INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **35**: 1453–1464 (2015)
Published online 11 June 2014 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.4068

RMetS
Royal Meteorological Society

Interpolation of daily rainfall using spatiotemporal models and clustering

A. F. Militino,[a*] M. D. Ugarte,[a] T. Goicoa[a] and M. Genton[b]

[a] *Department of Statistics and Operations Research, Public University of Navarre, Pamplona, Spain*
[b] *Computer, Electrical and Mathematical Sciences and Engineering (CEMSE), King Abdullah University Science and Technology (KAUST), Thuwal, Kingdom of Saudi Arabia*

ABSTRACT: Accumulated daily rainfall in non-observed locations on a particular day is frequently required as input to decision-making tools in precision agriculture or for hydrological or meteorological studies. Various solutions and estimation procedures have been proposed in the literature depending on the auxiliary information and the availability of data, but most such solutions are oriented to interpolating spatial data without incorporating temporal dependence. When data are available in space and time, spatiotemporal models usually provide better solutions. Here, we analyse the performance of three spatiotemporal models fitted to the whole sampled set and to clusters within the sampled set. The data consists of daily observations collected from 87 manual rainfall gauges from 1990 to 2010 in Navarre, Spain. The accuracy and precision of the interpolated data are compared with real data from 33 automated rainfall gauges in the same region, but placed in different locations than the manual rainfall gauges. Root mean squared error by months and by year are also provided. To illustrate these models, we also map interpolated daily precipitations and standard errors on a 1 km^2 grid in the whole region.

KEY WORDS kriging; state-space models; statistical models; thin-plate splines

Received 5 November 2013; Revised 31 March 2014; Accepted 9 May 2014

1. Introduction

Interpolation of daily rainfall data provides important information for precision agriculture, and for climatological, meteorological and hydrological studies for which daily precipitation data from non-observed locations are needed. Such interest in spatiotemporal modelling of rainfall is not new (Cox and Isham, 1988), and many researches worked on methods to interpolate daily rainfall data, whereas hydrologist or climatologists have recognized the inherent difficulty in this task (Kleiber *et al.*, 2012). For example, PRISM (Parameter-elevation Regressions on Independent Slopes Model) is a climate analysis system that uses point data, a digital elevation model (DEM), and other spatial datasets to generate gridded estimates of annual, monthly and event-based climatic parameters (Daly *et al.*, 1994). Numerical and deterministic approximations of daily precipitation are already frequent in practice (Martin-Vide, 2004), with deterministic interpolations becoming very popular because of their simplicity and ease of programming in commercial software. However, these approximations are not necessarily based on statistical models, consequently inference is not possible and precision assessment becomes a difficult task. For example, Weymouth *et al.* (1999) solved the problem in Australia by using a modular Barnes successive corrections scheme. They estimated rainfall at a grid-point by using weighted averages of surrounding observations and conducted their analysis on a 25 km^2 grid with a correlation scale of 80 km. They tested the accuracy of their method comparing their results with the full climatological dataset collected by the National Climate Center of Australia. Tait *et al.* (2006) used spline interpolations of daily rainfall based on a mean annual rainfall surface, rather than elevation, as an independent variable in an interpolation of daily rainfall data in New Zealand. They used spatial but not temporal interpolations. Xie *et al.* (2007) analysed daily climatology data by interpolating data from climatological stations as averaged over a 20-year period from 1978 to 1997. The daily climatology data were adjusted to correct the bias caused by orographic effects. Analysis of the total daily precipitation was calculated by multiplying the daily climatology data by the daily ratio of precipitation. Carrera-Hernández and Gaskin (2007) interpolated daily data using different kriging models. They used spatial information only. Herrera *et al.* (2012) presented a spatial interpolation of precipitation data using a 20 km^2 grid across Spain using data from 50 years. They used kriging, thin-plate splines and inverse weighting methods but no temporal information was added to the interpolation methods. Simple and descriptive methods of interpolation such as inverse distance-weighted interpolation and a local weighted regression method in which elevation and distance are the explanatory variables used by

* Correspondence to: A. F. Militino, Department of Statistics and Operations Research, Public University of Navarre, Pamplona, Spain. E-mail: militino@unavarra.es

Kurtzman *et al.* (2009). Their results showed that spatially variable, physically based parametrization of the distance weighting function can improve the spatial interpolation of daily precipitation data. Newlands *et al.* (2011) evaluated the precision of three interpolation methods (TP splines, a weighted-truncated Gaussian filter, and a hybrid inverse distance and natural-neighbour interpolation) to interpolate daily precipitation and temperature data in Canada using available historical records from daily station time-series data. They compared the performance of the three methods using cross-validation and averaging daily precipitation over the period 1961–1990. The specific interpolation of a particular day was not provided. A hidden Markov model for space–time evolution of daily rainfall was proposed by Ailliot *et al.* (2009). They used the expectation–maximization (EM) algorithm for the estimation process, but the focus was on evolution more than on prediction at new locations. Kokic *et al.* (2011) demonstrated the potential advantage of using a linear, mixed-effect state-space model for statistical downscaling of climate variables compared to the frequently used approach of linear regression. They demonstrated the utility of their method by predicting annual rainfall and temperature means in Australia. With lead times of 1–10 years, their state-space approach was able to predict observed seasonal temperature and rainfall means with substantially greater precision than climatology, multivariate linear regression or a standard linear state-space approaches. However, extensions of this model to monthly or daily data have not been made. Moral (2010) compared different spatial interpolation methods based on kriging techniques with monthly precipitation data from 136 rainfall stations from Extremadura, Spain. The inclusion of altitude improved the predictions considerably. Jeong *et al.* (2012) aimed at modelling daily temperatures and precipitation using covariates and historical data. They compared multiple regression models, ordinary least squares estimates, robust regression, ridge regression and artificial neural networks to identify an appropriate transfer function in statistical downscaling models to capture daily precipitation occurrence and amounts. This comparison was made with data from 25 observation sites located in five Canadian provinces from 1960 until 2000.

The use of simulation techniques may also be useful to data interpolation. Precipitation simulation dates back to the 1950s and many simulation techniques have been developed since then. These techniques are typically based on the assumption of mixed distributions of rainfall in a parametric or nonparametric framework (Li *et al.*, 2012) and the simulations are based on gauge stations. Recently, extensions to non-gauge stations have been made by borrowing information from the closest stations (Mehrotra *et al.*, 2012). Vaze *et al.* (2011) used simulation techniques to reproduce observed historical annual and seasonal mean rainfalls, observed annual rainfall series, and observed daily rainfall distributions across Southeast Australia. Infilling particular records have been also of interest for some authors. Disaggregated rainfall and infilling missing values can also be done using spatiotemporal models (Cowpertwait *et al.*, 2006). Daily stochastic spatiotemporal precipitation generators have been developed, although daily generators produce daily rainfall only from locations with observations. When interpolating without observations, deterministic methods are commonly used. Recently Kleiber *et al.* (2012) provided gridded simulations by using kriging to interpolate the model parameters necessary for the simulations.

The aim of this paper is to provide daily rainfall estimates anywhere in a specific region and on any day within a particular year. These predictions could be easily incorporated as input into more sophisticated models to study the climate, agriculture or forestry. We illustrate the results with historical rainfall data from a network of 87 manual rainfall gauges over a period of 21 years. To make predictions of daily rainfall, we use planar coordinates and elevation as covariates. Using this temporal and spatial information, precise and accurate interpolation of rainfall data in 1 km^2 network is achieved. Specifically, we compare three alternative stochastic models and validate the predictions from actual data collected from rainfall gauge stations. These alternative models include a kriging model with an external drift of historical data, called the spatiotemporal (ST) kriging model, a thin-plate (TP) spline model with the same external drift as the ST-kriging model and a dynamic state (S)-space model that simultaneously incorporates spatial and dynamic temporal dependence. We also analyse the models using clusters because great variability could occur on certain days due to the presence of local storms or dramatically changing weather.

The rest of the paper is organized as follows. Section 2 presents the data and the methods. It explains the application and reviews the models used in the interpolation process. Section 3 presents the results. Finally, we draw conclusions in Section 4.

2. Data and methods

The data used in this study are drawn from the records of 87 manual stations with daily precipitation observations for a 21-year period (1990–2010). There was less than 1% missing data from these stations, because manual stations experience fewer failures than do automated stations. Missing data have been replaced by historical means of 5 days, yet other periods can be used. The Agriculture and Environmental Department of Navarre provided the main data as well as another set of daily precipitation records collected from the automated rainfall gauges for validation purposes. From the automated rainfall gauge stations, we chose data from 33 rainfall gauges that had no missing data in 2010. The locations of the automated and manual rainfall gauge stations differed and were spread across the region of interest, called Navarre, a province located in the north of Spain. The left plot of Figure 1 shows the locations of the 87 manual stations in Navarre as squares and the locations of the 33 automated ones, along with an elevation scale in meters. Universal Transversal Mercator (UTM) original coordinates of Navarre are scaled to kilometres and translated into a new coordinate origin defined as the minimum of X and Y, respectively, to express distances in kilometres instead of metres. The right plot of the same figure presents a relief map of Navarre. Navarre is a region of roughly 10 000 km^2 located in North Central Spain. Elevations vary between 200 and 2500 m in the highest zone of the Pyrenees, located in northeastern Navarre. Valleys and mountains are ubiquitous in the north, and small hills are common in the central part of the province. South Navarre is mainly flat, with one of the biggest desert regions of Spain, called Bardenas Reales, the second largest European dessert with 42 500 ha. In this zone, the climate is continental with hot and dry summers and cold winters. The northwest Navarre is humid with an average annual temperature between 11 and 14.5 °C and average rainfall between 1400 and 2500 mm. The altitude of northeast Navarre lies between 1459 and 2438 m. The average annual temperature ranges from 7 to 12 °C and rainfall between 900 and 2200 mm in this area of the province. Close to Pamplona, the capital of Navarre, rainfall ranges between 700 and 1400 mm and the average annual temperature fluctuates between 10 and 13 °C. The central area of Navarre has a temperate Mediterranean climate, with an average rainfall of 450–750 mm and average temperatures between 12.5 and 14 °C, which suggests a tendency towards a continental climate. To the west is an area called Tierra Estella, where climate changes are frequent between the mountainous northern area, under the influence of the Atlantic Ocean with 1100–1500 mm of rainfall and an average temperature between 9 and 11 °C in the area of Urbasa-Andía. The southern plains of the province are under a Mediterranean influence with 500–800 mm of rain and 11.5–13.5 °C on average temperatures. In the south, called the Ribera region, the climate is of a Mediterranean continental nature, typical of the Ebro depression, with dry summers, temperatures with large annual variations, little and irregular rainfall (less than 500 mm year^{-1}) and the frequent presence of the northerly wind. Clearly, the climate varies a great deal across the province.

For introducing the stochastic models to be used in the ST interpolation, we assume that $\mathbf{z}_{st} = (z(s_1, t_1), z(s_1, t_2), \ldots, z(s_1, t_n), \ldots, z(s_n, t_1), \ldots, z(s_n, t_T))$ is the spatiotemporal process that has been observed at n geographical locations, $s_1, \ldots, s_n$, (87 in this case), at time t_j, from $j = 1, \ldots, T$, varying from 1 January 1990 until 31 December 2010. We analyse three models: a ST-kriging model, a TP-spline model and a S-space model. A brief description of these models is presented below.

2.1. The spatiotemporal (ST) kriging model

Kriging is the most common statistical procedure for spatial interpolation. Specifically, the assumption upon which kriging is based on is the decomposition of the stochastic process represented by accumulated daily rainfall that can be done through the sum of a linear trend and a stochastic error process. The linear trend is given by a linear combination of the planar coordinates and the orthometric elevation, $\mathbf{h}_s$, all of which are time-invariant. To include the time dimension we introduce a new covariate, $\mathbf{a}_{st}$, computed as the average precipitation of each 5-day period within each month between 1990 and 2010. The value of $\mathbf{a}_{st}$ is the same for all the days in the same period. We therefore have 72 different average rainfall values, although alternative periods are possible (Militino *et al.*, 2003). The total daily rainfall on a fixed day t, and at location s is then modelled as

$$\mathbf{z}_{st} = \mu_{st} + \epsilon_{st} = \beta_{0,t}\mathbf{1} + \beta_{1,t}\mathbf{x}_s + \beta_{2,t}\mathbf{y}_s + \beta_{3,t}\mathbf{h}_s + \beta_{4,t}\mathbf{a}_{st} + \epsilon_{st}, \tag{1}$$

where μ_{st} is the linear trend, $\mathbf{1}$ is a vector of ones, $\mathbf{x}_s$ and $\mathbf{y}_s$ are the spatial coordinates in $\mathbb{R}^2$ and $\epsilon_{st} \sim N_n(0, \Sigma(d))$. The spatial covariance structure is accounted for in the model error. $\Sigma(d)$ can be estimated between known alternatives as Matérn, exponential or spherical covariance matrices (Militino and Ugarte, 2001; Apanasovich *et al.*, 2012). The covariance matrix depends on the distance $d = \|s_i - s_j\|$, and it is invariant to translations. The model is therefore second-order stationary and isotropic. We used model (1) to make predictions for every day of the year 2010 with the 87 automated rainfall sampled stations. To validate the model, we used data from the 87 manual rainfall gauges and the non-sampled set of 33 automated rainfall gauges.

We used the R package geoR (see Ribeiro and Diggle, 2001) to estimate the unknown parameters ($\beta_{0,t}, \beta_{1,t}, \beta_{2,t}, \beta_{3,t}, \beta_{4,t}$ and those from Σ), the predictions, $z^k_{s_0 t_0}$, and the standard errors in the new location, $\mathbf{s}_0$, on a fixed day, t_0, by iterative likelihood methods. Convergence is achieved in all cases in a few seconds. However, additional programming is necessary for calculating and mapping the predictions and standard errors across the region. To achieve this aim, we use a regular square grid of 1 km^2 covering Navarre province to determine the average orthometric height and average rainfall for each 5-day period since 1990, necessary for defining the spatiotemporal covariate, $\mathbf{a}_{st}$. This required assigning to every point of the 2590 points of the grid information on the historical data corresponding to the closest rainfall gauge station. Closest rainfall stations are defined as those with lower Euclidean distance to the specific point on the grid.

2.2. The thin-plate (TP) spline model

A cubic spline is a curve made up of sections of cubic polynomials joined together so that they are continuous in value, as well as in their first and second derivatives (Ruppert *et al.*, 2003). The TP spline is the two-dimensional analogue of the cubic spline in one dimension. In particular, a second-order TP spline is the result of minimizing the residual sum of squares, $\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{z}_{s_i t} - f\left(\mathbf{x}_i, \mathbf{y}_i\right)\right)^2$, subject to a constraint that depends on the level of smoothness. The minimization problem (Duchon, 1977) is expressed as:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{z}_{s_i t} - f\left(\mathbf{x}_i, \mathbf{y}_i\right)\right)^2 + \lambda \int\int \left[\left(\frac{\partial^2 f}{\partial x^2}\right)^2 + \left(\frac{\partial^2 f}{\partial y^2}\right)^2 + 2\left(\frac{\partial^2 f}{\partial x^2 y^2}\right)^2\right] \mathrm{d}x\mathrm{d}y \tag{2}$$

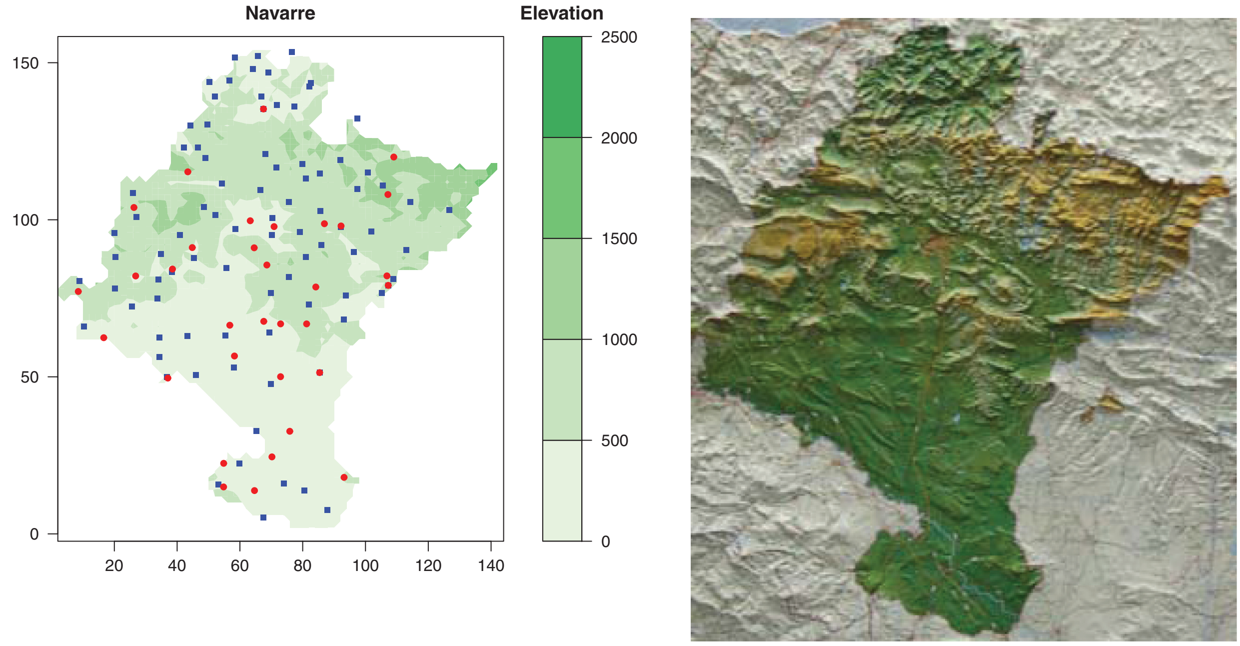

Figure 1. Left: Map of the locations of the manual rainfall gauges (squares) and automatic rainfall gauges (dots) located in Navarre, Spain, with elevation. Right: a relief map of Navarre.

The smoothing parameter, λ, must be chosen appropriately if the right balance is to be struck between maximizing the model's goodness of fit as measured by the first term and the model's wiggliness as measured by the second one. The value $\lambda = 0$ corresponds to no smoothness constraints and then data are interpolated, and $\lambda = \infty$ corresponds to just fitting the polynomial base model by ordinary least squares (Wood, 2003).

The TP-spline model used in this paper is an additive model similar to the ST-kriging model (1), where we do not necessarily assume a linear relationship of the total rainfall with the planar coordinates, but a smoother relationship fitted by splines and expressed by $f(\mathbf{x}_s, \mathbf{y}_s)$. The TP-spline model is given by

$$\mathbf{z}_{st} = f\left(\mathbf{x}_s, \mathbf{y}_s\right) + \beta_{1t}\mathbf{h}_s + \beta_{2t}\mathbf{a}_{st} + \epsilon_{st} \tag{3}$$

where $\epsilon_{st} \sim N_n(0, \Sigma(d))$, and $\mathbf{h}_s$ and $\mathbf{a}_{st}$ are the same covariates defined in Section 2.1 The spline is obtained as a weighted average of the observed data because the optimal estimate of $f(\mathbf{x}_s, \mathbf{y}_s)$ turns out to be linear in the observations. The solution to this minimization problem is identical to the universal kriging predictor of $f(\mathbf{x}_s, \mathbf{y}_s)$ under a certain intrinsic random function model that yields the second-order TP-smoothing spline as the optimal predictor (Stein, 1991). The R package *fields* (Fields Development Team, 2006) estimates second-order TP-smoothing spline models by fitting a surface to irregularly spaced data. The smoothing parameter is chosen by generalized cross-validation or by restricted maximum likelihood. Convergence is achieved for 365 days in a few seconds and additional programming is necessary for calculating and mapping the predictions and their standard errors across the region, in a similar way for the ST-kriging model.

The kriging estimator is not only the best linear unbiased estimator but also provides the solution to a particular variational problem. It can be interpreted as the generalized smoothing spline where the roughness penalty is determined by the covariance function of a spatial process (Nychka, 1995). In other words, the kriging method provides predictions that are linear combination of basis functions described by the covariance function. Matheron (1980) showed that for a given set of data, the interpolating spline is equivalent to kriging but with the covariance function given by $\Sigma(d) = |d|^2 \log(d)$. Therefore, it is expected that the ST-kriging model will provide similar predictions as those obtained with the TP-spline model.

2.3. The state (S)-space model

The S-space model is a spatiotemporal linear model that simultaneously accounts for spatial and temporal dependence. It is given by a transition equation and a state equation:

$$\begin{aligned} \mathbf{z}_{st} &= \beta_{0,t}1 + \beta_{1,t}\mathbf{x}_s + \beta_{2,t}\mathbf{y}_s + \beta_{3,t}\mathbf{h}_s + v_t + \epsilon_{st} \\ v_t &= v_{t-1} + \eta_t \end{aligned} \tag{4}$$

where $\mathbf{z}_{st}$ is the spatiotemporal process of rainfall, the error process $\epsilon_{st} \sim N_n(0, \Sigma(d))$ and $\Sigma(d)$ is a spatial covariance matrix similar to the other models. The unobservable latent temporal process, v_t, shows the temporal dynamics of data through a Markovian random walk. Finally, $\eta_t \sim N(0, \sigma_\eta)$, quantifying the uncertainty of the state estimate given the n observations. The transition equation incorporates the spatial dependence and the state equation takes into account the temporal dependence. Therefore, this S-space model can be interpreted as a ST-kriging model with a separable spatiotemporal covariance function.

The S-space model is implemented in the R package Stem. It was originally applied to predict air concentrations or to deal with error measurements in instruments (Fasso *et al.*, 2007). This package uses the function Stem.Estimation to carry out the iterations of the EM algorithm until convergence. Each iteration calls the function Kalman to carry out both the E-step and the M-step. The exponential covariance function is also assumed for $\Sigma(d)$. A similar S-space model is used by Amisigo and van de Giesen (2005) to estimate model parameters and missing data of river basin runoff values. A maximization process of the likelihood is performed by using the Kalman filter for which we need the initial values: μ_0 as the initial mean for the normal distribution, and v_0 as the initial state of the latent variable v_t. Additional programming is necessary for calculating new predictions and producing the maps of Navarre. In this case, we do not need any historical information about the rainfall at new locations because the state equation is independent of the location and the temporal information is dynamically introduced into the model through the state equation.

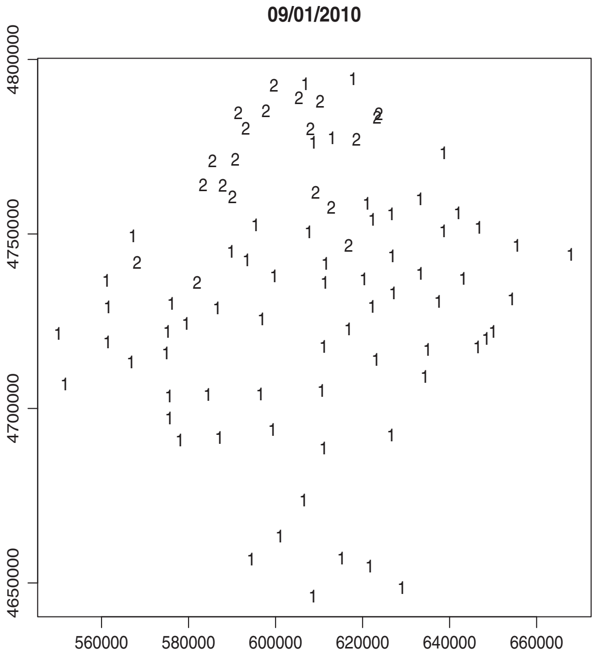

Figure 2. Clusters on 9 January 2010.

3. Results

ST-kriging and TP-spline models can provide predictions of accumulated daily rainfall in every location of Navarre as long as we know the UTM coordinates, the elevation and the historical average accumulated rainfall for 5-day increments used as the covariate. However, the S-space model needs planar coordinates and elevation as static covariates, because the temporal information is dynamically introduced into the model through a random walk in the state equation. For comparing the performance of the three alternative models, we calculated daily predictions during 2010 at the 87 manual rainfall stations and at the 33 automated stations located in different places.

By looking at daily predictions, we can check how on some specific days rain is present in small areas but is completely absent in the rest of Navarre (Militino *et al.,* 2001). In these cases, we can introduce clusters. Clusters are defined from the original sample set on the manual gauge stations. *K*-means clustering (Hartigan, 1975) is a method that aims to partition the locations on the sampled manual rainfall stations into k groups such that the sum of the squares from points to the assigned cluster centres is minimized. In this study, only two clusters with a minimum of ten stations are considered, because of the limited number of rainfall gauge stations. Observed and historical rainfall are the variables of classification for characterizing the clusters. Other alternatives that include altitude were considered, but they did not outperform the proposed method. Therefore, rainfall gauge stations in the same cluster are more similar to the observed and historical rainfall than to the spatial proximity, i.e. clusters are not necessarily contiguous because they do not incorporate the spatial position of the rainfall stations, yet when representing them on a map, it is common to observe that the points in the same cluster are geographically close. Figure 2 shows the clustering results of 9 January 2010. We can see that the majority of the rainfall stations belonging to cluster 1 are located in the south-central Navarre, whereas cluster 2 is mainly located in the north. It is a cluster scheme very common in the region, because the central and southern areas are more continental in climate than in the north, which is more humid. However, there are a few rainfall stations labelled as cluster 1 but located between all the stations of cluster 2 and vice versa. This is also common because of the presence of valleys between the mountainous regions.

When defining two clusters on a particular day, we need to fit two models, one for each cluster, for the three models considered in the paper, yet the analysed models work exactly in the same way as if there were no clusters. Moreover, additional programming is required for mapping predictions, because for any point on the 1 km^2 grid, we need to identify which sampled rainfall station is the closest. If the specific location where we are going to predict is closer to a sampled location that belongs to cluster 1, we will use this model for the prediction. In ST kriging and TP splines, we also borrow the historical covariate of the closest station. We also analysed the alternative of providing a covariate through inverse distance weighting from a neighbour of five stations but this method did not outperform the previous one. Mapping predictions and standard errors in Navarre require matching the results obtained simultaneously on a particular day with the appropriate cluster for every point on the grid.

Figures 3 and 4 summarize overall predictions for 2010. Obviously, there is a high proximity between real data and estimated predictions at the manual stations for the three models, specially in ST kriging, but the overall performance at the automated rainfall stations is very good for all models. The use of clusters occasionally allows a better approximation.

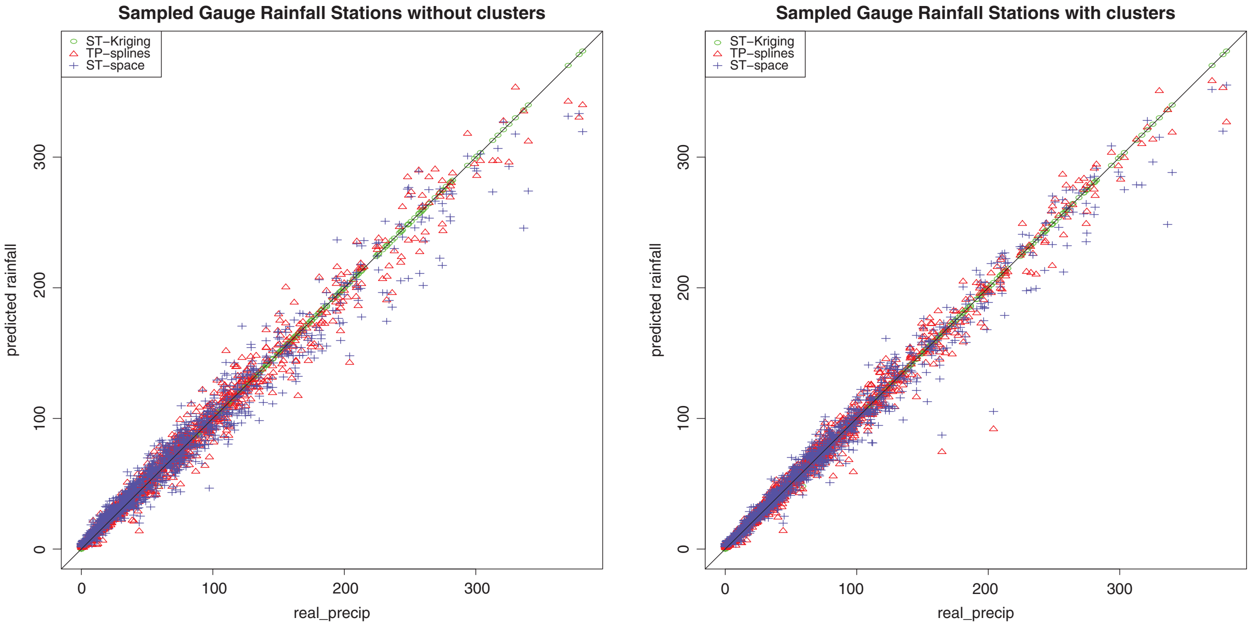

Figure 3. Monthly predictions of accumulated daily rainfall *versus* real data for the 87 manual gauge stations without clusters (left) and with clusters (right) obtained with the three models.

For numerical comparisons of the results, we present Tables 1 and 2 for the manual rainfall stations and Tables 3 and 4 for automated ones, both without and with clustering, respectively. These tables show the monthly and yearly observed and predicted accumulated daily rainfall and the empirical square root of the total mean squared error (MSE) by month and year obtained with the three spatiotemporal models. The MSE estimates the distance between real data and predictions; lower values are therefore preferable. As usual, the square root is calculated to preserve the same units of the variable under study. The square root of the MSE by month is given by:

$$\mathrm{RMSE}_{\mathrm{month}}\left(\hat{z}\right)^{j}=\sqrt{\frac{1}{n}\sum_{s=s_1}^{s_n}\sum_{t=1}^{T_{\mathrm{month}}}\left(z_{st}-\hat{z}_{st}^{j}\right)^{2}}$$

where month represents the 12 months of the year, s indicates the rainfall station, n can be 87 or 33 depending on manual or automated stations, t indicates the day and T_{month} is the total number of days of every month. The index j takes the three values k, ts and ss of the analysed models, representing ST-kriging, TP-spline and S-space

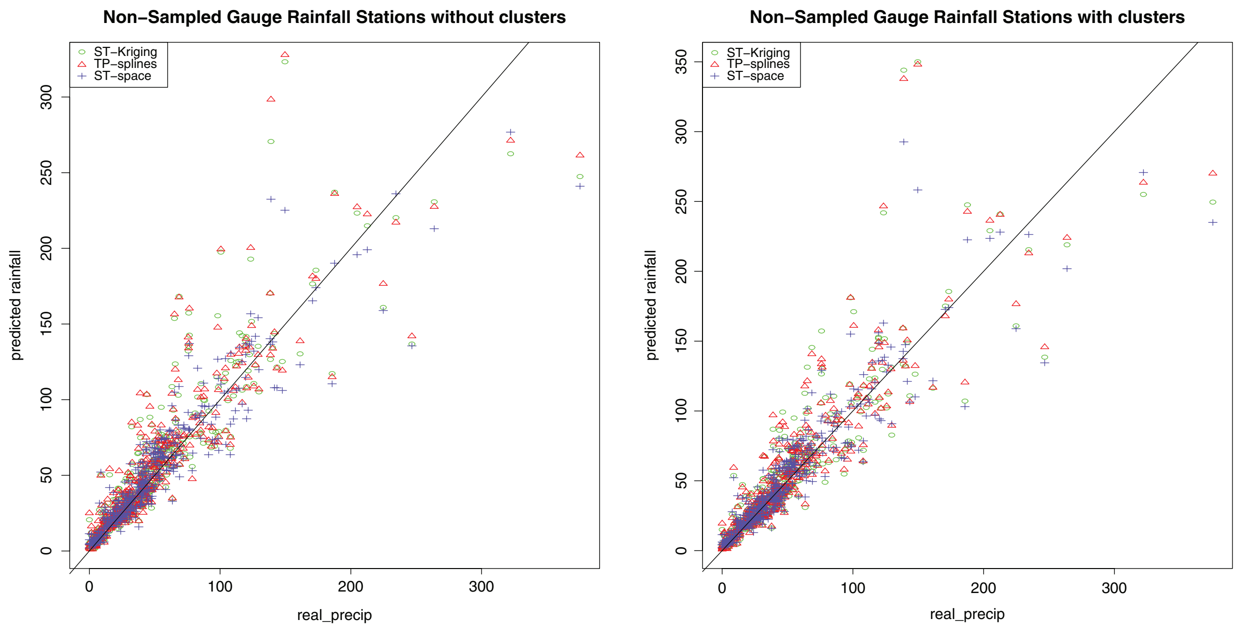

Figure 4. Monthly predictions of accumulated daily rainfall *versus* real data for the 33 automated gauge stations without clusters (left) and with clusters (right) obtained with the three models.

Table 1. Observed and predicted total rainfall and mean squared error by month (M) and year (Total) with the three statistical models without clusters in 87 manual rainfall gauges in 2010.

M	Observed	Predictions without clusters			RMSE		
		ST kriging	TP splines	S-space	ST kriging	TP splines	S-space
J	12854.31	12852.70	12895.17	12837.81	0.10	8.92	12.03
F	6756.01	6754.50	6783.19	6744.20	0.10	6.88	8.23
M	4433.81	4432.20	4449.41	4445.84	0.10	3.06	5.38
A	4478.20	4476.10	4493.52	4478.58	0.10	4.87	6.83
M	8086.32	8075.29	8121.85	8075.12	0.47	7.84	10.04
J	7898.80	7897.90	7962.74	7873.82	0.00	10.43	12.26
J	2776.01	2770.10	2809.07	2821.95	0.22	6.52	9.68
A	1842.24	1835.42	1880.31	1864.23	0.22	4.12	5.17
S	3376.80	3372.90	3414.61	3373.15	0.17	7.78	6.20
O	8180.60	8180.65	8231.80	8194.42	0.00	8.88	9.41
N	12694.94	12694.92	12731.00	12696.71	0.00	12.01	11.80
D	7513.51	7513.68	7535.79	7527.72	0.00	9.93	9.49
Total	80891.55	80856.36	81308.46	80933.53	0.62	27.78	31.88

Table 2. Observed and predicted total rainfall and mean squared error by month (M) and year (Total) with the three statistical models with clusters in 87 manual rainfall gauges in 2010.

M	Observed	Predictions with clusters			RMSE		
		ST kriging	TP splines	S-space	ST kriging	TP splines	S-space
J	12854.31	12852.71	12873.44	12848.30	0.10	6.41	10.48
F	6756.01	6754.51	6768.52	6719.22	0.10	5.41	6.81
M	4433.81	4432.21	4440.92	4428.11	0.10	2.84	4.38
A	4478.20	4476.10	4487.68	4522.03	0.10	3.83	6.32
M	8086.32	8075.33	8094.73	8036.70	0.47	7.23	8.38
J	7898.80	7897.90	7948.32	7844.90	0.00	8.87	11.69
J	2776.01	2770.11	2786.14	2773.72	0.22	4.40	2.60
A	1842.24	1835.45	1861.37	1885.01	0.22	3.23	5.17
S	3376.80	3372.90	3416.29	3370.01	0.17	6.56	5.87
O	8180.60	8180.60	8220.89	8183.33	0.00	7.01	8.59
N	12694.94	12694.94	12709.27	12710.59	0.00	9.41	9.10
D	7513.51	7513.51	7548.29	7515.44	0.00	5.98	7.37
Total	80891.55	80856.27	81155.86	80837.36	0.62	21.69	26.48

prediction models, respectively. The total MSE for the whole year is defined as the square root of the sum of the MSE (RMSE) by month. It is defined as:

$$\text{RMSE}\left(\hat{z}\right)^{j} = \sqrt{\sum_{\text{month}=1}^{12} \text{MSE}_{\text{month}}\left(\hat{z}\right)^{j}}$$

Theoretically, ST kriging should have a RMSE equal to zero because it is an exact interpolator, but for initiating the estimation process in all the models, a minimum of 2 mm needs to be observed across the whole region. Otherwise, a zero prediction will be given without running any programme. Therefore, in ST kriging, small differences are expected between observed and predicted data.

Tables 1 and 2 present the sum of the predictions by month and year for the 87 manual sampled stations and the RMSE calculated with the three models without and with clusters, respectively. There is a great proximity between predicted and observed data in all models by month and by year, particularly for the TP and S-space models that are not exact interpolators. When using clusters, the differences between predictions and observed data are even smaller, particularly for the S-space model. For the 87 sampled stations for the whole year, the observed rainfall is equal to 80891.55 mm, and the total predicted rainfall by the S-space model without clusters is equal to 80933.53 mm whereas using clusters it is equal to 80837.36 mm. The yearly and many of the monthly RMSEs of the S-space model are the biggest among the three models, with and without clusters, but this result is expected when predicting sampling data, because kriging is an exact interpolator and the thin-spline model is quite similar to ST kriging. Table 2 shows how the use of clusters in sampled data improves the predictions and reduces the RMSE, except for kriging because in manual sampled locations it provides the same predictions as the observed data with or without clusters.

When using data from automated non-sampled rainfall stations for predictions, the S-space model outperforms the other two models, with or without clusters (see Tables 3 and 4). Once again, monthly and yearly totals are quite similar to the observed rainfall in all models with and without clusters, and the S-space model provides the closest yearly prediction to the real data. That is, the

Table 3. Observed and predicted total rainfall and mean squared error by month (M) and year (Total) with the three statistical models without clusters in 33 automated rainfall gauges in 2010.

M	Observed	Predictions without clusters			RMSE		
		ST kriging	TP splines	S-space	ST kriging	TP splines	S-space
J	2984.00	3601.02	3592.84	3260.04	24.79	25.18	20.31
F	2010.50	2282.62	2284.64	2075.86	16.21	16.23	14.02
M	1409.60	1457.79	1467.67	1370.80	8.39	8.71	8.84
A	1240.40	1265.72	1313.85	1293.51	8.66	8.88	8.57
M	1856.10	2002.09	2026.51	1891.21	22.68	22.90	20.47
J	2113.30	2309.69	2282.66	2142.95	17.54	18.15	16.60
J	921.30	977.88	999.63	1050.98	14.63	15.67	15.36
A	272.70	334.93	368.75	339.14	5.91	6.22	5.94
S	722.90	812.44	804.89	731.71	11.73	11.92	11.80
O	2235.40	2481.07	2513.92	2349.28	19.27	18.64	17.04
N	2826.30	2870.49	2948.60	2774.39	23.17	23.26	21.25
D	1800.60	1963.77	2032.74	1948.00	17.07	16.86	15.67
Total	20393.10	22324.81	22560.55	21207.75	58.57	59.20	53.41

Table 4. Observed and predicted total rainfall and root mean squared error by month (M) and year (Total) with the three statistical models with clusters in 33 automated rainfall gauges in 2010.

M	Observed	Predictions with clusters			RMSE		
		ST kriging	TP splines	S-space	ST kriging	TP splines	S-space
J	2984.00	3473.42	3469.59	3300.93	25.69	26.82	21.51
F	2010.50	2194.46	2239.63	2049.88	16.11	16.97	14.54
M	1409.60	1419.17	1434.03	1339.08	8.39	8.73	8.73
A	1240.40	1296.47	1300.18	1330.65	9.86	9.90	10.38
M	1856.10	1956.58	1959.43	1840.84	22.97	23.06	20.88
J	2113.30	2249.81	2255.34	2158.80	20.38	19.90	18.50
J	921.30	972.95	1011.89	974.36	15.43	16.67	15.36
A	272.70	336.73	344.49	378.16	6.23	6.17	6.31
S	722.90	806.23	788.03	731.70	12.00	12.11	11.80
O	2235.40	2484.99	2437.77	2405.55	19.27	18.64	17.04
N	2826.30	2830.45	2866.54	2782.83	24.59	24.63	22.86
D	1800.60	1934.92	1977.40	1913.99	17.45	17.62	16.93
Total	20393.10	21956.18	22084.32	21206.77	61.08	61.94	56.13

total observed rainfall in 2010 from the 33 stations is equal to 20393.10 mm. The prediction from the same 33 stations without clusters is equal to 22324.81, 22560.55 and 21207.75 mm and the prediction with clusters is equal to 21956.18, 22084.32 and 21206.77 mm for ST-kriging, TP-spline and S-space models, respectively. The yearly results in Table 3 show that the S-space RMSE is lower than the TP-spline and ST-kriging RMSE. Therefore, for new rainfall gauges, the S-space model is the best model because it provides the lowest RMSE with regard to ST kriging or TP splines, with or without clusters. The kriging and the TP-spline model have lower MSEs in the sampled data, but the S-space model provides the best predictions for new observations.

Figures 5 and 6 provide maps of the predictions and standard errors calculated for Navarre for 9 January 2010 without and with clusters, respectively. Standard errors are calculated for all the models by the specific packages. Estimation procedures are different, but the S-space model provides lower ranges of variation than the other models, i.e. between 1.4 and 2.8 mm without clusters and between 1 and 4 mm with clusters. ST kriging gives standard errors between 2 and 4 mm without clusters or between 0 and 5 mm with clusters. The TP splines provide standard errors between 0 and 7 mm in both cases. Maps of predictions and standard errors provide different patterns for the three models and above all when looking at the presence of clusters. In the predictions, the north–west direction has a darker colour, indicating higher precipitation. Ranges vary between 0 and 20 mm in many maps, except for ST kriging and TP splines with clusters. The numbers printed in the figures indicate the prediction and standard error of an arbitrary point.

Finally, a continuous ranked probability score (CRPS) has been calculated to assess the prediction performance of the analysed models (Gneiting and Raftery, 2007; Gneiting *et al.*, 2007). If the predictive distribution is normal with mean μ and variance σ^2, the daily CRPS$((N(\mu, \sigma^2), z_{st})$ is defined as

$$\mathrm{CRPS}\left(N\left(\mu,\sigma^2\right), z_{st}\right) = \sigma\left(\frac{z_{st}-\mu}{\sigma}\left(2\Phi\left(\frac{z_{st}-\mu}{\sigma}\right)-1\right) + 2\phi\left(\frac{z_{st}-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right)$$

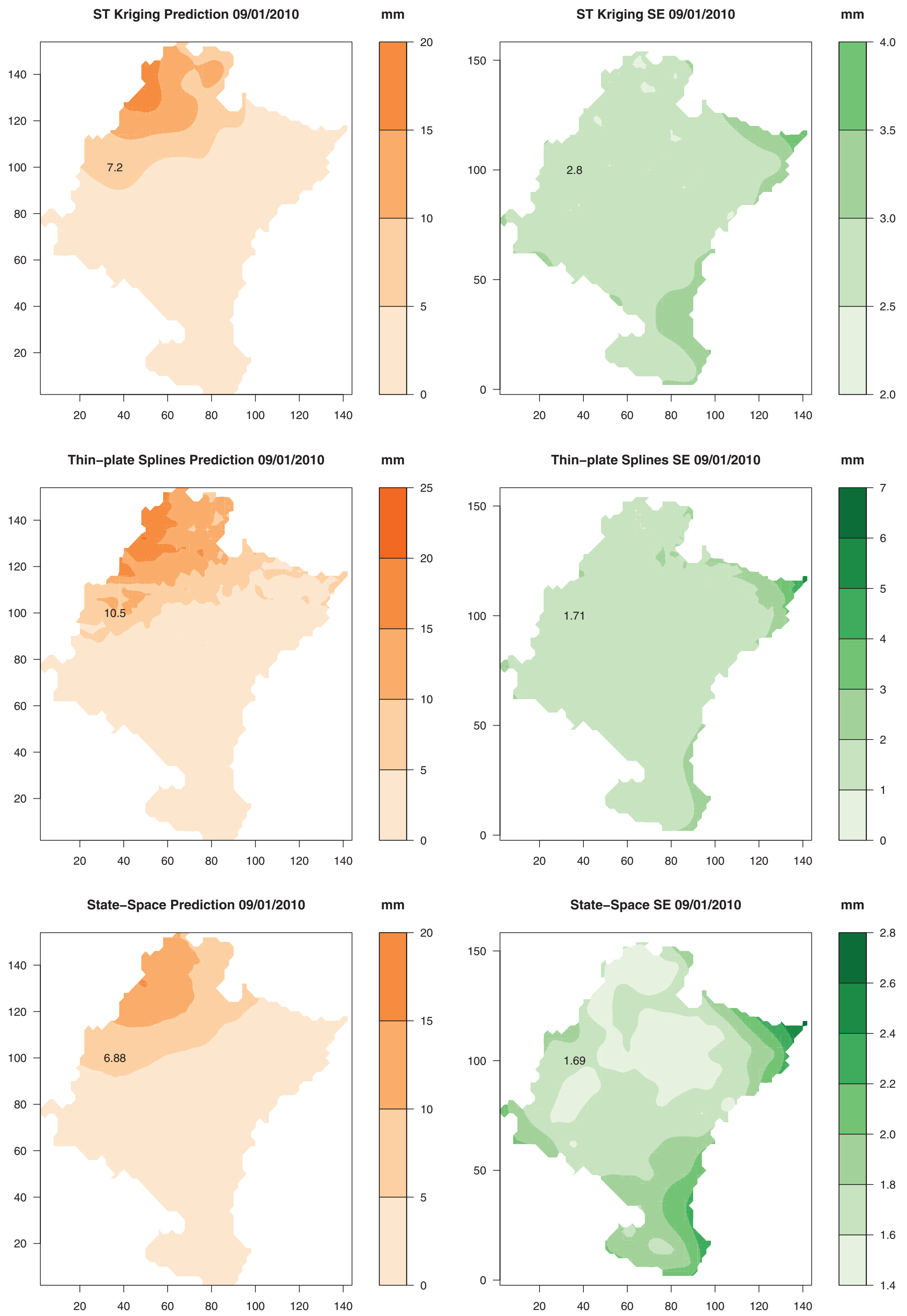

Figure 5. On the left, spatiotemporal kriging, thin-plate splines and state-space predictions for 9 January 2010 in Navarre. On the right, the corresponding maps of standard errors. Models have been fitted without considering any cluster.

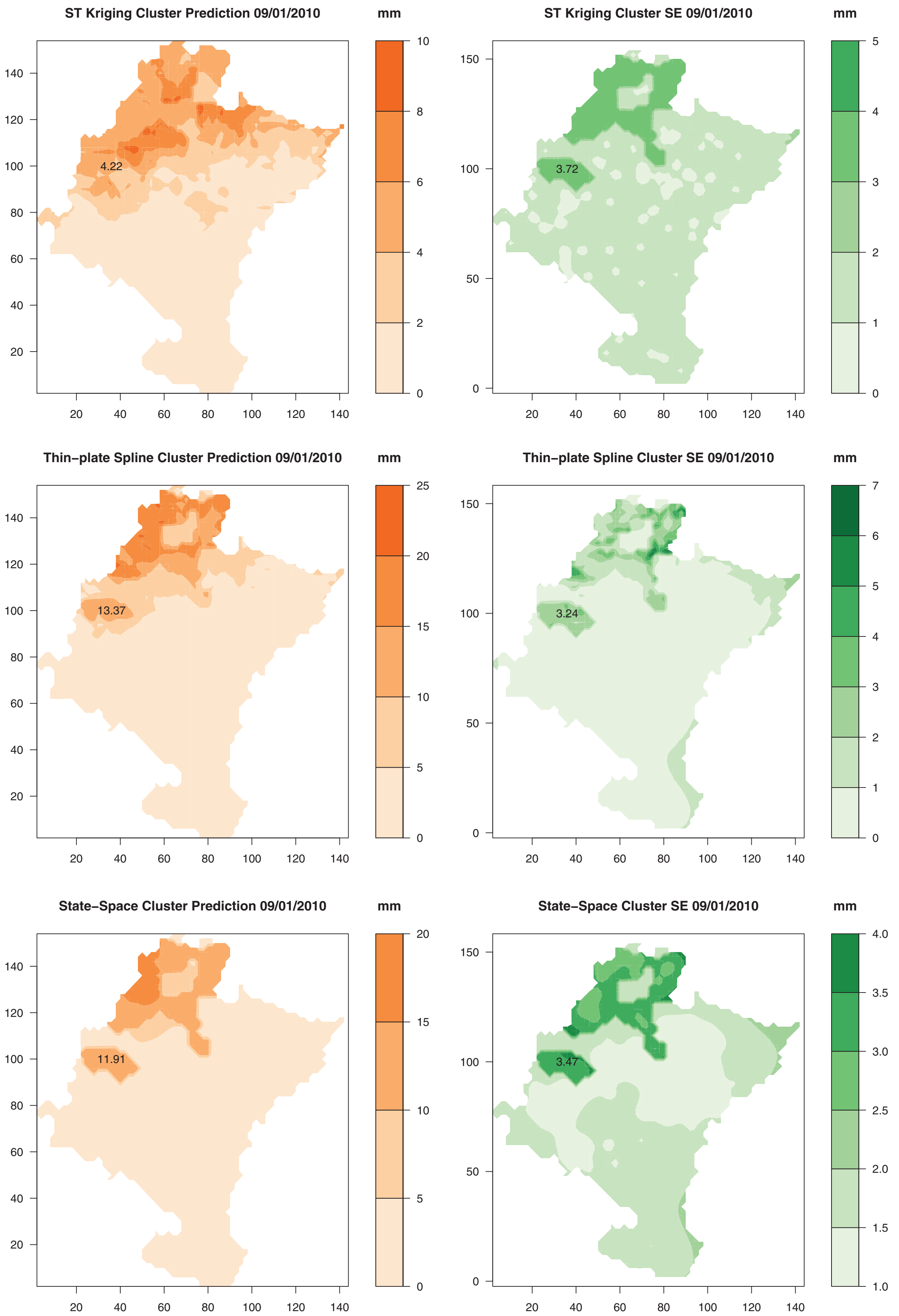

Figure 6. On the left, spatiotemporal kriging, thin-plate splines and state-space predictions for 9 January 2010 in Navarre. On the right, the corresponding maps of standard errors. Models have been fitted considering clusters.

Table 5. Continuous ranked probability score.

	Average CRPS manual			Average CRPS automated		
	ST kriging	TP splines	S-space	ST kriging	TP splines	S-space
Without clusters	–	4.81	6.45	10.12	10.74	9.66
With 2 var. in clusters	–	3.81	4.77	10.67	10.55	8.65

where z_{st} is the observed rainfall value, and ϕ and Φ denote the probability density function and the cumulative density function of the standard normal distribution, respectively. The average CRPS is defined as:

$$\text{CRPS} = \frac{1}{nT}\sum_{s=1}^{n}\sum_{t=1}^{T}\text{CRPS}\left(N\left(\mu,\sigma^2\right), z_{st}\right)$$

Results in Table 5 are in accordance with Tables 1–4. At the manual gauge stations, the ST-kriging and the TP splines outperform the S-space model, and the use of clusters improves the prediction performance in all models. At the automated gauge stations, S-space models outperform the other two models with and without clusters. The use of clusters improves the predictive performance in all gauge stations, except with the ST-kriging model, although the differences are very small. Once again the S-space model provides a good performance although the ST-kriging and the TP-spline models are very competitive.

4. Conclusions

Interpolation of rainfall data is a necessary task in precision agriculture, environmental studies or meteorology. Methods based on inverse distance weighting, kriging, cokriging or TP splines are common tools for spatial interpolation, but usually they do not incorporate the temporal dimension. Undoubtedly, the simultaneous use of space and time in interpolation models increases the difficulty of providing predictions because spatiotemporal models are not as widely used as spatial or temporal models. In this work, we propose two natural and simple extensions to kriging and TP splines to incorporate time dependence into the statistical model. Firstly, a new variable is defined as the average of the historical rainfall at the same rainfall gauge stations in neighbourhoods of 5 days, yet other numbers of days can also be used. Secondly, this covariate is introduced as a new explanatory variable into the model, in addition to the planar coordinates and the elevation. The third model analysed in this paper is the space-state model, which incorporates time dependence as a random walk in a new state equation of a hierarchical model. This approach offers great flexibility and incorporates time as a new dynamic variable without averaging neighbourhood information. It is a very natural and simple way of introducing a spatiotemporal model, yet some improvements can be made using clusters.

Daily rainfall can be difficult to interpolate, particularly for days with frequent storms or changing weather. Therefore, we propose to define clusters according to the accumulated and historical rainfall. In these cases, models need to be estimated and predictions need to be made according to these clusters. Model fitting was performed using the free statistical software R (R Development Core Team, 2012) but additional programming is necessary to yield predictions at unobserved locations, to define covariates for non-sampled locations, such as elevation and historical data that are taken from the closest sampled station, to map them, and to estimate the mean squared prediction error. This code is available from the authors.

To check the performance of these models, we used data from the sampled manual rainfall stations and an additional set of 33 automated rainfall gauge stations, located at different sites. We compared the predicted rainfall calculated with the six models, three with clusters and another three without clusters, with the observed daily rainfall during the 365 days of the year 2010. The performance of the three models is clearest when the predictions are made for new data. From the precision point of view, the S-space model is the best because in new locations its RMSE is the lowest with or without clusters. From the point of view of accuracy, the predictions are the closest to the real values with and without clusters, too. Using clusters does not guarantee better predictions in new locations, nor the lowest RMSE, but on specific days it can be appropriate. Unfortunately, these results cannot be clearly shown in the predictions made over the sample data due to the exact interpolation property of kriging and the similarity of the TP splines with kriging. Moreover, the CRPS ranks the S-space model as the best model when predictions are made from non-sampled stations. The S-space model performs the best, but the TP splines and ST-kriging models are still very competitive and can be useful when dealing with precision agriculture.

Acknowledgements

The authors would like to thank the associate editor and the referees for the constructive comments that led to improvement the paper. We also thank Francisco Cerveto and the Department of 'Desarrollo Rural y Medio Ambiente' of the Government of Navarre. This research was supported by the Spanish Ministry of Science and Education (project MTM2011-22664 jointly sponsored with Feder grants). Research reported in this publication was also supported by King Abdullah University of Science and Technology (KAUST).

References

Ailliot P, Thompson C, Thomson P. 2009. Space-time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *Appl. Stat.* **58**: 405–426.

Amisigo BA, van de Giesen NC. 2005. Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series. *Hydrol. Earth Syst. Sci.* **9**: 209–224.

Apanasovich TV, Genton MG, Sun Y. 2012. A valid Matern class of cross-covariance functions for multivariate random fields with any number of components. *J. Am. Stat. Assoc.* **107**: 180–193.

Carrera-Hernández JJ, Gaskin SJ. 2007. Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. *J. Hydrol.* **336**: 231–249.

Cowpertwait PSP, Lockie T, Davis MD. 2006. A stochastic spatial-temporal disaggregation model for rainfall. *J. Hydrol. (NZ)* **45** (1): 1–12.

Cox DR, Isham V. 1988. A simple spatial-temporal model of rainfall. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **415**(1849): 317–328.

Daly C, Neilson RP, Phillips DL. 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *J. Appl. Meteorol.* **33**: 140–158.

Duchon J. 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive Theory of Functions of Several Variables*, Schempp W, Zeller K (eds). Springer: Berlin and Heidelberg, Germany, 85–100.

Fasso A, Cameletti M, Nicolis O. 2007. Air quality monitoring using heterogeneous networks. *Environmetrics* **18**: 245–264.

Fields Development Team. 2006. *Fields: Tools for Spatial Data*, National Center for Atmospheric Research: Boulder, CO. Retrieved May 23, 2014. http://www.image.ucar.edu/˜nychka/Fields.

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**(477): 359–378.

Gneiting T, Genton M, Guttorp P. 2007. Geostatistical space-time models, stationarity, separability and full symmetry. In *Statistical Methods for Spatio-Temporal Systems*, Finkenstadt B, Held L, Isham V (eds). Chapman and Hall/CRC Press: Boca Raton, FL, 151–175.

Hartigan JA. 1975. *Clustering Algorithms*. John Wiley & Sons, Inc.: New York, NY.

Herrera S, Gutiérrez JM, Ancell R, Pons MR, Frías MD, Fernández J. 2012. Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02). *Int. J. Climatol.* **32**: 74–85.

Jeong DI, St-Hilaire A, Ouarda TBMJ, Gachon P. 2012. Comparison of transfer functions in statistical downscaling models for daily temperature and precipitation over Canada. *Stoch. Environ. Res. Risk Assess.* **26**: 633–653.

Kleiber W, Katz RW, Rajagopalan B. 2012. Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resour. Res.* **48**: W01523, DOI: 10.1029/2011WR011105.

Kokic P, Crimp S, Howden M. 2011. Forecasting climate variables using a mixed-effect state-space model. *Environmetrics* **22**: 409–419.

Kurtzman D, Navon S, Morin E. 2009. Improving interpolation of daily precipitation for hydrologic modelling: spatial patterns of preferred interpolators. *Hydrol. Processes* **23**: 3281–3291.

Li C, Singh VP, Mishra AK. 2012. Simulation of the entire range of daily precipitation using a hybrid probability distribution. *Water Resour. Res.* **48**: W03521, DOI: 10.1029/2011WR011446.

Martin-Vide J. 2004. Spatial distribution of daily precipitation concentration index in Peninsular Spain. *Int. J. Climatol.* **24**: 959–971.

Matheron G. 1980. Splines and kriging: their formal equivalence. Internal Report, Centre de Geostatistique: Ecole des Mines de Paris, Fontainebleau, 20 p.

Mehrotra R, Westra S, Sharma A, Srikanthan R. 2012. Continuous rainfall simulation: 2. A regionalized daily rainfall generation approach. *Water Resour. Res.* **48**: 1–16 W01536.

Militino AF, Ugarte MD. 2001. Assessing the covariance function in geostatistics. *Stat. Probab. Lett.* **52**(1): 199–206.

Militino AF, Palacios MB, Ugarte MD. 2001. Robust predictions of rainfall in Navarra, Spain. In *Geoenv III – Geostatistics for Environmental Applications*. Kluwer Academic Publishers: Dordrech, The Netherlands, 79–90.

Militino AF, Palacios MB, Ugarte MD. 2003. Robust trend parameters in a multivariate spatial linear model. *Test* **12**(2): 445–457.

Moral FJ. 2010. Comparison of different geostatistical approaches to map climate variables: application to precipitation. *Int. J. Climatol.* **30**: 620–631.

Newlands NK, Davidson A, Howardc A, Hilld H. 2011. Validation and inter-comparison of three methodologies for interpolating daily precipitation and temperature across Canada. *Environmetrics* **22**: 205–223.

Nychka D. 1995. Splines as local smoothers. *Ann. Stat.* **23**(4): 1175–1197.

R Development Core Team. 2012. *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna. ISBN 3-900051-07-0. Retrieved May 23, 2014. http://www.R-project.org.

Ribeiro PJ, Diggle PJ. 2001. geoR: a package for geostatistical analyses. *R-News* **1**(2): 15–18.

Ruppert D, Wand MP, Carroll RJ. 2003. *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press: Cambridge, UK.

Stein M. 1991. A kernel approximation to the kriging predictor of a spatial process. *Ann. Inst. Stat. Math.* **43**(1): 61–75.

Tait A, Henderson R, Turner R, Zheng X. 2006. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *Int. J. Climatol.* **26**: 2097–2115.

Vaze J, Teng J, Chieu FHS. 2011. Assessment of GCM simulations of annual and seasonal rainfall and daily rainfall distribution across south-east Australia. *Hydrol. Processes* **25**: 1486–1497.

Weymouth G, Mills GA, Jones D, Ebert EE, Manton MJ. 1999. A continental-scale daily rainfall analysis system. *Aust. Meteorol. Mag.* **48**(3): 169–179.

Wood SN. 2003. Thin plate smoothing splines. *J. R. Stat. Soc. B* **65**(Part 1): 95–114.

Xie P, Yatagai A, Chen M, Hayasaka T, Fukudhima Y, Liu C, Yang S. 2007. A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeorol.* **8**: 608–626.

