Brazilian Journal of Probability and Statistics
2016, Vol. 30, No. 4, 653–675
DOI: 10.1214/15-BJPS297

Spatio-temporal dynamic model and parallelized ensemble Kalman filter for precipitation data

Luis Sánchez[a], **Saba Infante**[b], **Victor Griffin**[a] **and Demetrio Rey**[a]

[a]*Carabobo University*
[b]*Technical University of North*

Abstract. This paper presents a spatiotemporal dynamic model which allows Bayesian inference of precipitation states in some Venezuelan meteorological stations. One of the limitations that are reported in digital databases is the reliability of the records and the lack of information for certain days, weeks, or months. To complete the missing data, the Gibbs algorithm, a Markov Chain Monte Carlo (MCMC) procedure, was used. A feature of precipitation series is that their distribution contains discrete and continuous components implying complicated dynamics. A model is proposed based on a discrete representation of a stochastic integro-difference equation. Given the difficulty of obtaining explicit analytical expressions for the predictive posterior distribution, approximations were obtained using a sequential Monte Carlo algorithm called the parallelized ensemble Kalman filter. The proposed method permits the completion of the missing data in the series where required and secondly allows the splitting of a large database into smaller ones for separate evaluation and eventual combination of the individual results. The objective is to reduce the dimension and computational cost in order to obtain models that are able to describe the reality in real time. It was shown that the obtained models are able to predict spatially and temporally the states of rainfall for at least three to four days quickly, efficiently and accurately. Three methods of statistical validation were used to evaluate the performance of the model and showed no significant discrepancies. Speedup and efficiency factors were calculated to compare the speed of calculation using the parallelized ensemble Kalman filter algorithm with the speed of the sequential version. The improvement in speed for four pthread executions was greatest.

1 Introduction

Rainfall data is observed in many locations at many different times and usually has a chaotic behaviour that varies in space and time. This phenomenon can be characterized by complex mathematical models that can be solved by approaches based on simulations of physical models; in practice they are used to predict, forecast, and make inferences about a response variable in a spatial domain. The results of these models can help solve important problems in areas such as agriculture,

Key words and phrases. Precipitation modeling, missing data, spatio temporal models, stochastic integro-difference equation, parallelized ensemble Kalman filter.
Received February 2014; accepted July 2015.

ecology, public health, environmental science, meteorology, hydrology and other scientific areas. Hierarchical dynamic models are appropriate tools to deal with these processes due to their flexibility and because they take into consideration sources of variability such as the correlation that arises between the covariates over time, their frequencies in space at different scales and the interactions between them. These could be spatial series, pure time series without a spatial dimension or dynamic fields in space and time (Fasso and Cameletti (2007)). There are two basic paradigms for modeling spatiotemporal phenomena described by Wikle and Hooten (2006). The first method is descriptive and comes from the classical techniques of geostatistics which consider different ways to model the covariance functions in space and time (e.g., Cressie and Huang (1999), Banerjee, Carlin and Gelfand (2004), Gneiting (2002), Stein (2005), Fernández-Casal, González-Manteiga and Febrero-Bande (2003), Jones and Zhang (1997) and Ma (2003)). The second method is dynamic and takes into account the variability that cannot be explained by the covariates. They use a variable with a dependency structure in the form of a dynamic model represented in state space form. There is an extensive literature on these models (e.g., Wikle and Hooten (2006), Cressie and Wikle (2011), Sigrist, Künsch and Stahel (2012), Cocchi, Greco and Trivisano (2007), Cameletti, Ignaccolo and Bande (2010), Cameletti et al. (2012) and Sahu (2011)).

In the context of Bayesian inference, some hierarchical models for environmental data have been implemented through MCMC methods. For example, see the work of Sigrist, Künsch and Stahel (2012), Lasinio, Sahu and Mardia (2007), Sahu, Yip and Holland (2011), Sahu (2011), Banerjee and Fuentes (2011), Calder et al. (2011), Cressie and Wikle (2011), Amisigo and van de Giesen (2005), Sansó and Guenni (1999a, 1999b, 2000), Hernández, Guenni and Sansó (2009, 2011) and Stroud et al. (2010). The tuning of these models involve matrix operations for which the computational cost of each iteration and the number of iterations of the MCMC algorithms increase dramatically with the size of the data sets. To get around this problem, we propose the use of a recursive filtering algorithm based on sequential Monte Carlo techniques (SMC). A filtering algorithm is the process used to obtain the best statistical estimate of a system modeled from partial observations of the true signal from nature (Majda and Harlim (2012)). The technique consists of a system of two predictor-corrector steps, adjusting the a priori estimate and updating the posterior distribution using Bayes theorem. The updated estimate is introduced into the model as an initial condition for the future forecast. This approach is known as a data assimilation technique. In the data assimilation community, there are many well developed algorithms based on hierarchical Bayesian spatio-temporal modeling (see Berliner, Milliff and Wikle (2003), Cressie and Wikle (2011) and Sánchez and Infante (2013) among others). The general objective of this work is to propose a dynamic model to filter the available information in real time and predict, spatial and/or temporarily, the distribution of rainfall signals across partially observed samples. A first objective is to complete the missing data using a statistical technique to augment data. A second objective is to write

the model in state space form and parameterize it for a set of weighted base functions where the state equation represents the unknown system and the observation equation represents the partially observed measurements. A third objective is to implement a parallelized version of the ensemble Kalman Filter (Evensen (2009)), to estimate and predict the weather states.

The rest of the paper is organized as follows. In Section 2, the spatio-temporal hierarchical model, the rainfall model, and the procedure to generate the missing data are defined. In Section 3, the computational scheme of the ensemble Kalman filter is defined, and the parallelization method is described. In Section 4, validation criteria are given. Section 5 shows the results and Section 6 provides discussions and conclusions.

2 Spatio temporal hierarchical models

Consider a real-valued spatio-temporal process

$$\{y_t(s) : s = (x, y)^T \in \mathbf{s} \subset \mathbf{R}^2\}, \tag{2.1}$$

where $\mathbf{s}$ is the spatial domain under study that can be finite or countably infinite. A discretized version of the process can be represented as $\{y_1(s_1), \ldots, y_t(s_i), \ldots\}$, where $y_t(s_i)$ represents the value of an underlying scientific process at time t and location s_i, $t = 1, \ldots, T$, $i = 1, \ldots, n$. Let $\mathbf{y}_t = (y_t(s_1), \ldots, y_t(s_n))^T$. Let $\mathbf{w}_t = (w_t(r_1), \ldots, w_t(r_m))^T$ be the vector containing the observed data values at spatial locations r_j at time t, for $j = 1, \ldots, m$. The two sets of spatial locations, $\{s_i : i = 1, \ldots, n\} \subseteq \mathbf{s}$ and $\{r_j : j = 1, \ldots, m\} \subseteq \mathbf{s}$, need not be the same. We are interested in predicting $\mathbf{y}_t$ the unobserved process based in $\mathbf{w}_t$ the observed process. The state-space representation for the prediction can be written as a model of the form:

$$\mathbf{y}_t = \mathcal{M}_t(\mathbf{y}_{t-1}) + \mathbf{u}_t; \qquad \mathbf{u}_t \sim N(\mathbf{0}, \mathbf{Q_t}), \tag{2.2}$$

$$\mathbf{w}_t = \mathcal{H}_t(\mathbf{y}_t) + \mathbf{v}_t; \qquad \mathbf{v}_t \sim N(\mathbf{0}, \mathbf{R_t}), \tag{2.3}$$

$$\mathbf{y}_0 \sim N(\boldsymbol{\mu}_0, \mathbf{P}_0), \tag{2.4}$$

where (2.2) is the state equation, (2.3) is called the measurement equation, and (2.4) is the initial state equation. $\mathbf{u}_t = (u_t(s_1), \ldots, u_t(s_n))^T$, denotes the random noise which is spatially colored, temporally white and Gaussian with mean zero and a covariance matrix $\mathbf{Q}_t$. $\mathcal{M}_t(\cdot)$ is a nonlinear operator mapping the state space into itself. Let $\mathcal{H}_t$ be the transition matrix that maps the state space into the observation space at time t. Let $\mathbf{v}_t = (v_t(r_1), \ldots, v_t(r_m))^T$ be the random measurement noise which has zero mean, is uncorrelated in time, and has Gaussian covariance matrix $\mathbf{R}_t$. The main goal is to estimate the hidden states $(y_t(s_1), \ldots, y_t(s_n))^T$, given observed measures $(w_t(r_1), \ldots, w_t(r_m))^T$.

The aim of the state estimation problem utilizing the Bayesian approach is to find the joint posterior distribution of the states $\mathbf{y}_{1:T}$ given the observations $\mathbf{w}_{1:T}$, denoted by $P(\mathbf{y}_{1:T}|\mathbf{w}_{1:T})$ or the conditional probability density function (filtered distribution) $P(\mathbf{y}_t|\mathbf{w}_{1:T})$, where: $\mathbf{y}_{1:T} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)^T$, and $\mathbf{w}_{1:T} = (\mathbf{w}_1, \ldots, \mathbf{w}_T)^T$. There are three types of estimation that can be distinguished within the state estimation problem according to the relation between the time instants t_k and t_T; the filtering problem with $k = T$, the prediction problem with $k > T$, and the smoothing problem with $k < T$. We are interested in finding estimators such as $E(\mathbf{y}_t|\mathbf{w}_1, \ldots, \mathbf{w}_{k=T})$, $E(\mathbf{y}_t|\mathbf{w}_1, \ldots, \mathbf{w}_{k>T})$, $E(\mathbf{y}_t|\mathbf{w}_1, \ldots, \mathbf{w}_{k<T})$, or the maximum posterior probability (MAP) estimator, among others (see West and Harrinson (1997), Cressie and Wikle (2011) or Sánchez and Infante (2013) for details). In practice, the calculation of the posterior distribution has complications when the dimensionality of the system increases, requiring efficient computational strategies to approximate it. This article proposes the implementation of the ensemble Kalman filter algorithm (EnKF) (Evensen (1994)) using a parallel programming approach to estimate the dynamics of rainfall in some weather stations of Venezuela.

2.1 Model for rainfall

This dynamic temporal-space hierarchical model is now used for the daily rainfall series or monthly averages of some weather stations of Venezuela. A typical feature of the precipitation is that its distribution is skewed. It consists of a discrete component indicating the occurrence of precipitation and a continuous component specifying the amount. The continuous and the discrete part are either modelled separately or together (Sansó and Guenni (2000), Hernández, Guenni and Sansó (2009), Sigrist, Künsch and Stahel (2012)).

The rainfall model given in (Sigrist, Künsch and Stahel (2012)) establishes that the process $y_t(s)$ at time t on site s depends on a latent normal variable $w_t(s)$ through

$$y_t(s) = \begin{cases} 0, & \text{if } w_t(s) \leq 0, \\ w_t(s)^{\lambda}, & \text{if } w_t(s) > 0, \end{cases}$$

where $\lambda > 0$. A power transformation is needed since rainfall is skewed and does not follow a truncated normal distribution. The latent variable $w_t(s)$ can be interpreted as a precipitation potential.

An alternative way of expressing the previously given rainfall model, is to use the space–time representation of the model based upon a stochastic integro-difference equation (IDE) developed by Wikle et al. (2001), Wikle (2002), Xu, Wikle and Fox (2005), Dewar (2007), Dewar, Scerri and Kadirkamanathan (2011), Scerri et al. (2011), Scerri, Dewar and Kadirkamanathan (2009), Wikle and Hooten (2006), Wikle and Holan (2011) and Calder et al. (2011) among others. The integro-difference linear equation of first order is defined as

$$z_t(s) = \int_{\mathbf{s}} K(s, r) z_{t-1}(r)\, dr + e_t(s), \tag{2.5}$$

where t denotes time and $s, r \in \mathbf{s}$ denote spatial locations in the n-dimensional spatial region under investigation. The spatial field at time t and location s denoted by $z_t(s)$ is related to the previous field via the convolution integral (2.5). The Kernel of the integral, $K(s,r) : \mathbf{R}^n \to \mathbf{R}$, is known as the spatial mixture. The spatial field is subject to disturbance $e_t(s)$, a normally distributed, zero-mean white noise process such that $e_t(s) \sim N(0, \sigma^2)$, for all t, s, with covariance defined by

$$\operatorname{cov}\big(e_t(s), e_{t+\tau}(r)\big) = \begin{cases} \sigma^2 \delta(s-r), & \tau = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{2.6}$$

for all $\tau \in \mathbf{Z}$, δ denotes the Dirac delta function.

Given that the system $z_t(s)$ is unknown, consider that the dynamic field is observed at every time instant t at m fixed spatial locations by

$$\mathbf{w}_t = \mathbf{z}_t + \boldsymbol{\eta}_t, \tag{2.7}$$

where $\mathbf{w}_t = (w_t(s_1), \ldots, w_t(s_m))^T$, $\mathbf{z}_t = (z_t(s_1), \ldots, z_t(s_m))^T$, and $\boldsymbol{\eta}_t = (\eta_t(s_1), \ldots, \eta_t(s_m))^T$, with $\eta_t(s_i) \sim N(0, \Sigma_\eta)$ for $i = 1, \ldots, m$.

The system given in (2.5) and (2.7) can be approximated in state-space form using a set of weighted basis functions (Dewar (2007)),

$$\begin{aligned} \mathbf{z}_t(s) &= \boldsymbol{\phi}(s)^T \mathbf{y}_t(s), \\ K(s,r) &= \mathbf{a}^T \boldsymbol{\psi}(s,r), \end{aligned} \tag{2.8}$$

where the unknown parameter vector $\mathbf{a} \in \mathbf{R}^{n_a}$ weights the kernel basis functions $\boldsymbol{\psi} : \mathbf{R}^{n_s} \to \mathbf{R}^{n_a}$ and the state vector $\mathbf{y}_t \in \mathbf{R}^n$ weights the vector of the field basis functions $\boldsymbol{\phi} : \mathbf{R}^{n_s} \to \mathbf{R}^n$ at time t. It is required that the sets of basis functions $\boldsymbol{\psi}(s,r) = (\psi_1(s,r), \ldots, \psi_{n_a}(s,r))^T$ and $\boldsymbol{\phi}(s) = (\phi_1(s), \ldots, \phi_{n_y}(s))^T$ are linearly independent.

Using Lemma 4 given in (Dewar (2007)), for a white noise process $e_t(s)$ with covariance as defined in (2.6), the integral

$$\boldsymbol{\lambda}_t = \int_{\mathbf{s}} \boldsymbol{\phi}(s) e_t(s)\, ds \tag{2.9}$$

is vector valued with $E(\boldsymbol{\lambda}_t) = 0$, and $\operatorname{cov}(\boldsymbol{\lambda}_t) = E(\boldsymbol{\lambda}_t \boldsymbol{\lambda}_t^T) = \sigma^2 \boldsymbol{\psi}_y$, where $\boldsymbol{\psi}_y = \int_{\mathbf{s}} \boldsymbol{\phi}(s) \boldsymbol{\phi}(s)^T\, ds$.

The spatio-temporal IDE model defined by the equations (2.5), (2.7), and decomposed as in (2.8) can be written as a parameterized model in state-space of the form (Dewar (2007))

$$\begin{aligned} \mathbf{y}_{t+1} &= \mathbf{A}\mathbf{y}_t + \mathbf{u}_t, \qquad & \mathbf{u}_t &\sim N\big(\mathbf{0}, \sigma^2 \boldsymbol{\psi}_y^{-1}\big), \\ \mathbf{w}_t &= \mathbf{C}\mathbf{y}_t + \mathbf{v}_t, \qquad & \mathbf{v}_t &\sim N\big(\mathbf{0}, \sigma_v^2 \mathbf{I}\big), \end{aligned} \tag{2.10}$$

where $\mathbf{A} \in \mathbf{R}^{n_y \times n_y}$ is constructed using an orthogonal set of basis functions to facilitate computation and dimension reduction. The matrix $\mathbf{A}$ is obtained as $\mathbf{A} = \Psi_y^{-1}\Psi_a$, where

$$\Phi_a(s) = \int_s \boldsymbol{\psi}(s,r)\boldsymbol{\phi}^T(r)\,dr,$$
$$\Psi_y = \int_s \boldsymbol{\phi}(s)\boldsymbol{\phi}^T(s)\,ds,$$
$$\Psi_a = \int_s \boldsymbol{\phi}(s)\mathbf{a}^T\boldsymbol{\Phi}_a(s)\,ds.$$

The observation matrix, $\mathbf{C} = (\phi(s_1), \phi(s_2), \ldots, \phi(s_{n_y}))^T$, can be used to indicate the location of the stations where each $\phi(s)$ represents a Gaussian radial basis function of the form:

$$\big[\phi(s)\big]_j \approx \exp\left\{-\frac{(s - [\mu_{\text{fbr}}]_j)^2}{\sigma^2_{\text{fbr}}}\right\},$$

where $[\mu_{\text{fbr}}]_j$, and σ^2_{fbr} are the mean and variance of the jth radial basis function, and s is the spatial location. The Kernel $\psi(s,r)$, is represented by

$$\big[\psi(s,r)\big]_i \approx \exp\left\{\frac{-(r - (s + [\mu_{\text{kfbr}}]_i))^2}{\sigma^2_{\text{kfbr}}}\right\},$$

where $[\mu_{\text{kfbr}}]_i$ and σ^2_{kfbr} represent the mean and variance of the ith radial basis function that defines the Kernel, and where r and s are different spatial locations. We consider a spatial mixing Kernel with Gaussian basis functions because they are invariant across space and time (Dewar, Scerri and Kadirkamanathan (2011)). Considering that $\phi_i(s) \sim N(\mu_i, \mathbf{C}_i)$, $\phi_j(s) \sim N(\mu_j, \mathbf{C}_j)$, we have

$$\Psi_y = \Psi_{ij} = \int \phi_i(s)\phi_j(s)\,ds \approx \pi^{n/2}|\mathbf{C}_{ij}^{-1}|^{1/2}\exp\{-r_{ij}\},$$

where

$$\mathbf{C}_{ij} = [\mathbf{C}_i + \mathbf{C}_j],$$
$$r_{ij} = (\mu_i - \mu_j)^T(\mathbf{C}_i + \mathbf{C}_j)^{-1}\mathbf{C}_i\mathbf{C}_j(\mu_i - \mu_j)$$

and

$$\Phi_a(s) = \int \big[\psi(s,r)\big]_i\big[\phi(r)\big]_j\,dr$$
$$\approx \frac{\pi^{1/2}}{|\sigma^2_{\text{fbr}} + \sigma^2_{\text{kfbr}}|^{1/2}}\exp\left\{-\frac{(s - [\mu_{\text{fbr}}]_j - [\mu_{\text{kfbr}}]_i)^2}{\sigma^2_{\text{fbr}} + \sigma^2_{\text{kfbr}}}\right\},$$

which represents the Kernel of a normal distribution with mean $[\mu_{\text{fbr}}]_j - [\mu_{\text{kfbr}}]_i$ and variance $\sigma^2_{\text{fbr}} + \sigma^2_{\text{kfbr}}$.

The hierarchical space temporal model proposed in this article is a combination of the rainfall model given in (Sigrist, Künsch and Stahel (2012)) and the following model proposed in (Dewar (2007)):

$$\begin{aligned} \theta_0 &\sim N(\mu_0, \Sigma_0), \\ \boldsymbol{\theta}_{t+1} &= \mathbf{A}\boldsymbol{\theta}_t + \mathbf{u}_t, \qquad \mathbf{u}_t \sim N\bigl(\mathbf{0}, \sigma^2 \boldsymbol{\psi}_\theta^{-1}\bigr), \\ \mathbf{w}_t &= \mathbf{C}\boldsymbol{\theta}_t + \mathbf{v}_t, \qquad \mathbf{v}_t \sim N\bigl(\mathbf{0}, \sigma_v^2 \mathbf{I}\bigr), \\ \mathbf{y}_t &= \bigl[\max(0, \mathbf{w}_t)\bigr]^\lambda, \end{aligned} \tag{2.11}$$

where the matrices $\mathbf{A}$ and $\mathbf{C}$ are obtained using the methodology given in (Dewar (2007)) and the maximum and power in the last equation are for each spatial component of the vector. The IDE model describes the spatio-temporal dynamics in discrete time and continuous space. The evolution is governed by a kernel mixture whose form describes the dynamical nature of the system response. It is an attractive representation in the spatio-temporal context since it simultaneously interpolates spatially and predicts temporarily permitting the consideration of unknown fields in unobserved points of interest. This technique has been used in situations such as bacteria growth, wave propagation, population growth, dispersion models, changes in the spatial patterns of real estate prices and complex models for studying climatic change (see Scerri, Dewar and Kadirkamanathan (2009), Wikle and Holan (2011) and Calder et al. (2011)). It provides tools for estimation and prediction using large data sets in space–time facilitating computation and dimension reduction.

2.2 Generation of missing data with the Gibbs sampler

Since the Venezuelan meteorological stations data base has missing data, it is necessary first of all to simulate the missing data using the Gibbs algorithm. Once the data is complete, the model proposed in (2.11) will be implemented. The Gibbs sampler for sampling from a truncated multivariate normal can be described as follows.

A random variable $\mathbf{w}$ is said to follow a truncated multivariate normal distribution ($\mathbf{w} \sim TN_d(\mu, \Sigma; a, b)$) subject to linear inequality constraints if its probability density function is

$$f_w(w, \mu, \Sigma, a, b) = \frac{\exp\{-(1/2)(w-\mu)^T \Sigma^{-1}(w-\mu)\}}{\int_a^b \exp\{-(1/2)(w-\mu)^T \Sigma^{-1}(w-\mu)\}\, dw} I_{\{a \le w \le b\}}.$$

Suppose that $\mathbf{w}_t = \mathbf{w}$, $\mathbf{w}_t^{\text{observed}} = \mathbf{w}_{\text{obs}}$ is observed data and $\mathbf{w}_t^{\text{missing}} = \mathbf{w}_{\text{miss}}$ is unobserved data. Then using the Proposition 2 given in Li and Ghosh (2013), we partition $\mathbf{w}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ as

$$\mathbf{w} = \begin{pmatrix} \mathbf{w}_{\text{obs}} \\ \mathbf{w}_{\text{miss}} \end{pmatrix}, \qquad \mu = \begin{pmatrix} \boldsymbol{\mu}_{\text{obs}} \\ \boldsymbol{\mu}_{\text{miss}} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{\text{obs,obs}} & \Sigma_{\text{obs,miss}} \\ \Sigma_{\text{miss,obs}} & \Sigma_{\text{miss,miss}} \end{pmatrix},$$

where $\mathbf{w}_{\text{obs}}$ is a random vector of dimension d_1, $\mathbf{w}_{\text{miss}}$ is a random vector of dimension d_2, and $d_1 + d_2 = d$. Using the fact that the conditional density of a multivariate normal distribution is also a multivariate normal, then the conditional distribution $\mathbf{w}_{\text{miss}}|\mathbf{w}_{\text{obs}}$ has a truncated normal distribution given by

$$\mathbf{w}_{\text{miss}}|\mathbf{w}_{\text{obs}} \sim TN\big(\mu_{\text{miss|obs}}, \Sigma_{\text{miss|obs}}, \mathbf{R}(w_{\text{obs}})\big),$$

where

$$\boldsymbol{\mu}_{\text{miss|obs}} = \mu_{\text{miss}} + \Sigma_{\text{miss,obs}}\Sigma_{\text{obs,obs}}^{-1}(w_{\text{obs}} - \mu_{\text{obs}}),$$

$$\boldsymbol{\Sigma}_{\text{miss|obs}} = \Sigma_{\text{miss,miss}} - \Sigma_{\text{miss,obs}}\Sigma_{\text{obs,obs}}^{-1}\Sigma_{\text{obs,miss}}$$

and

$$\mathbf{R}(w_{\text{obs}}) = \big\{\mathbf{w}_{\text{miss}} \in \mathbf{R}^{d_2} : a \leq \mathbf{R}(w_{\text{obs}}, \mathbf{w}_{\text{miss}}) \leq b\big\}.$$

The main goal is to generate independent random variables with the Gibbs sampler, sampling from a truncated normal distribution (Robert (1995), Kotecha and Djuric (1999), Wilhelm (2013), Li and Ghosh (2013)). We can then construct a Markov Chain which continuously draws from $f(w_{-i}|w_i) = f(w_i|w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_d)$ subject to $a_i \leq w_i \leq b_i$.

Considering $\mathbf{w}_{\text{obs}} = w_i$, $\mathbf{w}_{\text{miss}} = w_{-i}$, $\boldsymbol{\mu}_{\text{obs}} = \mu_i$, $\boldsymbol{\mu}_{\text{miss}} = \mu_{-i}$, $\Sigma_{\text{obs,obs}} = \Sigma_{i,i}$, $\Sigma_{\text{obs,miss}} = \Sigma_{i,-i}$, $\Sigma_{\text{miss,obs}} = \Sigma_{-i,i}$, and $\Sigma_{\text{miss,miss}} = \Sigma_{-i,-i}$, then $\mathbf{w}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given by

$$\mathbf{w} = \begin{pmatrix} w_i \\ w_{-i} \end{pmatrix} \sim N(\mu, \Sigma), \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_i \\ \mu_{-i} \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{i,i} & \Sigma_{i,-i} \\ \Sigma_{-i,i} & \Sigma_{-i,-i} \end{pmatrix}$$

then the distribution of w_{-i} conditional on w_i is normal $(w_{-i}|w_i) \sim N(\mu_{-i|i}, \Sigma_{-i|i})$, with mean

$$\mu_{-i|i} = \mu_{-i} + \Sigma_{-i,i}\Sigma_{i,i}^{-1}(w_i - \mu_i) \tag{2.12}$$

and variance

$$\Sigma_{-i|i} = \Sigma_{-i,-i} - \Sigma_{-i,i}\Sigma_{i,i}^{-1}\Sigma_{i,-i} = H_{ii}^{-1}. \tag{2.13}$$

Let $w^{(j)}$ denote the sample drawn at the jth MCMC iteration. The steps of the Gibbs sampler for generating n samples $w^{(1)}, \ldots, w^{(n)}$ are:

Step 1. Since the conditional variance $\Sigma_{-i|i}$ is independent from the observation $w_{-i}^{(j)}$, we can precalculate it before running the Markov chain.

Step 2. Choose a starting value $\mathbf{w^{(0)}} = (w_1^{(0)}, \ldots, w_d^{(0)})$ of the Markov chain.

Step 3. In each round $j = 1, \ldots, n$, we go from $i = 1, \ldots, d$, and sample from the conditional density

$$w_{-i}^{(j)}|w_1^{(j)}, \ldots, w_{i-1}^{(j-1)}, w_{i+1}^{(j-1)}, \ldots, w_d^{(j-1)} \sim N(\mu_{-i|i}, \Sigma_{-i|i}),$$

where the mean and variance of the normal distribution are given by (2.12), and (2.13).

Step 4. Draw a uniforme random variate $u \sim \text{Unif}(0, 1)$.

Step 5. Draw a normal random variate $y \sim N(\mu, \sigma^2)$ and a univariate truncated random variate $w \sim TN(\mu, \sigma^2, a, b)$. For each realization y we can find a w such as $P(Y \leq y) = P(W \leq w)$,

$$\frac{\Phi((w-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} = \Phi\left(\frac{y-\mu}{\sigma}\right) = u, \tag{2.14}$$

where $\Phi(\cdot)$ denotes its cumulative standard normal distribution function.

Step 6. Draw w_{-i} from conditional univariate truncated normal distribution ($w_{-i} \sim TN(\mu_{-i|i}, \Sigma_{-i|i}, a_i, b_i)$), using the inverse method

$$w_{-i} = \mu_{-i|i} + \sigma_{-i|i}\Phi^{-1}\left\{u\left[\Phi\left(\frac{b_i - \mu_{-i|i}}{\sigma_{-i|i}}\right) - \Phi\left(\frac{a_i - \mu_{-i|i}}{\sigma_{-i|i}}\right)\right] + \Phi\left(\frac{a_i - \mu_{-i|i}}{\sigma_{-i|i}}\right)\right\}.$$

3 Ensemble Kalman filter

The ensemble Kalman filter (EnKF) (Evensen (2009, 1994, 2003), Evensen and van Leeuwen (1996)), is a sequential Monte Carlo algorithm used to approximate the forecasting and filtering distributions in nonlinear high-dimensional state-space models. The state distribution is represented at each time period by an equally weighted sample of states. The ensemble is propagated forward through time using the equation (2.2) and is updated using the equation (2.3) when new data arrive. The algorithm proceeds as follows. Let $\{\theta^a_{t,i}, i = 1, \ldots, n\}$ and $\{\theta^b_{t,i}, i = 1, \ldots, n\}$ denote the forecast and filtered ensemble at time t, respectively. Let $\hat{\theta}^b_t$ and $\hat{P}^b_t$ denote the mean and covariance from the state filtered distribution. The algorithm is initialized at time $t = 0$ by drawing $\theta^a_{0,i} \sim N(\hat{\theta}_0, \hat{P}^b_0)$, for $i = 1, \ldots, n$. The ensemble is then propagated forward through time, alternating between the forecast and update. Starting with the filtering ensemble at time $t - 1$, the one-step-ahead forecasts at time t are obtained using the evolution equation. The algorithm is summarized as follows:

Step 1. *Propagation*. We begin by creating n initial ensemble members, say $\theta^a_{0,i}$ drawn from the normal distribution $\{\theta^a_{0,i} \sim N(\hat{\theta}_0, \hat{P}^b_0), i = 1, \ldots, n\}$. This is accomplished by first factoring $\hat{P}^b_0 = S_0 S_0^T$, and defining,

$$\theta^a_{0,i} = \hat{\theta}_0 + S_0 n^i_0, \qquad i = 1, \ldots, n, \text{ where } n^i_0 \sim N(0, I). \tag{3.1}$$

The ensemble mean is given by

$$\hat{\theta}^b_0 = \frac{1}{n}\sum_{i=1}^{n} \theta^a_{0,i}. \tag{3.2}$$

Similarly, the ensemble covariance is given by

$$\hat{P}_0^b = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{0,i}^a - \hat{\theta}_0^b)(\theta_{0,i}^a - \hat{\theta}_0^b)^T. \tag{3.3}$$

Step 2. *Ensemble forecast*. Inductively consider the time instant t. Given $(\hat{\theta}_t^b, \hat{P}_t^b)$, let $\hat{P}_t^b = \hat{S}_t\hat{S}_t^T$. Create an ensemble

$$\theta_{t,i}^a = \hat{\theta}_t^b + \hat{S}_t n_t^i, \qquad i = 1, \ldots, n, \text{ where } n_t^i \sim N(0, I). \tag{3.4}$$

The n members of the ensemble forecast at time $t+1$ are generated

$$\theta_{t+1,i}^b = \mathcal{M}_t(\theta_{t,i}^a) + u_{t+1}^i, \qquad i = 1, \ldots, n, u_{t+1}^i \sim N(0, Q_{t+1}). \tag{3.5}$$

The unbiased estimator of the sample mean is then given by

$$\hat{\theta}_{t+1}^b = \frac{1}{n}\sum_{i=1}^{n}\theta_{t+1,i}^b. \tag{3.6}$$

The forecast error is estimated by

$$e_{t+1}^i = \theta_{t+1,i}^b - \hat{\theta}_{t+1}^b. \tag{3.7}$$

The unbiased estimator of the sample covariance is then given by

$$\hat{P}_{t+1}^b = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{t+1,i}^b - \hat{\theta}_{t+1}^b)(\theta_{t+1,i}^b - \hat{\theta}_{t+1}^b)^T + Q_{t+1}. \tag{3.8}$$

In practice, it is not common to approximate $\hat{P}_t^b$; instead

$$\hat{P}_{t+1}^{\mathrm{cr}} = \frac{1}{n-1}\sum_{i=1}^{n}(\theta_{t+1,i}^b - \hat{\theta}_{t+1}^b)\big[\mathcal{H}_{t+1}(\theta_{t+1,i}^b) - \mathcal{H}_{t+1}(\hat{\theta}_{t+1}^b)\big]^T \tag{3.9}$$

and

$$\begin{aligned}\hat{P}_{t+1}^{\mathrm{pr}} = \frac{1}{n-1}\sum_{i=1}^{n}&\big[\mathcal{H}_{t+1}(\theta_{t+1,i}^b) - \mathcal{H}_{t+1}(\hat{\theta}_{t+1}^b)\big]\\ &\times\big[\mathcal{H}_{t+1}(\theta_{t+1,i}^b) - \mathcal{H}_{t+1}(\hat{\theta}_{t+1}^b)\big]^T\end{aligned} \tag{3.10}$$

are estimated where $\hat{P}_{t+1}^{\mathrm{cr}}$ is the (sample) cross covariance between the background ensemble and its predicted projection onto the observation space, while $\hat{P}_{t+1}^{\mathrm{pr}}$ is the (sample) covariance of the predicted projection of the background ensemble onto the observation space.

Step 3. *Data assimilation*. If observations are available at time t, then we update the ensemble using the perturbed observations algorithm as described above (Stroud et al. (2010)). We first generate synthetic observations from the measurement equation

$$w_{t+1}^i = \mathcal{H}_{t+1}(\theta_{t+1,i}^b) + v_{t+1}^i, \qquad i = 1, \ldots, n, v_{t+1}^i \sim N(0, R_{t+1}). \tag{3.11}$$

Then, w_{t+1}^i are updated with the rainfall model given in (Sigrist, Künsch and Stahel (2012))

$$y_{t+1}^i = \left[\max(0, w_{t+1}^i)\right]^\lambda, \qquad i = 1, \ldots, n. \tag{3.12}$$

This provides samples from the joint state and observation forecast distribution $P(\theta_t, y_t | y_{1:t-1})$. The update is completed using Bayes linear fit,

$$\hat{\theta}_{t+1}^a = \hat{\theta}_{t+1}^b + \mathbf{K}_{t+1}(w_{t+1} - y_{t+1}^i) \tag{3.13}$$

and

$$\hat{P}_{t+1}^a = \hat{P}_{t+1}^b - \mathbf{K}_{t+1}(\hat{P}_{t+1}^{\mathrm{cr}})^T, \tag{3.14}$$

where

$$\mathbf{K}_{t+1} = \hat{P}_{t+1}^{\mathrm{cr}}(\hat{P}_{t+1}^{\mathrm{pr}} + R_{t+1})^{-1} \tag{3.15}$$

is the Kalman gain matrix.

4 Criteria for validation of models

To compare the quality of predictions and forecasts obtained from the fitted model, we use validation criteria as in (Bakar (2011)). These validation criteria are:

1. Root mean squared error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\widehat{y}_t^j - y_t^j)^2} \tag{4.1}$$

2. Mean absolute error

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n}|(\widehat{y}_t^j - y_t^j)| \tag{4.2}$$

3. Relative bias

$$\mathrm{RB} = \frac{1}{n\overline{y}}\sum_{j=1}^{n}(\widehat{y}_t^j - y_t^j), \tag{4.3}$$

where

- y_t^j is the true state for the jth simulation, $j = 1, \ldots, n$.
- $\widehat{y}_t^j$, is the predicted posterior mean of the adjusted data of day t for the jth signal, $j = 1, \ldots, n$.
- $\overline{y}$ is the arithmetic mean of the observations.

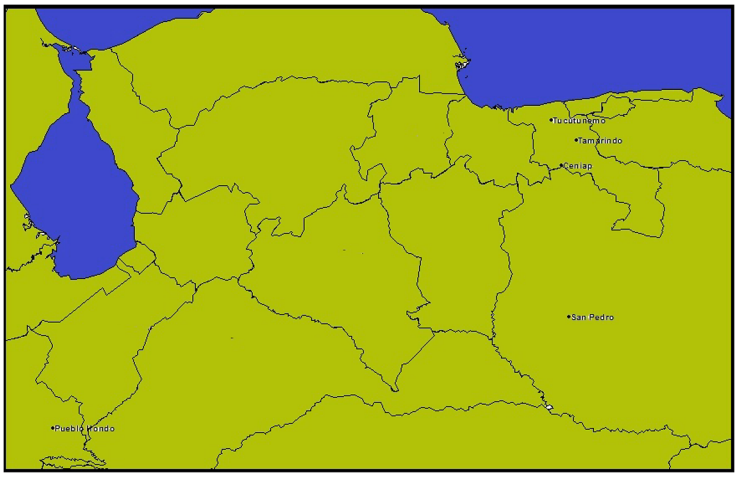

Figure 1 *Weather stations*: *Aragua*, *Guárico and Táchira*.

5 Results

To illustrate the methodology proposed in this paper, we consider series of daily mean precipitation from January 2011, until May 2012 for five weather stations in Venezuela; three located in Aragua state (Ceniap, Tamarindo, and Tucutunemo), one in the Guárico state (San Pedro) and another located in Táchira state (Hondo Pueblo) (see Figure 1 for station locations). The data is available in http://agrometeorologia.inia.gob.ve/. The Gibbs algorithm, as defined in Section 2.2, was implemented to generate missing data of daily precipitation; it was programmed in one Intel Core i7 CPU 3.6 GHz machine with 16 GB RAM running 64 Bit Debian Linux, using the ANSI **C** programming environment. To calibrate the parameters of the Gibbs algorithm, the first 400 samples were discarded; the samples were generated from Time $= 401$ up to Time $= 890$ to complete the data. The total execution time of the Gibbs algorithm was 405,062 μs for the five weather stations.

The prior distributions generate truncated normal random variables using the following initial parameters:

- The prior mean

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

- The prior covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 100 & 0.1 \\ 0.1 & 100 \end{pmatrix}.$$

- The lower and upper limits for the truncated Normal $a = 0$, $b = 10{,}000$.

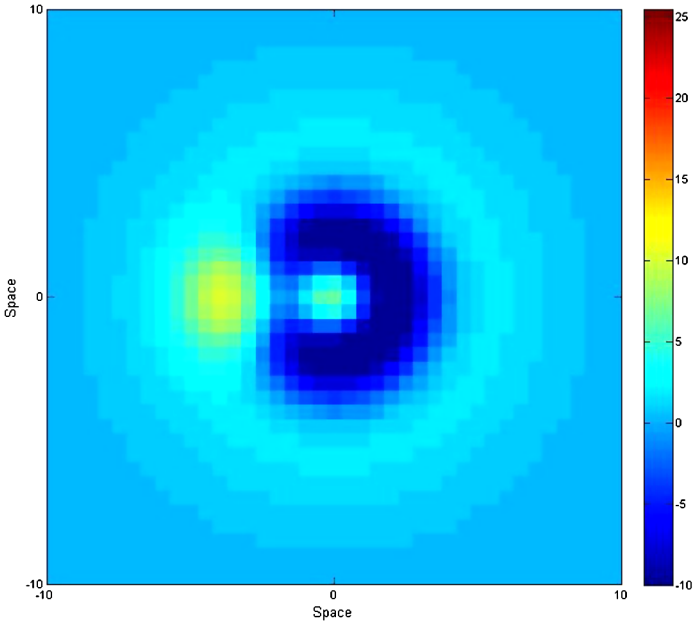

Figure 2 *Stochastic field generated by* 4 *basis functions.*

- Time: 890 days.
- Samples, $n = 100$.

Once the completion of the missing data in the weather stations is performed, the prior distributions given the parameters in the model are initialized (2.10). A normal Kernel of four basis functions is used whose parameters are Local amplitude 80, surround amplitude -80, lateral amplitude 5, and anisotropic amplitude 15 (Aram et al. (2007)). An approximated stochastic field by the four selected basis functions is shown in Figure 2.

The EnKF algorithm running in parallel for the precipitation data of each weather station was implemented using the POSIX threads library (*Pthread*) of the ANSI **C** programming environment, Appendix. Due to the great distances between the stations, the data from each weather station is considered independent of the others so that each station could be analysed separately. A covariance structure would permit a joint analysis of several weather stations with spatial interpolation however this was not studied in this article.

In order to test the procedure for making predictions, the series from January 2011 until May 2012 were used to predict the 30 days of June 2012. To initialize the PEnKF in the stations of Aragua state, the priors chosen after performing many runs were the following:

- Ceniap station $\sigma_u^2 \psi_0^{-1} = 100$, $\sigma_v^2 = 10$, $A_0 = 1, y_0 = 1.3$, $C_0 = 1.0001$, $\mu_{\text{fbr}} = \mu_{\text{kfbr}} = 1.3$, and $\sigma_{\text{fbr}} = \sigma_{\text{kfbr}} = 1000$, $\lambda = \frac{5}{3}$.

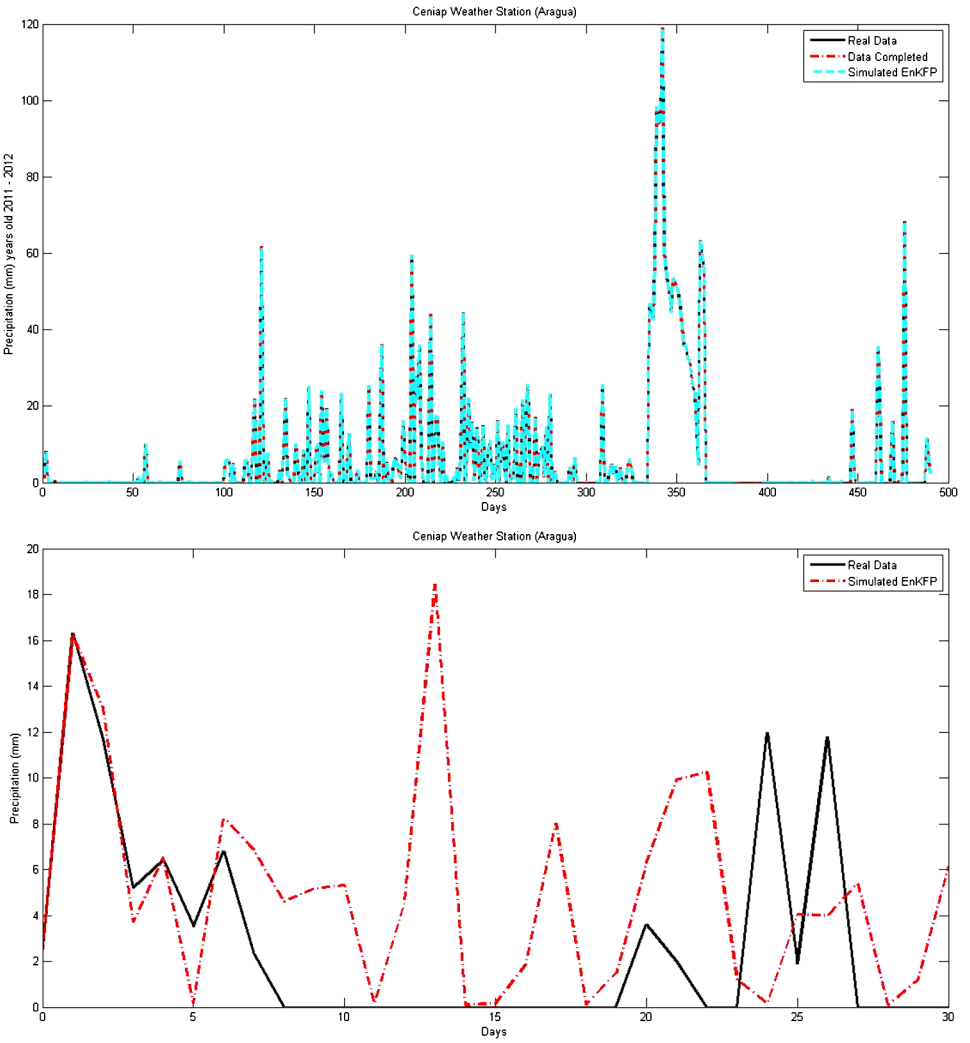

Figure 3 *Actual, augmented, and reconstructed data* (*top panel*), *prediction and actual data* (*bottom panel*), *Ceniap station.*

- Tamarindo station $\sigma_u^2\psi_0^{-1} = 100$, $\sigma_v^2 = 10$, $A_0 = 1$,$y_0 = 96.25$, $C_0 = 0$, $\mu_{\text{fbr}} = \mu_{\text{kfbr}} = 96.25$, and $\sigma_{\text{fbr}} = \sigma_{\text{kfbr}} = 1000$, $\lambda = \frac{5}{3}$.
- Tucutunemo station $\sigma_u^2\psi_0^{-1} = 100$, $\sigma_v^2 = 10$, $A_0 = 1$,$y_0 = 142.11$, $C_0 = 0$, $\mu_{\text{fbr}} = \mu_{\text{kfbr}} = 142.11$, and $\sigma_{\text{fbr}} = \sigma_{\text{kfbr}} = 1000$, $\lambda = \frac{5}{3}$.

The Figures 3, 4, 5 in the top panel show three graphs in which the real data sets are indicated with black color; the completed data obtained by the Gibbs algorithm denoted in red color, and the reconstructed data by the parallelized ensemble Kalman filter (PEnKF) denoted by cyan color. In the bottom panel of Figures 3,

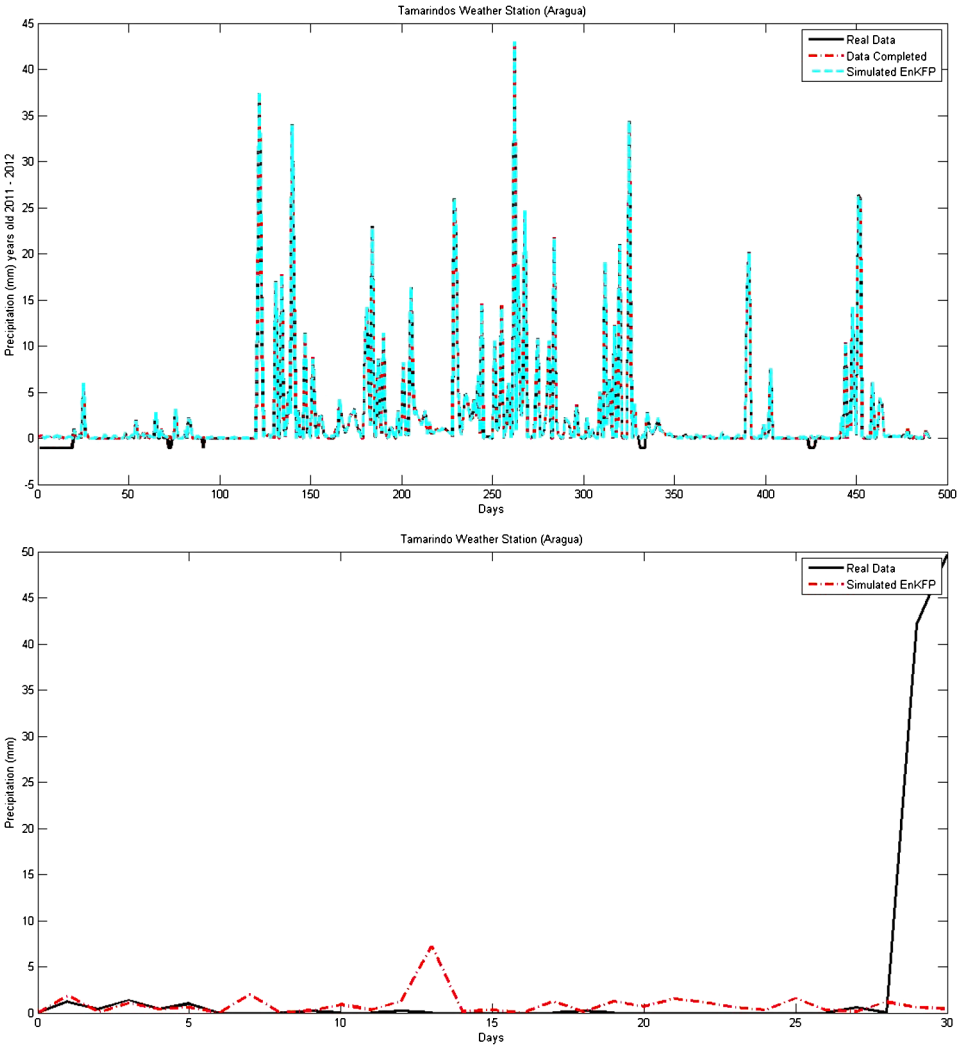

Figure 4 *Actual, augmented, and reconstructed data* (*top panel*), *prediction and actual data* (*bottom panel*), *Tamarindo station.*

4, 5 the daily predictions are shown by the PEnKF together with the series of true values of precipitation; reliable predictions are seen for the first four days whereas for the fifth day, a notable difference between the predicted values and the true values are observed.

For the San Pedro station in the Guárico state, the following specifications were taken:

- San Pedro station $\sigma_u^2\psi_0^{-1} = 100$, $\sigma_v^2 = 10$, $A_0 = 1$,$y_0 = 48.23$, $C_0 = 1$, $\mu_{\text{fbr}} = \mu_{\text{kfbr}} = 48.23$, and $\sigma_{\text{fbr}} = \sigma_{\text{kfbr}} = 0.1$, $\lambda = \frac{5}{3}$.

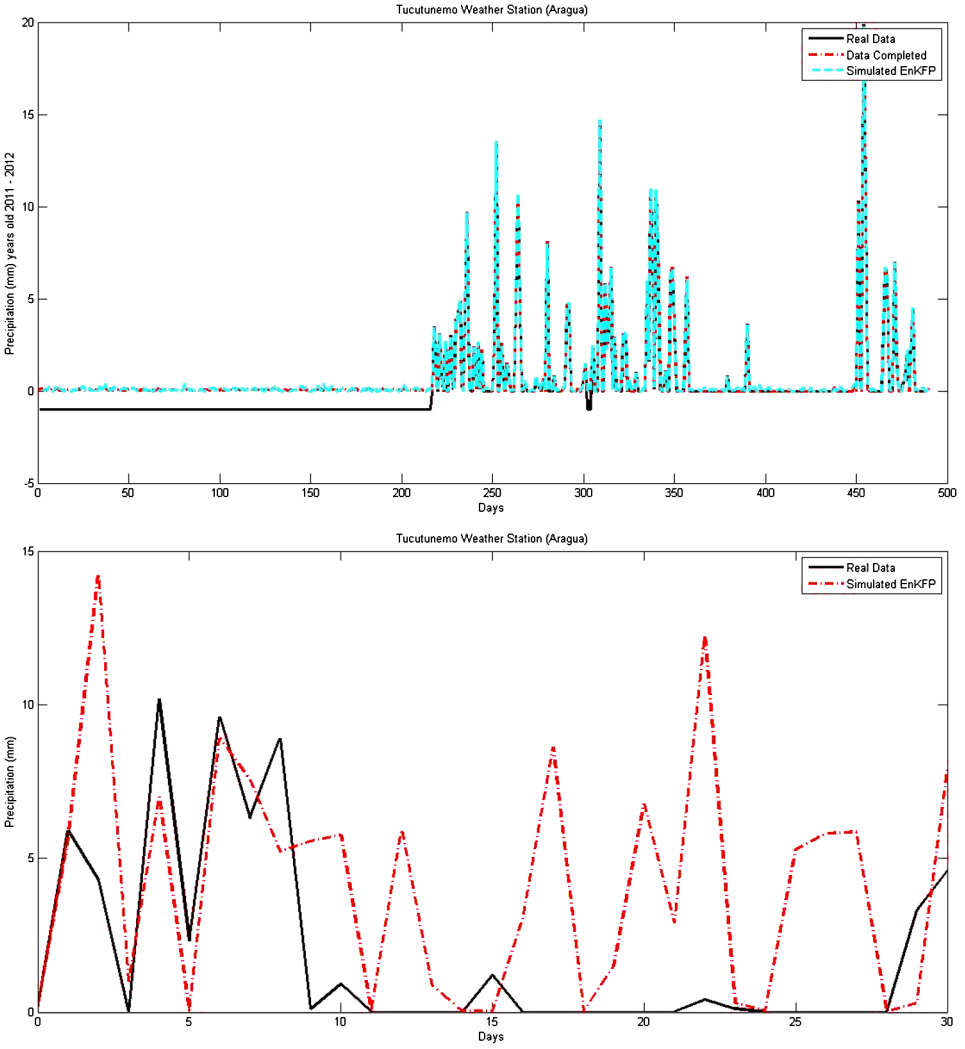

Figure 5 *Actual, augmented, and reconstructed data* (*top panel*), *prediction and actual data* (*bottom panel*), *Tucutunemo station*.

In Figure 6, the top panel shows a graph of the actual data in black; the augmented data in red, and the reconstructed data for the PEnKF denoted in cyan, showing that the PEnKF has a good performance in the reconstruction of the real system. In Figure 6, the bottom panel shows a graph of the daily predictions by the PEnKF, together with the series of true values of precipitation. Just as the predictions obtained for the Aragua state, the predictions are good to the fourth day.

Finally, for the Hondo Pueblo station, located in the state of Táchira, the specifications of the prior distributions are

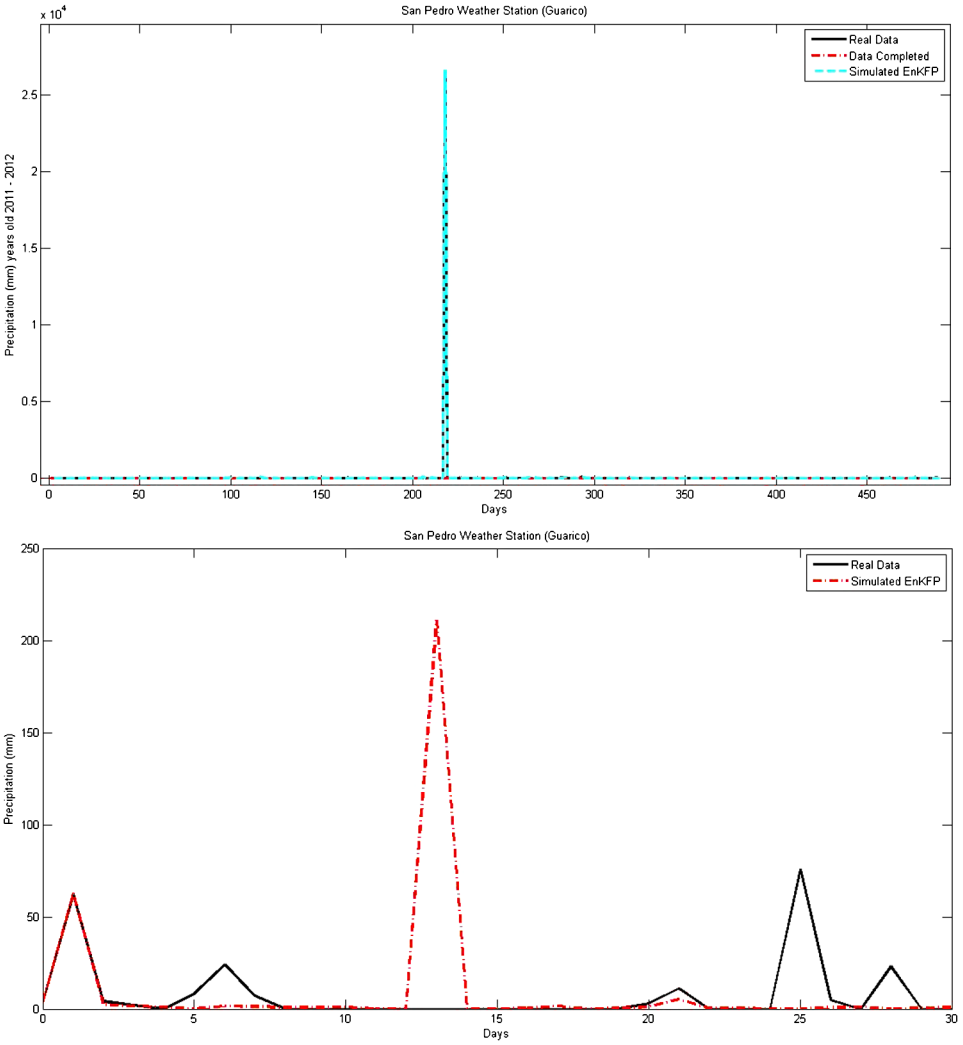

Figure 6 *Actual*, *augmented*, *reconstructed data* (*top panel*), *prediction and actual data* (*bottom panel*), *San Pedro station*.

- Pueblo Hondo station $\sigma_u^2 \psi_0^{-1} = 1$, $\sigma_v^2 = 10$, $A_0 = 1$, $y_0 = 0$, $C_0 = 1$, $\mu_{\text{fbr}} = \mu_{\text{kfbr}} = 0$, and $\sigma_{\text{fbr}} = \sigma_{\text{kfbr}} = 1$, $\lambda = \frac{5}{3}$.

In Figure 7 the top panel shows, similarly to the previous graphics, the augmented real data, as well as the reconstructed data. The bottom panel of Figure 7 shows the daily predictions and real values of precipitation. As in the previous cases, the predictions are good until the fourth day.

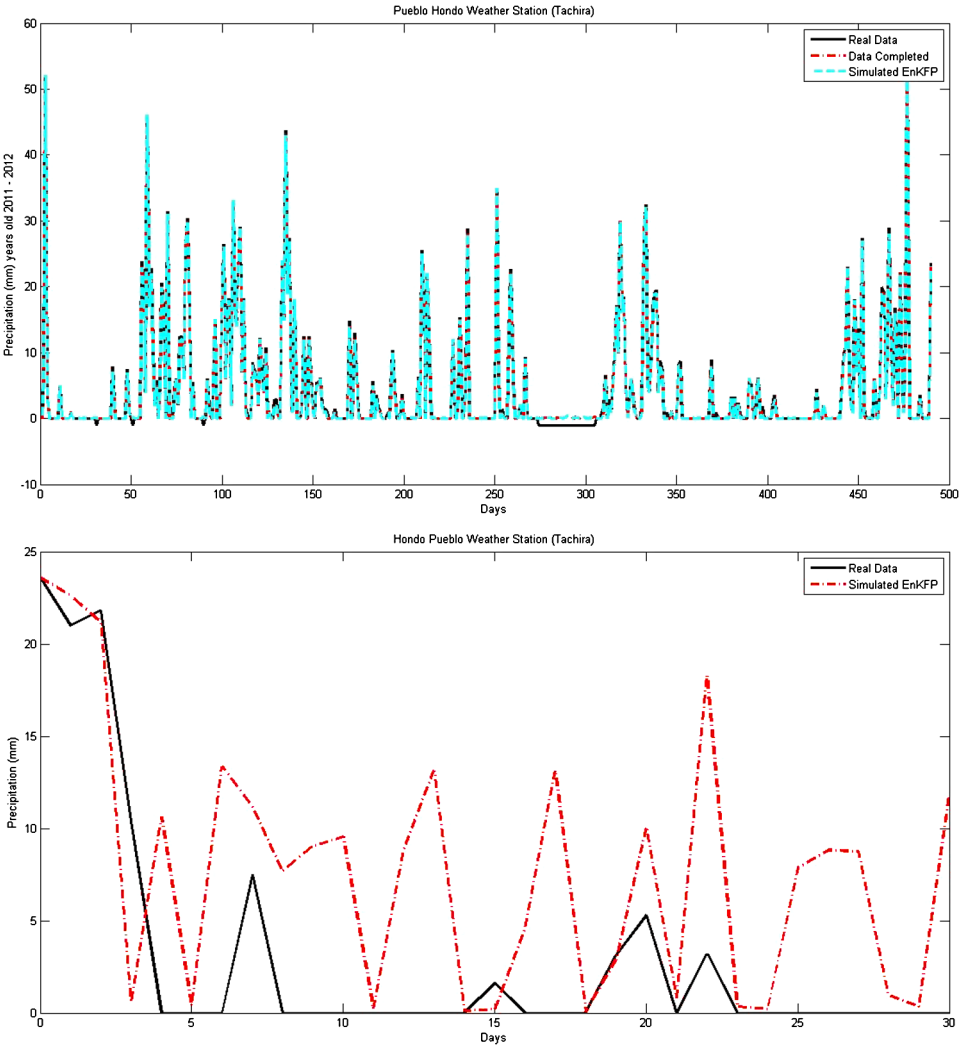

Figure 7 *Actual, augmented, reconstructed data* (*top panel*), *prediction and actual data* (*bottom panel*), *Pueblo Hondo station*.

Table 1 shows the three validation criteria to measure the quality of prediction of the proposed model. One can see that for the measures considered, the errors are low, and show little variability among them.

Table 2 shows the speedup and efficiency factors of the parallelized EnKF algorithm for different pthreads numbers. To measure the time it takes to execute the algorithm, the *gettimeofday* function is used, which is located in the library *sys/time.h*. The execution time of the sequential EnKF algorithm is 6,364,039 μs

Table 1 *Validation criteria to measure the quality of prediction of the proposed model*

Measures	Pueblo Hondo	San Pedro	Ceniap	Tamarindo	Tucutunemo
RMSE	0.2905	0.1037	0.1819	0.1461	0.2125
MAE	0.2082	0.1630	0.2116	0.1740	0.2410
RB	0.2250	0.1217	0.2019	0.1764	0.2525

Table 2 *The speedup and efficiency factors of the parallelized EnKF algorithm for different pthreads numbers*

Pthread	Time (μs)	Speedup	Efficiency
$p = 2$	5,860,180	1.085	0.54
$p = 3$	4,879,710	1.304	0.43
$p = 4$	5,042,585	1.262	0.315
$p = 8$	4,960,206	1.283	0.160
$p = 16$	7,098,805	0.896	0.056
$p = 32$	7,344,040	0.866	0.027

for the 5 weather stations, using 520 precipitation states and 500 members of the ensemble.

It is seen that by increasing the number of pthreads greater than 4, the efficiency is not increased so that it may be concluded that not all the pthreads are executing useful work. Perhaps the loss of efficiency is due to the increased cost of communication among processors, and to the delays in the communications and synchronizations with the non-parallelizable processes. Considering this, the ideal case would be $P = 2$ which provides a 8% increase in speed with respect to the sequential algorithm.

6 Discussion and conclusions

The traditional way to forecast rainfall is by using numerical prediction models. These models are posed in terms of systems of nonlinear differential equations that simulate the dynamics of the atmosphere. The problem encountered with these methods is that they require a lot of computation, the models are very complex and they are not freely available. Recently, other statistical techniques using prediction models and algorithms based on Monte Carlo Markov Chains (MCMC) have been developed. These computational methods are also tedious, but predictions are made at a much cheaper cost than the numerical models. They are also available in several libraries of free distribution, implemented in the **R** language platform. Statistical models based on MCMC techniques, and those proposed in this paper are inspired by the sequential Monte Carlo algorithms (SMC). They are

useful in situations where numerical models are not available or when it is required to obtain finer predictions with different temporal resolutions than those obtained by the numerical prediction model. The main contributions of this paper include the proposal of using a computational technique to complete missing data of daily rainfall at different weather stations and the implementation of a sequential Monte Carlo parallelized algorithm on a stochastic integro-difference equation model for data that varies in space and time. The parametrization of the model uses radial basis functions which reduce the computational cost and the size of the problem when working with large data sets that contain many variables. It was shown that the proposed methodology is able to predict the unknown states of rainfall, both spatially and temporally, quickly, efficiently and accurately for the first three to four days. To evaluate the performance of the model, three statistical validation methods are used: the square root of the mean squared error, the mean absolute error, and the relative bias. The three measures showed small errors with low variability among them. To evaluate the algorithm PEnKF, the speedup factor and the efficiency factor were used. For up to four threads, the algorithm executes faster than the sequential version.

Appendix: Paralleling the ensemble Kalman filter

The main program receives a set of daily rainfall at the weather stations which is stored independently in local variables. The program executes the ensemble Kalman filter in parallel using multiple threads for each local variable. That is to say, the main program, after storing the individual data of each weather station, creates p threads; each thread receives the means and variances of the base functions as parameters $\phi(s)$ and $\psi(s, r)$. When all the threads have finished calculating the ensemble Kalman filter for each local variable, the principal program shows the result. The high level algorithm, as defined in Sánchez, Infante, Marcano and Griffin (2015), is as follows:

a. *Initialization*.

- The daily precipitation data from weather stations is read.
- The precipitation data is assigned to independent variables.
- The mean and variance of the base functions $\phi(s)$ and $\psi(s, r)$ is read.

b. *Creating Pthread*.

- For each variable, p threads are created.
- The threads execute the EnKF.
- The EnKF generates the calculation of the posterior mean and covariance.

c. *Ending*.

- The runtime for the entire program is calculated.

- The criteria for validation of models are calculated.
- The speedup and efficiency factors for p threads are calculated.

To measure the speed of calculation using the parallelized ensemble Kalman filter algorithm, the speedup and efficiency factors, as defined in Wilkinson and Allen (2005), are used.

References

Amisigo, B. A. and van de Giesen, N. C. (2005). Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily river flow series. *Hydrology and Earth System Sciences* **9**, 209–224.

Aram, P., Freestone, D., Dewar, M., Grayden, D., Kadirkamanathan, V. and Scerr, K. (2007). Estimation of integro-difference equation based spatio-temporal systems. Available at https://github.com/mikedewar?tab=repositories.

Bakar, K. (2011). Bayesian Analysis of Daily Maximum Ozone Levels. Thesis for the degree of Doctor of Philosophy. Available at http://eprints.soton.ac.uk.

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. *Monographs on Statistics and Applied Probability* Chapman and Hall: New York.

Banerjee, S. and Fuentes, M. (2011). Bayesian modeling for large spatial data sets. Research report 2011-47. Division of Biostatistics, University of Minnesota. Wires Computational Statistics.

Berliner, L., Milliff, R. and Wikle, C. (2003). Bayesian hierarchical modeling of air–sea interaction. *Journal of Geophysics* **108**, 3104.

Cameletti, M., Ignaccolo, R. and Bande, S. (2010). Comparing air quality statistical models. Techical Report 40, Graspa working paper (accepted for publication at environmetrics).

Cameletti, M., Lindgren, F., Simpson, D. and Rue, H. (2012). *Spatio-temporal modeling of particulate matter concentration through the SPDE approach. 2, 109 - 131.*

Calder, C., Berrett, C., Shi, T., Xiao, N. and Munroe, D. K. (2011). Modeling space time dynamics of aerosols using satellite data and atmospheric transport model output. *Journal of Agricultural, Biological, and Environmental Statistics* **16**, 495–512. MR2862295

Cocchi, D., Greco, F. and Trivisano, C. (2007). Hierarchical space–time modelling of pollution. *Atmospheric Environment* **41**, 532–542.

Cressie, N. and Huang, H. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**, 1330–1340. MR1731494

Cressie, N. and Wikle, C. (2011). *Statistics for Spatio-Temporal Data*. New York: Wiley. MR2848400

Dewar, M. (2007). A framework for modelling dynamic spatiotemporal systems. Ph.D. Thesis.

Dewar, M., Scerri, K. and Kadirkamanathan (2011). Data-driven spatio-temporal modeling using the integro-difference equation. *IEEE Transactions on Signal Processing* **57**, 1. MR2674787

Evensen, G. (2009). *Data Assimilation: The Ensemble Kalman Filter*, 2nd ed. Berlin: Springer.

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research* **99**, 10143–10162.

Evensen, G. (2003). The ensemble Kalman filter: Theoretical formulation and practical implementation. *Ocean Dynamics* **53**, 343–367.

Evensen, G. and van Leeuwen, P. (1996). Assimilation of geosat altimeter data for the agulhas current using the ensemble Kalman filter with a quasi-geostrophic model. *Monthly Weather Review* **24**, 85–96.

Fasso, A. and Cameletti, M. (2007). A general spatio-temporal model for environmental data. Technical Report No. 27, Graspa–The Italian Group of Environmental Statistics.

Fernández-Casal, R., González-Manteiga, W. and Febrero-Bande, M. (2003). Flexible spatio-temporal stationary variogram models. *Statistics and Computing* **13**, 127–136. MR1963329

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association* **97**, 590–600. MR1941475

Hernández, A., Guenni, L. and Sansó, B. (2009). Extreme limit distribution of truncated models for daily rainfall. *Environmetrics* **20**, 962–980. MR2838498

Hernández, A., Guenni, L. and Sansó, B. (2011). Características de la precipitación extrema en algunas localidades de Venezuela. *Interciencia* **36**, 185–191.

Jones, R. H. and Zhang, Y. (1997). Models for continuous stationary space–time processes. In *Modelling Longitudinal and Spatially Correlated Data* (T. G. Gregoire, D. R. Brillinger, P. J. Diggle, E. Russek-Cohen, W. G. Warren and R. D. Wolfinger, eds.) 289–298. New York: Springer. MR1603158

Kotecha, J. H. and Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate Gaussian random variables. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1757–1760. Los Alamitos: IEEE Computer Society.

Lasinio, J., Sahu, S. and Mardia, K. (2007). Modeling rainfall data using a Bayesian Kriged–Kalman model. In *Bayesian Statistics and Its Applocations* (S. K. Upadhya, U. Singh and D. K. Dey, eds.). London: Anshan Ltd.

Li, Y. and Ghosh, S. (2013). Efficient sampling methods for truncated multivariate normal and Student-t distributions subject to linear inequality constraints. Technical reports 2649, NC State Department of Statistics.

Ma, C. (2003). Families of spatio-temporal stationary covariance models. *Journal of statistical planning and inference*. To appear. MR2000096

Majda, A. and Harlim, J. (2012). *Filtering Complex Turbulent Systems*. Cambridge: Cambridge University Press. MR2934167

Robert, C. (1995). Simulation of truncated normal variables. *Statistics and Computing* **5**, 121–125.

Sánchez, L. and Infante, S. (2013). Reconstruction of chaotic dynamic systems using non-linear filters. *Chilean Journal of Statistic* **4**, 1–19. MR3054223

Sánchez, L. Infante, S. Marcano, J. and Griffin, V. (2015). Polinomial Chaos based on the parallelized ensamble Kalman filter to estimate precipitation states *Statistics, Optimization and Information Computing* **1**, 79–95.

Sahu, S. (2011). Hierarchical Bayesian models for space–time air pollution data. In *Handbook of Statistics–Time Series Analysis, Methods and Applications* (C. Rao, ed.). *Handbook of Statistics* **30**. Holland: Elsevier Publishers.

Sahu, S. K., Yip, S. and Holland, D. M. (2011). A fast Bayesian method for updating and forecasting hourly ozone levels. *Environmental and Ecological Statistics* **18**, 185–207. MR2783689

Sansó, B. and Guenni, L. (1999a). Stochastic model for tropical rainfall at a single location. *Journal of Hydrology* **214**, 64–73.

Sansó, B. and Guenni, L. (1999b). Venezuelan rainfall data analysis using a Bayesian space–time model. *Journal of the Royal Statistical Society Series C Applied Statistics* **48**, 345–362.

Sansó, B. and Guenni, L. (2000). A non-stationary multi-site model for rainfall. *Journal of the American Statistical Association* **95**, 1089–1100. MR1821717

Scerri, K., Dewar, M. and Kadirkamanathan, V. (2009). Estimation and model selection for an IDE-based spatio-temporal model. *IEEE Transactions on Signal Processing* **57**, 482–492. MR2603377

Scerri, K., Dewar, M., Parham, A., Freestone, D., Kadirkamanathan, V. and Grayden, D. (2011). Balanced reduction of an IDE-based spatio-temporal model computation tools. The second international on computational logics, algebras, programming, tools, and benchmarking, 7–122.

Sigrist, F., Künsch, H. and Stahel, W. (2012). A dynamic nonstationary spatio-temporal model for short term prediction of precipitation. *Annals of Applied Statistics* **6**, 1452–1477. MR3058671

Stein, M. (2005). Space–time covariance functions. *Journal of the American Statistical Association* **100**, 310–321. MR2156840

Stroud, J., Stein, M., Lesht, B., Schwar, D. and Beletsky, D. (2010). An ensemble Kalman filter and smoother for satellite data assimilation. *Journal of the American Statistical Association* **105**, 978–990. MR2752594

West, M. and Harrinson, J. (1997). *Bayesian Forescasting and Dynamic Models*, 2nd ed. New York: Springer. MR1482232

Wilhelm, S. (2013). Truncated multivariate normal and student t distribution. Available at http://cran.r-project.org/web/packages/tmvtnorm/tmvtnorm.pdf.

Wilkinson, B. and Allen, M. (2005). *Parallel Programming: Techniques and Application Using Networked Workstations and Parallel Computers*, 2nd ed. New York: Prentice-Hall.

Wikle, C. K. and Holan, S. H. (2011). Polynomial nonlinear spatio-temporal integro-difference equation models. *Journal of Time Series Analysis* **32**, 339–350. doi:10.1111//j.1467-9892.2011.00729.x. MR2841788

Wikle, C. and Hooten, M. (2006). Hierarchical Bayesian spatio temporal models for population spread. In *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods* (J. S. Clark and A. Gelfand, eds.). London: Oxford University Press.

Wikle, C. (2002). A kernel-based spectral model for non-Gaussian spatio-temporal processes. *Statistical modelling: An international journal* **2**, 299–314. MR1951587

Wikle, C., Milliff, R., Nychka, D. and Berliner, L. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *Journal of the American Statistical Association* **96**, 382–397. MR1939342

Xu, K., Wikle, C. K. and Fox, N. I. (2005). A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. *Journal of the American Statistical Association* **100**, 1133–1144. MR2236929

L. Sánchez
Department of Mathematics
Faculty of Education
University Carabobo
Venezuela
E-mail: sluis@uc.edu.ve

S. Infante
Department of Biotechnology
Faculty of Engineering
Technical University of North
Ecuador
E-mail: sinfante@uc.edu.ve

V. Griffin
Department of Mathematics
Faculty of Science and Technology
University Carabobo
Venezuela
E-mail: vgriffin@uc.edu.ve

D. Rey
Institute of Mathematics and Applied Calculus
University Carabobo
Venezuela
E-mail: drey@uc.edu.ve