Rainfall modeling using latent Gaussian random fieldsan estimation method

Anastassia Baxevani, Jan Lennartsson

Pub. Inst. Stat. Univ. Paris
59, fasc.1-2, 2015, 7-24

Rainfall modeling using latent Gaussian random fields - an estimation method

Anastassia Baxevani*, Jan Lennartsson†

University of Cyprus and University of Gothenburg†*

Abstract: We propose a method for estimating the parameters in a latent Gaussian field used for modeling daily rainfall. For the rainfall variable, a monotonic transformation is applied to achieve marginal normality, thus, defining a latent variable, with zero rainfall values corresponding to censored values below a threshold. Methodology is presented for model estimation and validation illustrated using accumulated daily rainfall data from a network of 14 stations in the southern Sweden. Performance of the model is judged through its ability to accurately reproduce a series of temporal and spatial dependence measures.

1. Introduction

Spatio-temporal variability of rainfall is an important source of variability that must be properly taken into account. Daily and hourly stochastic rainfall models provide useful supporting roles in the analysis of risk and vulnerability within hydrological and hydraulic systems. These roles include the generation of synthetic rainfall records when there are none, extrapolation of short observed records, and temporal downscaling of observed records.

Typically daily rainfall at a single site is represented as a mixture of two distributions in a parametric, nonparametric or semi-parametric framework. One is a discrete binary part modeling the wet or dry state any given day and the other is a continuous distribution modeling the nonzero precipitation amounts on wet days.

A lot of effort has been devoted to statistically model high precipitation amounts which led to using distribution families that are right skewed. Different distributions have been used, such as exponential [29], gamma [21], mixed exponential [31] and truncated and power transformed normal distributions [4] and [20]. These distributions behave reasonably well in terms of reproducing average characteristics of precipitation but none of them succeeds in representing extremes. Besides parametric models nonparametric approaches have also been employed. Synthetic rainfall is sequentially sampled from historical records with replacement. Several limitations especially with respect to extremes have been recognized, inherent of the sampling scheme, see [14] and corrected via nonparametric kernel density estimator. Reproducing the entire range of daily rainfall has been studied in [30], [14], [23] and [19] in which compound distributions are used for modeling precipitation amount.

AMS 2000 subject classifications: Primary 62P12, 62M30; secondary 62G60

Keywords and phrases: rainfall modelling, space-time latent Gaussian random field, covariance structure, anamorphosis transformation

Recent interest in the field has moved away from individual location models to models that are capable of generating spatially and temporally correlated rainfall fields. This is especially difficult, given the highly variable nature of precipitation. It is essentially crucial to capture the domain aggregated behavior of rainfall intensity and the dry and wet spells that play an important role in hydraulic applications. There are a number of approaches to spatio-temporal modeling. The techniques include resampling, see [7], use of weather stations, see [18] and [10], or generalized linear models, [32]. Recently, [8] presented RainSim, a stochastic rainfall field generator where rainfall fields are sampled from a spatial-temporal Neyman-Scott rectangular pulses process. Hidden Markov models were used for occurrence in [18] and for intensity [1]. [2] and [27] on the other hand assumed that both parts of the precipitation process - occurrence and intensity - can be modeled using the same latent Gaussian process. Various well known transformation functions have been suggested, for example [3] use a quadratic power function, [27] use a power function, and [2] use a power-exponential function to transform the Gaussian value to the desired intensities. Recently, [22] applied a two-part transformation function, with one part being the inverse of a standard normal distribution and the marginal part be given by a gamma distribution and [5] used a similar model but with a hybrid gamma with generalized Pareto for the marginal distribution.

In this paper we assume that a latent Gaussian variable can be used to model the rainfall with dry conditions corresponding to censored values below a given threshold and wet conditions corresponding to transformed values above the threshold. A methodology is presented for estimating the parameters of the latent Gaussian field. The model is fitted to data from a network of 14 stations in southern Sweden. Validation is through the models ability to reproduce a series of temporal and spatial dependence measures.

2. Mathematical formulation of the model

The aim of this model is to statistically describe properties of rainfall fields. We will assume that at each location in space and day of the year, the rainfall amount accumulated over that day is a realization of a random variable. We suppose that we have a data set that consist of N distinct realizations of the precipitation process at K locations over a period of D days. This data set is denoted $\{y_n(\mathbf{s},t), n = 1,\dots,N; \mathbf{s} = (x_i, y_i), i = 1,\dots,K, t = 1,\dots,D\}$. We shall assume that this dataset constitutes a set of independent realizations of a transformed Gaussian random field (G.r.f) $\{Y(\mathbf{s},t)\}$, which relates to a latent G.r.f $\{Z(\mathbf{s},t)\}$ through

$$Y(\mathbf{s},t) = \begin{cases} \psi(Z(\mathbf{s},t)), & \text{if } Z(\mathbf{s},t) > \nu \\ 0, & \text{if } Z(\mathbf{s},t) \leq \nu, \end{cases} \tag{2.1}$$

where ψ is a non-decreasing function, usually referred to as anamorphosis transformation, used to transform the Gaussian values to the right skewed marginal distribution of the rainfall amounts and $\nu \in \mathbb{R}$ a threshold usually set to be zero.

The finite dimension distributions of a G.r.f Z are uniquely characterized by the mean value μ_Z and the covariance structure C_Z which are defined by:

$$\mu_Z(\mathbf{s},t) = \mathbb{E}[Z(\mathbf{s},t)] \tag{2.2}$$

and

$$C_Z(\mathbf{s},\mathbf{s}+\mathbf{h},t,t+\tau) = \mathbb{E}[(Z(\mathbf{s},t)-\mu_Z(\mathbf{s},t))\,(Z(\mathbf{s}+\mathbf{h},t+\tau)-\mu_Z(\mathbf{s}+\mathbf{h},t+\tau))]. \tag{2.3}$$

Hence from (2.1), is easy to see that the rainfall field Y is uniquely characterized by μ_Z, C_Z and the anamorphosis transformation ψ.

Notice that in the above scheme, we can also obtain the occurrence of wet events, i.e. days with positive rainfall, $O(\mathbf{s},t)$ by means of the latent G.r.f. by

$$O(\mathbf{s},t) = \begin{cases} 1, & \text{if } Z(\mathbf{s},t) > \nu \\ 0, & \text{if } Z(\mathbf{s},t) \leq \nu. \end{cases} \tag{2.4}$$

3. Model fitting

The aim is then to describe guidelines for fitting the model to the data. We can think of this procedure as consisting of two essentially independent parts: choosing an appropriate anamorphosis function and estimating the first two moments of the random field. The question of modeling the cumulative distribution function (cdf) of the rainfall intensity has been widely developed in the literature. In this work we shall concentrate on the problem of estimating the parameters of the latent G.r.f. Z. Before we proceed any further, we would like to emphasize that the usual procedure of fitting a mean and a covariance function to raw data does not apply anymore, since we do not observe realizations of the G.r.f. Z but of the censored transformed r.f. Y.

3.1. Mean estimation

There is a clear link between the mean function of the latent G.r.f. Z and the rainfall data Y through the frequency of wet days:

$$P(\text{wet } t \text{ day at location } \mathbf{s}) = P(Y(\mathbf{s},t) > 0) = P(Z(\mathbf{s},t) > \nu) = \Phi\left(\frac{\mu_Z(\mathbf{s},t)-\nu}{\sqrt{\mathbb{V}arZ(\mathbf{s},t)}}\right), \tag{3.1}$$

where $\Phi(\cdot)$ denotes the cdf of a standard normal random variable and ν is the threshold level in (2.1). For simplicity from now on we shall set $\nu = 0$ and assume the field Z has been standardized, i.e. $\mathbb{V}arZ(\mathbf{s},t) = 1$. Hence, (3.1) simplifies to

$$P(\text{wet } t \text{ day at location } \mathbf{s}) = \Phi(\mu_Z(\mathbf{s},t)). \tag{3.2}$$

Therefore, the mean function μ_Z is estimated by inverting (3.2) once estimates of the probability in the right hand side of (3.2) are obtained.

3.2. Covariance estimation

We turn now to the problem of estimating the covariance structure in the latent G.r.f. One way of estimating the covariance coefficients at a particular time lag for a censored Gaussian variable is by numerically maximizing the likelihood of the observed bivariate histogram of the censored latent variable, see [15] or by maximizing a modified version of the likelihood for censored values, see [13]. [16] proposed a method for estimating the covariance function of the latent field which comprises of computing the empirical covariance of raw data, then fitting a positive definite function to it, computing the inverse of this function through the Hermite polynomial expansion of the anamorphosis function, fitting a positive definite function to it and then reversing it again with the use of the same Hermite polynomial expansion of the anamorphosis function to finally obtain an estimate of the desired covariance. We propose an alternative method of moments approach, by inverting the theoretical expression for the mean of the censored cross product. We turn to this next.

The following relation holds for a bivariate Gaussian random variable $\mathbf{Z} = (Z_i, Z_j)$ with mean $\mu = (\mu_i, \mu_j)$ and correlation ρ_{ij} (unit variances):

$$\mathbb{E}[Z_i^+ Z_j^+] = \int_0^\infty g(x; \mu, \rho_{ij})\, dx, \tag{3.3}$$

where $Z_i^+ = max(Z_i, 0)$ and the function g is given by:

$$\begin{aligned} g(x; \mu, \rho_{ij}) &= x\phi(x - \mu_i) \left[(\rho_{ij}(x - \mu_j) + \mu_i)\, \Phi\left(\frac{\mu_i + \rho(x - \mu_j)}{\sqrt{1 - \rho_{ij}^2}} \right) + \right. \\ &+ \left. \sqrt{1 - \rho_{ij}^2}\, \phi\left(\frac{\mu_i + \rho(x - \mu_j)}{\sqrt{1 - \rho_{ij}^2}} \right) \right]. \end{aligned} \tag{3.4}$$

In the specific case of $\mu_1 = \mu_2 = 0$ the expectation in (3.3) simplifies to the well known relation

$$\mathbb{E}[Z_i^+ Z_j^+] = \frac{1}{2\pi} \left(\rho_{ij} \left(\frac{\pi}{2} + \arcsin(\rho_{ij}) \right) + \sqrt{1 - \rho_{ij}^2} \right). \tag{3.5}$$

Therefore using (3.3) and (3.4), correlation $\rho_{ij}(\tau)$ between locations $\mathbf{s}_i$ and $\mathbf{s}_j$ and time lag τ (we have assumed stationarity of the covariance structure at least over the temporal component) can be estimated by minimizing

$$min_\rho \left| \overline{z_i^+ \cdot z_j^+} - \int_0^\infty g(x; \mu_{ij}, \rho_{ij})\, dx \right|, \tag{3.6}$$

where $\overline{z_i^+ \cdot z_j^+}$ denotes the average of the product of the transformed censored values at locations indicated by the subindices for the given time lag τ and μ_{ij} denotes

the vector $(\mu(\mathbf{s}_i,t),\mu(\mathbf{s}_j,t+\tau))$. Note that function g is not a simple function of pairwise correlations which can be analytically inverted. The integral in (3.6) needs to be computed using numerical integration. Moreover, a method for obtaining the estimates z^+ of the censored latent G.r.f. Z is also required.

Once we have computed the empirical correlations $\hat{\rho}_{ij}(\tau)$ (which, for variances equal unity, coincide with $\widehat{Cov}(Z(\mathbf{s}_i,t),Z(\mathbf{s}_j,t+\tau))$, the covariance parameters are estimated through the following method of moments minimization:

$$min_{\boldsymbol{\eta}} \sum_t \sum_{i\neq j} n_{obs}(i,j,\tau)\left(\hat{\rho}_{ij}(\tau) - C_Z((\mathbf{s}_i,\mathbf{s}_j,t,t+\tau;\boldsymbol{\eta}))\right)^2, \tag{3.7}$$

where C_Z is a parametric covariance with $\boldsymbol{\eta}$ denoting the set of parameters of C_Z and n_{obs} is the number of observations used in the estimation and which varies with location and time. Hereafter we shall denote by $\hat{x}$ the estimate of any quantity x. Notice that there is no theoretical restriction on the type of model covariance function to be used, except for temporal stationarity. The only other restrictions are imposed by the numerical complexity of the resulting minimization procedure.

3.3. Anamorphosis transformation

As already mentioned, the problem of modeling the marginal distribution has attracted a lot of attention. Over the years, rainfall intensity has been modeled using a Box-Cox transformation, see [6], a quadratic power transformation, see [15] and [13], or a power-exponential transformation of the censored Gaussian distribution, see [2]. In [23] a generalized Pareto (GP) distribution modeled heavy rainfall above a high level. [22] used a gamma distribution, while mixtures of exponential distributions were used in [31]. A gamma distribution alone, although flexible enough, is not quite adequate to model the tail of the distribution since it underestimates large values, see e.g. [22]. The fit improves when the so-called hybrid gamma and GP distribution is used, see [24], [14] and [5]. This hybrid distribution is a result of coupling a gamma distribution with a GP distribution and has its origin in the one introduced by [9], where Gaussian and GP distributions were stitched together.

Since in this work our main focus is in modeling the structure of the latent G.r.f we take as anamorphosis transformation the composite of the empirical distribution of the rainfall intensity with a censored Gaussian random variable. That is:

$$\psi(x) = (F^{emp})^{-1} \circ \Phi_{\mu_Z}(x), \quad x \in \mathbb{R} \tag{3.8}$$

where F^{emp} is the empirical cdf of the rainfall intensity and Φ_{μ_Z} is the cdf of a censored normal random variable with mean value μ_Z and variance unity, that is given by

$$\Phi_{\mu_Z}(x) = \begin{cases} \frac{\Phi(x-\mu_Z)-\Phi(-\mu_Z)}{\Phi(\mu_Z)}, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0, \end{cases} \tag{3.9}$$

where ϕ and Φ are the pdf and cdf of a standard normal random variable respectively. Therefore the transformation ψ in (2.1) is semi-parametric. The motivation for this choice of anamorphosis function is that we do not want any additional variability that would be caused by parameter estimation of the marginal distribution. For more on the anamorphosis transformation see [11].

4. Implementation and Example

4.1. Data

We consider a small network of $K = 14$ rainfall stations, denoted from now on by $\mathbf{s} = (x_i, y_i), i = 1, \ldots, K$, located in southern Sweden and covering an area from the West coast to the East coast in the Baltic sea. The stations have been numbered, see Figure 1. The data consist of accumulated daily rainfall during the month of July (we assume the rainfall process can then be thought as stationary in time) for $N = 51$ years from 1961- 2011. Less than 10% of observations were missing from each station. The observation network is quite dense with distance between the stations ranging from 30 km (distance between stations 10 and 11) to 290 km (stations 4 and 12). The climate in the area is dominated by the effects of the South Swedish highland, an area situated more than 200 meter above the ocean with the prevailing western winds resulting to orographic precipitation in the surrounding areas.

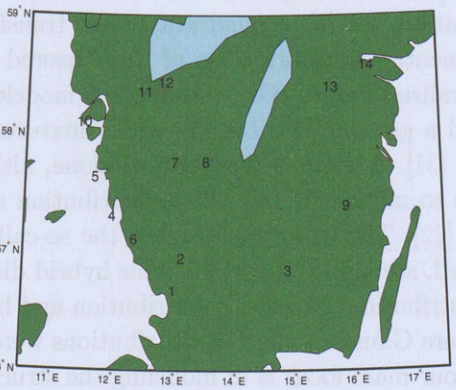

FIGURE 1. *The locations of the available weather stations.*

4.2. Marginal distribution

A simple unbiased estimator for any cdf is given by the corresponding empirical cdf. Denote $F_{\mathbf{s}}^{emp}$, for $\mathbf{s} = (x_i, y_i), i = 1, \ldots, K$ the estimator that is defined as

$$F_{\mathbf{s}}^{emp}(x) = \frac{1}{N_{\mathbf{s}}} \sum_{n=1}^{N_{\mathbf{s}}} \mathbb{I}_{\{y_n(x_{\mathbf{s}}) \leq x\}}, \tag{4.1}$$

where $\mathbb{I}_A$ denotes the indicator function of a set A, i.e. a function that equals unity when property A is satisfied and 0 otherwise and $N_{\mathbf{s}}$ is the number of available data at location $\mathbf{s}$.

It is easy to show that for each location, $F_{\mathbf{s}}^{emp}(x)$ is an unbiased estimator of the marginal cdf $F_{\mathbf{s}}(x)$, for all $x \in \mathbb{R}$. The computation of each $F_{\mathbf{s}}^{emp}$ results in K estimates for the empirical cdf, one for each location. Then the shape of each one of them enables us to choose an appropriate parametric or semi-parametric family of distributions. As we have mentioned before, we have decided not to perform this step at this time since we are mainly interested in evaluating the procedure of obtaining the distribution of the latent field. In [5], the authors have modeled the empirical distribution using a hybrid gamma with a generalized Pareto distribution and in [23] for the same data set, the authors used the composite of the empirical distribution with generalized Pareto for the excesses.

It remains to estimate the distribution of the censored Gaussian random variable in (3.9). For this we need to fist obtain estimates of the mean value μ_Z, which is the topic of the next subsection.

4.3. Mean function

Estimation of the mean value μ_Z is straightforward by simply inverting (3.2). An unbiased estimator of the probability of wet events is given by:

$$\hat{p}_{\mathbf{s},t} = \hat{P}(\text{wet } t \text{ day at location } \mathbf{s}) = \frac{1}{N_{\mathbf{s}}} \sum_{t=1}^{N_{\mathbf{s}}} \mathbb{I}_{\{y_n(\mathbf{s},t)>0\}}, \tag{4.2}$$

where $N_{\mathbf{s}}$ is defined as in (4.1).

Hence, a natural estimator of the mean function at location $\mathbf{s}$ and day t is

$$\hat{\mu}_Z(\mathbf{s},t) = \Phi^{-1}(\hat{p}_{\mathbf{s},t}). \tag{4.3}$$

The proportion of wet events over day t at location $\mathbf{s}$, $\hat{p}_{\mathbf{s},t}$, is estimated using (4.2) and then an estimate of μ_Z is obtained using (4.3). In Figure 2 we present the proportion of wet days for one example location the city of Halmstad (station 5 in Figure 1).

The mean estimates, as can be seen in Figure 2, exhibit seasonality which together with the spatial variability suggested different mean estimates at different locations and days of year. In order to interpolate the mean estimates in space, i.e., also at locations where there are no observations available, estimates of the mean $\mu_Z(\mathbf{s},\cdot)$ could be regressed on small-order harmonics, spatial coordinates and altitude. Additional covariates could also include distance to the coast as well as climate model output or broad scale atmospheric conditions. So we write:

$$\hat{\mu}(\mathbf{s},t) = \hat{\mu}_0(\mathbf{s}) + \sum_{j=1}^{J} \hat{\mu}_{j,1}(\mathbf{s}) \sin\left(\frac{2\pi j}{365} t + \hat{\mu}_{j,2}(\mathbf{s})\right), \tag{4.4}$$

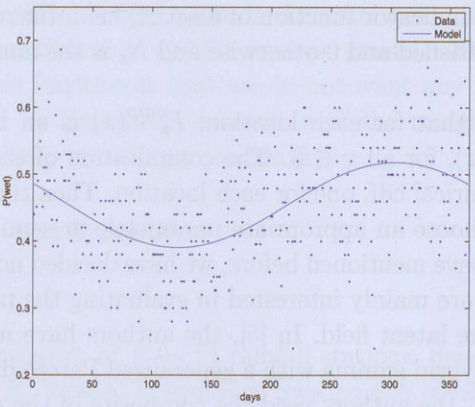

FIGURE 2. *Empirical and model estimates of* $\Phi^{-1}(\hat{p}_{\mathbf{s}_1,t})$.

with J not exceeding 2, and the regression coefficients being space dependent. Parameter estimation is done by the Weighted Least Square (WLS) method, while the number of covariates included is determined using the Bayesian Information Criterion (BIC) see [26]. The BIC gave 1 as optimal value of trigonometric terms in (4.4) for about half of the stations and 2 for the rest. We have decided to use $J = 1$ for all stations since the gain for using the most complex model was not substantial. This resulted to:

$$\hat{\mu}(\mathbf{s},t) = \hat{\mu}_0(\mathbf{s}) + \hat{\mu}_1(\mathbf{s}) \sin\left(\frac{2\pi j}{365}t + \hat{\mu}_2(\mathbf{s})\right). \tag{4.5}$$

Then, the parameter estimates were interpolated in space by regressing them on location covariates,

$$\hat{\mu}_i(\mathbf{s}) = \delta'_{\mu_i} w(\mathbf{s}), \quad i = 0,1,2 \tag{4.6}$$

where $\delta_{\mu_i} = (\delta_{\mu_i,0}, \delta_{\mu_i,1}, \delta_{\mu_i,2}, \delta_{\mu_i,3})'$ with $\delta_{\mu_i,0}$ being the intercept and $w(\mathbf{s})$ the covariates (1, latitude, longitude, altitude). The two steps in the estimation procedure could probably be combined in one by building a composite likelihood, although this was not explored any further. A different approach was adapted in [22], where the authors assumed that the parameter estimates were themselves a realization of some random field.

4.4. Covariance

We turn now to the problem of estimating the spatio-temporal covariance structure C_Z. In order to use the procedure outlined in section 3.2, we need estimates of the censored values, z^+, which are obtained by transforming the observed rainfall amounts y, as follows:

$$z^+ = \Phi^{-1}_{\hat{u}_7} \circ F^{emp}_{\mathbf{s}}(y), \quad y > 0 \tag{4.7}$$

with $F_{\mathbf{s}}^{emp}(\cdot)$ being the empirical cdf of the rainfall intensity during each month estimated according to the procedure in section 4.2 and $\Phi_{\hat{\mu}_Z}$ as in (3.9) with the estimate of the mean value according to section 4.3 Then, performing the minimization procedure in (3.6), using the MATLAB routine *fminsearch*, we obtain estimates of $\hat{\rho}_{ij}(\tau)$ for different time lags. The resulting covariances for different values of τ can be seen in Figure 3.

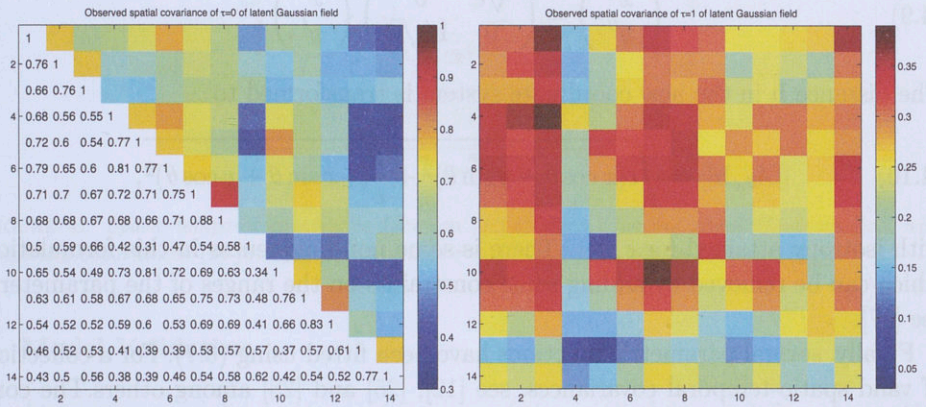

FIGURE 3. *Spatial covariance structure for the latent Gaussian field for July for* $\tau = 0$ *(*Left*) and* $\tau = 1$ *(*Right*).*

The correlation between different locations during the same day, $\tau = 0$ varies between 0.3 (for stations 5 and 9) up to 0.88 (between stations 7 and 8) as can be seen in Figure 3 (Left). In Figure 3 (Right), we plotted the correlation between the different stations but for one day delay, i.e. $\hat{\rho}_{ij}(1)$, which, as expected, are weaker and vary between 0.05 and 0.4. As can be seen in Figure 3 (Right), these correlations are no longer symmetric, i.e $\hat{\rho}_{ij}(1) \neq \hat{\rho}_{ji}(1)$. This is due to the dynamics involved, i.e. the motion of the weather systems.

To these correlation estimates we shall fit a parametric covariance function using the minimization procedure in (3.7). For simplicity, we shall consider only fully symmetric covariance functions, although as we have seen in Figure 3 (Right) the data does not appear to be fully symmetric. On the other hand the assumption of stationarity does not seem to be very restrictive, the area of southern Sweden seems to be homogeneous enough. The spatial anisotropy can be partly corrected by forming an anisotropic covariance function by applying to the isotropic covariance a non-Euclidean measure formed as Euclidean distance in a linearly transformed spatial coordinate system. Additionally, the covariance function should also include some kind of dynamics, a feature that is not accounted for in the present choice of covariance structure but should be further investigated.

Before we proceed any further the spatial coordinate system should be trans-

formed by rotating the old isotropic axis x, y by an angle θ, forming new coordinates

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \tag{4.8}$$

and then dilating them by a factor ϵ to form the new coordinates

$$\begin{pmatrix} x'' \\ y'' \end{pmatrix} = \begin{pmatrix} \sqrt{\epsilon} & 0 \\ 0 & 1/\sqrt{\epsilon} \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}. \tag{4.9}$$

The distance $\mathbf{h}$ in the new coordinate system is transformed to

$$||A_{\theta,\epsilon}\mathbf{h}|| = \sqrt{\epsilon(x\cos\theta + y\sin\theta)^2 + \frac{1}{\epsilon}(-x\sin\theta + y\cos\theta)^2}, \tag{4.10}$$

with isotropy attained for $\epsilon = 1$. There is some non-uniqueness in this formulation which can be removed by adding some constraints on the ranges of the parameters, see [17].

Finally, several parametric functions have been fitted using (3.7). For a collection of valid spatio-temporal covariances, see [12], [25] and [28] among others.The constant $n_{obs}(i, j, \tau)$ in (3.7) equals the number of observations used in the estimation of the empirical covariance between station $\mathbf{s}_i$ on a given day and station $\mathbf{s}_j$ after τ days. The minimization is performed for each time lag τ separately. The covariance structure that gave the best fit in terms of Weighted Least Squares (WLS) criterion is the following,

$$C_Z(\mathbf{h}, \tau) = \eta\{||\mathbf{h}|| = 0, \tau = 0\} + \frac{(1-\eta)}{a|\tau| + 1} e^{-\frac{b||\mathbf{h}||^2}{a|\tau|+1}}, \tag{4.11}$$

where $\eta > 0$. The function $\eta\{\cdot, \cdot\}$ models the nugget effect which allows for a discontinuity at zero, and is used to account for the undistinguishable micro-scale variability and measurement error. The nonnegative parameters a and b are the scaling parameters of time and space respectively, controlling the degree of dependence. The fitting of the covariance structure resulted to a set of parameters $[a, b, \eta, \theta, \epsilon]$.

To illustrate the fitting of the spatio-temporal covariance function, Figure 4 shows the fitted spatial covariance function for different time lags together with the empirical covariances for July. Our method of moments approach to estimating the parameters of the covariance shows reasonable performance. A simulation study showed that the method of moments approach performed considerably better than the modified maximum likelihood (MLE) used in [13] in terms of bias of the estimated correlation. The modified MLE severely overestimated the correlation when the dependence between censored bivariate Gaussian data was strong, but performed reasonably well when data were independent. On the other hand, the method of moments approach performed equally well for any type of dependence between data.

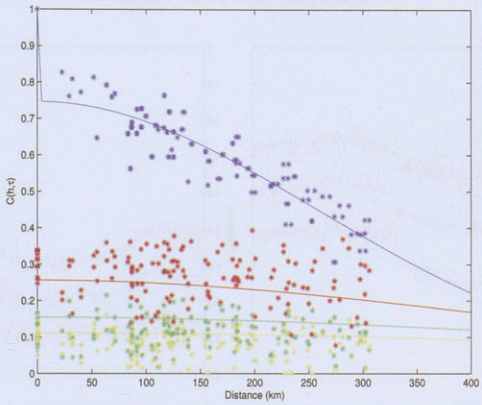

FIGURE 4. *Spatio-temporal covariance function (4.11)* (lines) *for the latent Gaussian field with empirical covariances* ($\star$) *for July for time lags* $\tau = 0, 1, 2, 3$ *from top to bottom.*

5. Model Validation

In this section we validate the performance of the method for modeling the spatio-temporal variability of the rainfall fields. For this, we generated 100 trajectories of the 51 years of data using the proposed algorithm and then examined the results.

5.1. Simulations

To simulate rainfall data at site $\mathbf{s}$ on day t, we perform the following steps.

- A realization of a normal random field $Z(\mathbf{s}, t)$ with mean function $\hat{\mu}_Z$ and covariance $\widehat{C}_Z$ is generated.
- For every location and day there is zero rainfall if $Z(\mathbf{s}, t) \leq 0$.
- For location and day with positive rainfall the simulated intensity is set to $Y(\mathbf{s}, t) = \hat{\psi}(Z(\mathbf{s}, t))$.

5.2. Temporal model

As is well known, the previous day's property of dry or wet greatly influences next day's weather, see e.g. [23]. For this reason, transition probabilities with previous dry and wet day are estimated. Figure 5 illustrates these observed and simulated transition probabilities for the city of Halmstad (station 5 in Figure 1) together with a pointwise 90%-confidence interval based on 100 simulations. As it can be seen the probability of next day to be wet is about 0.2 higher if present day is wet than if it is dry. The stochastic model slightly underestimates the transition probability of a wet day given the previous day is wet and it overestimates the corresponding probability given previous day is dry. Since these probabilities depend mainly on the

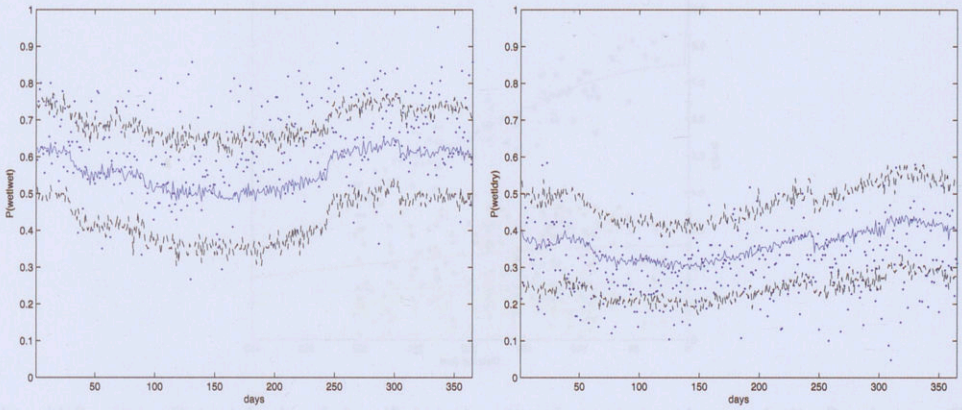

FIGURE 5. *The observed,* (dots) *and simulated* (line) *proportions of wet day given previous wet* (Left) *and dry* (Right) *day at Halmstad, with a pointwise* 90%*-confidence interval based on 100 simulations.*

one day-lag correlation between the different stations, this suggests that the fit of the correlation could be better improved.

We turn next to the dry/wet behavior of the obtained stochastic model. We remind that the distribution of the length of the wet and dry spells, i.e. the time there is positive rainfall or no rainfall respectively, is an essential feature of any stochastic model for rainfall. Notice that these characteristics depend solely on the latent process since they coincide with the time the process spends above (below) the zero level. The empirical distribution of wet (*Left*) and dry (*Right*) spells can be seen in Figure 6 together with the model distribution and a 90% confidence interval superimposed. The length of the wet spell is very well replicated for any spell length,

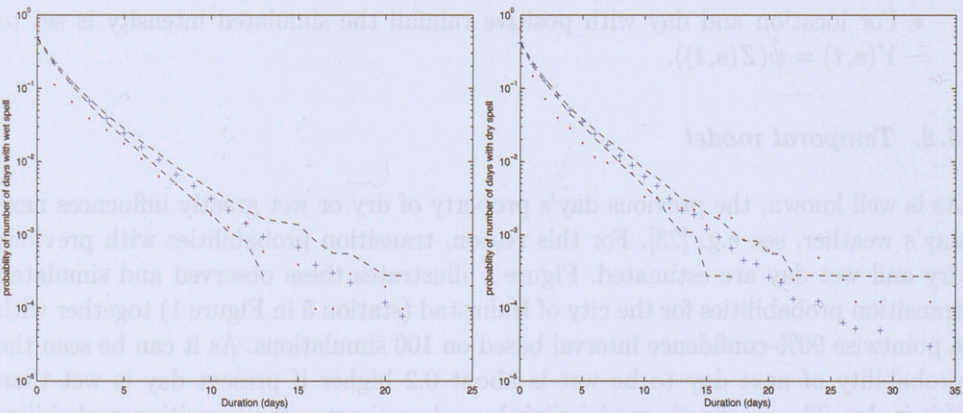

FIGURE 6. *The empirical distribution (dots) of number consecutive wet* (Left) *and dry* (Right) *days, and the ones based on simulations (+) with a* 90% *confidence interval superimposed, at Halmstad.*

and the same is true for most of the dry spell lengths.

5.3. Spatial model

A positive feature of the defined stochastic model is in correlating wet events across space. In order to illustrate the fit of spatial dependence we have considered all pairs of stations and then computed the observed and model proportions of simultaneously wet and simultaneously dry days. The results are gathered in Figure 7. The latent

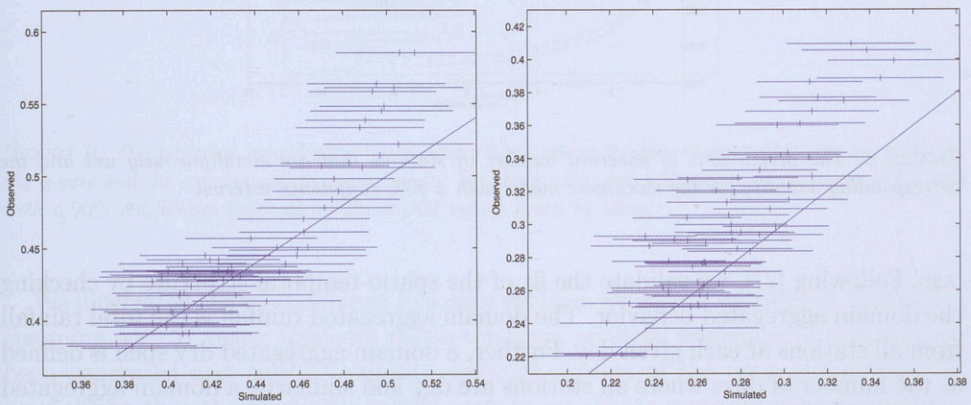

FIGURE 7. *The proportions of simultaneously dry* (Left) *and simultaneously wet days* (Right). *The proportions of observations are given on y-axis with the corresponding simulated proportions on the x-axis with a* 90% *confidence interval based on 100 simulations given by lines.*

Gaussian field seems to replicate well the simultaneous wet behavior.

Another interesting feature of the spatial dependence structure is whether it properly replicates the number of wet stations each day. Figure 8 displays the observed proportion of number of stations with positive rainfall. Days where data are missing for at least one of the stations were removed. The stochastic model seems to replicate quite well the observed frequency of total number of stations with simultaneous wet events. It is interesting to notice that the probability to have no observed rainfall at any of the stations is about 0.22, while the probability to observe precipitation at all stations is about 0.17 and higher than the probability to observe positive rainfall in any subset of locations.

In general we feel that the model replicates spatial aspects of the precipitation process well.

5.4. Spatio-temporal model

The most difficult feature of multisite rainfall data is the spatio-temporal dependence structure. We have already seen in Figure 4 that the parametric covariance function of the latent Gaussian field fits well the empirical covariance for at least a few time

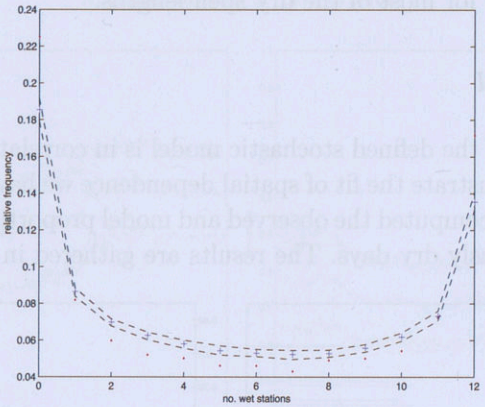

FIGURE 8. *The proportions of observed number of stations that are simultaneously wet and the corresponding quantity for the stochastic model with a 90% confidence interval.*

lags. Following [22], we validate the fit of the spatio-temporal structure by checking the domain aggregated behavior. The domain aggregated rainfall is the total rainfall from all stations at each given day. Further, a domain aggregated dry spell is defined as the number of days where all stations are dry and similarly, a domain aggregated wet spell is the number of days where at least one station is wet.

An alternative way to the spatio-temporal aspects of the generated dry/wet behavior is by examining pairwise lagged rainfall probabilities defined as follows:

$$P(Y(\mathbf{s}_i, t-1) = 0, Y(\mathbf{s}_j, t) > 0) \text{ and } P(Y(\mathbf{s}_i, t-1) > 0, Y(\mathbf{s}_j, t) = 0). \tag{5.1}$$

Each point in Figure 9 represents observed and model proportions of either one of the pairwise lagged rainfall occurrences. The stochastic model seems to overestimate the quantity for small proportions in July. The deviation is only of order 0.02 and unlikely to have a significant impact in practice.

6. Conclusions

We have presented a general methodology for modeling spatially correlated fields of daily rainfall. The method relies on a latent Gaussian random field that drives both the rainfall occurrence and the rainfall intensity processes, with the rainfall intensity being modeled as transformed Gaussian.

The mean function was estimated by inverting the proportions of rainfall at each location and then seasonally varying parameters were spatially interpolated using regression with Fourier components and location covariates respectively. A parametric covariance function was used to model the observed spatio-temporal correlations with the fitting performed using a method of moments approach and certain relations that hold for censored Gaussian moments.

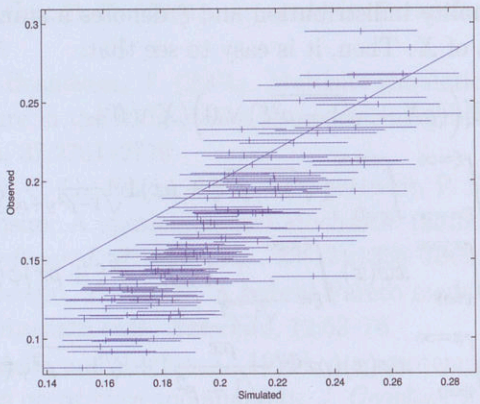

FIGURE 9. *The pairwise lagged occurrence proportions where the observed proportions are given on the y-axis and the corresponding quantities for the stochastic generator are represented on the x-axis with a 90% confidence interval for the model values given by lines.*

We illustrated the methodology in a data set from Sweden, a network of 14 stations spanning over 51 years. Realistic spatio-temporal artificial sequences of rainfall have been generated and used to validate different aspects of the proposed model. We have shown that the stochastic model seem to be able to reproduce the dependence between different days and stations quite well. Also, the method shows a good ability to replicate dry and wet behavior. The possibility to include other covariates such as climate model output or broad scale atmospheric conditions in order to model the corresponding spatial extrapolation of the parameters should be investigated. Finally non-isotropic spatial covariance structures should also be fitted.

7. Appendix section

In this section we derive formula (3.3). Let X, Y be two standard (mean zero and variance unity) Gaussian random variables with correlation ρ. Then we have the following representation:

$$Y \stackrel{\mathcal{D}}{=} \rho X + \sqrt{1-\rho^2}\xi,$$

where $\stackrel{\mathcal{D}}{=}$ denotes equality indistribution and ξ denotes a standard normal random variable independent of X. Then, it is easy to see that:

$$\begin{aligned}
\mathbb{E}[Y^+X^+] &= \mathbb{E}[\Big((\rho X + \sqrt{1-\rho^2}\xi) \vee 0\Big)(X \vee 0)] \\
&= \int_{\xi=-\infty}^{\xi=\infty}\int_{x=0}^{x=\infty} (\sqrt{1-\rho^2}\xi + \rho x)\mathbf{1}_{\sqrt{1-\rho^2}y+\rho x \geq 0} x\varphi(x)\varphi(\xi)dxd\xi \\
&= \int_{x=0}^{x=\infty} x\varphi(x)\int_{\xi=-\frac{\rho x}{\sqrt{1-\rho^2}}}^{\xi=\infty} (\sqrt{1-\rho^2}\xi + \rho x)\varphi(\xi)d\xi dx \\
&= \int_{x=0}^{x=\infty} x\varphi(x)(\rho x\Phi(\frac{\rho x}{\sqrt{1-\rho^2}}) + \sqrt{1-\rho^2}\varphi(\frac{\rho x}{\sqrt{1-\rho^2}}))dx,
\end{aligned}$$

where $X^+ = max\{0, X\}$. By letting

$$g(x;\rho) = x\big(x\rho\Phi(\frac{\rho x}{\sqrt{1-\rho^2}}) + \sqrt{1-\rho^2}\varphi(\frac{\rho x}{\sqrt{1-\rho^2}})\big)$$

the covariance of non-negatively truncated versions of X and Y is given by

$$\text{Cov}(X^+, Y^+) = \int_0^\infty g(x,\rho)\varphi(x)dx - \mathbb{E}[X^+]\mathbb{E}[Y^+]. \tag{7.1}$$

It is easy to generalise for the case of non zero means by modifying function g to

$$g(x;\mu_X,\mu_Y,\rho) = x\varphi(x-\mu_X)\big((\rho(x-\mu_Y)+\mu_X)\Phi(\frac{\mu_X+\rho(x-\mu_Y)}{\sqrt{1-\rho^2}}) + \sqrt{1-\rho^2}\varphi(\frac{\mu_X+}{\sqrt{}}$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

References

[1] Ailliot, P., Thompson, C., and Thomson, P. (2009). Space time modelling of precipitation by using a hidden Markov model and censored Gaussian distributions. *J. R. Stat. Soc*, 58(3):405–426.

[2] Allard, D. and Bourotte, M. (2015). Disaggregating daily precipitations into hourly values with a transformed censored latent Gaussian process. *Stochastic Environmental Research and Risk Assesment*, 29(2):453–462.

[3] Allcroft, D. J. and Glasbey, C. A. (2003). A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation. *Biomathematics and statistics*, 52:487–498.

[4] Bardossy, A. and Plate, E. J. (1992). Space-time model for daily rainfall using athmospheric circulation patterns. *Water resources reseach*, 28(5):1247–1259.

[5] Baxevani, A. and Lennartsson, J. (2015). A spatio-temporal weather generator based on a censored latent Gaussian field. *Water Resources Research*, DOI:10.1002/2014WR016455.

[6] Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. R. Stat. Soc.*, 26(2):211–252.

[7] Buishand, T. and Brandsma, T. (2001). Multisite simulation of daily precipitation and temperature in the Rhine basin by nearest-neighbor resampling. *Water Resources Research*, 37:2761–2776.

[8] Burton, A., Kilsby, C. G., Fowler, H. J., Cowpertwait, P. S. P., and O'Connell, P. E. (2008). Rainsim: A spatiotemporal stochastic rainfall modelling system. *Environ. modell. SoftwareJ. R. Stat. Soc.*, 23(12):1356–1369.

[9] Carreau, J. and Bengio, Y. (2009). A hybrid Pareto model for asymmetric fat-tailed data: The univariate case. *Extremes*, 12:53–76.

[10] Charles, S., Bates, B., and Hughes, J. (1999). A spatiotemporal model for downscaling precipitation occurrence and amounts. *J. Geophys. Res.*, 104(D24):31657–31669.

[11] Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty.* Wiley series in probability and statistics, 2nd edition.

[12] Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association*, 94(448):1330–1340.

[13] Durban, M. and Glasbey, C. A. (2001). Weather modelling using a multivariate latent Gaussian model. *Agricultural and Forest Meteorology*, 109:187–201.

[14] Furrer, E., M. and Katz, R. (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water resources research*, 44(12):doi:10.1029/2008WR007316.

[15] Glasbey, C. A. and Nevison, I. M. (1997). Rainfall modelling using a latent Gaussian variable. *Modelling longitudal and spatially correlated data*, 122:233–242.

[16] Guillot, G. (1999). Approximation of Sahelian rainfall fields with meta-Gaussian random functions Part 1: model definition and methodology. *Stochastic Environmental Research and Risk Assesment*, 13:100–112.

[17] Haskard, K. A. (2007). *An anisotropic Matern spatial covariance model: REML estimation and properties.* PhD thesis, University of Adelaide, School of Agriculture, Food and Wine.

[18] Hughes, J., Guttorp, P., and Charles, S. (1999). A non-homogeneous hidden Markov model for precipitation occurrence. *Applied Statistics*, 48(1):15–30.

[19] Hundecha, Y., Pahlow, M., and Schumann, A. (2009). Modeling of daily precipitation at multiple locations using a mixture of distributions to characterize the extremes. *Water Resources Research*, 45(W12412,dio:10.1029/2008WR007543).

[20] Hutchinson, M. F. (1995). Stochastic space-time weather models from ground-based data. *Agricultural and Forest Meteorology*, 73:237–264.

[21] Katz, R. (1977). Precipitation as a chain-dependant process. *Journal of Applied Meteorology*, 16:671–676.

[22] Kleiber, W., Katz, R. W., and Rajagopalan, B. (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water*

resources research, 48(W01523).

[23] Lennartsson, J., Baxevani, A., and Chen, D. (2008). Modelling precipitation in Sweden using multiple step Markov chains and a composite model. *Journal of Hydrology*, 363(1–4):42–59.

[24] Li, C., Singh, V., and Mishra, A. (2012). Simulation of the enire range of daily precipitation using a hybrid probability distribution. *Water Resources Research*, 48(W03521):doi:10.1029/2011 WR011446.

[25] Ma, C. (2003). Families of spatio-temporal stationary covariance models. *J. Stat. Plann. Inference*, 116:489–501.

[26] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

[27] Sigrist, F., Kunsch, H. R., and Stahel, W. A. (2012). A dynamic non-stationary spatio-temporal model for short term prediction of precipitation. *Annals of applied statistics.*

[28] Stein, M. (2005). Space time covariance functions. *JASA*, 100:310–321.

[29] Todorovic, P. and Woolhiser, D. A. (1975). A stochastic model of n-day precipitation. *Journal of Applied Meteorology*, 14:17–24.

[30] Vrac, M. and Naveau, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water Resources Research*, 43(W07402,doi:10.1029/2006WR005308).

[31] Wilks, D. S. (1998). Multi-site generalization of a daily stochastic precipitation generation model. *Journal of Hydrology*, 210:178–191.

[32] Yang, C., Chandler, R. E., Isham, V. S., and Wheater, H. S. (2005). Spatio-temporal rainfall simulation using generalized linear models. *Water Resources Research*, 41(W11415).

Anastassia Baxevani
Department of Mathematics and Statistics
University of Cyprus, Cyprus

Jan Lennartsson
Department of Mathematical Statistics
University of Gothenburg, Sweden
e-mail: baxevani@ucy.ac.cy
jan.lennartsson@gmail.com