Bayesian Nonparametric Modeling For Integro-Difference Equations

Robert Richardson, Athanasios Kottas and Bruno Sansó *

Abstract

Integro-Difference Equations (IDEs) provide a flexible framework for dynamic modeling of spatio-temporal data. The choice of kernel in an IDE model relates directly to the underlying physical process modeled, and it can affect model fit and predictive accuracy. We introduce Bayesian nonparametric methods to the IDE literature as a means to allow flexibility in modeling the kernel. We propose a mixture of normal distributions for the IDE kernel, built from a spatial Dirichlet process for the mixing distribution, which can model kernels with shapes that change with location. This allows the IDE model to capture non-stationarity with respect to location and to reflect a changing physical process across the domain. We address computational concerns for inference that leverage the use of Hermite polynomials as a basis for the representation of the process and the IDE kernel, and incorporate Hamiltonian Markov chain Monte Carlo steps in the posterior simulation method. An example with synthetic data demonstrates that the model can successfully capture location-dependent dynamics. Moreover, using a data set of ozone pressure, we show that the spatial Dirichlet process mixture model outperforms several alternative models for the IDE kernel, including the state of the art in the IDE literature, that is, a Gaussian kernel with location-dependent parameters.

KEY WORDS: Dirichlet process mixtures; Hamiltonian Markov chain Monte Carlo; Hermite polynomials; Spatial Dirichlet process.

*Robert Richardson (richardson@stat.byu.edu) is Assistant Professor, Department of Statistics, Brigham Young University. Athanasios Kottas (thanos@soe.ucsc.edu) and Bruno Sansó (bruno@soe.ucsc.edu) are Professors of Statistics, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

1 Introduction

Data arising from spatio-temporal processes is collected frequently in a variety of fields, including ecology and environmental sciences. Methods to model such data must simultaneously provide for spatial and temporal dependence. This has been accomplished via spatio-temporal covariance structures (Cressie and Huang, 1999) and scientifically motivated partial differential equation (PDE) models (Hooten and Wikle, 2008). While these two methods are connected (Heine, 1955), each has different advantages and limitations. Determining classes of spatial covariance functions leads to a variety of useful techniques, such as kriging (Cressie, 1993) and convolution modeling (Higdon, 1998) and the interpretation of these representations as second order features is straightforward, even for spatio-temporal extensions. While recent work has focused on flexible non-separable covariance functions (Ma, 2003), these methods still lag behind more descriptive modeling techniques and do not scale well with the size of the data. Scientific evidence can motivate explicit PDE representations of spatio-temporal processes (Jones and Zhang, 1997), but if the nature of the relationship is uncertain, restricting to a specific form can result in unneeded bias and a poor model fit.

Dynamic spatio-temporal models maintain the positive traits of both of these modeling techniques while not limited in the same ways (Wikle and Hooten, 2010). Specific formulations can model physical characteristics of spatio-temporal processes while allowing for a great deal of second order flexibility. A hierarchical dynamic spatio-temporal model (Cressie and Wikle, 2011) includes a level for data observations, $Y_t(s)$, and a level for process evolution:

$$Y_t(s) = X_t(s) + \varepsilon_t(s), \quad \text{and} \quad X_t(s) = \mathcal{M}(X_{t-1}(s)) + \omega_t(s), \tag{1}$$

where $\varepsilon_t(s)$ is independent observational noise and $\omega_t(s)$ is potentially correlated process noise. Here, s refers to the location at which the process is measured and is two-dimensional in many applications, but may also be one- or three-dimensional in others. The specific model choice for the process evolution, represented by $\mathcal{M}(\cdot)$ in (1), can be defined by scientific justification or other specific process characteristics. A natural choice is the integro-difference equation (IDE)

model formulation. An IDE process model is expressed as

$$X_t(s) = e^{\lambda} \int k(u \mid s, \boldsymbol{\theta}) X_{t-1}(u) \, \mathrm{d}u + \omega_t(s), \tag{2}$$

where $k(u \mid s, \boldsymbol{\theta})$ is a redistribution kernel with parameter vector $\boldsymbol{\theta}$. The IDE kernel weights the contribution of the process at location u and time $t-1$ to the process at location s and time t. The parameter λ controls growth or decay of the process; in this paper, we set $\lambda = 0$.

The use of IDEs in stochastic modeling began in ecology with the works of Neubert et al. (1995) and Kot et al. (1996) on modeling the spread of an invasive organism. IDEs were then expanded for use in general spatio-temporal models in Wikle and Cressie (1999). Brown et al. (2000) and Storvik et al. (2002) explore properties of the IDE model, such as stationarity and separability, and make connections to other space-time modeling approaches. IDE modeling has been further elaborated by using spatially dependent kernel parameters (Wikle, 2002; Xu et al., 2005) and non-linear interaction terms (Wikle and Holan, 2011).

Despite no inherent restrictions of kernel choice implied by the IDE structure, the ubiquitous choice for a kernel is the Gaussian distribution. Richardson et al. (2017) show that additional physical characteristics of the process can be modeled by allowing the kernel to extend past the Gaussian distribution with respect to tail behavior. In particular, asymmetric Laplace and stable family kernel IDE models were shown to substantially outperform the Gaussian kernel IDE model in prediction. Building upon this idea, we explore nonparametric prior models for the kernel. Specifically, we model the IDE kernel with a mixture of normals built from a location-dependent mixing distribution, to which we assign a spatial Dirichlet process (SDP) prior. Hence, the prior model supports non-Gaussian IDE kernels where shape depends on location. The result is a non-separable, non-stationary dynamic model with a high degree of flexibility in the underlying physical processes being modeled. We will show how a nonparametric kernel IDE model can capture effects existing models cannot, such as location-dependent long-tail dependence or skewness. We also study the special case of the model for processes that are stationary with respect to location, based on a Dirichlet process (DP) mixture of normals prior

for the kernel with a mixing distribution that does not change with location. The computational complexities of the model are handled with Hermite polynomial approximations to the IDE kernel density and Hamiltonian Markov chain Monte Carlo (HMCMC) posterior simulation methods.

In this paper, we focus our attention to one-dimensional space for the underlying process. Section 5 includes discussion on the challenge of two-dimensional extensions, and on possible strategies to account for the added complications. The univariate case has merit on its own from a practical point of view, since there are several applications where, for instance, environmental variables are measured across time by altitude, depth or length. As an example, the application we use to illustrate the methodology involves ozone pressure data collected by releasing a balloon in the air that takes ozone pressure measurements at certain intervals throughout its flight.

The outline of the paper is as follows. Section 2 develops the methodology, including the model formulation for the IDE kernel, the basis function expansion used for the hierarchical model representation of the data, and the posterior simulation method based on HMCMC techniques. Section 3 presents a simulated data example to demonstrate the capacity of the SDP mixture kernel IDE model to capture dynamics that change by location through distinct kernels. In Section 4, we apply the DP and SDP mixture models to the ozone pressure data, including comparison with parametric kernel IDE models. Finally, Section 5 concludes with a discussion.

2 Methods

2.1 Model Formulation

An important advantage to using kernels with location-dependent parameters involves stationarity of the resulting process, $X_t(s)$, for given time point t. Brown et al. (2000) show that IDE models are stationary in space, but the work was limited to kernels with parameters which are independent of location. The result in Appendix A shows that the process is non-stationary with respect to location under the more general case with a location-dependent IDE kernel.

Our key modeling objective is to achieve the flexibility provided by IDE models with location-dependent kernels, while also allowing more general kernel distributional shapes than the Gaus-

sian. To this end, we use a normal mixture model for the IDE kernel with a location-dependent mixing distribution, which is assigned a nonparametric prior. In particular, we use an SDP (Gelfand et al., 2005; Kottas et al., 2008), which provides a prior probability model for non-Gaussian, non-stationary processes. The SDP prior is defined through a base stochastic process, which can be taken to be an isotropic Gaussian process, and a scalar parameter $\alpha > 0$ which controls how close an SDP realization is to the base process. The SDP is defined by extending the DP constructive definition (Sethuraman, 1994). More specifically, an SDP realization has the (almost surely) discrete representation $G_D = \sum_{l=1}^{\infty} w_l \delta_{\theta_{l,D}}$, where each $\theta_{l,D} = \{\theta_l(s) : s \in D\}$ is an independent realization from the base Gaussian process defined over region D. (Here, δ_a denotes a point mass at a.) The weights are defined through stick-breaking, which involves latent variables $\xi_1, \xi_2, ...,$ drawn independently from a Beta$(1, \alpha)$ distribution: $w_1 = \xi_1$ and $w_l = \xi_l \prod_{i=1}^{l-1}(1 - \xi_i)$, for $l \geq 2$. The SDP implies a DP prior for the distribution associated with any finite set of locations. In particular, for a generic location $s \in D$, the implied $G_s = \sum_{l=1}^{\infty} w_l \delta_{\theta_l(s)}$ is a realization from a DP prior with precision parameter α, and base distribution given by the normal distribution induced by the Gaussian process at location s.

Because the weights are stochastically decreasing, the sum can be truncated to a finite value L, such that $w_L = 1 - \sum_{l=1}^{L-1} w_l = \prod_{i=1}^{L-1}(1 - \xi_i)$. Such truncation facilitates implementation of our model. The truncation level L can be chosen to any desired level of accuracy, using standard DP properties. For instance, a simple way to specify L is through the expectation for the partial sum of the original DP weights, $\mathrm{E}(\sum_{l=1}^{L} w_l \mid \alpha) = 1 - \{\alpha/(\alpha + 1)\}^L$. This expression can be averaged over the prior for α to estimate the marginal prior expectation $\mathrm{E}(\sum_{l=1}^{L} w_l)$, which is used to specify L given a tolerance level for the approximation.

Denote by $\phi(\cdot \mid \mu, \sigma_0^2)$ the density of the normal distribution with mean μ and variance σ_0^2. Then, the SDP mixture model for the IDE kernel is given by

$$k(u \mid s, G_s, \sigma_0^2) = \sum_{l=1}^{L} w_l \, \phi(u \mid s + \mu_l(s), \sigma_0^2) \tag{3}$$

where, for any finite set of locations $s_1, ..., s_n$, the distribution of the corresponding vector of

atoms, $(\mu_l(s_1), ..., \mu_l(s_n))$, is n-dimensional normal arising as the finite dimensional distribution of the base Gaussian process. Note that the IDE kernel is location-dependent through the location-dependent mixing distribution rather than simply through extending the parameter of a parametric IDE kernel to a process. The practical implication is that the kernel can change its shape in a more flexible fashion than only through the first or second moment of the Gaussian kernel IDE model. Regarding the implied IDE process, equation (2) can be written as

$$X_t(s) = \sum_{l=1}^{L} w_l \int \phi(u \mid s + \mu_l(s), \sigma_0^2) X_{t-1}(u)\, \mathrm{d}u \, + \, \omega_t(s),$$

which is a weighted sum of L Gaussian kernel IDE models with location-dependent means.

2.2 Basis Function Expansion

An IDE model is typically decomposed using a common orthonormal basis for the process and the kernel (Wikle, 2002). For a given set of basis functions, $\psi_i(s)$, the process is written as $X_t(s) = \sum_{i=1}^{\infty} a_i(t)\psi_i(s)$ and the kernel as $k(u \mid s, \boldsymbol{\theta}_s) = \sum_{j=1}^{\infty} b_j(s, \boldsymbol{\theta}_s)\psi_j(u)$. Any orthonormal basis will then lead to the representation of the process at time $t+1$ as $X_{t+1}(s) = \sum_{i=1}^{\infty} a_i(t) b_i(s, \boldsymbol{\theta}_s)$. When fitting this model, the basis functions are predetermined and fixed, and their number truncated to a finite value K. The coefficients of the process, $a_i(t)$, are random variables contributing to the stochasticity of the process $X_t(s)$. The coefficients of the kernel decomposition are deterministic given the distributional family of the kernel and its parameters.

Using these representations, a hierarchical IDE model can be rewritten as

$$\begin{aligned} Y_t(s) &= \sum_{i=1}^{K} a_i(t)\psi_i(s) \, + \, \varepsilon_t(s) && (4) \\ \sum_{i=1}^{K} a_i(t)\psi_i(s) &= \sum_{i=1}^{K} a_i(t-1) b_i(s, \boldsymbol{\theta}_s) \, + \, \omega_t(s). && (5) \end{aligned}$$

The number of basis functions used for the series expansion approximations is directly related to the accuracy of the approximation to the kernel and inversely related to computational speed.

Physicist's Hermite polynomials are defined as $H_n(x) = \left(2x - \frac{d}{dx}\right)^n \cdot 1$. These polynomials are

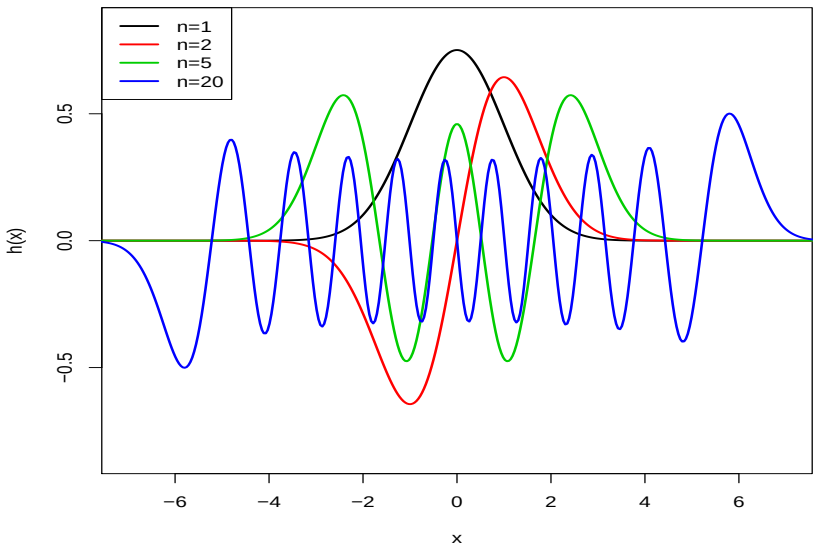

Figure 1: A sample of Hermite basis functions.

orthogonal with respect to the weight function $\exp(-x^2)$. The specific inner product for Hermite polynomials is $\int_{-\infty}^{\infty} H_m(x)H_n(x)\exp(-x^2)\mathrm{d}x = \sqrt{\pi}2^n n!\delta_{nm}$. The new polynomials $h_n(x) = H_n(x)\exp(-x^2/2)(\sqrt{\pi}2^n n!)^{-1/2}$, for $n = 0, 1, 2, ...$, are called Hermite functions. They satisfy the orthogonality property $\int_{-\infty}^{\infty} h_m(x)h_n(x)\mathrm{d}x = \delta_{nm}$, thus providing a basis which can be used for the series expansion in IDE modeling. When using Hermite functions in a series expansion to approximate the SDP normal mixture IDE kernel in (3), the coefficient corresponding to the n-th Hermite function is given by

$$b_n(s, \boldsymbol{\theta}_s) = \sum_{l=1}^{L} w_l \frac{1}{\sqrt{(\sqrt{\pi}2^n n!)\,(1+\sigma_0^2)}} \exp\left(-\frac{\mu_l^2(s)}{2(1+\sigma_0^2)}\right) \sum_{k=0}^{n} H_{n,k} m_{k,l}(s). \tag{6}$$

Here, the parameter vector, $\boldsymbol{\theta}_s$, includes the weights, w_l, and location-dependent means, $\mu_l(s)$, for each atom l, as well as the variance σ_0^2. Moreover, $H_{n,k}$ is the k-th coefficient in the n-th Hermite polynomial, and $m_{k,l}(s)$ is the k-th raw moment of a normal distribution with mean $\mu_l(s)/(\sigma_0^2+1)$ and variance $\sigma_0^2/(\sigma_0^2+1)$. Then, decomposing the kernel results in $k(u \mid s, \boldsymbol{\theta}_s) = \sum_{n=0}^{K} b_n(s, \boldsymbol{\theta}_s)h_n(u)$, and decomposing the process yields $X_t(s) = \sum_{m=0}^{K} a_m(t)h_m(s)$.

Number of basis functions	Suggested range
10	(-4,4)
20	(-5,5)
30	(-6.5,6.5)
40	(-8,8)
50	(-9,9)

Table 1: Suggested ranges are given for the corresponding number of basis functions.

Figure 1 shows that the squared exponential weight function causes the function to decrease to 0 far from the origin, so Hermite functions are unable to approximate functions that extend past a certain domain. To avoid this, the locations of the data should be scaled using a standard linear transformation. Resulting estimates of kernel densities can be scaled back without consequence. These new spatial locations can be restricted within a certain range, but relative distances between adjacent points remain the same. The actual distances will change, however, requiring the range parameters in the process covariance structure to be adjusted. There is a tradeoff between computational burden and efficiency when choosing the new range. Some suggested ranges are given in Table 1. These ranges are found by graphical exploration using a normal kernel with standard deviation 0.25. This value for the standard deviation is chosen as a baseline. Densities with larger/smaller standard deviation need fewer/more basis functions.

Figure 2 shows how the Hermite basis and Fourier basis differ when approximating the normal density using a small number of basis functions. The difference can become even more dramatic when the kernel width becomes smaller. In our experience, the Hermite basis performs better than the Fourier basis when approximating the Gaussian kernel. One reason for this is that the exponential weight of the Hermite functions matches the exponential tails of the Gaussian kernel, while the Fourier basis is composed of functions with sinusoidal tails. Also, using the Fourier basis requires specification of a region which needs to be expanded past the data to account for its periodic nature. This extra space on the boundary requires additional Fourier basis functions to accurately approximate the Gaussian kernel.

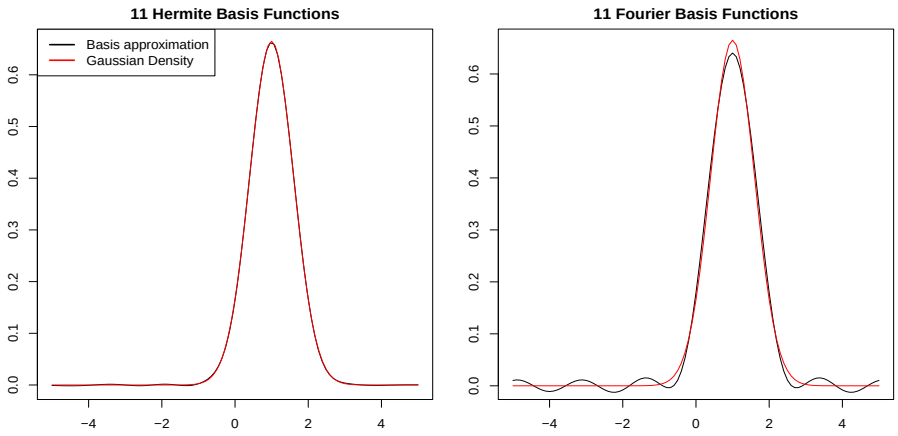

Figure 2: The approximation to a normal density with mean 1 and variance 0.6^2 is compared using 11 Hermite basis functions and 11 Fourier basis functions over a range of -4 to 4.

2.3 Posterior Inference

The IDE model we use for posterior inference is designed as in Wikle (2002), with the main difference being the distribution used for the kernel. Denote by $(s_{t,1}, ..., s_{t,n_t})$ the locations at which observations are available at time t, for $t = 1, ..., T$, and by $\boldsymbol{Y}_t = (Y_t(s_{t,1}), ..., Y_t(s_{t,n_t}))'$ the corresponding data vector. Then, using a basis function representation for the IDE model as in equations (4) and (5), the hierarchical model for the data can be generically written as

$$\boldsymbol{Y}_t \mid \boldsymbol{a}_t, \sigma^2 \sim \mathrm{N}(\boldsymbol{\Psi}_t \boldsymbol{a}_t, \sigma^2 \boldsymbol{I}_{n_t}), \quad t = 1, ..., T \tag{7}$$

$$\boldsymbol{a}_t \mid \boldsymbol{a}_{t-1}, \tau^2, \boldsymbol{\theta} \sim \mathrm{N}(\boldsymbol{G}_t \boldsymbol{B}_{\theta,t} \boldsymbol{a}_{t-1}, \tau^2 \boldsymbol{G}_t \boldsymbol{V}_t \boldsymbol{G}_t'), \quad t = 1, ..., T \tag{8}$$

$$\boldsymbol{\theta} \mid \gamma \sim p(\boldsymbol{\theta} \mid \gamma), \quad \sigma^2, \tau^2 \sim p(\sigma^2)p(\tau^2). \tag{9}$$

Here, $\boldsymbol{a}_t$ are the latent state variables representing the stochastic basis coefficients of the process, the (i,j)th element of $\boldsymbol{\Psi}_t$ is $\psi_i(s_{t,j})$ (the i-th basis function evaluated at $s_{t,j}$), the (i,j)-th element of $\boldsymbol{B}_{\theta,t}$ is $b_i(s_{t,j}, \boldsymbol{\theta})$ (the i-th basis coefficient of the kernel at the location $s_{t,j}$), and $\boldsymbol{G}_t = (\boldsymbol{\Psi}_t' \boldsymbol{\Psi}_t)^{-1} \boldsymbol{\Psi}_t'$. The length of the state vector is equal to the number of basis functions. The observational variance is $\sigma^2 \boldsymbol{I}_{n_t}$ and the variance of the process level is $\tau^2 \boldsymbol{G}_t \boldsymbol{V}_t \boldsymbol{G}_t'$, where $\boldsymbol{V}_t$ is a spatial covariance matrix. The parameters in the model are σ^2, τ^2, $\boldsymbol{\theta}$, and possibly γ, if hyper-priors are used. Note that for location-dependent kernels, $\boldsymbol{\theta}$ collects all parameters

that define each $\boldsymbol{\theta}_s$. We use inverse gamma priors for σ^2 and τ^2 with parameters $\alpha_\sigma, \beta_\sigma$ and α_τ, β_τ, respectively, where the mean of an inverse gamma distribution with parameters α and β is $\beta/(\alpha-1)$, provided $\alpha > 1$. A prior must also be specified for $\boldsymbol{a}_0$, the coefficients of the basis expansion of the time zero process, and this is taken to be multivariate normal with mean $\boldsymbol{m}_0$ and variance $\boldsymbol{C}_0$. The locations used to determine $\boldsymbol{\Psi}_t$ in (7) are the observed locations of the data. The number of observations and their locations need not be the same for each time point, which is why $\boldsymbol{\Psi}_t$ potentially changes over time even though the basis functions themselves stay the same. Also, the locations used to determine the matrices in equation (8) can be different to those of the data locations. For computational convenience, a new grid is defined at locations $r_1, ..., r_n$ for all t. We can thus drop the time index in the evolution matrix, setting $\boldsymbol{G}_t \equiv \boldsymbol{G} = \boldsymbol{(\Psi'\Psi)^{-1}\Psi'}$, where $\boldsymbol{\Psi}$ and $\boldsymbol{B}_{\theta,t} = \boldsymbol{B}_\theta$ are determined using the new grid. This produces substantial computational time and memory savings for long times series of data.

2.3.1 Implementation details for the SDP mixture kernel IDE model

For the SDP mixture IDE kernel model, the full parameter set represented by $\boldsymbol{\theta}$ includes the stick-breaking random variables that define the weights, $\{\xi_l : l = 1, ..., L-1\}$, the location-dependent mixture component means $\{\mu_l(s) : l = 1, ..., L\}$, and the common mixture component variance σ_0^2. Because the means are only needed for a finite number of locations, each $\mu_l(s)$ requires a multivariate normal random vector corresponding to the SDP base Gaussian process. However the dimension of this vector may still be very large and computationally burdensome.

To reduce the dimensionality of the problem, we represent the SDP base Gaussian process as a discrete process convolution (Higdon, 1998). Thus, based on a function k_ζ and a grid, $u_1, ..., u_q$, we set $\mu_l(s) = \mu_\zeta + \sum_{i=1}^q k_\zeta(u_i, s)\zeta_l(u_i)$, where $\zeta_l(u_i) \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\zeta^2)$. The hyperparameters μ_ζ and σ_ζ^2 represent the mean and scale of the base Gaussian process, and the process convolution kernel k_ζ controls the correlation structure. We fix k_ζ to a certain form, but include μ_ζ and σ_ζ^2 as parameters to be estimated. A discretized version of the kernel convolution is used by assigning the mean process at locations $s_1, ..., s_n$ to $(\mu_l(s_1), ..., \mu_l(s_n))' = \boldsymbol{K}_\mu\boldsymbol{\zeta}_l + \mu_\zeta\boldsymbol{1}$, where $\boldsymbol{\zeta}_l = (\zeta_l(u_1), ..., \zeta_l(u_q))'$, and $\boldsymbol{K}_\mu$ is a $n \times q$ matrix that maps the smaller grid to the larger grid based

on the discrete process convolution. The value for q must be chosen to balance the amount of parameter size reduction and the amount of local variation recovered by the approximation.

The model specified in equation (7) – (9) corresponds to a conditional dynamic linear model. As such it is straightforward to obtain the full conditional distributions needed to sample the model parameters. Specifically, conditional on $\boldsymbol{\theta}$, σ^2, and τ^2, we use Forward Filtering Backwards Sampling (Frühwirth-Schnatter, 1994) to draw a sample for $\boldsymbol{a}_0, ..., \boldsymbol{a}_T$. Conditional on sampled state vectors, the posteriors for σ^2 and τ^2 are

$$\sigma^2 \mid \cdot \sim \text{IG}\left(\alpha_\sigma + \frac{nT}{2}, \beta_\sigma + \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{Y}_t - \boldsymbol{\Psi}\boldsymbol{a}_t)'(\boldsymbol{Y}_t - \boldsymbol{\Psi}\boldsymbol{a}_t)\right) \tag{10}$$

$$\tau^2 \mid \cdot \sim \text{IG}\left(\alpha_\tau + \frac{KT}{2}, \beta_\tau + \frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{a}_t - \boldsymbol{G}\boldsymbol{B}_\theta\boldsymbol{a}_{t-1})'\boldsymbol{V}^{-1}(\boldsymbol{a}_t - \boldsymbol{G}\boldsymbol{B}_\theta\boldsymbol{a}_{t-1})\right). \tag{11}$$

To obtain samples of the set of kernel parameters, $\boldsymbol{\theta}$, we observe that the conditional posterior of $\boldsymbol{\theta}$ is proportional to $\prod_{t=1}^{T} p(\boldsymbol{a}_t \mid \boldsymbol{a}_{t-1}, \tau^2, \boldsymbol{\theta})p(\boldsymbol{\theta} \mid \gamma)$. For applications with a complicated kernel, such as the SDP mixture, learning the parameters may be very difficult using standard Metropolis-Hastings. For situations such as these, we have found that a HMCMC algorithm (Neal, 2011) may be used successfully to sample from the posterior distribution of $\boldsymbol{\theta}$.

Hamiltonian Markov chain Monte Carlo introduces latent variables for each parameter. These represent the momentum of the path the parameter follows in the HMCMC chain. The momentum and position of the proposal change according to Hamiltonian dynamics. We use HMCMC to sample the latent variables $\{\zeta_l(u_i) : l = 1, ..., L;\ i = 1, ..., q\}$. There are Lq of these latent parameters, but the HMCMC is split up into L blocks, where each $(\zeta_l(u_1), ..., \zeta_l(u_q))$ is updated individually. Because the latent kernel convolution variables are linearly related to the parameters, an application of the chain rule shows that $\partial b_n(s_j, \boldsymbol{\theta})/\partial\zeta_l(u_i) = \sum_{m=1}^{q} k_\zeta(u_i, s_m)\partial b_n(s_j, \boldsymbol{\theta})/\partial\mu_l(s_m)$. This property can be used to calculate the Hamiltonian dynamics required for the HMCMC on the latent parameters. Further details are provided in Appendix B. HMCMC significantly speeds up convergence to the posterior when compared to standard Metropolis-Hastings. Our implementation involves the most basic of HMCMC algorithms. Girolami and Calderhead (2011)

extends HMCMC to employ the Riemann manifolds of the parameter space to improve the mixing. Welling and Teh (2011) and Chen et al. (2014) speed up convergence to the posterior in HMCMC by using stochastic gradients.

The kernel parameters σ_0^2 and $\{\xi_l : l = 1, ..., L-1\}$ can be sampled using standard Metropolis-Hastings steps. The hyperparameters μ_ζ and σ_ζ^2 are readily sampled under (conditionally) conjugate priors, that is, a normal prior for μ_ζ and an inverse gamma prior for σ_ζ^2.

2.3.2 Implementation of related IDE models

To demonstrate the effectiveness of the SDP mixture of normals kernel, we will also consider the location-dependent normal kernel (Wikle, 2002). For the SDP mixture kernel, the mean is mixed with the SDP, but the variance is constant throughout space $\big(k(u \mid s, \boldsymbol{\theta}) = \int \phi(u \mid \mu, \sigma_0^2) \mathrm{d}G_s(\mu)\big)$, whereas the location-dependent normal kernel uses processes for the mean and variance $\big(k(u \mid s, \boldsymbol{\theta}) = \phi(u \mid \mu(s), \sigma_0^2(s))\big)$. Note that even though the parameter σ_0^2 is constant through space in the SDP model, the mixture results in the variance of the IDE kernel being dependent on the location.

For the location-dependent normal kernel IDE model, we apply independently to the mean, $\mu(s)$, and log variance, $\log(\sigma_0^2(s))$, the process convolution approximation, discussed in the previous section. The mean process is approximated using a convolution kernel in exactly the same way the individual mean components are approximated in the SDP mixture model. The log variance is approximated with a convolution kernel, where latent variables $\boldsymbol{\eta} = (\eta(u_1), ..., \eta(u_q))'$ are *i.i.d.* from $\mathrm{N}(0, \sigma_\eta^2)$. The log variance is then set to $(\log(\sigma_0^2(s_1)), ..., \log(\sigma_0^2(s_n)))' = \boldsymbol{K}_\sigma \boldsymbol{\eta} + \mu_\eta \mathbf{1}$. The specific convolution kernel and knot locations can be different for the mean process and the log variance process. Here, we use independent Gaussian processes for the mean and log variance, but there may be merit in using a bivariate Gaussian process. The sampling procedure is similar to the one used for the SDP mixture kernel model. The latent process variables for the convolution approximation to the mean and log variance can be estimated using HMCMC. The additional hyperparameters, μ_η and σ_η^2, can be treated similarly to μ_ζ and σ_ζ^2. The overall number of parameters is reduced because there are no stick-breaking weights to estimate and

only one mean process. Conditional on the IDE kernel parameters, sampling of the state variables and of the process and observational error variances proceeds similarly to the SDP mixture model.

As a simplified version of the SDP kernel IDE model described in Section 2.1, we can construct a DP mixture kernel IDE model by substituting μ_l for $\mu_l(s)$. The atoms μ_l are drawn from a base distribution G_0 instead of a process. The result is a flexible kernel for processes which are stationary with respect to location for any given time point. The kernel is still a weighted average of Gaussian IDE kernels, but it no longer varies by location. Thus, there is no need for the convolution process approximation. Much of the remaining analysis is the same, but with a drastically reduced parameter set for the IDE kernel.

Each kernel choice produces a draw from the posterior distribution in roughly the same amount of time. However, given that it is defined through a larger number of parameters, convergence is reached more slowly for the SDP mixture kernel. In general, the location-dependent normal kernel model converges faster than the nonparametric models. Also, in settings with a large number of locations, working with the prior covariance structure in the SDP mixture kernel and location-dependent Gaussian kernel can be computationally demanding.

3 Simulated Data Example

To demonstrate how the SDP mixture model performs in the IDE setting, we consider a simulated data set (shown in Figure 3). Data following an IDE model are simulated at 300 locations and 30 time points. The kernel distribution is normal (with mean 0 and variance 100) for locations 1 to 100, asymmetric Laplace (with mean 5, variance 61, and skewness parameter 2/3) for locations 101 to 200, and stable (with location parameter 0, scale parameter 3, skewness parameter 0.5, and stability parameter 1.3) for the last 100 locations. The true IDE kernel densities corresponding to the three regions are included in Figure 4.

The SDP mixture IDE model is fitted to the data. To facilitate estimation and interpretation, a smooth Gaussian process is assumed for the discrete process convolution on the atoms. A

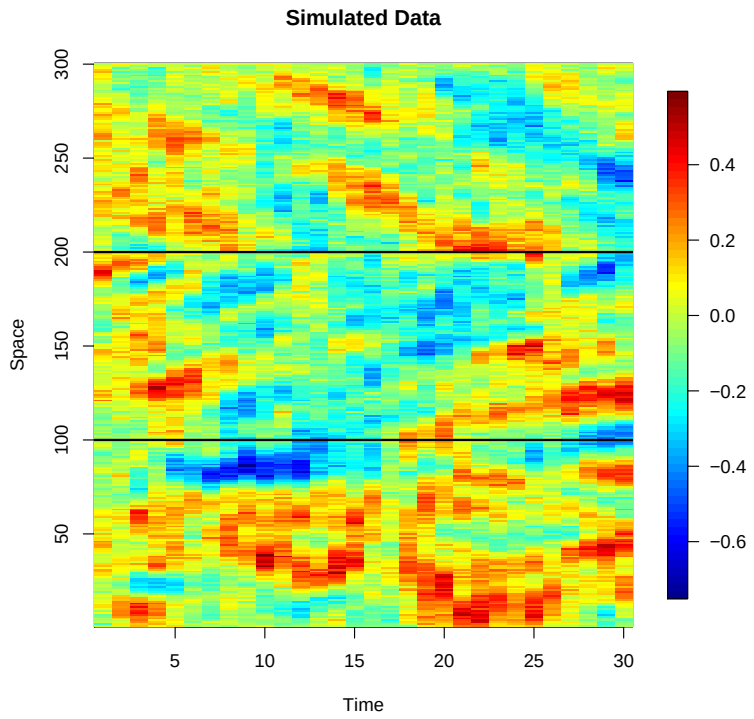

Figure 3: Data simulated from a location-dependent kernel. The three partitioned areas each use different kernels; refer to Section 3 for details.

Matern convolution kernel is used with smoothness parameter 2.5 and an effective range of 40, and q is chosen to be smaller than n. The simulation uses 300 data locations, but q is set to 50. This forces a level of smoothness that may not be appropriate in general, but in this application, a smooth transition of kernel shapes from one location to the next is expected. We use relatively diffuse priors for the various model hyperparameters. Empirical evidence shows that the posterior distribution is insensitive to diffuse priors, but when overly dispersed priors are used the convergence is delayed significantly. The hyperparameters μ_ζ and σ^2_ζ are given $\mathrm{N}(0, .5)$ and $\mathrm{IG}(4, 3)$ priors, respectively. The prior for σ^2_0 is a standard exponential distribution, and an $\mathrm{IG}(3, 2)$ prior is assigned to both σ^2 and τ^2. The matrix V is a Matern correlation matrix with smoothness parameter 1.5 and an effective range of 0.5. Twenty Hermite functions are used for the basis, and the locations are scaled to lie between -5 and 5. The number of basis functions is chosen to be as small as possible to control computation time while still large

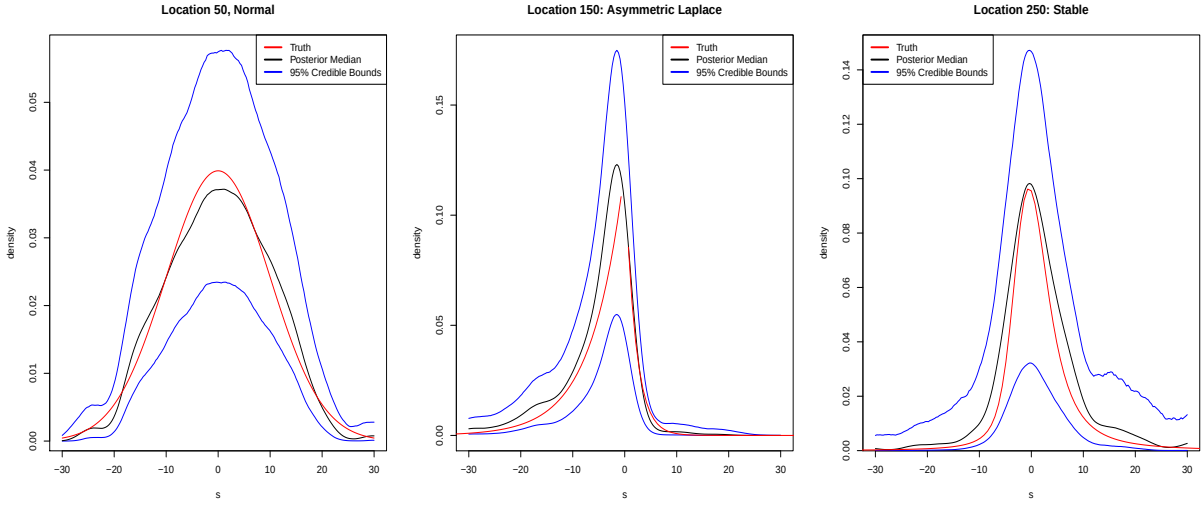

Figure 4: Simulated data. Posterior point and 95% interval estimates for the IDE kernel at the midpoints of the three regions.

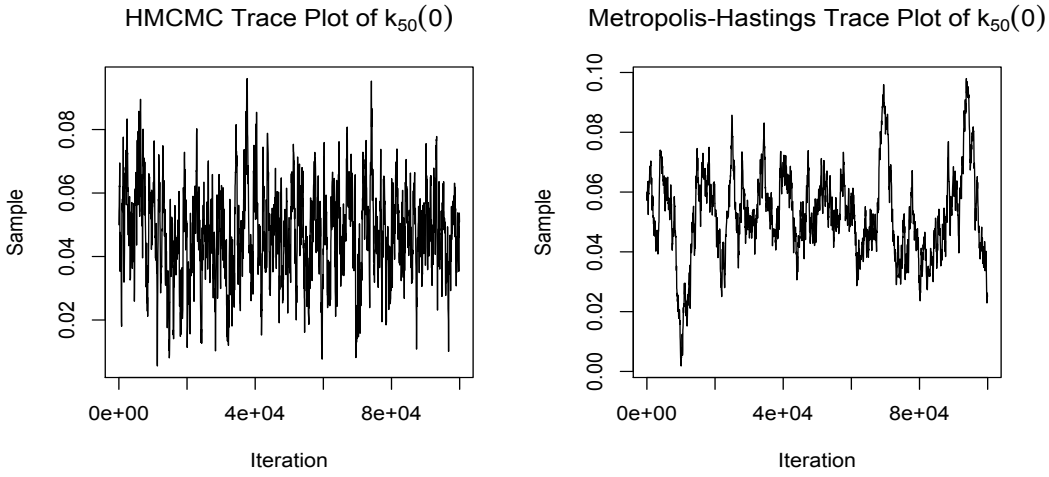

Figure 5: Simulated data. Trace plots for the IDE kernel density at location 50 evaluated at 0, using HMCMC (left panel) and Metropolis-Hastings (right panel) for the SDP atoms.

enough to support an accurate approximation to the IDE kernel. To simplify computation, the SDP precision parameter α is fixed at 2.5. Finally, the truncation level is set to $L = 30$, which, for $\alpha = 2.5$, yields $\mathrm{E}(\sum_{l=1}^{30} w_l) = 0.99996$ (see Section 2.1). The same values for α and L were used for the DP and SDP mixture models in the data analyses of Section 4.

Posterior estimates for the IDE kernel at the midpoints of the three regions (locations 50,

150, and 250) are shown in Figure 4. The model recovers successfully the three distinct shapes of the underlying IDE kernel. The computational methods employed have allowed accurate estimation of the SDP mixture kernel IDE model. Figure 5 demonstrates the advantage of using HMCMC on the SDP atoms. The model converges when using HMCMC after 10,000 iterations. Using Metropolis-Hastings, it is not clear if it has converged at all through 100,000 iterations, and if it has, the chain shows a strong autocorrelation. The Gelman and Rubin test statistic (Gelman and Rubin, 1992) using HMCMC has a 95% confidence interval of $[1.01, 1.05]$, while using standard Metropolis Hastings it has a 95% confidence interval of $[1.09, 1.21]$, which also suggests that HMCMC improves convergence. All the samplers were programmed in C++ and leveraged the threaded OpenBLAS and LAPACK routines for matrix computations on a Linux workstation with 128 GB of RAM and two Intel Xeon CPU E5-2690 v3 @ 2.60GHz processors. In this example, the system time to obtain 100,000 posterior samples was 8 hours.

4 Ozone Data Analysis

To illustrate the potential of the IDE model with DP and SDP mixture kernels, we analyze a data set of ozone pressure and compare the model fits. Specifically, we will show that the SDP mixture kernel IDE model performs significantly better in prediction than any of the parametric kernels previously studied, including a location-dependent normal kernel.

The data are collected at the Koldewey Station near the North Pole. From this station, balloons are released into the air to collect ozone pressure in millipascals (mPa) at somewhat regular intervals. Due to the randomness of how the balloon rises in the air, the locations of these measurements change for each time point, so there is not a specific grid where the data lies. We restrict our attention to lower atmosphere ozone for 10 years from October 1996 to October 2006. The measurements for the analysis are biweekly with a total of 260 time points. The data are indexed by height, a one-dimensional location. Ozone pressure is displayed in Figure 6 by altitude and time. This data set has been analyzed in Richardson et al. (2017), using IDEs with parametric kernels.

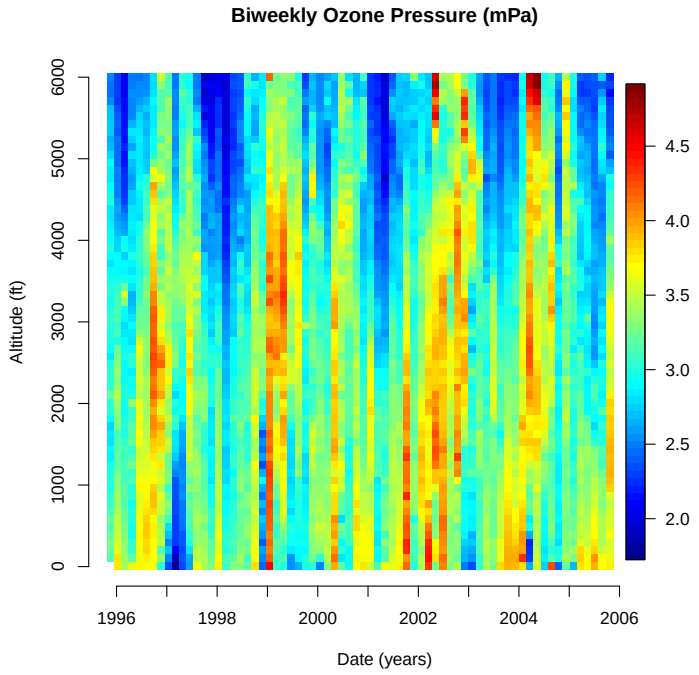

Figure 6: Biweekly ozone pressure measured on a vertical profile, plotted across altitude (0 to 6,000 feet) and over time (October 1996 to October 2006).

Two harmonics are added to the IDE model to account for seasonality, $\boldsymbol{Z}_{ti} = (Z_{ti}^{(1)}, Z_{ti}^{(2)})$, for $i = 1, 2$. These variables evolve through a rotation matrix with frequency z_i, resulting in a process that has a cyclical forecast function with period of $2\pi/z_i$ (West and Harrison, 1997, Chp. 8). By including two harmonics we can account for seasonal variability with two different periods. We denote ozone measurements for time t and location s by $Y_t(s)$ and the data vector at time t by $\boldsymbol{Y}_t = (Y_t(s_{t,1}), ..., Y_t(s_{t,n_t}))$. The model for a general kernel with parameter set $\boldsymbol{\theta}$ is

$$
\begin{aligned}
\boldsymbol{Y}_t \mid \boldsymbol{a}_t, \sigma^2 &\sim \mathrm{N}(\boldsymbol{\Psi}_t \boldsymbol{a}_t + Z_{t1}^{(1)} + Z_{t2}^{(1)}, \sigma^2 \boldsymbol{I}), \quad t = 1, ..., T \\
\boldsymbol{a}_t \mid \boldsymbol{a}_{t-1}, \tau^2, \boldsymbol{\theta} &\sim \mathrm{N}(\boldsymbol{G}\boldsymbol{B}_\theta \boldsymbol{a}_{t-1}, \tau^2 \boldsymbol{G}\boldsymbol{V}\boldsymbol{G}'), \\
\begin{pmatrix} Z_{ti}^{(1)} \\ Z_{ti}^{(2)} \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} \cos(z_i) & \sin(z_i) \\ -\sin(z_i) & \cos(z_i) \end{pmatrix} \begin{pmatrix} Z_{t-1,i}^{(1)} \\ Z_{t-1,i}^{(2)} \end{pmatrix}, \boldsymbol{V}_Z \right), \quad i = 1, 2 \\
\sigma^2, \tau^2, \boldsymbol{V}_Z &\sim p(\sigma^2) p(\tau^2) p(\boldsymbol{V}_Z) \\
\boldsymbol{\theta} \mid \gamma &\sim p(\boldsymbol{\theta} \mid \gamma), \qquad \gamma \sim p(\gamma).
\end{aligned}
$$

The matrices $\boldsymbol{\Psi}_t$, $\boldsymbol{B}_\theta$, and $\boldsymbol{G}$ are derived from the basis function choice and the kernel choice, as described in Section 2.3. The matrix $\boldsymbol{V}_Z$ is a fixed covariance matrix. To perform conditionally linear filtering for this model with the seasonal variables, we augment the state vector to $(\boldsymbol{a}_t', Z_{t1}', Z_{t2}')'$ and augment the process level evolution matrix as a block diagonal. The model parameters σ^2, τ^2, and $\boldsymbol{V}_Z$ are treated similarly for each model. The parameters σ^2 and τ^2 are given IG$(3, 3)$ priors. The matrix $\boldsymbol{V}_Z$ is given an inverse Wishart prior $(IW(10, 10I))$, which results in a conjugate posterior distribution conditional on the state vector. Inference results under the SDP mixture kernel model are affected by overly dispersed priors, but are insensitive to a wide variety of informative priors. For the filtering, priors must be defined for $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$, the mean vector and covariance matrix of the state vector $\boldsymbol{a}_0$. The model may be sensitive to the specification of $\boldsymbol{m}_0$, so some care must be taken to inform a prior for the time 0 process which lies relatively near the data. For this analysis, $\boldsymbol{m}_0$ is specified to give the prior mean of the time 0 process a constant value of 3, because ozone pressure levels will typically vary between 2 and 4. The covariance matrix $\boldsymbol{C}_0$ is specified as $4\boldsymbol{I}$, which is considerably diffuse for Hermite basis coefficients.

We will score the one-step ahead posterior predictive distributions using energy scores (Gneiting et al., 2008), defined as

$$\hat{es}(F, y) = \frac{1}{m}\sum_{i=1}^{m} ||\boldsymbol{y}^{(i)} - \boldsymbol{y}|| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} ||\boldsymbol{y}^{(i)} - \boldsymbol{y}^{(j)}||, \tag{12}$$

where $\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(m)}$ are samples from F, the posterior predictive distribution of the particular model, and $\boldsymbol{y}$ denotes the data vector. Energy scores are calculated for each time point, resulting in 260 energy scores for each model.

4.1 Dirichlet Process Mixture Kernel

Richardson et al. (2017) compared three IDE models for this data set, using the normal, asymmetric Laplace, and stable distribution for the kernel, in all cases with parameters that do not change with location. It was found that the stable kernel performed better in prediction and

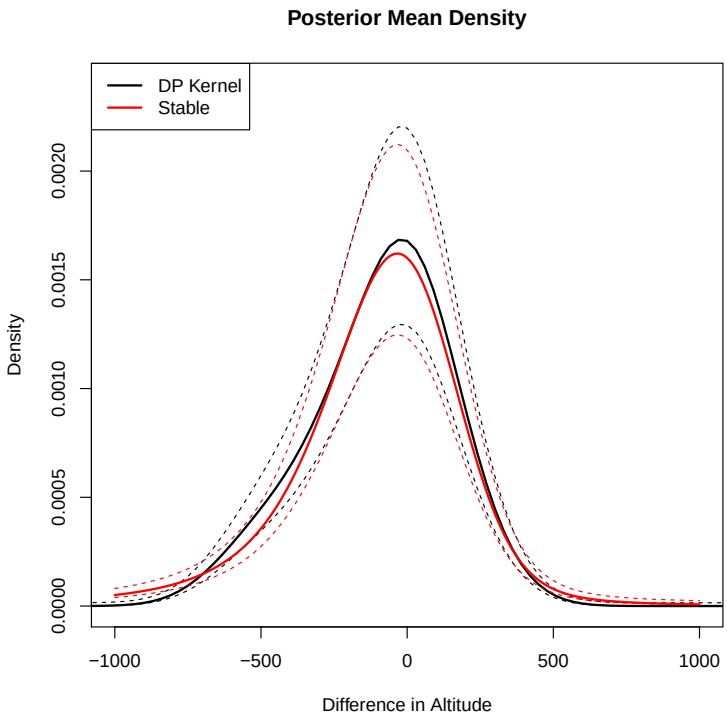

Figure 7: Ozone data. The posterior mean and 95% interval estimates for the IDE kernel under the stable distribution and DP normal mixture models.

scoring, in particular, the IDE kernel was estimated to be left skewed.

The DP mixture model for the kernel can represent heavy tailed and skewed distributions, but it can go beyond the stable family in representing a variety of other features as well. Thus, for comparison purposes, we consider an IDE model with a DP mixture kernel that does not vary with location. The DP centering distribution generating the atoms μ_l is taken to be $\mathrm{N}(\mu_\zeta, \sigma_\mu^2)$, and we place $\mathrm{N}(0, 100^2)$ and $\mathrm{IG}(2.5, 300)$ hyperpriors on μ_ζ and σ_μ^2, respectively. Again, the model is insensitive to a wide array of informative priors, but using priors which are too diffuse can delay convergence of the posterior simulation algorithm.

Figure 7 shows the posterior estimates of the IDE kernel under the models based on the stable distribution and the DP normal mixture. Interestingly, both point and interval estimates are similar, and thus the more general nonparametric IDE kernel model confirms the earlier findings under parametric models. The main difference is the thickness of the left tail. The DP mixture

kernel model scored lower than the stable distribution model 64% of all time points, suggesting that there is an advantage to using the DP mixture, but in this case, it is perhaps not a significant advantage considering the extra required parameters. Of course, if the underlying process is non-stationary with respect to location, then these results will be affected. To study this, in the next section, we extend the analysis to IDE kernels with location-dependent parameters.

4.2 SDP Mixture Kernel

Including all models considered, both with and without location-dependent parameters for the kernel, 6 different IDE models are applied to the ozone pressure data: normal, asymmetric Laplace, stable, DP mixture of normals, location-dependent normal, and SDP mixture of normals. The size of the kernel parameter set, $\boldsymbol{\theta}$, for this analysis varies from 2 for the normal to over 1,000 for the SDP mixture. For the location-dependent normal kernel and the SDP mixture kernel, the locations are scaled to -6.5 and 6.5 and 30 Hermite functions are used for the basis, but for convenience of interpretation, the priors are presented in terms of the original scale. The posterior distribution of the kernel parameters is robust to many different prior choices. The priors used for the analysis shown here are chosen to be somewhat diffuse. The prior mean μ_ζ has a $\mathrm{N}(0, 50^2)$ hyerprior and the variance of the SDP mixture kernel, σ_0^2, has a $\mathrm{Gamma}(10, .1)$ prior. An $\mathrm{IG}(3, 100)$ prior is placed on σ_ζ^2, an $\mathrm{IG}(3, 4)$ prior is used for σ_η^2, and a $\mathrm{N}(5, 1)$ prior is used for μ_η.

After careful tuning of the HMCMC and Metropolis-Hastings steps, 20,000 samples were collected from the posterior distribution of each model.

For the more computationally demanding SDP mixture model, the system time was approximately 12 hours. Convergence was assessed using methods found in the "boa" package in R (Smith, 2007), including the Gelman and Rubin test statistic (Gelman and Rubin, 1992) and the Geweke test statistic (Geweke et al., 1991), and found the results to be acceptable.

The posterior mean estimates of the kernel density for the SDP mixture model change across location, as shown in Figure 8. The kernel shifts skewness from left to right throughout the range of the data. The kernels in higher altitudes have, in general, heavier tails than the kernels

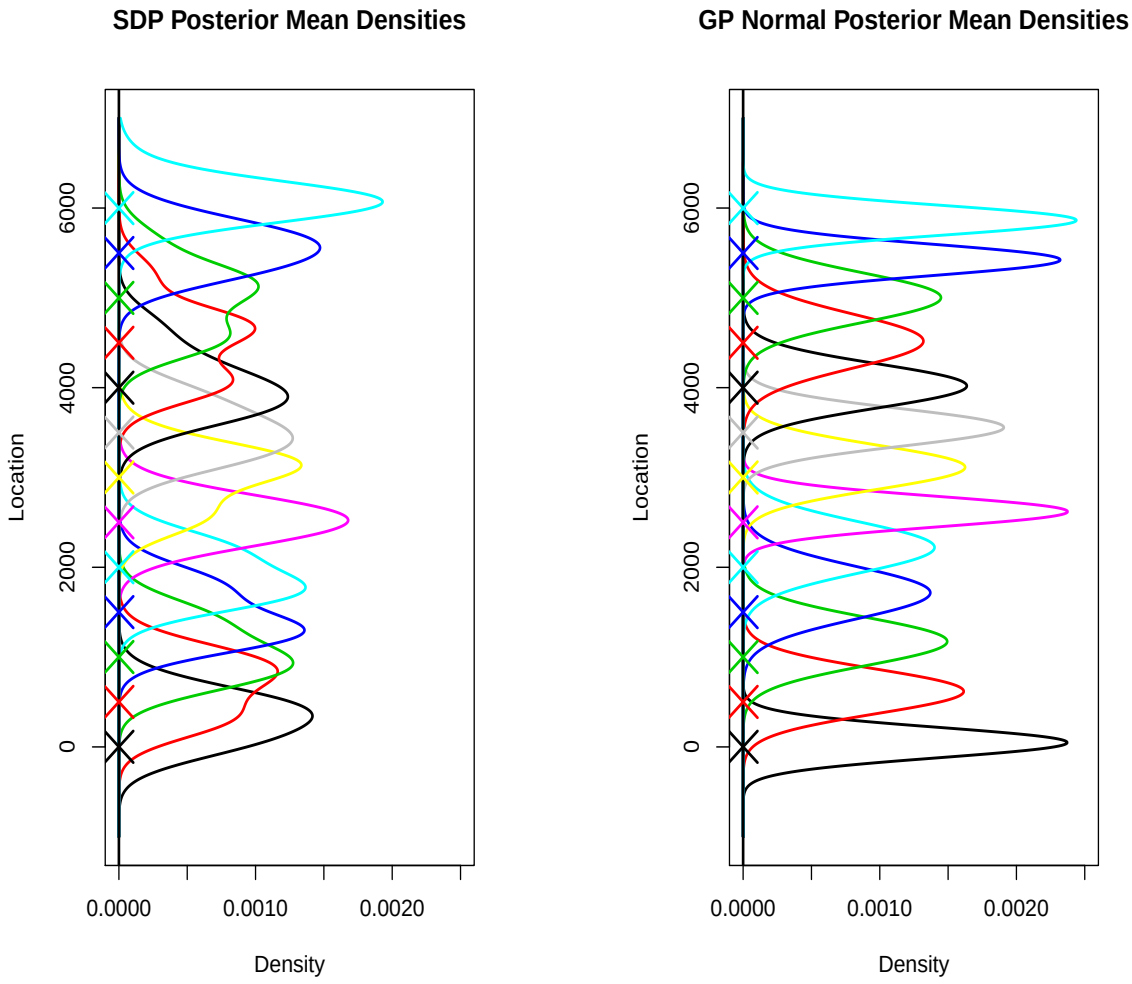

Figure 8: Ozone data. Posterior mean estimates of the kernel densities under the spatial DP mixture model (left) and the location-dependent Gaussian model (right). The "X" on the y-axis show the location associated with the kernel of the matching color.

in lower altitudes. There is also suggestion of bimodalilty at some locations. While bimodality may be difficult to interpret in a physical sense, it is a feature which would be impossible to recreate using a less flexible kernel. Figure 8 includes also the estimates at the same locations under the location-dependent normal kernel, and for further comparison, Figure 9 shows how the kernel expected value and variance vary across different locations. There is a clear association, although the variance under the SDP mixture is consistently larger than the location-dependent normal. One-step ahead prediction profiles for 3 different time points are shown for all 6 models

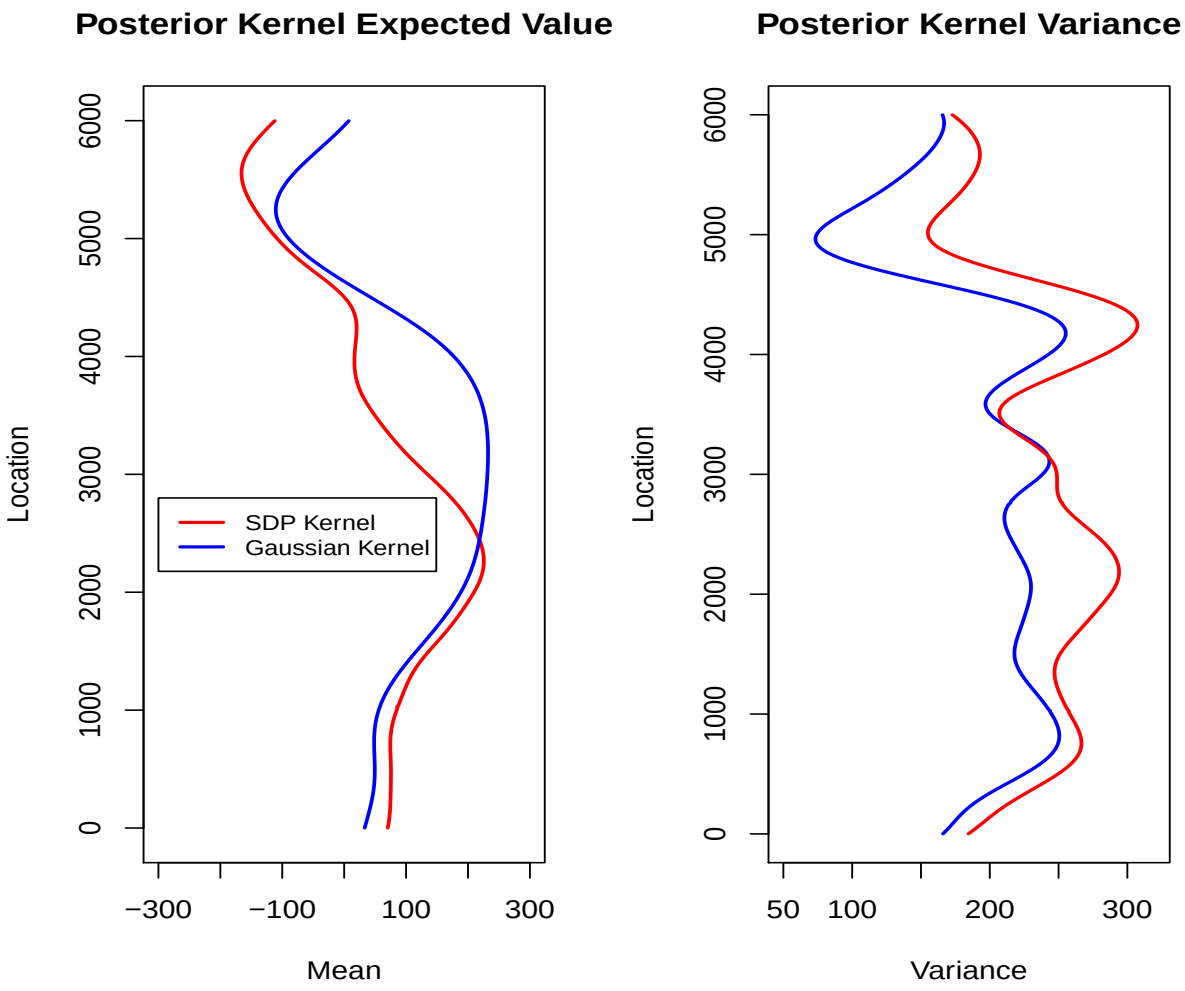

Figure 9: Ozone data. Posterior mean estimates for the expected value and variance of the SDP mixture kernel and location-dependent Gaussian kernel across the locations of the data.

in Figure 10. We note that the model predictive power improves with increasing flexibility of the kernel.

For each time point we calculate the energy scores from Equation (12) and compare. For 222 of the 260 time points, the SDP mixture of normals kernel IDE model has the lowest energy score of all the models. The location-dependent normal kernel IDE model has the lowest score for 20 of the remaining time points. The stable and the DP mixture kernels scored lowest 9 times each, whereas the stationary normal and asymmetric Laplace never did. The stable and

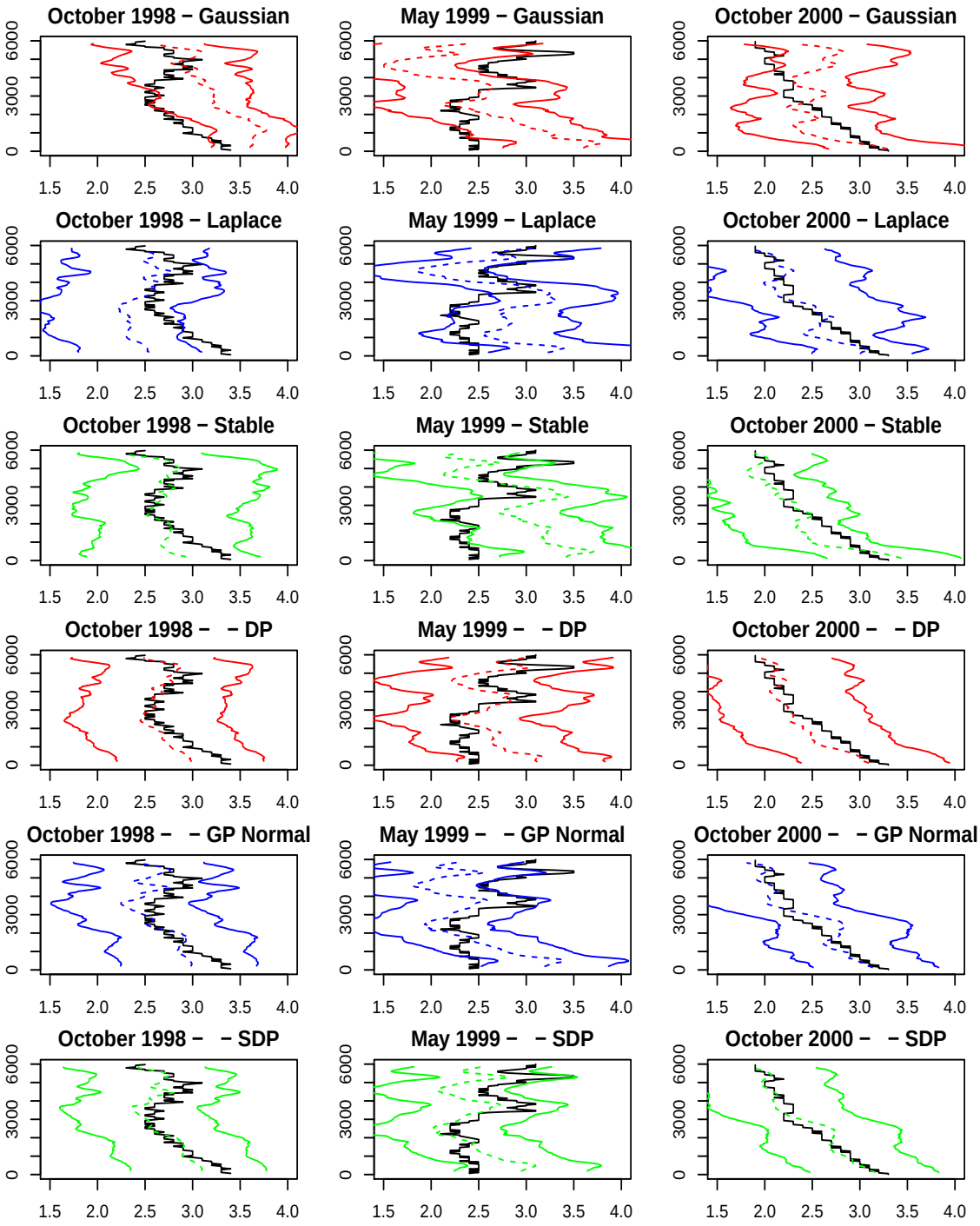

Figure 10: Ozone data. Profiles for one-step ahead predictions for 3 time points under each of the 6 models. For all the plots the x-axis is ozone pressure estimate (in millipascals) and the y-axis is altitude (in feet).

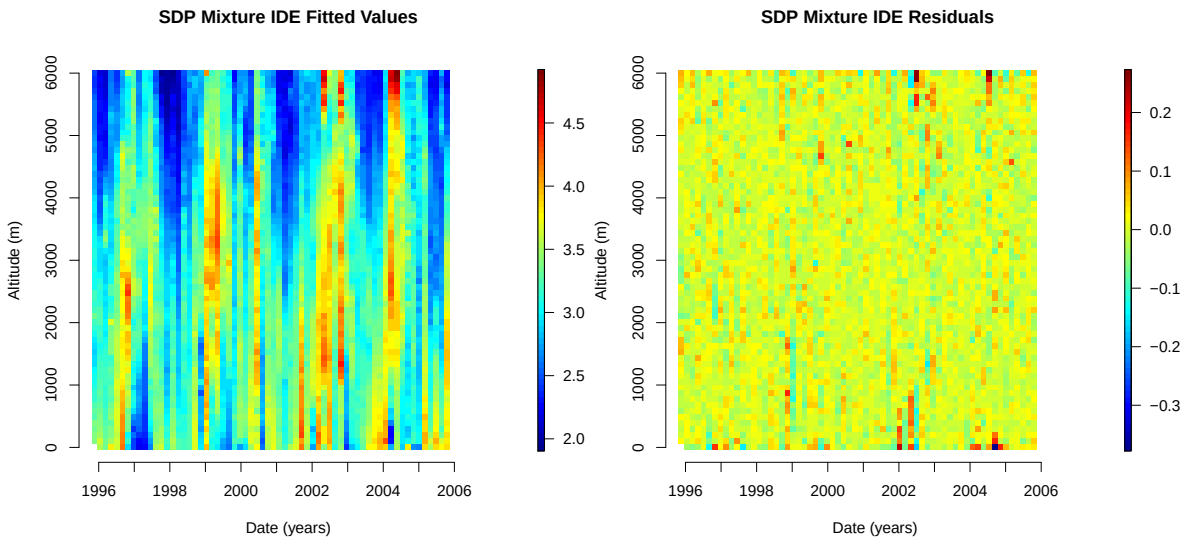

Figure 11: Ozone data. The left panel shows the fitted values for the SDP mixture kernel IDE model. The right panel shows residuals from the same model.

DP mixture models did have lower scores than the location-dependent normal kernel 12% of the time. From the profiles and the scoring procedures, it is clear that using location-dependent parameters is advantageous despite the difficulty of learning the complicated models. Also, the SDP mixture model clearly performs the best. Figure 11 shows the fitted values and residuals for the SDP mixture model.

The frequencies of the harmonics were chosen by comparing model fits when using the parametric kernels, but the non-stationary models or the DP mixture kernel models may require different harmonics or more of them. By using harmonics, the resulting forecast function includes a cyclical sinusoidal element. The amplitude and phase of this forecast function for the SDP kernel IDE model can show how the harmonics affect the model. According to West and Harrison (1997), we find the amplitude from the state variables as $\sum_{i=1}^{2} \sqrt{\sum_{j=1}^{2} Z_{ti}^{(j)2}}$. The amplitude of the first harmonic averages 0.108 for all time points and decreases slightly from 1996 to 2006. The second harmonic averages 0.76 and increases slightly over the time span. The values for the phase shift are $\arctan(-Z_{ti}^{(2)}/Z_{ti}^{(1)})$. The posterior means for the phase randomly

vary about 0.

5 Conclusion

We have explored the full potential of IDE models by using Bayesian nonparametric priors for the IDE kernel density. For applications involving one-dimensional space, the SDP mixture of normals kernel is able to capture a variety of spatio-temporal effects that can not be recovered using other kernels. For example, in the ozone example, we see that the SDP based kernel has a skewness that is strongly dependent on location. Also, in certain regions the kernel can have heavier tails than in others. In fact, the proposed model has the ability to capture complicated location-dependent tail behavior.

Scoring procedures and profile plots show that, for prediction, non-stationary IDE models perform better than stationary ones, with the SDP mixture kernel model outperforming the model built on location-dependent normal kernels. Based on careful consideration of the properties of different basis functions, we develop a model that uses Hermite basis functions. This appears to be a good basis choice for a mixture of normals, providing high approximation accuracy with a reduced number of functions. In order to explore the posterior distribution of the parameters that define the kernels efficiently, we propose to sample the atoms of the SDP mixture in blocks, using HMCMC updates.

In principle, the extension to two dimensions is straightforward, but the computational challenges make its use impractical. The main issue is that the number of basis functions that need to be used in two dimensions increases substantially. Also, draws from the posterior distribution require MCMC methods that search in two dimensions instead of one, complicating convergence and mixing. A simplification is to consider a kernel that is not spatially varying, using a DP mixture kernel IDE model: $k(\boldsymbol{u} \mid \boldsymbol{s}, G, \boldsymbol{\Sigma}_0) = \sum_{l=1}^{L} w_l \, \phi_2(\boldsymbol{u} \mid \boldsymbol{s} + \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_0)$, where ϕ_2 denotes the bivariate normal density, and the covariance matrix $\boldsymbol{\Sigma}_0$ is diagonal. In the context of IDE models for stationary processes, this choice offers more flexibility than the Gaussian kernel, and such a model can be estimated more consistently than the more general SDP mixture model.

Simulations have shown that this simplified model can capture the general shape of non-Gaussian IDE kernels that do not change with location.

Current work explores the bivariate stable distribution (Nolan, 2003) as a flexible, two-dimensional space alternative to the Gaussian kernel. The bivariate stable is governed by an unknown measure which controls characteristics such as the orientation, shape, and spread of the distribution. By placing a flexible prior on this measure, the result is a semiparametric kernel which can support a much wider variety of shapes than the normal density, while keeping at check the computational requirements of the estimation methods.

Acknowledgments

This research is part of the first author's Ph.D. dissertation completed at University of California, Santa Cruz. A. Kottas was supported in part by the National Science Foundation under award DMS 1310438. B. Sansó was supported in part by the National Science Foundation under award DMS 1513076. The authors wish to thank an Associate Editor and two reviewers for constructive feedback and for comments that improved the presentation of the material in the paper.

Appendix A Stationarity of IDE Processes

Brown et al. (2000) show that IDE models are stationary in space when the IDE kernel has parameters which are not spatially varying. The following lemma handles the more general case with spatially varying IDE kernels.

Lemma 1. *Consider an IDE process, $X_t(s)$, with a stationary initial process $X_0(s)$, and with a kernel that belongs to a location family of distributions. The covariance function of the process is non-stationary with respect to location for all $t > 0$, when the kernel parameters depend on the location s.*

Proof. Assume $\text{Cov}[X_{t-1}(s), X_{t-1}(r)] = \rho(|s - r|)$, a stationary covariance function, and that

the error process $\omega_t(s)$ is also stationary with covariance function $\gamma(|s-r|)$. Then,

$$\begin{aligned}\text{Cov}[X_t(s), X_t(s+r)] &= \text{Cov}\left[\int k(u \mid s, \boldsymbol{\theta}_s) X_{t-1}(u)\mathrm{d}u + \omega_t(s), \int k(v \mid s+r, \boldsymbol{\theta}_{s+r}) X_{t-1}(v)\mathrm{d}v + \omega_t(s+r)\right] \\ &= \int\int k(u \mid s, \boldsymbol{\theta}_s) k(v \mid s+r, \boldsymbol{\theta}_{s+r}) \rho(|u-v|)\mathrm{d}u\mathrm{d}v + \gamma(r).\end{aligned}$$

Using the transformations $\eta = u - v$ and $w = v - s$, and the assumption of a kernel that belongs to a location family, i.e., $k(u \mid s, \boldsymbol{\theta}_s) = k(u - s \mid \boldsymbol{\theta}_s)$, the covariance function can be written as

$$\text{Cov}[X_t(s), X_t(s+r)] = \int\int k(\eta + w \mid \boldsymbol{\theta}_s) k(w - r \mid \boldsymbol{\theta}_{s+r}) \rho(|\eta|)\mathrm{d}\eta\mathrm{d}w + \gamma(r).$$

The covariance is a function of s and r, which implies non-stationarity in space. □

When the parameter vector does not depend on the location, i.e., $\boldsymbol{\theta}_s = \boldsymbol{\theta}$, the location s disappears from the covariance. This is consistent with previous work which shows that Gaussian kernel IDE models are stationary when the parameters do not change with location.

Appendix B Hamiltonian Markov chain Monte Carlo

Hamiltonian Markov chain Monte Carlo (HMCMC) involves taking the derivative with respect to unknown parameters of the negative log of the target function, which in this case is the posterior (Neal, 2011). Letting $\boldsymbol{W} = \tau^2 \boldsymbol{G}\boldsymbol{V}\boldsymbol{G}$, the negative log of the relevant parts of the posterior is

$$-l(\boldsymbol{\theta}) = -\log(p(\boldsymbol{\theta})) + \frac{1}{2}\sum_{t=1}^{T} \left(\boldsymbol{a}_t - \boldsymbol{G}'\boldsymbol{B}_\theta \boldsymbol{a}_{t-1}\right)' \boldsymbol{W}^{-1} \left(\boldsymbol{a}_t - \boldsymbol{G}'\boldsymbol{B}_\theta \boldsymbol{a}_{t-1}\right),$$

which can be expanded out as

$$-l(\boldsymbol{\theta}) = -\log(p(\boldsymbol{\theta})) + \frac{1}{2}\sum_{t=1}^{T} \left(\boldsymbol{a}_t'\boldsymbol{W}^{-1}\boldsymbol{a}_t - 2\boldsymbol{a}_t\boldsymbol{W}^{-1}\boldsymbol{G}'\boldsymbol{B}_\theta\boldsymbol{a}_{t-1} + \boldsymbol{a}_{t-1}'\boldsymbol{B}_\theta'\boldsymbol{G}\boldsymbol{W}^{-1}\boldsymbol{G}'\boldsymbol{B}_\theta\boldsymbol{a}_{t-1}\right).$$

Using matrix calculus, the derivative is

$$\frac{\partial(-l(\boldsymbol{\theta}))}{\partial\theta_i} = -\frac{\partial(-\log(p(\boldsymbol{\theta})))}{\partial\theta_i} + \frac{1}{2}\sum_{t=1}^{T} -2\boldsymbol{a}_t\boldsymbol{W}^{-1}\boldsymbol{G}'\frac{\partial\boldsymbol{B}_\theta}{\partial\theta_i}\boldsymbol{a}_{t-1} + 2\,tr\left(\boldsymbol{a}_{t-1}\boldsymbol{a}'_{t-1}\boldsymbol{B}_\theta\boldsymbol{G}\boldsymbol{W}^{-1}\boldsymbol{G}'\frac{\partial\boldsymbol{B}_\theta}{\partial\theta_i}\right),$$

where $\frac{\partial\boldsymbol{B}_\theta}{\partial\theta_i}$ is a element-wise derivative of $\boldsymbol{B}_\theta$ with respect to θ_i, and θ_i is the i-th element of the parameter vector, $\boldsymbol{\theta}$.

Let $E(\boldsymbol{\theta})$ refer to the negative log posterior. A step size ϵ and number of iterations, L, must be defined prior to the algorithm. Latent variables, p_i, are introduced for each parameter as independent normal variables with zero mean and variance M_i. Then, one "leapfrog" step given current iteration $(\boldsymbol{\theta}^b, \mathbf{p}^b)$ is

$$\begin{aligned}
\mathbf{p}^{(b+\epsilon/2)} &= \mathbf{p}^b - \frac{\epsilon}{2}\frac{\partial E}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^{(b)}) \\
\boldsymbol{\theta}^{(b+\epsilon)} &= \theta^b + \epsilon\frac{\mathbf{p}^{(b+\epsilon/2)}}{\mathbf{m}} \\
\mathbf{p}^{(b+\epsilon)} &= \mathbf{p}^{(b+\epsilon/2)} - \frac{\epsilon}{2}\frac{\partial E}{\partial\boldsymbol{\theta}}(\boldsymbol{\theta}^{(b+\epsilon)})
\end{aligned}$$

The parameters leapfrog L times ending at new proposals for the posterior. The function $H(\boldsymbol{\theta}, p)$ is defined to be the sum of $E(\boldsymbol{\theta})$ and $K(p) = \frac{1}{2}\sum\frac{p_i^2}{M_i}$. The new values $(\boldsymbol{\theta}^{(b+1)}, \mathbf{p}^{(b+1)})$ are accepted with probability $\min(1, \exp\left(H(\boldsymbol{\theta}^{(b+1)}, \mathbf{p}^{(b+1)}) - H(\boldsymbol{\theta}^b, \mathbf{p}^b)\right)$. If the new value is rejected, it is set to the previous values. The method must be tuned to accept and reject at reasonable rates, perhaps accepting between 40 and 60% of proposed samples. Both L and ϵ can be tuned, where $L \times \epsilon$ is closely associated with acceptance rates.

For a normal distribution, the derivative $\frac{\partial B_\theta}{\partial\theta_i}$ is found by taking element-wise derivatives of the coefficients found in equation (6). The required derivative with respect to the mean parameter using Hermite basis functions is

$$\frac{\partial b_n}{\partial\mu} = \frac{1}{\sigma^2}\frac{1}{\sqrt{(\sqrt{\pi}2^n n!)\,(1+\sigma^2)}}\exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right)\sum_{k=0}^{n} H_{n,k}\left(\mu m_k - m_{k+1}\right).$$

Again, $H_{n,k}$ is the k-th coefficient in the n-th Hermite polynomial and m_k is the k-th raw moment of a normal distribution with mean $\mu/(\sigma^2+1)$ and variance $\sigma^2/(\sigma^2+1)$. In terms of the basis coefficients the derivative can be written as

$$\frac{\partial b_n}{\partial \mu} = \frac{1}{\sigma^2}\left(\mu b_n - b_{n+1} + \frac{1}{\sqrt{(\sqrt{\pi}2^n n!)\,(1+\sigma^2)}}\exp\left(-\frac{\mu^2}{2(1+\sigma^2)}\right)\right). \tag{13}$$

Mixtures of normals result in more complex calculations than the normal, but the basis coefficients of the mixture as a whole are given through the sum of the individual basis coefficients. The derivative of the basis coefficients needed for the HMCMC is the weighted sum of the derivatives of the basis coefficients of the normal components.

References

Brown, P. E., Roberts, G. O., Kåresen, K. F., and Tonellato, S. (2000), "Blur-generated non-separable space–time models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62, 847–860.

Chen, T., Fox, E. B., and Guestrin, C. (2014), "Stochastic Gradient Hamiltonian Monte Carlo." in *ICML*, pp. 1683–1691.

Cressie, N. (1993), *Statistics for Spatial Data*, New York: John Wiley & Sons.

Cressie, N. and Huang, H.-C. (1999), "Classes of nonseparable, spatio-temporal stationary covariance functions," *Journal of the American Statistical Association*, 94, 1330–1339.

Cressie, N. and Wikle, C. K. (2011), *Statistics for spatio-temporal data*, New York: John Wiley & Sons.

Frühwirth-Schnatter, S. (1994), "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, 15, 183–202.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian nonparametric spatial modeling with Dirichlet process mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical science*, 7, 457–472.

Geweke, J. et al. (1991), *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, vol. 196, Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Girolami, M. and Calderhead, B. (2011), "Riemann manifold langevin and hamiltonian monte carlo methods," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 123–214.

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008), "Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds," *Test*, 17, 211–235.

Heine, V. (1955), "Models for two-dimensional stationary stochastic processes," *Biometrika*, 42, 170–178.

Higdon, D. (1998), "A process-convolution approach to modelling temperatures in the North Atlantic Ocean," *Environmental and Ecological Statistics*, 5, 173–190.

Hooten, M. B. and Wikle, C. K. (2008), "A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the Eurasian Collared-Dove," *Environmental and Ecological Statistics*, 15, 59–70.

Jones, R. H. and Zhang, Y. (1997), "Models for continuous stationary space-time processes," in *Modelling longitudinal and spatially correlated data*, eds. Gregoire, T. G., Brillinger, D. R., Diggle, P. J., Russek-Cohen, E., Warren, W. G., and Wolfinger, R. D., Springer, pp. 289–298.

Kot, M., Lewis, M. A., and van den Driessche, P. (1996), "Dispersal data and the spread of invading organisms," *Ecology*, 77, 2027–2042.

Kottas, A., Duan, J. A., and Gelfand, A. E. (2008), "Modeling disease incidence data with spatial and spatio temporal Dirichlet process mixtures," *Biometrical Journal*, 50, 29–42.

Ma, C. (2003), "Nonstationary covariance functions that model space–time interactions," *Statistics & Probability Letters*, 61, 411–419.

Neal, R. M. (2011), "MCMC using Hamiltonian dynamics," *Handbook of Markov Chain Monte Carlo*, 2.

Neubert, M. G., Kot, M., and Lewis, M. A. (1995), "Dispersal and pattern formation in a discrete-time predator-prey model," *Theoretical Population Biology*, 48, 7–43.

Nolan, J. (2003), *Stable distributions: models for heavy-tailed data*, New York: Birkhauser.

Olver, F. W. (2010), *NIST handbook of mathematical functions*, Cambridge University Press.

Richardson, R., Kottas, A., and Sansó, B. (2017), "Flexible Integro-Difference Equation Modeling for Spatio-Temporal Data," To appear in *Computational Statistics and Data Analysis*.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Smith, B. J. (2007), "boa: an R package for MCMC output convergence assessment and posterior inference," *Journal of Statistical Software*, 21, 1–37.

Storvik, G., Frigessi, A., and Hirst, D. (2002), "Stationary space-time Gaussian fields and their time autoregressive representation," *Statistical Modelling*, 2, 139–161.

Welling, M. and Teh, Y. W. (2011), "Bayesian learning via stochastic gradient Langevin dynamics," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 681–688.

West, M. and Harrison, J. (1997), *Bayesian Forecasting and Dynamic Models*, New York: Springer Verlag, 2nd ed.

Wikle, C. K. (2002), "A kernel-based spectral model for non-Gaussian spatio-temporal processes," *Statistical Modelling*, 2, 299–314.

Wikle, C. K. and Cressie, N. (1999), "A dimension-reduced approach to space-time Kalman filtering," *Biometrika*, 86, 815–829.

Wikle, C. K. and Holan, S. H. (2011), "Polynomial nonlinear spatio-temporal integro-difference equation models," *Journal of Time Series Analysis*, 32, 339–350.

Wikle, C. K. and Hooten, M. B. (2010), "A general science-based framework for dynamical spatio-temporal models," *Test*, 19, 417–451.

Xu, K., Wikle, C. K., and Fox, N. I. (2005), "A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities," *Journal of the American Statistical Association*, 100, 1133–1144.