

## A Comparison of Spatio-Temporal Bayesian Models for Reconstruction of Rainfall Fields in a Cloud Seeding Experiment

<sup>1</sup>Sujit K. Sahu, <sup>2</sup>Giovanna Jona Lasinio, <sup>3</sup>Arianna Orasi and <sup>4</sup>Kanti V. Mardia

<sup>1</sup>School of Mathematics, University of Southampton, Highfield, SO17 1BJ, UK.

<sup>2</sup>DSPSA, University of Rome "La Sapienza", P.le Aldo Moro 5, 00185 Rome Italy

<sup>3</sup>Italian Environmental Protection Agency (APAT), Via Curtatone 3, 00185 Rome, Italy

<sup>4</sup>University of Leeds, Leeds, West Yorkshire, LS2 9JT, UK

---

**Abstract:** In response to the drought experienced in Southern Italy a rain seeding project has been setup and developed during the years 1989-1994. The initiative was taken with the purpose of applying existing methods of rain enhancement technology to regions of south Italy including Puglia. The aim of this study is to provide statistical support for the evaluation of the experimental part of the project. In particular our aim is to reconstruct rainfall fields by combining two data sources: rainfall intensity as measured by ground raingauges and radar reflectivity. A difficulty in modeling the rainfall data here comes from rounding of many recorded raingauges. The rounding of the rainfall measurements make the data essentially discrete and models based on continuous distributions are not suitable for modeling these discrete data. In this study we extend two recently developed spatio-temporal models for continuous data to accommodate rounded rainfall measurements taking discrete values with positive probabilities. We use MCMC methods to implement the models and obtain forecasts in space and time together with their standard errors. We compare the two models using predictive Bayesian methods. The benefits of our modeling extensions are seen in accurate predictions of dry periods with no positive prediction standard errors.

**Key words:** Bayesian inference, Gibbs Sampler; Kalman Filter; Kriging; Markov chain Monte Carlo; rainfall modeling, cloud seeding operation, Spatial Temporal Modeling; State-Space Model.

---

### INTRODUCTION

In this article our aim is to model a dataset on rain enhancing experiment through seeding operations conducted during the years 1989-1994 in the dry regions of south Italy starting with Puglia. The primary purpose is to reconstruct rainfall fields using spatio-temporal data obtained from a network of ground raingauges and data from a C-band digital weather radar. Radar raingauges are increasingly used to reconstruct rainfall fields since they are able to provide spatially continuous images of precipitation for short and regular time intervals; ground raingauges, on the other hand, provide more accurate and direct estimates of rainfall intensity.

Statistical spatio-temporal models are appropriate for our purposes due to the presence of spatio-temporal correlations in rainfall and radar data. Spatio-temporal modeling of rainfall data has received considerable attention in recent literature<sup>[1-4]</sup>. Many authors have also considered modeling the relationships between the radar reflectance and rainfall intensity. For example, Brown *et al.*<sup>[5]</sup> use multivariate time series models with state space representation incorporating continuous radar readings as covariates. Cassiraga *et al.*<sup>[6]</sup> model cross-correlation between radar and rainfall data using experimental surface variogram. Cornford<sup>[7]</sup> uses

probabilistic models which treat the radar readings as noisy realizations of the underlying true precipitation process and builds a forecasting model for short-term predictions. Jordan *et al.*<sup>[8]</sup> develop a stochastic space-time model for rainfall using the variations in the reflectivity-rainfall intensity (Z-R) relationships.

The Puglia region of South Italy is known to be very dry and the total amount of annual rainfall is usually small. As a result typically there is a huge number of zero rainfalls in any rainfall data set for this region. Moreover, after a rain seeding operation many rainfall amounts recorded by the available raingauges were rounded to the nearest 10th of a millimeter giving rise to essentially discrete data.

The occurrences of many discrete amount of rainfall in the data exclude the use of many currently available methods and models cited above since those are essentially developed to model continuous rainfall measurements. However, some authors<sup>[1-3]</sup> have developed methods for handling zero rainfalls using censoring mechanisms. Here the problem is to extend the model to accommodate more than one discrete rainfall value occurring with non-zero probability.

Another objective of this study is to develop methods for relating radar reflectance and rainfall intensity to reconstruct rainfall fields in the presence of the discrete rainfall amounts. In this study we do this by

explicitly regressing rainfall data on the radar measurements in a spatio-temporal model. We believe that this method is novel for data obtained from a rain seeding experiment conducted in a very dry region such as southern Italy.

We consider two recently developed hierarchical Bayesian modeling approaches. The first approach is a separable and stationary Gaussian spatio-temporal model developed by Sahu *et al.*<sup>[9]</sup> for monitoring some air pollution levels. The second approach is a hierarchical space-time Bayesian Kriged-Kalman filtering (BKKF) model. The spatial prediction surface of the BKKF model is built using the well known method of kriging for optimum spatial prediction and the temporal effects are analyzed using the models underlying the Kalman filtering method.

We extend both the models to accommodate rounded rainfall measurements taking discrete values with positive probabilities. The full Bayesian models are implemented using MCMC techniques which enable us to obtain the optimal Bayesian forecasts in time and space. We compare the two modeling approaches using the mean-square error of predictions and some formal Bayesian model selection criteria.

**The dataset:** Our data come from the rain enhancement project carried on in the South of Italy (Fig. 1) during the period 1989-1994. This is a very dry region and the total amount of annual rainfall is usually very small (approximately 80 millimeter on the average per year during the study period). We consider a rainfall seeding operation conducted at 5:00AM on April 11, 1992 when 44 out of total 80 ground raingauges recorded amount of rainfall in 10 minutes interval; in addition data from a C-band digital weather radar, scanning the whole area every five minutes, are available.



Fig. 1: Operational raingauges in south Italy

In this study we consider the data recorded every ten minutes from 5:10AM in the morning until 9:30AM. A subsequent seeding operation was performed at 9:30AM and we do not include the data recorded after 9:30AM on that day since our aim is to devise methodology for evaluating a single seeding

operation. Data were available for many other seeding operations performed on other days. However, our previous investigation<sup>[10,11]</sup> in modeling those data have found a number of insurmountable problems in modeling the full data set, for example, (a) there was a large number of missing values due to many malfunctioning automatic raingauges (in some cases there were only 10 raingauges working properly); (b) there were extreme variability in meteorological conditions affecting the amounts of rainfall during different seeding operations.

On April 11, 1992 there were  $N=44$  working raingauges in the study region. Let  $s_i, i=1, \dots, N$  denote the UTM  $x$  and  $y$ -coordinate of the locations. Out of these 44 sites we choose to set aside data from six sites for validation purposes. The validation sites were chosen judiciously so that those covered the entire study region. Thus we model data from the remaining 38 sites which will be denoted by  $s_1, \dots, s_n$  where  $n=38$ . The 44 locations together with a predictive grid of 2710 locations are shown in Fig. 2. We aim to perform spatial predictions in the grid.

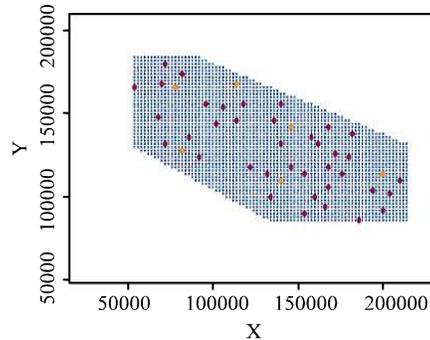


Fig. 2: A predictive grid of 2710 locations together with the 38 modeling sites (raingauges) in red color and the 6 validation sites (raingauges) in orange color

Each site had temporal raingauge data observed at every 10 minutes and there are  $T=27$  observations at each site covering the time period from 10 minutes past 5 AM to 30 minutes past 9 AM. There are no missing data and the 27 observations at each site are equally spaced in time. Let  $z(s_i, t)$  denote the observed log amount of rainfall (in millimeter) at site  $s_i$  and at time  $t$ , where  $i=1, \dots, n$  and  $t=1, \dots, T$ . Thus we have 1026 ( $=38 \times 27$ ) log rainfall measurements for modeling purposes. We denote the validation data by  $z(s_i, t)$ ,  $i=n+1, \dots, N$ ,  $t=1, \dots, T$  where  $N=44$  and  $T=27$ . Thus we have 162 ( $6 \times 27$ ) validation data points.

Together with the rainfall we have data from a C-band digital weather radar, scanning the whole area every five minutes. We shall model log rainfall data using log of the radar measurements as a covariate because there exist a well known linear relationship

between them. Let  $r(s_i, t)$  denote the log of the radar measurements at site  $s_i$  and at time  $t$ . Radar reflectivity is expressed in units of dBZ. This is a measure of the power scattered back to the radar by precipitation particles in the atmosphere. The power is a function of the distribution of raindrops size given by  $Z(dBZ) = 10 \log_{10} \left( \sum D^6 \right)$  where the summation of the drop diameters ( $D$ ) takes place over the volume of space sampled by the radar.

**Exploratory analysis:** Although the amount of rainfall variable is a continuous random variable, the readings from the rain gauges were rounded to the nearest 10th of one millimeter. Moreover, the zero rainfalls were already replaced by 0.02 for obvious benefits in working with the log-scale. Table 1 provides the frequencies of the amount of rainfall. Due to this discreteness in the observations we shall use a censoring mechanism when modeling these data as continuous observations.

Table 1: Frequency table of rainfall measurements in millimeter

Amount of rainfall	Frequency
0.02	90
0.1	136
0.2	117
0.3	174
0.4	136
0.5	54
0.6	78
0.7	54
0.8	42
0.9	37
1	42
>1	66
Total	1026

There exist a well known linear relationship between logarithm of radar data and the logarithm of the actual amount of rainfall known as the Marshal and Palmer law<sup>[12]</sup>. We visualize the relationship in Fig. 3 where we have plotted the mean log rainfall for each distinct value of log-radar values. An approximate, though rather weak, linear relationship is seen in the graph.

In Fig. 4 we provide the mean amount of rainfall (over 27 measurements) in each of the 44 sites. There is no evidence of spatial trend for the site means. In fact we shall illustrate with the exponential covariance function for simplicity.

The site means show evidence of spatial variation. We investigate this using an empirical variogram of the data. We first obtain the residuals after fitting a regression line with  $r(s_i, t)$  as one covariate. We also remove any temporal variation and trend present in the residuals by explicit modeling or by creating successive differences. Let  $W(s_i, t)$  denote the residuals. We suppose that  $W(s_i, t), t=1, \dots, T$  are independent

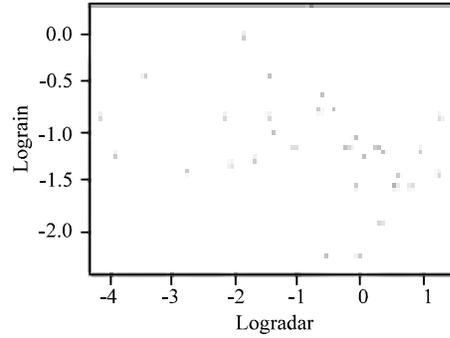


Fig. 3: Average log rain versus log radar at the 38 modeling sites

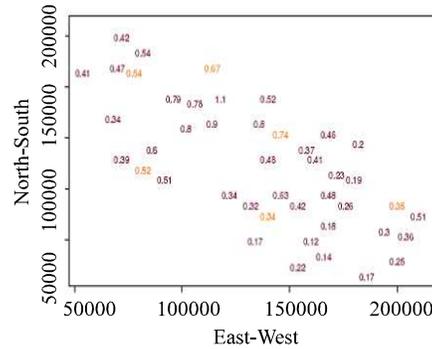


Fig. 4: Sites means at the 38 modeling sites (red) and 6 validation sites (orange)

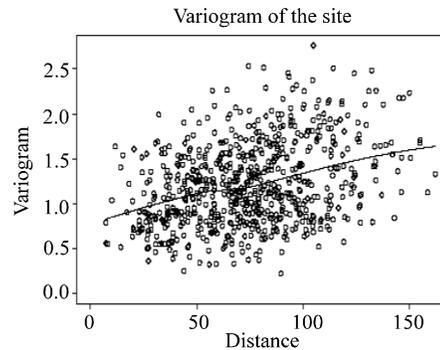


Fig. 5: Variogram cloud and a smoothed variogram of the residuals for the 38 modeling locations using the 27 time points as replications

replications at location  $s_i, i=1, \dots, n$  since we have detrended the data. We now consider the average variogram defined by

$$\gamma(d_{ij}) = (2T)^{-1} \sum_t E \{ [W(s_i, t) - W(s_j, t)]^2 \}$$

where  $d_{ij}$  is the distance between the spatial locations  $s_i$  and  $s_j$ . The quantity  $\gamma(d_{ij})$  is estimated by

$$\hat{\gamma}(d_{ij}) = (2T)^{-1} \sum_t [w(s_i, t) - w(s_j, t)]^2.$$

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 Fig. 5 we provide the variogram cloud and we super-impose a smooth loess curve (as obtained using the S-Plus function loess). This plot justifies our choice of the exponential spatial covariance function.

All the 44 sites have been classified into two areas called the target and control. The classification comes from the experimental design applied in the rain seeding project. The seeding operation is carried out in the target area, however, rainfall is measured both in the target area and the control area. For the April 11 experiment the target area was Bari and the control area was Canosa, marked as C in Fig. 1.

In order to investigate the temporal variation in the data set we show the time series plots of data for each of the 38 sites in Fig. 6. The Figure does not show a large amount of temporal variation. Moreover, there is not much difference between the time series plots of the sites in target and control areas; we have performed significance testing using linear models to confirm this conclusion. Thus our modeling approaches earlier will not differentiate between the sites from target and control areas.

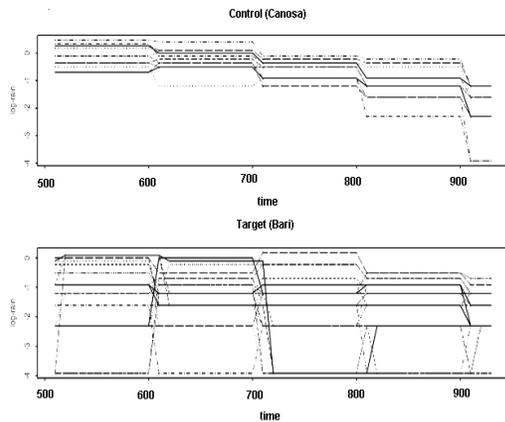


Fig. 6: Time series for the 38 modeling locations (log-rain)

**Spatio-temporal models**

**Latent variables to model discrete data:** As seen in Table 1 the amount of rainfall had been rounded to the nearest 10th of a millimeter. These discrete values occur with non-zero probabilities, but the actual rainfall is a continuous measurement falling between two discrete endpoints. This is a very common problem in modeling rainfall data with zero rainfall<sup>[1,2]</sup>. A common approach is to model the zeros by the values of a latent continuous variable below a threshold value (censoring).

The problem of discreteness is more severe for the current data as there are many rounded discrete values occurring with non-zero probabilities. Thus we extend the censoring mechanism for latent variable to include multiple discrete values. Let  $X(s,t)$  denote a continuous latent variable and let there be  $k$  particular values of log rainfall  $\lambda_1, \lambda_2, \dots, \lambda_k$  which may occur with positive

probabilities. Let  $c_1, \dots, c_k$  be constants such that  $\lambda_i < c_i, i=1,\dots,k$ . We suppose that the observed data point at a site  $s$  at time  $t$  is given by

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

In our data set we have  $k=11$  and the values of  $\lambda_1, \dots, \lambda_k$  are the logarithm of the values in the first column of Table 1, i.e.  $\lambda_1 = \log(0.02), \lambda_2 = \log(0.1), \dots, \lambda_{11} = \log(1)$ . We choose the constants  $c_1, \dots, c_k$  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. We suppose that the latent random variable  $X(s,t)$  for any observed rainfall bigger than 1 millimeter on the original scale is the actual log amount of rainfall. Henceforth, we model the latent variables  $X(s,t)$  rather than the observations  $Z(s,t)$  some of which have been rounded.

**A Gaussian spatio-temporal random effect model:**

We first assume the following hierarchical model:

$$X(s_i,t) = \mu(s_i,t) + v(s_i,t) + \varepsilon(s_i,t), i=1,\dots,n, t=1,\dots,T, (1)$$

where  $\mu(s_i,t)$  is given below,  $v(s_i,t)$  is an independent zero mean spatio-temporal process and the error term  $\varepsilon(s,t)$  is a white noise process assumed to follow  $N(0, \sigma_\varepsilon^2)$  independently. The function  $\mu(s_i,t)$  is given by

$$\mu(s_i,t) = \beta_0 + \beta_1 r(s_i,t). (2)$$

As mentioned before, for  $v(s_i,t)$ , we adopt a separable covariance structure<sup>[13]</sup>. That is,

$$Cov\{v(s_i,t), v(s_j,t')\} = \sigma_v^2 \rho_s(s_i - s_j, \phi_s) \rho_t(t - t', \phi_t). (3)$$

In addition, the  $\rho$ 's are taken to be exponential correlation functions, i.e.,  $\rho(d, \phi) = \exp(-\phi \cdot d)$ , as we have decided to assume previously. We define  $\Sigma_s$  and  $\Sigma_t$  to be square matrices of order  $n$  and  $T$ , respectively with elements  $\rho_s(s_i - s_j, \phi_s)$  and  $\rho_t(t - t', \phi_t)$ .

The prior distributions for  $\sigma_\varepsilon^2$  and  $\sigma_v^2$  are assumed to be the inverse gamma distribution with parameters  $a$  and  $b$ ,  $IG(a,b)$  with mean  $b/(a-1)$ . We take  $a=2$  and  $b=1$  to have a proper but diffuse prior distribution with mean 1 and infinite variance. The regression parameters  $\beta_0$  and  $\beta_1$  are all given normal prior distributions with mean 0 and variance  $10^4$ .

**A Bayesian Kriged-Kalman model:** We follow Sahu and Mardia<sup>[9]</sup> to construct a BKKF model for the

rainfall data. In so doing we extend their approach to account for discreteness in the data and also add the radar data as a regressor. Let  $\mathbf{X}_t=(X(s_1,t), \dots, X(s_n,t))'$  denote the  $n$ -dimensional latent random vector at time  $t$ ;  $t=1, \dots, T$ . The first modeling assumption is the hierarchical model:

$$\mathbf{X}_t = \mathbf{Y}_t + \boldsymbol{\varepsilon}_t \tag{4}$$

where  $\mathbf{Y}_t=(Y(s_1,t), \dots, Y(s_n,t))'$  is an unobserved but scientifically meaningful process (signal) and  $\boldsymbol{\varepsilon}_t$  is a white noise process. Thus we assume that the components of  $\boldsymbol{\varepsilon}_t$  are i.i.d. normal random variables with mean zero and unknown variance  $\boldsymbol{\sigma}_\varepsilon^2$ .

The space-time process  $\mathbf{Y}_t$  is given by

$$\mathbf{Y}_t = H\boldsymbol{\alpha}_t + \boldsymbol{\beta}_1 \mathbf{r}_t + \boldsymbol{\gamma}_t \tag{5}$$

where the matrix  $H$  of order  $n \times p$  is defined below,  $\boldsymbol{\alpha}_t$  is the state vector of dimension  $p$ ,  $\mathbf{r}_t = (r(s_1,t), \dots, r(s_n,t))'$  and the error term  $\boldsymbol{\gamma}_t$  is assumed to be zero mean Gaussian with covariance matrix  $\boldsymbol{\Sigma}_\gamma$  which has elements

$$\sigma(\mathbf{s}_i, \mathbf{s}_j) = Cov(\boldsymbol{\gamma}(\mathbf{s}_i, t), \boldsymbol{\gamma}(\mathbf{s}_j, t)) \tag{6}$$

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

The prior distributions for  $\boldsymbol{\sigma}_\varepsilon^2$  and  $\boldsymbol{\sigma}_\gamma^2$  are assumed to be the inverse gamma distribution with parameters  $a$  and  $b$ ,  $IG(a, b)$ . As previously we choose  $a=2$  and  $b=1$  to have a proper but diffuse prior distribution with mean 1 and infinite variance. The regression parameter  $\boldsymbol{\beta}_1$  is given the flat normal prior distributions with mean 0 and variance  $10^4$ .

The matrix  $H$  is obtained by using what are known as principal kriging functions<sup>[9,13]</sup>. In this implementation we take the first column of  $H$  to be the unit vector,  $\mathbf{1}$ . The other columns are obtained as follows: We first obtain

$$\mathbf{B} = \boldsymbol{\Sigma}_\gamma^{-1} - \frac{1}{\mathbf{1}' \boldsymbol{\Sigma}_\gamma^{-1} \mathbf{1}} \boldsymbol{\Sigma}_\gamma^{-1} \mathbf{1} \mathbf{1}' \boldsymbol{\Sigma}_\gamma^{-1}.$$

We now perform the spectral decomposition of  $\mathbf{B}$ ,

$$\mathbf{B} = \mathbf{U} \mathbf{E} \mathbf{U}' , \mathbf{B} \mathbf{u}_i = e_i \mathbf{u}_i,$$

where  $\mathbf{U}=(\mathbf{u}_1, \dots, \mathbf{u}_n)$  and  $\mathbf{E} = \text{diag}(e_1, \dots, e_n)$  and we assume without loss of generality that the eigenvalues are in non-decreasing order,  $e_1 = 0 < e_{2+1} \leq \dots \leq e_n$ . Finally, the matrix  $H$  is taken as

$$H = (1, e_2 \boldsymbol{\Sigma}_\gamma \mathbf{u}_2, \dots, e_p \boldsymbol{\Sigma}_\gamma \mathbf{u}_p). \tag{7}$$

We assume that

$$\boldsymbol{\alpha}_t = \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \tag{8}$$

and  $\boldsymbol{\eta}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ . To complete the modeling

hierarchies we suppose that  $\boldsymbol{\alpha}_0 \sim N(\mathbf{0}, C_a \mathbf{I})$  and with a large value of  $C_a$  where  $\mathbf{I}$  is the identity matrix. For  $Q_\eta = \boldsymbol{\Sigma}_\eta^{-1}$  we suppose that it has the Wishart prior distribution,

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

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

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

if  $x$  is a  $p \times p$  positive definite matrix<sup>[14]</sup>. (Here  $\text{tr}(A)$  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<sup>[9]</sup> we take  $b_\eta$  to be the 0.01 times the identity matrix.

**Strategies for model choice:** Many graphical diagnostic methods are used to perform diagnostic checking and model validation<sup>[13]</sup>. Several validation statistics are also available<sup>[15]</sup>. In this article we shall use the following methods for model choice and validation.

**Model choice:** To compare between two different models we shall use the following criterion based on Gelfand and Ghosh<sup>[16]</sup> and Laud and Ibrahim<sup>[17]</sup>.

$$PMCC = \sum_{i=1}^n \sum_{t=1}^T \left[ \text{Var}\{Z(\mathbf{s}_i, t)_{\text{rep}}\} + \left\{ Z(\mathbf{s}_i, t)_{\text{obs}} - E\{Z(\mathbf{s}_i, t)_{\text{rep}}\} \right\}^2 \right], \tag{9}$$

where  $Z(\mathbf{s}_i, t)_{\text{rep}}$  is a future observation corresponding to  $Z(\mathbf{s}_i, t)_{\text{obs}}$  under the assumed model. The first term in  $PMCC$  is a penalty term for prediction and the second is a goodness-of-fit term (GOF).

**Validation:** Recall that we have set aside the observations  $z(\mathbf{s}_i, t)$ ,  $i=n+1, \dots, N$ ,  $t=1, \dots, T$  for validation purposes. Let  $z_{\text{orig}}(\mathbf{s}_i, t)$  and  $\hat{z}_{\text{orig}}(\mathbf{s}_i, t)$  denote respectively the predicted and observed amount of rainfall on the original scale corresponding to  $z(\mathbf{s}_i, t)$  for each  $i=n+1, \dots, N$  and  $t=1, \dots, T$ , i.e.  $z_{\text{orig}}(\mathbf{s}_i, t) = \exp\{z(\mathbf{s}_i, t)\}$ . A simple measure of validation is the  $MSE$  given by

$$MSE = \frac{1}{(N-n)T} \sum_{i=n+1}^N \sum_{t=1}^T \{z_{\text{orig}}(\mathbf{s}_i, t) - \hat{z}_{\text{orig}}(\mathbf{s}_i, t)\}^2. \tag{10}$$

The above validation measure does not take into account the spatial and temporal dependence between the observations. Hence we adopt the validation

criterion developed by Sahu and Mardia<sup>[9]</sup>. Let  $\mathbf{Z}_{orig}$  denote the vector of 162 observations on the original scale for which we seek validation and  $\hat{\mathbf{Z}}_{orig}$  denote the predictions on the original scale and  $\hat{\Sigma}$  denote the 162 dimensional estimated covariance matrix of  $\hat{\mathbf{Z}}_{orig}$ . The validation criterion developed by Sahu and Mardia<sup>[9]</sup> is given by:

$$D^2 = (\mathbf{Z}_{orig} - \hat{\mathbf{Z}}_{orig})' \Sigma^{-1} (\mathbf{Z}_{orig} - \hat{\mathbf{Z}}_{orig}). \quad (11)$$

The quantity  $D^2$  will increase if there are large discrepancies between the predictions based on the model,  $\hat{\mathbf{Z}}_{orig}$  and the observed data,  $\mathbf{Z}_{orig}$ . The observed value of  $D^2$  can be referred to the theoretical values of the  $\chi^2$  distribution with 162 degrees of freedom. In our illustration we shall compare using both *MSE* and  $D^2$ . Using the *MSE* we can compare previous results obtained by Orasi and Jona Lasinio<sup>[10]</sup>.

**Prediction details:** Our aim is to predict the amount of rainfall for all locations in a grid of  $m=2710$  sites at any given time point  $t=1, \dots, T$ . The radar values (covariate) for these locations are available. Moreover, we need the prediction details to carry out the cross-validation for the six sites for which we have set aside data. Here, we provide the prediction details for a location  $s'$  at a time point  $t'$  for the two modeling approaches presented earlier.

The MCMC methods are first implemented for sampling from the posterior distributions. Subsequently, the predictive distributions are sampled by composition. The draws from the posterior enable draws from the predictive distribution of  $X(s', t')$ . This predictive distribution is model dependent and the details for obtaining draws from it are given below.

The sampled values from the predictive distribution of  $X(s', t')$  are then used to construct the predictive distribution of  $Z(s', t')$ . To implement this step we simply invert the censoring relationship given earlier, i.e. we choose the appropriate value of  $Z(s', t')$  by seeing the position of  $X(s', t')$  in the set of ordered values  $c_1 < c_2 < \dots < c_k$ . Finally, to obtain the predictions on the original scale, we simply work with the exponential of the predictive realizations  $Z(s', t')$ .

**Predictive distribution for the Gaussian random effects model:** Using (1) and (2), for a new location  $s'$  at time  $t'$ ,  $X(s', t')$  is conditionally independent of  $\mathbf{z}$  given  $v(s', t')$ , with its distribution given by

$$X(s', t') \sim N\left(\mu(s', t') + v(s', t'), \sigma_\epsilon^2\right). \quad (12)$$

The posterior predictive distribution of  $X(s', t')$  is obtained by integrating over the unknown parameters in (12) with respect to the joint posterior distribution. MCMC samples from the posterior distribution enable us to perform the integration.

Note that in (12) we require a new  $v(s', t')$

conditional on the posterior samples at the observed locations  $s_1, \dots, s_n$  and at the time points  $t_1, \dots, t_T$ . Let  $\mathbf{V}$  denote the vector of  $v(s, t)$  in  $n$  locations and  $T$  time points and  $\otimes$  denote the Kronecker product. For this we have:

$$\begin{pmatrix} v(s', t') \\ \mathbf{V} \end{pmatrix} \sim N \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma_v^2 \bullet \begin{pmatrix} 1 & \Sigma'_s(s-s') \otimes \Sigma'_t(t-t') \\ \Sigma_s(s-s') \otimes \Sigma_t(t-t') & \Sigma_s \otimes \Sigma_t \end{pmatrix} \right]$$

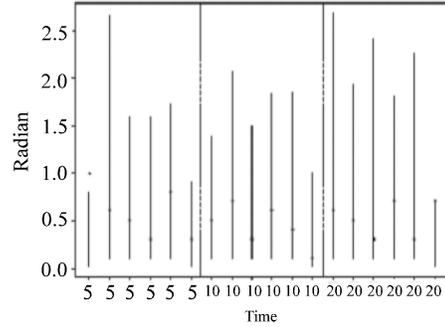


Fig. 7: The 90% prediction intervals of rainfall for times 5(5:50AM), 10(6:40AM) and 20(8:20AM) at each of six validation sites. Observed data are plotted as points in the plot

where  $\Sigma_s(s-s')$  is an  $n \times 1$  column vector with the  $i$ th entry given by  $\sigma(s_i - s') = \rho_s(s_i - s; \phi_s)$  and  $\Sigma_t(t-t')$  is a  $T \times 1$  column vector with the  $k$ th entry given by  $\sigma(t_k - t') = \rho_t(t_k - t'; \phi_t)$ . Hence,

$$v(s', t') / \mathbf{V} \sim N \left( \sum_{j=1}^n \sum_{k=1}^T B_{jk}(s', t') v(s_j, t_k), \sigma_v^2 C(s', t') \right) \quad (13)$$

where

$$B_{jk}(s', t') = \sum_{i=1}^n \sum_{l=1}^T \sigma(s_i - s') \sigma(t_l - t') (\Sigma_s^{-1})_{ij} (\Sigma_t^{-1})_{lk} \quad (14)$$

and

$$C(s', t') = 1 - \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^T \sum_{k=1}^T \sigma(s - s') \sigma(t - t') (\Sigma_s^{-1})_{ij} (\Sigma_t^{-1})_{lk} \sigma(s - s') \sigma(t - t'). \quad (15)$$

The conditional mean and variance are very computationally expensive to calculate due to the dimensionality of  $\mathbf{V}$  and the very large number of sites  $s'$  for which we require predictions. However, by fixing the decay parameters  $\phi$ , the quantities  $B_{jk}(s', t')$  and  $C(s', t')$  given in (14) and (15) need only be calculated once; no updating is required in the MCMC.

**Predictive distribution for the BKKF model:** We use the models (4) and (5) to predict at location  $s'$  and at

time  $t$ . We first obtain the spatial covariance matrix  $\Sigma_\gamma^*$  of order  $n+1$  using the assumed covariogram (6). That is,

$$\Sigma_\gamma^* = \begin{pmatrix} \Sigma_\gamma & \Sigma_{12}(\mathbf{s}') \\ \Sigma'_{12}(\mathbf{s}') & \sigma_\gamma^2 \end{pmatrix}$$

where  $\Sigma_{12}(\mathbf{s}')$  is the  $n$ -dimensional vector with elements  $\sigma(\mathbf{s}_i, \mathbf{s}')$ ,  $i=1, \dots, n$ . Based on the  $n+1$  spatial locations  $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_n$  and  $\mathbf{s}'$  we derive the  $(n+1) \times p$  matrix  $H^*$  using (7) where we replace  $\Sigma_\gamma$  by  $\Sigma_\gamma^*$ .

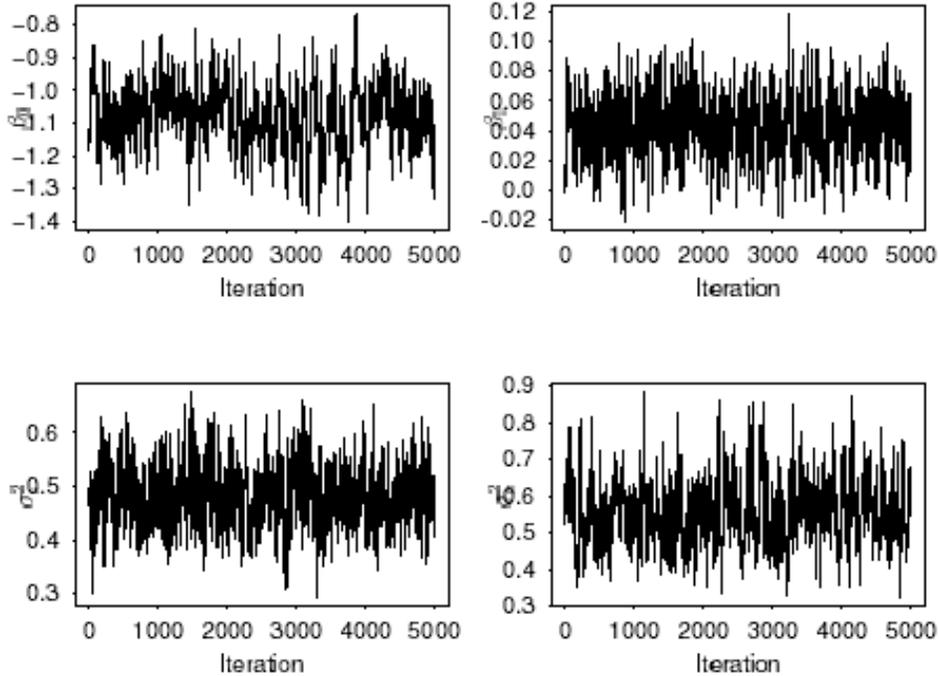


Fig. 8: Trace plot of parameters for the adopted model

Let us partition the matrix  $H^*$  as follows:

$$H^* = \begin{pmatrix} H_1^* \\ H_2^* \end{pmatrix}$$

where  $H_1^*$  is  $n \times p$  and  $H_2^*$  is  $1 \times p$ . We now have that

$$\begin{pmatrix} \mathbf{Y}_t \\ Y(\mathbf{s}', t) \end{pmatrix} \sim N(H^* \alpha_t + \beta_1(\mathbf{r}_t', r(\mathbf{s}', t))', \Sigma_\gamma^*)$$

using the model assumption (5). From this multivariate normal distribution we obtain that

$$Y(\mathbf{s}', t) | \theta \sim N \left( \begin{pmatrix} H_2^* \alpha_t + \beta_1 r(\mathbf{s}', t) + \Sigma'_{12}(\mathbf{s}') \Sigma_\gamma^{-1} \\ \mathbf{Y}_t - H_1^* \alpha_t - \beta_1 \mathbf{r}(\mathbf{s}', t), \\ \sigma_\gamma^2 - \Sigma'_{12}(\mathbf{s}') \Sigma_\gamma^{-1} \Sigma_{12}(\mathbf{s}') \end{pmatrix}, \Sigma_\gamma^* \right) \quad (16)$$

using standard methods. Now using the model assumption (4) we have that

$$X(\mathbf{s}', t) | \theta \sim N(Y(\mathbf{s}', t), \sigma_\epsilon^2), \quad (17)$$

where  $Y(\mathbf{s}', t)$  follows (16) conditionally on  $\theta$ . The draws from the posterior distribution of  $\theta$  enable draws from  $Y(\mathbf{s}', t) | \theta$  as given in (16). Given these and the corresponding draw from the posterior distribution of  $\sigma_\epsilon^2$  we obtain samples from the predictive distribution

of  $X(\mathbf{s}', t)$  as given in (17).

### Analysis

**Model choice and validation:** The spatio-temporal models described earlier require suitable choices of the smoothing parameters  $\phi$ . We adopt the validation MSE criterion (10) to choose these parameters. We calculate the MSE for all models corresponding to a grid of smoothing parameters and then choose the parameter value for which we have the minimum MSE. For the random effect model described earlier we require two smoothing parameter values: one for the spatial and the other for temporal correlation. Using a two-dimensional grid-search we obtain the optimal  $\phi$  values 0.05 and 1 for the spatial and temporal processes. For the Kriged-Kalman model described earlier we also obtain the optimal  $\phi$  to be 0.05. We have also compared several other model fitting statistics and validation criteria, e.g. the  $D^2$  criterion (11) for different values of the smoothing parameters near these optimal values. Those also pointed to the same optimal values and were not very sensitive to changes around those optimal values.

We use the predictive model choice criterion (9) to choose between the random effect model and the Kriged-Kalman model. Table 2 lists the values of model choice criterion for different models. According

to the PMCC the random effect models are seen to be better than the BKKF models. Moreover, the censoring mechanism detailed previously is seen to be worthwhile since the random effect model (1) with this is better than the model without the censoring mechanism implemented. The random effect model is also better than the simple fixed effect model without the spatio-temporal process. Thus our best model is the Gaussian spatio-temporal random effect model with the latent variables to handle discrete data. We have checked the residual plots (not shown) for this model for all the 38 modeling sites. The plots do not show any recognizable pattern and the model seem to be adequate for the data.

We now return to the validation data for six sites which we have set-side. We consider validation at three time points 5:50AM, 6:40AM and 8:20AM. The 90% prediction intervals are plotted in Fig. 7. Only one out of 18 observation falls outside its prediction interval. As a result we conclude that the model has performed well in re-constructing the rainfall fields. For the remainder of this study we shall use this model for analysis.

Table 2: PMCC values for different models; Penalty is the first term in (9) and GOF is the second term

Model	Penalty	GOF	PMCC
Random effect (censored)	751.5	420.1	1171.6
Random effect (not censored)	903.5	435.8	1339.3
Fixed effect	1095.7	1225.6	2321.3
BKKF ( $p=5$ ) (censored)	986.0	1036.2	2022.2
BKKF ( $p=5$ ) (not censored)	1141.5	1032.1	2173.6
BKKF ( $p=15$ ) (censored)	951.1	960.1	1911.2
BKKF ( $p=15$ ) (not censored)	1096.1	954.9	2051.0

**Parameter estimates:** The MCMC trace plots of parameters of the adopted Gaussian random effects model is given in Fig. 8. The MCMC algorithm converges rapidly and mixes well.

Table 3 provides the parameter estimates for the adopted random effect model. The regression coefficient  $\beta_1$  for the radar measurements is seen to be significant and positive as expected. Furthermore, the estimated values of  $\beta_0$  and  $\beta_1$  are consistent with the approximate linear relationship we have seen in the scatter plot given in Fig. 3. The random effect variance,  $\sigma_V^2$  is slightly larger than the error variance  $\sigma_e^2$  which shows that the random effects explain more variation than the pure error.

Table 3: Parameter estimates for the chosen random effect model

Parameter	Mean	sd	95% interval
$\beta_0$	-1.080	0.088	(-1.249, -0.912)
$\beta_1$	0.044	0.020	(0.004, 0.085)
$\sigma_e^2$	0.485	0.059	(0.373, 0.602)
$\sigma_V^2$	0.537	0.093	(0.365, 0.731)

**Predictions: Reconstructing the rainfall fields:** We use the prediction details discussed to reconstruct the

rainfall fields at different time points. We obtain the rainfall maps at time 5, 10 and 20 for illustration in Fig. 9-11. These time points correspond to 5:50AM, 6:40AM and 8:20AM respectively. From these figures we see precipitation moving from the south-west to the north-east. The south-east part of the region has remained consistently dry. The standard deviations of the predictions increases with the predicted amount of rainfall, although they are smaller near the observation sites. For the dry regions at any time point the standard deviations are approximately zero which implies that accurate predictions are possible. Effective modeling of discrete data using the censoring mechanism developed here has made this possible.

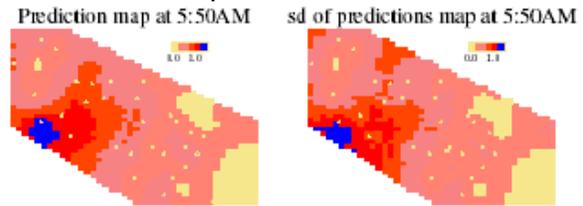


Fig. 9: Predictions and their standard errors at 5:50AM

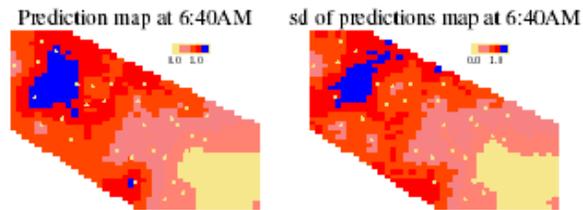


Fig. 10: Predictions and their standard errors at 6:40AM

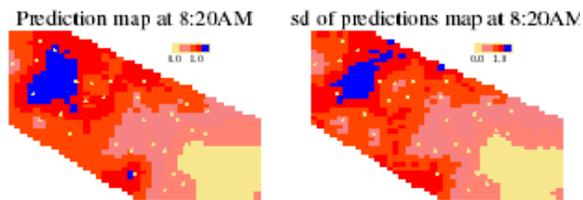


Fig. 11: Predictions and their standard errors at 8:20AM

## DISCUSSION

In this study we have compared two competitive spatio-temporal modeling approaches for rainfall data obtained from a cloud seeding experiment in the regions of south Italy including Puglia. The Gaussian random effect model is seen to perform better than the BKKF model. Each model has, a priori, a good reason to be chosen. The BKKF is a model that naturally extend in a Bayesian space-time setting the geostatistical approach which is usually considered quite sensible when treating rainfall-radar data<sup>[6,11,18-20]</sup>.

It is able to easily account for non separable behavior of the space-time process and allow to include seasonal effects and covariates in an easily interpretable manner. The Gaussian random effect model with separable space-time covariance structure and the fixed effect model can be seen as alternatives to this approach when there is a reasonable suspect that the space-time process is indeed separable. Furthermore, our data are essentially discrete and this fact is not accounted for in any of the above models. We have developed a censoring method using a latent variable to handle multiple discrete (rounded) rainfall amounts and introduced it in both the models. The benefits of modeling of discrete data using the censoring mechanism are seen in more accurate predictions of dry periods with no positive prediction standard errors. Notice that the BKKF performs better than the fixed effect model but worse than the random effect one with or without censoring. A possible reason for this is that the spatio-temporal process is indeed separable and the BKKF is not a suitable model for such processes.

#### ACKNOWLEDGEMENT

This work has been partially supported by the MIUR COFIN project 2002-2004 and it has been part of the project CLIMAGRI (UCEA) subproject 3.

#### REFERENCES

1. Allcroft, D.J. and C.A. Glasbey, 2003. A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *J. Royal Statistical Society, Series C, Applied Statistics*, 52: 487-498.
2. Sansò, B. and L. Guenni, 1999. Venezuelan rainfall data analysed by using a Bayesian space-time model. *J. Royal Statistical Society, Series C, Applied Statistics*, 48: 345-362.
3. Sansò, B. and L. Guenni, 2000. A nonstationary multisite model for rainfall. *J. Am. Stat. Assoc.*, 95: 1089-1100.
4. Stroud, J.R., P. Müller and B. Sansò, 2001. Dynamic models for Spatio-temporal data. *J. Royal Stat. Soc., B*, 63: 673-689.
5. Brown, P.E., P.J. Diggle, M.E. Lord and P.C. Young, 2001. Space-time calibration of radar rainfall data. *J. Royal Stat. Soc., Series C, Applied Statistics*, 50: 221-241.
6. Cassiraga, E.F., C. Guardiola-Albert and J.J Gomez-Hernandez, 2004. Automatic modeling of cross-covariances for rainfall estimation using raingage and radar data. *GEOENV IV - Geostatistics for environmental applications: Proc. Quantitative Geology and Geostatistics 13: 391-399*, Kluwer Academic Publishing, Dordrecht.
7. Cornford, D., 2004. A Bayesian state space modelling approach to probabilistic quantitative precipitation forecasting. *J. Hydrol.*, 288: 92-104.
8. Jordan, P.W., A.W. Seed. and P.E. Weinmann, 2003. A stochastic model of radar measurement errors in rainfall accumulations at catchment scale. *J. Hydrometeorol.*, 4: 841-855.
9. Sahu, S.K. and K.V. Mardia, 2005. A bayesian kriged-kalman model for short-term forecasting of spatio-temporal processes. *J. Royal Stat. Soc., Series C, Applied Statistics*, 54: 223-244.
10. Orasi, A. and G. Jona Lasinio, 2004. Comparison of calibration methods for the reconstruction of space-time rainfall fields in Southern Italy during a rain enhancement experiment. Technical Report, University of Rome.
11. Orasi, A., G. Jona Lasinio and C. Ferrari, 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/orasij\\_onaferrari.pdf](http://w3.uniroma1.it/dspsa/Rapporti_tecnici/orasij_onaferrari.pdf)
12. Marshall, J. and W. Palmer, 1948. The distribution of raindrops with size. *J. Meteorol.*, 5: 165-166.
13. Mardia, K.V., C.R. Goodall, E.J. Redfern and F.J. Alonso, 1998. The kriged kalman filter (with discussion). *Test*, 7: 217-252.
14. Mardia, K.V., J.T. Kent and J.M. Bibby, 1979. *Multivariate Analysis*. London: Academic Press.
15. Carroll, S.S. and N. Cressie, 1996. A comparison of geostatistical methodologies used to estimate snow water equivalent. *Water Resources Bulletin*, 32: 267-278.
16. Gelfand, A.E. and S.K. Ghosh, 1998. Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, 85: 1-11.
17. Laud, P.W. and J.G. Ibrahim, 1995. Predictive Model Selection. *J. Royal Stat. Soc., B*, 57: 247-262.
18. Seo, D., W. Krajewski and D. Bowles, 1990a. Stochastic interpolation of rainfall data from raingages and radar using cokriging. 1. design of experiments. *Water Resour. Res.*, 26: 469-477.
19. Seo, D., W. Krajewski, D. Bowles and A. Azimi-Zonooz, 1990b. Stochastic interpolation of rainfall data from raingages and radar using cokriging. 2. results. *Water Resour. Res.*, 26: 915-924.
20. 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.), Vol. 2 of *Geostatistics Wollongong '96*, pp: 1306-1315. E.Y. Baafi and N.A. Schofield eds.