J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

Bayesian Modeling of Discrete-Time Point-Referenced Spatio-Temporal Data

Suman Guha[1*] *and Sourabh Bhattacharya*[2]

Abstract | Discrete-time point-referenced spatio-temporal data are obtained by collecting observations at arbitrary but fixed spatial locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$ at regular intervals of time $t := 1, 2, \ldots, T$. They are encountered routinely in meteorological and environmental studies. Gaussian linear dynamic spatio-temporal models (LDSTMs) are the most widely used models for fitting and prediction with them. While Gaussian LDSTMs demonstrate good predictive performance at a wide range of scenarios, discrete-time point-referenced spatio-temporal data, often being the end product of complex interactions among environmental processes, are better modeled by nonlinear dynamic spatio-temporal models (NLDSTMs). Several such nonlinear models have been proposed in the context of precipitation, deposition, and sea-surface temperature modeling. Some of the above-mentioned models, although are fitted classically, dynamic spatio-temporal models with their complex dependence structure, are more naturally accommodated within the fully Bayesian framework. In this article, we review many such linear and nonlinear Bayesian models for discrete-time point-referenced spatio-temporal data. As we go along, we also review some nonparametric spatio-temporal models as well as some recently proposed Bayesian models for massive spatio-temporal data.

Keywords: *Bayesian spatio-temporal modeling, Gaussian process, Space–time covariance function, Massive spatio-temporal data, Nonlinear spatio-temporal model, Posterior predictive distribution*

1 Introduction

Modeling of spatio-temporal data has received much attention in recent years. Particularly, the rise in global temperature being a major environmental concern; scientists are now taking a keen interest in developing appropriate statistical models to study spatio-temporal data associated with climatic phenomena[28,46,62,64,68,71]. Other closely related events, that are also drawing much attention toward spatio-temporal modeling, are rainfall[15,66] and precipitation (mist, snowfall, sulfate, nitrate[60], etc.) across different regions. Apart from meteorology, challenging spatio-temporal data also arise from environmental and ecological science. To mention a few, studies on the ground level concentration of ozone[11,24,37,43], SO2[34,41], NO2[1], and PM-related air pollution[57], species distribution over a region[14], change in land-usage pattern over time[22], etc. Predominantly, these spatio-temporal data are observed at discrete time points and indexed continuously in space. Such spatio-temporal data which are obtained by collecting observations at arbitrary, but fixed spatial locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$ at regular intervals of time $t := 1, 2, \ldots, T$ are referred to as discrete-time point-referenced spatio-temporal data. Although in reality, data are available only at finitely many spatial locations (generally called monitoring sites), it is conceptually always useful to assume the existence of time-series at every spatial location. Pertained to that, then the data can be conceptually thought of as a partial

[1] Department of Statistics, Presidency University Kolkata, 86/1 College Street, Kolkata 700073, India.
[2] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata, 203, B. T. Road, Kolkata 700108, India.
*suman.stat@presiuniv.ac.in, bst0404@gmail.com

Springer

realization of $Y(\boldsymbol{s},t)$, which is a stochastic process indexed by $(\boldsymbol{s},t)$ and for each $t,\boldsymbol{s} \in D_{\boldsymbol{s}}(t)$, a continuous subset of $\mathbb{R}^d$. Most often $d = 2$ or 3, although sometimes $D_{\boldsymbol{s}}(t)$ can even be a continuous subset of some nonlinear manifold like $\mathbb{S}^2$ (the ordinary sphere). This situation is encountered when the measurements are taken at a global scale. The problem facing the statistician is to develop an appropriate model, infer about the spatio-temporal process $Y(\boldsymbol{s},t)$, and possibly predict at new sites and future points of time based on this partial realization. Note that any model applicable to them can also be extended to data collected at non-equispaced time points. Here, for the sake of simplicity, we confine ourselves mainly to the models for equispaced time points.

There exist two different perspectives from which one can develop a model for discrete-time point-referenced spatio-temporal data. One is the marginal approach, in which one develops a joint distribution for $Y(\boldsymbol{s},t)$. The other approach relies on specifying the conditional distribution of the current process given the past process realizations. Although theoretically, it is equivalent specifying the conditional distribution or the marginal distribution, one being derivable from the other, the conditional approach being closer to the etiology of the phenomena under study, is preferred over the marginal one. For more discussions on this issue, interested readers may look into the book by[17].[A]

2 Marginal Approach

In this case, one proposes a joint distribution for $Y(\boldsymbol{s},t)$. Under the assumption of Gaussianity, this amounts to specifying a spatio-temporal mean and covariance function. Therefore, the simplest prototype for such a marginal model is a Gaussian process on the plane with mean function $\mu(\boldsymbol{s},t)$ and covariance function $c_Y((\boldsymbol{s},t),(\boldsymbol{s}',t'))$. Often, one makes simplifying assumptions regarding the covariance function $c_Y((\boldsymbol{s},t),(\boldsymbol{s}',t'))$ like separability, stationarity, isotropy, etc., so that it is parametrized by fewer number of parameters. Similarly, the mean function is often parametrized by only a few parameters accommodating polynomial or periodic components in an additive manner. In that case, the model can be represented as

$$Y(\boldsymbol{s},t) := \mu(\boldsymbol{s},t) + \eta(\boldsymbol{s},t), \tag{1}$$

where the mean function is usually of the form $\mu(\boldsymbol{s},t) := \sum_{i=1}^{k} \beta_i^{(1)} \phi_i(\boldsymbol{s},t)$, $\phi_i(\boldsymbol{s},t)$ usually being a polynomial or trigonometric function in space–time. A broad review of spatio-temporal mean formulations given in[50], separated models into two classes, depending on whether the mean function is viewed as deterministic or stochastic. Dimitrakopoulos and Luo[23], on the other hand, classified the mean function into three alternative types: traditional polynomial functions, Fourier expressions, and combinations of the two.

In Eq. (1), $\eta(\boldsymbol{s},t)$ is a centered Gaussian process indexed by space–time with covariance function $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$. In choosing the form of $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$, one needs to look at the spatial dimension and the temporal dimension from different perspectives. A simple generalization of a spatial covariance function to a covariance function on $\mathbb{R}^3$ needs not be realistic. This is because distance in time is very different from distance in space, and moreover, time which always unfolds forward is intrinsically different from space that does not have any preferred direction. Reinsel et al.[59] proposed a regional-effects marginal model for the analysis of stratospheric ozone data for which

$$c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t')) : = \begin{cases} \sigma_\gamma^2 + \sigma_\kappa^2 + \sigma_\delta^2 \text{ if } \boldsymbol{s} = \boldsymbol{s}', t = t' \\ \sigma_\gamma^2 + \sigma_\kappa^2 \text{ if } \boldsymbol{s},\boldsymbol{s}' \in D_s^{(k)}, t = t', \\ k = 1,2,\ldots,K \\ \sigma_\gamma^2 \text{ if } \boldsymbol{s} \in D_s^{(k)}, \boldsymbol{s}' \in D_s^{(l)}, t = t', k \neq l, \\ k = 1,2,\cdots,K, l = 1,2,\ldots,K \\ 0 \text{ if } t \neq t'. \end{cases}$$

Here, $D_s^{(k)}$ denotes the k-th region, σ_γ^2 denotes the time-specific variability, σ_κ^2 denotes the region-specific variability, and σ_δ^2 denotes the variance of the nugget process. Later, Bloomfield et al.[10] extended it to a model accommodating spatio-temporal random effects. More realistic specifications for $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$ emerged subsequently. As a natural simplifying assumption separability is assumed, under which $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$ factors as $c_\eta^{(1)}(\boldsymbol{s},\boldsymbol{s}')c_\eta^{(2)}(t,t')$. If we further assume isotropy both in terms of space and time, it becomes $\sigma_\eta^2 \rho_\eta^{(1)}(\|\boldsymbol{s}-\boldsymbol{s}'\|)\rho_\eta^{(2)}(|t-t'|)$. Setting $\rho_\eta^{(2)}(|\tau|) := e^{-\lambda_t|\tau|}$ yields autoregressive dependence with respect to time and for eliciting spatial dependence, the corresponding spatial analogue $\rho_\eta^{(1)}(\|\boldsymbol{h}\|) := e^{-\lambda_s\|\boldsymbol{h}\|}$ is often used. More varied spatial dependence structures can be elicited through a generalization of the spatial

[A] Cressie and Wikle[17] covered a wide range of materials on spatio-temporal modeling that serves as the main resource for many of the models that we cite subsequently.

exponential covariance function to the Matérn covariance function.

Although the separable form of $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$ is convenient for computation and easier interpretation, it limits the nature of space–time interaction. Realistic nonseparable spatio-temporal covariance functions obtained through mixing broaden the scope. Another way to construct nonseparable spatio-temporal covariance functions is through the frequency domain approach. Cressie and Huang[16] introduced a flexible class of spatio-temporal covariance functions that allow for interaction in space–time. Another flexible class of spatio-temporal covariance functions was proposed by[35]. The covariance functions proposed by[35] are attractive, since they do not require closed-form Fourier inversion. One member of this class of covariance functions is

$$c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t')) := \frac{\sigma_\eta^2}{(q|t-t'|^{2\alpha}+1)^\beta} \exp\left\{-\left(\frac{p\|\boldsymbol{s}-\boldsymbol{s}'\|^{2\gamma}}{(q|t-t'|^{2\alpha}+1)^{\beta\gamma}}\right)\right\},$$

where setting the value of β to 0 yields a separable covariance function. Using the frequency domain approach, Stein[69] also provided a class of nonseparable covariance functions. However, unlike the covariance functions proposed by[35], here separability does not arise as a special or limiting case.

Also, attempts have been made to relax the assumption of isotropy and stationarity. In what is regarded as a landmark paper in spatial statistics, Sampson and Guttorp[63] introduced an approach to nonstationarity through deformation. Another approach to nonstationarity is obtained via kernel convolution. This approach is attributed to the two papers[26,40]. However, the kernel mixing form of[40] is fundamentally different from that of[26]. Further efforts were given to derive spatio-temporal covariance functions that are both nonstationary and nonseparable. The covariance function proposed by[27] is both nonstationary and nonseparable. Bruno et al.[11], on the other hand, proposed a deformation-based nonstationary and nonseparable covariance function to model tropospheric ozone data.

Spatio-temporal data are often observed with additional information in terms of covariates. The covariate data, which are also indexed by space–time, are often a partial realization of another spatio-temporal stochastic process. However, modeling the covariates is not the prime goal and the statistician is interested in the conditional distribution of $Y(\boldsymbol{s},t)$ given the value of covariates, i.e., $\boldsymbol{Z}(\boldsymbol{s},t)$ where $\boldsymbol{Z}(\boldsymbol{s},t) \in \mathbb{R}^l$. Minor modification of Eq. (1) to

$$Y(\boldsymbol{s},t) := \mu(\boldsymbol{s},t) + \beta_1^{(2)} Z_1(\boldsymbol{s},t) + \cdots + \beta_l^{(2)} Z_l(\boldsymbol{s},t) + \eta(\boldsymbol{s},t), \tag{2}$$

incorporates the covariates into the spatio-temporal model for $Y(\boldsymbol{s},t)$ (see[13]). Gelfand et al.[33] proposed a spatio-temporal hedonic model for house prices as $Y(\boldsymbol{s},t) := \boldsymbol{Z}(\boldsymbol{s},t)^T\boldsymbol{\beta}(\boldsymbol{s},t) + \alpha(t) + w(\boldsymbol{s}) + \epsilon(\boldsymbol{s},t)$ where $Y(\boldsymbol{s},t)$ is the log selling price, $\alpha(t)$ is the common time effect for all the locations, $w(\boldsymbol{s})$ is the spatial effect, $\epsilon(\boldsymbol{s},t)$ is a Gaussian white noise, and $\boldsymbol{Z}(\boldsymbol{s},t) \in \mathbb{R}^l$ contains useful covariate information. This form allows spatio-temporally varying coefficients, which is perhaps more than what is required. Therefore, $\boldsymbol{\beta}(\boldsymbol{s},t) := \boldsymbol{\beta}$ is frequently adopted. Setting $\boldsymbol{\beta}(\boldsymbol{s},t) := \boldsymbol{\beta}(t)$ yields an extension of the model proposed by[49].

2.1 Bayesian Inference, Bayesian Kriging and Forecasting

Traditional approaches to modeling of discrete-time point-referenced spatio-temporal data have their roots in geostatistics. While these methods use tools from classical statistics, with the advent of Markov Chain Monte Carlo (MCMC) and a plethora of other Bayesian computational algorithms, Bayesian models for spatio-temporal data are rapidly gaining popularity among practitioners. While the computation associated with the Bayesian approach requires tuning and human intervention, the subsequent step on inference is relatively straightforward. To carry out Bayesian inference, one needs to elicit prior distributions associated with the parameters. Often, the priors for individual parameters are elicited independently, and then, the joint prior distribution is the product of them. For example, let us consider the model specified by Eq. (2). Assume that $\boldsymbol{\beta} := (\beta_1^{(1)}, \ldots, \beta_k^{(1)}, \beta_1^{(2)}, \ldots, \beta_l^{(2)})$. Then, an $MVN(\boldsymbol{0}, c^2\boldsymbol{I})$ distribution, that is, a multivariate normal (MVN) distribution with mean vector zero and covariance matrix $c^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix, is taken as the prior for the vector-valued parameter $\boldsymbol{\beta}$. The value of c^2 is set to some large number like 1000 to make the prior non-informative. The advantage of using an MVN prior is that the full conditional distribution of $\boldsymbol{\beta}$ given the remaining parameters and the data is again an MVN distribution, simulation from which is straightforward. $\sigma_\eta^2 \sim IG(a_\eta, b_\eta)$ (IG stands for Inverse-gamma) is a prior that leads to IG full conditional distribution of σ_η^2 given the remaining parameters and the data. On the contrary, there do not exist such priors for the remaining parameters associated with $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$ that could lead to closed-form full conditional distributions. Separability is assumed and very often the form $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t')) := \sigma_\eta^2 e^{-\lambda_s\|\boldsymbol{s}-\boldsymbol{s}'\|}e^{-\lambda_t|t-t'|}$ is considered. One assumes $\lambda_s \sim \text{Gamma}(a_s,b_s)$ and $\lambda_t \sim \text{Gamma}(a_t,b_t)$. Samples from the posterior are then obtained by implementing a Metropolis Hastings within Gibbs type of algorithm where $\boldsymbol{\beta}$ and σ_η^2 are simulated by Gibbs steps and λ_s and λ_t are simulated by Metropolis Hastings (MH) steps. To facilitate the Metropolis Hastings within Gibbs algorithm, one transforms to $\theta_s := \log(\lambda_s)$ and $\theta_t := \log(\lambda_t)$, and subsequently uses Gaussian proposal distributions on the transformed parameters θ_s and θ_t. Once the posterior samples are obtained, the first few thousand samples are discarded as burn-in and posterior inference is carried out based on the post burn-in samples. Let us assume that the post burn-in posterior samples associated with the parameter λ_s are given as $\lambda_s^{(B+1)}, \lambda_s^{(B+2)}, \cdots$. Then, based on that, a histogram or kernel density estimator is calculated and plotted. The median obtained from the histogram/kernel density estimator gives a point estimate $\hat{\lambda}_s$ of λ_s and the $\frac{\alpha}{2}$th and $(1-\frac{\alpha}{2})$th quantiles give a $100(1-\alpha)\%$ equal tail credible interval. The equal tail credible interval is generally longer in length than the highest posterior density (HPD) interval but relatively easier to compute. If the posterior samples show high autocorrelation, then thinning is used, and generally, 1 out of every 5 or 10 samples are retained, based on which then all the posterior inferences are carried out. Inference for the remaining parameters can be carried out in a similar manner. Our experience tells that among all the parameters, tuning the proposal variance for the smoothness parameters is the most difficult. Improper tuning associated with the smoothness parameter causes the erratic movement of the Markov chain in the state-space, thereby leading to convergence failure.

Sometimes, a reparametrization of the original model with respect to new parameters may improve the convergence of the MCMC algorithm. Banerjee et al.[3] considered an alternative parametrization of $c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$ as $c_\eta((\boldsymbol{s},i),(\boldsymbol{s}',j)) := \sigma_\eta^2 e^{-\lambda_s\|\boldsymbol{s}-\boldsymbol{s}'\|}\left(\frac{\psi^{|i-j|}}{1-\psi^2}\right)$ for equispaced integer-valued time points $t := 1,2,\cdots,T$. Then, they used $\lambda_s \sim \text{Gamma}(a_s,b_s)$ and $\psi \sim U(0,1)$ as priors for the smoothness parameters. The choice of the uniform priors guaranteed that the model does not allow for negative correlation and no positive value for ψ is favoured over one another. To facilitate the Metropolis Hastings step, they considered the transformations $\theta_s := \log(\lambda_s)$ and $\theta_t := \log(\frac{\psi}{1-\psi})$, and updated the Markov Chain in the transformed space using Gaussian proposal distributions.

Often, the spatio-temporal covariance function is specified including the nugget effect. Nugget effect is a phenomenon present in many spatio-temporal datasets and represents short-scale randomness or noise. In that case, the covariance function takes the form $c_Y((\boldsymbol{s},t),(\boldsymbol{s}',t')) := c_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t')) + \sigma_\epsilon^2 I(\boldsymbol{s}=\boldsymbol{s}', t=t')$ and the model equation becomes

$$Y(\boldsymbol{s},t) := \mu(\boldsymbol{s},t) + \beta_1^{(2)}Z_1(\boldsymbol{s},t) + \cdots + \beta_l^{(2)}Z_l(\boldsymbol{s},t) + \eta(\boldsymbol{s},t) + \epsilon(\boldsymbol{s},t), \tag{3}$$

where $\epsilon(\boldsymbol{s},t)$ is iid in space–time and incorporates the nugget effect into the model. For the parameter σ_ϵ^2, one usually assumes an $IG(a_\epsilon,b_\epsilon)$ prior.

A critical part of any Bayesian modeling and more so in the Bayesian spatio-temporal modeling is the selection of the hyperparameters associated with the prior distributions. The exact values of the hyperparameters are to be set based on prior beliefs and care should be taken in choosing the values so as to ensure that a wide range of spatio-temporal structures can be accommodated within the proposed Bayesian model. The general pragmatic solution of selecting proper, but weakly informative priors for each scaling and smoothness parameter, often starting with conjugate inverse-gamma distributions for scale parameters such as σ_η^2 and σ_ϵ^2, is usually effective. However, very vague priors on such parameters may lead to essentially improper posterior distributions, thereby leading to MCMC convergence failure. Therefore, some care is still required to maintain posterior propriety without unduly limiting prior ranges of parameters (see the discussion in[75]). Waller[75] also discussed other notions of non-informative priors in this context. Quoting from him[6], Berger et al.[5] move toward objective Bayesian spatial analysis by considering reference and Jeffreys' priors for variance–covariance parameters in a Gaussian random field with no nugget effect. The authors illustrate that the Gaussian random field structure raises several interesting aspects not previously encountered in the reference prior literature. In particular, the model provides an example where Jeffreys' prior applied independently to each component yields an improper posterior. In addition, a popular approximation in the derivation of reference priors does not hold for the multivariate Gaussian data arising in a Gaussian random field. The results of[6], Berger et al.[5] provide tantalizing information regarding assignment of prior distributions for spatial covariance parameters, but also point to the need for additional development".

After we infer about the unknown parameters, the next step is to carry out spatio-temporal prediction. If the prediction is carried out at one of the data locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n$ for future time lags, then it is referred to as forecasting. On the other hand, if the prediction is done at a new spatial location $\boldsymbol{s}^*$ other than the data locations, then it is known as Kriging. Under the Bayesian framework, both Kriging and forecasting are done by simulating from the posterior predictive distributions. Often Kriging is done at multiple spatial locations. Now, we take a deeper look at the posterior predictive distribution under model (3). Let us assume that $\boldsymbol{Y}^*$ consists of $Y(\boldsymbol{s}, t)$ at the space–time coordinates associated with the prediction problem and $\boldsymbol{Y}$ consists of $Y(\boldsymbol{s}, t)$ at the space–time coordinates, from where the measurements have been taken. Assume that $\boldsymbol{y}$ represents the observed data. Also, assume that $E(\boldsymbol{Y}^*) := \boldsymbol{\mu}_1$, $E(\boldsymbol{Y}) := \boldsymbol{\mu}_2$, $\mathrm{Var}(\boldsymbol{Y}^*) := \boldsymbol{\Sigma}_{11}$, $\mathrm{Var}(\boldsymbol{Y}) := \boldsymbol{\Sigma}_{22}$ and $\mathrm{Cov}(\boldsymbol{Y}^*, \boldsymbol{Y}) := \boldsymbol{\Sigma}_{12}$. Then, the posterior predictive density takes the following form:

$$p\big(\boldsymbol{Y}^* \mid \boldsymbol{Y} = \boldsymbol{y}\big) = \int p\big(\boldsymbol{Y}^* \mid \boldsymbol{y}, \boldsymbol{\theta}\big) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}.$$

Drawing samples from the posterior predictive distribution is straightforward. If $\boldsymbol{\theta}^{(B+1)}, \boldsymbol{\theta}^{(B+2)}, \cdots$ denote the post burn-in samples associated with all the unknown parameters from the joint posterior distribution and then simulate serially from $p(\boldsymbol{Y}^* \mid \boldsymbol{y}, \boldsymbol{\theta}^{(B+1)}), p(\boldsymbol{Y}^* \mid \boldsymbol{y}, \boldsymbol{\theta}^{(B+2)}), \ldots,$ where $p(\boldsymbol{Y}^* \mid \boldsymbol{y}, \boldsymbol{\theta})$ is an MVN distribution with mean vector $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_2)$ and covariance matrix $\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. If the number of spatio-temporal coordinates where prediction is sought is large, then sometimes instead of simulating from the aforesaid MVN, one simulates from the MVN with mean vector $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_2)$ and covariance matrix $\mathrm{Diag}(\boldsymbol{\Sigma}_{11} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$ to reduce the computational burden. Here, $\mathrm{Diag}(\boldsymbol{A})$ refers to the diagonal matrix whose diagonal elements are the same as that of the matrix $\boldsymbol{A}$. Then, the resulting samples $\boldsymbol{y}^{*(B+1)}, \boldsymbol{y}^{*(B+2)}, \ldots$ constitute draws from the posterior predictive distribution $p(\boldsymbol{Y}^* \mid \boldsymbol{Y} = \boldsymbol{y})$. Once samples from $p(\boldsymbol{Y}^* \mid \boldsymbol{Y} = \boldsymbol{y})$ are available, we can perform point prediction, interval prediction, and density prediction in a straightforward manner.

While considering estimation and prediction under the Bayesian framework, we have not discussed missing data. In reality, however, missing data are quite common in spatio-temporal datasets. Mechanical disturbances and electronic malfunctions in measuring devices may lead to situations where a large proportion of data may be missing at some specific site. Missing data problems, however, can be handled straightforwardly in the aforesaid Bayesian spatio-temporal models, by the data augmentation technique. The underlying idea is to treat the missing data problem as a problem in prediction.

2.2 Non-Gaussian Response

So far, we have only considered the Gaussian process as the model for discrete-time point-referenced spatio-temporal data. Indeed, MVN distribution being mathematically more amenable than other multivariate distributions, and having many interesting theoretical properties, is the first choice to statisticians. However, sometimes, the very nature of the spatio-temporal dataset compels one to work with a non-Gaussian stochastic process. For example, pollutant measurements are positive numbers and are often transformed by a logarithmic function before modeling by a Gaussian process[38]. However, the transformation of the originally measured data accounts for estimated parameters that are less interpretable and so, in this case working directly with a non-Gaussian process is a better alternative. In the analysis of most spatio-temporal processes in environmental studies, observations present skewed distributions, with a heavy right or left tail. Pertained to that, Schmidt et al.[67] proposed a skew-Gaussian spatio-temporal process for fitting monthly average temperature data observed during 2001–2011 at different monitoring sites in the south of Brazil, under the Bayesian framework. Skew-t spatio-temporal process, which recently gained interest, can be used to model spatio-temporal data that are not only skewed but also contains outlying observations. More general non-Gaussian spatio-temporal processes can be constructed using a scale mixture of Gaussian processes and log-Gaussian processes as outlined in[25]. This model was later extended in[12] where the scale process is allowed to vary as a function of spatio-temporal covariates.

Another class of non-Gaussian spatio-temporal processes/models arises from the need for modeling extremes of spatio-temporally dependent data. At the heart of these spatio-temporal models lies the theory of the max-stable process. Max-stable spatio-temporal processes capture the local behavior of spatio-temporal extremes accurately. Davis et al.[21] and Huser and Davison[45] outlined classical fitting of these max-stable spatio-temporal process-based models in environmental applications. While these processes arise as asymptotically justified extensions of the generalized extreme value distribution for modeling univariate extremes when pointwise maxima are taken over spatio-temporal domains, a drawback of max-stable spatio-temporal process-based models is that they are computationally intensive to fit, limiting the number of space–time locations one can feasibly handle. Moreover, this class of processes can only capture asymptotic dependence, thereby limiting its scope. On the other hand, conditional models fitted under a fully Bayesian framework offer a very flexible tool due to the ease with which random effects are incorporated in them. The application of such a Bayesian spatio-temporal extreme value model was discussed in[74].

2.3 Categorical Spatio-temporal Data

When the spatio-temporal observation is categorical, none of the aforesaid models work. For example, $Y(\boldsymbol{s},t)$ can be a binary or count variable for observed spatio-temporal locations or species data providing presence/absence or abundance, respectively. In that case, the model specified by Eq. (1) can be generalized to

$$Y(\boldsymbol{s},t) \overset{ind}{\sim} \exp\left[\gamma\left\{y(\boldsymbol{s},t)(\mu(\boldsymbol{s},t)+\eta(\boldsymbol{s},t)) - \zeta(\mu(\boldsymbol{s},t)+\eta(\boldsymbol{s},t))\right\} + h(y(\boldsymbol{s},t),\gamma)\right],$$

to accommodate distributions that belongs to the exponential family. Here, $\zeta(\cdot)$ is a known, twice differentiable function, $h(\cdot,\cdot)$ is a known function and γ is a scalar dispersion parameter. Prior elicitation and posterior inference can be done similarly as outlined in Sect. 2.1.

2.4 Multivariate Spatio-temporal Modeling

Examples of multivariate point-referenced spatial data arise in environmental monitoring stations, where one might take measurements on multiple pollutants together, for example, ozone, NO, CO, PM2.5, etc. With such data dependence among measurements at a particular location as well as across locations is natural. Gelfand et al.[31] considered a univariate spatio-temporal data on monthly maximum temperature measurements where monthly (maximum) precipitation measurement is an important covariate. They argued that the multivariate spatio-temporal model that attempts to explain precipitation and temperature jointly gives better inference about all types of variability and association. Both of the above-mentioned problems require multivariate spatio-temporal modeling. Marginal specification of multivariate spatio-temporal model requires a multivariate stochastic process. Suppose we have p-variate observations at each spatio-temporal coordinate. Then, a simple multivariate spatio-temporal model for this data can be expressed by the following equation:

$$\boldsymbol{Y}(\boldsymbol{s},t) := \boldsymbol{\mu}(\boldsymbol{s},t) + \boldsymbol{\eta}(\boldsymbol{s},t),$$

where $\boldsymbol{\mu}(\boldsymbol{s},t)$ now denotes a p-variate function indexed by space–time and $\boldsymbol{\eta}(\boldsymbol{s},t)$ denotes a p-variate centered Gaussian process with cross-covariance function $\boldsymbol{c}_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$. As it is with the univariate Gaussian process, the construction of valid cross-covariance function $\boldsymbol{c}_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$, which is a $p\times p$ matrix-valued function, now poses a challenge. There is substantial theoretical literature regarding the creation of valid specifications for $\boldsymbol{c}_\eta((\boldsymbol{s},t),(\boldsymbol{s}',t'))$. A valid cross-covariance function can be formulated starting with p independent spatio-temporal Gaussian processes as follows. If the p independent centered Gaussian processes are denoted by $w_1(\boldsymbol{s},t), w_2(\boldsymbol{s},t),\cdots,w_p(\boldsymbol{s},t)$, the associated covariance functions are denoted by $c_1((\boldsymbol{s},t),(\boldsymbol{s}',t'))$, $c_2((\boldsymbol{s},t),(\boldsymbol{s}',t')),\cdots,c_p((\boldsymbol{s},t),(\boldsymbol{s}',t'))$, respectively, and $\boldsymbol{w}(\boldsymbol{s},t) := \left(w_1(\boldsymbol{s},t), w_2(\boldsymbol{s},t),\ldots,w_p(\boldsymbol{s},t)\right)^T$ denotes the associated p-variate Gaussian process on space–time, then $\boldsymbol{\eta}(\boldsymbol{s},t) := \boldsymbol{A}\boldsymbol{w}(\boldsymbol{s},t)$ yields a p-variate Gaussian process with cross-covariance function

$$\boldsymbol{c}_\eta\left((\boldsymbol{s},t),(\boldsymbol{s}',t')\right) = \sum_{i=1}^{p} c_i\left((\boldsymbol{s},t),(\boldsymbol{s}',t')\right)\boldsymbol{a}_i\boldsymbol{a}_i^T,$$

$\boldsymbol{a}_i$ denoting the ith column vector of the $p\times p$ nonsingular matrix $\boldsymbol{A}$. Another approach to the creation of valid spatio-temporal cross-covariance function is through kernel convolution. As outlined in[30], if $w(\boldsymbol{s},t)$ denotes a centered Gaussian process with covariance function $c_w((\boldsymbol{s},t),(\boldsymbol{s}',t'))$, $k_1\begin{pmatrix}\boldsymbol{h}\\ \tau\end{pmatrix}, k_2\begin{pmatrix}\boldsymbol{h}\\ \tau\end{pmatrix},\ldots,k_p\begin{pmatrix}\boldsymbol{h}\\ \tau\end{pmatrix}$ denote p square-integrable kernel functions on

$\mathbb{R}^2 \times \mathbb{Z}$ with $k_i \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} = 1$ and the p-variate spatio-temporal process $\boldsymbol{\eta}(\boldsymbol{s}, t)$ is defined component-wise by the relation $\eta_i(\boldsymbol{s}, t) := \int k_i \begin{pmatrix} \boldsymbol{s} - \boldsymbol{u} \\ t - v \end{pmatrix} w(\boldsymbol{u}, v) \mathrm{d}\boldsymbol{u} \mathrm{d}v$ for $i := 1, 2, \ldots, p$, then that yields a p-variate centered Gaussian process $\boldsymbol{\eta}(\boldsymbol{s}, t)$, with spatio-temporal cross-covariance function

$$c_\eta\big((\boldsymbol{s}, t), (\boldsymbol{s}', t')\big)_{i,j} = \int k_i \begin{pmatrix} \boldsymbol{s} - \boldsymbol{u} \\ t - v \end{pmatrix} k_j \begin{pmatrix} \boldsymbol{s}' - \boldsymbol{u}' \\ t' - v' \end{pmatrix} c_w((\boldsymbol{u}, v), (\boldsymbol{u}', v')) \mathrm{d}\boldsymbol{u} \mathrm{d}v \mathrm{d}\boldsymbol{u}' \mathrm{d}v'; \; i, j := 1, 2, \ldots, p.$$

Here, $\boldsymbol{c}_\eta((\boldsymbol{s}, t), (\boldsymbol{s}', t'))_{i,j}$ denotes the (i, j)th entry of the cross-covariance matrix $\boldsymbol{c}_\eta((\boldsymbol{s}, t), (\boldsymbol{s}', t'))$. Finally, the multivariate spatio-temporal model will be complete after the addition of the nugget effect as $\boldsymbol{Y}(\boldsymbol{s}, t) := \boldsymbol{\mu}(\boldsymbol{s}, t) + \boldsymbol{\eta}(\boldsymbol{s}, t) + \boldsymbol{\epsilon}(\boldsymbol{s}, t)$ where $\boldsymbol{\epsilon}(\boldsymbol{s}, t)$ now denotes a p-variate centered Gaussian process that is independent with respect to space and time and $Var(\boldsymbol{\epsilon}(\boldsymbol{s}, t))$ is a $p \times p$ diagonal matrix Σ_ϵ with positive diagonal entries.

3 Conditional Approach

Akin to the marginal approach, several models for the conditional distribution of $Y(\boldsymbol{s}, t)$ given the past process realizations have been proposed and most of them are based on the idea of linear Gaussian state-space models in time-series. These models are generally referred to as Gaussian linear dynamic spatio-temporal models (LDSTMs). Berliner[7] advocated a hierarchical representation for general conditional models of similar kinds as follows:

$$\begin{aligned} &\text{Data Model:}[Y(\boldsymbol{s}, t) \mid X(\boldsymbol{s}, t), \theta_D];\\ &\text{Process Model :}[X(\boldsymbol{s}, t) \mid \theta_P]. \end{aligned}$$

When these conditional models are fitted under the Bayesian framework, a third hierarchy is introduced

$$\text{Parameter Model:}[\theta_D, \theta_P].$$

Following this hierarchical representation, a Gaussian LDSTM can be expressed as:

$$\begin{aligned} \text{Data Model :}\boldsymbol{Y}_t &:= \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{\epsilon}_t;\\ \text{Process Model :}\boldsymbol{X}_t &:= \boldsymbol{G}_t \boldsymbol{X}_{t-1} + \boldsymbol{\eta}_t; \qquad (4)\\ \boldsymbol{X}_0 &\sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \end{aligned}$$

where $\boldsymbol{Y}_t$ is an n-dimensional random vector representing $Y(\boldsymbol{s}, t)$ observed on time t at spatial locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n$. Also, $\boldsymbol{\epsilon}_t \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon)$ is a random noise process associated with the observation equation and is iid with respect to time. The random vector $\boldsymbol{X}_t$ denotes a multivariate state process indexed by time t, $\boldsymbol{\eta}_t \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma}_\eta)$ is a random noise process associated with the evolutionary equation and is iid with respect to time and $\boldsymbol{X}_0$ denotes the initial state process at time 0. Typically, one assumes that $\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t$ and the initial state process $\boldsymbol{X}_0$ are jointly independent. The matrices $\boldsymbol{F}_t$ of order $n \times s$ and $\boldsymbol{G}_t$ of order $s \times s$ are unknown and time-varying. One also assumes that $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ are unknown parameters that are to be estimated from the data.

Gaussian LDSTM as specified above is too general and suffers from the non-identifiability problem. It can, however, be made identifiable by imposing suitable restrictions on the parameters. If one assumes that the propagator matrix is time-invariant, i.e., $\boldsymbol{G}_t := \boldsymbol{G}$, the matrix $\boldsymbol{F}_t$ is known and of full column rank, then the parameters $\boldsymbol{\epsilon}_t, \boldsymbol{\eta}_t$ and $\boldsymbol{G}$ are rendered identifiable. For more discussion on non-identifiability issues associated with Gaussian LDSTMs, see[73]. Often, the matrix $\boldsymbol{F}_t$ consists of known spatial basis functions giving rise to what is referred to as the spatio-temporal random effects model (STRE). The random effects are incorporated in the model through the state process $\boldsymbol{X}_t$.

Gaussian LDSTMs, with their complex dependence structure, can be more naturally accommodated within the fully Bayesian framework (see[43,56,61,65,70,72]). In a fully Bayesian approach, priors are assigned to the parameters governing the state and the observation equations, thereby producing a rich dependence structure across space and time. However, the drawback is high computational overhead, which may prove to be insurmountable if the number of spatial locations is large. To address this problem, dimension-reduced state-space models have been proposed. Dimension reduction problem can be overcome through parametrization of the propagator matrix. When the dimension of the state process is s, the propagator matrix, if fully unspecified, contains a whopping s^2 unknown parameters causing calculation of the posterior distribution difficult. Parametrizations based on a propagator matrix that is identity (which corresponds to a random walk) or diagonal matrix significantly reduce the computational load. Sparse, yet less-restrictive parametrizations are obtained by considering a propagator matrix based on discretization of partial differential equations or integro-difference equations[80] or lagged nearest-neighbors[77]. The noise covariance matrix $\boldsymbol{\Sigma}_\eta$ is also parametrized using exponential or Matérn covariance function on space. Alternatively, dimension reduction can be relative to the state/latent process itself, for example, through spectral representation/basis-function expansion. In that case, we can decompose the s-dimensional state process as

$$X_t := \Phi\alpha_t + \nu_t,$$

where α_t is a p_α dimensional state process with $p_\alpha << s$. The matrix Φ consists of basis functions, for which there are many different choices. One may consider them to be orthogonal functions like Fourier, Hermite polynomials, eigenvectors from a specified or estimated covariance function, etc., or non-orthogonal functions like wavelets, splines, bisquare, discrete kernel convolutions, etc. With this dimension reduced specification, now the process model is given by

$$\alpha_t := M\alpha_t + \delta_t.$$

STRE models, which form an important subclass of LDSTMs, can be further generalized to spatio-temporal mixed effects (STME) model as follows:

$$Y_t := F_t X_t + Z_t \beta_t + \epsilon_t, \tag{5}$$

where β_t is a time-varying unknown parameter and Z_t is an $n \times l$ dimensional covariate matrix associated with time t, n being the number of observed data locations and l denoting the number of covariates.

On a similar note, a different subclass of LDSTMs arrives through the following equations:

$$\begin{aligned} \text{Data Model}: Y_t &:= Z_t \beta_t + \epsilon_t; \\ \text{Process Model}: \beta_t &:= G_t \beta_{t-1} + \eta_t; \\ \beta_0 &\sim N(\mu_0, \Sigma_0). \end{aligned} \tag{6}$$

The design matrix Z_t associated with time t consists of deterministic functions of space and important covariate information. The observed data Y_t arise from a linear model that is changing with time. This model in a multivariate time-series context is referred to as a dynamic linear model (DLM) (see[31,58]). The scope of the model is broadened further as[29] allowed the columns of Z_t to vary as random functions of space. The aforesaid models are dynamically specified, but the dynamics remain invariant across space. A significant extension can be achieved by allowing the coefficient vector β_t to vary across space. This obviously implies a substantial increase in the number of parameters and may even lead to the non-identifiability problem. Putting restrictions on the parameter space keeps the problem manageable. Paez et al.[58] proposed such a spatially varying coefficient dynamic model where they took $\beta_t(s) := \bar{\gamma}_t + \gamma_t(s)$. The common trend $\bar{\gamma}_t$ evolves dynamically with time, but the spatio-temporal part $\gamma_t(s)$ are iid Gaussian processes in space. Gelfand et al.[31] considered a more general model in the environmental context where both $\bar{\gamma}_t$ and $\gamma_t(s)$ vary dynamically with respect to time. They applied their model in the context of spatio-temporal regression. The spatially varying coefficients associated with these LDSTMs can be represented as

$$\begin{aligned} \beta_t(s) &:= \bar{\gamma}_t + \gamma_t(s); \\ \bar{\gamma}_t &:= \bar{G}_t \bar{\gamma}_{t-1} + \bar{\omega}_t; \\ \gamma_t &:= G_t \gamma_{t-1} + \omega_t. \end{aligned} \tag{7}$$

A number of authors proposed an apparently different form for the evolution of $\beta_t(s)$. They assumed that

$$\beta_t(s) := \int K(u, s)\beta_{t-1}(u)du + \eta_t(s),$$

where $K(u, s)$ is a redistribution kernel that determines how the coefficients at the previous time point influence the coefficients at present. Typically, spatially nearer coefficients at the previous time point get more weight in determining the coefficients at present. Although it appears to be very different from the evolutionary equations considered so far, a discrete convolution-based approximation to the integral would yield that usual expression.

3.1 Hierarchical Specification and Prior Structure

Now, we consider the LDSTM specified by Eq. (4) with time-invariant propagator matrix G to illustrate how prior elicitation and Bayesian inference can be carried out with it. Following the suggestion of[17], one assumes independent priors for the unknown parameters $\Sigma_\epsilon, \Sigma_\eta, G, \mu_0$ and Σ_0. Cressie and Wikle[17] assumed $\Sigma_\epsilon \sim IW(\nu_\epsilon C_\epsilon, \nu_\epsilon)$, $\Sigma_\eta \sim IW(\nu_\eta C_\eta, \nu_\eta)$, and $m \equiv vec(G) \sim MVN(\mu_m, \Sigma_m)$ where IW stands for inverse Wishart distribution and $vec(G)$ represents the vectorized representation of the matrix G. Regarding the prior hyperparameters $\nu_\epsilon, C_\epsilon, \nu_\eta, C_\eta, \mu_m, \Sigma_m$, we assume that they are fixed and known. The remaining two parameters μ_0 and Σ_0 are also assumed to be fixed and known. With this prior specification as mentioned in[17], then one can carry out posterior inference in a straightforward manner using a Gibbs sampler. The Gibbs sampler is implemented by simulating from the full conditional distributions in the following order. Simulate from $[X_0 \mid \cdot], [X_t \mid \cdot]$ for $t := 1, 2, \cdots, T-1, [X_T \mid \cdot]$ and then simulate from $[\mathbf{\Sigma}_\epsilon \mid \cdot], \left[\mathbf{\Sigma}_\eta \mid \cdot\right], [\boldsymbol{m} \mid \cdot]$. The full conditional distributions associated with state process being MVNs, the full conditional distributions associated with $\mathbf{\Sigma}_\epsilon$ and $\mathbf{\Sigma}_\eta$ being IWs, and the full conditional distribution associated with $\boldsymbol{m}$ being MVN, implementation of the Gibbs sampler is straightforward. Simulation from the posterior predictive distribution is also straightforward as it amounts to simulation from an MVN, and therefore, Kriging and forecasting can be performed easily.

Alternatively, one can consider the noise covariance matrices $\mathbf{\Sigma}_\epsilon$ and $\mathbf{\Sigma}_\eta$ to be induced by parametric spatial covariance functions and hence can elicit priors for scale and smoothness parameters associated with them. In that case, the full conditional distributions associated with $\mathbf{\Sigma}_\epsilon$ and $\mathbf{\Sigma}_\eta$ would not have closed-form expressions. Consequently, the Gibbs sampler cannot be directly implemented and a hybrid Metropolis Hastings within Gibbs algorithm is used. Alike the marginal models, prior hyperparameter selection requires special attention also in the conditional approach. In the light of that, ignoring parametric covariance specification and instead using a prior distribution on the entire cone of positive definite matrices, perhaps using a conjugate IW distribution may seem attractive as this leads to a closed-form full conditional distribution. However, note that the inverse Wishart posterior no longer limits attention to specific spatio-temporal covariance structures of interest, which may be considered somewhat undesirable. Moreover, Daniels and Kass[18] outlined questionable performance of the IW prior in some hierarchical modeling settings. As a result, there has been renewed interest in non-conjugate Bayesian analysis of covariance matrices specified through parametric spatial covariance functions.

Finally, the LDSTMs specified previously can be extended easily for providing a Bayesian model for multivariate spatio-temporal data. Paez et al.[58] modified the DLM stated via Eq. (6) to $\boldsymbol{Y}_t := \boldsymbol{Z}_t \beta_t + \boldsymbol{\epsilon}_t, \boldsymbol{\beta}_t := \boldsymbol{G}_t \beta_{t-1} + \boldsymbol{\eta}_t$ and $\boldsymbol{\beta}_0 \sim N(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$ where the only difference with Eq. (6) is that now $\boldsymbol{Y}_t, \beta_t, \beta_0, \boldsymbol{\epsilon}_t$ and $\boldsymbol{\eta}_t$ are matrices and the two noise process distributions and the initial distribution all are matrix normal. The prior choice is similar to the univariate model, and so, Bayesian inference and prediction can be carried out in a similar manner.

3.2 Nonlinear Dynamic Spatio-temporal Models

So far, we have considered conditional spatio-temporal models whose temporal evolution can be described by linear equations. However, real-life environmental processes are complex and require much more sophistication in model specification. In particular, processes like precipitation, deposition, etc. are driven by complex interactions among atmospheric processes and are best represented by nonlinear models. LDSTMs, as considered in Sect. 3 would be unsuitable in such situations. Apart from atmospheric processes, sometimes, sea-surface temperature data is modeled by nonlinear dynamic spatio-temporal models (NLDSTMs) owing to the complex dynamics of sea waves. Also, many processes in the context of growth curve modeling exhibit state-dependent or density-dependent growth, e.g., $\partial Y/\partial t = Yg(Y; \boldsymbol{\theta})$ for some nonlinear growth function $g(\cdot)$ (e.g., logistic, Ricker, Beverton-Holt, etc.). In addition, many processes exhibit what is sometimes referred to as nonlinear advection, e.g., in one spatial dimension, $\partial Y/\partial t = Y \partial Y/\partial s_1$ (see[17]). More general nonlinear dynamic spatio-temporal models are required to accommodate such processes, among others. Sanso and Guenni[66] considered such a truncation-based NLDSTM for modeling of a Venezuelan rainfall dataset, given as follows:

$$Y_t(\boldsymbol{s}) := \begin{cases} X_t(\boldsymbol{s})^{b_t} & \text{if } X_t(\boldsymbol{s}) > 0; \\ 0 & \text{if } X_t(\boldsymbol{s}) \le 0, \end{cases}$$

where $X_t(\boldsymbol{s})$ is the state process and b_t is a time-varying parameter associated with the truncation equation. A more general formulation for this model can be written as (see equation 7.39 on page 380 of[17])

$$Y_t(\boldsymbol{s}) := a_t(\boldsymbol{s}) + h_t(\boldsymbol{s}) X_t(\boldsymbol{s})^{b_t(\boldsymbol{s})} + \epsilon_t(\boldsymbol{s}), \quad (8)$$

where $\epsilon_t(\boldsymbol{s})$ is a spatio-temporal noise process. The above model assumes that the observational equation that connects the observed process $Y_t(\boldsymbol{s})$ with the state process $X_t(\boldsymbol{s})$ is nonlinear. Instead of that, nonlinearity can be introduced into the evolutionary equation. In that case, we would have

$$X_t(\boldsymbol{s}) := \Psi(X_{t-1}(\boldsymbol{s})) + \eta_t(\boldsymbol{s}), \tag{9}$$

where $\Psi(\cdot)$ is some appropriate nonlinear function. Cressie and Wikle[17] discussed nonlinear state-dependent models $X_t(\boldsymbol{s}) := \Psi_t(X_{t-1}(\boldsymbol{s}))X_{t-1}(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s})$ which are time-varying versions of it. Often the nonlinear function $\Psi_t(X_{t-1}(\boldsymbol{s}))X_{t-1}(\boldsymbol{s})$ is taken to be a threshold function as follows:

$$X_t(\boldsymbol{s}) := \Psi_t(X_{t-1}(\boldsymbol{s}))X_{t-1}(\boldsymbol{s}) + \eta_t(\boldsymbol{s}) : \\ = \begin{cases} G_1 X_{t-1}(\boldsymbol{s}) + \eta_{1,t}(\boldsymbol{s}) & \text{if } f_1(\gamma_t) \in c_1; \\ \vdots & \\ G_K X_{t-1}(\boldsymbol{s}) + \eta_{K,t}(\boldsymbol{s}) & \text{if } f_K(\gamma_t) \in c_K, \end{cases}$$

where $f_k(\gamma_t)$ is a function of a time-varying parameter γ_t, and $c_k; k := 1, 2, \ldots, K$, is the condition under which the kth equation is to be followed (see equation 7.69 on page 406 of[17]). An example of such a model was given by[9] with regard to long-lead forecasting of tropical Pacific sea-surface temperature. Hughes and Guttorp[44] used such a model in an atmospheric application and[42] employed them for the analysis of an ecological dataset. Another interesting class of NLDSTM, that[79] (also see[17]) referred to as the General Quadratic Nonlinear (GQN) Model, is given by

$$X_t(\boldsymbol{s}_i) := \sum_{j=1}^{n} a_{ij} X_{t-1}(\boldsymbol{s}_j) \\ + \sum_{k=1}^{n}\sum_{l=1}^{n} b_{i,kl} X_{t-1}(\boldsymbol{s}_k) g(X_{t-1}(\boldsymbol{s}_l); \boldsymbol{\theta}^G) \\ + \eta_t(\boldsymbol{s}_i); \text{ for } i := 1, \ldots, n,$$

where $\sum_{j=1}^{n} a_{ij} X_{t-1}(\boldsymbol{s}_j)$ is a linear combination of the process at the previous time and $\sum_{k=1}^{n}\sum_{l=1}^{n} b_{i,kl} X_{t-1}(\boldsymbol{s}_k) g(X_{t-1}(\boldsymbol{s}_l); \boldsymbol{\theta}^G)$ contains quadratic interactions of the process and potentially some transformation of the lagged process, at the previous time. The model is flexible enough to accommodate nonlinear transformations of the process through the function $g(\cdot)$, which might depend upon the unknown parameter vector $\boldsymbol{\theta}^G$. GQN constitutes a very rich class of models and many complex process models including the one considered by[8] for the so-called quasi-geostrophic flow in the ocean are special cases of it. Wikle and Holan[78] considered an extension of GQN by incorporating higher order polynomial interaction terms in the following way:

$$X_t(\boldsymbol{s}_i) := \sum_{j_1=1}^{n} a^{(1)}_{i,j_1} X_{t-1}(\boldsymbol{s}_{j_1}) \\ + \sum_{j_2=1}^{n}\sum_{j_1=1}^{n} a^{(2)}_{i,j_1 j_2} X_{t-1}(\boldsymbol{s}_{j_2}) g(X_{t-1}(\boldsymbol{s}_{j_1}); \boldsymbol{\theta}_{g_1}) \\ + \sum_{j_3=1}^{n}\sum_{j_2=1}^{n}\sum_{j_1=1}^{n} a^{(3)}_{i,j_1 j_2 j_3} X_{t-1}(\boldsymbol{s}_{j_3}) X_{t-1}(\boldsymbol{s}_{j_2}) \\ g(X_{t-1}(\boldsymbol{s}_{j_1}); \boldsymbol{\theta}_{g_2}) \\ \vdots \\ + \sum_{j_p=1}^{n}\cdots\sum_{j_2=1}^{n}\sum_{j_1=1}^{n} a^{(p)}_{i,j_1 j_2 \ldots j_p} X_{t-1}(\boldsymbol{s}_{j_p}) \ldots \\ X_{t-1}(\boldsymbol{s}_{j_2}) g(X_{t-1}(\boldsymbol{s}_{j_1}); \boldsymbol{\theta}_{g_p}) \\ + \eta_t(\boldsymbol{s}_i).$$

It is called the General Polynomial Nonlinear (GPN) model. For an excellent overview of such NLDSTMs, the reader may look into chapter 7 of[17]. NLDSTMs attempt to provide a way out when the usual LDSTMs turn out to be too naive for the phenomena under study, but that too comes with a cost. The issues of dimensionality and efficient parametrization present the most significant challenges for statistical modeling of LDSTMs and these issues get even more critical for NLDSTMs. However, what is more daunting is that without very precise knowledge of the underlying dynamics, it is almost impossible to elicit an appropriate nonlinear model from a large class of probable nonlinear functions. A selected nonlinear model, that is unsuitable for the physical process under study, would show grossly poor predictive performance. Seemingly irrelevant departure from reality at the level of model specification cumulates over time and the outcome may be devastating.

As happens with the marginal approach, members of the exponential family of distributions can also be accommodated within the conditional approach to use them as an appropriate model for spatio-temporal categorical data. For example, a common choice for spatio-temporal species abundance data is a dynamic spatio-temporal Poisson model

$$Y_t(\boldsymbol{s}) \overset{ind}{\sim} Poi(X_{t-1}(\boldsymbol{s})), \tag{10}$$

where $X_t(\boldsymbol{s})$ may vary dynamically with respect to time.

4 Non-parametric Bayes in Spatio-temporal Data

As already mentioned, a focal issue for the NLDSTMs is the selection of the form of nonlinearity, and such a task is highly non-trivial under little knowledge about the underlying process. In that case, a non-parametric or semiparametric spatio-temporal model would be a better alternative. Much to our surprise, there is little work on non-parametric spatio-temporal models. Marginal models with non-parametrically specified mean function have been applied in the investigation of dynamics of spatial pollution surface. In that case, one models the spatio-temporal measurement or its transformation using a spatio-temporal generalized additive model (STGAM). A typical STGAM looks like the following:

$$Y(\boldsymbol{s},t) := \mu + m_{\boldsymbol{s}}(s_1,s_2) + m_t(t) + \epsilon(\boldsymbol{s},t).$$

The STGAM models the data without assuming linear or specific nonlinear form for $m_{\boldsymbol{s}}(.)$ and $m_t(.)$, thereby modeling the spatial and temporal trend non-parametrically. However, a conditional model being closer to the etiology of the process is a more eluding one. In this regard, Lu et al.[52] proposed an adaptively varying coefficient spatio-temporal model that can be expressed as

$$Y_t(\boldsymbol{s}) := a(\boldsymbol{s},\alpha(\boldsymbol{s})^T\boldsymbol{X}^T(\boldsymbol{s})) + \boldsymbol{b}_1(\boldsymbol{s},\alpha(\boldsymbol{s})^T\boldsymbol{X}^T(\boldsymbol{s}))\boldsymbol{X}(\boldsymbol{s}) + \epsilon_t(\boldsymbol{s}),$$

where $Y_t(\boldsymbol{s})$ is the spatio-temporal variable of interest, $a(\boldsymbol{s},z)$ and $\boldsymbol{b}_1(\boldsymbol{s},z)$ are unknown scalar and d-dimensional vector functions, respectively, $\alpha(\boldsymbol{s})$ is an unknown d-dimensional coefficient vector, $\epsilon_t(\boldsymbol{s})$ is a noise process which, for each fixed $\boldsymbol{s}$, forms a sequence of independent and identically distributed random variables over time and $\boldsymbol{X}_t(\boldsymbol{s}) := (X_{t1}(\boldsymbol{s}), X_{t2}(\boldsymbol{s}), \cdots, X_{td}(\boldsymbol{s}))^T$ consists of time-lagged values of $Y_t(\boldsymbol{s})$ in a neighbourhood of $\boldsymbol{s}$ and, possibly, some covariate values. The model was fitted using a classical two-stage procedure. The main difference between this model and the STGAM is that the former corresponds to the conditional approach and assumes flexible non-parametric form for the conditional structure of the space–time model, whereas the latter is associated with the marginal approach and stresses more on flexible modeling of the trend in space and time and the interaction terms.

Guha and Bhattacharya[36], on the other hand, proposed a Gaussian process-based non-parametric Bayesian dynamic spatio-temporal model, that they referred to as the Gaussian random functional dynamic spatio-temporal model (GRFDSTM) where both the observational and evolutionary equations are random functions. Their model has the following form:

$$\begin{aligned} &Y_t(\boldsymbol{s}) := f(X_t(\boldsymbol{s})) + \epsilon_t(\boldsymbol{s}),\\ &X_t(\boldsymbol{s}) := g(X_{t-1}(\boldsymbol{s})) + \eta_t(\boldsymbol{s})\\ &X_0(\cdot) \sim \mathrm{GP}(\mu_0(\cdot), c_0(\cdot,\cdot));\ f(\cdot), g(\cdot) \sim \mathrm{GRF}(\cdot,\cdot), \end{aligned} \tag{11}$$

where $\boldsymbol{s} \in \mathbb{R}^2$ and $t \in \{1,2,3,\ldots\}$. In the above description, "GP" stands for "Gaussian process" and "GRF" stands for "Gaussian random function". Here, $X_0(\cdot)$ is a spatial Gaussian process on $\mathbb{R}^2$; $\epsilon_t(\cdot)$ and $\eta_t(\cdot)$ are temporally independent and identically distributed spatial Gaussian processes on $\mathbb{R}^2$, and $f(\cdot)$ and $g(\cdot)$ are Gaussian random functions on $\mathbb{R}$. They are all independent of each other. Hence, one no longer has to decide about the specific functional forms; instead, all one needs to do is to ensure that the probabilistic laws for the random functions are so chosen that they give enough probability to sets of functions that seem potentially appropriate for the data at hand. Moreover, unlike the LDSTMs or NLDSTMs where the functional form is fixed, a random functional form is more adaptable to the data and expected to represent the true underlying process, which may be complex and highly nonlinear, more realistically. The reason behind choosing Gaussian processes as probabilistic laws for the random functions is that Gaussian processes are good natural priors for non-parametric regression and classification problems and under increasingly dense observations, the true shape of the arbitrary function or the classifier can be captured accurately, a posteriori. The GRFDSTM as specified above involves Gaussian processes and therefore requires Cholesky decomposition of large covariance matrices, making it computationally demanding.

Alternatively, based on kernel convolution of order-based dependent Dirichlet process (ODDP), Das and Bhattacharya[19] constructed a nonstationary, non-parametric space–time process, which is fitted under the Bayesian framework using a transdimensional transformation-based Markov Chain Monte Carlo method. Their approach is attractive, being very fast for moderate size spatio-temporal data.

5 Massive Spatio-temporal Data

With the growing capabilities of Geographic Information Systems (GIS) and user-friendly software, statisticians today routinely encounter geographically referenced data containing observations from a large number of spatial locations and time points. Conditional spatio-temporal models discussed throughout this article often involve inversion and Cholesky decomposition of large covariance matrices, thereby rendering them infeasible in such settings. The traditional path to modeling of such massive spatio-temporal datasets is to use low-rank spatio-temporal models, which assume that the observed spatio-temporal process is driven by a much lower dimensional spatio-temporal state process[32]. The generic equation associated with such a low-rank process is of the form $\boldsymbol{X}_t := \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\nu}_t$ where $\boldsymbol{X}_t$ is the original state process that drives the spatio-temporal observed process which yields the observed data and $\boldsymbol{\alpha}_t$ is a p_α dimensional low-rank state process with $p_\alpha << s$. That way, the computational cost can be reduced from the order of $\sim O(n^3T^3)$ flops to $\sim O(nT)$ flops. Here, flops stand for floating-point operations and are a unit for measuring the amount of computation performed by an algorithm.

Katzfuss and Cressie[48], in an attempt to model a massive spatio-temporal CO_2 measurement data over the globe, used a low-rank spatio-temporal model with spatially heterogeneous variability. They fitted the model under the Bayesian framework using sparsity-inducing and shrinkage-inducing prior for the propagator matrix of the basis-function coefficients, i.e., $\boldsymbol{\Phi}$. The data size being 61, 236 their model demonstrated an impressive performance. Alternatively, the dimension-reduction can be achieved through a kernel convolution of the form $X_t(\boldsymbol{s}) := \int K(\boldsymbol{u}, \boldsymbol{s})X_{t-1}(\boldsymbol{u})d\boldsymbol{u} + \eta_t(\boldsymbol{s})$. Indeed, Lemos and Sansó[51] considered a discrete process convolution that approximates the integral, and based on that, they proposed a model for a massive sea-surface temperature dataset. Their model provided an effective way of reducing the computational burden required for inference, by considering only nearby spatial locations in the formation of the discrete convolution approximation. On the contrary, Banerjee et al.[4] proposed a class of low-rank models motivated by Kriging ideas. Although the model was proposed in the context of massive spatial data, but it can be implemented to massive spatio-temporal data too. Let us briefly illustrate their idea.

Consider a set of knots $(\boldsymbol{s}_1^*, t_1^*), (\boldsymbol{s}_2^*, t_2^*), \cdots, (\boldsymbol{s}_m^*, t_m^*))$ in space–time, which may, but need not, be a subset of the entire collection of nT observed space–time locations $(\boldsymbol{s}_1, 1), (\boldsymbol{s}_2, 1), \cdots, (\boldsymbol{s}_n, T)$. If the original model for $Y(\boldsymbol{s}, t)$ is a Gaussian process with mean function 0 and covariance function $c_\eta((\boldsymbol{s}, t), (\boldsymbol{s}', t'))$, then the low-rank process is a Gaussian process with mean function 0 and covariance function $c_\eta((\boldsymbol{s}, t), (\boldsymbol{s}', t')) - \boldsymbol{\Sigma}_{\boldsymbol{s},t}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{s}',t'}$. Here, $Var(\boldsymbol{Y}^*) := \boldsymbol{\Sigma}_{22}$ where $\boldsymbol{Y}^*$ consists of $Y(\boldsymbol{s}, t)$ associated with the m spatio-temporal knot points, $Cov(Y(\boldsymbol{s}, t), \boldsymbol{Y}^*) := \boldsymbol{\Sigma}_{\boldsymbol{s},t}$ and $Cov(\boldsymbol{Y}^*, Y(\boldsymbol{s}', t')) := \boldsymbol{\Sigma}_{\boldsymbol{s}',t'}$. The name originates from the fact that this low-rank process is obtained by considering the prediction of $Y(\boldsymbol{s}, t)$ and $Y(\boldsymbol{s}', t')$, pretending as if $Y(\boldsymbol{s}, t)$ associated with the m knot points are the actual observed data. Unfortunately, for nT large, the low-rank process is not a good approximation to the original process from which the data is being generated, and moreover, under strong spatio-temporal dependence among nearby space–time coordinates, spatio-temporal models based on them perform poorly even for moderate values of nT. Instead, bias-adjusted low-rank models tend to perform better but at the cost of increased computation.

Other approaches to the modeling of massive spatio-temporal data hinge on the assumption of sparsity. Sparse methods include covariance tapering. The idea of covariance tapering is based on the fact that many entries in the $nT \times nT$ covariance matrix associated with the nT space–time coordinates are close to zero and associated space–time coordinate pairs could be considered as essentially independent. Usually, sparsity in the $nT \times nT$ covariance matrix associated with the nT space–time coordinates is induced using compactly supported spatio-temporal covariance functions, thereby producing only a few nonzero entries. This is effective for parameter estimation and prediction of the response, but it has not been fully developed or explored for more general inference on the state process. Sparsity can also be induced by considering an inverse covariance matrix associated with the nT space–time coordinates, that contains only a few nonzero entries. However, unlike low-rank processes, these do not, necessarily, extend to new random variables at arbitrary spatio-temporal coordinates. In this regard, the work of[20] is very important. They offered an alternative computationally efficient strategy by constructing a sparsity-inducing spatio-temporal process, also induced by a given Gaussian process in space–time. Unlike the covariance tapering-based sparsity methods, their method produces a sparse process over whole space–time, thereby giving a valid probability model for $Y(\boldsymbol{s}, t)$ for any combination of $\boldsymbol{s}$ and t. In their method, they replaced the full likelihood by an approximate likelihood which can be computed in $O(nT)$ flops thereby making it applicable to massive spatio-temporal datasets. In that article, Datta et al.[20] successfully applied their model to a massive pollution dataset, using a fully Bayesian framework.

Different methods of massive spatio-temporal data can be clubbed together to what can be referred to as the multi-resolution approximation. See[47], who recently studied a multi-resolution dynamic spatio-temporal model in the context of sediment movements. With the advent of powerful computers, there is a sharp rise in the number of articles on massive spatio-temporal data and it is impossible to provide a comprehensive review of all such existing methods. However, in a recent article, Heaton et al.[39] compared different classical and Bayesian methods for massive data, albeit in the purely spatial context. Banerjee[2], on the other hand, discussed the relation between many different Bayesian methods for massive data, again in the purely spatial context.

6 Discussion and Conclusion

In this article, we have considered Bayesian spatio-temporal modeling of discrete-time point-referenced spatio-temporal data. Starting with the marginal models, we have discussed different specifications of mean and covariance functions and have demonstrated how a marginal spatio-temporal model can be fitted under the Bayesian framework. We also have touched upon marginal spatio-temporal modeling of non-Gaussian data and categorical data with a discussion on how marginal spatio-temporal models can be extended to multivariate settings. Then, we have considered the conditional specification approach and reviewed some linear and nonlinear conditional spatio-temporal models, their Bayesian implementation, and related issues. Following that discussion, we also have mentioned some non-parametric spatio-temporal models. Finally, we have provided a brief exposure to spatio-temporal modeling approaches for massive spatio-temporal data. However, we have not discussed the very important problem of spatio-temporal misalignment and related Bayesian models neither have we reviewed machine learningbased Bayesian models, which is a rapidly emerging subfield of spatio-temporal statistics. Some recent articles in this direction are[53–55,76] and[81].

Publisher's Note

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

Acknowledgements

We would like to thank an anonymous referee whose suggestions lead to an improvement of the review article.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

Received: 2 September 2021 Accepted: 8 February 2022
Published online: 25 March 2022

References

1. Amini H, Taghavi-Shahri SM, Henderson SB, Hosseini V, Hassankhany H, Naderi M, Ahadi S, Schindler C, Künzli N, Yunesian M (2016) Annual and seasonal spatial models for nitrogen oxides in Tehran, Iran. Sci Rep 6(1):1–11
2. Banerjee S (2017) High-dimensional Bayesian geostatistics. Bayesian Anal 12(2):583
3. Banerjee S, Carlin B.P, Gelfand A.E (2003) Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC
4. Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B (Statistical Methodology) 70(4):825–848
5. Berger J, De Oliveira V, Sansó B, Ren C, Wittig T (2003) Objective Bayesian analysis of spatially correlated data (vol 96, pg 1361, 2001). J Am Stat Assoc 98(463):779–779
6. Berger JO, De Oliveira V, Sansó B (2001) Objective Bayesian analysis of spatially correlated data. J Am Stat Assoc 96(456):1361–1374
7. Berliner LM (1996) Hierarchical Bayesian time series models. In: Maximum entropy and Bayesian methods. edited by Kenneth M. Hanson and Richard N. Silver, Springer:15–22
8. Berliner LM, Milliff RF, Wikle CK (2003) Bayesian hierarchical modeling of air-sea interaction. J Geophys Res Oceans 108(C4):1–18
9. Berliner LM, Wikle CK, Cressie N (2000) Long-lead prediction of pacific ssts via Bayesian dynamic modeling. J Clim 13(22):3953–3968
10. Bloomfield P, Oehlert G, Thompson ML, Zeger S (1983) A frequency domain analysis of trends in Dobson total ozone records. J Geophys Res Oceans 88(C13):8512–8522
11. Bruno F, Guttorp P, Sampson PD, Cocchi D (2009) A simple non-separable, non-stationary spatiotemporal model for ozone. Environ Ecol Stat 16(4):515–529
12. Bueno RS, Fonseca TC, Schmidt AM (2017) Accounting for covariate information in the scale component of spatio-temporal mixing models. Spatial Stat 22:196–218
13. Carroll R, Chen R, George E, Li T, Newton H, Schmiediche H, Wang N (1997) Ozone exposure and

population density in Harris county, Texas. J Am Stat Assoc 92(438):392–404
14. Chakraborty A, Gelfand AE, Wilson AM, Latimer AM, Silander JA Jr (2010) Modeling large scale species abundance with latent spatial processes. Ann Appl Stat 4(3):1403–1429
15. Cox D.R, Isham V (1988) A simple spatial-temporal model of rainfall. Proc R Soc Lond A Math Phys Sci 415(1849):317–328
16. Cressie N, Huang HC (1999) Classes of nonseparable, spatio-temporal stationary covariance functions. J Am Stat Assoc 94(448):1330–1339
17. Cressie N, Wikle CK (2015) Statistics for spatio-temporal data. Wiley
18. Daniels MJ, Kass RE (1999) Nonconjugate Bayesian estimation of covariance matrices and its use in hierarchical models. J Am Stat Assoc 94(448):1254–1263
19. Das M, Bhattacharya S (2014) Nonstationary, nonparametric, nonseparable Bayesian spatio-temporal modeling using kernel convolution of order based dependent Dirichlet process. arXiv preprint arXiv:1405.4955
20. Datta A, Banerjee S, Finley AO, Hamm NA, Schaap M (2016) Nonseparable dynamic nearest neighbor gaussian process models for large spatio-temporal data with an application to particulate matter analysis. Ann Appl Stat 10(3):1286
21. Davis RA, Klüppelberg C, Steinkohl C (2013) Statistical inference for max-stable processes in space and time. J R Stat Soc Ser B (Statistical Methodology) 75(5):791–819
22. Deng JS, Wang K, Hong Y, Qi JG (2009) Spatio-temporal dynamics and evolution of land use change and landscape pattern in response to rapid urbanization. Landsc Urban Plan 92(3–4):187–198
23. Dimitrakopoulos R, Luo X (1997) Joint space-time modeling in the presence of trends. In: Geostatistics Wollongong'96 1: 138–149
24. Dou Y, Le ND, Zidek JV (2010) Modeling hourly ozone concentration fields. Ann Appl Stat 4(3):1183–1213
25. Fonseca TC, Steel MF (2011) Non-gaussian spatiotemporal modelling through scale mixing. Biometrika 98(4):761–774
26. Fuentes M (2002) Spectral methods for nonstationary spatial processes. Biometrika 89(1):197–210
27. Fuentes M, Chen L, Davis JM (2008) A class of nonseparable and nonstationary spatial temporal covariance functions. Environmetrics 19(5):487–507
28. Furrer R, Sain SR (2009) Spatial model fitting for large datasets with applications to climate and microarray problems. Stat Comput 19(2):113–128
29. Gamerman D, Lopes HF, Salazar E (2008) Spatial dynamic factor analysis. Bayesian Anal 3(4):759–792
30. Gelfand AE, Banerjee S (2017) Bayesian modeling and analysis of geostatistical data. Annu Rev Stat Appl 4:245–266
31. Gelfand AE, Banerjee S, Gamerman D (2005) Spatial process modelling for univariate and multivariate dynamic spatial data. Environmetrics 16(5):465–479
32. Gelfand A.E, Diggle P, Guttorp P, Fuentes M (2010) Handbook of spatial statistics. CRC Press
33. Gelfand AE, Ecker MD, Knight JR, Sirmans C (2004) The dynamics of location in home price. J Real Estate Finance Econ 29(2):149–166
34. Giannitrapani M, Bowman A, Scott M, Smith R (2006) Sulphur dioxide in Europe: statistical relationships between emissions and measured concentrations. Atmos Environ 40(14):2524–2532
35. Gneiting T (2002) Nonseparable, stationary covariance functions for space-time data. J Am Stat Assoc 97(458):590–600
36. Guha S, Bhattacharya S (2014) Gaussian random functional dynamic spatio-temporal modeling of discrete time spatial time series data. arXiv preprint arXiv:1405.6531
37. Guttorp P, Meiring W, Sampson PD (1994) A space-time analysis of ground-level ozone data. Environmetrics 5(3):241–254
38. Haas TC (1990) Lognormal and moving window methods of estimating acid deposition. J Am Stat Assoc 85(412):950–963
39. Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, Guhaniyogi R, Gerber F, Gramacy RB, Hammerling D, Katzfuss M et al (2019) A case study competition among methods for analyzing large spatial data. J Agric Biol Environ Stat 24(3):398–425
40. Higdon D (1998) A process-convolution approach to modelling temperatures in the north Atlantic ocean. Environ Ecol Stat 5(2):173–190
41. Holland DM, De OV, Cox LH, Smith RL (2000) Estimation of regional trends in sulfur dioxide over the eastern united states. Environmetrics 11(4):373–393
42. Hooten MB, Wikle CK (2007) Shifts in the spatio-temporal growth dynamics of shortleaf pine. Environ Ecol Stat 14(3):207–227
43. Huerta G, Sansó B, Stroud JR (2004) A spatiotemporal model for Mexico city ozone levels. J Roy Stat Soc Ser C (Appl Stat) 53(2):231–248
44. Hughes JP, Guttorp P (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resour Res 30(5):1535–1546
45. Huser R, Davison AC (2014) Space-time modelling of extreme events. J R Stat Soc Ser B Stat Methodol 76(2):439–461
46. Jun M, Knutti R, Nychka DW (2008) Spatial analysis to quantify numerical model bias and dependence: How many climate models are there? J Am Stat Assoc 103(483):934–947

47. Jurek M, Katzfuss M (2021) Multi-resolution filters for massive spatio-temporal data. J Comput Graph Statist 30(4):1–16
48. Katzfuss M, Cressie N (2012) Bayesian hierarchical spatio-temporal smoothing for very large datasets. Environmetrics 23(1):94–107
49. Knight JR, Dombrow J, Sirmans C (1995) A varying parameters approach to constructing house price indexes. Real Estate Econ 23(2):187–205
50. Kyriakidis PC, Journel AG (1999) Geostatistical space-time models: a review. Math Geol 31(6):651–684
51. Lemos RT, Sansó 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
52. Lu Z, Steinskog DJ, Tjøstheim D, Yao Q (2009) Adaptively varying-coefficient spatiotemporal models. J R Stat Soc Ser B (statistical methodology) 71(4):859–880
53. McDermott PL, Wikle CK (2017) An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. Stat 6(1):315–330
54. McDermott PL, Wikle CK (2019) Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. Entropy 21(2):184
55. McDermott PL, Wikle CK (2019) Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. Environmetrics 30(3):e2553
56. Milliff R.F, Bonazzi A, Wikle C.K, Pinardi N, Berliner L.M (2011) Ocean ensemble forecasting. part i: ensemble Mediterranean winds from a Bayesian hierarchical model. Q J R Meteorol Soc 137(657):858–878
57. Paciorek CJ, Yanosky JD, Puett RC, Laden F, Suh HH (2009) Practical large-scale spatio-temporal modeling of particulate matter concentrations. Ann Appl Stat 3(1):370–397
58. Paez MS, Gamerman D, Landim FM, Salazar E (2008) Spatially varying dynamic coefficient models. J Stat Plan Inference 138(4):1038–1058
59. Reinsel G, Tiao GC, Wang M, Lewis R, Nychka D (1981) Statistical analysis of stratospheric ozone data for the detection of trends. Atmos Environ (1967) 15(9):1569–1577
60. Sahu SK, Gelfand AE, Holland DM (2010) Fusing point and areal level space-time data with application to wet deposition. J Roy Stat Soc Ser C (Appl Stat) 59(1):77–103
61. Sahu SK, Mardia KV (2005) A Bayesian kriged Kalman model for short-term forecasting of air pollution levels. J Roy Stat Soc Ser C (Appl Stat) 54(1):223–244
62. Sain SR, Furrer R, Cressie N (2011) A spatial analysis of multivariate output from regional climate models. Ann Appl Stat 5(1):150–175
63. Sampson PD, Guttorp P (1992) Nonparametric estimation of nonstationary spatial covariance structure. J Am Stat Assoc 87(417):108–119
64. Sang H, Jun M, Huang JZ (2011) Covariance approximation for large multivariate spatial data sets with an application to multiple climate model errors. Ann Appl Stat 5(4):2519–2548
65. Sans Ó B, Schmidt A.M, Nobre A.A (2008) Bayesian spatio-temporal models based on discrete convolutions. Can J Stat 36(2):239–258
66. Sanso B, Guenni L (1999) Venezuelan rainfall data analysed by using a Bayesian space-time model. J R Stat Soc Ser C (Appl Stat) 48(3):345–362
67. Schmidt AM, Gonçalves KC, Velozo PL (2017) Spatiotemporal models for skewed processes. Environmetrics 28(6):e2411
68. Smith RL, Tebaldi C, Nychka D, Mearns LO (2009) Bayesian modeling of uncertainty in ensembles of climate models. J Am Stat Assoc 104(485):97–116
69. Stein ML (2005) Space-time covariance functions. J Am Stat Assoc 100(469):310–321
70. Stroud JR, Müller P, Sansó B (2001) Dynamic models for spatiotemporal data. J R Stat Soc Ser B (Statistical Methodology) 63(4):673–689
71. Tebaldi C, Sansó B (2009) Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. J R Stat Soc A Stat Soc 172(1):83–106
72. Tonellato SF (2001) A multivariate time series model for the analysis and prediction of carbon monoxide atmospheric concentrations. J R Stat Soc Ser C (Appl Stat) 50(2):187–200
73. Tonellato SF et al (2005) Identifiability conditions for spatio-temporal Bayesian dynamic linear models. Metron-Int J Stat 63(1):81–101
74. Turkman KF, Turkman MA, Pereira J (2010) Asymptotic models and inference for extremes of spatio-temporal data. Extremes 13(4):375–397
75. Waller LA (2005) Bayesian thinking in spatial statistics. Handb Stat 25:589–622
76. Wikle CK (2019) Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. J Agric Biol Environ Stat 24(2):175–203
77. Wikle CK, Berliner LM, Cressie N (1998) Hierarchical Bayesian space-time models. Environ Ecol Stat 5(2):117–154
78. Wikle CK, Holan SH (2011) Polynomial nonlinear spatio-temporal integro-difference equation models. J Time Ser Anal 32(4):339–350
79. Wikle CK, Hooten MB (2010) A general science-based framework for dynamical spatio-temporal models. TEST 19(3):417–451
80. 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
81. Zammit-Mangion A, Wikle CK (2020) Deep integro-difference equation models for spatio-temporal forecasting. Spatial Stat 37:100408

Suman Guha received his M.Stat. degree with specialization in Mathematical Statistics and Probability (MSP) on 2009 from Indian Statistical Institute, Kolkata and Ph.D. Statistics on 2020 from Indian Statistical Institute, Kolkata. He started working as an Assistant Professor of Statistics from 2014 at Bethune College, Kolkata and currently he is an Assistant Professor at the Department of Statistics, Presidency University, Kolkata.

Sourabh Bhattacharya received his M.Stat. degree with specialization in Mathematical Statistics and Probability (MSP) on 2001 from Indian Statistical Institute, Kolkata and Ph.D. Statistics on 2004 from Trinity College Dublin, Ireland followed by a postdoc at ISDS, Duke University and University of Sheffield. He started working as an Assistant Professor of Statistics from 2007 at ISRU, Indian Statistical Institute, Kolkata and currently he is an Associate Professor at the same department.