INTERNATIONAL JOURNAL OF CLIMATOLOGY
*Int. J. Climatol.* **18**: 1031–1047 (1998)

# COMPARISON OF GEOSTATISTICAL METHODS FOR ESTIMATING THE AREAL AVERAGE CLIMATOLOGICAL RAINFALL MEAN USING DATA ON PRECIPITATION AND TOPOGRAPHY

EULOGIO PARDO-IGÚZQUIZA*
*TAMSAT, Department of Meteorology, University of Reading, Reading RG6 6BB, Berkshire, England*

*Received 7 August 1997*
*Revised 22 February 1998*
*Accepted 28 February 1998*

ABSTRACT

The results of estimating the areal average climatological rainfall mean in the Guadalhorce river basin in southern Spain are presented in this paper. The classical Thiessen method and three different geostatistical approaches (ordinary kriging, cokriging and kriging with an external drift) have been used as estimators and their results are compared and discussed. The first two methods use only rainfall information, while cokriging and kriging with an external drift use both precipitation data and orographic information (easily accessible from topographic maps). In the case study presented, kriging with an external drift seems to give the most coherent results in accordance with cross-validation statistics. If there is a correlation between climatological rainfall mean and altitude, it seems logical that the inclusion of topographic information should improve the estimates. Kriging with an external drift has the advantage of requiring a less demanding variogram analysis than cokriging. © 1998 Royal Meteorological Society.

KEY WORDS: techniques; Thiessen method; ordinary kriging; cokriging; kriging with an external drift; rainfall; areal average; Guadalhorce river basin; Spain; geostatistical methods

## 1. INTRODUCTION

Most of the water received by a river basin occurs as rainfall events over the basin. Different quantities are of interest, from total rainfall in a single storm to long-term aggregates over time. Among the latter, mean annual rainfall is used widely in different applications, and varies from year-to-year. The mean of the annual rainfall means for all the years available in the historical record is known as the climatological mean and gives the expected annual rainfall. However the climatological mean is usually known only at the locations of raingauge stations; to estimate it at unsampled locations over the whole basin (and the expected total water input by rainfall), a procedure is required to calculate the average spatial climatological rainfall mean. To calculate average spatial statistics in climatology, the Thiessen method (Thiessen, 1911) was the standard for about half a century until the development of geostatistics in the middle 1960s (Matheron, 1963). The geostatistical method of spatial estimation, known generically as kriging, takes into account the spatial correlation between experimental data through the variogram function. Kriging improves on the Thiessen method in two ways:

(i) More precise estimates due to using the information from surrounding samples.
(ii) It minimises the estimation variance which in turn provide a measure of the uncertainty of the estimates.

* Correspondence to. Department of Meteorology, University of Reading, Reading RG6 6BB, UK; tel.: +44 118 9316705; fax: +44 118 9318905; e-mail: E.Pardo@reading.ac.uk

Contract grant sponsor: European Community Training Fellowship (part of the Training and Mobility of Researches (TMR) programme)

CCC 0899–8418/98/091031–17$17.50

Thus kriging has currently become the standard method of estimating average rainfall values at sites with no records. Early applications of kriging in rainfall estimation were described by Delhomme and Delfiner (1973). Since then there have been many applications and different kinds of kriging have also been developed. For example, multivariate techniques to deal with at least two variables that are coregionalized. This is the case where the rainfall pattern in an area is influenced by physiography, i.e. orographic rainfall. The number of raingauges is usually small and their spatial distribution sparse while, topographical information is widely available. Then, if the rainfall is coregionalized with altitude, the topographical data may be used to improve the precision of the rainfall estimates.

The aim of this paper is to show the results of the application of different geostatistical methods to estimate the areal average climatological mean using precipitation data and topography. A review of the geostatistical techniques and the results of their application to the Guadalhorce river basin in the south of Spain are presented and discussed in the following sections.

## 2. THEORY REVIEW ON GEOSTATISTICAL METHODS

Given a river basin of area $A$ (Figure 1(a)), the expected annual input $Q$ by rainfall is given by:

$$Q = \int_A p(x)\,\mathrm{d}x \tag{1}$$

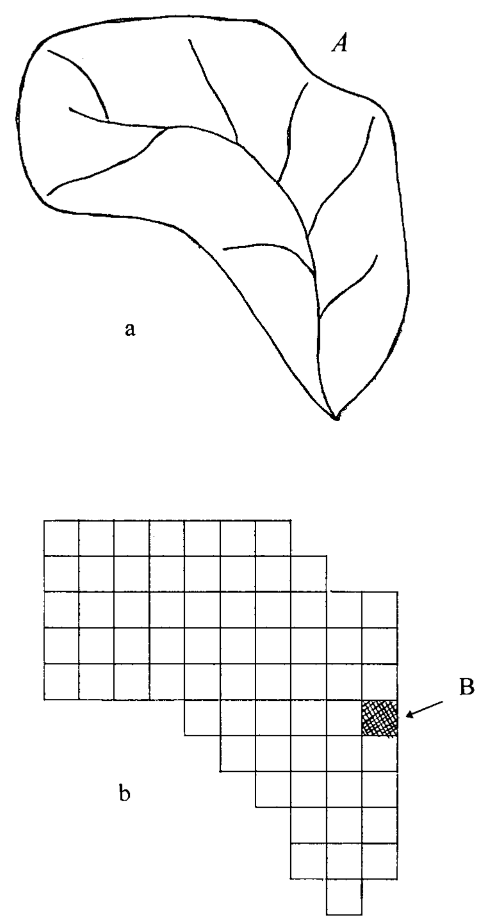


Figure 1. River basin (a) and discretized river basin (b)

where $p(x)$ is the value of the climatological rainfall mean at spatial location $x$ ($x$ denoting the coordinates of a point in the plane). $p(x)$ is only known at those points where raingauge stations are located, and then it must be estimated at the non- sampled points. In practice, the integral (1) is evaluated using the discrete sum approximation:

$$Q = \sum_{i=1}^{N} p_B(x_i) \tag{2}$$

where the area $A$ of the river basin has been discretized in $N$ cells of equal area $B$ (Figure 1(b)); and $p_B(x_i)$ is the areal average climatological rainfall mean over cell $B$ with centroid $x_i$.

It is clear that $p_B(x_i)$ is an areal average estimate (also referred to as a block estimate):

$$p_B(x_i) = \frac{1}{B} \int_{B(x_i)} p(x)\,\mathrm{d}x \tag{3}$$

where $B$ is called the support of the estimate, being:

$$A = \sum_{i=1}^{N} B(x_i) \tag{4}$$

the whole region, and $B(x_i)$ denotes the support $B$ at the location $x_i$. Block kriging provides estimates of Equation (3) from the experimental data with punctual support (for example data measured in raingauge stations):

$$\hat{p}_B(x_i) = \sum_{j=1}^{n} \lambda_j \cdot p(x_j) \tag{5}$$

with $p(x_j)$ being the climatological mean at location $x_j$.

In equation (5) only rainfall data (the climatological mean) are used. There are different possibilities for performing such an estimation using geostatistics; there are also different ways of incorporating the topographical information for the purpose of improving estimates $\hat{p}_B(x_i)$, and then $Q$. The methods of estimation are briefly explained below.

### *2.1. Thiessen method*

The Thiessen polygons method is equivalent to the nearest neighbor method; all the points inside the polygon have the same value which is equal to the closest sample value; $p(x)$ in Equation (3) is then estimated as the nearest experimental point. In practice, this is solved by the well-known method of drawing Voronoï polygons around each experimental data. However, an easier procedure has been adopted here as it is more suitable for computation. The support $B$ is subdivided into smaller square units and each small unit is estimated by the nearest neighbor method. If each subunit is sufficiently small this procedure gives a good approximation to the classical triangulation method.

### *2.2. Ordinary kriging (OK)*

In ordinary kriging the estimate $\hat{p}_B(x_o)$ is given by equation (5) where the weights $\lambda_i$ are obtained by solving to the ordinary kriging system (Journel and Huijbregts, 1978):

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \cdot \gamma_p(x_i, x_j) + \mu = \bar{\gamma}_p(x_i, B(x_o)) \\ \qquad\qquad\qquad\qquad\qquad ; \; i = 1, \ldots, n \\ \sum_{j=1}^{n} \lambda_j = 1 \end{cases} \tag{6}$$

And the estimation variance is given by:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \lambda_i \cdot \bar{\gamma}_p(x_i, B(x_o)) + \mu \tag{7}$$

where $\mu$, Lagrange multiplier; $\gamma_p(x_i, x_j)$, variogram function between points $x_i$ and $x_j$; $\bar{\gamma}_p(x_i, B(x_o))$, mean variogram function between point $x_i$ and support $B$ with centroid $x_o$; and $\gamma_p$, denotes variogram of the climatological rainfall mean.

The two previous methods use only rainfall data (the climatological mean). Information on altitude may be included in the estimation procedure in different ways, two of them being cokriging and kriging with an external drift.

### 2.3. Cokriging (CoK)

The CoK estimate of $p_B(x_o)$ is given by:

$$\hat{p}_B(x_o) = \sum_{i=1}^{n} \lambda_i \cdot p(x_i) + \sum_{j=1}^{m} \eta_j \cdot y(x_j) \tag{8}$$

where $\{y(x_j); j = 1, \ldots, m\}$ are experimental data on altitude. The weights $\lambda_i$, $\eta_j$ are obtained as solution of the CoK system:

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \cdot \gamma_p(x_i, x_j) + \sum_{k=1}^{m} \eta_k \cdot \gamma_{pY}(x_i, x_k) + \mu_1 = \bar{\gamma}_p(x_i, B(x_o)) \quad i = 1, \ldots, n \\ \sum_{j=1}^{n} \lambda_j \cdot \gamma_{pY}(x_k, x_j) + \sum_{t=1}^{m} \eta_t \cdot \gamma_Y(x_k, x_t) + \mu_2 = \bar{\gamma}_{pY}(x_k, B(x_o)) \quad k = 1, \ldots, m \\ \sum_{j=1}^{n} \lambda_j = 1 \\ \sum_{j=1}^{m} \eta_j = 0 \end{cases} \tag{9}$$

with estimation variance:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \lambda_i \cdot \bar{\gamma}_p(x_i, B(x_o)) + \sum_{k=1}^{m} \eta_k \cdot \bar{\gamma}_{pY}(x_k, B(x_o)) + \mu_1 \tag{10}$$

where: $\mu_1$, $\mu_2$, Lagrange parameters; $\gamma_p(h)$, the variogram function of the climatological mean; $\gamma_Y(h)$, variogram of altitude; $\gamma_{pY}(h)$, cross-variogram of climatological mean and altitude.

The bar over the variogram denotes the mean variogram between a point and a support $B$.

### 2.4. Kriging with an external drift (KED)

In KED the mathematical expectation of the climatological rainfall mean random function $P(x)$ is expressed as a linear function of altitude $Y(x)$:

$$E[P(x)] = a_1 + a_2 Y(x) \tag{11}$$

The estimate $\hat{p}_B(x_o)$ has the form given in Equation (4), with the weights $\lambda_i$ obtained as solutions of the KED system (Ahmed and De Marsily, 1987; Hudson and Wackernagel, 1994):

$$\begin{cases} \sum_{j=1}^{n} \lambda_j \cdot \gamma_p(x_i, x_j) + \mu_1 + \mu_2 y(x_i) = \bar{\gamma}(x_i, B(x_o)) \quad i = 1, \ldots, n \\ \sum_{j=1}^{n} \lambda_j \cdot y(x_j) = y(x_o) \\ \sum_{j=1}^{n} \lambda j = 1 \end{cases} \tag{12}$$

with estimation variance:

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \lambda_i \cdot \bar{\gamma}_p(x_i, B(x_o)) + \mu_1 + \mu_2 y(x_o) \tag{13}$$

where all the notation has been defined previously.

It is clear from Equation (12) how the values of the coefficients $a_1$ and $a_2$ are not needed; it is not necessary to know them in order to apply KED. In order to apply KED it is necessary to know the value of the mean altitude of the block centered at each experimental location, as well the mean altitude of the block to be estimated. While CoK requires the variograms of climatological mean and rainfall as well as the crossvariogram, KED and OK require only the variogram of the climatological mean.

## 3. CASE STUDY

The study area is the Guadalhorce river basin, located between longitude 4° and 6° west and between latitude 36° and 38° north (Figure 2). The area of the basin is discretized to squares of 4 km side, and is 2864 $Km^2$. There are 51 raingauge stations located within or around the river basin, as depicted in Figure 3; in the figure the coordinates are given in km with an arbitrary origin.

The climatological mean (mean annual rainfall calculated from historical records covering more than 20 years in all cases), and the altitude are known for every raingauge station. Figure 4 shows the climatological mean plotted against altitude; the linear correlation coefficient is 0.63. On average, the climatological mean increases with altitude statistically. The physical explanation of this correlation coefficient is complex because there are different meteorological factors that influence the rainfall pattern in this basin, among them the wind patterns and the effect of the Mediterranean Sea in the south of the basin (Figure 2).

Figure 5 shows a contour map of the mean altitude over blocks of $4 \times 4$ km calculated from a topographical map by averaging the point altitudes of a regular grid of points over each block. The highest altitudes are located at the western and eastern borders of the river basin.

The expected annual resources input in the basin is obtained by the evaluation of $\hat{p}_B(x_i)$ for $N = 179$ blocks of $4 \times 4$ $Km^2$ in Equation (2).

The contour maps of the estimates using Thiessen polygons, OK, CoK and KED are shown in Figure 6(a–d), respectively. In the figures, the climatological rainfall mean is given in meters and the isohyet interval is 50 mm (0.05 m). The average areal climatological mean over the whole river basin is 586.9, 591.1, 606.5 and 632.6 mm for the four methods, respectively. Thus the estimates of the expected annual input of water resources are 1681, 1693, 1738 and 1812 $hm^3$ obtained for Thiessen, OK, CoK and KDE, respectively. In decreasing order of the magnitude of their estimates the estimators are KED, CoK, OK and Thiessen; KED gives 4% more annual input resources than CoK, 7% more than OK and 8% more than the Thiessen method; CoK gives 3% more input resources than OK and the Thiessen method. The difference between OK and the Thiessen method is less than 1%. Figure 6 shows a great similarity between the Thiessen estimates (Figure 6(a)) and OK estimates (Figure 6(d)). The isohyet map of KED (Figure

 

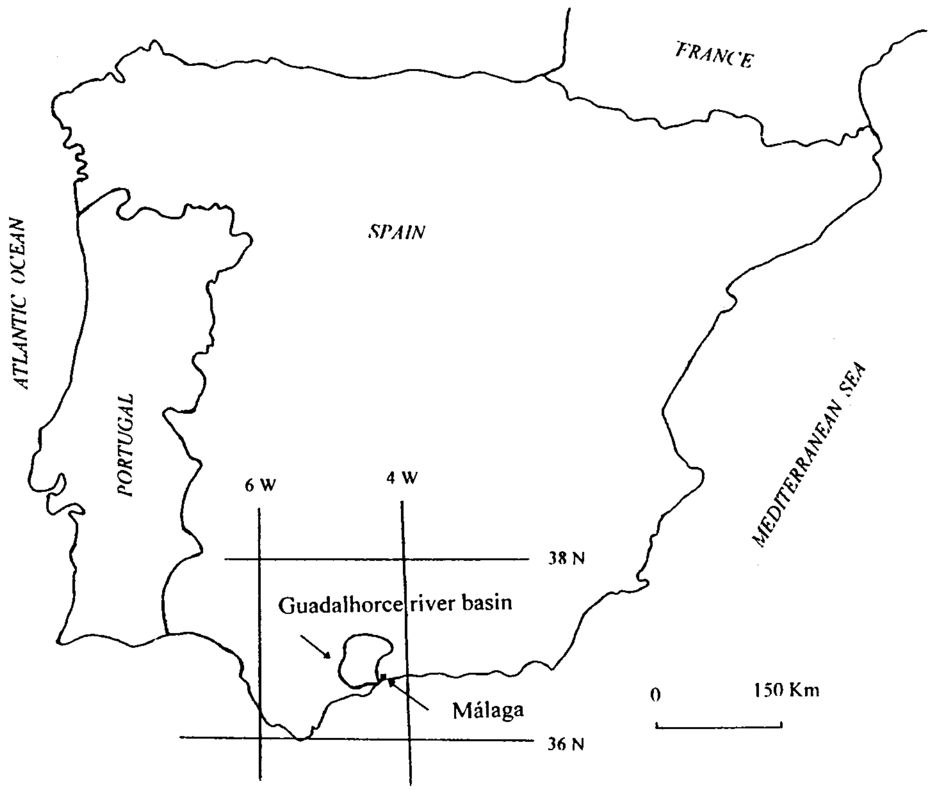


Figure 2. Location of Guadalhorce river basin in the south of Spain

6(d)) has greater similarity with the topography (Figure 5) and the CoK isohyet map also resembles the topography but to a lesser extent than KED. For the application of CoK and KED, the values used have been 309 of mean altitudes over blocks $4 \times 4$ km which corresponds to a $20 \times 17$ grid of blocks over the area represented in Figure 3, after discounting the blocks over the sea with altitude 0.

The experimental variograms of climatological mean and altitude and the experimental cross-variogram between climatological mean and altitude, as well as the fitted models used in the different kriging systems, are shown in Figure 7(a–c). The correspondent models are:

$$\gamma_p(h) = 0.0313 \cdot Sph(h)_{40} \tag{14}$$

$$\gamma_Y(h) = 0.0733 \cdot Sph(h)_{40} \tag{15}$$

$$\gamma_{pY}(h) = 0.0345 \cdot Sph(h)_{40} \tag{16}$$

where $c \cdot Sph(h)_a$ denotes a spherical variogram with sill $c$ and range $a$. These variograms have been estimated by fitting a model to the nonparametric variogram estimates shown in Figure 7. The variograms have been modelled by the same spherical structure in order to apply the linear model of coregionalization in CoK (Journel and Huijbregts, 1978). The models fitted to the nonparametric estimates (Equations 14–16) are in agreement well with the parameters estimated by maximum likelihood (Pardo-Igúzquiza, 1997), the results of which are shown in Table I.

From the results in Table I several considerations may be extracted:

- The Akaike information criterion (Akaike, 1974) suggests that the models without nugget effect for both regionalized variables, climatological mean and topography are more plausible; and, even when the nugget effect is estimated, its value is very small in relation to the variance.

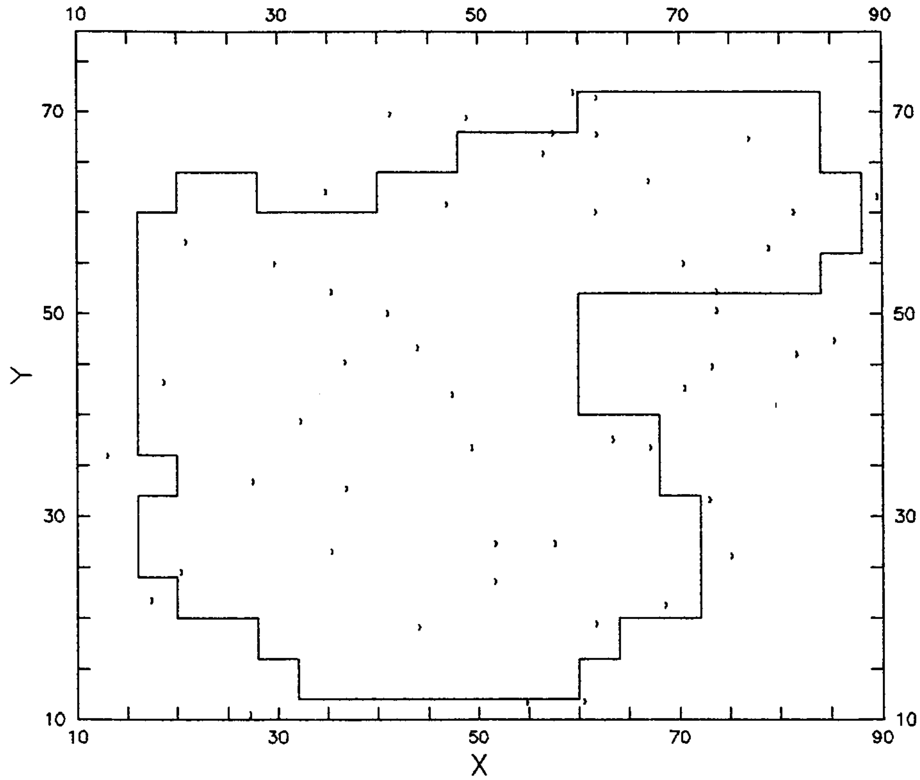


Figure 3. Guadalhorce river basin and location of the 51 raingauge stations

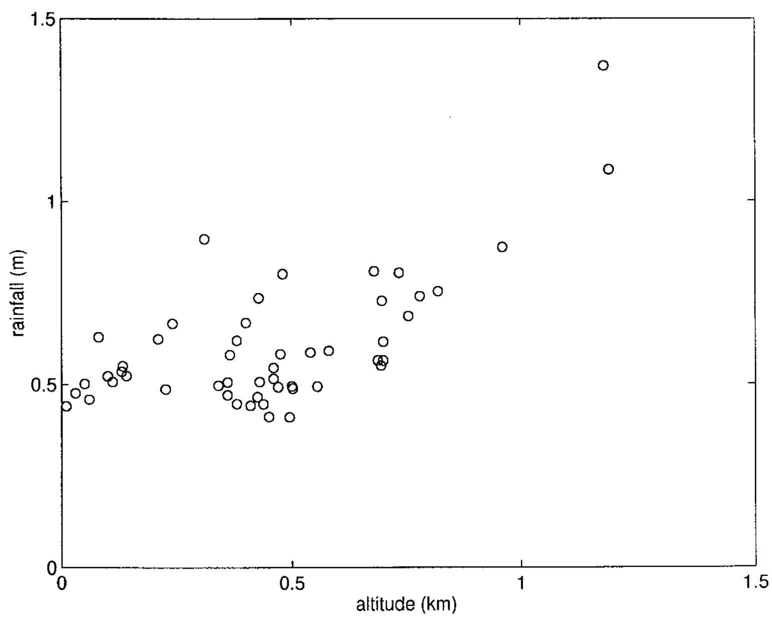


Figure 4. Climatological rainfall mean versus altitude

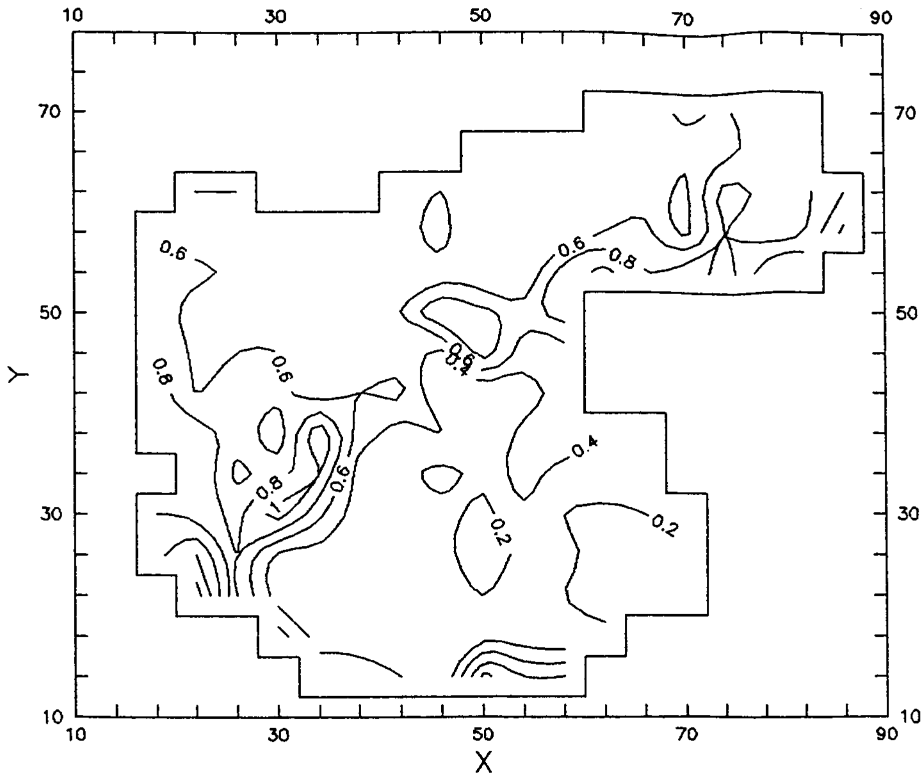


Figure 5. Cartography of the mean altitude over block (4 × 4 Km) support

- The maximum likelihood estimates for the range are 39.3 km with standard error of 5.33 km, and 34.7 km with standard error of 4.5 km for climatological rainfall and altitude, respectively. The maximum likelihood estimates for the variance are 0.044 $m^2$ with standard error of 0.0087 $m^2$ and 0.059 $km^2$ with standard error of 0.012 $km^2$.
- The model Equation (14) for the climatological rainfall variogram is coherent with the model estimated by maximum likelihood. For topography the model suggested by maximum likelihood has a smaller range but the model given in Equation (15) is still within the 95% confidence interval for the range estimated by maximum likelihood.

The sample size (51 locations in Figure 3) is large from a hydrological point of view; to have 51 raingauges inside a $80 \times 78 = 6240$ $km^2$ (area of the rectangle of Figure 3) is a high density network. From a statistical point of view, it is still an adequate number of data; Pardo-Igúzquiza (1997) shows how with more than 30 data, the maximum likelihood estimator gives good results for the inference of spatial covariance parameters. What is more important, a standard error is attached to the estimated parameters.

The uncertainty of the climatological rainfall mean estimates are assessed by the estimation variance, which is not available for the Thiessen method. A representation of the square root of the estimation variance (the S.D. of the estimation error) known as kriging error, or standard error, has been represented in Figure 8(a–c) for OK, CoK and KED, respectively. The standard error of the global average (the average for the whole river basin) climatological mean is 79.86mm, 79.88 mm and 79.14 mm for OK, CoK and KED, respectively (see Appendix 1). The standard errors for the global average are almost the same, but when we consider the local standard errors, as can be seen from Figure 8(a–c), the differences are noticeable. KED gives a larger kriging error than CoK and OK, while OK gives a kriging error larger than CoK. The latter case is perfectly logical because CoK uses additional information on altitude as well as the variogram of altitude, and the cross- variogram between climatological mean and altitude; a decrease of the uncertainty in the estimates is to be expected.

 

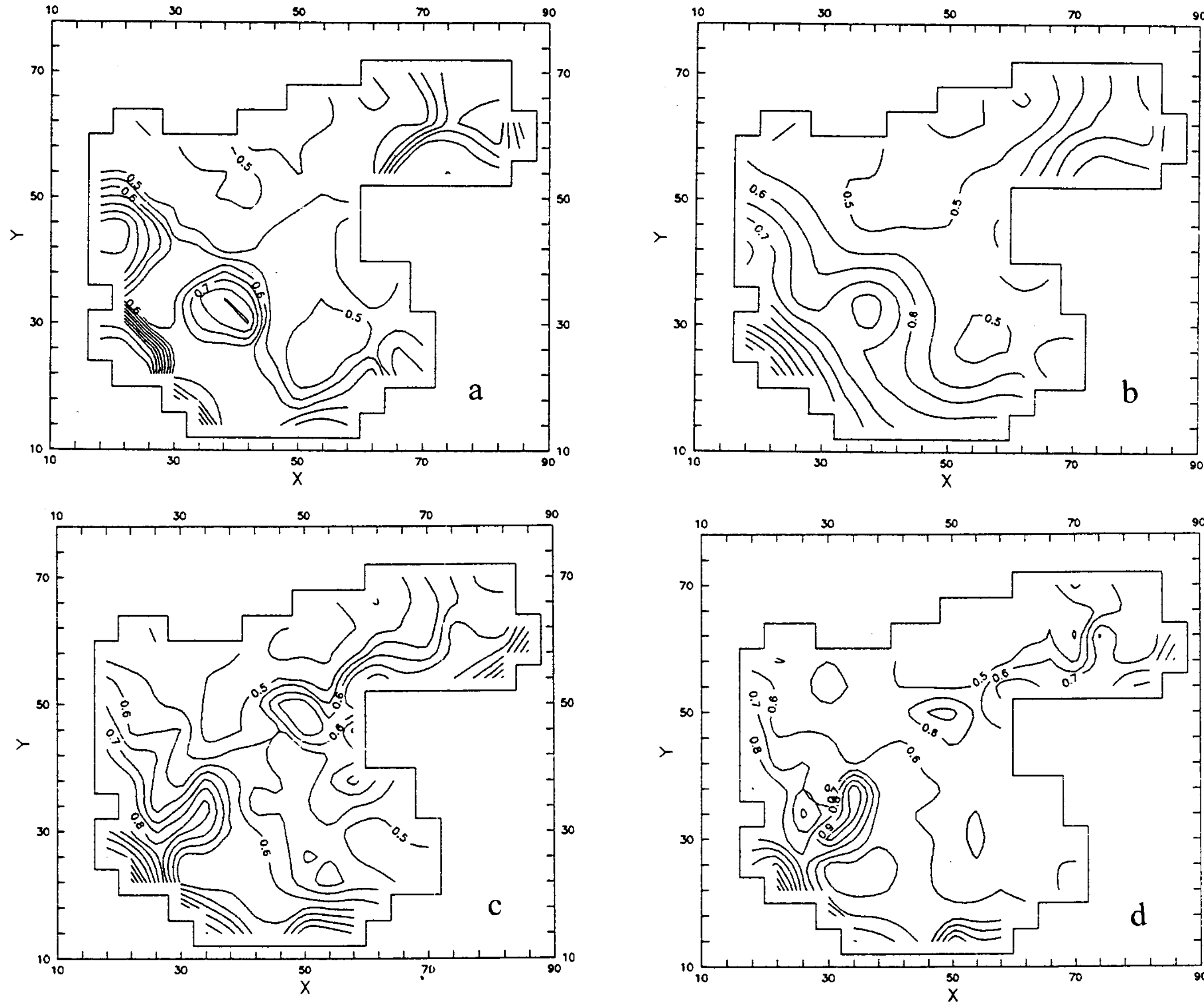


Figure 6. Cartography of the areal average climatological rainfall mean estimated by: (a) Thiessen; (b) OK; (c) CoK; and (d) KED

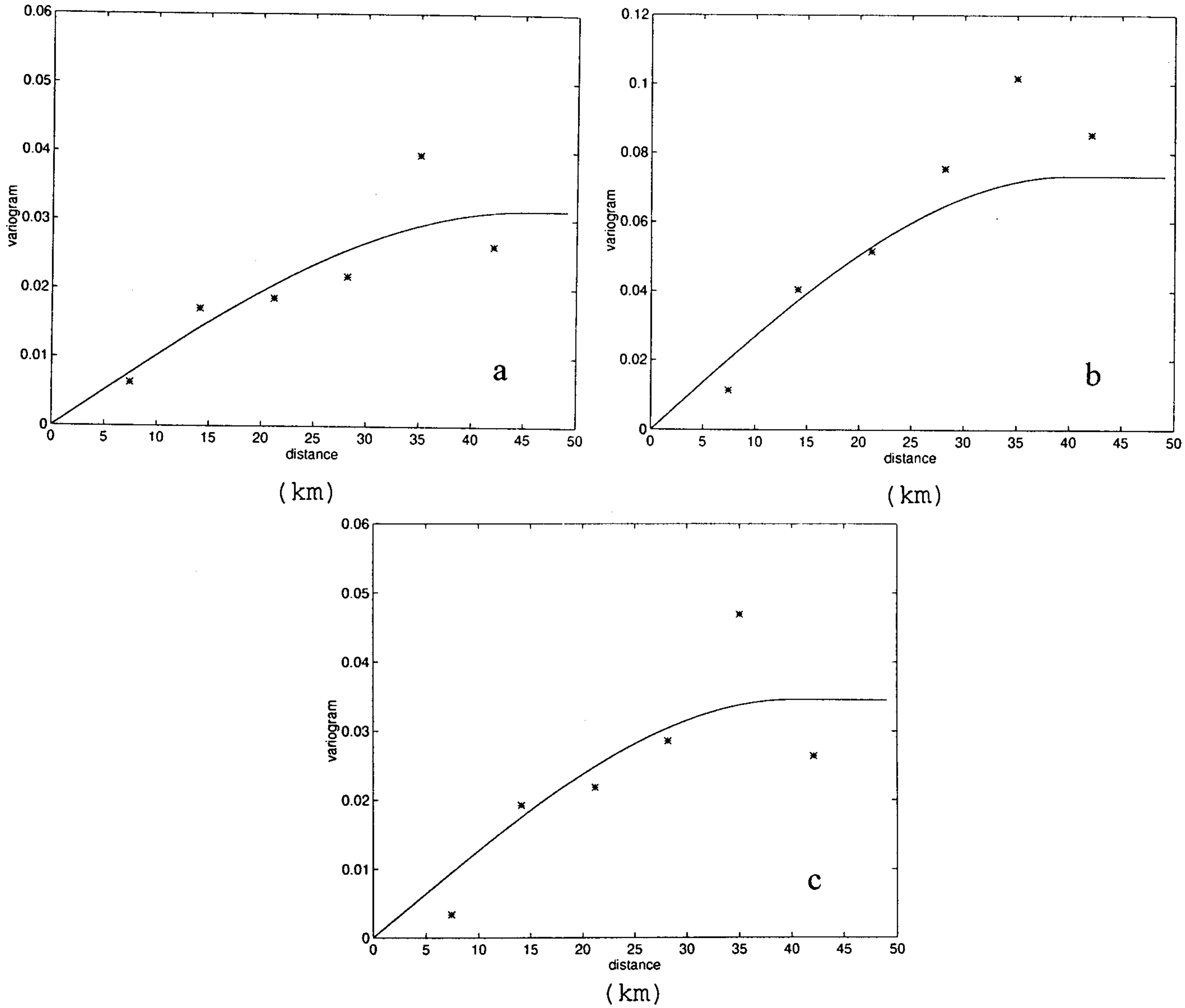


Figure 7. Experimental (*) and model (continuous line) for: (a) climatological mean; (b) altitude; and (c) cross-variogram between climatological mean and altitude

Table I. Variogram parameters estimated by maximum likelihood

| Regionalized variable | Range | S.E. | Nugget effect | S.E. | Variance | S.E. | NLLF | AIC |
|---|---|---|---|---|---|---|---|---|
| Rainfall | 39.3 | 5.3 | NA | NA | 0.0441 | 0.0087 | −32.732 | −59.464 |
| Rainfall | 59.4 | 16.6 | 0.0041 | 0.0012 | 0.0409 | 0.0012 | −33.910 | −59.820 |
| Altitude | 34.7 | 4.5 | NA | NA | 0.0594 | 0.0117 | −22.247 | −38.494 |
| Altitude | 34.9 | 5.2 | 0.0000 | 0.0014 | 0.0598 | 0.0118 | −22.244 | −36.488 |

S.E.: standard error; NLLF: negative loglikelihood function; NA: not applicable.
Range in km; variance and nugget effect in $m^2$ and $km^2$ for rainfall and nugget effect, respectively.
AIC: Akaike information parameter; $AIC = -\log L + 2k = NLLF + 2k$, where $L$: likelihood function, NLLF: negative loglikelihood function and $k$ is the number of parameters fitted to the model (equal to 3 (mean, range, variance) and 4 when the nugget effect is fitted (mean, range, variance, nugget effect).

In the KED system given in Equation (12), the variogram of the climatological mean $\gamma_p(h)$ is used, though the variogram that should be used is the variogram of the unknown residuals $r(x_i) = p(x_i) - a_1 - a_2 y(x_i)$. This approximation does not modify the estimates of $p(x)$, but overestimates the variance of the estimation error (Ahmed and De Marsily, 1987).

## 4. CROSS-VALIDATION RESULTS

Cross-validation is frequently used to validate or compare variogram models. The procedure, as is well-known, consists of the estimation of each experimental value using the remainder of the experimental information (i.e. excluding itself). This makes the calculation of the true error in the estimation procedure possible:

$$\varepsilon_i = \hat{p}(x_i) - p(x_i) \tag{17}$$

where: $\varepsilon$ is the error; $\hat{p}(x)$ is the estimated value and $p(x)$ is the true value. Thus, repeating this estimation for the number of experimental data $n$, the following cross-validation statistics may be calculated.

Mean error:

$$ME = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \tag{18}$$

mean squared error

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 \tag{19}$$

mean standardized square error

$$MSSE = \frac{1}{n} \sum_{i=1}^{n} \frac{\varepsilon_i^2}{\hat{\sigma}_i^2} \tag{20}$$

It is assumed that if the variogram model is correct, ME should be almost zero, MSE a small value and MSSE close to 1. It should be noted that ME is useless, as $ME = 0$ is imposed in the kriging system by the unbiased conditions, no matter which variogram model is used. The value of cross-validation has been criticized by some authors (Journel and Rossi, 1989). If cross-validation works efficiently, it can be used for estimation purposes. This is the minimum interpolation error method of Lebel and Bastin (1985), which entails choosing the values of the variogram parameters that give minimum MSE and MSSE equal to 1. Pardo-Igúzquiza (1998) shows that when five variogram estimators are compared, the one using cross-validation (the previous method) gives the worst results. Therefore cross-validation must be used with caution, small differences in the cross-validation results may be given just by sampling variability. The cross-validation statistics for OK, CoK and KED with the 51 experimental data of climatological

 

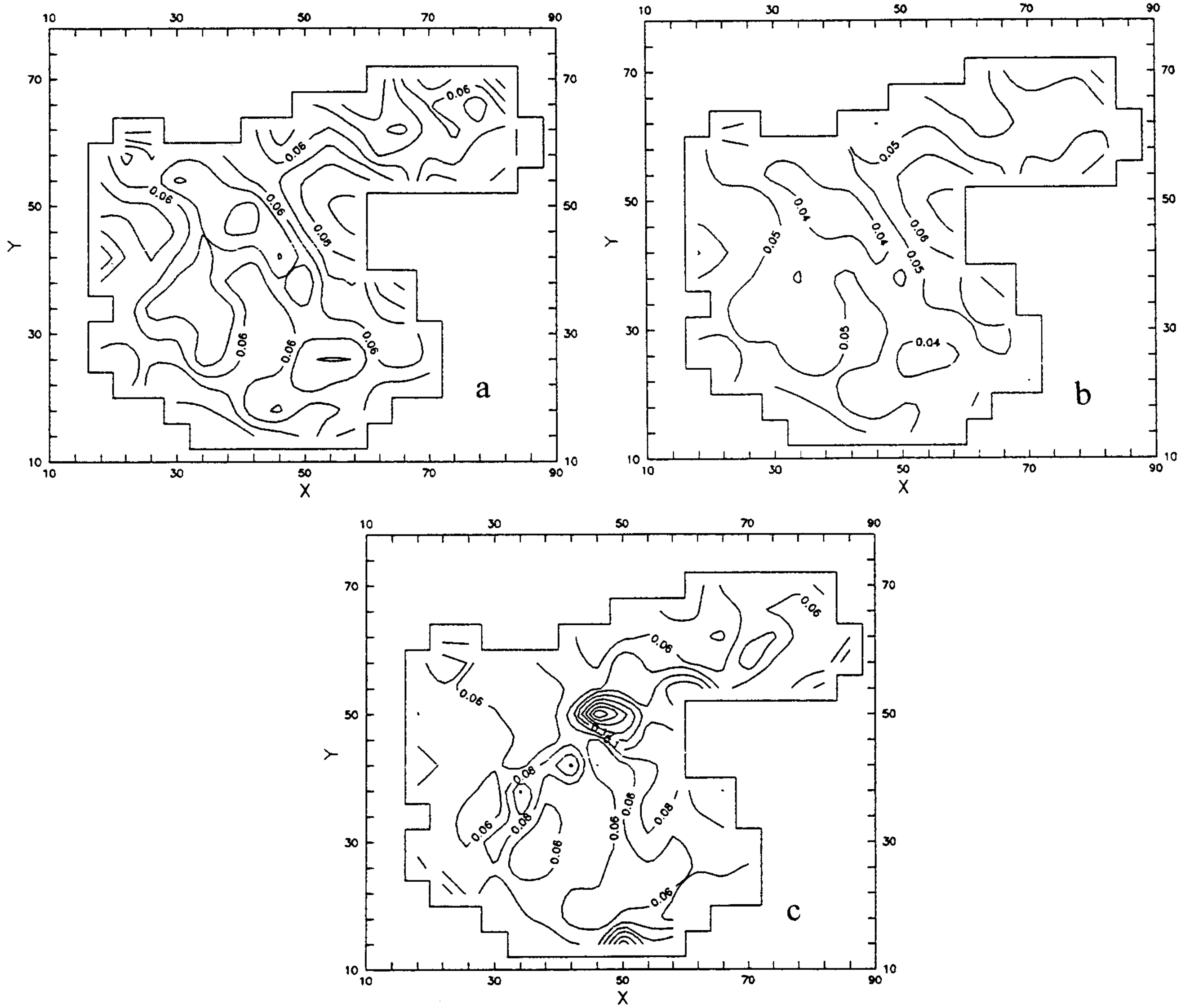


Figure 8. Cartography of the kriging error for: (a) OK; (b) CoK; and (c) KED

Table II. Cross-validation statistics

| | ME | MSE | MSSE |
|---|---|---|---|
| OK | 0.0015 | 0.0147 | 1.7939 |
| CoK | −0.0099 | 0.0137 | 2.7606 |
| KED | −0.0021 | 0.0097 | 1.2008 |

mean are shown in Table II. The KED approach gives the most congruent results, with the smallest MSE and the MSSE closest to 1. The values of MSSE for OK and CoK suggest that the sill of the variograms was underestimated, but that the estimation variance obtained by KED was not.

If the distribution of the errors is Gaussian (which is generally the case, at least approximately by the central limit theorem), 95, 68, 38 and 19% confidence intervals for the climatological rainfall mean estimates may be constructed as $[\hat{p} \pm 2\hat{\sigma}]$, $[\hat{p} \pm \hat{\sigma}]$, $[\hat{p} \pm 0.5\hat{\sigma}]$ and $[\hat{p} \pm 0.25\hat{\sigma}]$, where $\hat{p}$ is the estimate and $\hat{\sigma}$ is its kriging error. It can then be calculated in which percentages from among the previous random intervals the true values are contained. Because there are 51 experimental data, the expected number of times that the random intervals will contain the true values are 48, 35, 19 and 10 for the previous intervals, respectively. The results for OK, CoK and KED are shown in Table III. The best results are obtained with KED.

Other representations of the cross-validation results are given in Figures 9 and 10. In Figure 9(a–d), the estimated values versus the true values of climatological mean are represented for Thiessen, OK, CoK and KED, respectively. The points should be around the 45° line. Again the best results are obtained for KED with a linear correlation coefficient of 0.83. The linear correlation coefficient for Thiessen, OK and CoK are 0.75, 0.74 and 0.77, respectively. Figure 10(a–d) show the errors given by equation (17) versus the estimated values. The magnitude of the errors should be independent of the magnitude of the estimated values. All the methods yield the larger errors when the magnitude of the estimate is very high, but KED gives roughly the best results.

## 5. CONCLUSIONS

Among the geostatistical methods used in the estimation of the areal average climatological rainfall mean on a river basin in the south of Spain, ordinary kriging uses only values on the rainfall while cokriging and kriging with external drift use both information on rainfall and on topography. Cokriging has more requirements than kriging with an external drift, as two direct variogram models and a crossvariogram model need to be fitted instead of only one direct variogram model in kriging with an external drift and ordinary kriging. Considering the cross-validation results, kriging with an external drift seems to give the most coherent results in the case study presented; therefore, the estimates of the expected total resources input over the Guadalhorce river basin given by kriging with an external drift can be considered the most appropriate.

Table III. Results from random confidence intervals

| | $[\hat{z} \pm 2\hat{\sigma}]$ (%) | $[\hat{z} \pm \hat{\sigma}]$ (%) | $[\hat{z} \pm 0.56\hat{\sigma}]$ (%) | $[\hat{z} \pm 0.25\hat{\sigma}]$ (%) |
|---|---|---|---|---|
| Theoretical | 45 (95) | 35 (68) | 19 (38) | 10 (19) |
| OK | 42 (82) | 29 (56) | 18 (35) | 8 (15) |
| CoK | 41 (80) | 24 (47) | 15 (29) | 11 (21) |
| KED | 47 (92) | 32 (62) | 20 (39) | 11 (21) |

 

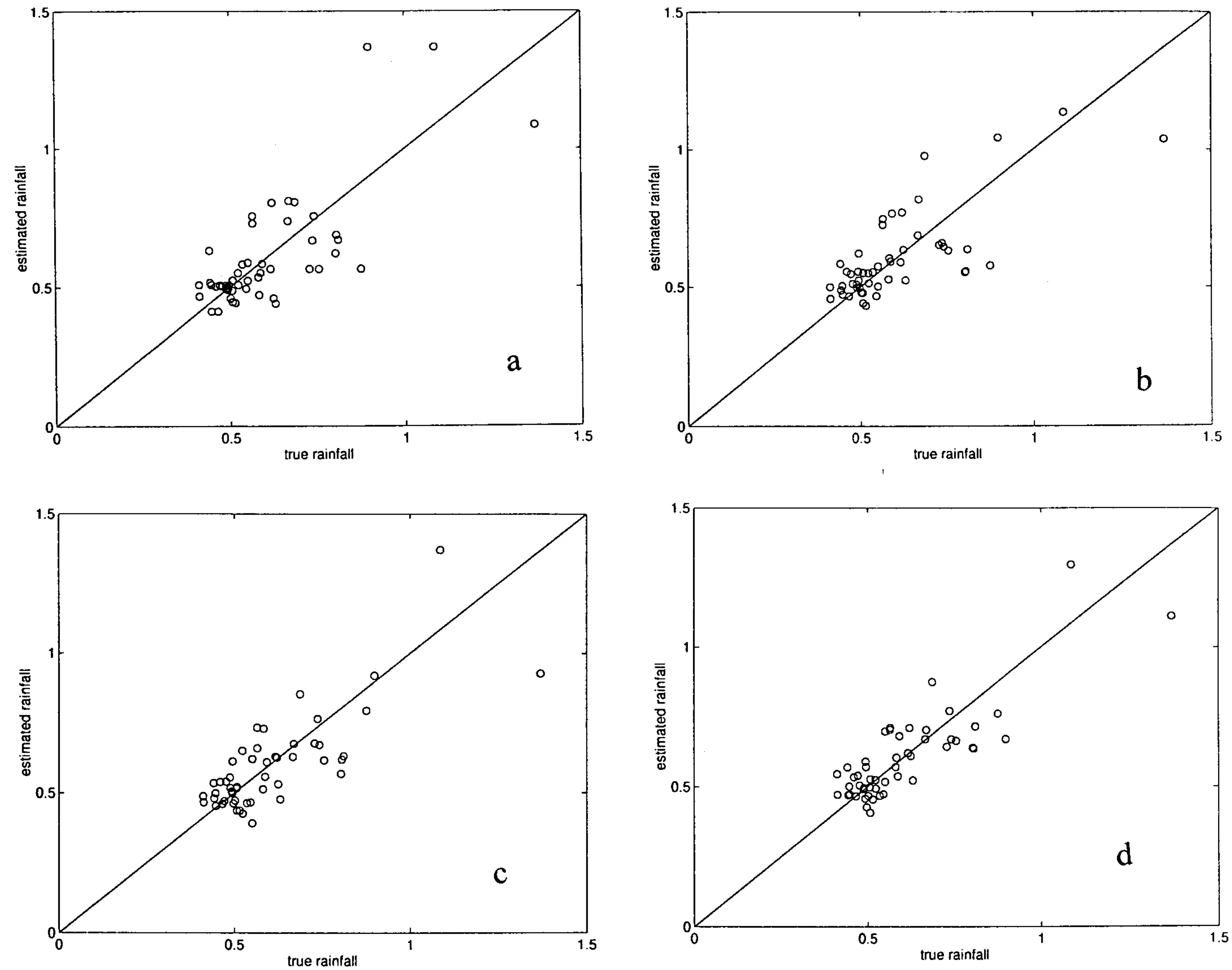


Figure 9. Estimated value versus true value for: (a) Thiessen method; (b) OK; (c) CoK; and (d) KED

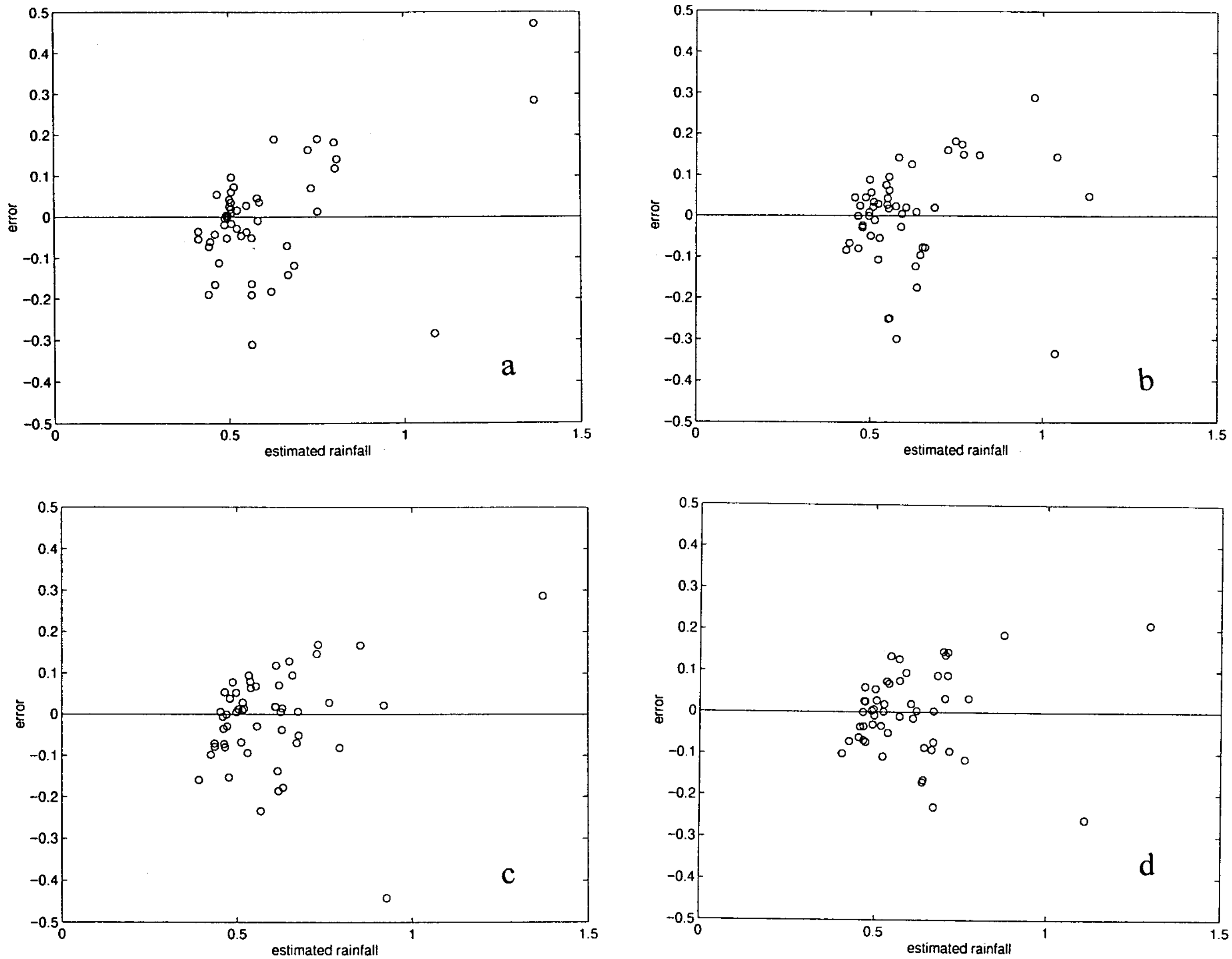


Figure 10. Error vs. estimated value for: (a) Thiessen method; (b) OK; (c) CoK; and (d) KED

## ACKNOWLEDGEMENTS

The work of the author has been supported by a European Community Training Fellowship in the Department of Meteorology (Reading University, UK) as part of the Training and Mobility of Researches (TMR) programme. The author would like to thank the reviewers for providing very constructive comments.


## APPENDIX A. CALCULATION OF THE ESTIMATION VARIANCE OF THE GLOBAL MEAN FROM COMBINATION OF LOCAL ESTIMATION VARIANCES

Figure A1 shows an area $V$ composed of $N$ elementary units of equal area $u_i$:

$$V = \sum_{i=1}^{N} u_i \tag{A1}$$

Over the whole area $V$, there is defined a regionalized variable $Z$ that is only known in $n$ experimental locations. There are $n$ observations that are assumed as realizations of $n$ random variables $\{Z_i, i = 1, \ldots, n\}$. Let $\bar{Z}_V$ be the global mean that is given by

$$\bar{Z}_V = \frac{1}{N} Z_V = \frac{1}{N} \sum_{i=1}^{N} Z_i^u \tag{A2}$$

where $Z_V = \Sigma_{i=1}^{N} Z_i{}^u$ and $Z_i^u$ is the mean value of $Z$ in the $i$-thm cell of size (or support) $u$ and estimated as a weighted sum of the experimental data on point support:

$$\hat{Z}_i^u = \sum_{k=1}^{n} \lambda_k^i \cdot Z_k \tag{A3}$$

where $\lambda_k^i$ is the weight given to the $k$-thm experimental data in the estimation the mean value of $Z$ over the support $u_i$ and calculated as the solution of the kriging system. Kriging also provides the variance of the estimation error:

$$s_i^2 = \mathrm{var}(R_i^u) = \mathrm{var}(\hat{Z}_i^u - Z_i^u) = E\{[\hat{Z}_i^u - Z_i^u]^2\} \tag{A4}$$

Although an unbiased estimate of global mean in Equation (A2) can be easily found from the local estimates in Equation (A3) as:

$$\hat{\bar{Z}}_V = \frac{1}{N} \hat{Z}_V = \frac{1}{N} \sum_{i=1}^{N} \hat{Z}_i^u \tag{A5}$$

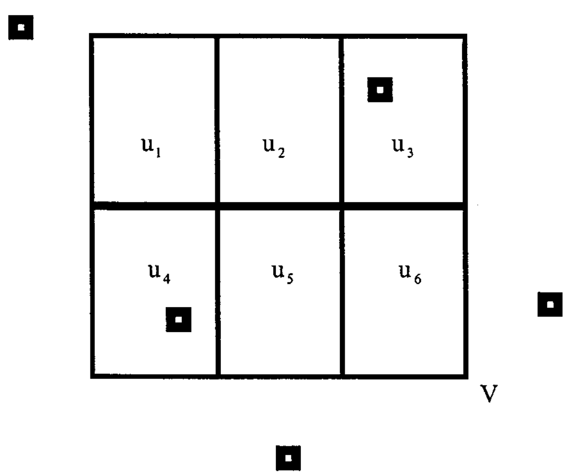


Figure A1. Global domain $V$ as sum of $N = 6$ equal elementary domains of support $u$ and $n = 5$ experimental values (■) of punctual support the calculation of the variance of the estimation error of the global mean is not so straightforward, i.e. because the estimation variances of the different $u$ units are correlated then:

$$\mathrm{var}(R_V) = \mathrm{var}(\hat{\bar{Z}}_V - \bar{Z}_V) = E\{[\hat{\bar{Z}}_V - \bar{Z}_V]^2\} \neq \frac{1}{N}\sum_{i=1}^{N} s_i^2 \tag{A6}$$

The calculation of $\mathrm{var}(R_V)$ is a little more cumbersome that just the arithmetic mean of estimation variances. It is clear that

$$\mathrm{var}(R_V) = \frac{1}{N^2}\mathrm{var}(R_V^*) = \frac{1}{N^2}\mathrm{var}(\hat{Z}_V - Z_V) = \frac{1}{N^2}E\{[\hat{Z}_V - Z_V]^2\} \tag{A7}$$

where $\mathrm{var}(X)$ is variance of the variable $X$. It then requires only some algebra to show that:

$$\mathrm{var}(R_V^*) = \bar{C}_{VV} + \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{n}\sum_{\ell=1}^{n} \lambda_k^i \cdot \lambda_\ell^j \cdot C_{k\ell} - 2\sum_{i=1}^{N}\sum_{k=1}^{n} \lambda_k^i \bar{C}_{kV} \tag{A8}$$

where $\bar{C}_{VV}$ is the mean covariance inside the support $V$; $C_{k\ell}$ is the covariance between $Z_k$ and $Z_\ell$; $\bar{C}_{kV}$ is the mean covariance between $Z_k$ and the support $V$. If the covariance model is pure nugget effect, i.e. there is no correlation between the experimental samples, for the covariance functions:

$$\begin{cases} \bar{C}_{VV} = 0 \\ \bar{C}_{kV} = 0 \\ C_{kl} = 0 \ \text{ if } \ k \neq l \\ C_{kl} = \mathrm{var}(Z) \ \text{ if } \ k \neq l \end{cases} \tag{A9}$$

and for the weights:

$$\lambda_k^i = \frac{1}{n} \qquad \forall i, k \tag{A10}$$

which gives

$$\mathrm{var}(R_V^*) = N^2 \cdot n \cdot \frac{\mathrm{var}(Z)}{n^2} \tag{A11}$$

and then

$$\mathrm{var}(R_V) = \frac{\mathrm{var}(Z)}{n} \tag{A12}$$

which is the classical statistical result that gives the variance of the mean of a random variable.

The fact that the observations are correlated increases the standard error (squared root of the estimation variance) of the mean. In the example presented in the text, the standard error for the global mean, evaluated by any of the kriging procedures presented, is around 80 mm while using standard theory considering that they are independent observations we have, with $n = 51$, and $\mathrm{var}(Z) = 0.0313$, that the standard error is $\sqrt{0.0313/51} \cong 0.025$ m $= 25$ mm.

## REFERENCES


Ahmed, S. and De Marsily, G. 1987. 'Comparison of geostatistical methods for estimating transmissivity using data on transmissivity and specific capacity', *Water Resourc. Res.*, **23**(9), 1717–1737.

Akaike, H. 1974. 'A new look at the statistical model identification', *IEEE Trans. Auto. Control*, **79**, 53–71.

Delhomme, J.P. and Delfiner, P. 1973. 'Application du krigeage a l'optimisation d'une campagne pluviometrique en zone aride', in *Proceedings*, *Symposium on Design of Water Resources Projects with Inadequate Data. Vol. 2*, UNESCO, Madrid, pp. 191–210.

Hudson, G. and Wackernagel, H. 1994. 'Mapping temperature using kriging with external drift: theory and an example from Scotland', *Int. J. Climatol.*, **14**, 77–91.

Journel, A.G. and Huijbregts, Ch. 1978. *Mining Geostatistics*. Academic Press, New York, 600 p.

Journel, A.G. and Rossi, M.E. 1989. 'When do we need a trend model in kriging?', *Math. Geol.*, **21**(7), 715–739.

Lebel, T. and Bastin, G. 1985. 'Variogram identification by the mean-square interpolation error method', *J. Hydrol.*, **77**, 31–56.

Matheron, G. 1963. 'Principles of geostatistics', *Econ. Geol.*, **58**, 1246–1266

Pardo-Igúzquiza, E., 1997. 'MLREML: a computer program for the inference of spatial covariance parameters by maximum likelihood and restricted maximum likelihood', *Comput. Geosci.*, **23**(2), 153–162.

Pardo-Igúzquiza, E. 1998. 'Maximum likelihood estimation of spatial covariance parameters', *Math. Geol.*, **30**(1), 95–108.

Thiessen, A.H. 1911. 'Precipitation for large areas', *Mon. Wea. Rev.*, **39**, 1082–1084.


 