Modeling rainfall data using a Bayesian Kriged-Kalman model

Giovanna Jona Lasinio[1], Sujit K. Sahu[2], and Kanti V. Mardia[3*]

July 12, 2005

SUMMARY

A suitable model for analyzing rainfall data needs to take into account variation in both space and time. The method of kriging is a popular approach in spatial statistics which makes predictions for spatial data. Kalman filtering using dynamic models is often used to analyze temporal data. These approaches have been combined in a classical framework termed kriged Kalman filter (KKF) model. In the combined model, the kriging predictions dictate the optimal regression surface for incorporating spatial structure and the dynamic linear model framework is used to learn about temporal factors such as trends, autoregressive components and cyclical variations. In this article we consider a full Bayesian KKF (BKKF) model for rainfall data and its MCMC implementation. The MCMC techniques provide unified estimation of spatio-temporal effects and allow optimal predictions in time and space. The methods are illustrated with two real data examples. Using many well known validation methods we highlight the advantages of the BKKF model.

Some key words: Gibbs Sampler; Kalman Filter; Kriging; Markov chain Monte Carlo; Spatial Temporal Modeling; State-Space Model.

1 Introduction

Rainfall data observed at many locations over a number of time points typically covary in space and time. These spatio-temporal data types arise in many other contexts such as environmental pollution monitoring and surveillance, disease mapping, and economic monitoring of real estate prices to name a few. As a result the spatio-temporal models and data handling techniques used for data from other applications can be used to analyze rainfall data. Often the primary interests in analyzing spatio-temporal data are to smooth

*[1] =University of Rome "La Sapienza", Giovanna.Jonalasinio@uniroma1.it, [2] =University of Southampton, S.K.Sahu@maths.soton.ac.uk, and [3] =University of Leeds, K.V.Mardia@leeds.ac.uk. This work has been partially supported by the MIUR COFIN project 2004-2006 Statistical methods for forecasting improvement in environmental sciences

and predict time evolution of the response variable over a certain spatial domain. In recent years there has been a tremendous growth in the statistical models and techniques to solve such problems.

Typical rainfall data show strong spatial and seasonal effects and some other covariates, such as the elevation of the observation locations, may also significantly influence the rainfall amounts. Weather radar data are a useful covariate for short-term, e.g. hourly or daily, predictions, see e.g. Brown *et al.* (2001) and Sahu *et al.* (2005). In this paper we model monthly rainfall data observed over several years. Using the models we aim to produce annual maps for comparing different years and thereby detect long-term trend in rainfall.

Several statistical models exist for modeling precipitation; the literature is quite extensive. We only discuss a few papers that inspired the current work, the interested reader is referred to the references therein. In Rodriguez-Iturbe *et al.* (1987, 1988) stochastic point processes based models in space and time are used. A different approach is considered in Smith (1994) where, as in Stern and Coe (1984), they distinguish between processes for wet and dry periods and introduce a positively skewed distribution for the amount of rainfall, conditionally on a wet period.

The amount of rainfall is a continuous random variable, however, the measuring process often implies the rounding of observed values and the occurrences of zero rainfall with positive probability. Many authors have developed methods for handling zero rainfalls using censoring mechanisms, see for example, Dunn (2003), Allcroft and Glasbey (2003), and Sansó and Guenni (1999, 2000). Sahu *et al.* (2005) extend the modeling to account for several rounded discrete rainfall values occurring with non-zero probabilities by incorporating a latent continuous random variable. In their approach a part of the support of the latent variable is categorized to account for the discrete values and the remaining part is left for modeling the continuous rainfall values. The same approach is adopted here.

Geostatistical approaches are usually considered quite sensible when treating rainfall data and radar rainfall data. The use of co-kriging and kriging with external drift in rainfall fields reconstruction started in the early 1990's (Seo *et al.*, 1990a, 1990b, and Raspa *et al.*, 1997) and since then geostatistical models have been used very often in both classical and Bayesian framework (see, e.g. Cassiraga *et al.*, 2004 and Orasi *et al.*, 2005).

The work in the general area of space time modeling has a long history, see Sahu and Mardia (2005b) for a recent review. In a discussion paper Mardia *et al.* (1998) have introduced a combined approach on kriged-Kalman filter (KKF) modeling. Recent papers within this broad framework include Kent and Mardia (2002), Kyriakidis and Journel (1999), Sahu and Mardia (2005a), and Wikle and Cressie (1999).

Kent and Mardia (2002) provide a unified approach to spatio-temporal modeling through the use of drift and/or correlation in space and/or time to accommodate spatial continuity. For drift functions, they have emphasized the use of so called principal kriging functions, and for correlations they have discussed the use of a first order Markov structure in time combined with spatial blurring. Here we adopt one of their strategies but in a full Bayesian framework.

We work here with a process which is continuous in space and discrete in time. The underlying spatial drift is modeled by the principal kriging functions and the time component in observed sites is modeled by a vector random-walk process. The dynamic random-walk process models stochastic trend and the resulting Bayesian analysis essentially leads to Kalman

filtering which is a computational method to analyze dynamic time series data, see e.g. Mardia *et al.* (1998) and Sahu and Mardia (2005a). In addition, the proposed models are presented in a hierarchical framework. This allows the inclusion of a 'nugget' term in the spatial part of the model. Furthermore, we incorporate model components for many interesting spatio-temporal effects, for example, seasonal effects which influence the rainfall and the effect of additional covariates including elevation. We introduce model terms which describe and incorporate these interesting spatio temporal phenomena.

Sahu *et al.* (2005) compare two approaches in modeling radar-rainfall data obtained from a controlled cloud seeding operation. There we successfully use a Gaussian random effect (GRE) model with a separable space-time covariance structure, that performs better then BKKF. In the present work we do not adopt the GRE model as the two monthly rainfall data sets are non-stationary in both space and time and a non-separable model must be used. That is why BKKF is a more sensible choice as it can easily handle non-separable and non-stationary phenomena.

The full Bayesian model is hierarchical, non-linear and incorporates a huge number of parameters. The model is fitted and used for forecasting in a unified computational framework using Markov chain Monte Carlo (MCMC) methods. The MCMC methods allow the estimation of principal kriging functions of space and replace the task of Kalman filtering using random-walk model in time. In addition, all the parameters are estimated using the implemented Markov chain. Optimal spatial predictions and temporal forecasts using predictive Bayesian techniques are also obtained as a byproduct of the coded MCMC methods.

The plan of the remainder of this article is as follows. In Section 2 we describe the Italian rainfall data. Section 3 extends the hierarchical BKKF model to include seasonality and covariates. Section 4 illustrates the methods with two examples. The paper ends with a few summary remarks in Section 5.

2 Italian Rainfall data

In this section we describe a large dataset of monthly rainfall (mm) recorded by 226 raingauges located in south-west Italy (see Figure 1) from 1972 to 1980. The rainguage locations are given by their latitudes and longitudes. Moreover, the elevation information in meters above the sea level is also available. In what follows we shall denote by $Z(\mathbf{s}_i, t)$ $i = 1, \ldots, 226$ and $t = 1, \ldots, 108$ (12 months in 9 years) the amount of rainfall at site $\mathbf{s}_i$ and time t. The dataset comes from the historical data warehouse of the APAT (Italian Environmental Protection Agency). During this period the network was managed by the *Ufficio Compartimentale di Catanzaro* and it covered an area spanning more then 350 kilometers in the North-South direction and about 100 kilometers in the East-West direction. This network covers an area with strong morphological differences: on the southern part there is the Mediterranean on the two sides and a chain of mountain (Sila) in the middle; the northern part is a relatively dry region with few hills and large plains in between, see Figure 2. This figure has been constructed by linearly interpolating elevation information from 226 rainguage locations using the routines `interp` in the software packages `S-Plus` and `R`.

Out of the 226 available rainguages we set aside six sites for validating our model. These

validation sites are carefully chosen using the spatial optimal design utility[1] of the package `fields` in the software `R`. This function finds the set of points on a discrete grid (the coordinate set, i.e. the set of 226 rainguage sites) which minimize a geometric space-filling criterion based, by our choice, on the great circle distance, see the `R` help manual for more in this regard.

There are 23,760 monthly data points from the 220 modeling sites over the 9 year period. There are no missing data and simple summary statistics are reported in Table 1. In the data set about 7% observations are zero values and the rainfall values under one millimeter have been rounded to the 10th of a millimeter. Hence, we shall adopt the latent variable approach detailed in Section 3.1. See also Sahu *et al.* (2005) for more details in this regard.

Time	Min.	1^{st} Q.	Median	Mean	3^{rd}Q.	Max.
monthly data	0.00	19.80	58.00	85.36	120.60	1528.00
annual data	77	698.2	1024.2689	943.5	1272.15	3049

Table 1: Summary statistics of monthly and annual rainfall amount (in mm)

There is considerable variability and non-stationarity in the data due to variation in space and time. Monthly data on the raw scale show a clear mean-variance relationship (see figure 3 (a)) which is almost eliminated by log-transforming observed values (see figure 3 (b)). We shall model on the log scale since it allows us to model the mean and variance independently.

To see monthly seasonal variations we provide a boxplot of log-rainfall values in Figure 4. The month of July is the driest while December and January are the wettest months on average. In Figure 5 (a)-(c) we provide scatter-plots (with linear regression lines superimposed) of log-rainfall and three possible covariates: elevation, latitude and longitude. The plots indicate that latitude is not going to be a significant covariate while longitude can be a worthwhile covariate to include in the model. The strongest covariate information is provided by elevation and this is in accord with the well-known hydrologists' knowledge that elevation is one of the most discriminating factors for raingauge classification. We model elevation in kilometers rather than in meters simply to avoid large values.

The site means show evidence of spatial variation and non-stationarity as can be seen in Figure 6 where averages of log-rainfall, classified according to their quantiles are shown in the map.

We obtain an empirical variogram of the data to investigate spatial variation. We first obtain the residuals after fitting a linear model with month as a factor variable and elevation and longitude as continuous covariates. Let $W(\mathbf{s}_i, t)$ denote the residuals. We suppose that $W(\mathbf{s}_i, t), t = 1, \ldots, T$ are independent replications at location $\mathbf{s}_i, i = 1, \ldots, n$ since we have de-trended the data. We now consider the average variogram defined by

$$\gamma(d_{ij}) = \frac{1}{2T} \sum_{t=1}^{T} E[\{W(\mathbf{s}_i, t) - W(\mathbf{s}_j, t)\}^2]$$

where d_{ij} is the distance between the spatial locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The quantity $\gamma(d_{ij})$ is

[1] `cover.design` in the software `R`, version 2.01, http://www.R-project.org

estimated by

$$\hat{\gamma}(d_{ij}) = \frac{1}{2T}\sum_{t=1}^{T}\{w(\mathbf{s}_i,t) - w(\mathbf{s}_j,t)\}^2.$$

The empirical variogram cloud is obtained by plotting $\hat{\gamma}(d_{ij})$ against d_{ij} for the $n(n-1)/2$ possible pairs of locations. In Figure 7 we provide a smooth loess curve[2] obtained from the variogram cloud which reveals a clear nugget effect. It also shows oscillations at higher distances which may point to possible non-stationarity. The underlying shape of the variogram can be approximated by an exponential covariogram which we assume in Section 3.

The Figures 6, 7 and the time series plots in the validation plot in Figure 9 all point to non-stationary variation in time and space. That is why we describe a non-stationary model with non-separable covariance structure in the next section for these data.

3 The BKKF model

The general model discussed here is for rainfall data recorded at n sites $\mathbf{s}_i$, $i = 1,\ldots,n$, over a period of T equally spaced time points. Let $\mathbf{Z}_t = (Z(\mathbf{s}_1,t),\ldots,Z(\mathbf{s}_n,t))^T$ denote the n-dimensional observation vector at time point t; $t = 1,\ldots,T$. The data may also include covariate information which can be time varying, see Section 3.3 for modeling details.

3.1 Latent variables to model discrete values

As mentioned in Section 2 often rainfall data are rounded and zero rainfall corresponding to dry periods occurs with positive probability. We follow Sahu *et al.* (2005) to model these discrete values with a continuous latent variable, denoted by $\mathbf{X}_t = (X(\mathbf{s}_1,t),\ldots,X(\mathbf{s}_n,t))'$ at time t; $t = 1,\ldots,T$.

Let there be k particular values of log rainfall $\lambda_1, \lambda_2, \ldots, \lambda_k$ which may occur with positive probabilities. Let $c_1,\ldots,c_k$ be constants such that $\lambda_i < c_i$, $i = 1,\ldots,k$. We suppose that the observed log-rainfall value at a site $\mathbf{s}$ at time t is given by

$$\log Z(\mathbf{s},t) = \begin{cases} \lambda_1 & \text{if } X(\mathbf{s},t) < c_1, \\ \lambda_2 & \text{if } c_1 \le X(\mathbf{s},t) < c_2, \\ \vdots & \vdots \\ \lambda_k & \text{if } c_{k-1} \le X(\mathbf{s},t) < c_k, \\ X(\mathbf{s},t) & \text{otherwise.} \end{cases}$$

Here we set $k = 11$ and the values of $\lambda_1,\ldots,\lambda_k$ are $\lambda_1 = \log(0.02)$, $\lambda_2 = \log(0.1)$, ..., $\lambda_{11} = \log(1)$. We choose the constants $c_1,\ldots,c_k$ to be the logarithms of the numbers 0.05, 0.15, ..., 1.05 which are the mid-points of the successive intervals formed of the values 0, 0.1, 0.2 and so on.

The latent random variable $X(\mathbf{s},t)$ for any observed rainfall bigger than 1 millimeter on the original scale is the actual log amount of rainfall. The values of $X(\mathbf{s},t)$ corresponding to the k discrete values of $Z(\mathbf{s},t)$ are simulated from the hierarchical model (1) given below, and

[2]the curve is obtained using the R2.01 function `loess` with smoothing parameter equal to 0.1

are restricted to lie in the intervals implied by the above relationships. That is, for example if $\log Z(\mathbf{s},t) = \lambda_1$ then the corresponding $X(\mathbf{s},t)$ will be simulated in the interval $(-\infty, c_1)$. The inverse relationship between $X(\mathbf{s},t)$ and $Z(\mathbf{s},t)$ is needed for predictive purposes and is straightforward to obtain from the above discussion.

3.2 Hierarchical models

We follow Sahu and Mardia (2005a) to construct a hierarchical BKKF model for the latent variable. We assume the hierarchical model:

$$X(\mathbf{s}_i,t) = Y(\mathbf{s}_i,t) + \epsilon(\mathbf{s}_i,t) \tag{1}$$

where $\mathbf{Y}_t = (Y(\mathbf{s}_1,t),\ldots,Y(\mathbf{s}_n,t))'$ is an unobserved but scientifically meaningful process (signal) and $\boldsymbol{\epsilon}_t = (\epsilon(\mathbf{s}_1,t),\ldots,\epsilon(\mathbf{s}_n,t))$ is a white noise process. Thus we assume $\epsilon(\mathbf{s}_i,t)$ are i.i.d. normal random variables with mean zero and unknown variance σ^2_ϵ.

The space-time process $\mathbf{Y}_t$ is modeled as the sum of a mean process and a spatially colored process:

$$Y(\mathbf{s}_i,t) = \mu(\mathbf{s}_i,t) + \gamma(\mathbf{s}_i,t) \tag{2}$$

where the mean process $\mu(\mathbf{s}_i,t)$ is described below and $\boldsymbol{\gamma}_t = (\gamma(\mathbf{s}_1,t),\ldots,\gamma(\mathbf{s}_n,t))$ is assumed to be zero mean Gaussian with covariance matrix Σ_γ which has elements

$$\sigma(\mathbf{s}_i,\mathbf{s}_j) = \text{Cov}\left(\gamma(\mathbf{s}_i,t),\ \gamma(\mathbf{s}_j,t)\right) \tag{3}$$

for $i,j = 1,\ldots,n$. We assume exponential covariance structure, i.e. $\sigma(\mathbf{s}_i,\mathbf{s}_j) = \sigma^2_\gamma \exp(-\phi\, d_{ij})$ where d_{ij} is the distance between sites $\mathbf{s}_i$ and $\mathbf{s}_j$.

3.3 Models for the mean process

Now we turn to modeling the mean process, $\boldsymbol{\mu}_t = (\mu(\mathbf{s}_1,t),\ldots,\mu(\mathbf{s}_n,t))'$. It is comprised of three terms: (i) a kriged-Kalman filter term as described in Sahu and Mardia (2005a), (ii) a term for modeling the seasonal effects, and (iii) a term to adjust covariate effects. We first describe the last two terms.

As a first attempt the seasonal effects can be modeled by the monthly indicators. However, in many practical situations as in our second example on Venezuelan rainfall data, this may not be sufficient and a more complex set of seasonal harmonics are required. In these situations the seasonal effects are modeled by Fourier representations since those provide adequate flexible models, see e.g. West and Harrison (1997, Chapter 8). Let m be the known periodicity of the data and define $K = m/2$ if m is even and $(m-1)/2$ otherwise. Then, in this paper, we are going to use two different seasonal term, one for each example. For the Italian data analyzed in Section 4.1 we apply the following simple model

$$S_t(\mathbf{s}) = \sum_{j=1}^{12} \rho_j(\mathbf{s})\delta_j(t)$$

were $\delta_j(t)$ are monthly indicators and the unknown coefficients $\rho_j(\mathbf{s})$ may depend on the location $\mathbf{s}$. For the Venezuelan example in Section 4.2 the seasonal term is given by:

$$S_t(\mathbf{s}) = \sum_{r=1}^{K} [c_r(\mathbf{s}) \cos(2\pi tr/m) + d_r(\mathbf{s}) \sin(2\pi tr/m)]$$

where the unknown coefficients $c_r(\mathbf{s})$ and $d_r(\mathbf{s})$ may depend on the site $\mathbf{s}$. When m is even we let $d_r(\mathbf{s}) = 0$ so that $m-1$ free seasonal parameters are kept in the model; the remaining parameter is obtained through the requirement that the seasonal effects cancel each other, i.e. they sum to zero. If there are no justifications for having spatially varying seasonal effects we can work with the simpler model $\rho_j(\mathbf{s}) = \rho_j$, $c_r(\mathbf{s}) = c_r$ and $d_r(\mathbf{s}) = d_r$, as we do in our examples.

If there are J covariates with values $w(\mathbf{s}, t, j)$ for the jth covariate ($j = 1, \ldots, J$) at site $\mathbf{s}$ and at time t, the regression model

$$\sum_{j=1}^{J} \beta_{jt} w(\mathbf{s}, t, j)$$

can be used, where β_{jt} denote the time-varying regression coefficient. This allows for the possibility of including time varying covariate effects which can be significant at certain times and not significant in others. Letting $\beta_{jt} = \beta_j$ for all t in the above, we obtain a simpler model where a constant (over time) regression effect is present.

The mean process is now assumed to be:

$$\mu(\mathbf{s}_i, t) = \sum_{j=1}^{p} h_{\mathbf{s}_i j} \alpha_{tj} + S_t(\mathbf{s}) + \sum_{j=1}^{J} \beta_{jt} w(\mathbf{s}_i, t, j) \tag{4}$$

where the quantities $h_{\mathbf{s}_i,j}$ are defined below, $\boldsymbol{\alpha}_t = (\alpha_{t1}, \ldots, \alpha_{tp})'$ is the state vector of dimension p. We assume that $\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t$, and $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \Sigma_\eta)$. For the initial value we assume that: $\boldsymbol{\alpha}_0 \sim N(\mathbf{0}, C_\alpha I)$ with a large value of C_α where I is the identity matrix.

Let H be the matrix of order $n \times p$ with elements $h_{\mathbf{s}_i,j}$. The matrix H is constructed by using what are known as principal kriging functions, see Mardia *et al.* (1998) and Sahu and Mardia (2005a) for full details. In this implementation we take the fist column of H to be the unit vector, $\mathbf{1}$. The other columns are obtained as follows: We first obtain

$$B = \Sigma_\gamma^{-1} - \frac{1}{\mathbf{1}'\Sigma_\gamma^{-1}\mathbf{1}} \Sigma_\gamma^{-1} \mathbf{1}\mathbf{1}' \Sigma_\gamma^{-1}.$$

We now perform the spectral decomposition of B, $B = UEU'$, $B\mathbf{u}_i = e_i \mathbf{u}_i$, where $U = (\mathbf{u}_1, \ldots, \mathbf{u}_n)$ and $E = \text{diag}(e_1, \ldots, e_n)$, and we assume without loss of generality that the eigenvalues are in non-decreasing order, $e_1 = 0 < e_{2+1} \leq \cdots \leq e_n$. Finally, the matrix H is taken as

$$H = (\mathbf{1}, e_2 \Sigma_\gamma \mathbf{u}_2, \ldots, e_p \Sigma_\gamma \mathbf{u}_p). \tag{5}$$

3.4 Prior distributions

The full Bayesian BKKF model is completed by assuming suitable prior distributions for the parameters. The prior distributions for σ_ϵ^2 and σ_γ^2 are assumed to be the inverse gamma distribution with parameters a and b, $IG(a,b)$. We take $a=2$ and $b=1$ to have a proper but diffuse prior distribution with mean 1 and infinite variance. Let $\boldsymbol{\theta}$ denote the regression co-efficients $\boldsymbol{\beta}$ and the seasonal parameters defining $S_t(\mathbf{s})$. We assume that $\boldsymbol{\theta} \sim N(\mathbf{0}, C_\theta I)$ for a large value of C_θ, 10^4 say, to have a flat normal prior distribution for the regression co-efficients.

For $Q_\eta = \Sigma_\eta^{-1}$ we assume the conjugate Wishart prior distribution,

$$Q_\eta \sim W_p(2a_\eta, 2b_\eta)$$

where $2a_\eta$ is the prior degrees of freedom ($\geq p$) and b_η is a known positive definite matrix. We say that $\mathbf{X}$ has the Wishart distribution $W_p(m,R)$ if its density is proportional to

$$|R|^{m/2}|x|^{\frac{1}{2}(m-p-1)}e^{-\frac{1}{2}\mathrm{tr}(Rx)}$$

if x is a $p \times p$ positive definite matrix, see e.g. Mardia *et al.* (1979, page 85). (Here $\mathrm{tr}(A)$ is the trace of a matrix A.) To obtain diffuse but proper prior distributions we choose $a_\eta = p/2$ and following Sahu and Mardia (2005a) we take b_η to be the 0.01 times the identity matrix.

3.5 Joint posterior distribution

The above model has the following set of parameters: the error variance, σ_ϵ^2, the latent process, $\mathbf{Y}_t$, the spatial variance, σ_γ^2, the dynamic parameters, $\boldsymbol{\alpha}_t$ and their precision matrix Q_η, and the seasonal and the regression coefficients in the last two terms in (4), $\boldsymbol{\theta}$. Let $\boldsymbol{\xi}$ denote these parameters. The log of the joint posterior density of $\boldsymbol{\xi}$, denoted by $\pi(\boldsymbol{\xi}|\mathbf{z}_1,\ldots,\mathbf{z}_T)$, is given by (upto a constant of proportionality):

$$\begin{aligned}
&-\tfrac{Tn}{2}\log(\sigma_\epsilon^2) - \tfrac{1}{2\sigma_\epsilon^2}\textstyle\sum_{t=1}^T (\mathbf{x}_t - \mathbf{y}_t)'(\mathbf{x}_t - \mathbf{y}_t) - \tfrac{T}{2}\log|\Sigma_\gamma| - \tfrac{1}{2}\sum_{t=1}^T (\mathbf{y}_t - \boldsymbol{\mu}_t)'\Sigma_\gamma^{-1}(\mathbf{y}_t - \boldsymbol{\mu}_t) \\
&-\tfrac{T}{2}\log|\Sigma_\eta| - \tfrac{1}{2}\textstyle\sum_{t=1}^T (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1})'\Sigma_\eta^{-1}(\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{t-1}) + \tfrac{1}{2}(2a_\eta - p - 1)\log|Q_\eta| - \tfrac{1}{2}\mathrm{tr}(2b_\eta Q_\eta) \\
&-\tfrac{1}{b\sigma_\epsilon^2} - (a+1)\log(\sigma_\epsilon^2) - \tfrac{1}{b\sigma_\gamma^2} - (a+1)\log(\sigma_\gamma^2) - \tfrac{1}{2C_\alpha}\boldsymbol{\alpha}_0'\boldsymbol{\alpha}_0 - \tfrac{1}{2C_\theta}\boldsymbol{\theta}'\boldsymbol{\theta}.
\end{aligned}$$

All complete conditional distributions are standard and are straightforward to obtain, see Sahu and Mardia (2005a) for more details. The posterior predictive distributions are used to make Bayesian predictions in both space and time.

4 Examples

4.1 Italian Rainfall Data

We now return to modeling the Italian rainfall data described in Section 2. The computation problem here is huge since there are 23,760 monthly data points from 220 sites over 9 years. We have implemented the full BKKF model using MCMC methods and performed the usual

MCMC diagnostic analyses, the Bayesian sensitivity analyses and model choice methods for choosing and tuning various simulation parameters and the spatial smoothing parameter ϕ in the BKKF model. Those are omitted for brevity, the interested reader can see Sahu and Mardia (2005a) and Sahu *et al.* (2005) for more details.

Using a grid search we choose the value 0.05 for ϕ. The value 0.05 provides an effective range of 60 kilometers which is a reasonable choice given that the study region is roughly 300 kilometers by 100 kilometers. Lastly, we work with the state vector of dimension $p = 10$ in Equation (5). This choice is again based on predictive performance of the model for different values of p. In the paragraphs below we discuss the main results. We

- show the residual plots for assessing model adequacy,
- validate the model by performing out of sample predictions,
- present the parameter estimates,
- illustrate monthly prediction maps,
- obtain annual prediction maps.

We have checked all the residual plots (not shown) for our model for all the 220 modeling sites. In Figure 8 we provide the residuals for four randomly chosen sites out of the 220 modeling ones. Residuals do not show any cause for concern and we can state that the model seems to be adequate for the data. This discussion is continued below where we compare observed data surfaces and the fitted surfaces for a typical month later in this section.

We now return to the validation data for six sites which we have set aside. We consider validation at all the 108 time points. Figure 9 provides the validations plots. Overall few points seem to be poorly predicted. Indeed if we set an error threshold of $\pm$ 150mm for the monthly data less than 10% of the 648 predicted points are poorly estimated with a prevalence of over estimation. However, we emphasize that all the observed values are within the 95% prediction intervals. We do not show the prediction intervals because all the lower limits are zero (rainfall cannot be negative) and the upper limits are high due to large variability on the original rainfall measurement scale.

Table 2 provides the parameter estimates for our model. The regression co-efficient for elevation is seen to be positively significant as expected, see Figure 5(a). The co-efficient for longitude is negatively significant showing that western regions are wetter than the eastern sites on average. The error variance, σ^2_ϵ is roughly equal to the nugget effect, see Figure 7(b). Monthly indicators turn out to be not significant and that is why we have not included their estimates in Table 2. This, in our opinion, is an indication that the random walk part of the model is enough to account for most of the temporal variation. Furthermore, we have computed the variance of the residuals categorized by month. Subsequently we have found the ratio of the maximum and minimum variances to be 2.72 which shows that the residuals do not vary by month, i.e. the residuals are homoscedastic.

In Figure 10 three spatial surfaces[3] illustrate the model for monthly data. In Figure 10 (a) a rough linear interpolation of observed data shows small scale variability (roughness) of

[3]Surfaces are obtained using the `interp` function and the library `map` in Splus

Parameter	Mean	sd	95% interval
Elevation	0.475	0.035	(0.407, 0.542)
Longitude	–0.013	0.005	(–0.022, –0.003)
σ^2_γ	0.925	0.036	(0.855, 0.997)
σ^2_ϵ	1.389	0.024	(1.341, 1.436)

Table 2: Parameter estimates.

the data. Figure 10 (b) reports the same interpolation performed on the fitted values. As expected the fitted surface is smoother then the observed one, but it re-produces the observed rainfall field fairly well. This behavior is common to all months, even when very little rain is recorded. Figure 10 (c) shows the standard error surface for the fitted values. All values are on the original scale (mm).

From the fitted monthly values on the original scale we build annual maps of total rainfall amount simply by summation. The maps are presented in Figure 11 and 12. In panel (a) of the first figure observed values for 1976 (the wettest year) are compared to fitted values (b) for the same year. The same comparison is performed in the second figure for 1977 which happens to be the driest year. Again the fitted surfaces are smoother then the observed ones and there is general agreement between the two. There are many similarities between the two fitted surfaces for the two years. This is generally expected since rainfall patterns in space do not change rapidly, that is wet areas remain wet in the following year and so on. However, there are dis-similarities between the two fitted maps, for example the south-east corner is considerable drier in 1977 (driest year) than 1976 (wettest year). The observed data maps agree with these findings.

4.2 Venezuelan Rainfall Data

The previous example provides many detailed analyses of the BKKF model fitted to the Italian rainfall data. The objective of the second example on Venezuelan rainfall data is to compare our methods with some currently available techniques. We consider a data set analyzed by Sansó and Guenni (1999, 2000), and Stroud *et al.* (2001). The data set consists of monthly rainfall (mm) for 16 years from 80 stations in the Venezuelan state of Guarcio. See Sansó and Guenni (1999) for some preliminary analyses of the data.

We use data for first 15 years from 40 randomly chosen stations to fit the proposed model and use the observations from the remaining one year and 40 stations to validate the model. Of course, better estimation and prediction can be achieved by using all the data (for 16 years and from 80 stations) to fit the models. But here our purpose is to demonstrate the effectiveness of the methods using cross-validation in both space and time.

By considering an annual periodicity ($m = 12$), Sansó and Guenni (2000) have modeled the data so that the mean and variance for a particular month remains the same over all the years at a particular site. We do not impose this restriction in our model. We, however, adopt their time varying transformation approach. Let $z(\mathbf{s}, t)$ denote the observed rainfall

at time t in the site $\mathbf{s}$. We suppose that

$$z(\mathbf{s},t) = \begin{cases} x(\mathbf{s},t)^{\beta_t} & \text{if } x(\mathbf{s},t) > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\beta_t = \beta_l$ whenever $t = l$ modulo m. The vector of transformed variables $\mathbf{X}_t$ for $n = 40$ stations are assumed to follow the hierarchical model (1). Sansó and Guenni used a time varying variance $\sigma^2_{\epsilon,t}$ with $\sigma^2_{\epsilon,t} = \sigma^2_{\epsilon,l}$ whenever $t = l$ modulo m. However, their analysis shows that β_t and $\sigma^2_{\gamma,t}$ are highly correlated and it is not satisfactory to estimate both simultaneously. This is why we do not assume a time varying variance parameter.

The seasonal harmonics described in Section 3 are significant here since the data show strong periodicity, see Figure 13. This figure shows the fitted means and forecasts for a randomly chosen site which was used in fitting. The model based estimates generally agree with the observed data points, validating the forecasting abilities of the proposed methods. The error intervals are not super-imposed since those were not very informative, though all the observations fell within the two limits. The lower limits were mostly zeros giving a straight line for the lower interval.

The fitted means and forecast plot from the Sansó and Guenni model is shown in Figure 14. The forecasts from the Sansó and Guenni model are slightly better than those from the proposed model while the fitted means are better under the proposed model. This is expected since the proposed model is dynamic and incorporates more number of parameters than the Sansó and Guenni model. A better model fit is achieved using more number of parameters while forecasting using a smaller number of parameters is seen to be better.

The main difference between the proposed model and the Sansó and Guenni model is in estimation of spatial characteristics. Our model is expected to achieve better results in predicting in the spatial domain as it includes the important principal kriging functions. In order to see this we calculate the weighted distance criterion D^2 proposed in Sahu and Mardia (2005a) to compare the models both spatially and temporally. The distance value for the proposed model is 3547 which should be compared to a theoretical cut-off point from the χ^2 distribution with 7680 ($= 192 \times 40$) degrees of freedom. Since the observed value is quite small compared to the degrees of freedom we conclude that the model is performing a good job in prediction. The weighted distance for the model proposed by Sansó and Guenni is estimated to be 4996, a value much higher than the distance under the proposed model. Thus the new model performs much better prediction at new locations. This is expected since the proposed model uses the optimal ('kriging') spatial prediction surface while the earlier model does not.

5 Discussion

In this article we have proposed the BKKF model for analyzing rainfall data. Our approach solves in a unified framework the problems of: accounting for discreteness in the data, spatial variation, temporal variation and joint space-time non-separable variation. It easily allows us to incorporate different type of seasonal effects which depend on climatic and morphological conditions typical of each region. The model has been fully implemented through MCMC on two different data sets, and have shown advantages over some current methods. The

MCMC implementation of the model allows us to predict in space and forecast in time using Bayesian methods.

Acknowledgments We would like to thank Dr. Attilio Colagrossi of the *Servizio Raccolta e Gestione Dati* (APAT) for kindly giving us the data and for the many useful discussions.

REFERENCES

Allcroft, D. J. and Glasbey, C. A. (2003). A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Journal of the Royal Statistical Society*, Series C, *Applied Statistics*, **52**, 487–498.

Brown, P. E., Diggle, P. J., Lord, M. E. and Young, P. C. (2001) Space-time calibration of radar rainfall data. *Applied Statistics*, **50**, 221–241.

Cassiraga, E. F., Guardiola-Albert, C. and Gomez-Hernandez, J. J. (2004) Automatic modeling of cross-covariances for rainfall estimation using reengauge and radar data. GEOENV IV - Geostatistics for environmental applications: Proceedings quantitative geology and geostatistics 13, pp 391–399, Kluwer Academic Publishing, Dordrecht.

Dunn, P. (2003). Precipitation occurrence and amount can be modelled simultaneously. Working Paper, Series SC-MC-0305, Faculty of Sciences, USQ.

Kent, J. T. K. and Mardia, K. V. (2002) Modelling Strategies for Spatial-Temporal Data. In *Spatial Cluster Modelling*. (Eds A. Lawson and D. Denison), London: Chapman and Hall, pp 214–226.

Kyriakidis, P. C. and Journel, A. G. (1999) Geostatistical space-time models: A review. *Mathematical Geology*, **31**, 651–684.

Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979) *Multivariate Analysis*. London: Academic Press.

Mardia K.V., Goodall C., Redfern E.J., and Alonso F.J. (1998) The Kriged Kalman filter (with discussion). *Test*, **7**, 217–252.

Orasi, A., Jona Lasinio, G. and Ferrari, C. (2005) Comparison of calibration methods for the reconstruction of space-time rainfall fields in Southern Italy. Submitted.
Available from http://w3.uniroma1.it/dspsa/Rapporti_tecnici/orasijonaferrari.pdf

Raspa, G., M. Tucci, and R. Bruno (1997). Reconstruction of rainfall fields by combining ground raingauges data with radar maps using external drift method (Kluwer Academic Publishers ed.), Volume 2 of Geostatistics Wollongong 96, pp. 1306–1315. E.Y. Baafi and N.A. Schofield eds.

Rodriguez-Iturbe, I., D. Cox, and V. Isham (1987). Some models for rainfall based on stochastic point processes. Proceedings of the Royal Society of London Series A **410**, 269–288.

Rodriguez-Iturbe, I., D. Cox, and V. Isham (1988). A point process model for rainfall: further developments. Proceedings of the Royal Society of London, Series A **417**, 283–298.

Sahu, K.S., Jona Lasinio G., Orasi, A., and Mardia, K.V. (2005) A comparison of spatio-temporal Bayesian models for reconstruction of rainfall fields in a cloud seeding experiment. Submitted

Sahu, S. K. and Mardia, K. V. (2005a) A Bayesian Kriged-Kalman model for short-term forecasting of air pollution levels. *Journal of the Royal Statistical Society*, Series C, *Applied Statistics*, **54**, 223–244.

Sahu, S. K. and Mardia, K. V. (2005b) Recent Trends in Modeling Spatio-Temporal Data. Technical report, University of Southampton.

Sansó, B. and Guenni, L. (1999) Venezuelan rainfall data analysed by using a Bayesian space-time model. *Applied Statistics*, **48**, 345–362.

Sansó, B. and Guenni, L. (2000) A nonstationary multisite model for rainfall. *Journal of the American Statistical Association*, **95**, 1089–1100.

Smith, R. (1994) Spatial modelling of rainfall data. In *Statistics for the environment 2: Water Related Issues*, eds. V. Barnett, and K. Feridum Turkman, Wiley: New York, pp. 19–42.

Stern, R. and R. Coe (1984) A model fitting analysis of rainfall data. *Journal of the Royal Statistical Society*, Series A 147, 1–34.

Stroud, J. R., Müller, P. and Sansó, B. (2001) Dynamic models for Spatio-temporal data. *Journal of the Royal Statistical Society*, B, **63**, 673–689.

West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. New York: Springer.

Wikle, C. K. and Cressie, N. (1999) A dimension-reduced approach to space-time Kalman filtering. *Biometrika*, **86**, 815–829.

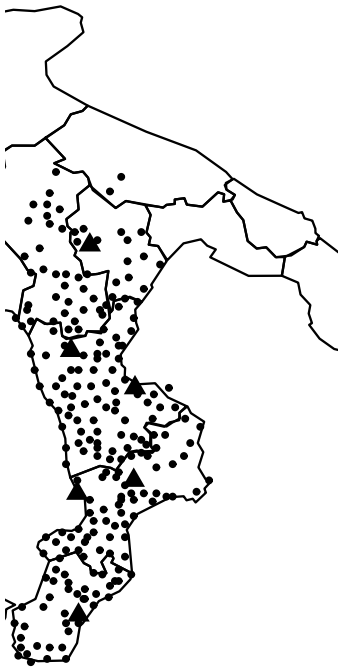


Figure 1: The Catanzaro raingauges network in south Italy. Dots are modeling sites and solid triangles are validation locations.

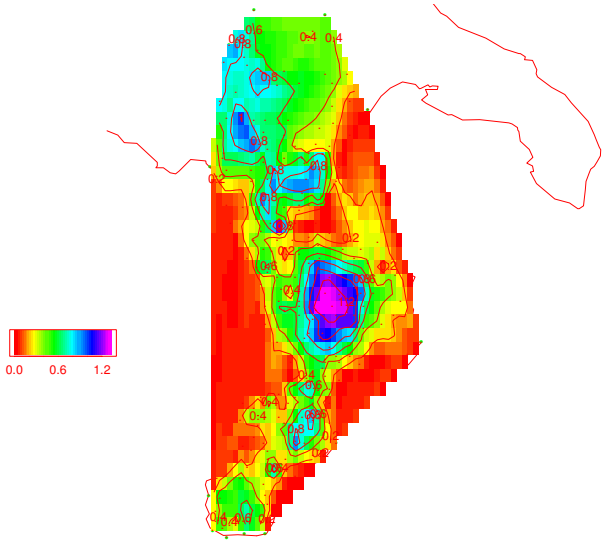
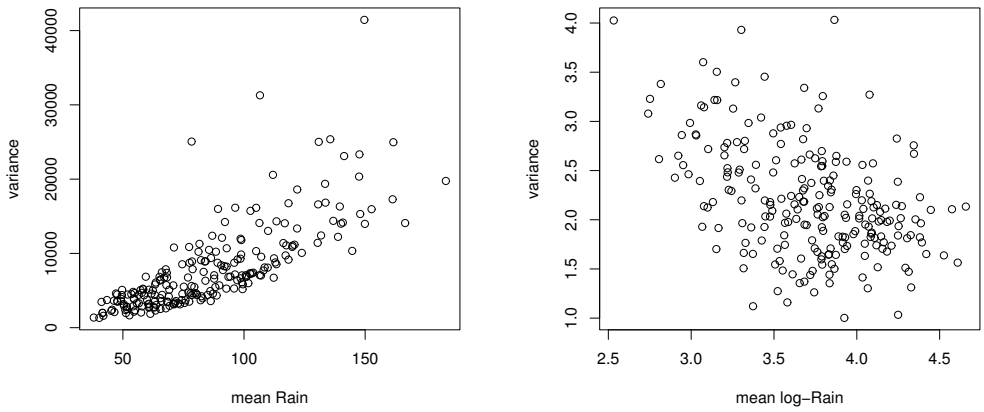


Figure 2: Elevation surface (km).



(a) original scale (b) log-scale

Figure 3: Mean versus variance plots on: (a) original scale, and (b) log-scale.

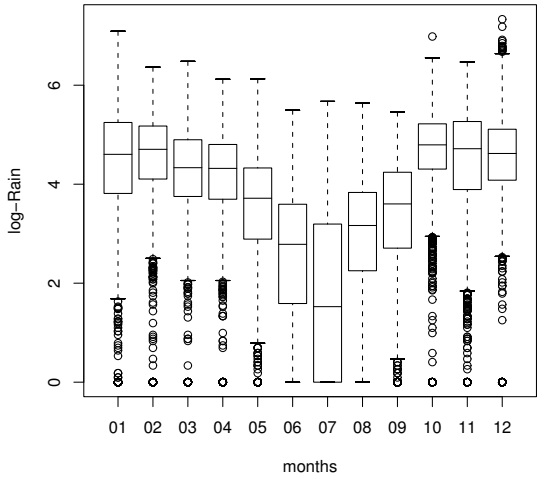
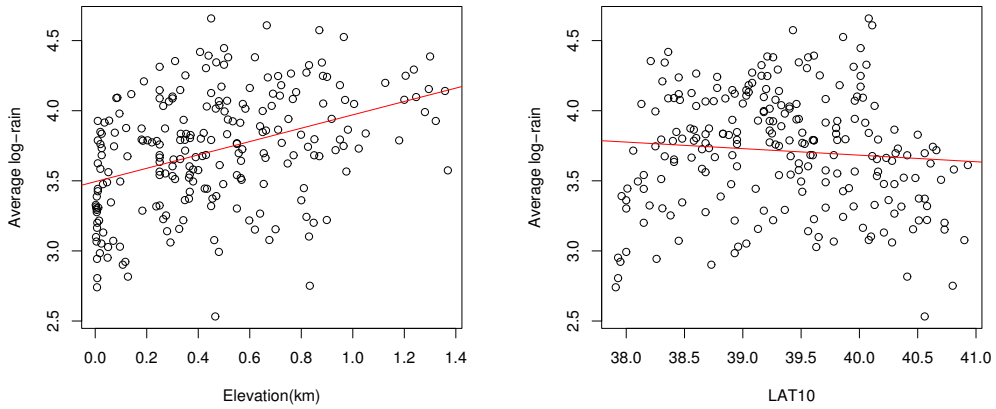
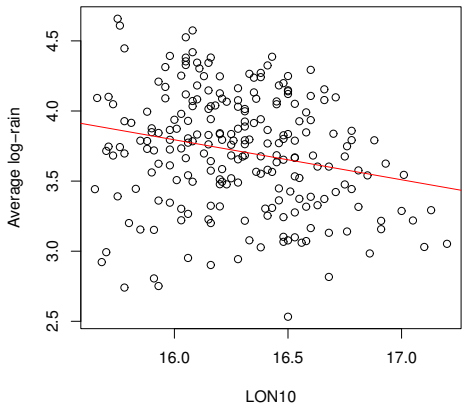


Figure 4: Boxplot of log-rain measurements against months.



(a)

(b)



(c)

Figure 5: Plots of log-rainfall averages of 226 sites and regression lines against (a) Elevation above sea level, (b) Latitude, (c) Longitude.

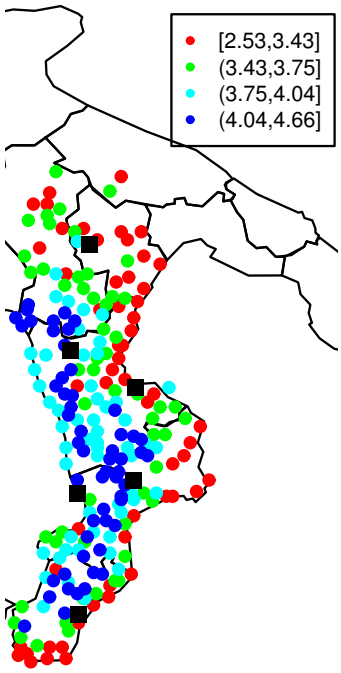


Figure 6: Map of log-rainfall site averages (1972-1980) classified according to their quantiles. Black squares denote validation sites.

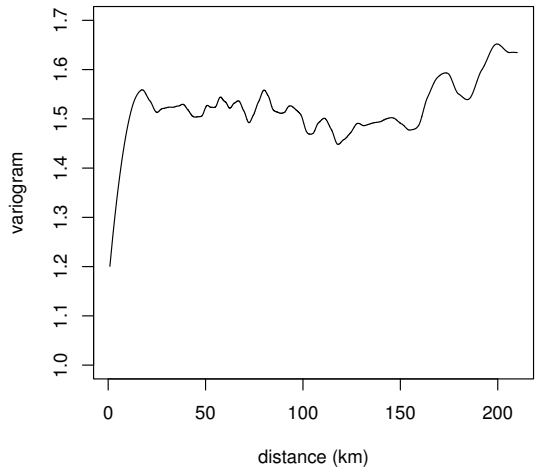


Figure 7: A smoothed variogram of residuals.

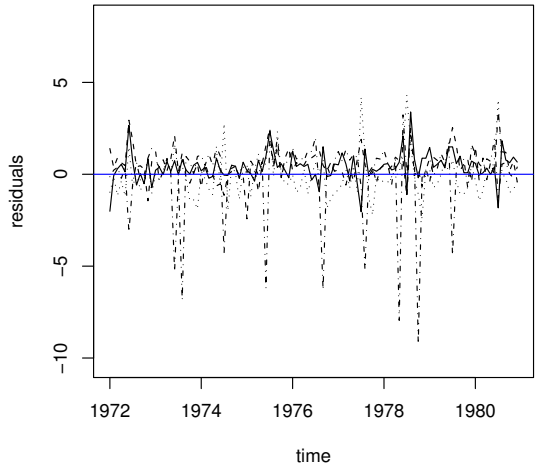


Figure 8: Time series plot of residuals for 4 randomly chosen sites, out of the 220 modeling ones.

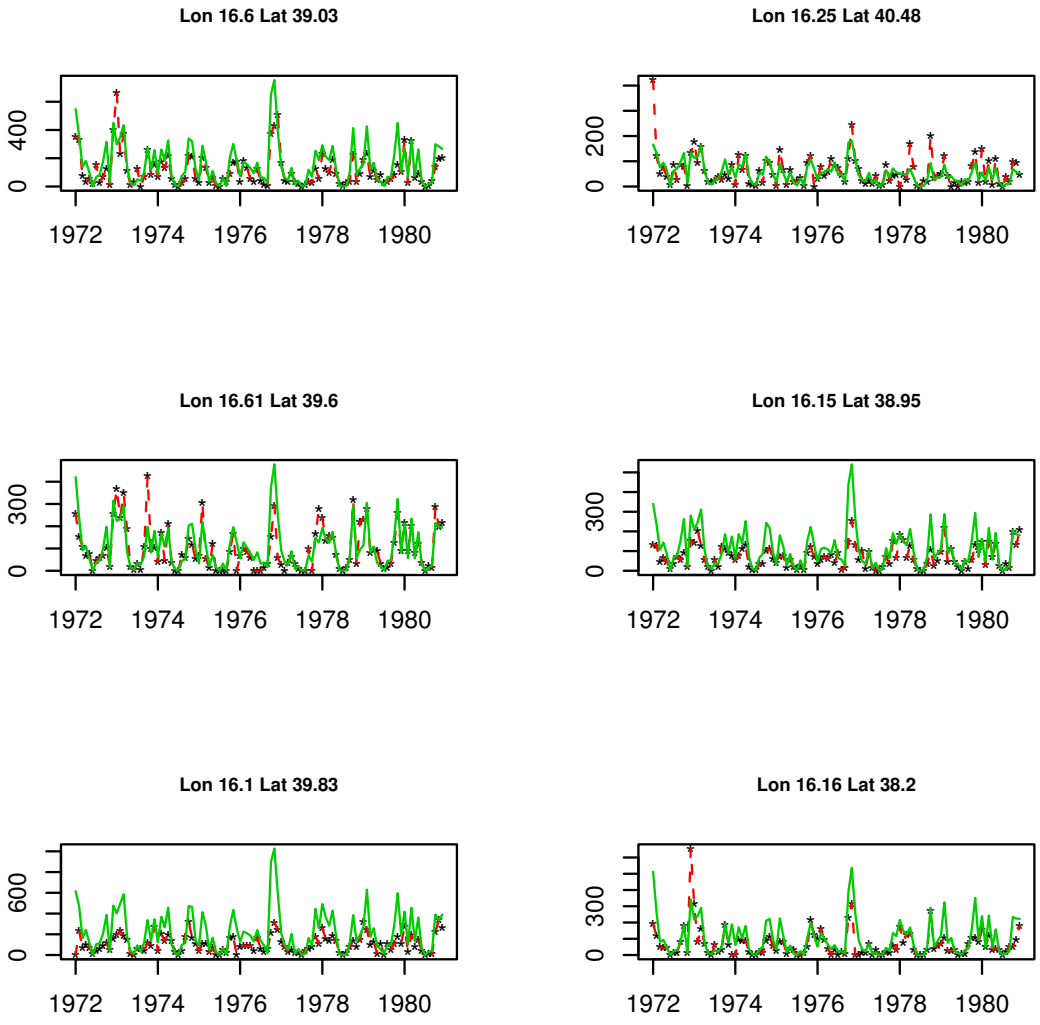
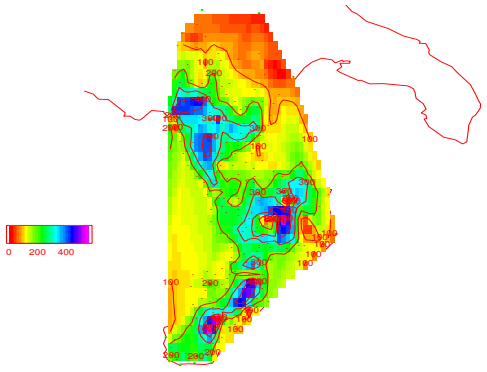
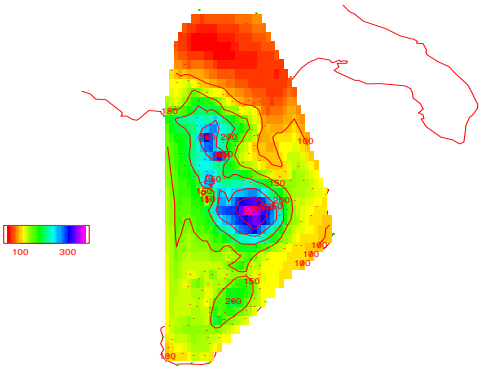


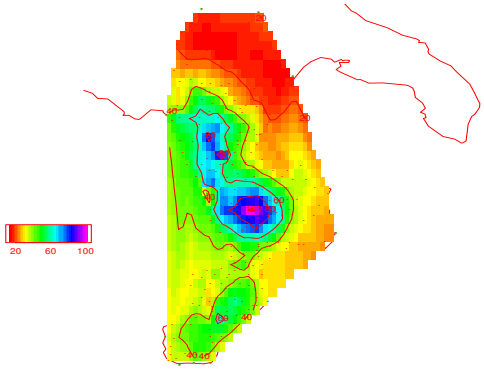
Figure 9: The observed and predicted values at 6 validation sites. Observed values are shown in red (points and dotted lines) and predicted values are shown in green (solid lines).



(a)

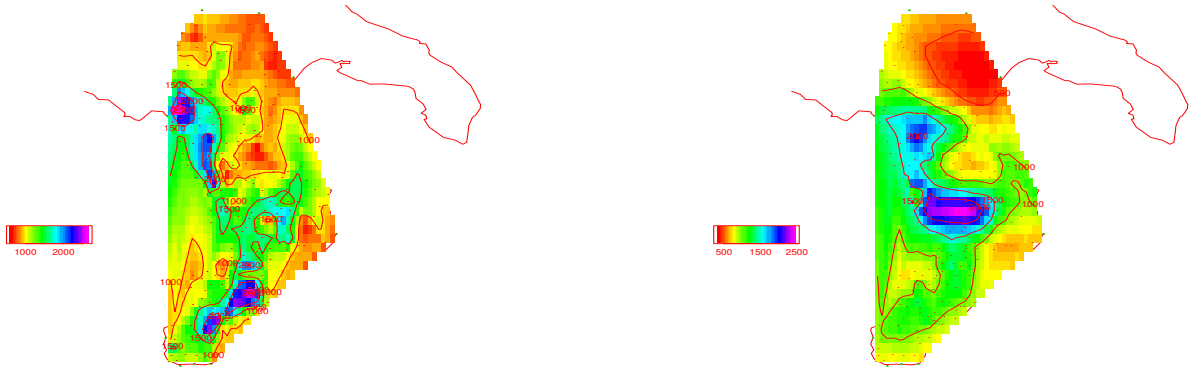


(b)



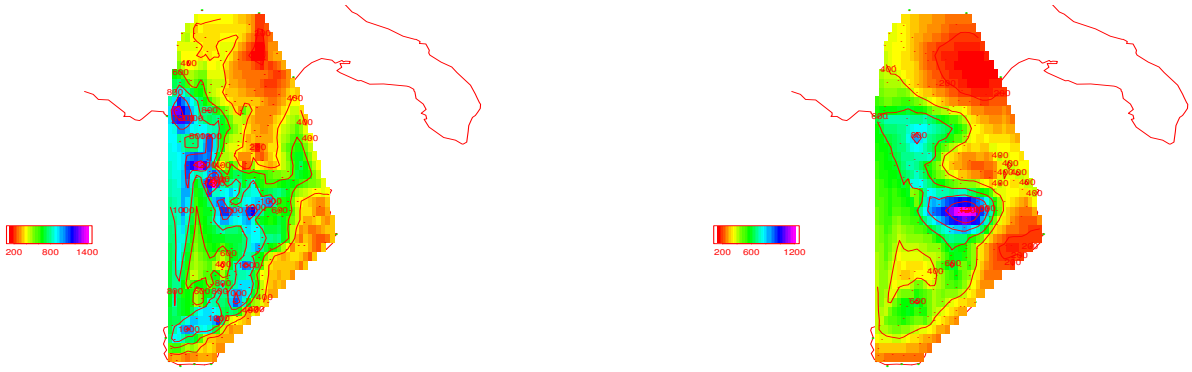
(c)

Figure 10: Fitted maps for December 1976.(a) Observed values (mm), (b) fitted values (mm) and (c) standard errors.



(a) (b)

Figure 11: The fitted annual maps in 1976: the wettest year. (a) Observed total rainfall amount (mm), (b) fitted values (mm).



(a) (b)

Figure 12: The fitted annual maps in 1977: the driest year. (a) Observed total rainfall amount (mm), (b) fitted values (mm).

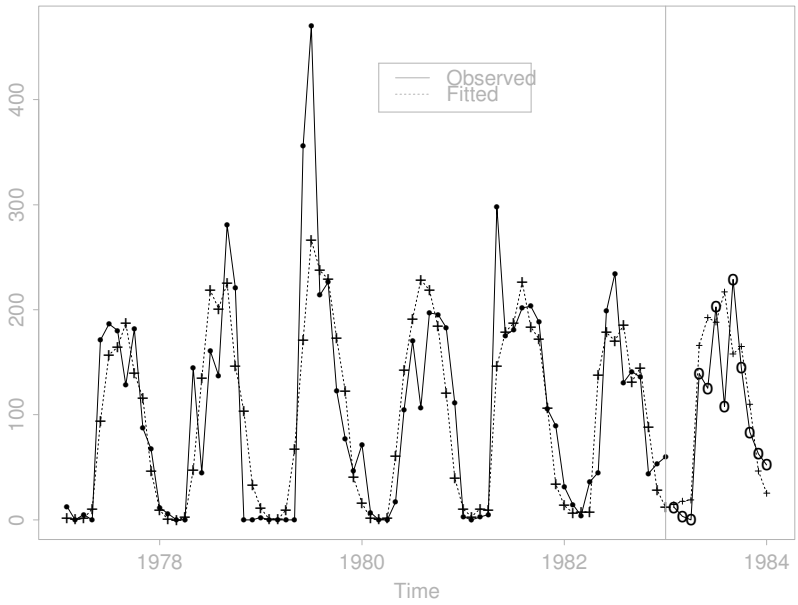


Figure 13: The fitted means and forecasts for a randomly chosen site using the proposed method.

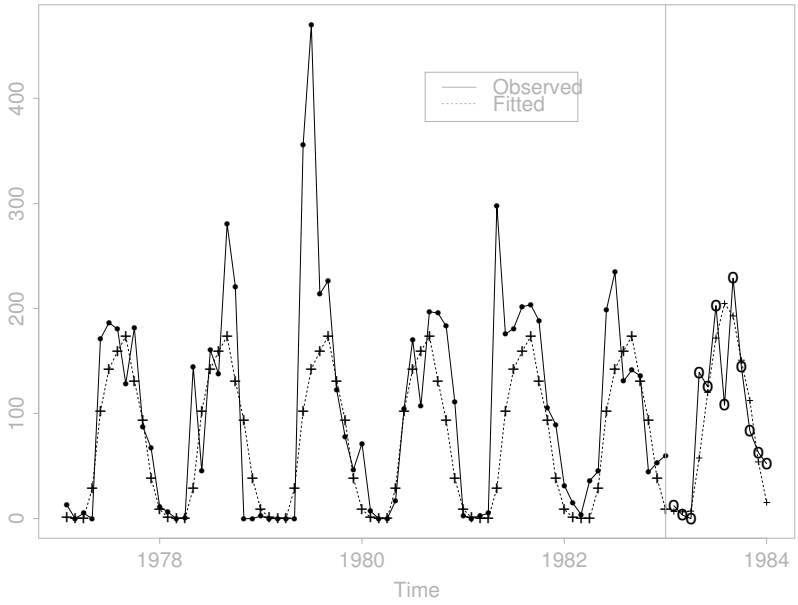


Figure 14: The fitted means and forecasts using the Sansó and Guenni method for the same site as in Figure 13.