A Bayesian Approach to Compare Observed Rainfall Data to Deterministic Simulations

Bruno Sansó†

Centro de Estadística y Software Matemático, Universidad Simón Bolívar, Caracas Venezuela and Department of Applied Mathematics and Statistics, University of California, Santa Cruz, U.S.A.

E-mail: *bruno@ams.ucsc.edu,* URL: *www.ams.ucsc.edu/∼bruno*

Lelys Guenni

Centro de Estadística y Software Matemático, Universidad Simón Bolívar, Caracas Venezuela

E-mail: *lbravo@cesma.usb.ve,* URL: *www.cesma.usb.ve/∼lbravo*

Summary. We compare ground rainfall with purely deterministic Regional Climate Model (RCM) simulations within a Bayesian framework. A truncated normal model is fitted to the observed ground data to represent spatial variability. The predictive posterior distribution of the spatially aggregated rainfall is obtained by using a Markov chain Monte Carlo method and compared to the RCM simulations. Also, the predictive posterior distribution of the RCM output is downscaled using the truncated normal model and obtaining pointwise rainfall estimates from aereal observations which are compared to the ground observations. These two procedures allow to determine if the differences between the two sources of information are compatible with the variability predicted by the spatial model. Also, point rainfall estimates at locations without rainfall measurements conditioned on RCM observations can be obtained. We considered a set of data from an area in Nebraska for which time is considered fixed and rainfall is accumulated monthly.

†*Address for correspondence:* Bruno Sansó, School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064 USA.

1. Introduction

Global atmospheric general circulation models (GCMs) are deterministic models used to generate meteorological data from large scale atmospheric variables in order to study climate processes and natural climate variability. The resolution of GCMs is usually limited to around 300 km. Following Murphy (1999) a "dynamic downscaling" of the output of GCMs can be produced by a Regional Climate Model (RCM). RCMs produce detailed meteorological data by using large scale information as initial, lateral and surface boundary conditions usually derived from the GCMs (Giorgi and Mearns, 1999). These are purely deterministic models that attempt to simulate relatively short-term atmospheric and land-surface processes and the interactions between the two, at a spatial resolution much higher than that produced by GCMs.

According to Richard et al. (2002), RCMs are used for a variety of applications: downscaling climate change projections from GCMs; assessing the implications of regional change due to changing regional factors such as land use; reconstruction of regional scale paleoclimatology, among others. The authors also highlight the need for a careful validation of the RCMs by observing that "When RCMs are forced with operational atmospheric objective analyses or re-analyses they should reproduce observed present day climate, i.e., its mean behaviors, variability and extremes." Validating the RCMs presents the challenge of comparing information at different scales. Outputs from RCMs correspond to a grid where cells are of around 50×50 km. Usually for validation purposes, gridded values are taken as point values and compared to the nearest measured point data (Mearns et al., 1999, Liston and Pielke, 2000).

Under present climate conditions RCMs are the state of the art in representing current climate feedback mechanisms between large scale information and local conditions, but they do not attempt to describe the weather of a specific area at a specific point in time. So the problem of calibrating a RCM consists of deciding whether the model is reproducing climate conditions that are compatible with given observations as oppose to requiring a close agreement of simulations and readings obtained on the ground. In this paper we propose a method of calibration based on the exploration of the predictive distribution obtained from a statistical model linking observations with model output. Such a model has two byproducts. First, the spatial resolution of RCMs can still be too coarse for applications over small geographical areas. Learning the statistical links between large and local scales allows to perform a stochastic downscaling of RCM output. Second, RCM outputs contain valuable information on a scale much larger than that provided by ground based data. Thus, simulations can be used to drive stochastic models which represent the statistical behavior of the ground based network, as a way of improving statistical predictions in locations without point measurements.

In this paper we use a Bayesian framework to represent the space variability of rainfall observed on the ground and its links to RCM predictions. Fuentes and Raftery (2001) consider a similar problem for SO_2 concentrations. A related problem is that considered in Gelfand, Zhu and Carlin (2000) where spatial data with different support are studied. To illustrate the type of data that we consider in this paper, we present in Figure 1 the plots that correspond to monthly rainfall over a region in Nebraska during May 1989. The left panel shows the output of the RCM, the right panel is obtained by a piecewise interpolation of the ground observations using the function `interp` in `Splus`. The left panel should correspond to a system under climate conditions equivalent to those that produced the observed data on the right panel. Notice that the RCM produces a substantial overestimation of the rainfall

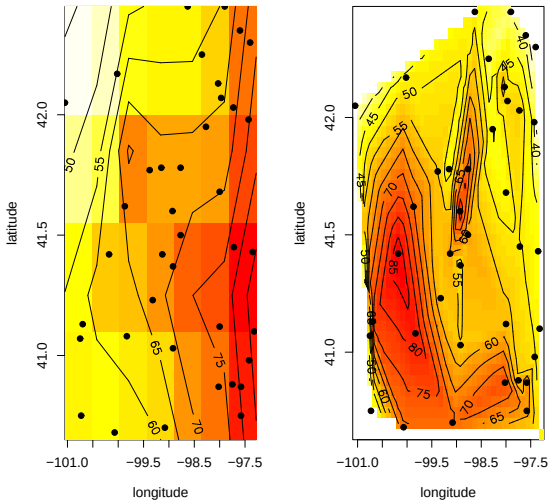

Fig. 1. Monthly rainfall over a region in Nebraska in May 1989: estimated by a RCM model (left panel); Ground observations during the same period interpolated using the `interp` function in `Splus` (right panel). The dots denote the locations of the ground stations (not used to obtain the output of the RCM model).

in the eastern part of the region and some underestimation in the western part. Our goal is to decide if such differences are acceptable under the typical variability of the observed data.

We proceed by first fitting an interpolating model to the observed rainfall. We use a truncated normal model in a similar fashion to Sansó and Guenni (1999). This, together with the Markov chain Monte Carlo method (MCMC) that we propose for the exploration of the distribution of the parameters in the model, is presented in Section 2. In Section 3 we create a link between the ground observations and the RCM output by assuming that the latter is a, possibly biased, realization of the mean aereal grid process that corresponds to the pointwise model for the former. We also present the additions and modifications to the previously discussed MCMC that are relevant for the new setting. After exploring the posterior distribution of model parameters it is possible to obtain a description of the predictive distribution. Such distribution provides a full assessment of the behaviour of the samples that would be produced by the model after taking into account all available information. The comparison of the RCM output with the model prediction is considered in Section 4 were the model proposed in Sections 2 and 3 is applied to rainfall corresponding to April and May 1989 over a region in Nebraska. The last section presents a discussion of the results and points toward further developments.

2. Specification of the model for ground rainfall

We consider data collected at a number of rainfall stations at a given point in time. We assume that the area under study has been gridded and cells on the grid, say S_i, are numbered from 1 to M. Such cells correspond to the output of an RCM producing one aereal rainfall prediction per cell, say $a_1, \ldots, a_M$. We denote the locations of the stations within the ith cell by $s_{ij}, j = 1, \ldots, n_i$ and the corresponding observation by r_{ij}.

The comparison of the ground observations with the RCM simulations requires that the pointwise data be spatially aggregated. This can be done by interpolating a random field to the observations and then integrating over the corresponding cell grids, as in Rodríguez-Iturbe and Mejía (1974). Stein (1992) warns about the problems of making inference about a random field like rainfall, which is naturally truncated for negative values and accumulates positive probability at zero, without taking into account the truncation. In view of this, we consider a model based on a truncated normal as in Sansó and Guenni (1999). We assume that

$$r_{ij} = \begin{cases} w_{ij}^{\beta} & w_{ij} > 0 \\ 0 & w_{ij} \leq 0 \end{cases}$$

where, conditional on some unobserved variables z_{ij}, $w_{ij} = z_{ij} + \eta_{ij}$, η_{ij} are uncorrelated normal errors with zero mean and variance ν^2. Denoting w and z the vector of, respectively, all w_{ij} and z_{ij} we consider the hierarchical model

$$\begin{array}{ll} w = z + \eta & \eta \sim N(0, \nu^2 I) \\ z = X\alpha + \varepsilon & \varepsilon \sim N(0, \sigma^2 V_\lambda) \ , \end{array} \tag{1}$$

where $N(\cdot, \cdot)$ denotes a normal distribution, V_λ is a covariance matrix such that $\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{kl})$ is a function of the distance between s_{ij} and s_{kl} which depends on a parameter, say, λ; $\alpha \in \mathbb{R}^p$ are unknown regression parameters and X is an $n \times p$ matrix defined by $X_{ij} = f_j(s_i)$ where $f(s) = (f_1(s), \ldots, f_p(s))'$ are known location-dependent covariates. Note that equation (1) corresponds to a hierarchical representation of the usual Gaussian random field model with a nugget effect (Cressie, 1993), which is convenient for simulation purposes as will become clear later.

We can express the relationship between w_{ij} and r_{ij} as

$$w_{ij} = \begin{cases} r_{ij}^{1/\beta} & \text{if } r_{ij} > 0 \\ v_{ij} & \text{if } r_{ij} = 0 \\ u_{ij} & \text{if } r_{ij} \text{ is missing} \end{cases}$$

where v_{ij} and u_{ij} are unknown quantities, with the restriction that $v_{ij} < 0$. The reason for considering missing observations will become clear in Section 3. Letting $\rho^2 = \nu^2/\sigma^2$ we have that

$$p(w|\beta, \nu^2, \rho^2, \lambda, z, v, u) \propto \frac{(\rho^2)^{N/2} |V_\lambda|^{-1/2}}{(\nu^2)^N} \left(\prod_{r_{ij}>0} \frac{r_{ij}^{1/\beta - 1}}{\beta} \right)$$
$$\exp\left\{ -\frac{1}{2\nu^2} \left(||w - z||^2 - \rho^2 (z - X\alpha)' V_\lambda (z - X\alpha) \right) \right\} \tag{2}$$

where $N = \sum_{i=1}^{M} n_i$. Following the classical terminology we observe that the nugget is given by ν^2 and the sill is equal to $\nu^2 + \sigma^2 = \nu^2(1 + 1/\rho^2)$, thus the nugget to sill ratio is equal

to $(1+1/\rho^2)^{-1}$. This quantity will be of interest for the specification of the parameters of the prior of ρ^2.

2.1. Exploration of the posterior distribution

Obtaining a MCMC for the posterior distribution of the parameters in (2) is complicated by the lack of conditional conjugacy. This can be handled by the use of a Metropolis-Hastings scheme (Gamerman, 1997). Another difficulty is that samples of the parameters corresponding to the covariance matrix of a spatial model (λ in our case) are usually subject to slow mixing of the chain and choosing the appropriate jumping distribution is not simple. We tackle these problems by considering Metropolis-Hastings steps for blocks of parameters, after performing partial marginalisations of some of the full conditionals.

Let R denote the observed rainfall data, then consider a block of parameters given by ρ^2, α and λ. We observe that

$$p(\alpha,\rho^2,\lambda|z,\nu^2,R) \propto p(\alpha,\rho^2,\lambda)|V_\lambda|^{-1/2}(\nu^2)^{-N}(\rho^2)^{N/2} \exp\left\{-\frac{\rho^2}{2\nu^2}\left((\alpha-\hat{\alpha})'(X'V_\lambda^{-1}X)(\alpha-\hat{\alpha})+z'B_\lambda z\right)\right\}$$

where $\hat{\alpha}$ is the solution of $(X'V_\lambda^{-1}X)\hat{\alpha} = X'V_\lambda^{-1}z$ and $B_\lambda = (I-P_\lambda)V_\lambda^{-1}$ where $P_\lambda = V_\lambda^{-1}X(X'V_\lambda^{-1}X)X'$.

Assume that the prior $p(\alpha,\rho^2,\lambda) \propto p(\rho^2)p(\lambda)$, where $p(\rho^2)$ is a gamma distribution with parameters a_ρ and b_ρ. Then we have that $p(\alpha,\rho^2,\lambda|z,\nu^2,R) = p(\alpha|\rho^2,\lambda,z,\nu^2,R)$ $p(\rho^2|\lambda,z,\nu^2,R)p(\lambda|z,\nu^2,R)$ where

$$p(\alpha|\rho^2,\lambda,z,\nu^2,R) \propto N\left(\hat{\alpha},\frac{\nu^2}{\rho^2}(X'V_\lambda^{-1}X)\right) \quad ,$$

$$p(\rho^2|\lambda,z,\nu^2,R) \propto \exp\left\{-\frac{\rho^2}{2\nu^2}S_\lambda^2\right\}(\rho^2)^{(n-p)/2}p(\rho^2) \quad ,$$

with $S_\lambda^2 = z'B_\lambda z = (z-X\hat{\alpha})'V_\lambda^{-1}(z-X\hat{\alpha})$ and

$$p(\lambda|z,\nu^2,R) \propto \left(S_\lambda^2+2\nu^2 b_\rho\right)^{(n-p)/2+a_\rho}|V_\lambda|^{-1/2}|X'V_\lambda^{-1}X|^{-1/2}p(\lambda) \quad .$$

To obtain a sample of λ, ρ^2 and α we propose the following Metropolis-Hastings scheme: given the current sample $(\lambda_t,\rho_t^2,\alpha_t)$ obtain a proposal $(\lambda_*,\rho_*^2,\alpha_*)$ by sampling first $\log(\lambda_*)$ from a $N(\log(\lambda_t),\kappa^2)$, then ρ_*^2 from a $G((n-p)/2+a_\rho,(S_{\lambda_*}^2/2\nu^2)+b_\rho)$, where $G(\cdot,\cdot)$ denotes a gamma density, and last, sample α_* from $N(\hat{\alpha}_*,(\nu^2/\rho_*^2)(X'V_{\lambda_*}^{-1}X))$. Accept the sample with probability

$$\min\left\{1,\frac{p(\alpha_*,\rho_*^2,\lambda_*)}{p(\alpha_t,\rho_t^2,\lambda_t)}\,\frac{p(\lambda_t)p(\rho_t^2|\lambda_t)p(\alpha_t|\rho_t^2,\lambda_t)}{p(\lambda_*)p(\rho_*^2|\lambda_*)p(\alpha_*|\rho_*^2,\lambda_*)}\frac{\lambda_*}{\lambda_t}\right\} = \min\left\{1,\frac{p(\lambda_*)}{p(\lambda_t)}\frac{\lambda_*}{\lambda_t}\right\} \quad ,$$

where all distributions have to be taken conditional on z,ν^2 and R. Notice that the acceptance probability depends only on the (marginalised) full conditional of λ and thus no samples of ρ^2 and α have to be effectively computed if λ_* is rejected. In our experience this schemes produces a chain with better mixing properties than the one that is obtained by

independent sampling of the full conditionals of α, ρ^2 and λ, with the bonus of a reduced sampling effort.

The full conditional of β can be sampled using another Metropolis step. We observe that, if $p(\nu^2)$ is taken as an inverse gamma distribution with parameters a_ν and b_ν, that we denote as $IG(a_\nu, b_\nu)$, then the full conditional of ν^2 is proportional to

$$\exp\left\{-\frac{1}{2\nu^2}||w-z||^2\right\}(\nu^2)^{-N}p(\nu^2) \propto IG(N+a_\nu, ||w-z||^2+b_\nu) \quad .$$

The only term that involves u and v in the joint posterior distribution of all parameters is $e^{-||z-w||^2/(2\nu^2)}\mathbf{1}_{\{v<0\}}$, which is a truncated normal kernel. Thus, denoting by z_1 the sub-vector of z that corresponds to u and z_2 the sub-vector that corresponds to v, we sample u from a $N(z_1, \nu^2 I)$ and v from a $N(z_2, \nu^2 I)$ truncated so that all variates are negative.

To obtain a sample from the full conditional of z we observe that

$$\exp\left\{-\frac{1}{2\nu^2}\left(||w-z||^2-(z-X\alpha)'V_\lambda^{-1}(z-X\alpha)\right)\right\} \propto \exp\left\{-\frac{1}{2}(z-m)'C^{-1}(z-m)\right\}$$

where

$$m = \nu^2 C\left(w+\rho^2 V_\lambda^{-1}X\alpha\right) \quad \text{and} \quad C = \rho^2\left(I+\rho^2 V_\lambda^{-1}\right) \quad .$$

This corresponds to a multivariate normal distribution with mean m and covariance matrix C.

3. A model for the RCM output

In order to compare the RCM simulations to the ground observations, let $a(S)$ be the process of aereal rainfall indexed on the area S and $r(s)$ the process of rainfall indexed on the location s, then, for the ith cell,

$$a(S_i) = \frac{1}{|S_i|}\int_{S_i} r(s)ds \approx \sum_{j=1}^{k_i}\gamma_{ij}r(s_{ij}) = \sum_{i=1}^{k_i}\gamma_{ij}w_{ij}^\beta \mathbf{1}_{\{w_{ij}>0\}} \tag{3}$$

for some integer k_i and some constants γ_{ij}. Kriging methods like the ones proposed in Cressie (1993) will have γ_{ij} depending on λ, but a simpler approximation is to let $\gamma_{ij} = 1/k_i$. Note that, the number k_i of sites needed to obtain a meaningful expression for the mean of the aereal rainfall could be larger than n_i, the number of available observations in the cell i, thus several of the components of w will be treated as missing and correspond to u_{ij}. We shall assume that $a_i = \eta + a(S_i) + \zeta_i$ with $\zeta_i \sim N(0, \tau^2)$ and that this equation is approximated by $a_i = \eta + \sum_{i=1}^{k_i}\gamma_{ij}w_{ij}^\beta \mathbf{1}_{\{w_{ij}>0\}} + \zeta_i$, where η represents a, possible, systematic bias of the RCM simulations with respect to the actual aereal rainfall process. Let $A = (a_1, \ldots, a_M)$, then

$$p(A|w, \tau^2, \eta, \beta) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{M}\left(a_i-\eta-\sum_{j=1}^{k_i}\gamma_{ij}w_{ij}^\beta \mathbf{1}_{\{w_{ij}>0\}}\right)^2\right\}\left(\tau^2\right)^{-M/2} \quad . \tag{4}$$

From expressions (2) and (4) it is possible to obtain the joint likelihood, say, $p(A, w|\theta)$ of all the data, where θ denotes all the parameters in the model. We can also interpret equation (4) as a probabilistic restriction imposed on some of the parameters of model (2) due to the observations obtained from the RCM, or as a data driven prior for the unknown variables u_{ij}.

The exploration of the posterior distribution of θ from the combination of $p(A, w|\theta)$ with a prior distribution $\pi(\theta)$ yields several products that are useful for the comparison of R and A. In the first place we have an estimation of η, that can reveal and quantify a systematic bias in the RCM output. Second, denoting the predictive posterior distribution of the RCM simulations as $p(a|A, R)$ we can assess the predictive behaviour of the model, after observing both sources of data, in comparison with the simulation output. On the other hand, we can obtain ground rainfall predictions driven by RCM output using the predictive posterior distribution of the pointwise rainfall, say $p(r|A, R)$.

3.1. Downscaling aereal rainfall to point observation

The proposed model allows for the downscaling of the aereal RCM observations. In fact, suppose that all rainfall observations r_{ij} are missing. Then we have a model where all components of w are equal to u. We can run the Markov chain to obtain estimations of the missing values. This is a way of obtaining pointwise estimated values of rainfall from aereal observations. Therefore, following this approach, downscaling can be done by simulation of the posterior predictive distribution $p(r|A, R)$ when the components of w are equal to u. This method can be used to calibrate the output of the RCM by comparing the estimates with the observed pointwise values. It can also be used to perform simulation studies of the pointwise rainfall corresponding to different sets of initial and boundary conditions of the RCM. More explicitly, suppose a new RCM simulation is run under different boundary conditions (like doubled CO2 concentrations) and A^* is obtained. Then a downscaled version of A^*, say R^*, can be obtained from the predictive posterior $p(r^*|R, A, A^*)$. This requires samples of the posterior $p(\theta|A^*, A, R)$ which can be re-sampled from $p(\theta|A, R)$.

3.2. Exploration of the posterior distribution

Many of the components of θ have already been considered in Section 2.1 and their full conditionals are unchanged. In order to sample the remaining parameters let

$$Q = \sum_{i=1}^{M} \left(a_i - \eta - \sum_{j=1}^{k_i} \gamma_{ij} w_{ij}^{\beta} \mathbf{1}_{\{w_{ij}>0\}} \right)^2 . \tag{5}$$

The full conditional of β is proportional to

$$\left(\prod_{r_{ij}>0} \frac{r_{ij}^{1/\beta - 1}}{\beta} \right) \exp\left\{ -\frac{1}{2\nu^2} ||w - z||^2 - \frac{1}{2\tau^2} Q \right\} p(\beta) . \tag{6}$$

To obtain a proposal, sample from a normal with mean equal to the logarithm of the current value of β, then use expression (6) to obtain the acceptance probability. The full conditional of τ^2 is given by $e^{-Q/(2\tau^2)} p(\tau^2)$. If an inverse gamma prior is considered, the full conditional will correspond to an inverse gamma.

Getting samples of u and v is complicated by the fact that Q may depend on u and v. We consider a Metropolis-Hastings step using a jumping distribution based on the distributions considered in Section 2.1, i.e. we sample proposals u^* from a $N(z_1, \nu^2 I)$ and v^* from a $N(z_2, \nu^2 I)$ truncated so that all variates are negative. Cancelations in the formula of the acceptance probability of these proposed variates show that these proposals are accepted with a probability equal to $e^{-(Q^*-Q)/(2\tau^2)}$, where Q^* denotes the evaluation of (5) using u^* and v^*.

To obtain samples of η let $H_i = a_i - \sum_{j=1}^{k_i} \gamma_{ij} w_{ij}^{\beta} \mathbf{1}_{\{w_{ij}>0\}}$ and $\overline{H} = 1/M \sum_{i=1}^{M} H_i$. Assuming that the η follows the prior $N(0, h^2)$, then the full conditional of η will be

$$N\left(\frac{\overline{H}h^2}{h^2 + \tau^2/M}, \frac{h^2\tau^2/M}{h^2 + \tau^2/M}\right).$$

4. The data

Data are taken from simulations of a climate version of the Regional Atmospheric Modeling System (RAMS) called ClimRAMS, which was used to simulate diurnal, seasonal, and annual cycles of atmospheric and hydrologic variables and their interactions within the central United States during 1989 (Liston and Pielke, 2000). The model was nested from a 200-km grid covering the coterminous Unites States into a 50-km grid covering the Great Plains and Rocky Mountain states of Kansas, Nebraska, South Dakota, Wyoming and Colorado. A region from the original 50-km grid was selected. This region of 28 ($= M$) 40×50-km grid cells is located in Nebraska, where a high density of point rainfall measurements are available from the National Climatic data Center (NCDC) Summary-of-the-Day (SOD) data set. The number of rainfall stations for the selected area is 39.

We analysed two sets of data, one corresponding to May 1989, which was presented in figure 1. There are no zero values in this data set and there seems to be an interesting spatial variability over the area. The values predicted by the ClimRAMS model are fairly comparable with the ones obtained on the ground but there is a relevant overestimation of rainfall in the east and, in particular, in the south-central and south-east portion of the region.

The other set of data corresponds to April 1989. April's data are shown in figure 2. This is a much drier month than May, although only three of the stations recorded zero rainfall. The comparison of the left panel with the results of the ClimRAMS simulation shows a substantial overestimation of the model in the whole region with respect to the ground observations in the right panel. The north-east corner of the region, where we observe some underestimation, is the only exception.

4.1. The priors

We carefully considered the choice of the hyperparameters of the priors to avoid very flat posteriors that create estimation problems. We observed that the likelihood is very flat as a function of β, for values greater than 0.9. From previous experience fitting the truncated normal model to rainfall data we chose as a prior for β a gamma distribution with mean 3 concentrated around 2.5 and 3.5. This is in agreement with the empirical evidence that the cubic root of the observed rainfall is approximately normal. The prior for the nugget is specified by a_ν and b_ν which were taken to obtain a gamma density with mean 0.003 and

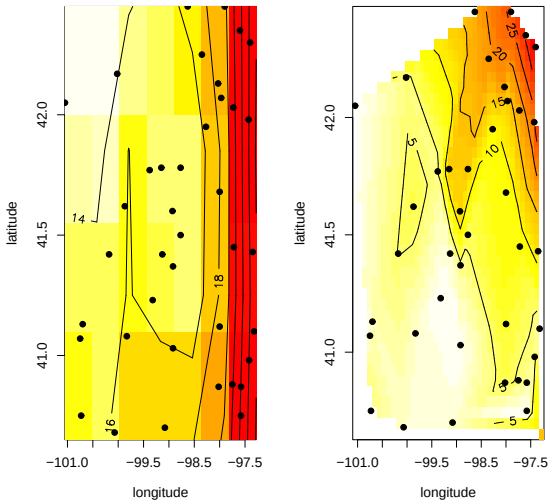

Fig. 2. Monthly rainfall over a region in Nebraska in April 1989: estimated by a RCM model (left panel); Ground observations during the same period interpolated using the `interp` function in `Splus` (right panel). The dots denote the locations of the ground stations (not used to obtain the output of the RCM model).

standard deviation 0.001, so that, a priori, w is subject to relative errors of about 0.001. This specification did not seem to have much influence on the posterior inference.

On the other hand, the values of a_ρ and b_ρ, the hyperparameters of ρ^2, are fairly influential. From the discussion following equation (1) we observe that fixing the value of the nugget to sill ratio at $1/20$ implies $\rho^2 = 1/19$. We used this value as the mean of the prior for ρ^2, with a standard deviation of 10^{-2}. We observed that the values of the nugget to sill ratio affect the flatness of the predictive surface of rainfall, higher values producing flatter surfaces; nevertheless, values smaller than $1/20$ did not seem to change predictions substantially.

We considered a spatial correlation given by

$$\mathrm{cov}(\varepsilon_{ij}, \varepsilon_{kl}) = \frac{\nu^2}{\rho^2} \frac{1}{2^{\lambda_2 - 1}\Gamma(\lambda_2)} \left(\frac{||s_{ij} - s_{kl}||}{\lambda_1} \right)^{\lambda_2} \mathcal{K}_{\lambda_2} \left(\frac{||s_{ij} - s_{kl}||}{\lambda_1} \right)$$

where $\mathcal{K}_{\lambda_2}(\cdot)$ is the modified Bessel function of second kind and order λ_2. λ_1 and λ_2 are positive parameters corresponding, respectively, to the spatial range and the smoothness of the correlation. This is known as the Matèrn class (see for example, Stein, 1999, chap. 2) and has been advocated as a fairly flexible class of isotropic correlations. Other correlation functions could be easily considered within the framework that we present

A gamma density was taken as the prior for λ_1. The hyperparameters were chosen so that the prior expectation of λ_1 equals 0.3 and the standard deviation 0.54, values that are compatible with the empirical estimate of the spatial correlation function. The prior for λ_2 was taken as a normal with mean one and standard deviation $1/3$. This choice is

justified by the argument in Whittle (1954), according to which $\lambda_2 = 1$ should be the default value. These priors did not seem to be influential on the posterior inference. Berger, De Oliveira and Sansó (2001) discuss the problem of using a proper prior for λ to guarantee the propriety of its posterior for a Gaussian model with no nugget. We tried the reference but proper prior that they obtain, but it did not seem to work for model (2), since the chain showed a very erratic behaviour. More theoretical work is needed for a full understanding of the problem.

4.2. Posterior inference

We started by fitting model (2) to both sets of data. In both cases we run the MCMC described in Section 2.1 for 30,000 iterations after a burn-in of 2,000. Assessment of the convergence of the chains was done using the `R` functions provided in `BOA`, a software for the analysis of MCMC output developed by Brian Smith (`brian-j-smith@uiowa.edu`). Most inferences presented here were produced with a sub-sample of 200 interations with large lags between them.

Originally we took f as a second order polynomial in latitude and longitude, so that $\alpha \in \mathbb{R}^6$. We found that posterior intervals of 95% probability for $\alpha_i, i \geq 2$ contained the value zero, for both sets of data, so we resorted to the simplified model with constant spatial mean. To assess the goodness of fit we split the stations randomly in four blocks, three of ten stations and one of nine stations. We fitted the data leaving one block out at a time and obtained the posterior predictive density (PPD) of each station. In Figures 3 and 4 we compare those distributions to the actual observed value for a group of six stations chosen at random. We observe that, for most stations, the actual observed value of rainfall is pretty central to the predictive distribution, corresponding to a likely value. A similar behaviour was observed for the remaining stations. A surface obtained using the whole set of observations, corresponding to the median of the joint PPD of a 40×40 regular grid, can be seen on the left hand panels of Figure 5. For both months we observe patterns that are similar to the crude interpolations in the right panels of Figures 2 and 1. We conclude that the model produces an acceptable fit of the data.

A comparison between the ClimRAMS simulations and the PPD of the aereal rainfall is presented in Figures 6 and 7, for nine randomly chosen grid cells. We observe that for very few of the cells the value simulated by the RCM could not be considered as a likely sample from the PPD. This behaviour is representative of that observed in most cells for both months.

Figure 8 shows the fields of the standardized differences between the ClimRAMS output and samples from the PPD of the aereal rainfall. The standardization was done dividing by the posterior standard deviations of the aereal rainfall. The error field in the left panel corresponds to May and the one in the right panel to April. We notice that, for both months, there are clusters of large and small values that point at, respectively, systematic overestimation and underestimation of the ClimRAMS model. The left panel (May) shows a clear overestimation of the ClimRAMS model increasing from west to east and the right panel (April) shows overestimation increasing from south to north.

The analysis was completed by fitting the model specified by equations (2) and (4) using both the ground observations and the ClimRAMS output. A regular grid of 20×20 was used to approximate the integral in (3). Increasing the number of points in the grid will produce a better approximation of the aereal random field but a compromise has to be taken to make the MCMC computationally feasible. In fact, for a 20×20 grid the matrix

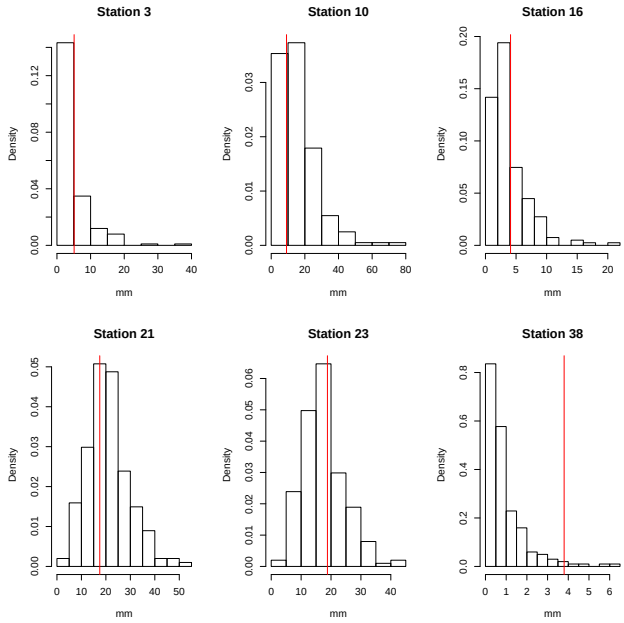

Fig. 3. Histograms of samples of the PPD of ground rainfall for April 1989 at six randomly chosen stations. Vertical lines correspond to actual observed values (not used in the model fit).

Table 1. Posterior median and quartiles of the bias parameter for the May and April 1989 data

	1st Quartile	Median	3rd Quartile
April	7.84	9.31	10.78
May	4.87	7.61	10.34

V_λ is of 400×400, already implying a heavy computational burden. The median, first and third quartiles of the posterior distribution of η are reported in Table 1. Given that the averages of the ClimRAMS simulated values are 17.5mm and 63.8mm for April and May respectively, we observe that the median biases are 53% and 12% of the averages.

A byproduct of the comparison is given by the surfaces on the right panels of Figure 5. These correspond to the median rainfall predictions based on the ground readings using a (4) as prior based on the ClimRAMS simulations. We observe similar patterns in the left and right panels of both figures. The simulations have little effect other than slightly smoothing the field since ground observations are driving the predictions.

4.3. Downscaled pointwise rainfall

Using the output from the RCM as the only source of data in the model we obtained a description of the field that corresponds to the pointwise observations from the estimated values of the us, assuming that all r_{ij} are missing. The results for both months are shown in figure 9. As expected, these plots resemble very closely the ones observed in the left panels

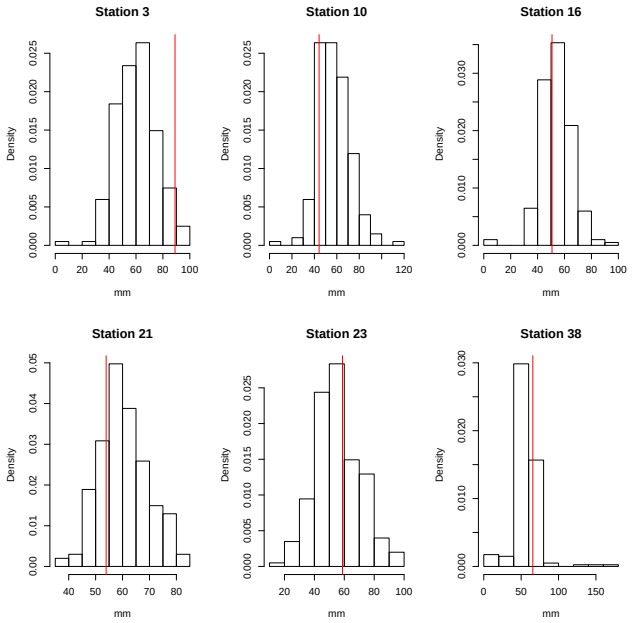

Fig. 4. Histograms of samples of the PPD of ground rainfall for May 1989 at six randomly chosen stations. Vertical lines correspond to actual observed values (not used in the model fit).

of figures 1 and 2 respectively. Comparisons between the observed values and the PPD of the downscaled ClimRAMS outputs are considered in Figures 10 and 11 for the same stations considered in Figures 3 and 4. We observe that, in most cases, the observations correspond to extreme values of the predictive densities. A similar behaviour was observed for the other stations.

5. Discussion and Conclusions

Downscaling methods are recognized as essential tools to get meteorological inputs at scales where most applications in climate impact studies are required. The use of nested RCM driven by GCM outputs is a powerful tool to provide high resolution simulations of climate at a regional scale. In this application we have raised three issues regarding RCM simulations. The first is the need to perform a coherent assessment of the RCM output with observed data. We adopt a predictive viewpoint stating that predictive distributions of the aggregated rainfall, based on ground observations, should be compatible with RCM simulations (bottom up) and ground observations should be compatible with the predictive distribution of downscaled RCM output (top down). Besides, it is possible to quantify the bias between the two sources of information by estimating the parameters of a statistical model blending the two sources of information.

The second issue is that RCM output can be used to guide rainfall predictions on the ground by providing large scale climate information. The last issue is that the comparison between RCM output and ground observations can be used for the downscaling of the

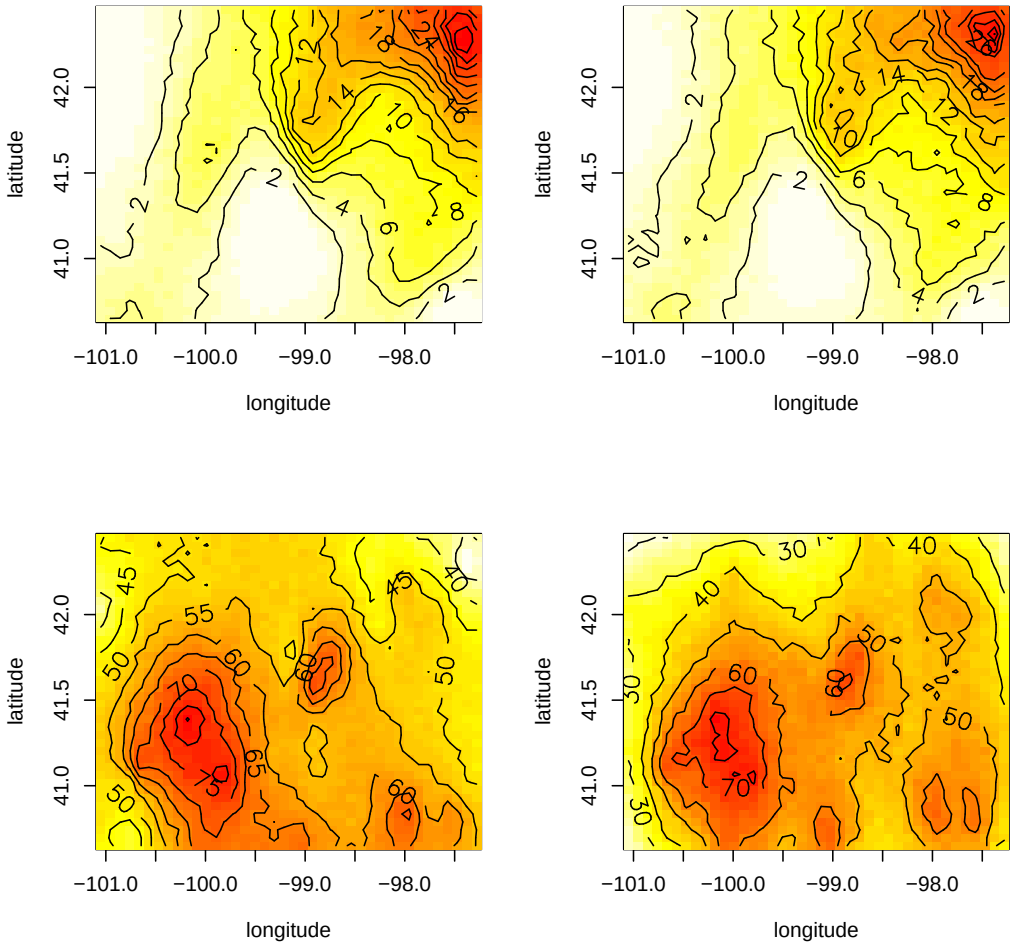

Fig. 5. Posterior median surfaces of rainfall for April and May 1989. Left panels correspond to the model that uses only ground observations. Right panels correspond to the use ground data and ClimRAMS output. Top panels correspond to April and bottom panels to May.

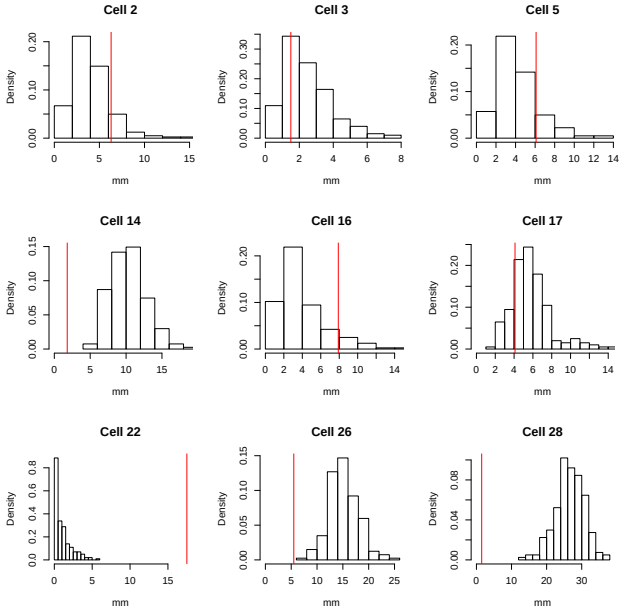

Fig. 6. Histograms of samples of the PPD of aereal rainfall for April 1989 at nine randomly chosen grid cells. Vertical lines correspond to values simulated by the RCM.

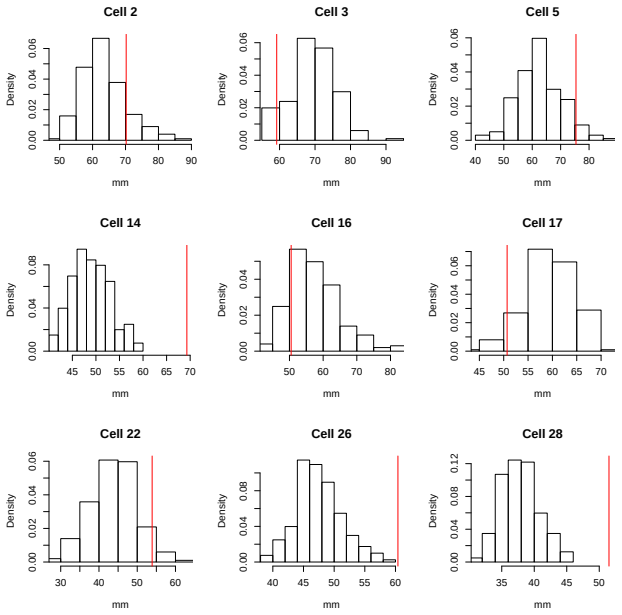

Fig. 7. Histograms of samples of the PPD of aereal rainfall for May 1989 at nine randomly chosen grid cells. Vertical lines correspond to values simulated by the RCM.

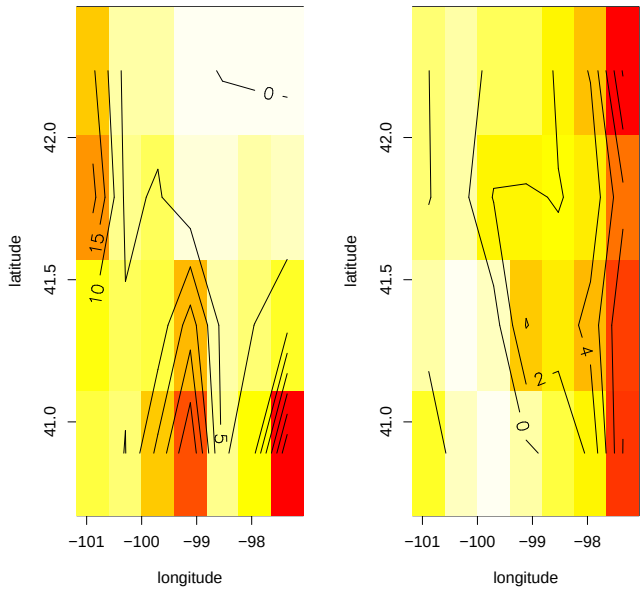

Fig. 8. Standardized differences between ClimRAMS output and estimated posterior aereal rainfall over each of the ClimRAMS cells. Left panel corresponds to April and right panel to May.

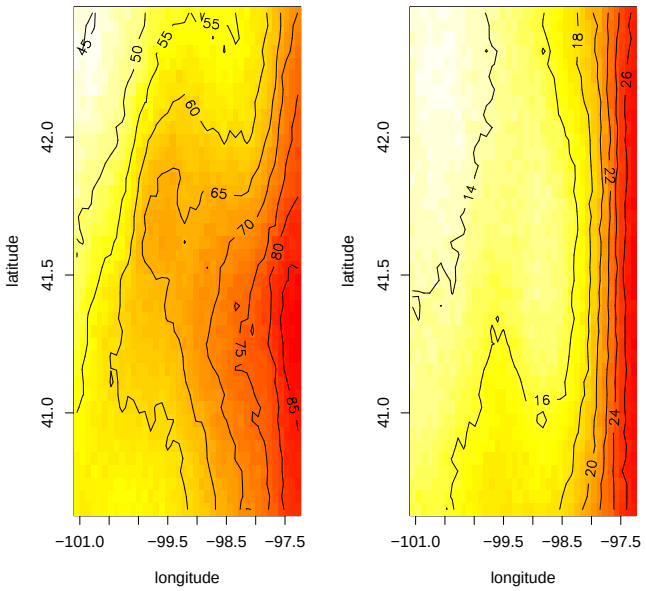

Fig. 9. Predicted pointwise rainfall fields using the RCM output only. Left panel corresponds to May and right panel to April.

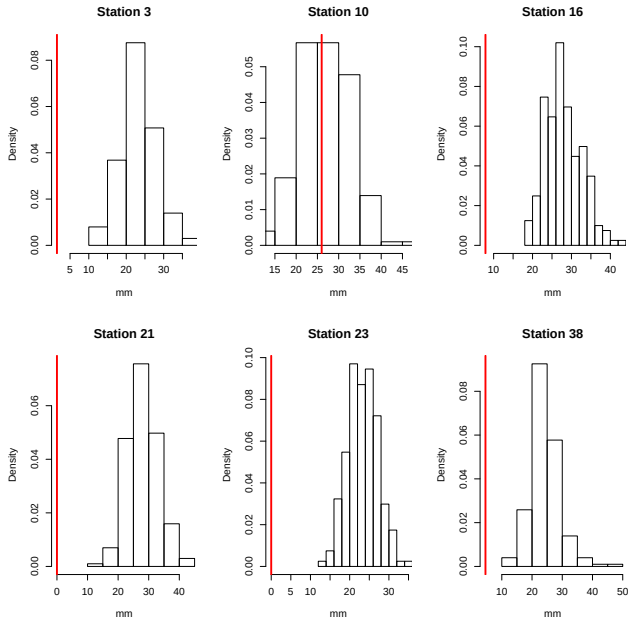

Fig. 10. PPD of ground rainfall downscaled from ClimRAMS simulations for April 1989 at six randomly chosen stations. Vertical lines correspond to actual observed values.

simulations, providing an output with higher spatial resolution. If, for example, a study of the impact of doubling CO2 concentrations on the climate of a given region is performed using a RCM, then a downscaled version of the output can be estimated after obtaining the posterior distribution of the parameters that correspond to current climate conditions.

There are several directions for development of the methods presented in this paper. We have considered data corresponding to one time step (month). A more precise assessment of the climate conditions requires the use of series of observations in time. Our statistical model handles only spatial variability and should be extended to consider space and time. A criticism of the spatial model is also due. We have purposely selected a region with a low geographical complexity and have aggregated the observations monthly. This provides a high level of regularity that allows for an acceptable fit even with a relatively simple spatial model. Elaborations to this model can be considered, but a trade off has to be done with computational complexity and the possibility to capture complex spatial correlations after a non-linear transformation. A second criticism relates to the distributional assumptions for RCM simulations. Equation (4) assumes that conditioning on the approximation of the aereal process and the bias removes all spatial correlations. This is a simplifying assumption upon which elaborations can be made, in particular if several simulations of the RCM corresponding to the same time step are available, something that has become more common with the increase of computing power.

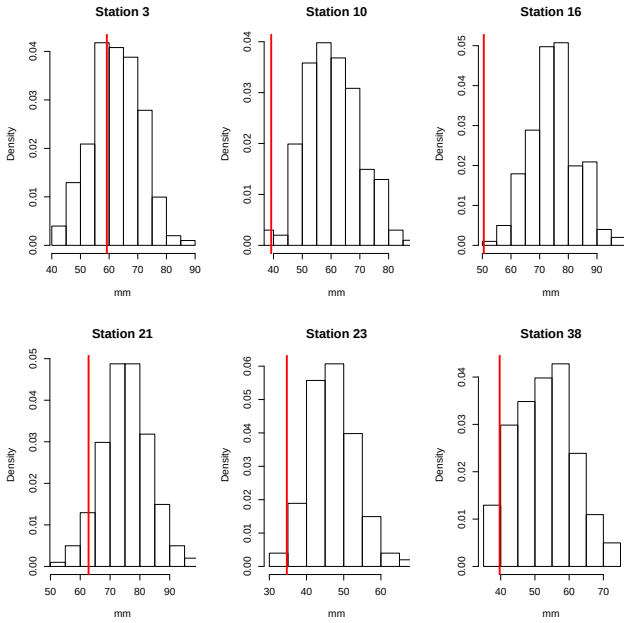

Fig. 11. PPD of ground rainfall downscaled from ClimRAMS simulations for May 1989 at six randomly chosen stations. Vertical lines correspond to actual observed values.

Acknowlegements

We are grateful to Roger Pielke Sr. and Glen Liston for kindly providing the ClimRAMS data. This research was partially funded by grant 97-000592 from Consejo Nacional de Investigaciones Científicas y Teconológicas, Venezuela.

References

Berger, J., De Oliveira, V. and Sansó, B. (2001) Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96**, 1361–1374.

Cressie, N. A. C. (1993) *Statistics for Spatial Data, Revised Edition*. New York: John Wiley and Sons.

Fuentes, M. and Raftery, A. (2001) Model validation and prediction from combined spatial data: via Bayesian melding. *Tech. rep.*, Statistics Department, North Carolina State University.

Gamerman, D. (1997) *Markov Chain Monte Carlo*. London, UK: Chapman and Hall.

Gelfand, A., Zhu, L. and Carlin, B. (2000) On the change of support problem for spatio-temporal data. *Tech. Rep. 2000-011*, School of Public Health, University of Minnesota.

Giorgi, F. and Mearns, L. O. (1999) Introduction to special section: Regional climate modeling revisited. *Journal of Geophysical Research - Atmospheres*, **104 (D6)**, 6335–6352.

Liston, G. E. and Pielke, R. A. (2000) A climate version of the regional atmospheric modeling system. *Journal of Climate*. Submitted.

Mearns, L. O., Bogardi, I., Giorgi, F., Matyasovszky, I. and Pakecki, M. (1999) Comparison of climate change scenarios generated from regional climate model experiments and statistical downscaling. *Journal of Geophysical Research - Atmospheres*, **104 (D6)**, 6603–6622.

Murphy, J. (1999) An evaluation of statistical and dynamical techniques for downscaling local climate. *Journal of Climate*, **12**, 2256–2284.

Richard, J., Kirtman, B., Laprise, R., von Storch, H. and Wergen, W. (2002) Atmospheric regional climate models (RCMs): A multiple purpose tool? *Tech. rep.* Report of the joint WGNE/WGCM ad hoc panel on Regional Climate Modelling `http://w3g.gkss.de/Mitarbeiter/Storch/pdf/RCM.report.040302.pdf`.

Rodríguez-Iturbe, I. and Mejía, J. (1974) On the transformation of point rainfall to aereal rainfall. *Water Resources Research*, **10**, 729–735.

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

Stein, M. (1992) Prediction and inference for truncated spatial data. *Journal of Computational and Graphical Statistics*, **1**, 91–110.

— (1999) *Interpolation of Spatial Data*. New York, USA: Springer-Verlag.

Whittle, P. (1954) On stationary processes in the plane. *Biometrika*, **41**, 434–449.