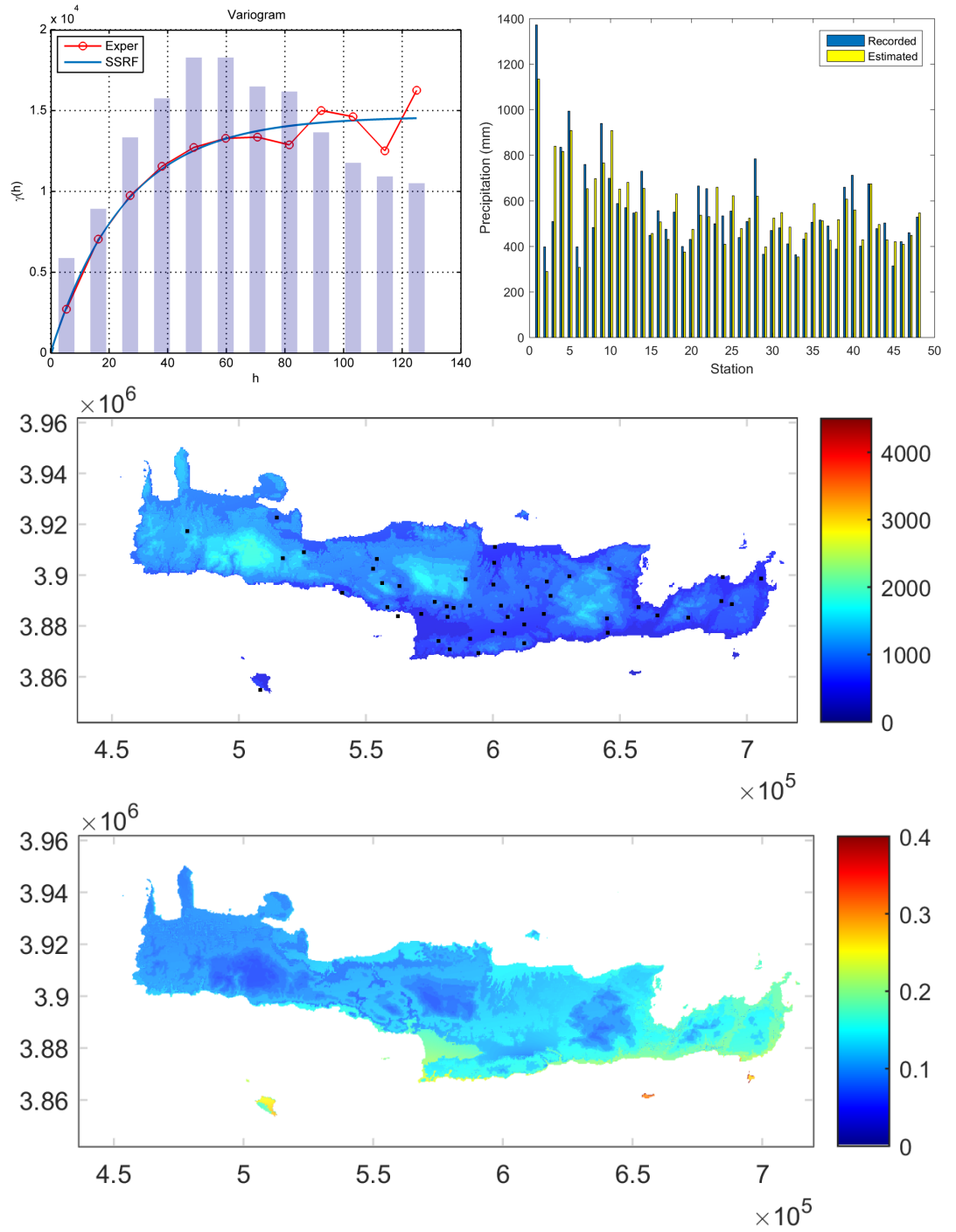


Figure 21: Year 1985 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 4.5399e - 05 \text{ mm}^2$, $\eta_0 = 3.6757e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 25.2580$.

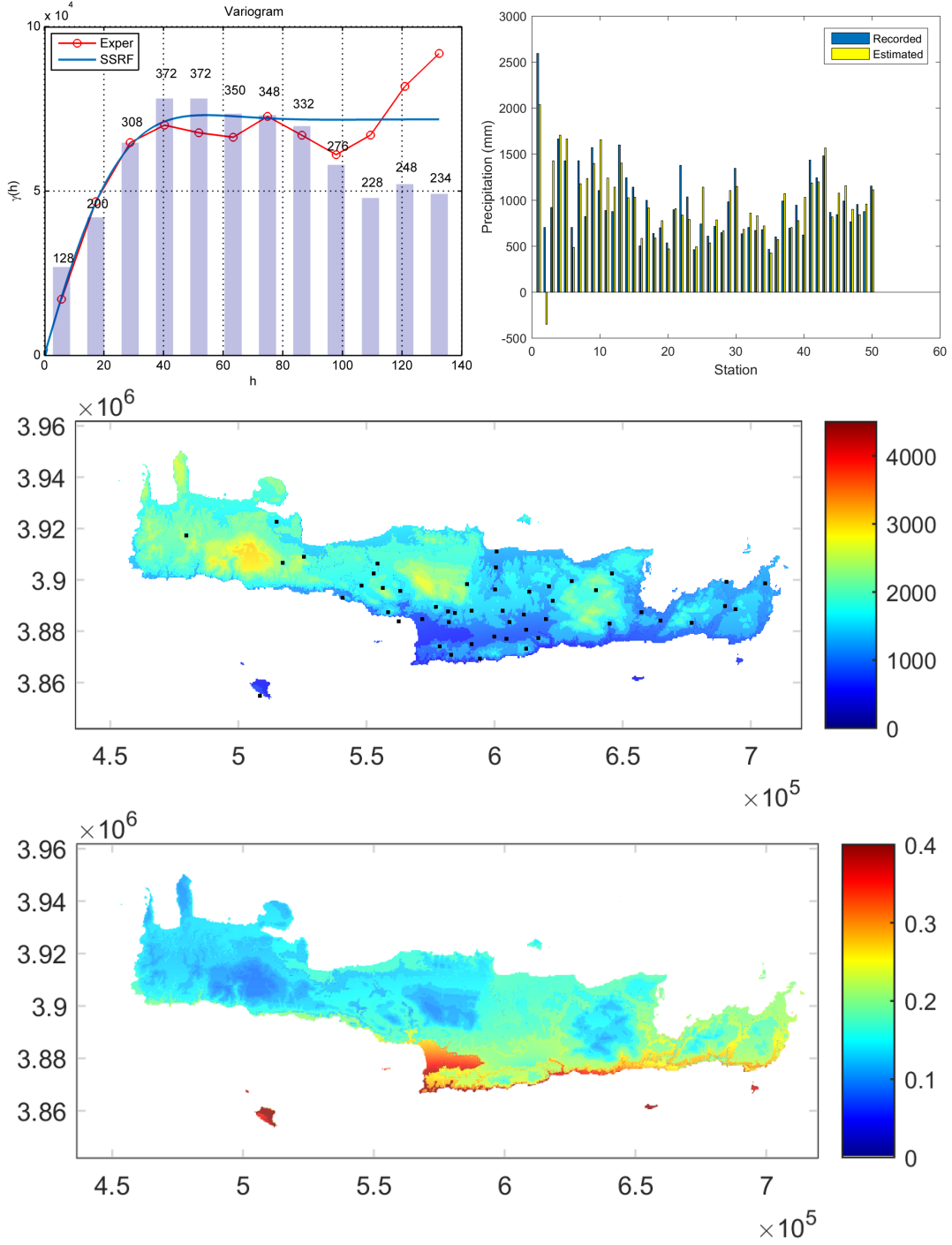


Figure 22: Year 1986 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 0.0018 \text{ mm}^2$, $\eta_0 = 9.0470e + 05 \text{ mm}^2$, $\eta_1 = -0.9929$, $\xi = 11.5590$.

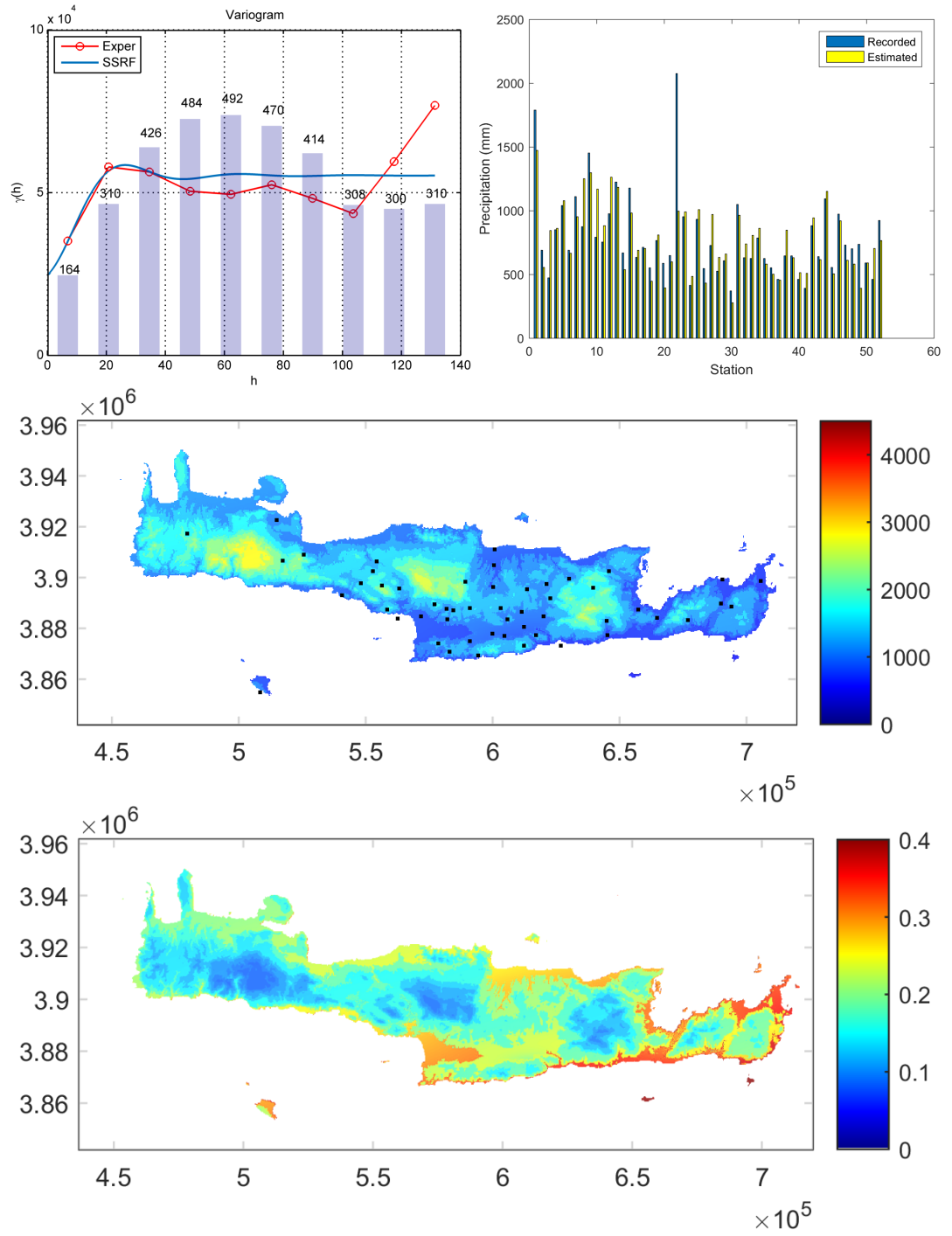


Figure 23: Year 1987 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.4680e + 04 \text{ mm}^2$, $\eta_0 = 1.2462e + 05 \text{ mm}^2$, $\eta_1 = -1.8947$, $\xi = 5.9949$.

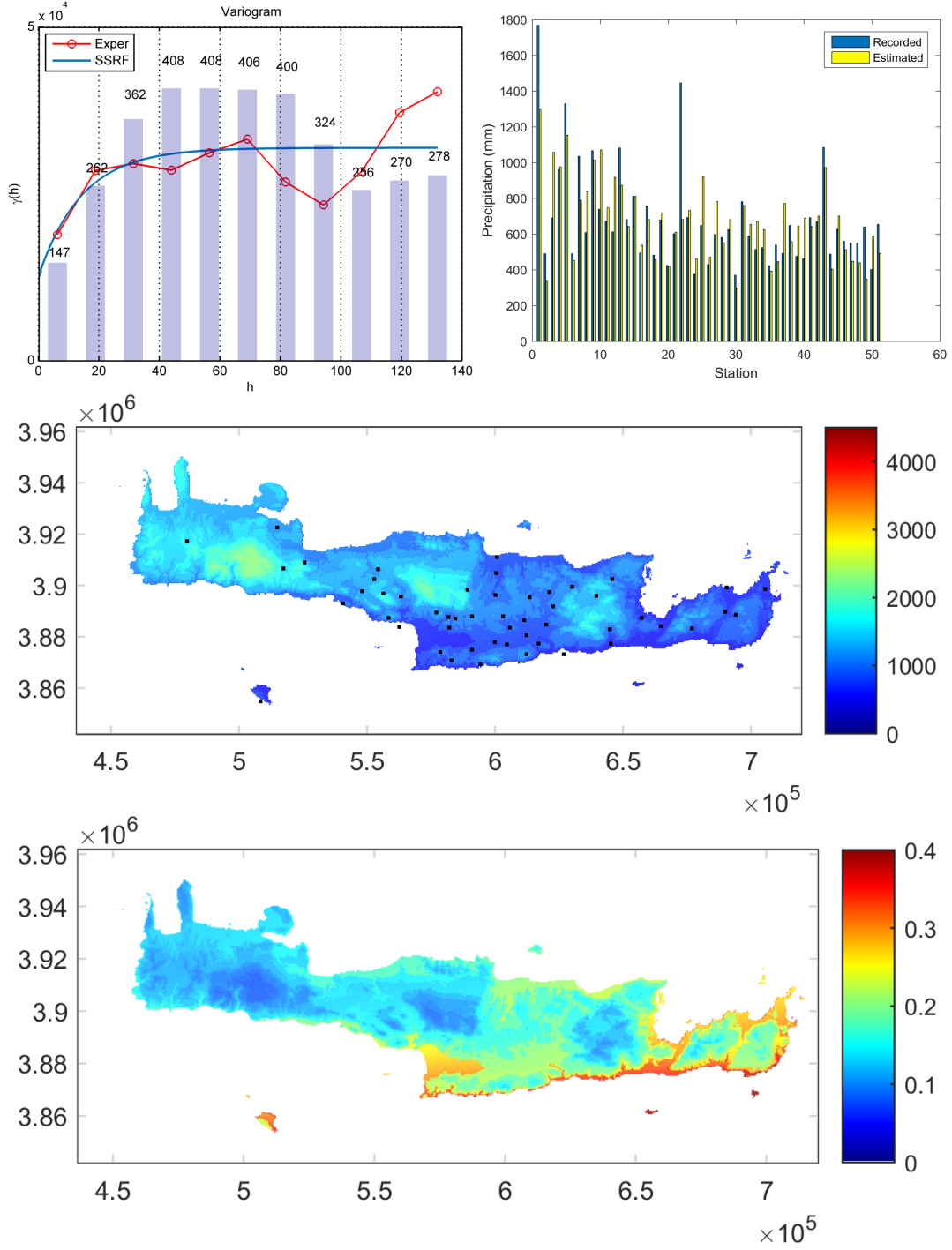


Figure 24: Year 1988 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.2608e + 04 \text{ mm}^2$, $\eta_0 = 4.8478e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 14.3913$.

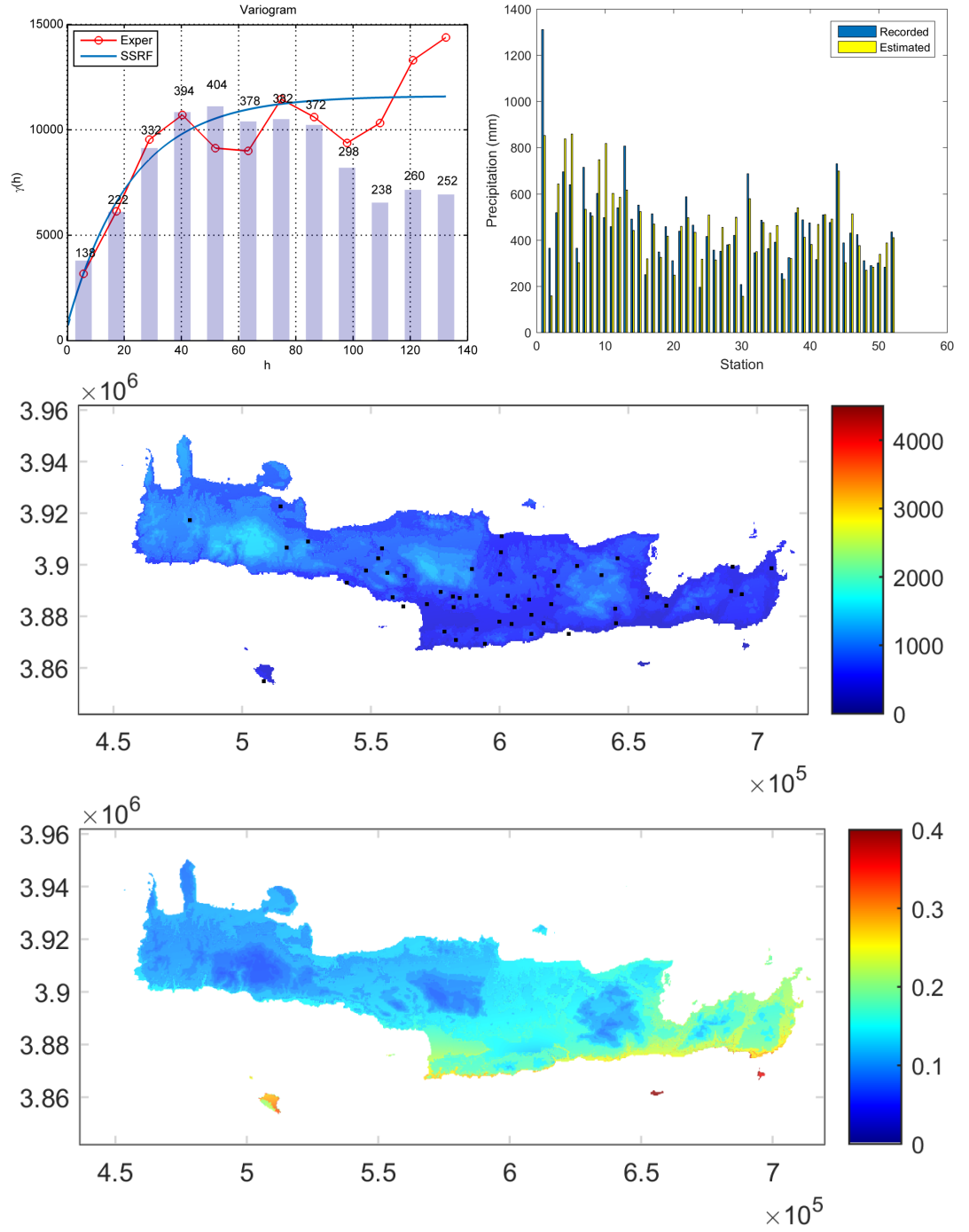


Figure 25: Year 1989 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 717.0774 \text{ mm}^2$, $\eta_0 = 2.7366e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 22.3578$.

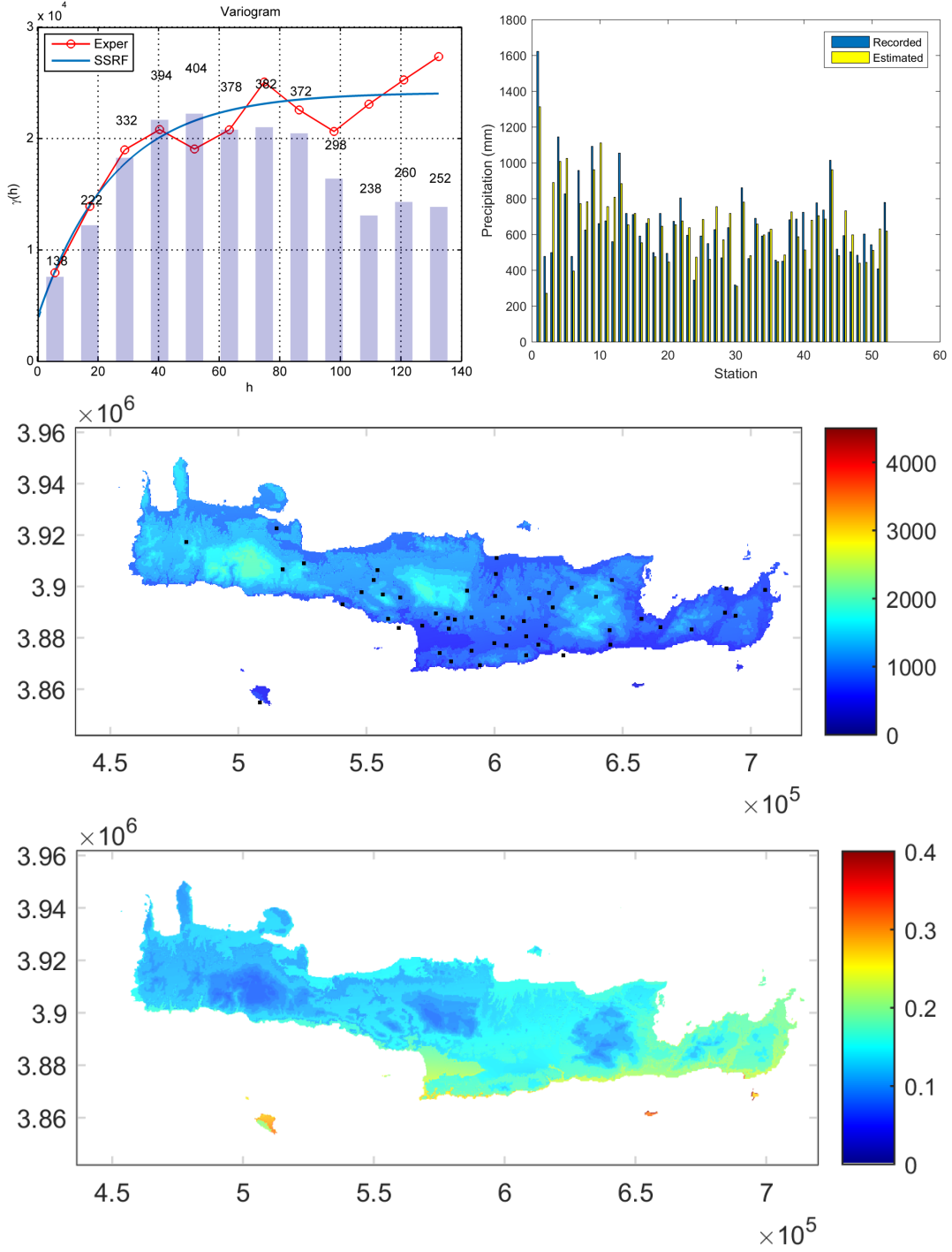


Figure 26: Year 1990 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 3.9497e + 03 \text{ mm}^2$, $\eta_0 = 5.0704e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 25.0299$.

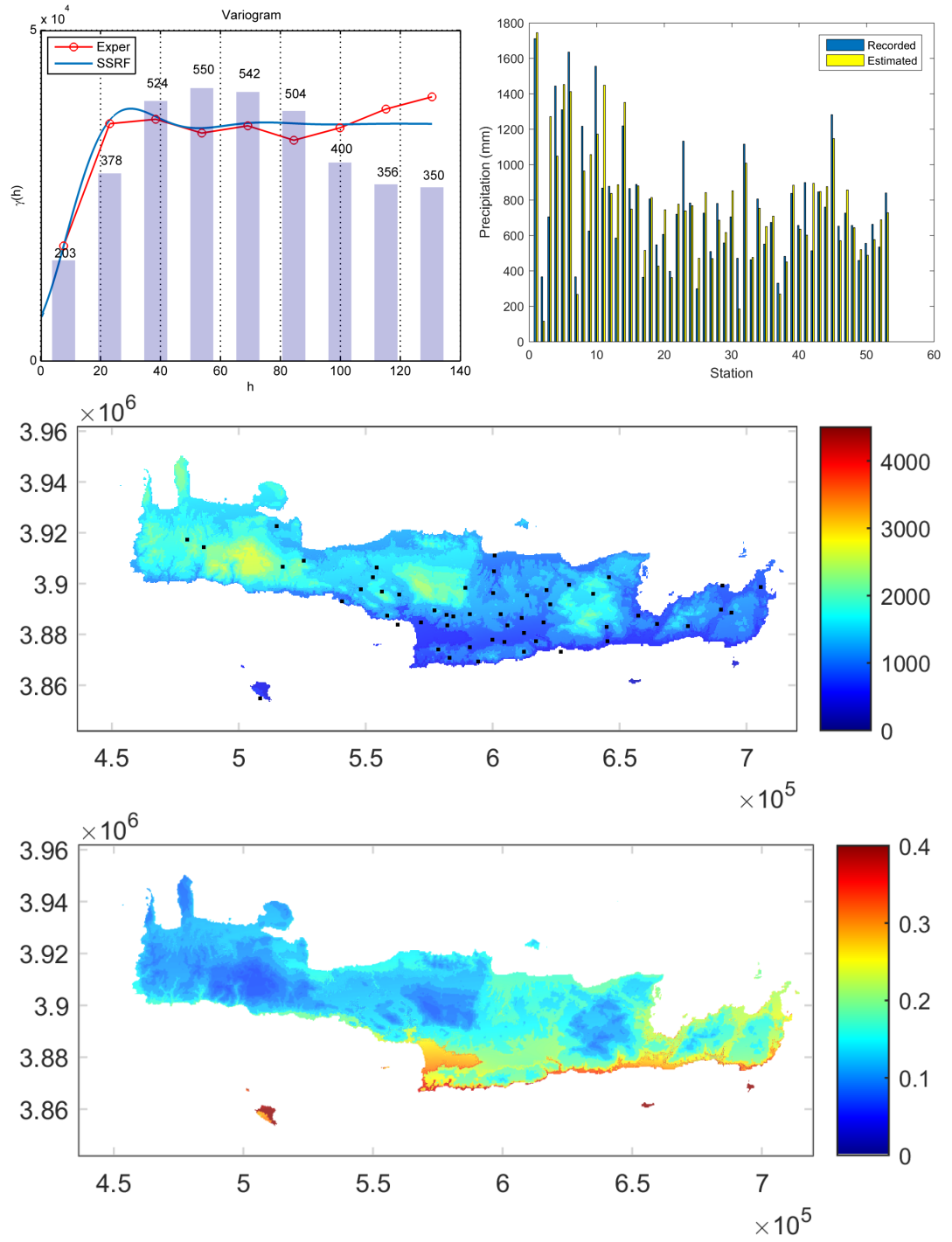


Figure 27: Year 1991 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 6.6308e + 03 \text{ mm}^2$, $\eta_0 = 1.6522e + 05 \text{ mm}^2$, $\eta_1 = -1.7973$, $\xi = 6.8429$.

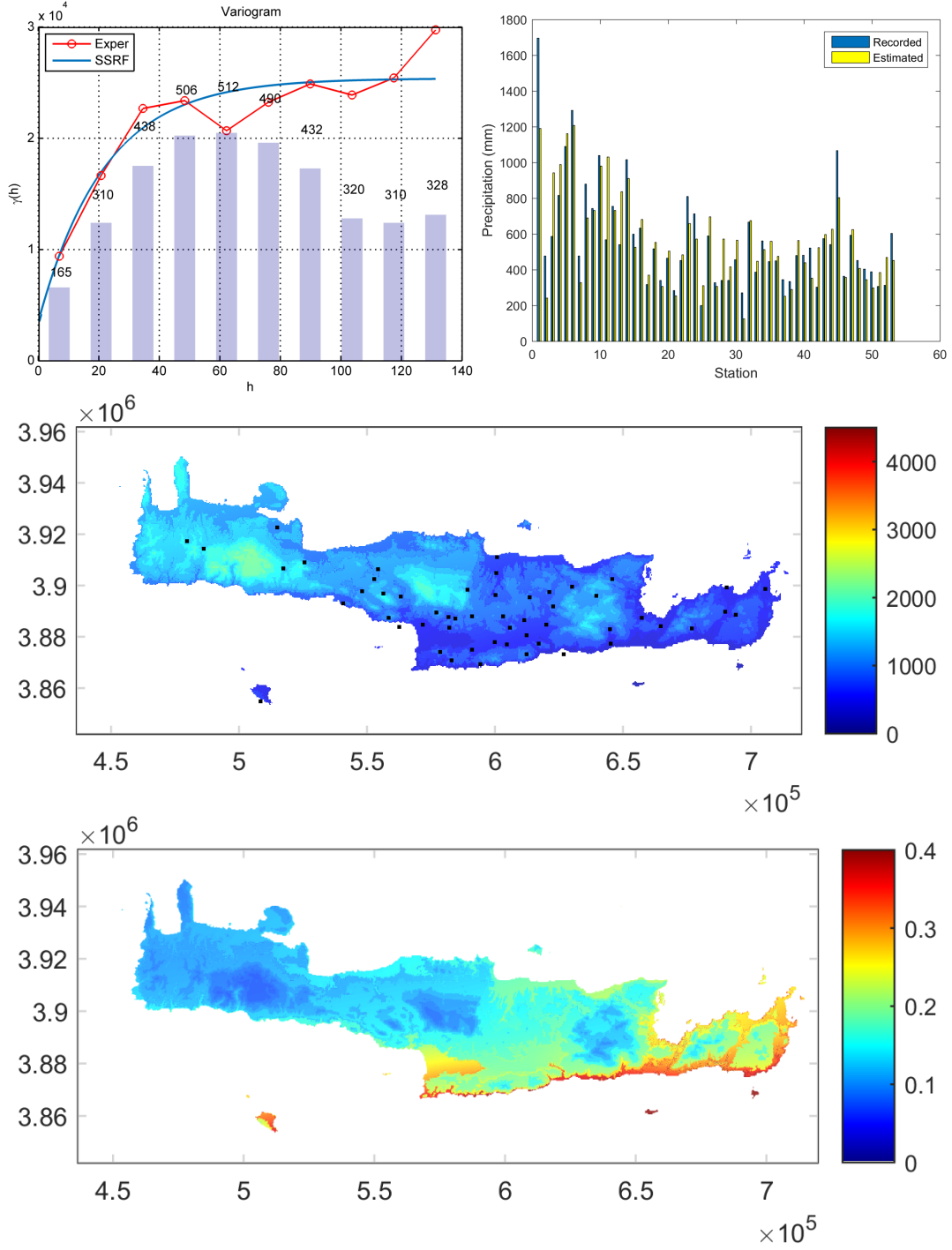


Figure 28: Year 1992 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 3.5278e + 03 \text{ mm}^2$, $\eta_0 = 5.4916e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 21.7160$.

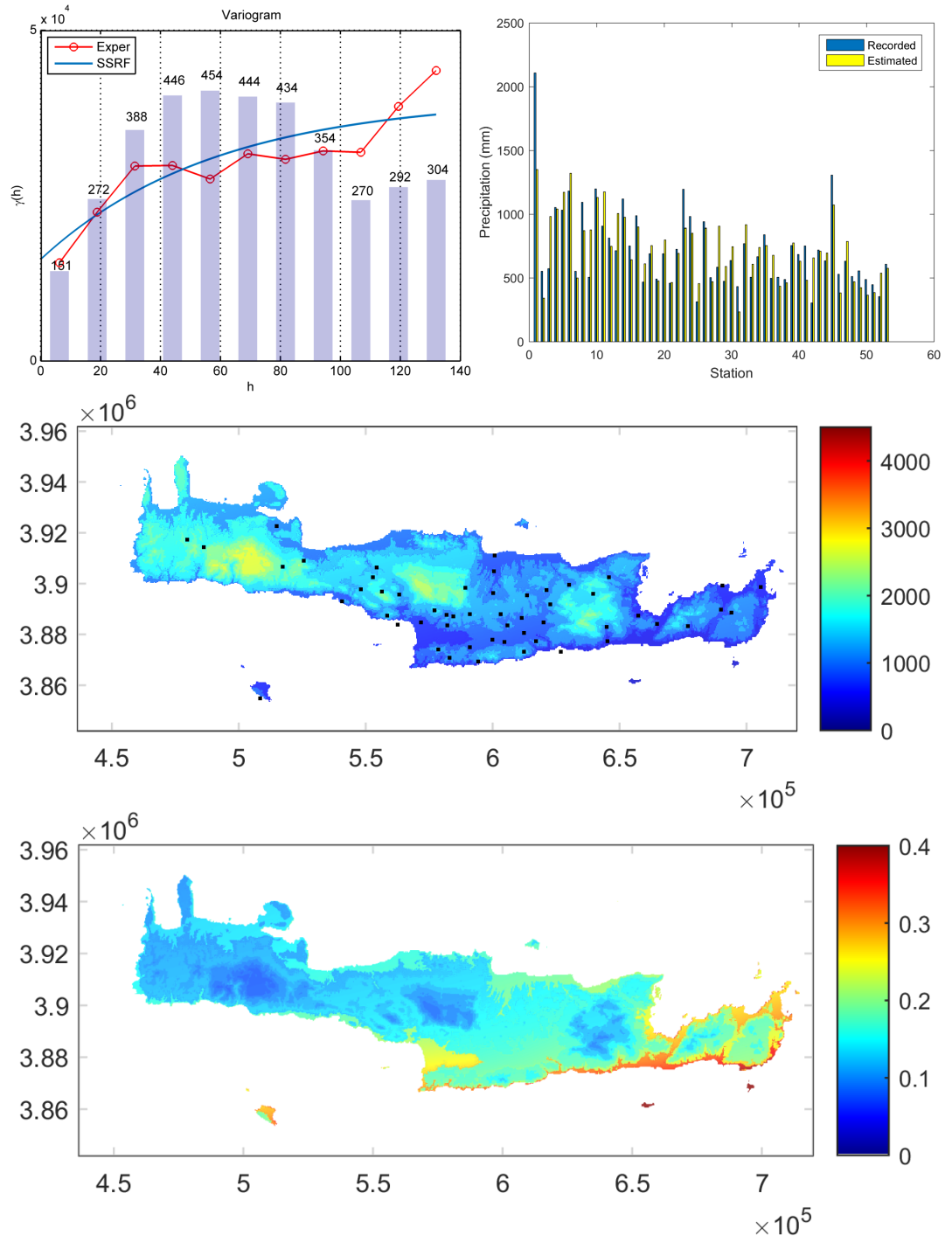


Figure 29: Year 1993 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.5363e + 04 \text{ mm}^2$, $\eta_0 = 6.1321e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 58.0159$.

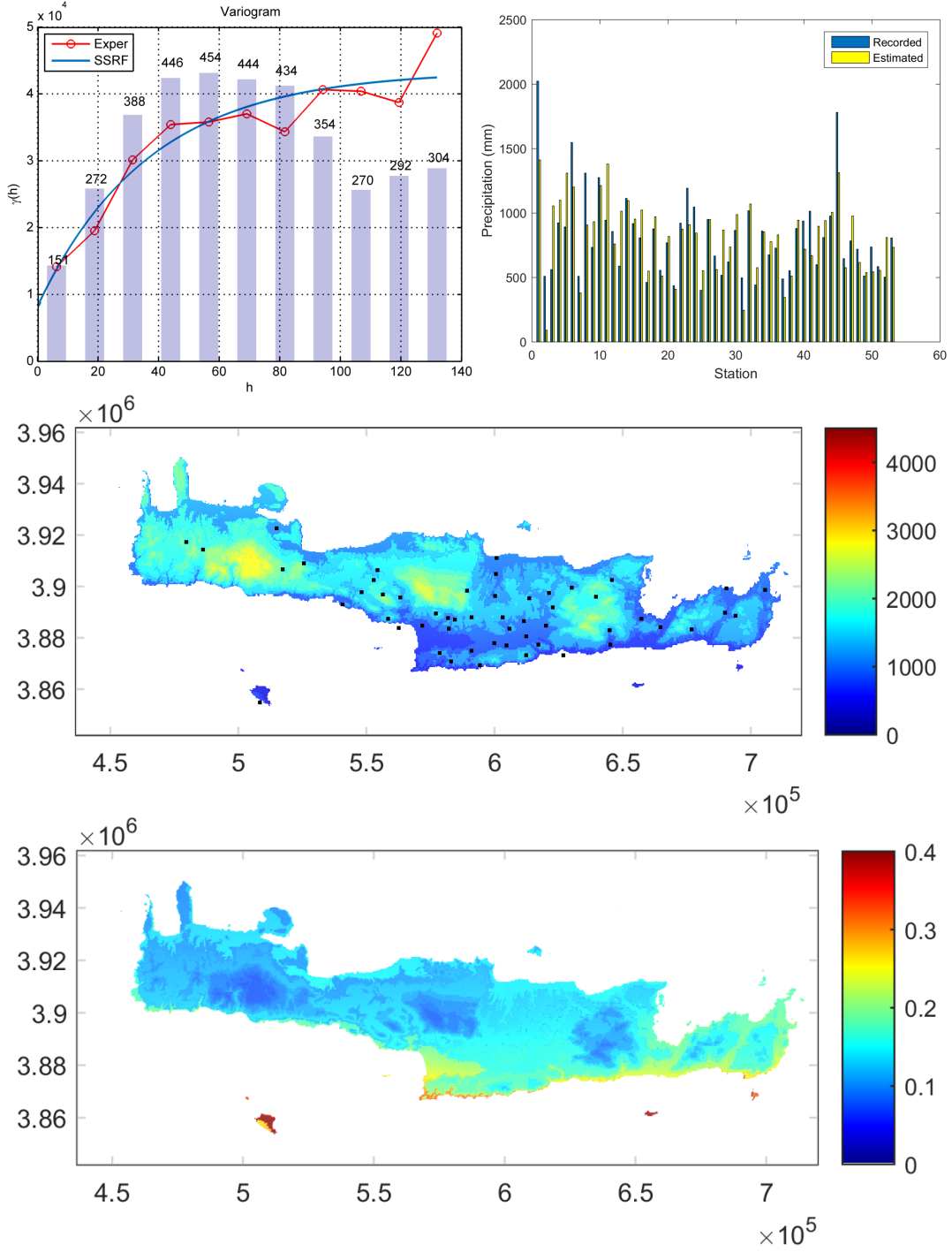


Figure 30: Year 1994 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 8.2364e + 03 \text{ mm}^2$, $\eta_0 = 8.8440e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 36.7642$.

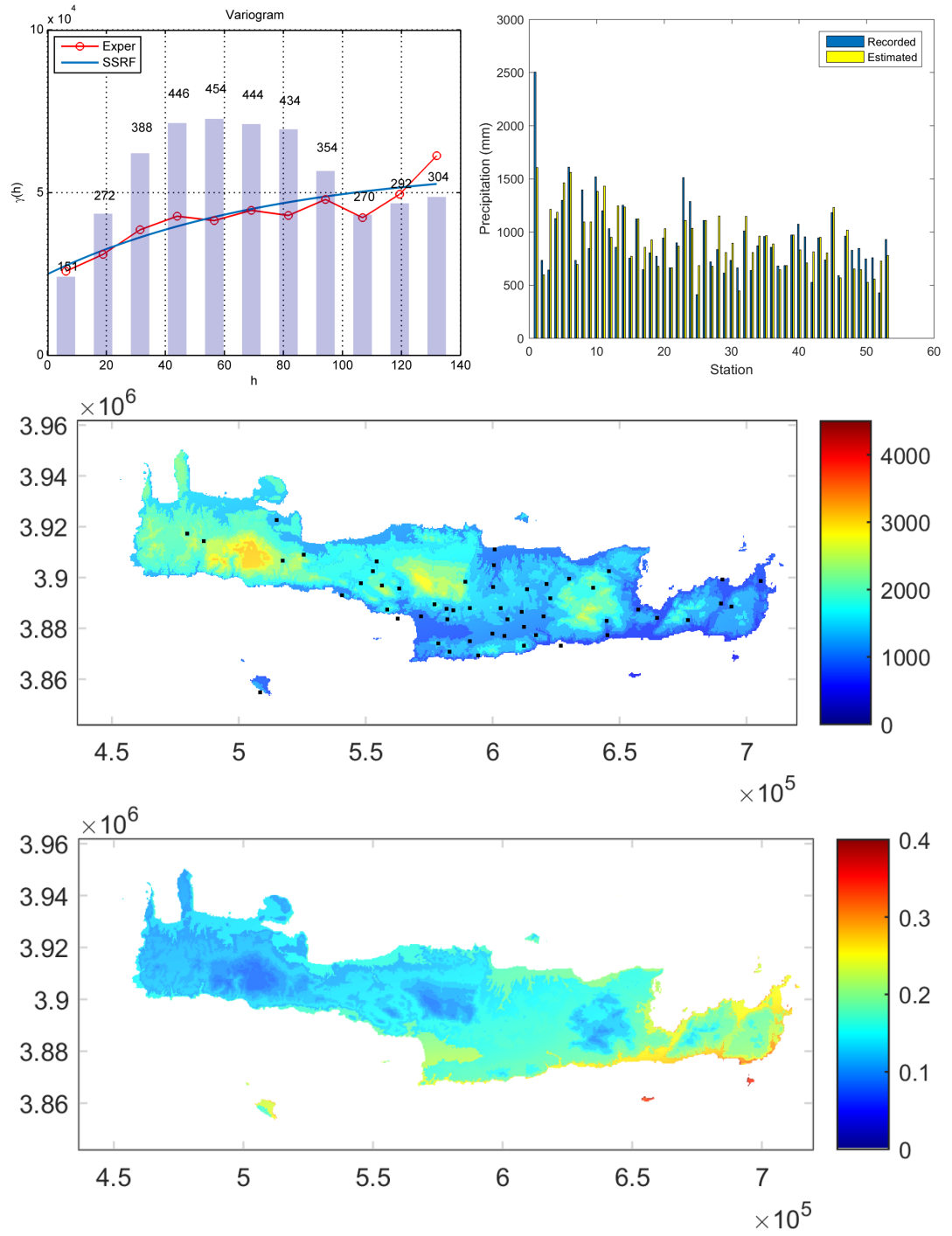


Figure 31: Year 1995 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.4940e + 04 \text{ mm}^2$, $\eta_0 = 8.5090e + 5 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 77.2552$.

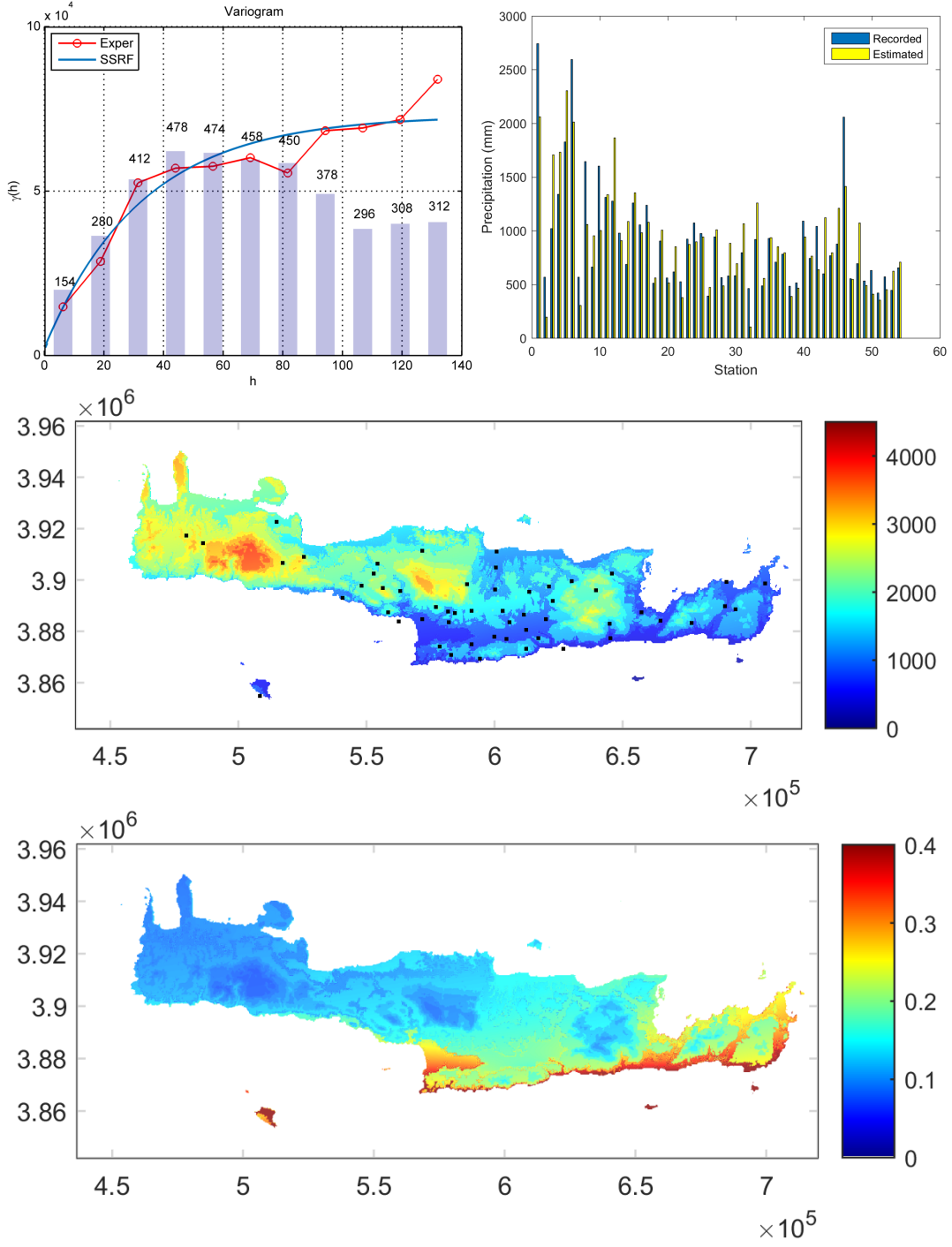


Figure 32: Year 1996 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.3365e + 03 \text{ mm}^2$, $\eta_0 = 1.7743e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 32.6777$.

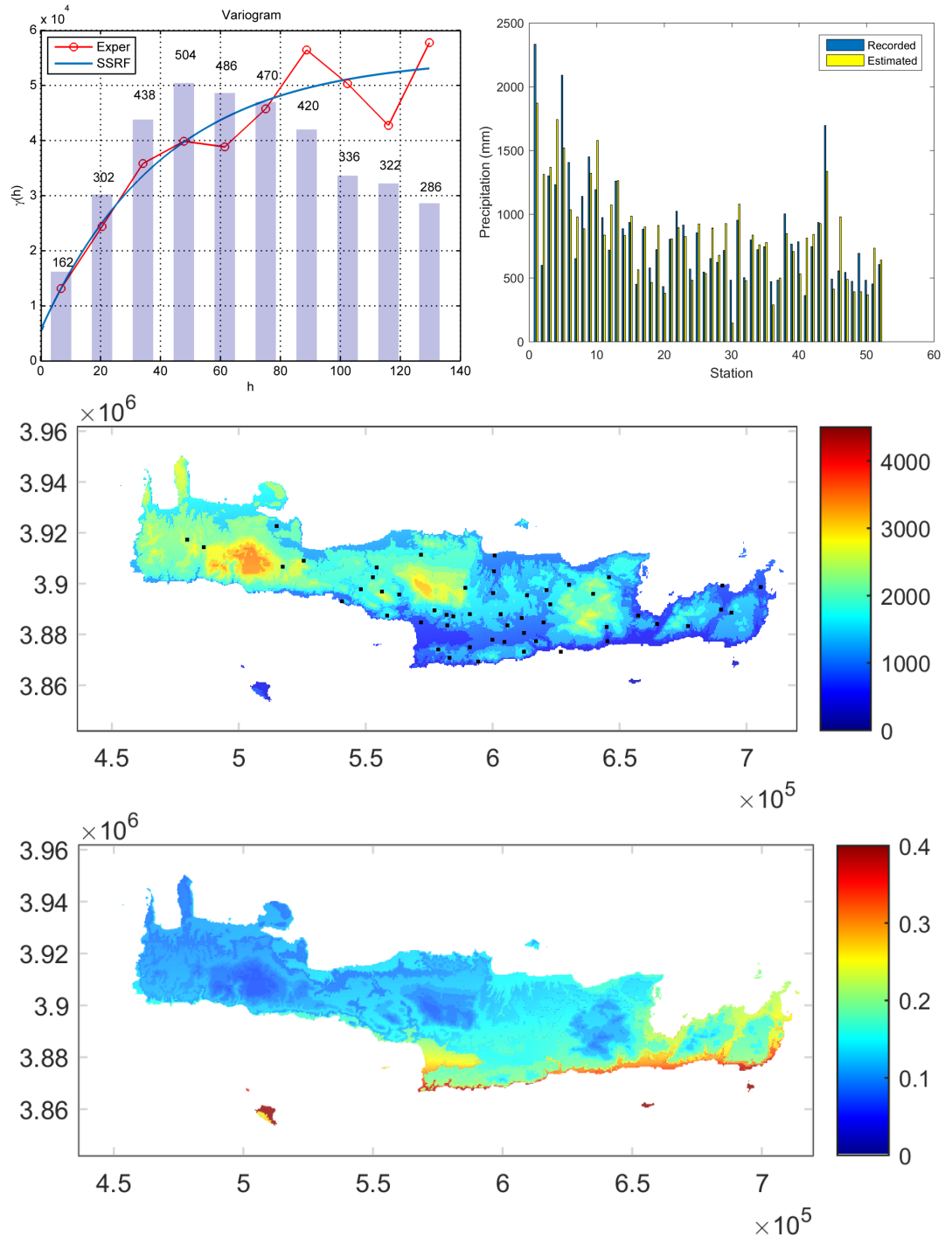


Figure 33: Year 1997 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 5.5623e + 03 \text{ mm}^2$, $\eta_0 = 1.2460e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 40.9075$.

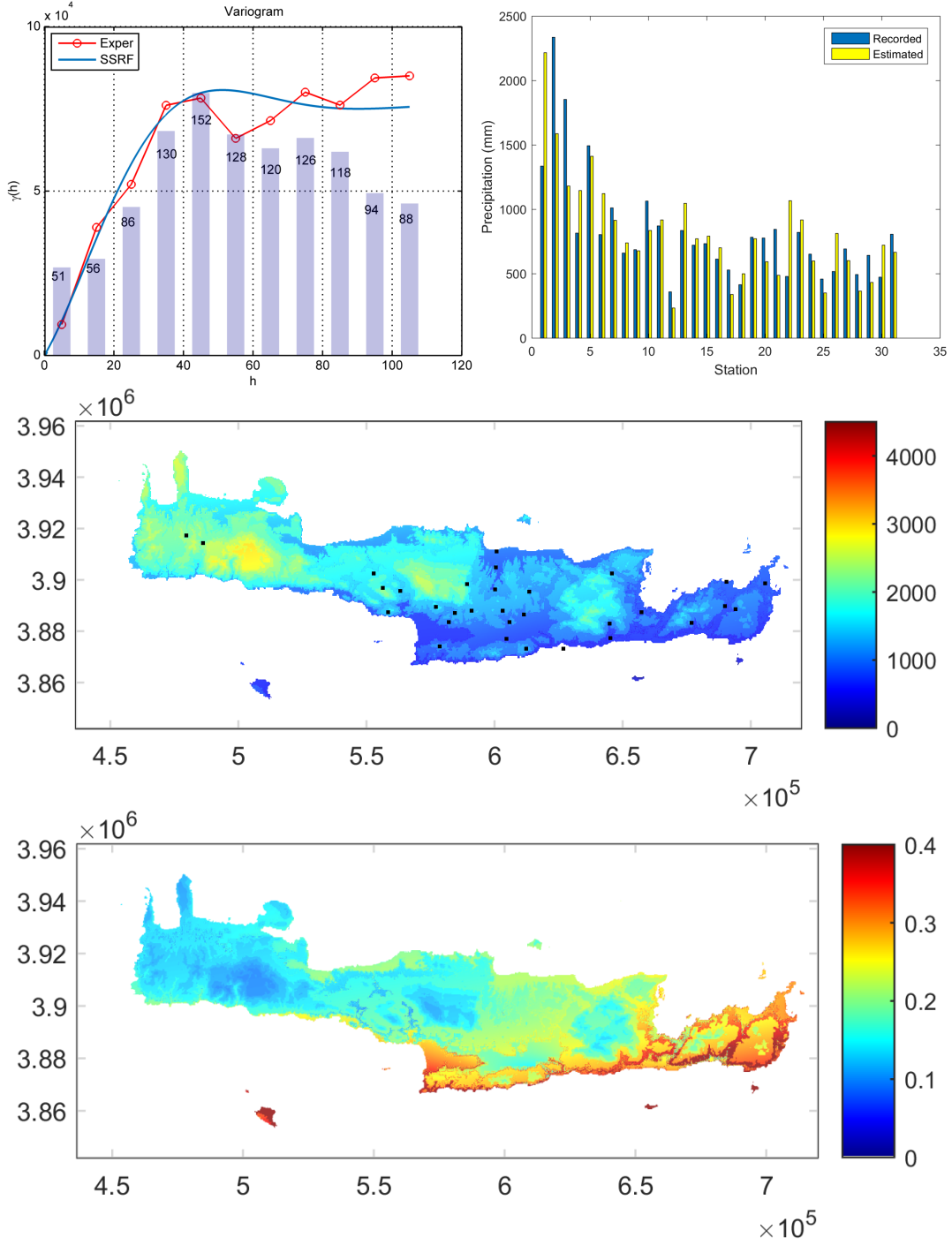


Figure 34: Year 1998 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 6.3451e - 04 \text{ mm}^2$, $\eta_0 = 5.3467e + 05 \text{ mm}^2$, $\eta_1 = -1.6870$, $\xi = 11.6252$.

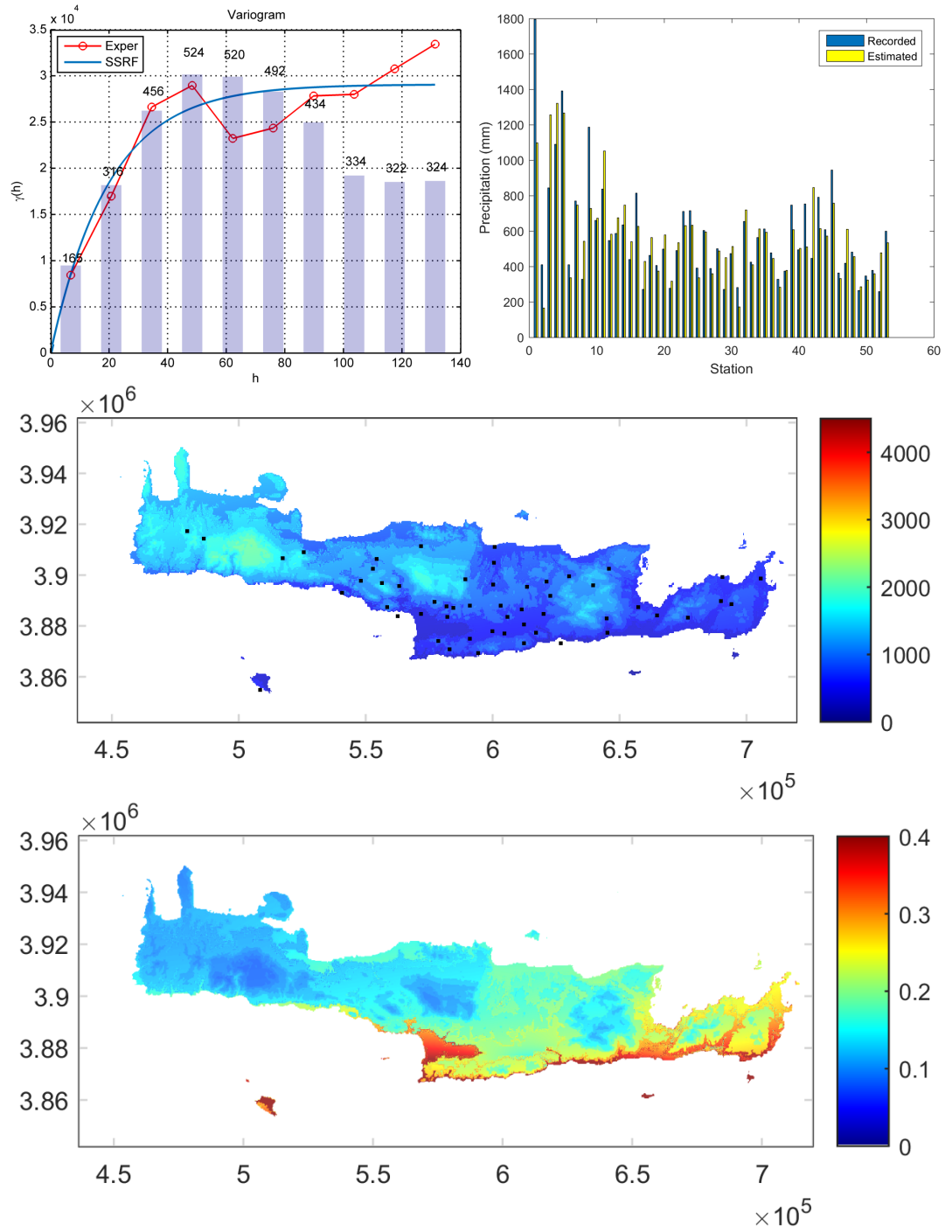


Figure 35: Year 1999 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 6.9015e - 04 \text{ mm}^2$, $\eta_0 = 7.3055e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 19.8601$.

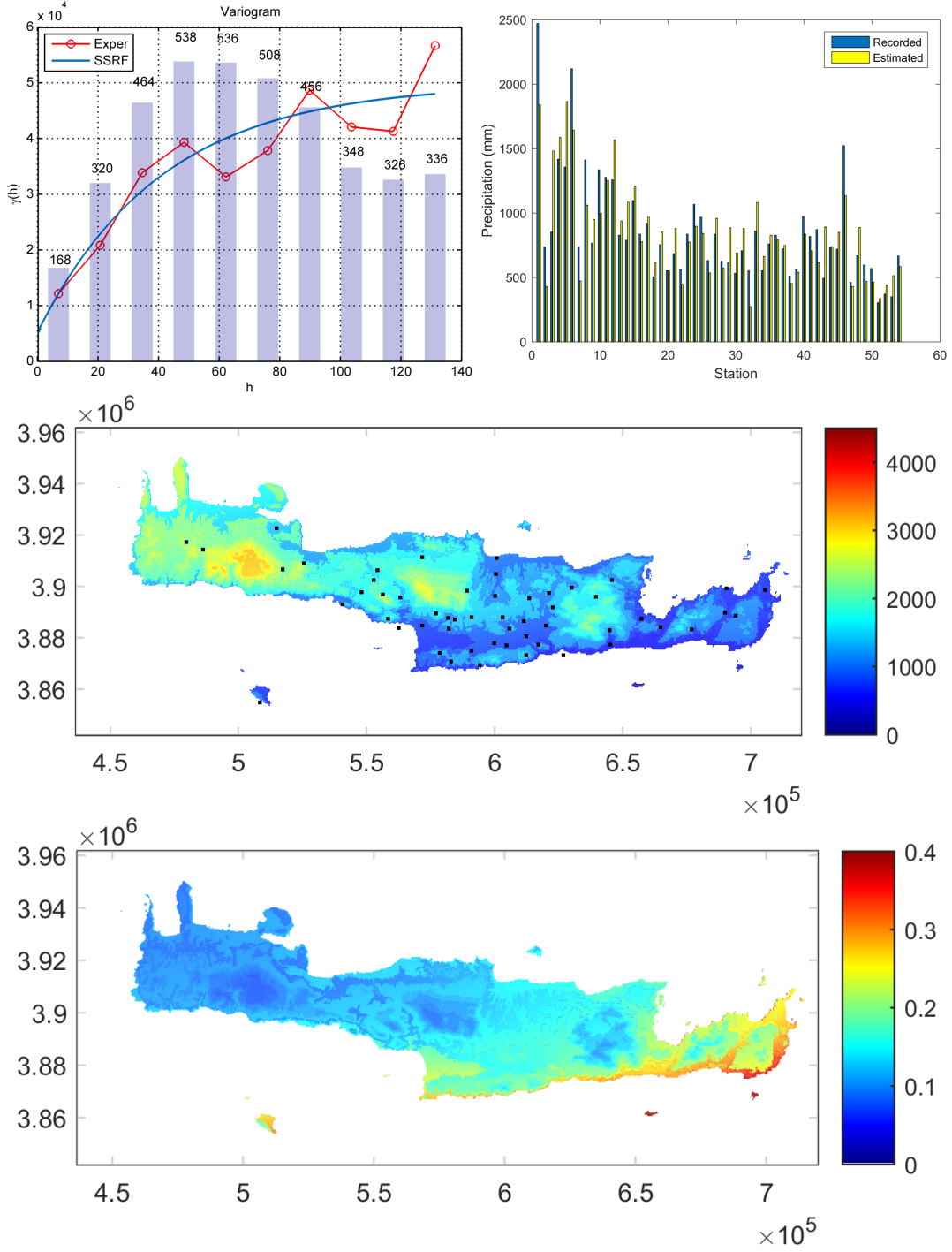


Figure 36: Year 2000 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 5.2191e + 03 \text{ mm}^2$, $\eta_0 = 1.1207e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 41.0061$.

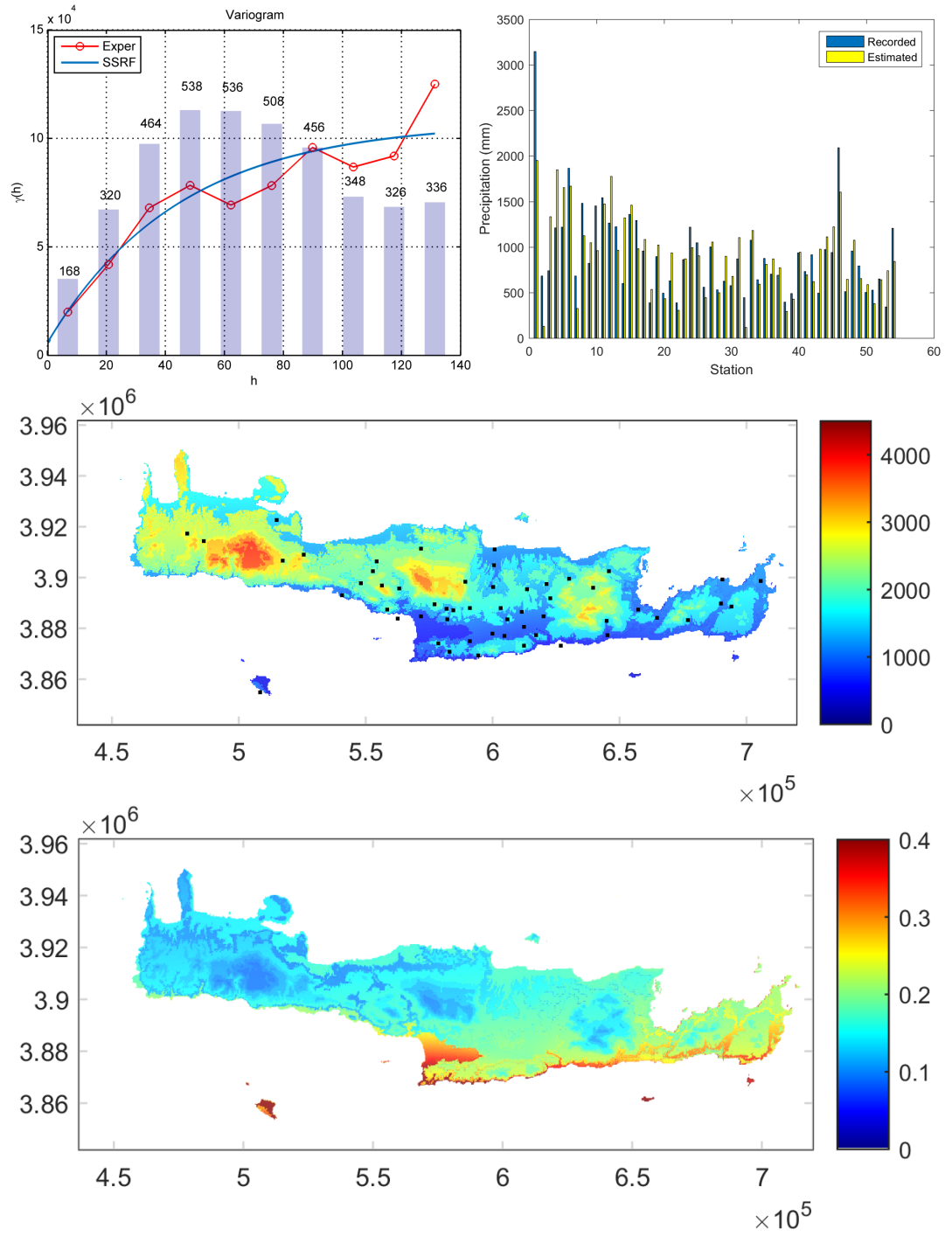


Figure 37: Year 2001 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 5.7736e + 03 \text{ mm}^2$, $\eta_0 = 2.5562e + 06 \text{ mm}^2$, $\eta_1 = 1.9929$, $\xi = 44.4091$.

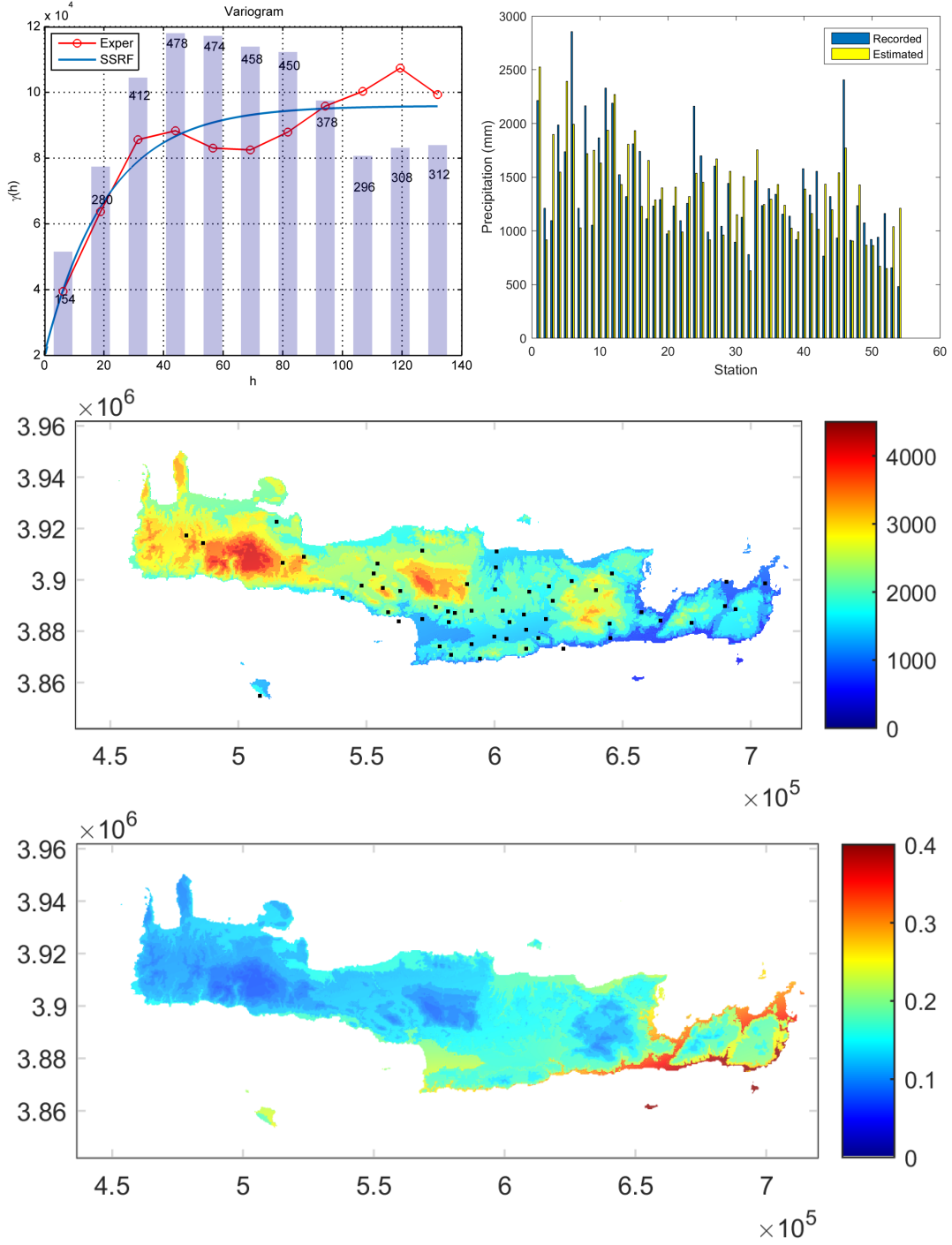


Figure 38: Year 2002 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.0380e + 04 \text{ mm}^2$, $\eta_0 = 1.8979e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 20.9565$.

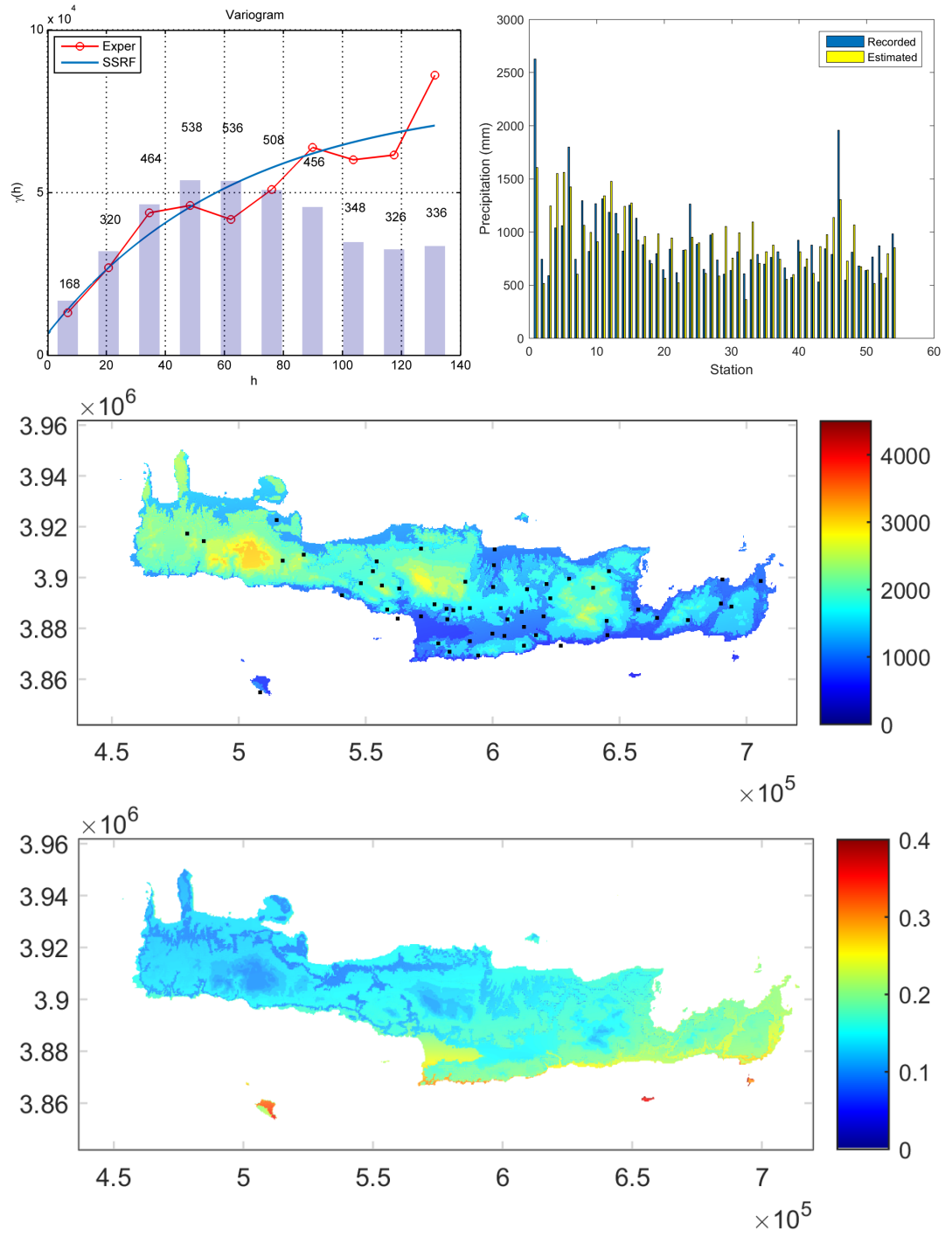


Figure 39: Year 2003 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 6.6351e + 03 \text{ mm}^2$, $\eta_0 = 1.8439e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 63.9489$.

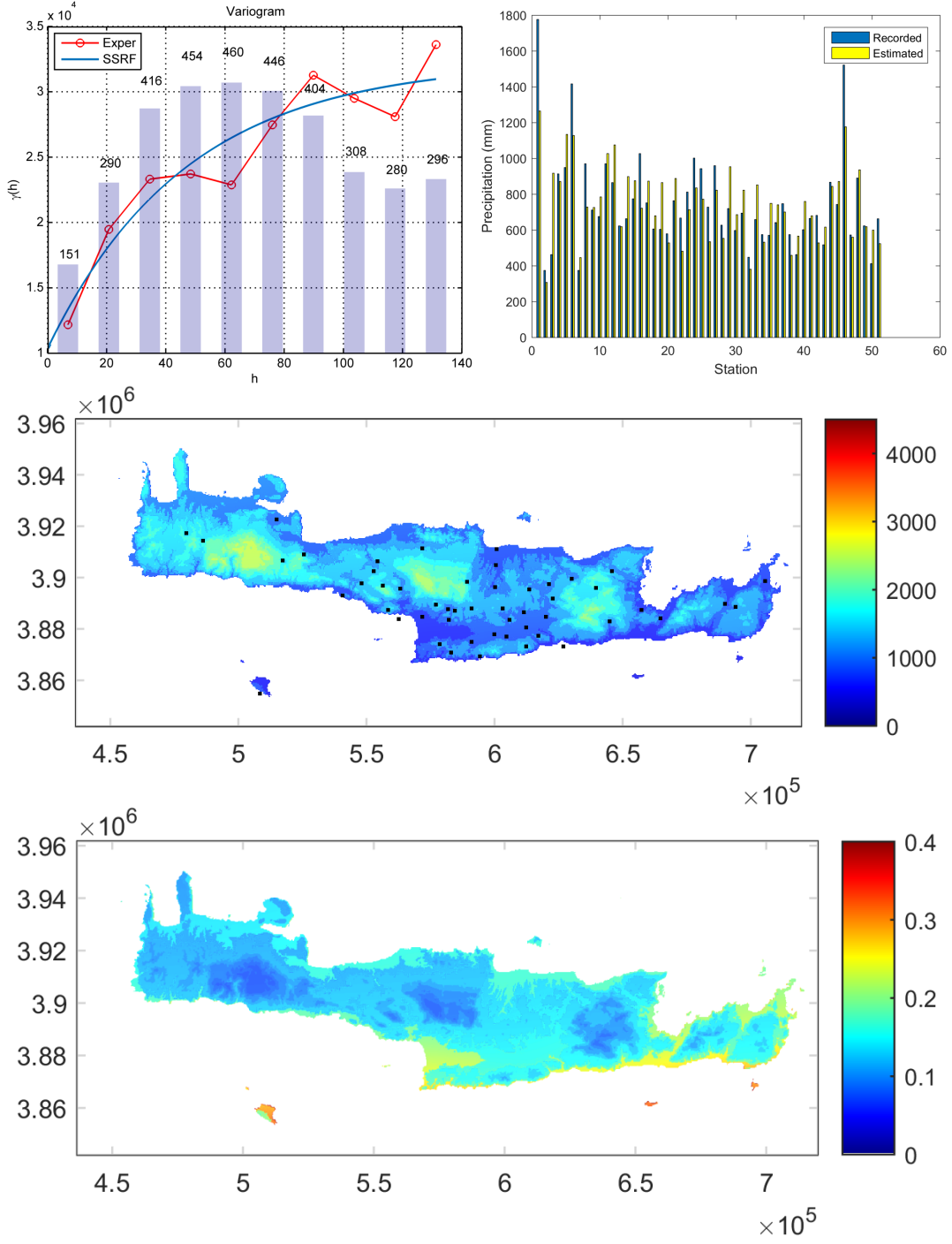


Figure 40: Year 2004 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.0361e + 04 \text{ mm}^2$, $\eta_0 = 5.5118e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 46.8156$.

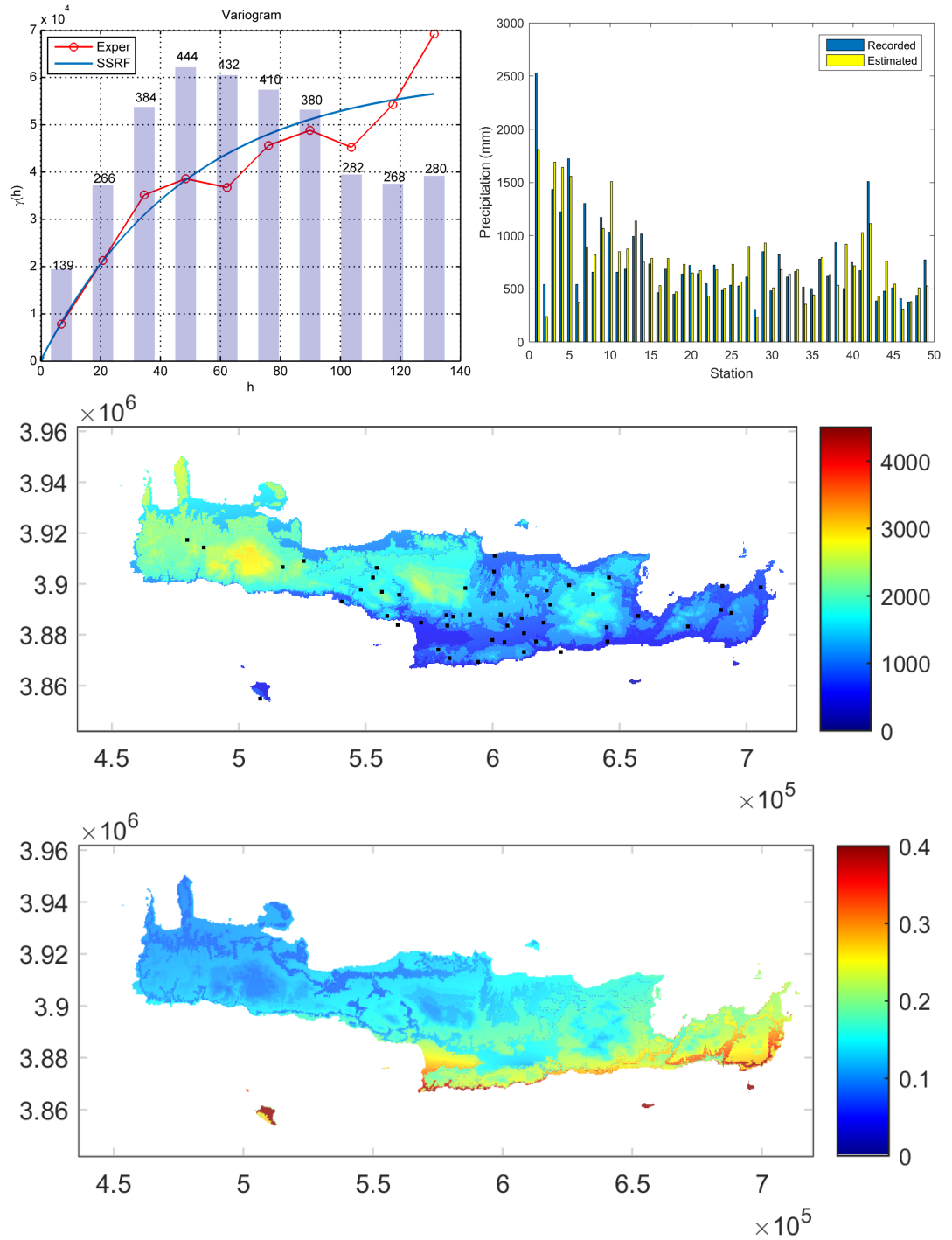


Figure 41: Year 2005 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 94.6634 \text{ mm}^2$, $\eta_0 = 1.5193e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 48.3514$.

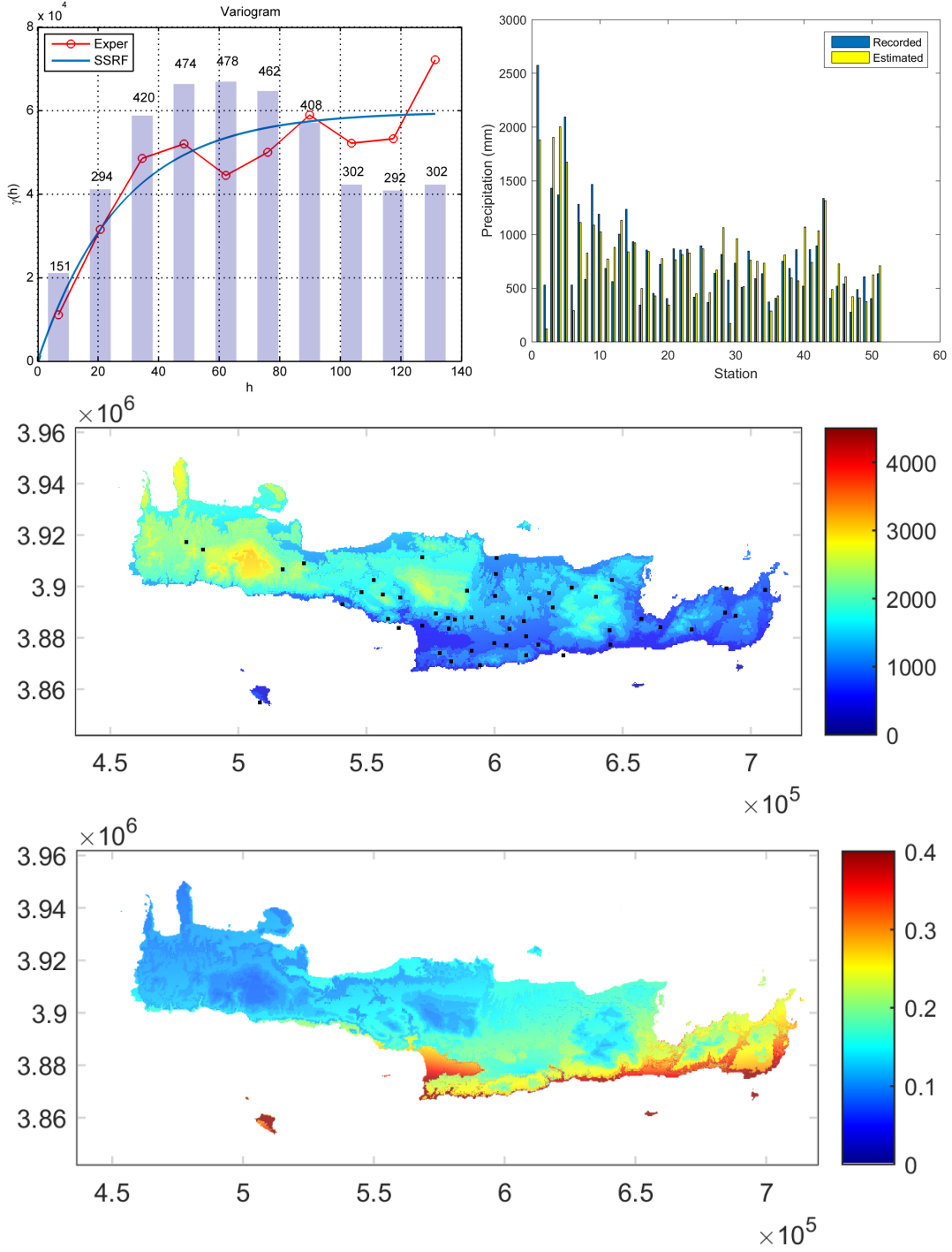


Figure 42: Year 2006 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 9.6712e - 06 \text{ mm}^2$, $\eta_0 = 1.5007e + 06 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 27.5556$.

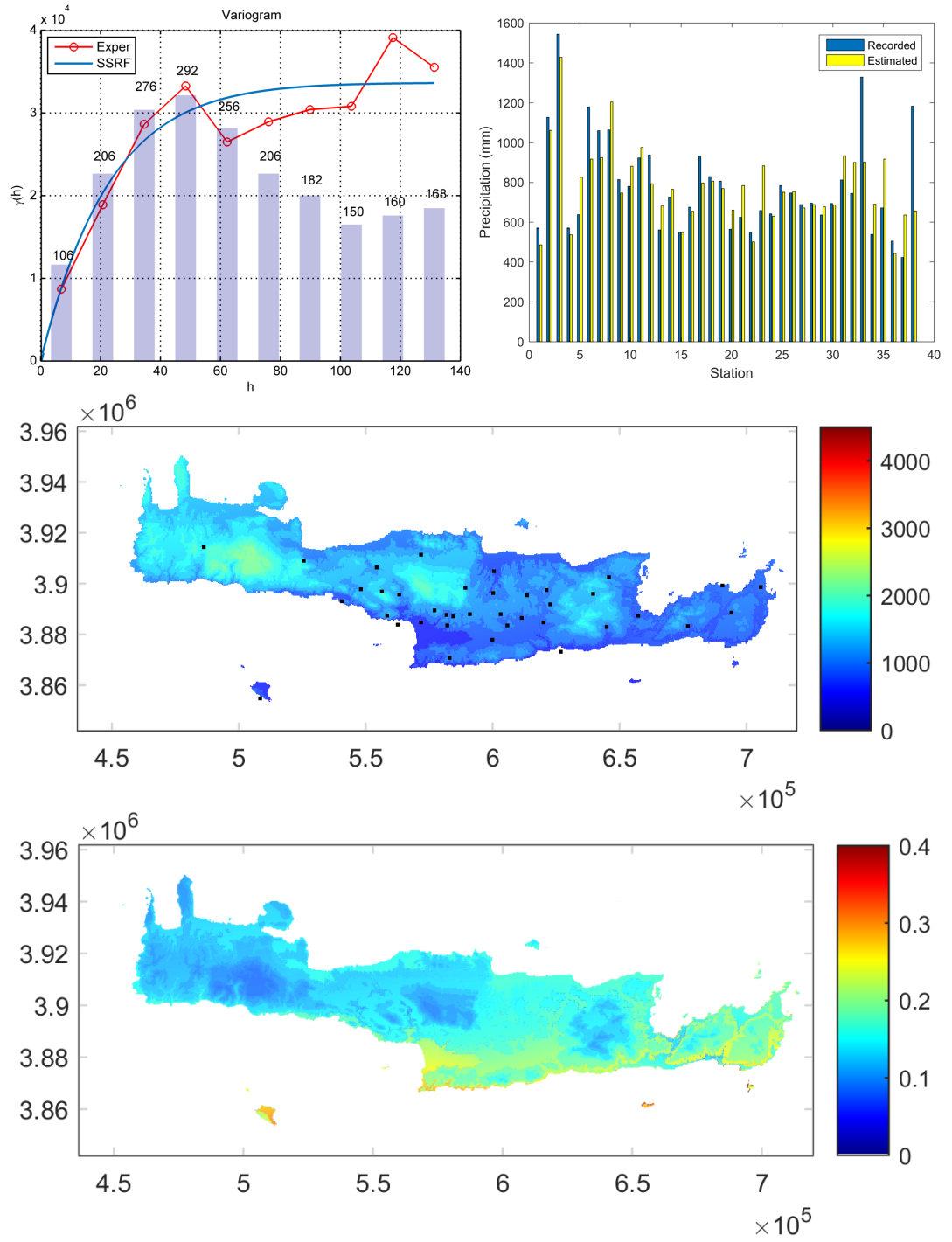


Figure 43: Year 2007 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 0.0011 \text{ mm}^2$, $\eta_0 = 7.8444e+05 \text{ mm}^2$, $\eta_1 = 1.4447$, $\xi = 21.2302$.

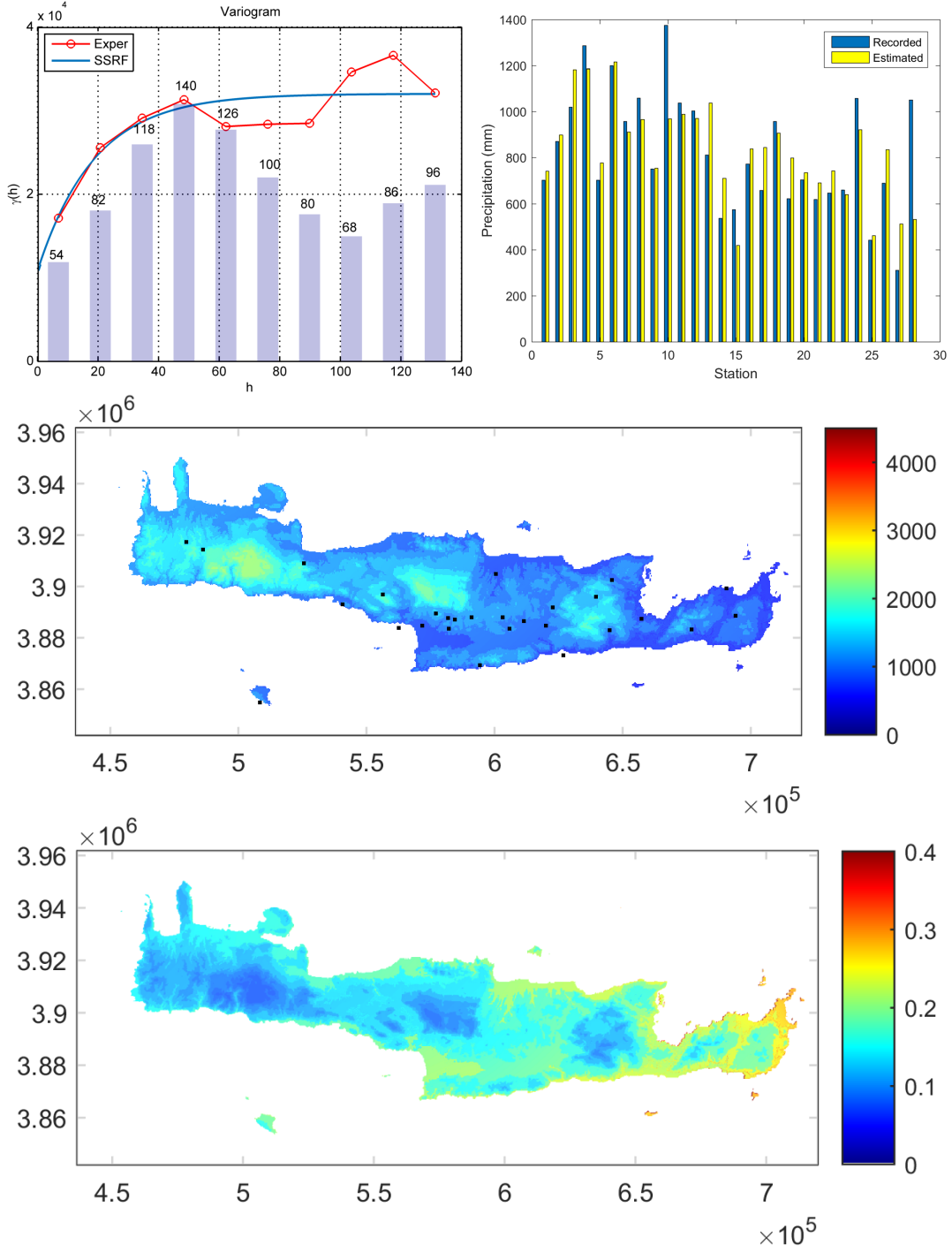


Figure 44: Year 2008 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 1.0835e + 04 \text{ mm}^2$, $\eta_0 = 5.3203e + 05 \text{ mm}^2$, $\eta_1 = 1.9999$, $\xi = 18.7451$.

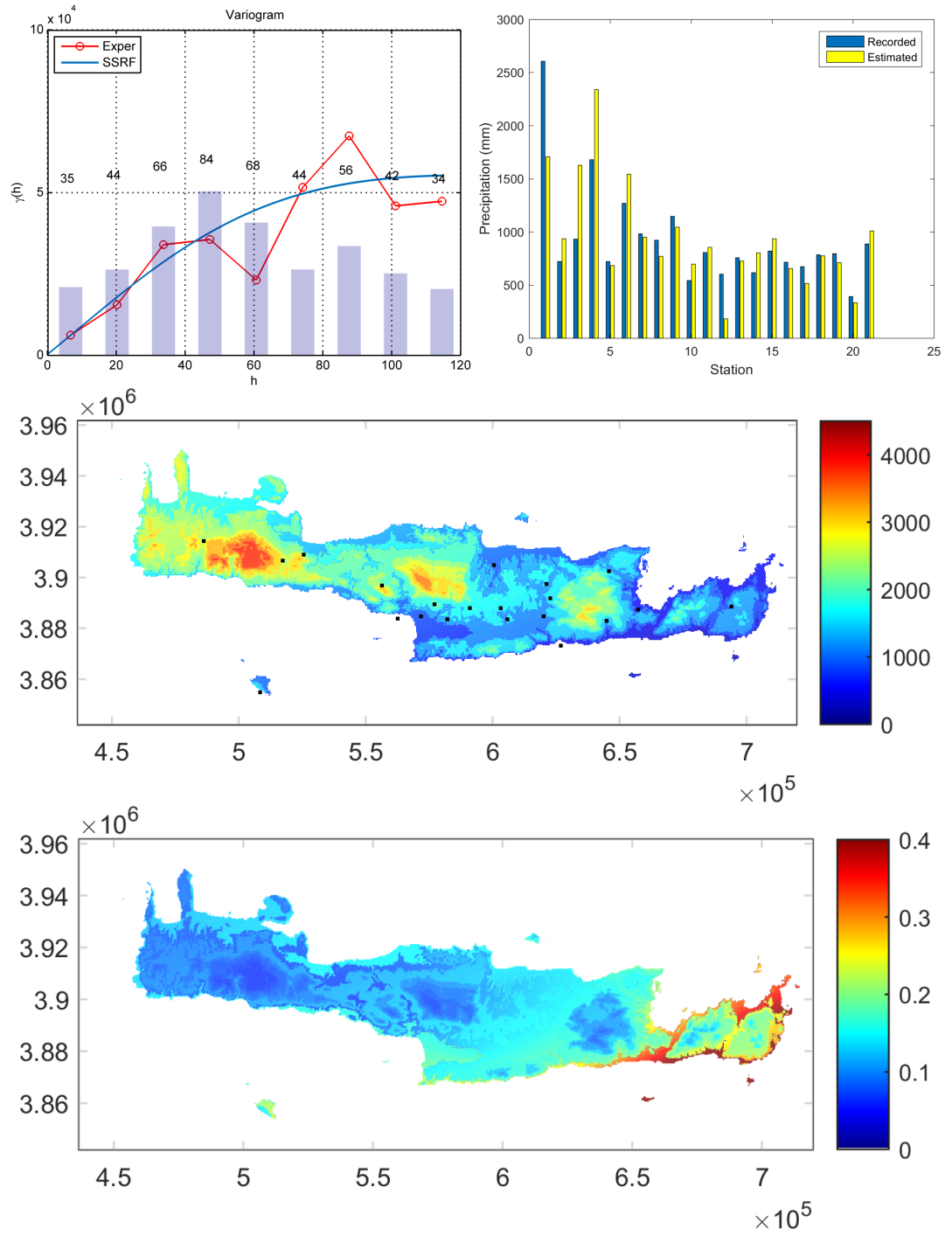


Figure 45: Year 2009 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 325.7335 \text{ mm}^2$, $\eta_0 = 5.8068e + 05 \text{ mm}^2$, $\eta_1 = -1.2541$, $\xi = 27.1204$.

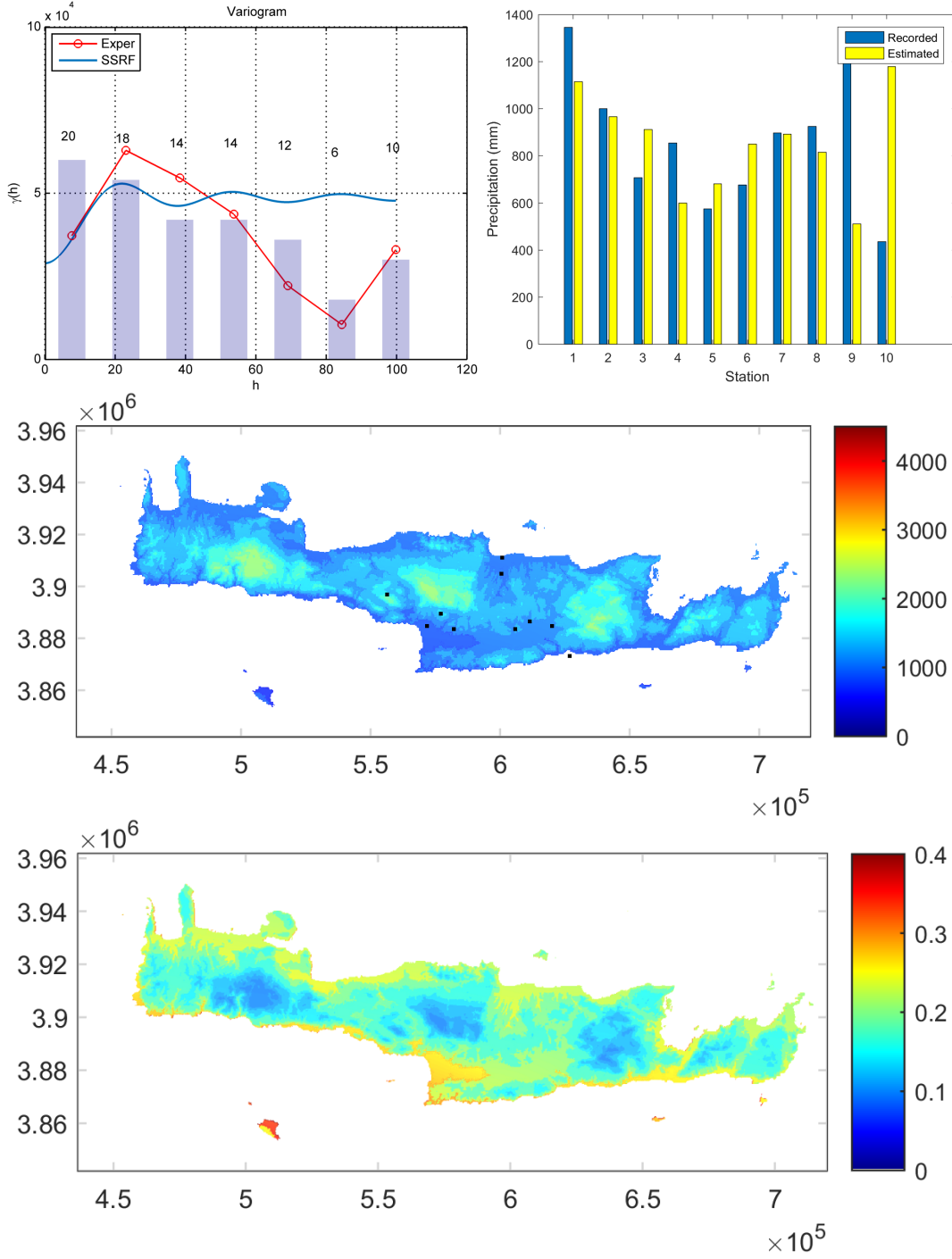


Figure 46: Year 2010 annual precipitation. The Spartan variogram parameters are: nugget variance $\sigma_n^2 = 2.8934e + 04 \text{ mm}^2$, $\eta_0 = 1.4142e + 03 \text{ mm}^2$, $\eta_1 = -1.9999$, $\xi = 4.8863$.

References

- Projected changes in heavy precipitation in winter (left) and summer (right), 2015. URL www.eea.europa.eu/data-and-maps/figures/projected-changes-in-20-year-2. Provider:EURO-CORDEX initiative. [xiii](#), [7](#)
- AGRESTE. Agreste infos rapides-grandes cultures et fourrages, n °6/7. Les Prairies Vues par ISOP en Septembre 2011 [Cassagne, J.P. (ed.)]. 4 pp., Ministère de l'Agriculture, de l'Alimentation, de la Pêche, de la Ruralité et de l'Aménagement du Territoire, Secrétariat Général, Montreuil, France, 2011. [6](#)
- M.M. Ali. Stochastic ordering and kurtosis measure. *Journal of the American Statistical Association*, 69:543–545, 1974. [35](#)
- K.P. Balanda and H.L. Macgillivray. Kurtosis: A critical review. *The American Statistician*, 42(2):111–119, 1988. doi: 10.1080/00031305.1988.10475539. [35](#)
- 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. [25](#)
- 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. [6](#)
- 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. [13](#), [14](#), [16](#), [75](#), [123](#)

- M. Berterretche, A.T. Hudak, W.B. Cohen, T.K. Maersperger, S.T. 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. [58](#)
- S. Bochner, M. Tenenbaum, and H. Pollard. *Lectures on Fourier Integrals*. Annals of mathematics studies. Princeton University Press, 1959. ISBN 9780691079943. [37](#)
- V.G. Bondur. Satellite monitoring of wildfires during the anomalous heat wave of 2010 in Russia. *Izvestiya, Atmospheric and Oceanic Physics*, 47(9):1039–1048, 2011. [6](#)
- 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. [62](#)
- E.J. Burke, R.H.J. Perry, and S.J. Brown. An extreme value analysis of UK drought and projections of change in the future. *Journal of Hydrology*, 388: 131–143, 2010. doi: 10.1016/j.jhydrol.2010.04.035. [19](#)
- P. Burman. A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989. [61](#), [62](#)
- K.S. Chantzoulakis, N.V. Paranychianakis, and A.N. Angelakis. Water resources management in the island of crete, greece, with emphasis on the agricultural use. *Water Policy*, 3:193–205, 2001. [xxiii](#), [69](#)
- J. Chatterton, C. Viavattene, J. Morris, E. Penning-Rowsell, and S. Tapsell. The costs of the summer 2007 floods in England. *Environmental Agency*, page 41 pp., 2010. Environment Agency (EA). [6](#)
- J.P. Chilès and P. Delfiner. *Geostatistics: modeling spatial uncertainty*. Wiley series in probability and statistics. Wiley, 2012. ISBN 9780470183151. [27](#), [38](#), [39](#), [48](#), [50](#), [53](#)

- Chowdhury. *Dr. Chowdhury's Guide to Planet Earth: "The Water Cycle"*. West ed. retrieved 2006-10-24 edition, 2005. [7](#)
- G. Christakos. On the problem of permissible covariance and variogram models. *Water Resources Research*, 20(2):251–265, 1984. [48](#)
- G. Christakos. *random field models in earth sciences*. Academic Press, Chapel Hill, North Carolina, 1992. ISBN 0-12-174230-X. [27](#), [56](#)
- J.H. Christensen, B. Hewitson, A. Busuioc, A. Chen, X. Gao, and et al. Regional climate projections. In *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge and New York: Cambridge University Press., solomon s, qin d, manning m, chen z, marquis m, et al. edition, 2007. [12](#)
- P. Ciais, Reichstein M., Viovy N., Granier A., Ogee J., Allard V., Aubinet M., Buchmann N., Bernhofer C., Carrara A., Chevallier F., De Noblet N., Friend A.D., Friedlingstein P., Grunwald T., Heinesch B., Keronen P., Knohl A., Krinner G., Loustau D., Manca G., Matteucci G., Miglietta F., Ourcival J.M., Papale D., Pilegaard K., Rambal S., Seufert G., Soussana J.F., Sanz M.J., Schulze E.D., Vesala T., and Valentini R. Europe-wide reduction in primary productivity caused by the heat and drought in 2003. *Nature*, 437(7058):529–533, 2005. [6](#)
- S. Coles. *An Introduction to Statistical Modeling of Extreme Values*. Springer, London, 2001. ISBN 1-85233-459-2. [20](#), [22](#), [29](#)
- G.E. Condra. Drought: Its effect and measures of control in Nebraska. Technical report, Nebraska Conservation Bull. 25, Conservation and Survey Division, University of Nebraska, Lincoln, NE, 1944. 43 pp. [10](#)
- D. Coumou and S. Rahmstorf. A decade of weather extremes. *Nature Climate Change*, 2(7):491–496, 2012. [6](#)
- N.A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and statistics. Wiley, revised edition edition, 1993. ISBN 9780471002550. [32](#), [33](#)

- M. Daufresne, P. Bady, and J.F. Fruget. Impacts of global changes and extreme hydroclimatic events on macroinvertebrate community structures in the French Rhône River. *Oecologia*, 151(3):544–559, 2007. 6
- A.C. Davison, R. Huser, and E. Thibaud. Geostatistics of dependent and asymptotically independent extremes. *Mathematical Geosciences*, 45(5):511–529, 2013. ISSN 1874-8953. doi: 10.1007/s11004-013-9469-y. 123
- S. Dean and B. Illowsky. Descriptive statistics: Skewness and the mean, median, and mode. Technical report, OpenStax-CNX, 2012. 34
- L. Devroye. *Non-Uniform Random Variate Generation*. Springer-Verlag New York Inc, 1986. 16, 18, 23
- R.E. Dickinson, R.M. Errico, F. Giorgi, and G.T. Bates. A regional climate model for western United States. *Clim. Change*, 15:383–422, 1989. 12
- N. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, second edition, 1981. 57
- B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382):316–331, 1983. 61
- esri. USGS EarthExplorer, 2015. URL <http://www.esri.com/>. 71
- D. Founda and C. Giannakopoulos. The exceptionally hot summer of 2007 in Athens, Greece - a typical summer in the future climate. *Global and Planetary Change*, 67(3–4):227–236, 2009. 6
- H.J. Fowler, S. Blenkinsop, and C. Tebaldi. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, 27:1547–1578, 2007. 13
- M. Fuentes. Spectral methods for nonstationary processes. *Biometrika*, 89:197–210, 2002. 43
- S. Ghosh. Modelling bivariate rainfall distribution and generating bivariate correlated rainfall data in neighbouring meteorological subdivisions using copula. *Hydrological Processes*, 24:3558–3567, 2010. doi: 10.1002/hyp.7785. 19

- W.J. Gibbs and J.V. Maher. Rainfall deciles as drought indicators. *Bureau of Meteorology Bulletin*, 1967. Commonwealth of Australia, Melbourne. 11
- I.I. Gikhman and A.V. Skorokhod. *Introduction to the theory of random processes*. Dover Publications, INC, Mineola, New York, 1996. ISBN 0-486-69387-2. 28
- M. Gilli and E. K llezi. An application of Extreme Value Theory for measuring financial risk. *Computational Economics*, 27(1):1–23, 2006. 19
- F. Giorgi and L.O. Mearns. Introduction to special section: regional climate modeling revisited. *J. Geophys. Res. (Atmospheres)*, 1999. doi: 10.1029/98JD02072. 12
- Google Earth. Crete island, Greece, 2015. URL <http://www.earth.google.com>. SIO, NOAA, U.S. Navy, NGA, GEBCO. DigitalGlobe 2015. xiv, 68, 73
- P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Applied geostatistics series. Oxford University Press, 1997. ISBN 9780195115383. URL <http://books.google.gr/books?id=CW-7tHAaVR0C>. 27, 33, 34, 41, 53, 56, 57
- P. Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129, 2000. 13
- R. Grossman, G. Seni, J. Elder, N. Agarwal, and H. Liu. *Ensemble methods in data mining: Improving Accuracy Through Combining Predictions*. Morgan & Claypool, 2010. doi: 10.2200/S00240ED1V01Y200912DMK002. 60
- S.L. Grotch and M.C. MacCracken. The use of general circulation models to predict regional climatic change. *Journal of Climate*, 4(3):296–303, 1991. doi: 10.1175/1520-0442. 12
- P. Guttorp and T. Gneiting. Studies in the history of probability and statistics XLIX: On the Mat rn correlation family. *Biometrika*, 93(4):989–995, 2006. 43
- M.S. Handcock and J.R. Wallis. An approach to statistical spatio-temporal modeling of meteorological fields. *J. Am. Statist. Assoc.*, 89:368–378, 1994. 43

- D.L. Hartmann, Klein Tank A.M.G., Rusticucci M., Alexander L.V., Bronnimann S., Charabi Y., Dentener F.J., Dlugokencky E.J., Easterling D.R., Kaplan A., Soden B.J., Thorne P.W., Wild M., and Zhai P.M. *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 D., Qin and G.-K., Plattner and M., Tignor and S.K., Allen and J., Boschung and A., Nauels and Y., Xia and V., Bex and P.M., Midgley (eds.)]. [xiii](#), [xxiii](#), [3](#), [4](#), [5](#), [116](#), [122](#)
- R.R. Heim. A review of twentieth century drought indices used in the United States. *Bull. American Meteorological Society*, 83:1149–1165, 2002. [9](#)
- 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, 2014. [67](#)
- T. Hengl. *A practical guide to geostatistical mapping of environmental variables*. Office for Official Publications of the European Communities, Luxembourg, 2007. [57](#)
- T. Hengl, G.B. M. Geuvelink, and A. Stein. Comparison of kriging with external drift and regression–kriging. Technical report, ITC, 2003. URL http://www.itc.nl/library/Academic_output/. [57](#)
- T. Hengl, G.B.M. Heuvelink, and D.G. Rossiter. About regression–kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301–1315, 2007. [57](#), [59](#)
- D.T Hristopulos. Spartan Gibbs random field models for geostatistical applications. *SIAM Journal on Scientific Computing*, 24(6):2125–2162, 2003. [43](#), [44](#)
- D.T. Hristopulos. Applied geostatistics. Technical report, Technical University of Crete, 2012. [xiv](#), [29](#), [32](#), [34](#), [36](#), [38](#), [39](#), [41](#), [46](#), [47](#)

- 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. 43, 44
- M. Hulme. A 1951-80 global land precipitation climatology for the evaluation of general circulation models. *Climate Dynamics*, 7:57–72, 1992. 12
- 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., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. 4
- IPCC. Guidance on the use of data: What is a GCM?, 2015. URL <http://www.ipcc-data.org>. 11, 12
- E.H. Isaaks and R.M. Srivastava. *Applied Geostatistics*. Oxford University Press, 1989. ISBN 9780195050134. URL <https://books.google.gr/books?id=vC2dcXFLI3YC>. 31, 33, 64, 66
- K.P. Jagdish and B.R. Campbell. *Handbook of the normal distribution*. Marcel Dekker, INC, New York and Basel, 1982. ISBN 0-8247-1541-1. 15
- P. Jones, C. Harpham, T. Osborn, and M. Salmon. Temperature, 2015. URL www.cru.uea.ac.uk/cru/data/temperature/. 3
- A.G. Journel. Fundamentals of geostatistics in five lessons. *American Geophysical Union*, page 45, 1989. 41
- A.G. Journel and C.J. Huijbregts. *Mining Geostatistics*. Blackburn Press, 2003. ISBN 9781930665910. 49
- 2008 JRC. Forest Fires in Europe 2007. JRC Scientific and Technical Reports, Report No. 8, EUR 23492, European Commission, Joint Research Centre (JRC) 77 pp., Institute for Environment and Sustainability, Luxembourg, 2008. Publications Office of the European Union. 6

- I. Kaplansky. A common error concerning kurtosis. *Journal of the American Statistical Association*, 40(230):259–259, 1945. doi: 10.1080/01621459.1945.10501856. URL <http://www.tandfonline.com/doi/abs/10.1080/01621459.1945.10501856>. 35
- C.A. Karavitis, S. Alexandris, D.E. Tsesmelis, and G. Athanasopoulos. Application of the Standardized Precipitation Index (SPI) in Greece. *Water*, 3: 787–805, 2011. doi: 10.3390/w3030787. 14
- 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*, 2004. W09304. 11
- 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. 61
- 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., C.B. Field, D.J. Dokken, M.D. Mastrandrea, K.J. Mach, T.E. Bilir, M. Chatterjee, K.L. Ebi, Y.O. Estrada, R.C. Genova, B. Girma, E.S. Kissel, A.N. Levy, S. MacCracken, P.R. Mastrandrea, and L.L. White (eds.)]. xxiii, 1, 5, 6
- 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. 49
- D.P. Kroese, T. Taimre, and Z.I. Botev. *Handbook of Monte Carlo Methods*. Wiley Series in Probability and Statistics, John Wiley and Sons, New York, 2011. 14, 16, 17, 21

- P.V. Kumar, M. Bindi, A. Crisci, and G. Maracchi. Detection of variations in precipitation at different time scales of twentieth century at three locations of Italy. *Weather and Climate Extremes*, 2:7–15, 2013. doi: 10.1016/j.wace.2013.10.005. 25
- K.-C Li. Asymptotic optimality for c_p, c_l , cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987. 61
- D. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, Inc., New York, 1969. 52
- J. Luterbacher, Dietrich D., Xoplaki E., Grosjean M., and Wanner H. European seasonal and annual temperature variability, trends, and extremes since 1500. *Science*, 303(5663):1499–1503, 2004. 6
- G. Matheron. Principles of geostatistics. *Economic Geology*, 58:1246–1266, 1963. doi: 10.2113/gsecongeo.58.8.1246. 49
- G.J. McCabe and M.D. Dettinger. Relations between winter precipitation and atmospheric circulation simulated by the geophysical fluid dynamics laboratory general circulation model. *International Journal of Climatology*, 15(6):625–638, 1995. ISSN 1097-0088. doi: 10.1002/joc.3370150604. URL <http://dx.doi.org/10.1002/joc.3370150604>. 12
- T. McKee, N. Doesken, and J. Kleist. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, page 179184, 1993. Boston. 11, 14, 16, 17, 25, 75
- G.J. McLachlan, K.-A. Do, and C. Ambrose. *Analyzing microarray gene expression data*. Wiley, 2004. 62
- S. Meier. Planar geodetic covariance functions. *Rev. Geophys. Space Phys.*, 19: 673–686, 1981. 43
- J.M. Mejía and I. Rodríguez-Iturbe. On the synthesis of random field sampling from the spectrum: An application to the generation of hydrological spatial processes. *Water Resour. Res.*, 10:705–711, 1974. 43

- L. Mitas and H. Mitasova. *Geographical Information Systems Abridged: Principles, Techniques, Management and Applications*, chapter 34. Spatial Interpolation, pages 481–492. Wiley, 2005. URL http://www.geos.ed.ac.uk/~gisteac/gis_book_abridged/. 49
- F.J. Moral. Comparison of different geostatistical approaches to map climate variables: application to precipitation. *International Journal of Climatology*, 30:620–631, 2010. doi: 10.1002/joc.1913. 13
- C. Morice. HadCRUT4, 2015. URL www.metoffice.gov.uk/hadobs/hadcrut4/. 4
- S. Niemeyer. New drought indices. *Options Mditerranennes. Srie A: Sminaires Mditerranens*, 80:267–274, 2008. 11
- F.C. Odo, S.U. Offiah, and P.E. Ugwuoke. Weibull distribution–based model for prediction of wind potential in Enugu, Nigeria. *Advances in Applied Science Research*, 3(2):1202–1208, 2012. 19
- R.A. Olea. *Geostatistics for Engineers and Earth Scientists*. Springer US, 1999. ISBN 9780792385233. URL <http://books.google.gr/books?id=bKoD2mMORHUC>. 27, 32, 38, 42
- R.J. Oosterbaan. *Drainage Principles and Applications*. International Institute for Land Reclamation and Improvement (ILRI) Publication 16, Wageningen, The Netherlands, second edition, 1994. ISBN 90 70754 3 39. 19
- W.C. Palmer. Meteorological drought. *Research Paper No 45*, 1965. US Weather Bureau, Washington, DC. 11
- W.C. Palmer. Keeping track of crop moisture conditions, nationwide: The new Crop Moisture Index. *Weatherwise 21*, pages 156–161, 1968. 11
- A. Papoulis and S.U. Pillai. *Probability, Random Variables and Stochastic Process*. McGraw-Hill Inc., New York, fourth edition, 2002. 28
- N. Phillips. The general circulation of the atmosphere: a numerical experiment. *Quarterly Journal of the Royal Meteorological Society*, 82:123–164, 1956. 12

- R.R. Picard and R.D. Cook. Cross-validation of regression models. *J. Amer. Statist. Assoc.*, 79(387):575–583, 1984. 62
- M. Pitt. The Pitt Review: Lessons learned from the 2007 floods. final report. *Cabinet Office*, page 500 pp., 2008. 6
- D.R.J. Pleydell, F. Raoul, F. Tourneux, Danson F.M., A.J. Graham, P.S. Craig, and P. Giraudoux. Modelling the spatial distribution of echinococcus multilocularis infection in foxes. *Acta Tropica*, 91(3):253–265, 2004. 58
- Land Processes Distributed Active Archive Center (LP DAAC). USGS EarthExplorer, 2015. URL <http://lpdaac.usgs.gov>. 71
- C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. Technical report, MA: MIT Press, Cambridge, 2006. 43
- Region of Crete Information Bull. Region of crete (2002) sustainable management of water resources in crete., November 2002. pp. 24 (in Greek). 71, 73
- B. Revich and D.A. Shaposhnikov. Climate change, heat waves, and cold spells as risk factors for increased mortality in some regions of russia. *Studies on Russian Economic Development*, 23(2):195–207, 2012. 6
- J. Rivoirard. On the structural link between variables in kriging with external drift. *Mathematical Geology*, 34(7):797–808, 2002. 57
- J.M. Robine, S.L.K. Cheung, S. Le Roy, H. Van Oyen, C. Griffiths, J.P. Michel, and F.R. Herrmann. Death toll exceeded 70,000 in Europe during the summer of 2003. *Comptes Rendus Biologies*, 331(2):171–178, 2008. 6
- 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. 24
- M. Seeger. Gaussian processes for machine learning. *Int. J. Neural Syst.*, 14: 69–106, 2004. 43
- 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.

- Proceedings of the Western Snow Conference, Reno (NV)*, 1982. Colorado State Univ., Fort Collins (CO). 11
- M.A. Sharma and J.B. Singh. Use of probability distribution in rainfall analysis. *New York Science Journal*, 3(9):40–49, 2010. 19
- I. Shiklomanov. *Water in Crisis: A Guide to the World's Fresh Water Resources*. Oxford University Press, New York, 1993. xiii, 7, 8
- J. Shohat. Inequalities for moments of frequency functions and for various statistical constants. *Biometrika*, 21(1–4):361–375, 1929. doi: 10.2307/2332566. 35
- A.Z. Shvidenko, D.G. Shchepashchenko, E.A. Vaganov, A.I. Sukhinin, S.S. Maksyutov, I. McCallum, and I.P. Lakyda. Impact of wildfire in Russia between 1998-2010 on ecosystems and the global carbon budget. *Doklady Earth Sciences*, 441(2):1678–1682, 2011. 6
- M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999. 43
- M. Stone. Cross-validatory choice and assessment oif statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. 61
- I. Strangeways. *Precipitation: Theory, Measurement and Distribution*. Cambridge University Press, 2006. ISBN 9780511535772. URL <http://dx.doi.org/10.1017/CB09780511535772>. Cambridge Books Online. 8
- V. Sudhir and S. Richard. 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. 61
- A. Sylvain and C. Alain. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. ISSN 1935–7516. doi: 10.1214/09-SS054. 59, 62

- G. Tsakiris, D. Pangalou, and H. Vangelis. Regional drought assessment based on the Reconnaissance Drought Index (RDI). *Water Resources Management*, 21 (5):821–833, 2007. [11](#)
- S. Tushaus. Bayesian statistics yields new insights into topographically influenced precipitation. *physicstoday*, 2014. [13](#)
- 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. [36](#), [58](#)
- 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. [43](#), [58](#), [69](#)
- N. Vivekanandan. Comparison of parameter estimation procedures of Gumbel and Fréchet distributions for modelling annual maximum rainfall. *Journal of Engineering & Technology Research*, 1(1):1–9, 2013. ISSN 2315-9286. [19](#)
- A.-E. Vrochidou. Spatiotemporal drought analysis for the island of crete. Master’s thesis, Environmental Engineering of Technical University of Crete, Chania, Greece, 2009. (in Greek). [68](#), [71](#)
- 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. [11](#), [14](#), [16](#), [75](#)
- A.-E.K. Vrochidou, M.G. Grillakis, and I.K. Tsanis. Drought assessment based on multi-model precipitation projections for the island of crete. *Journal of Earth Science & Climatic Change*, 4:158, 2013. doi: 10.4172/2157-7617.1000158. [116](#), [122](#)
- C. Walck. *Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists*. Privately Published, December 1996. 9070763. [14](#), [42](#)
- L.V. Watrous. *Lasithi: A History of Settlement on a Highland Plain in Crete*. Princeton, NJ: American School of Classical Studies in Athens., 1982. [68](#)

- K. Weghorst. The reclamation drought index: Guidelines and practical applications. *Bureau of Reclamation, Denver (CO)*, 1996. 11
- wikipedia. List of european islands by area, 2015a. URL https://en.wikipedia.org/wiki/List_of_European_islands_by_area. 67
- wikipedia. Fisher-tippett-gnedenko theorem, 2015b. URL https://en.wikipedia.org/wiki/Fisher-Tippett-Gnedenko_theorem. 19
- R.L. Wilby, H. Hassan, and K. Hanaki. Statistical downscaling of hydrometeorological variables using general circulation model output. *Journal of Hydrology*, 205:1–19, 1998. 13
- D.A. Wilhite and M.H. Glantz. Understanding the drought phenomenon: the role of definitions. *Water International*, 10(3):111–120, 1985. doi: 10.1080/02508068508686328. 9
- C.L. Williams. Data analysis and introductory statistical inference with statistical formulae and tables with sas implementations, 2000. 35
- WMO. Climate research: Climate models, 2015. URL <https://www.wmo.int.xiii>, 11, 12
- B. Wolffenbuttel and T. van der Meijden. Mountains of crete, 2015. URL www.discoveronfoot.com. 68
- H. Yanping, A.H. Monahan, C.G. Jones, A. Dai, S. Biner, D. Caya, and K. Winger. Probability distributions of land surface wind speeds over North America. *Journal of Geophysical Research*, 115:1–19, 2010. doi: 10.1029/2008JD010708. 19
- M. Yemefack, D.G. Rossiter, and R. Njomgang. Multi-scale characterization of soil variability within an agricultural landscape mosaic system in southern cameroon. *Geoderma*, 125(1–2):117–143, 2005. 58
- A. Zargar, R. Sadiq, B. Naser, and F.I. Khan. A review of drought indices. *Environmental Reviews*, 19:333–349, 2011. 10

- P. Zhang. Model selection via multifold across validation. *Ann. Statist.*, 21(1): 299–313, 1993. [61](#), [62](#)
- M. Žuković 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. [43](#)