Computational Statistics (2020) 35:153–173
https://doi.org/10.1007/s00180-019-00905-y

ORIGINAL PAPER

A scalable Bayesian nonparametric model for large spatio-temporal data

Zahra Barzegar[1] · **Firoozeh Rivaz**[1]

Received: 8 June 2018 / Accepted: 8 June 2019 / Published online: 12 June 2019
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract
The Bayesian nonparametric (BNP) approach is an effective tool for building flexible spatio-temporal probability models. Despite the flexibility and attractiveness of this approach, the resulting spatio-temporal models become computationally demanding when datasets are large. This paper develops a class of computationally efficient and easy to implement BNP models for large spatio-temporal data. To be more specific, we introduce a random distribution for the spatio-temporal effects based on a stick-breaking construction in which the atoms are modeled in terms of a basis system. In this framework, a low rank basis approximation and a vector autoregressive process are used to model spatial and temporal dependencies, respectively. We demonstrate that the proposed model is an extension of the Gaussian low rank model with similar computational complexity, hence it offers great scalability for large spatio-temporal data. Through a simulation study, we assess the performance of the proposed model. For illustration, we then analyze a set of data comprised of precipitation measurements.

Keywords Large datasets · Stick-breaking process · Non-stationarity · Non-Gaussianity

1 Introduction

With the advancement of technology in collecting data, spatio-temporal analysts often encounter large amounts of observations from many spatial locations over time. This type of data have applications in different sciences such as climatology, ecology,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00180-019-00905-y) contains supplementary material, which is available to authorized users.

✉ Firoozeh Rivaz
f_rivaz@sbu.ac.ir

Zahra Barzegar
z_barzegar@sbu.ac.ir

[1] Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Tehran, Iran

environmental health, and atmospheric science. Modeling and analysis of large spatio-temporal data raise several challenges in the applications. Since such datasets are often observed on a large spatial domain, they often show different behaviors at each time point such as non-stationary in the spatial covariance structure and non-Gaussianity in the marginal spatial distribution. On the other hand, the computational burden of statistical analysis is a great problem. For example, the computational complexity of the inverse and the determinant of the covariance matrix that are needed in both inference and prediction, are of cubic order with the total number of observations.

In recent years computational issues attracted many spatial statisticians and several approaches have been proposed; examples include Gaussian predictive processes (Banerjee et al. 2008), fixed rank kriging (Cressie and Johannesson 2008), covariance tapering (Furrer et al. 2006), Gaussian Markov random fields (Rue and Tjelmeland 2002; Rue and Held 2005), kernel convolution (Higdon 1998; Lemos and Sanso 2009) and approximation of the spatial likelihood (Vecchia 1988; Stein et al. 2004). Latter, some of these methods have been extended to the spatio-temporal setting. One of the widely used techniques is to approximate the spatial process by a basis system which leads to a low rank (LR) model. The LR representation of the process is the baseline of many classes of spatial and spatio-temporal models such as predictive process, kernel convolution and spatial and spatio-temporal random effects model (Cressie et al. 2010). The popularity of the LR models in analyzing large spatial and spatio-temporal datasets have been raised by ability to reduce computational cost and inducing flexible correlation structures. To be more specific, they are not restricted to stationary assumption, and the order of computational complexity to calculate exact predictions is linear with the total number of observations. Stein (2014) made some criticism about the ability of low rank models for approximating the likelihood of spatial processes with parametric covariances in certain low-noise situations. Although, high frequency or discontinuous basis functions can address this criticism (Bradley et al. 2015), Bradley et al. (2016) have shown that a carefully selected reduced rank set of basis functions can produce as good or better predictions than those based on the full rank alternatives. A common assumption in LR models is that the coefficients associated with the bases are considered as Gaussian random variables. While convenient, such assumption may be overly restrictive in many applied problems. As mentioned before, large spatio-temporal data often exhibit non-Gaussianity feature such as heavy tails, skewness and multimodality. An efficient strategy that allows us to mitigate the effect of Gaussianity assumption on inference and prediction is to use the Bayesian nonparametric (BNP) approach.

There is a substantial literature proposing BNP approach for constructing flexible models in different settings, such as density estimation (e.g. Escobar and West 1995; Ghosal et al. 1999; Pati et al. 2013; Cavatti Vieira et al. 2013; Canale and Scarpa 2016), spatial data analysis (Gelfand et al. 2005; Duan et al. 2007; Petrone et al. 2009; Reich and Fuentes 2012; Hosseinpouri and Khaledi 2016), regression (Walker and Mallick 1999; Hanson and Johnson 2002; Schörgendorfer et al. 2013) and time series analysis (Nieto-Barajas et al. 2008; Griffin and Steel 2011; Di Lucca et al. 2013; Nieto-Barajas and Contreras-Cristán 2014; Kalli and Griffin 2018). However, there are limited works available on spatio-temporal BNP modeling. Duan et al. (2007) introduced a model which induces temporal dependence in the usual way with autoregressive models where the error term follows a generalized spatial DP (SDP). More

precisely, they considered common dynamic model for spatio-temporal data and let the error term of the evolution equation follows a generalized SDP. Alternatively, Warren et al. (2012) extended the spatial kernel DP (SKDP) model of Reich and Fuentes (2007) to the spatio-temporal setting. Their model includes space and time simultaneously in both the weights and the atoms of the stick-breaking prior through the use of kernel functions. Recently, Gutiérrez et al. (2016) introduced a time dependent BNP model which is a nonparametric mixture of spatial parametric kernels. They induced temporal dependence through the mixing measure based on a stick-breaking construction with time varying weights and fixed atoms where the weights follow a Markovian process. Although the above BNP models provide useful frameworks to account for important features of data, they suffer from a large computational burden when the number of spatial locations is large. More precisely, at each iteration of the MCMC algorithm for the model proposed by Duan et al. (2007), we need to inverse an $n_t \times n_t$ matrix which requires $O(n_t^3)$ flops where n_t is the number of spatial locations at time point t. Therefore, when the number of spatial locations is large, the computational complexity is not negligible. A similar cost is incurred in the model proposed by Gutiérrez et al. (2016) and Warren et al. (2012).

In this paper, we aim to develop a flexible and easy to implement spatio-temporal BNP model in a computationally feasible way. More precisely, we are interested to introduce a model while relax the Gaussianity assumption, it preserves the properties of the low rank models. For this purpose, we propose a spatio-temporal model that is a nonparametric mixture of Gaussian kernels, where the mixing measure is a stick-breaking process that induces the spatio-temporal dependence through the atoms in terms of a basis system. Specifically, a low rank basis approximation and a vector autoregressive process are used to model spatial and temporal dependencies, respectively. We demonstrate that the proposed model is also an extension of the Gaussian low rank (GLR) model with the similar computational complexity. More precisely, the computational cost of our model increases linearly by the total number of observations, similar to the GLR, hence it offers great scalability for large to very large spatio-temporal data. Moreover, the proposed model induces a closed form expression for the mean and the covariance functions in a similar fashion to the low rank model.

This modeling approach can be considered as a temporal extension of the model proposed by Gelfand et al. (2005). To be more specific, Gelfand et al. (2005) introduced a novel SDP mixture model where spatial dependence is induced through the atoms. Indeed, the atoms are considered as a realization of a stationary Gaussian process. Their model allows for a large amount of flexibility, since the resulting random spatial process is neither Gaussian nor stationary. The proposed dependent model not only has similar properties to their model, but also it induces a closed form expression for the first and second order structures of the process, as pointed before. However, our modeling framework introduces models marginally rather than jointly in the Gelfand et al. (2005). Another way for constructing a temporal extension of the SDP, is the approach was proposed by Duan et al. (2007). As we pointed out before, when the spatio-temporal datasets are large, the computational burden of statistical analyses is not negligible.

The remainder of this paper is as follows. Section 2 presents the proposed model with its properties. Prior specification, posterior implementation and prediction are

described in Sect. 3. The computational complexity of our model is investigated in Sect. 4. Section 5 evaluates the performance of the proposed model in accordance with the GLR via simulated examples. Section 6 illustrates the approach through a real data set. Finally, in Sect. 7 we present a short discussion and some future directions.

2 A Bayesian nonparametric spatio-temporal model

Let's assume that the spatio-temporal random field $Z(.,.) = \{Z(\mathbf{s},t); \mathbf{s} \in D \subseteq \mathbb{R}^d, t \in \{1,2,\ldots\}\}$ is observed at locations $\mathbf{s}_{i,t}$ at times t for $i = 1,\ldots,n_t$ and $t = 1,2,\ldots,T$. Also, let the sampling model be

$$Z(\mathbf{s}_{it},t) = Y(\mathbf{s}_{it},t) + \varepsilon(\mathbf{s}_{it},t), \tag{1}$$

with

$$Y(\mathbf{s}_{it},t) = \mu(\mathbf{s}_{it},t) + \nu(\mathbf{s}_{it},t), \tag{2}$$

where $\mu(\mathbf{s}_{it},t)$ models the large-scale variability and is assumed to be a linear function of p regressors $\mathbf{f}_t'(.) = (f_{t,1}(.),\ldots,f_{t,p}(.))$ with the unknown vector of coefficients $\beta_t \in \Re^p$. Further, the random effects $\varepsilon(\mathbf{s}_{it},t)$ and $\nu(\mathbf{s}_{it},t)$ are a pure error process with distribution $\varepsilon(\mathbf{s}_{it},t) \sim N(0,\sigma_\varepsilon^2)$, and a spatio-temporal process which follows a random probability measure $G_{\mathbf{s}_{i,t},t}$, respectively.

For the random measure $G_{\mathbf{s}_{i,t},t}$, we admit a spatio-temporal stick-breaking process (ST-SBP) in which the spatial and temporal dependencies are induced through the atoms. More specifically,

$$G_{\mathbf{s}_{i,t},t} = \sum_{l=1}^{\infty} p_l \delta_{\omega_l(\mathbf{s}_{i,t},t)}, \tag{3}$$

where $\delta_{\omega_l(\mathbf{s}_{i,t},t)}$ denotes the point mass at $\omega_l(\mathbf{s}_{i,t},t)$. The weights $\{p_l\}_{l\geq 1}$ with their infinite sum of one, follow a stick-breaking process, i.e. $p_1 = V_1$, and for $l > 1$, $p_l = V_l \prod_{i<l}(1-V_i)$, with $V_l \overset{iid}{\sim} Beta(1,\alpha)$, $\alpha > 0$. The atoms $\{\omega_l(\mathbf{s}_{i,t},t)\}_{l\geq 1}$ are considered as realizations of a spatio-temporal random field. When the number of spatial locations at each time point t, i.e. n_t, is large, a computationally efficient approach to handle this kind of datasets is the low rank model. In this framework, the spatial and temporal dependencies are modeled through the low-rank basis approximation and a vector autoregressive process, respectively. More precisely,

$$\omega_l(\mathbf{s}_{i,t},t) = \mathbf{B}_t'(\mathbf{s}_{i,t})\theta_{l,t} + \xi(\mathbf{s}_{i,t},t), \tag{4}$$

where $\mathbf{B}_t'(\mathbf{s}_{i,t}) = (B_{1,t}(\mathbf{s}_{i,t}),\ldots,B_{r_t,t}(\mathbf{s}_{i,t}))$ represents a set of r_t ($r_t \ll n_t$) known spatio-temporal basis functions and $\theta_{l,t}' = (\theta_{1,l,t},\ldots,\theta_{r_t,l,t})$ is a zero-mean normal random vector with a covariance matrix $\mathbf{W}_t$. Also, $\xi(\mathbf{s}_{i,t},t)$ is the approximation error introduced by the dimension reduction. It is modeled as a white-noise Gaussian process in space and time with mean zero and variance σ_ξ^2, independent of $\theta_{l,t}$. Now, assume that a Markovian evolution for the random coefficient $\theta_{l,t}$ as

$$\theta_{l,t} = \mathbf{H}_t \theta_{l,(t-1)} + \zeta_{l,t}, \quad l = 1, 2, \ldots, \quad t = 1, 2, \ldots, T$$
$$\zeta_{l,t} \sim N(0, \mathbf{W}_t), \tag{5}$$

where $\theta_{l,0} \sim N(\mathbf{m}_0, \mathbf{C}_0)$, the $r_t \times r_t$ matrix $\mathbf{H}_t$ which is the evolution matrix, measures the dynamic dependence of the process $\{\theta_{l,t}\}_{t\geq 1}$, and the covariance matrix $\mathbf{W}_t$ controls the magnitude of the change at time t. Also, the innovation vector $\{\zeta_{l,t}\}_{l,t}$ are assumed to be independent of $\{\theta_{l,t}\}_{l,t}$.

According to the Eq. (4), we see that the problem of modeling a spatio-temporal process is reduced to that of modeling the coefficients associated with the bases. The main feature of the proposed model (hereafter, referred to as BNP-LR) is that it includes GLR as a limiting case. To be more specific, by letting $\alpha \to 0$, BNP-LR reduces to GLR. Specifically, by letting α to zero, the V_j, $j = 1, 2, \ldots$ are tend to be selected from a Beta distribution with mean 1 and variance zero. In this case, by the definition of weights, the p_1 tends to be one and the other weights, p_l; $l > 1$ tend to be zero. Moreover, we can rewrite $G_{\mathbf{s}_{i,t},t}$ as $\mathbf{B}_t'(\mathbf{s}_{i,t})G_t$ where

$$G_t = \sum_{l=1}^{\infty} p_l \delta_{\theta_{l,t}},$$

and $\theta_{l,t}$ follows the model in Eq. (5). Since, for any measurable set A, $E(G_t(A)) = G_{0t}(A)$ where G_{0t} is a normal distribution with the mean of $\mathbf{H}_t\theta^*_{t-1}$ and the covariance of $\mathbf{W}^*_t$ [induced by a vector autoregressive process as in (5)], hence

$$E(G_{s_{i,t},t}(A)) = \mathbf{B}_t'(\mathbf{s}_{i,t})G_{0t}(A).$$

In fact, the base measure of our proposed model is the same as the random effect's distribution of GLR.

The finite-dimensional conditional distribution of $\mathbf{Z}_t = \big(Z(\mathbf{s}_{1,t}, t), \ldots, Z(\mathbf{s}_{n_t,t}, t)\big)'$ given β_t, G_t and σ^2_ε is

$$\sum_{l=1}^{\infty} p_l N_{n_t}\left(\mathbf{Z}_t | \mathbf{F}_t\beta_t + \mathbb{B}_t\theta_{l,t} + \xi_t, \sigma^2_\varepsilon I_{n_t}\right) \tag{6}$$

where $\mathbb{B}_t = (\mathbf{B}_t(\mathbf{s}_{1,t}), \ldots, \mathbf{B}_t(\mathbf{s}_{n_t,t}))'$, $\mathbf{F}_t = (\mathbf{f}_t(\mathbf{s}_{1,t}), \ldots, \mathbf{f}_t(\mathbf{s}_{n_t,t}))'$ and N_{n_t} denotes n_t-dimensional normal distribution. It is easy to verify that any finite-dimensional distribution function of $\{Z(\mathbf{s}_{i,t}, t)\}_{i,t}$ satisfies the following Kolmogorov's conditions of symmetry and consistency:

(1) *Symmetry* For any permutation $\pi = (\pi_1, \ldots, \pi_{n_t})$ of the set $\{1, 2, \ldots, n_t\}$,

$$P_{n_t}(Z(\mathbf{s}_{\pi_1,t}, t) < z_{\pi_1}, \ldots, Z(\mathbf{s}_{\pi_{n_t},t}, t) < z_{\pi_{n_t}})$$
$$= P_{n_t}(Z(\mathbf{s}_{1,t}, t) < z_1, \ldots, Z(\mathbf{s}_{n_t,t}, t) < z_{n_t}).$$

(2) *Consistency* For any $n_t > 1$,

$$P_{n_t+1}(Z(\mathbf{s}_{1,t}, t) < z_1, \ldots, Z(\mathbf{s}_{n_t,t}, t) < z_{n_t}, Z(\mathbf{s}_{n_t+1,t}, t) < \infty) = P_{n_t}(Z(\mathbf{s}_{1,t}, t) < z_1, \ldots, Z(\mathbf{s}_{n_t,t}, t) < z_{n_t}).$$

Details are presented in the supplemental Appendix B. The mean and covariance functions of BNP-LR are (in the sequel of the paper, for simplicity, we consider $r_t = r$ for $t = 1, \ldots, T$)

$$E(Z(\mathbf{s}_{i,t}; t)) = \mathbf{f}_t'(\mathbf{s}_{i,t})\beta_t + \mathbf{B}_t'(\mathbf{s}_{i,t}) \left(\left(\prod_{r=1}^{t} \mathbf{H}_{t-r+1} \right) \mathbf{m}_0 \right),$$

and

$$\begin{aligned} &Cov(Z(\mathbf{s}_{i,t}, t), Z(\mathbf{s}_{i',t+k}, t+k)) \\ &= \frac{1}{1+\alpha} \mathbf{B}_t'(\mathbf{s}_{i,t}) Var(\theta_t^*)(\mathbf{H}_{t+k}\mathbf{H}_{t+k-1} \ldots \mathbf{H}_{t+1})' \\ &\quad \times \mathbf{B}_{t+k}(\mathbf{s}_{i',t+k}) + \sigma_\xi^2 I(i = i', k = 0) + \sigma_\varepsilon^2 I(i = i', k = 0), \end{aligned} \tag{7}$$

respectively, where $Var(\theta_t^*)$ is given by

$$\begin{aligned} Var(\theta_t^*) = &\left(\prod_{r=1}^{t} \mathbf{H}_{t-r+1} \right) \mathbf{C}_0 \left(\prod_{r=1}^{t} \mathbf{H}_{t-r+1} \right)' \\ &+ \sum_{r=1}^{t-1} \left(\prod_{s=1}^{t-r} \mathbf{H}_{t-s+1} \right) \mathbf{W}_r \left(\prod_{s=1}^{t-r} \mathbf{H}_{t-s+1} \right)' + \mathbf{W}_t. \end{aligned} \tag{8}$$

Details have been provided in the supplemental Appendix A. Interestingly, these functions have closed form expressions with similar structures as in GLR. Since $\alpha > 0$, the covariance function of BNP-LR is strictly less than that of GLR's. This is not surprising because the nonparametric mixture models provide conditions that allow observations to come from different models. This leads to higher uncertainty in model specification and therefore reduction of the overall dependency with respect to the GLR. A computational problem of BNP-LR is that the stick-breaking prior (3) is infinite mixture. One way to make the stick-breaking procedure applicable in practice is to approximate it with a finite number of mixtures. An alternative approach for posterior sampling which does not involve an approximation of the nonparametric prior can be designed based on the slice sampling presented in Walker (2007) and Kalli et al. (2011), that described in the next section.

When BNP-LR model is adopted as the sampling model, two issues need to be considered for their effect on model fitting, prediction results and the computational burden. These issues are the basis functions and the evolution matrix $\mathbf{H}_t$. We address them in the following two subsections.

2.1 Basis functions

Selecting an appropriate set of basis functions depends on the three following matters:

- The type of basis functions,
- The rank of basis functions (the number of knots),
- The location of knots.

In general there are two class of basis functions: orthogonal (e.g., Fourier, orthogonal polynomials and empirical orthogonal functions (EOFs)) and non-orthogonal (kernel based) (Gelfand et al. 2010). As each of these basis functions has advantages and disadvantages, there is no consensus about the best type of basis function in the literature. However, it is recommended to use multiresolutional one to capture different scales of spatial variations (Cressie and Johannesson 2008). One obvious class of multiresoltional basis functions is bisquare given by

$$B_{j(l)}(\mathbf{s}) \equiv \begin{cases} \{1 - (||\mathbf{s} - \mathbf{u}_{j(l)}||/r_l)^2\}^2 & ||\mathbf{s} - \mathbf{u}_{j(l)}|| \leq r_l \\ 0 & otherwise \end{cases} \tag{9}$$

where $\{u_{j(l)}\}_j$ are the center points of l_{th} resolution and $||.||$ denotes the Euclidean distance. Additionally, the radius of a bisquare basis function of a particular resolution, r_l, is defined as 1.5 times the shortest distance between the center points of that resolution (Cressie and Johannesson 2008). Besides to multiresolution property, bisquare basis functions have another appealing features such as they can be evaluated at any point in spatial domain without needing interpolation. Although, the proposed BNP-LR model is applicable with any kind of basis functions, we use bisquare basis functions in the analysis of both simulated and real datasets.

A key component in the low-rank representation is the number of the basis functions. It is clear that a large number of knots is preferable for prediction, but more knots generally lead to increased computational cost. Several remedies have been proposed in the literature to tackle this problem. The main idea in this context is based on implementing the analysis over different number of knots and evaluating the stability of results using a criterion function such as spatially averaged predictive variance (Gelfand et al. 2012) and Akaike information criterion (AIC) (Bradley et al. 2011). Alternatively, the number and the location of knots could be considered random and estimated similar to the method in Katzfuss (2013). However, this approach imposes extra computation (Heaton et al. 2014). In this paper, by using deviance information criterion (DIC) which simultaneously controls goodness of fit and model complexity, we select optimal number of basis functions in the real data example.

The only remaining issue is to select the location of knots for each resolution. It is clear that the location of the basis functions should be covered the entire spatial domain. Also to capture the boundary effects, it is recommended that some knots are selected outside the study region. For fairly equally distributed data locations, one possibility is to select locations on a grid overlaid on the domain (Banerjee et al. 2010). For highly irregularly distributed locations, one can use the clustering algorithm such as k-means for selecting the locations of knots (Kaufman and Rousseeuw 1990). Another strategy could be to use design-based approach (Gelfand et al. 2012). When the basis functions

are multiresolution, the locations of the basis functions from different resolutions should not be coincident. Two suitable approaches for selecting the locations of knots at each resolution are quadtree structure (Nguyen et al. 2012) and Discrete Global Grid (DGG) (Sahr et al. 2003; Katzfuss and Cressie 2011). These two approaches typically imply a space-covering design. In this study we use quadtree design idea to select the location of basis functions.

2.2 The evolution matrix

Since the properties of the base measure (3) are related to the evolution matrix $\mathbf{H_t}$'s, appropriate determination of it is necessary. To avoid identifiability problem, we assume the matrix $\mathbf{H_t}$ and $\mathbf{W}_t$ are time invariant. Several parameterizations for $\mathbf{H}$ have been proposed in the literature. The simplest parameterization is to assume that $\mathbf{H} = \mathbf{I}$, corresponding to a multivariate random walk (Gelfand et al. 2005). Another parameterizations are $\mathbf{H} = \rho\mathbf{I}$ and $\mathbf{H} = diag(\rho_1, \ldots, \rho_r)$ (Xu et al. 2005). Alternatively, the evolution matrix $\mathbf{H}$ could be considered totally unknown and estimated similar to the method in Katzfuss and Cressie (2011). According to our experiments (not shown here), an unknown evolution matrix leads to some identifiability problems. In the spatio-temporal setting presented here, the number of basis functions r is larger than the number of times. Therefore, without any prior information, the elements of $\mathbf{H}$ are not identifiable. Even with a block-diagonal matrix for $\mathbf{H}$, where blocks are evolution matrices for each resolution, we observed non-identifiability of $\mathbf{H}$. Hence, in the analysis of real data, the evolution matrix is considered to be $diag(\rho_1, \ldots, \rho_r)$.

3 Posterior inference and prediction

The most widely used posterior inference methods in BNP models are Markov Chain Monte Carlo (MCMC) methods. Our method is based on a Gibbs MCMC algorithm with slice sampling steps (Walker 2007) that incorporates the forward filtering backward sampling ($FFBS$) algorithm (Carter and Kohn 1994; Frühwirth-Schnatter 1994 and West and Harrison 1997). In a Gibbs sampler, each unknown variable is updated from its full conditional distributions, that is proportional to the joint distribution of all variables. Due to the conditional independency, the joint distribution can be written as the product of the data model, the process model and the parameter (prior) model. To specify the data model, we introduce latent $\phi_1, \ldots, \phi_T$ where $\phi_t = h$ if the lth $\theta_{\cdot,t}$ is drawn from the hth mixture component. Therefore, conditionally on ϕ_t, observations from time point t, $\mathbf{Z}_t$, is drawn from the hth mixture component. Then, conditionally on ϕ_t observations from time point t; $\mathbf{Z}_t$ is drawn from a single multivariate normal distribution. We consider an augmented model given β_t, G_t and σ_ε^2 as

$$f(\mathbf{Z}_t, u_t, \phi_t) = I(u_t < p_{\phi_t}) N\left(\mathbf{Z}_t | \mathbf{F}_t\beta_t + \mathbb{B}_t\theta_{\phi_t,t} + \xi_t, \sigma_\varepsilon^2 I_{n_t}\right) \tag{10}$$

where u_t is a uniform random variate on $(0, p_{\phi_t})$.

To compute the full conditional distributions of parameters we utilize data augmentation to incorporate $\Theta_{1:T} = (\Theta_1, \ldots, \Theta_T)$. Let $\mathbb{Z} \equiv (\mathbf{Z}_1', \ldots, \mathbf{Z}_T')'$ be the vector of all observations. By defining Δ^P as the process model parameters, including $\beta_{1:T}$, $p_{1:T}, \phi_{1:T}, \sigma_\xi^2, \alpha, \mathbf{W}, u_{1:T}$ and Ξ^P as the vector of $(\xi_1', \ldots, \xi_T')'$, the joint distribution can be writen as

$$[\mathbb{Z}, \Theta_{1:T}, \Delta^P, \Xi^P] = [\mathbb{Z}|\Theta_{1:T}, \Delta^P, \Xi^P][\Theta_{1:T}, , \Xi^P|\Delta^P][\Delta^P]$$

where the notation $[X|Y]$ and $[X]$ stand for the conditional probability density function of X given Y and the marginal density function of $[X]$, respectively. Also, we use $[X|.]$ to denote the conditional distribution of variable X given other variables including the data.

The specification of our Bayesian hierarchical model is completed by placing priors on the remaining parameters and determining $[\Delta^P]$. Due to the independency assumption, the joint distribution of Δ^P is the product of the distribution of each parameter. To avoid identifiability problem, the measurement error variance σ_ε^2 is assumed to be known and determined based on the emprical variogram (Kang et al. 2010). Also, we assume that the evolution matrix $\mathbf{H}$ is known. However as pointed out before, in the applied example we set $\mathbf{H} = diag(\rho_1, \ldots, \rho_r)$, and for each diognal element we consider a uniform prioir on $(-1, 1)$. Moreover, Metropolis–Hastings (MH) updates (Metropolis et al. 1953; Hastings 1970) is employed to sample $\rho_1, \ldots, \rho_r$. For other parameters, we adopt independent vague normals for the elements of the regression parameters, $\beta_t, t = 1, \ldots, T$, and a vague inverse Gamma prior $IG(a_0, b_0)$ with the mean of $b_0/(a_0 - 1)$ for σ_ξ^2. Also a gamma prior with the mean of c_0/d_0 is considered for α. We assign an inverse Wishart prior to $\mathbf{W}$, i.e. $\mathbf{W} \sim IW(\nu_0, \Psi)$ where ν_0 denotes the degree of freedom and Ψ is a scale positive definite matrix.

The iterations of the Gibbs sampler will consist of sampling from closed form distributions which are presented below.

For each $t, \quad t = 1, \ldots, T$, update u_t from the uniform distribution on the interval $(0, p_{\phi_t})$.

For each $h, h = 1, \ldots, N$,

$$V_h|. \sim Beta\left(1 + M_h, \alpha + \sum_{k>h} M_k\right),$$

where $N := max_t\{N_t\}$, N_t is the largest ϕ_t such that $u_t < p_{\phi_t}$. This is equivalent to determine N_t such that $\sum_{i=1}^{N_t} p_i > 1 - u_t$.

Other full conditionals are

$$\phi_t|. \sim \sum_{A_p(u_t)} p_h f(\mathbf{Z}_t|\Delta^P, \xi_t, \theta_{h,t})\delta_h(\phi_t),$$

$$\beta_t|. \sim N\left(\left[\frac{\mathbf{F}_t'\mathbf{F}_t}{\sigma_\varepsilon^2} + V_0^{-1}\right]^{-1}\left[\frac{\mathbf{F}_t'\left[\mathbf{Z}_t - \mathbb{B}_t'\theta_{\phi_t,t} - \xi_t\right]}{\sigma_\varepsilon^2} + V_0^{-1}\beta_0\right],\right.$$

$$\left[\frac{\mathbf{F}_t'\mathbf{F}_t}{\sigma_\varepsilon^2}+V_0^{-1}\right]^{-1}\Bigg),$$

$$\mathbf{W}|.\sim IW(\nu_0+TN+1,\Psi+\sum_{t=1}^{T}\sum_{h=1}^{N}(\theta_{h,t}-\mathbf{H}\theta_{h,t-1})(\theta_{h,t}-\mathbf{H}\theta_{h,t-1})'),$$

$$\alpha|.\sim G\left(N+c_0-1,d_0-log\left(1-\sum_{h=1}^{N-1}p_h\right)\right),$$

$$\xi_t|.\sim N_{n_t}\left(\left(\frac{1}{\sigma_\varepsilon^2}+\frac{1}{\sigma_\xi^2}\right)^{-1}\frac{1}{\sigma_\varepsilon^2}\left(\mathbf{Z}_t-\mathbf{F}_t\beta_t-\mathbb{B}_t'\theta_{\phi_t,t}\right),\left(\frac{1}{\sigma_\varepsilon^2}+\frac{1}{\sigma_\xi^2}\right)^{-1}\right),$$

and

$$\sigma_\xi^2|.\sim IG\left(\left(\sum_{t=1}^{T}n_t-1\right)/2,\sum_{t=1}^{T}\xi_t'\xi_t\right),$$

where $A_p(u_t)=\{j;u_t<p_j\}$ and M_h is the number of ϕ_t's that are equal with h. The details of full conditionals are presented in the supplemental Appendix C. Since all full conditionals are standard distributions, sampling from them is straightforward.

To sample from the full conditional distribution of the latent process $\Theta_{1:T}=(\Theta_1,\ldots,\Theta_T)$, we use the $FFBS$ algorithm. The first stage of $FFBS$ is the recursive application of the forward filter in time for $t=1,\ldots,T$. More precisely, for each t, $t=1,\ldots,T$ and h, $h=1,\ldots,N$, $\theta_{h,t}$'s can be sampled from $N(\mathbf{m}_{h,t},\mathbf{C}_{h,t})$ where

$$\mathbf{m}_{h,t}=\mathbf{a}_{h,t}+\mathbf{A}_{h,t}\mathbf{e}_{h,t},$$

and

$$\mathbf{C}_{h,t}=\mathbf{R}_{h,t}-\mathbf{A}_{h,t}\mathbf{Q}_{h,t}\mathbf{A}_{h,t}',$$

with $\mathbf{A}_{h,t}=\mathbf{R}_{h,t}\mathbb{B}_t\mathbf{Q}_{h,t}^{-1}$, $\mathbf{e}_{h,t}=\mathbf{Y}_t-\mathbf{f}_{h,t}$, $\mathbf{f}_{h,t}=\mathbf{F}_t\beta_t+\mathbb{B}_t\mathbf{a}_{h,t}$, $\mathbf{Q}_{h,t}=\mathbb{B}_t'\mathbf{R}_{h,t}\mathbb{B}_t+\sigma_\varepsilon^2\mathbf{I}+\sigma_\xi^2\mathbf{I}$, $\mathbf{a}_{h,t}=\mathbf{H}\mathbf{m}_{h,t-1}$ and $\mathbf{R}_{h,t}=\mathbf{H}\mathbf{C}_{h,t-1}\mathbf{H}'+\mathbf{W}$. It should be noted that for every h, $\mathbf{m}_{h,0}=\mathbf{m}_0$ and $\mathbf{C}_{h,0}=\mathbf{C}_0$ are known. The second stage of the $FFBS$ starts by sampling $\theta_{h,T}$ from its full conditional distribution, i.e. $N(\mathbf{m}_{h,T},\mathbf{C}_{h,T})$, for each $h=1,\ldots,N$, and then continues with the recursive backward sampling from $N(\mathbf{d}_{h,t},\mathbf{D}_{h,t})$ for $t=T-1,T-2,\ldots,2,1$, where $\mathbf{d}_{h,t}=\mathbf{m}_{h,t}+\mathbf{E}_{h,t}(\theta_{h,t+1}-\mathbf{a}_{h,t+1})$, $\mathbf{D}_{h,t}=\mathbf{C}_{h,t}-\mathbf{E}_{h,t}\mathbf{R}_{h,t+1}\mathbf{E}_{h,t}'$ and $\mathbf{E}_{h,t}=\mathbf{C}_{h,t}\mathbf{H}\mathbf{R}_{h,t+1}^{-1}$.

One of the primary goals of spatio-temporal data analysis is to predict the underlying process at new locations and/or times. Under the proposed model in Sect. 2, the predictive distribution of $Y(\mathbf{s}_0,t_0)$ is

$$f(Y(\mathbf{s}_0,t_0)|\mathbb{Z})=\int f(Y(\mathbf{s}_0,t_0)|\Theta_{t_0},\Delta_{t_0}^P)f(\Theta_{t_0}|\Delta_{t_0}^P)\pi(\Delta_{t_0}^P|\mathbb{Z})d\Delta_{t_0}^P d\Theta_{t_0}.$$

where $\Delta_{t_0}^P$ denotes the parameters of the aforementioned Δ^P at time point t_0. Given parameter estimates, the Rao-Blackwellized estimate of the predictive distribution $Y(\mathbf{s_0},\mathbf{t_0})$ is

$$\frac{1}{L}\sum_{l=1}^{L} f(Y(\mathbf{s}_0, t_0)|\Theta_{t_0}^{(l)}, \Delta_{t_0}^{P(l)}),$$

where $f(Y(\mathbf{s}_0, t_0)|\Theta_{t_0}^{(l)}, \Delta_{t_0}^{P(l)})$ is $N(f'_{t_0}(\mathbf{s}_0)\beta_{t_0}^{(l)} + \mathbb{B}'_{t_0}(\mathbf{s}_0)\theta^{(l)}_{\phi_{t_0}^{(l)},t_0}, \sigma_\xi^{2(l)})$.

4 Computational complexity of the BNP-LR

In this section, we evaluate the computational feasibility of our proposed model. The computational complexity is reported in terms of the number of observed time points, the number of regression covariates, the dimension of basis functions, the number of observed locations at each time point t, n_t, and the number of the mixing components N in the BNP-LR model.

At each iteration of the MCMC algorithm, we need to implement the FFBS algorithm. Following Sect. 3, in the forward filtering step, it is required to calculate $\mathbb{B}_t\mathbf{Q}_{h,t}^{-1}$ and $\mathbf{A}_{h,t}\mathbf{Q}_{h,t}\mathbf{A}'_{h,t}$ where $\mathbf{Q}_{h,t} = \mathbb{B}'_t\mathbf{R}_{t,h}\mathbb{B} + (\sigma_\varepsilon^2 + \sigma_\xi^2)\mathbf{I}$. Using Sherman-Morrison-Woodbury formula to invert $\mathbf{Q}_{h,t}$, we have

$$\begin{aligned}\mathbb{B}_t\mathbf{Q}_{h,t}^{-1} &= \mathbb{B}_t(((\sigma_\varepsilon^2 + \sigma_\xi^2)I)^{-1} - ((\sigma_\varepsilon^2 + \sigma_\xi^2)I)^{-1}\mathbb{B}'_t(\mathbf{R}_{t,h}^{-1} \\ &\quad + \mathbb{B}_t((\sigma_\varepsilon^2 + \sigma_\xi^2)^{-1}\mathbb{B}'_t)\mathbb{B}_t((\sigma_\varepsilon^2 + \sigma_\xi^2)I)^{-1}) \\ &= (\sigma_\varepsilon^2 + \sigma_\xi^2)^{-1}\mathbb{B}_t - (\sigma_\varepsilon^2 + \sigma_\xi^2)^{-2}\mathbb{B}_t\mathbb{B}'_t(\mathbf{R}_{t,h}^{-1} + (\sigma_\varepsilon^2 + \sigma_\xi^2)^{-1}\mathbb{B}_t\mathbb{B}'_t). \quad (11)\end{aligned}$$

This reduces computational complexity from $O(n_t^3)$, to $O(r^2 n_t)$. In addition, by defining $\mathbf{M}_{t,h} = \mathbf{A}_{t,h}\mathbf{B}'_t$, we have

$$\begin{aligned}\mathbf{A}_{t,h}\mathbf{Q}_{t,h}\mathbf{A}'_{t,h} &= \mathbf{A}_{t,h}(\mathbb{B}'_t\mathbf{R}_{t,h}\mathbb{B}_t + (\sigma_\varepsilon^2 + \sigma_\xi^2)I)\mathbf{A}'_{t,h} \\ &= \mathbf{A}_{t,h}\mathbb{B}'_t\mathbf{R}_{t,h}\mathbb{B}_t\mathbf{A}'_{t,h} + (\sigma_\varepsilon^2 + \sigma_\xi^2)\mathbf{A}_{t,h}\mathbf{A}'_{t,h} \\ &= \mathbf{M}_{t,h}\mathbf{R}_{t,h}\mathbf{M}'_{t,h} + (\sigma_\varepsilon^2 + \sigma_\xi^2)\mathbf{A}_{t,h}\mathbf{A}'_{t,h} \quad (12)\end{aligned}$$

which has the computation cost of $O(r^2 n_t)$, leading to $O(Nr^2 n_t)$ for N components of ST-SBP. The other stage of $FFBS$ algorithm costs $O(r^3)$ floating operations. Other calculations at each iteration of the MCMC algorithm need only $O(p^3) + O(r^3)$ where $O(p^3)$ is required to invert the covariance matrix in the posterior density of the regression parameters and $O(r^3)$ is induced by calculating the determinant in the posterior density of $\mathbf{W}_t$. Therefore, since p, r, and N are fixed, the computational complexity at each iteration of the MCMC algorithm is of order $O(n)$ with n the total number of observations in space and time.

5 Simulation study

To assess the performance of the proposed model in terms of model fitting and prediction accuracy, we conducted a simulation study comparing our BNP-LR model with

the Gaussian LR model in two different situations. More precisely, in the Example 1, the data are generated from a full rank non-stationary and non-Gaussian process while in the next example, we generate data from low rank Gaussian and non-Gaussian models.

5.1 Example 1

We generate 15 datasets at 1000 randomly selected locations within $[0, 50] \times [0, 50]$ square for $t = 1, \ldots, 10$ from model (1) where $Y(\mathbf{s}_{i,t}, t)$ follows a two component mixture of independent Gaussian processes. The kth process, $k = 1, 2$, is a Gaussian process with constant mean μ_k and covariance function $\sigma(\mathbf{s}_{i,t})\sigma(\mathbf{s}_{j,t}) \exp(-\tau_k||\mathbf{s}_{i,t} - \mathbf{s}_{j,t}||)$, for each $i, j = 1, 2, \ldots, 1000$ and $t = 1, 2, \ldots, 10$, where at time t the first process is sampled with probability λ_t and the second process is sampled with probability $1 - \lambda_t$. Specifically, the time dependency is induced by λ_t and we set $\lambda_t = \frac{\delta_t}{\delta_t + \gamma_t}$, $t = 1, \ldots, T$, such that $\delta_t|\delta_{t-1} \sim LN(\log(\delta_{t-1}), 0.25)$ and $\gamma_t|\gamma_{t-1} \sim LN(\log(\gamma_{t-1}), 0.25)$ with $\delta_0 = \gamma_0 = 1.5$, where LN denotes log-normal distribution. Moreover, we set $\mu_1 = -2$, $\mu_2 = 2$, $\tau_1 = \tau_2 = 0.0025$, and $\sigma_\varepsilon^2 = 0.25$. Similar to Gelfand et al. (2005), $\sigma^2(\mathbf{s})$ is modeled as $\max(\varsigma(\mathbf{s}), 1)$ with $\varsigma(\mathbf{s}) = \sigma^2\{(\text{lat}(\mathbf{s}) - \text{midlat})^2 + (\text{lon}(\mathbf{s}) - \text{midlon})^2\}$, where $\sigma^2 = 0.05$, and $\text{lat}(\mathbf{s})$ and $\text{lon}(\mathbf{s})$ denote the latitude and longitude for location $\mathbf{s}$ and $\text{midlat} = (\max \text{lat}(\mathbf{s}) + \min \text{lat}(\mathbf{s}))/2$ and $\text{midlon} = (\max \text{lon}(\mathbf{s}) + \min \text{lon}(\mathbf{s}))/2$. It is worth mentioning that this simulation strategy yields a non-stationary process which is non-Gaussian.

After generating the datasets, we randomly select 10% of observations at each time point $t = 1, \ldots, 10$ as test data. Then, $n_t = 900$ samples at each time point are used for the model estimation and prediction at the locations of remaining samples.

To apply BNP-LR and GLR models, we consider a bisquare basis function with two resolutions, one with 25 knots and the other one with 16 knots. Also, we set $\mu(s_{i,t}, t) = \beta$. For priors, we consider $\beta \sim N(0, 10^3)$, $\sigma_\xi^2 \sim IG(0.01, 0.01)$, $\mathbf{W} \sim IW(r, I_r)$, and for the stick-breaking parameter α, we choose $G(2, 1)$. According to our sensitivity analysis, we see that estimations of parameters are robust in the face of moderate changes in the prior of α. For each datasets, the MCMC algorithm was run with a total number of 20,000 iterations based on BNP-LR and GLR where we use a simple random walk, $\mathbf{H}_t = \mathbf{I}$, for describing the evolution of atoms. The posterior inferences are based on the last 15,000 iterations. To reduce the correlation between samples after burn in time, the lag value was taken to be 5.

To compare GLR and BNP-LR models, we compute the root mean squared prediction error (RMSPE) and the deviance information criterion (DIC) as introduced by Spiegelhalter et al. (2002) and defined as $\text{DIC} = \bar{D} + p_D$. Here $\bar{D}$ is the posterior mean of the deviance and p_D is the effective number of parameters. Smaller RMSPE indicates better predictions. Also, the smaller the DIC, the better the trade-off between model fit and complexity.

Figure 1 shows the distribution of the two criteria across the 15 sets of holdout samples. In general, the BNP-LR model performs better than the GLR under two criteria. However, in term of DIC criterion, the BNP-LR has overwhelming performance com-

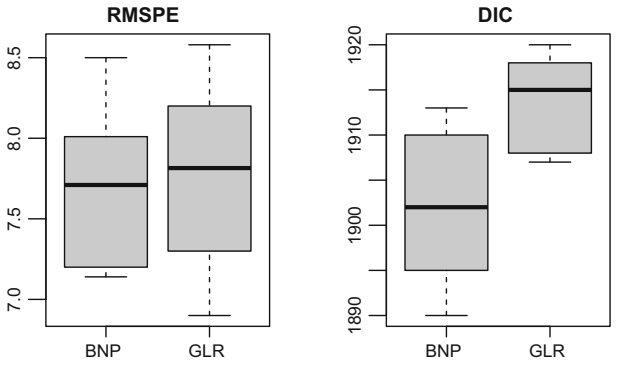


Fig. 1 Box plots of the RMSPE and DIC for two models fit to the simulated data in Example 1, summarized for each of 15 holdout replicates

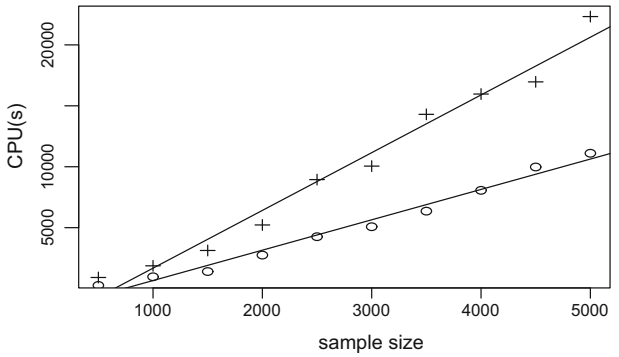


Fig. 2 Computational complexity (CPU time measured in seconds on a 1.8 GHz Intel R i7-8550U computer) of the proposed model with 5 mixture components ("o") and 10 mixture components ("+") as a function of sample size, n

pared to the GLR. The p_D of BNP-LR and GLR models averaged over 15 simulated datasets are 1788 and 1792, respectively.

As mentioned before, the computational complexity of the proposed model is of the order $O(Nr^2n)$ with n as the total number of observations. We carried out a small simulation study to demonstrate it, numerically. The data were generated as above mechanism with different sample sizes, 50×10, $100 \times 10, \ldots, 450 \times 10$ where the first and second numbers indicate the number of locations and time points, respectively. To fit the BNP-LR model, we used the bisquare basis function with 41 knots from two resolutions, one with 25 knots and the other one with 16 knots. Also, two values were considered for the number of mixture components, N, to assess its impact on computational complexity. Figure 2 shows the CPU times of running 1000 iterations of the MCMC algorithm for $N = 5$ and $N = 10$. As observed, there is a linear pattern between CPU times and sample sizes. As it is expected, the computational cost increases with the number of mixture components.

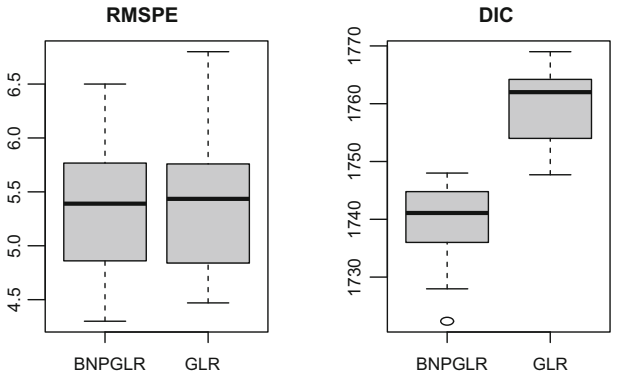


Fig. 3 Box plots of the RMSPE and DIC for two models fit to the simulated data from the model (13), summarized for each of 15 holdout replicates

5.2 Example 2

In this example, we simulate several datasets from Gaussian and non-Gaussian low-rank models to assess the BNP-LR and GLR. First, we generate 15 datasets at 1000 randomly selected locations on the square $[0, 50] \times [0, 50]$ for $t = 1, \ldots, 10$ from the following mechanism. For each location $\mathbf{s}_{i,t}, i = 1, 2, \ldots, 1000$ and $t = 1, 2, \ldots, 10$, we generate $Z(\mathbf{s}_{i,t}, t)$ from a mixture of two constant normal distributions with fixed locations and scales but time varying weights, that is

$$\begin{aligned} &Z(\mathbf{s}_{i,t}, t)|\beta, \eta_t, \sigma_\varepsilon^2 \sim N(\mu(\mathbf{s}_{i,t}, t) + B'(\mathbf{s}_{i,t})\eta_t, \sigma_\varepsilon^2), \\ &\eta_t \sim \lambda_t N(.|\mu_1, I) + (1 - \lambda_t)N(.|\mu_2, I), \end{aligned} \tag{13}$$

where $\mu(\mathbf{s}_{i,t}, t) = 2$, $\mu_1 = (1.5, \ldots, 1.5)'$, $\mu_2 = (-1.5, \ldots, -1.5)'$ and $\sigma_\varepsilon^2 = 0.25$. To induce time dependency in the weights, we model λ_t similar to the previous example. We note that the above mechanism to generate data is not included in our specification in Sect. 2. In fact, the data generation mechanism is not in the class of models defined in this paper.

Further feature that we consider is a bisquare basis function with two resolutions, one with 25 knots and the other one with 16 knots. After the data were generated, we randomly omit 10% of observations at each time point and use them as test data.

We applied the MCMC algorithm described in Sect. 4 with $\mathbf{H}_t = \mathbf{I}$ for estimation. We also fit the GLR model to the training data. Figure 3 presents the distribution of obtained RMSPE and DIC. It's evident that, while there is not much difference between the prediction performance of two models, the proposed model outperforms the GLR in terms of DIC, even though the data generation mechanism is not a member of the class of our models.

In what follows, we assess the goodness of fit and predictive performance of BNP-LR in the case that the data come from GLR. for this purpose, we generate 15 datasets at 1000 locations that are selected randomly on the square $[0, 50] \times [0, 50]$ for each

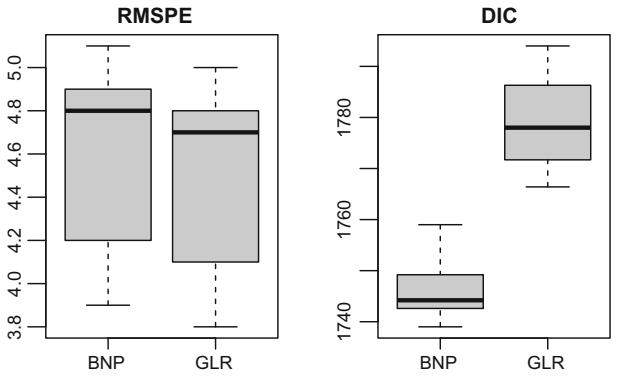


Fig. 4 Box plots of the RMSPE and DIC for two models fit to the simulated data from model (14), summarized for each of 15 holdout replicates

time point t, $t = 1, 2, \ldots, 10$ from GLR model. Specifically, datasets are generated from

$$\begin{aligned} Y(\mathbf{s}_{i,t}, t)|\beta, \eta_t, \sigma_\varepsilon^2 &\sim N(\mu(\mathbf{s}_{i,t}, t) + B'(\mathbf{s}_{i,t})\eta_t, \sigma_\varepsilon^2), \\ \eta_t &\sim N(\mathbf{H}_t\eta_{t-1}, \mathbf{W}_t), \end{aligned} \tag{14}$$

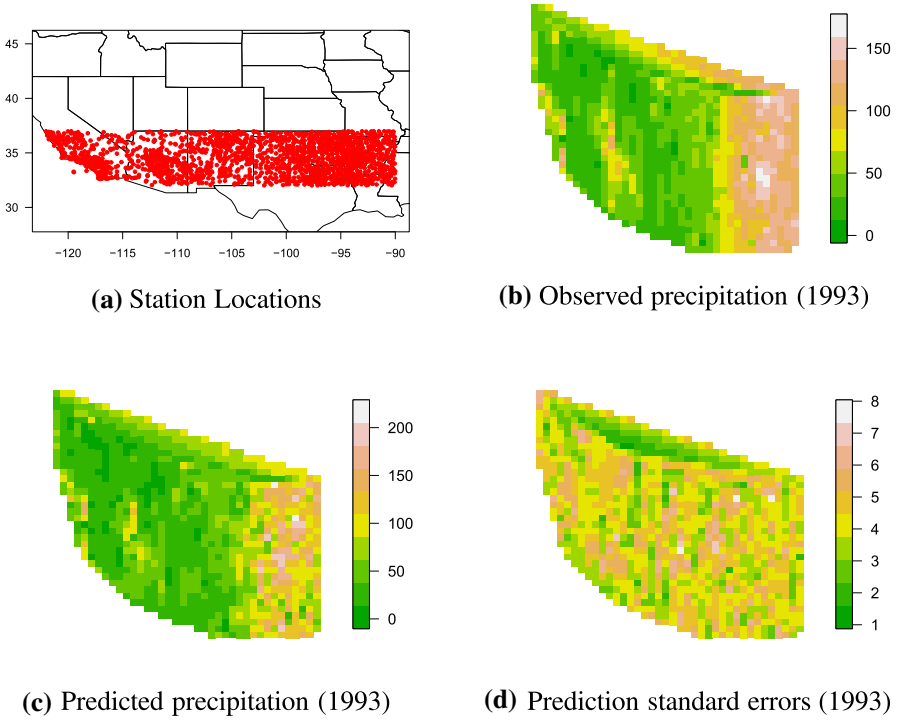
where the basis functions are bisquare with two resolutions, one with 25 knots and the other one with 16 knots. Moreover we set $\mathbf{H}_t = \mathbf{I}$, and a fixed mean $\mu(\mathbf{s}_{i,t}, t) = 2$. Similar to the above, 10% of generating observations at each time point are holdout for prediction model assessment. Then the GLR and BNP-LR models are fitted on simulated data. The distribution of the obtained RMSPE and DIC is shown in Fig. 4.

The results indicate that the mean square prediction error of applying BNP-LR on these data provides similar prediction results to GLR model. In addition, the BNP-LR model outperforms the GLR model in terms of DIC, even though when the true model of data is GLR.

6 Application to precipitation data

Precipitation is one of the important meteorological elements that influence various activities. Accurate knowledge of precipitation levels is a fundamental requirement for understanding and managing the climate changes. In this section, we aim to assess the effectiveness of the proposed BNP-LR model in prediction of annual precipitation. To this end, we apply the BNP-LR to a precipitation dataset, provided by the Institute for Mathematics Applied to Geosciences (http://www.image.ucar.edu/Data/US.monthly.met/).

The dataset is annual total precipitation in the region $D = [-122, -90] \times [32, 37]$, between the years 1982 and 1993 from 1976 stations. The data available to us had



(a) Station Locations

(b) Observed precipitation (1993)

(c) Predicted precipitation (1993)

(d) Prediction standard errors (1993)

Fig. 5 **a** Location of stations (red dots). **b** Annual total precipitation reported from observed stations. **c** Predicted annual total precipitation using BNP-LR model for year 1993, **d** estimated prediction standard errors (color figure online)

missing observations for each year; leaving about 1000–1250 observations from 1976 stations.

Figure 5a shows the location of stations across the study area that includes California, Arizona, New Mexico, Oklahoma, Arkansan, large part of Texas, and small part of Nevada, Louisiana, and Missouri states. This reveals the climate diversity in the study area, in such a way that Northern Arizona and New Mexico have a semi-desert climate, while California has a Mediterranean climate. On the other hand, northern Texas and Oklahoma have temperate climate while western Texas has semi-arid climate. The eastern Arizona has hot semi-arid climate and the western Arizona has hot desert climate. Then it seems unreasonable to assume stationarity. Additionally, the p values of the stationary test proposed by Bandyopadhyay and Rao (2017) at each time point are 0.030, 0.020, 0.007, 0.003, 0.001, 0.096, 0.002, 0.012, 0.015, and 0.132 which confirm spatial non-stationarity over time except for time instants 6 and 10. Additionally, Fig. 6 shows the histogram of measurements for sites $\mathbf{s}_1 = (-112.25, 33.45)$, $\mathbf{s}_2 = (-92.72, 35.98)$, $\mathbf{s}_3 = (-108.00, 36.83)$, $\mathbf{s}_4 = (-107.23, 34.12)$, $\mathbf{s}_5 = (-100.25, 35.23)$, $\mathbf{s}_6 = (-96.85, 32.40)$, $\mathbf{s}_7 = (-108.25, 36.70)$, $\mathbf{s}_8 = (-106.39, 35.92)$ and $\mathbf{s}_9 = (-109.15, 33.80)$. As observed, different distributional behaviors including multimodality are witnessed over spatial

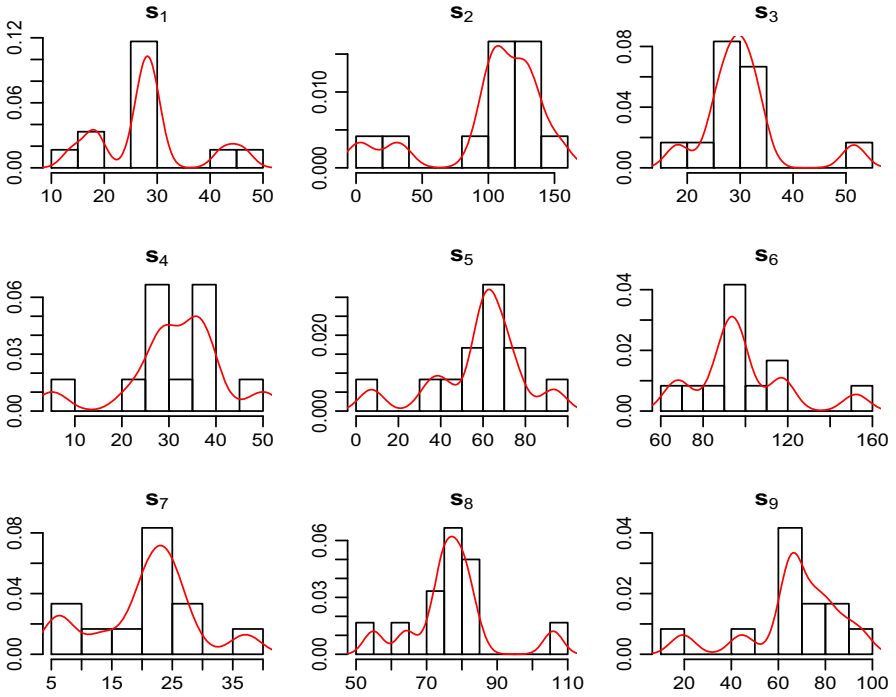


Fig. 6 Histograms of observations for several sites

domain which makes it challenging to choose an appropriate transformation. Therefore we apply our model to the observed data with following features.

Since explanatory analysis of the data showed a significant linear relation between annual total precipitation and elevation (in meters, divided by 100), then the mean function is assumed to be a first order linear function of elevation.

In the sequel, it needs to construct a set of appropriate basis functions. To this end, three resolutions of bisquare basis functions with 36, 16 and 9 knots are chosen to model the spatio-temporal random effect. As it was mentioned previously, the number of basis functions were selected based on DIC. According to our investigation (not shown here), there we no significant difference between DIC based on $r = 36+16+9$ knots and more number of knots. The respective radius of each resolution are $r_1 = 259{,}643$, $r_2 = 346{,}190$ and $r_3 = 649{,}104$ km.

The prior distributions are determined in the way described in the previous section. Additionally, for the evolution matrix, we consider a diognal form as $\mathbf{H} = diag(\rho_1, \ldots, \rho_r)$ where $\rho_1, .., \rho_r$ are unknown autoregressive parameters and r is the number of knots.

In what follows, we evaluate the predictive performance of our BNP-LR model in smoothing, filtering and forecasting with the GLR. To assess the performance of the proposed model in smoothing and filtering, we create 10 datasets by randomly taking 90% of the available data as the training data and the rest as prediction locations. For each dataset, we ran 20,000 MCMC iterations where the burn in time was 5000.

Table 1 Forecasting results for BNP-LR and GLR models

	BNP-LR	GLR
$RMSPE$	37.17	61.31
DIC	5430.18	7396.30

Again, the lag value was taken to be 5 in order to reduce the correlation between samples after burn in time. The convergence of the MCMC was verified through the autocorrelations and visual inspection of the trace plots. Evidence reveals no obvious convergence problem.

Also, the GLR model was fitted to the data and the averaged RMSPE over the 10 previously created datasets was calculated. The relative RMSPE is 1.73 for GLR model (53.0) compared with BNP-LR (30.64). Also, the relative DIC over 10 datasets is 1.43 for GLR (7494) compared with the BNP-LR (5248.8).

To evaluate the BNP-LR model in forecasting, we split the data into two parts: the first 11 years as training data and the last year as testing data.Similar to the above, the MCMC samples were obtained. Two models comparison criteria are shown in Table 1. Based on two criteria, the BNP-LR model outperforms the GLR model. More precisely, relative RMSPE and DIC are 1.65 and 1.36 for GLR model compared with BNP-LR model, respectively.

Figure 5c, d shows the posterior mean surface and standard error of the precipitation for the year 1993, based on the first 11 time data. According to the Fig. 5c the proposed model detected the behavior of underlying process with an acceptable accuracy.

7 Discussion

In this paper, we have proposed a computationally convenient BNP model for analyzing large spatio-temporal datasets. The core of the nonparametric component is the introduction of a spatio-temporal stick-breaking process to model the random effect where the spatio-temporal dependence is induced through the atoms. A low rank basis approximation and a vector autoregressive process are used to model spatial and temporal dependencies, respectively.

Our model is easy to implement and flexible in adapting to non-normal datasets. The BNP-LR as an extension of the GLR, inherits the advantageous properties of GLR. In other words, it preserves the mean and the covariance structures, as well as the cost of computation. In particular, the computational cost of BNP-LR increases linearly with the total number of observations. This feature of our model is its main advantage over other spatio-temporal BNP models. Our experiments with simulated data indicate that our BNP-LR offers improved performance over GLR in model fitting and prediction.

Based on the applied example, it seems that annual precipitation highly depend on elevation (Fig. 7). Therefore, accounting for elevation in the correlation structure in addition to the mean structure may result in improved inference and prediction. Our further investigation will focus on introducing a spatio-temporal BNP model to

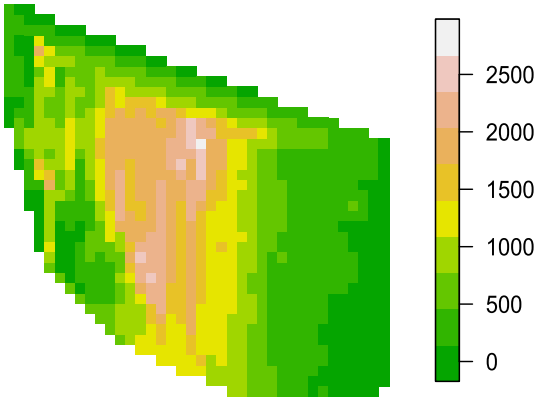


Fig. 7 Elevation map (in meters)

incorporate covariate information in the covariance structure in a computationally feasible way.

Further extensions of BNP-LR can also be envisioned. For instance, in addition to the atoms, we can supplement the spatio-temporal dependency with stick-breaking weights. Constructing such BNP models with varying weights is well-justified in some applications since it would grant substantial flexibility albeit with more burdensome computation. Another issue with this extension would be that it cannot be regarded as an extension of the flexible GLR.

Acknowledgements The Editor, and two referees are gratefully acknowledged. Their precise comments and constructive suggestions have substantially improved the manuscript.

References

Bandyopadhyay S, Rao SS (2017) A test for stationarity for irregularly spaced spatial data. J R Stat Soc Ser B (Stat Method) 79(1):95–123

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 70(4):825–848

Banerjee S, Finley AO, Waldmann P, Ericsson T (2010) Hierarchical spatial process models for multiple traits in large genetic trials. J Am Stat Assoc 105(490):506–521

Bradley JR, Cressie N, Shi T (2011) Selection of rank and basis functions in the spatial random effects model. In: Proceedings of the 2011 joint statistical meetings. American Statistical Association, Alexandria, pp 3393–3406

Bradley JR, Cressie N, Shi T (2015) Comparing and selecting spatial predictors using local criteria. Test 24(1):1–28

Bradley JR, Cressie N, Shi T (2016) A comparison of spatial predictors when datasets could be very large. Stat Surv 10:100–131

Canale A, Scarpa B (2016) Bayesian nonparametric location–scale–shape mixtures. Test 25(1):113–130

Carter CK, Kohn R (1994) On Gibbs sampling for state space models. Biometrika 81:541–553

Cavatti Vieira C, Loschi RH, Duarte D (2015) Nonparametric mixtures based on skew-normal distributions: an application to density estimation. Commun Stat Theory Methods 44(8):1552–1570

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 70(1):209–226

Cressie N, Shi T, Kang EL (2010) Fixed rank filtering for spatio-temporal data. J Comput Graph Stat 19(3):724–745

Di Lucca MA, Guglielmi A, Müller P, Quintana FA (2013) A simple class of Bayesian nonparametric autoregression models. Bayesian Anal (Online) 8(1):63
Duan JA, Guindani M, Gelfand AE (2007) Generalized spatial Dirichlet process models. Biometrika 94(4):809–825
Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. J Am Stat Assoc 90(430):577–588
Finley AO, Banerjee S, Gelfand AE (2012) Bayesian dynamic modeling for large space–time datasets using Gaussian predictive processes, vol 14. Springer, Berlin
Frühwirth-Schnatter S (1994) Data augmentation and dynamic linear models. J Time Ser Anal 15(2):183–202
Furrer R, Genton MG, Nychka D (2006) Covariance tapering for interpolation of large spatial datasets. J Comput Graph Stat 15(3):502–523
Gelfand AE, Kottas A, MacEachern SN (2005) Bayesian nonparametric spatial modeling with Dirichlet process mixing. J Am Stat Assoc 100(471):1021–1035
Gelfand AE, Diggle P, Guttorp P, Fuentes M (eds) (2010) Handbook of spatial statistics. CRC Press, Cambridge
Gelfand AE, Banerjee S, Finley A (2012) Spatial design for knot selection in knot-based dimension reduction models. In: Mateu JM, Mueller W (eds) Spatio-temporal design: Advances in efficient data acquisition. Wiley, pp 142–169
Ghosal S, Ghosh JK, Ramamoorthi RV (1999) Posterior consistency of Dirichlet mixtures in density estimation. Ann Stat 27(1):143–158
Griffin JE, Steel MF (2011) Stick-breaking autoregressive processes. J Econom 162(2):383–396
Gutiérrez L, Mena RH, Ruggiero M (2016) A time dependent bayesian nonparametric model for air quality analysis. Comput Stat Data Anal 95:161–175
Hanson T, Johnson WO (2002) Modeling regression error with a mixture of Polya trees. J Am Stat Assoc 97(460):1020–1033
Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109
Heaton MJ, Katzfuss M, Berrett C, Nychka DW (2014) Constructing valid spatial processes on the sphere using kernel convolutions. Environmetrics 25:2–15
Higdon D (1998) A process-convolution approach to modelling temperatures in the North Atlantic Ocean. Environ Ecol Stat 5(2):173–190
Hosseinpouri M, Khaledi MJ (2019) An area-specific stick breaking process for spatial data. Stat Pap 60(1):199–221
Kalli M, Griffin JE (2018) Bayesian nonparametric vector autoregressive models. J Econom 203(2):267–282
Kalli M, Griffin JE, Walker SG (2011) Slice sampling mixture models. Stat Comput 21(1):93–105
Kang EL, Cressie N, Shi T (2010) Using temporal variability to improve spatial mapping with application to satellite data. Can J Stat 38(2):271–289
Katzfuss M (2013) Bayesian nonstationary spatial modeling for very large datasets. Environmetrics 24(3):189–200
Katzfuss M, Cressie N (2011) Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics 23(1):94–107
Kaufman L, Rousseeuw P (1990) Finding groups in data, vol 16. Wiley, New York
Lemos RT, Sanso B (2009) A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature. J Am Stat Assoc 104(485):5–18
Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E (1953) Equation of state calculations by fast computing machines. J Chem Phys 21(6):1087–1092
Nieto-Barajas LE, Contreras-Cristán A (2014) A Bayesian nonparametric approach for time series clustering. Bayesian Anal 9(1):147–170
Nieto-Barajas L, Müller P, Ji Y, Lu Y, Mills G (2008) Time series dependent Dirichlet process. Preprint
Nguyen H, Cressie N, Braverman A (2012) Spatial statistical data fusion for remote sensing applications. J Am Stat Assoc 107(499):1004–1018
Pati D, Dunson DB, Tokdar ST (2013) Posterior consistency in conditional distribution estimation. J Multivar Anal 116:456–472
Petrone S, Guindani M, Gelfand AE (2009) Hybrid Dirichlet mixture models for functional data. J R Stat Soc Ser B (Stat Methodol) 71(4):755–782

Reich BJ, Fuentes M (2007) A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. Ann Appl Stat 1:249–264
Reich BJ, Fuentes M (2012) Nonparametric Bayesian models for a spatial covariance. Stat Methodol 9(1–2):265–274
Rue H, Held L (2005) Gaussian Markov random fields: theory and applications. CRC Press, Cambridge
Rue H, Tjelmeland H (2002) Fitting Gaussian Markov random fields to Gaussian fields. Scand J Stat 29(1):31–49
Sahr K, White D, Kimerling AJ (2003) Geodesic discrete global grid systems. Cartogr Geogr Inf Sci 30(2):121–134
Schörgendorfer A, Branscum AJ, Hanson TE (2013) A Bayesian goodness of fit test and semiparametric generalization of logistic regression with measurement data. Biometrics 69(2):508–519
Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit (with Discussion). J Roy Stat Soc B 64:583–639
Stein ML (2014) Limitations on low rank approximations for covariance matrices of spatial data. Spat Stat 8:1–19
Stein ML, Chi Z, Welty LJ (2004) Approximating likelihoods for large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 66(2):275–296
Vecchia AV (1988) Estimation and model identification for continuous spatial processes. J R Stat Soc Ser B (Methodol) 50(2):297–312
Walker SG (2007) Sampling the Dirichlet mixture model with slices. Commun Stat Simul Comput 36(1):45–54
Walker SG, Mallick BK (1999) Semiparametric accelerated life time model. Biometrics 55:477–483
Warren J, Fuentes M, Herring A, Langlois P (2012) Bayesian spatial–temporal model for cardiac congenital anomalies and ambient air pollution risk assessment. Environmetrics 23(8):673–684
West M, Harrison J (1997) Bayesian forecasting and dynamic models, 2nd edn. Springer, New York
Xu K, Wikle CK, Fox NI (2005) A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. J Am Stat Assoc 100(472):1133–1144

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.