Journal Pre-proof

A hierarchical bi-resolution spatial skew- t model

Felipe Tagle, Stefano Castruccio, Marc G. Genton

PII: S2211-6753(19)30149-6
DOI: <https://doi.org/10.1016/j.spasta.2019.100398>
Reference: SPASTA 100398

To appear in: *Spatial Statistics*

Received date: 11 April 2019
Revised date: 19 November 2019
Accepted date: 27 November 2019

Please cite this article as: F. Tagle, S. Castruccio and M.G. Genton, A hierarchical bi-resolution spatial skew- t model. *Spatial Statistics* (2019), doi: <https://doi.org/10.1016/j.spasta.2019.100398>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Elsevier B.V. All rights reserved.

A Hierarchical bi-Resolution Spatial Skew- t Model

Felipe Tagle¹, Stefano Castruccio¹ and Marc G. Genton²

November 19, 2019

Abstract

Advances in Gaussian methodology for spatio-temporal data have made it possible to develop sophisticated non-stationary models for very large data sets. The literature on non-Gaussian spatio-temporal models is comparably sparser and strongly focused on distributing the uncertainty across layers of a hierarchical model. This choice allows to model the data conditionally, to transfer the dependence structure at the process level via a link function, and to use the familiar Gaussian framework. Conditional modeling, however, implies an (unconditional) distribution function that can only be obtained through integration of the latent process, with a closed form only in special cases. In this work, we present a spatio-temporal non-Gaussian model that assumes an (unconditional) skew- t data distribution, but also allows for a hierarchical representation by defining the model as the sum of a small and a large scale spatial latent effect. We provide semi-closed form expressions for the steps of the Expectation-Maximization algorithm for inference, as well as the conditional distribution for spatial prediction. We demonstrate how it outperforms a Gaussian model in a simulation study, and show an example of application to precipitation data in Colorado.

Keywords: EM algorithm; hierarchical model; latent process; non-Gaussian distribution; skew- t distribution; spatial statistics

¹Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, United States.
E-mail: ftagleso@nd.edu, scastruc@nd.edu

²Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.
E-mail: marc.genton@kaust.edu.sa

This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No: OSR-2015-CRG4-2640.

1 Introduction

Gaussian models have always been at the center of investigations in spatio-temporal statistics, as the normal assumption implies a unparalleled degree of analytical tractability and computational convenience. In many applications, however, the data often exhibit features that are inconsistent with this assumption, and therefore require more flexible non-Gaussian models.

Different solutions for non-Gaussian, spatio-temporal data are available depending on the extent of the departure from the normal assumption. When the non-Gaussianity is not too severe, the simplest and more widespread solution is a non-linear transformation of the data that modifies the marginal distribution to resemble the normal distribution, allowing for the subsequent use of the traditional methodology (Cressie 1993). These *trans-Gaussian* random fields consist of a simple, marginal transformation of the data, such as a logarithm, and sometimes allow to explicitly express the degree of skewness and kurtosis in the model; see Xu & Genton (2017) and references therein.

For severe departures from Gaussianity (e.g., count data), a more sophisticated solution is to rely on a generalized linear model framework that assumes data independence conditionally on a process, while the process is assumed to be Gaussian and captures the spatio-temporal dependence. This *model-based geostatistics* approach (Diggle et al. 1998), whose cornerstone principle is the stratification of the spatio-temporal uncertainty across layers of a (Bayesian) hierarchical model, has recently seen a surge in popularity with the advent of the Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)) and the R-INLA project (Lindgren & Rue 2015). In recent years, the INLA approach has received a strong interest from the statistical community (and beyond), because of its ability to sparsify the conditional dependence structure and therefore to achieve inference for large spatial data. Although this modeling approach has arguably many advantages, it suffers, by design, of a lack of explicit parametrizations of the non-Gaussian features. Since the focus of the modeling is on a latent process and the data

are only modeled conditionally, the (unconditional) data distribution can be expressed only through integrals, by marginalizing over the latent process as well as the prior uncertainty of its hyperparameters (in the case of a Bayesian model). These integrals rarely have a closed form, and hence hamper model interpretability should one be interested in the unconditional properties of the model.

In this work, we propose a new non-Gaussian model focused on the unconditional properties of the data, which also allows for a convenient hierarchical representation for inference purposes. Therefore, we retain the flexibility, as well as (some of) the computational convenience of hierarchical models while also achieving an explicit expression of the moments, hence improving the model interpretability. The model relies on a domain partition into regions, where a common latent process across each region is assumed and estimated. Domain partition in spatio-temporal statistics is a strategy typically employed as a means to perform approximate inference. Indeed, when the data set is too large the likelihood is often misspecified through a composite approach. In the case of non-Gaussian models such as max-stable processes, composite likelihood approaches are necessary as full likelihood inference for more than eleven points is impossible (Castruccio et al. 2016). In this work, the choice of partitioning was not dictated by computational needs, and instead of misspecifying a model we opted for defining a model with a partition, and defined a global model to allow non-Gaussian modeling in space with explicit expression of the unconditional properties of the process.

The model relies on a perturbation of the Gaussian distribution, introduced by Azzalini (1985) in the univariate case: the skew-normal. This distribution generalizes the normal distribution by explicitly accounting for the skewness with a parameter; it has been actively studied and developed for over three decades. Azzalini & Dalla Valle (1996) proposed a multivariate extension, whereas Azzalini & Capitanio (2003) proposed an equivalent formulation for the Student's t distribution. A full review can be found in Genton (2004) and most recently in Azzalini & Capitanio (2014). While some work has focused on extending these multivariate distributions to

processes (Zhang & El-Shaarawi 2010, Mahmoudian 2017, Bevilacqua et al. 2018, Kim & Mallick 2004, Kim et al. 2004, Allard & Naveau 2007, Schmidt et al. 2017), the existence of a process is not guaranteed in general (Genton & Zhang 2012, Minozzo & Ferracuti 2012). Skewed distributions cannot be obtained from a central limit theorem or any other large sample results, they are obtained by perturbing the normal distribution to account for some degree of skewness and kurtosis, and they also emerge naturally in selective population sampling, when a latent variable is involved in the sampling and does not allow a completely random sample (Arellano-Valle et al. 2006).

Skew- t models have been widely used for applications involving multivariate and sometimes spatial data, but to our knowledge, no work has highlighted the potential of this distribution as a baseline for establishing a spatial hierarchical model. Here we show how this reformulation allows for a bi-resolution model with a partition of the spatial domain into multiple regions, each one with a different stationary behavior, linked with a latent large-scale spatial effect.

We perform a frequentist inference, and develop a Monte Carlo Expectation-Maximization (MCEM) algorithm, a choice also used in the spatial skew-normal model developed by Zhang & El-Shaarawi (2010). MCEM is also often used for inference involving skewed distributions in a non-spatial context (e.g., Lin & Lee (2008), Lin (2010), Lachos et al. (2010); see Lachos et al. (2018) for a review). In this work, we derive the semi-explicit expressions of the E and M step, as well as a closed form of the conditional distribution at unobserved locations, in order to allow spatial prediction.

The remainder of the paper is organized as follows: Section 2 introduces the model, Section 3 presents the E and the M step of the algorithm for inference, as well as the conditional distribution at unsampled locations for prediction. Section 4 presents a simulation study where the proposed model is compared with a Gaussian model. Section 5 shows an application to Colorado precipitation data. Section 6 concludes with a discussion.

2 Model Definition

For this section, we assume we have a purely spatial process. We consider a scalar, non-Gaussian random field $\{Y(\mathbf{s}), \mathbf{s} \in D\}$, where $D \subset \mathbb{R}^2$, and a partition $D = \cup_{r=1}^R \mathcal{D}_r$ such that $\mathcal{D}_r \cap \mathcal{D}_{r'} = \emptyset$ if $r \neq r'$. The partition should be determined by areas where higher moment characteristics are similar, as we will show in the application. We propose a model for $Y(\mathbf{s})$ for which the process, within each region, is multivariate skew- t , with the parameterization of Azzalini & Capitanio (2003).

For each region $\mathcal{D}_r$, and for each point $\mathbf{s} \in \mathcal{D}_r$, the following model applies

$$Y(\mathbf{s}) = \sigma_r \frac{\rho_r U_{0,r} + \lambda_r U_{1,r} + \eta_r(\mathbf{s})}{\sqrt{Z_r}}, \quad (1)$$

where $U_{1,r}$ has a standard half-normal distribution, Z_r a $\text{Gamma}(\nu_r/2, \nu_r/2)$ distribution, where the first argument is the scale, while the second the rate, and $\eta_r(\mathbf{s})$ is a stationary Gaussian process independent across r , with mean zero and correlation function that depends on parameters $\boldsymbol{\psi}_r$, with an associated correlation matrix $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}(\boldsymbol{\psi}_r)$; and $\sigma_r \geq 0$, $\rho_r \geq 0$, $\lambda_r \in \mathbb{R}$. Setting $\rho_r = 0$ reduces the model to that of Tagle et al. (2019), in which each region has a multivariate skew- t distribution, but evolves independently from the others. Here, we introduce the random vector $\mathbf{U}_0 = (U_{0,1}, \dots, U_{0,R})^\top$, assumed to follow a mean zero multivariate normal distribution with correlation matrix $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\psi}_0)$, with each component being assigned to each region, with the aim of inducing an inter-regional spatial dependence. The vector $\mathbf{U}_0$ hence takes on the role of a large-scale effect, with components interacting additively with the other terms in the numerator and capturing the fine-scale variability, jointly resulting in a multivariate skew-normal distribution. A location parameter could be added in model (1), and in fact the application in Section 5 allows for a changing latent parameter for each region. However, for the ease of analytical tractability and computational convenience in the simulation studies, we chose not to allow for this parameter at this stage.

Model (1) assumes that, conditional on $U_{0,r}$ and $U_{1,r}$, points across different regions are independent, and hence simulations from this model will have discontinuities at the boundaries and there is conditional independence among points for different regions. While this is arguably a suboptimal feature of the model globally this can however be mitigated by constraining the latent process $U_{0,r}$ to be very smooth, by fixing the smoothness of the corresponding covariance matrix Σ_0 .

The additive form ensures that the numerator of (1) remains a multivariate skew-normal distribution given the closure of the skew-normal distribution under convolutions with a normal random variable (Azzalini & Capitanio 2014). It is straightforward to show that, for a collection of points $\mathbf{s}_1, \dots, \mathbf{s}_n \in \mathcal{D}_r$, the numerator has a skew-normal distribution with a probability density function (pdf) given by

$$\rho_r U_{0,r} + \lambda_r U_{1,r} + \eta_r(\mathbf{s}) \sim 2(\sqrt{1 + \lambda_r^2 + \rho_r^2}) \phi_n(\mathbf{y}; \mathbf{\Omega}_r) \Phi(\mathbf{\alpha}_r^\top \mathbf{y}), \quad (2)$$

where ϕ_n is the pdf of an n -dimensional normal distribution with mean zero and covariance $\mathbf{\Omega}_r$, and Φ is the cumulative distribution function of the univariate standard normal distribution. Furthermore,

$$\mathbf{\alpha}_r = \frac{\delta_r}{1 - \delta_r^2} \frac{\mathbf{1}_n^\top \mathbf{\Omega}_r^{-1}}{\left(1 + \frac{\delta_r^2}{1 - \delta_r^2} \mathbf{1}_n^\top \mathbf{\Omega}_r^{-1} \mathbf{1}_n\right)^{1/2}},$$

$$\mathbf{\Omega}_r = (1 - \delta_r^2) \left(\mathbf{\Sigma}(\boldsymbol{\psi}_r)^* + \frac{\delta_r^2}{1 - \delta_r^2} \mathbf{1}_n \mathbf{1}_n^\top \right),$$

$\mathbf{\Sigma}(\boldsymbol{\psi}_r)^*$ is a correlation matrix with terms $\mathbf{\Sigma}(\boldsymbol{\psi}_r)_{ij}^* = (\mathbf{\Sigma}(\boldsymbol{\psi}_r)_{ij} + \rho_r^2)/(1 + \rho_r^2)$, and $\delta_r = \frac{\lambda_r}{\sqrt{1 + \lambda_r^2 + \rho_r^2}}$.

Dividing by a Gamma distribution results in the multivariate skew- t distribution.

Let $\mathbf{s}$ be one point of the collection $\mathbf{s}_1, \dots, \mathbf{s}_n$ with corresponding skewness parameter α drawn from the vector $\mathbf{\alpha}_r$. For simplicity of notation, we further drop the index r across all terms. The proposed skew- t formulation allows to explicitly compute the moments. From (1) and (2), we

can write the k -th moment as

$$E[Y(\mathbf{s})^k] = \sigma^k(1 + \lambda^2 + \rho^2)^{k/2} E[Z^{-k/2}] E[Y^*(\mathbf{s})^k], \quad (3)$$

where $Y^*(\mathbf{s}) = \{\rho U_0 + \lambda U_1 + \eta(\mathbf{s})\} / \sqrt{1 + \lambda^2 + \rho^2}$ which has a standard skew-normal distribution.

Since Z is a Gamma distribution, it can be easily shown that

$$E[Z^{-k/2}] = \frac{(\nu/2)^{k/2} \Gamma\{(\nu - k)/2\}}{\Gamma(\nu/2)},$$

whereas the k -th moment of a standard skew- t distribution can be obtained from Azzalini & Capitanio (2014), pg. 32, as

$$E[Y^*(\mathbf{s})^k] = \sqrt{\frac{2}{\pi}} \frac{\text{sgn}(\alpha)}{\alpha^{k+1}} B_k(\alpha^{-2}),$$

where, for $h > 0$,

$$B_k(h) = \frac{k-1}{h} B_{k-2}(h) + \frac{\beta_{k-1}}{h(1+h)^{k/2}}, \quad k = 2, 3, \dots$$

and β_k is the k -th moment of the standard normal distribution, $\beta_k = (k-1)!!$, for $k = 2, 4, 6, \dots$ and 0 otherwise.

The k -moment formula (3) allows to obtain the mean and variance as

$$E[Y(\mathbf{s})] = \sigma(1 + \lambda^2 + \rho^2)^{1/2} \frac{(\nu/2)^{1/2} \Gamma\{(\nu - 1)/2\}}{\Gamma(\nu/2)} \text{sgn}(\alpha) \sqrt{\frac{2}{\pi(1 + \alpha^{-2})}},$$

$$\text{Var}[Y(\mathbf{s})] = \sigma^2(1 + \lambda^2 + \rho^2) \text{sgn}(\alpha) - E[Y(\mathbf{s})]^2.$$

The covariance for points $\mathbf{s}_1$ and $\mathbf{s}_2$ belonging to the same region $\mathcal{D}_r$ can be expressed as

$$\text{Cov}(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = \sigma^2 \frac{\nu_r}{\nu - 2} \{\rho^2 + \lambda^2 + C_\psi(h)\} - E[Y(\mathbf{s}_1)]^2, \quad (4)$$

where $C_{\psi_r}(h)$ is the correlation function associated with the matrix Σ_r . If, instead, $\mathbf{s}_1 \in \mathcal{D}_{r_1}$ and

151 $\mathbf{s}_2 \in \mathcal{D}_{r_2}$ with $r_1 \neq r_2$

$$\text{Cov}(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = \sigma_{r_1} \sigma_{r_2} \rho_{r_1} \rho_{r_2} \frac{(\nu_{r_1}/2)^{1/2} \Gamma\{(\nu_{r_1} - 1)/2\}}{\Gamma(\nu_{r_1}/2)} \frac{(\nu_{r_2}/2)^{1/2} \Gamma\{(\nu_{r_2} - 1)/2\}}{\Gamma(\nu_{r_2}/2)} \Sigma_{0;r_1, r_2}. \quad (5)$$

152 The representation in (1) involves the latent variables $U_{0,r}$, $U_{1,r}$ and Z_r , $r = 1, \dots, R$, which
 153 we choose to estimate with an EM algorithm (Dempster et al. 1977). This approach aims to
 154 maximize the model likelihood in the presence of latent processes by alternating between an
 155 expectation step to estimate $\mathbf{U}_0$, and a maximization step to estimate all the model parameters,
 156 see McLachlan & Krishnan (2007) for details. During the first step, $(U_{0,r}, U_{1,r}, Z_r)^\top$, $r = 1, \dots, R$
 157 and functions thereof are replaced by their conditional expectations given the data and param-
 158 eter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^\top, \dots, \hat{\boldsymbol{\theta}}_R^\top, \hat{\boldsymbol{\psi}}_0^\top)^\top$, with $\boldsymbol{\theta}_r = (\sigma_r, \rho_r, \lambda_r, \boldsymbol{\psi}_r^\top)^\top$. In the second step, $\hat{\boldsymbol{\theta}}$ is
 159 updated in the traditional maximum-likelihood sense based on the maximization of the associ-
 160 ated log-likelihood. Since a parameter set is required to conduct the first E-step, an initial set of
 161 values is provided, which is then subsequently updated by alternating between both steps until
 162 convergence is achieved.

163 Our choice of an EM approach was mostly computational and dictated by the availability
 164 of closed form expressions for some of the parameters during the Maximization step, as the
 165 next section shows. A full Bayesian model is a possible alternative, but we could not find any
 166 conjugate prior for the parameters, so we would have needed another MCMC step within the
 167 Gibbs sampling, thus adding additional strain to the already burdensome inference approach.

168 3 EM Algorithm

169 Henceforth $\mathbf{Y}_r$, $r = 1, \dots, R$ represents the vector of random variables corresponding to the
 170 observations in region $\mathcal{D}_r$, and $|\mathcal{D}_r| = n_r$, so that $\sum_{r=1}^R n_r = n$. The proposed model in (1) has

the following hierarchical representation,

$$\begin{aligned} \mathbf{Y}_r | U_{0,r} = u_{0,r}, U_{1,r} = u_{1,r}, Z_r = z_r &\stackrel{iid}{\sim} N_{n_r} \left(\frac{\sigma_r}{\sqrt{z_r}} (\rho_r u_{0,r} + \lambda_r u_{1,r}) \mathbf{1}_{n_r}, \frac{\sigma_r^2}{z_r} \mathbf{\Sigma}_r \right), \\ \mathbf{U}_0 &\sim N_R(\mathbf{0}, \mathbf{\Sigma}_0), \\ U_{1,R} &\stackrel{iid}{\sim} HN(0, 1), \\ Z_r &\stackrel{iid}{\sim} \text{Gamma}(\nu_r/2, \nu_r/2), \end{aligned}$$

where HN refers to the half-normal distribution, and $\mathbf{1}_{n_r}$ is a $n_r \times 1$ vector with all entries being equal to one. We now assume to have independent temporal replicates, as will be the case in the application. We thereby consider the vector $\mathbf{y}_t = (\mathbf{y}_{1,t}^\top, \dots, \mathbf{y}_{R,t}^\top)^\top$, $t = 1, \dots, T$, as well as $\mathbf{u}_{0,t}$, $\mathbf{u}_{1,t}$, and $\mathbf{z}_{1,t}$, which are defined conformably (in an abuse of notation, we henceforth use lower-case letters to denote both realizations and random quantities). Their joint distribution, for each t , follows from the above representation:

$$\begin{aligned} p(\mathbf{y}_t, \mathbf{u}_{0,t}, \mathbf{u}_{1,t}, \mathbf{z}_t | \boldsymbol{\theta}) &= p(\mathbf{y}_t | \mathbf{u}_{0,t}, \mathbf{u}_{1,t}, \mathbf{z}_t, \boldsymbol{\theta}) \times p(\mathbf{u}_{0,t} | \boldsymbol{\theta}) \times p(\mathbf{u}_{1,t} | \boldsymbol{\theta}) \times p(\mathbf{z}_t | \boldsymbol{\theta}) \\ &= \prod_{r=1}^R \frac{1}{(2\pi)^{n_r/2} |\sigma_r^2 / z_r \mathbf{\Sigma}_r|^{1/2}} \exp \left\{ -\frac{z_{r,t}}{2\sigma_r^2} \mathbf{x}_{r,t}^\top \mathbf{\Sigma}_r^{-1} \mathbf{x}_{r,t} \right\} \\ &\times \frac{1}{(2\pi)^{R/2} |\mathbf{\Sigma}_0|^{1/2}} \exp \left\{ -\frac{1}{2} \mathbf{u}_{0,t}^\top \mathbf{\Sigma}_0^{-1} \mathbf{u}_{0,t} \right\} \times \frac{(\nu_r/2)^{\nu_r/2}}{\Gamma(\nu_r/2)} z_{r,t}^{\frac{\nu_r+n_r}{2}-1} \times \sqrt{\frac{2}{\pi}} \exp \left\{ -\frac{u_{1,r,t}^2}{2} \right\} \end{aligned} \quad (6)$$

where $\mathbf{x}_{r,t} = \mathbf{y}_{r,t} - \sigma_r / \sqrt{z_{r,t}} (\rho_r u_{0,t} + \lambda_r u_{1,t}) \mathbf{1}_{n_r}$. We can further aggregate the time- t vectors, as $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_T^\top)^\top$, with conformable definitions for the latent variables, allowing us to express the corresponding log-likelihood as

$$\ell(\boldsymbol{\theta} | \mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) = \sum_{r=1}^R \ell_r(\boldsymbol{\theta} | \mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) - \frac{T}{2} \log |\mathbf{\Sigma}_0| - \frac{1}{2} \sum_{t=1}^T \mathbf{u}_{0,t}^\top \mathbf{\Sigma}_0^{-1} \mathbf{u}_{0,t}, \quad (7)$$

where

$$\begin{aligned} \ell_r(\boldsymbol{\theta} | \mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) &= T \frac{\nu_r}{2} \log \left(\frac{\nu_r}{2} \right) - T \log \Gamma \left(\frac{\nu_r}{2} \right) - \frac{T n_r}{2} \log(\sigma_r) - \frac{T}{2} \log |\mathbf{\Sigma}_r| \\ &+ \left(\frac{\nu_r + n_r}{2} - 1 \right) \sum_{t=1}^T \log(Z_{r,t}) - \frac{1}{2} \sum_{t=1}^T \left(u_{1,r,t}^2 + \nu_r z_{r,t} + \frac{z_{r,t}}{\sigma_r^2} \mathbf{x}_{r,t}^\top \mathbf{\Sigma}_r^{-1} \mathbf{x}_{r,t} \right). \end{aligned}$$

The log-likelihood involves the inversion of the matrices $\mathbf{\Sigma}_r$ and $\mathbf{\Sigma}_0$, as well as the computation of their respective determinants, which are typically problematic when the size of the dataset becomes large. However, our approach of regionalizing the spatial domain allows for a judicious choice of n_r to ensure the computation feasibility of said operations on $\mathbf{\Sigma}_r$, while the dimension

of Σ_0 is determined by the number of regions R for which $R \ll n$ is assumed to hold.

3.1 E-step

The E-step proceeds by computing the conditional expectation, commonly denoted by the Q function,

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{[k]}) = \mathbb{E} \left[\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) | \mathbf{y}, \hat{\boldsymbol{\theta}}^{[k]} \right]$$

based on the value of $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{[k]}$ at the k -th iteration, in order to obtain a new estimate of the latent processes $\mathbf{u}_0$, $\mathbf{u}_1$ and $\mathbf{z}$. Technical details on how to simulate the full conditional of $\mathbf{z}$ and the closed form expressions for $\mathbf{u}_0$ and $\mathbf{u}_1$ are provided in the Supplementary Material.

3.2 M-step

Once the estimates for the latent processes $\hat{\mathbf{u}}_0^{[k]}$, $\hat{\mathbf{u}}_1^{[k]}$ and $\hat{\mathbf{z}}^{[k]}$ are available, they are plugged in (7) and a new parameter vector $\hat{\boldsymbol{\theta}}^{[k+1]}$ is obtained as

$$\hat{\boldsymbol{\theta}}^{[k+1]} = \arg \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta} | \mathbf{y}, \hat{\mathbf{u}}_0^{[k]}, \hat{\mathbf{u}}_1^{[k]}, \hat{\mathbf{z}}^{[k]}).$$

Details of the maximization of each of the parameters, i.e. $\nu_r, \sigma_r, \rho_r, \lambda_r, \Sigma_r$ and Σ_0 , are provided in the Supplementary Material.

3.3 Prediction

We carry out prediction in the classical sense; first estimating the parameter set $\boldsymbol{\theta}$ based on a collection of observations and subsequently, making use of the plug-in predictive distribution that treats the estimated parameters as truth to generate predictions at another set of locations (Diggle & Ribeiro 2007). Let us first fix some notation. Let $\mathbf{y} = (\mathbf{y}_{\text{pr}}^\top, \mathbf{y}_{\text{ob}}^\top)^\top$, denote the vector that stacks the vector of desired predictions $\mathbf{y}_{\text{pr}}$ at n_{pr} locations, and a vector of $\mathbf{y}_{\text{ob}}$ observations at n_{ob} locations. The plug-in predictive distribution of $\mathbf{y}_{\text{pr}}$ is then given by

$$p(\mathbf{y}_{\text{pr}} | \mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}}) = \int p(\mathbf{y}_{\text{pr}} | \mathbf{y}_{\text{ob}}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}, \hat{\boldsymbol{\theta}}) p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{z} | \mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}}) d\mathbf{u}_0 d\mathbf{u}_1 d\mathbf{z}.$$

Denote with Σ the covariance matrix of $\mathbf{y}$, and its partition for the observed and predicted part of the model

$$\Sigma = \begin{pmatrix} \Sigma_{\text{ob,ob}} & \Sigma_{\text{ob,pr}} \\ \Sigma_{\text{pr,ob}} & \Sigma_{\text{pr,pr}} \end{pmatrix}.$$

Applying standard results from multivariate normal theory we obtain

$$p(\mathbf{y}_{\text{pr}} | \mathbf{y}_{\text{ob}}, u_0, u_1, z, \hat{\boldsymbol{\theta}}) \sim N_{n_{\text{pr}}} \left\{ \frac{\rho u_0 + \lambda u_1}{\sqrt{z}} \mathbf{1}_{n_{\text{pr}}} + \Sigma_{\text{pr,ob}} \Sigma_{\text{ob,ob}}^{-1} \left(\mathbf{y}_{\text{ob}} - \frac{\rho u_0 + \lambda u_1}{\sqrt{z}} \mathbf{1}_{n_{\text{ob}}} \right), \right. \\ \left. \frac{\sigma^2}{z} (\Sigma_{\text{pr,pr}} - \Sigma_{\text{pr,ob}} \Sigma_{\text{ob,ob}}^{-1} \Sigma_{\text{ob,pr}}) \right\} \quad (8)$$

On the other hand, the density $p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{z} | \mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}})$ coincides with that of the E-step of the EM algorithm, replacing here the k -th parameter set with the parameter set $\hat{\boldsymbol{\theta}}$. Thus, in practice, we may draw samples from $p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{z} | \mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}})$ for as many regions as required, as indicated in Section 3.1, followed by samples for each regional set of prediction locations, according to (8).

4 Simulation Study

Here we examine the predictive performance of the proposed skew- t model, which we hereby denote as SKT, against a skew-Normal (SN) with $Z_r = 1$ in (1), a bi-resolution Gaussian model (biGau) with $Z_r = 1$ and $\lambda_r = 0$ in (1), and a Gaussian model (GAU) with $Z_r = 1, \lambda_r = 0$ and $\rho_r = 0$. We test these four models against samples generated from a Tukey g and h random field (Xu & Genton 2017), a trans-Gaussian process with transformation

$$\tau_{g,h}(z) = \begin{cases} g^{-1} \{ \exp(gz) - 1 \} \exp\left(\frac{hz^2}{2}\right), & g \neq 0, \\ z \exp(hz^2/2), & g = 0. \end{cases}$$

The transformation $\tau_{g,h}(z)$ allows to control skewness and kurtosis separately using the two parameters g and h , and are here sampled randomly for each region from a range of 0.3 to 0.5 for the former, and 0.03 to 0.06 for the latter. Using the biGau model as the underlying model to which the Tukey g and h transformation was subsequently applied, the resulting parameter values yielded skewness and kurtosis values ranging from approximately 1.25 to 3, and 6.5 to

25, respectively. We further assume $\phi_0 = 2$, for the range parameter governing the large-scale correlation function, for which region centroids are used to determine intra-region distances. The range parameter for each region is fixed at $\phi_r = 0.1$. We also fix $\sigma_r = 1$, and assume $\nu_r = 7$. A total of ten simulations is performed, each with $T = 150$ spatial replicates. We consider a spatial design composed of $R = 15$ regions, each consisting of a 1×1 square, arranged in a 4-by-4 grid. Within each region, 15 points are chosen at random, for a total of 225 spatial locations (see Fig. S1 in Supplementary Material). After the Tukey g and h has been simulated, all four models are fit and we compare differences in prediction accuracy as measured by the continuous ranked probability score (CRPS, Gneiting et al. (2007)).

Inference for the SKT, SN and biGau model is achieved using the EM algorithm outlined in Section 3 (or simplified versions thereof in the case of SN and biGau), with $M = 1000$ MCEM iterations for each spatial replicate t . (Inference for the GAU model is achieved by maximum likelihood.) We consider multiple initial parameter values, and do not observe convergence to local minima. Figure S2 displays the evolution of $\hat{\theta}^{[k]}$ and the Q-function for 500 iterations of the average of the ten simulations; parameter estimates can be seen stabilizing beyond the 200-th iteration, and close to their true values. The estimation of ν_r appears to be particularly noisy. A quantification of the (asymptotic) standard deviation would require EM-specific approaches such as the Supplemented Expectation Maximization (SEM, Meng & Rubin (1991)) algorithm. However, this task was computationally infeasible given the large number of temporal replicates and the number of repeated simulations. SEM was however performed in the application from the next section, and the results highlight how the standard deviation in the estimation of ν_r is more than one order of magnitude larger than all other parameters.

Table 1 provides a comparison of the out-of-sample average CRPS values for the four models for each of the 15 regions. The SKT model outperforms all other three models. An increase in model performances is highlighted when allowing two resolutions, i.e. from GAU (0.61) to biGau and SN (0.57), and an additional improvement is achieved when allowing for a skewed behavior

for the SKT model (0.56).

Table 1: Out-of-sample average continuous-ranked probability scores based on 10 spatial replicates for the SKT, SN and biGau and GAU models for the 15 regions, and their overall mean.

Model	Region															Mean
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
SKT	0.49	0.56	0.47	0.55	0.56	0.62	0.67	0.56	0.60	0.54	0.50	0.65	0.62	0.47	0.57	0.56
SN	0.54	0.57	0.50	0.55	0.59	0.66	0.52	0.55	0.62	0.65	0.58	0.56	0.55	0.65	0.52	0.57
biGau	0.54	0.57	0.50	0.55	0.59	0.66	0.52	0.55	0.62	0.65	0.58	0.57	0.55	0.65	0.52	0.57
GAU	0.55	0.60	0.51	0.61	0.55	0.74	0.61	0.61	0.80	0.68	0.59	0.55	0.63	0.63	0.44	0.61

5 Application

We use the Colorado climatological data obtained from the Geophysical Statistics Project at the National Center for Atmospheric Research (www.cgd.ucar.edu/stats/Data/US.monthly.met). The dataset provides monthly data for temperature maxima and minima, as well as precipitation, for the years 1895 to 1997. Here, we focus solely on annual total precipitation amounts (in millimeters), and consider only the 137 stations from the original 397 that have observations over the final 40 years, see Figure 1. We chose a subset of stations to allow an uninterrupted 40 years long record of annual data. We could in principle have added more, but that would have required adding missing data whose estimation is beyond the scope of this work. We refrain from any transformation of the data (e.g., log-transformed), keeping them to their original scale, except for a prior standardization. The lack of transformation implies an always positive quantity, while all the models we consider allow for values across the entire real line. Figure S1 in the supplementary material shows the aggregated histogram across all locations and years. As apparent from the plot, no data point was exactly equal to zero, and vast majority of the data are far away from zero as well, hence allowing us to approximate the probability of a negative event to zero. The median skewness across locations is -0.22 ($Q1=-0.69$, $Q3=0.27$), while the median kurtosis is $3.30(-3.34, 3.95)$, hence an overall non-Gaussian, skewed and heavy tailed behavior that justifies the use of a ST distribution introduced in the previous sections. We also found no indication

of temporal dependence. We calculated the autocorrelation function at each site, along with its asymptotic standard deviation, and we found that no site had a significant lag-1 value. Thus, at annual level it is perfectly reasonable to assume temporal independence. A modified model that would account for temporal dependence and possibly lack of space-time separability would likely be necessary for lower levels of aggregation, i.e. daily or sub-hourly temporal scales.

We extend the model in (1) to account for the non-zero location parameter, which we assume to be the same across all points belonging to the same region. The location transformation adds the term ξ_r to the conditional mean of the sampling distribution, which implies replacing the $\mathbf{y}_{r,t}$ terms with $\mathbf{y}_{r,t} - \xi_r \mathbf{1}_{n_r}$ in the formulas of the EM algorithm, and the additional update as part of the M-step,

$$\xi_r^{(k)} = \frac{\text{tr} \left(\Sigma_r^{-1} \left[1/\sigma_r^2 \sum_{t=1}^T \langle z_{r,t} \rangle (\mathbf{1}_{n_r} \mathbf{y}_{r,t}^\top + \mathbf{y}_{r,t} \mathbf{1}_{n_r}^\top) - 2/\sigma_r \sum_{t=1}^T (\langle \sqrt{z_{r,t}} u_{0,r,t} \rangle \rho_r + \langle \sqrt{z_{r,t}} u_{1,r,t} \rangle \lambda_r) \mathbf{1}_{n_r} \mathbf{1}_{n_r}^\top \right] \right)}{2 \text{tr} \left(\Sigma_r^{-1} \left[1/\sigma_r^2 \sum_{t=1}^T \langle z_{r,t} \rangle \mathbf{1}_{n_r} \mathbf{1}_{n_r}^\top \right] \right)}$$

We distinguish six regions, shown color-coded in Fig. 1, covering the Eastern plains, the foothills, the high peaks and the adjacent Southern and Western areas. As part of an exploratory analysis, we fit a multivariate skew- t distribution, with the help of the R package `sn` (Azzalini 2018), independently to each region, and find that the resulting parameter estimates exhibit consistent non-zero skewness in all regions, as well as heavy-tails in most, lending justification to the use of the SKT model. We fit the SKT model, but given the small number of regions ($R = 6$), the EM algorithm was unstable in the identification of the large-scale range parameter ϕ_0 , as it tended towards increasingly larger values, in order to reproduce the inter-regional dependence structure. We thus fixed it at a suitably large value, $\phi_0 = 10,000$. Estimates for the remaining parameters are shown in Table S1, while those for the GAU model fit are shown in Table S2. The evolution of the Q-function and parameter estimates of each M-step iteration are shown in Fig. S3.

Figure 2 shows the empirical correlation for each region as a function of the distance and the model-implied correlation function evaluated, using estimated parameters. We see overall

Figure 1: Location of precipitation stations, each region being assigned a different marker, with region number plotted at the centroid. The overlaid elevation map (in meters) is from the United States Geological Survey GTOPO30 digital elevation model.

good agreement between the fitted and empirical correlation, although the terminal value of the correlation function in region 1 seems sensibly too high, underscoring the tension between inducing inter-regional spatial dependence and skewness, and non-vanishing intra-regional spatial correlation. Figure 3 examines the inter-regional spatial dependence, contrasting the median of the correlations between points located in different regions and the model implied inter-regional correlation. From this figure it is apparent how the model overall adequately captures the inter-regional dependence as well.

We contrast the in-sample predictive performance between both models, using the first 10 spatial replicates of the 40 year period. Table 2 presents the CRPS values for each model, where it is seen that the SKT model offers better predictive accuracy. The results are dependent on the number R and location of the clusters, so a simulation study assessing the sensitivity against these factors could be performed. However, such a task is currently too computationally intensive as it would require re-running the model for different configurations. As a general rule of thumb, the regions should not be too small to have unstable estimates, and not too large to avoid losing

Figure 2: Empirical correlations (blue dots) for each of the six regions denoted in Fig. 1, and the respective correlation functions (red curves) of the SKT model (4) evaluated using parameter estimates.

information about local spatial structure. Moreover, if one is interested in a particular subregion, a clustering approach with more points in that region can be envisioned, and as long as prediction is not performed too close to the edges of the regions, no artifacts from the discontinuity should be apparent.

Table 2: Predictive performances of GAU and SKT model in terms of average continuous-ranked probability scores for the first ten replicated of the 40 year period.

Model	Region						Mean
	1	2	3	4	5	6	
SKT	0.39	0.36	0.33	0.45	0.41	0.36	0.38
GAU	0.51	0.50	0.47	0.60	0.50	0.60	0.53

6 Conclusions

In this work, we proposed a new model for non-Gaussian spatial data that does not rely on the classical conditional independence assumption of hierarchical models. We showed how the skew- t has a representation that can be used as a baseline for a new class of hierarchical models, and is

Figure 3: Matrix of kernel density estimates of the correlations between points belonging to the respective regions, along with median correlation (dashed blue) and the inter-regional correlation (solid red) based on Eq. (5), evaluated at the parameter estimates.

suitable for spatial data that have similar high-order moments.

The proposed model has a closed form expression for all moments, as well as a covariance between points. We chose to perform inference in a frequentist setting via an MCEM algorithm, with semi-closed form expressions for both the E and the M steps. The skew- t model also allows for a continuous underlying process within each region, and hence prediction can be performed conditionally on the observed data using the estimated parameters from the aforementioned MCEM algorithm. The proposed MCEM algorithms requires simulations of $\mathbf{u}_{0,t}$, $\mathbf{u}_{1,t}$ and $\mathbf{z}_t$ for all t in the M-step, so it is computational burdensome. However, full conditionals for $\mathbf{u}_{0,t}$, $\mathbf{u}_{1,t}$ are available and full MCMC is necessary only for $\mathbf{z}_t$. In order to fit larger data sets with this model, a promising direction of research to achieve approximate but faster computation is the use of Stochastic Approximation EM (SAEM, Jank (2006), Ordoez et al. (2018)), which relies on a Taylor approximation of the Q function.

The model could be generalized to account for spatial covariates. In the application we have

provided a common location parameter ξ_r across each region, so the model would also allow for a changing mean. The location parameter could also be generalized to allow for spatial covariates, i.e. $\xi_r = \mathbf{x}_r^\top \boldsymbol{\beta}$, for some design vector $\mathbf{x}_r$ and some parameters $\boldsymbol{\beta}$ to be estimated. This specification could also be made more flexible with a location specific scale parameter, but that would require a very large sample size to avoid identifiability issues.

The extent of the generalizability of an (additive) hierarchical representation of the skew- t is an open question. In principle, a similar representation could be sought for skew-elliptical models, but the algebra of the E and M steps, or full conditionals in a Bayesian version of this model, are expected to be non-trivial. Additional extensions to scale mixtures of skew-normal (Cabral et al. 2012) would be even more challenging as a closed form expression is not guaranteed, hence the computation would be considerably more burdensome. Besides, there would be subjectivity in the choice of the mixing distribution and for an application a sensitivity study should be performed.

Scalability is also a potential issue when applying such a model to considerably larger spatial data sets. Additional work is needed to seek solutions to either reparametrize the latent parameter space from the current dimension vector of size $3R$, or a Bayesian version of the model with either a Laplace or Hamiltonian approximation of the posterior.

Lastly, the time component here is used only to provide replicates and improve the inference, thus implicitly assuming space-time separability. This assumption is likely not realistic for data sets at high temporal resolution, and will thus require a new formulation that allows a temporally nonstationary model with a marginal structure identical or similar to the one presented in this work.

References

Allard, D. & Naveau, P. (2007), ‘A new spatial skew-normal random field model’, *Communications in Statistics: Theory and Methods* **36**, 1821–1834.

- 350 Arellano-Valle, R. B., Branco, M. D. & Genton, M. G. (2006), ‘A unified view on skewed distri-
 351 butions arising from selections’, *The Canadian Journal of Statistics / La Revue Canadienne*
 352 *de Statistique* **34**(4), 581–601.
- 353 Azzalini, A. (1985), ‘A class of distributions which includes the normal ones’, *Scandinavian*
 354 *Journal of Statistics* **12**, 171–178.
- 355 Azzalini, A. (2018), *The R package sn: The Skew-Normal and Related Distributions such as the*
 356 *Skew-t (version 1.5-3)*., Università di Padova, Italia.
 357 **URL:** <http://azzalini.stat.unipd.it/SN>
- 358 Azzalini, A. & Capitanio, A. (2003), ‘Distributions generated by perturbation of symmetry with
 359 emphasis on a multivariate skew t-distribution’, *Journal of the Royal Statistical Society: Series*
 360 *B (Statistical Methodology)* **65**(2), 367–389.
- 361 Azzalini, A. & Capitanio, A. (2014), *The Skew-Normal and Related Families*, Cambridge Uni-
 362 versity Press, IMS Monograph Series.
- 363 Azzalini, A. & Dalla Valle, A. (1996), ‘The multivariate skew-normal distribution’, *Biometrika*
 364 **83**(4), 715–726.
- 365 Bevilacqua, M., Caamaño, C., Arellano-Valle, R. B. & Morales-Onñate, V. (2018), ‘On spatial
 366 (skew) t processes and applications’, *arXiv preprint arXiv:1812.06310* .
- 367 Cabral, C. R. B., Lachos, V. H. & Prates, M. O. (2012), ‘Multivariate mixture modeling using
 368 skew-normal independent distributions’, *Computational Statistics & Data Analysis* **56**(1), 126
 369 – 142.
 370 **URL:** <http://www.sciencedirect.com/science/article/pii/S016794731100243X>
- 371 Castruccio, S., Huser, R. & Genton, M. G. (2016), ‘High-order composite likelihood inference
 372 for max-stable distributions and processes’, *Journal of Computational and Graphical Statistics*
 373 **25**(4), 1212–1229.
 374 **URL:** <https://doi.org/10.1080/10618600.2015.1086656>
- 375 Cressie, N. A. (1993), *Statistics for Spatial Data*, Wiley.

- 376 Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), ‘Maximum likelihood from incomplete
377 data via the em algorithm’, *Journal of the Royal Statistical Society: Series B (Statistical
378 Methodology)* **39**, 1–38.
- 379 Diggle, P. J., Tawn, J. & Moyeed, R. (1998), ‘Model-based geostatistics’, *Journal of the Royal
380 Statistical Society: Series C (Applied Statistics)* **47**(3), 299–350.
- 381 Diggle, P. & Ribeiro, P. (2007), *Model-based Geostatistics*, Springer, New York.
- 382 Genton, M. G. (2004), *Skew-Elliptical Distributions and Their Applications: A Journey Beyond
383 Normality*, Chapman and Hall / CRC, Boca Raton, FL.
- 384 Genton, M. G. & Zhang, H. (2012), ‘Identifiability problems in some non-gaussian spatial random
385 fields’, *Chilean Journal of Statistics* **3**, 171–179.
- 386 Gneiting, T., Balabdaoui, F. & Raftery, A. E. (2007), ‘Probabilistic forecasts, calibration
387 and sharpness’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*
388 **69**(2), 243–268.
- 389 Jank, W. (2006), ‘Implementing and diagnosing the stochastic approximation EM algorithm’,
390 *Journal of Computational and Graphical Statistics* **15**(4), 803–829.
- 391 Kim, H., Ha, E. & Mallick, B. (2004), Spatial prediction of rainfall using skew-normal processes,
392 in M. G. Genton, ed., ‘Skew-Elliptical Distributions and Their Applications: A Journey Beyond
393 Normality’, CRC Press, Boca Raton, FL, chapter 16, pp. 279–289.
- 394 Kim, H. & Mallick, B. K. (2004), ‘A Bayesian prediction using the skew-Gaussian processes’,
395 *Journal of Statistical Planning and Inference* **120**, 85–101.
- 396 Lachos, V. H., Cabral, C. R. B. & Zeller, C. B. (2018), Scale mixtures of skew-normal distribu-
397 tions, in ‘Finite Mixture of Skewed Distributions’, Springer, pp. 15–36.
- 398 Lachos, V. H., Ghosh, P. & Arellano-Valle, R. B. (2010), ‘Likelihood based inference for skew-
399 normal independent linear mixed models’, *Statistica Sinica* **20**, 303–322.
- 400 Lin, T. I. (2010), ‘Robust mixture modeling using multivariate skew t distributions’, *Statistics
401 and Computing* **20**(3), 343–356.

- Lin, T. I. & Lee, J. C. (2008), ‘Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data’, *Statistics in Medicine* **27**(9), 1490–1507.
- Lindgren, F. & Rue, H. (2015), ‘Bayesian spatial modelling with r-inla’, *Journal of Statistical Software* **63**(19), 1–25.
- Mahmoudian, B. (2017), ‘A skewed and heavy-tailed latent random field model for spatial extremes’, *Journal of Computational and Graphical Statistics* **26**(3), 658–670.
- McLachlan, G. & Krishnan, T. (2007), *The EM Algorithm and Extensions*, Vol. 382, John Wiley & Sons.
- Meng, X.-L. & Rubin, D. B. (1991), ‘Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm’, *Journal of the American Statistical Association* **86**(416), 899–909.
- Minozzo, M. & Ferracuti, L. (2012), ‘On the existence of some skew-normal stationary processes’, *Chilean Journal of Statistics* **3**, 157–170.
- Ordoez, J. A., Bandyopadhyay, D., Lachos, V. H. & Cabral, C. R. (2018), ‘Geostatistical estimation and prediction for censored responses’, *Spatial Statistics* **23**, 109 – 123.
- Rue, H., Martino, S. & Chopin, N. (2009), ‘Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2), 319–392.
- Schmidt, A. M., Gonçalves, K. & Velozo, P. L. (2017), ‘Spatio-temporal models for skewed processes’, *Environmetrics (with discussion)* **28**(6), 1821–1834.
- Tagle, F., Castruccio, S., Crippa, P. & Genton, M. G. (2019), ‘A non-Gaussian spatio-temporal model for daily wind speeds based on a multi-variate skew-t distribution’, *Journal of Time Series Analysis* **40**, 312–326.
- Xu, G. & Genton, M. G. (2017), ‘Tukey g -and- h random fields’, *Journal of the American Statistical Association* **112**, 1236–1249.
- Zhang, H. & El-Shaarawi, A. (2010), ‘On spatial skew-gaussian processes and applications’, *Environmetrics* **21**(1), 33–47.