



Università degli Studi di Cagliari

DOTTORATO DI RICERCA
INGEGNERIA DEL TERRITORIO

Ciclo XXVII

Extreme rainfall regime characterization in Sardinia
using daily rainfall data

Settore scientifico disciplinari di afferenza

ICAR/02 Costruzioni idrauliche e marittime e idrologia

Presentata da:	Dott. Matteo Hellies
Coordinatore Dottorato	Prof. Roberto Deidda
Tutor	Prof. Roberto Deidda

Esame finale anno accademico 2014 – 2015



Unione europea
Fondo sociale europeo



REGIONE AUTÒNOMA DE SARDIGNA
REGIONE AUTONOMA DELLA SARDEGNA



University of Cagliari
Faculty of Engineering and Architecture
Department of Civil Engineering, Environmental Engineering
and Architecture (DICAAR)
Doctoral Degree Course in Land Engineering
Cycle XXVII

Extreme rainfall regime characterization in Sardinia

using daily rainfall data

Doctoral candidate Dott. Matteo Hellies

Tutor Prof. Roberto Deidda

Matteo Hellies gratefully acknowledges Sardinian Regional Government for the financial support of her PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 - Axis IV Human Resources Objective 1.3, Line of Activity 1.3.1).

Abstract

For the design of hydraulic structures for flood conveyance and discharge, or protection of territory against flood is fundamental the knowledge of the "extreme rainfall regime" in the area where the hydraulic structures must be set up.

Indeed the design flood is commonly evaluated as output of rainfall-runoff models that receive as input the quantitative description of a rainfall extreme event with a given exceedance probability.

This dissertation assesses the performance of different statistical approaches in characterizing extreme rainfall in the island of Sardinia (Italy).

After a detailed review of the theoretical bases of existing methodologies, we compare the results obtained from the use of:

a) a Generalized Extreme value (GEV) distribution model, and a Two component Extreme Value (TCEV) distribution model, both applied to yearly maxima of daily rainfall, and b) a Generalized Pareto (GP) distribution model applied to rainfall excesses above a properly specified threshold. For the latter purpose, we use the Multiple Threshold Method (MTM) developed by Deidda(2010), which demonstrate good performance also in the case of quantized records.

In order to describe the spatial variation of TCEV, GEV and GP model parameters a regional approach based on homogeneous regions, and two versions of Kriging (a commonly used geostatistical approach) i.e. ordinary Kriging (OK), and Kriging for uncertain Data (KUD), are compared.

The obtained results are very promising, pointing towards the use of: a) a GEV distribution model for yearly rainfall maxima, and a KUD model to describe the spatial variation of model parameters, and b) a GP model for rainfall excesses and either an OK or a KUD model for the spatial variation of model parameters. The reason why the OK and KUD approaches lead to the same results in the GP case, is attributed to the robustness of the MTM method.

Contents

I	Introduction and study area	1
1	Introduction	3
2	Study area and rainfall data	11
2.1	Study area	11
2.2	Database	13
II	Theory and methods	19
3	Principles of Statistics	21
3.1	Basic concepts of probability	21
3.2	Discrete random variable	22
3.3	Continuous random variables	22
3.4	Population moments and their sample estimators	23
3.5	Methods of estimation	24
3.5.1	Maximum Likelihood Estimation	26
3.5.2	Moment Method Estimators	26
3.5.3	Minimize a Loss Function	27
3.6	Probability Weighted Moments and L-moments	27
3.6.1	Similarities and differences between simple moments and L-moment	30
3.7	Plotting position rules	31
3.8	Return Period	32
4	Extreme Value theory	33
4.1	The block maxima (BM) approach	34
4.1.1	The generalized extreme value (GEV) distribution	35
4.2	The peaks over threshold (POT) approach	41
4.2.1	The general Pareto (GP) distribution	42
4.2.2	Threshold selection	47

4.2.3	Multiple Threshold Method (MTM)	48
5	Regional and geostatistical approaches	55
5.1	Regional frequency analysis	56
5.1.1	Index-rainfall method	57
5.1.2	GEV growth curve	57
5.1.3	Identification of homogeneous regions	60
5.1.4	Heterogeneity measures	61
5.1.5	TCEV model	64
5.2	Geostatistical analysis	66
5.2.1	Spatial interpolation with the kriging technique	66
5.2.2	Sample variogram	69
5.2.3	Ordinary kriging and kriging for uncertain data	71
6	Error metrics	75
6.1	Square statistics of Cramer-von Mises' family	75
6.2	Quantiles errors	76
6.3	Cross-validation	78
III	Results and conclusions	79
7	BM results	81
7.1	Local analysis results	83
7.2	Regional analysis results	86
7.2.1	Preliminary analysis	86
7.2.2	Hypothesis of a unique homogeneous region	91
7.2.3	Identification of new homogeneous regions	96
7.2.4	Comparison between regional configurations	129
7.3	Geostatistical analysis results	132
7.4	Regional and geostatistical comparison	144
7.4.1	Spatial distribution of errors	148
7.5	BM summary	153
8	POT results	155
8.1	Local analysis results	156
8.2	Geostatistical analysis results	163
8.3	Spatial distribution of errors	172
8.4	POT and BM comparison	175
9	Conclusions	179

List of Figures

2.1	Digital Elevation Model of Sardinia (Italy). Black and blue circles indicate the location of the stations used for the analyses, the dark circles represent the 229 stations with at least 50 complete years of observations. The inset shows the location of the island relative to the Italian peninsula.	14
2.2	Spatial distribution of the annual rainfall. The black circles indicate the location of the 229 stations used for the spatial interpolation with ordinary kriging.	15
2.3	Main classes of rainfall spatial patterns in the Sardinian region; adapted from Chessa et al. (1999); see main text for details. .	16
2.4	Number of stations sorted by number of full years of observations of daily precipitation.	16
2.5	Spatial distribution of the 256 rainfall stations, with at least 30 complete years of observations. Each station is assigned a unique code number for identification purposes.	17
4.1	The sketch depicts some relations among the cumulative distribution functions (CDFs) $F(x) = Pr\{X \leq x X \geq 0\}$, $F_0(x) = Pr\{X \leq x X > 0\}$, and $F_u(x) = Pr\{X \leq x X > u\}$. Cartesian axes of $F(x)$ are drawn with a thin line and characteristic values are reported on the left side, while the axes of $F_0(x)$ and $F_u(x)$ are drawn with dashed and solid thick lines, respectively, with values reported on the right side (from Deidda, 2010). . .	49

4.2	Station 008 : example of MTM application on a daily rainfall time series collected by a tipping-bucket rain gauge with a 0.2 mm resolution. The first plot from top displays the fraction of the values exceeding different thresholds u in a range from 0 to 30 mm. The second plot from top displays the $\xi(u)$ estimates with increasing threshold u : the ξ_0^M MTM estimate is the median value (horizontal red line) within the range of thresholds between 2.5 and 12.5 mm suggested for practical applications. In the third plot the α_0^M MTM estimate is obtained as the median value of the reparameterized α_0^C estimates conditioned on the ξ_0^M MTM estimate, while in the fourth plot the ζ_0^M MTM estimate is obtained by the ζ_0^C estimates conditioned on both ξ_0^M and α_0^M MTM estimates.	53
6.1	The proposed error metrics measure discrepancies between the theoretical frequency distribution and the empirical one (red lines), and between the observed values and the theoretical quantiles (green lines).	78
7.1	Comparison between pairs of L-moment ratios (L-skewness, L-kurtosis) calculated on annual maxima of daily precipitation for the 229 stations with more than 50 complete years in record (circles) and theoretical pairs for some distributions widely used in statistical hydrology, represented by lines of different strokes. Big marks denote the regional values of the same statistics.	82
7.2	Cumulative distribution function of the GEV parameters estimates obtained with SM, ML and PWM techniques.	84
7.3	Station 002: Empirical cumulative distribution functions (calculated with Hazen's plotting position) of annual maxima of daily precipitation, compared with theoretical TCEV and GEV distributions, whose parameters are locally estimated through SM, ML and PWM techniques.	85
7.4	Classification by quartiles of L-moment ℓ_1 (average) estimates, measured in mm, for the 229 stations with at least 50 complete years of observations.	87
7.5	Classification by quartiles of the coefficients L-CV (top) and CV (bottom) estimates for the 229 stations with at least 50 complete years of observations.	88

7.6	Classification by quartiles of the coefficients L-skewness (top) and skewness (bottom) estimates for the 229 stations with at least 50 complete years of observations.	89
7.7	Classification by quartiles of the coefficients L-kurtosis (top) and kurtosis (bottom) estimates for the 229 stations with at least 50 years of observations.	90
7.8	Hypothesis of a single homogeneous zone for the whole Sardinia. On the left are shown scatterplots with pairs of statistics L-CV, L-skew and L-kurt calculated on the observations of each of the 229 considered sites (the corresponding regional averages are marked in red). On the right are shown, for comparison, the same L-moment ratios calculated on one of the 10,000 synthetic series generated for each of the 229 sites using a Monte Carlo procedure.	93
7.9	Hypothesis of a single homogeneous zone for the whole Sardinia. Left: in black it's shown the CDF of dispersion measures V , V_2 and V_3 (top to bottom) obtained from 10'000 synthetic generations for each station, the vertical red lines represent the corresponding sample values. Middle: in black it's shown the CDF of standard deviations of statistics L-CV, L-skewness and L-kurtosis (top to bottom) obtained from 10'000 synthetic generations for each station, the vertical red lines represent the corresponding sample values. Right: test of uniformity of the exceedance probability for statistics L-CV, L-skewness and L-kurtosis (top to bottom) calculated from the CDF obtained from 10'000 synthetic generations for each station.	95
7.10	Spatial distribution of the 7 homogeneous regions obtained by cluster analysis with metric L-CV. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	97
7.11	Hypothesis A: Spatial distribution of the 5 homogeneous regions obtained by cluster analysis with metric L-CV. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	99
7.12	Hypothesis A: Cluster A₁ Same description of the Figures 7.8 and 7.9	100
7.13	Hypothesis A: Cluster A₂ Same description of the Figures 7.8 and 7.9	101
7.14	Hypothesis A: Cluster A₃ Same description of the Figures 7.8 and 7.9	102

7.15	Hypothesis A: Cluster A₄ Same description of the Figures 7.8 and 7.9	103
7.16	Hypothesis A: Cluster A₅ Same description of the Figures 7.8 and 7.9	104
7.17	Hypothesis B: Spatial distribution of the 4 homogeneous regions obtained by cluster analysis with metric L-CV. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	105
7.18	Hypothesis B: Cluster B₃ Same description of the Figures 7.8 and 7.9	106
7.19	Left: scatterplots of the L-moments ratios t and t_3 , obtained through a Monte Carlo procedure. The length of each time series is equal to 50, 70 and 90 years (top to bottom). Middle: scatterplots of the standard deviation of the L-moments ratios t and t_3 , obtained by a Monte Carlo procedure. The length of each time series is equal to 50, 70 and 90 years (top to bottom). Right: scatterplots of the ratio between the standard deviation of the L-moments ratios t and t_3 , obtained by a Monte Carlo procedure. The length of each time series is equal to 50, 70 and 90 years (top to bottom).	108
7.20	Spatial distribution of the 8 homogeneous regions obtained by cluster analysis with metrics L-CV and L-skew properly weighted. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	109
7.21	Hypothesis C: Spatial distribution of the 5 homogeneous regions obtained by cluster analysis with metrics L-CV and L-skew properly weighted. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	111
7.22	Hypothesis C: Cluster C₁ Same description of the Figures 7.8 and 7.9	112
7.23	Hypothesis C: Cluster C₂ Same description of the Figures 7.8 and 7.9	113
7.24	Hypothesis C: Cluster C₃ Same description of the Figures 7.8 and 7.9	114
7.25	Hypothesis C: Cluster C₄ Same description of the Figures 7.8 and 7.9	115
7.26	Hypothesis C: Cluster C₅ Same description of the Figures 7.8 and 7.9	116

7.27	Hypothesis D: Spatial distribution of the 4 homogeneous regions obtained by cluster analysis with metrics L-CV and L-skew properly weighted. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	117
7.28	Hypothesis D: Cluster D₃ Same description of the Figures 7.8 and 7.9	118
7.29	Hypothesis E: Spatial distribution of the 5 homogeneous regions obtained from hypotheses A and C through empirical aggregation. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	121
7.30	Hypothesis E: Cluster E₁ Same description of the Figures 7.8 and 7.9	122
7.31	Hypothesis E: Cluster E₂ Same description of the Figures 7.8 and 7.9	123
7.32	Hypothesis E: Cluster E₃ Same description of the Figures 7.8 and 7.9	124
7.33	Hypothesis E: Cluster E₄ Same description of the Figures 7.8 and 7.9	125
7.34	Hypothesis E: Cluster E₅ Same description of the Figures 7.8 and 7.9	126
7.35	Hypothesis F: Spatial distribution of the 4 homogeneous regions obtained from hypotheses B and D through empirical aggregation. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.	127
7.36	Hypothesis F: Cluster F₃ Same description of the Figures 7.8 and 7.9	128
7.37	Comparison between pairs of L-moment ratios (L-skewness, L-kurtosis) calculated on annual maximum daily precipitation for the 229 stations with more than 50 complete years in record (circles), partitioned according to the cluster allocation (symbols of different color for each cluster, consistent with Figures 7.11, 7.17, 7.21, 7.27, 7.29 and 7.35) and theoretical pairs for some distributions widely used in statistical hydrology, represented by lines of different strokes. Big marks denote, for each cluster, the regional values of the same statistics. From top to bottom and from left to right are reported cases related to the hypothesis A , B , C , D , E ed F of division into homogeneous regions.	130

- 7.38 Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the shape parameter κ . The ordinate shows the mean absolute error (*MAE*) in the interpolation of the parameter κ (using cross-validation) in each of the 229 stations with at least 50 complete years of observations (excluding time to time the estimate of the station considered), in function of the number of nearest stations (on the abscissa) used for the kriging system. The empty circles represent the results with the OK, the asterisks refer to the KUD. The different colors refer to the minimum number of years to select the stations to be used for the interpolations. 134
- 7.39 Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the dimensionless scale parameter σ^* . The symbolism is the same as for Figure 7.38, but *MAE* is calculated on the interpolations of the σ^* parameter using cross-validation. 135
- 7.40 Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the index-rainfall m . The symbolism is the same as for Figure 7.38, but *MAE* is calculated on the interpolations of m using cross-validation. 136
- 7.41 Comparison between sample variograms, based on local estimates of GEV growth curve parameters κ , σ^* , the index-rainfall m (from top to bottom), and some theoretical variograms. 139
- 7.42 Representation of the spatial distribution of the shape parameter of the GEV distribution, κ . The map is obtained from a regular grid with 1 km resolution. 140
- 7.43 Representation of the spatial distribution of the dimensionless scale parameter of the GEV distribution, σ^* . The map is obtained from a regular grid with 1 km resolution. 141
- 7.44 Representation of the spatial distribution of the index-rainfall, m . The map is obtained from a regular grid with 1 km resolution. 142
- 7.45 Representation of the spatial distribution of the shape parameter of the GEV distribution, κ . The map on the left is obtained using the ordinary kriging (OK), while the map on the right is obtained using kriging for uncertain data (KUD). 143

-
- 7.46 Map of daily rainfall depth h_T (mm) exceeded with return period $T=200$ yr. Left: the result obtained using the regional approach (case **F** with 4 clusters). Right: the result obtained using the geostatistical approach. 147
- 7.47 Spatial distribution of the error metric $ME(5)$ for the cases: GEV with local parameters (top left), GEV with regional parameters related to hypothesis **F** (top right), GEV with parameters estimated by kriging on a regular grid at 1 km resolution (bottom left), TCEV with index-rainfall updated to 2008 (bottom right). The same index-rainfall has been used also for the other distributions, in order to ensure a fair comparison. 149
- 7.48 Same representations used in Figure 7.47, but on the error metric $MEr(5)$ 150
- 7.49 Same representations used in Figure 7.47, but on the error metric $ME(5)$ calculated after the **cross-validation** procedure. 151
- 7.50 Same representations used in Figure 7.47, but on the error metric $MEr(5)$ calculated after the **cross-validation** procedure. 152
- 8.1 Comparison between pairs of L-moment ratios (L-skewness, L-kurtosis) calculated on daily rainfall depths exceeding 5 mm, for the 256 stations with more than 30 complete years in record (circles) and theoretical pairs for some distributions widely used in statistical hydrology, represented by lines of different strokes. Big marks denote the average values of the same statistics. 156
- 8.2 CDF of rainy data collected by station 007. The zoom shows how many data are rounded off to discrete values. 157
- 8.3 Cumulative distribution function of the MTM-GP parameters: ξ_0^M , α_0^M and ζ_0^M , from top to bottom. Estimates are obtained with SM, ML and PWM techniques. 160
- 8.4 Empirical cumulative distribution functions (calculated with Hazen's plotting position) of daily precipitation, compared with theoretical MTM-GP distributions, whose parameters are locally estimated through SM, ML and PWM techniques. Top: station 001. Bottom: station 323 161

-
- 8.5 Top: Representation of the spatial distribution of the shape parameter of the MTM-GP distribution, ξ_0^M , obtained using SM, PWM and ML techniques, from left to right. Bottom: Map of the rainfall depth h_T (mm) exceeded with return period $T=200$ yr using MTM-GP estimates obtained with SM, PWM and ML techniques, from left to right. 162
- 8.6 Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the shape parameter ξ_0^M . The ordinate shows the mean absolute error (MAE) in the interpolation of the parameter ξ_0^M (using cross-validation) in each of the 229 stations with at least 50 years of observations (excluding time to time the estimate of the considered station), in function of the number of nearest stations (on the abscissa) used for the kriging system. The empty circles represent the results with the ordinary kriging, the asterisks refer to the kriging with uncertain data. The different colors refer to the minimum number of years to select the stations to be used for the interpolations. 164
- 8.7 Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the scale parameter α_0^M . The symbolism is the same as for Figure 8.6, but MAE are calculated on the interpolations of the α_0^M , parameter using cross-validation. 165
- 8.8 Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the ζ_0^M parameter. The symbolism is the same as for Figure 8.6, but MAE are calculated on the interpolations of ζ_0^M , using cross-validation. 166
- 8.9 Comparison between sample variograms, based on local estimates of parameters ξ_0^M , α_0^M and ζ_0^M (from top to bottom), and some theoretical variograms. 168
- 8.10 Representation of the spatial distribution of the MTM-GP shape parameter, ξ_0^M . The map is obtained from a regular grid with 1 km resolution. 169
- 8.11 Representation of the spatial distribution of the MTM-GP scale parameter, α_0^M . The map is obtained from a regular grid with 1 km resolution. 170
- 8.12 Representation of the spatial distribution of the MTM-GP parameter, ζ_0^M . The map is obtained from a regular grid with 1 km resolution. 171

-
- 8.13 Spatial distribution of the error metric $ME(5)$ for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution (right). 173
- 8.14 Spatial distribution of the error metric $MEr(5)$ for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution (right). 173
- 8.15 Spatial distribution of the error metric $ME(5)$, for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution, calculated after the **cross-validation** procedure (right). . . . 174
- 8.16 Spatial distribution of the error metric $MEr(5)$, for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution, calculated after the **cross-validation** procedure (right). . . . 174
- 8.17 Comparison between rainfall depths h_T exceed with different time return periods: 10yr (top left), 50yr (top right), 100yr (bottom left) and 200yr (bottom right) for each of the 229 stations with at least 50 complete years of observations. In the x-axis, for each figure, it is reported the rainfall depths obtained with the MTM-GP geostatistical model. In the y-axis it is reported the rainfall depths obtained with the GEV geostatistical model. 176
- 8.18 Map of daily rainfall depth h_T (mm) exceeded with return period $T=200$ yr. Left: the result obtained using the BM approach and the GEV geostatistical model. Right: the result obtained using the POT approach and the MTM-GP geostatistical model. 177

List of Tables

3.1	Symbols and names used for the statistics calculated with simple moments and L-moments, theoretical and sample.	29
7.1	Comparison between: local fits of GEV distribution (whose parameters are estimated by ML, SM, PWM methods) and regional fits of TCEV distribution. The averages of error metrics are calculated over the 229 stations with at least 50 complete years of observations.	85
7.2	Hypothesis of unique homogeneous zone: regional values of L-moment ratios calculated on 229 stations, and corresponding estimates of Kappa distribution.	91
7.3	Dispersion measures and heterogeneity measures, evaluated on the 229 stations, under the hypothesis of a unique homogeneous region for the whole Sardinia island.	92
7.4	L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h , for the 7 homogeneous zones obtained with L-CV metric.	96
7.5	Hypothesis A: partition in 5 homogeneous regions obtained through cluster analysis with the L-CV metric. L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h are reported for each homogeneous region.	98
7.6	Hypothesis B: partition in 4 homogeneous regions obtained through cluster analysis with the L-CV metric. L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h are reported for each homogeneous region.	98
7.7	L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h , for the 8 homogeneous zones obtained with weighted L-CV and L-skewness metrics.	110

7.8	Hypothesis C: partition in 5 homogeneous regions obtained through cluster analysis with weighted L-CV and L-skewness metrics. L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h are reported for each homogeneous region.	110
7.9	Hypothesis D: partition in 4 homogeneous regions obtained through cluster analysis with weighted L-CV and L-skewness metrics. L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h are reported for each homogeneous region.	119
7.10	Hypothesis E: partition in 5 homogeneous regions obtained from hypotheses A and C through empirical aggregation. L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h are reported for each homogeneous region.	120
7.11	Hypothesis F: partition in 4 homogeneous regions obtained from hypotheses B and D through empirical aggregation. L-moment ratios, heterogeneity measures H , H_2 , H_3 and <i>Kappa</i> distribution's parameter h are reported for each homogeneous region.	120
7.12	Estimation of GEV growth curve parameters (κ , σ^* and μ^*) for every cluster of each of the six hypothesis of partition in homogeneous regions.	131
7.13	Comparison between the performances of regional fits with the GEV distribution (with parameters estimated with PWM method) and regional fits with the TCEV distribution. Mean of error metrics calculated over the 229 stations with at least 50 complete years of observations.	132
7.14	Comparisons between the performances of local and regional fits using: local GEV distribution, regional fit with TCEV distribution, regional fits (hypotheses E and F) with GEV distribution and kriging fit. Average of error metrics calculated over the 229 stations with at least 50 complete years of observations.	146
7.15	Same results presented in Table 7.14, but obtained with the cross-validation procedure for the regional GEV estimations and the kriging estimations.	146

8.1	Comparison between performances of local fits with the MTM-GPD model, with parameters estimated with ML, SM and PWM techniques. The averages of error metrics are calculated over the 256 stations with at least 30 complete years of observations.	159
8.2	Comparisons between the performances of local and kriging fit. Averages of error metrics calculated over the 256 stations with at least 30 complete years of observations.	167
8.3	Comparisons between the performances of kriging fit using the GEV and MTM-GP models. Average of error metrics calculated over the 229 stations with at least 50 complete years of observations.	178

Part I

Introduction and study area

Chapter 1

Introduction

Background

The design of hydraulic structures for flood conveyance and discharge, or protection of territory against flood, is based on “design floods” characterized by a given (small) exceedance probability. The determination of the design flood requires the knowledge of the “extreme rainfall regime” in the area where the hydraulic structures must be set up. Indeed the design flood is commonly evaluated as output of rainfall-runoff models that receive as input the quantitative description of a rainfall extreme event with the same exceedance probability required for the design flood. The importance of an accurate assessment and management of flood risk is well known among scientists and practitioners worldwide. Adequate dimensioning of engineering works and reliable evaluation of flood risk are strictly connected with the accuracy in the estimation of extreme rainfall events that are used to determine the design floods.

Considering daily rainfall time series, the ordinary rainfall regime describes the everyday rainfall characteristics, such as wet/dry occurrences and the probabilistic distribution of non-zero rainfall depth on dry days, while the extreme rainfall regime describes the situation of maxima among observations within time blocks of fixed size. This approach is usually referred to as Block Maxima (BM) approach. In hydrological applications the size of the time block is usually assumed to be one year, thus studies on extremes are commonly performed on annual maxima series.

Using the BM approach sample sizes are generally small since they correspond to the number of years of observations at each site. This means that, whatever parametric probability distribution is selected among the many proposed to interpret observed maxima, statistical inference of unknown pa-

rameters is affected by estimation errors. Distributions with at least three parameters are usually needed to correctly reproduce the coefficient of variation (CV) and the skewness (Skew) of rainfall extremes. This can cause some uncertainty in the estimation procedure, because the higher the number of parameters, the higher the parameter estimation error.

The General Extreme Value (GEV) distribution is the limiting distribution of the maximum of a series of independent and identically distributed observations (Fisher and Tippett, 1928; Gnedenko, 1943; Coles, 2001), and it is widely adopted for modeling annual maxima. Mathematically the GEV distribution is very attractive because its inverse has a closed form, and its parameters are easily estimated using regular moments or L-moments. The GEV distribution reduces to a Gumbel distribution when the shape parameter tends to zero. The Gumbel distribution has been the prevailing model of extreme rainfall for half a century, but recent works (Koutsoyiannis, 2004a,b; Papalexiou and Koutsoyiannis, 2013) show that that GEV distribution with shape parameter values corresponding to heavy-tailed Fréchet distribution is a more consistent choice. The success of the Gumbel distribution is often due to the small size of samples, usually less than 50 annual maxima, that tends to hide the heavy tail behavior.

Alternatively to the BM analysis, studies of extremes based on Peaks Over Threshold (POT) have been proposed (Van Montfort and Witter, 1986; Rosbjerg et al., 1992; Madsen et al., 1997b,a). A POT analysis is carried out on the observations above a certain threshold value with the aim to enlarge the size of the sample and consequently improve the accuracy of extreme estimates. Nevertheless, the POT approach has found only a few applications in the study of floods and even less applications to rainfall characterization (Coles et al., 2003; De Michele and Salvadori, 2005; Deidda and Puliga, 2006, 2009; Deidda, 2010, and some others). The majority of the studies on rainfall extremes are based on the BM approach, mainly because only annual maxima are easily available for past rainfall occurrences. The Generalized Pareto distribution (GP) is widely adopted for modeling the peaks over a threshold. Papalexiou et al. (2013) compared, using more than use 15000 daily rainfall records from around the world with record lengths from 50 to 172 yr, the upper part of the empirical distributions with four common distribution tails: Pareto, Lognormal, Weibull and Gamma. They found that the Pareto and Lognormal distributions (heavier-tailed distributions) perform better with respect to the Weibull and Gamma distributions (lighter tailed distributions).

In Deidda and Puliga (2006), the authors, analyzed several time series of daily precipitation collected in Sardinia, the study area of this research project. Using the L-moment ratios diagram proposed by Hosking (1990)

they found that the GP is the best distribution to describe daily rainfall exceedances above a fixed threshold. It is also important to note that Deidda and Puliga (2006); Deidda (2007) highlighted the problems arising in estimating GP parameters from heavily rounded-off data. In fact, the presence of a large portion of heavily rounded off daily rainfall data poses some problems in fitting parametric distributions and in interpreting results of goodness of fit tests. As highlighted by Deidda (2007), this problem may concern a large number of rainfall time series whose data have been collected by non recording gauges. Recently Deidda (2010) proposed a multiple threshold model (MTM) for fitting the GP distribution to rainfall series (in what follows this model is referred to as MTM-GP). Using this model, it is possible to overcome problems related to the choice of the optimal threshold and the presence of highly rounded-off data.

Several papers in the hydrological literature investigate whether the BM or the POT approaches give the best quantile estimates (e.g. Madsen et al., 1997b; Martins and Stedinger, 2001; Villarini et al., 2011). The POT approach is assumed to be more precise than the BM approach with an one-year block, especially for short data series, because it uses more data and the annual maxima are not always true extremes. It remains, however, the crucial decision of how to choose a threshold value in POT analysis (e.g. Tanaka and Takara, 2002), which is the the main reason why this approach is less widespread than the BM. Also, Bezak et al. (2014) confirm these results, and, in addition, they find a better performance of the method of L-moments relative to conventional moments and maximum likelihood.

Extreme rainfall regimes are usually characterized by the statistical analysis of annual maxima collected in different sites, and then merged together using regionalisation procedures. With this approach it is possible to use information at different gauged sites to compensate for short records at a single site, and to obtain rainfall quantiles at locations where no measurements are available. In summary, a regional approach assumes that the mean could vary from point to point and should be estimated “at-site” (i.e. on the records of each gage), while dimensionless moments (CV, Skew, etc.) are constant in large homogeneous regions and, thus, can be estimated by merging standardized records from different gages inside each region. In such a way, the regional approach obviously reduces the estimation error, since the parameter related to the mean of the rainfall field can be estimated with good accuracy even from small local samples “at-site”, while the other parameters (depending on dimensionless moments) are estimated from larger merged samples. Nevertheless, the hypothesis of constant dimensionless moments may be unrealistic within presumed homogeneous areas with complex to-

pography. Moreover, it poses boundary problems, as discussed in Schaefer (1990), since it implies a jump from one probability distribution to another when crossing the boundaries between regions.

In Italy, the most documented methodology of regionalization is based on the Two-Component Extreme Value (TCEV) distribution (Cannarozzo et al., 1995; Ferro and Porto, 1997), which is a four parameters distribution. In the last years it is slowly being replaced by probabilistic models based on the GEV distribution. In Sardinia the last regional frequency analysis of annual maxima of daily precipitation has been performed more than 15 years ago (Deidda and Piga, 1998; Deidda et al., 2000) using a probabilistic model based on the TCEV distribution. In that work, the authors distinguished three sub-zones in the island with similar characteristics regarding the probabilistic annual maximum daily rainfall model. No regional frequency analysis were carried out using a POT approach in Sardinia.

Madsen et al. (1997a) performed a comparison between two regional models, one based on the BM approach and the other on the POT approach. They used the GEV distribution for the first and the GP for the latter, and found that the POT/GP model is in general more efficient in regions with a positive shape parameter (heavier-tailed distributions), whereas in regions with a negative shape parameter (lighter tailed distributions) the BM/GEV model is preferable.

Although the regional approach is commonly adopted worldwide (see e.g. Fitzgerald, 1989; Hosking and Wallis, 1993, 1997) it exhibits at least two limitations. The first one is that the variation of the parameters due to physical heterogeneity and topography is not reproduced inside an homogeneous region, since they are kept (except for the mean) constant. Any variability, even if observed, is due to sampling, and the acceptance tests of homogeneous regions are built on this principle. On the contrary, abrupt changes of parameter values at the boundaries of adjacent homogeneous regions are possible, which cannot be explained through physical interpretations. The second limitation concerns practical applications, when more than one homogeneous region fall within the same basin. In that case the assignment of parameter values is ambiguous and problematic. This ambiguity is intensified by uncertainties in the definition of hypothetical borders between homogeneous regions, due to both the low density of observation points and to the process of grouping stations to homogeneous regions. The latter can create different configurations of homogeneous regions, according to the specific grouping criterion used.

In fact, in the regional approach there are several subjective choices in the identification of homogeneous regions: aggregation criteria (eg. metrics used for cluster analysis, morphometric criteria, climate, etc.), homogeneity checks

(statistical tests, metrics), inevitable subjective choices in merging/splitting clusters and reassigning stations.

Most of the drawbacks of the regional approach can be bypassed using a geostatistical approach, that allows for continuous representations of the spatial distributions of interest (mean rainfall intensity, distribution parameters, etc). This approach allows to represent local peculiarities (that may be induced by several factors, like exposure, morphometric, climate and microclimate) better than the regional one. In addition the geostatistical approach overcomes the problems associated with abrupt discontinuities of the parameters of the probability distributions at the border between contiguous homogeneous regions. The kriging is a geostatistical technique initially proposed by Krige (1951), and improved by Matheron (1963), and even now it is probably the most widely used technique for spatial interpolation. Prudhomme and Reed (1999) use ordinary kriging and modified residual kriging to map the median of the annual maximum daily rainfall (RMED) in Scotland. Ceresetti et al. (2012) used kriging with an external drift and neural network techniques to interpolate the locally estimated parameters of the distributions of two extreme-value models (BM and POT) throughout the Cèvennes-Vivarais region. They found that the best results are obtained when combining the POT method with kriging. Blanchet and Lehning (2010), with the objective of mapping snow depth return levels in Switzerland, interpolated the local estimates of GEV distribution parameters using several methods: inverse distance, linear regression models, spline-based regression models and kriging. Among these, kriging performed best.

Until now no geostatistical approach has been proposed and applied in Sardinia for the frequency distribution of daily rainfall data.

Thesis's objectives

This research aims to improve the characterization of extreme rainfall regime in Sardinia using daily rainfall time series provided by the National Hydrographic Service (SI). In order to achieve this objective, the analysis is conducted using both the BM and POT approaches. The frequency analysis is performed using both a regional approach and a geostatistical model.

The first phase of the research project regards the study of the annual maxima of daily precipitation. One of the objectives is to test the hypothesis that the GEV distribution is a valid model for the description of annual maxima of daily precipitation in Sardinia, as suggested by several studies conducted in other parts of the world and by extreme value theory. The GEV model outputs are compared with those obtained using the TCEV model. If this hypothesis is confirmed, it will lead to considerable practical advantages. In fact, the GEV distribution has three parameters, unlike the TCEV distribution that has four parameters, and, as previously said, the higher the number of parameters the higher the parameter estimation errors.

Another goal of this research is to understand if the regional approach, which is widely used in hydrology, is still a valid approach to describe the spatial distribution of extreme rainfall regimes, or whether it is better to set aside this approach in favor of a geostatistical model. The latter consists a better representation of local peculiarities, and is not affected by boundary problems.

The second and last phase of the research project regards the study of daily precipitation depths over a certain threshold, with the objective to develop a regional/geostatistical approach based on the MTM-GP model recently proposed by Deidda (2010). Applying this model to a large database, like the one at our disposal, we can better understand its strengths and weaknesses. In particular, the sensibility of the model regards the parameter estimation technique, and the presence of a large number of series containing rounded-off recordings. If the MTM-GP model proves to be robust, and provides reliable estimates, it will be extremely useful to investigate the spatial pattern of rainfall signature in the context of regional/geostatistical analyses. In fact, using a POT approach and a GP distribution, the scale parameter depends not only on the local climatic conditions but also on the at-site optimum threshold. Instead the parameters of the MTM-GP model do not depend on the threshold used for GP fitting, but only on the local climatic features. So the MTM-GP model is more suitable for regional/geostatistical analysis than a simple GP model.

In addition, the results of this research will be summarized using simple formulations for practitioners in professional engineering that are involved in

hydraulic designs.

Thesis structure

The thesis is organized as follows:

Chapter 2 describes the study area, which is the island of Sardinia, and the database.

Chapter 3 reviews some principles of statistics, in order to have a better understanding of the material presented in the following chapters. Particular attention is given to the description of the L-moment and L-moment ratios.

Chapter 4 summarizes the key background material regarding extreme value theory. In particular, the GEV distribution, the GP distribution and the MTM-GP model are presented..

Chapter 5 briefly reviews the theory of regional frequency analysis based on the index-rainfall method, and the theory of geostatistical analysis using two different types of kriging techniques.

Chapter 6 reports the error metrics used for validation of the obtained results.

Chapter 7 reports the results of the frequency analysis of annual maxima of daily precipitation. Both the regional and geostatistical approaches are used and compared.

Chapter 8 reports the results of the frequency analysis of daily precipitation. Only the geostatistical approach is used.

Chapter 9 is dedicated to the conclusions.

Chapter 2

Study area and rainfall data

2.1 Study area

The study area is Sardinia (Figure 2.1), an island located in the Mediterranean sea, about 400 km west off of the Italian peninsula (inset in Figure 2.1), between 32°N and 41°N latitude and 8°E and 10°E longitude.

The surface of the island is $\sim 24\,000$ km²; its topography is rather complex, as shown in the Digital Elevation Model (DEM) reported in Figure 2.1. A long mountain range is located in the East part of the island, running from north to South, with highest elevation of 1834 m; and a smaller mountain range is located in the south East zone. Between them, the Campidano plain is formed.

The climate is Mediterranean, characterized by dry summers and rainfall mainly occurring during the period from September to May. A climatic North–South gradient is present, due to the latitude development of the island. In addition to this, especially in autumn, there is a longitudinal precipitation gradient, due to the presence of the major mountain range and to the hot and moist air current coming from North Africa.

The Central-East and the South-West areas of the island experience the highest extreme events, due to the interaction of the two mountain ranges with the hot and moist currents coming from Africa. Soil structure and utilization make this area highly vulnerable to flash floods and landslides.

Figure 2.2 shows the distribution of annual rainfall averages obtained from 50-yr-long rainfall records collected by 229 rain gauges operating at daily resolution. The map has been obtained using ordinary kriging, see section 5.2.3. A strong relation between the annual rainfall depth and elevation is revealed when comparing Figures 2.1 and Figure 2.2. In areas of lower elevation, the total rainfall is about 450 mm per year, reaching 1190 mm at

the highest mountains. The regional mean is 728 mm per year. The East zone of the island is also characterized by the highest frequency of severe events, as previously reported.

Chessa et al. (1999) applied different cluster analysis techniques to study the winter (from September to May) rainfall regimes over Sardinia and the linkages to synoptic circulation. The analysis was performed using daily rainfall depths from 114 gauges and spatial fields of meteorological variables at 5° resolution, provided by the National Center for Atmospheric Research (NCAR) analysis. These authors identified three main clusters of rainfall spatial patterns in the island, reported in Figure 2.3, associated with different dominant synoptic conditions. Clusters 1 and 2 are characterized by a limited negative gradient of rainfall intensity from SW to NE (cluster 1) and from NW to SE (cluster 2). In both cases, the Sardinian-Corse Mountain System leads to lower precipitation amount in the eastern part of the island, and the dominant synoptic patterns are characterized by north-westerly flows (the Mistral wind) bringing large frontal systems. Cluster 3 is completely different and is characterized by a strong East-West negative rainfall gradient, with synoptic circulation associated with Atlantic flow passing over Northern Africa and crossing the southern part of the Mediterranean sea. Under these conditions, moist air at lower levels of the atmosphere is transported towards Sardinia by south-easterly winds (the Sirocco wind) while, simultaneously, cold air arrives at upper levels from the North. This potential instability state is further enhanced by the orographic barrier in the eastern part of the island and by the mountain ranges in the south (Figure 2.1). Under this type of synoptic conditions, precipitation events of high intensity (frequently on the order of 300 mm and sometimes of more than 500 mm accumulated in 24 h, with peaks exceeding 100 mm in less than 1 h) have been observed, especially during the autumn season when the sea temperature is relatively high. These storms have caused severe floods in the territories located along the eastern coast of the island and close to Cagliari (the main town), with significant property damage and loss of lives (Chessa et al., 2004).

2.2 Database

The frequency analysis has been performed using daily rainfall depths provided by the National Hydrographic Service (SI). The SI has more than 400 rain gauges, mostly mechanical and manually operated, distributed throughout the whole Sardinia.

In this research, the data we analyze have been obtained through integration of a database characterized by observations from 1922 to 1980, used in the past by the University of Cagliari, and data from 1922 to 2008, supplied by the Hydrographic Agency District of Sardinia (ADIS). The database containing the observations from 1922 to 1980 has been used more than 15 years ago for a regional frequency analysis based on the TCEV distribution (see section 5.1.5), within the VAPI project. The VAPI Project aimed to establish a uniform procedure for the evaluation of natural flood discharges in Italy, with the purpose of providing a tool and a guide for researchers and technicians to understand the phenomena involved in the production of natural floods, as well as to make predictions about future flood values in an unregulated section of the basin.

Data from the two aforementioned databases were compared considering the common period of observation, in order to eliminate inconsistencies.

The final database consists of 441 time series with a period of observation between 1920 and 2008. Among these stations, we selected 229 with at least 50 complete years of observations to estimate magnitudes affected by high uncertainty, and 256 stations with at least 30 complete years of observations to estimate magnitudes affected by low uncertainty.

The spatial distribution of these stations is illustrated in Figure 2.1, between them the 229 stations with at least 50 complete years of observations are marked with dark circles. Observing Figure 2.4, which connects the large number of stations with the lengths of the periods of observation, it is clear that this choice is a good compromise between the number of stations and the length of the series. Figure 2.5 shows the location of the 256 stations and the codes used to identify them in Appendices and figures.

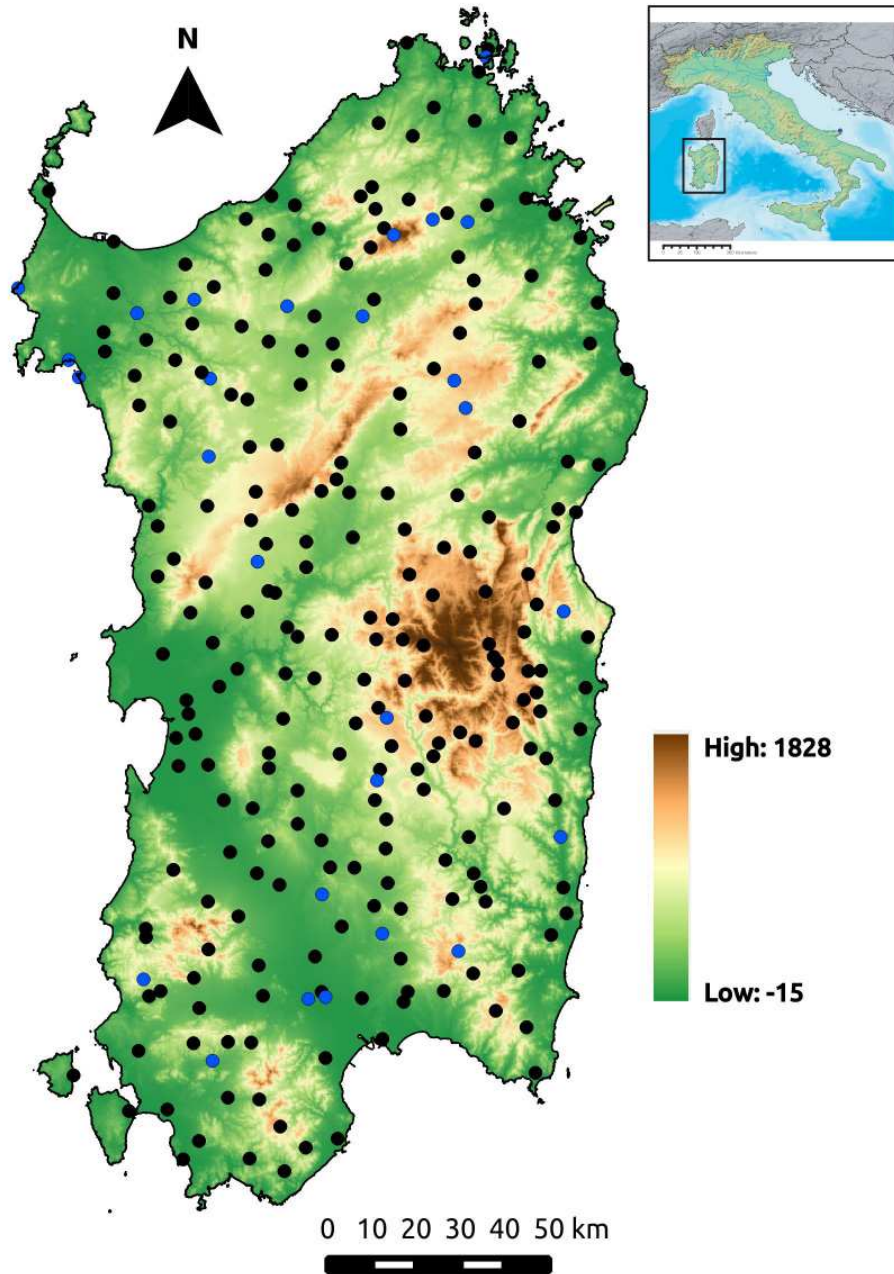


Figure 2.1: Digital Elevation Model of Sardinia (Italy). Black and blue circles indicate the location of the stations used for the analyses, the dark circles represent the 229 stations with at least 50 complete years of observations. The inset shows the location of the island relative to the Italian peninsula.

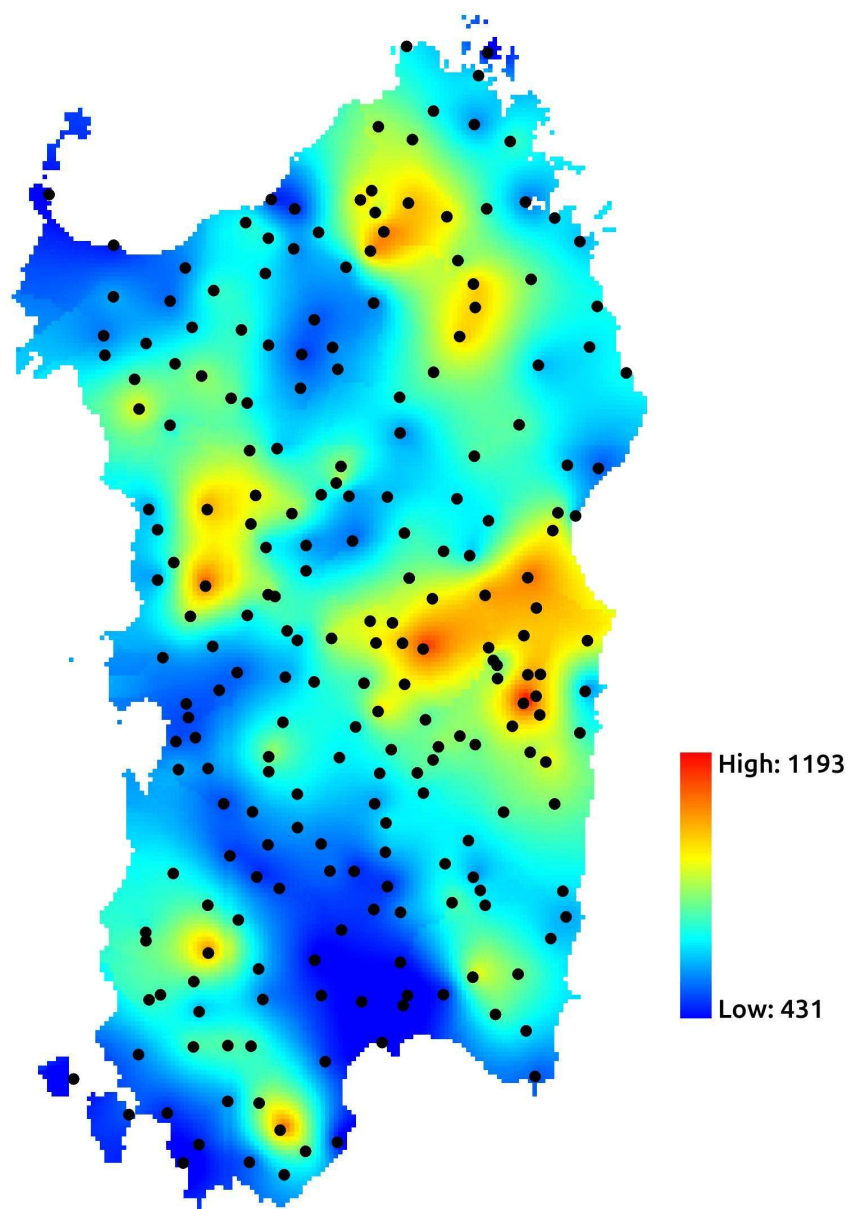


Figure 2.2: Spatial distribution of the annual rainfall. The black circles indicate the location of the 229 stations used for the spatial interpolation with ordinary kriging.

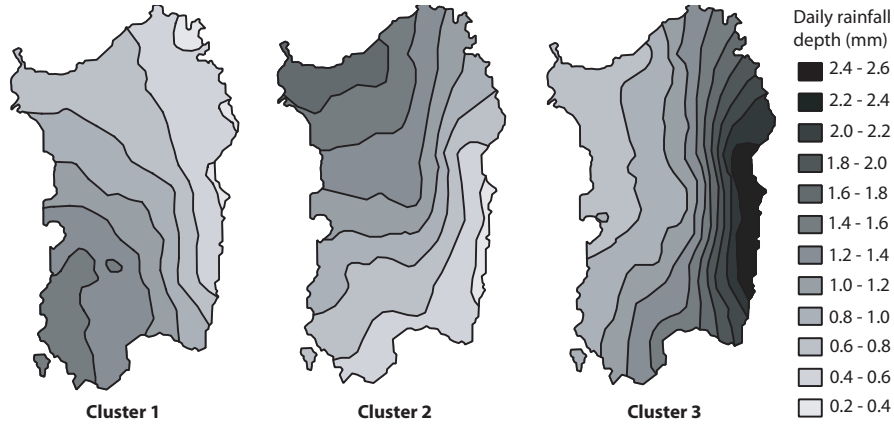


Figure 2.3: Main classes of rainfall spatial patterns in the Sardinian region; adapted from Chessa et al. (1999); see main text for details.

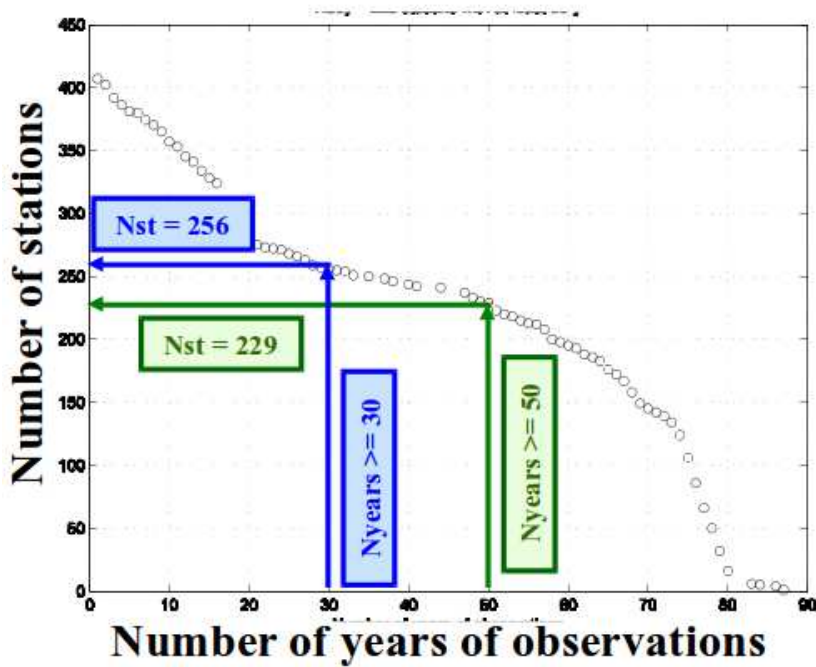


Figure 2.4: Number of stations sorted by number of full years of observations of daily precipitation.

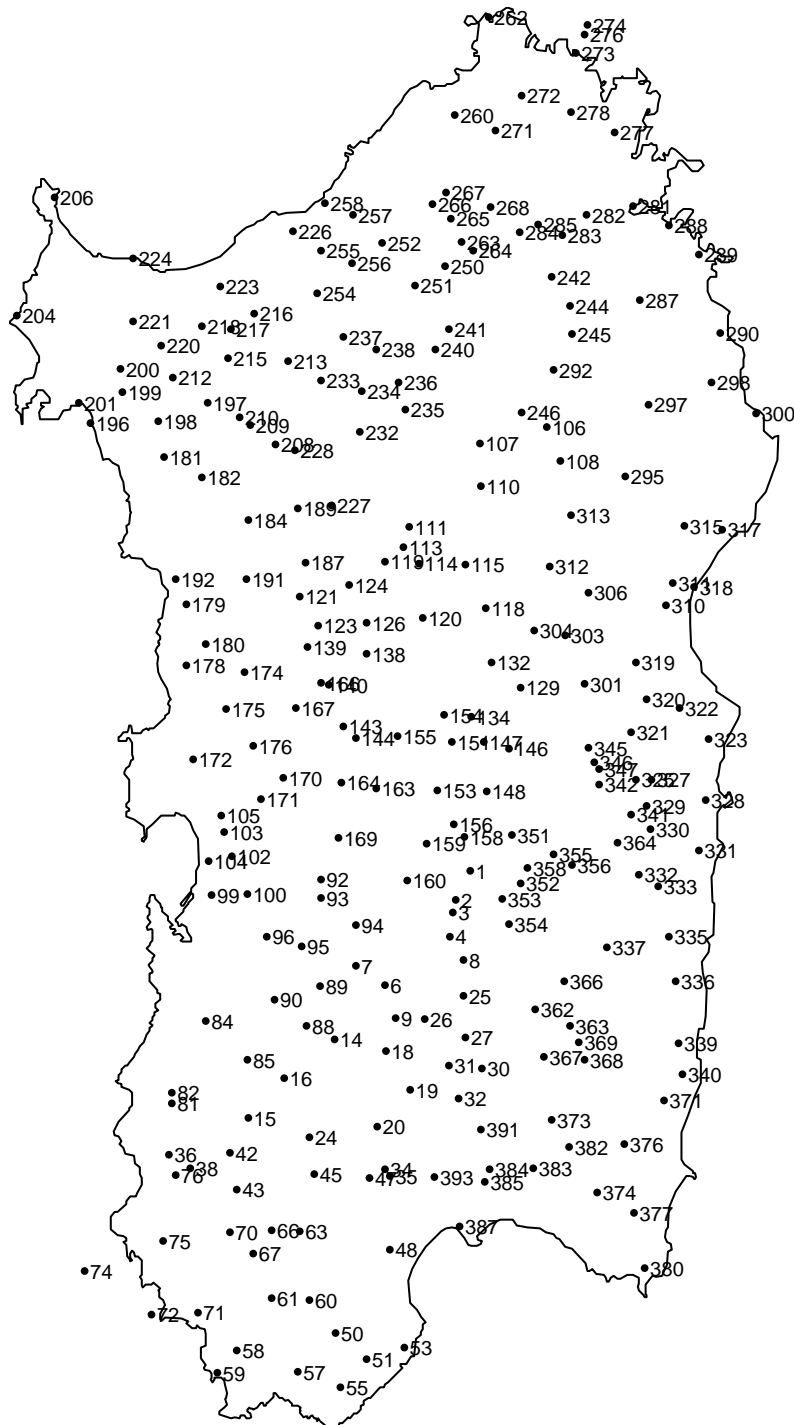
Station codes – Nyears \geq 30 – Nstaz = 256

Figure 2.5: Spatial distribution of the 256 rainfall stations, with at least 30 complete years of observations. Each station is assigned a unique code number for identification purposes.

Part II

Theory and methods

Chapter 3

Principles of Statistics

3.1 Basic concepts of probability

Kolmogorov's approach to probability theory is based on the notion of measure, which maps sets into numbers. The objects of probability theory, the events, to which probability is assigned, are thought of as sets.

Based on Kolmogorov's axiomatization, probability theory is based on three fundamental concepts:

1. A non-empty set Ω , sometimes called the *basic set*, *sample space* or the *certain event*, whose elements ω are known as outcomes or states.
2. A set Σ known as σ -*algebra* or σ -*field* whose elements E are subsets of Ω , known as events. Ω and \emptyset are both members of Σ , and, in addition:
 - if E is in Σ then the complement $\Omega - E$ is in Σ ;
 - the union of countably many sets in Σ is also in Σ .
3. A function P called *probability* that maps events to real numbers, assigning each event E (member of Σ) a number between 0 and 1.

The triplet (Ω, Σ, P) is called *probability space*.

A *random variable* X is a function that maps outcomes to numbers. More formally, a real single-valued function $X(\omega)$, defined on the basic set Ω , is called a random variable if for each choice of a real number α the set $X < \alpha$ for all ω for which the inequality $X(\omega) < \alpha$ holds, belongs to Σ .

We must be careful that a random variable is not a number but a function. Intuitively, we could think of a random variable as an object that represents simultaneously all possible states and only them. A particular value that a random variable may take in a random experiment, else known as a *realization*

of the variable is a number. Usually we denote a random variable by an upper case letter, e.g. X , and its realization by a lower case letter, e.g. x . The two should not be confused.

A random variable X can be *continuous* or *discrete*.

3.2 Discrete random variable

The *probability function* of a discrete random variable X is a function $f(x)$ that associates to each value x_i its given probability:

$$f(x_i) = P(X = x_i)$$

The *distribution function* of a discrete random variable X is a function $F(x)$ that associates each value x_i to the probability that the random variable be lower or equal to this value:

$$F(x_i) = P(X \leq x_i) = f(x_1) + f(x_2) + \dots + f(x_i)$$

The domain of $F(x)$ is not identical to the range of the random variable X ; rather it is always the set of real numbers. The distribution function is a non-decreasing function, for this is also called as *cumulative distribution function (CDF)*.

The range of the variable, along with its probability function, or distribution function, is called *probability distribution* of the variable.

3.3 Continuous random variables

Continuous random variables, unlike the discrete variables, can take any real value in an interval. Therefore, the number of values that they can have is not only infinite, but uncountable.

Thus, in the case of continuous variables, it makes no sense to measure probability of individual values, but we will associate probabilities to intervals.

The *probability density function (PDF)* of a continuous random variable is a function that fulfills the following properties:

$$f(x) \geq 0 \quad \forall x \in \mathbb{R}$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

The probability that random variable X will be in an interval $[a, b]$ can be calculated as:

$$\int_a^b f(x)dx$$

The *cumulative distribution function* is defined as in the discrete case, it measures the cumulated probability.

$$F(x_i) = P(X \leq x_i) = \int_{-\infty}^{x_i} f(x)dx$$

For continuous random variables, the inverse function F^{-1} of $F(x)$ exists. Consequently, the equation $u = F(x)$ has a unique solution for x , that is $x_u = F^{-1}(u)$. The value x_u , which corresponds to a specific value u of the distribution function, is called *u-quantile* of random variable X .

3.4 Population moments and their sample estimators

The *population moments* are numerical values that allow to know the shape of the probability function, as an alternative to the probability function but, nevertheless, we could have more than a distribution with the same moments (at least, some of them).

The general definition of a moment is $E[h(x)]$, where $E[\cdot]$ means expectation and $h(x)$ is any function of the random variable. Most usual moments are:

- Mean, or mathematical expectation:

$$\mu = E[X] = \sum_{i=1}^n x_i f(x_i) \quad (3.1)$$

- Variance:

$$\sigma^2 = E[X - E[X]]^2 = E[X^2] - (E[X])^2 = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2 \quad (3.2)$$

- coefficient of Skewness:

$$\gamma_1 = \frac{E[X - \mu]^3}{\sigma^3} \quad (3.3)$$

- coefficient of kurtosis:

$$\gamma_2 = \frac{E[X - \mu]^4}{\sigma^4} \quad (3.4)$$

- Standard deviation:

$$\sigma = +\sqrt{\sigma^2} \quad (3.5)$$

- coefficient of variation:

$$\gamma_1 = \frac{\sigma}{\mu} \quad (3.6)$$

The mean describes the location or central tendency of a random variable while the standard deviation describes its spread. The coefficient of variation is a dimensionless measure of the variability of X . The coefficient of skewness describes the relative asymmetry of a distribution, the coefficient of kurtosis describes the thickness of a distribution's tail.

Let (x_1, x_2, \dots, x_n) be a set of observations, the unbiased estimators of the first simple moments (mean, variance, and coefficient of skewness) are:

$$m_x = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3.7)$$

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (3.8)$$

$$s_x^3 = \frac{1}{(n-1)(n-2)s_x^3} \sum_{i=1}^n (x_i - \bar{x})^3 \quad (3.9)$$

3.5 Methods of estimation

An *estimator* is a function of the sample:

$$\hat{\theta} = h(X_1, X_2, \dots, X_n) \quad (3.10)$$

Therefore the estimator is a random variable with a distribution that depends on the distribution model of the sample. Usually the main target of an estimator is to try to get knowledge about some parameters in a model distribution.

An *estimate* is a realization of the estimator for a particular sample. Given an estimator, there are as many estimates as the number of sample we have.

The *sampling distribution* is the probability distribution of the random variable estimator.

Theoretically, we could have more than an estimator for a parameter, the choice is conditioned by certain properties of an estimator:

- UNBIASEDNESS

$$E(\hat{\theta}) = \theta$$

When an estimator is not unbiased the bias is defined as:

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta \quad (3.11)$$

- CONSISTENCY

An estimator is consistent if, as the sample size increases, it converges in probability to the true value of the parameter:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \varepsilon) = 1 \quad \forall \varepsilon > 0 \quad (3.12)$$

Sufficient conditions for the consistency:

$$\begin{aligned} \lim_{n \rightarrow \infty} bias(\hat{\theta}_n) &= 0 \\ \lim_{n \rightarrow \infty} Var(\hat{\theta}_n) &= 0 \end{aligned} \quad (3.13)$$

- EFFICIENCY

The efficient estimator is defined as the estimator which have the minimum mean square error:

$$\begin{aligned} MSE(\hat{\theta}_0) &= E\left[(\hat{\theta} - \theta)^2\right] \\ MSE(\hat{\theta}_0) &= [bias(\hat{\theta})]^2 + Var(\hat{\theta}) \end{aligned} \quad (3.14)$$

An biased estimator might be more efficient than an unbiased one, it depends on the variance.

- SUFFICIENCY

A statistic $\hat{\theta}$ is said to be a sufficient statistic for θ if and only if the conditional distribution of the sample X given $\hat{\theta} = y$ does not depend on θ . This means that when the value of the statistic is given, every other information is irrelevant for θ

There are three main methods to get estimators which fulfill good properties:

1. Maximum likelihood estimators (ML)
2. Moment method estimators: simple moments (SM) or probability weighted moments (PWM)
3. Estimator which minimizes a loss function

3.5.1 Maximum Likelihood Estimation

The *likelihood function* is the density function of the random sample given the dataset. This function only depends on the θ parameters of the distribution model; thus we can make a first important consideration: this estimation method requires a priori knowledge of the probability distribution of the population. The likelihood function is given by:

$$L(\theta) = \prod_{i=1}^n f(x_i; \theta) \quad (3.15)$$

The maximum likelihood estimator, $\theta^{\hat{ML}}$, is the value of θ that maximizes the likelihood function, equation (3.15). In practice it is common to maximize the log-likelihood function, as the logarithm is a monotonically increasing function.

The first order condition in the maximization problem consist in setting the partial derivative of equation (3.15) equal to zero; so we obtain a system where the number of equations is equal to the number of unknown parameters, (the length of the vector θ).

Proprieties of $\theta^{\hat{ML}}$:

1. Consistency.
2. Unbiased or asymptotically unbiased.
3. Efficiency or asymptotic efficiency.
4. Asymptotically normally distributed.
5. Invariant, example if $\theta^{\hat{ML}}$ is an estimator of θ , then $\sqrt{\theta^{\hat{ML}}}$ is an estimator of $\sqrt{\theta}$
6. Sufficiency

3.5.2 Moment Method Estimators

Is based on the estimate of the theoretical moments with their empirical counterpart. In general, if we want to estimate a vector of k parameters $\theta = (\theta_1, \theta_2, \dots, \theta_k)^T$, we can express them as a function of the k moments of the population:

$$\begin{aligned} \hat{\theta}_1 &= g_1(E(X), E(X^2), \dots, E(X^k)) \\ &\vdots \\ &\vdots \\ \hat{\theta}_k &= g_k(E(X), E(X^2), \dots, E(X^k)) \end{aligned} \quad (3.16)$$

These estimators are, at least, consistent but the rest of the properties depends on the distribution model.

3.5.3 Minimize a Loss Function

In this case we choose the estimators which minimize a previous defined loss function. For example the *ordinary least square method* minimize the least squares between observed data and fitted data.

$$\hat{\theta}^{LS} = \text{ArgMin}_{\theta} \sum_{i=1}^n (y_i - f(x_i, \theta))^2 \quad (3.17)$$

3.6 Probability Weighted Moments and L-moments

The Probability weighted moments (PWM) are the precursors of L-moments. Specifically the PWM of a continuous random variable X , with cumulative distribution function F , is defined as:

$$M_{p,r,s} = E[X^p F(X)^r (1 - F(X))^s] = \int_0^1 (x(F))^p F^r (1 - F)^s dF \quad (3.18)$$

where $x(F)$ represents the inverse of F , and p, r, s are real numbers. When $r = s = 0$ equation (3.18) reduces to a simple moment of order p .

Usually PWM are used by setting $p = 1$ and one of the other two exponents equal to zero, so we obtain:

$$\alpha_s = M_{1,0,s} = \int_0^1 x(f)(1 - F)^s dF, \quad \beta_r = M_{1,r,0} = \int_0^1 x(F)F^r dF \quad (3.19)$$

When adapting a parametric distribution to the sample data, it is common practice to estimate the parameters equaling the sample moments with those of the theoretical model. The PWM can be used in the same manner, with different theoretical and practical advantages.

The following linear combinations have been proposed for an unbiased estimate of the sample PWM:

$$\begin{aligned} a_0 &= \frac{1}{n} \sum_{j=1}^n x_j; & a_s &= \frac{1}{n} \sum_{j=1}^n \frac{(n-j)(n-j-1)\dots(n-j-s+1)}{(n-1)(n-2)\dots(n-s)} x_j \\ b_0 &= \frac{1}{n} \sum_{j=1}^n x_j; & b_s &= \frac{1}{n} \sum_{j=1}^n \frac{(j-1)(j-2)\dots(j-r)}{(n-1)(n-2)\dots(n-r)} x_j \end{aligned} \quad (3.20)$$

where a_s and b_r are sample estimates of α_s and β_r , x_j are the sample values sorted in ascending order, and n is the sample size.

Some linear combinations of the PWM can be interpreted as measures of position, scale and shape of a probability distribution. These are the L-moments defined by Hosking (1990); Hosking and Wallis (1997), and are given by:

$$\begin{aligned}\lambda_1 &= \alpha_0 & &= \beta_0 \\ \lambda_2 &= \alpha_0 - 2\alpha_1 & &= 2\beta_1 - \beta_0 \\ \lambda_3 &= \alpha_0 - 6\alpha_1 + 6\alpha_2 & &= 6\beta_2 - 6\beta_1 + \beta_0 \\ \lambda_4 &= \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 & &= 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0\end{aligned}\tag{3.21}$$

In practice, the L-moment values are calculated using the estimates of the parameters defined in equation (3.20)

$$\begin{aligned}\ell_1 &= a_0 & &= b_0 \\ \ell_2 &= a_0 - 2a_1 & &= 2b_1 - b_0 \\ \ell_3 &= a_0 - 6a_1 + 6a_2 & &= 6b_2 - 6b_1 + b_0 \\ \ell_4 &= a_0 - 12a_1 + 30a_2 - 20a_3 & &= 20b_3 - 30b_2 + 12b_1 - b_0\end{aligned}\tag{3.22}$$

and in general:

$$\lambda_{r+1} = (-1)^r \sum_{k=0}^r p_{r,k}^* a_k = \sum_{k=0}^r p_{r,k}^* b_k\tag{3.23}$$

Define as L-moment ratios the quantities:

$$\tau_r = \lambda_r / \lambda_2, \quad r = 3, 4, \dots\tag{3.24}$$

The L-moments ratio measure the shape of the function regardless of the measurement scale. Specifically τ_3 and τ_4 are estimates of the *skewness coefficient* and *Kurtosis coefficient*, labeled as L-skewness and L-kurtosis (or L-skew, L-kurt). Their values are between -1 and $+1$.

We indicate L-CV the analogous of the coefficient of variation (CV), defined as:

$$\tau = \frac{\lambda_2}{\lambda_1}\tag{3.25}$$

L-moments λ_1 , λ_2 , and L-moment ratios τ , τ_3 and τ_4 are the most commonly used quantities to characterize the properties and shape of distribution functions up to four parameters. Table 3.1 summarizes the symbols and the names used for the statistics calculated with simple moments and L-moments.

	theoretical moments	sample moments		theoretical L-moments	sample L-moments
Mean	μ	\bar{x}	L-location	λ_1	ℓ_1
Standard deviation	σ	S	L-scale	λ_2	ℓ_2
CV	CV	$\hat{C}V$	L-CV	τ	t
Skewness	γ	g	L-skewness	τ_3	t_3
Kurtosis	κ	k	L-kurtosis	τ_4	t_4

Table 3.1: Symbols and names used for the statistics calculated with simple moments and L-moments, theoretical and sample.

L-moments and L-moment ratios fundamental properties are:

- *Existence*: if the distribution mean exists, then all of the L-moments exist.
- *Uniqueness*: if the distribution exists, then it is uniquely defined by the L-moments; two different distributions cannot have the same L-moments.
- *Numerical values*
 - λ_1 can take any value.
 - $\lambda_2 \geq 0$
 - For a distribution with only positive values $0 \leq \tau < 1$
 - $\|\tau_r\| < 1$ for $r \geq 3$. More restrictive bounds can be found for individual τ_r quantities. For example, bounds for τ_4 given τ_3 are:

$$\frac{1}{4}(5\tau_3^2 - 1) \leq \tau_4 < 1 \quad (3.26)$$

for a distribution that takes only positive values, bounds for τ_3 given τ are:

$$2\tau - 1 \leq \tau_3 < 1 \quad (3.27)$$

- *Linear transformation*: let be X and Y two random variables with L-moments $\lambda_r^{(x)}$ and $\lambda_r^{(y)}$ respectively; and suppose that $Y = aX + b$, then:

$$\lambda_1^{(y)} = a\lambda_1^{(x)} + b \quad (3.28)$$

$$\lambda_2^{(y)} = |a|\lambda_2^{(x)} \quad (3.29)$$

$$\tau_r^{(y)} = (\text{sign } a)^r \tau_r^{(x)}, \quad r \geq 3 \quad (3.30)$$

- *Symmetry*: let be X a symmetric random variable with mean μ , such that $P[X \geq \mu + x] = P[X \leq \mu - x]$ for each x . Then all the odd L-moment ratios are zero, i.e. $\tau_r = 0$, for $r = 3, 5, \dots$

The sample L-moments are an *unbiased* estimate of the theoretical L-moments.

Instead the estimate of the L-moments ratios *are not unbiased*, but their bias is very small for large samples. For small samples the bias can be calculated through Monte Carlo simulations.

3.6.1 Similarities and differences between simple moments and L-moment

Referring to the statistics mentioned in Table 3.1, now we recall some similarities and differences between the estimators based on simple moments and L-moments (Hosking and Wallis, 1997)

- λ_1 is μ .
- $\sigma \geq \sqrt{3}\lambda_2$; equality hold only in case of uniform distribution.
- $\hat{CV} \geq \left(\frac{3n}{n+1}\right)^{1/2} t$
- For symmetric distributions both γ and τ_3 are equal to zero. In general doesn't exist a simple relationship between the two coefficients. γ is more sensible to the extreme values, so for distribution with a heavy tail γ can go to infinity, while τ_3 maintains a low value, and in any case below unity.
- τ_4 gives less weight to extreme values of the distribution with respect to k

Additional considerations:

- For some probability distributions characterized by heavy tails, for example the GEV distribution, theoretical simple moments higher than a certain order diverge. Rather, the theoretical L-moments always exist, given that the mean value is finite.
- Simple moments are unbounded, whereas L-moment ratios have a natural bound $|\tau_r| < 1$, which facilitates the interpretation, just think for example to the coefficient of asymmetry.

- Simple moments give greater weight to the tails of the distributions, since they depend on $[x(F)]^r$, which grows faster than F^r per $F \rightarrow 1$ (extreme values).
- The simple moments are most affected by outliers.
- Identify the distribution from which the sample was drawn is more immediate with the L-moments than with simple moments. A useful diagnostic is the **L-moment ratios diagram** proposed by Hosking (1990), which compares the possible pairs of L-skew and L-kurt for the most widely used distributions in hydrology.

3.7 Plotting position rules

Plotting position rules are introduced to describe the cumulative distribution function of an observed sample. If n is the sample length, the most known plotting position rule is obtained allocating a relative frequency equal to $1/n$ to each observation, that determines the following empirical cumulative distribution function:

$$F_n(x_i) = Pr \{X \leq x_i\} = \frac{i}{n} \quad (3.31)$$

where x_i is the i -th observation in the ordered sample (ascending order: $x_1 \leq x_2 \leq \dots \leq x_i \leq \dots \leq x_n$ and X is the random variable of interest. This rule of plotting position attributes a cumulative frequency equal to $1/n$ to the highest observed value. That means assuming that the probability to observe values higher than the maximum in the sample is equal to zero. This characteristic limits the application of this plotting position rule in the study about extreme events, in which higher values than those recorded in the past are frequently observed, once the analysis on the available observations in a particular historic moment is completed. The following group of distributions is proposed in order to overcome the limitations:

$$F_n(x_i) = \frac{i - \alpha}{n + 1 - 2\alpha} \quad (3.32)$$

where the coefficient α can vary between 0 and 1. The equation (3.32) includes two notable plotting position rules commonly used in the study of extremes: the Weibull's when $\alpha = 0$,

$$F_n(x_i) = \frac{i}{n + 1} \quad (3.33)$$

And Hazen's when $\alpha = 0.5$,

$$F_n(x_i) = \frac{i - 0.5}{n} \quad (3.34)$$

3.8 Return Period

Assume an event such that its probability of occurrence in a unit period of time (normally one year) is p . Assume also that occurrences of such an event in different periods are independent. Then as time passes, we have a sequence of equally likely Bernoulli experiments, so the time (measured in unit periods) to the first occurrence is a Geometric random variable $G_e(p)$ with mean value $1/p$. This motivates the following definition (Stedinger et al., 1993):

“Let A be an event, and t the random time between two consecutive occurrences of A events. The mean value, T , of the random variable t is called the *return period* of event A ”.

If T is measured in years: x_T is the threshold that is exceeded in one year with a probability of $1/T$. (One or more exceedances!)

If T is very large ($T \gg 1$ year) this is equivalent to saying that x_T is exceeded on average once in T years.

If $F(x)$ is the CDF of the yearly maxima of a random variable, the return period is related to the exceedance probability of the value x by:

$$T_A = \frac{1}{1 - F(x)} \quad (3.35)$$

The importance of return periods in engineering is due to the fact that many design criteria are defined in terms of return periods.

Chapter 4

Extreme Value theory

Although the fundamental probabilistic theory of extreme values has been well developed for a long time (e.g., Gumbel (1958)), the statistical modeling of extremes remains a subject of active research. The most current text available on the theory of extreme values is Coles (2001). A review of the use of the statistics of extremes in hydrology and the characteristics of hydrological extremes can be found in Katz et al. (2002). In the same article future developments in the methodology of the statistics of extremes are suggested too.

Defining the extreme values as the maxima within non-overlapping time blocks (BM), it can be proved that, if there exists a limiting distribution of the maxima, this distribution belongs to the General Extreme Value (GEV) family (Fisher and Tippett, 1928; Gnedenko, 1943). Conversely, when looking at the exceedances above a high threshold (POT), it can be proved that the Generalized Pareto (GP) is the expected distribution (Pickands, 1975), a comprehensive treatment of the model is given by Davison and Smith (1990).

In section 4.1 we review basic of extreme value theory related to the BM approach and give a brief overview of the GEV distribution, while theory relative to the POT approach and the GP distribution is reported in section 4.2. Section 4.2.3 describes in detail the Multiple Threshold Method proposed by Deidda (2010) to infer the parameters of the GP distribution underlying the exceedances of daily rainfall records over a wide range of thresholds. The model has been used in Zoglat et al. (2014) where an integrated approach to detect the optimal threshold and estimate the shape parameter of the GP underlying the exceedances is proposed, and in Serinaldi and Kilsby (2014) for smoothing the GP shape parameter fluctuations due to threshold selection.

4.1 The block maxima (BM) approach

Denote by X_i a sequence of independent random variables having a common distribution function F . In what follows we let's focus on the statistical behavior of:

$$M_n = \max \{X_1, \dots, X_n\}$$

that represents the maximum over n time units of observations. For example, if n is the number of observations in one year, then M_n corresponds to the annual maximum.

In applications, X_i usually represents a set of values of a process measured at a regular time scale, for example, daily rainfall

In theory, under independence of X_i , the distribution of M_n can be derived exactly for all possible values of n :

$$\begin{aligned} Pr \{N_n \leq z\} &= Pr \{X_1 \leq z, \dots, X_n \leq z\} \\ &= Pr \{X_1 \leq z\} \cdots Pr \{X_n \leq z\} \\ &= \{F(z)\}^n \end{aligned} \quad (4.1)$$

In practice, this is not possible because the function F is unknown. It is possible to use standard techniques to estimate F from observed values, and then replace in equation (4.1), but unfortunately a very small discrepancy in the estimate of F can lead to a large error in the estimation of F^n .

An alternative approach is to accept that F is unknown and to look for approximate families of models for F^n , which can be estimated solely on the basis of extreme data.

It is therefore necessary to study the behavior of F^n for $n \rightarrow \infty$. But this alone is not enough, because for each $z < z_+$, where z_+ is the smallest value of z such that $F(z) = 1$, $F^n(z) \rightarrow 0$ for $n \rightarrow \infty$. Therefore the distribution of M_n degenerates thus to a point mass concentrated in z_+ . In order to overcome this problem it is possible to rescale variable M_n :

$$M_n^* = \frac{M_n - b_n}{a_n} \quad (4.2)$$

where $\{a_n > 0\}$ and $\{b_n\}$ are sequences of constants. Appropriate choice of $\{a_n\}$ and $\{b_n\}$ stabilize the location and scale of M_n^* as n increases avoiding the difficulties that arise with the original variable M_n .

The key result, known as the extremal type theorem, gives the entire range of possible limiting distributions for M_n , where the limit is taken as $n \rightarrow \infty$. Different aspect of this results were proved by Fisher and Tippett (1928) and Gnedenko (1943)

Theorem 4.1.1 (Extremal Types Theorem, Fisher and Tippet (1928))

If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that:

$$Pr \left\{ \frac{M_n - b_n}{a_n} \leq z \right\} \rightarrow G(z) \quad \text{as } n \rightarrow \infty$$

where $G(z)$ is a non-degenerate distribution function, then $G(z)$ belongs to one of the following families:

$$I : G(z) = \exp \left\{ - \exp \left[- \left(\frac{z-b}{a} \right) \right] \right\} \quad -\infty < z < \infty \quad (4.3a)$$

$$II : G(z) = \begin{cases} 0 & z \leq b \\ \exp \left\{ - \left(\frac{z-b}{a} \right)^{-\alpha} \right\} & z > b \end{cases} \quad (4.3b)$$

$$III : G(z) = \begin{cases} \exp \left\{ - \left[- \left(\frac{z-b}{a} \right)^\alpha \right] \right\} & z < b \\ 1 & z \geq b \end{cases} \quad (4.3c)$$

Taken together, these three classes of distributions are known as **extreme value distributions**, that correspond to the distribution families: Gumbel, Fréchet and Weibull. Each family has a location and scale parameter, b and a respectively, the Fréchet and Weibull families have also a shape parameter α .

4.1.1 The generalized extreme value (GEV) distribution

However it is inconvenient to have to work with three possible limiting families. The Gumbel, Fréchet and Weibull families can be combined into a single family of models, having a common distribution function of the form (Jenkinson, 1955)

$$F(x; \mu, \sigma, \kappa) = \begin{cases} \exp \left\{ - \left[1 + \kappa \left(\frac{x - \mu}{\sigma} \right) \right]^{-1/\kappa} \right\} & \kappa \neq 0 \\ \exp \left\{ - \exp \left[- \left(\frac{x - \mu}{\sigma} \right) \right] \right\} & \kappa = 0 \end{cases} \quad (4.4)$$

defined on $\{1 + \kappa(x - \mu)/\sigma > 0\}$. $\kappa \in (-\infty, \infty)$ is the shape parameter, which determines the rate of tail decay, $\sigma > 0$ is the scale parameter, and $\mu \in (-\infty, \infty)$ is the location parameter.

Equation (4.4) represents the **generalized extreme value** (GEV) family of distributions. The three families of GEV distributions are listed in the following:

- $\kappa = 0$ **Type I - Gumbel distribution.** The distribution has only two parameters (μ, σ) , that control scale and location. The distribution is unbounded on the left and on the right side: $x \in (-\infty, \infty)$.
- $\kappa > 0$ **Type II - Fréchet distribution.** The distribution is bounded on the left and have a right tail: $x \in (\mu - \sigma/\kappa, \infty)$. In this case, conventional moments of order greater than or equal to $1/\kappa$ diverge (i.e. if $\kappa > 1/3$ the ordinary skewness is infinite)
- $\kappa < 0$ **Type III - Weibull distribution.** The distribution is bounded on the right and have a left tail: $x \in (-\infty, \mu - \sigma/\kappa)$.

Papalexiou and Koutsoyiannis (2013) analyzed the annual maxima of more than 1500 worldwide rainfall series with length varying from 40 to 163 years. The authors highlighted that the GEV shape parameter estimates depend on the record length and that essentially the parameter varies in the interval $(0, 0.23)$ and propose that in the case where data suggest a GEV distribution with negative shape parameter, this should not be used.

The probability distribution function of the GEV model is:

$$f(x; \mu, \sigma, \kappa) = \frac{1}{\sigma} \exp[-(1 + \kappa)y - \exp(-y)] \quad (4.5)$$

where

$$y = \begin{cases} \frac{1}{\kappa} \ln \left[1 + \kappa \left(\frac{x - \mu}{\sigma} \right) \right] & \kappa \neq 0 \\ \frac{x - \mu}{\sigma} & \kappa = 0 \end{cases} \quad (4.6)$$

From the inversion of equation (4.4) is possible to obtain the quantile of the GEV distribution:

$$x(F) = \begin{cases} \mu - \frac{\sigma}{\kappa} \{1 - (-\ln F)^{-\kappa}\} & \kappa \neq 0 \\ \mu - \sigma \ln(-\ln F) & \kappa = 0 \end{cases} \quad (4.7)$$

Generally $F = 1 - \frac{1}{T}$, so we can evaluate the events with a magnitude given by the time return period T , in years.

The choice of block size is a critical issue. Blocks too small are likely to lead to a poor approximation by the limit model in Theorem 4.1.1. This would lead to biases in the estimation of parameters and, consequently, in extrapolation. Large blocks would generate few block maxima, leading to a large estimation variance. It is therefore necessary to find a balance between the bias and the size of variances. Pragmatic considerations often lead to the adoption of blocks of length equal to one year, resulting in a series of annual maxima data. Furthermore with this choice seasonal inhomogeneity problems are avoided.

Several techniques have been proposed for the estimation of parameters of the GEV distribution, equation (4.4), including:

- maximum likelihood (ML);
- simple moments (SM);
- probability weighted moments (PWM);

In the following paragraphs we review the ML, SM and the PWM (L-moments) estimation methods for the GEV distribution.

Maximum Likelihood (ML) of GEV

The ML parameter estimates are those that maximize the logarithm of the likelihood function, namely $L(\mathbf{x}; \mu, \sigma, \kappa) = \sum_{i=1}^n \log f(x_i; \mu, \sigma, \kappa)$:

$$L(\mathbf{x}; \mu, \sigma, \kappa) = \begin{cases} -n \log \sigma - \frac{(1 + \kappa)}{\kappa} \sum_{i=1}^n \log \left[1 + \kappa \left(\frac{x_i - \mu}{\sigma} \right) \right] & \kappa \neq 0 \\ - \sum_{i=1}^n \left[1 + \kappa \left(\frac{x_i - \mu}{\sigma} \right) \right]^{-1/\kappa} & \kappa \neq 0 \\ -n \log \sigma - \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right) - \sum_{i=1}^n \exp \left[- \left(\frac{x_i - \mu}{\sigma} \right) \right] & \kappa = 0 \end{cases} \quad (4.8)$$

Parameter estimates $(\hat{\kappa}, \hat{\sigma}, \hat{\mu})$ can be obtained by maximization of the ML function $L(\mathbf{x}; \mu, \sigma, \kappa)$ in equation (4.8), with the obvious constraint $1 + \kappa \left(\frac{x_i - \mu}{\sigma} \right) > 0$.

In case one or more parameters are known, the same equation should be maximized keeping constant the known parameters.

A potential difficulty with the use of likelihood methods for the GEV distribution concerns the regularity conditions that are required for the usual

asymptotic properties associated with the maximum likelihood estimator to be valid (Coles, 2001). Such conditions are not satisfied by the GEV model because the end-points of the GEV distribution are functions of the parameter values: $\mu - \sigma/\xi$ is an upper end-point of the distribution when $\xi < 0$, and a lower point when $\xi > 0$. This violation of the usual regularity conditions means that the standard asymptotic likelihood results are not automatically applicable. Smith (1985) studied this problem in detail and obtained the following results:

- $\xi > -0.5$, maximum likelihood estimators are regular, in the sense of having the usual asymptotic properties;
- $-1 < \xi < -0.5$, maximum likelihood estimators are generally obtainable, but do not have the standard asymptotic properties;
- $\xi < -1$, maximum likelihood estimators are unlikely to be obtainable.

The case $\xi \leq -0.5$ corresponds to distributions with a very short bounded upper tail. This situation is rarely encountered in applications of extreme value modeling, so the theoretical limitations of the maximum likelihood approach are usually no obstacle in practice.

Simple Moments (SM) of GEV

Theoretical expectations of the first (μ_x), second (σ_x), third (γ_x) simple moments of the GEV distribution are provided in the following for $\kappa \neq 0$:

$$\mu_x = \mu - \frac{\sigma}{\kappa} [1 - \Gamma(1 - \kappa)] \quad (4.9a)$$

$$\sigma_x = \frac{\sigma}{|\kappa|} \{ \Gamma(1 - 2\kappa) - [\Gamma(1 - \kappa)]^2 \}^{1/2} \quad (4.9b)$$

$$\gamma_x = \text{sign}(\kappa) \frac{\Gamma(1 - 3\kappa) - 3\Gamma(1 - \kappa)\Gamma(1 - 2\kappa) + 2[\Gamma(1 - \kappa)]^3}{\{ \Gamma(1 - 2\kappa) - [\Gamma(1 - \kappa)]^2 \}^{3/2}} \quad (4.9c)$$

where $\text{sign}(\kappa)$ is plus or minus 1 depending on the sign of κ , and $\Gamma(\cdot)$ is the *gamma function*. The first three moments exist only for $\kappa < 1/3$.

Theoretical simple moments of the Gumbel distribution ($\kappa = 0$) are the following:

$$\mu_x = \mu + \gamma\sigma \approx \mu + 0.577215665 \sigma \quad (4.10a)$$

$$\sigma_x = \frac{\pi}{\sqrt{6}} \sigma \quad (4.10b)$$

$$\gamma_x = 1.139547 \quad (4.10c)$$

where $\gamma = 0.577215665\dots$ is the Euler constant.

The shape parameter κ should be estimated by equations (4.9c) and (4.10c), after substitution of the expected moment γ_x with the sample moment g_x , equation (3.7). Unfortunately, equation (4.10c) cannot be inverted in order to obtain an estimate of κ . Numerical methods able to find the zero of the following equation must be adopted and applied:

$$f(\kappa) = \begin{cases} g - \text{sign}(\kappa) \frac{\Gamma(1 - 3\kappa) - 3\Gamma(1 - \kappa)\Gamma(1 - 2\kappa) + 2[\Gamma(1 - \kappa)]^3}{\{\Gamma(1 - 2\kappa) - [\Gamma(1 - \kappa)]^2\}^{3/2}} & \kappa \neq 0 \\ g - 1.139547 & \kappa = 0 \end{cases} \quad (4.11)$$

where the constraint $\kappa < 1/3$ must be considered within the zero finding function.

Once an estimate of $\hat{\kappa}$ is obtained by previous equation, scale (σ) and position (μ) parameters can be estimated substituting the expected moments (μ_x, σ_x) with sample moments (m_x, s_x) in equations (4.9a,b) and (4.10a,b):

$$\hat{\sigma} = \begin{cases} \frac{s|\hat{\kappa}|}{\{\Gamma(1 - 2\hat{\kappa}) - [\Gamma(1 - \hat{\kappa})]^2\}^{1/2}} & \hat{\kappa} \neq 0 \\ \frac{\sqrt{6}}{\pi}s \approx 0.7796968s & \hat{\kappa} = 0 \end{cases} \quad (4.12)$$

$$\hat{\mu} = \begin{cases} m + \frac{\hat{\sigma}}{|\hat{\kappa}|} [1 - \Gamma(1 - \hat{\kappa})] & \hat{\kappa} \neq 0 \\ m - \gamma\hat{\sigma} \approx m - 0.577215665\hat{\sigma} & \hat{\kappa} = 0 \end{cases} \quad (4.13)$$

where $\gamma = 0.577215665\dots$ is the Euler constant.

In the case the shape parameter κ is known, estimates of scale σ and location μ parameters can be obtained using only equations (4.12) and (4.13).

In the case the shape κ and scale σ parameters are known, estimates of the location parameter μ can be obtained using only the last equation (4.13).

Probability Weighted Moments (PWM) of GEV

PWMs are more popular than ML in applications to extreme hydrological events, because they require less computational effort and, also, demonstrate better performance when applied to small samples.

L-moments of the GEV distribution are defined for $\kappa < 1$ (see e.g. Hosking and Wallis, 1997, page 196). For $\kappa \neq 0$ we have:

$$\lambda_1 = \mu - \frac{\sigma}{\kappa} [1 - \Gamma(1 - \kappa)] \quad (4.14a)$$

$$\lambda_2 = -\frac{\sigma}{\kappa} (1 - 2^\kappa) \Gamma(1 - \kappa) \quad (4.14b)$$

$$\tau_3 = \frac{2(1 - 3^\kappa)}{(1 - 2^\kappa)} - 3 \quad (4.14c)$$

where $\Gamma(\cdot)$ is the *gamma function*.

For the Gumbel distribution ($\kappa = 0$) theoretical L-moments and L-moment ratios are:

$$\lambda_1 = \mu + \gamma\sigma \quad (4.15a)$$

$$\lambda_2 = \sigma \ln 2 \quad (4.15b)$$

$$\tau_3 = \frac{\ln(9/8)}{\ln 2} \approx 0.1699 \quad (4.15c)$$

where $\gamma = 0.577215665\dots$ is the Euler constant.

Substituting expected L-moments ($\lambda_1, \lambda_2, \tau_3$) with sample L-moments (ℓ_1, ℓ_2, t_3) in equations (4.14a,b,c) and (4.15a,b,c) allows to derive $\hat{\kappa}, \hat{\sigma}, \hat{\mu}$ estimators.

To estimate κ , equation (4.14c) should be solved for κ . Unfortunately no explicit solution exists, thus Hosking et al. (1985) gave the following approximation that has accuracy better than 9×10^{-4} for $-0.5 \leq \kappa \leq 0.5$.

$$\hat{\kappa} = -7.8590c - 2.9554c^2 \quad \text{dove} \quad c = \frac{2}{3 + t_3} - \frac{\ln 2}{\ln 3} \quad (4.16)$$

Once an estimate of $\hat{\kappa}$ is obtained by previous equation, scale (σ) and position (μ) parameters can be estimated as:

$$\hat{\sigma} = \begin{cases} \frac{-\ell_2 \hat{\kappa}}{(1 - 2^{\hat{\kappa}}) \Gamma(1 - \hat{\kappa})} & \hat{\kappa} \neq 0 \\ \frac{\ell_2}{\ln 2} & \hat{\kappa} = 0 \end{cases} \quad (4.17)$$

$$\hat{\mu} = \begin{cases} \ell_1 + \frac{\hat{\sigma}}{\hat{\kappa}} [1 - \Gamma(1 - \hat{\kappa})] & \hat{\kappa} \neq 0 \\ \ell_1 - \gamma \hat{\sigma} \approx m - 0.577215665 \hat{\sigma} & \hat{\kappa} = 0 \end{cases} \quad (4.18)$$

where $\gamma = 0.577215665\dots$ is the Euler constant.

In the case when the shape parameter κ is known, estimates of scale σ and location μ parameters can be obtained using only equations (4.17) and (4.18).

In the case when the shape κ and scale σ parameters are known, estimates of the location parameter μ can be obtained using only the last equation (4.18).

4.2 The peaks over threshold (POT) approach

Modeling only maxima is a wasteful approach to extreme value analysis if other data on extremes are available. An alternative approach is to model all observations exceeding a specific high threshold.

Let X_1, X_2, \dots be a sequence of independent and identically distributed random variables, having marginal distribution F . We define extreme events those of the X_i that exceed some high threshold u . A stochastic behavior of extreme events is given by the conditional probability:

$$\Pr \{X > u + y | X > u\} = \frac{F(u + y)}{1 - F(u)}, \quad y > 0 \quad (4.19)$$

Equation (4.19) has no practical applications, because usually we do not know the parent distribution F , so the distribution must be approximated.

Theorem 4.2.1 *Let X_1, X_2, \dots be a sequence of independent random variables, with common distribution F , and let*

$$M_n = \max \{X_1, \dots, X_n\}$$

Denote an arbitrary term in the X_i sequence by X , and suppose that F satisfies theorem 4.1.1, so that for large n

$$\Pr \{M_n \leq z\} \approx G(z)$$

where $G(z)$ represent the GEV distribution. Then, for large enough u , the distribution function of $(X - u)$, conditional on $X > u$, is approximately

$$1 - \left(1 + \xi \frac{x - u}{\alpha}\right)^{-1/\xi} \quad (4.20)$$

defined on $x - u > 0$ and $(1 + \xi \frac{x - u}{\alpha}) > 0$

The case $\xi = 0$ is interpreted by taking the limit $\xi \rightarrow 0$ in equation (4.20). This lead to:

$$1 - \exp\left(-\frac{x-u}{\alpha}\right) \quad (4.21)$$

which correspond to an Exponential distribution with parameter $1/\alpha$.

The family of distributions defined by equations (4.20) and (4.21) is called the **generalized Pareto family**.

Theorem 4.2.1 is the key result for modelling threshold exceedances and it implies that, if G is the approximating distribution of block maxima, then there is a corresponding approximate distribution for threshold exceedances that belongs to the generalized Pareto family.

4.2.1 The general Pareto (GP) distribution

As previously said equations (4.20) and (4.21) can be combined into a single family of models having a common distribution function called the **generalized Pareto (GP) distribution**. This is the expected distribution of the exceedances over a high threshold, regardless the original data distribution. The CDF has the following form:

$$F(x; u, \alpha, \xi) = \begin{cases} 1 - \left(1 + \xi \frac{x-u}{\alpha}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left(-\frac{x-u}{\alpha}\right) & \xi = 0 \end{cases} \quad (4.22)$$

where ξ is the shape parameter, α the scale parameter, and the threshold u is the location parameter. Using the parameter of the GEV we have that

$$\alpha = \sigma + \xi(u - \mu) \quad (4.23)$$

Moreover, the parameters of the GP distribution of threshold excesses are uniquely determined by those of the associated GEV distribution of block maxima. In particular the parameter ξ is equal to that of the corresponding GEV distribution. Choosing a different, but still large, block size n would affect the values of the GEV parameters, but not those of the corresponding GP model of threshold excesses: ξ is invariant to block size, while the calculation of α in (4.23) is unperturbed by the changes in μ and σ which are self-compensating.

Varying the ξ value, the shape and the extreme properties of GP changes:

- $\xi > 0$ the distribution has a long right tail (so is often referred to as heavy tailed distribution). In this case, conventional moments of order greater or equal to $1/\xi$ diverge (i.e. if $\xi > 1/2$ the ordinary variance is infinite).
- $\xi = 0$ the distribution has the ordinary exponential form.
- $\xi < 0$ the distribution is short tailed with an upper bound value ($u - \alpha/\xi$)

the probability distribution function is:

$$f(x; u, \alpha, \xi) = \begin{cases} \frac{1}{\alpha} \left(1 + \xi \frac{x - u}{\alpha}\right)^{-\frac{1+\xi}{\xi}} & \xi \neq 0 \\ \frac{1}{\alpha} \exp\left(-\frac{x - u}{\alpha}\right) & \xi = 0 \end{cases} \quad (4.24)$$

Deidda and Puliga (2009) compared the performance of SM, ML and PWM techniques in estimate the GP parameters, using both synthetic continuous series and observed daily precipitation series characterized by a strong presence of rounded-off data. They found that ML technique provides the best result when applied to synthetic continuous series, followed by PWM and SM. Instead using the observed series SM technique provides the best fit, followed by ML and PMW. All three have however, a high bias when applied to real data, sometimes even on the same order of magnitude with the parameter estimates.

In the following paragraphs we review the ML, SM and the PWM (L-moments) estimation methods for the GP distribution. Sample data x_i are first transformed into exceedances over threshold u by

$$y_i = x_i - u$$

Maximum Likelihood estimation of GP

The ML parameter estimates are those that maximize the logarithm of the likelihood function, namely $L(\mathbf{y}; \alpha, \xi) = \sum_{i=1}^n \log f(y_i; \alpha, \xi)$:

$$L(\mathbf{y}; \alpha, \xi) = \begin{cases} -n \log \alpha - \frac{(1 + \xi)}{\xi} \sum_{i=1}^n \log \left(1 + \xi \frac{y_i}{\alpha}\right) & \xi \neq 0 \\ -n \log \alpha - n \frac{\bar{y}}{\alpha} & \xi = 0 \end{cases} \quad (4.25)$$

Let us note that the second equation (4.25) can also be obtained as limit function of the first equation (4.25) for $\xi \rightarrow 0$, thus the continuity of equations (4.25) in $\xi = 0$ is assured.

ML estimates $\hat{\alpha}$ and $\hat{\xi}$ should be obtained by maximizing equations (4.25), and thus by finding the zeros of $\delta L/\delta \xi = 0$ and $\delta L/\delta \alpha = 0$. Unfortunately, for the general case in which both α and ξ are unknown, a simple closed form of these zeros does not exist: thus numerical maximization is required. The only exception is for the exponential distribution (case $\xi = 0$), in such a case it is simple to find the zero of the derivative ($\delta L/\delta \alpha = 0$) of the second equation (4.25) that is:

$$\hat{\alpha} = \bar{y} \quad (4.26)$$

If both ξ and α are unknown, numerical maximization of equations (4.25) with respect to α and ξ (i.e. using Newton-Raphson optimization algorithms) brings to the optimal choice of GP parameters. Dealing with small samples, Hosking and Wallis (1987) observed that a local maximum of equations (4.25) may not exist, leading to the failure of the search algorithms. To overcome this kind of problems, Grimshaw (1993) introduced an optimized technique for ML estimation that is based on the simple transformation $\theta = -\xi/\alpha$. The log-likelihood function becomes:

$$L(\mathbf{y}; \theta) = \begin{cases} -n - \sum_{i=1}^n \log(1 - \theta y_i) - n \log \left[-\frac{1}{n\theta} \sum_{i=1}^n \log(1 - \theta y_i) \right] & \theta \neq 0 \\ -n - n \log(\bar{y}) & \theta = 0 \end{cases} \quad (4.27)$$

Let us note that the second equation (4.27) can also be obtained as limit function of the first equation (4.27) for $\theta \rightarrow 0$, thus the continuity of equations (4.27) in $\theta = 0$ is assured.

Let us suppose that a local maximum exists and let $\hat{\theta}$ be the value that maximizes the log-likelihood function (4.27). Parameter estimates are given by:

$$\begin{cases} \hat{\xi} = \frac{1}{n} \sum_{i=1}^n \log(1 - \hat{\theta} y_i) & ; & \hat{\alpha} = -\frac{\hat{\xi}}{\hat{\theta}} & & \text{if } \hat{\theta} \neq 0 \\ \hat{\xi} = 0 & & ; & \hat{\alpha} = \bar{y} & & \text{if } \hat{\theta} = 0 \end{cases} \quad (4.28)$$

If ξ is known and α is unknown, solution $\hat{\alpha}$ is given by the univariate maximization of equations (4.25), where ξ is fixed to the known value. In

case of an exponential distribution ($\xi = 0$) the estimator is given by equation (4.26). In case $\xi \neq 0$, numerical (univariate) maximization of the first equation (4.25) is required.

If ξ is unknown and α is known, solution $\hat{\xi}$ is given by the numerical univariate maximization of equations (4.25), where α is fixed to the known value. Continuity of equations (4.25) in $\xi = 0$ was already discussed.

Simple Moments (SM) of GPD

The shape and scale parameters of the GP distribution are estimated introducing the sample mean m and standard deviation s of the exceedances y_i above a chosen threshold into the following theoretical expectations of the first (μ) and second (σ) simple moments of the GP distribution ($\xi < 0.5$):

$$\begin{aligned}\mu &= \frac{\alpha}{1 - \xi} \\ \sigma &= \frac{\alpha^2}{(1 - \xi)^2(1 - 2\xi)}\end{aligned}\tag{4.29}$$

If both ξ and α are unknown, substituting the expected moments (μ, σ) with sample moments (m, s) in equations (4.29) allows to derive $\hat{\alpha}$ and $\hat{\xi}$ estimators:

$$\begin{aligned}\hat{\alpha} &= \frac{1}{2}m \left(1 + \frac{m^2}{s^2}\right) \\ \hat{\xi} &= \frac{1}{2} \left(1 - \frac{m^2}{s^2}\right)\end{aligned}\tag{4.30}$$

If ξ is known and α is unknown, substituting the first expected moment (μ) with the first sample moment (m) in the first equation (4.29), where ξ is known, allows to derive $\hat{\alpha}$ estimator:

$$\hat{\alpha} = m(1 - \xi)\tag{4.31}$$

If ξ is unknown and α is known, substituting the first expected moment (μ) with the first sample moment (m) in the first equation (4.29), where α is known, allows to derive $\hat{\xi}$ estimator:

$$\hat{\xi} = 1 - \frac{\alpha}{m}\tag{4.32}$$

Probability Weighted Moments (PWM) of GPD

Hosking and Wallis (1987) derived some relationships to estimate parameters of GP distribution by the sample PWM's. Specifically the first two expected L-moments (λ_1, λ_2) of the GP are the following:

$$\begin{aligned}\lambda_1 &= \frac{\alpha}{1 - \xi} \\ \lambda_2 &= \frac{\alpha}{(1 - \xi)(2 - \xi)}\end{aligned}\tag{4.33}$$

If both ξ and α are unknown, substituting expected L-moments (λ_1, λ_2) with sample L-moments (ℓ_1, ℓ_2) in equations (4.33) allows to derive $\hat{\alpha}$ and $\hat{\xi}$ estimators:

$$\begin{aligned}\hat{\alpha} &= \frac{\ell_1(\ell_1 - \ell_2)}{\ell_2} \equiv \frac{2a_0a_1}{a_0 - 2a_1} \\ \hat{\xi} &= 2 - \frac{\ell_1}{\ell_2} \equiv 2 - \frac{a_0}{a_0 - 2a_1}\end{aligned}\tag{4.34}$$

where, in the right-hand side, sample L-moments ℓ_1, ℓ_2 are expressed as linear combination of probability weighted moments a_0, a_1 provided by equation (3.21) for $\xi < 1$.

If ξ is known and α is unknown, substituting the first expected L-moment (λ_1) with the first sample L-moment (ℓ_1) in the first equation (4.33), where ξ is known, allows to derive $\hat{\alpha}$ estimator:

$$\hat{\alpha} = \ell_1(1 - \xi) \equiv a_0(1 - \xi) \equiv m(1 - \xi)\tag{4.35}$$

Let us note that PWM and SM estimators are the same in this case: compare equation (4.35) and equation (4.31).

If ξ is unknown and α is known, substituting the first expected L-moment (λ_1) with the first sample L-moment (ℓ_1) in the first equation (4.33), where α is known, allows to derive $\hat{\xi}$ estimator:

$$\hat{\xi} = 1 - \frac{\alpha}{\ell_1} \equiv 1 - \frac{\alpha}{a_0} \equiv 1 - \frac{\alpha}{m}\tag{4.36}$$

Let us note that PWM and SM estimators are the same in this case: compare equation (4.36) and equation (4.32).

4.2.2 Threshold selection

What can be assumed as an optimal threshold for rainfall observations is still an open question without a definite answer. The issue of threshold selection is analogous to the choice of block size in block maxima approach, implying a balance between bias and variance. In this case, too low thresholds are likely to violate the asymptotic bases of the model, leading to biases; while too high thresholds will produce few excesses, leading to high variance of the estimates.

The standard practice is to adopt a threshold as low as possible, subject to the limit model, providing a reasonable approximation.

Several methods are available in literature for this purpose, for example:

- Mean residual life plot.
- Parameter stability plot.
- Failure to reject method.
- Dispersion index plot.
- Rules of thumb.
- Square error method.
- Likelihood ratio test.

Most of these methodologies are summarized in Lang et al. (1999); Zoglat et al. (2014). Each of them has strong limitations when dealing with large databases or with regionalization problems. In fact the following drawbacks are present:

- Graphical methods \Rightarrow no use for large databases.
- Computational problem.
- Different methods can lead to different results \Rightarrow different thresholds.
- The presence of roughly rounded-off data, strongly influence the results of some methods, (e.g. the Failure to reject method, see Deidda and Puliga (2006)).
- Dependence on the threshold \Rightarrow not the best indicators of climatological spatial patterns.

So we decided to use the Multiple Threshold Method (MTM) proposed by Deidda (2010), which overcomes all the drawbacks previously listed. In particular the problem related to the presence of roughly rounded-off data. In fact the detection of an optimum threshold becomes even more difficult, if not impossible, using available methods on heavily quantized records (Deidda and Puliga, 2006, 2009; Deidda, 2007).

4.2.3 Multiple Threshold Method (MTM)

This method was developed by Deidda (2010) to infer the parameters of the GP distribution underlying the exceedances of daily rainfall records over a wide range of thresholds.

Given a set of rainy and non rainy values x at daily or any other fixed time scale is possible to describe the marginal distribution of the process by the following CDF:

$$F(x) = Pr\{X \leq x | X \geq 0\} = (1 - \zeta_0) + \zeta_0 F_0(x) \quad x \geq 0 \quad (4.37)$$

where $\zeta_0 = Pr\{X > 0 | X \geq 0\}$ represents the probability of occurrence of rainy days, while $F_0(x) = Pr\{X \leq x | X > 0\}$ is the CDF of only rainy values.

Commonly used distribution functions $F_0(x)$ of strictly positive rainfall records include the exponential, Gamma (Pearson III), log-Gamma (log-Pearson III), skewed normal (i.e. a normal distribution fitted to the Box-Cox transformed data), and lognormal.

Define now the $F_u(x)$ as the CDF of the records above a given threshold u :

$$F_u(x) = Pr\{X \leq x | X > u\}$$

In general the parameter estimates of $F_u(x)$ differs from those of $F_0(x)$, even if $F_0(x)$ and $F_u(x)$ belong to the same family. This because the distribution of very small values may not be clearly definite and may depart from the distribution of the bulk of higher records. The author derived relationships to parametrize equation (4.37) with threshold-invariant parameters by assuring a perfect overlapping with the distribution $F_u(x)$ for any $x > u$, regardless the value of the threshold u . The GP distribution was used as $F_u(x)$. So we called this model MTM-GP.

Before describing the MTM-GP is better to illustrate the relationships among $F(x)$, $F_0(x)$, and $F_u(x)$, as described in Deidda (2010), in order to obtain a perfect overlapping among these CDFs for any $x > u$ as sketched in Figure 4.1.

Using simple arguments of probability is possible to write:

$$F_u(x) = 1 - Pr\{X > x | X > u\} = 1 - \frac{Pr\{X > x | X \geq 0\}}{Pr\{X > u | X \geq 0\}} = 1 - \frac{1 - F(x)}{1 - F(u)}$$

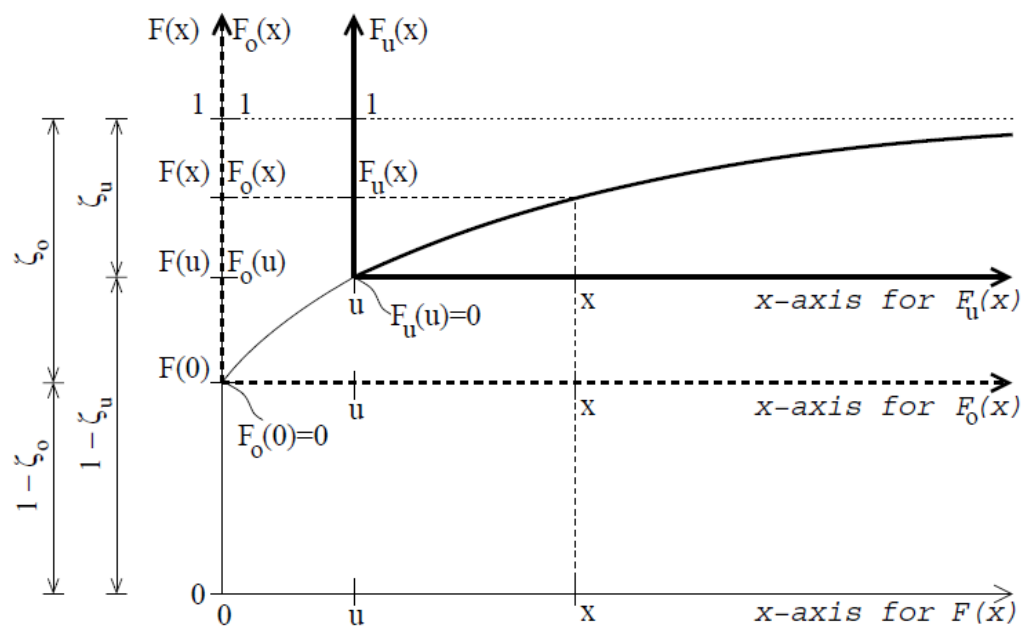


Figure 4.1: The sketch depicts some relations among the cumulative distribution functions (CDFs) $F(x) = Pr\{X \leq x|X \geq 0\}$, $F_0(x) = Pr\{X \leq x|X > 0\}$, and $F_u(x) = Pr\{X \leq x|X > u\}$. Cartesian axes of $F(x)$ are drawn with a thin line and characteristic values are reported on the left side, while the axes of $F_0(x)$ and $F_u(x)$ are drawn with dashed and solid thick lines, respectively, with values reported on the right side (from Deidda, 2010).

for $x > u$.

These equalities lead to the following relationship between $F(x)$ and $F_u(x)$ for any $x > u$:

$$F(x) = (1 - \zeta_u) + \zeta_u F_u(x) \quad x > u \quad (4.38)$$

where

$$\zeta_u = Pr\{X > u | X \geq 0\} = 1 - F(u)$$

represents the probability to observe excesses of u . We note that since $F_u(u) = \lim_{x \rightarrow u^+} F_u(x) = 0$, equation (4.38) becomes valid for any $x \geq u$ and thus includes also equation (4.37) as a special case for $u = 0$.

Using similar arguments we can write:

$$F_u(x) = 1 - \frac{Pr\{X > x | X > 0\}}{Pr\{X > u | X > 0\}} = 1 - \frac{1 - F_0(x)}{1 - F_0(u)}$$

in order to obtain a relationship between $F_0(x)$ and $F_u(x)$:

$$F_0(x) = F_0(u) + [1 - F_0(u)] F_u(x) \quad x \geq u \quad (4.39)$$

Finally, computing equations (4.37) and (4.38) for $x = u$, eliminating $F(u)$ among the equations, and putting $F_u(u) = 0$ we obtain:

$$\zeta_u = \zeta_0 [1 - F_0(u)] \quad (4.40)$$

The same equations can be derived by the following proportions in Figure 4.1

$$\frac{1 - F(x)}{1 - F(u)} = \frac{1 - F_0(x)}{1 - F_0(u)} = \frac{1 - F_u(x)}{1 - F_u(u)} \quad x > u \quad (4.41)$$

The probability ζ_u to observe an exceedance of the threshold u is estimated as:

$$\zeta_u = \frac{N_u}{N} \quad (4.42)$$

where N_u is the number of records above the threshold u and N is the sample size (including the zeros).

Now let us assume that also $F_0(x)$ is a GP distribution with threshold $u = 0$ and parameters α_0 and ξ , and that it can be expressed by equation (4.22) with $u = 0$.

Substituting $F_0(x)$ and $F_u(x)$ from equation (4.22) into equation (4.39) we can easily obtain:

$$\alpha_0 = \alpha_u - \xi_u u \quad \forall \xi_u \quad (4.43)$$

where the subscript u is used to label parameter estimates (including ξ) on the basis of the threshold used. Thus, if a suitable threshold has been selected, by virtue of equation (4.43) the α_0 reparametrization should be invariant for any higher threshold (even if α_u changes with u).

Computing now $F_0(u)$ from equation (4.22), i.e. putting first $u = 0$ and then computing for $x = u$, substituting $F_0(u)$ in equation (4.40), and (optionally) using equation (4.22) we obtain:

$$\zeta_0 = \begin{cases} \zeta_u \left(1 + \xi_u \frac{u}{\alpha_0}\right)^{1/\xi} & = \zeta_u \left(1 - \xi_u \frac{u}{\alpha_u}\right)^{-1/\xi} & \xi_u \neq 0 \\ \zeta_u \exp \frac{u}{\alpha_0} & = \zeta_u \exp \frac{u}{\alpha_u} & \xi_u = 0 \end{cases} \quad (4.44)$$

This last equation states that the ζ_0 reparameterization is threshold-invariant, although the probability ζ_u of exceeding u obviously decreases as u increases.

The **threshold-invariant GP parameterization** is obtained by substituting $F_0(x)$ from equation (4.22) into equation (4.37), and using α_0 and ζ_0 values obtained from equations (4.43) and (4.44):

$$F(x; \zeta_0, \alpha_0, \xi_0) = \begin{cases} 1 - \zeta_0 \left(1 + \xi \frac{x}{\alpha_0}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - \zeta_0 \exp\left(-\frac{x}{\alpha_0}\right) & \xi = 0 \end{cases} \quad (4.45)$$

Assuming x as an i.i.d. random variable, the distribution function of annual maxima $G(x)$ is related to $F(x)$ and the yearly return period T by the relation

$$G(x) = F(x)^n = 1 - \frac{1}{T} \quad (4.46)$$

where $n = 365.25$ is the average number of days in a year. From the inversion of equation (4.45) and using equation (4.46) we obtain the expression for the T -year return period quantile:

$$x_T = \begin{cases} \frac{\alpha_0}{\xi} \left\{ \left[\frac{1 - \left(1 - \frac{1}{T}\right)^{\frac{1}{n}}}{\zeta_0} \right]^{-\xi} - 1 \right\} & \xi \neq 0 \\ -\alpha_0 \ln \left[\frac{1 - \left(1 - \frac{1}{T}\right)^{\frac{1}{n}}}{\zeta_0} \right] & \xi = 0 \end{cases} \quad (4.47)$$

As remarked by Deidda (2010), equation (4.45) perfectly overlaps any GP distribution fitted on the exceedances over thresholds larger than the optimum one u^* : the only minor drawback is that there can be small departures from records smaller than u^* , but this does not affect extreme quantile estimation using equation (4.47). Concerning the choice of the optimum threshold u^* it should be selected large enough to reliably consider the distribution of the exceedances closely approximated by a GP distribution, but low enough to keep small the estimation variance.

The MTM improves the fitting on irregularly discretized records, as often happens in presence manually collected rainfall measurements. In Deidda (2010) the performances of the MTM model is superior compared to those of standard single threshold fitting on regularly discretized data. This is very important because Deidda (2007) highlighted that many time series collected by the Sardinian Hydrological Survey contain anomalous quantities of daily rainfall records rounded off at unexpected resolutions of 0.5, 1 and 5 mm/d. Furthermore, the three parameters in equation (4.45) do not depend on the threshold used for GP fitting, but only on the local climatic features: this property is particularly helpful to investigate the spatial pattern of rainfall signature in regional analyses.

MTM-GPD estimates

The MTM-GP estimates are obtained by the following hierarchical procedure:

1. ξ^M estimate. Identify suitable values of equally spaced threshold candidates $u^* < u_1 < \dots < u_n$. Take the MTM estimate ξ^M of the shape parameter as the median of the ξ estimates on the suggested range of thresholds.
2. α_0^M estimate. In order to filter out the variability of the α_0^M estimates driven by the fluctuations of ξ we estimate again the α_u values conditioned to ξ^M estimate obtained at step 1 and use again the reparameterization in equation (4.43) with the new α_u estimates and $\xi = \xi^M$ constant. Results from equation (4.43) are now denoted as α_0^C to remark that they are conditioned on ξ^M . The MTM estimate α_0^M of the scale parameter is the median of the new α_0^C estimates within the range of thresholds.
3. ζ_0^M estimate. In a similar way we can reduce the variability of ζ_0 by introducing the ζ_u estimates provided by equation (4.42) together with the MTM estimates ξ^M and α_0^M (obtained at step 1 and 2) into equation (4.44). Results from equation (4.44) are now denoted as ζ_0^C

to remark again that they are conditioned to ξ^M and α_0^M . The MTM estimate ζ_0^M is the median of the new ζ_0^C estimates within the range of thresholds.

Figure 4.2 shows an example of the MTM procedure on a daily rainfall time series of our database (station 008). The figure graphically shows the hierarchical procedure previously described.

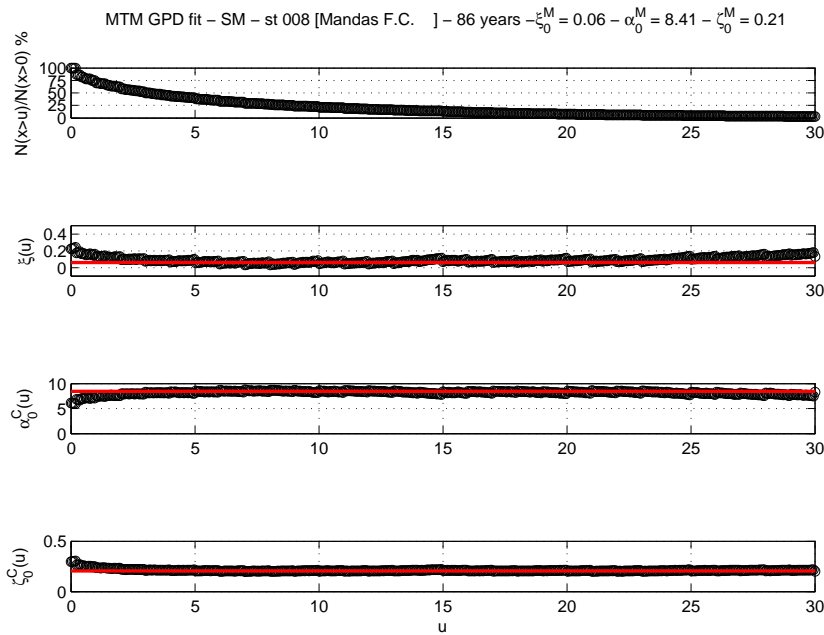


Figure 4.2: Station 008 : example of MTM application on a daily rainfall time series collected by a tipping-bucket rain gauge with a 0.2 mm resolution. The first plot from top displays the fraction of the values exceeding different thresholds u in a range from 0 to 30 mm. The second plot from top displays the $\xi(u)$ estimates with increasing threshold u : the ξ_0^M MTM estimate is the median value (horizontal red line) within the range of thresholds between 2.5 and 12.5 mm suggested for practical applications. In the third plot the α_0^M MTM estimate is obtained as the median value of the reparameterized α_0^C estimates conditioned on the ξ_0^M MTM estimate, while in the fourth plot the ζ_0^M MTM estimate is obtained by the ζ_0^C estimates conditioned on both ξ_0^M and α_0^M MTM estimates.

Chapter 5

Regional and geostatistical approaches

Initially, research on the statistical distribution of extreme rainfall events focused on obtaining the most accurate estimates at measurement sites, based on long series of observations (typically longer than 30 years). Now, one of the main challenges in this area is the spatial representation of rainfall extremes in order to obtain estimates at ungauged sites.

In order to achieve this two different approaches are generally used.

The first approach merges information from different gauged sites, according to a selected procedure, to compensate for short records at a single gauged site, and to obtain rainfall quantiles at locations where no measurements are available.

The second approach infers the parameters of the selected distribution model at each station, and then the return levels, or the distribution parameters, are spatially interpolated over the region. Different interpolation techniques can be used, like linear regression-based methods, inverse distance weighting, spline, kriging.

The regional approach is described in section 5.1 while the geostatistical approach is described in section 5.2.

5.1 Regional frequency analysis

Determination of the distribution of the annual maximum of daily precipitation from a single site is generally affected by large sample uncertainties. For this reason regionalization techniques have been proposed, with the purpose of using also the statistical information of the neighboring sites in order to obtain more robust estimates.

Regional frequency analysis consists in grouping the sites in homogeneous regions, choosing a frequency distribution, and then in estimating the rainfall quantiles at the sites of interest.

Several methods are commonly used for the regionalization of hydrological variables such as rainfall or floods. Multivariate techniques, such as a cluster analysis (CA), principal component analysis (PCA) and factorial analysis (FA), are very common methods for classification (Beaudoin and Rousselle, 1982; Karl et al., 1982; Mallants and Feyen, 1990; Van Regenmortel, 1995; Baeriswyl and Rebetez, 1997; Comrie and Glenn, 1998; Munoz-Diaz and Rodrigo, 2004; Pineda-Martinez et al., 2007).

Hosking and Wallis (1997) developed several tests for judging the degree of homogeneity of a group of sites and for choosing and estimating a regional distribution. This methodology is widely used for regional rainfall/flood frequency analysis, e.g. Alila (1999) applied L-moments for regionalization of 5 min to 24 hours annual rainfall extremes in Canada using the GEV distribution; Trefry et al. (2005) used this methodology to estimate intensity duration frequency (IDF) curves using two index-rainfall model, one for the annual maximum series and the other for the partial duration series, using a GEV and GP distribution respectively; Satyanarayana and Srinivas (2008) used large-scale atmospheric variables to the identification of homogeneous using a cluster analysis and the homogeneity tests described in Hosking and Wallis (1997) .

In this research we used a regional frequency analysis based on the **index-rainfall** method, described in sections 5.1.1 with a GEV growth curve described in section 5.1.2. For the identification of homogeneous regions we used the cluster analysis and the homogeneity tests proposed by Hosking and Wallis (1997), and reported in sections 5.1.3 and 5.1.4. The L-moment ratio diagram (Hosking, 1990) guided us in the identification of regional distributions. Section 5.1.5 briefly describes the Two-Component Extreme Value (TCEV) distribution. In the results section, we compare the outcomes from using the regional GEV model, with those from the TCEV model reported in Deidda and Piga (1998).

5.1.1 Index-rainfall method

The index-flood procedure (Dalrymple, 1960) is a simple regionalization technique with a long history in hydrology and flood frequency analysis. It uses data sets from several sites in an effort to construct more reliable flood-quantile estimators. By analogy in applications to precipitation data is called *index-rainfall* procedure.

Denoted by:

- N : number of sites;
- n_i : sample size at the i -th site
- x_i^j : j -th observation at the i -th site, $j = 1, \dots, n_i$
- $x_i(F)$: frequency distribution quantile at the i -th site.

The concept underlying the index-rainfall method is that the distribution of rainfall at different sites in a *homogeneous region* is the same for each site, except for a scale parameter which varies from site to site. This multiplicative factor is called *index-rainfall*, we denoted it by m_i . So the quantile at site i can be written as:

$$x_i(F) = m_i y(F), \quad i = 1, \dots, N \quad (5.1)$$

where $y(F)$ is the *regional growth curve*, a dimensionless quantile function common to every site of the homogeneous region. While the function $y(F)$ is invariant within each homogeneous region, the index-rainfall m_i varies locally and can easily be estimated with spatial mapping procedures. Generally the mean of the random variable used as index-rainfall, but any other position index, such as mode or the median can be used instead.

5.1.2 GEV growth curve

Let X be a random variable distributed according to a GEV distribution, equation (4.4), with parameters κ , σ and μ . Introduce the dimensionless variable $y = x/m$, where m is the sample mean of X , also known as index-rainfall. With simple algebra, from equation (4.4) it is possible to obtain the distribution function for the new variable y :

$$F(y; \mu^*, \sigma^*, \kappa) = \begin{cases} \exp \left\{ - \left[1 + \kappa \left(\frac{y - \mu^*}{\sigma^*} \right) \right]^{-1/\kappa} \right\} & \kappa \neq 0 \\ \exp \left\{ - \exp \left[- \left(\frac{y - \mu^*}{\sigma^*} \right) \right] \right\} & \kappa = 0 \end{cases} \quad (5.2)$$

where the shape parameter κ does not change, while the dimensionless scale parameter and position parameter are respectively equal to:

$$\sigma^* = \frac{\sigma}{m} \quad \text{and} \quad \mu^* = \frac{\mu}{m} \quad (5.3)$$

The growth curve is similar to equation (4.7), but with the new dimensionless parameters:

$$y(F) = \begin{cases} \mu^* - \frac{\sigma^*}{\kappa} \{1 - (-\ln F)^{-\kappa}\} & \kappa \neq 0 \\ \mu^* - \sigma^* \ln(-\ln F) & \kappa = 0 \end{cases} \quad (5.4)$$

Known the rain index-rainfall m in the specific location of interest, the quantile of the dimensional variable x is equal to:

$$x(F) = m y(F) = m \begin{cases} \mu^* - \frac{\sigma^*}{\kappa} \{1 - (-\ln F)^{-\kappa}\} / & \kappa \neq 0 \\ \mu^* - \sigma^* \ln(-\ln F) & \kappa = 0 \end{cases} \quad (5.5)$$

The estimators of the new dimensionless parameters σ^* and μ^* are easily obtained by the estimators already described for the GEV distribution in section 4.1.1. In the following we recall the modified equations.

Maximum Likelihood (ML) of GEV growth curve

The ML parameter estimates are those that maximize the logarithm of the likelihood function, namely $L(\mathbf{y}; \mu^*, \sigma^*, \kappa) = \sum_{i=1}^n \log f(y_i; \mu^*, \sigma^*, \kappa)$:

$$L(\mathbf{y}; \mu^*, \sigma^*, \kappa) = \begin{cases} -n \log \sigma^* - \frac{(1+\kappa)}{\kappa} \sum_{i=1}^n \log \left[1 + \kappa \left(\frac{y_i - \mu^*}{\sigma^*} \right) \right] & \kappa \neq 0 \\ - \sum_{i=1}^n \left[1 + \kappa \left(\frac{y_i - \mu^*}{\sigma^*} \right) \right]^{-1/\kappa} & \kappa \neq 0 \\ -n \log \sigma^* - \sum_{i=1}^n \left(\frac{y_i - \mu^*}{\sigma^*} \right) - \sum_{i=1}^n \exp \left[- \left(\frac{y_i - \mu^*}{\sigma^*} \right) \right] & \kappa = 0 \end{cases} \quad (5.6)$$

Parameter estimates $(\hat{\kappa}, \hat{\sigma}^*, \hat{\mu}^*)$ can be obtained by maximization of the Maximum Likelihood function $L(\mathbf{x}; \mu, \sigma^*, \kappa^*)$ in equation (5.6), with the constraint $1 + \kappa \left(\frac{x_i - \mu^*}{\sigma^*} \right) > 0$.

In case one or more parameters are known, the same equation should be maximized keeping constant the known parameters.

Simple Moments (SM) of GEV growth curve

The shape parameter κ can be estimated by determining the numerical value that annuls equation (4.11), as already described in section 4.1.1. Having the estimate of $\hat{\kappa}$, the dimensionless parameters of scale σ^* and position μ^* can be estimated by replacing the expected theoretical moments (μ_x, σ_x) with the sample mean and sample standard deviation of samples (m, s) in equations (4.9a,b) and (4.10a,b), having preliminarily divided all members for the theoretical average μ_x :

$$\hat{\sigma}^* = \begin{cases} \frac{s|\hat{\kappa}|}{m \{ \Gamma(1 - 2\hat{\kappa}) - [\Gamma(1 - \hat{\kappa})]^2 \}^{1/2}} & \hat{\kappa} \neq 0 \\ \sigma^* = \frac{\sqrt{6}}{\pi} \frac{s}{m} \approx 0.7796968 s/m & \hat{\kappa} = 0 \end{cases} \quad (5.7)$$

$$\hat{\mu}^* = \begin{cases} 1 + \frac{\hat{\sigma}^*}{|\hat{\kappa}|} [1 - \Gamma(1 - \hat{\kappa})] & \hat{\kappa} \neq 0 \\ 1 - \gamma \hat{\sigma}^* \approx 1 - 0.577215665 \hat{\sigma}^* & \hat{\kappa} = 0 \end{cases} \quad (5.8)$$

where $\gamma = 0.577215665\dots$ is the Euler constant.

In the case the shape parameter κ is known, estimates of dimensionless scale σ^* and location μ^* parameters can be obtained using only equations (5.7) and (5.8).

In the case the shape parameter κ and the dimensionless scale parameter σ^* are known, estimates of the dimensionless location parameter μ^* can be obtained using only the last equation (5.8).

Probability Weighted Moments (PWM) of GEV growth curve

The estimate of the parameter of shape κ is obtained in this case by numerical inversion of the equations (4.14c) and (4.15c), or from the approximate expression already provided in equation (4.16), with the limitations discussed in the section 4.1.1.

Once an estimate of $\hat{\kappa}$ is obtained, the dimensionless parameters of scale σ^* and position μ^* can be estimated, reminding that $\ell_2 = \ell_1 t$, as:

$$\hat{\sigma}^* = \begin{cases} \frac{-t\hat{\kappa}}{(1 - 2\hat{\kappa}) \Gamma(1 - \hat{\kappa})} & \hat{\kappa} \neq 0 \\ \frac{t}{\ln 2} & \hat{\kappa} = 0 \end{cases} \quad (5.9)$$

$$\hat{\mu}^* = \begin{cases} 1 + \frac{\hat{\sigma}^*}{\hat{\kappa}} [1 - \Gamma(1 - \hat{\kappa})] & \hat{\kappa} \neq 0 \\ 1 - \gamma \hat{\sigma}^* \approx 1 - 0.577215665 \hat{\sigma}^* & \hat{\kappa} = 0 \end{cases} \quad (5.10)$$

where $\gamma = 0.577215665\dots$ is the Euler constant.

In the case when the shape parameter κ is known, estimates of dimensionless scale σ^* and location μ^* parameters can be obtained using only equations (5.9) and (5.10).

In the case when the shape parameter κ and the dimensionless scale parameter σ^* are known, estimates of the location parameter μ^* can be obtained using only the last equation (5.10).

5.1.3 Identification of homogeneous regions

The process of regionalization based on the index-rainfall is applicable only to homogeneous regions or areas, where the distribution of the dimensionless variable $F(y) = F\left(\frac{x}{m}\right)$ is the same at all sites. Unfortunately, this condition is hardly maintained in vast territories where the terrain and exposure to perturbations may introduce inhomogeneities. In these cases it is necessary to divide the sites into disjoint groups, taking care to aggregate sites in contiguous regions for easy extension of the results to ungauged locations.

In hydrology geographically contiguous regions have been used for a long time. But this methodology has been criticized because of its arbitrariness. In fact, the geographical proximity does not guarantee hydrological similarity. In this study we searched for *statistical similarity*, using the L-moment ratios estimated at gauged locations.

In the results presented in chapter 7 the possible combinations of sites in homogeneous areas were determined using hierarchical *cluster analysis*, ensuring that aggregation always occurs between adjacent groups. At each site, which initially is a single group, it's associated an array of data with the characteristics of the site. The groups are then subsequently aggregated according to the similarity of their vectors. The Ward's method, that minimizes the total variance of the system, was used as aggregation criterion, allowing only aggregations between contiguous clusters, according to a Delaunay triangulation performed on station sites.

Clusters should group sites with similar characteristics. Many algorithms measure the similarity between sites using distance measurements calculated in the space of the sites characteristics. Euclidean distance measurements are sensitive to the scale with which we measure the individual characteristics. Therefore, it is common practice to rescale the various magnitudes, so that

they have the same range of variability. This basically gives each site feature the same weight, although this may not be appropriate since some features are of greater interest than others.

There is no rule to determine the correct number of clusters in a certain geographical area, and even less to determine their distribution. It is necessary to find the right balance between the size of the regions and the degree of homogeneity. Regions that contain few sites leads to a marginal improvement in the estimation of quantiles with respect to local analyses, while regions with a large number of stations are unlikely to be homogeneous and thus some sites should be affected by strong distortion (bias) in the estimation of quantiles.

Generally it is not appropriate to use directly the result of the cluster analysis as homogeneous regions configuration. Subjective adjustments are often necessary to improve the physical consistency of the regions and thus diminish the heterogeneity. Among the possible adjustments we include:

- move a site or several sites from one region to another;
- delete a site, or a few sites from the data set;
- subdivide large regions in two or more regions;
- break a region and relocating its sites to other regions;
- merge two or more regions and redefine small groups.

5.1.4 Heterogeneity measures

Within a homogeneous region it is assumed that the observed time series in each site, although having different averages, have the same theoretical dimensionless statistics obtained with simple moments (CV, γ , k , ...) or L-moment ratios (t , t_3 , t_4 , ...), already introduced in Table 3.1. However, even in the absence of heterogeneity, given the limited series lengths, the sample moments can differ from site to site due to sampling variability. So it is necessary to discover whether the dispersion of the L-moment ratios within a hypothetical homogeneous region is equivalent to the dispersion that would be expected from sampling variability. In case such equivalence is confirmed we accept the hypothesis of homogeneity of the concerned region.

The heterogeneity measures are thus designed to quantify if the changes observed between sites could be interpreted or not like sampling variability, or if they should instead be attributable to real heterogeneity among the considered sites. For this purpose numerous simulations are performed using

the same probability distribution $F(y)$ for each site belonging to the suppose homogeneous region. The samples obtained by simulation have the same lengths of the observed samples. For each of them we obtain the distribution of the characteristics of interest (e.g. L-moment ratios), or at least the mean and standard deviation of these characteristics. To compare the observed and simulated dispersion, an appropriate metric is (Hosking and Wallis, 1997):

$$\frac{(\text{observed dispersion}) - (\text{mean of simulations})}{(\text{standard deviations of simulations})}$$

If we use the L-moment ratios as characteristic of interest, a large value of this metric indicates that the observed L-moment ratios are more dispersed with respect to the case in which the homogeneity hypothesis is true.

A distribution to generate synthetic data should be chosen. Here we use the *Kappa* distribution, for its ability to adapt to a wide range of distributions. The *Kappa* distribution has four parameters, ξ (position), α (scale), k and h , and its CDF is equal to:

$$F(x) = \left[1 - h \left\{ 1 - \frac{k(x - \xi)}{\alpha} \right\}^{\frac{1}{k}} \right]^{\frac{1}{h}} \quad (5.11)$$

The *Kappa* distribution includes as special cases: the Generalized Logistic distribution (GL) when $h = -1$, the GEV when $h = 0$ and the GP when $h = 1$. For the *Kappa* distribution, as well as for various other probability distributions used in statistical hydrology, Hosking and Wallis (1997) have provided the theoretical expressions of the parameters as a function of the location statistic ℓ_1 (average) and the L-moment ratios t , t_3 and t_4 . The authors suggest to calculate the parameters of the probability distributions to fit into homogeneous regions using not only the local average ℓ_1 (site dependent), but also the regional L-moment ratios calculated as follows:

$$t^R = \frac{\sum_{i=1}^N n_i t^{(i)}}{\sum_{i=1}^N n_i} \quad (5.12)$$

$$t_3^R = \frac{\sum_{i=1}^N n_i t_3^{(i)}}{\sum_{i=1}^N n_i} \quad (5.13)$$

$$t_4^R = \frac{\sum_{i=1}^N n_i t_4^{(i)}}{\sum_{i=1}^N n_i} \quad (5.14)$$

where n_i represent the sample size in the site i-th, $t^{(i)}$, $t_3^{(i)}$ and $t_4^{(i)}$ are the L-CV, L-skewed L-kurt estimated using the observed data at the site i-th, and N is the number of sites belonging to the homogeneous region.

To measure the degree of heterogeneity in a homogeneous region Hosking and Wallis (1997) suggest the use of **dispersion measures** (V , V_2 , V_3). These metrics measure the weighted standard deviations of the L-moment ratios t , t_3 and t_4 respect to the regional values provided by equations (5.12), (5.13) and (5.14):

$$V = \left\{ \frac{\sum_{i=1}^N n_i (t^{(i)} - t^R)^2}{\sum_{i=1}^N n_i} \right\}^{1/2} \quad (5.15)$$

$$V_2 = \frac{\sum_{i=1}^N n_i \{(t^{(i)} - t^R)^2 + (t_3^{(i)} - t_3^R)^2\}^{1/2}}{\sum_{i=1}^N n_i} \quad (5.16)$$

$$V_3 = \frac{\sum_{i=1}^N n_i \{(t_3^{(i)} - t_3^R)^2 + (t_4^{(i)} - t_4^R)^2\}^{1/2}}{\sum_{i=1}^N n_i} \quad (5.17)$$

Next, chosen a frequency distribution, with a Monte Carlo procedure a large number of simulations, denoted by N_{sim} , is performed for each of the N stations. These simulations are performed using the regional parameters obtained from the observed data. For each implementation, the N synthetic series have the same sample lengths n_i of the observed ones. The regions are homogeneous and the simulated series have no correlation. For each simulated region the statistics V , V_2 , V_3 are calculated. From simulations are calculated means and standard deviations of the N_{sim} values of V , V_2 , V_3 , denoted respectively with μ_V , μ_{V_2} , μ_{V_3} and σ_V , σ_{V_2} , σ_{V_3} .

Now it is possible to define the **heterogeneity measures** as:

$$H = \frac{(V - \mu_V)}{\sigma_V} \quad (5.18)$$

$$H_2 = \frac{(V_2 - \mu_{V_2})}{\sigma_{V_2}} \quad (5.19)$$

$$H_3 = \frac{(V_3 - \mu_{V_3})}{\sigma_{V_3}} \quad (5.20)$$

According to Hosking and Wallis (1997) the results obtained with the statistic H can be so interpreted:

- If $H < 1$ the area is acceptably homogeneous.
- If $1 \leq H \leq 2$ the area is possibly heterogeneous
- If $H \geq 2$ the area is considered heterogeneous.

Although the authors do not provide guidance on acceptance thresholds of H_2 and H_3 , it is reasonable to use the same intervals for these metrics too. It is important to specify that the authors suggest using these thresholds as guidelines rather than as regions of acceptance of a statistical test in the strict sense. If one wants to evaluate the results of measures H , H_2 and H_3 under this perspective, it is reasonable to assume that they are distributed according to a normal standard (being V , V_2 and V_3 error measures distributed according to a Gaussian), so for example the acceptance region with significance level of 5% would be $H \leq 1.64$.

5.1.5 TCEV model

The most widely used and documented methodology of regionalization in Italy is the *Valutazione delle Piene* (VAPI) procedure, promoted by the National Group for Defense against Hydrogeological Disasters, and based on the Two-Component Extreme Value (TCEV) distribution. The TCEV is a probabilistic distribution introduced by Rossi et al. (1984) in order to represent time series of maximum annual peak flow characterized by high asymmetry.

The TCEV model has already been used in a previous study for the characterization of Intensity Duration Frequency Curves (IDFs) in Sardinia (Deidda and Piga, 1998; Deidda et al., 2000). The TCEV probabilistic model is based on the hypothesis that extreme values of the considered hydrological quantities come from two different populations of random variables, caused by different meteoric phenomena (Rossi et al., 1984). The first population includes the most frequent and low-intensity ordinary events and it represents the basis component of the process, whereas the second population is characterized by the high-intensity and rare events and it indicates the extraordinary component. The two different climatic mechanisms are merged in a unique Poisson process in which the annual maximum precipitation value's distribution $F(x)$ is expressed through the product of two Gumbel's distributions, according to the relation:

$$F(x; \Lambda_1, \Lambda_2, \Theta_1, \Theta_2) = \exp \left\{ -\Lambda_1 \exp \left(-\frac{x}{\Theta_1} \right) - \Lambda_2 \exp \left(-\frac{x}{\Theta_2} \right) \right\} \quad (5.21)$$

Where Λ_1 is the average number of annual occurrences of the basis component, Λ_2 is the average number of annual occurrences of the extraordinary component, Θ_1 is the average intensity of the basis component and Θ_2 is the average intensity of the extraordinary component. The methods of hierarchi-

cal and regional estimation of TCEV parameters are reported in Fiorentino and Gabriele (1985).

In this study, results obtained using the GEV distribution and results obtained using the TCEV distribution were compared.

5.2 Geostatistical analysis

In several fields of applied sciences it is necessary to make estimations in correspondence to points where no measurements are available, utilizing in an efficient way measurements at discrete locations. The representation of maps and contour lines starting from punctual measures is an example of this kind of problems. Geostatistics is a branch of statistics that provides the opportunity to make correctly the estimations starting from measures in discrete points. The **kriging** is a geostatistical technique developed by Matheron (1963) and even now it is probably the most commonly used technique for spatial interpolations. Section 5.2 reports the basic concepts on this methodology.

Beguiria and Vicente-Serrano (2006) studied partial duration series of rainfall cluster maxima, fitting a GP distribution. They used georegression techniques in order to obtain a probability model in which the distribution parameters vary spatially, yielding a robust regional extreme value model.

Furcolo et al. (1995) proposed a regional model based on the TCEV distribution along with a geostatistical analysis of its parameters. The model takes into account the presence of deterministic and aleatory components on a different spatial scale.

In this research project we used ordinary kriging (OK) and kriging for uncertain data (KUD) in order to mapping the GEV parameters. OK is the simplest and most widespread version of kriging, KUD is used when some level of uncertainty is attached to the data to be interpolated. Both techniques are briefly described in section 5.2.3.

Panthou et al. (2012) used both techniques to map the extreme precipitation events in West Africa, and found that KUD has better performance than OK. Although the authors refer to the KUD formulation proposed by De Marsily (1986), which was corrected by Mazzetti and Todini (2009).

5.2.1 Spatial interpolation with the kriging technique

Suppose $z(\mathbf{x})$ is a realization of the random variable $Z(\mathbf{x})$ in the point \mathbf{x} , where \mathbf{x} represents the vector of spatial coordinates of a general point: in the straight line $\mathbf{x} = x_1$ or $\mathbf{x} = t$, in the plane $\mathbf{x} = [x_1, x_2]$, in the space $\mathbf{x} = [x_1, x_2, x_3]$. In the stochastic fields, the random variable $Z(\mathbf{x})$ is dependent on the other random variables relative to the other points in the space. Using the kriging, the estimation $\hat{z}(\mathbf{x}_0)$ in the point \mathbf{x}_0 is obtained through the linear combination of measures $z(\mathbf{x}_i)$, which are done in the N points of observation \mathbf{x}_i :

$$\hat{z}_0 = \hat{z}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i z(\mathbf{x}_i) \quad \text{for the r.v.} \quad \hat{Z}(\mathbf{x}_0) = \sum_{i=1}^N \lambda_i Z(\mathbf{x}_i) \quad (5.22)$$

In section 5.2.3 we'll see how to write the system of linear equations that permits to obtain unbiased estimation of weights λ_i , minimizing the variance of the estimation error. In succession, some properties of the stochastic fields are reported.

A stochastic field is defined *homogeneous* if the joined probability distribution $f(z(\mathbf{x}), z(\mathbf{x} + \mathbf{h}_1), z(\mathbf{x} + \mathbf{h}_2), \dots, z(\mathbf{x} + \mathbf{h}_m))$ for any subset of points $\{\mathbf{x}, \mathbf{x} + \mathbf{h}_1, \mathbf{x} + \mathbf{h}_2, \dots, \mathbf{x} + \mathbf{h}_m\}$ does not depend on \mathbf{x} , but it depends only on vectors $\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_m$.

Moreover, the field is *isotropic* if the joined probability distribution depends only on the modules of vectors $h_k = \|\mathbf{h}_k\|$ (distances) and it does not depend on their direction and spin.

In empirical applications, it is difficult to verify the above mentioned homogeneity and isotropy conditions for stochastic fields, so frequently the homogeneity and isotropy conditions in the weak-sense are looked for (**second order statistics**). We cite the definitions of mean and of covariance function for this aim:

$$\mu(\mathbf{x}) = E[Z(\mathbf{x})] = \int_{-\infty}^{\infty} z(\mathbf{x}) f(z(\mathbf{x})) dz(\mathbf{x}) \quad (5.23)$$

$$C(\mathbf{x}_1, \mathbf{x}_2) = \text{Cov}(Z(\mathbf{x}_1), Z(\mathbf{x}_2)) = E[\{Z(\mathbf{x}_1) - \mu(\mathbf{x}_1)\} \{Z(\mathbf{x}_2) - \mu(\mathbf{x}_2)\}] \quad (5.24)$$

The field is called *second-order homogeneous* if the mean and the covariance do not depend on \mathbf{x} but only on the vector $\mathbf{h} = \mathbf{x}_1 - \mathbf{x}_2$:

$$\mu(\mathbf{x}) = \mu \quad (5.25)$$

$$C(\mathbf{x}_1, \mathbf{x}_2) = C(\mathbf{h}) \quad (5.26)$$

Moreover, if the covariance depends only on the distance $h = \|\mathbf{h}\|$, the field is called *second-order isotropic*:

$$C(\mathbf{x}_1, \mathbf{x}_2) = C(h) \quad (5.27)$$

Sometimes the hypothesis of second-order stationarity with a finite variance is not satisfied; the experimental variance rises as the area of study

grows, or it can be, in some cases, infinite. In these cases, a less restrictive hypothesis, called *intrinsic hypothesis*, is introduced. The intrinsic hypothesis consists of the assumption that, even if the variance of Z is not finite, the variance of the increases of first order of Z is finite, and these increases are characterized by stationarity of second order. For this purpose, we introduce the *increases function* Y (random variable) and the *(semi-)variogram* function γ :

$$Y(\mathbf{x}_1, \mathbf{x}_2) = Z(\mathbf{x}_1) - Z(\mathbf{x}_2)$$

$$\gamma(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2}E [\{Z(\mathbf{x}_1) - Z(\mathbf{x}_2)\}^2] \quad (5.28)$$

The conditions of second-order homogeneity and second-order isotropy that previously characterized the mean and the covariance are now attributed to the increases function Y and to the variogram γ . In particular, the homogeneity condition is verified if these functions do not depend on \mathbf{x} but only on the vector $\mathbf{h} = \mathbf{x}_1 - \mathbf{x}_2$:

$$Y(\mathbf{x}_1, \mathbf{x}_2) = Y(\mathbf{h})$$

$$\gamma(\mathbf{x}_1, \mathbf{x}_2) = \gamma(\mathbf{h})$$

The isotropy condition is obviously verified if the same functions depend only on the distance $h = \|\mathbf{h}\|$:

$$Y(\mathbf{x}_1, \mathbf{x}_2) = Y(h)$$

$$\gamma(\mathbf{x}_1, \mathbf{x}_2) = \gamma(h)$$

For a homogeneous and isotropic field of second order, the following relation between the variogram function and the covariance function is valid:

$$\gamma(h) = C(0) - C(h) \quad (5.29)$$

Hence, if the covariance is known, the variogram is its mirror image with respect to the horizontal axis shifted vertically by the quantity $C(0)$. When the variance of Z is finite, the variogram tends to an asymptotic value equal to its variance, called *sill*, for big distances, whereas the distance at which a portion next to 1 of this value is reached is called *range* or *correlation length*. If the object of study phenomenon is not characterized by a finite variance, then the variogram grows indefinitely.

5.2.2 Sample variogram

The estimation of the function $\gamma(h)$ is based on the in site measures of the random variable Z . Suppose to have measures $z(\mathbf{x}_i)$ in N points \mathbf{x}_i con $i = 1, \dots, N$. For each couple of points $\mathbf{x}_i, \mathbf{x}_j$ the distance $h_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|$ and the quantity γ_{ij} are calculated:

$$\gamma_{ij} = \frac{1}{2}[z(\mathbf{x}_i) - z(\mathbf{x}_j)]^2$$

The number of pairs of points that can be identified is $N(N - 1)/2$. If $N(N - 1)/2$ points (h_{ij}, γ_{ij}) are represented on a Cartesian plane, they will shape a cloud of points (that is a scatter plot called **raw variogram**). In order to make its trend regular, a procedure similar to that utilized for the histograms is applied: the x-axis h is divided into K consecutive intervals in which the k -th interval is $[h_k^l, h_k^u)$: the union of K intervals entirely covers the semi-axis $h \geq 0$ without overlapping. Thus, the experimental variogram is drawn by points (one in each class k , $k = 1, \dots, K$) whose x-coordinate and y-coordinate are equal to:

$$h_k = \frac{1}{N_k} \sum_{s=1}^{N_k} h_{ij}^{(s)} \quad (5.30)$$

$$\gamma_k = \gamma(h_k) = \frac{1}{N_k} \sum_{s=1}^{N_k} \gamma_{ij}^{(s)} \quad (5.31)$$

Where, for each class k , the N_k couples of points $\mathbf{x}_i, \mathbf{x}_j$ that satisfy the condition $h_k^l \leq h_{ij} < h_k^u$ are indexed with s . Of course we have that $\sum_{k=1}^K N_k = N(N - 1)/2$.

At a later stage, the experimental variogram is approximated by parametric analytic functions. In the next paragraphs, some models employed to describe theoretical stationary and intrinsic not stationary variograms are cited.

Theoretical variograms: stationary models (finite variance)

Label the variogram and the covariance with $\gamma(h)$ and $C(h)$ respectively. The relation: $\gamma(h) = C(0) - C(h)$ is valid.

Gaussian model

$$C(h) = \sigma^2 \exp\left(-\frac{h^2}{L^2}\right) \quad (5.32)$$

$$\gamma(h) = \sigma^2 \left[1 - \exp \left(-\frac{h^2}{L^2} \right) \right] \quad (5.33)$$

where σ^2 is the variance of the field, L is a scale parameter. The correlation length (or range) is equal to $\alpha \approx 7L/4$ (when correlation is equal to $0.05\sigma^2$).

Exponential model

$$C(h) = \sigma^2 \exp \left(-\frac{h}{L} \right) \quad (5.34)$$

$$\gamma(h) = \sigma^2 \left[1 - \exp \left(-\frac{h}{L} \right) \right] \quad (5.35)$$

where σ^2 is the variance of the field, L is a scale parameter. The range is equal to $\alpha \approx 3L$. This model is very exploited in hydrological applications.

Spherical model

$$C(h) = \begin{cases} \sigma^2 \left(1 - \frac{3h}{2\alpha} + \frac{1}{2} \frac{h^3}{\alpha^3} \right) & 0 \leq h \leq \alpha \\ 0 & h > \alpha \end{cases} \quad (5.36)$$

$$\gamma(h) = \begin{cases} \sigma^2 \left(\frac{3h}{2\alpha} - \frac{1}{2} \frac{h^3}{\alpha^3} \right) & 0 \leq h \leq \alpha \\ \sigma^2 & h > \alpha \end{cases} \quad (5.37)$$

where σ^2 is the variance of the field, α is the range.

"Hole-effect" model

$$C(h) = \sigma^2 \left(1 - \frac{h}{L} \right) \exp \left(-\frac{h}{L} \right) \quad (5.38)$$

$$\gamma(h) = \sigma^2 \left[1 - \left(1 - \frac{h}{L} \right) \exp \left(-\frac{h}{L} \right) \right] \quad (5.39)$$

where σ^2 is the variance of the field, L is a scale parameter. These functions are not monotonic and can be used to represent pseudo-periodic unidimensional processes.

"Nugget-effect" model

$$C(h) = \sigma^2 \delta(h) = \begin{cases} 0 & h > 0 \\ \sigma^2 & h = 0 \end{cases} \quad (5.40)$$

$$\gamma(h) = \sigma^2 [1 - \delta(h)] = \begin{cases} \sigma^2 & h > 0 \\ 0 & h = 0 \end{cases} \quad (5.41)$$

where $\delta(h)$ is Kronecker's delta (it is equal to 1 for $h = 0$ and to 0 for $h \neq 0$), σ^2 is the variance of the field.

The name of this model derives from its first applications in geomining field. It describes the field variability at smaller scales than samples scales: it can be interpreted as the limit of other models (for instance the exponential or the Gaussian models) when the integral scale L tends to zero.

Theoretical variograms: not stationary models

We define only the variogram $\gamma(h)$, because the fields have infinite variance.

Power model

$$\gamma(h) = \theta h^s \quad (5.42)$$

where s is the power of the model, θ is a dimensionless parameter that guarantee that $\gamma(h)$ has the dimension of a variance.

Linear model

$$\gamma(h) = \theta h \quad (5.43)$$

This is a particular case of the Power Model in which $s = 1$.

Logarithmic model

$$\gamma(h) = A \log(h + 1) \quad (5.44)$$

Where A is a dimensional parameter that establishes that $\gamma(h)$ has the dimension of a variance.

5.2.3 Ordinary kriging and kriging for uncertain data

The kriging is a Best Linear and Unbiased Estimator (BLUE) and the words of the acronym have the following meanings:

- *Best*: minimum variance of the estimation error (efficiency condition).

- *Linear*: linear estimator (the estimation is obtained as linear combination of available measures).
- *Unbiased*: unbiased estimator (accuracy condition).

The accuracy condition is represented by the constraint between the coefficients $\sum_{i=1}^N \lambda_i = 1$, that is obtained from the following equation:

$$E \left[\hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0) \right] = 0$$

The efficiency condition is imposed minimizing the estimation error variance:

$$\min \sigma_E^2 = \min E \left[\left\{ \hat{Z}(\mathbf{x}_0) - Z(\mathbf{x}_0) \right\}^2 \right]$$

Analytic derivations can be found in the numerous textbooks about the topic (De Marsily, 1986; Kitanidis, 1997). In the next paragraphs are reported the linear equations' systems that give the values of λ_i for the ordinary kriging and for the kriging for uncertain data.

Ordinary kriging (OK)

Ordinary kriging (OK) is the most commonly used version of kriging. The mean component of the process is assumed to be spatially constant and is unknown. In the OK accuracy and efficiency conditions lead to the following $N + 1$ linear equations' system with unknown variables $\lambda_1, \lambda_2, \dots, \lambda_N, \nu$:

$$\begin{cases} \sum_{j=1}^N \lambda_j \gamma(\|\mathbf{x}_k - \mathbf{x}_j\|) + \nu = \gamma(\|\mathbf{x}_k - \mathbf{x}_0\|) & k = 1, \dots, N \\ \sum_{i=1}^N \lambda_i = 1 \\ \lambda_j \geq 0 \end{cases} \quad (5.45)$$

that can be written in matrix form $\mathbf{A}\mathbf{y} = \mathbf{b}$ where:

$$\mathbf{A} = \begin{bmatrix} 0 & \gamma(\|\mathbf{x}_1 - \mathbf{x}_2\|) & \cdots & \gamma(\|\mathbf{x}_1 - \mathbf{x}_N\|) & 1 \\ \gamma(\|\mathbf{x}_2 - \mathbf{x}_1\|) & 0 & \cdots & \gamma(\|\mathbf{x}_2 - \mathbf{x}_N\|) & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ \gamma(\|\mathbf{x}_N - \mathbf{x}_1\|) & \gamma(\|\mathbf{x}_N - \mathbf{x}_2\|) & \cdots & 0 & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \nu \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} \gamma(\|\mathbf{x}_1 - \mathbf{x}_0\|) \\ \gamma(\|\mathbf{x}_2 - \mathbf{x}_0\|) \\ \vdots \\ \gamma(\|\mathbf{x}_N - \mathbf{x}_0\|) \\ 1 \end{bmatrix}$$

The results of this linear equations' system allow to determine the weights λ_i that can be utilized for the estimation in the point \mathbf{x}_0 through the equation (5.22). The Lagrange multiplier ν , used to introduce accuracy condition, allows to calculate estimation error variance in the same point:

$$\sigma_E^2 = \sum_{i=1}^N \lambda_i \gamma(\|\mathbf{x}_i - \mathbf{x}_0\|) + \nu \quad (5.46)$$

The kriging problem is formulated by adding the non negativity constraints to each of the weights on top of the classical constraint that their sum equals one. The solution is first found without inequality constraints. If all the weights are non-negative, the solution is accepted. Alternatively, the negative weights are set to zero and the corresponding gauges are removed from the computation and a new solution is found based on the reduced set of gauges.

Kriging for uncertain data (KUD)

It is possible to verify that the application of ordinary kriging in the points where measures are conducted gives an estimation that exactly corresponds to the measured values in that points. When available measures are affected by uncertainty, this characteristic represents a limit, for instance for measure or samples errors. In order to overcome this limitation, De Marsily (1986) proposes a variation called **kriging for uncertain data** (KUD). Mazzetti and Todini (2009) found that the method proposed by the previous author was incorrect or only valid for an homoschedastic field (all the errors at the different sites have the same variance). Mazzetti and Todini (2009) modified and tested the methodology proposed by De Marsily (1986). The new linear equation system became:

$$\begin{cases} \sum_{j=1}^N \lambda_j \gamma^*(\|\mathbf{x}_k - \mathbf{x}_j\|) + \nu = \gamma^*(\|\mathbf{x}_k - \mathbf{x}_0\|) & k = 1, \dots, N \\ \sum_{i=1}^N \lambda_i = 1 \\ \lambda_j \geq 0 \end{cases} \quad (5.47)$$

where N is the number of gauges and:

$$\begin{cases} \gamma_{k,j}^* = \gamma_{k,j} + \frac{\sigma_k^2 + \sigma_j^2}{2} & \forall k, j = 1, \dots, N \cup k \neq j \\ \gamma_{k,0}^* = \gamma_{k,0} + \frac{\sigma_k^2}{2} & \forall k = 1, \dots, N \end{cases} \quad (5.48)$$

where σ_i^2 represents the variance of measure error relative to the i -th observation point.

So adding one half the sum of variances of gauges errors to the extra diagonal terms of the kriging matrix, and, at the same time, adding one half of the errors variance to the variogram between the gauges and the point to be estimated, is possible to account for errors in gauges.

The kriging for uncertain data is particularly useful and appropriate when some level of uncertainty is attached to the data to be interpolated, as in the case of statistical parameters produced by fitting the GEV distribution. KUD has the advantage of making the interpolation less sensitive to local sampling effects: these sampling effects are incorporated in the parameter variance of estimation error and the interpolation is no longer required to be exact at the point of measurements. However, it remains unbiased and still minimizes the interpolation error variance.

Chapter 6

Error metrics

Error metrics reported in the next paragraphs were employed in order to compare the accuracy of results from different methods to reproduce observed extremes. In particular, metrics based on square statistics and metrics based on quantiles, calculated on the highest observed values, are considered. The first family of metrics measure the distance between the estimated frequency distribution and the empirical distribution of the sample. The second family of metrics only consider the observed higher-intensity events in the stations, for instance the first 5 maxima, in order to evaluate the goodness of fit in the right tail of the distributions, because of its importance in the characterization of extreme values. Some of these metrics need the definition of plotting position rules, that have been discussed in section 3.7. The metrics described in the next section are created using Hazen's plotting position.

Figure 6.1 clearly illustrates the difference between the two families of error metrics. The red lines denotes the distances between the theoretical frequency distribution and the empirical one. The green lines denotes the distances between the observed values and the theoretical quantiles.

6.1 Square statistics of Cramer-von Mises' family

The square statistics of Cramer-von Mises' family measure the discrepancy between the empirical cumulative distribution function, which is labeled $F_n(x)$ and the theoretical distribution to test: $F(x)$. $F_n(x)$ can be calculated with one of the plotting position rules mentioned in section 3.7. Parameters of $F(x)$ can be known or unknown. This family of square statistics functions is described as:

$$Q = n \int_{-\infty}^{\infty} [F_n(x) - F(x)]^2 \psi(x) dF(x) \quad (6.1)$$

where $\psi(x)$ represents a function of weights.

The Cramer-von Mises statistic W^2 is obtained from equation (6.1) considering $\psi(x) = 1$, whereas the Anderson-Darling statistic A^2 is obtained considering $\psi(x) = \{F(x)[1 - F(x)]\}^{-1}$. Hence, A^2 gives larger weight to both distribution's tails with respect to the central part, whereas utilizing the statistic W^2 each part of the distribution is equally weighted. Finally, utilizing Hazen's plotting position, see equation 3.34, for $F_n(x)$, the integral in equation (6.1) produces the following expressions for the statistics W^2 and A^2 :

$$W^2 = \frac{1}{12n} + \sum_{i=1}^n \left(F(x_i) - \frac{2i-1}{2n} \right)^2 \quad (6.2)$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^n (2i-1) [\log(F(x_i)) + \log(1 - F(x_{n+1-i}))] \quad (6.3)$$

where x_i represents the i -th observation of the ordered sample in ascending order, whereas F represents the cumulative distribution function to test.

A^2 and W^2 are generally used to construct goodness of fit tests, whereas in this study they were employed to evaluate which distribution best fits the observed data. The smaller the A^2 and W^2 values are, the better the fit is.

6.2 Quantiles errors

The A^2 and W^2 metrics give us an overall measure of the deviation between the sample frequency distribution and the estimated theoretical distribution, considering the whole sample (including maxima annual precipitation of not critical years). However, in hydrological applications interest is mostly in the good reproduction of quantiles belonging to the right tail of the distribution (higher values). For this purpose, we introduce the errors on the quantile estimates.

Consider a sample of n observations sorted in ascending order, we can introduce error metrics, for each observation x_i . These metrics are based on the distance between the observation x_i^o and the corresponding quantile x_i^d . The latter is obtained through an inversion of the considered theoretical cumulative distribution function for a specific probability value. This value has been calculated using Hazen's plotting position rule in equation (3.34):

$$E_i = x_i^d - x_i^o \quad (6.4)$$

$$AE_i = |x_i^d - x_i^o| \quad (6.5)$$

$$Er_i = \frac{x_i^d - x_i^o}{x_i^o} \quad (6.6)$$

$$AEr_i = \left| \frac{x_i^d - x_i^o}{x_i^o} \right| \quad (6.7)$$

In order to evaluate the errors on the highest M observations, we can calculate the averages:

$$ME(M) = \frac{1}{M} \sum_{k=1}^M E_{n-k+1} \quad (6.8)$$

$$MAE(M) = \frac{1}{M} \sum_{k=1}^M AE_{n-k+1} \quad (6.9)$$

$$MEr(M) = \frac{1}{M} \sum_{k=1}^M Er_{n-k+1} \quad (6.10)$$

$$MAEr(M) = \frac{1}{M} \sum_{k=1}^M AEr_{n-k+1} \quad (6.11)$$

The metrics ME and MEr indicate how much the quantiles obtained through the inversion of the theoretical probability distribution underestimate or overestimate the M highest observed values. In particular, the errors are relative (dimensionless) in the metric MEr , which means that they are divided by the observed value. If interest is in the magnitude of the error, without considering the fact that it could be an underestimation or overestimation, the only metrics to consider are MAE and $MAEr$. These metrics are obtained by introducing absolute values in the metrics ME and MEr . The latter is a relative error, so is a dimensionless measure.

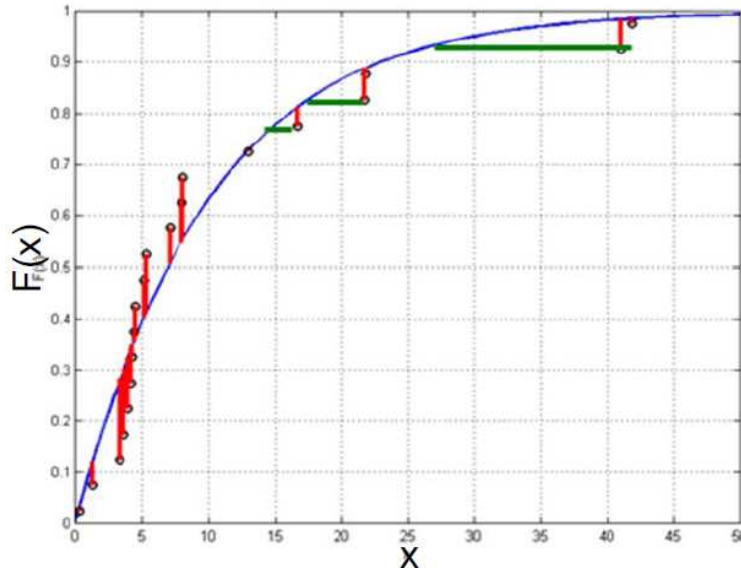


Figure 6.1: The proposed error metrics measure discrepancies between the theoretical frequency distribution and the empirical one (red lines), and between the observed values and the theoretical quantiles (green lines).

6.3 Cross-validation

The error metrics already mentioned were evaluated also in with a cross-validation procedure (also called leave-one-out method or jack-knife method) in order to pragmatically test and compare the performances of the different methods. The following procedure is going to be used.

Let N be the number of considered stations. One station is excluded and the estimation procedure, utilizing only the remaining $N - 1$ stations, is applied. The result is used to estimate the frequency distribution at the point where the excluded station is located. The error between this frequency distribution and the empirical frequency distribution of the observed sample in the excluded station is calculated. The same procedure is iteratively applied for each of the $N - 1$ remaining stations. At the end of the procedure the error metrics for each of the N station points are obtained. Both are used to calculate synthetic indexes of goodness of fit, and to analyse eventual spatial patterns of the error distribution.

Part III

Results and conclusions

Chapter 7

BM results

As mentioned in the previous chapters, in this study we used the GEV distribution to describe annual maxima of daily precipitation. The obtained results have been compared with those of a previous regional study about extreme precipitations in Sardinia, based on the TCEV distribution described in section 5.1.5. The results of such a model are reported in Deidda and Piga (1998) and Deidda et al. (2000), and in the next pages are labelled as “TCEV-1980”, in order to highlight that the database used in that study included observation till 1980. In order to have an up-to-date comparison, the TCEV estimates have been updated with the new database described in section 2.2, and this updated model is labeled as “TCEV-2008”. As all the procedures use the index-rainfall methodology, some comparisons were made directly between growth curves or between relative quantiles, calculated by employing the same index-rainfall for all the distributions. In this way any error in the estimation of the index-rainfall has a multiplicative effect on the calculated quantiles.

Descriptive statistics in Table 3.1 were calculated by using the annual maxima of daily precipitation observed in the stations point described in Chapter 2. After these computations, it is important to introduce the diagnostic diagram proposed by Hosking (1990) that reports the theoretical pairs of L-moment ratios (L-skewness, L-kurtosis) for some distributions widely used in statistical hydrology. In Figure 7.1 are reported the empirical L-moment ratios calculated considering the annual maxima daily precipitations observed in the 229 stations with more than 50 complete year of observations. Despite a high sample dispersion, the line that links the theoretical couples relative to the GEV distribution is the most barycentric and interpolating among the considered ones. In Figure 7.1, the point which corresponds to regional statistics (t_3^R, t_4^R) , calculated through equations 5.13 and 5.14, is reported too. In this case, the sample error is reduced and the point lies

on the line relative to the GEV distribution. This result suggest the use of the GEV distribution to describe our data, as expected considering the well-known theorems about extreme values asymptotic distributions (Coles, 2001; Castillo, 1988) reported in section 4.1.

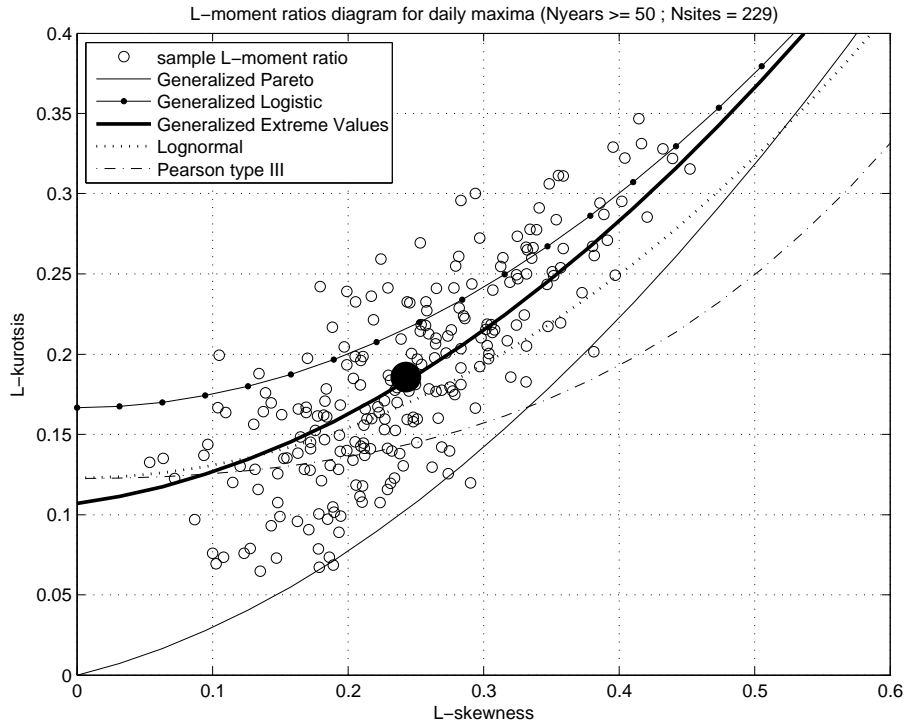


Figure 7.1: Comparison between pairs of L-moment ratios (L-skewness, L-kurtosis) calculated on annual maxima of daily precipitation for the 229 stations with more than 50 complete years in record (circles) and theoretical pairs for some distributions widely used in statistical hydrology, represented by lines of different strokes. Big marks denote the regional values of the same statistics.

7.1 Local analysis results

First off the GEV parameters (κ, σ, μ) were locally estimated, for each observed time series of annual maxima of daily precipitation. We found some negative values of the shape parameter, but very close to 0, so we used the constraint $\kappa \geq 0$, as suggested by Papalexiou and Koutsoyiannis (2013).

In Figure 7.2 we report the empirical cumulative distribution functions (CDFs) for the estimates of the GEV parameters, obtained through maximum likelihood (ML), simple moments (SM) and probability weighted moments (PWM) methods. The top panel of Figure 7.2 shows the CDF of the the shape parameter κ . ML and PWM techniques provided similar results, while the SM estimates used to be lower than those obtained with the other two methods. This discrepancy, and the fact that the SM estimates are always lower than 0.33 is due to the fact that conventional moments of order greater than or equal to $1/\kappa$ diverge (i.e. if $\kappa > 1/3$ the ordinary skewness is infinite). So SM method is not suitable to estimate the parameters of the GEV distribution, or any other distribution with 3 or more parameters. The central panel of Figure 7.2 shows the CDF of the the scale parameter σ , while the bottom panel is dedicated to the CDF of the the position parameter μ . We can observe that the three differer estimation methods give similar results.

For each of the 229 stations with at least 50 complete years of observations, empirical cumulative distribution functions (calculated with Hazen plotting position) of annual maxima of daily precipitation were compared with theoretical GEV distributions, whose parameters were locally estimated through ML, SM and PWM methods and with theoretical TCEV distributions (“TCEV-1980” and “TCEV-2008”), see for example Figure 7.3.

For each station, the error metrics $MEr(5)$, A^2 and W^2 , described in Chapter 6 have been calculated. As a reminder, the error metric $MEr(5)$ provides a percentage estimation of how much the theoretical distribution overestimates or underestimates, on average, the highest 5 observations. Instead the metrics A^2 and W^2 give an indication about the goodness of fit of the theoretical distribution to the whole set of observed data. Comparing these metrics for each station we noted that the GEV fit with PWM estimator is generally better. Moreover, we observed a better fit of the GEV distribution with respect to the TCEV distribution. But this could be due to the fact that we used local estimates for the GEV distribution, whereas for the TCEV distribution regional estimates were used.

A synthesis of the performances of local fits is presented in Table 7.1. In detail, using the local estimates of GEV parameters (κ, σ, μ) , the error metrics described in section 6 were calculated for each of the 229 stations.

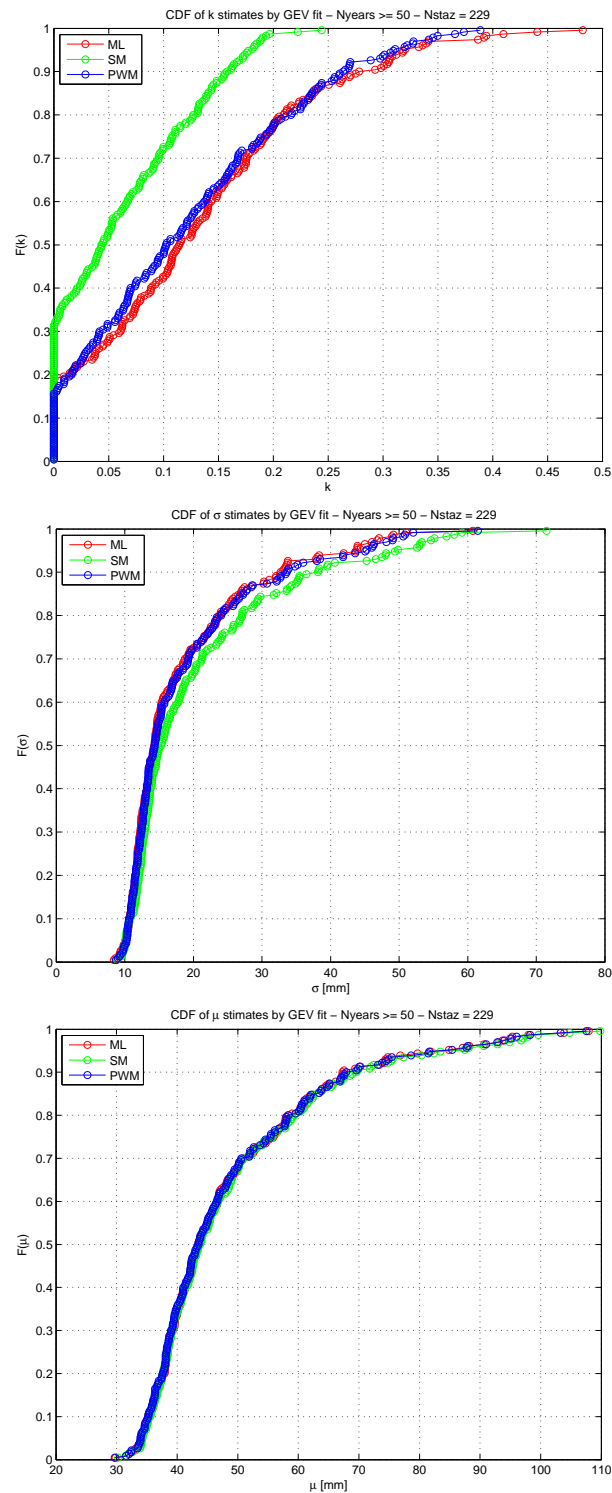


Figure 7.2: Cumulative distribution function of the GEV parameters estimates obtained with SM, ML and PWM techniques.

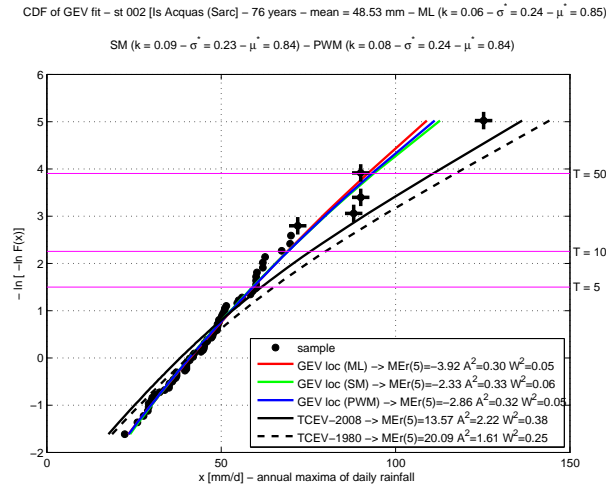


Figure 7.3: Station 002: Empirical cumulative distribution functions (calculated with Hazen's plotting position) of annual maxima of daily precipitation, compared with theoretical TCEV and GEV distributions, whose parameters are locally estimated through SM, ML and PWM techniques.

The averages of these metrics are reported in Table 7.1, where we can observe that the estimates obtained using the PWM are characterized by the best result in terms of metrics A^2 and W^2 . Furthermore they are characterized by lower average absolute error than the one obtained through ML estimates, and as the same order of magnitude as that obtained with SM estimates. This result, together with findings of past studies, lead us to adopt the PWM estimations.

	MAE(5) [mm]	MAEr(5) [-]	A^2 [-]	W^2 [-]
GEV local (ML)	12.101	0.078	0.324	0.051
GEV local (SM)	10.201	0.067	0.533	0.080
GEV local (PWM)	10.496	0.070	0.315	0.047
TCEV-1980	18.524	0.132	1.521	0.244
TCEV-2008	16.762	0.115	0.922	0.138

Table 7.1: Comparison between: local fits of GEV distribution (whose parameters are estimated by ML, SM, PWM methods) and regional fits of TCEV distribution. The averages of error metrics are calculated over the 229 stations with at least 50 complete years of observations.

7.2 Regional analysis results

7.2.1 Preliminary analysis

The regional analysis requires the identification of homogeneous regions. Each homogeneous region includes measuring sites whose observations are supposed to have the same dimensionless moments with rank higher than 1. In this case, it is possible to use the same regional growth curve $y(F)$ for each site belonging to the same homogeneous region. If the dimensionless quantile obtained by the growth curve is multiplied for the local index-rainfall, it is possible to obtain the local quantile, see equation 5.1.

A careful analysis of the sample statistics has shown that it is almost impossible to find exactly the same values in wide groups of stations. Every regionalization procedure consists in reiteration of two phases: in the first ones, an hypothesis of possible groups or homogeneous regions is proposed, whereas in the second ones statistical tests are applied in order to verify if the variation of the considered statistic within each region can be interpreted as sampling variability.

An analysis of the spatial distribution of the sample statistics in the regional territory could be helpful in defining configurations of the hypothetical homogeneous regions. We chose to classify the results considering the quartiles in order to point out possible differences in the spatial distributions. Higher-rank statistics are characterized by growing sample variability which conceals a possible territorial continuity of the examined characteristics. Figure 7.4 shows the distribution of the annual maxima of daily precipitation average. This metric has been modeled through spatial interpolation techniques in order to find the index-rainfall value in unobserved sites.

According to Hosking and Wallis (1997), homogeneity tests were implemented through Monte Carlo's techniques, generating 10'000 synthetic samples having the same length of the observed ones. We used a *Kappa* distribution (that includes the GEV distribution and others), whose regional parameters were held constant for all the sites within the same hypothesized homogeneous region. Comparing the dispersion of the calculated statistics over synthetic series with respect to that calculated in the observed ones it is possible to verify whether the latter is due to sampling variability linked to the finite number of observations or not. In the next paragraphs we test the hypothesis that the whole Sardinia island could be treated as a unique homogeneous region. Moreover, some hypotheses of possible groupings of observation sites in hypothetical smaller homogeneous zones have been examined, since statistical tests rejected the simplistic hypothesis of a unique homogeneous region.

Figure 7.5 shows the L-CV's and CV dispersion. It is evident that the highest values are located in the East part of the island which is characterized by the most intense rainfalls. Figure 7.6 and Figure 7.7 show the dispersion of the L-moment ratios t_3 (L-skewness), t_4 (L-kurtosis) and the third-rank and fourth-rank simple moments (skewness and kurtosis, respectively).

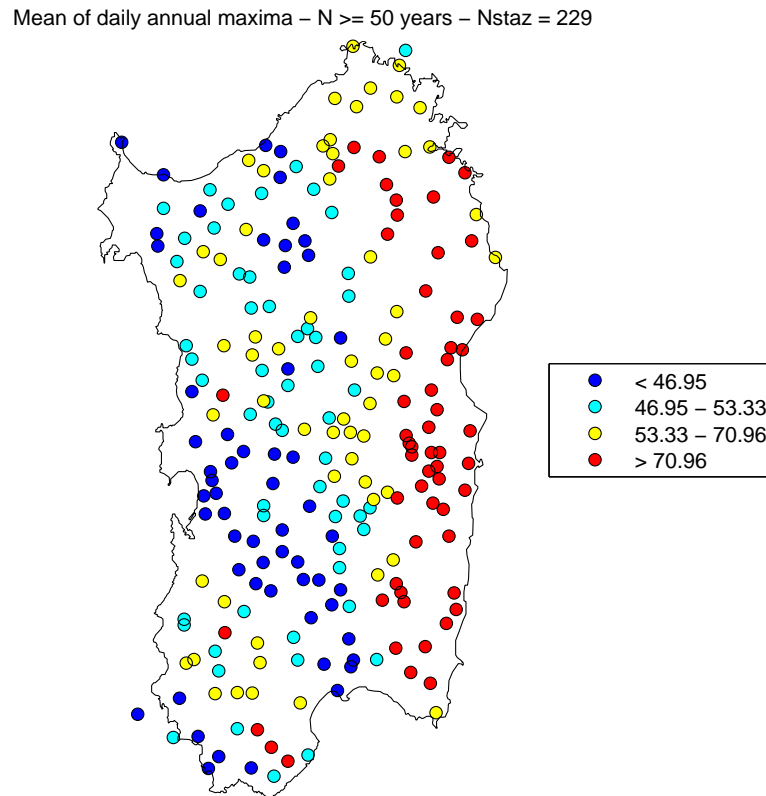
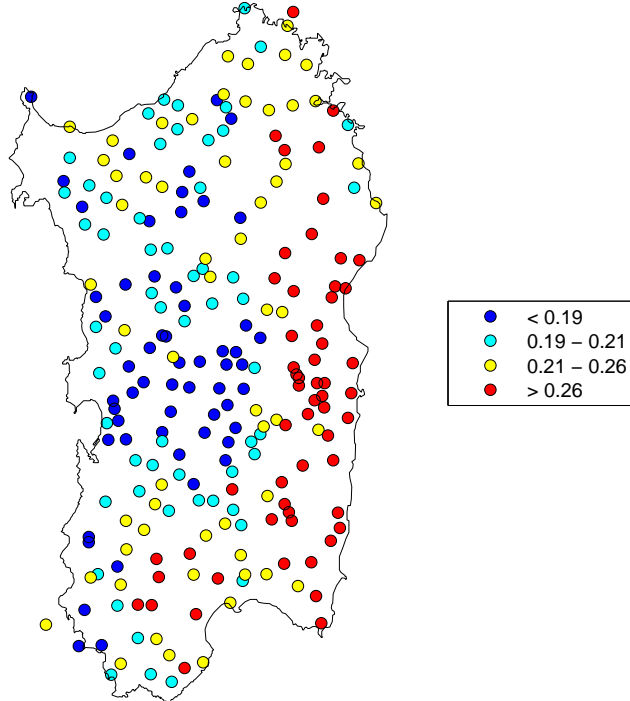


Figure 7.4: Classification by quartiles of L-moment ℓ_1 (average) estimates, measured in mm, for the 229 stations with at least 50 complete years of observations.

L-CV (t) of daily annual maxima – N \geq 50 years – Nstaz = 229



CV of daily annual maxima – N \geq 50 years – Nstaz = 229

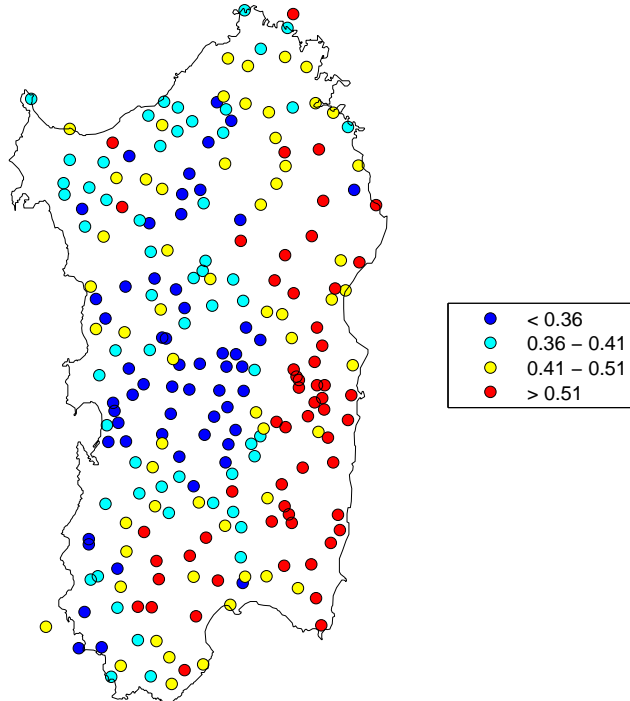
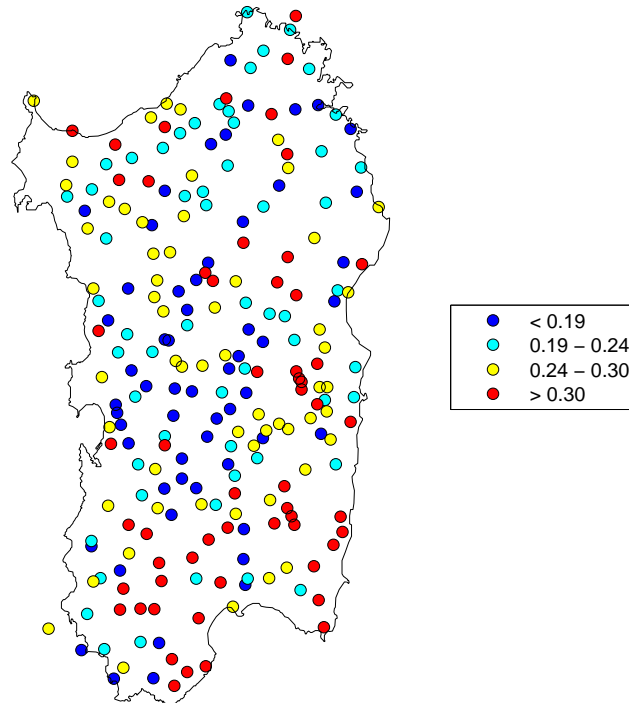


Figure 7.5: Classification by quartiles of the coefficients L-CV (top) and CV (bottom) estimates for the 229 stations with at least 50 complete years of observations.

L-skewness (t_3) of daily annual maxima – N \geq 50 years – Nstaz = 229



Skewness of daily annual maxima – N \geq 50 years – Nstaz = 229

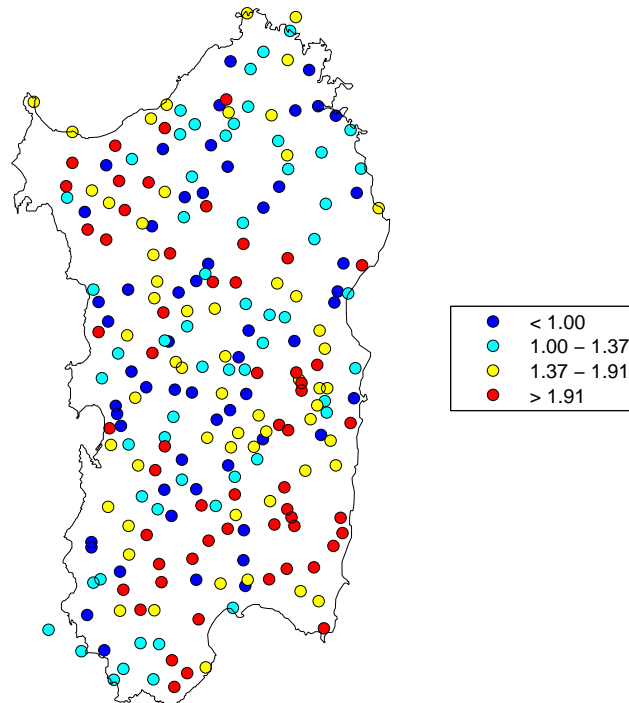
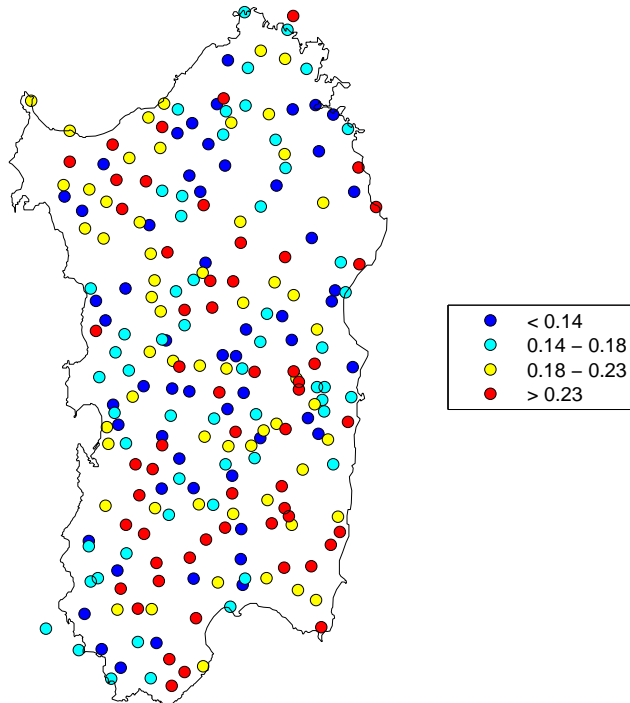


Figure 7.6: Classification by quartiles of the coefficients L-skewness (top) and skewness (bottom) estimates for the 229 stations with at least 50 complete years of observations.

L-kurtosis (t_4) of daily annual maxima – N \geq 50 years – Nstaz = 229



Kurtosis of daily annual maxima – N \geq 50 years – Nstaz = 229

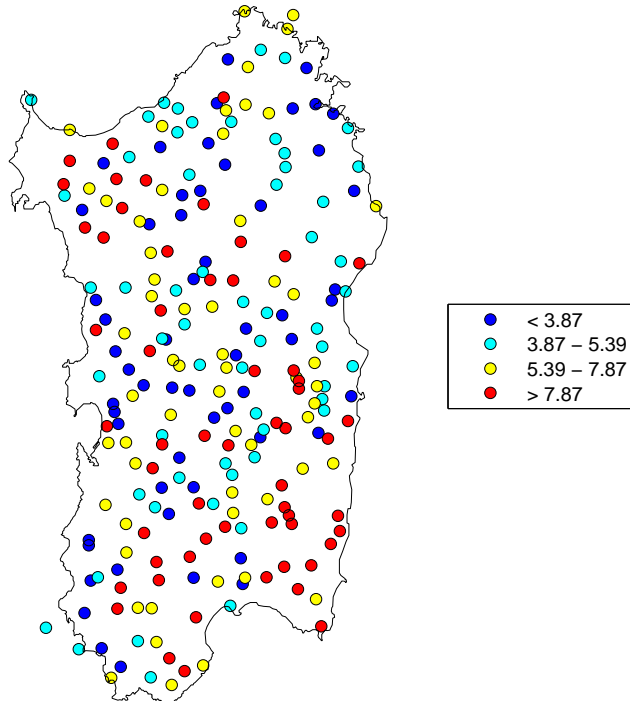


Figure 7.7: Classification by quartiles of the coefficients L-kurtosis (top) and kurtosis (bottom) estimates for the 229 stations with at least 50 years of observations.

7.2.2 Hypothesis of a unique homogeneous region

The first and simplest hypothesis that we examined was the assumption that the whole Sardinia territory could be considered as a unique homogeneous region. According to this hypothesis, regional estimates of L-moment ratios were obtained through equations (5.12), (5.13) and (5.14), using the data from the selected 229 rainfall stations with at least 50 complete years of observations. Since *Kappa* distribution's parameters ξ and α depend on the sample mean (l_1), each time series have been divided in advance for its average, coherently with the index-rainfall procedure, see section 5.1.1. This procedure does not invalidate the results of homogeneity analysis as they are based on the L-moment ratios that are independent from the average. Moreover, the approach based on the index-rainfall aims to determine the regional growth curve which is characterized by a mean equal to 1 in accordance with the equation (5.1). It is true when the expected value is used as index-rainfall, as done here. Table 7.2 shows the regional values of the L-moment ratios, and of the *Kappa* distribution parameters, obtained through the estimation procedure reported in Hosking and Wallis (1997, pages 202-204).

Kappa distribution, whose regional parameters are shown in Table 7.2, has been used to generate 10'000 synthetic series, for each of the 229 sites, through Monte Carlo's procedure. Each synthetic series has the same length of the observed one. We used the synthetic series to evaluate, for each statistic of interest (for instance L-CV, L-skewness, L-kurtosis), if the sample dispersion within the sites is due to the limited size of the samples, under the hypothesis that the considered sites belong to the same homogeneous region.

For example, the diagrams on the left of Figure 7.8 show the dispersions of L-moment ratios L-CV (t), L-skewness (t_3) and L-kurtosis (t_4). In these diagrams, each point represents a statistic couple calculated on one of the 229 observed time-series. The analogous diagrams on the right of Figure 7.8 show the couples of the same statistics extracted from one of the 10'000 synthetic time-series obtained through Monte Carlo's procedure in the same observation site. In each diagram, the regional values of statistics (just re-

$\ell_1^{(R)}$	$t^{(R)}$	$t_3^{(R)}$	$t_4^{(R)}$	$\xi^{(R)}$	$\alpha^{(R)}$	$k^{(R)}$	$h^{(R)}$
1	0.225	0.243	0.186	0.804	0.284	0.119	-0.004

Table 7.2: Hypothesis of unique homogeneous zone: regional values of L-moment ratios calculated on 229 stations, and corresponding estimates of *Kappa* distribution.

ported in Table 7.2) are represented through a cross symbol. Through the comparison of the diagrams on the left and those on the right, it is possible to observe that sample L-CV dispersion is markedly higher than L-CV dispersion obtained from the synthetic series. Similar observations can be proposed considering L-skewness values, but the differences in the dispersion values are less pronounced. The dispersion of L-kurtosis is similar for observed and synthetic samples.

Results presented in Figure 7.8 clearly show that observed dispersions (in particular for the L-CVs, but also for the L-skewnesses) cannot be interpreted as statistical fluctuations due to sample variability. Hence, we decided to apply homogeneity tests, since these discrepancies were evident. First of all, values of dispersion measures V , V_2 and V_3 were calculated through equations (5.15), (5.16) and (5.17), considering the hypothesis of unique homogeneous region for the whole Sardinian territory. Results are reported in the first row of Table 7.3.

In the same way we calculated, for each of the 229 stations, 10'000 values of the same dispersion measures. Every synthetic series had the same length of the observed one. Using these values we estimated the means (μ_V , μ_{V_2} and μ_{V_3}) and standard deviations (σ_V , σ_{V_2} and σ_{V_3}) of the dispersion measures, and the heterogeneity measures H , H_2 e H_3 , through equations (5.18), (5.19) and (5.20). The obtained results are reported in Table 7.3. It is possible to observe that heterogeneity measures assume values that, according to Hosking and Wallis (1997), characterized heterogeneous region. This result confirms what deduced from Figure 7.8.

In order to conduct a visual comparison that could support results reported in Table 7.3, in the graphs on the left of Figure 7.9, we show CDFs of 10'000 dispersion measures V , V_2 and V_3 , obtained by synthetic series and we compared them with sample values of the same dispersion measures, represented with a large vertical line. The visual analysis clearly shows that the

dispersion measures	V	V₂	V₃
regional value	0.040	0.081	0.086
μ_{V_*}	0.0175	0.0579	0.0735
σ_{V_*}	0.0008	0.0026	0.0032
heterogeneity measures	H	H₂	H₃
	25.25	8.80	3.92

Table 7.3: Dispersion measures and heterogeneity measures, evaluated on the 229 stations, under the hypothesis of a unique homogeneous region for the whole Sardinia island.

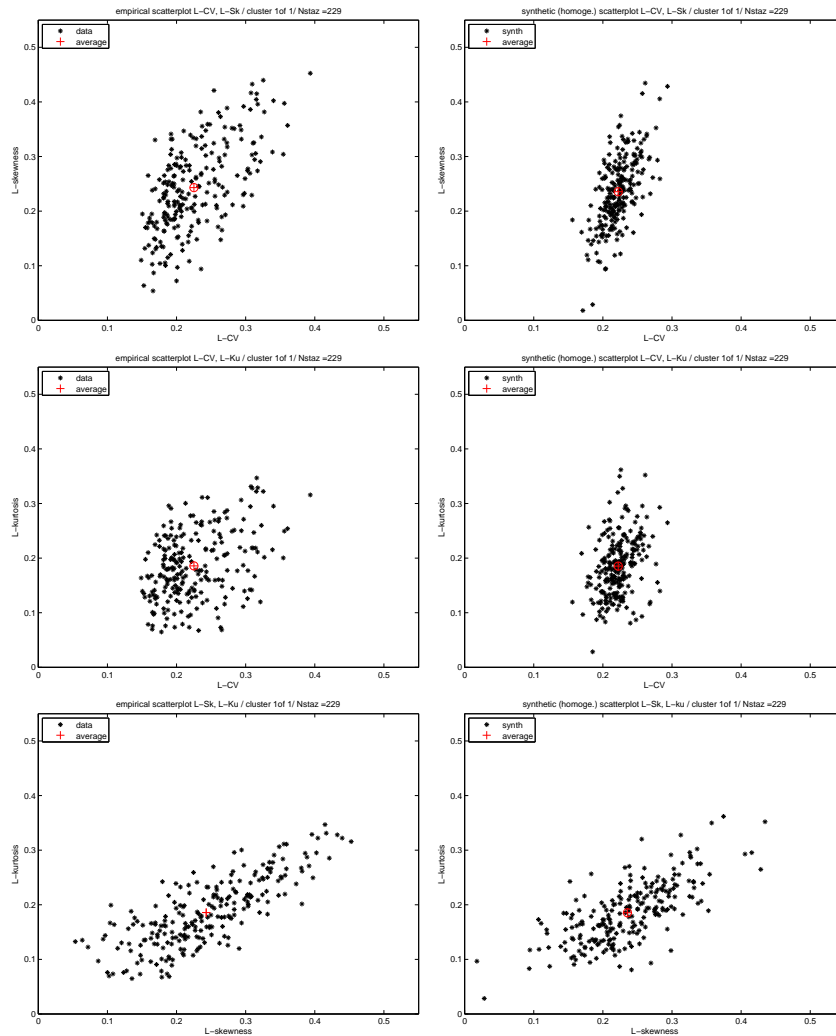


Figure 7.8: **Hypothesis of a single homogeneous zone for the whole Sardinia.**

On the left are shown scatterplots with pairs of statistics L-CV, L-skew and L-kurt calculated on the observations of each of the 229 considered sites (the corresponding regional averages are marked in red).

On the right are shown, for comparison, the same L-moment ratios calculated on one of the 10,000 synthetic series generated for each of the 229 sites using a Monte Carlo procedure.

sample values are higher than the values for the synthetic series.

However, since V_2 and V_3 statistics give the same weight to L-moment ratios' dispersions of different ranks, we consider appropriate to investigate the dispersion of each L-moment ratios singularly. In regional procedures, it is desirable to held constant the majority of parameters of the considered probabilistic distribution. Moreover, since in general parameters hierarchically depend on moments (or L-moments) of different rank, it is useful to investigate if the dispersions of L-moment ratios can be interpreted simply as sample variability. For instance, if fluctuations of moments of rank equal or higher than three (L-skewness and L-kurtosis) are not due to heterogeneity between sites, it is possible to assume a unique value for k and h parameters of the *Kappa* distribution in the considered region.

Considering these objective, in the graphs at the core of Figure 7.9, we compared the CDFs of standard deviation of L-CV, L-skewness and L-kurtosis statistics (from top to bottom). The values were obtained from 10'000 synthetic series, and standard deviation of the same statistic calculated on the 229 observed time-series is reported too (large vertical line). It is evident that L-kurtosis dispersion present a different behavior with respect to that of L-CV and L-skewness, whose standard deviations are higher than the corresponding standard deviations obtained by synthetic time series. Indeed we observed that the empirical standard deviation of L-kurtosis is located in the bulk the distribution and it could pass an homogeneity test with a 5% significance level.

Similar conclusions can be deduced from the diagrams on the right of Figure 7.9 in which are reported the results of a uniformity analysis of L-CV's, L-skewness' and L-kurtosis' exceeding probabilities, represented by a rank histogram. These probabilities were calculated from CDFs obtained from 10'000 synthetic series for each station. In a homogeneous region, it is expected that whatever statistic of interest (for instance L-CV) has the same probability to occupy one of the 10'000 ranks associated to the same statistic calculated on the 10'000 synthetic series of the same station. A natural normalization of ranks between 0 and 1, with respect to the number of synthetic series, is the exceeding probability, which is expected to be uniformly distributed between 0 and 1 when the set of the all stations of a homogeneous region is considered. This uniformity analysis clearly confirms that only L-kurtosis statistic distribution is consistent with the hypothesis of homogeneous region with a 5% significance level.

Two conclusions can be drawn from these results. The first is that it is not necessary to adopt distributions with more than three parameters. We expect that a three-parameter distribution, such as the GEV is sufficient for our goals. Moreover, it is evident that the *Kappa* distribution assumes a

GEV shape when the h parameter is zero, and that the sample value determined on the 229 considered stations is close to zero (Table 7.2), this also suggests the implementation of the GEV distribution. The second conclusion is that the assumption of a unique homogeneous zone for the whole territory of Region Sardinia is not statistically acceptable, so it is not right to assume a probabilistic distribution characterized by the same scale and shape parameter for all the sites.

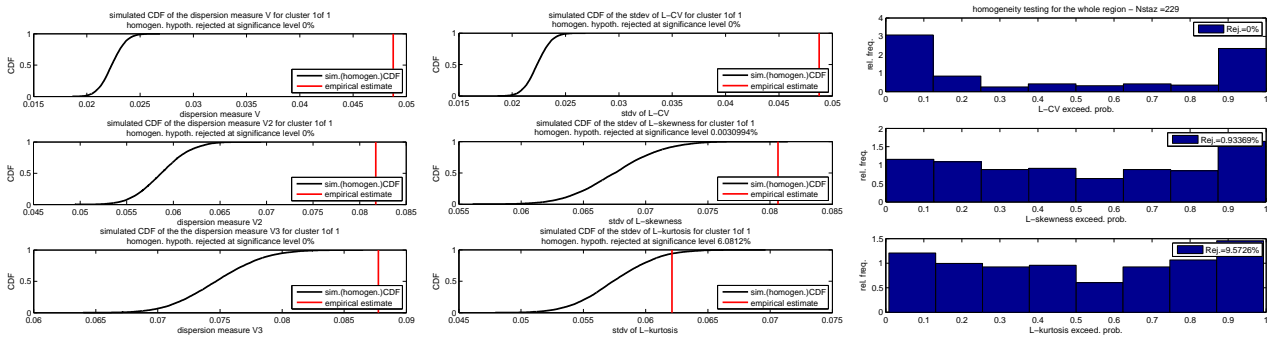


Figure 7.9: **Hypothesis of a single homogeneous zone for the whole Sardinia.**

Left: in black it's shown the CDF of dispersion measures V , V_2 and V_3 (top to bottom) obtained from 10'000 synthetic generations for each station, the vertical red lines represent the corresponding sample values.

Middle: in black it's shown the CDF of standard deviations of statistics L-CV, L-skewness and L-kurtosis (top to bottom) obtained from 10'000 synthetic generations for each station, the vertical red lines represent the corresponding sample values.

Right: test of uniformity of the exceedance probability for statistics L-CV, L-skewness and L-kurtosis (top to bottom) calculated from the CDF obtained from 10'000 synthetic generations for each station.

7.2.3 Identification of new homogeneous regions

Several hypothesis about the partition of the regional territory into homogeneous zones were formulated and the tests about statistical likelihood were conducted. The aggregation of sites in hypothetical homogeneous regions was done using a hierarchical cluster analysis based on the Ward method, with the L-CV and L-skewness as metrics, since these metrics show the highest territorial dispersion. In order to have compact sets of stations, a condition of territorial continuity based on Delaunay's triangulation, which allows only aggregations between contiguous stations, was implemented.

Clusters with L-CV metrics

Results of hierarchical cluster analysis with the Ward method, using the L-CV (t) as a metric, are shown below.

The minimum number of regions that can be considered statistically homogeneous on the basis of the heterogeneity measure H is equal to 7, the corresponding aggregations are shown in Figure 7.10. In Table 7.4, we can notice that all the clusters are homogeneous according to the heterogeneity measure H . Moreover, in the last column, PMW estimation of *Kappa* distribution's parameter h , for each cluster, is reported. Parameter estimates are $\simeq 0$, this suggest the use of the GEV distribution. This configuration was considered as starting point for further configurations.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
1	40	0.18	0.18	0.16	1.59	1.54	1.24	-0.02
2	40	0.21	0.25	0.19	-0.37	0.99	1.00	-0.09
3	8	0.29	0.36	0.26	-0.97	-1.03	-0.83	-0.05
4	48	0.29	0.29	0.20	-1.19	0.28	0.71	0.07
5	63	0.20	0.23	0.18	-0.65	-0.91	-0.44	-0.08
6	25	0.23	0.22	0.17	0.30	1.59	1.20	0.08
7	5	0.36	0.37	0.25	-1.28	-0.89	-1.04	0.20

Table 7.4: L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h , for the 7 homogeneous zones obtained with L-CV metric.

First of all, stations number 42, 81 and 82, see Figure 2.4, have been moved from cluster 1 to cluster 2. Moreover, on one side clusters 2 and 5 and clusters 3 and 4 on the other side were merged, creating a configuration with 5 clusters, labeled as configuration **A**, described in Figure 7.11. Table 7.5 shows that all the regions are statistically homogeneous according to statistics H ,

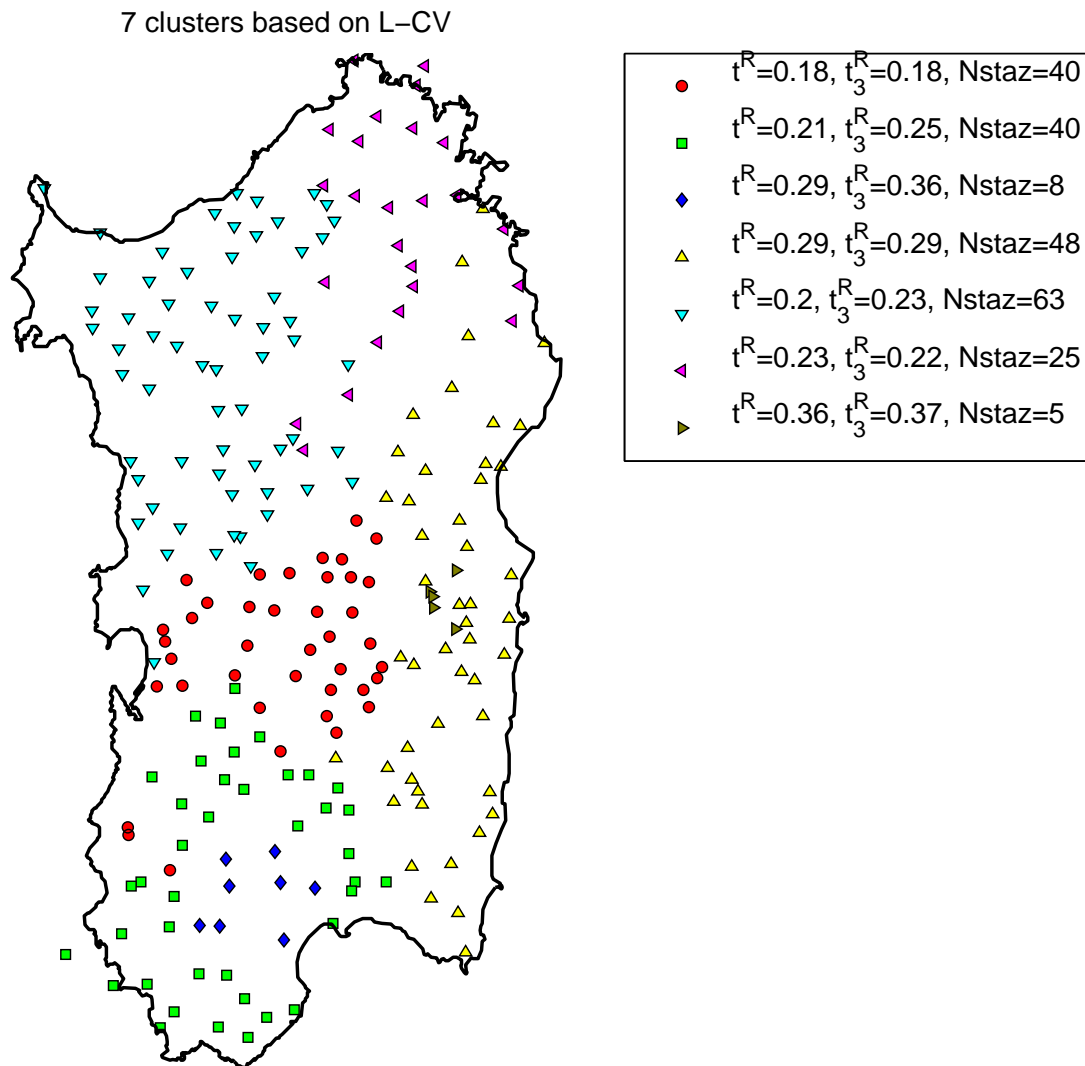


Figure 7.10: Spatial distribution of the 7 homogeneous regions obtained by cluster analysis with metric L-CV. The legend shows the regional values of the L-moment ratios t^R, t_3^R and the number of stations for each cluster.

H_2 and H_3 . The results of the other comparisons and of homogeneity tests for the 5 hypothesized homogeneous regions are reported in Figure 7.12, 7.13, 7.14, 7.15 and 7.16. In each Figure, scatter plots between L-moment ratios are shown in the left part and homogeneity tests in the right part.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
A ₁	36	0.18	0.18	0.16	1.68	1.30	1.04	-0.05
A ₂	107	0.20	0.24	0.19	0.92	0.47	0.60	-0.07
A ₃	56	0.29	0.30	0.21	-1.61	0.65	0.93	0.05
A ₄	25	0.23	0.22	0.17	0.29	1.63	1.22	0.08
A ₅	5	0.36	0.37	0.25	-1.29	-0.91	-1.07	0.20

Table 7.5: **Hypothesis A**: partition in 5 homogeneous regions obtained through cluster analysis with the L-CV metric. L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h are reported for each homogeneous region.

The new configuration **B** was obtained merging cluster **A₅**, which is constituted just by 5 stations characterized by high asymmetry, with cluster **A₃**, creating the new cluster **B₃**. The spatial distribution of the stations in the four clusters configuration **B** is shown in Figure 7.17, whereas results of the homogeneity tests for the new cluster **B₃** are reported in Figure 7.18. In Table 7.6, regional values of L-moment, heterogeneity measures described by Hosking and Wallis (1997), and the regional value of *Kappa* distribution's parameter h are reported for each cluster.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
B ₁	36	0.18	0.18	0.16	1.64	1.28	1.03	-0.05
B ₂	107	0.20	0.24	0.19	0.95	0.47	0.60	-0.07
B ₃	61	0.29	0.30	0.21	-0.08	0.85	0.62	0.05
B ₄	25	0.23	0.22	0.17	0.29	1.63	1.24	0.08

Table 7.6: **Hypothesis B**: partition in 4 homogeneous regions obtained through cluster analysis with the L-CV metric. L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h are reported for each homogeneous region.

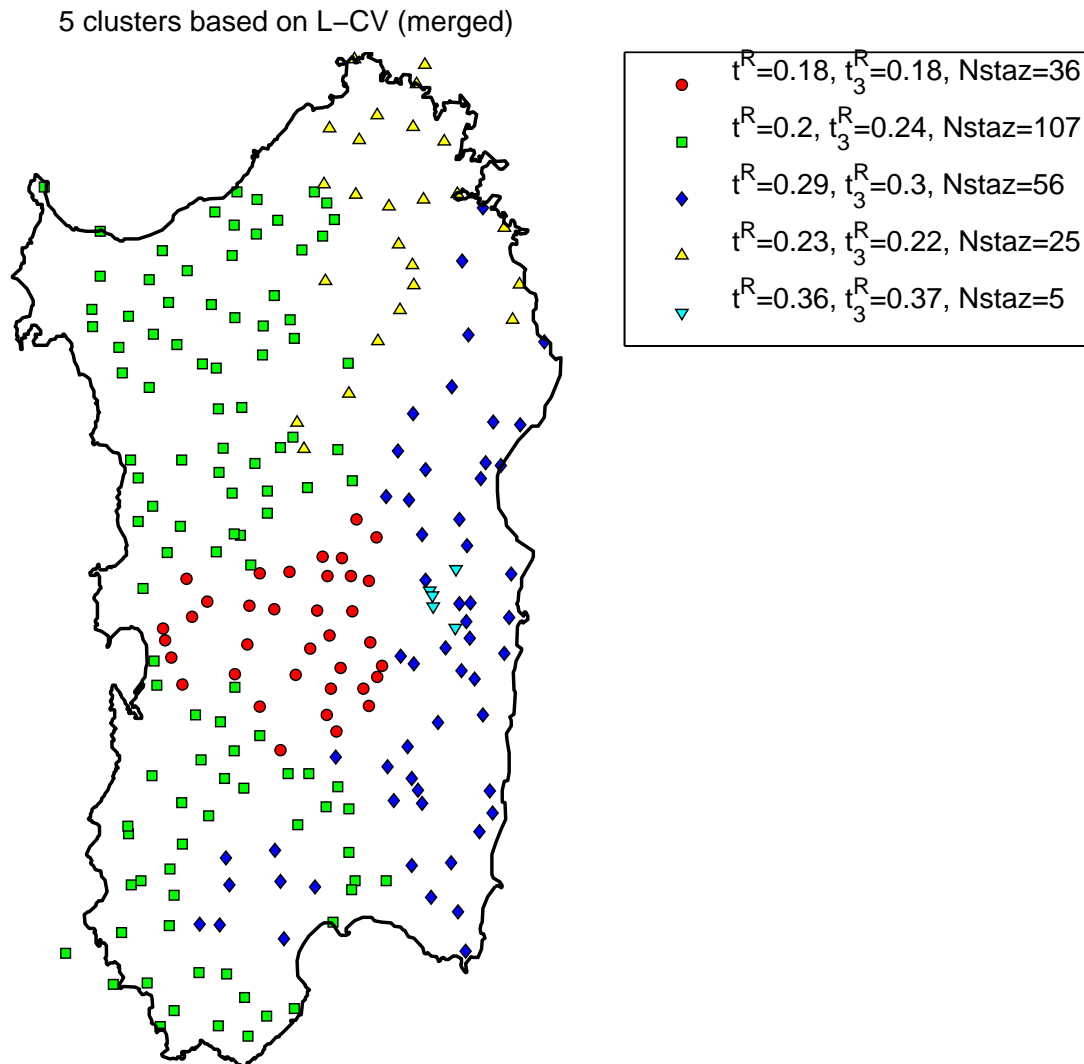


Figure 7.11: **Hypothesis A**: Spatial distribution of the 5 homogeneous regions obtained by cluster analysis with metric L-CV. The legend shows the regional values of the L-moment ratios t^R, t_3^R and the number of stations for each cluster.

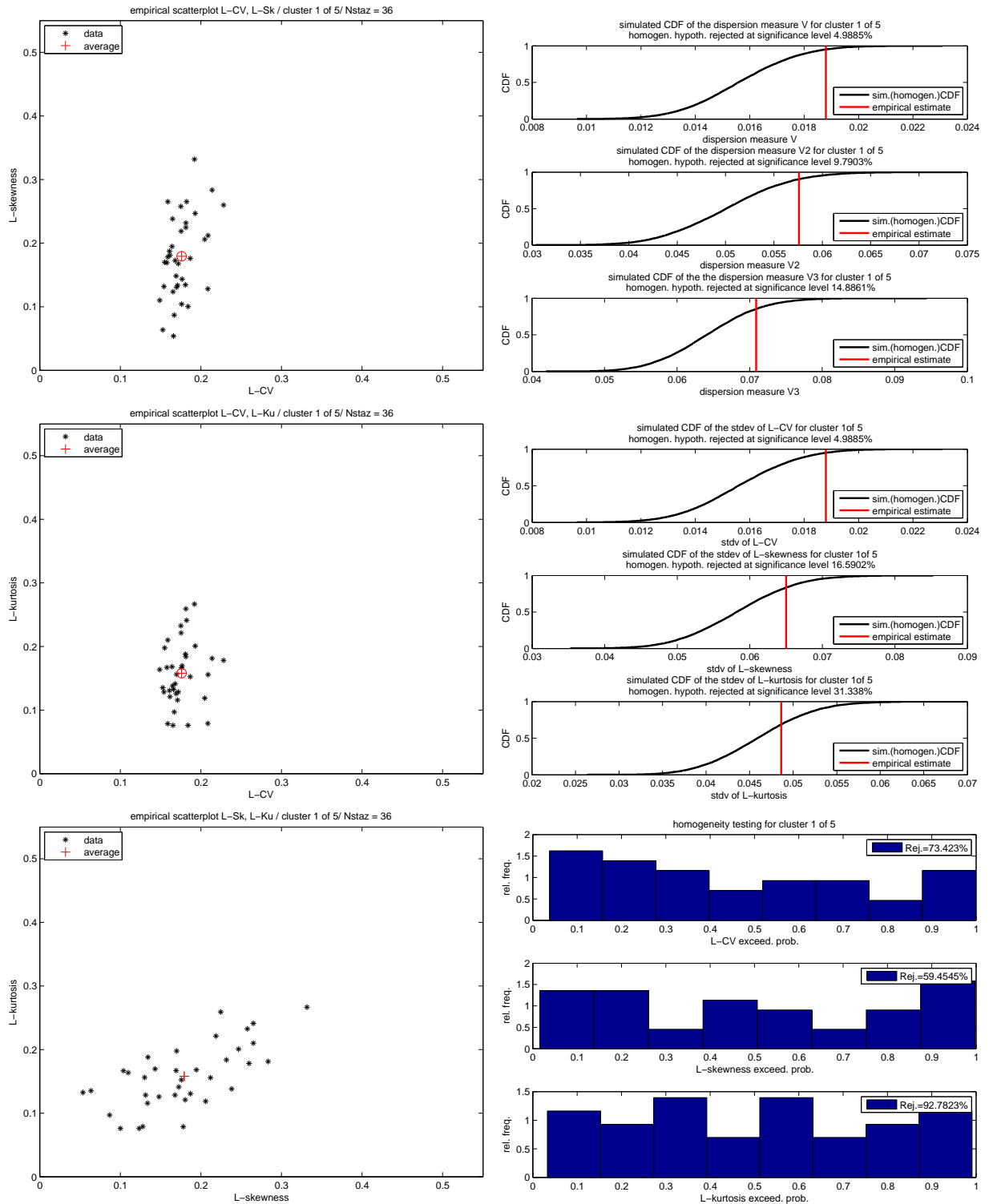


Figure 7.12: Hypothesis A: Cluster A₁
 Same description of the Figures 7.8 and 7.9

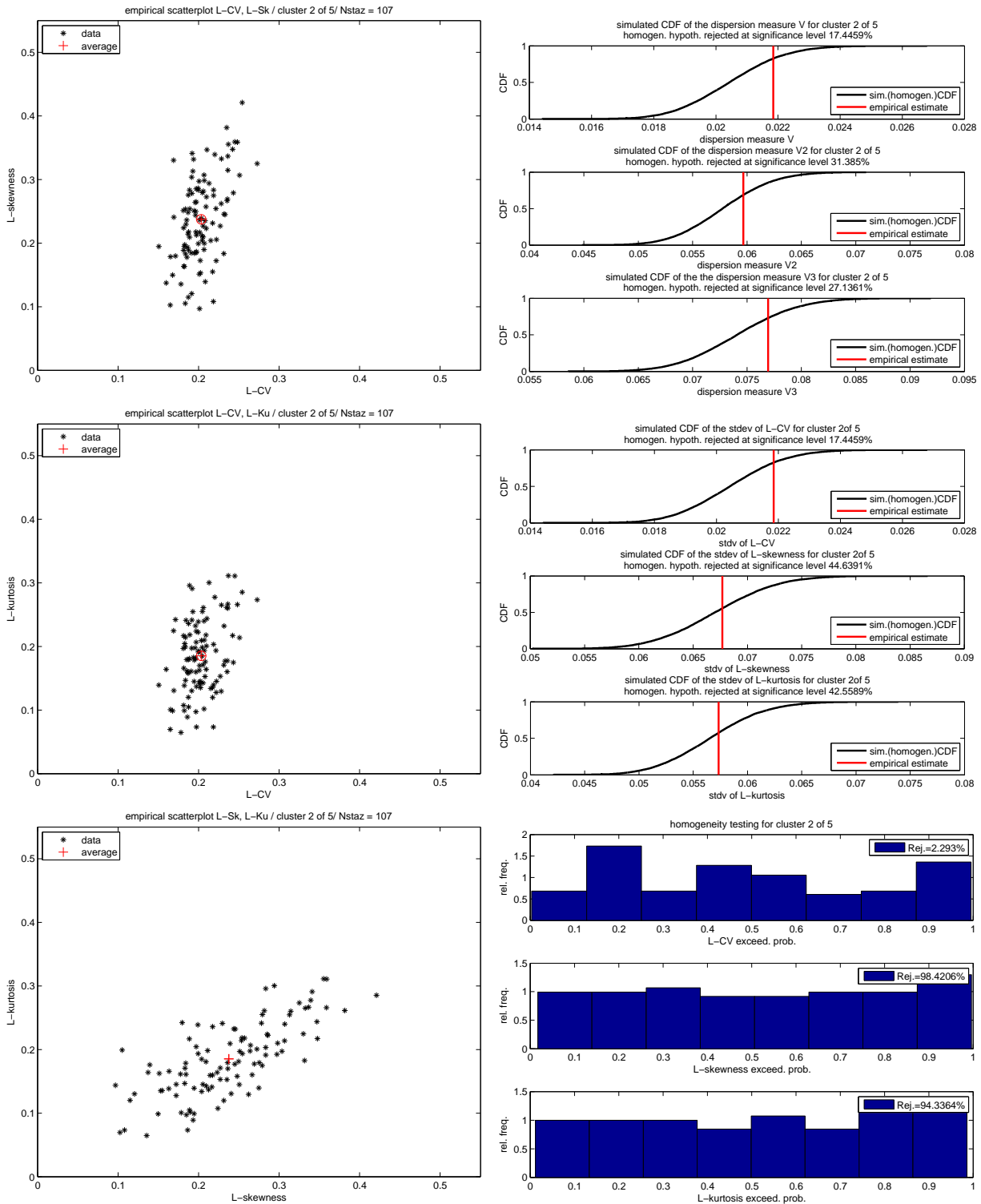


Figure 7.13: Hypothesis A: Cluster A₂
 Same description of the Figures 7.8 and 7.9

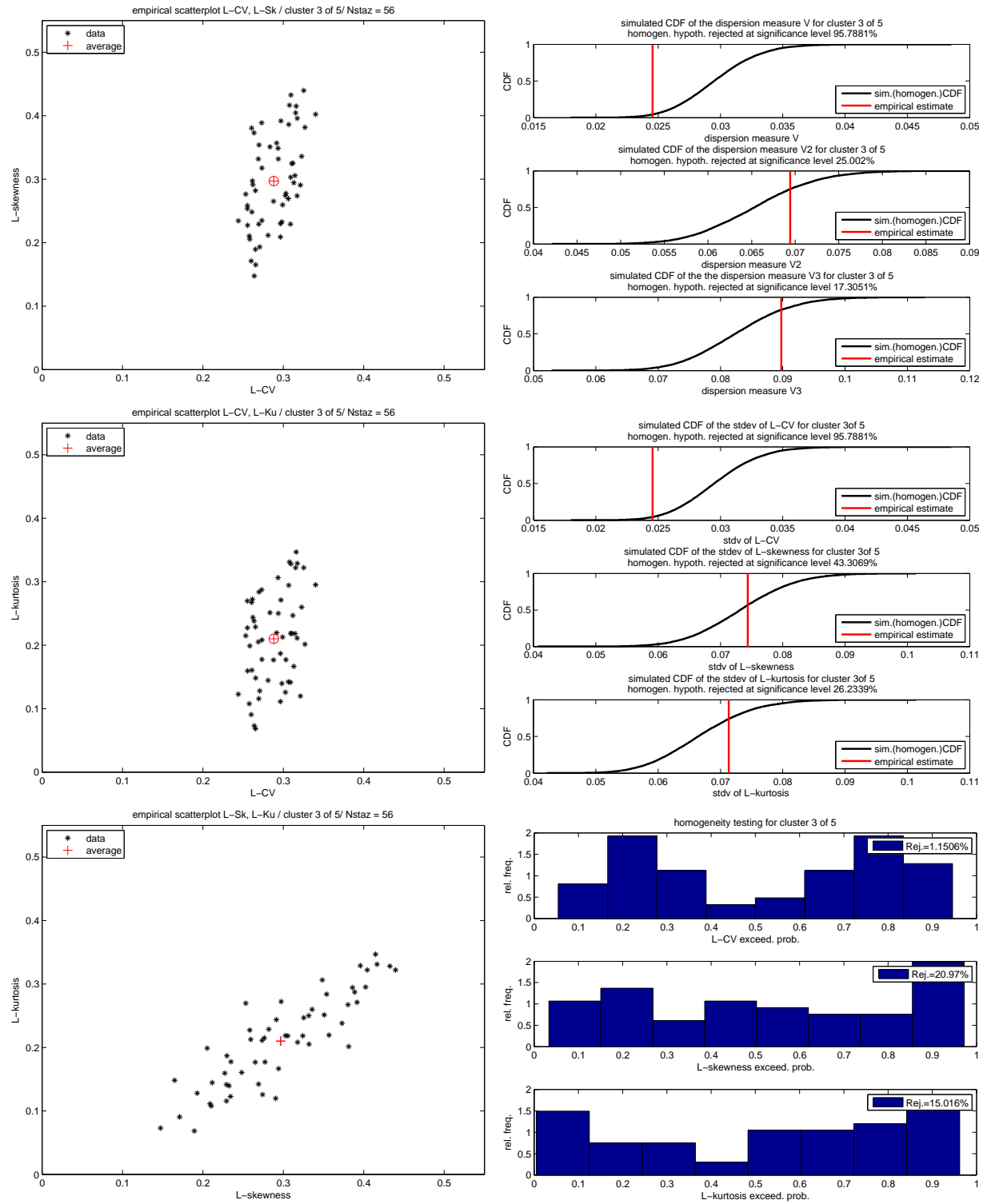


Figure 7.14: Hypothesis A: Cluster A₃
 Same description of the Figures 7.8 and 7.9

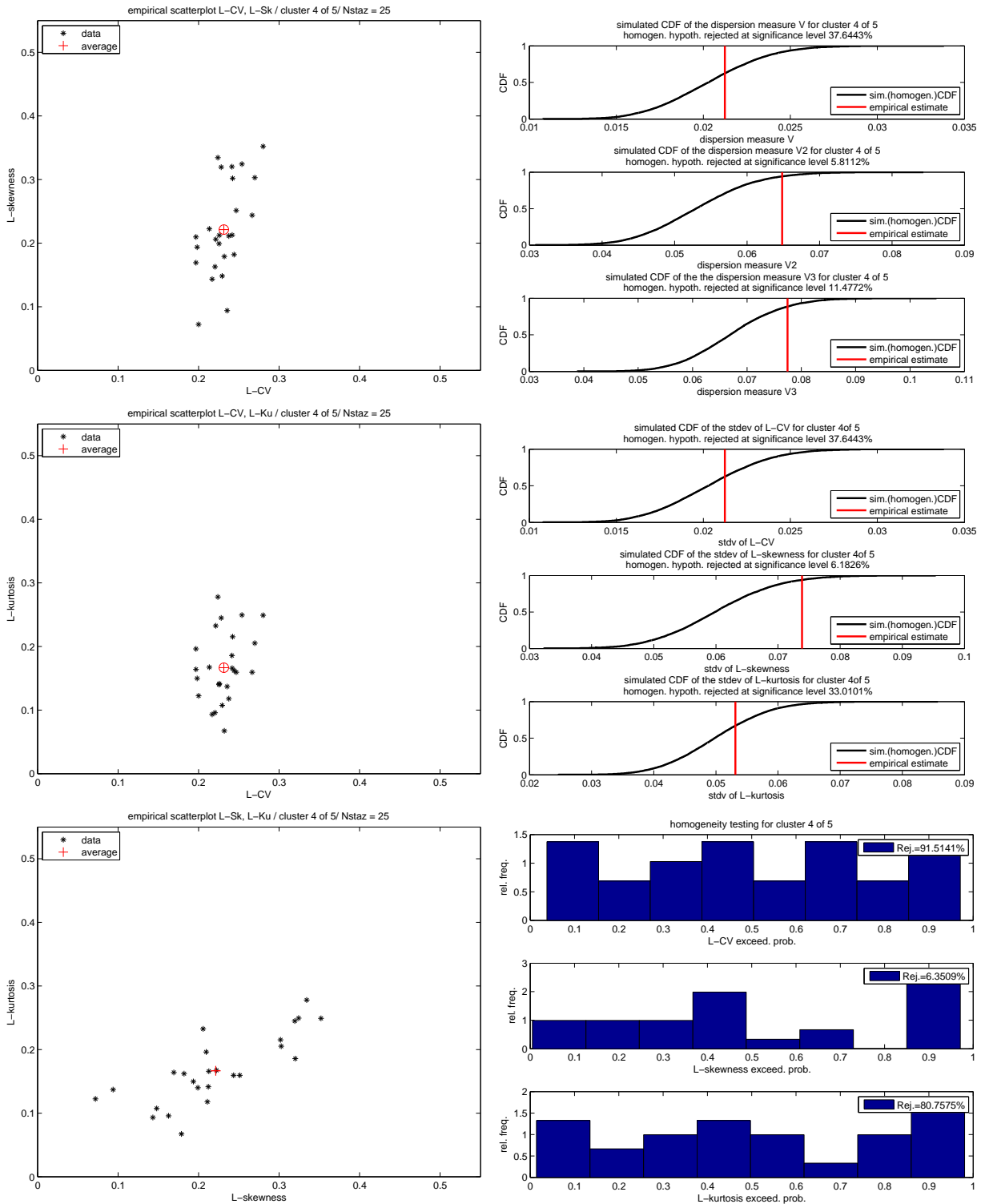


Figure 7.15: Hypothesis A: Cluster A₄
 Same description of the Figures 7.8 and 7.9

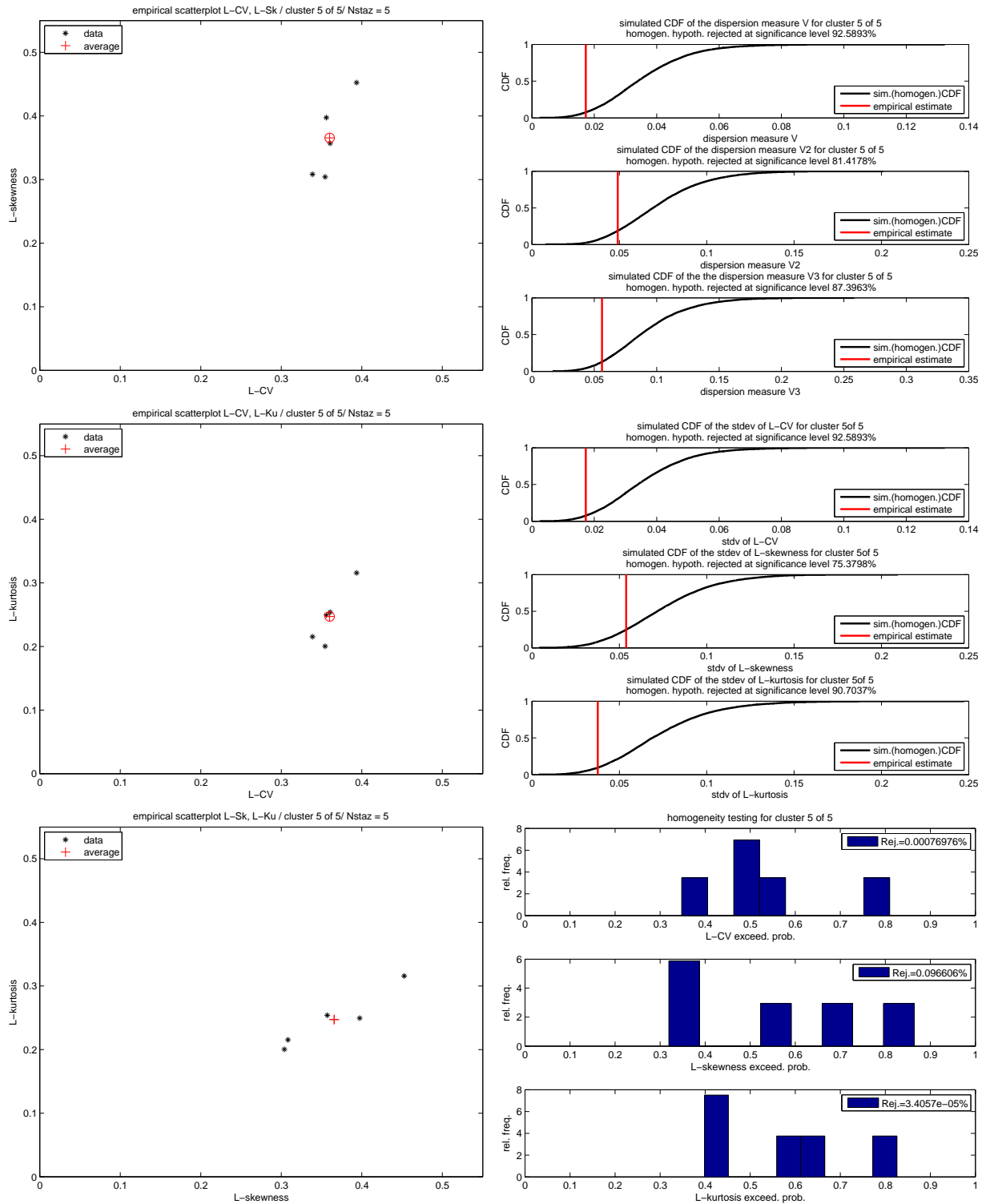


Figure 7.16: Hypothesis A: Cluster A₅
 Same description of the Figures 7.8 and 7.9

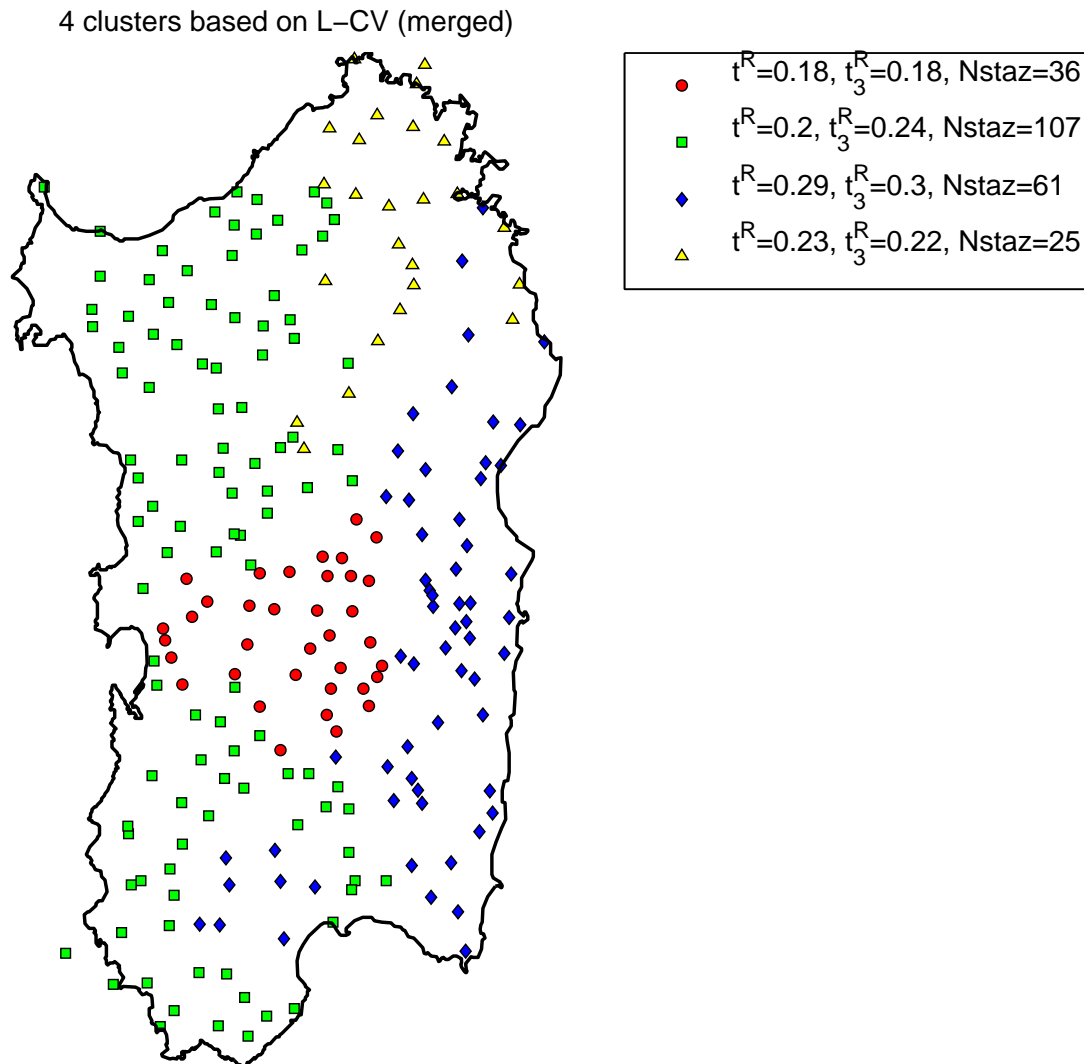


Figure 7.17: **Hypothesis B**: Spatial distribution of the 4 homogeneous regions obtained by cluster analysis with metric L-CV. The legend shows the regional values of the L-moment ratios t^R, t_3^R and the number of stations for each cluster.

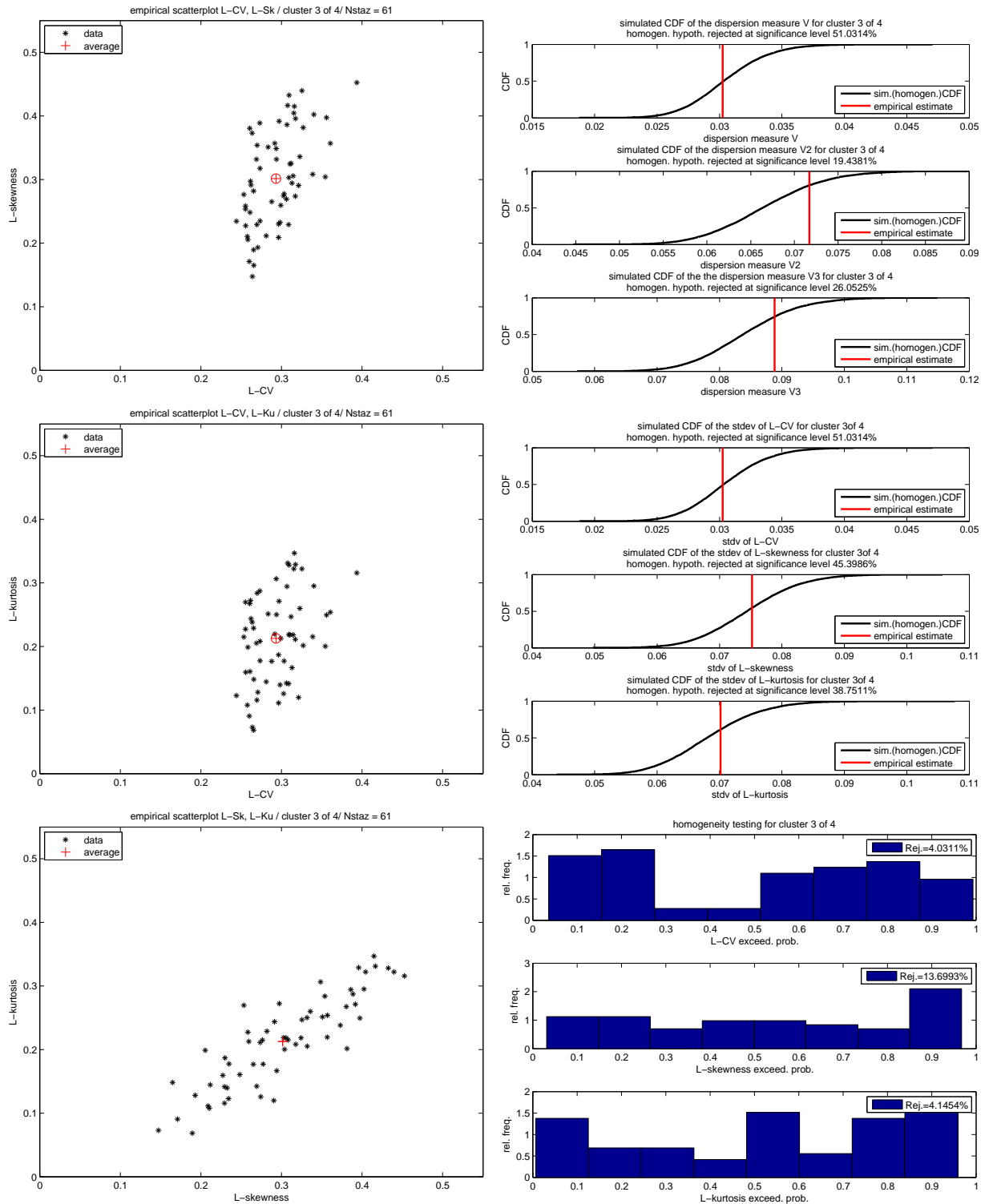


Figure 7.18: Hypothesis B: Cluster B₃
Same description of the Figures 7.8 and 7.9

Cluster with L-skewness and weighted L-CV metrics

A new hierarchical cluster analysis has been conducted using Ward's method. This time we used as metric a measure obtained through a weighted combination of L-CV (t) and L-skewness (t_3). This metrics, with the index-rainfall characterize three-parameter distributions like the GEV distribution. The weights are determined proportionally to the ratio between dispersion of t and t_3 , obtained with Monte Carlo simulations following this procedure: for every station we simulated 10'000 time series using a *Kappa* distribution with parameter estimates equal to the regional ones. We used the simplifying hypothesis that each series had a predetermined length, equal for all the series. Cases in which the length of historic series of annual maxima are equal to 50, 70 and 90 years respectively were simulated. L-moment ratios were estimated from the samples. In the left part of Figure 7.19, we can notice that the dispersion of t_3 is higher than the dispersion of t . The same phenomenon can be seen if the corresponding standard deviations are considered, middle part of Figure 7.19. The ratio between standard deviations of t and of t_3 is on average equal to 0.33 (right part of Figure 7.19). According to these results, the utilized metric for the cluster analysis was defined employing the weights for t and t_3 in ratio of 3 to 1.

The minimum number of regions that can be considered statistically homogeneous according to the statistic H is 8, their spatial distribution is reported in Figure 7.20. In Table 7.7 we can notice that all the clusters are homogeneous according to Hosking and Wallis heterogeneity measures. Now *Kappa* distribution's parameter h moves away from zero value in several clusters, but this phenomenon comes up in the zones with a reduced number of stations, so it can be due to sample uncertainty. For instance, in clusters 2 and 4, which are composed only by 16 and 13 stations respectively, there are the highest absolute values of h , that are equal to -0.48 and 0.30. This configuration has been considered as a starting point for further configurations.

First of all, stations number 38, 42, 81, 82 and 104 of clusters 1 and 2 were moved from a cluster to another. Furthermore, clusters 2, 4 and 5 and clusters 3 and 7 were joined, creating a new configuration with 5 clusters that we called configuration **C** and that is illustrated in Figure 7.21. In Table 7.8, it is evident that all the regions are statistically homogeneous according to H , H_2 and H_3 statistics. Moreover, parameter h is now close to zero for each of the 5 homogeneous regions. The results of the homogeneity tests for each of the 5 hypothesized homogeneous regions are reported in Figures 7.22, 7.23, 7.24, 7.25 and 7.26. In each figure, scatter plots between L-moment ratios and homogeneity tests are shown in the left and right part, respectively.

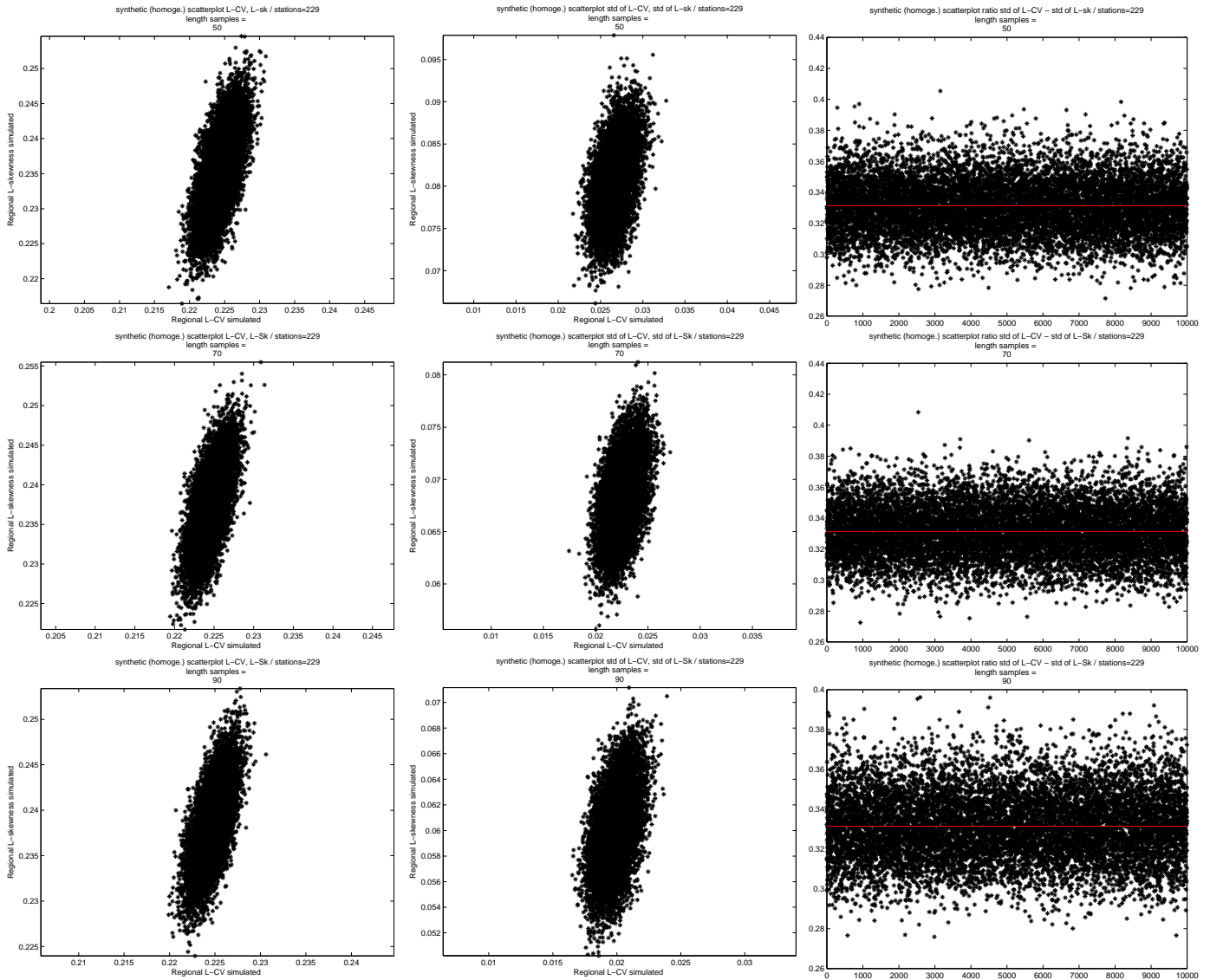


Figure 7.19: Left: scatterplots of the L-moments ratios t and t_3 , obtained through a Monte Carlo procedure. The length of each time series is equal to 50, 70 and 90 years (top to bottom).

Middle: scatterplots of the standard deviation of the L-moments ratios t and t_3 , obtained by a Monte Carlo procedure. The length of each time series is equal to 50, 70 and 90 years (top to bottom).

Right: scatterplots of the ratio between the standard deviation of the L-moments ratios t and t_3 , obtained by a Monte Carlo procedure. The length of each time series is equal to 50, 70 and 90 years (top to bottom).

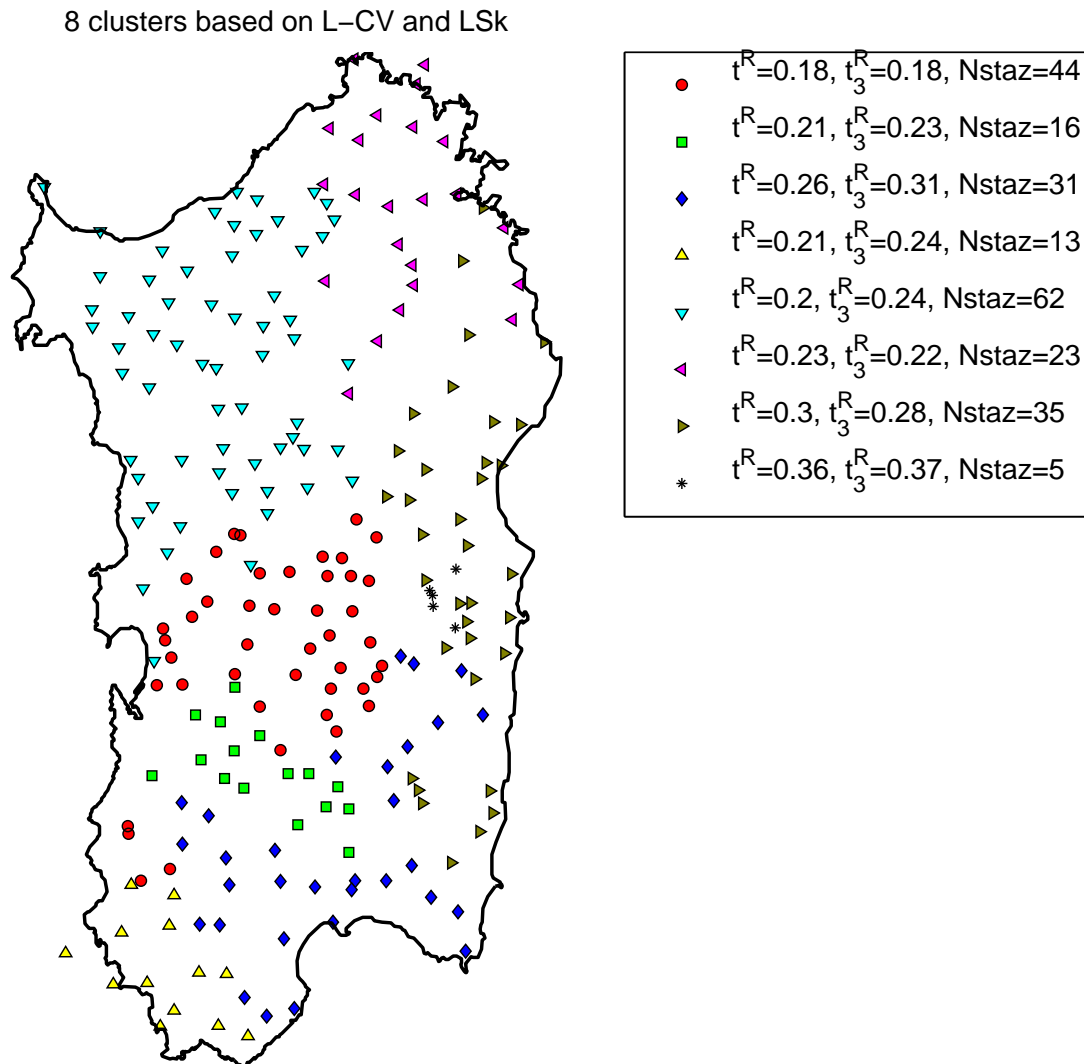


Figure 7.20: Spatial distribution of the 8 homogeneous regions obtained by cluster analysis with metrics L-CV and L-skew properly weighted. The legend shows the regional values of the L-moment ratios t^R, t_3^R and the number of stations for each cluster.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
1	44	0.18	0.18	0.16	1.46	1.18	0.92	-0.04
2	16	0.21	0.23	0.20	-2.06	0.21	0.49	-0.48
3	31	0.26	0.31	0.23	-0.77	-0.31	-0.81	-0.21
4	13	0.21	0.24	0.16	-0.31	0.85	0.23	0.30
5	62	0.20	0.24	0.18	-0.66	-0.93	-0.34	-0.06
6	23	0.23	0.22	0.16	0.56	1.22	0.68	0.10
7	35	0.30	0.28	0.19	-1.22	0.59	1.44	0.19
8	5	0.36	0.37	0.25	-1.28	-0.91	-1.08	0.20

Table 7.7: L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h , for the 8 homogeneous zones obtained with weighted L-CV and L-skewness metrics.

Another new configuration, called **D**, has been obtained joining the cluster **C₅**, which is constituted by only 5 stations characterized by high asymmetry, with the cluster **C₃**, determining the new cluster **D₃**. This configuration is reported in Figure 7.27, whereas results about the homogeneity tests for the new cluster **D₃** are reported in Figure 7.28. Regional values of L-moment statistics and heterogeneity measures and the regional value of *Kappa* distribution's parameter h , are reported, for each cluster, in Table 7.9.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
C₁	39	0.18	0.18	0.16	1.32	0.88	0.76	-0.08
C₂	96	0.20	0.24	0.18	-0.74	-0.12	0.50	-0.05
C₃	66	0.28	0.30	0.21	0.15	1.01	0.76	0.02
C₄	23	0.23	0.22	0.16	0.54	1.21	0.68	0.10
C₅	5	0.36	0.37	0.25	-1.27	-0.91	-1.05	0.20

Table 7.8: **Hypothesis C**: partition in 5 homogeneous regions obtained through cluster analysis with weighted L-CV and L-skewness metrics. L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h are reported for each homogeneous region.

5 clusters based on L-CV and LSk (merged)

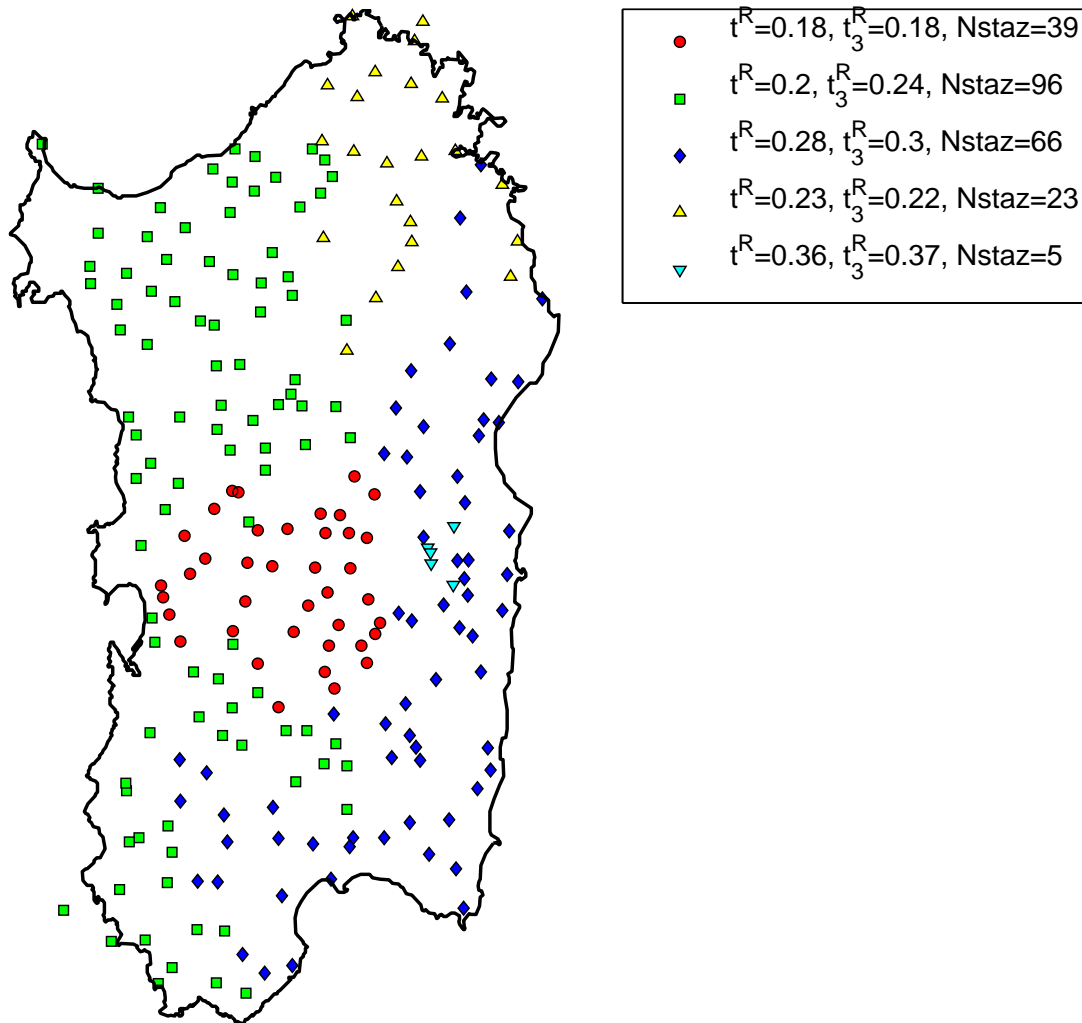


Figure 7.21: **Hypothesis C**: Spatial distribution of the 5 homogeneous regions obtained by cluster analysis with metrics L-CV and L-skew properly weighted. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.

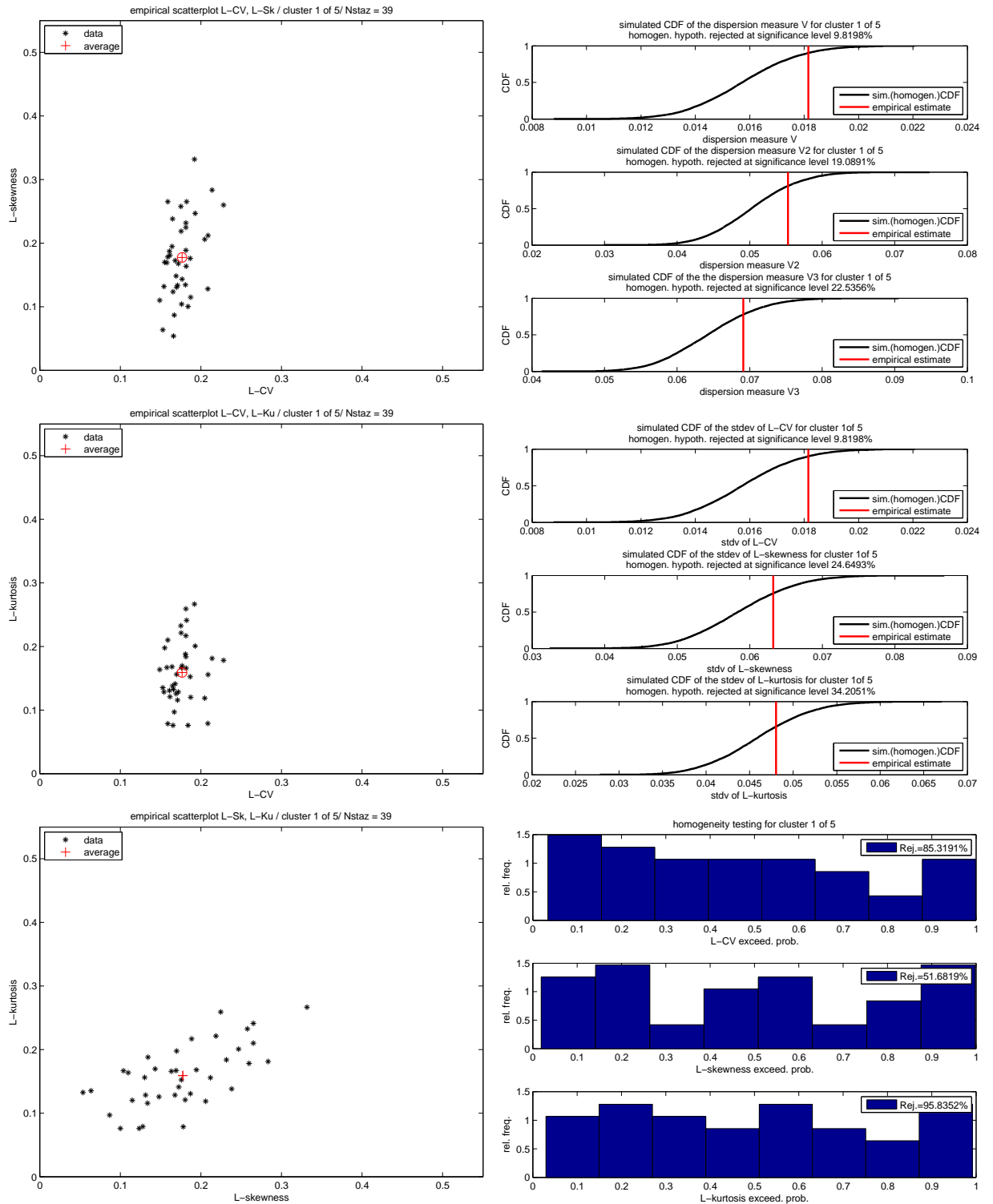


Figure 7.22: Hypothesis C: Cluster C_1
 Same description of the Figures 7.8 and 7.9

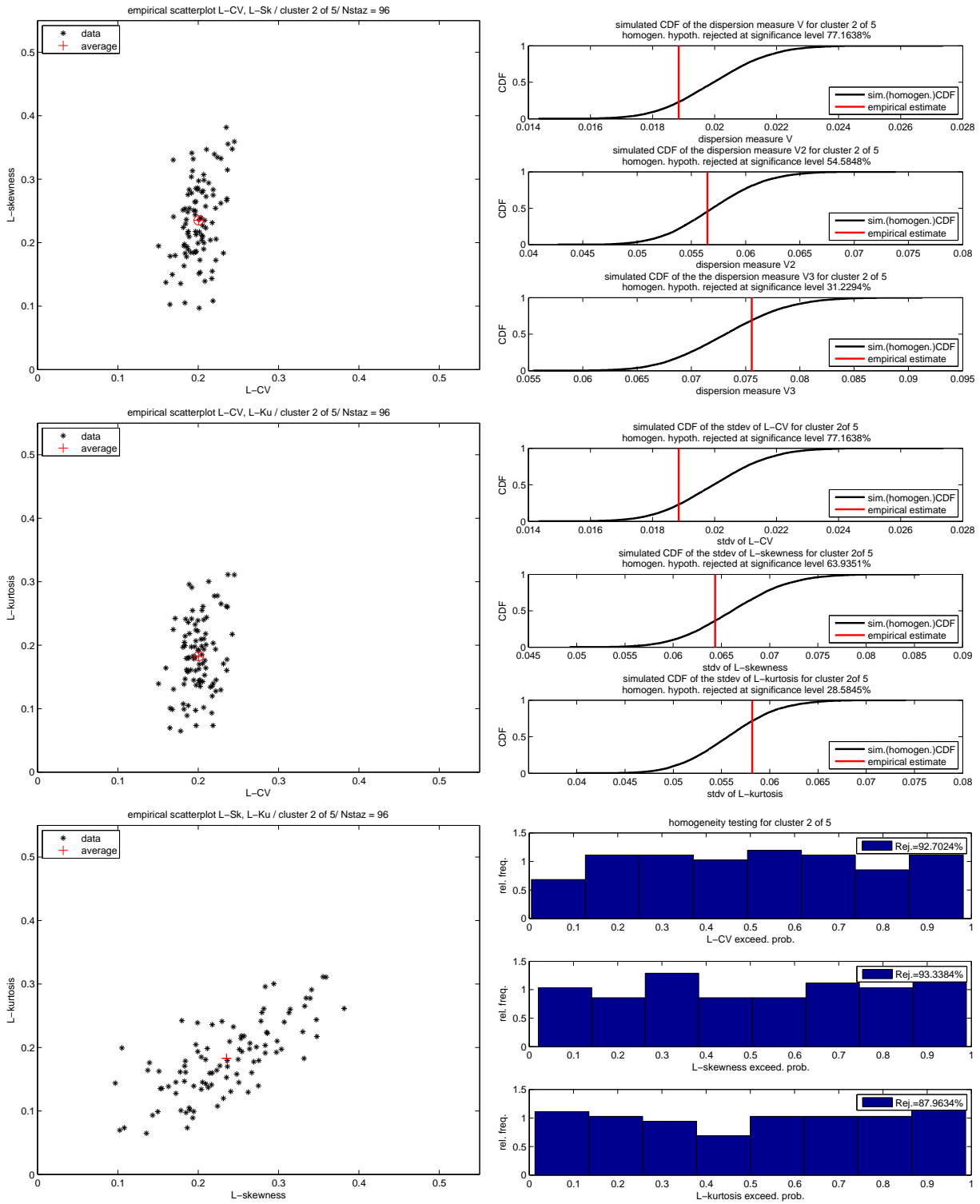


Figure 7.23: Hypothesis C: Cluster C_2
 Same description of the Figures 7.8 and 7.9

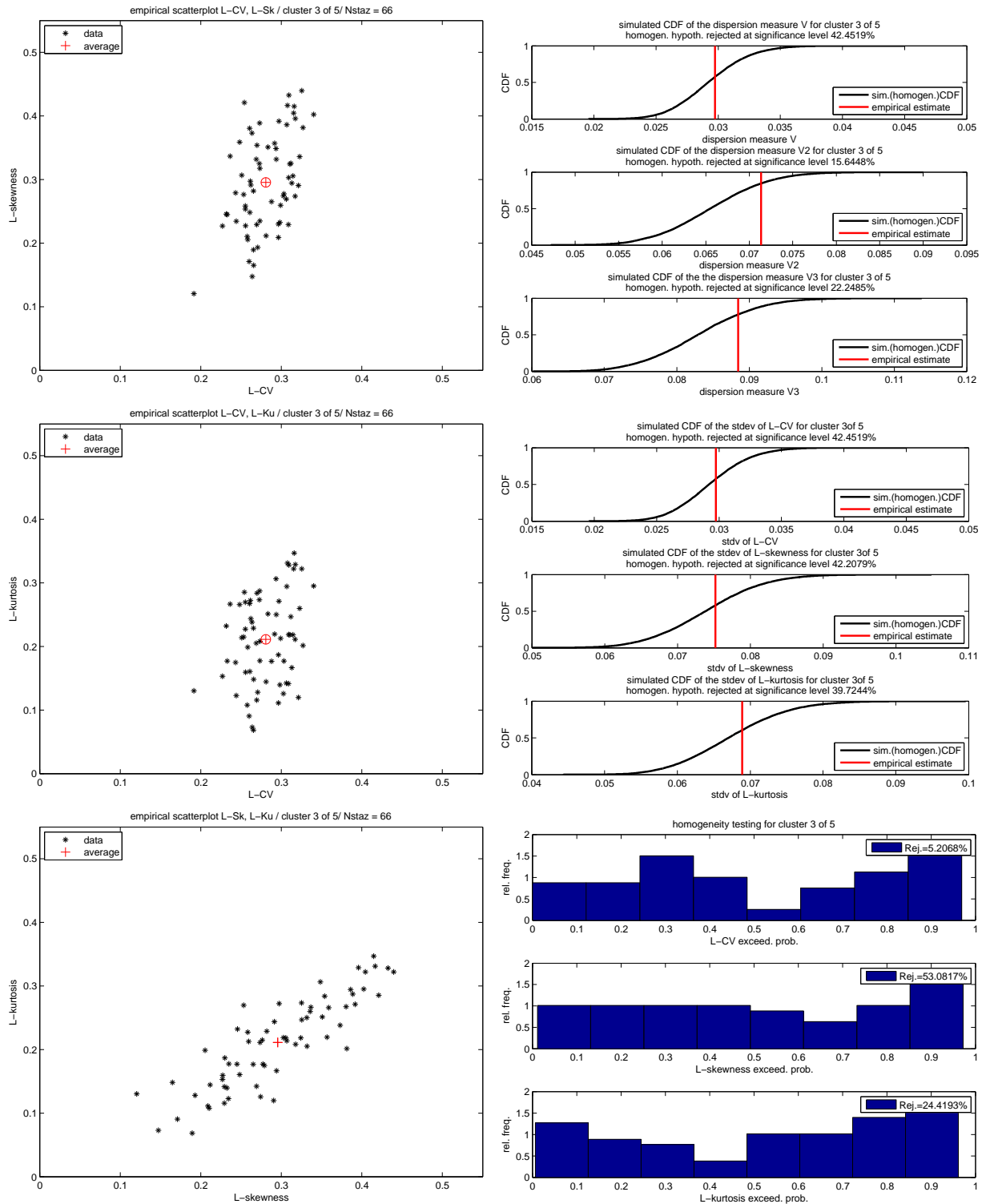


Figure 7.24: Hypothesis C: Cluster C_3
 Same description of the Figures 7.8 and 7.9

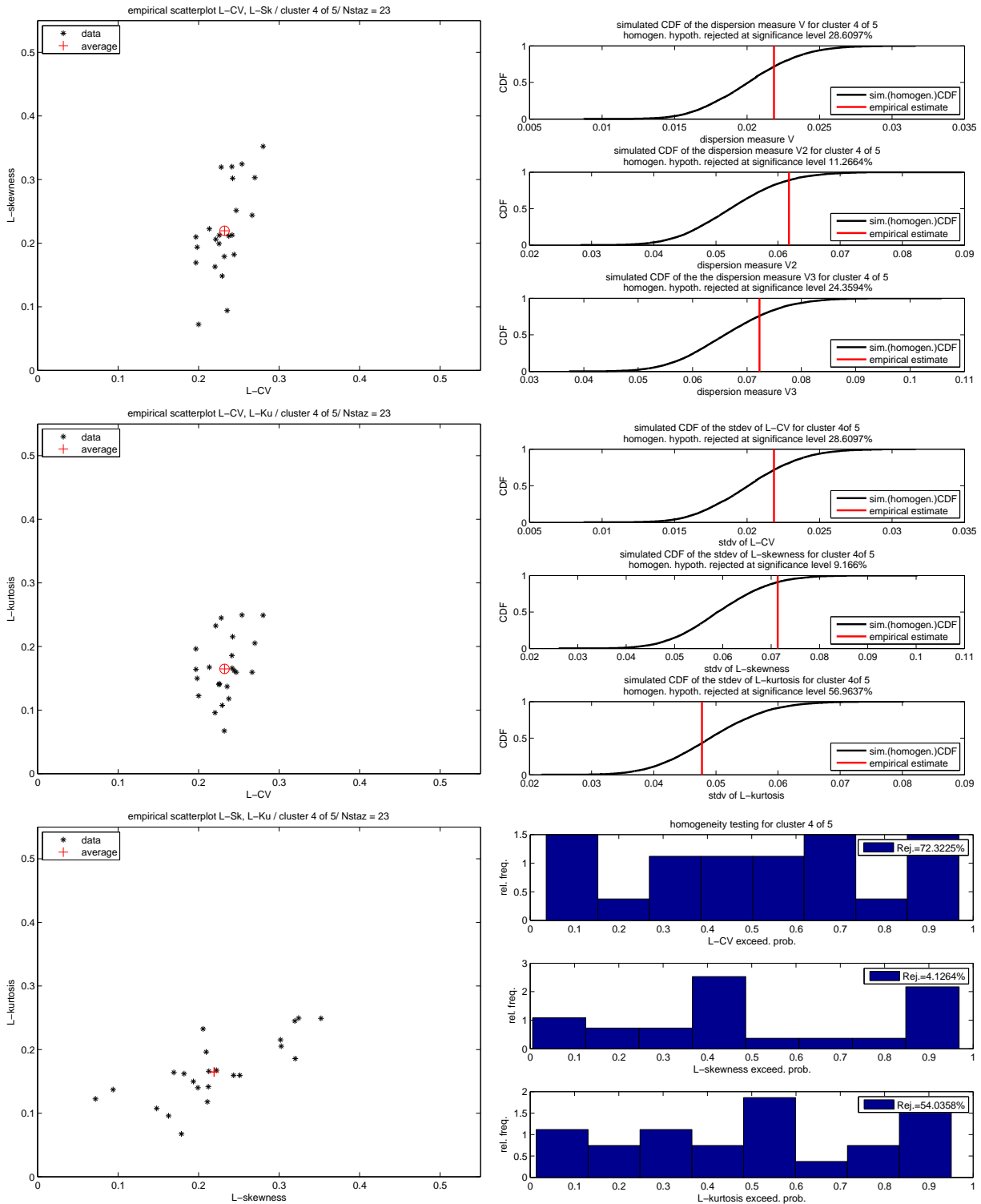


Figure 7.25: Hypothesis C: Cluster C₄
 Same description of the Figures 7.8 and 7.9

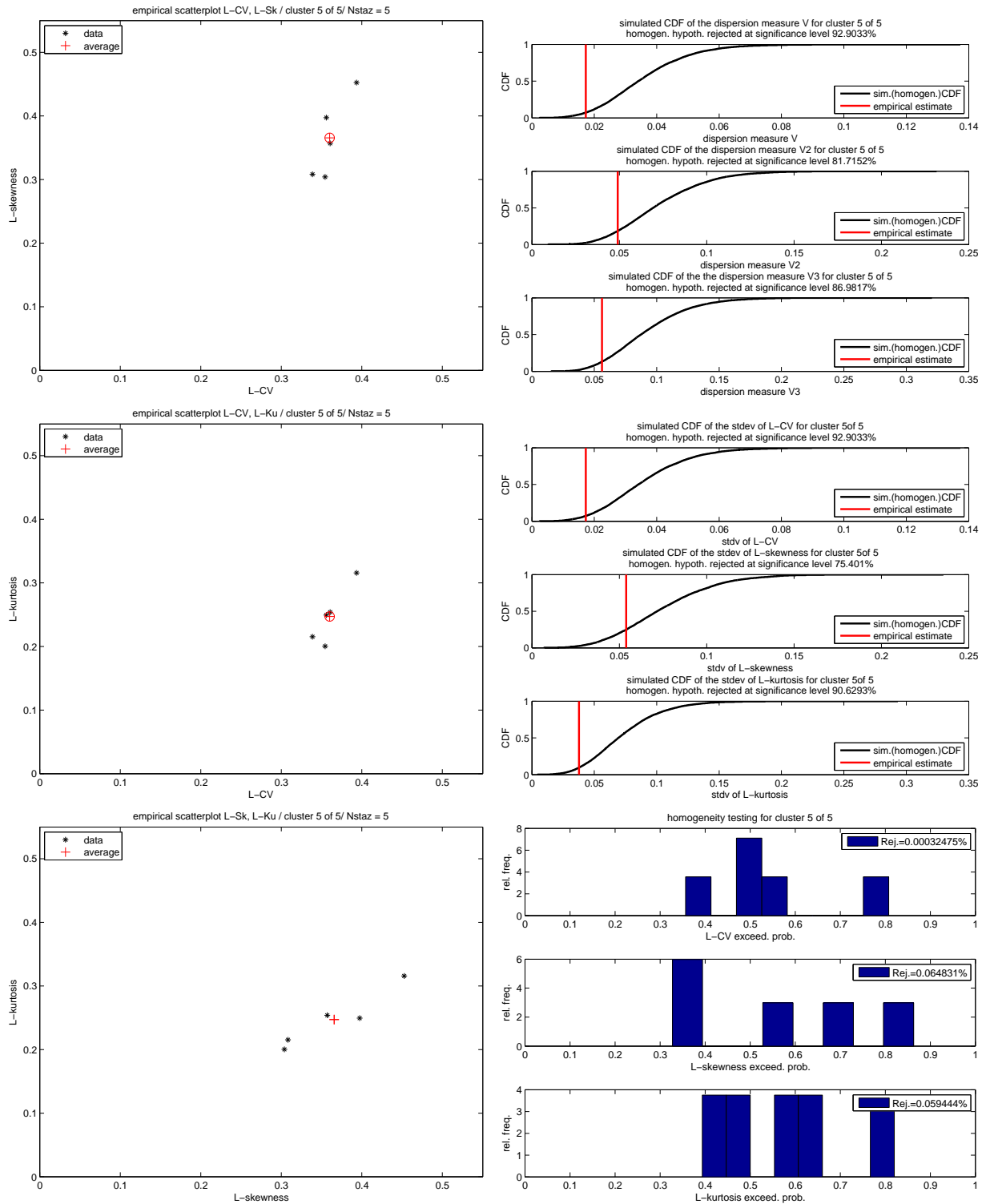


Figure 7.26: Hypothesis C: Cluster C_5
 Same description of the Figures 7.8 and 7.9

4 clusters based on L-CV and LSk (merged)

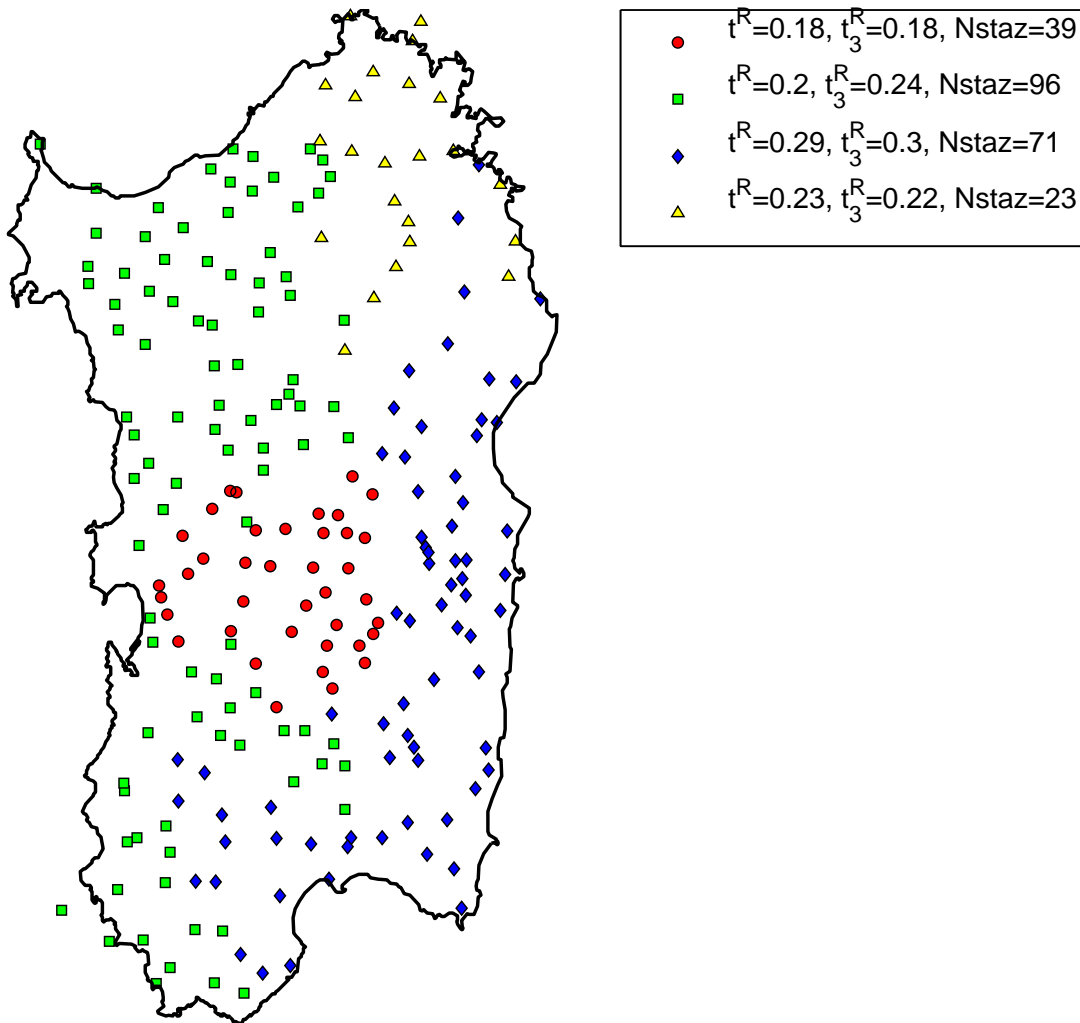


Figure 7.27: **Hypothesis D**: Spatial distribution of the 4 homogeneous regions obtained by cluster analysis with metrics L-CV and L-skew properly weighted. The legend shows the regional values of the L-moment ratios t^R , t_3^R and the number of stations for each cluster.

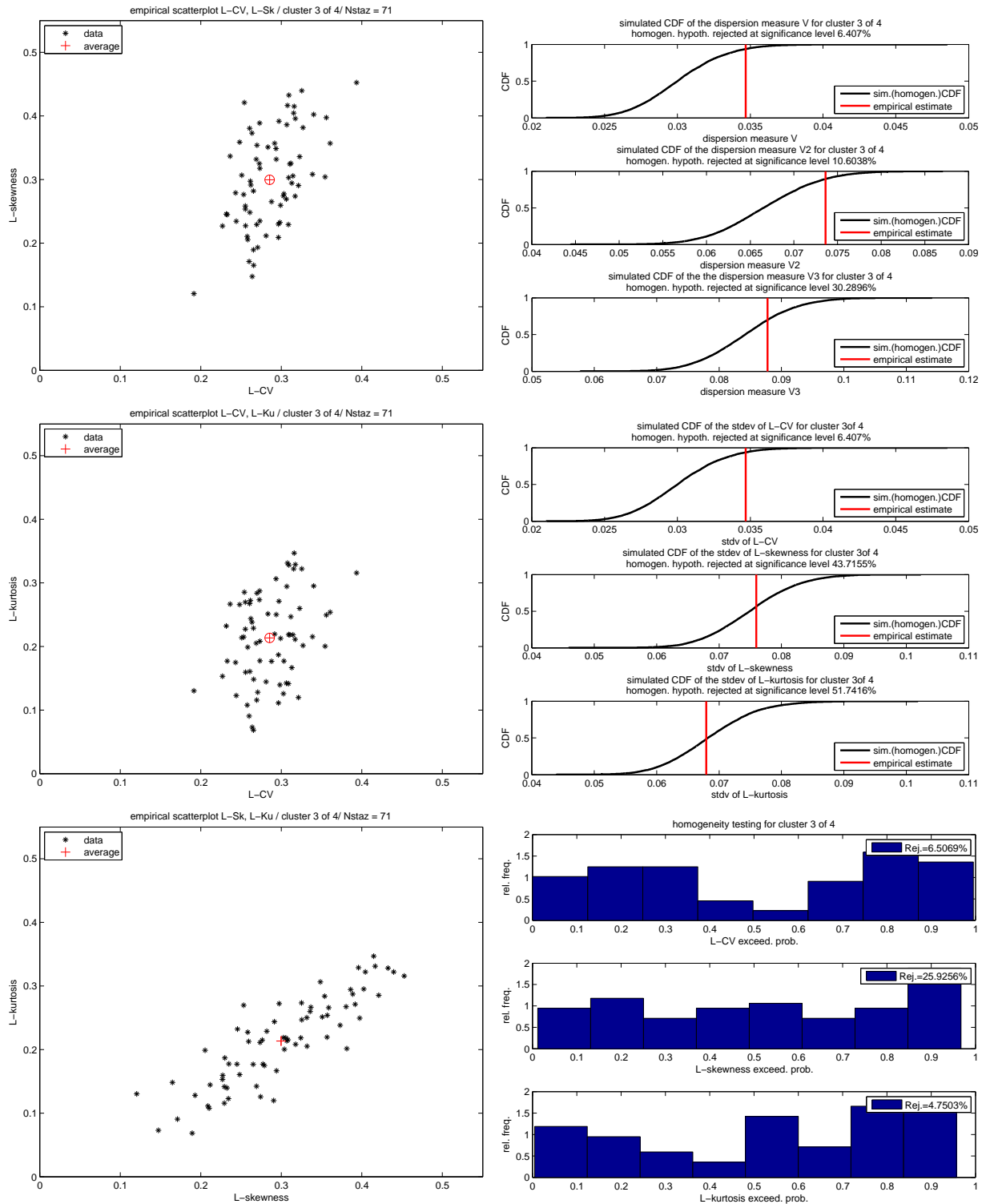


Figure 7.28: Hypothesis D: Cluster D₃
 Same description of the Figures 7.8 and 7.9

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
D ₁	39	0.18	0.18	0.16	1.32	0.88	0.76	-0.08
D ₂	96	0.20	0.24	0.18	-0.74	-0.12	0.48	-0.05
D ₃	71	0.29	0.30	0.21	1.58	1.26	0.50	0.02
D ₄	23	0.23	0.22	0.16	0.53	1.21	0.68	0.10

Table 7.9: **Hypothesis D**: partition in 4 homogeneous regions obtained through cluster analysis with weighted L-CV and L-skewness metrics. L-moment ratios, heterogeneity measures H , H_2 , H_3 and $Kappa$ distribution's parameter h are reported for each homogeneous region.

Empirical clusters

Observing configurations **A** and **C** (or similarly configurations **B** and **D**), it is evident that these configurations are very similar. So we decided to look for a common configuration that minimizes error metrics (between the empirical distribution function and the GEV distribution with regional parameters). In order to reach this goal, some stations were manually moved getting a final configuration with 5 homogeneous regions, labeled configuration **E**. Similarly to previous cases, starting from case **E** and incorporating the 5 stations of the eastern zone in the adjacent cluster, a new configuration with 4 homogeneous regions has been obtained, labeled configuration **F**. The spatial disposition of the two configurations is shown in Figures 7.29 and 7.35. In Tables 7.10 and 7.11, for each cluster, average values of L-moment statistics, heterogeneity measures H , H_2 , H_3 , and the regional value of $Kappa$ distribution's parameter h are reported. The results of homogeneity checks, for each of the 5 homogeneous regions of configuration **E**, are reported in Figures 7.30, 7.31, 7.32, 7.33 and 7.34. Figure 7.36 shows the results of homogeneity checks for the new cluster **F₃** obtained through the union of clusters **E₃** and **E₅**.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
E₁	39	0.18	0.18	0.16	1.31	0.87	0.76	-0.08
E₂	99	0.20	0.23	0.18	-0.54	-0.28	0.23	-0.04
E₃	62	0.28	0.30	0.21	-1.23	0.49	0.49	0.03
E₄	24	0.23	0.22	0.17	0.28	1.48	1.04	0.06
E₅	5	0.36	0.37	0.25	-1.30	-0.92	-1.07	0.20

Table 7.10: **Hypothesis E**: partition in 5 homogeneous regions obtained from hypotheses **A** and **C** through empirical aggregation. L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h are reported for each homogeneous region.

region	n stations	t	t ₃	t ₄	H	H ₂	H ₃	h
F₁	39	0.18	0.18	0.16	1.32	0.86	0.75	-0.08
F₂	99	0.20	0.23	0.18	-0.56	-0.28	0.24	-0.04
F₃	67	0.29	0.30	0.22	0.30	0.72	0.20	0.04
F₄	24	0.23	0.22	0.17	0.27	1.49	1.03	0.06

Table 7.11: **Hypothesis F**: partition in 4 homogeneous regions obtained from hypotheses **B** and **D** through empirical aggregation. L-moment ratios, heterogeneity measures H , H_2 , H_3 and *Kappa* distribution's parameter h are reported for each homogeneous region.

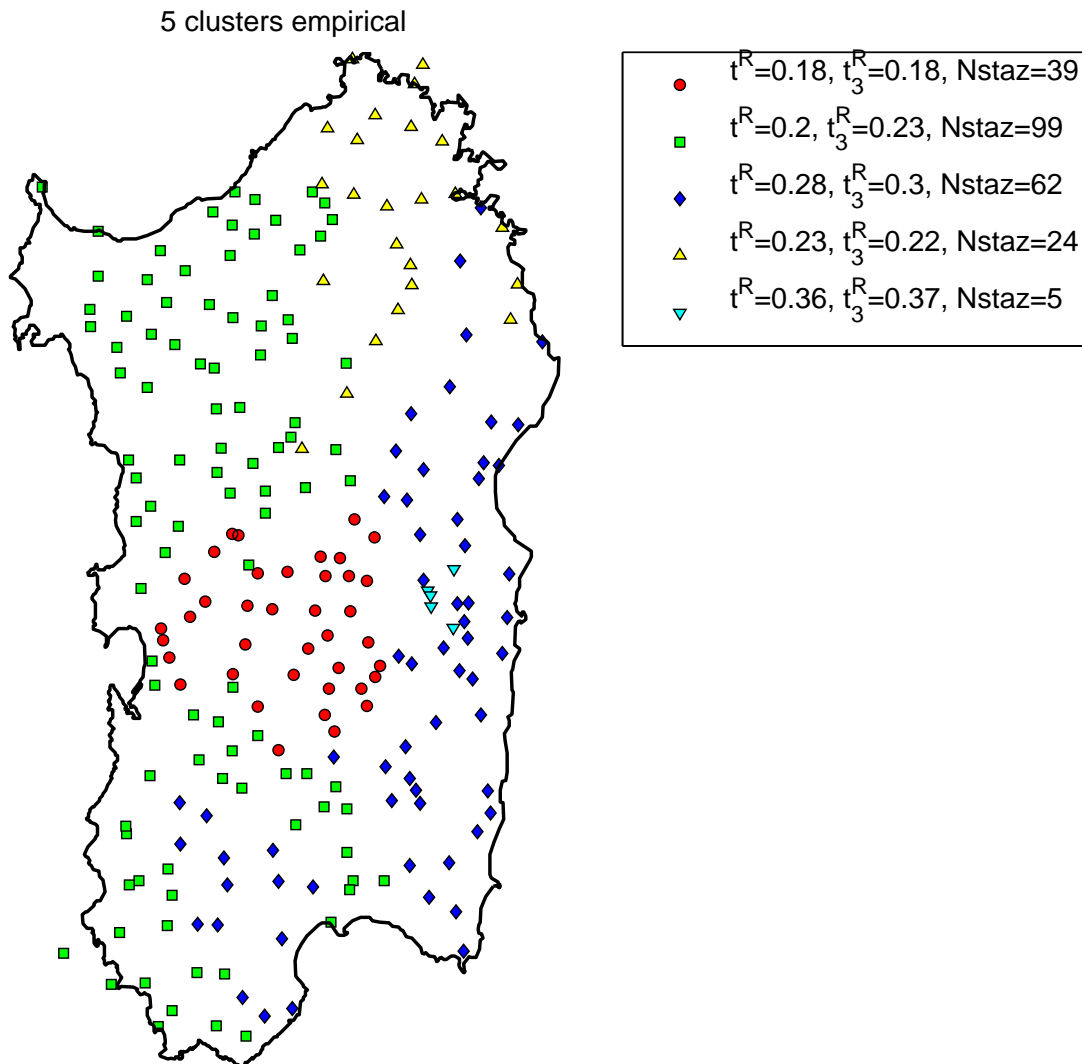


Figure 7.29: **Hypothesis E**: Spatial distribution of the 5 homogeneous regions obtained from hypotheses A and C through empirical aggregation. The legend shows the regional values of the L-moment ratios t^R, t_3^R and the number of stations for each cluster.

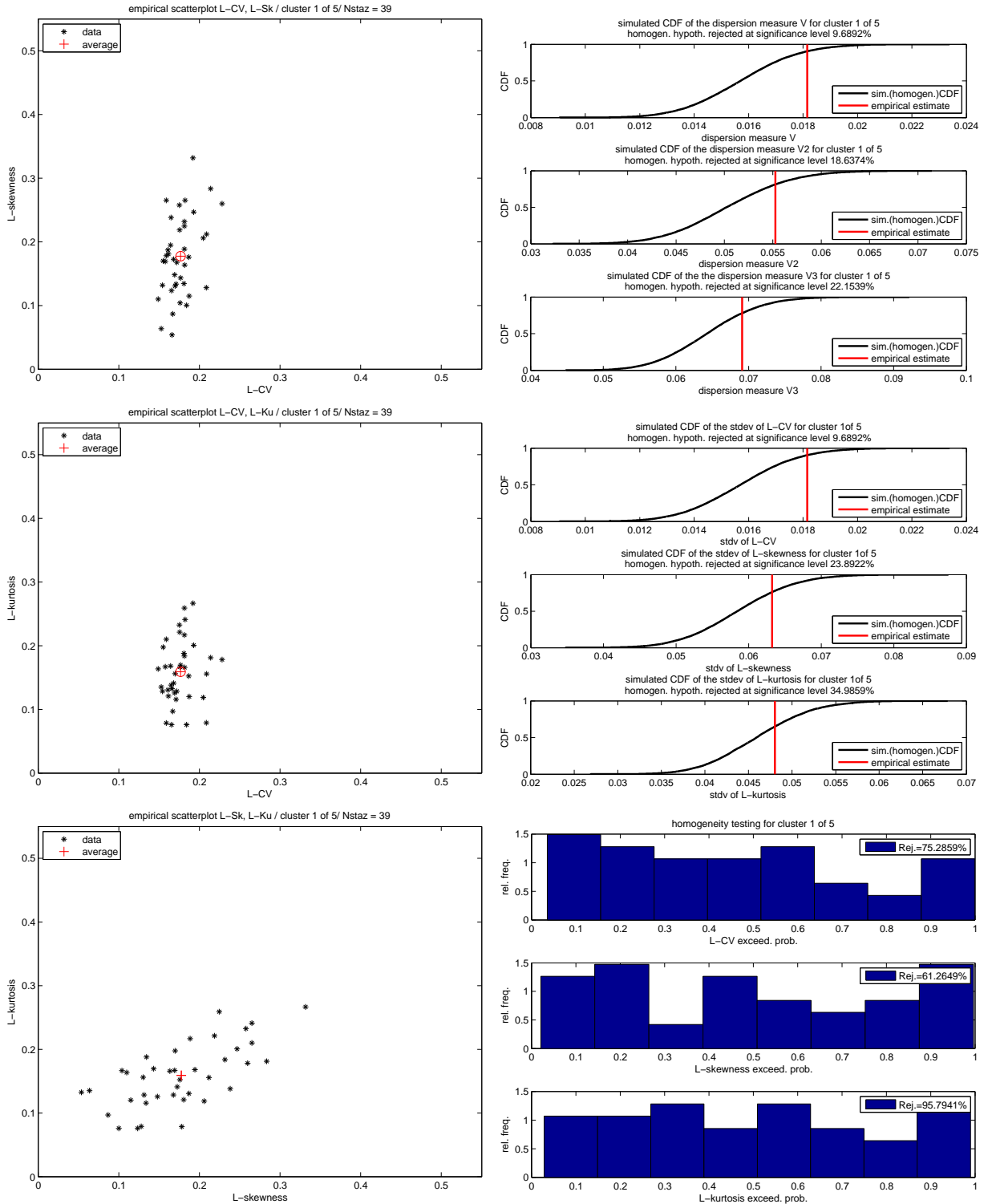


Figure 7.30: Hypothesis E: Cluster E_1
 Same description of the Figures 7.8 and 7.9

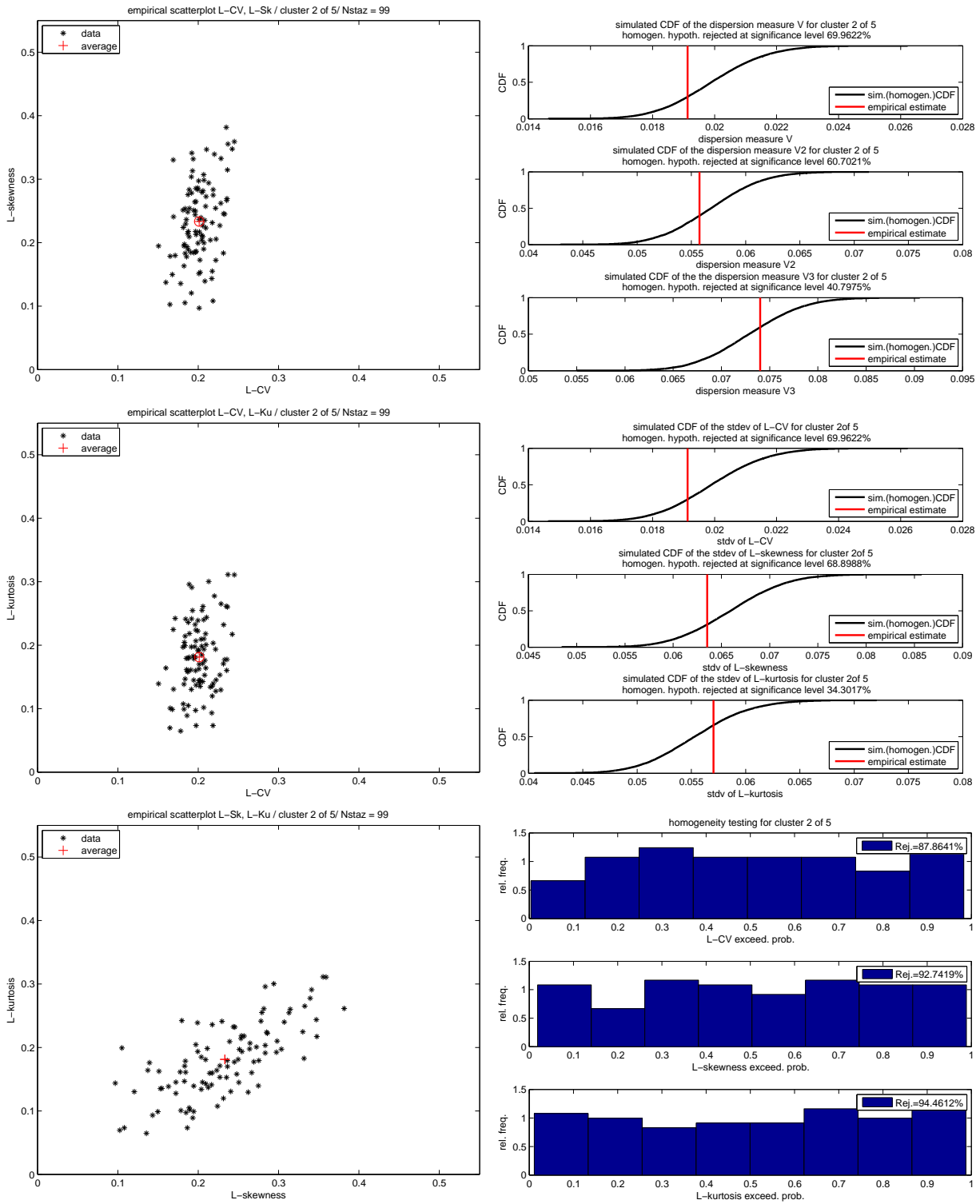


Figure 7.31: Hypothesis E: Cluster E₂
 Same description of the Figures 7.8 and 7.9

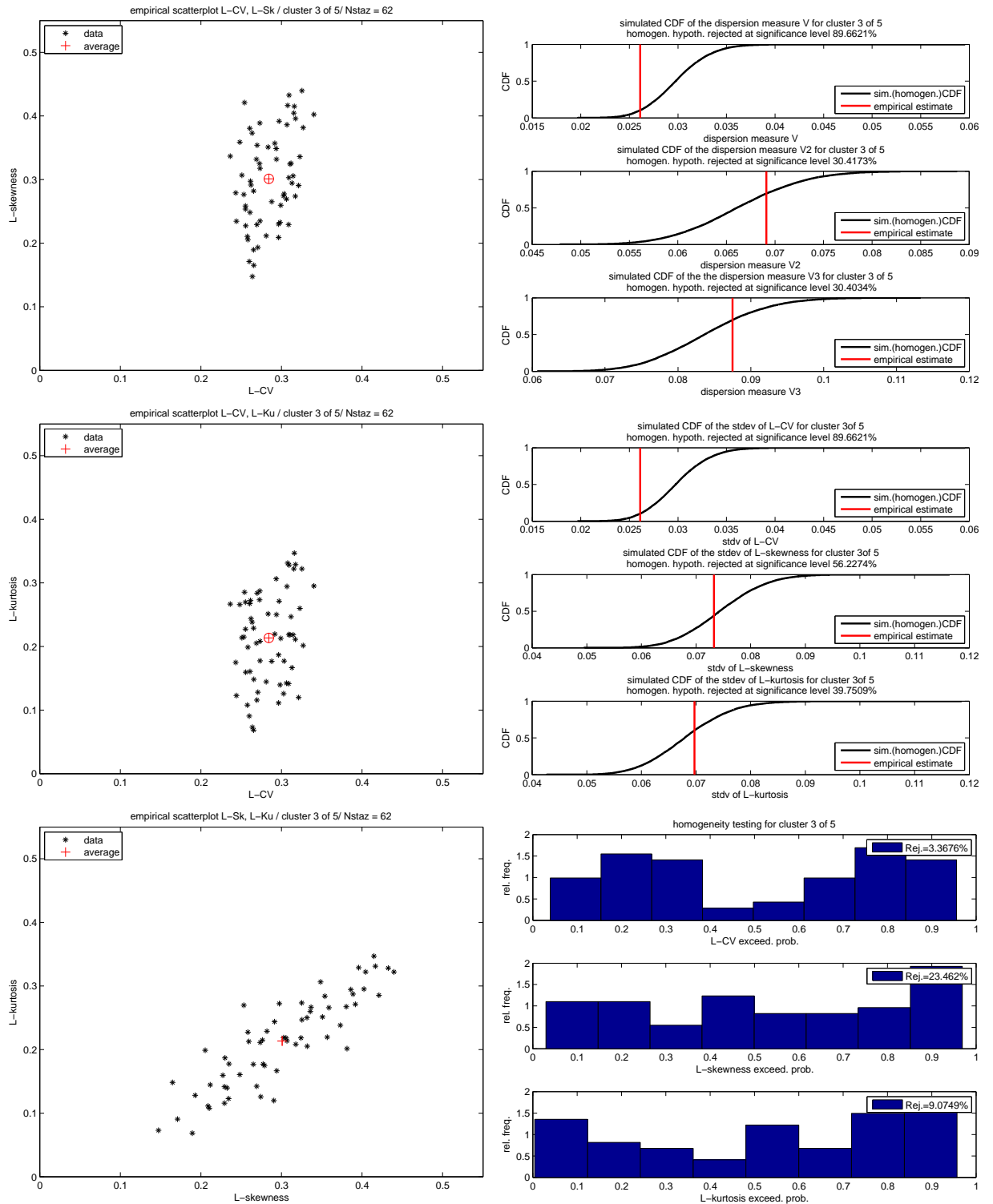


Figure 7.32: Hypothesis E: Cluster E_3
 Same description of the Figures 7.8 and 7.9

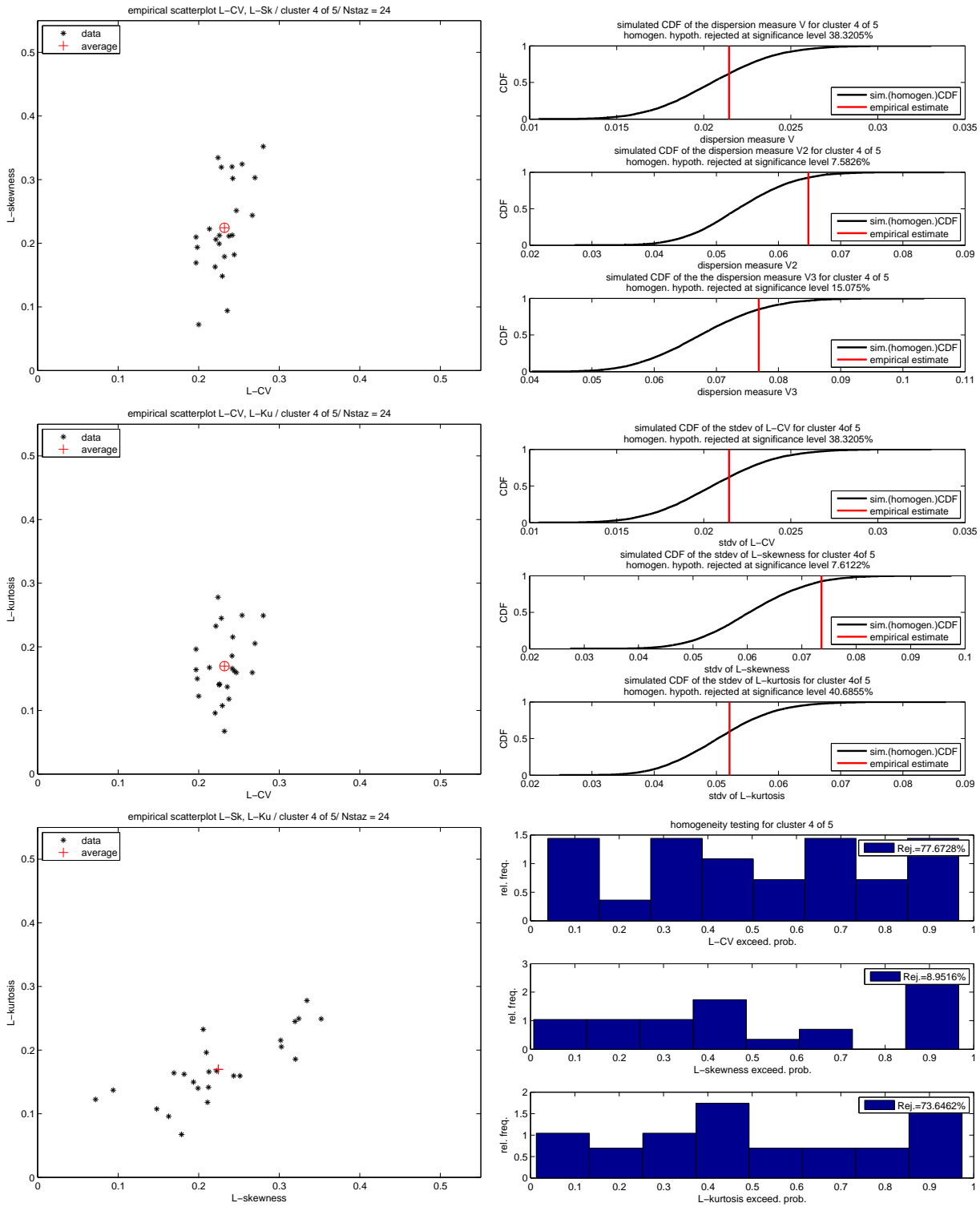


Figure 7.33: Hypothesis E: Cluster E₄
 Same description of the Figures 7.8 and 7.9

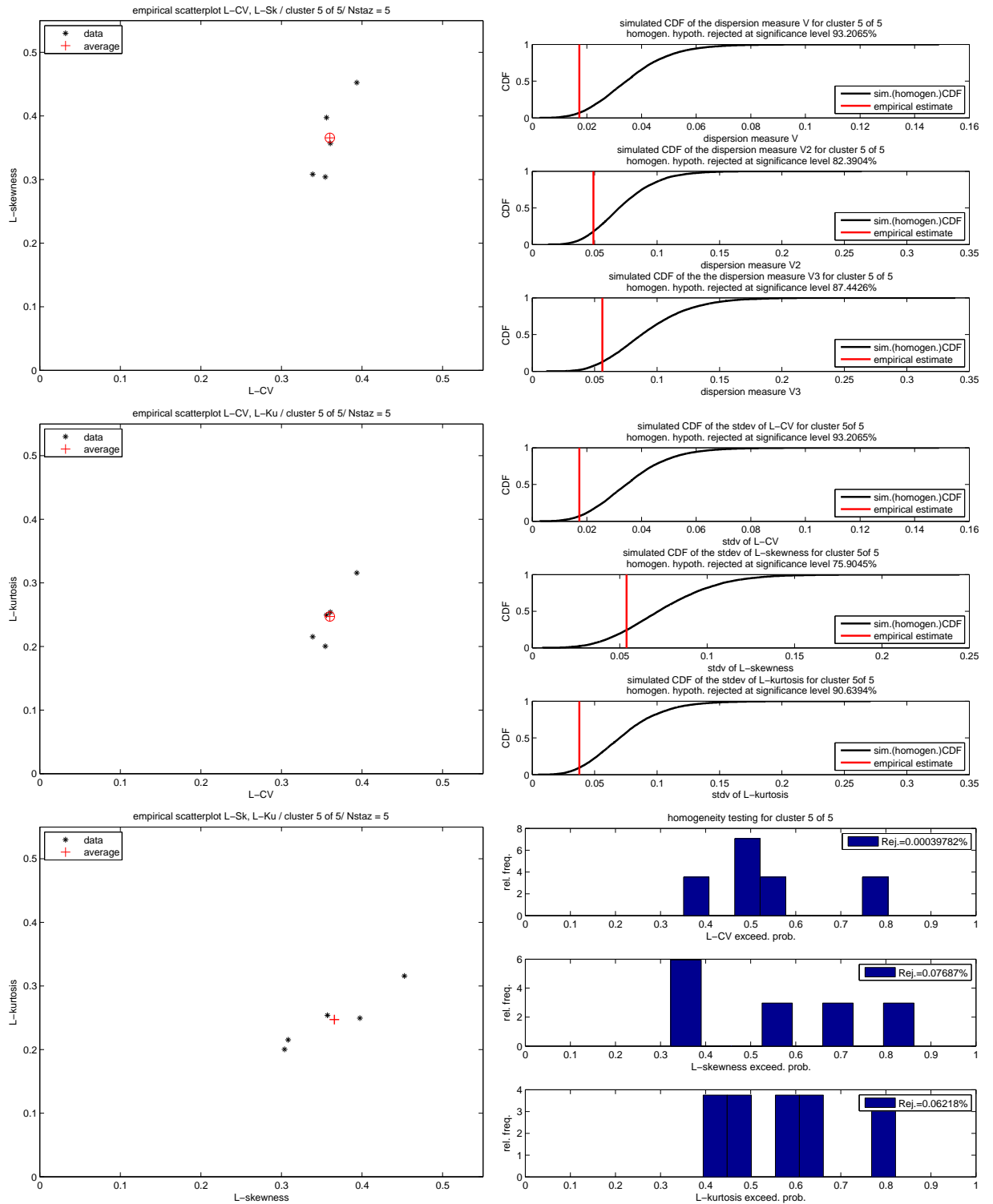


Figure 7.34: Hypothesis E: Cluster E₅
 Same description of the Figures 7.8 and 7.9

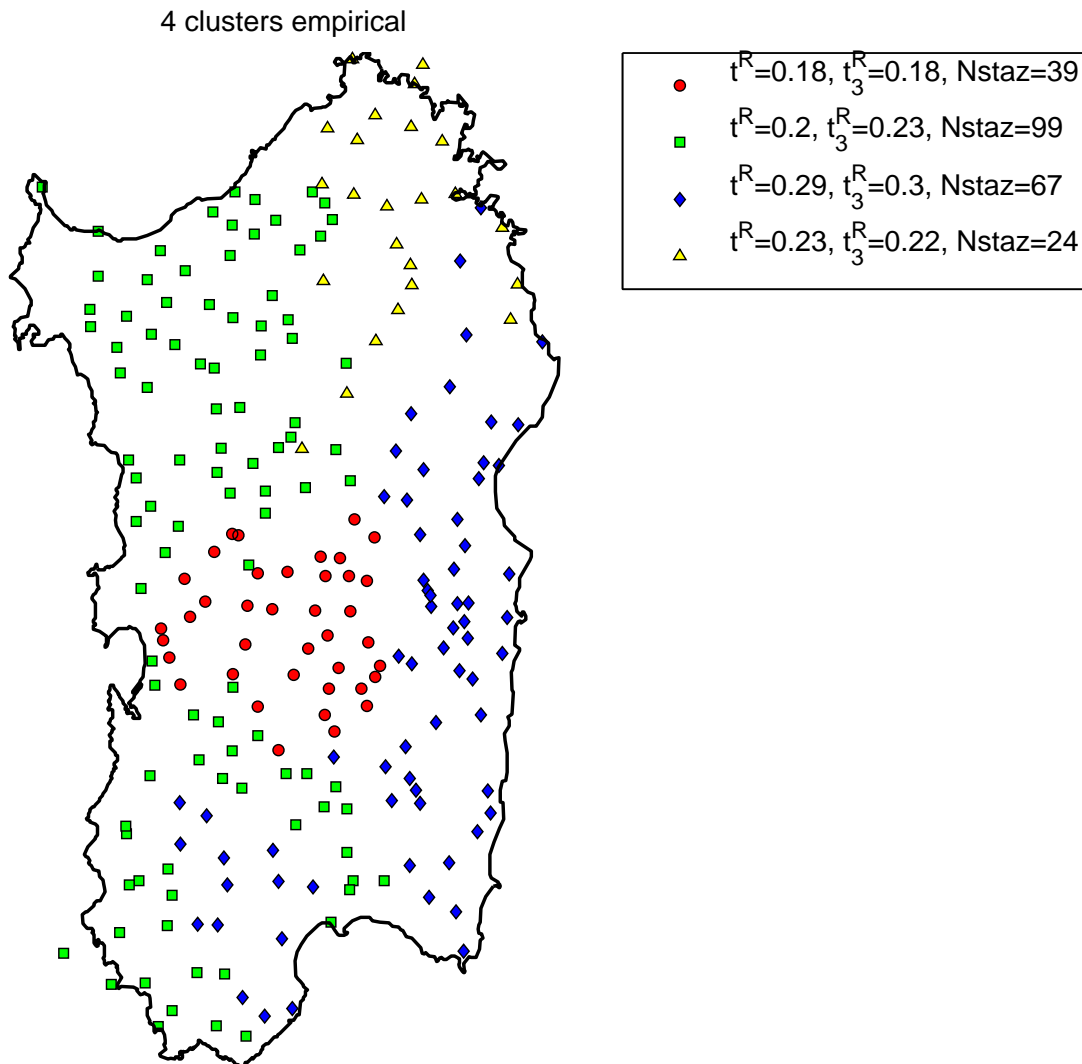


Figure 7.35: **Hypothesis F**: Spatial distribution of the 4 homogeneous regions obtained from hypotheses B and D through empirical aggregation. The legend shows the regional values of the L-moment ratios t^R, t_3^R and the number of stations for each cluster.

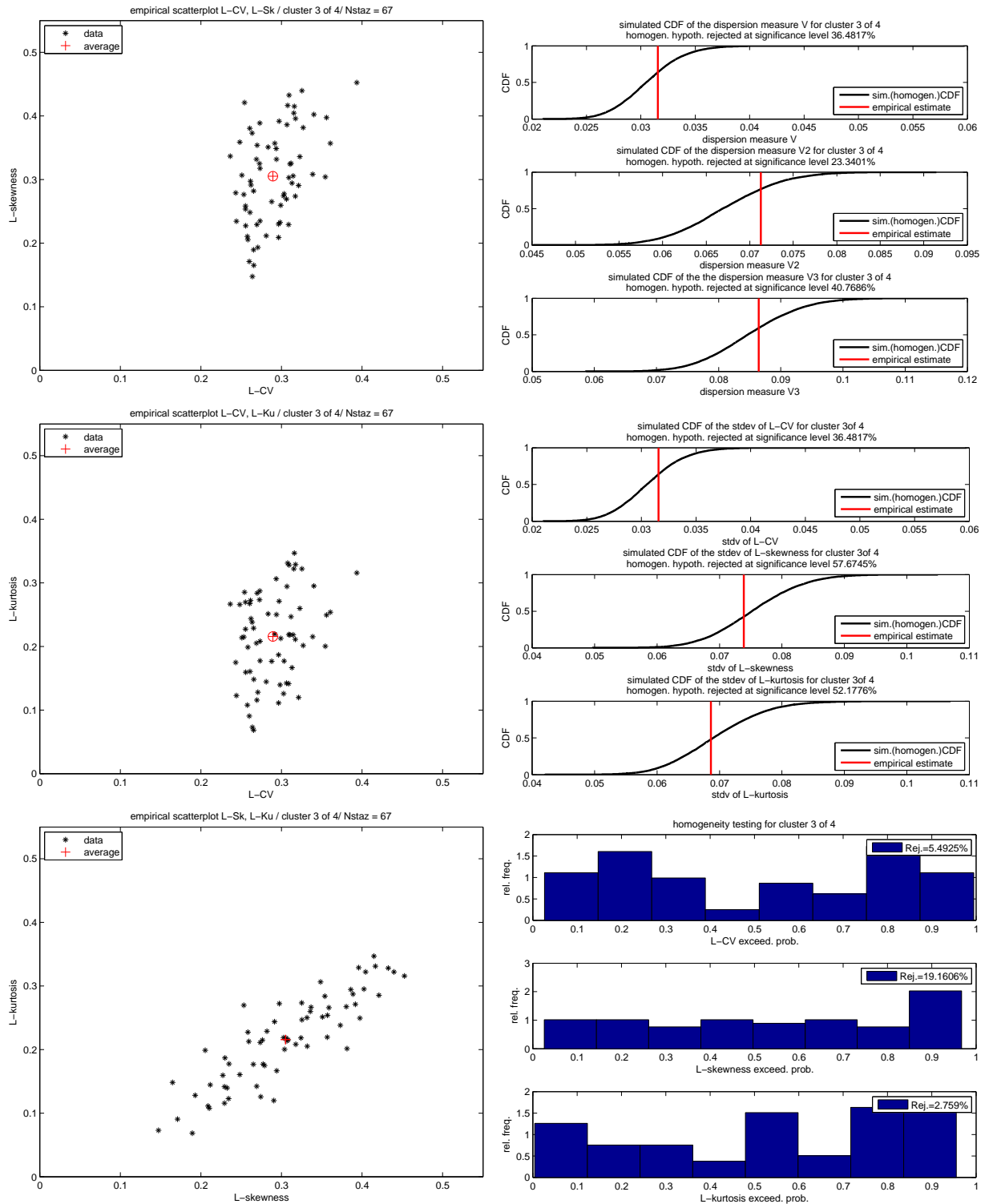


Figure 7.36: Hypothesis F: Cluster F_3
 Same description of the Figures 7.8 and 7.9

7.2.4 Comparison between regional configurations

Figure 7.37 shows the L-moment ratios diagrams (Hosking, 1990) for each one of the six hypothesis of partition in homogeneous regions described in the previous section. Each different color according to the cluster allocation described in Figures 7.11, 7.17, 7.21, 7.27, 7.29 and 7.35. The line that links the possible theoretical couples relative to the GEV distribution is the most barycentric and interpolating among the considered ones, for each cluster of each of the six hypothesis of partition in homogeneous regions. The point which corresponds to regional statistics (t_3^R, t_4^R) calculated through equations 5.13 and 5.14 is reported for each cluster. In this case, the sample error is reduced and the point lies on the line relative to the GEV distribution. These results suggest the use the GEV distribution for each cluster of each of the six hypothesis of partition in homogeneous regions.

Dimensionless parameters of the regional GEV growth curve were estimated adopting the regional PWM estimators mentioned in paragraph 5.1.2. In particular, regional L-moment ratios were calculated with equations (5.12-5.14) and parameters κ , σ^* and μ^* with equations (4.16), (5.9) and (5.10). The estimates of these parameters are shown in Table 7.12 for each homogeneous region of the six configuration. Values of σ and μ were obtained multiplying σ^* and μ^* by the index-rainfall.

Analysing Table 7.12 and the Figures 7.11, 7.17, 7.21, 7.27, 7.29 and 7.35 we noted that the first cluster of each configuration ($\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1, \mathbf{D}_1, \mathbf{E}_1, \mathbf{F}_1$) is always characterized by values of κ very close to zero and by the lowest σ^* values. This means that the first cluster, that goes from the center of Sardinia towards the west, is characterized by a GEV distributions that degenerates to a Gumbel distribution. The highest values of the parameters κ and σ^* are in the clusters in the east and south-east zone of the island, which is characterized by the highest events. High values of the shape and scale parameter mean that the distributions has a heavy right tail. So the quantile grows more quickly respect to the Gumbel distribution, with the same-exceeding probability.

A synthesis of the performances of regional GEV fits with the different hypothesis of partition in homogeneous regions is presented in Table 7.13. The table reports the mean of the error metrics described in section 6, evaluated on the 229 stations with more than 50 complete years of observations. The regional GEV fits are reported and compared with the regional TCEV fits (with the same local index-rainfall in “TCEV-2008”). The local case clearly presents values of the error metrics lower than those in the regional approach. Among them, error metrics evaluated on the GEV model fit the observed data better than the TCEV model (also in the “TCEV-2008” case). In particular,

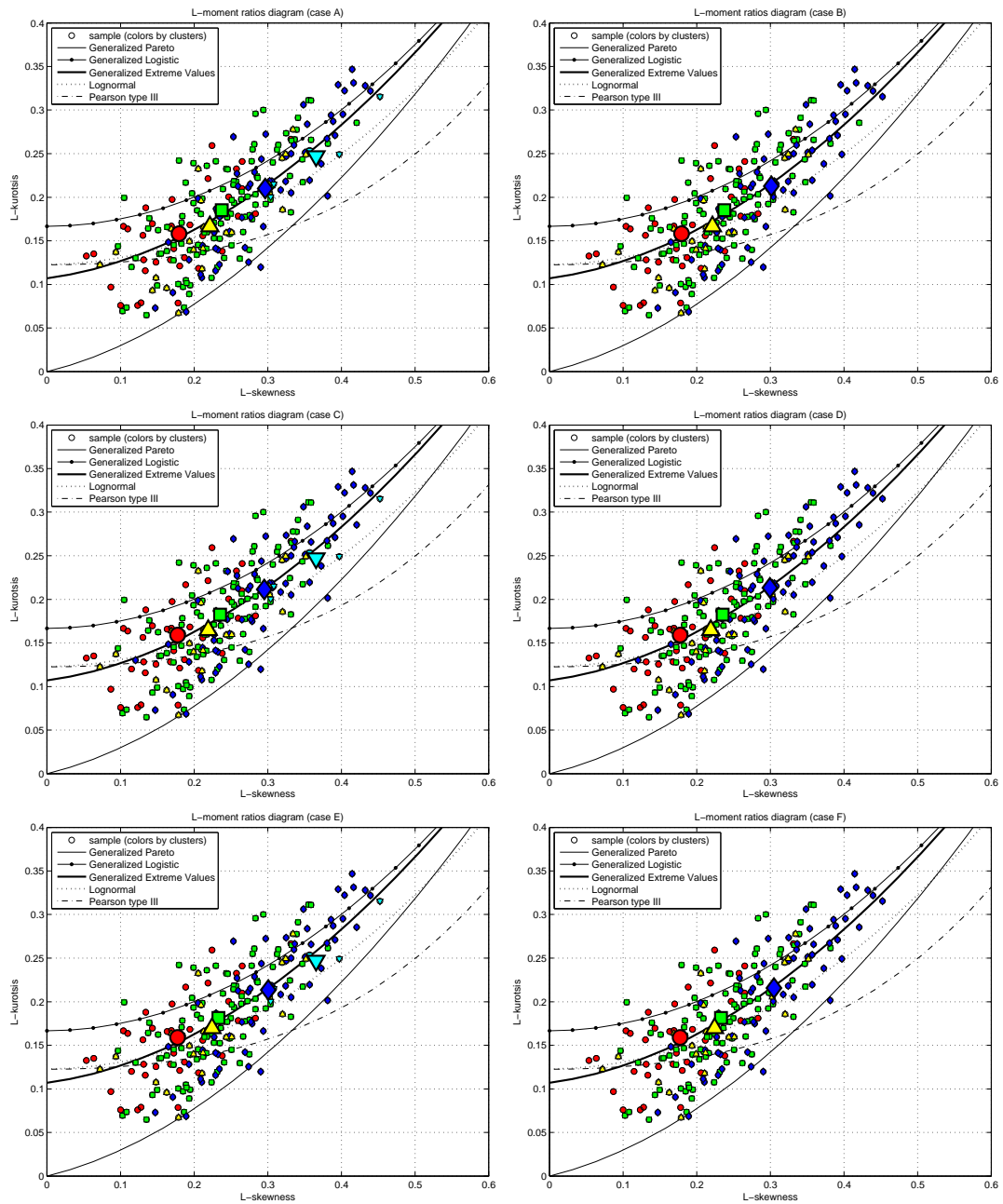


Figure 7.37: Comparison between pairs of L-moment ratios (L-skewness, L-kurtosis) calculated on annual maximum daily precipitation for the 229 stations with more than 50 complete years in record (circles), partitioned according to the cluster allocation (symbols of different color for each cluster, consistent with Figures 7.11, 7.17, 7.21, 7.27, 7.29 and 7.35) and theoretical pairs for some distributions widely used in statistical hydrology, represented by lines of different strokes. Big marks denote, for each cluster, the regional values of the same statistics. From top to bottom and from left to right are reported cases related to the hypothesis **A**, **B**, **C**, **D**, **E** ed **F** of division into homogeneous regions.

cluster	κ	σ^*	μ^*
A₁	0.015	0.251	0.852
A₂	0.103	0.264	0.818
A₃	0.189	0.338	0.728
A₄	0.079	0.309	0.796
A₅	0.284	0.368	0.645
B₁	0.015	0.251	0.852
B₂	0.103	0.264	0.818
B₃	0.196	0.341	0.723
B₄	0.079	0.309	0.796
C₁	0.012	0.252	0.851
C₂	0.100	0.262	0.821
C₃	0.187	0.330	0.736
C₄	0.076	0.312	0.796
C₅	0.284	0.369	0.645
D₁	0.012	0.252	0.851
D₂	0.100	0.262	0.821
D₃	0.193	0.333	0.730
D₄	0.076	0.311	0.796
E₁	0.012	0.252	0.851
E₂	0.097	0.263	0.820
E₃	0.195	0.331	0.731
E₄	0.083	0.308	0.795
E₅	0.284	0.369	0.645
F₁	0.012	0.252	0.851
F₂	0.097	0.263	0.820
F₃	0.201	0.334	0.726
F₄	0.083	0.308	0.795

Table 7.12: Estimation of GEV growth curve parameters (κ , σ^* and μ^*) for every cluster of each of the six hypothesis of partition in homogeneous regions.

regional GEV fit better both the whole set of observed data (metrics A^2 and W^2) and the 5 highest observed values for each station (metrics $MAE(5)$ and $MAEr(5)$).

Between the 6 hypotheses of partition in homogeneous zones, the ones which presents the best results using the regional GEV estimation are configurations **E** and **F**, (5 and 4 empirical clusters respectively). The difference in the average relative error (metric $MAEr$) between this two hypotheses is very small, so the choice between them can be based on other considerations. If fact, if one is interested in use the smallest number of regions that can be considered homogeneous configuration **F** represent the best choice. Instead if one is more interested in taking into account the local peculiarity of cluster **E₅** configuration **E** must be considered.

	MAE(5) [mm]	MAEr(5) [-]	A ² [-]	W ² [-]
GEV local (PWM)	10.496	0.070	0.315	0.047
TCEV-2008	16.762	0.115	0.922	0.138
TCEV-1980	18.524	0.132	1.521	0.244
GEV reg. (A: 5cl L-CV)	15.045	0.099	0.682	0.103
GEV reg. (C: 5cl L-CV L-Sk)	14.707	0.098	0.689	0.104
GEV reg. (E: 5cl empirical)	14.564	0.096	0.670	0.101
GEV reg. (B: 4cl L-CV)	15.619	0.101	0.714	0.108
GEV reg. (D: 4cl L-CV L-Sk)	15.369	0.100	0.731	0.109
GEV reg. (F: 4cl empirical)	15.172	0.098	0.708	0.106

Table 7.13: Comparison between the performances of regional fits with the GEV distribution (with parameters estimated with PWM method) and regional fits with the TCEV distribution. Mean of error metrics calculated over the 229 stations with at least 50 complete years of observations.

7.3 Geostatistical analysis results

In order to overcome the limitation of the regional approach, (see the introduction, section 1) the opportunity to represent with continuity the spatial distribution of the GEV growth curve parameters was investigated. For this purpose, the kriging technique was utilized, see section 5.2. In detail, interpolations were done for the shape parameter κ , the dimensionless scale parameter σ^* and the index-rainfall m . Once these quantities are known, it is possible to estimate the dimensionless position parameter μ^* from equation (5.10), and the scale and position parameters (σ and μ) of the GEV canonic form utilizing transformations reported in equations (5.3).

As already reaffirmed, the ordinary kriging (OK) exactly reproduces the observed values in the measurement points. But we knew a priori that the estimations of the parameters we wanted to interpolate are affected by remarkable uncertainty linked to estimators variance in small samples. For this reason we used kriging for uncertain data (KUD), that determines a smoothing of the interpolating surface, coherently with estimator variance in each point of observation. This choice was supported by theoretical reasons and by results of a preliminary analysis, which showed improvements when the KUD is used instead of the OK in the spatial interpolation of the GEV dimensionless parameters κ , σ^* and the index-rainfall m . These preliminary analysis showed that the optimal number of adjacent stations in order to write the liner equations systems of kriging was equal to 9 and that the more plausible variogram was the exponential one. We observed also that if the

number of adjacent stations exceed the value of $8 \div 9$ KUD perform worse than the OK when applied to estimates κ and m . Moreover, the minimum optimal number of years of observation to select the historic series for the spatial interpolations was searched too. In fact, a low threshold permits to utilize a larger number of stations, so to better describe spatial trends, but the estimations on the stations with few years of observation are affected by big errors. On the other hand, increasing the threshold provides more precise estimations but the reduction of the number of stations can cause the loss of some local peculiarities. We definitely decided to use all the 229 stations with at least 50 complete years of observations for the interpolations of parameters k and σ^* , and all the 256 stations with at least 30 complete years of observations for interpolate the index-rainfall m , which is affected by lower sample uncertainty. The preliminary analysis that lead the just described choices were conducted with the cross-validation procedure described in section 6.3. In particular, the value of the considered parameter in the point of measure of each station was iteratively interpolated, excluding from time to time the parameter estimation coming from the same station. For each aspect we analyzed (such as the optimal number of stations for the kriging system, the minimum length of series we need to consider, the estimation method based on ordinary kriging or on kriging for uncertain data, and so on), the overall performances were evaluated calculating the mean absolute error between the value of the locally estimated parameter and that interpolated with the cross-validation procedure. A synthesis of the results of this preliminary analysis is given in Figures 7.38, 7.39 and 7.40 for each of the three parameters of interest.

Considering the results of these preliminary analyses, we decided to use the KUD for the interpolation of each parameter, on a regular grid with spatial step equal to 1 kilometer. That grid covers the whole territory of Sardinia. We utilized the estimations in the nearest 9 stations and the exponential variogram. Despite the implementation of KUD, we sporadically observed noise around some grid points that could interrupt the local monotonicity of parameters' spatial variations. An accurate analysis permitted to understand that these effects could be originated by the use of a predetermined number of nearest stations in the interpolations. Thus, moving between adjacent points of the grid, one or more stations used in the estimations can be substituted by other nearer stations. We eliminated these small disturbances applying a simple moving average with a 9×9 km window. In other words, the interpolated value in each point of the grid has been replaced by the average of the interpolated values in the point and in the setting of points around it, with a distance of 1, 2, 3 and 4 km. The numerical comparison between the grids obtained through the kriging and those obtained after the moving average

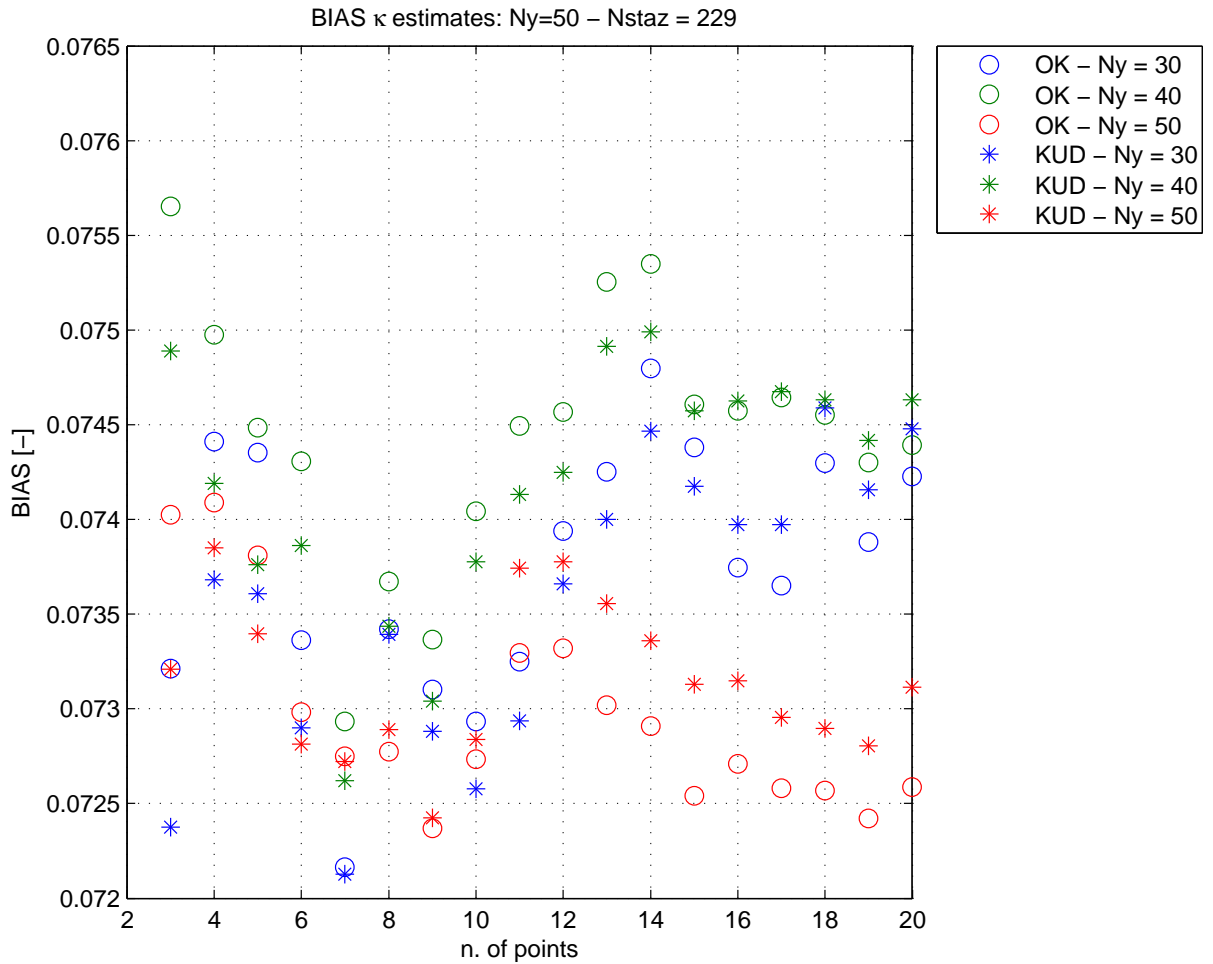


Figure 7.38: Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the shape parameter κ . The ordinate shows the mean absolute error (MAE) in the interpolation of the parameter κ (using cross-validation) in each of the 229 stations with at least 50 complete years of observations (excluding time to time the estimate of the station considered), in function of the number of nearest stations (on the abscissa) used for the kriging system. The empty circles represent the results with the OK, the asterisks refer to the KUD. The different colors refer to the minimum number of years to select the stations to be used for the interpolations.

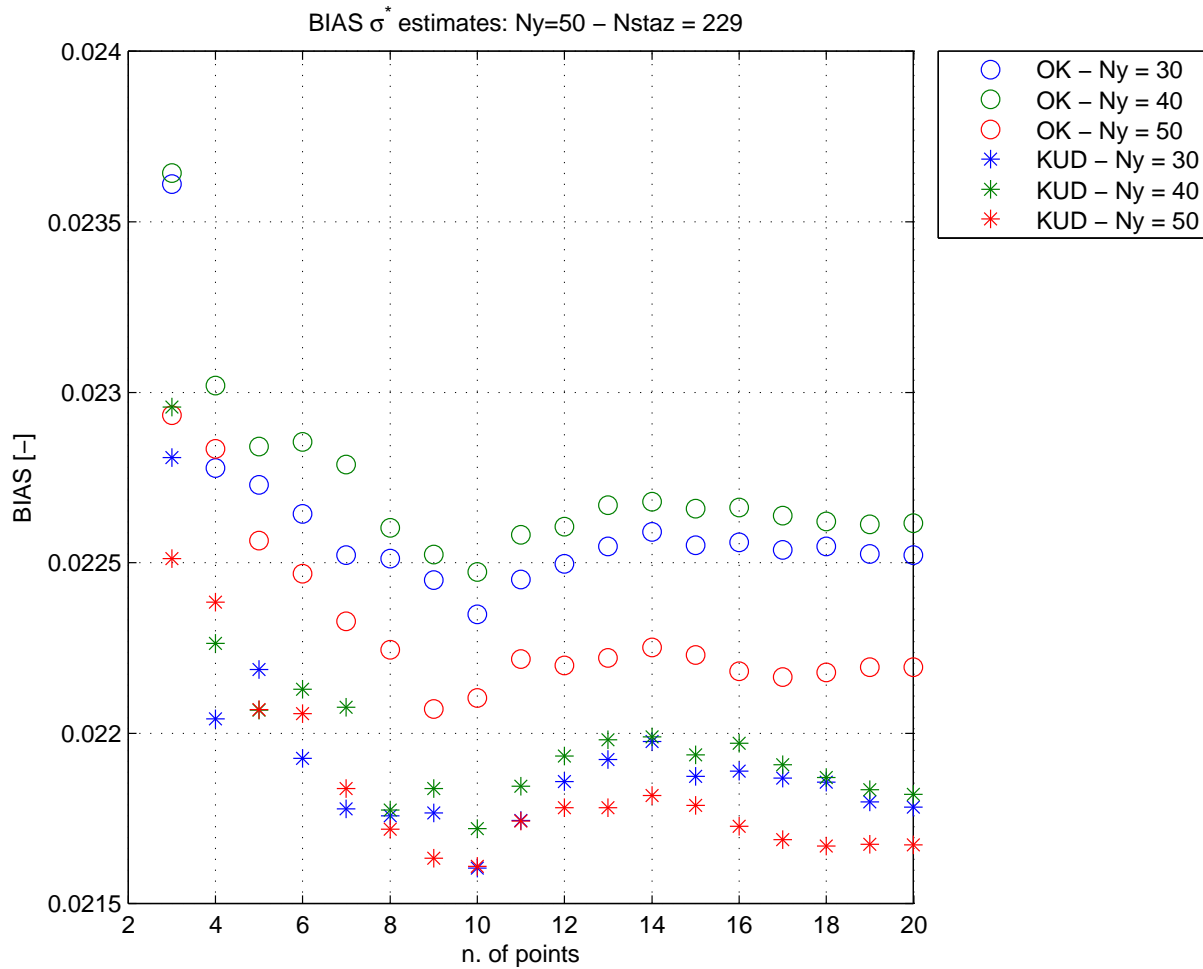


Figure 7.39: Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the dimensionless scale parameter σ^* . The symbolism is the same as for Figure 7.38, but *MAE* is calculated on the interpolations of the σ^* parameter using cross-validation.

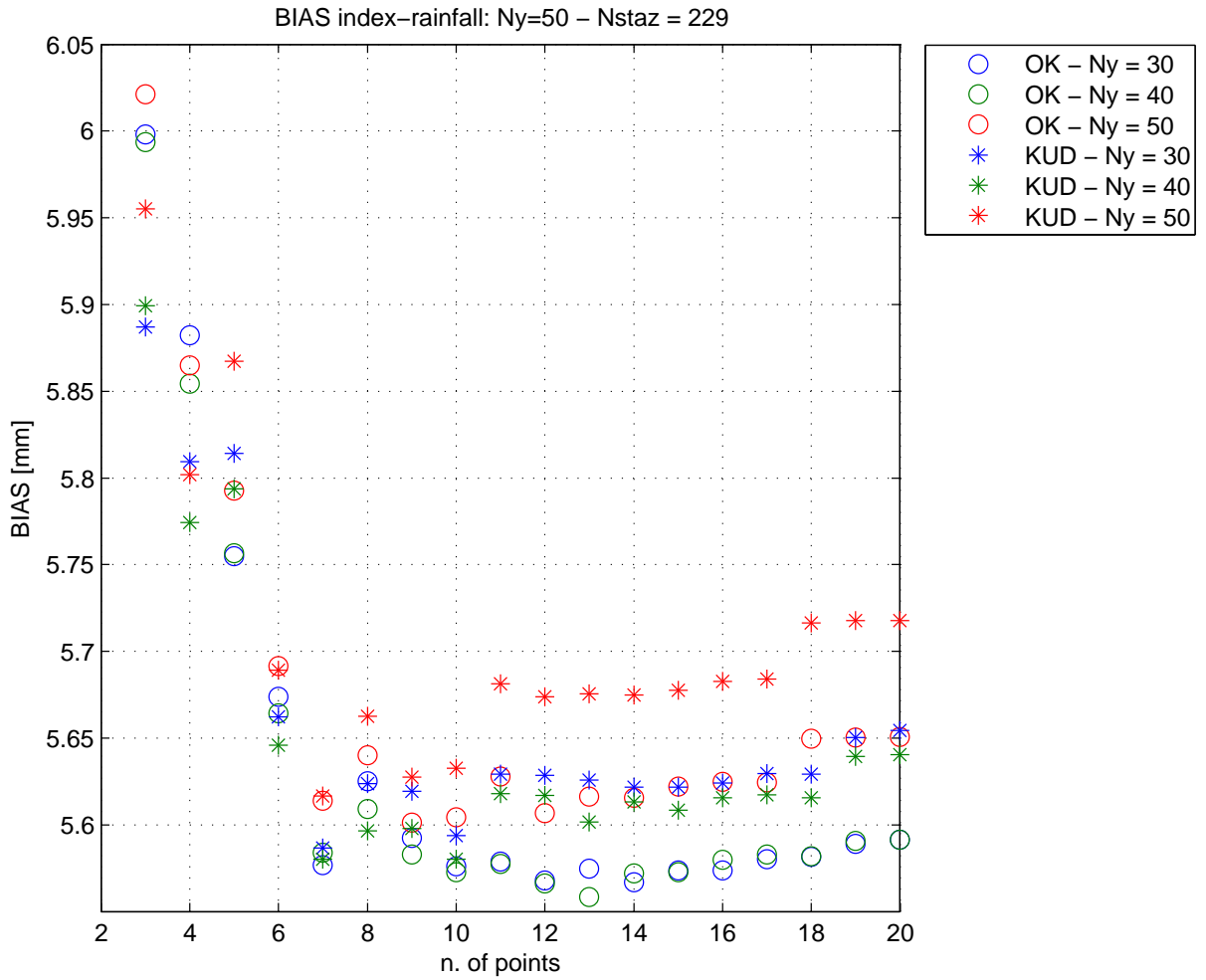


Figure 7.40: Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the index-rainfall m . The symbolism is the same as for Figure 7.38, but MAE is calculated on the interpolations of m using cross-validation.

highlights negligible variations in the peaks, whereas the visual comparisons do not show variations in the spatial trends, but only a benefit in terms of continuity in the distribution of parameters.

The determination of regular grids with 1 km resolution was made using the hierarchical procedure described in succession. The measures that needed to be interpolated with equation (5.22) were obtained with PWM estimators, at different estimation levels, whereas the measure error variances that appear in equations (5.48) have been determined with a Monte Carlo's procedure, hypothesizing that observed samples are realizations of a GEV distribution. The utilized procedure is the following:

- 1 For each station, GEV parameters were locally estimated.
- 2 For each station, 10'000 synthetic samples were generated from a GEV distribution using the local parameters estimated at point 1. Each sample had a length equal to the number of years of observation of the respective station. Hence, the GEV parameters were estimated again on each synthetic sample and their variance was estimated. The variance of GEV parameter κ was used in the KUD at further point.
- 3 A **regular grid** with spatial step of 1 km with the values of the shape parameter κ was created. This grid was obtained applying KUD to interpolate local estimations of parameter κ , obtained at point 1 (for the 229 stations with at least 50 complete years of observations) which were characterized by the estimation error variance obtained at point 2. Comparing several variograms, represented in Figure 7.41 (top), we decided to adopt the exponential model. At a later stage, a spatial moving average with a moving 9x9 sub-grid was applied. The final result is presented in Figure 7.42, which shows the spatial distribution of the shape parameter κ .
- 4 For each station we estimated the GEV scale and position parameters, σ and μ , conditioned to the values of κ obtained from the regular grid described at point 3 through a bilinear interpolation between the four points that contain the considered station. Dividing this estimates for the local mean m (index-rainfall), dimensionless parameters σ^* and μ^* were obtained from equations (5.3).
- 5 For each station, 10'000 synthetic samples were generated from a GEV distribution using the local parameters estimated at point 4. Each sample had a length equal to the number of years of observation of the respective station. The GEV parameters, σ and μ , were calculated

again (with the constraint of the estimation of κ from the grid, as at point 4) for each synthetic sample and their variance was estimated. The variance of the parameter σ^* was used in the KUD at further point.

- 6 A **regular grid** with 1 km resolution, containing the values of dimensionless scale parameter σ^* was created. This grid was obtained applying KUD to interpolate local estimations of parameter σ^* , obtained at point 4 (for the 229 stations with at least 50 complete years of observations), which were characterized by the estimation error variance obtained at point 5. We choose to adopt the exponential model comparing several variograms represented in Figure 7.41 (middle). At a later stage, a spatial moving average with a moving 9x9 sub-grid was applied. The final result is presented in Figure 7.43, which shows the spatial distribution of the dimensionless scale parameter σ^* .
- 7 For each station, the values of parameters k and σ^* were obtained from the grids given at points 3 and 6, utilizing the bilinear interpolation between the four grid points that contain the considered station. The corresponding estimation of dimensionless position parameter μ^* was obtained from equation (5.10). GEV parameters, σ and μ , were calculated from equations (5.3) through the local mean (index-rainfall) m .
- 8 For each station, 10'000 synthetic samples were generated from a GEV distribution using the local parameters estimated at point 7. Each sample had a length equal to the number of years of observation of the respective station. The average (index-rainfall) of each synthetic sample was calculated and their variance was estimated. The variance of the index-rainfall was used in the KUD at further point.
- 9 A **regular grid** with 1 km resolution, characterized by the values of the daily index-rainfall m was obtained. This grid was created applying KUD to interpolate local estimations of index-rainfall calculated for the 256 stations with at least 30 complete years of observations, which were characterized by the estimation variance obtained at point 8. We choose to utilize the exponential model, comparing several variograms, as show in Figure 7.41 (bottom). After that, a spatial moving average with a moving 9x9 sub-grid was applied. The final result is presented in Figure 7.44, which shows the spatial distribution of the index-rainfall.

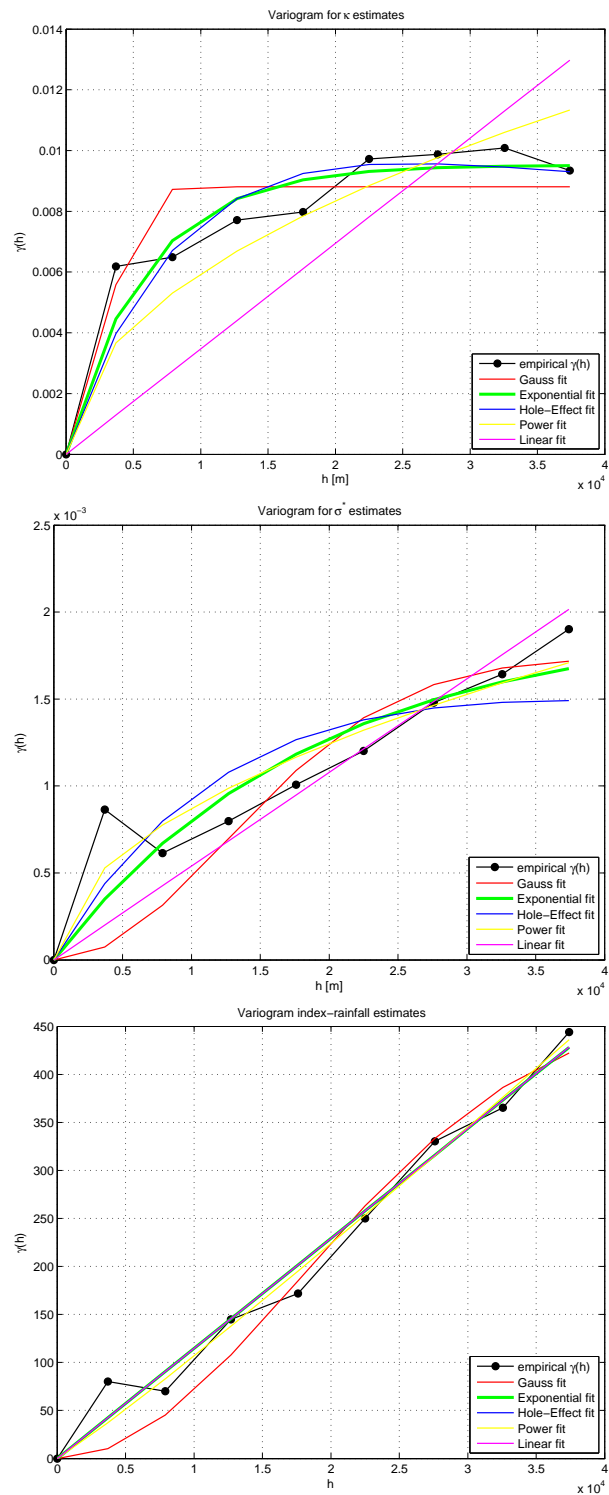


Figure 7.41: Comparison between sample variograms, based on local estimates of GEV growth curve parameters κ , σ^* , the index-rainfall m (from top to bottom), and some theoretical variograms.

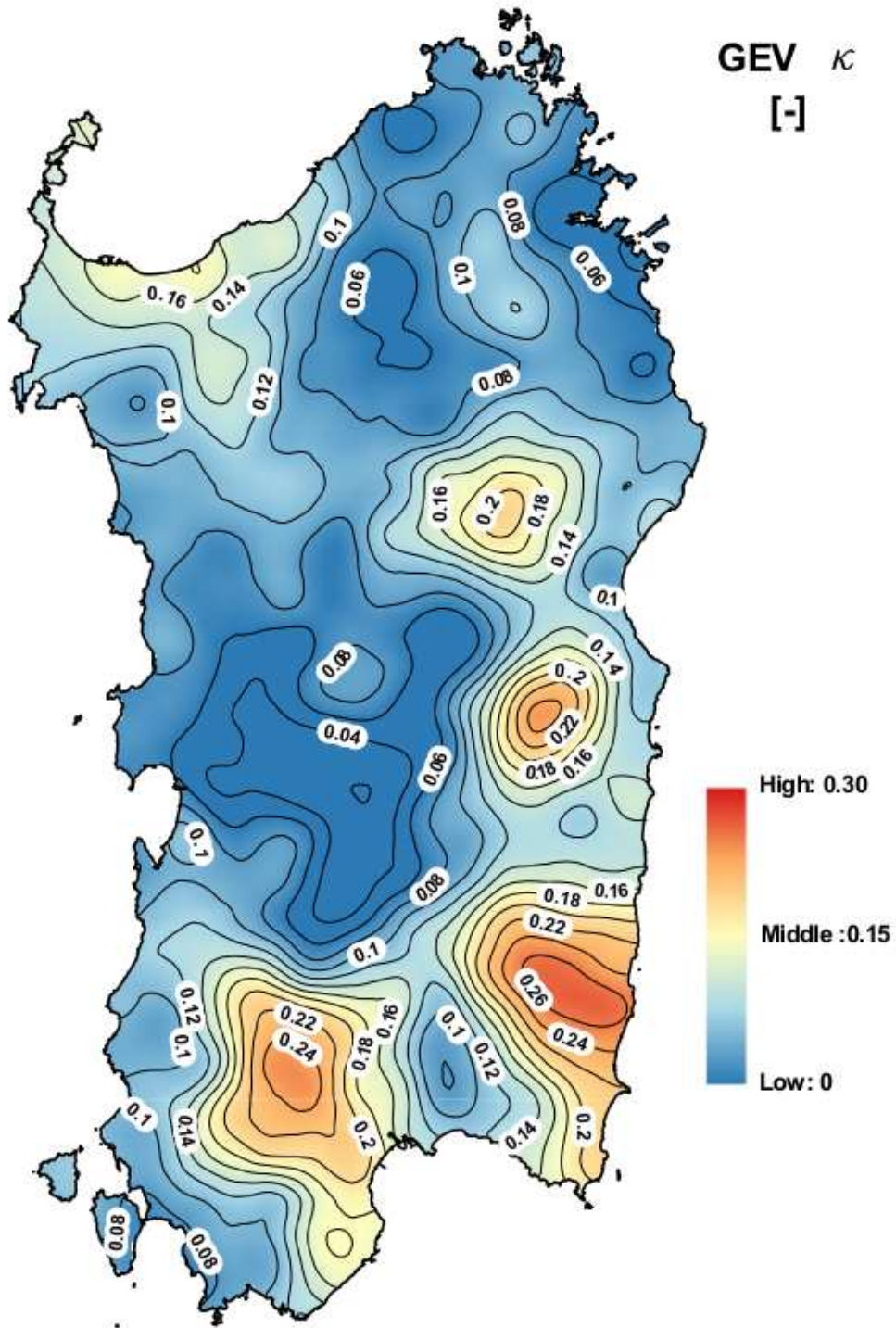


Figure 7.42: Representation of the spatial distribution of the shape parameter of the GEV distribution, κ . The map is obtained from a regular grid with 1 km resolution.

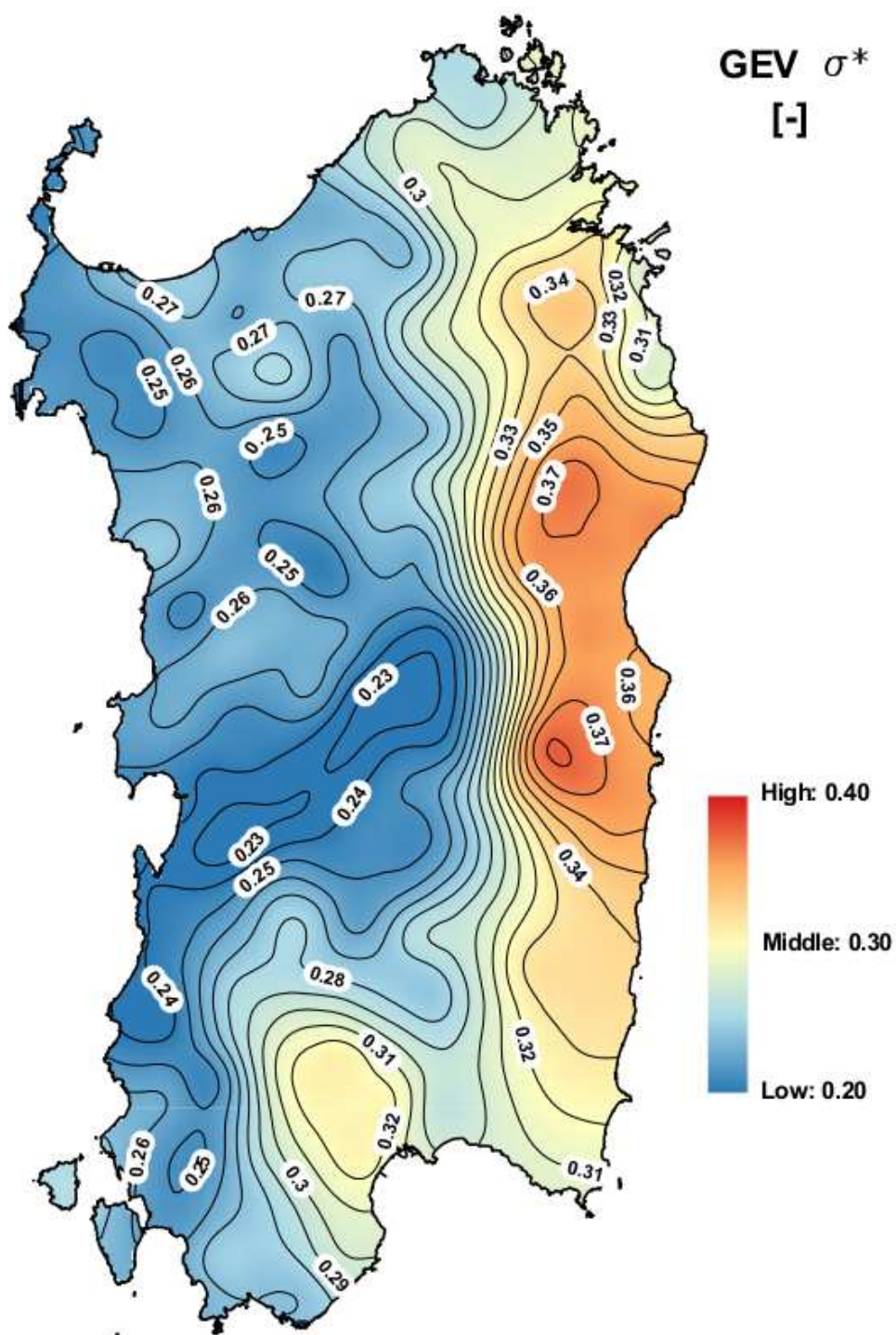


Figure 7.43: Representation of the spatial distribution of the dimensionless scale parameter of the GEV distribution, σ^* . The map is obtained from a regular grid with 1 km resolution.

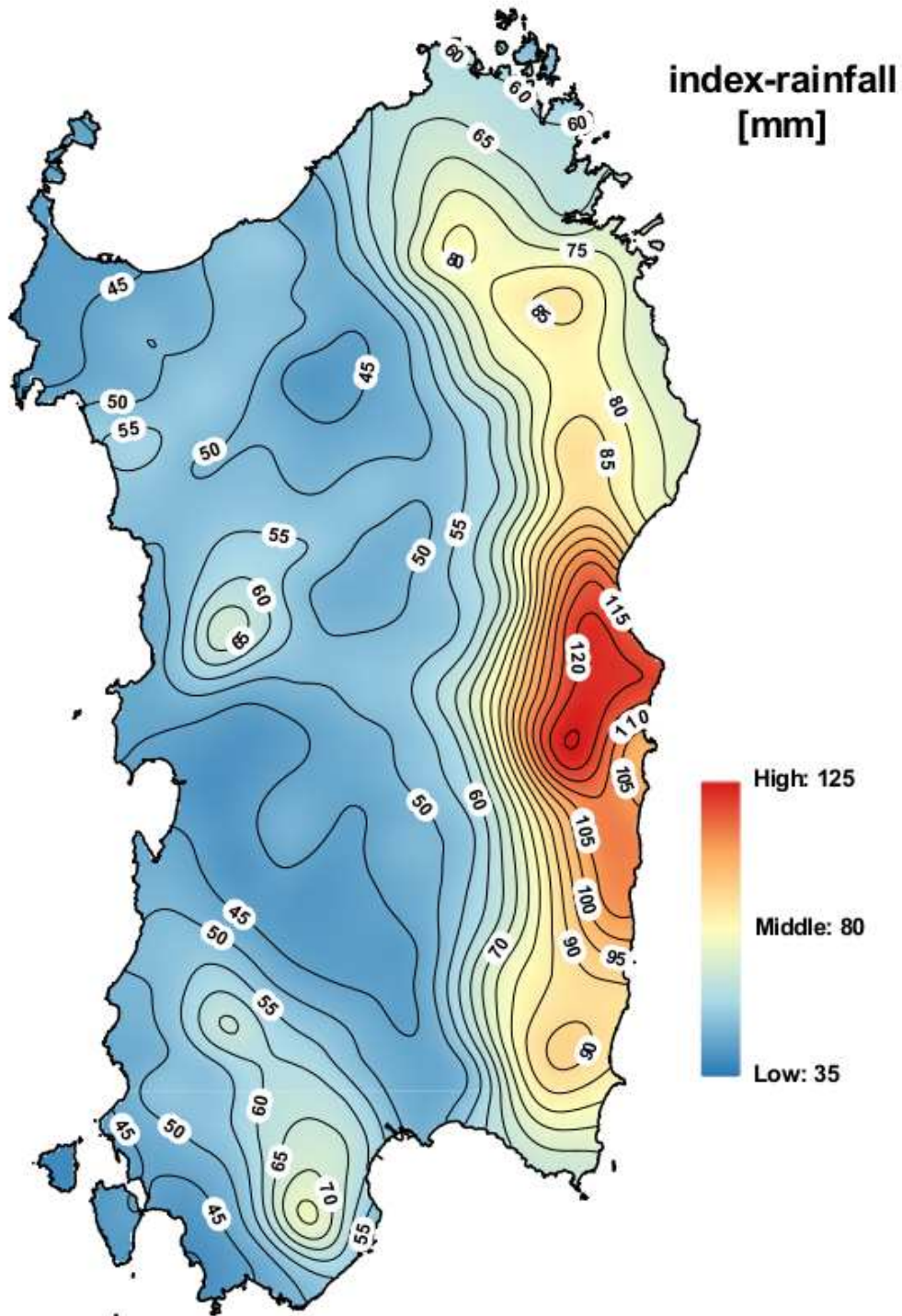


Figure 7.44: Representation of the spatial distribution of the index-rainfall, m . The map is obtained from a regular grid with 1 km resolution.

In order to show the difference between OK and KUD, in Figure 7.46 are reported the spatial interpolations of the shape parameter κ obtained using both techniques. The map on the left has been obtained using OK, the one on the right has been obtained using KUD. In both maps the moving average with a moving 9x9 sub-grid was applied. The figure highlights how the KUD determines a smoothing of the interpolating surface, coherently with estimator variance in each point of observation.

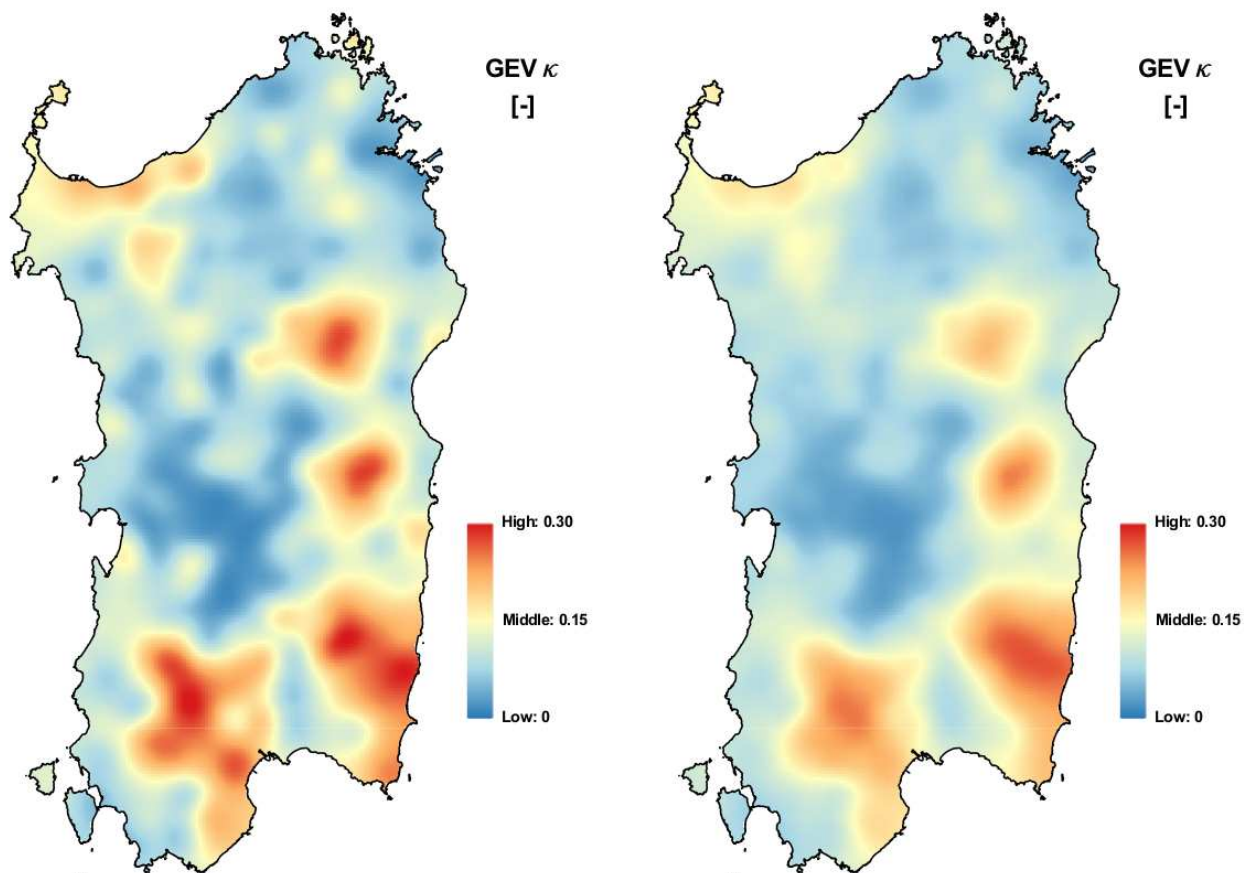


Figure 7.45: Representation of the spatial distribution of the shape parameter of the GEV distribution, κ . The map on the left is obtained using the ordinary kriging (OK), while the map on the right is obtained using kriging for uncertain data (KUD).

7.4 Regional and geostatistical comparison

In this section we compare the performance of local, regional and geostatistical fits. Regarding the geostatistical fits, for each station, the values of parameters κ and σ^* were obtained from grids with 1-km resolution, previously described, while the values of the corresponding parameter μ^* were obtained from equation (5.10). In order to make a fair comparison between the performances of the different approaches, the values of parameters σ and μ have been calculated from equation (5.3) using as index-rainfall the local mean instead of the grid value.

The results of the comparisons are reported in Attachment 1, where we can find, for each of the 229 stations with more than 50 complete years of observations, a comparison between the empirical CDFs and the theoretical ones obtained using: local estimates of GEV parameters, GEV regional estimates (hypotheses **E** and **F** of partition in homogeneous regions), estimates of GEV parameters from kriging grid, and regional estimations from “TCEV-2008” model. In the legend, for each station, are reported: the error metric $MEr(5)$ which gives a percentage estimation of how much the theoretical distribution overestimate or underestimate, in average, the 5 biggest observations, and the metrics A^2 and W^2 , which give indication about the fit of theoretical distributions to the whole set of observed data.

A synthetic comparison of the performances of the different approaches (utilizing the same locally estimated index-rainfall) is presented in Table 7.14, where average values of some error metrics described in section 6 are reported. From the results presented in Table 7.14, we can conclude that the best fit is obtained utilizing the GEV distribution with parameters interpolated from the kriging grid with 1-km step. The best performances are proved by the lowest values of all the considered error metrics.

In order to compare in a more realistic and significant way the several proposed approaches, the results presented in Table 7.14 were calculated again applying the cross-validation procedure mentioned in paragraph 6.3. However, before analyzing new results, it must be stated that the cross-validation procedure was fully implemented just in the geostatistical approach. This integral application of the cross-validation procedure was possible thanks to the complete automatization of all the phases of the hierarchical procedure of estimation with KUD. Regarding the regional estimates the cross-validation procedure has been partially applied because they need the preliminary grouping of stations in homogeneous regions. The determination of homogeneous regions need manual operations to unify clusters with similar characteristics and reduce the total number of aggregations. In the context of cross-validation, the computerization of these procedures, that need the

manual control (and sometimes the manual intervention), is difficult to obtain. So, in the case of GEV regional estimates, we hypothesized to ignore the homogeneous region of the temporarily removed station, and we attributed to it the same homogeneous region of the nearest station (and the parameters of the regional growth curve determined with all the useful stations). Consequently, for most of the stations, regional estimations were the same in the cross-validation procedure, except for few stations at the border between homogeneous regions that can be attributed to a group different from the original one. Despite this comparison in the cross-validation procedure is clearly penalizing for the geostatistical approach respect to the regional approach, the superiority of the first is confirmed, as reported in Table 7.15. The table shows the average values of the error metrics calculated for the 229 stations with at least 50 years of observation, using the cross-validation procedure.

Figure 7.46 illustrates the map of daily rainfall depth h_T (mm) exceeded with return period $T=200$ yr using the two different approaches. The map on the left of Figure 7.46 shows the result obtained using the regional approach (case **F** with 4 clusters). It can be noted how the abrupt discontinuity of the regional growth curve, in proximity of the borders between homogeneous regions, have a strong impact in the quantile estimates. In fact there are high “jumps” in the rainfall depth, in correspondence of the boundaries of the homogeneous regions. The map on the right shows the result obtained using the geostatistical approach. Is evident that the boundaries problem disappear and there is a better representation of the local peculiarities. The results reported in Table 7.15, and the analysis of Figure 7.46, lead to the conclusion of the supremacy of the geostatistical approach compared to the regional approach.

	MAE(5) [mm]	MAEr(5) [-]	A² [-]	W² [-]
GEV local (PWM)	10.496	0.070	0.315	0.047
TCEV-2008	16.762	0.115	0.922	0.138
GEV reg. (E: 5cl empirical)	14.564	0.096	0.670	0.101
GEV reg. (F: 4cl empirical)	15.172	0.098	0.708	0.106
GEV kriging	12.768	0.085	0.483	0.074

Table 7.14: Comparisons between the performances of local and regional fits using: local GEV distribution, regional fit with TCEV distribution, regional fits (hypotheses **E** and **F**) with GEV distribution and kriging fit. Average of error metrics calculated over the 229 stations with at least 50 complete years of observations.

	MAE(5) [mm]	MAEr(5) [-]	A² [-]	W² [-]
GEV local (PWM)	10.496	0.070	0.315	0.047
TCEV-2008	17.346	0.119	1.065	0.154
GEV reg. (E: 5cl empirici)	16.658	0.109	0.858	0.128
GEV reg. (F: 4cl empirici)	16.495	0.109	0.880	0.132
GEV kriging	15.351	0.103	0.688	0.105

Table 7.15: Same results presented in Table 7.14, but obtained with the **cross-validation** procedure for the regional GEV estimations and the kriging estimations.

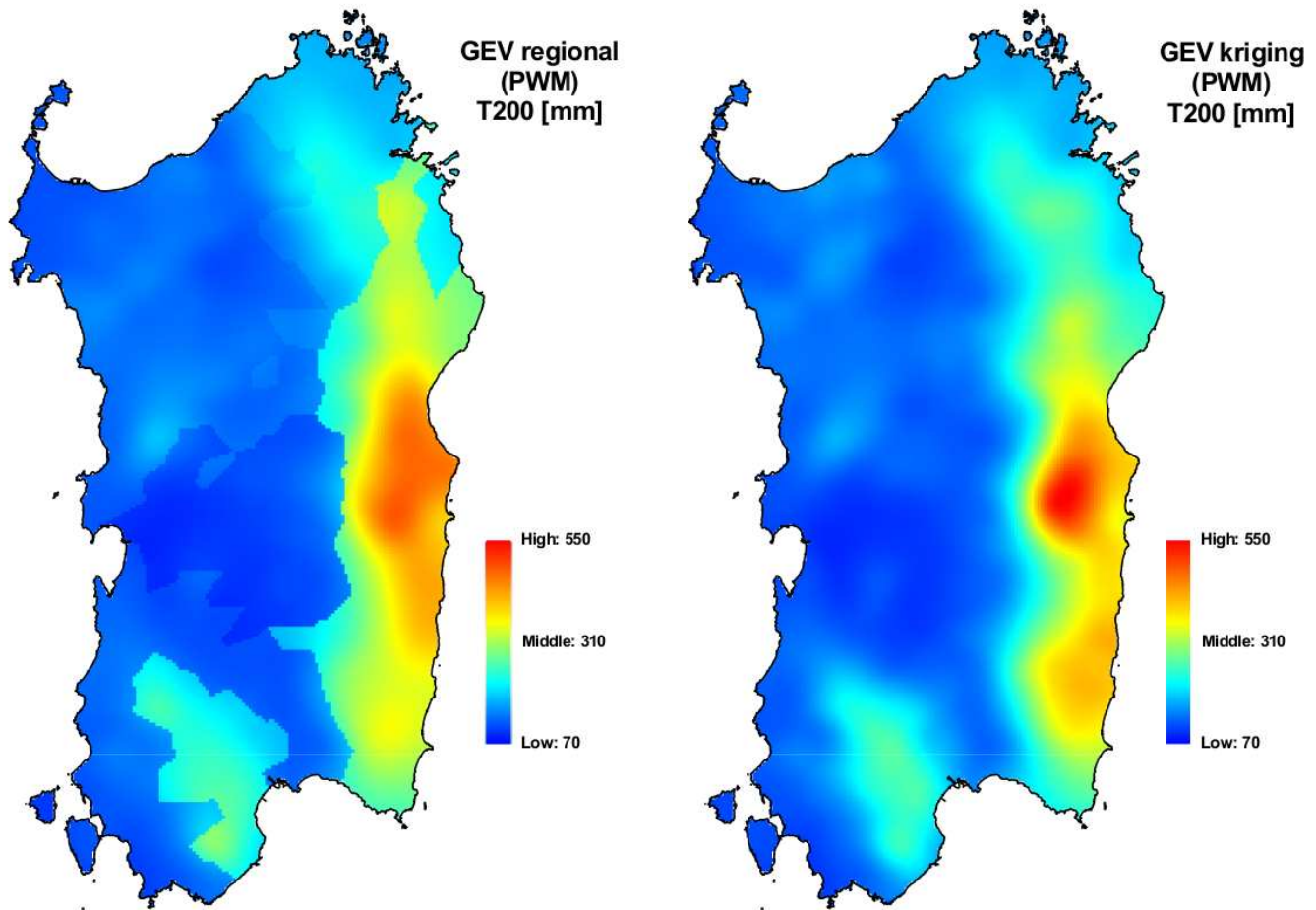


Figure 7.46: Map of daily rainfall depth h_T (mm) exceeded with return period $T=200$ yr.

Left: the result obtained using the regional approach (case **F** with 4 clusters).

Right: the result obtained using the geostatistical approach.

7.4.1 Spatial distribution of errors

The spatial distribution of error metrics ME and MEr , which characterize the eventual bias, is very interesting. We have not considered them so far because of little significance if averaged on many stations. Their spatial distribution can give indication about the spread of areas with an overestimation or an underestimation of the extreme events. The first metric (ME) provides a measure of how much, in average, the quantiles of theoretical CDF tend to overestimate/underestimate the extreme observed data (the highest 5 values for each station) and it is expressed in mm, whereas the second metric (MEr) represents the same quantity in relative and dimensionless terms, because it is divided by the observed data.

In Figure 7.47, the spatial distribution of the metric $ME(5)$ is reported for some approaches described in the previous paragraphs: a) local GEV; b) regional GEV (hypothesis **F** of partition in 4 homogeneous regions); c) GEV with parameters obtained from the kriging grid with 1-km step; d) regional TCEV-2008. All the distributions were determined using the same index-rainfall, locally calculated, in order to have a fair comparison. The analysis of Figure 7.47 clearly shows that the local case gives the best result (even if it is characterized by errors), but it cannot be taken into consideration because it is just locally valid. Contrarily, the TCEV case presents the most significant overestimations/underestimations (often higher than ± 40 mm), whereas GEV fits, both with regional parameters (hypothesis **F**), and with parameters interpolated by kriging grid, present lower error values. All the adopted approaches (including the GEV local fits) are characterized by larger errors in the East and South-East part of the island. The errors are concentrated in the areas characterized by more intense precipitations (see the spatial distribution of the index-rainfall given in Figure 7.44).

Observing the spatial distribution of the relative error metric $MEr(5)$, represented in Figure 7.48, a more uniform distribution of errors appears. The Figure confirms the best performances of GEV fits with spatially interpolated parameters.

These analyses have been repeated utilizing the cross-validation procedure. Results are reported in Figures 7.49 and 7.50, that, despite they show bigger errors, confirm the considerations just expressed and the superiority of GEV fits with parameters interpolated by kriging grid.

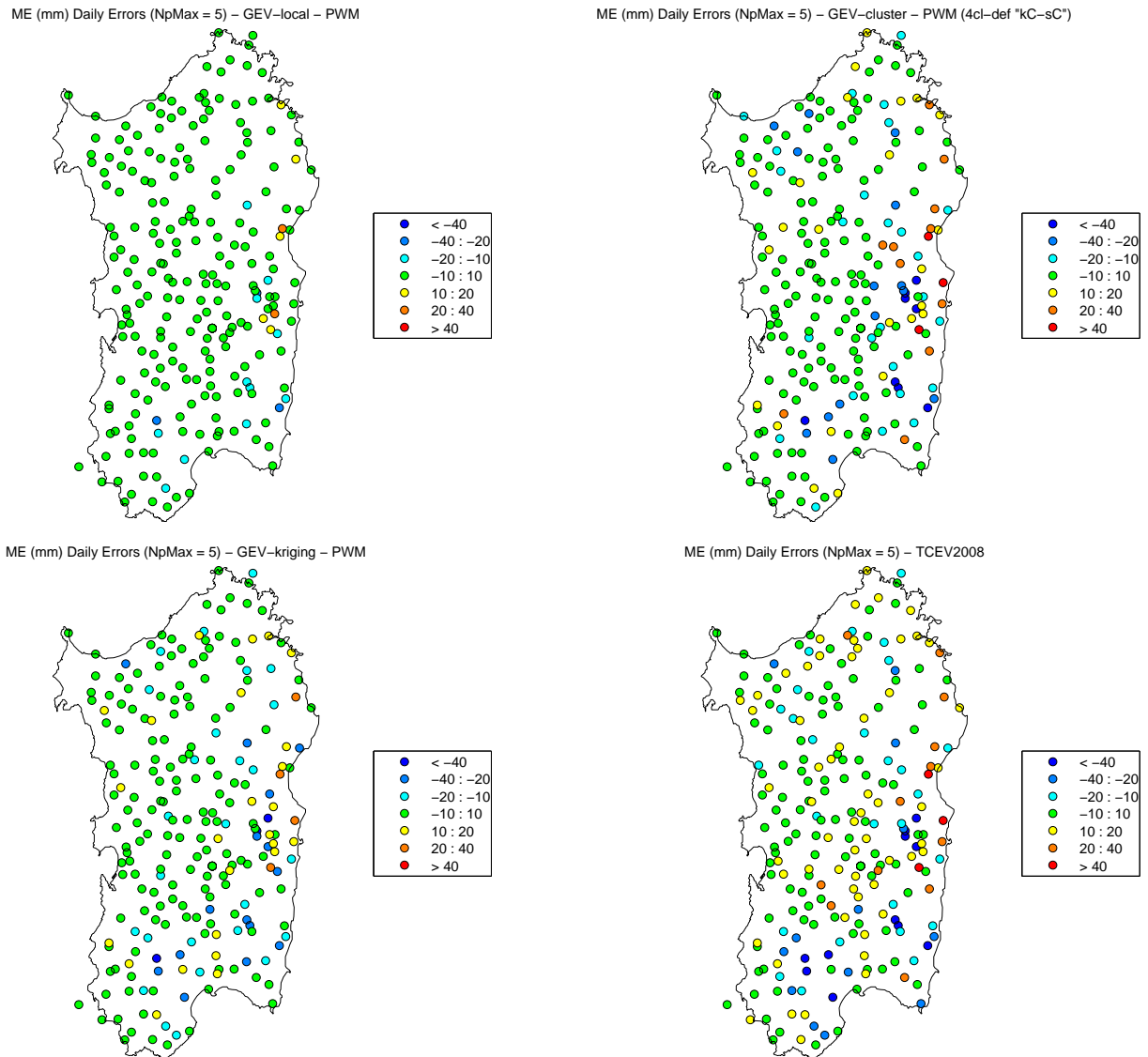


Figure 7.47: Spatial distribution of the error metric $ME(5)$ for the cases: GEV with local parameters (top left), GEV with regional parameters related to hypothesis **F** (top right), GEV with parameters estimated by kriging on a regular grid at 1 km resolution (bottom left), TCEV with index-rainfall updated to 2008 (bottom right). The same index-rainfall has been used also for the other distributions, in order to ensure a fair comparison.

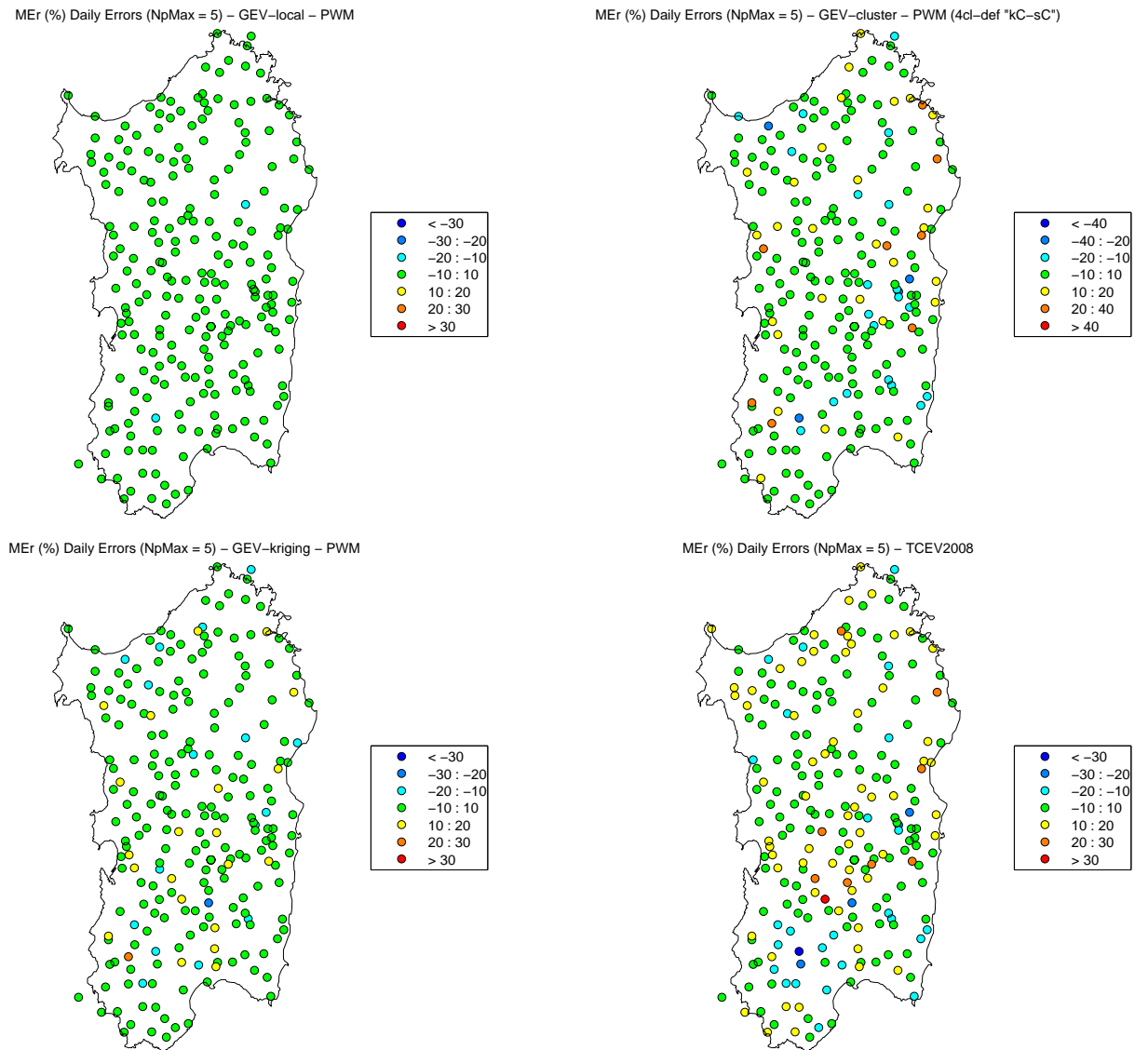
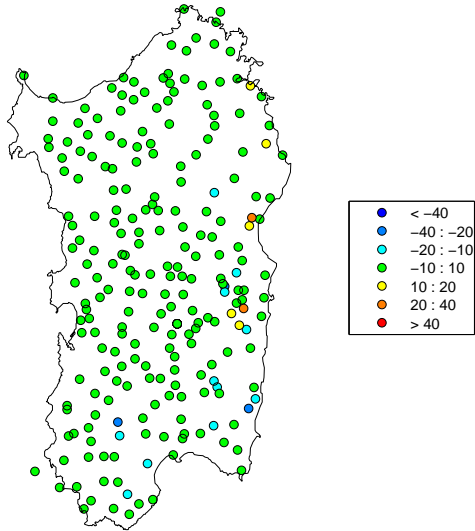
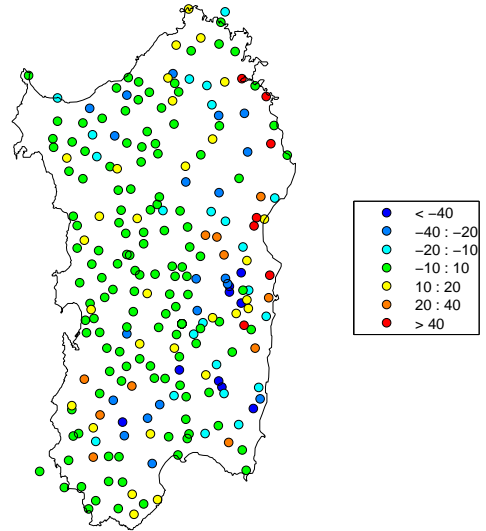


Figure 7.48: Same representations used in Figure 7.47, but on the error metric $MEr(5)$.

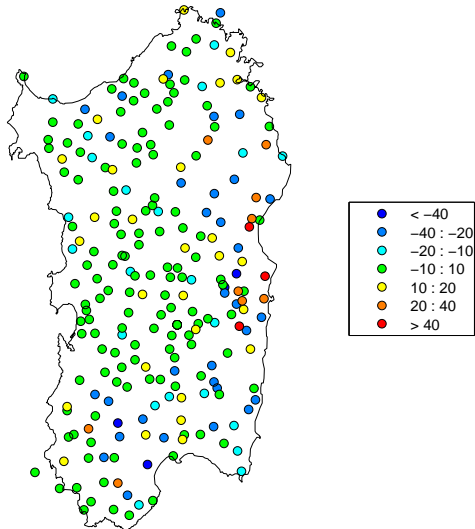
ME (mm) Daily Errors (NpMax = 5) – GEV–local – PWM



ME (mm) Daily Errors (NpMax = 5) – GEV–cluster – PWM (4cl–def "kC–sC")



ME (mm) Daily Errors (NpMax = 5) – GEV–kriging – PWM



ME (mm) Daily Errors (NpMax = 5) – TCEV2008

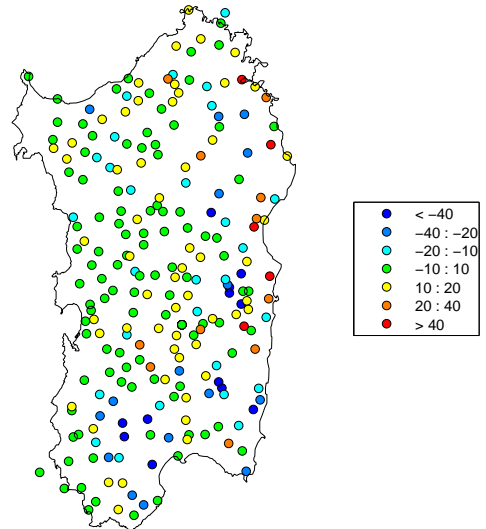


Figure 7.49: Same representations used in Figure 7.47, but on the error metric $ME(5)$ calculated after the **cross-validation** procedure.

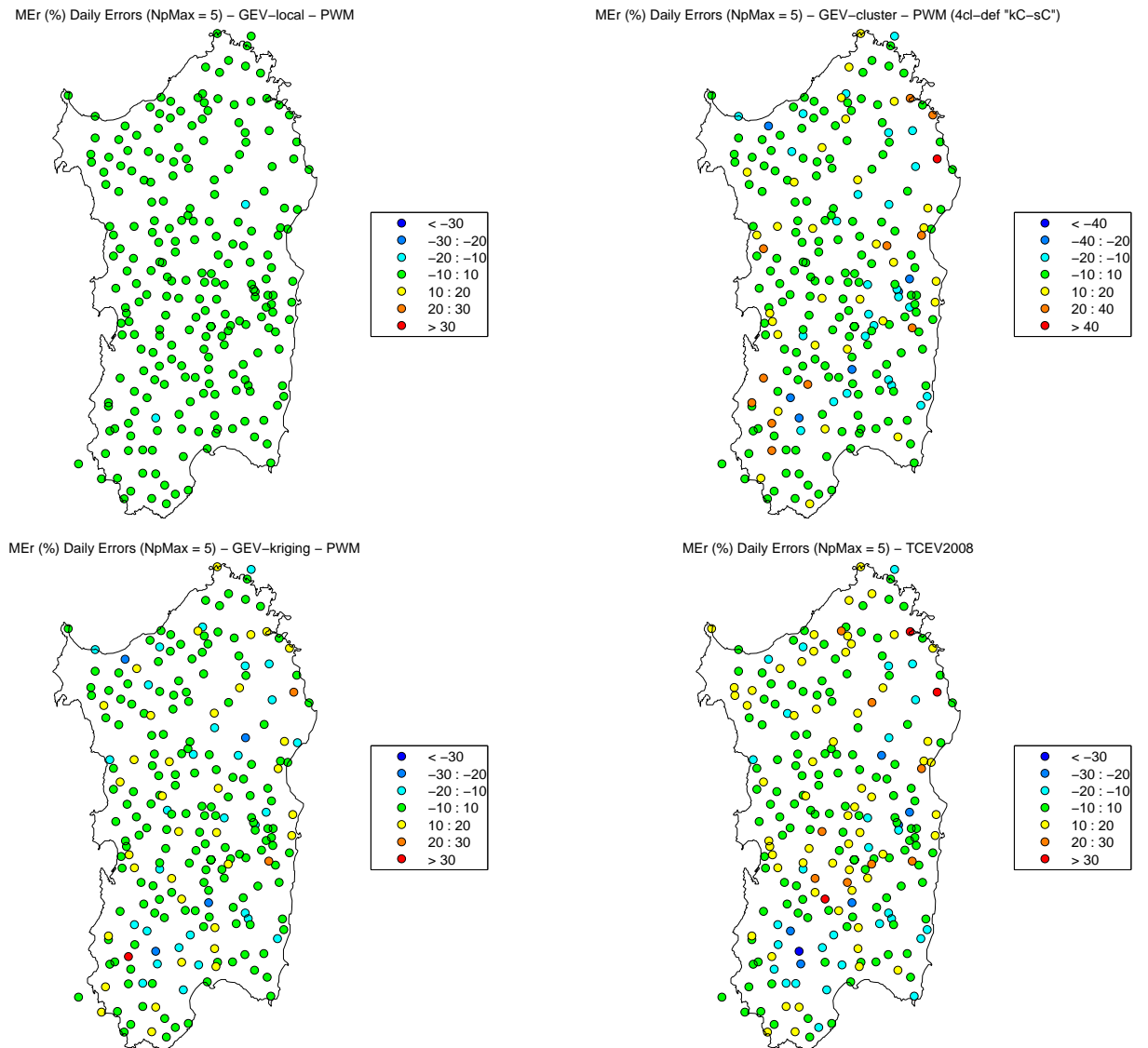


Figure 7.50: Same representations used in Figure 7.47, but on the error metric $MEr(5)$ calculated after the **cross-validation** procedure.

7.5 BM summary

This section summarizes the results of the BM approach.

We adopted the GEV as the probabilistic distributions that best fit the empirical distribution of annual maxima of daily precipitation observed in Sardinia. We chose to adopt the GEV distribution because it best represents the observed data, see L-moment ratios diagrams in Figures 7.37 and 7.1. Moreover, with respect to TCEV distribution, adopted in a previous regional study (Deidda and Piga, 1998; Deidda et al., 2000), the GEV distribution can be determined through three parameters, that make it more manageable and more simple applicable in engineering practice, also for the availability of explicit expressions to calculate the quantiles.

The TCEV distribution estimates reported in Deidda and Piga (1998); Deidda et al. (2000) have been updated using the new database of daily precipitations available till 2008, in order to make a fair comparison. The comparisons between the different approaches were executed evaluating the fits with two different classes of error metrics. The metrics of the first class measure square errors between cumulative distribution functions, whereas those of the second class measure the errors in reproducing the highest observed values.

The regional parameters of the GEV distribution for each homogeneous regions in each of the 6 hypotheses of subdivision Sardinian territory were estimated. All the error metrics indicate that the best fit is obtained with regional GEV distribution in the hypothesis **F** of partition in 4 homogeneous regions.

The opportunity to represent with continuity the spatial variability of GEV parameters was explored too. Regular grids with 1-km step were determined with the technique of kriging for uncertain data.

The comparison between regional and geostatistical approaches has been conducted with the method of cross-validation. The results indicated that the best fit is obtained using GEV parameters given by the grids obtained through kriging technique.

Furthermore there are other reasons to prefer the geostatistical approach to the regional one. In fact, with a regional approach eventual variations of parameters due to physical heterogeneity and to topography may not be reproduced inside an homogeneous region, since the parameters (except the index-rainfall) are held constant. Moreover with the regional approach there is uncertainty in the definition of any hypothetical borders between homogeneous regions, due to both the inadequate density of observation points and to the process of assignment of stations to homogeneous regions. Another drawback to the regional approach, is that the identification of the homo-

geneous region depends on subjective choices, as for example the choice of the aggregation criteria, of the homogeneity check, etc. In addition this approach may generate difficulties in practical applications when more than one homogeneous region fall within the same basin

Using a geostatistical approach is possible to overcome all these drawbacks. There are less subjective choice (only the type of spatial interpolation to apply), is possible to represent the local peculiarities that can be induced by the exposure, general morphometric factors, climate and microclimate, and overcomes the problems associated with abrupt spatial discontinuity in probability distribution parameters at the border between continuous homogeneous regions.

So, for the study of daily rainfall time series using the peaks over threshold approach, whose results are reported in Chapter 8, we decided to use only the geostatistical approach.

Chapter 8

POT results

In this section are reported the results obtained applying the MTM-GP model, see section 4.2.3, to the daily rainfall time series. We used the 256 stations with at least 50 complete years of observations. We decided to use more stations respect to the previous chapter because the MTM-GPD parameter estimation procedure is very robust. This consents to use data from rain gauges with less years of observation, without compromising the goodness of the estimate. Furthermore, with a POT approach the sample size increases, because we are considering more data for every year of observation, not just the maxima as in the BM approach.

Considering the results obtained from the analysis of annual maxima of daily precipitation, we decided to use only the geostatistical approach. In fact, as already said in section 7.5, the geostatistical approach is able to represent the local peculiarities that can be induced by exposure, general morphometric factors, climate and microclimate. Furthermore the geostatistical approach overcomes the problems associated with spatial discontinuity of probability distribution parameters at boundaries between contiguous homogeneous regions.

Figure 8.1 shows the L-moment ratios diagram (Hosking, 1990) for the 256 stations with more than 50 complete years in record, using only daily rainfall depths exceeding 5 mm. The line that links the possible theoretical couples relative to the GP distribution is the most barycentric and interpolating among the considered lines. These results suggest the use of the GP distribution to represent the daily rainfall depths.

Many time series of our database contain anomalous quantities of daily rainfall records rounded off at unexpected resolutions of 0.5, 1 and 5 mm/d. An example of this situation is illustrated in Figure 8.2, which shows a zoom of the empirical CDF of daily rainy data collected by station 007. We can notice a large amount of data rounded every 5 mm. A deeper inspection of

the figure reveals also that many values are rounded off every 1 mm. The MTM-GP model overcomes the problems related to the presence of high rounded-off data discussed in section 4.2.3 and in Deidda and Puliga (2006); Deidda (2007).

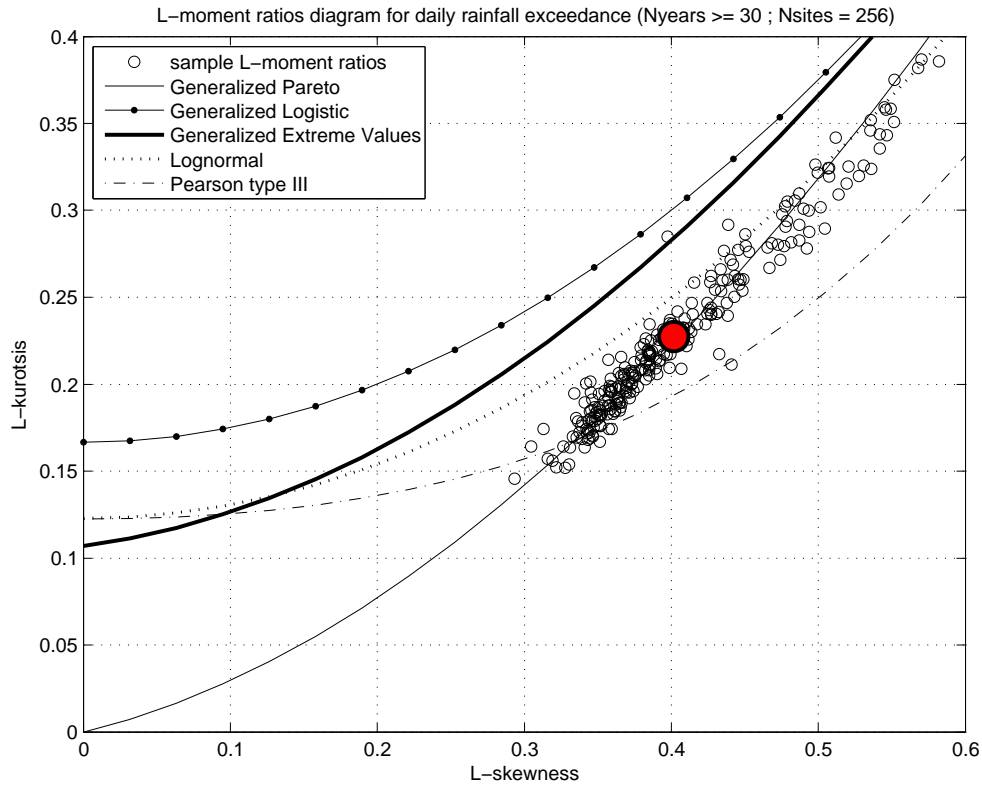


Figure 8.1: Comparison between pairs of L-moment ratios (L-skewness, L-kurtosis) calculated on daily rainfall depths exceeding 5 mm, for the 256 stations with more than 30 complete years in record (circles) and theoretical pairs for some distributions widely used in statistical hydrology, represented by lines of different strokes. Big marks denote the average values of the same statistics.

8.1 Local analysis results

In a first phase, parameters $(\xi_0^M, \alpha_0^M, \zeta_0^M)$ of the MTM-GP model, described by equation (4.45), were locally estimated for each daily precipitation time series considered in this study by following the estimation procedure described

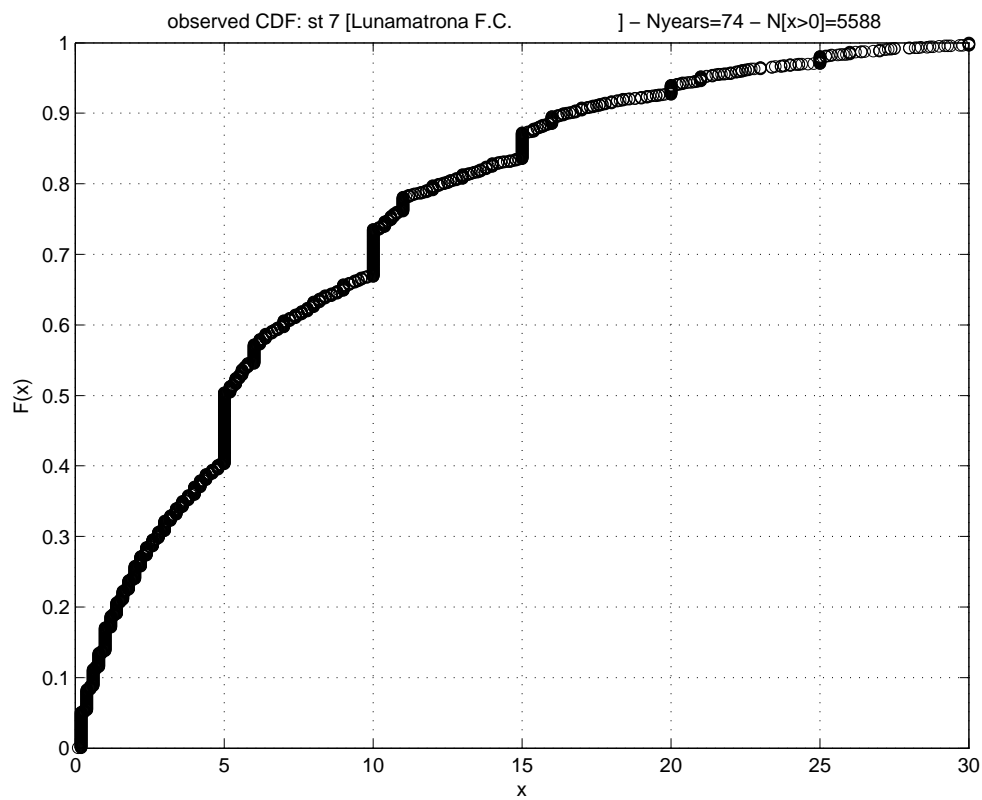


Figure 8.2: CDF of rainy data collected by station 007. The zoom shows how many data are rounded off to discrete values.

in section 4.2.3. We applied the MTM-GP model in a range of thresholds between 2.5 and 12.5 mm, with a step of 0,06 mm.

The choice of the threshold interval initially was derived from a visual analysis of the figures representing the hierarchical estimation procedure of the MTM-GP parameters, like Figure 4.2. From the visual analysis, for each of the 256 stations, we noticed that for values of the threshold u lower than 2.5 mm the parameter $\xi_0(u)$ is not stable, and that for value of u larger than 15/20 mm there is a departure from the median of the values, due to the increasing variance associated with the small number of exceedance. So it was possible to fix a range of threshold between 2 and 20 mm. We tested different ranges of threshold within this interval, the range 2.5-12.5 mm showed the best performance in term of goodness of fit test. This result agrees with the range proposed in Deidda (2010). In addition, as suggested by the author, this range corresponds to joining two intervals of thresholds of 5 mm in size and centered on 5 and 10 mm, where we often observed jumps in the estimates, due to the anomalous rounding-off data.

In Figure 8.3 are reported the empirical cumulative function for the locally estimates of the MTM-GP parameters, obtained through maximum likelihood (ML), simple moments (SM) and probability weighted moments (PWM) methods. In figure on top is reported the CDF of ξ_0^M estimates, in the middle the CDF of α_0^M estimates and in the bottom the CDF of ζ_0^M estimates. The main differences are in the estimates of the shape parameter ξ_0^M . SM estimates are always less than 0.4, so we don't have problems related to the fact that conventional moments of order greater than or equal to $1/\xi_0^M$ degenerate. PWM and ML estimates are higher than SM estimates for a certain number of stations, and the differences are of so significant that they could cause big difference in the estimation of quantiles too. An example of this situation is shown in Figure 8.4 where the empirical CDF of two different stations and the theoretical ones obtained with the MTM-GP model are reported. In the legend, for each station, the error metric $MEr(5)$, A^2 and W^2 are reported. On the top, station 001, the different estimation techniques provides similar estimates of the MTM-GP parameters. On the bottom, station 323, the different methods provides very different estimates, and the ML and PWM techniques tend to overestimate the shape parameter, and, consequently to strongly overestimate the extreme events.

The spatial distribution of the stations in which there is the greatest discrepancy between SM and ML (or PWM) estimates is not random. Most of them are in the East part of the island, characterized by the most extreme events. Their spatial distribution is shown in Figure 8.5, which anticipates some of the results described in next section. The picture shows, in the top part, the spatial distribution of the shape parameter whose estimates were

obtained through SM, PWM and ML techniques, from left to right. In most of regional territory there are not relevant differences, but in a limited zone in the Est part of the island marked differences emerge. A big difference in the estimates of the shape parameter means a big difference in the quantiles estimates too. We can observed that in the bottom part of Figure 8.5, where maps of rainfall depth exceeded with return period of 200yr are reported.

The main cause of this situation is that SM estimates are influenced by the highest (and the lowest) observed values more than PWM an ML's. So SM estimates tend to give better fits to the tails of the distribution, while PWM and ML estimates give a best fit with respect to the whole distribution of observed data. The MTM-GP model describes the distribution of all daily rainy and non rainy values, not only the extremes (like the GEV distribution), so we have a high number of data for each station, and the bulk of the empirical distribution may significantly differ with respect to the tails. In conclusion, if one is interested in the estimates of extreme events, SM technique provides the best result, because it tends to describe better the tails of the distribution, being more influenced by "outliers".

A synthesis of the performances of local fits with the MTM-GP model is reported in Table 8.1. In detail, by using the local estimates of parameters ($\xi_0^M, \alpha_0^M, \zeta_0^M$), the error metrics described in paragraph 6 were calculated for each one of the 256 stations. The averages of these metrics are reported in Table 8.1, where we can observe that the estimates obtained by using the SM are characterized by the best result in terms of quantile error, metrics $MAE(5)$ and $MAEr(5)$, but perform worse respect to the metrics A^2 and W^2 . The difference in the quantile errors is very high, using ML or PWM estimates the errors are about double than those obtained using SM estimates.

This results reported in the Table, and the observations previously made, lead us to adopt the SM estimates.

	MAE(5) [mm]	MAEr(5) [-]	A² [-]	W² [-]
MTM-GPD local (ML)	39.131	0.184	10.759	1.632
MTM-GPD local (SM)	15.794	0.097	12.634	2.067
MTM-GPD local (PWM)	30.535	0.163	9.855	1.447

Table 8.1: Comparison between performances of local fits with the MTM-GPD model, with parameters estimated with ML, SM and PWM techniques. The averages of error metrics are calculated over the 256 stations with at least 30 complete years of observations.

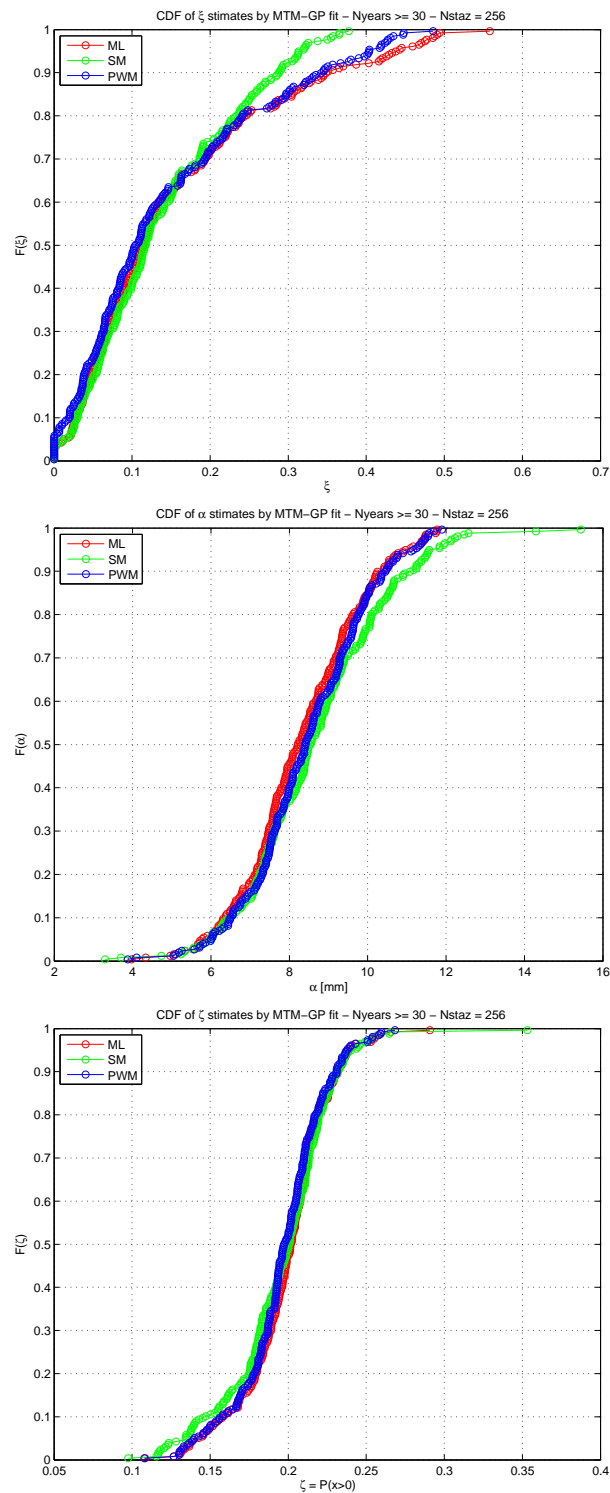


Figure 8.3: Cumulative distribution function of the MTM-GP parameters: ξ_0^M , α_0^M and ζ_0^M , from top to bottom. Estimates are obtained with SM, ML and PWM techniques.

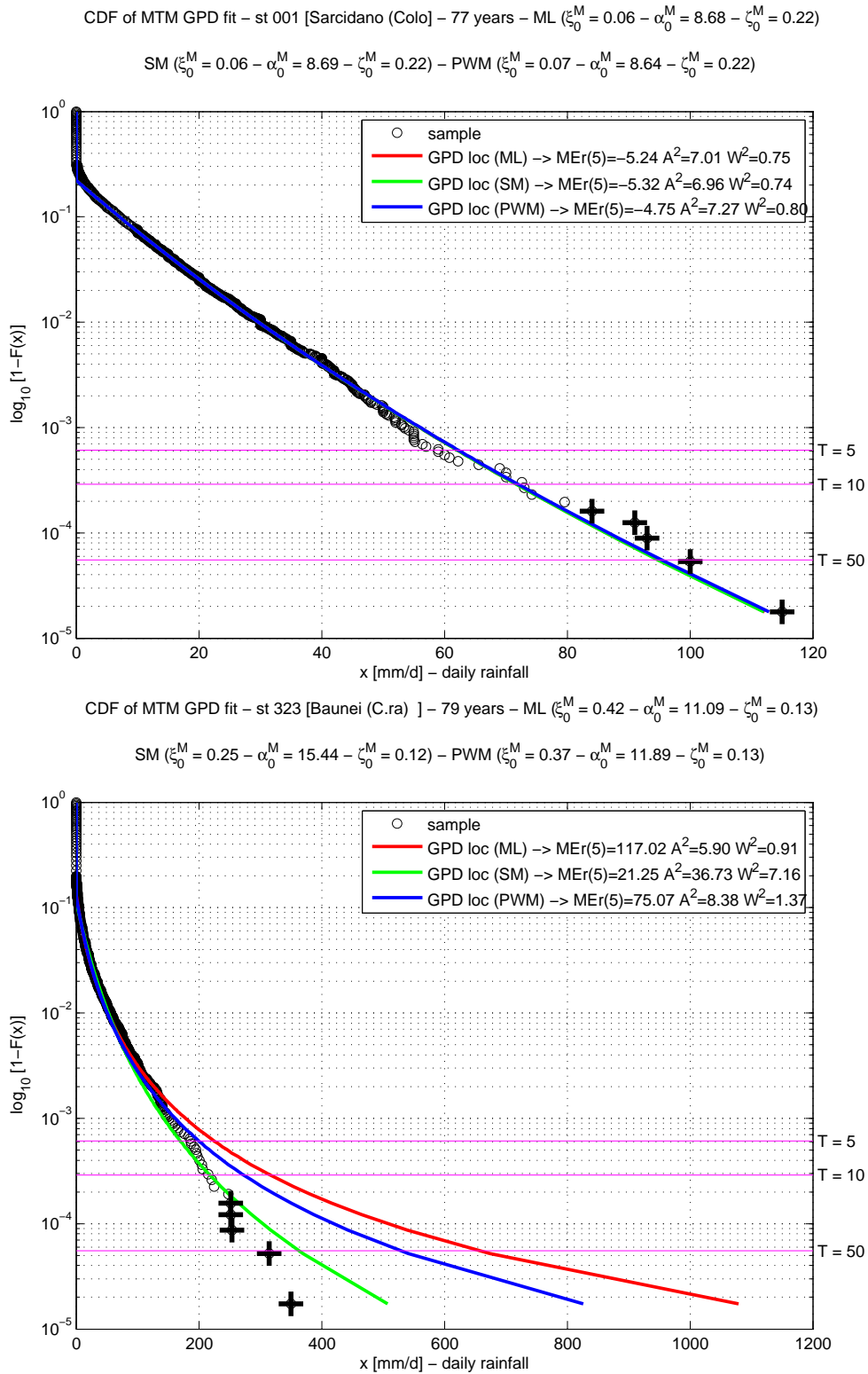


Figure 8.4: Empirical cumulative distribution functions (calculated with Hazen’s plotting position) of daily precipitation, compared with theoretical MTM-GP distributions, whose parameters are locally estimated through SM, ML and PWM techniques.

Top: station 001. Bottom: station 323

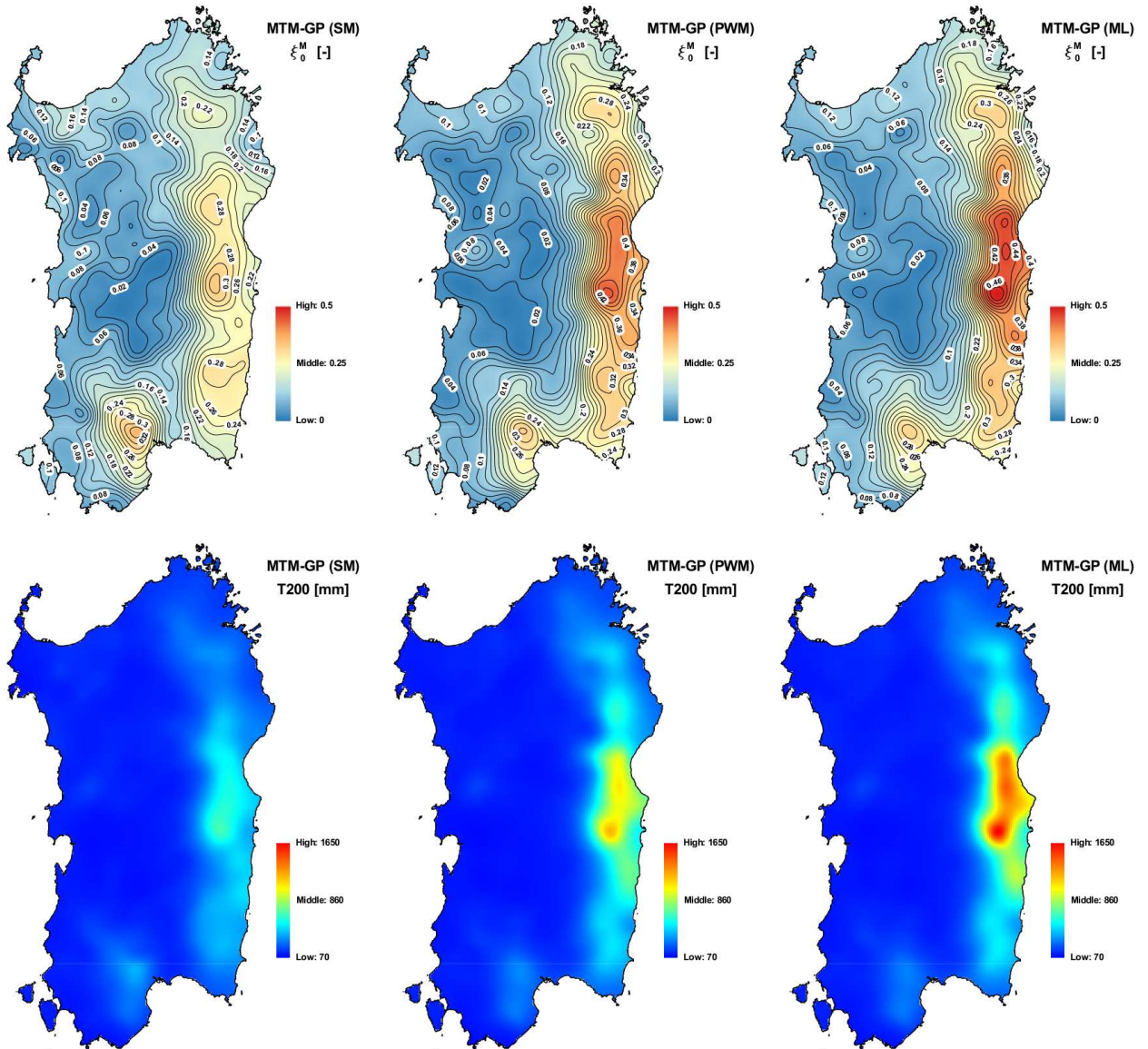


Figure 8.5: Top: Representation of the spatial distribution of the shape parameter of the MTM-GP distribution, ξ_0^M , obtained using SM, PWM and ML techniques, from left to right.

Bottom: Map of the rainfall depth h_T (mm) exceeded with return period $T=200$ yr using MTM-GP estimates obtained with SM, PWM and ML techniques, from left to right.

8.2 Geostatistical analysis results

The opportunity to represent with continuity the spatial distribution of the MTM-GP model parameters have been investigated. For this purpose, the kriging technique has been used, as described in section 5.2.

A preliminary analysis has been performed, in order to search the optimal number of adjacent stations to use, the minimum optimal number of years of observation in order to select the historical series, and the type of kriging technique to use.

The results of these analyses, showed that there is no difference in using the OK or the KUD for the spatial interpolation of the parameters of the MTM-GP model. That's because the parameters are defined as the median of the estimates in the selected range of threshold, so the estimates are very robust, and the estimation variance is close to zero. We chose 10 as the optimal number of adjacent stations in order to write the liner system of equations for kriging. The more plausible variogram for the spatial interpolation of the shape parameter ξ_0^M is the linear model, for the spatial interpolation of the other two parameters the more plausible variogram is the power model. We decided to use all the 256 stations with at least 30 complete years of observations for the spatial interpolations of the MTM-GP parameters. The preliminary analysis that lead the just described choices were conducted with the cross-validation method. For each aspect we analyzed, the overall performances were evaluated calculating the mean absolute error between the value of the locally estimated parameter and the one obtained by interpolation during the cross-validation procedure. A synthesis of the results of this preliminary analysis is reported in Figures 8.6, 8.7 e 8.8 for each of the three parameters of interest.

Considering the results of these preliminary analyses, all the spatial interpolations were conducted using the OK, on a regular grid with spatial step equal to 1 kilometre (the same grid used for the spatial interpolation of the GEV parameters).

Figure 8.9 shows the comparison of several variograms for the spatial interpolation of the shape parameter ξ_0^M (top), the scale parameter α_0^M (middle) and the ζ_0^M parameter (bottom).

As described in section 7.3 we sporadically observed noise around some grid points that could interrupt the local monotonicity of parameters' spatial variations. In order to smooth this phenomena we applied the same moving average with a 9x9 km window described in section 7.3. The comparison between the grids obtained through the kriging and those obtained after the moving average highlights negligible variations in the peaks, but only a benefit in terms of continuity in the distribution of parameters.

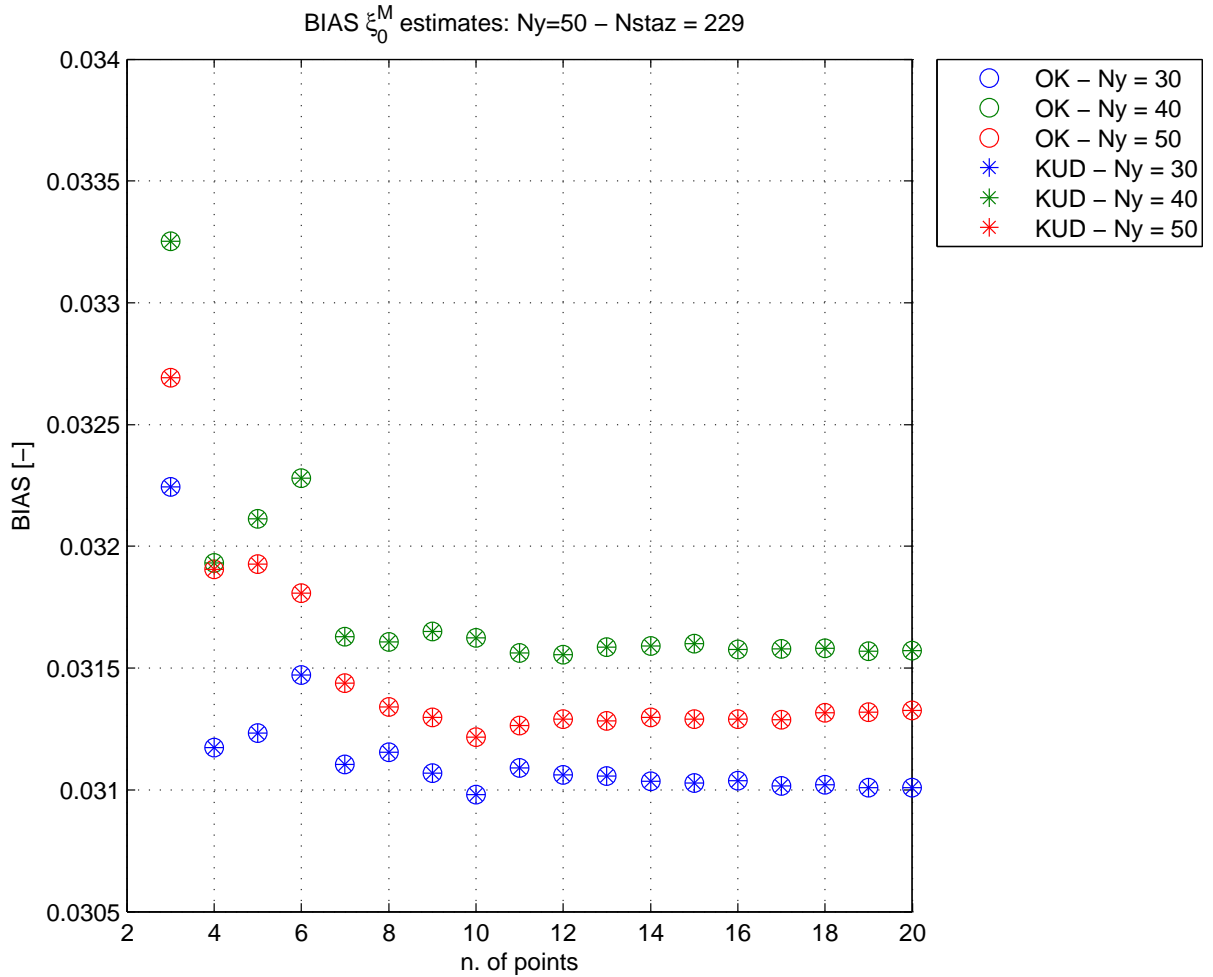


Figure 8.6: Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the shape parameter ξ_0^M . The ordinate shows the mean absolute error (MAE) in the interpolation of the parameter ξ_0^M (using cross-validation) in each of the 229 stations with at least 50 years of observations (excluding time to time the estimate of the considered station), in function of the number of nearest stations (on the abscissa) used for the kriging system. The empty circles represent the results with the ordinary kriging, the asterisks refer to the kriging with uncertain data. The different colors refer to the minimum number of years to select the stations to be used for the interpolations.

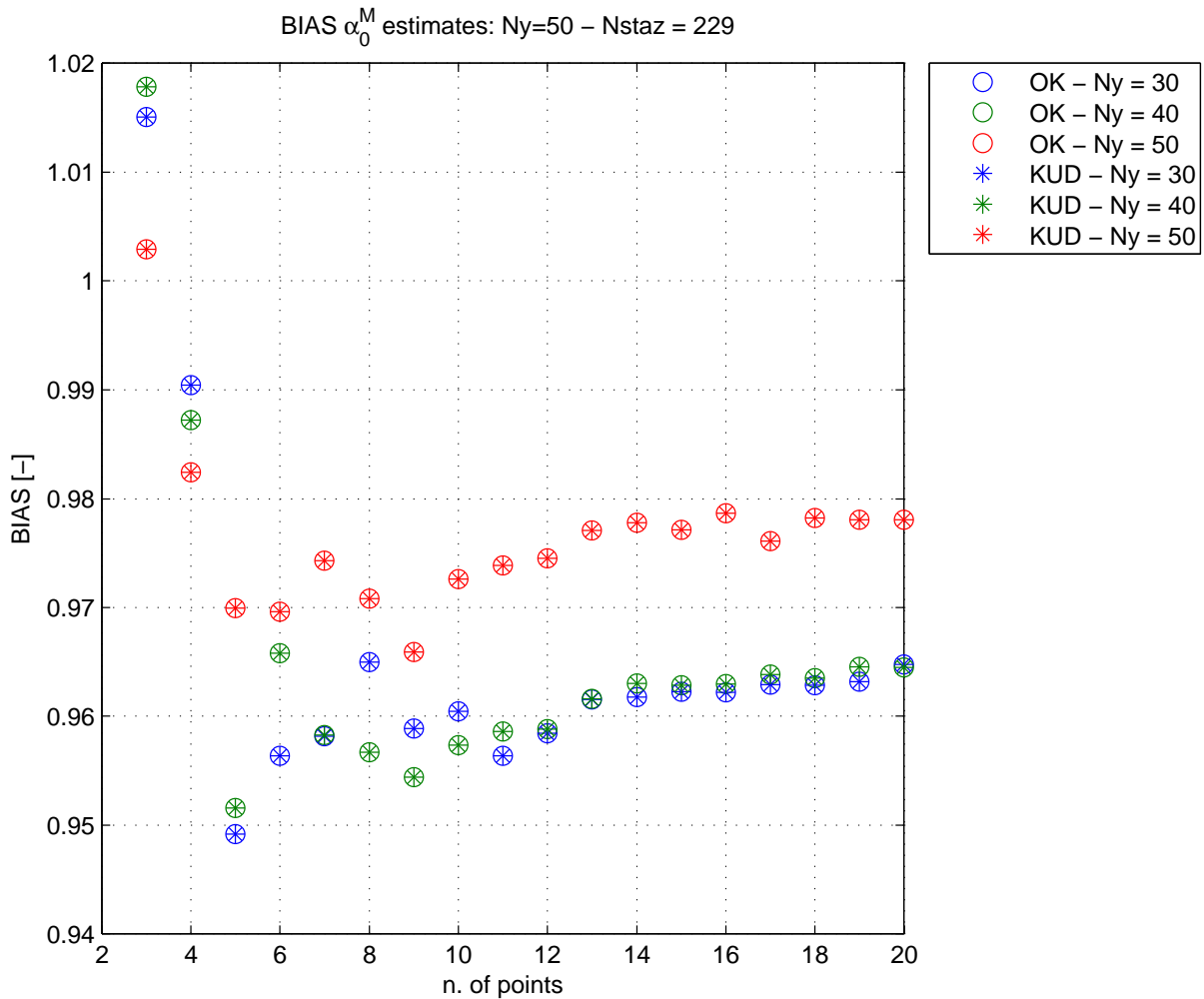


Figure 8.7: Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the scale parameter α_0^M . The symbolism is the same as for Figure 8.6, but MAE are calculated on the interpolations of the α_0^M , parameter using cross-validation.

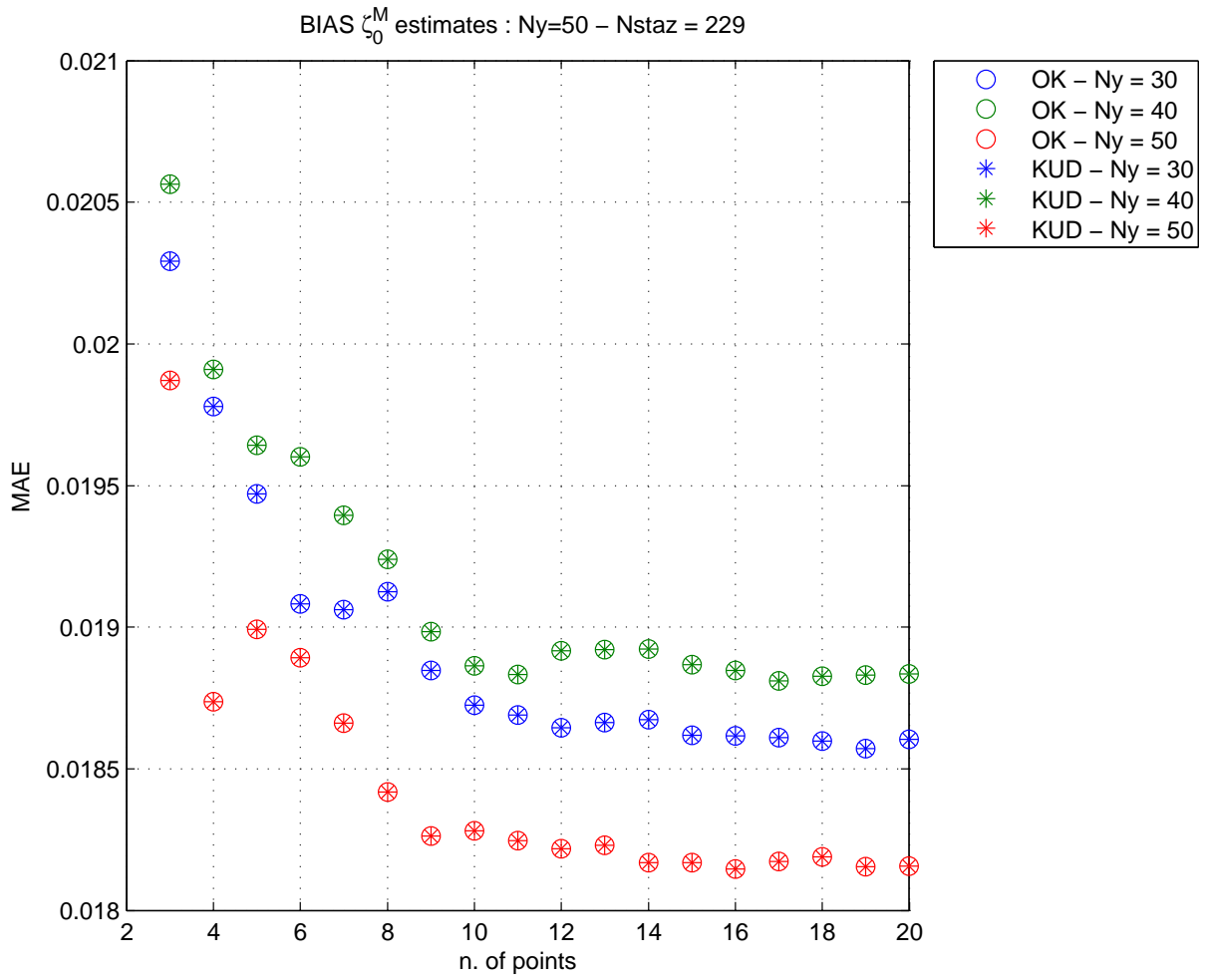


Figure 8.8: Results of the preliminary analysis aimed at determining the optimal conditions for the spatial interpolation of the ζ_0^M parameter. The symbolism is the same as for Figure 8.6, but MAE are calculated on the interpolations of ζ_0^M , using cross-validation.

In Figures 8.10, 8.11 and 8.12 the the spatial distributions of the MTM-GP parameters: ξ_0^M , α_0^M and ζ_0^M are reported.

A synthesis of the performances of local and geostatistical fits of the MTM-GP model, using SM estimates, are presented in Table 8.2. The errors of the geostatistical approach were calculated again by applying the cross-validation procedure.

	MAE(5) [mm]	MAEr(5) [-]	A² [-]	W² [-]
MTM-GP local (SM)	15.794	0.097	12.634	2.067
MTM-GP kriging	16.569	0.106	18.935	3.366
MTM-GP kriging (cross-val)	20.720	0.139	21.647	3.831

Table 8.2: Comparisons between the performances of local and kriging fit. Averages of error metrics calculated over the 256 stations with at least 30 complete years of observations.

In Attachment 2, for each of the 256 stations with more than 30 complete years of observations, are reported the empirical CDFs of daily rainfall and theoretical ones obtained by using: local estimates of MTM-GP parameters, estimates of MTM-GP parameters from kriging grid. In the legend, for each station, it's reported: the error metric $MEr(5)$ which gives a percentage estimation of how much the theoretical distribution overestimate or underestimate, in average, the biggest 5 observations, and the metrics A^2 and W^2 , which give an indication about the fit of theoretical distributions to the whole set of observed data.

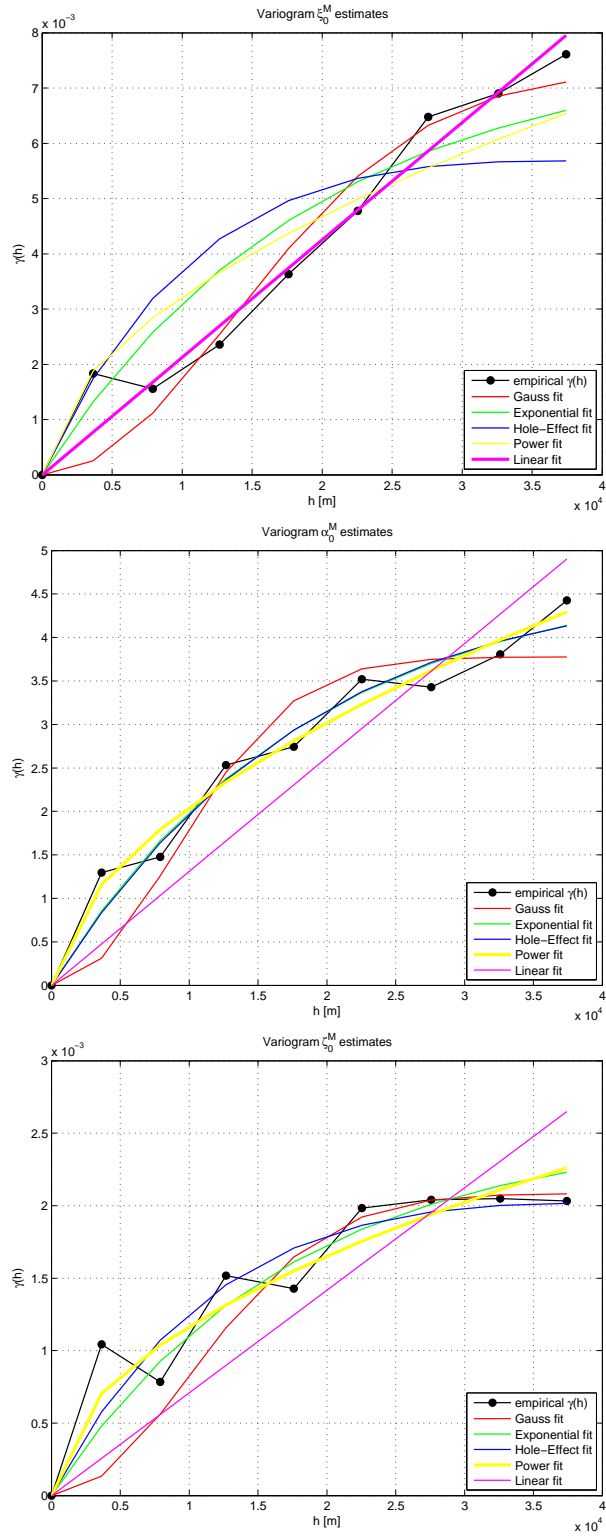


Figure 8.9: Comparison between sample variograms, based on local estimates of parameters ξ_0^M , α_0^M and ζ_0^M (from top to bottom), and some theoretical variograms.

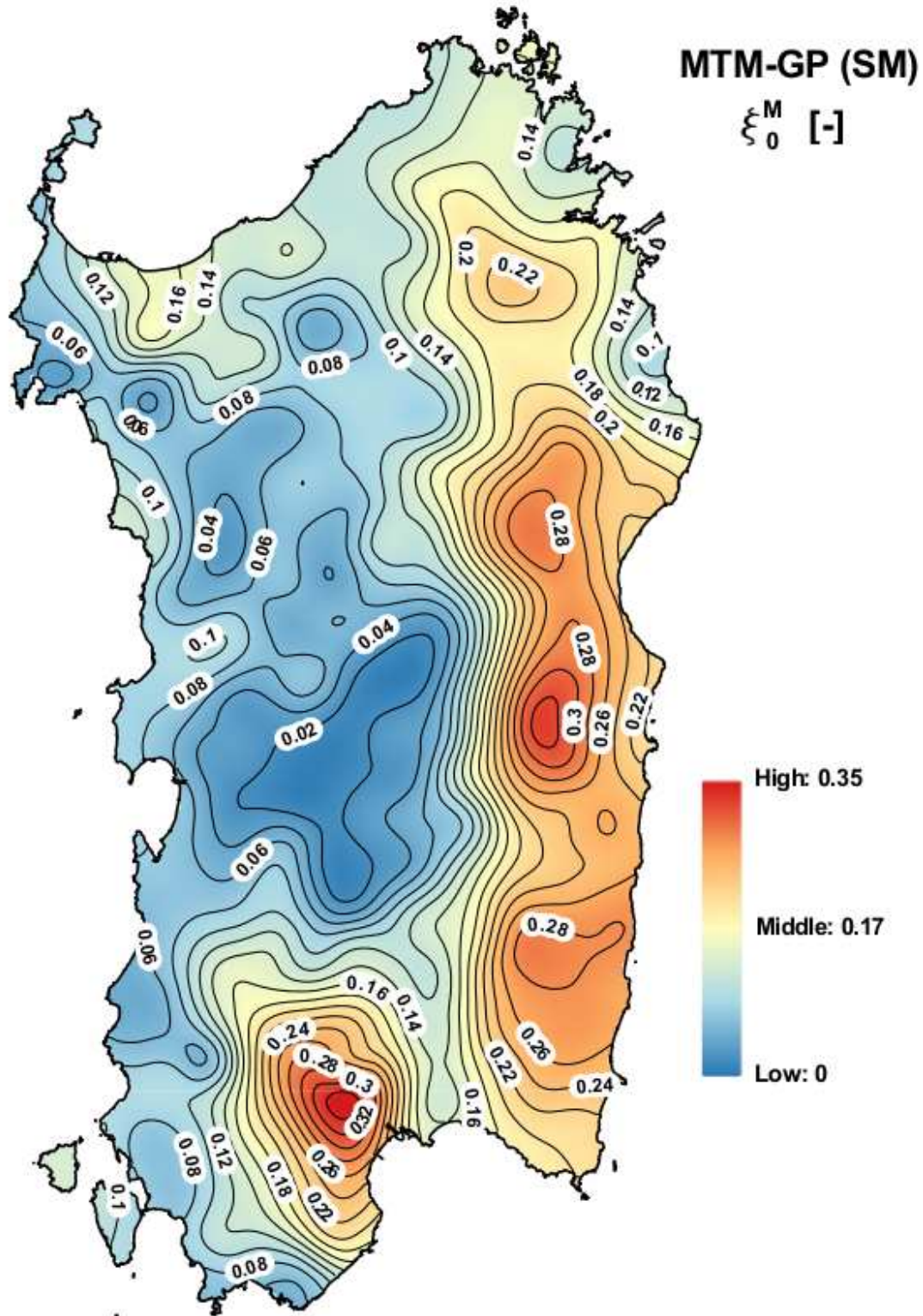


Figure 8.10: Representation of the spatial distribution of the MTM-GP shape parameter, ξ_0^M . The map is obtained from a regular grid with 1 km resolution.

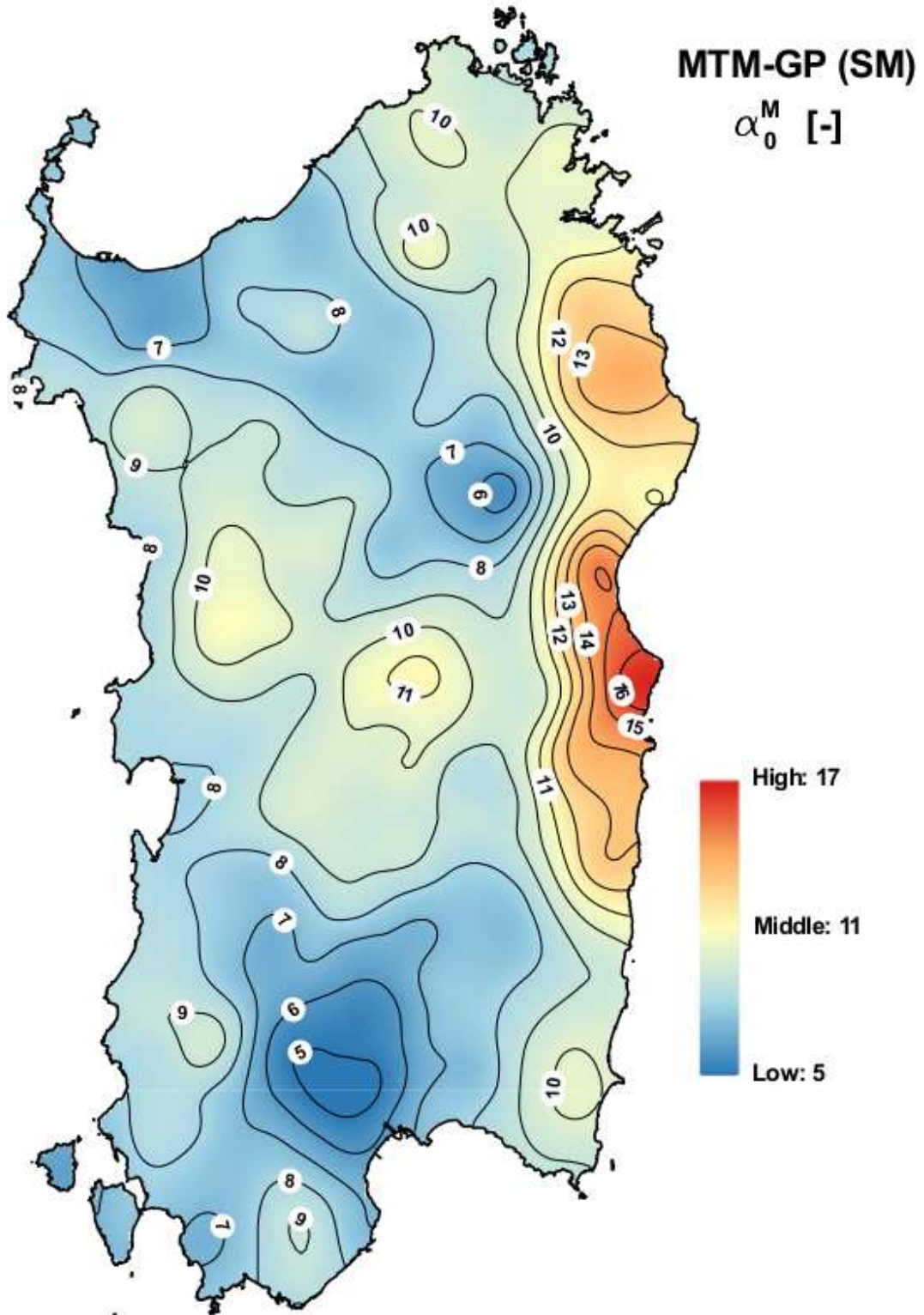


Figure 8.11: Representation of the spatial distribution of the MTM-GP scale parameter, α_0^M . The map is obtained from a regular grid with 1 km resolution.

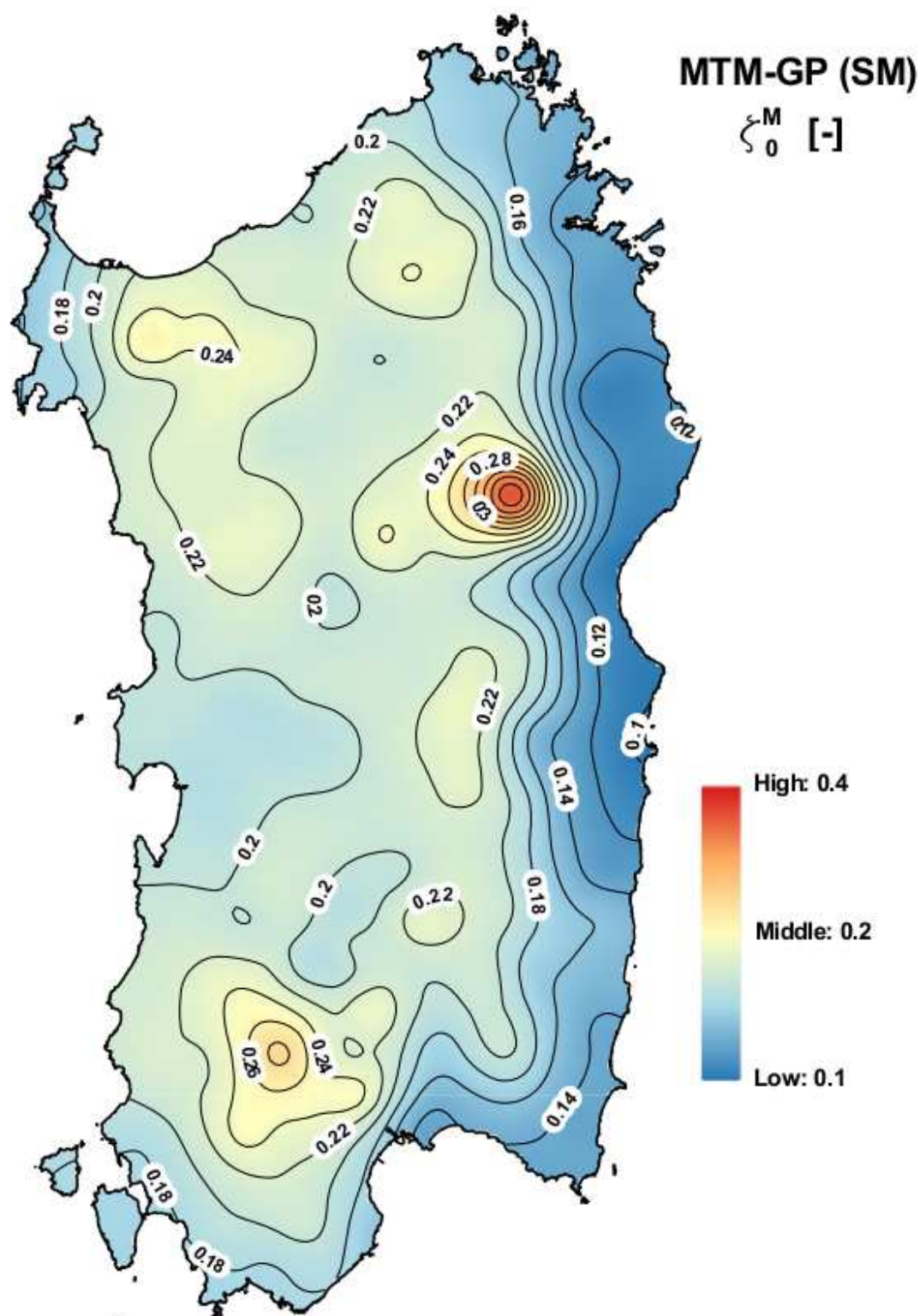


Figure 8.12: Representation of the spatial distribution of the MTM-GP parameter, ζ_0^M . The map is obtained from a regular grid with 1 km resolution.

8.3 Spatial distribution of errors

The spatial distribution of error metrics ME and MEr , which characterize the eventual bias, is reported in this section. Their spatial distribution gives indication about the spread of areas with an overestimation or an underestimation of extreme events. The first metric (ME) provides a measure of how much, on average, the quantiles of theoretical CDF tend to overestimate/underestimate extreme observations (the 5 highest values for each station) and it is expressed in mm, whereas the second metric (MEr) represents the same quantity in relative and dimensionless terms, because it is divided by the observed data.

In Figure 8.13, the spatial distribution of the metric $ME(5)$, which is calculated over the 5 highest values, for each of the 256 stations with at least 30 complete years of observations, is reported for local MTM-GP estimations and MTM-GP estimations with parameters obtained from the kriging grid with 1-km step.

The larger errors are located in the eastern part of the island, characterized by more intense precipitations (for instance, look at the spatial distribution of index-rainfall given in Figure 7.44). It is more significant to observe the spatial distribution of the relative error metric $MEr(5)$ represented in Figure 8.14, that shows a more uniform distribution of errors.

Also this analysis has been repeated utilizing cross-validation technique. Results are reported in Figures 8.15 and 8.16, that confirm the just expressed considerations.

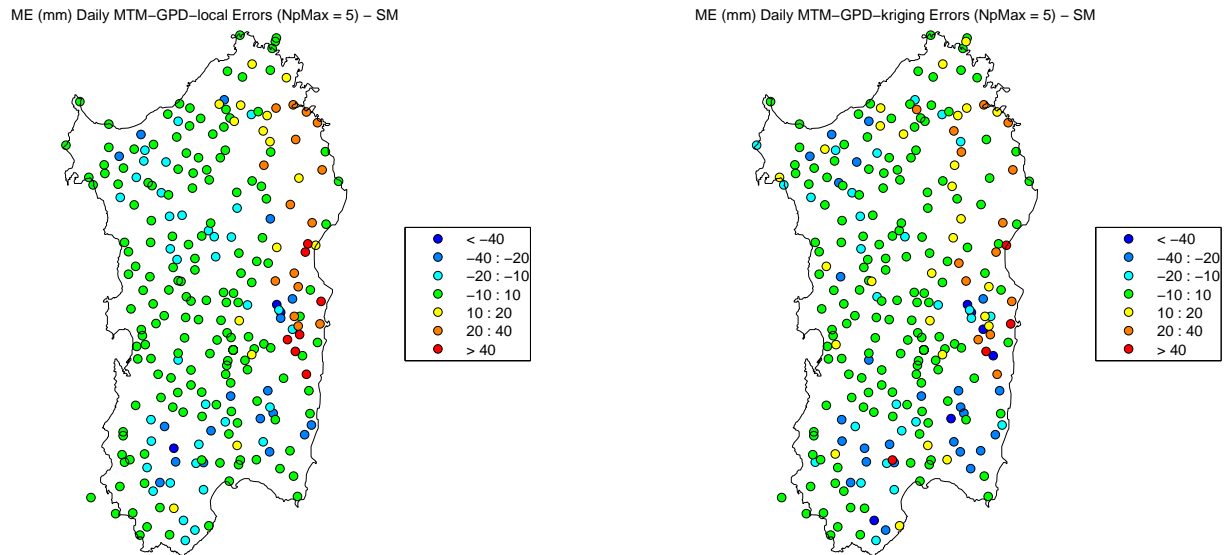


Figure 8.13: Spatial distribution of the error metric $ME(5)$ for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution (right).

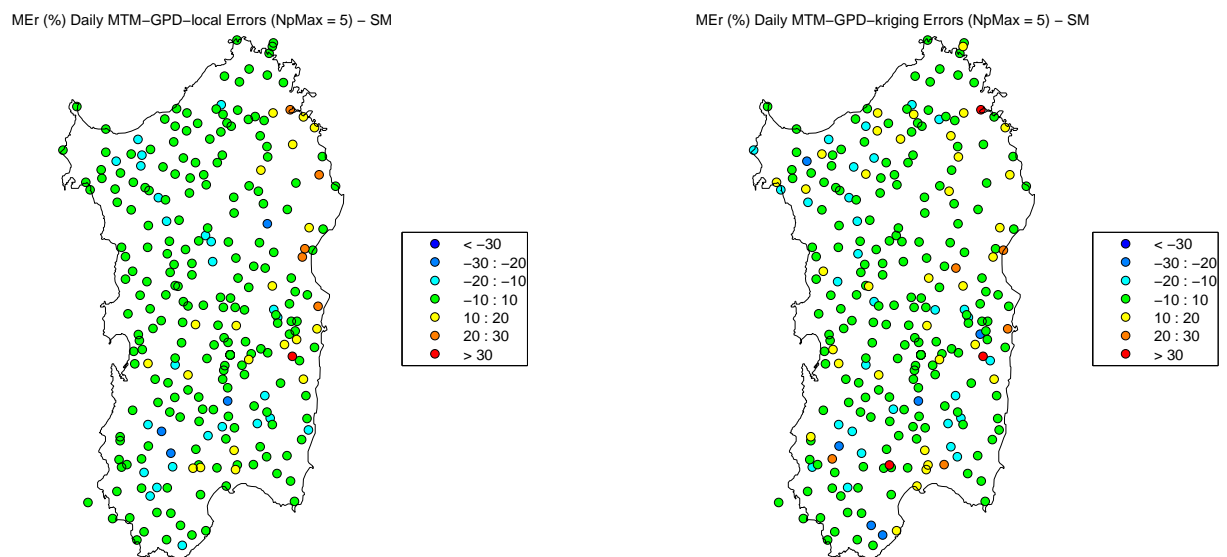


Figure 8.14: Spatial distribution of the error metric $MEr(5)$ for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution (right).

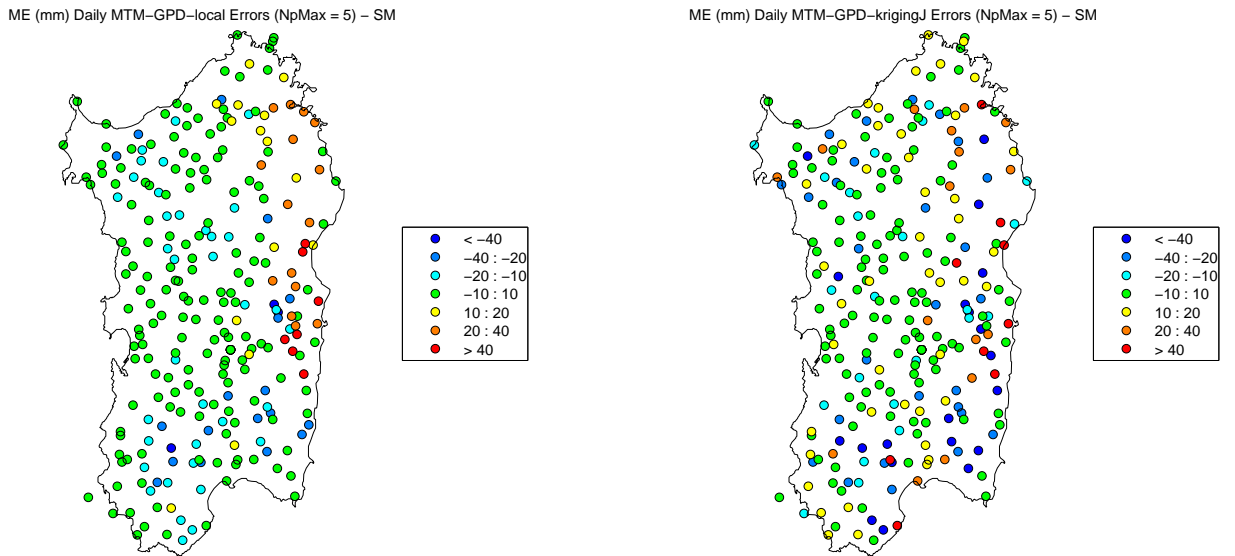


Figure 8.15: Spatial distribution of the error metric $ME(5)$, for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution, calculated after the **cross-validation** procedure (right).

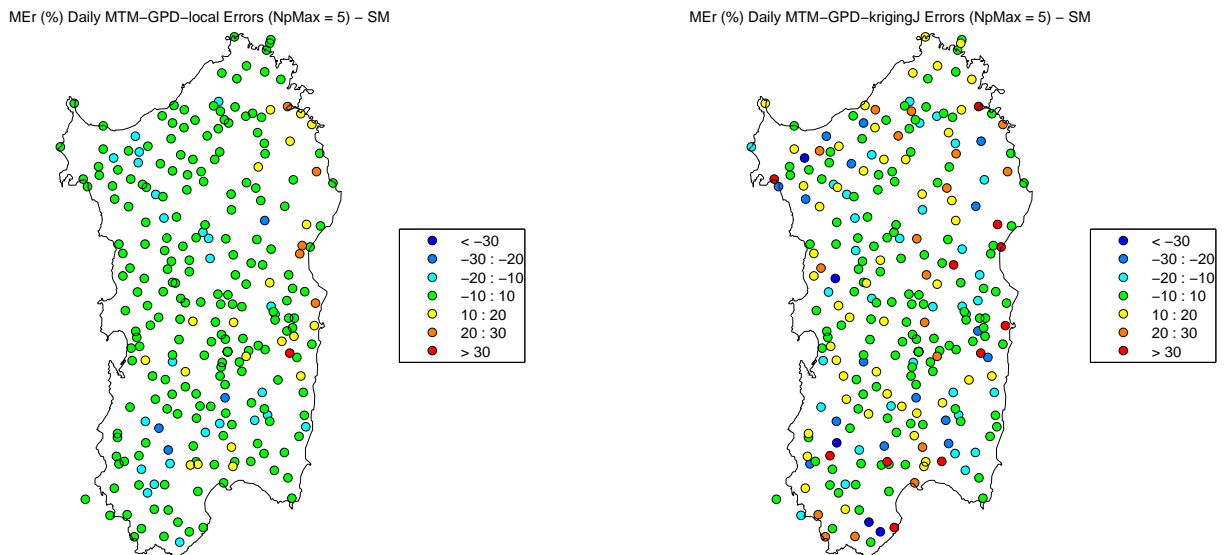


Figure 8.16: Spatial distribution of the error metric $MEr(5)$, for the cases: MTM-GP with local parameters (left), MTM-GP with parameters estimated by kriging on a regular grid at 1 km resolution, calculated after the **cross-validation** procedure (right).

8.4 POT and BM comparison

Figure 8.17 shows the comparison between rainfall depths h_T exceeded with different time return periods (10,50,100 and 200 years), obtained with the POT and BM approaches. In the x-axis, for each sub-figure, are reported the quantiles obtained with geostatistical MTM-GP model, whose parameters were estimated through the SM technique. In the y-axis are reported the quantiles obtained with geostatistical GEV model whose parameters were estimated through the PWM technique. The quantiles are estimated in measurement point with at least 50 complete years of observations, the same used for the estimates of the GEV growth curve parameters.

We can notice how, when time return period increases MTM-GP model tends to give higher quantiles than GEV model, for a certain number of stations. This happens because estimates of the shape parameter of the MTM-GP model are slightly higher than those of GEV growth curve in the East part of Sardinia.

Figure 8.18 shows the map of the rainfall depth h_T (mm) exceeded with return period $T=200$ yr using the two different models. The comparison between the maps shows that the location of the stations discussed before is effectively in the East part of the island. In fact, in that zone we can observe how MTM-GP model (POT approach) provides higher quantiles than GEV model (BM approach).

In order to compare the results of the two models, we decided to study the goodness of fit of the GEV and MTM-GP distributions respect to the observed time series of annual maxima of daily rainfall, measured on the rain gauges with at least 50 complete years of observations. This comparison is possible thanks to equation (4.46) which relates the CDF of annual maxima $G(x)$ to the CDF of daily rainfall $F(x)$. In fact from equation (4.46) it is possible to write:

$$F(x) = [G(x)]^{\frac{1}{n}} \quad (8.1)$$

where $n = 365.25$ is the average number of days in a year. In equation (8.1) we used as $G(x)$ the empirical distribution function (calculated with Hazen's plotting position) of annual maxima of daily precipitation. Known the values of probability $F(x)$, through equation (8.1), we obtained the relative quantiles by using the kriging estimates of the MTM-GP model. These goodness of fit of these estimates were tested by using the error metrics $ME(5)$, $MEr(5)$, $MAEr(5)$ and $MAEr(5)$, calculated on the 5 highest observed values for each station.

Regarding the estimates obtained with the BM approach, we use the GEV model whose parameters were estimated using the kriging procedure,

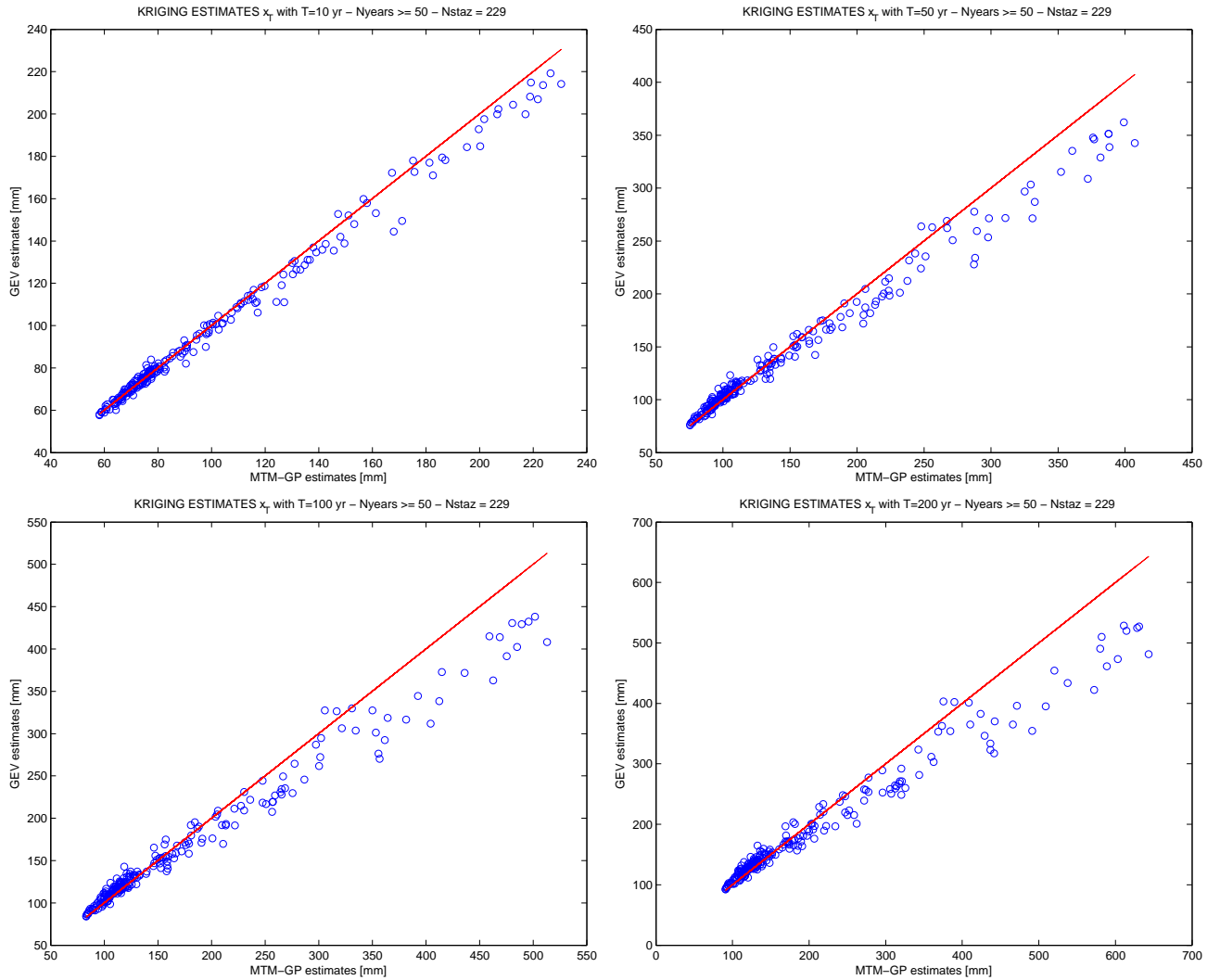


Figure 8.17: Comparison between rainfall depths h_T exceed with different time return periods: 10yr (top left), 50yr (top right), 100yr (bottom left) and 200yr (bottom right) for each of the 229 stations with at least 50 complete years of observations. In the x-axis, for each figure, it is reported the rainfall depths obtained with the MTM-GP geostatistical model. In the y-axis it is reported the rainfall depths obtained with the GEV geostatistical model.

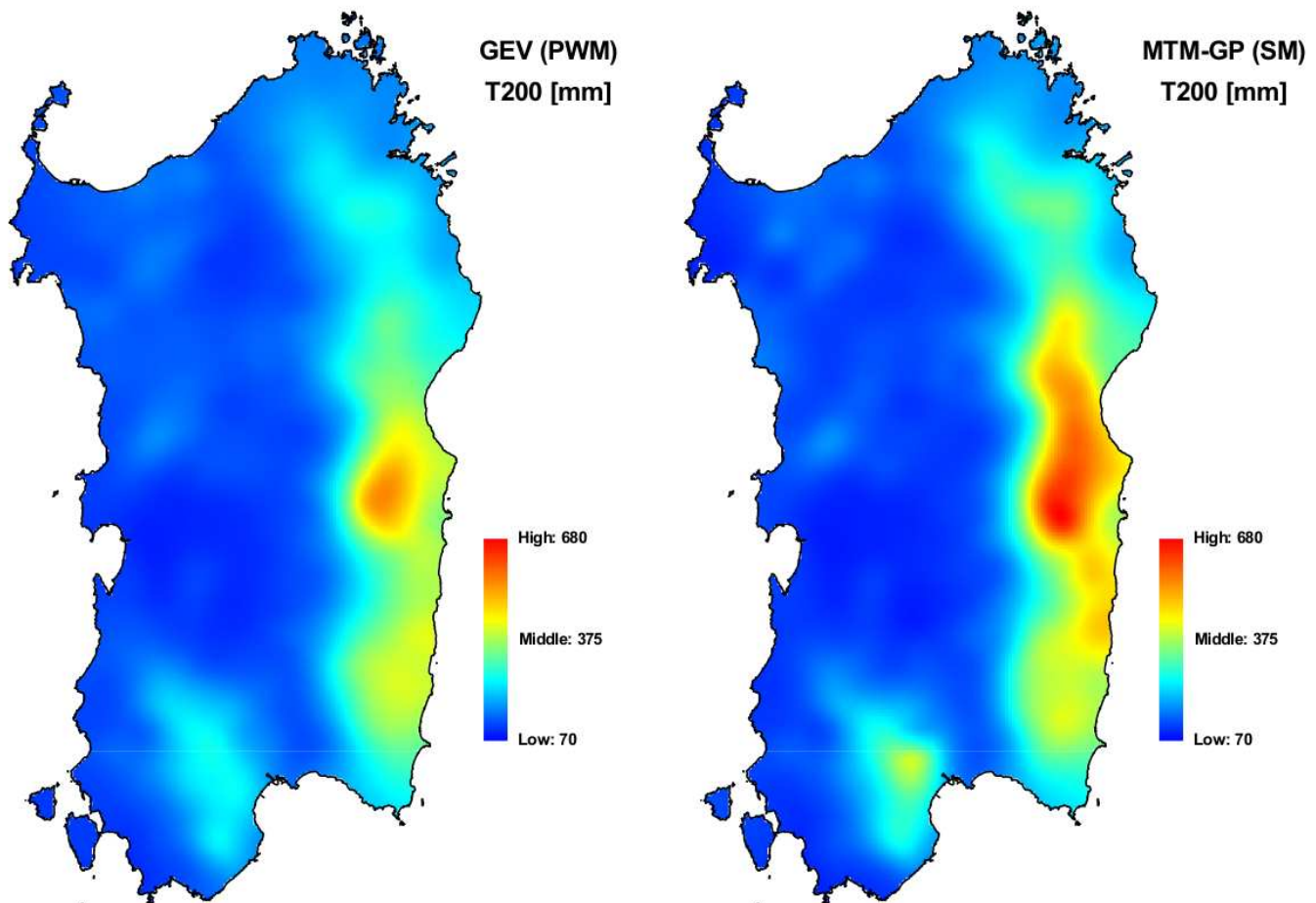


Figure 8.18: Map of daily rainfall depth h_T (mm) exceeded with return period $T=200$ yr. Left: the result obtained using the BM approach and the GEV geostatistical model. Right: the result obtained using the POT approach and the MTM-GP geostatistical model.

and with the index-rainfall obtained from the grid illustrated in Figure 7.44 (in section 7.4 we used as index-rainfall the local mean instead of the grid values, in order to make a fair comparison between regional and geostatistical approaches).

A synthetic comparison of the performances of the BM and POT approaches is presented in Table 8.3, where average values of error metrics calculated for the 229 stations with more than 50 complete years of observation are reported. From the results presented in Table 8.3, we can conclude that the MTM-GP geostatistical model has, on average, the same goodness of fit of the GEV geostatistical model.

	MAE(5) [mm]	MAEr(5) [-]
GEV kriging (PWM)	15.072	0.100
MTM-GP kriging (SM)	15.660	0.103

Table 8.3: Comparisons between the performances of kriging fit using the GEV and MTM-GP models. Average of error metrics calculated over the 229 stations with at least 50 complete years of observations.

Chapter 9

Conclusions

The main contributions of this research work are summarized in this final chapter. The primary objective of this thesis is to describe the extreme rainfall regime in Sardinia by using different approaches. We used daily rainfall time series from 256 rain gauges heterogeneously distributed throughout the whole Sardinian territory. Each time series has at least 30 complete years of observation, in the period from 1920 to 2008. Among these stations, 229 of them have at least 50 complete years of observation. The performance of the proposed methodologies have been evaluated using two different type of error metrics. The metrics of the first class measure square errors between cumulative distribution functions, whereas those of the second class measure the errors in reproducing the highest observed values.

Block Maxima (BM) approach

First off we used a block maxima (BM) approach, determining the probabilistic distributions that best fit the empirical distribution of annual maxima of daily precipitation. Analyzing the time series of the 229 stations with more than 50 complete years of observation, we choose to adopt the Generalized Extreme Values (GEV) distribution to represent the observed data. This choice was driven by the results of a preliminary analysis conducted using the four parameters distribution *Kappa*, and the L-moment ratios diagram. We observed that the *Kappa* distribution degenerated in a GEV distribution; and the L-moment ratios diagram suggests the use of the GEV distribution compared to other distributions widely used in statistical hydrology (Generalized Pareto, Generalized Logistic, Lognormal, and Pearson type III).

The estimates of the GEV parameters were performed by using maximum likelihood (ML), simple moments (SM) and probability weighted moments (PWM) estimation techniques. Regarding the shape parameter estimates,

ML and PWM provided similar results, while the SM estimates were generally lower than those obtained with the other two methods, and always less than 0.30. That's because SM estimates of κ depend on the third moment (skewness) and are unable to give estimates $> 1/3$, see section 4.1.1. So SM technique is not suitable for estimating the shape parameter of the GEV distribution, especially if observed data present high skewness.

We obtained the best local fit with the PWM estimates and decided to use them in the regional and geostatistical analysis.

Regional model

We performed a regional analysis based on the index-rainfall methodology, widely used in hydrology. The concept underlying the index-rainfall method is that the distribution of rainfall at different sites in a homogeneous region is the same for each site, except for a scale parameter, the index-rainfall, which varies from site to site. We used the locally observed mean as index-rainfall. The performance of the GEV regional model was compared with that obtained by using the TCEV distribution, which is commonly applied in Italy for the regional frequency analysis of annual maxima of daily precipitation. The regional estimates of the TCEV parameters were taken from Deidda and Piga (1998); Deidda et al. (2000) and updated using the new database of daily precipitations available till 2008, in order to make a fair comparison.

Several hypothesis of subdivision in homogeneous regions were obtained by using cluster analysis and the homogeneity tests described in section 5.1. However each of them led, after some empirical modifications, to a common configuration with 5/4 homogeneous regions, see Figure 7.29 and 7.35. The GEV regional model fits observed data better than the TCEV model. This is an important result because the GEV distribution can be determined through three parameters, which make it more manageable and more simply applicable in engineering practice respect to the TCEV distribution that has four parameters. Furthermore the fewer parameters the lower estimation errors.

Geostatistical model

With a geostatistical approach we represented with continuity the spatial variability of the GEV growth curve parameters locally estimated with the PWM technique and the index-rainfall. Regular grids with 1-km step were determined with the technique of kriging for uncertain data (KUD). The KUD is a very useful tool for spatial interpolation of metrics affected by high uncertainty, like GEV parameters, and provides smoother maps than those obtained with the ordinary kriging (OK). By using this grid we were able

to obtain maps of annual maxima of daily rainfall depth depending on the non-exceedance probability (time return period).

Regional vs Geostatistical

The comparisons between regional and geostatistical approaches were conducted by using the cross-validation methodology. The results indicates that the best fit is obtained using the GEV parameters given by the grids obtained through the KUD. We must highlight that this result is not globally valid. In fact, changing some decisions during the identification of homogeneous regions, the gap between the two approaches could decrease, or we could even obtain opposite results. Moreover the same methods applied to different parts of the world could lead to different result, especially in limited extension areas characterized by slight spatial heterogeneity. This highlights one of the weaknesses of the regional approach: the presence of subjective components that strongly influence the final result. For example subjective choices in the identification of homogeneous regions, related to: aggregation criteria, homogeneity check, inevitable subjective choices in merging/splitting clusters and reassigning stations. Instead, by using a geostatistical approach the only subjective component concerns the choice of the spatial interpolation technique to use. Furthermore, and this is always true, the geostatistical approach represents local peculiarities better than the regional one and overcomes the problems associated with abrupt discontinuities in the parameters of probability distribution on the border between contiguous homogeneous regions, see Figure 7.46.

Peaks Over Threshold (POT) approach

In a second phase we used a peaks over threshold (POT) approach, determining the probabilistic distributions that best fit the empirical distribution of daily precipitation in Sardinia, with particular interest for the tail of distribution. We used the 256 stations with more than 30 complete years of observation. The L-moment ratios diagram in Figure 8.1 shows that the best candidate to represent the daily rainfall depths higher than 5 mm is the Generalized Pareto (GP) distribution. We proposed a model based on the reparametrization of the GP distribution. The parameters of the new distribution are threshold invariant, and their estimates were obtained by using the multiple threshold method (MTM) proposed by Deidda (2010). So we labeled this distribution as “MTM-GP”. This model describes the distribution of all $x \geq 0$.

We decided to use this model mainly for three reason.

The first one concerns the problems related to the choice of the optimum threshold to fit the GP distribution. Several methods are available in literature for this purpose, but a general consensus has not been reached yet and proposed methods can lead to different results. In addition most of them have strong limitations when dealing with large databases because they are graphical models or present computational problems.

The second reason is the presence of roughly rounded-off records in many time series of our database. This makes the determination of the optimum threshold even more difficult, if not impossible, as highlighted in Deidda and Puliga (2006); Deidda (2007).

The third one is the fact that by using a GP distribution four parameters for each site should be determined: the optimal threshold u^* , the shape ξ and scale α_{u^*} parameters and the probability ζ_{u^*} to observe exceedances over the threshold. But α_{u^*} and ζ_{u^*} estimates are not the best indicators of climatological spatial patterns, because of their dependence on the optimal threshold u^* . The MTM-GP distribution overcome this problems because its parameters don't depend on the optimum threshold so they only reflect the climatic signature. In addition the MTM is particularly able to filter out the deviations from threshold invariance which are artificially driven by the presence of roughly rounded-off records. Another important property of the MTM-GP is that its estimates are not affected by small errors in the location of the optimum threshold, as it conversely happens for the single-threshold standard fit.

The estimates of the parameters were obtained, as in the BM analysis, by using SM, ML and PWM estimation techniques. We observed that the SM is the only possible estimation technique, if the interest is focused in the tail of the distribution. In fact, for each station we get thousands of data, and sometimes it can happen that the tail of distribution strongly differs from the bulk. By using SM technique the shape parameter estimates are more conditioned to the extreme values compared to the other two techniques. PWM and ML tends to strongly overestimate the values of the shape parameter in the East zone of Sardinia, which means a strong overestimation in the quantiles too.

So the MTM-GP parameters were locally estimated by using SM technique and successively spatially interpolated by using the same geostatistical approach we had previously applied for the spatial interpolation of the GEV growth curve parameters. Using the MTM-GP we observed that the kriging for uncertain data provides no advantage over the ordinary kriging. This is due to the bigger sample size compared to the BM approach, and to the robustness of the MTM estimates. In fact the MTM estimates are obtained through the median value of parameters values obtained within a

prefixed range of thresholds (Figure 4.2). The MTM-GP model provided robust quantile estimates, the average error in estimating the 5 highest daily rainfall values observed for each station is $\simeq 11\%$ using the MTM-GP model with kriging estimates.

BM vs POT

In the last part of the research project the goodness of fit of the GEV and MTM-GP models, using the geostatistical approach, were compared by measuring the errors in reproducing extreme events. Comparisons were made towards observed time series of annual maxima of daily rainfall, measured on the 229 rain gauges with at least 50 complete years of observations. This has been possible thanks to equation (4.46) which relates the distribution function of annual maxima $G(x)$ to the distribution function of daily rainfall $F(x)$. We found out that the two different models provide similar and satisfactory results. The average error in estimating the 5 highest observed annual maxima for each station is $\simeq 10\%$ with both approaches.

Bibliography

- Alila, Y.: A hierarchical approach for the regionalization of precipitation annual maxima in Canada, *Journal of Geophysical Research*, 10, 31,645–31,655, 1999.
- Baeriswyl, P. A. and Rebetez, M.: Regionalization of Precipitation in Switzerland by Means of Principal Component Analysis, *Theoretical and Applied Climatology*, 58, 31–41, 1997.
- Beaudoin, P. and Rousselle, J.: A STUDY OF SPACE VARIATIONS OF PRECIPITATION BY FACTOR ANALYSIS, *Journal of Hydrology*, 59, 123–138, 1982.
- Beguiria, S. and Vicente-Serrano, S.: Mapping the Hazard of Extreme Rainfall by Peaks over Threshold Extreme Value Analysis and Spatial Regression Techniques, *Journal of Applied Meteorology and Climatology*, 45, 108–124, 2006.
- Bezak, N., Brilly, M., and Sraj, M.: Comparison between the peaks-over-threshold method and the annual maximum method for flood frequency analysis, *Hydrological Sciences Journal*, 59, 59–977, doi:10.1080/02626667.2013.831174, 2014.
- Blanchet, J. and Lehning, M.: Mapping snow depth return levels: smooth spatial modeling versus station interpolation, *Hydrology and Earth System Sciences*, 14, 2527–2544, doi:doi:10.5194/hess-14-2527-2010, 2010.
- Cannarozzo, M., D'asaro, F., and Ferro, V.: Regional rainfall and flood frequency analysis for Sicily using the two component extreme value distribution, *Hydrological Sciences Journal*, 40, 19–42, 1995.
- Castillo, E.: *Extreme Value Theory in Engineering*, Academic Press, Inc., 1988.

- Ceresetti, D., Ursu, E., Carreau, J., Anquetin, S., Creutin, J. D., Gardes, L., Girard, S., and Molinie, G.: Evaluation of classical spatial-analysis schemes of extreme rainfall, *Natural Hazards and Earth System Sciences*, 12, doi: 10.5194/nhess-12-3229-2012, URL www.nat-hazards-earth-syst-sci.net/12/3229/2012/, 2012.
- Chessa, P. A., Cesari, D., and Delitala, A. M. S.: Mesoscale Precipitation and Temperature Regimes in Sardinia (Italy) and their Related Synoptic Circulation, *Theoretical and Applied Climatology*, 1999.
- Chessa, P. A., Ficca, G., Marrocu, M., and Buizza, R.: Application of a Limited-Area Short-Range Ensemble Forecast System to a Case of Heavy Rainfall in the Mediterranean Region, *Weather and Forecasting*, 19, 566–581, 2004.
- Coles, S.: *An Introduction to Statistical Modeling of Extreme Values*, Springer, 2001.
- Coles, S., Pericchi, L. R., and Sisson, S.: A fully probabilistic approach to extreme rainfall modeling, *Journal of Hydrology*, 273, 35–50, 2003.
- Comrie, A. C. and Glenn, E. C.: Principal components-based regionalization of precipitation regimes across the southwest United States and northern Mexico, with an application to monsoon precipitation variability, *Climate Research*, 10, 201–215, 1998.
- Dalrymple, T.: Flood-frequency analyses, *Manual of Hydrology: Part 3 Flood-Flow Techniques*, Geological Survey Water-Supply Paper, 1960.
- Davison, A. and Smith, R.: Models for Exceedances over High Thresholds, *Journal of the Royal Statistical Society*, 52, 847–856, 1990.
- De Marsily, G.: *Quantitative hydrogeology: groundwater hydrology for engineers*, Academic Press, Inc., 1986.
- De Michele, G. and Salvadori, G.: Some hydrological applications of small sample estimators of Generalized Pareto and Extreme Value distributions, *Journal of Hydrology*, 301, 37–53, 2005.
- Deidda, R.: An efficient rounding-off rule estimator: Application to daily rainfall time series, *Water Resources Research*, 43, doi:doi:10.1029/2006WR005409, 2007.

- Deidda, R.: A multiple threshold method for fitting the generalized Pareto distribution to rainfall time series, *Hydrology and Earth System Sciences*, 14, 2559–2575, 2010.
- Deidda, R. and Piga, E.: Curve di possibilità pluviometrica basate sul modello TCEV, *Informazione*, 81, 9–14, 1998.
- Deidda, R. and Puliga, M.: Sensitivity of goodness-of-fit statistics to rainfall data rounding off, *Phys, Chem. Earth*, 31, 1240–1250, 2006.
- Deidda, R. and Puliga, M.: Performances of some parameter estimators of the generalized Pareto distribution over rounded-off samples, *Physics and Chemistry of the Earth*, 34, 626–634, 2009.
- Deidda, R., Piga, E., and Sechi, G.: Analisi regionale di frequenza delle precipitazioni intense in Sardegna, *L'Acqua*, 5, 29–38, 2000.
- Ferro, V. and Porto, P.: REGIONAL ANALYSIS OF RAINFALL-DEPTH-DURATION EQUATION FOR SOUTH ITALY, *Journal of Hydrologic Engineering*, 4, 326–336, doi:10.1061/(ASCE)1084-0699(1999)4:4(326), 1997.
- Fiorentino, M. and Gabriele, S.: Distribuzione TCEV: metodi di stima dei parametri e proprietà statistiche degli stimatori, *Geodata*, 25, 1985.
- Fisher, R. and Tippett, L.: On the estimation of the frequency distributions of the largest or smallest member of a sample, *Proceedings of the Cambridge Philosophical Society*, 24, 180–190, 1928.
- Fitzgerald, D. L.: Single station and regional analysis of daily rainfall extremes, *Stochastic Hydrology and Hydraulics*, 3, 281–292, 1989.
- Furcolo, P., Villani, P., and Rossi, F.: STATISTICAL ANALYSIS OF THE SPATIAL VARIABILITY OF VERY EXTREME RAINFALL IN THE MEDITERRANEAN AREA, in: *U.S.- Italy Research Workshop on the Hydrometeorology, Impacts, and Management of Extreme Floods*, 1995.
- Gnedenko, B.: Sur la distribution limite du terme maximum d'une serie aleatoire, *Annals of Mathematics*, 44, 423–453, 1943.
- Grimshaw, S. D.: Computing Maximum Likelihood Estimates for the Generalized Pareto Distribution, *Technometrics*, 35, 185–191, 1993.
- Gumbel, E. J.: *Statistic of Extremes*, Columbia University Press, New York, 1958.

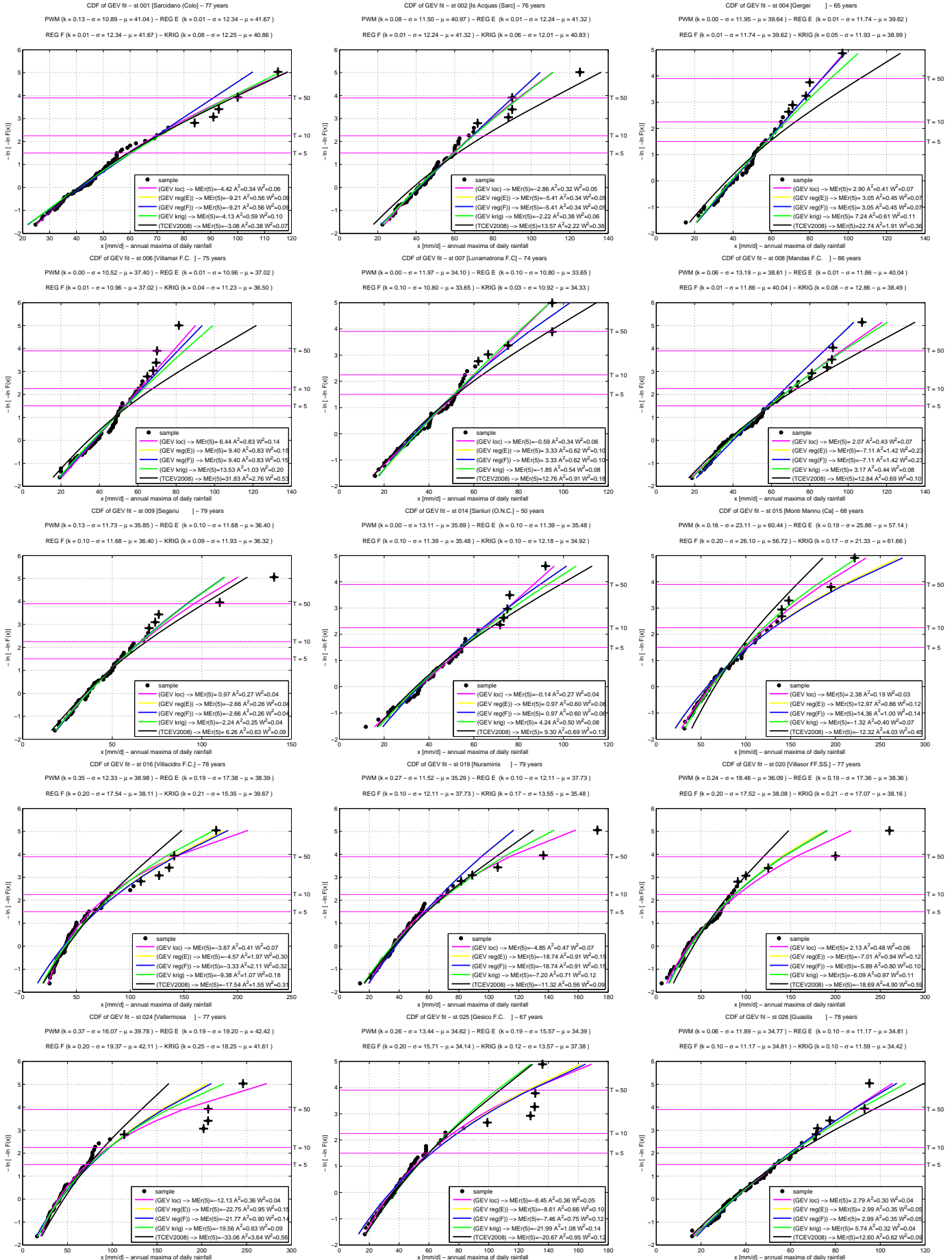
- Hosking, J.: L-moments: Analysis and estimation of distributions using linear combinations of order statistics, *J. R. Statist. Soc. B*, 52, 105–124, 1990.
- Hosking, J. and Wallis, J.: *Regional Frequency Analysis, an approach based on L-moments.*, Cambridge University Press, 1997.
- Hosking, J. and Wallis, J. R.: Parameter and Quantile Estimation for the Generalized Pareto Distribution, *Technometrics*, 29, 339–349, 1987.
- Hosking, J. and Wallis, J. R.: Some statistics useful in regional frequency analysis, *Water Resources Research*, 29, 271–281, correction, *Water Resour. Res.*, 31(1), 251, 1995, 1993.
- Hosking, J., Wallis, J., and Wood: Estimation of the Generalized Extreme-Value Distribution by the Method of Probability-Weighted Moments, *Technometrics*, 27, 251–261, 1985.
- Jenkinson, A. F.: The frequency distribution of the annual maximum (or minimum) value of meteorological elements., *Quarterly Journal of the Royal Meteorological Society*, 1955.
- Karl, T. R., Koscielny, A. J., and Diaz, H. F.: Potential Errors in the Application of Principal Component (Eigenvector) Analysis to Geophysical Data, *Journal of Applied Meteorology*, 21, 1183–1186, 1982.
- Katz, R. W., Parlange, M. B., and Naveau, P.: Statistics of extremes in hydrology, *Advances in Water Resources*, 25, 1287–1304, 2002.
- Kitanidis, P. K.: *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press, 1997.
- Koutsoyiannis, D.: Statistics of extremes and estimation of extreme rainfall: I. Theoretical investigation, *Hydrological Sciences Journal*, 49, 575–590, 2004a.
- Koutsoyiannis, D.: Statistics of extremes and estimation of extreme rainfall: II. Empirical investigation of long rainfall records, *Hydrological Sciences Journal*, 49, 591–610, 2004b.
- Krige, D.: A statistical approach to some basic mine valuation problems on the Witwatersrand, *Journal of Chemical, Metallurgical, and Mining Society of South Africa*, 1951.

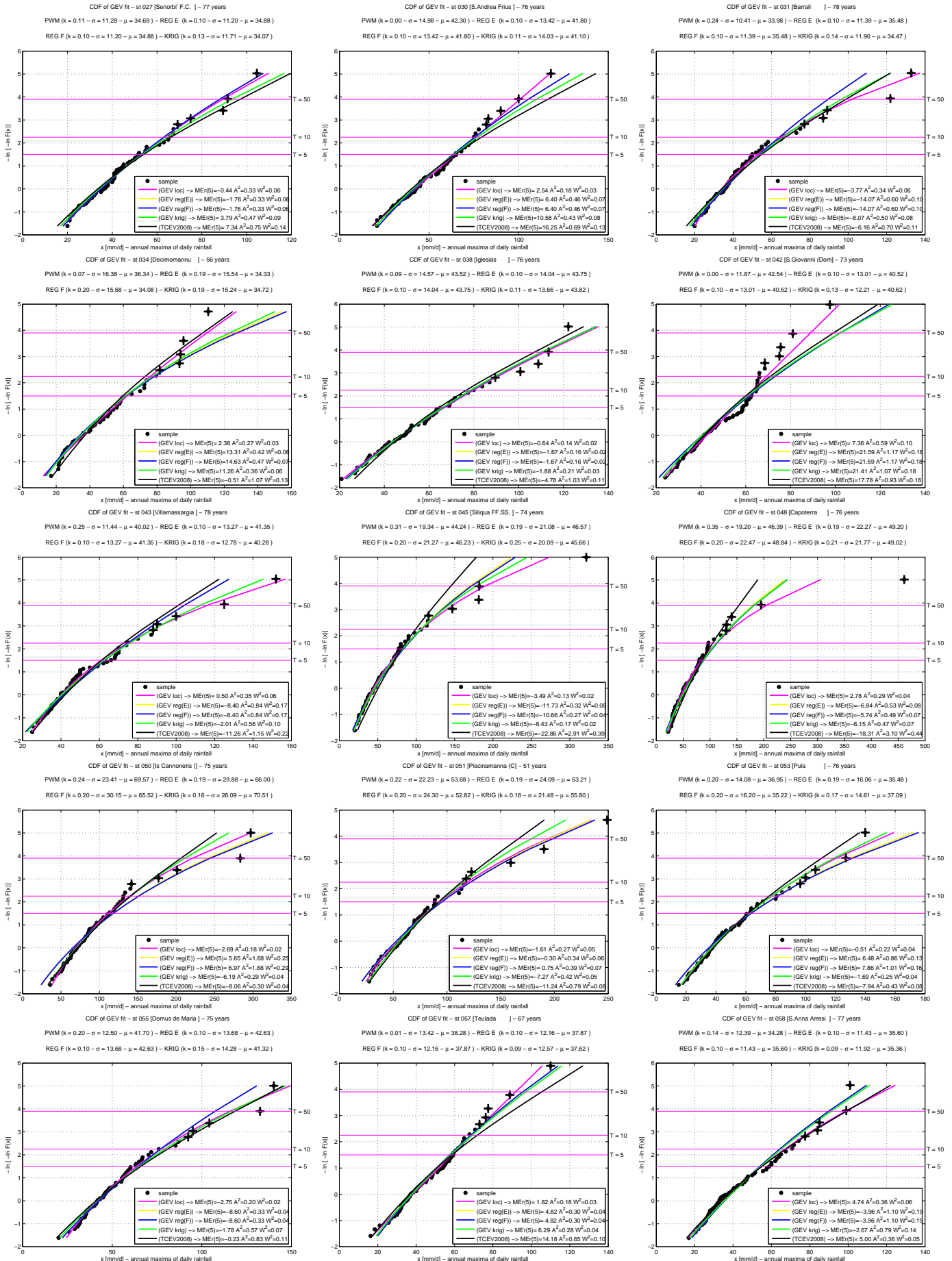
- Lang, M., Ouarda, T. B. M. J., and Bobée, B.: Towards operational guidelines for over-threshold modeling, *Journal of Hydrology*, 225, 103–117, 1999.
- Madsen, H., Pearson, C. P., and Rosbjerg, D.: Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events 2. Regional modeling, *Water Resources Research*, 33, 759–770, 1997a.
- Madsen, H., Rasmussen, P. F., and Rosbjerg, D.: Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events 1. At-site modeling, *Water Resources Research*, 33, 747–758, 1997b.
- Mallants, D. and Feyen, J.: Defining Homogeneous Precipitation Regions by Means of Principal Component Analysis, *Journal of Applied Meteorology*, 29, 892–901, 1990.
- Martins, E. S. and Stedinger, J. R.: Generalized maximum likelihood Pareto-Poisson estimators for partial duration series, *Water Resources Research*, 37, 2551–2558, 2001.
- Matheron, G.: Principles of geostatistics, *Economic Geology*, 58, 1246–66, 1963.
- Mazzetti, C. and Todini, E.: Combining Weather Radar and Raingauge Data for Hydrologic Applications, *Flood Risk Management: Research and Practice Extended Abstracts Volume*, pp. 1345–1348, doi:10.1201/9780203883020.ch159, 2009.
- Munoz-Diaz, D. and Rodrigo, F. S.: Spatio-temporal patterns of seasonal rainfall in Spain (1912-2000) using cluster and principal component analysis: comparison, *Annales Geophysicae*, pp. 1435–1448, 2004.
- Panthou, G., T. Vischel, T. Lebel, J. B. G. Q., and Ali, A.: Extreme rainfall in West Africa: A regional modeling, *Water Resources Research*, 48, doi:10.1029/2012WR012052, 2012.
- Papalexiou, S. and Koutsoyiannis, D.: Battle of extreme value distributions: A global survey on extreme daily rainfall, *Water Resources Research*, 49, 87–201, doi:10.1029/2012WR012557, 2013.
- Papalexiou, S., Koutsoyiannis, D., and Makropoulos, C.: How extreme is extreme? An assessment of daily rainfall distribution tails, *Hydrology and Earth System Sciences*, 17, 851–862, 2013.

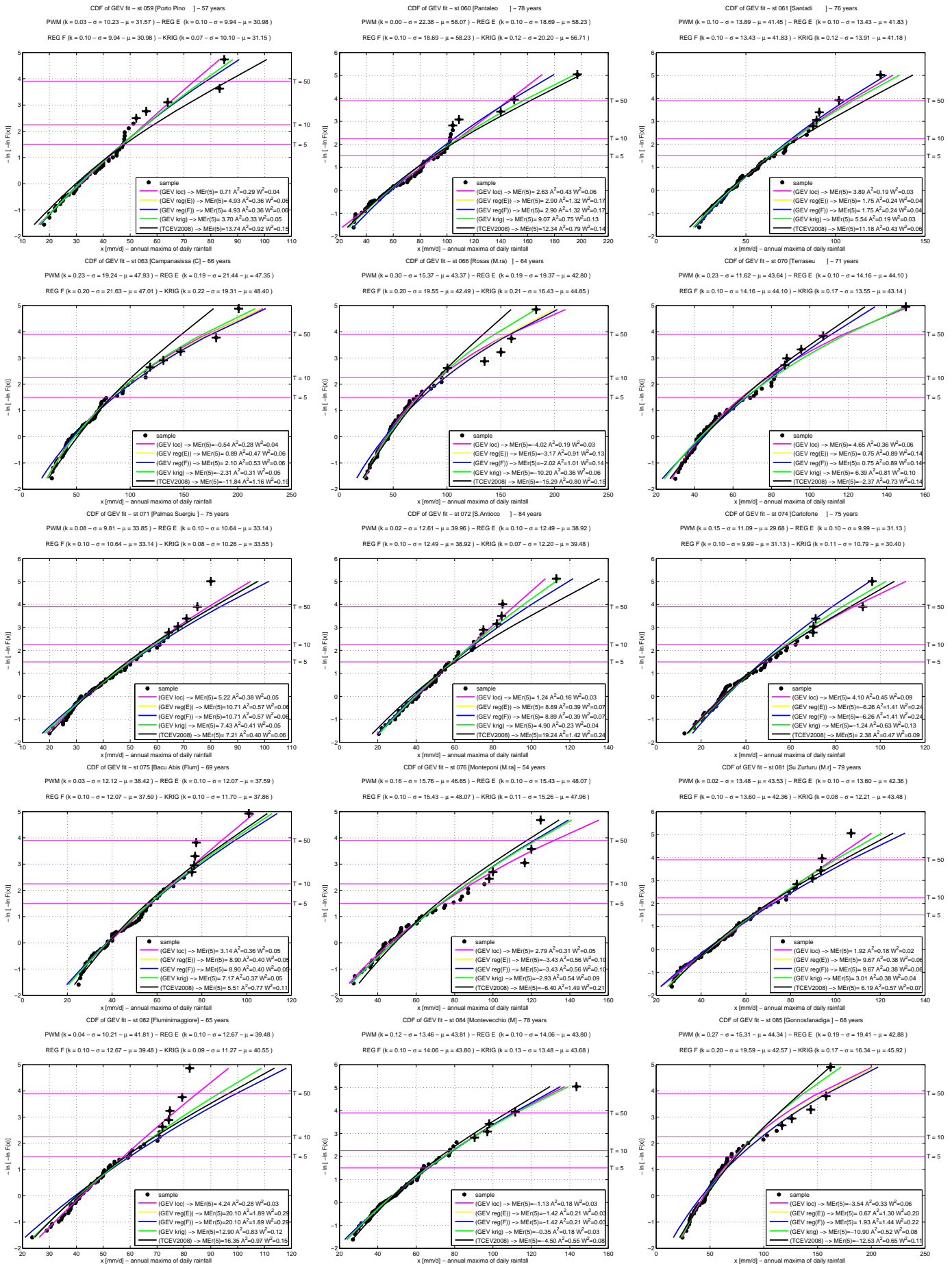
- Pickands, J.: Statistical inference using extreme order statistics, *Annals of Statistics*, 3, 119–131, 1975.
- Pineda-Martinez, L., Carbajal, N., and Medina-Roldan, E.: Regionlization and classification of bioclimatic zones in the central-northeastern region of Mexico using principal component analysis (PCA), *Atmosfera*, 20, 133–145, 2007.
- Prudhomme, C. and Reed, D.: MAPPING EXTREME RAINFALL IN A MOUNTAINOUS REGION USING GEOSTATISTICAL TECHNIQUES: A CASE STUDY IN SCOTLAND, *international Journal of Climatology*, 19, 1337–1356, 1999.
- Rosbjerg, D., Madsen, H., and Rasmussen, P. F.: Prediction in Partial Duration Series With Generalized Pareto-Distributed Exceedances, *Water Resources Research*, 28, 3001–3010, 1992.
- Rossi, F., Fiorentino, M., and Versace, P.: Two component extreme value distribution for flood frequency analysis, *Water Resources Research*, 20, 847–856, 1984.
- Satyanarayana, P. and Srinivas, V.: Regional frequency analysis of precipitation using large-scale atmospheric variables, *Journal of Geophysical Research*, 113, 2008.
- Schaefer, M. G.: Regional analyses of precipitation annual maxima in Washington State, *Water Resources Research*, 26, 119–131, 1990.
- Serinaldi, F. and Kilsby, C. G.: Rainfall extremes: Toward reconciliation after the battle of distributions, *Water Resources Research*, 50, 336–352, doi:10.1002/2013WR014211, 2014.
- Smith, R. L.: Maximum likelihood estimation in a class of nonregular cases, *Biometrika*, 72, 67–90, 1985.
- Stedinger, J., Vogel, R., and Foufoula-Georgiou, E.: *Frequency Analysis of Extreme Events*, McGraw-Hill Book Company, chapter 18, 1993.
- Tanaka, S. and Takara, K.: A study on threshold selection in POT analysis of extreme floods, in: *The Extremes of the Extremes: Extraordinary Floods*, edited by Snorrason, A; Finnsdottir, H. M. M., 271, pp. 299–304, 2002.
- Trefry, C., Jr., D. W. W., and Johnson, D.: Regional Rainfall Frequency Analysis for the State of Michigan, *Journal of Hydrologic Engineering*, 10, 437–449, 2005.

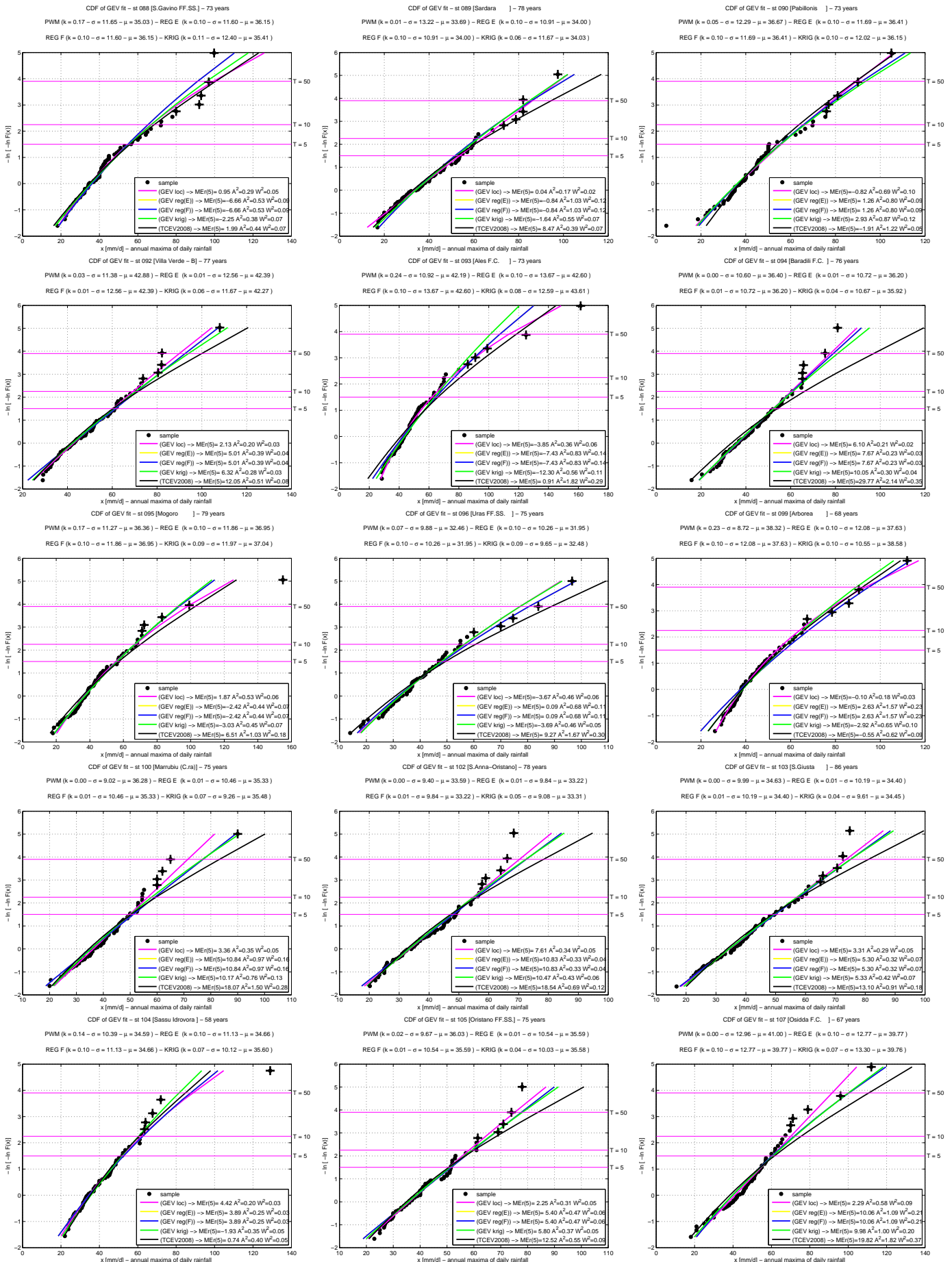
- Van Montfort, M. A. J. and Witter, J. V.: The Generalized Pareto distribution applied to rainfall depths, *Hydrological Sciences Journal*, 31, 151–162, 1986.
- Van Regenmortel, G.: REGIONALIZATION OF BOTSWANA RAINFALL DURING THE 1980s USING PRINCIPAL COMPONENT ANALYSIS, *International Journal of Climatology*, 15, 313–323, 1995.
- Villarini, G., Smith, J. A., Ntelekos, A. A., and Schwarz, U.: Annual maximum and peaks-over-threshold analyses of daily rainfall accumulations for Austria, *Journal of Geophysical Research*, 116, D05 103, doi:10.1029/2010JD015038, 2011.
- Zoglat, A., Adlouni, S. E., Badaoui, F., Amar, A., and Okou, C. G.: Managing Hydrological Risks with Extreme Modeling: Application of Peaks over Threshold Model to the Loukkos Watershed, Morocco, *Journal of Hydrologic Engineering*, 19, doi:10.1061/(ASCE)HE.1943-5584.0000996, 2014.

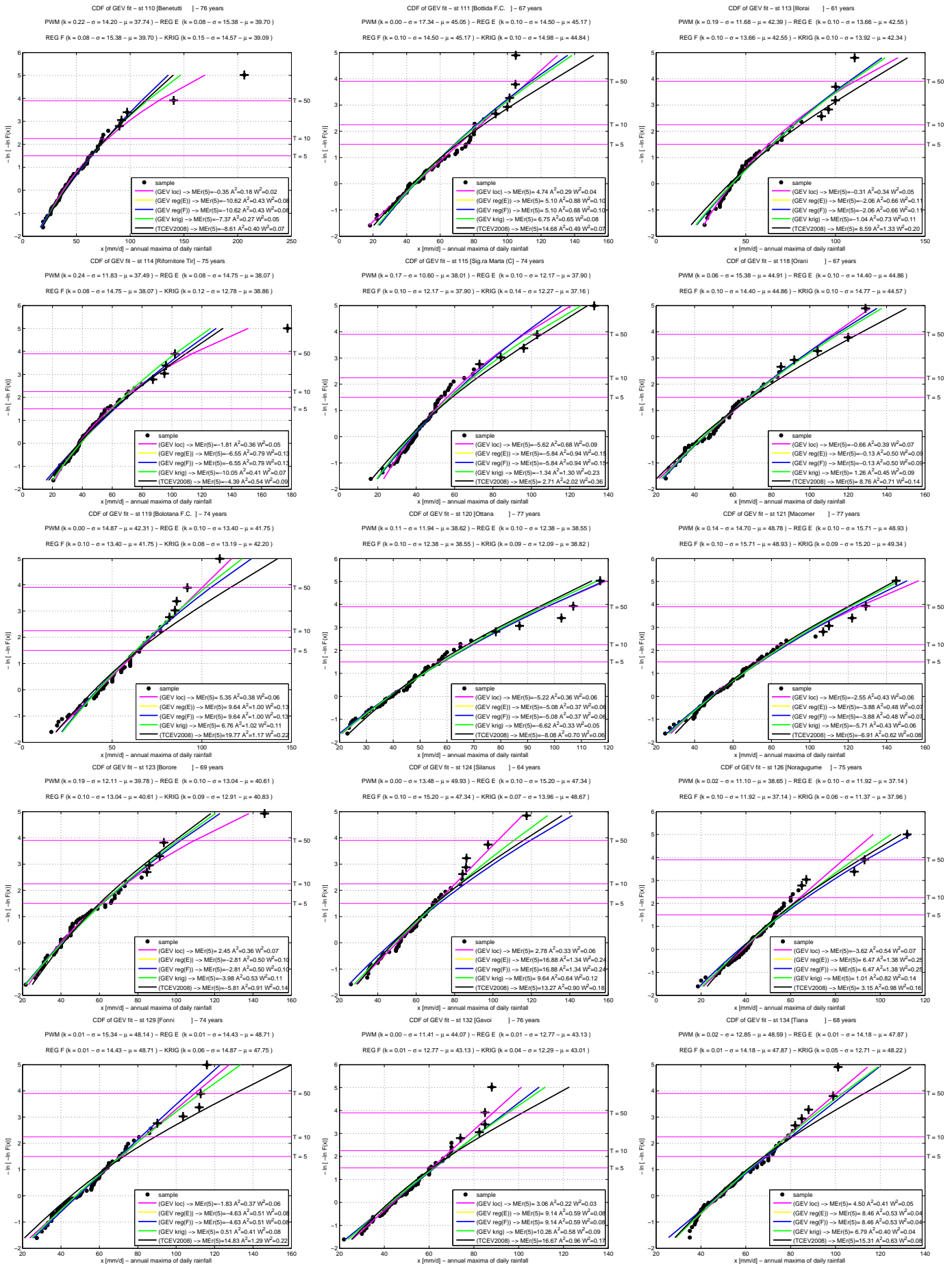
ATTACHMENTS

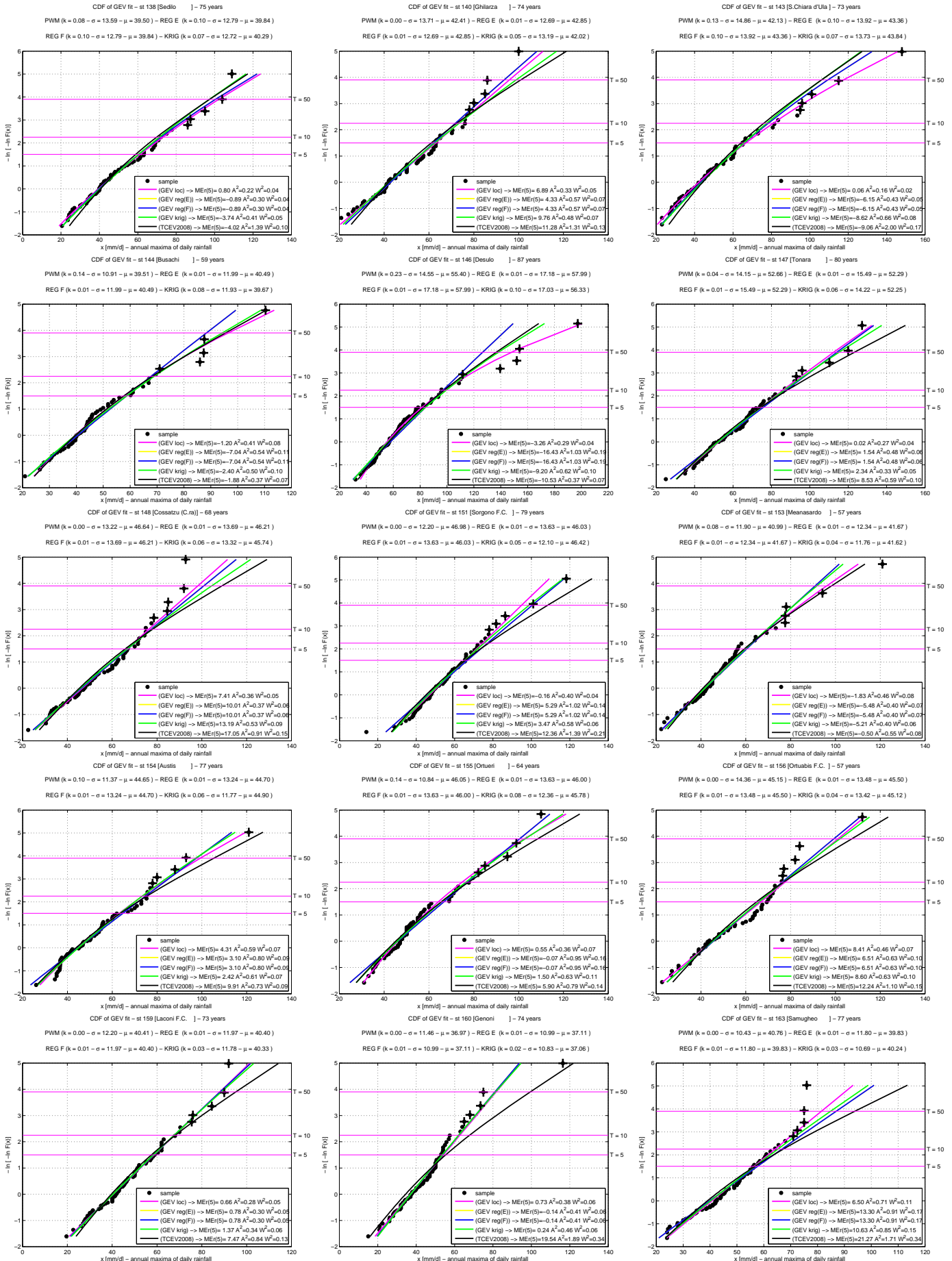


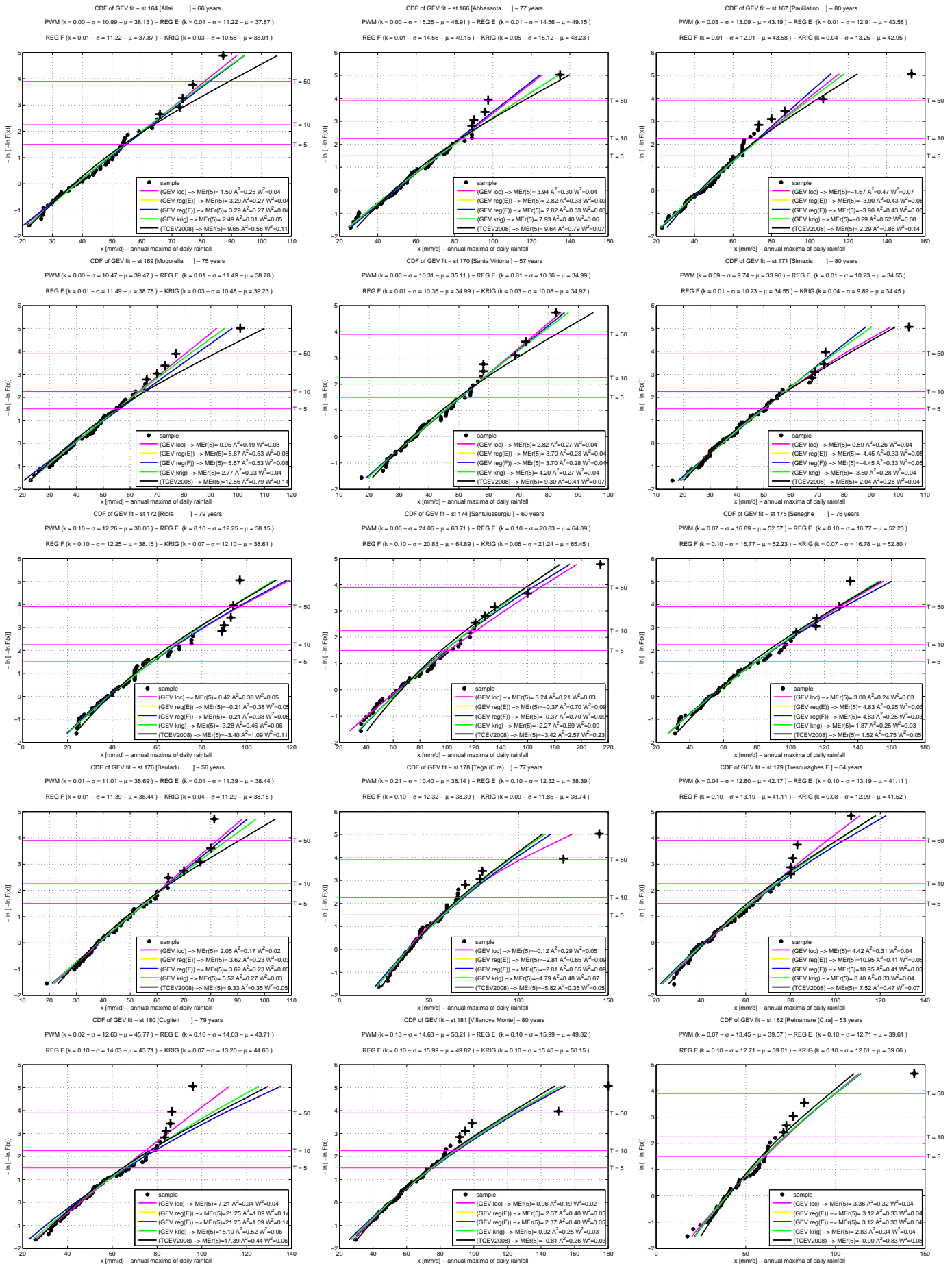


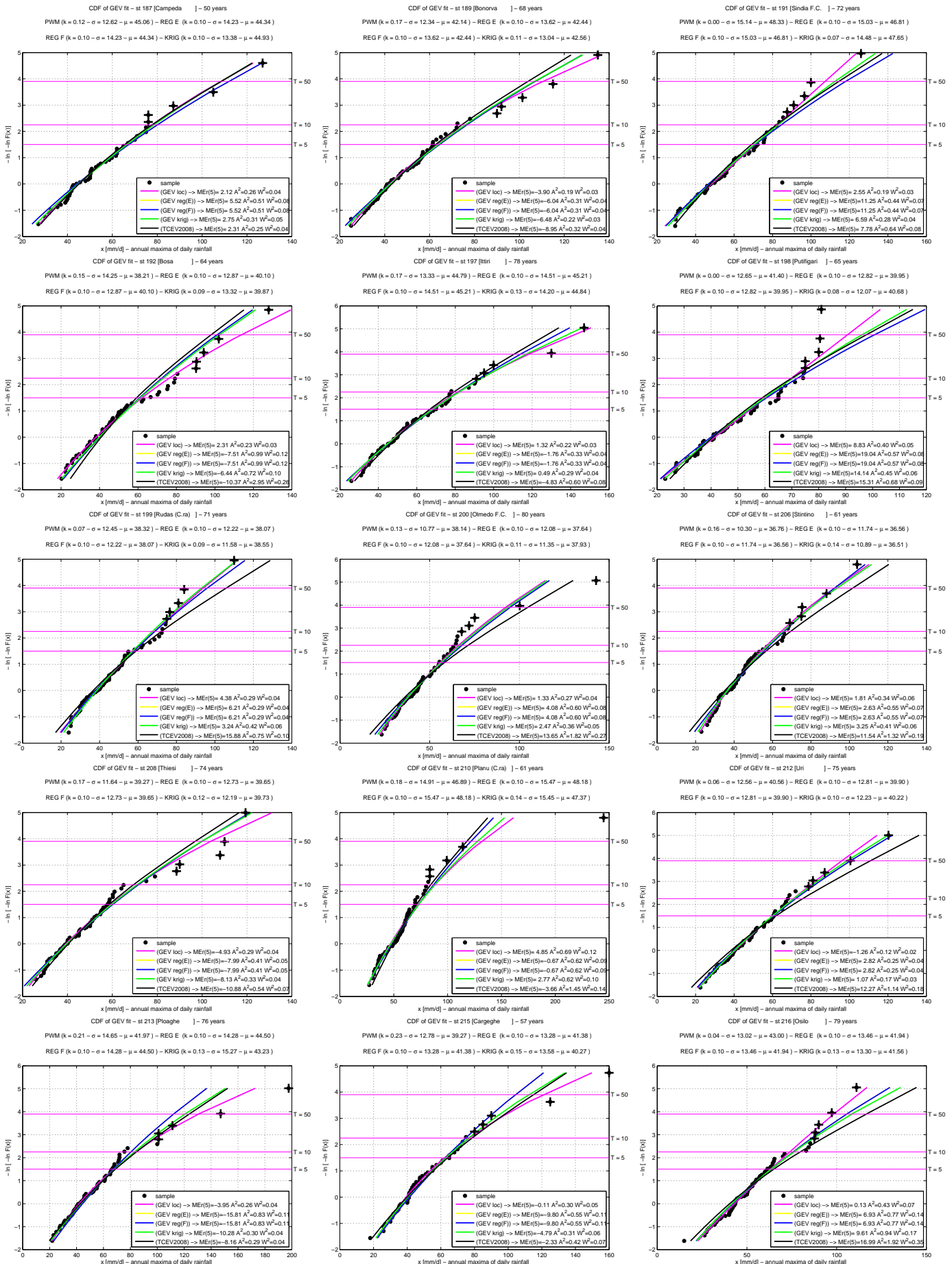


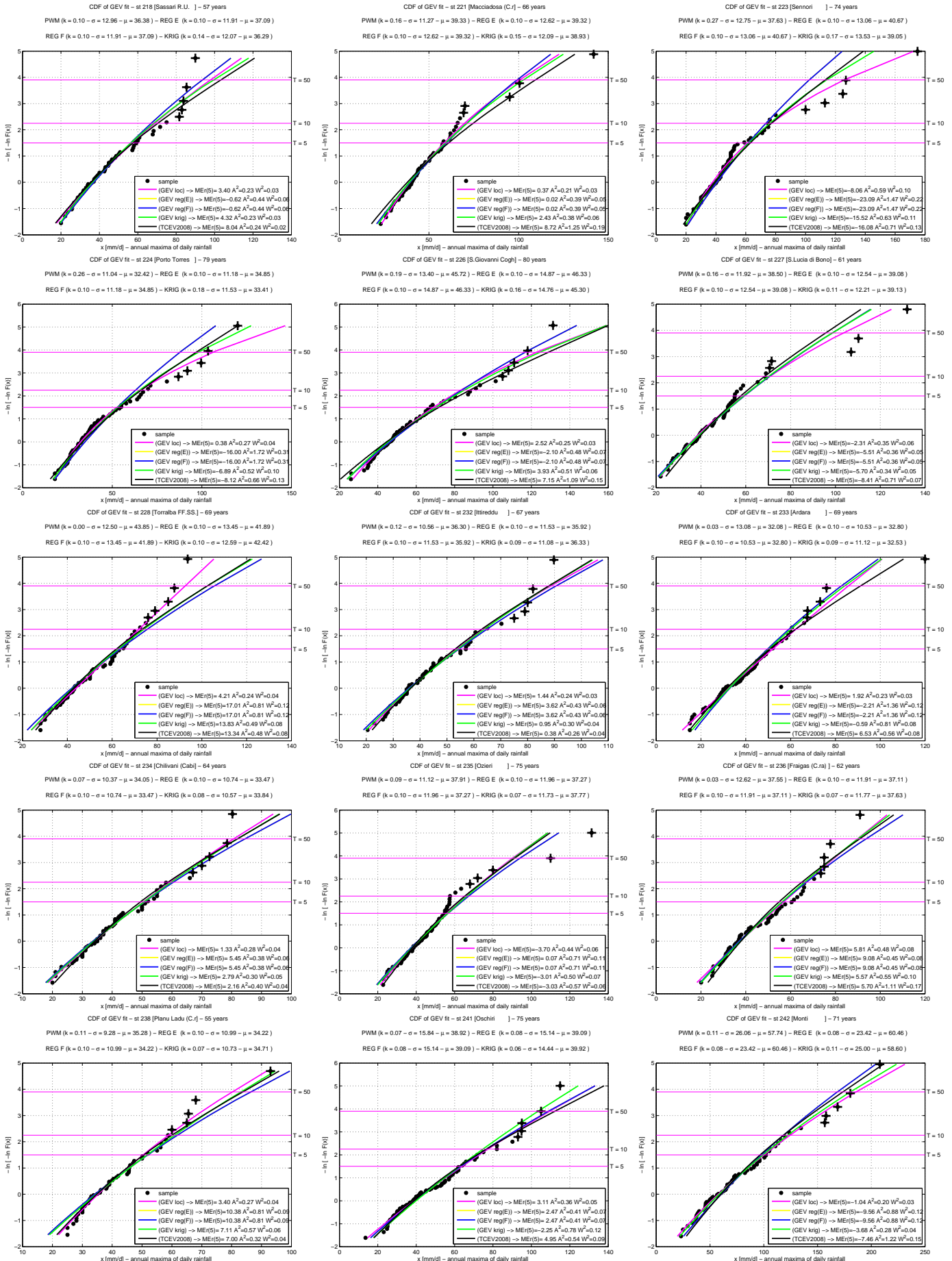


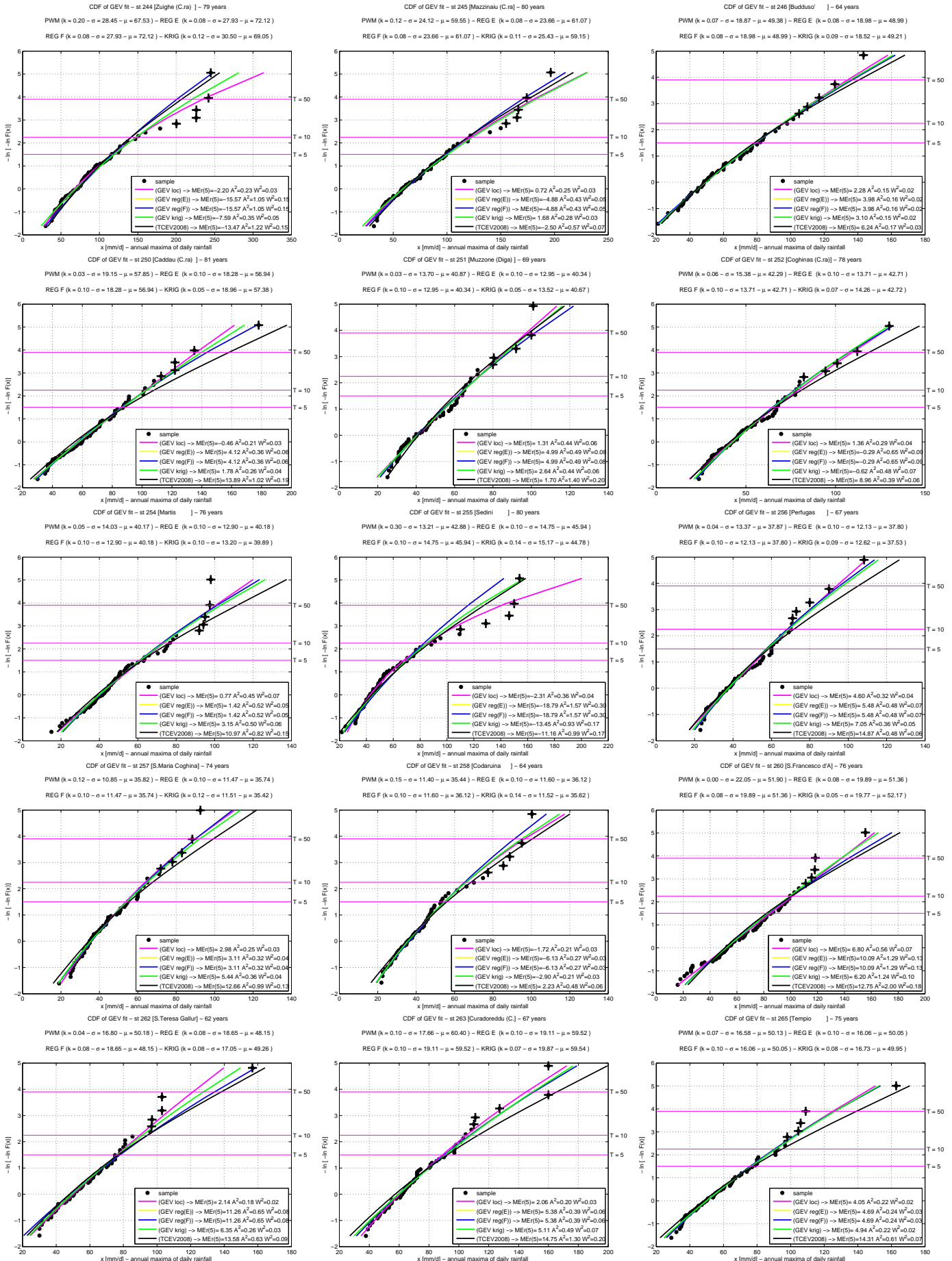


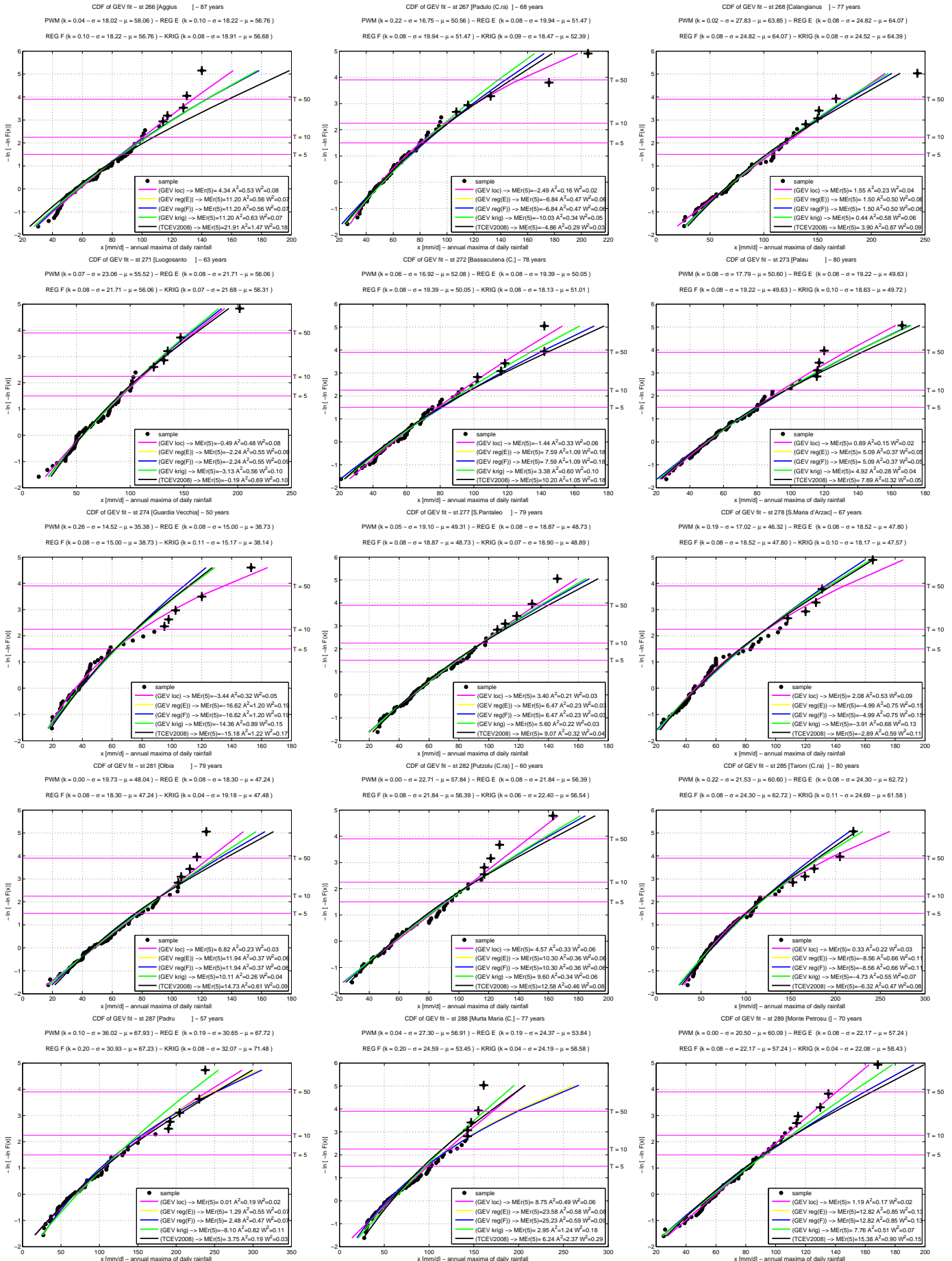


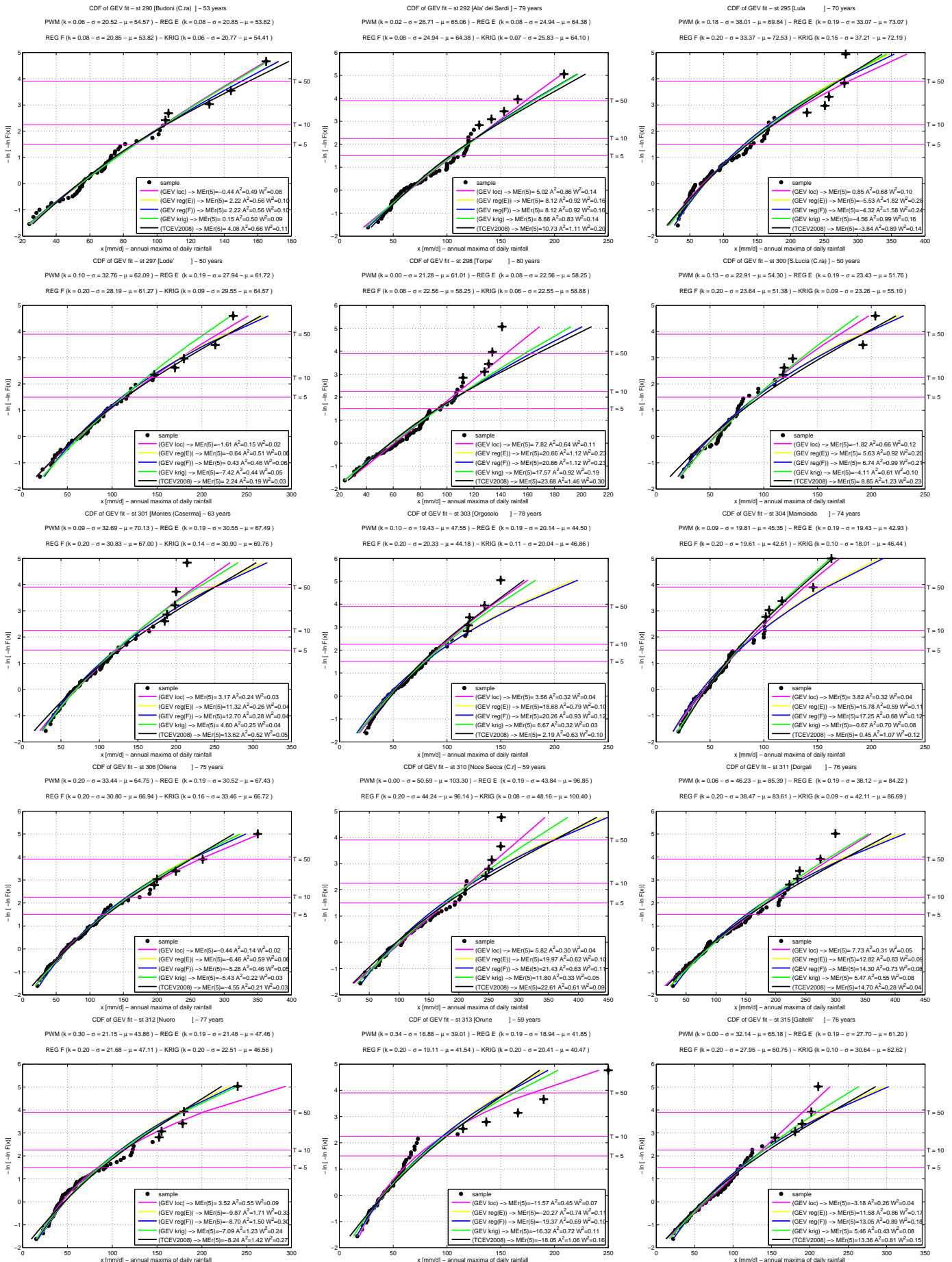


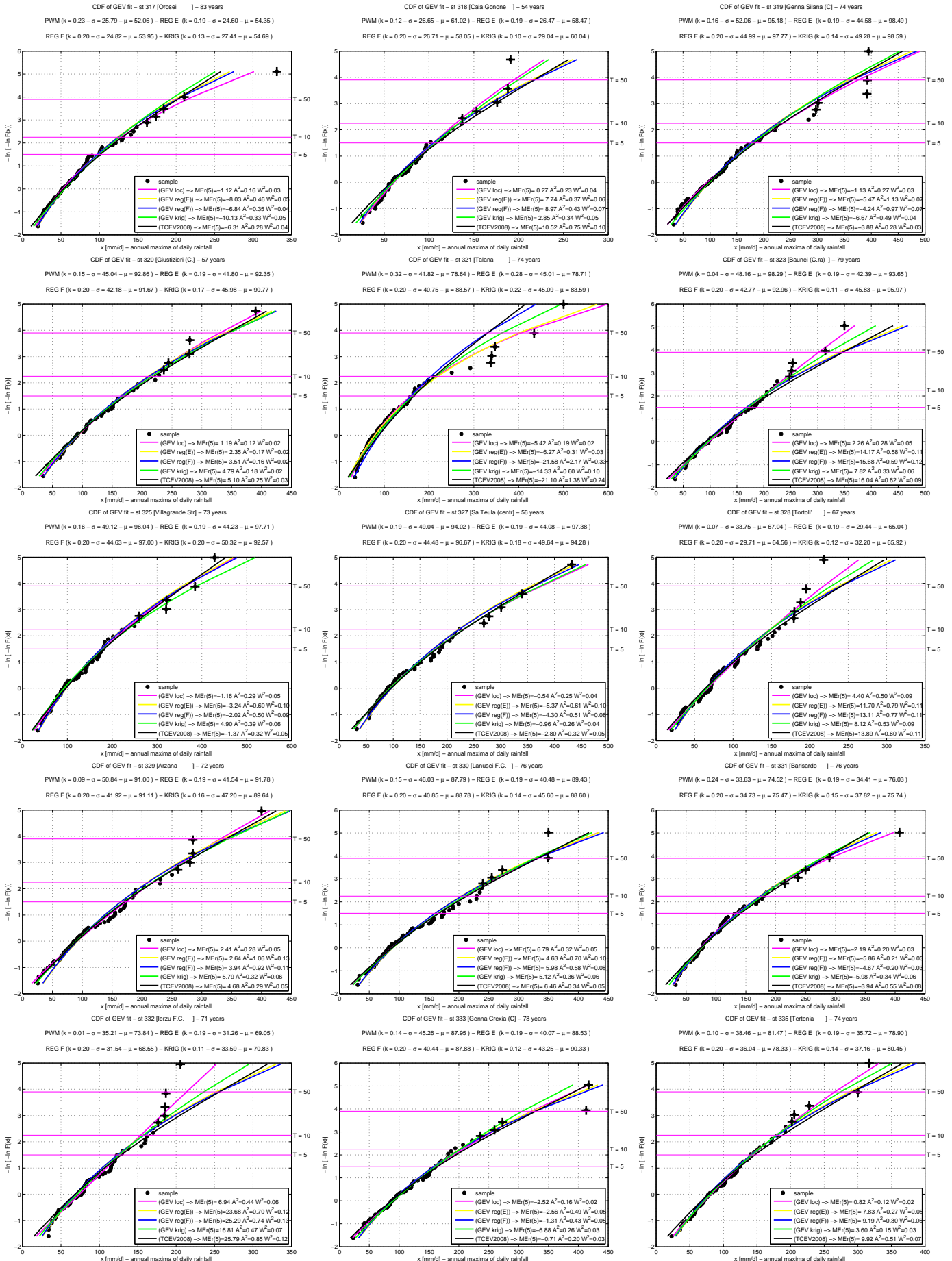


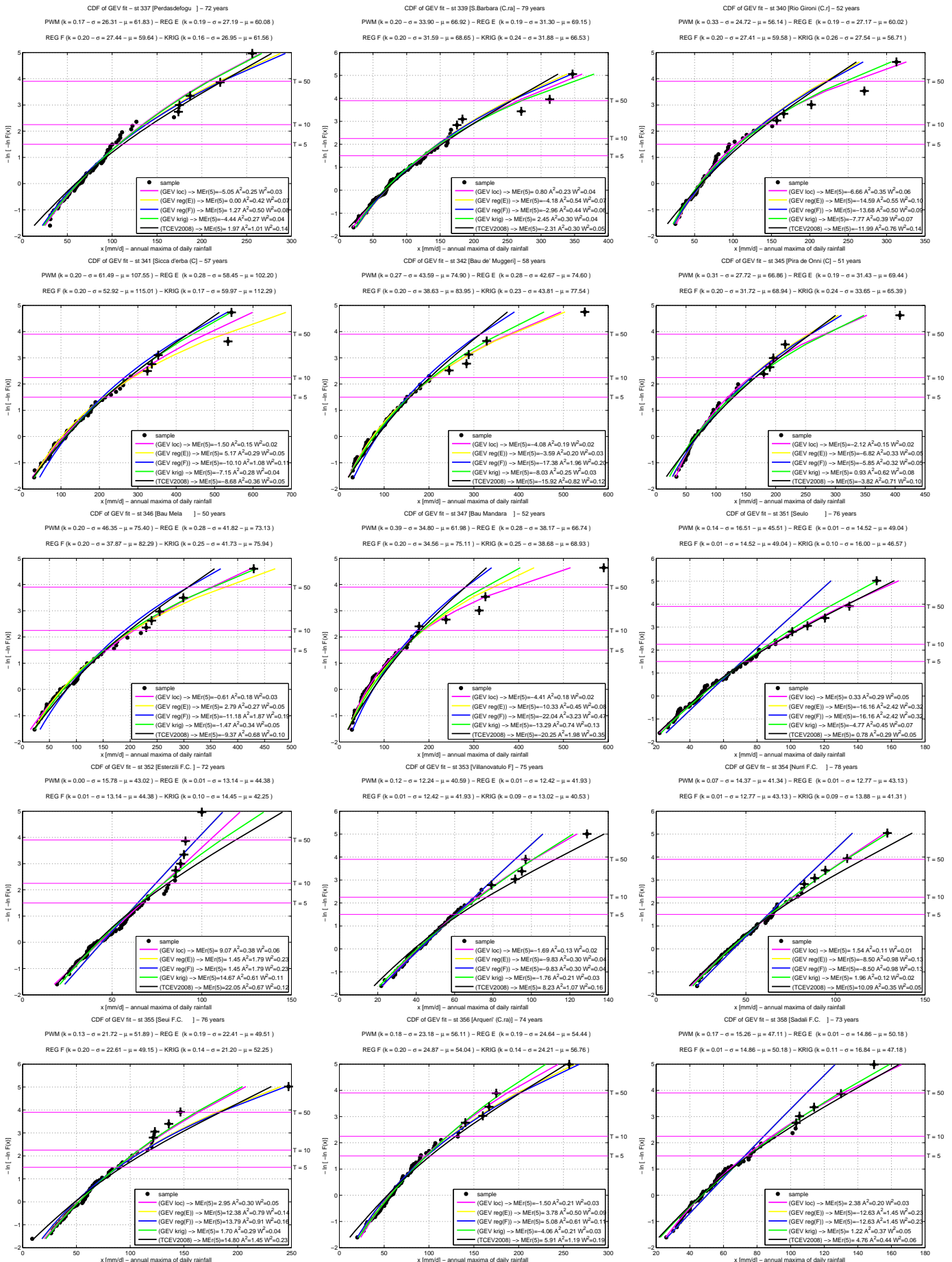


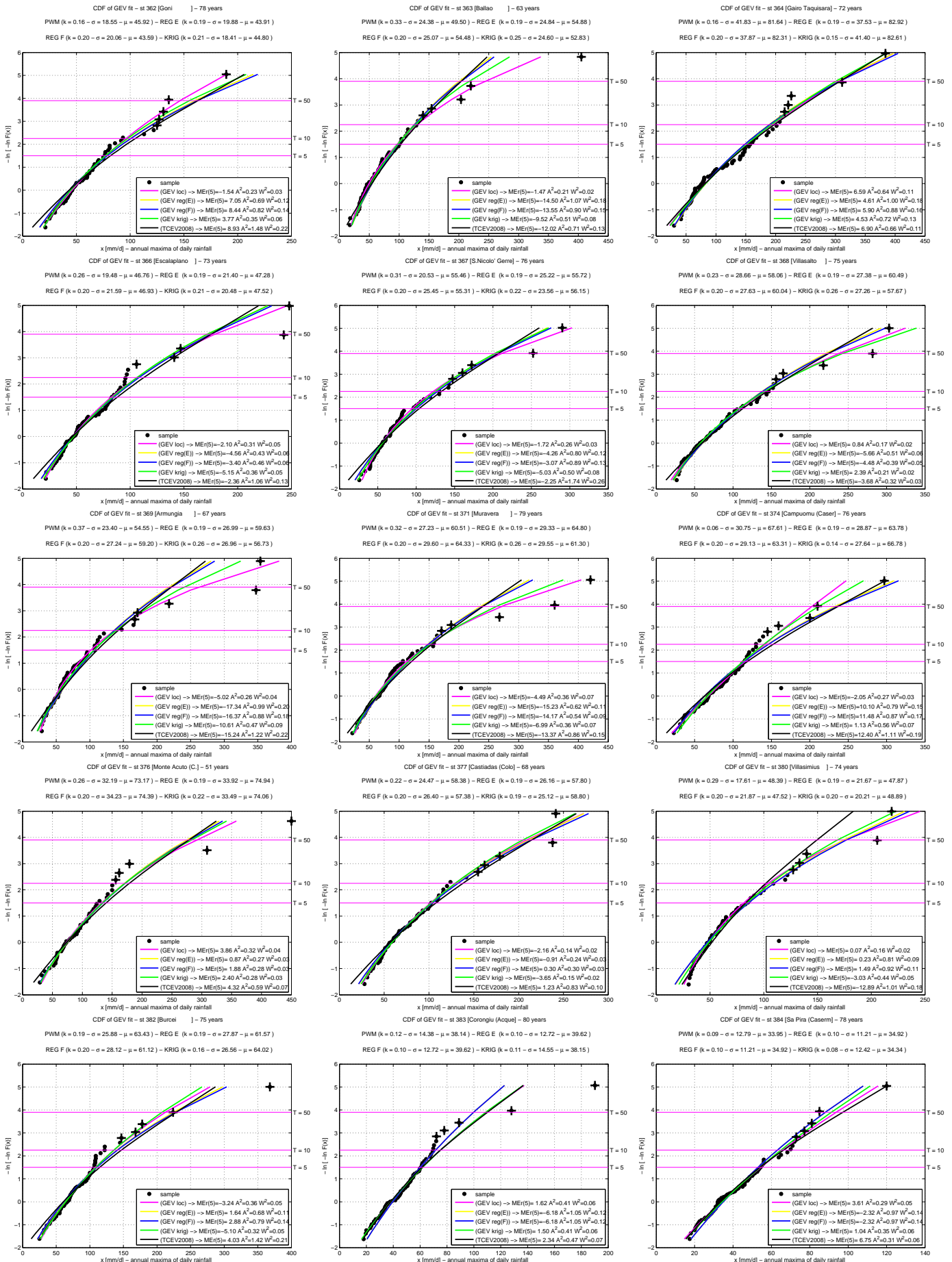


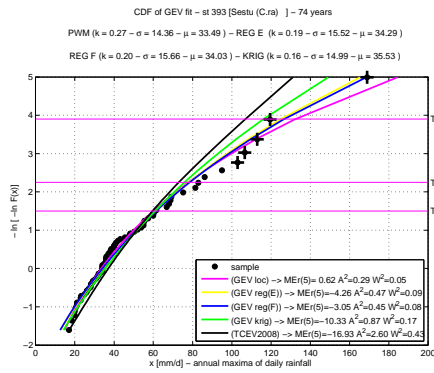
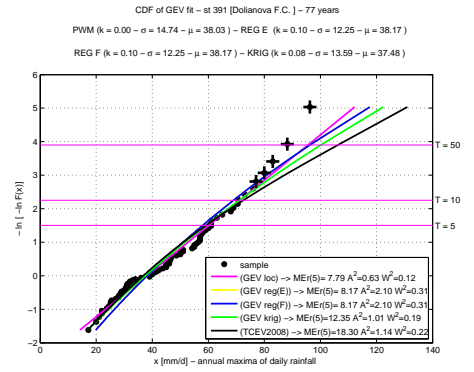
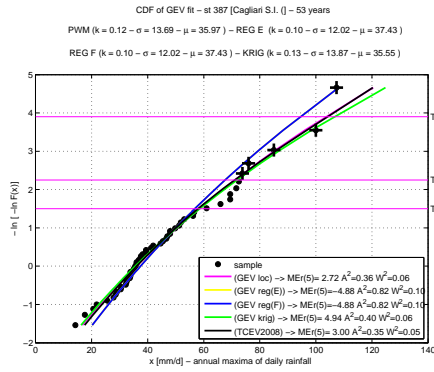
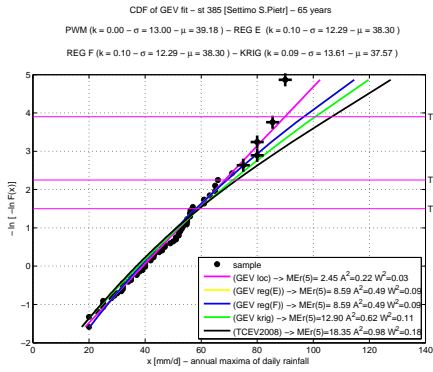




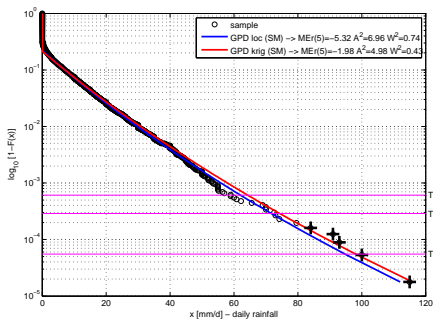




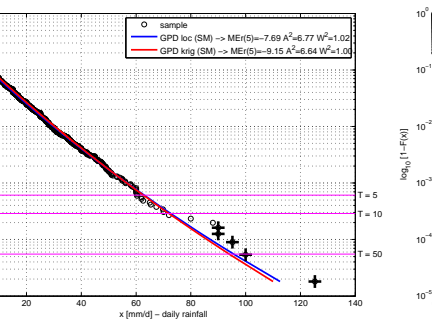




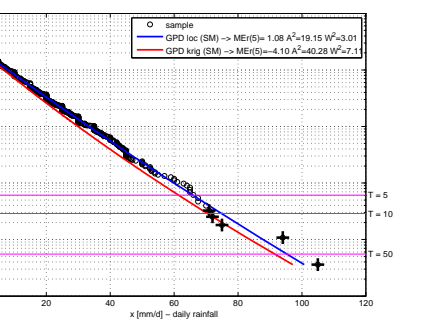
CDF of MTM GPD fit - st 001 [Saridano (Colo)] - 77 years
 SM ($\xi_0^M = 0.06 - \alpha_0^M = 8.69 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.06 - \alpha_0^M = 9.13 - \xi_0^M = 0.21$)



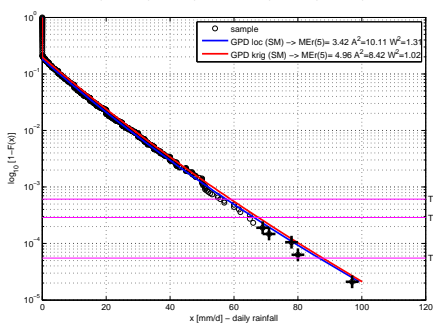
CDF of MTM GPD fit - st 002 [Is Jus Ascas (San)] - 76 years
 SM ($\xi_0^M = 0.06 - \alpha_0^M = 9.22 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.05 - \alpha_0^M = 9.30 - \xi_0^M = 0.21$)



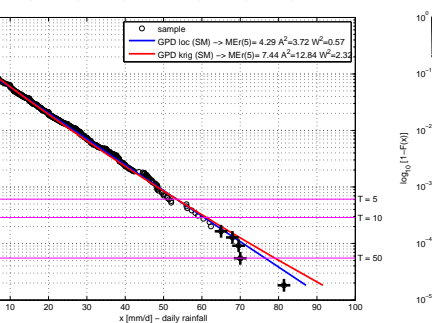
CDF of MTM GPD fit - st 003 [Iall] - 38 years
 SM ($\xi_0^M = 0.02 - \alpha_0^M = 10.54 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.04 - \alpha_0^M = 9.35 - \xi_0^M = 0.20$)



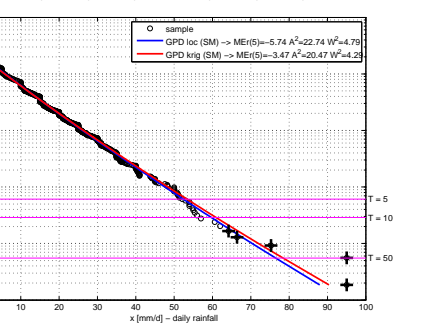
CDF of MTM GPD fit - st 004 [Gergei] - 65 years
 SM ($\xi_0^M = 0.05 - \alpha_0^M = 8.81 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.04 - \alpha_0^M = 9.03 - \xi_0^M = 0.20$)



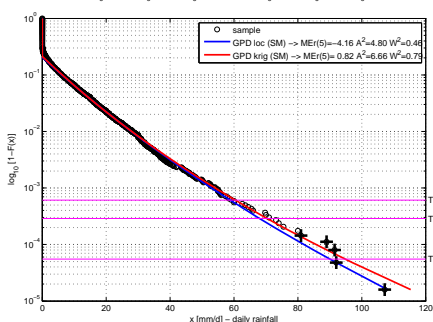
CDF of MTM GPD fit - st 006 [Villamar F.C.] - 75 years
 SM ($\xi_0^M = 0.00 - \alpha_0^M = 9.39 - \xi_0^M = 0.17$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 8.41 - \xi_0^M = 0.19$)



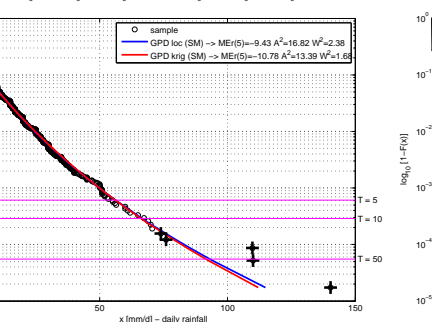
CDF of MTM GPD fit - st 007 [Lunamatrona F.C.] - 74 years
 SM ($\xi_0^M = 0.03 - \alpha_0^M = 8.36 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 8.45 - \xi_0^M = 0.20$)



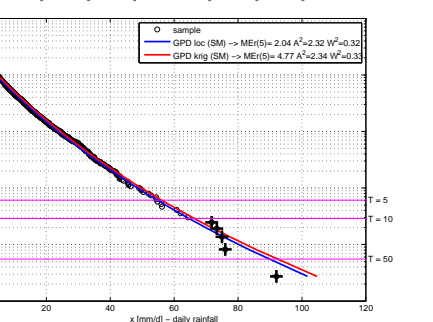
CDF of MTM GPD fit - st 008 [Mondis F.C.] - 86 years
 SM ($\xi_0^M = 0.06 - \alpha_0^M = 8.41 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.08 - \alpha_0^M = 8.07 - \xi_0^M = 0.21$)



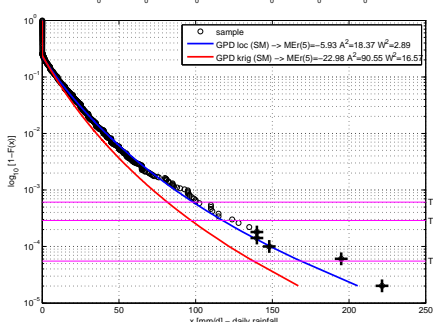
CDF of MTM GPD fit - st 009 [Segaria] - 79 years
 SM ($\xi_0^M = 0.12 - \alpha_0^M = 6.64 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.11 - \alpha_0^M = 7.01 - \xi_0^M = 0.20$)



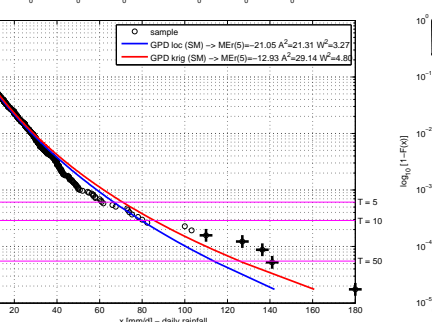
CDF of MTM GPD fit - st 014 [Sanluri (N.O.C.)] - 50 years
 SM ($\xi_0^M = 0.11 - \alpha_0^M = 7.02 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.11 - \alpha_0^M = 7.00 - \xi_0^M = 0.19$)



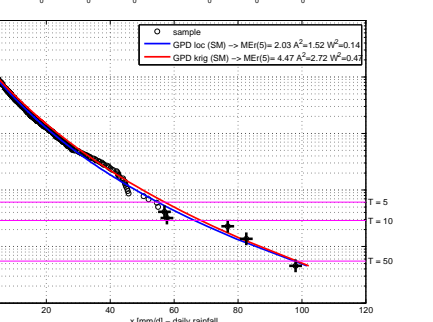
CDF of MTM GPD fit - st 015 [Monte Mannu (Ca)] - 68 years
 SM ($\xi_0^M = 0.14 - \alpha_0^M = 10.62 - \xi_0^M = 0.23$) - KRIG ($\xi_0^M = 0.13 - \alpha_0^M = 8.95 - \xi_0^M = 0.24$)



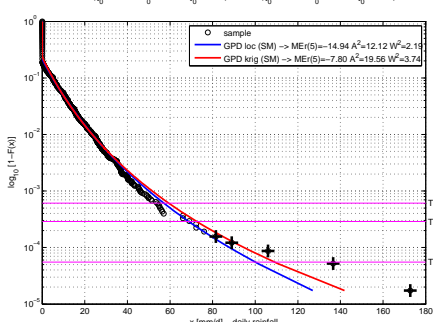
CDF of MTM GPD fit - st 016 [Villadrio F.C.] - 78 years
 SM ($\xi_0^M = 0.14 - \alpha_0^M = 7.39 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.17 - \alpha_0^M = 6.93 - \xi_0^M = 0.24$)



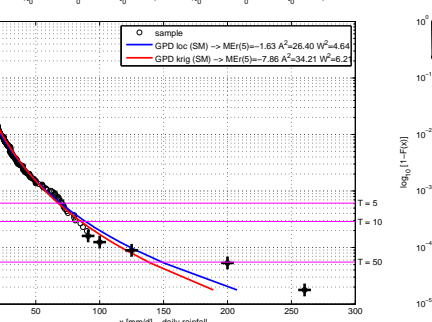
CDF of MTM GPD fit - st 018 [Serrenti] - 30 years
 SM ($\xi_0^M = 0.16 - \alpha_0^M = 5.77 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.14 - \alpha_0^M = 6.38 - \xi_0^M = 0.20$)



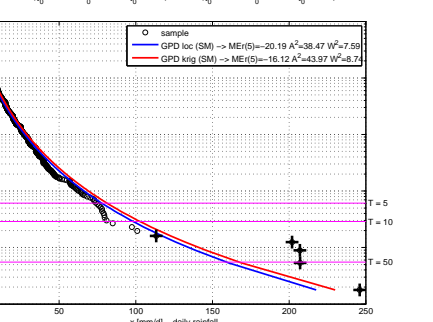
CDF of MTM GPD fit - st 019 [Nuraminis] - 79 years
 SM ($\xi_0^M = 0.15 - \alpha_0^M = 6.01 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.18 - \alpha_0^M = 5.52 - \xi_0^M = 0.23$)



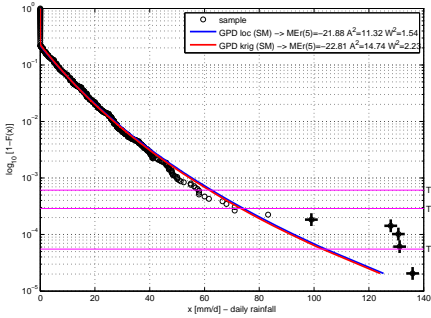
CDF of MTM GPD fit - st 020 [Vilassar FF.SS.] - 77 years
 SM ($\xi_0^M = 0.26 - \alpha_0^M = 5.23 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.24 - \alpha_0^M = 5.15 - \xi_0^M = 0.22$)



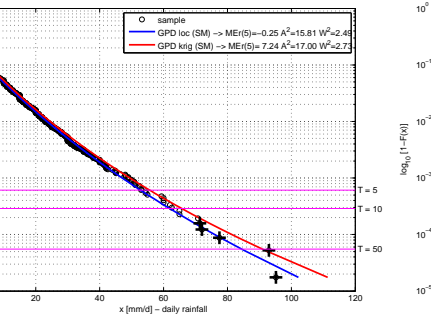
CDF of MTM GPD fit - st 024 [Villemosa] - 77 years
 SM ($\xi_0^M = 0.25 - \alpha_0^M = 5.70 - \xi_0^M = 0.24$) - KRIG ($\xi_0^M = 0.25 - \alpha_0^M = 5.53 - \xi_0^M = 0.28$)



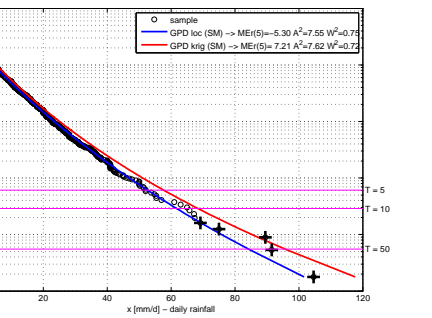
CDF of MTM GPD fit - st 025 [Gelsco F.C.] - 67 years
SM ($\xi_0^M = 0.12 - \alpha_0^M = 7.25 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.13 - \alpha_0^M = 7.03 - \xi_0^M = 0.23$)



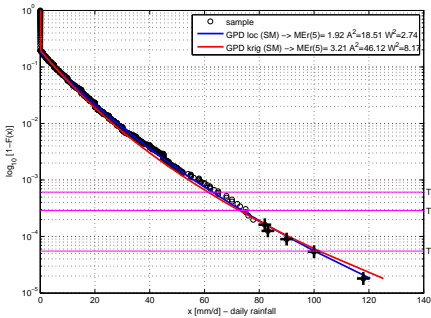
CDF of MTM GPD fit - st 026 [Guesia] - 78 years
SM ($\xi_0^M = 0.09 - \alpha_0^M = 7.05 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.11 - \alpha_0^M = 6.86 - \xi_0^M = 0.21$)



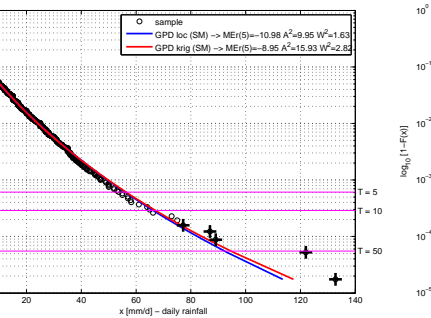
CDF of MTM GPD fit - st 027 [Senorbi F.C.] - 77 years
SM ($\xi_0^M = 0.09 - \alpha_0^M = 7.08 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.12 - \alpha_0^M = 6.88 - \xi_0^M = 0.21$)



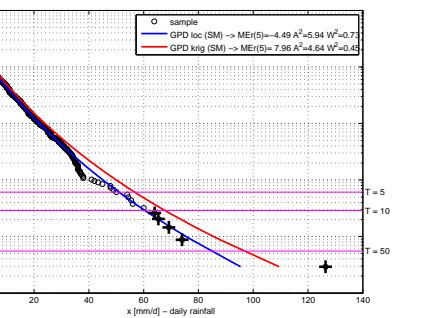
CDF of MTM GPD fit - st 030 [S.Andrea Fiusi] - 76 years
SM ($\xi_0^M = 0.09 - \alpha_0^M = 8.50 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.12 - \alpha_0^M = 7.37 - \xi_0^M = 0.20$)



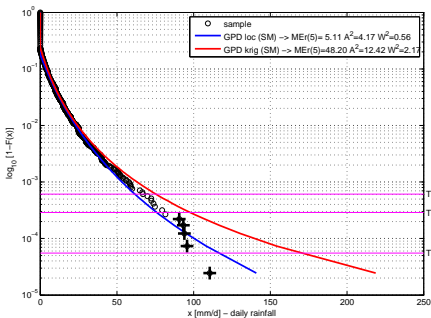
CDF of MTM GPD fit - st 031 [Barrali] - 78 years
SM ($\xi_0^M = 0.11 - \alpha_0^M = 6.91 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.13 - \alpha_0^M = 6.53 - \xi_0^M = 0.21$)



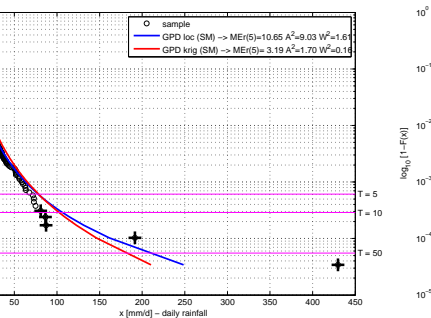
CDF of MTM GPD fit - st 032 [Donori F.C.] - 47 years
SM ($\xi_0^M = 0.12 - \alpha_0^M = 6.26 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.13 - \alpha_0^M = 6.49 - \xi_0^M = 0.20$)



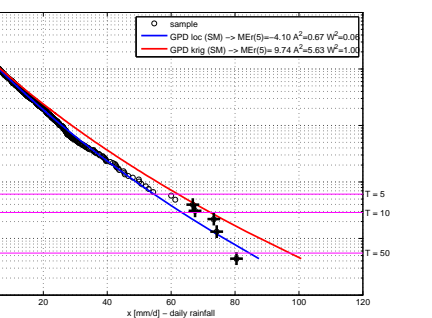
CDF of MTM GPD fit - st 034 [Decomianinu] - 56 years
SM ($\xi_0^M = 0.19 - \alpha_0^M = 5.99 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.29 - \alpha_0^M = 4.63 - \xi_0^M = 0.24$)



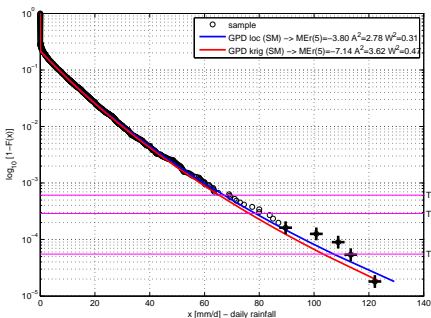
CDF of MTM GPD fit - st 035 [Decomianinu (V)] - 40 years
SM ($\xi_0^M = 0.38 - \alpha_0^M = 3.30 - \xi_0^M = 0.27$) - KRIG ($\xi_0^M = 0.31 - \alpha_0^M = 4.57 - \xi_0^M = 0.24$)



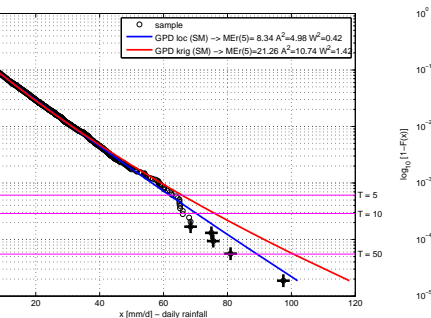
CDF of MTM GPD fit - st 036 [Bellicci (Priva)] - 31 years
SM ($\xi_0^M = 0.07 - \alpha_0^M = 7.43 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.08 - \alpha_0^M = 8.30 - \xi_0^M = 0.22$)



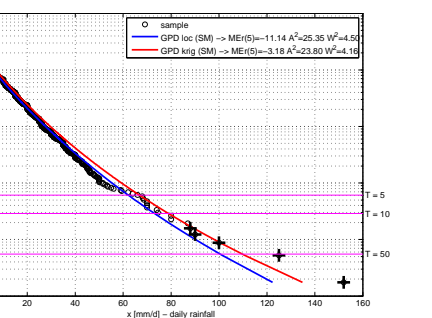
CDF of MTM GPD fit - st 038 [Iglesias] - 76 years
SM ($\xi_0^M = 0.09 - \alpha_0^M = 8.55 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.09 - \alpha_0^M = 8.51 - \xi_0^M = 0.22$)



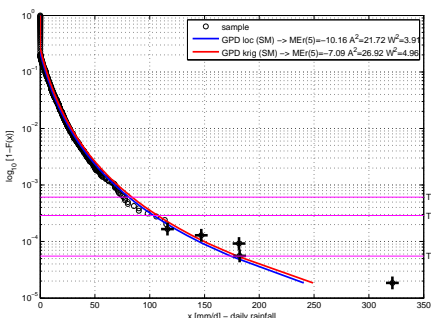
CDF of MTM GPD fit - st 042 [S.Giovanni (Dom)] - 73 years
SM ($\xi_0^M = 0.02 - \alpha_0^M = 10.08 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.06 - \alpha_0^M = 9.16 - \xi_0^M = 0.22$)



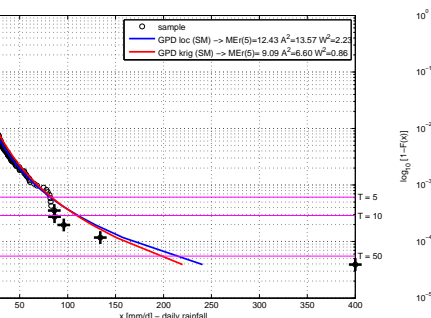
CDF of MTM GPD fit - st 043 [Vilamassargia] - 78 years
SM ($\xi_0^M = 0.10 - \alpha_0^M = 7.92 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.11 - \alpha_0^M = 7.87 - \xi_0^M = 0.23$)



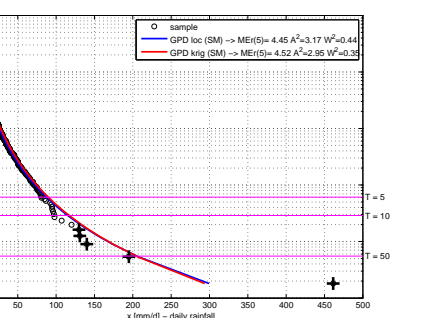
CDF of MTM GPD fit - st 045 [Sittqua FF.SS.] - 74 years
SM ($\xi_0^M = 0.27 - \alpha_0^M = 5.56 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.27 - \alpha_0^M = 5.39 - \xi_0^M = 0.26$)

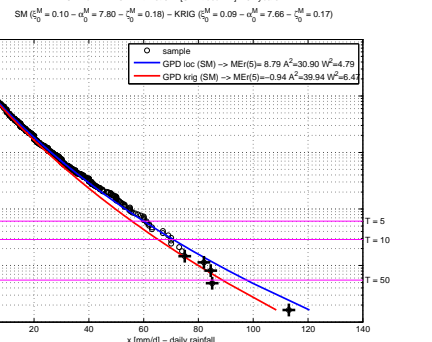
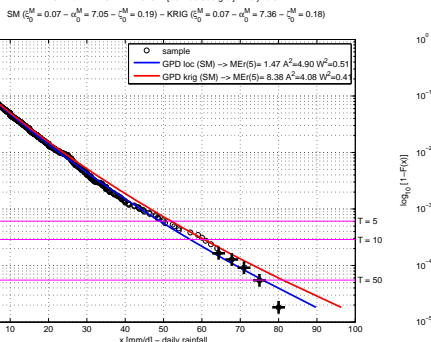
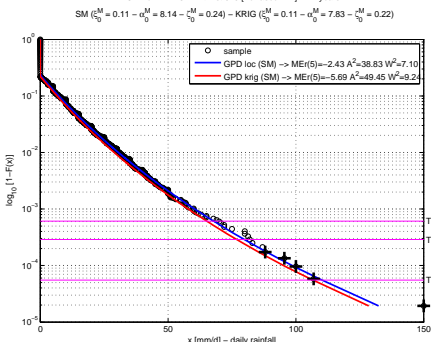
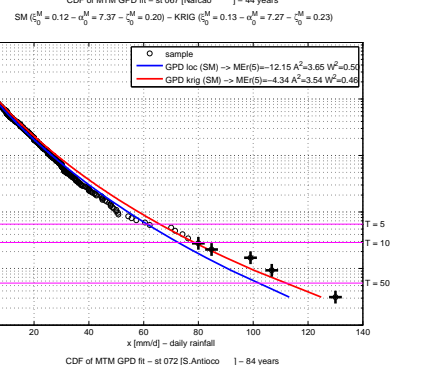
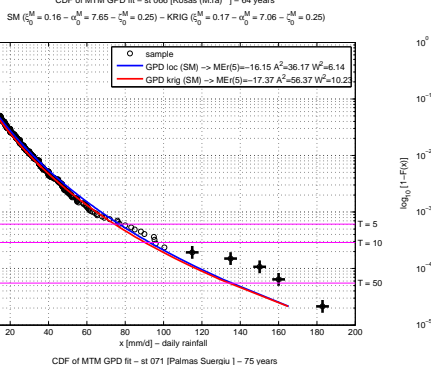
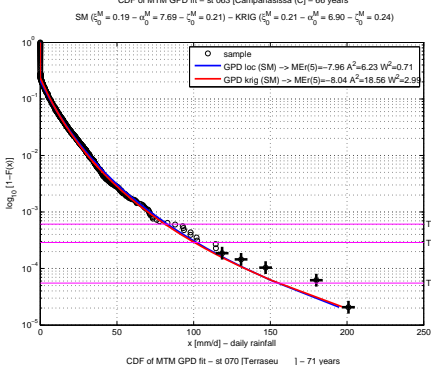
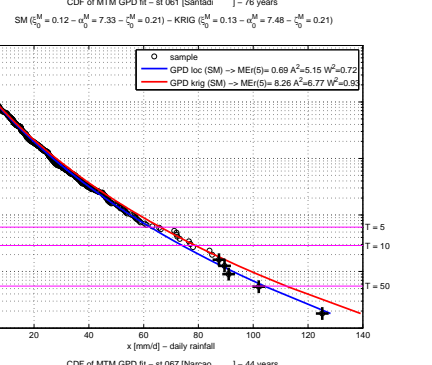
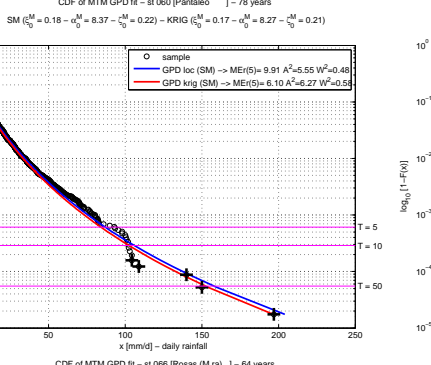
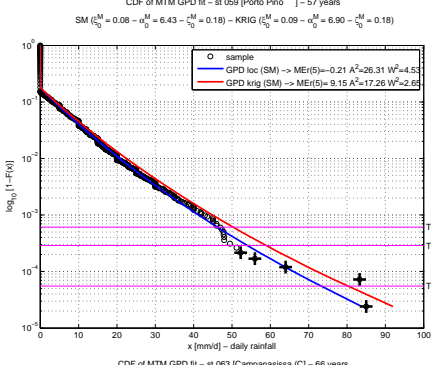
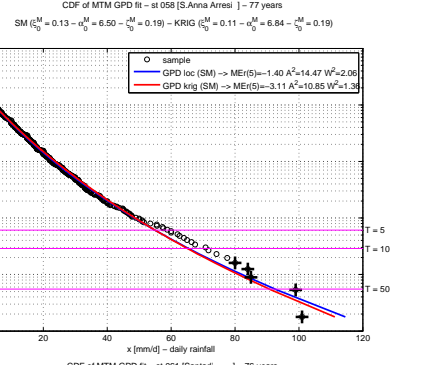
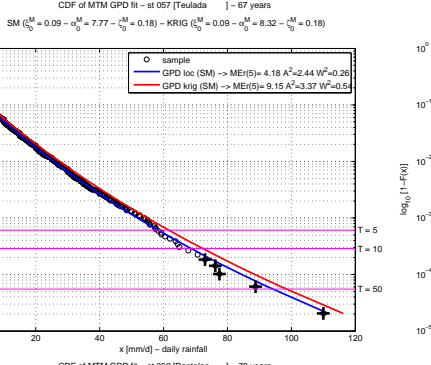
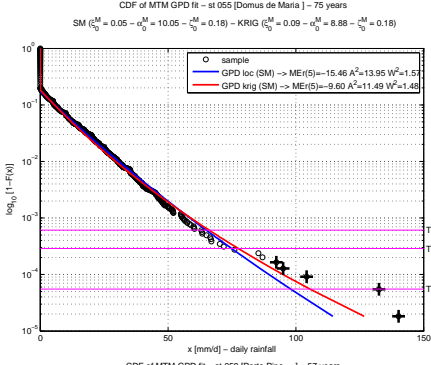
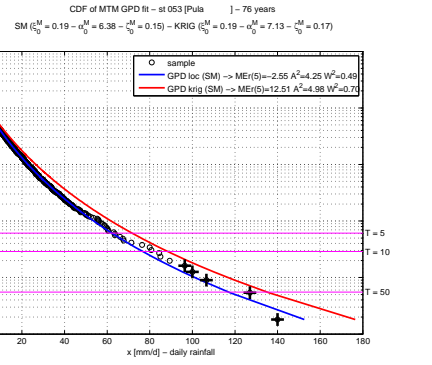
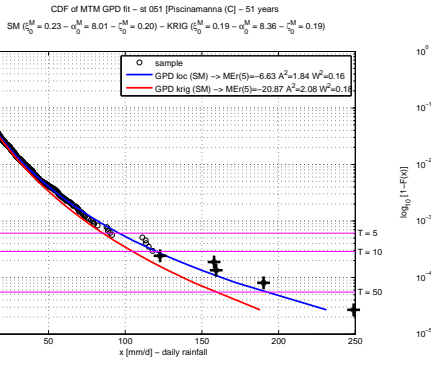
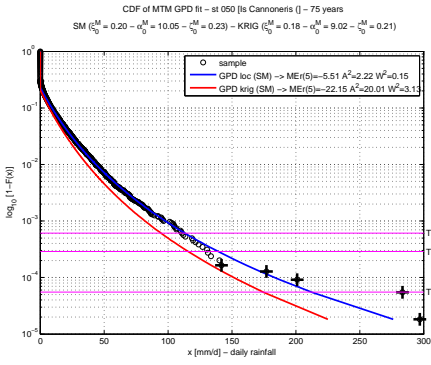


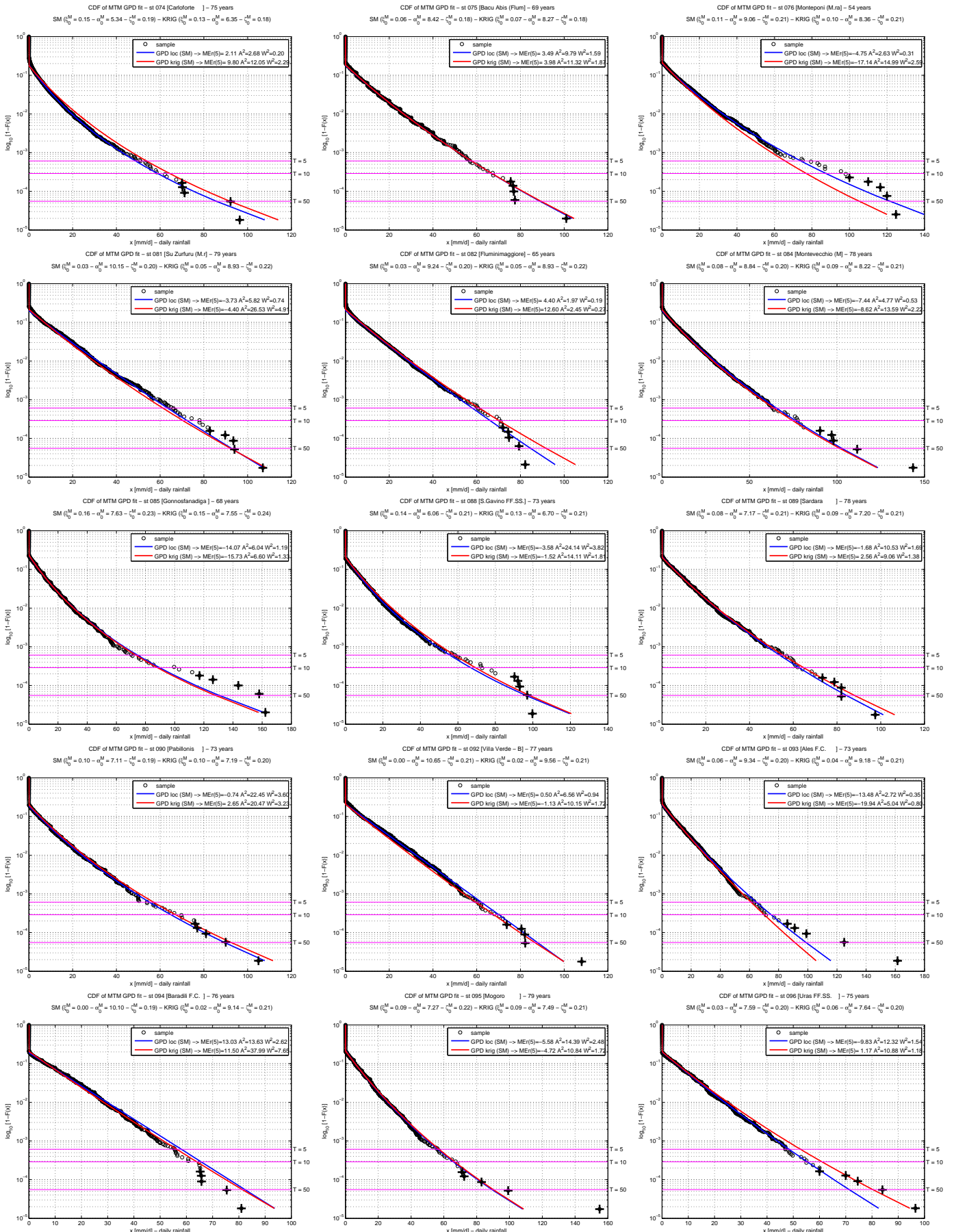
CDF of MTM GPD fit - st 047 [Uta (C.R.A.)] - 35 years
SM ($\xi_0^M = 0.37 - \alpha_0^M = 3.70 - \xi_0^M = 0.25$) - KRIG ($\xi_0^M = 0.32 - \alpha_0^M = 4.59 - \xi_0^M = 0.24$)



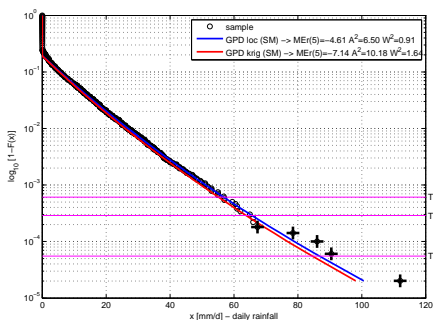
CDF of MTM GPD fit - st 048 [Capoterra] - 76 years
SM ($\xi_0^M = 0.31 - \alpha_0^M = 5.56 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.30 - \alpha_0^M = 5.81 - \xi_0^M = 0.20$)



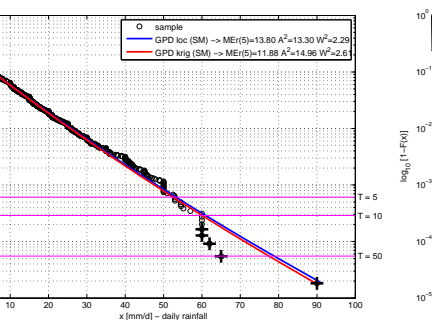




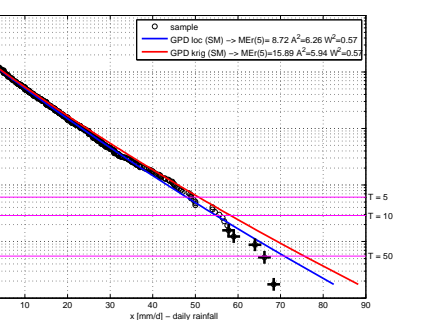
CDF of MTM GPD fit - st 099 [Aborea] - 68 years
SM ($\hat{\epsilon}_0^M = 0.06 - \hat{\sigma}_0^M = 8.30 - \hat{\sigma}_0^M = 0.19$) - KRIG ($\hat{\epsilon}_0^M = 0.06 - \hat{\sigma}_0^M = 8.04 - \hat{\sigma}_0^M = 0.19$)



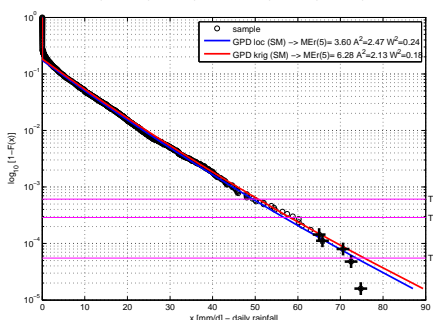
CDF of MTM GPD fit - st 100 [Marubiu (C.ra)] - 75 years
SM ($\hat{\epsilon}_0^M = 0.03 - \hat{\sigma}_0^M = 8.39 - \hat{\sigma}_0^M = 0.19$) - KRIG ($\hat{\epsilon}_0^M = 0.03 - \hat{\sigma}_0^M = 8.32 - \hat{\sigma}_0^M = 0.19$)



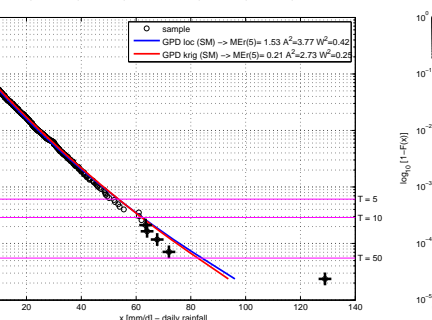
CDF of MTM GPD fit - st 102 [S.Annu-Ortiano] - 78 years
SM ($\hat{\epsilon}_0^M = 0.03 - \hat{\sigma}_0^M = 7.62 - \hat{\sigma}_0^M = 0.19$) - KRIG ($\hat{\epsilon}_0^M = 0.04 - \hat{\sigma}_0^M = 7.97 - \hat{\sigma}_0^M = 0.19$)



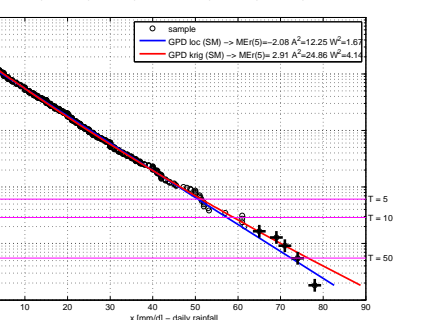
CDF of MTM GPD fit - st 103 [S.Giusta] - 86 years
SM ($\hat{\epsilon}_0^M = 0.04 - \hat{\sigma}_0^M = 7.88 - \hat{\sigma}_0^M = 0.17$) - KRIG ($\hat{\epsilon}_0^M = 0.04 - \hat{\sigma}_0^M = 7.95 - \hat{\sigma}_0^M = 0.18$)



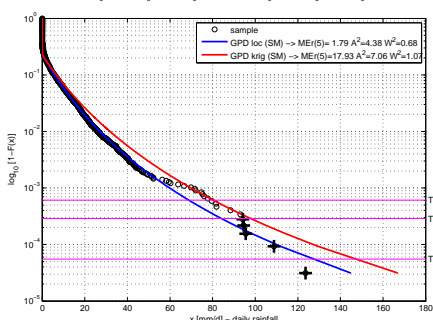
CDF of MTM GPD fit - st 104 [Sassu Idrovara] - 58 years
SM ($\hat{\epsilon}_0^M = 0.08 - \hat{\sigma}_0^M = 7.48 - \hat{\sigma}_0^M = 0.18$) - KRIG ($\hat{\epsilon}_0^M = 0.06 - \hat{\sigma}_0^M = 7.78 - \hat{\sigma}_0^M = 0.19$)



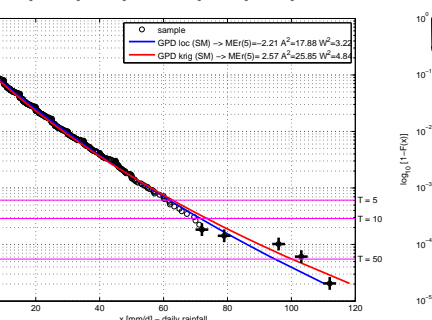
CDF of MTM GPD fit - st 105 [Ortiano FF.SS.] - 75 years
SM ($\hat{\epsilon}_0^M = 0.00 - \hat{\sigma}_0^M = 8.92 - \hat{\sigma}_0^M = 0.17$) - KRIG ($\hat{\epsilon}_0^M = 0.04 - \hat{\sigma}_0^M = 8.07 - \hat{\sigma}_0^M = 0.19$)



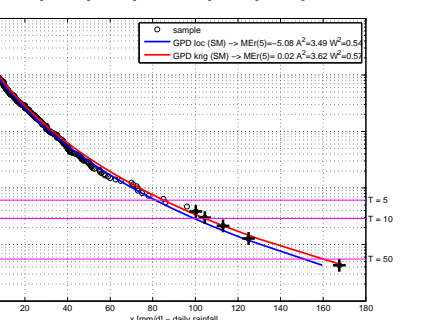
CDF of MTM GPD fit - st 106 [Soc Canales (C.)] - 44 years
SM ($\hat{\epsilon}_0^M = 0.19 - \hat{\sigma}_0^M = 6.33 - \hat{\sigma}_0^M = 0.23$) - KRIG ($\hat{\epsilon}_0^M = 0.18 - \hat{\sigma}_0^M = 7.60 - \hat{\sigma}_0^M = 0.23$)



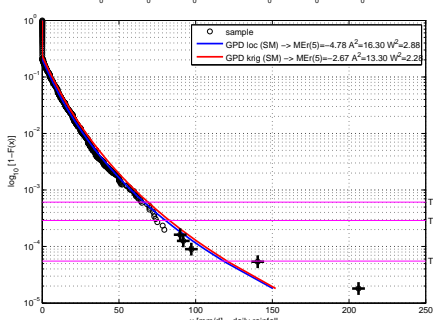
CDF of MTM GPD fit - st 107 [Osside F.C.] - 67 years
SM ($\hat{\epsilon}_0^M = 0.07 - \hat{\sigma}_0^M = 8.58 - \hat{\sigma}_0^M = 0.22$) - KRIG ($\hat{\epsilon}_0^M = 0.09 - \hat{\sigma}_0^M = 8.01 - \hat{\sigma}_0^M = 0.22$)



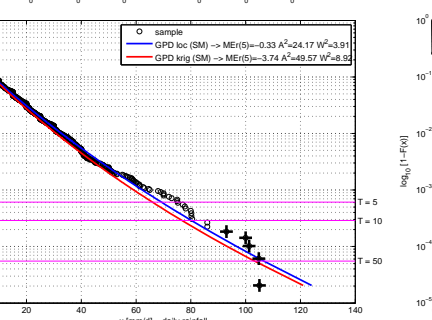
CDF of MTM GPD fit - st 108 [S.Giovanni- Bi] - 32 years
SM ($\hat{\epsilon}_0^M = 0.19 - \hat{\sigma}_0^M = 7.47 - \hat{\sigma}_0^M = 0.23$) - KRIG ($\hat{\epsilon}_0^M = 0.20 - \hat{\sigma}_0^M = 7.37 - \hat{\sigma}_0^M = 0.25$)



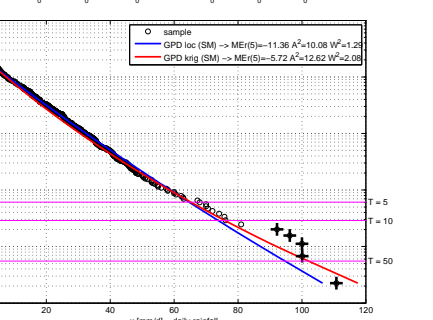
CDF of MTM GPD fit - st 110 [Benetutti] - 76 years
SM ($\hat{\epsilon}_0^M = 0.16 - \hat{\sigma}_0^M = 6.79 - \hat{\sigma}_0^M = 0.21$) - KRIG ($\hat{\epsilon}_0^M = 0.16 - \hat{\sigma}_0^M = 6.99 - \hat{\sigma}_0^M = 0.24$)



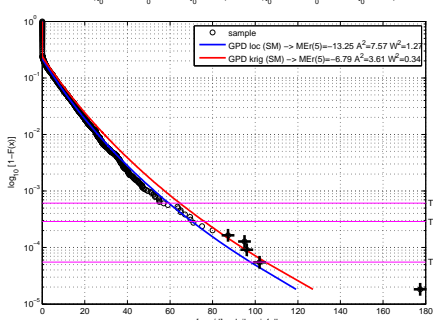
CDF of MTM GPD fit - st 111 [Botida F.C.] - 67 years
SM ($\hat{\epsilon}_0^M = 0.07 - \hat{\sigma}_0^M = 9.53 - \hat{\sigma}_0^M = 0.22$) - KRIG ($\hat{\epsilon}_0^M = 0.08 - \hat{\sigma}_0^M = 8.66 - \hat{\sigma}_0^M = 0.23$)



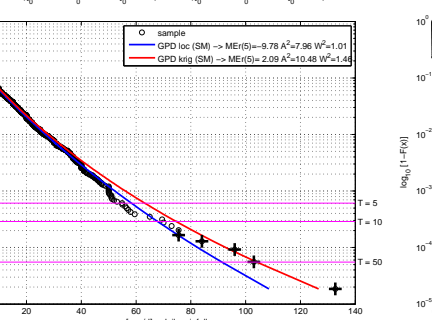
CDF of MTM GPD fit - st 113 [Ilorai] - 61 years
SM ($\hat{\epsilon}_0^M = 0.03 - \hat{\sigma}_0^M = 10.06 - \hat{\sigma}_0^M = 0.22$) - KRIG ($\hat{\epsilon}_0^M = 0.08 - \hat{\sigma}_0^M = 8.58 - \hat{\sigma}_0^M = 0.23$)



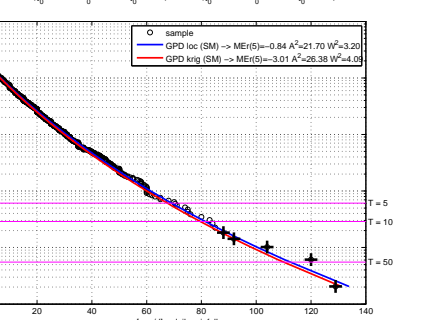
CDF of MTM GPD fit - st 114 [Rifonitore Tir] - 75 years
SM ($\hat{\epsilon}_0^M = 0.11 - \hat{\sigma}_0^M = 7.09 - \hat{\sigma}_0^M = 0.22$) - KRIG ($\hat{\epsilon}_0^M = 0.11 - \hat{\sigma}_0^M = 7.70 - \hat{\sigma}_0^M = 0.24$)

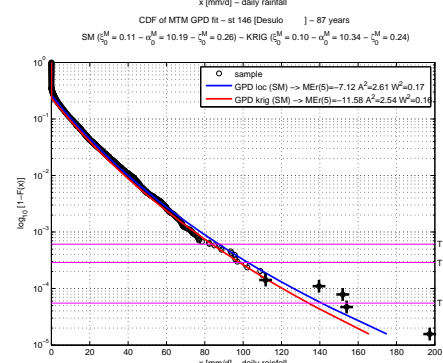
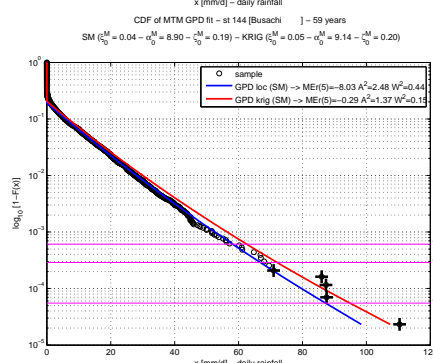
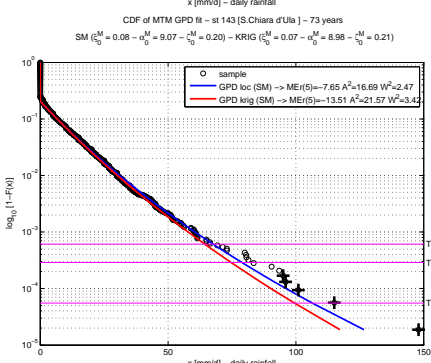
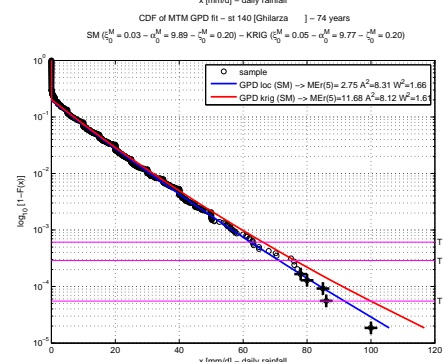
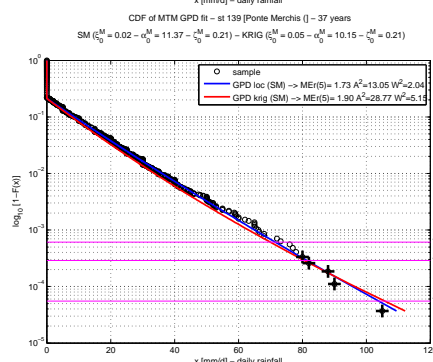
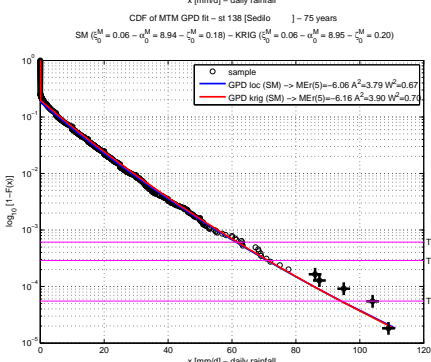
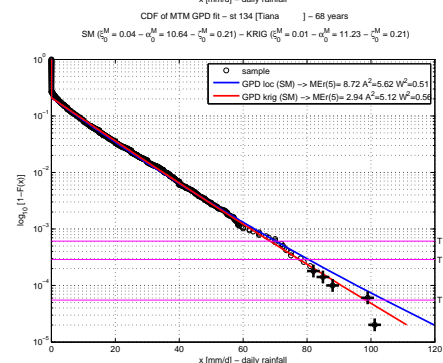
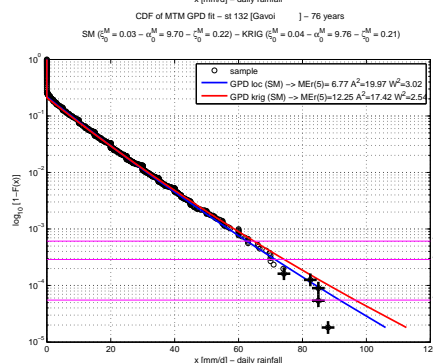
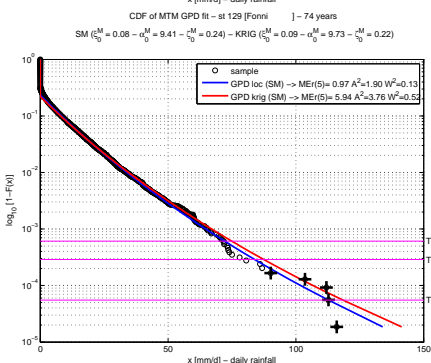
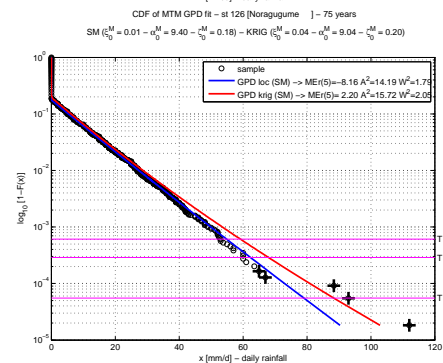
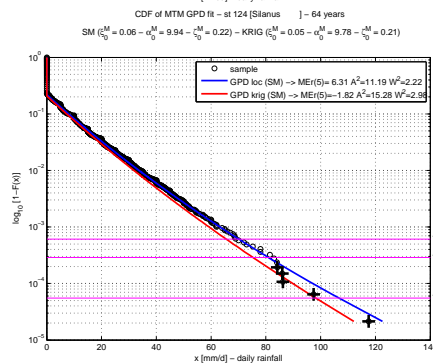
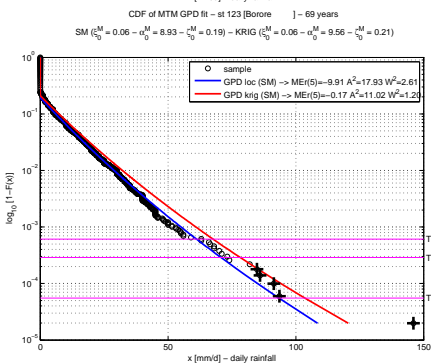
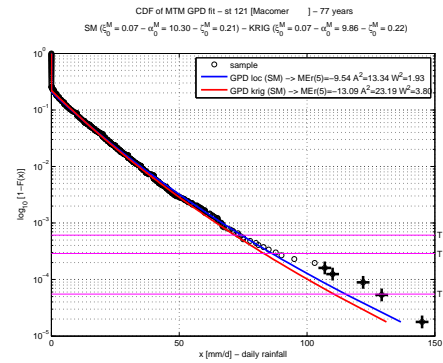
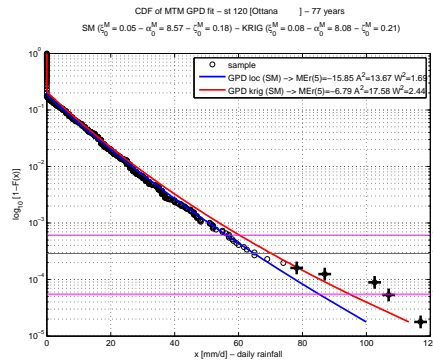
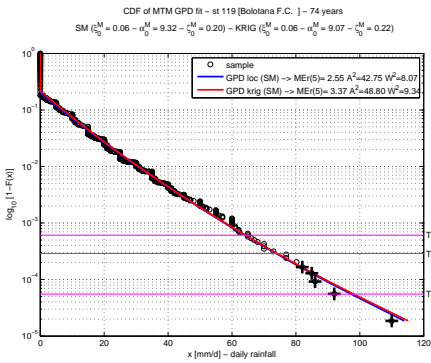


CDF of MTM GPD fit - st 115 [Siga Maria (C.)] - 74 years
SM ($\hat{\epsilon}_0^M = 0.08 - \hat{\sigma}_0^M = 7.89 - \hat{\sigma}_0^M = 0.21$) - KRIG ($\hat{\epsilon}_0^M = 0.12 - \hat{\sigma}_0^M = 7.44 - \hat{\sigma}_0^M = 0.23$)

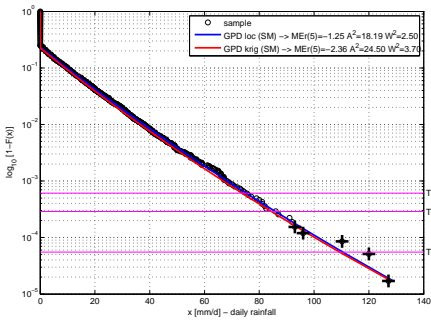


CDF of MTM GPD fit - st 118 [Orani] - 67 years
SM ($\hat{\epsilon}_0^M = 0.11 - \hat{\sigma}_0^M = 8.53 - \hat{\sigma}_0^M = 0.21$) - KRIG ($\hat{\epsilon}_0^M = 0.11 - \hat{\sigma}_0^M = 8.33 - \hat{\sigma}_0^M = 0.21$)

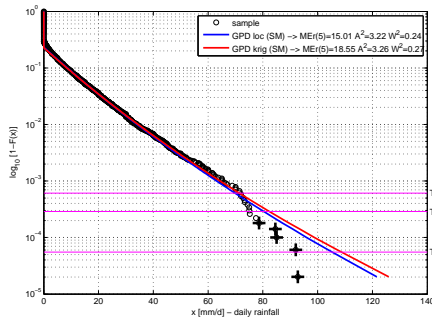




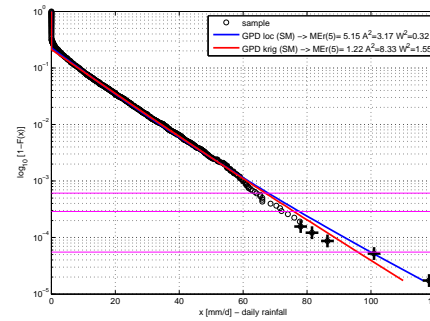
CDF of MTM GPD fit - st 147 [Torera] - 80 years
SM ($\xi_0^M = 0.04 - \alpha_0^M = 11.30 - \xi_0^M = 0.23$) - KRIG ($\xi_0^M = 0.04 - \alpha_0^M = 10.97 - \xi_0^M = 0.23$)



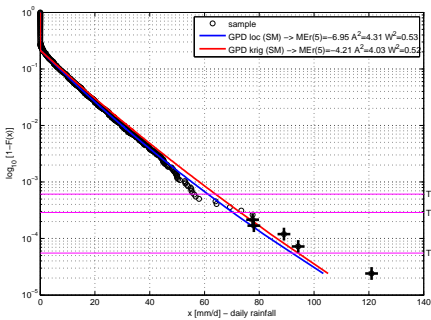
CDF of MTM GPD fit - st 148 [Cossato (C.re)] - 68 years
SM ($\xi_0^M = 0.06 - \alpha_0^M = 9.94 - \xi_0^M = 0.23$) - KRIG ($\xi_0^M = 0.06 - \alpha_0^M = 10.00 - \xi_0^M = 0.23$)



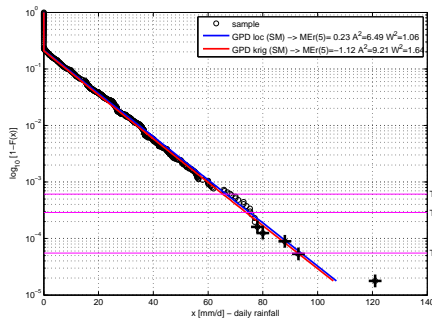
CDF of MTM GPD fit - st 151 [Sorgono F.C.] - 79 years
SM ($\xi_0^M = 0.04 - \alpha_0^M = 10.28 - \xi_0^M = 0.23$) - KRIG ($\xi_0^M = 0.01 - \alpha_0^M = 11.07 - \xi_0^M = 0.21$)



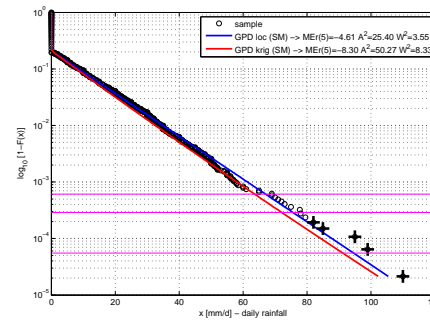
CDF of MTM GPD fit - st 153 [Meansardo] - 57 years
SM ($\xi_0^M = 0.05 - \alpha_0^M = 8.98 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 9.89 - \xi_0^M = 0.21$)



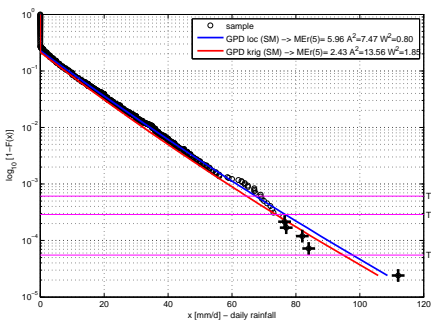
CDF of MTM GPD fit - st 154 [Austria] - 77 years
SM ($\xi_0^M = 0.00 - \alpha_0^M = 11.38 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.00 - \alpha_0^M = 11.09 - \xi_0^M = 0.21$)



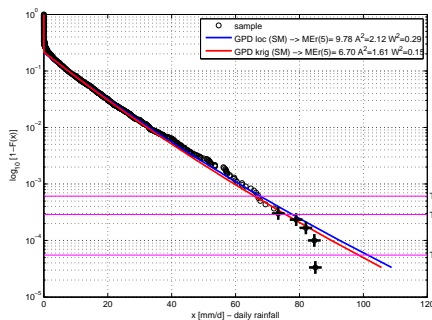
CDF of MTM GPD fit - st 155 [Ortueri] - 64 years
SM ($\xi_0^M = 0.00 - \alpha_0^M = 11.47 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.01 - \alpha_0^M = 10.50 - \xi_0^M = 0.21$)



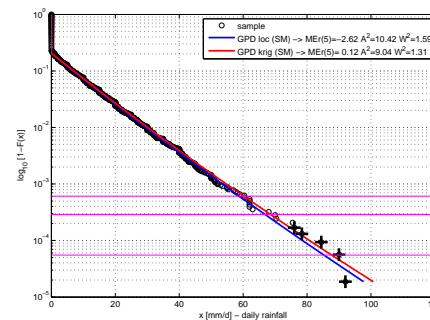
CDF of MTM GPD fit - st 156 [Oniavola F.C.] - 57 years
SM ($\xi_0^M = 0.02 - \alpha_0^M = 10.57 - \xi_0^M = 0.23$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 10.01 - \xi_0^M = 0.22$)



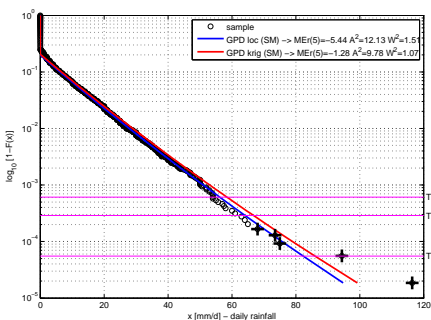
CDF of MTM GPD fit - st 158 [Santa Sofia] - 41 years
SM ($\xi_0^M = 0.05 - \alpha_0^M = 10.08 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.05 - \alpha_0^M = 9.79 - \xi_0^M = 0.22$)



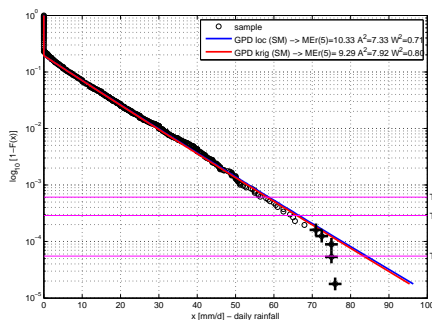
CDF of MTM GPD fit - st 159 [Leconi F.C.] - 73 years
SM ($\xi_0^M = 0.02 - \alpha_0^M = 9.49 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 9.58 - \xi_0^M = 0.21$)



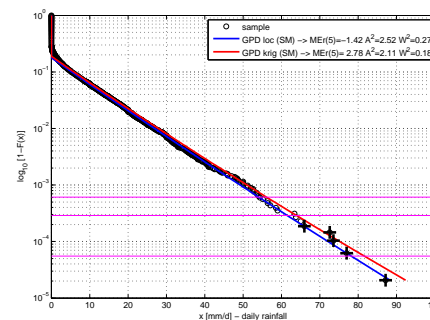
CDF of MTM GPD fit - st 160 [Genoni] - 74 years
SM ($\xi_0^M = 0.03 - \alpha_0^M = 8.95 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 9.12 - \xi_0^M = 0.21$)



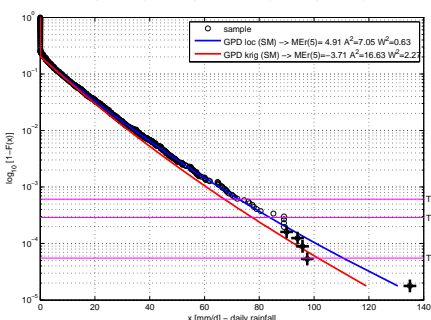
CDF of MTM GPD fit - st 163 [Samugheo] - 77 years
SM ($\xi_0^M = 0.01 - \alpha_0^M = 9.90 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.01 - \alpha_0^M = 9.85 - \xi_0^M = 0.19$)



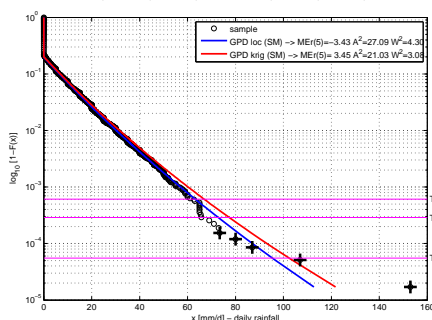
CDF of MTM GPD fit - st 164 [Alai] - 66 years
SM ($\xi_0^M = 0.01 - \alpha_0^M = 9.16 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.02 - \alpha_0^M = 9.21 - \xi_0^M = 0.19$)



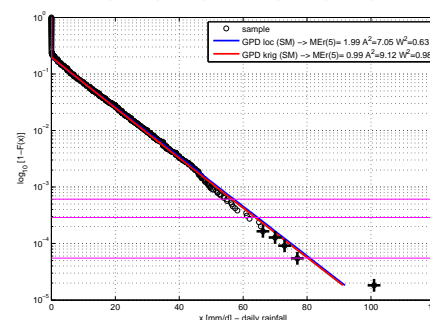
CDF of MTM GPD fit - st 166 [Abbasanta] - 77 years
SM ($\xi_0^M = 0.06 - \alpha_0^M = 10.46 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.05 - \alpha_0^M = 9.93 - \xi_0^M = 0.20$)



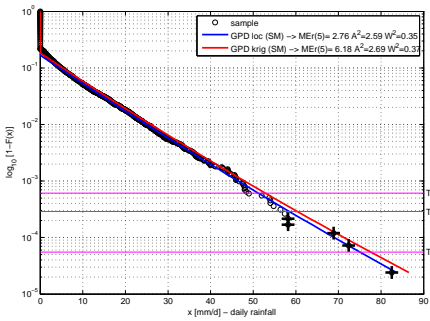
CDF of MTM GPD fit - st 167 [Paullistino] - 80 years
SM ($\xi_0^M = 0.05 - \alpha_0^M = 9.39 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.06 - \alpha_0^M = 9.59 - \xi_0^M = 0.20$)



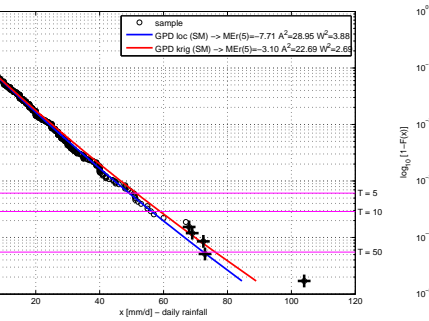
CDF of MTM GPD fit - st 168 [Mogrono] - 75 years
SM ($\xi_0^M = 0.00 - \alpha_0^M = 9.82 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.00 - \alpha_0^M = 9.66 - \xi_0^M = 0.20$)



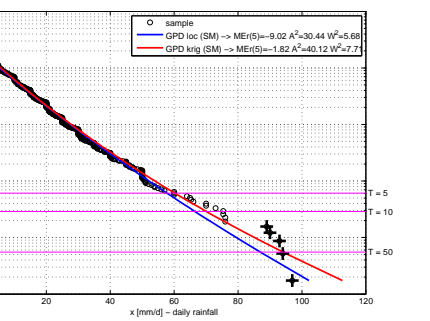
CDF of MTM GPD fit - st 170 [Santa Vittoria] - 57 years
SM ($\xi_0^M = 0.02 - \alpha_0^M = 8.59 - \xi_0^M = 0.17$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 8.56 - \xi_0^M = 0.18$)



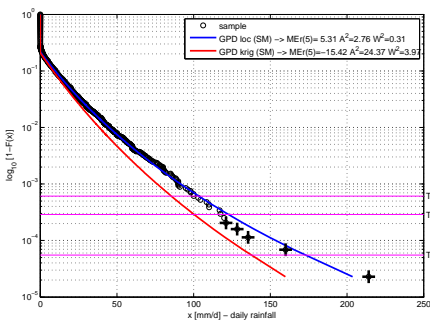
CDF of MTM GPD fit - st 171 [Sinnais] - 80 years
SM ($\xi_0^M = 0.03 - \alpha_0^M = 7.82 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 8.12 - \xi_0^M = 0.19$)



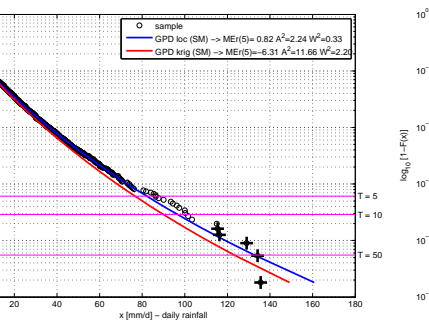
CDF of MTM GPD fit - st 172 [Roda] - 79 years
SM ($\xi_0^M = 0.05 - \alpha_0^M = 8.89 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.08 - \alpha_0^M = 8.33 - \xi_0^M = 0.20$)



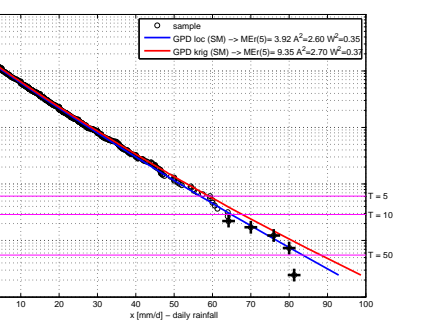
CDF of MTM GPD fit - st 174 [Sanluferrugli] - 60 years
SM ($\xi_0^M = 0.12 - \alpha_0^M = 11.97 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.10 - \alpha_0^M = 10.88 - \xi_0^M = 0.21$)



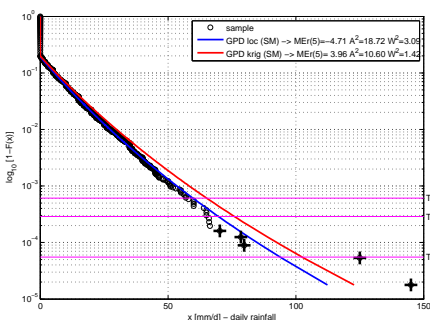
CDF of MTM GPD fit - st 175 [Seneghe] - 78 years
SM ($\xi_0^M = 0.10 - \alpha_0^M = 10.50 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.10 - \alpha_0^M = 9.82 - \xi_0^M = 0.20$)



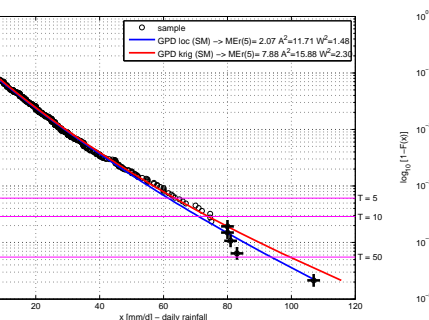
CDF of MTM GPD fit - st 176 [Bauladu] - 56 years
SM ($\xi_0^M = 0.03 - \alpha_0^M = 9.28 - \xi_0^M = 0.18$) - KRIG ($\xi_0^M = 0.04 - \alpha_0^M = 9.15 - \xi_0^M = 0.19$)



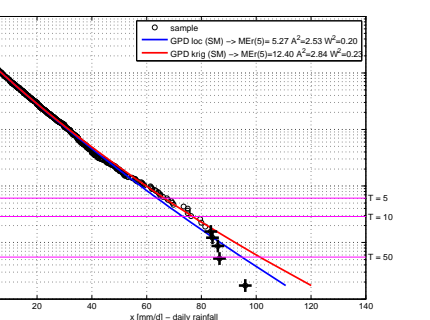
CDF of MTM GPD fit - st 178 [Tegge (C.ra)] - 77 years
SM ($\xi_0^M = 0.08 - \alpha_0^M = 8.22 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.08 - \alpha_0^M = 8.83 - \xi_0^M = 0.20$)



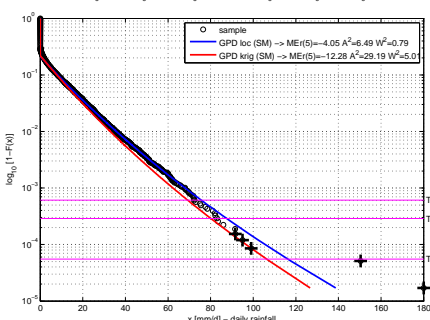
CDF of MTM GPD fit - st 179 [Tresnuraghes F.] - 64 years
SM ($\xi_0^M = 0.06 - \alpha_0^M = 9.05 - \xi_0^M = 0.19$) - KRIG ($\xi_0^M = 0.08 - \alpha_0^M = 8.58 - \xi_0^M = 0.21$)



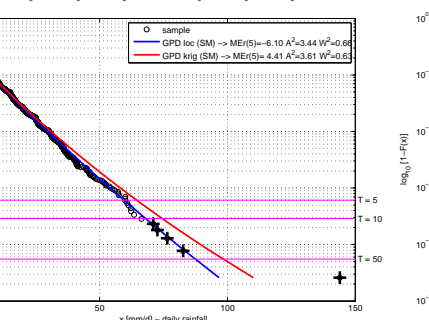
CDF of MTM GPD fit - st 180 [Cuglieri] - 79 years
SM ($\xi_0^M = 0.04 - \alpha_0^M = 9.81 - \xi_0^M = 0.20$) - KRIG ($\xi_0^M = 0.05 - \alpha_0^M = 9.56 - \xi_0^M = 0.20$)



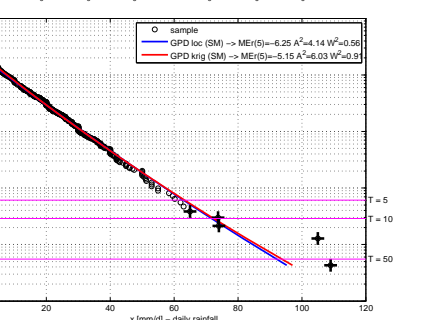
CDF of MTM GPD fit - st 181 [Vilanova Monte] - 80 years
SM ($\xi_0^M = 0.07 - \alpha_0^M = 10.51 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.07 - \alpha_0^M = 9.59 - \xi_0^M = 0.22$)



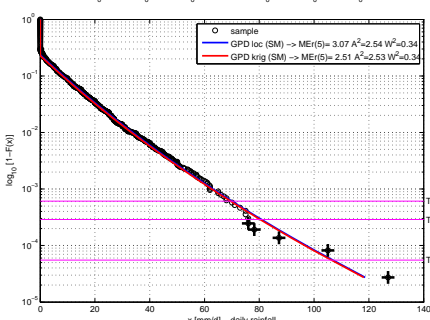
CDF of MTM GPD fit - st 182 [Reinamare (C.ra)] - 53 years
SM ($\xi_0^M = 0.03 - \alpha_0^M = 9.20 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.06 - \alpha_0^M = 9.03 - \xi_0^M = 0.22$)



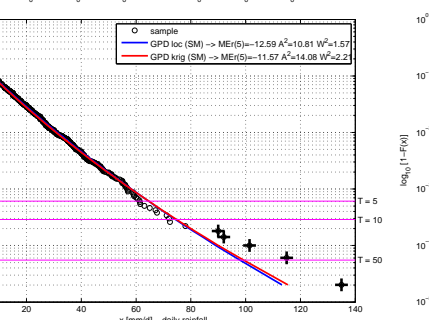
CDF of MTM GPD fit - st 184 [Pozzomaggiore] - 32 years
SM ($\xi_0^M = 0.03 - \alpha_0^M = 9.88 - \xi_0^M = 0.21$) - KRIG ($\xi_0^M = 0.04 - \alpha_0^M = 9.51 - \xi_0^M = 0.22$)



CDF of MTM GPD fit - st 187 [Campeda] - 50 years
SM ($\xi_0^M = 0.06 - \alpha_0^M = 9.62 - \xi_0^M = 0.24$) - KRIG ($\xi_0^M = 0.07 - \alpha_0^M = 9.61 - \xi_0^M = 0.23$)



CDF of MTM GPD fit - st 188 [Borona] - 68 years
SM ($\xi_0^M = 0.06 - \alpha_0^M = 9.19 - \xi_0^M = 0.22$) - KRIG ($\xi_0^M = 0.07 - \alpha_0^M = 8.93 - \xi_0^M = 0.22$)



CDF of MTM GPD fit - st 191 [Sindia F.C.] - 72 years
SM ($\xi_0^M = 0.00 - \alpha_0^M = 11.46 - \xi_0^M = 0.24$) - KRIG ($\xi_0^M = 0.03 - \alpha_0^M = 10.27 - \xi_0^M = 0.23$)

