Flexible Integro-Difference Equation Modeling for Spatio-Temporal Data

Robert Richardson, Athanasios Kottas and Bruno Sansó *

July 24, 2016

Abstract

The choice of kernel in an integro-difference equation (IDE) approach to model spatio-temporal data is studied. By using approximations to stochastic partial differential equations, it is shown that higher order cumulants and tail behavior of the kernel affect how an IDE process evolves over time. The asymmetric Laplace and the family of stable distributions are presented as alternatives to the Gaussian kernel. The asymmetric Laplace has an extra parameter controlling skewness, whereas the class of stable distributions includes parameters controlling both tail behavior and skewness. Simulations show that failing to account for kernel shape may lead to poor predictions from the model. For an illustration with real data, we consider ozone pressure measurements collected biweekly by radiosonde at varying altitudes. We compare the results obtained with the different kernel families and confirm that better model prediction may be achieved by electing to use a more flexible kernel.

Bayesian environmetrics; Dynamic spatio-temporal models; Integro-Difference Equation; Markov chain Monte Carlo.

*Robert Richardson (richardson@stat.byu.edu) is an Assistant Professor in the Department of Statistics at Brigham Young University, and Athanasios Kottas (thanos@ams.ucsc.edu) and Bruno Sansó (bruno@ams.ucsc.edu) are Professors in the Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA. This research was supported in part by the National Science Foundation under award SES 1024484.

1 Introduction

A spatio-temporal data set refers to data collected across a spatial field and over several time points. Climatological and environmental variables provide several common and abundant examples of data recorded in space and time. In addition to traditional examples of environmental space-time variables, such as temperature or precipitation, there is an increasing ability to store and monitor the dynamics of different types of georeferenced processes. Data for housing costs, crime rates, population growth, soil content, and disease incidence, are some of the many examples of variables that are of interest in areas as diverse as spatial econometrics, epidemiology, and geography, to mention a few.

The field of time series has produced a rich body of literature during at least the last 50 years (Hamilton, 1994; Shumway and Stoffer, 2011). Spatial statistics, despite the seminal work by Matheron (1963), was a fringe area as recently as the early 1990s (Cressie, 1993), but has since received a great deal of attention within the statistical community. Spatio-temporal models stem naturally from these areas, but a systematic treatment of spatio-temporal statistical models has only recently been developed (Cressie and Wikle, 2011). Compared to times series and spatial statistics, the fundamental challenge of spatio-temporal models is to capture the interactions between the spatial and temporal components.

Three general methods are currently used to analyze data from spatio-temporal processes of the form $\{X_t(s) : s \in \mathcal{S}, t \in \mathcal{T}\}$, where s indexes the spatial domain $\mathcal{S}$ and t indexes the time domain $\mathcal{T}$. The first involves an extension of the traditional approach to modeling random fields, focusing on the first and second moment of the process. The goal is to find general families of space-time correlation functions of the form $Cov(X_t(s), X_u(v)) = C(s, v, t, u)$, which are "smooth everywhere" and yet "allow different degrees of smoothness" (Stein, 2005). In this setting, both s and t are considered as continuous indexes. This lends flexibility to the models, but requires dealing with potentially large covariance matrices. This approach can thus have important computational drawbacks when large spatial domains or long time periods are considered.

A second common modeling approach for spatio-temporal data is an extension of deterministic dynamical models that incorporates stochastic components. This leads to stochastic partial differential equation (SPDE) models. For instance, Jones and Zhang (1997) consider the SPDE $\frac{\partial}{\partial t}X_t(s) - \beta\frac{\partial^2}{\partial s^2}X_t(s) + \alpha X_t(s) = \delta_t(s)$, where $\delta_t(s)$ is a zero mean error process. This SPDE is called a diffusion-injection equation and is just one of the various SPDE-based models commonly used for naturally occurring physical processes (Heine, 1955; Zheng and Aukema, 2010).

The third method is to obtain an explicit description of the dynamics of the process by specifying its evolution as a function of the spatial distribution of the process. A dynamic spatio-temporal model can be written as

$$X_t(s) = \mathcal{M}(X_{t-1}(s), s, \boldsymbol{\theta}) + \varepsilon_t(s), \quad t = 1, \ldots, T,$$

where $\mathcal{M}$ represents a specific model configuration, governing the transfer of information from time $t-1$ to time t. Here, $\boldsymbol{\theta}$ is a parameter vector, and $\varepsilon_t(s)$ is a zero mean noise process which may have a spatially dependent covariance structure. In these models, the process evolves as an entire spatial field over a discrete time component. Cressie and Wikle (2011) strongly support this approach, and suggest a "hierarchical dynamical spatio-temporal model" of the form

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{B}_t\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim N(\boldsymbol{0}, \boldsymbol{V}_t), \quad t = 1, \ldots, T \qquad (1) \\ \boldsymbol{X_t} &= \mathcal{M}_t(\boldsymbol{X}_{t-1}, \boldsymbol{\theta}) + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N(\boldsymbol{0}, \boldsymbol{W}_t), \quad t = 1, \ldots, T, \qquad (2) \end{aligned}$$

where $\boldsymbol{Y}_t$ is the vector of data, and $\boldsymbol{X}_t$ is a vector of latent variables representing an underlying process that is linked to $\boldsymbol{Y}_t$ through the incidence matrix $\boldsymbol{B}_t$. Moreover, $\boldsymbol{\varepsilon}_t$ and $\boldsymbol{\omega}_t$ are noise terms with specified covariances $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$, respectively.

A specific case of the model described by equations (1) and (2) is the integro-difference

equation (IDE) spatio-temporal model. We consider IDE models of the form

$$X_t(s) = e^{\lambda} \int k(s-u|\boldsymbol{\theta}) X_{t-1}(u) du + \omega_t(s), \tag{3}$$

where $k(\cdot)$ is a redistribution kernel with parameter vector $\boldsymbol{\theta}$, and $\omega_t(s)$ is an error process which may be spatially colored. This kernel weights the contribution of the process at time $t-1$ to the process at time t at location s. The scaling term λ controls the growth or decay of the process. Typically, the center of the kernel for each location is somewhere near s, resulting in nearby values being weighted more heavily than others. The spatial dependency in the IDE model arises from nearby observations sharing large contributions from many of the same observations of the previous time point. Thus, the spatial and temporal relationships interact with each other as the process evolves, producing a non-separable process. Furthermore, the kernel width affects the smoothness of the resulting process.

Originally used by ecologists studying the growth and spread of species (Kot et al., 1996), integro-difference equations were introduced for general spatio-temporal processes in Wikle and Cressie (1999). In Wikle (2002) the IDE kernel is specified parametrically through a Gaussian distribution with unknown location and scale parameters. The stochastic properties of the process that results from an IDE, such as stationarity and separability, are explored in Brown et al. (2000) and Storvik et al. (2002). An important extension where the mean of the kernel is spatially indexed is presented in Wikle (2002) and Xu et al. (2005).

Overall, the literature is dominated by IDE models based on Gaussian kernels. Though there is some mention of non-Gaussian kernels, it is without exploring the modeling benefits and inferential issues arising from the use of more general kernel families. Spatio-temporal data can have a variety of features that may not be represented well by a Gaussian kernel IDE model. As shown here, these features include dispersion, extra-diffusion, and flexibility in local behavior. In this paper, we focus on the exploration of the properties and the

development of inferential methods to deal with IDE models based on relatively simple non-Gaussian parametric families of kernels. Our main purpose is to show how using non-Gaussian kernels in IDE modeling can add value to spatio-temporal modeling. We will show that, for hierarchical models as in equations (1) and (2), an IDE with a kernel more flexible than the Gaussian can lead to improved model performance and prediction, and capture a wider array of process dynamics. We restrict the scope of this paper to one-dimensional space for ease in computation, but we expect that the same advantages arising from the use of non-Gaussian kernels in one dimension will also emerge when using non-Gaussian kernels in two dimensions.

The rest of the paper is organized as follows. In Section 2 we use two approximations of the IDE to differential equations to theoretically justify the use of more flexible kernels. Section 3 provides modeling techniques for the IDE model and presents two alternatives to the Gaussian kernel. Direct comparison of model fit and prediction is performed for each kernel choice in Section 4, for both real and synthetic data. Concluding remarks are made in Section 5, and the four appendices collect technical details.

2 Theoretical Foundations

This section aims to explain how the choice of kernel affects the process represented by the IDE. It will justify using more flexible kernels and will motivate how to select those kernels. Section 2.1 connects the IDE model to a partial differential equation (PDE) constructed from the cumulants of the kernel. Section 2.2 shows that the IDE model is a solution to a certain system of PDEs which is constructed from the hazard function of the kernel distribution.

2.1 High Order Cumulants PDE Representation

Brown et al. (2000) consider an IDE model where the time increment is infinitesimal. Using Taylor expansions, they show that the solution of the IDE in equation (3), when the kernel

is infinitely divisible and from a location family, satisfies the approximation

$$\frac{\partial X_t(s)}{\partial t} \approx \lambda X_t(s) - \mu \frac{\partial X_t(s)}{\partial s} + \frac{1}{2}\sigma^2 \frac{\partial^2 X_t(s)}{\partial s^2} + B_t(s) \tag{4}$$

where μ and σ^2 are, respectively, the mean and the variance of the kernel, and $B_t(s)$ is Brownian motion. A distribution, $\mathcal{F}$, is infinitely divisible if any random variable $X \sim \mathcal{F}$ can be written as $X = \sum_{i=1}^{n} X_i$, for any n, where X_i are identically distributed random variables (Steutel and Harn, 2003). Intuitively, the effect of an infinitely divisible kernel controlling the evolution of an IDE for one unit of time can be decomposed into the sum of the effects of n IDEs operating on $1/n$ units of time. Thus, infinite divisibility allows a discrete time IDE to be approximated by an SPDE, which is a continuous time model. The model in equation (4) depends on two parameters, μ and σ^2 that control, respectively, the advection and diffusion of the process $X_t(s)$. Thus, the SPDE approximation of the IDE sheds light on how the kernel parameters control the physical properties of the process $X_t(s)$.

Following the framework in Brown et al. (2000), we establish the following result on an SPDE representation for an IDE using cumulants of order higher than two. This is achieved by considering expansions of a Taylor series approximation beyond the first two terms.

Lemma 1. *Consider the IDE model* $X_t(s) = \int k(s-u|\boldsymbol{\theta})X_{t-1}(u)du$, *where the kernel belongs to an infinitely divisible, location family of distributions for which the first* $J+1$ *cumulants,* $\kappa_1, \ldots, \kappa_{J+1}$, *exist. Moreover, assume that* $\frac{\partial^j}{\partial s^j}X_{t-\delta}(s)$ *exists and that* $\left|\frac{\partial^j}{\partial s^j}X_{t-\delta}(s)\right|$ *is bounded above, for any (small)* $\delta > 0$ *and for* $j = 1, ...J+1$. *Then, the solution to the IDE equation can be approximated by the solution of the equation*

$$\frac{\partial X_t(s)}{\partial t} \approx \sum_{j=1}^{J} (-1)^j \frac{1}{j!} \kappa_j \frac{\partial^j X_t(s)}{\partial s^j}. \tag{5}$$

Proof. For an infinitely divisible location kernel $k(s-u|\boldsymbol{\theta})$, we define $k_{\frac{1}{n}}(s-u|\boldsymbol{\theta}_{\frac{1}{n}})$ as an n-fold self convolution, $k_{\frac{1}{n}}(x) * k_{\frac{1}{n}}(x) * ... * k_{\frac{1}{n}}(x) = k(x)$, and $\boldsymbol{\theta}_{\frac{1}{n}}$ as the adjusted parameter

set induced by the self-convolution. Let $\delta = 1/n$. Then, by representing the process at time $t - \delta$ as a Taylor series with J terms, we can write $X_t(s)$ as

$$\int k_\delta(u|\boldsymbol{\theta}_\delta) X_{t-\delta}(s-u)du = \int k_\delta(u|\boldsymbol{\theta}_\delta) \left[\sum_{j=0}^{J} (-1)^j \frac{1}{j!} u^j \frac{\partial^j}{\partial s^j} X_{t-\delta}(s) + R(u) \right] du$$

where $R(u) = \frac{(-u)^{J+1}}{(J+1)!} \frac{\partial^{J+1}}{\partial s^{J+1}} X_{t-\delta}(s^*(u))$, with $s^*(u)$ in a neighborhood of s that depends on u. Given that the derivatives of $X_{t-\delta}(s)$ are bounded above, we have that

$$\int k_\delta(u|\boldsymbol{\theta}_\delta) X_{t-\delta}(s-u)du = X_{t-\delta}(s) + \sum_{j=1}^{J} (-1)^j \frac{1}{j!} E_{k_\delta}\left[u^j\right] \frac{\partial^j}{\partial s^j} X_{t-\delta}(s) + \frac{O\left(E_{k_\delta}\left[u^{J+1}\right]\right)}{(J+1)!}$$

where E_{k_δ} is the expected value with respect to the distribution with density k_δ. Note that the last term in the above expression tends to zero as J becomes large. Whereas in Brown et al. (2000) $J = 2$, we consider a larger J obtaining the following approximation

$$\frac{X_t(s) - X_{t-\delta}(s)}{\delta} \approx \sum_{j=1}^{J} (-1)^j \frac{1}{j!} \delta^{-1} E_{k_\delta}\left[u^j\right] \frac{\partial^j}{\partial s^j} X_{t-\delta}(s)$$

where we have also rearranged terms and divided by δ.

To complete the proof, we need to show that, as $\delta \to 0$, $\delta^{-1} E_{k_\delta}[u^j] \to \kappa_j$, the j-th cumulant of the kernel distribution. Using the additivity property for cumulants of sums of independent random variables, $\kappa'_j = \delta \kappa_j$, where κ'_j is the j-th cumulant for the distribution with density k_δ. The other key result is the relationship between raw moments and cumulants (e.g., Papoulis and Pillai, 2002). In particular, $E_{k_\delta}[u^j] = \kappa'_j + h(\kappa'_1, ..., \kappa'_{j-1})$, where $h(\kappa'_1, ..., \kappa'_{j-1})$ is a polynomial function every term of which is the product of at least two of the κ'_m, $m = 1, ..., j-1$. Hence, $\delta^{-1} E_{k_\delta}[u^j] = \kappa_j + \delta^{-1} h(\delta\kappa_1, ..., \delta\kappa_{j-1})$, and the result is obtained since every term of $\delta^{-1} h(\delta\kappa_1, ..., \delta\kappa_{j-1})$ includes a factor of δ^ℓ with $\ell \geq 1$. □

As previously discussed, the first two cumulants of the kernel control the advection and diffusion of the resulting process. The third cumulant is known to control dispersion, which in

this context allows for extra variability in how the process behaves from one spatial location to the next. Lemma 1 suggests that kernels more flexible than the Gaussian can model more complicated dynamics.

2.2 Hazard Function PDE Representation

An alternative characterization of an IDE in terms of a PDE has been studied by ecologists dealing with the dispersal of organisms after their introduction in a foreign region (Neubert et al., 1995). A simple experiment would consist of a researcher placing a foreign species in the middle of an open field. After a specified period of time, they would return and measure how the plant spread over the field. They would use what is essentially an one time step IDE process to describe the behavior of how the organism spread.

Lemma 2. *Let k, S, and h be, respectively, the kernel density, and the corresponding survival and hazard functions defined as $S(s) = 1 - \int_{-\infty}^{s} k(u)du$ and $h(s) = k(s)/S(s)$. Then, setting the initial condition to $u_0(s) = X_{t-1}(s)$, the system of differential equations*

$$\frac{\partial u_\tau(s)}{\partial \tau} = -\frac{\partial u_\tau(s)}{\partial s} - h(\tau)u_\tau(s) \quad \textit{and} \quad \frac{\partial v_\tau(s)}{\partial \tau} = h(\tau)S(0)u_\tau(s) \tag{6}$$

has the solution

$$u_\tau(s) = X_{t-1}(s-\tau)\frac{S(\tau)}{S(0)} \quad \textit{and} \quad v_\tau(s) = \int X_{t-1}(s-u)k(u)du\ .$$

The proof of Lemma 2 is included in Appendix A. In ecology, the interpretation of $u_\tau(s)$ is that of a latent process representing the path of particulates in motion. The process $v_\tau(s)$ is a measure of the organisms once they have settled. The variable τ is an index of the path of the process in-between time steps. As τ travels from 0 to ∞, the process $X_t(s)$ moves from time t to time $t+1$, and the process $u_\tau(s)$ becomes 0 as all the particulates settle into locations contributing to $v_\tau(s)$. To make this a multi-step process we set the initial value for $u_\tau^{(t)}(s)$

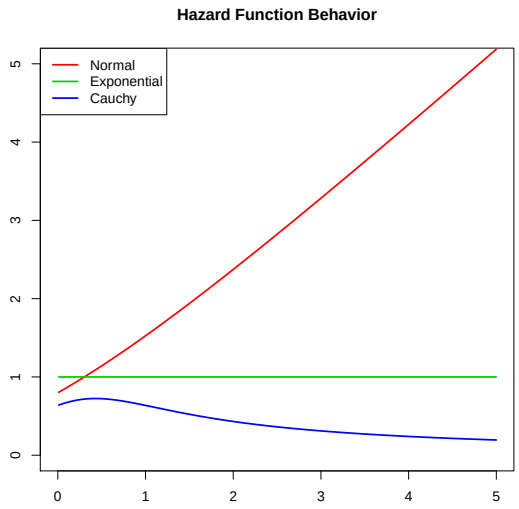

Figure 1: The hazard functions for the standard normal, standard Cauchy, and exponential distributions are shown. The Cauchy has polynomial tails that yield a decreasing hazard function. The normal distribution has a hazard function which is increasing, and the exponential hazard is constant.

equal to $X_{t-1}(s)$ and solve the series of differential equations $\{(u_\tau^{(t)}(s), v_\tau^{(t)}(s)) : t = 1, ..., T\}$ piece by piece.

From the above discussion, we can identify $v_\tau(s)$ with $X_t(s)$, implying that the dynamics of a process that satisfies an IDE with no random shocks, are regulated by the PDE in equation (6). This indicates that the behavior of an IDE process depends on the hazard function associated with the kernel. Tail behavior and hazard functions are directly related. This is illustrated in Figure 1 for three densities with different tails. Thus, we expect that a kernel with thick tails, such as a Cauchy, will produce solutions to the IDE that behave very differently than those that correspond to a Gaussian kernel IDE.

Lemmas 1 and 2 indicate that there is merit in using IDE kernels with more flexibility in high order cumulants and tail behavior than the Gaussian kernel. In line with the results considered in this section, we seek alternative kernels that belong to parametric infinitely divisible, location families of distributions that possess higher order cumulants and/or have more flexible tails than the normal. Two parametric families that offer flexibility along these

lines, without compromising tractability, will be presented in the next section.

3 Modeling Approaches for the IDE Kernel

An attractive characteristic of the IDE model is a dimension-reducing feature which naturally makes the analysis of the problem more computationally feasible (Wikle and Cressie, 1999). Relevant results are summarized in Section 3.1. Sections 3.2 and 3.3 propose the asymmetric Laplace and the stable family of distributions, respectively, as alternatives to the Gaussian kernel. A brief summary of each of these families of distributions is presented, followed in Section 3.4 by prior simulations from IDE models with these kernels. Note that, for all the methods and analyses presented, it is assumed that $\lambda = 0$. While λ models growth or decay of the process, we will use a hierarchical structure that allows for these trends in the first layer, so the extra parameter is unnecessary. Moreover, for the purposes of comparing model fits with different kernels it is sufficient to set λ equal to 0.

3.1 Basis Expansion for Model Fitting

We will use an orthogonal basis expansion for both the kernel and the process, where the basis functions, $\{\phi_1, \phi_2, ...\}$, are common to both. In particular,

$$X_t(s) = \sum_{i=1}^{\infty} \phi_i(s) a_i(t) \qquad \text{and} \qquad k(u - s|\boldsymbol{\theta}) = \sum_{j=1}^{\infty} b_j(s, \boldsymbol{\theta}) \phi_j(u), \tag{7}$$

where $a_i(t)$ are coefficients for the basis expansion of the process, and $b_j(s, \boldsymbol{\theta})$ are coefficients for the basis expansion of the kernel at location s. Using a set of basis functions where truncation is appropriate, both series in equation (7) may be truncated to the first L terms. The value for L should be sufficiently large for the basis expansion to accurately approximate both the kernel and the process. The number of basis functions required for an accurate representation is inversely related to the width of the kernel with respect to the overall range

of the data. For instance, a normal distribution with variance 1 may need 15 basis functions, but when the variance is .1, 30 basis functions may be needed. Also, kernel choice influences the number of basis functions required. For example, a kernel which is highly skewed will require more basis functions than a symmetric kernel. To ensure the correct number of basis functions are being used, the approximation should be checked graphically to ensure that it fits the distribution well. Recommendations for L under certain distributions will be noted as they are discussed.

Due to the orthogonality of the basis functions, the components of the integral in equation (3) can be replaced with the basis expansions in equation (7) and rewritten as follows: $\int k(s-u|\boldsymbol{\theta})X_t(u)du \approx \mathbf{a}_t'\mathbf{b}(s,\boldsymbol{\theta})$, where $\mathbf{a}_t = (a_1(t),...,a_L(t))'$ and $\mathbf{b}(s,\boldsymbol{\theta}) = (b_1(s,\boldsymbol{\theta}),...,b_L(s,\boldsymbol{\theta}))'$. Moreover, by placing $X_{t+1}(s) = \sum_{i=1}^{L}\phi_i(s)a_i(t+1)$ into the left side of equation (3), we obtain $\mathbf{a}_{t+1}'\boldsymbol{\phi}(s) = \mathbf{a}_t'\mathbf{b}(s,\boldsymbol{\theta}) + \omega_{t+1}(s)$, where $\boldsymbol{\phi}(s) = (\phi_1(s),...,\phi_L(s))'$.

The values for $b_j(s,\boldsymbol{\theta})$ are deterministic, given the choice of kernel and the corresponding parameters. The values for $a_i(t)$ are unknown and vary with time. Under the basis expansion, the process model can be summarized hierarchically as follows:

$$\boldsymbol{X}_t = \boldsymbol{\Phi}\mathbf{a}_t + \boldsymbol{\varepsilon}_t \tag{8}$$

$$\mathbf{a}_t = (\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{B}_\theta\mathbf{a}_{t-1} + \boldsymbol{\omega}_t, \tag{9}$$

where $\boldsymbol{X}_t = (X_t(s_1),...,.X_t(s_n))'$ is a realization from the process at time t, $\boldsymbol{B}_\theta = (\mathbf{b}(s_1,\boldsymbol{\theta})\ldots$ $\mathbf{b}(s_n,\boldsymbol{\theta}))$ is a matrix whose columns consist of the vectors of the kernel basis coefficients, and the (i,j)th element of $\boldsymbol{\Phi}$ is $\phi_i(s_j)$. The vector $\boldsymbol{\varepsilon}_t$ accounts for both observational error and truncation error from the basis approximation and $\boldsymbol{\omega}_t$ accounts for process error.

For a bounded spatial domain, say $[r_1, r_2]$, it is natural to consider the orthogonal family given by the Fourier basis. In such case, the kernel needs only be specified through its characteristic function. For example, a Gaussian kernel has Fourier coefficients $b_{2j-1}(s,\boldsymbol{\theta}) = r^{-1/2}\exp(-.5\rho_j^2\sigma^2)\cos(\rho_j(s+\mu))$, and $b_{2j}(s,\boldsymbol{\theta}) = r^{-1/2}\exp(-.5\rho_j^2\sigma^2)\sin(\rho_j(s+\mu))$, where

$r = r_2 - r_1$ and $\rho_j = 2\pi j/r$ is the spatial frequency. For both the Gaussian density and the two alternative IDE kernel densities proposed next, technical details concerning the approximation error of the Fourier basis decomposition are presented in Appendix D.

3.2 Asymmetric Laplace Distribution

The asymmetric Laplace is an infinitely divisible, location family distribution which allows for skewness and heavier tails than the normal. The distribution is characterized by its mode ξ, a scale parameter σ, and a parameter controlling the skewness and other shape properties, $\kappa > 0$. The density function is given by

$$k(x|\xi,\sigma,\kappa) = \frac{\sqrt{2}}{\sigma}\frac{\kappa}{1+\kappa^2}\begin{cases}\exp\left(-\frac{\sqrt{2}\kappa}{\sigma}|x-\xi|\right) & \text{if } x \geq \xi \\ \exp\left(-\frac{\sqrt{2}}{\sigma\kappa}|x-\xi|\right) & \text{if } x < \xi\end{cases}$$

which shows how the asymmetric Laplace can be formed from two exponentials with different intensities. When $\kappa = 1$, the distribution simplifies to the (symmetric) Laplace distribution. The property of infinite divisibility can be found in Kotz et al. (2001). Figure 2 shows different asymmetric Laplace densities for varying values of κ.

The asymmetric Laplace can be written as a mixture of normals with mean $\xi + \mu W$ and variance $\sigma^2 W$, where $\mu = 2^{-1/2}\sigma\left(\kappa^{-1} - \kappa\right)$ and W is an exponential distributed random variable with mean 1 (Kotz et al., 2001). This mixture representation yields the following result, the proof of which can be found in Appendix B.

Lemma 3. *The Fourier coefficients of the basis expansion for a kernel in the asymmetric Laplace family are:*

$$b_{2j-1}(s,\boldsymbol{\theta}) = \frac{(1+.5\rho_j^2\sigma^2)\cos(\rho_j(s+\xi)) + \rho_j\mu\sin(\rho_j(s+\xi))}{(-1-.5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2}$$

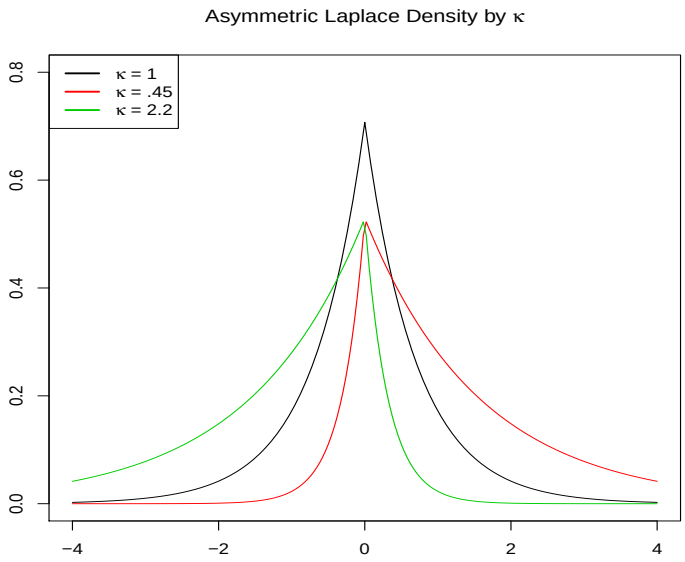

Figure 2: Asymmetric Laplace densities for different values of the skewness parameter κ. The distribution is symmetric when $\kappa = 1$ and can be highly skewed in either direction when κ is large or small.

and

$$b_{2j}(s, \boldsymbol{\theta}) = \frac{(1 + .5\rho_j^2\sigma^2)\sin(\rho_j(s+\xi)) - \rho_j\mu\cos(\rho_j(s+\xi))}{(-1 - .5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2} \quad .$$

Computationally, the non-differentiability of the density at its mode makes it harder to approximate using basis functions. To get a working approximation using a Fourier basis, the truncation point required is much larger for the asymmetric Laplace than it is for the Gaussian density, typically ranging from 30 to 100 basis functions. The more skewed the distribution is, the harder it becomes to approximate well.

3.3 Stable Distributions

Lemma 2 suggests that the kernel tail behavior will affect IDE evolution. To explore infinitely divisible kernels with tails that are substantially heavier than those of a Gaussian, we consider the family of stable distributions. A distribution belongs to the class of stable distributions if any linear combination of two random variables from a particular class of distributions

also belong to that same family. Thus, this is a subset of infinitely divisible distributions, as shown in Samorodnitsky and Taqqu (1997) and Nolan (2003). The stable distributions are governed by 4 parameters, $\mu \in \mathbb{R}$, $c > 0$, $\alpha \in (0, 2]$, and $\beta \in [-1, 1]$, and a wide range of skewness and tail behavior can be achieved by varying the parameters appropriately. A characteristic of the family of stable distributions is that, in general, it does not have an analytically available form for the density function, or moments. Specials cases of stable distributions include the Gaussian distribution when $\alpha = 2$, the Cauchy distribution when $\alpha = 1$ and $\beta = 0$, and the Levy distribution when $\alpha = 1/2$ and $\beta = 1$. The family is generally defined through its characteristic function, which for $\alpha \neq 1$ is given by

$$\psi(t|\mu, c, \alpha, \beta) = \exp\left\{it\mu - |ct|^{\alpha}(1 - i\beta sgn(t)\tan(\pi\alpha/2))\right\},$$

where $sgn(t)$ is equal to 1 when t is positive, -1 when t is negative, and 0 when t is 0. Figure 3 shows how the shape of the density changes with α and β. Note that α controls the tails and β controls the skewness, while μ and c are location and scale parameters, respectively. In Appendix B we derive the following result on the Fourier series expansion for IDE kernels from the stable family of distributions.

Lemma 4. *The Fourier coefficients of the basis expansion for a kernel in the family of stable distributions are:*

$$\begin{aligned} b_{2j-1}(s, \boldsymbol{\theta}) &= \cos\left(\rho_j(s + \mu) + |c\rho_j|^{\alpha}\beta sgn(\rho_j)\tan(\pi\alpha/2)\right)\exp\left(-|c\rho_j|^{\alpha}\right) \\ b_{2j}(s, \boldsymbol{\theta}) &= \sin\left(\rho_j(s + \mu) + |c\rho_j|^{\alpha}\beta sgn(\rho_j)\tan(\pi\alpha/2)\right)\exp\left(-|c\rho_j|^{\alpha}\right). \end{aligned}$$

The quality of this Fourier series approximation depends on the shape of the distribution. To avoid requiring a large truncation point, it is computationally convenient to restrict $\alpha \in (1, 2]$. This restricts the tail behavior to be between the Cauchy and the Gaussian distributions, but still ensures polynomial tail behavior for all values of $\alpha < 2$. For $\alpha < 1$, the required truncation level for a Fourier basis expansion increases tremendously. With

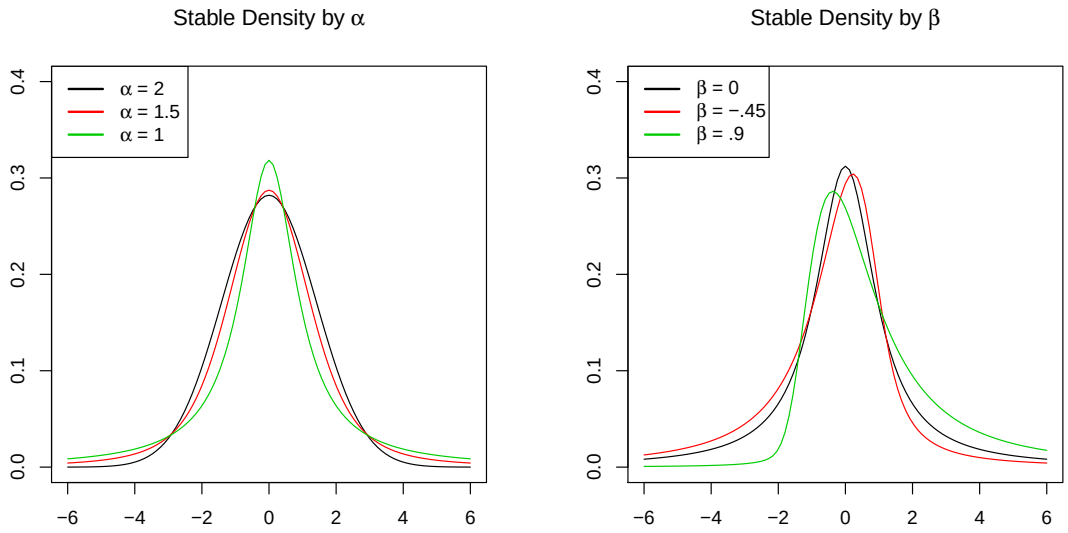

Figure 3: Densities from the stable class of distributions with $\mu = 0$ and $c = 1$, for different values of the stability parameter α (left panel) and skewness parameter β (right panel). Smaller α values result in heavier tails and β values far from 0 result in greater skewness. The left panel fixes $\beta = 0$ and the right panel fixes $\alpha = 1.1$.

$\alpha > 1$, the number of terms required is comparable to the normal, between 20 and 50, and thus computational expense will be similar. The high degree of flexibility in modeling the heaviness of the tails and the skewness combined with similar computational burden as the Gaussian kernel IDE makes the stable family a very attractive choice for the IDE kernel.

3.4 Prior Simulation

To empirically study how the various kernels affect the IDE model, we perform a series of prior simulations under four different kernels. The first of these kernels is normal with a mean of $-.67$ and a variance of 2. The second kernel is an asymmetric Laplace with parameters $\xi = .85$, $\sigma = .63$ and $\kappa = 3$, resulting in the same mean and variance as the normal kernel, but with a left skewed density. By matching the means and the variances of these two kernels, we can explore whether the first two moments dominate the IDE process or if a non-zero third moment results in different process realizations, as suggested by equation (5). The third

kernel is a stable distribution with parameters $\mu = .55$, $c = .77$, $\alpha = 1$ and $\beta = -.8$, which are chosen such that the resulting density is skewed and shaped to match the asymmetric Laplace. The final kernel is also a stable distribution with quartiles and a median which match the normal kernel, having heavy tails and no skewness; the corresponding parameters are $\mu = .33$, $c = .8$, $\alpha = 1$ and $\beta = 0$. This will test how tail behavior affects the IDE model for otherwise similar kernels. These simulations were conducted using the same process realization at time 0 and without any noise, to avoid confounding any process effects with random error.

Process realizations are simulated from the IDE model according to equation (3). The initial condition at $t = 0$ is a realization from a Gaussian process. The resulting IDE realizations are shown in Figure 4. These simulations show that, while the general trend is similar across each kernel choice, the localized features differ for each time point. The process for the IDE with more flexible kernels behaves as a more colorful version of the process using a Gaussian kernel, as can be seen best in the third column.

4 Illustrative Data Examples

The theory supports the use of the asymmetric Laplace and the stable family as possible extensions to the normal distribution for the IDE kernel. To see how these kernels compare in actual model performance, we apply the IDE model with all three kernels and compare the predictive results. In Section 4.1 two synthetic data sets will be fit and compared. In Section 4.2, the methods will be compared using real data collected by ozonesonde readings on ozone pressure.

4.1 Comparing Model Fits with Synthetic Data

To test the asymmetric Laplace and stable distributions against the normal, data is simulated under the IDE setting from two different kernels. The first is a mixture of normal

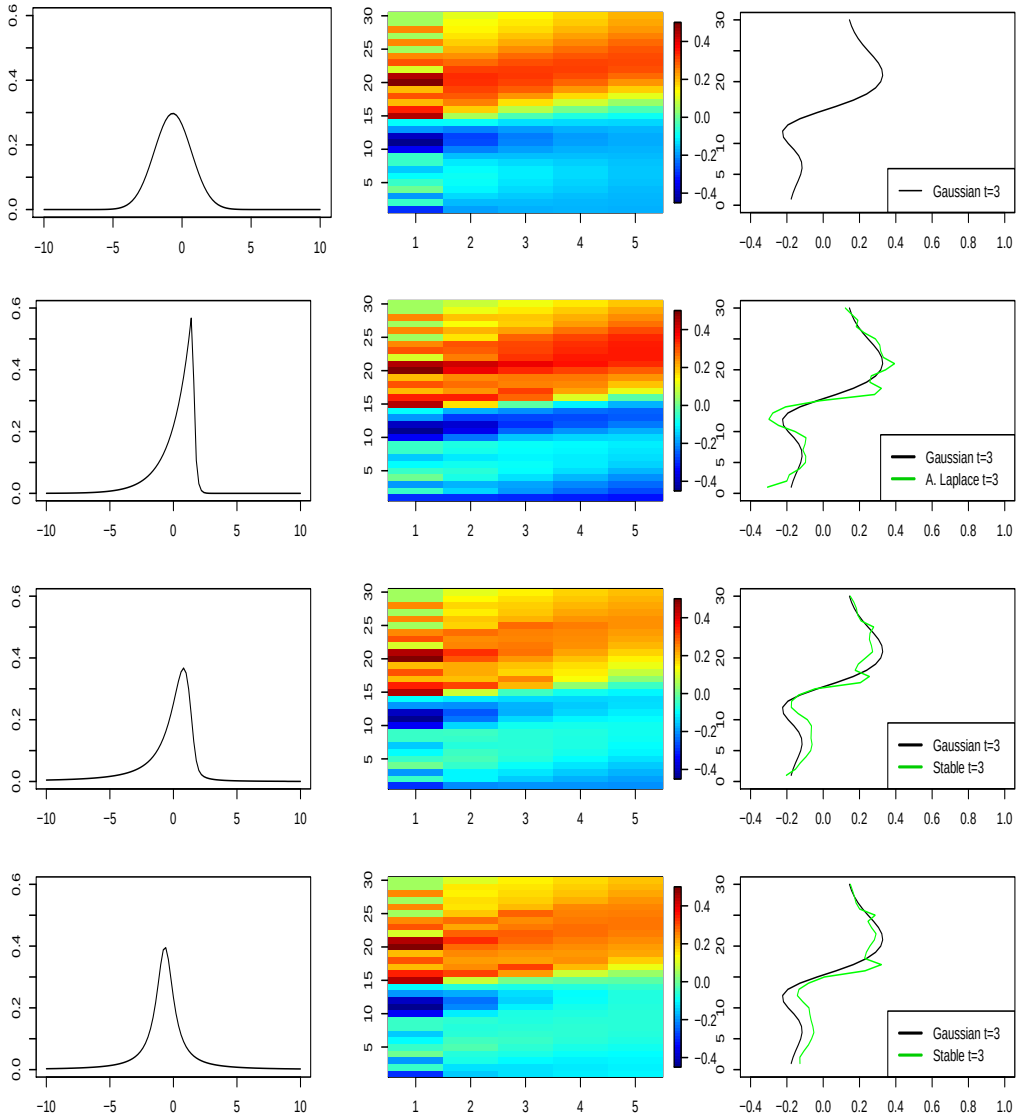

Figure 4: IDE prior simulations in one-dimensional space for four distinct kernels. From top to bottom the kernels are normal, asymmetric Laplace, stable with skewness, and stable with heavy tails and no skewness. The first column plots the IDE kernel density, the second column shows the simulated process for 5 time points, and the last column compares the spatial field between the particular kernel and the Gaussian kernel for the third time point.

distributions, $.35N(-3,1) + .25N(-1,1) + .15N(1,1) + .1N(3,1) + .1N(5,1) + .05N(7,1)$, which results in a skewed density with exponential tails. The second simulation is from an IDE with a Cauchy kernel, which is a special case of the stable with $\alpha = 1$ and $\beta = 0$. Each of these simulated data sets spans over 200 gridded spatial locations, between -40 and 40, and over 50 time points. We use independent observation errors arising from a normal

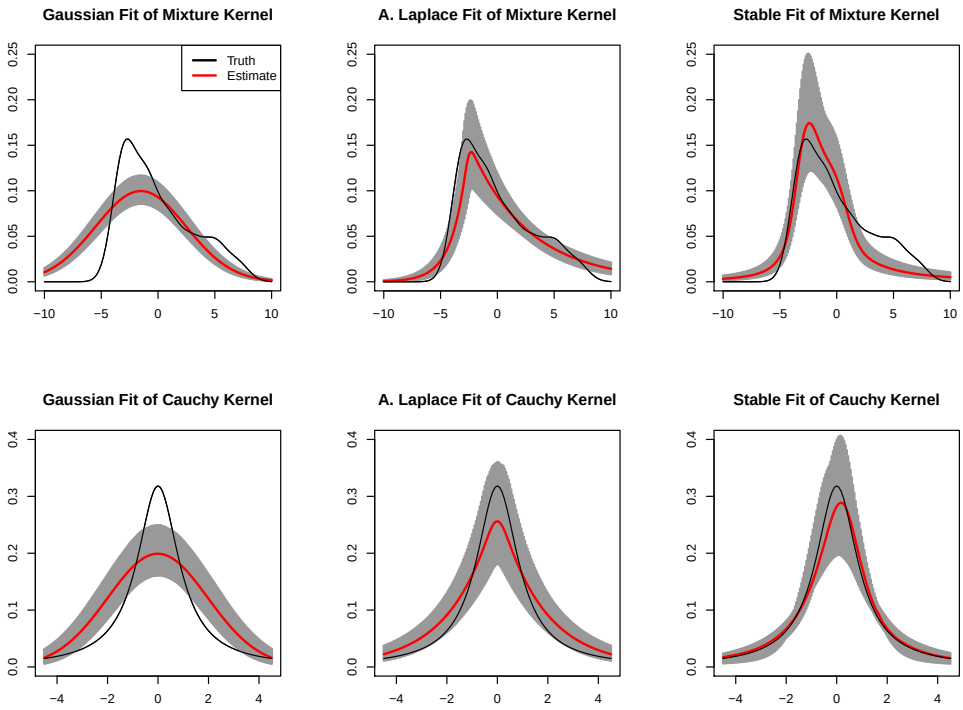

Figure 5: Synthetic data. Posterior mean and interval estimates for the IDE kernel density under the model with the Gaussian, asymmetric Laplace, and stable kernels. The top row corresponds to the data generated from an IDE model with a normal mixture for the kernel, and the bottom row to the simulated data based on an IDE model with a Cauchy kernel.

distribution with standard deviation .5, and a spatially correlated Gaussian error process with Matérn covariance function having smoothness parameter 1.5, range 3, and scale .25.

Because the kernel parameters are embedded within the structure of the evolution matrix, we use Metropolis-Hastings steps within a dynamic linear model framework to obtain samples from the posterior distribution of those parameters. Details are provided in Appendix C for the specific hierarchical model given in Section 4.2.

Posterior mean and pointwise interval estimates for the kernel densities are shown in Figure 5. The interval estimates are computed as the 95% credible interval of the density at each point over a grid, using the posterior samples for the kernel parameters. The Gaussian kernel is unsuccessful in recreating the truth. The more flexible kernels more appropriately capture the skewness and tail behavior of the underlying IDE kernel.

	True kernel	
Fitted kernel	Mixture	Cauchy
Gaussian	16%	0%
Asymmetric Laplace	70%	4%
Stable	14%	96%

Table 1: Synthetic data. The percentage of times each of the kernels had the lowest energy score for each of the simulated data sets. The asymmetric Laplace performed the best for the mixture kernel and the stable family performed the best for the Cauchy kernel.

We compare the predictions from Markov chain Monte Carlo (MCMC) runs using an energy score, following Gneiting (2002). This procedure allows for simultaneous scoring of a whole spatial field. The energy score is calculated as

$$\hat{es}(F, y) = \frac{1}{m}\sum_{i=1}^{m} ||y^{(i)} - y|| - \frac{1}{2m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} ||y^{(i)} - y^{(j)}||, \tag{10}$$

where $y^{(1)}, ..., y^{(m)}$ are samples from F, the posterior predictive distribution and y denotes the data vector. For each of the simulated data sets we compute energy scores for one step ahead out-of-sample predictions for 50 time points. Table 1 shows the percentage of times each of the kernels scored the lowest. The scoring indicates clearly which kernel performs the best in each case. The asymmetric Laplace outperforms the others for the skewed mixture, whereas the stable distribution outperforms the others for the Cauchy kernel. To offer an explanation, note that the polynomial tails of the stable may not match up well with the exponential tails of the mixture IDE kernel and the Gaussian could not capture the skewness, but the asymmetric Laplace is able to capture skewness and tail behavior better. Whereas with the Cauchy IDE kernel, only the stable distribution could match the polynomial tails.

To be able to compare model fits, the number of Fourier basis functions was fixed for all three models to be 51. The normal and stable distribution kernel IDE models produce similar results with around 25 basis functions, but the larger number of basis functions is necessary for the asymmetric Laplace kernel. Using 51 basis functions, it took about 3 hours to get 3,000 MCMC samples from the posterior distribution. Using 25 basis functions, reduces the

corresponding computing time to 1 hour.

For both of these simulation settings, we tested the hypothesis that $\alpha = 2$ for the stable density (which corresponds to the special case of a Gaussian density) by placing a prior on α that includes a point mass probability of $\Pr(\alpha = 2) = .2$. A high posterior probability of α being equal to 2 would suggest that the Gaussian IDE kernel is sufficient. For the normal mixture simulation setting, the posterior probability was $\Pr(\alpha = 2 \mid \text{data}) = .02$, and for the Cauchy simulation example it was $\Pr(\alpha = 2 \mid \text{data}) = .06$, thus providing further evidence that there is merit to using the stable IDE kernel distribution instead of the Gaussian.

4.2 Ozone Data

The study of ozone has provided an abundant source of environmental and statistical literature over the past decades. The effect of lower atmosphere ozone measurements has been seen to affect other climate variables such as concentration of certain pollutants and temperature (Robeson and Steyn, 1990). Other studies have shown how ozone concentrations affect crop yields and other agricultural variables (Heck et al., 1984). Understanding lower atmosphere ozone levels may help to understand and predict many other important variables which have a direct societal impact.

To study how the kernel choice may affect IDE model performance, we fit our proposed models to 10 years of low atmosphere ozone pressure data. These data are collected by ozonesondes, which are balloons that ascend into the atmosphere and record measurements at regular intervals. The data set we study includes biweekly ozone pressure from October 1996 to October 2006 collected at Koldewey Station near the North Pole. Details about this weather station and others related to it can be found at http://www.awi.de/en/home/.

The data are collected by releasing a balloon in the air which, at certain intervals throughout its flight, takes a measurement of ozone pressure in millPascals (mPa). The resulting data structure poses many issues for modelers. First, the locations at which the data are collected vary across time. The balloon usually takes measurements at regular intervals, but

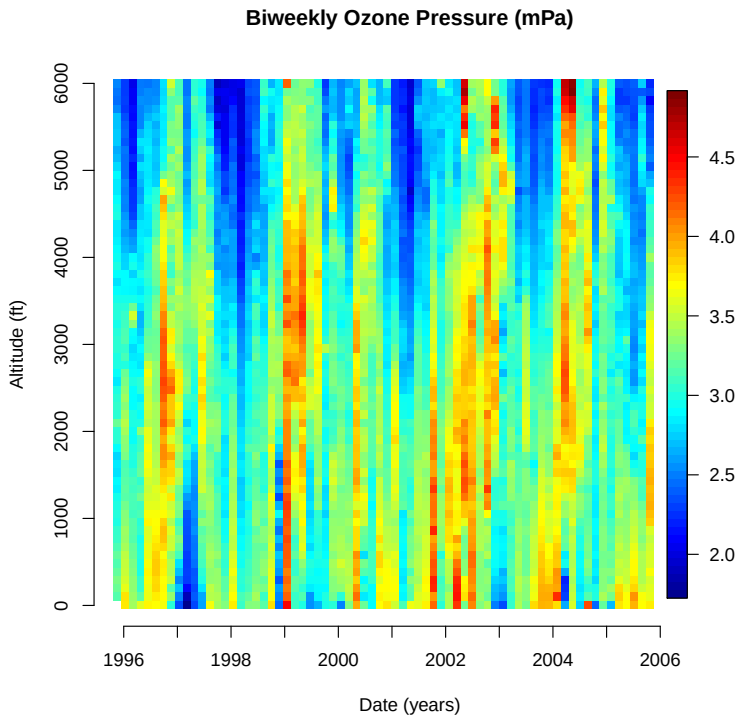

Figure 6: Biweekly ozone pressure measured on a vertical profile, plotted across altitude (0 to 6,000 feet) and over time (October 1996 to October 2006).

this rarely corresponds to a consistent pattern with surrounding time points. Another issue is that the data collecting mechanism would often fail to reach higher altitudes, leaving the entire upper half of the observation interval missing. Yet another major issue is that the data is somewhat irregular and hard to model using standard methods. For this particular illustration, we restrict our focus to the first 6,000 feet, which corresponds to lower-atmosphere ozone pressure. The data is collected almost every week, though several weeks are missing. Since this is an illustration, we opt to use biweekly data to avoid missing time points.

Because the balloon moves only in one direction, the domain for space is one-dimensional. The data is displayed in Figure 6 by altitude and time. There are a few stretches with outlying observations which are included in the analysis but are not shown so that the finer details of the data can be viewed, and also to help compare with the fitted values in Figure 10. In Figure 6, we note a potential seasonal trend. To account for this seasonality, we add two harmonics, $\boldsymbol{Z}_{ti} = (Z_{ti}^{(1)}, Z_{ti}^{(2)})$, for $i = 1, 2$. These variables will evolve through a

rotation matrix with frequency ζ_i. The resulting process has a cyclical forecast function with a period of $2\pi/\zeta_i$ (West and Harrison, 1997, Chp. 8). By including two harmonics we can account for seasonal variability with two different periods. It is possible to elaborate on this model by, for example, considering space-varying seasonal components and spatially-varying kernels. However, we opt to keep the model simpler to focus on the effects of using kernels with different skewness and tail behaviors. The full model is

$$\begin{aligned}
Y_t(s)|X_t(s), Z_{t1}^{(1)}, Z_{t2}^{(1)}, \sigma^2 &= X_t(s) + Z_{t1}^{(1)} + Z_{t2}^{(1)} + \varepsilon_t(s), \quad \varepsilon_t(s) \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad t = 1, ..., T \\
X_t(s)|\{X_{t-1}(s) : s \in D\}, \boldsymbol{\theta} &= \int_D k(s - u|\boldsymbol{\theta}) X_{t-1}(u) du + \omega_t(s) \\
\begin{pmatrix} Z_{ti}^{(1)} \\ Z_{ti}^{(2)} \end{pmatrix} &= \begin{pmatrix} \cos(\zeta_i) & \sin(\zeta_i) \\ -\sin(\zeta_i) & \cos(\zeta_i) \end{pmatrix} \begin{pmatrix} Z_{t-1,i}^{(1)} \\ Z_{t-1,i}^{(2)} \end{pmatrix} + \nu_t, \quad i = 1, 2 \\
\sigma^2 \sim \text{gamma}(a, b), \quad \boldsymbol{\theta} &\sim p(\boldsymbol{\theta}), \quad \nu_t | W_t^Z \sim N_2(0, W_t^Z),
\end{aligned}$$

where, for any points $s_1, ..., s_n$, the vector $(\omega_t(s_1), ..., \omega_t(s_n))$ has a normal distribution with a zero mean and some covariance structure, W_t. The matrices W_t and W_t^Z are modeled using discount factors. A discount factor, say δ, is a number between 0 and 1 that determines the amount of information lost through the process evolution in time (West and Harrison, 1997, Chp. 6). Note that the amplitude and phase of this cyclical effect will vary smoothly across time. The IDE kernel is chosen to be Gaussian, asymmetric Laplace, and then stable in three different model fits. The prior parameters a and b are fixed. An exploratory analysis was performed where the periods of the two harmonics were included as parameters in the model and a cluster of posterior mass around 6 months and 12 months was observed. The two harmonics were then fixed to have periods of 6 and 12 months, which results in $\zeta_1 = 2\pi/26$ and $\zeta_2 = 2\pi/13$, assuming 52 weeks per year.

For the kernel parameters and the observational variance, the posterior distributions are robust to a wide range of priors. We use a $N(0, 300^2)$ prior for the location parameter, and a gamma$(1, .01)$ prior for the scale parameter in each case. The skewness parameter κ in

the asymmetric Laplace received a gamma$(1, 1)$ prior. The stable parameters α and β were assigned scaled Beta$(2, 2)$ prior distributions to match their support. The posterior distribution was not sensitive to a wide array of slightly informative priors for the kernel parameters, but more dispersed priors seemed to delay convergence and mix poorly. The important prior specification is for $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$, which are the mean and covariance, respectively, of the basis coefficients for the time 0 process. Poor choices for these can greatly affect the posterior for the kernel parameters. Ozone pressure typically does not stray too far from the range of 2 to 4. Our best guess of the time 0 process is a constant function at 3. The basis coefficients that define $\boldsymbol{m}_0$ are $(3/\sqrt{2r}, 0, ..., 0)$, where r is the period of the Fourier transform used for the basis. We constructed $\boldsymbol{C}_0$ as a diagonal matrix with decreasing values down the diagonal so that variances of the higher order terms of the basis function expansion are close to 0. We present results based on 50,000 MCMC samples (after a burn-in period of 50,000 iterations) for every parameter of every model. 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).

To demonstrate the practical utility of non-Gaussian IDE kernels, we can study the kernel estimates, and the posterior distribution of the parameters which control skewness and heavy tails. The posterior mean estimates for the kernel density under each model are shown in Figure 7, and it can be clearly seen that the kernel tends to be asymmetric for models where that is allowed. The posterior distribution for κ in the asymmetric Laplace, and the stable parameters α an β are shown in Figure 8. Recall that κ controls the skewness of the asymmetric Laplace, and α and β control the tail behavior and skewness of the stable distribution. The credible intervals for each of these parameters are shown in Table 2. The asymmetric Laplace parameter κ includes 1 in the credible interval, suggesting that we can not rule out symmetry based on the parameter estimates. However, the credible interval for the stable distribution parameter β does not include 0, suggesting that the IDE kernel is not symmetric under the stable distribution model.

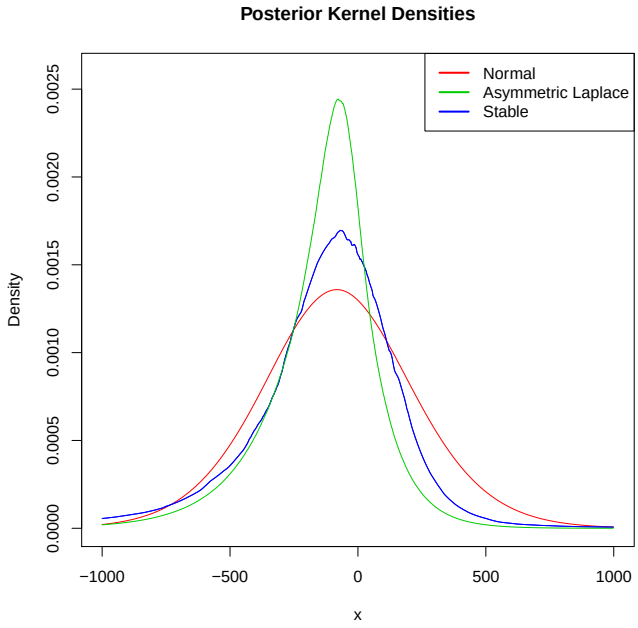

Figure 7: Ozone data. Posterior mean estimates for the IDE kernel under the Gaussian, asymmetric Laplace, and stable models.

Parameter	Median	2.5%	97.5%
κ	1.22	.69	1.75
α	1.48	1.26	1.80
β	$-.57$	$-.86$	$-.17$

Table 2: Ozone data. Posterior median and 95% credible intervals for certain parameters of the IDE models with non-Gaussian kernels.

Figure 9 shows profiles of the fitted values of one step ahead predictions for three observations from the data set using each kernel. Using such profiles, it can be seen that the Gaussian kernel IDE does not appropriately model ozone pressure in several regions. The stable distribution, however, seems to perform much better. For the Gaussian model, only two parameters determine the entire evolution matrix. Figure 9 shows that this is not enough to provide accurate one step ahead predictions. The stable distribution kernel uses four parameters and while the accuracy of prediction is certainly increased, the uncertainty increases as well, resulting in wider credible intervals. The model residuals for the stable distribution IDE model are shown in Figure 10.

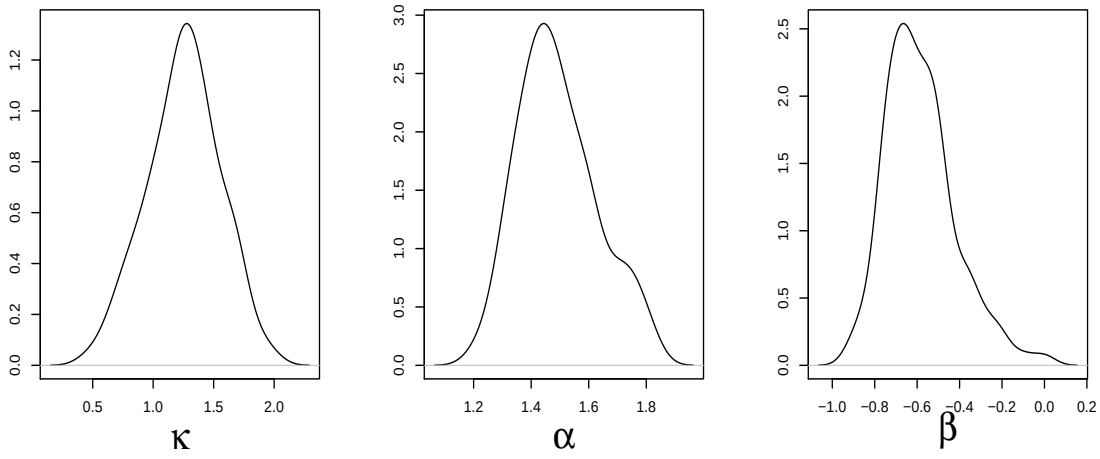

Figure 8: Ozone data. Posterior density for the skewness parameter κ of the asymmetric Laplace kernel (left panel). Posterior densities for parameters α and β which control the tails and the skewness of the stable kernel (middle and right panel).

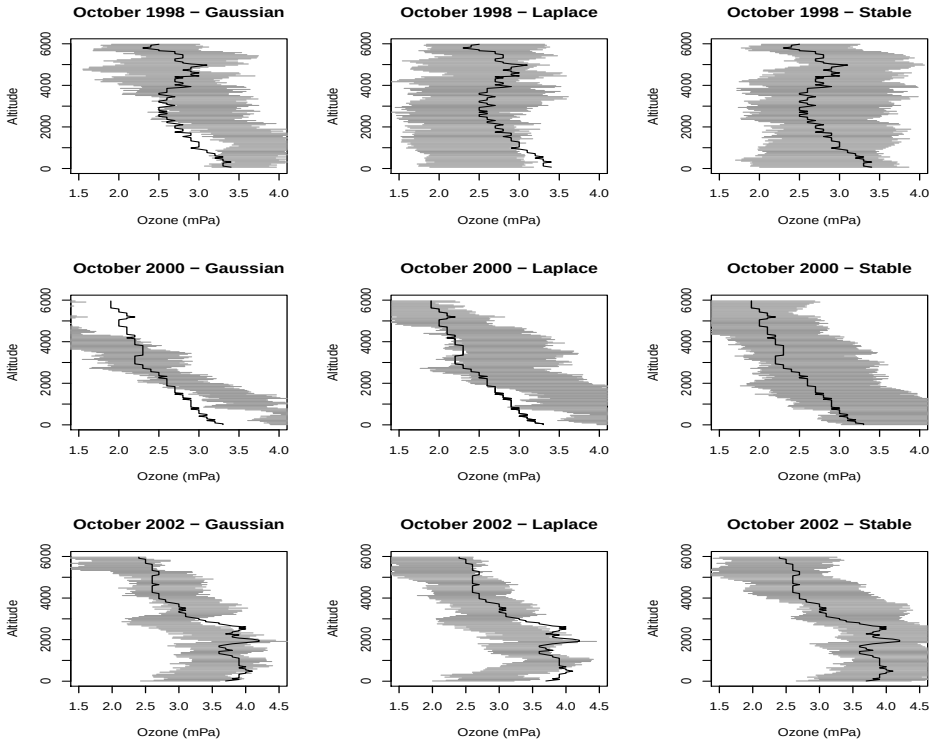

Figure 9: Ozone data. The profiles of ozone concentration are shown for three different months with one step ahead 95% predictive intervals shaded in for each of the three different kernels.

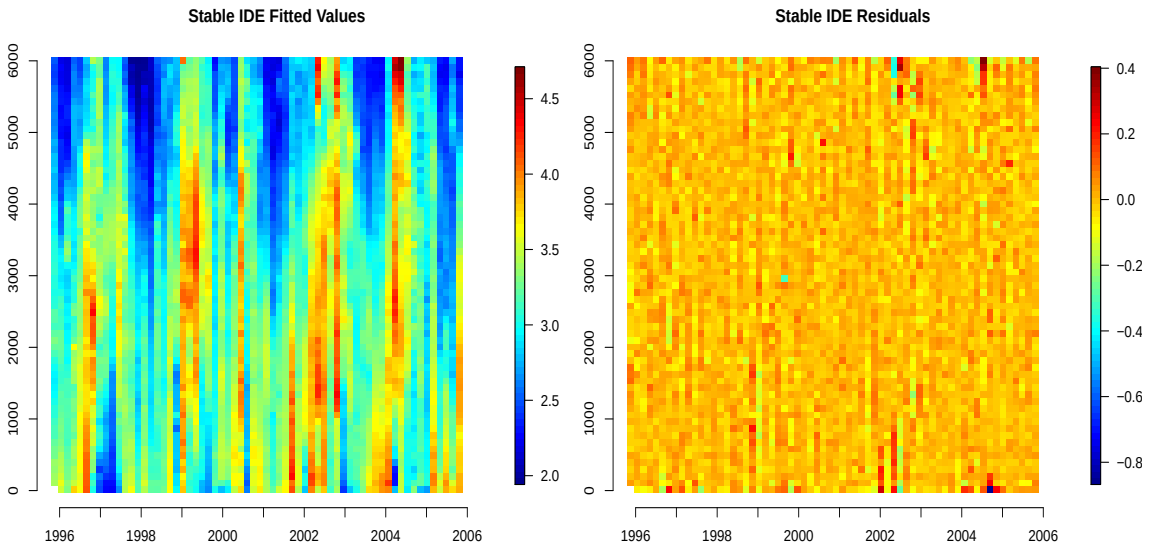

Figure 10: Ozone data. Fitted values (left) and residuals (right) are shown for every observation under the IDE model with the stable distribution kernel. The overall fit is good with the exception of a few outlying stretches.

The one step ahead predictions shown in Figure 9 are calculated for the data set for all three models. As in the simulated example, these predictions can be scored using the measure in equation (10). The results of the energy scores can help determine which model performs the best in one step ahead predictions. Table 3 shows each possible ordering for the scores and how often they occur. Recall that lower scores refer to a better fit. The stable distribution has the lowest energy score for 73% of the observations. Only 10% of the observations have the Gaussian kernel as the lowest score. These scores help discern the differences that the figures are not able to show. It is clear that the particular criterion favors the stable distribution IDE kernel over the Gaussian and asymmetric Laplace. As in the simulation examples, a separate analysis tested the hypothesis that $\alpha = 2$ in the stable distribution by placing a prior probability of $\Pr(\alpha = 2) = .2$. The posterior probability was $\Pr(\alpha = 2 \mid \text{data}) = .08$, again supporting the general stable distribution for the IDE kernel relative to the special case given by the Gaussian distribution.

To summarize these results, the posterior distribution of the stable kernel parameters suggest that normality and symmetry are poor assumptions for the IDE kernel. The posterior

Order	Frequency
S<AL<G	53%
AL<S<G	11%
G<S<AL	6%
S<G<AL	20%
AL<G<S	6%
G<AL<S	4%

Table 3: Ozone data. Each possible ordering for the scores of the IDE models under the three distinct kernels are shown with the percentage of observations that the scores followed that order. S refers to the stable distribution, AL to the asymmetric Laplace, and G to the Gaussian kernel.

distribution of the asymmetric Laplace does not rule out symmetry, but the estimate shown in Figure 7 for the asymmetric Laplace kernel is clearly asymmetric. Scoring procedures for the out of sample predictions suggest that the model with the stable kernel distribution performs the best in terms of predictive model accuracy.

5 Conclusions

Spatio-temporal data often present complicated space-time interactions that are difficult to model accurately. Under the IDE model framework, electing to use a kernel more flexible than the Gaussian, which is used in nearly all IDE modeling, provides better predictive accuracy and more potential for successfully capturing the spatio-temporal evolution of the field. Compared to Gaussian kernel densities, kernels with flexible tail behavior and potential skewness, facilitate more complicated transfer of dynamics from one time point to the next. In this paper, we have shown how the choice of kernel influences the process through theory, simulations, and data analysis. We have proposed two alternative kernel families with desirable theoretical and computational properties.

Computations for the models proposed in this paper are based on truncated expansions on Fourier bases. Based on both empirical evidence and the theoretical results developed in Appendix D, asymmetric Laplace kernels require a larger number of coefficients than stable

distribution kernels with $\alpha > 1$. The latter can be well approximated with a computational effort comparable to Gaussian kernels. Hence, when alternatives to the Gaussian IDE model are needed for large data sets, the stable distribution seems a more practical choice than the asymmetric Laplace. Additionally, because the stable family can capture a range of tail behavior as well as skewness, it is more flexible than the asymmetric Laplace.

The models proposed in this paper can be extended in at least three different ways. First, we can place a Gaussian process prior on the location parameter, along the lines of Wikle (2002). This will achieve full spatio-temporal non-stationarity. The remaining kernel parameters can also be spatially varying by using, for example, transformations of Gaussian processes. Second, we can further extend the flexibility of the kernel shape by considering non-parametric representations of the kernel. Within a Bayesian setting, we can consider a Dirichlet process mixture of normals prior on the kernel. Expressions such as (5) still hold true for such IDE models. Extensions to space-varying kernels can then be achieved by using a spatial Dirichlet process (Gelfand et al., 2005), resulting in a model whose kernel moments are completely flexible and change smoothly across a surface. Physical characteristics such as diffusion, advection, and dispersion will be spatially varying. Ongoing work shows that, for one-dimensional space problems, a spatial Dirichlet process mixture kernel IDE model can outperform Gaussian kernel IDE models with spatially varying parameters.

We have developed models for one dimensional spatial settings. While these are useful for the study of profiles that are very common in environmental variables, such as the ozone data considered in this paper, it is important to generalize the methodology for higher dimensional spaces. Conceptually, this extension is straightforward. Nevertheless, inference and computations for the families proposed in this paper are quite challenging. Starting from a multivariate characteristic function, it is possible to obtain the Fourier basis expansion in order to evaluate the IDE integral, along the lines of the one-dimensional case. For the asymmetric Laplace, a two-dimensional characteristic function is readily available (Kotz et al., 2001). However, the computational burden due to the large number of basis functions required for

a good approximation of the kernel is compounded by the dimensionality. A possible way to tackle this problem is to use a thresholding approach, where the number of basis functions is allowed to change from one MCMC iteration to the next. Such an approach requires careful exploration of the criteria to set to zero a given coefficient. For the stable family, multivariate generalizations are not immediate. The main difficulty stems from the fact that the characteristic function does not have a parametric form. In fact, the characteristic function for the multivariate stable distribution is $\exp\big(i\boldsymbol{\mu}'\boldsymbol{t} - \int |\boldsymbol{t}'\boldsymbol{s}|^{\alpha}\{1 - i\, sgn(\boldsymbol{t}'\boldsymbol{s})\tan(\pi\alpha/2)\}d\Gamma(\boldsymbol{s})\big)$, where Γ is a measure with compact support. Γ determines the main orientation, scale and skewness of the resulting stable distribution (Samorodnitsky and Taqqu, 1997). Future work will report on methods to model Γ and to implement a stable family kernel in a two-dimensional IDE setting.

Appendix A Proof of Lemma 2

Neubert et al. (1995) show that the IDE model can be written as a system of partial differential equations. That work is modified to account for asymmetric kernels and processes defined on the entire real line. The representation for the IDE process is

$$\frac{du}{d\tau} = -\frac{du}{ds} - h(\tau)u, \qquad \frac{dv}{d\tau} = h(\tau)S(0)u.$$

We confirm that this is an IDE representation using the method of characteristics. To use this method, we find curves where the PDE is trivial and then create functions of those curves based on the initial conditions. The characteristics curves can be found by solving the differential equations $d\tau = ds$ and $d\tau = -(h(\tau)u)^{-1}du$. The first PDE is simple to integrate both sides. The second can be solved for $u = C\exp\big[-\int h(\tau)d\tau\big]$.

According to the method of characteristics, the general solution can be written as $u = g(s-\tau)\exp\big[-\int h(\tau)dt\big]$. Neubert et al. (1995) assume the initial condition $u(s,0) = \delta(s)$ because all the organisms begin in one location, but in general we can use the initial condition

$u(s,0) = X_t(s)$, that is, our initial condition is the process at the previous time. Using properties of hazard functions, based on this initial condition, the general function for $u(s,\tau)$ is $X_{t-1}(s-\tau)S(\tau)/S(0)$ where $S(\cdot)$ is the survival function. Solving for v proceeds by integrating both sides of $\frac{dv}{d\tau} = X_{t-1}(s-\tau)\frac{S(\tau)}{S(0)}h(\tau)S(0)$, which then becomes

$$v(s,\tau) = \int X_{t-1}(s-\tau)k(\tau)d\tau.$$

If the initial condition for $u(x,0)$ is $X_{t-1}(s)$, then the solution for $v(s,\tau)$ is $X_t(s)$.

Appendix B Fourier Coefficients

Here, we develop the Fourier representations of the asymmetric Laplace and stable distributions. The asymmetric Laplace distribution can be written as a mixture of normals. If X is a standard normal and W is a standard exponential, then $Y = \xi + \mu W + \sigma\sqrt{W}X$ has an asymmetric Laplace distribution. Hence, conditionally on W, Y has a normal distribution with mean $\xi + \mu W$ and variance $\sigma^2 W$. This representation uses the parametrization $\mu = 2^{-1/2}\sigma\left(\kappa^{-1} - \kappa\right)$. To find appropriate basis function expansions for the asymmetric Laplace distribution in the IDE model, we can use the specific form of the normal distribution for $Y|W$ and mix over W. Recall that the $N(\mu,\sigma^2)$ kernel can be decomposed into $\sum_j b_j(s,\boldsymbol{\theta})\phi_j(u)$, where the basis functions are $\phi_{2j-1}(u) = \cos(\rho_j u)$ and $\phi_{2j}(u) = \sin(\rho_j u)$, and the coefficients are $b_{2j-1}(s,\boldsymbol{\theta}) = \exp\left(-.5\rho_j^2\sigma^2\right)\cos(\rho_j(s+\mu))$ and $b_{2j}(s,\boldsymbol{\theta}) = \exp\left(-.5\rho_j^2\sigma^2\right)\sin(\rho_j(s+\mu))$. Therefore, by mixing on W, the asymmetric Laplace distribution coefficients for the Fourier basis expansion can be found through

$$\begin{aligned} b_{2j-1}(s,\boldsymbol{\theta}) &= \int_0^\infty \exp\left(-.5\rho_j^2\sigma^2 W\right)\cos(\rho_j(s+\xi+\mu W))\exp\left(-W\right)dW \\ &= \frac{1}{(-1-.5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2}\left[(1+.5\rho_j^2\sigma^2)\cos(\rho_j(s+\xi)) + \rho_j\mu\sin(\rho_j(s+\xi))\right]. \end{aligned}$$

Similarly, we obtain

$$b_{2j}(s,\boldsymbol{\theta}) = \frac{1}{(-1-.5\rho_j^2\sigma^2)^2+(\rho_j\mu)^2}\left[(1+.5\rho_j^2\sigma^2)\sin(\rho_j(s+\xi)) - \rho_j\mu\cos(\rho_j(s+\xi))\right].$$

The stable family of distributions with $\alpha \neq 1$ has characteristic function of the form $\psi(t) = \exp\{it\mu - |ct|^\alpha(1 - i\beta sgn(t)\tan(\pi\alpha/2))\}$. Decomposing the characteristic function into its real and imaginary parts and then applying Euler's formula, we find the coefficients for the sine and cosine basis functions:

$$\psi(t) = \cos\left(t\mu + |ct|^\alpha\beta sgn(t)\tan(\pi\alpha/2)\right)\exp(-|ct|^\alpha) + i\sin\left(t\mu + |ct|^\alpha\beta sgn(t)\tan(\pi\alpha/2)\right)\exp(-|ct|^\alpha).$$

The real part of this equation corresponds to the cosine coefficients and the sine part refers to the sine coefficients in a Fourier transform.

Appendix C MCMC Details

This section details the procedure used for learning model parameters for the ozone data. As outlined in Cressie and Wikle (2011) and summarized in equations (1) and (2), we use a hierarchical dynamic linear model framework. Due to the seasonality in the ozone data, we include two harmonics at the process level. Combining a hierarchical model with the basis function decompositions in equation (7), we can write the model for data vector, $\boldsymbol{Y}_t$, as

$$\begin{aligned}
\boldsymbol{Y}_t|\mathbf{a}_t, Z_{t1}^{(1)}, Z_{t2}^{(1)}, \sigma^2 &\sim N(\boldsymbol{\Phi}\mathbf{a}_t + Z_{t1}^{(1)} + Z_{t2}^{(1)}, \sigma^2 I), \quad t = 1:T \\
\mathbf{a}_t|\mathbf{a}_{t-1}, \boldsymbol{\theta}, \boldsymbol{W}_t &\sim N((\boldsymbol{\Phi}'\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}'\boldsymbol{B}_\theta\mathbf{a}_{t-1}, \boldsymbol{W}_t), \\
\begin{pmatrix} Z_{ti}^{(1)} \\ Z_{ti}^{(2)} \end{pmatrix} \Bigg| \begin{pmatrix} Z_{t-1,i}^{(1)} \\ Z_{t-1,i}^{(2)} \end{pmatrix}, \boldsymbol{W}_t^Z &\sim N\left(\begin{pmatrix} \cos(\zeta_i) & \sin(\zeta_i) \\ -\sin(\zeta_i) & \cos(\zeta_i) \end{pmatrix}\begin{pmatrix} Z_{t-1,i}^{(1)} \\ Z_{t-1,i}^{(2)} \end{pmatrix}, \boldsymbol{W}_t^Z\right), \quad i = 1,2,
\end{aligned}$$

The kernel parameters, $\boldsymbol{\theta}$ are embedded in the matrix $\boldsymbol{B}_\theta$. To find posterior distributions for the parameters, we must use Metropolis-Hastings. The parameters of the model include

the latent variables $\mathbf{a}_t$ and $\boldsymbol{Z}_{ti}$, in addition to the parameter set $\boldsymbol{\theta}$ and variance terms σ^2, $\boldsymbol{W}_t$, and $\boldsymbol{W}_t^Z$. By augmenting the vectors and blocking the matrices in the model we can rewrite this in the simpler form

$$\begin{aligned}\boldsymbol{Y}_t &= \boldsymbol{F}_t\boldsymbol{X}_t + \boldsymbol{\varepsilon}_t^*, \quad \boldsymbol{\varepsilon}_t^* \sim N(\mathbf{0}, \boldsymbol{R}_t), \quad t = 1:T \\ \boldsymbol{X}_t &= \boldsymbol{G}_t\boldsymbol{X}_{t-1} + \boldsymbol{\omega}_t^*, \quad \boldsymbol{\omega}_t^* \sim N(\mathbf{0}, \boldsymbol{Q}_t).\end{aligned}$$

The state vector is now $\boldsymbol{X}_t = (\mathbf{a}_t', \boldsymbol{Z}_{t1}', \boldsymbol{Z}_{t2}')'$. The matrices here are also redefined as $\boldsymbol{F}_t = (\boldsymbol{\Phi}, \mathbf{1}, \mathbf{0}, \mathbf{1}, \mathbf{0})$, $R_t = \sigma^2 I$, $\boldsymbol{G}_t = \text{blockdiag}\left((\boldsymbol{\Phi}_t'\boldsymbol{\Phi}_t)^{-1}\boldsymbol{\Phi}_t'\boldsymbol{B}_\theta, J(\zeta_1), J(\zeta_2)\right)$, where $J(\zeta)$ is a rotation matrix with frequency ζ, and $\boldsymbol{Q}_t = \text{blockdiag}\left(\boldsymbol{W}_t, \boldsymbol{W}_t^Z\right)$.

Estimating the states is done using standard filtering formulas as found in Prado and West (2010). First the prior mean and covariance estimates for $\boldsymbol{X}_0$ must be set as $\boldsymbol{m}_0$ and $\boldsymbol{C}_0$. Given all information up to time t, denoted as $\boldsymbol{D}_t$, the posterior distributions of the state vectors $\boldsymbol{X}_t|\boldsymbol{D}_t$ are $N(\boldsymbol{m}_t, \boldsymbol{C}_t)$. These can be found using recursive formulas

$$\boldsymbol{m}_t = \boldsymbol{G}_t\boldsymbol{m}_{t-1} + \boldsymbol{K}_t\left(\boldsymbol{Y}_t - \boldsymbol{F}_t\boldsymbol{G}_t\boldsymbol{m}_{t-1}\right)), \qquad \boldsymbol{C}_t = \left(I - \boldsymbol{K}_t\boldsymbol{F}_t\right)\left(\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{Q}_t\right),$$

where $\boldsymbol{K}_t = \left(\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{Q}_t\right)\boldsymbol{F}_t'\left(\boldsymbol{F}_t\left(\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t' + \boldsymbol{Q}_t\right)\boldsymbol{F}_t' + \boldsymbol{R}_t\right)^{-1}$. The covariance matrices $\boldsymbol{W}_t$, and $\boldsymbol{W}_t^Z$ can be estimated using discount factors. This involves setting $\boldsymbol{Q}_t$ equal to $\frac{1-\delta}{\delta}\boldsymbol{G}_t\boldsymbol{C}_{t-1}\boldsymbol{G}_t$, for some fixed value for δ. More details are provided in Prado and West (2010). Given these covariances and the states, the parameters in the kernel and the variance σ^2 can be updated using Metropolis-Hastings steps. The distribution of $\boldsymbol{Y}_t|\boldsymbol{D}_t, \boldsymbol{\theta}, \sigma^2$ is $N(\boldsymbol{F}_t\boldsymbol{m}_t, \boldsymbol{F}_t\boldsymbol{C}_t\boldsymbol{F}_t' + \boldsymbol{R}_t)$. New values for $\boldsymbol{\theta}^*$ are proposed from a proposal distribution $q(\cdot)$. Typically, the variables were transformed so that the proposal distribution could be a normal distribution. The variance of this normal proposal distribution was tuned to an appropriate acceptance rate. If the value of the parameters at the previous iteration of the MCMC is

$\boldsymbol{\theta}^{(B-1)}$, then the new values will be accepted with probability

$$\min\left(\frac{p(\boldsymbol{Y}_t|\boldsymbol{D}_t,\boldsymbol{\theta}^*,\sigma^2)p(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^{(B-1)}|\boldsymbol{\theta}^*)}{p(\boldsymbol{Y}_t|\boldsymbol{D}_t,\boldsymbol{\theta}^{(B-1)},\sigma^2)p(\boldsymbol{\theta}^{(B-1)})q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(B-1)})},1\right)$$

and, otherwise, set $\boldsymbol{\theta}^{(B)}=\boldsymbol{\theta}^{(B-1)}$. A similar Metropolis-Hastings step is used to update σ^2.

Appendix D Fourier Series Approximation Error

Let $R(u|\boldsymbol{\theta},L)=k(u-s|\boldsymbol{\theta})-\sum_{j=1}^{L}b_j(s,\boldsymbol{\theta})\phi_j(u)$ represent the remainder of the error in the Fourier series approximation for the IDE kernel density. Such error depends on the location u, the parameter set $\boldsymbol{\theta}$, and the truncation level L. For all three density choices considered in this work, the L_2 norm of the approximation error, $\|R\|_2=\left(\int_{r_1}^{r_2}\left(R(u|\boldsymbol{\theta},L)\right)^2du\right)^{1/2}$, can be bounded above by a constant that depends on $\boldsymbol{\theta}$ and L. The expression for this bound can be used to study how the parameters affect the approximation error and how the error compares between the distributions.

As $L\to\infty$, there is no approximation error, that is, $k(u-s|\boldsymbol{\theta})=\sum_{j=1}^{\infty}b_j(s,\boldsymbol{\theta})\phi_j(u)$. Hence, $\|R\|_2=\left(\int_{r_1}^{r_2}\left(\sum_{j=L+1}^{\infty}b_j(s,\boldsymbol{\theta})\phi_j(u)\right)^2du\right)^{1/2}$, which, due to orthogonality, simplifies to $\left(\sum_{j=L+1}^{\infty}\{b_j(s,\boldsymbol{\theta})\}^2\right)^{1/2}$.

For the Gaussian density, the Fourier coefficients, $b_{2j-1}(s,\boldsymbol{\theta})$ and $b_{2j}(s,\boldsymbol{\theta})$ (given in Section 3.1), are each bounded by $r^{-1/2}\exp\left(-.5\rho_j^2\sigma^2\right)$, where $r=r_2-r_1$ and $\rho_j=2\pi j/r$. Therefore, the L_2 norm of the truncation error is bounded by

$$\begin{aligned}\|R\|_2 &\leq \left(\sum_{j=L+1}^{\infty}r^{-1}\exp\left(-(2\pi/r)^2\sigma^2j^2\right)\right)^{1/2}\leq\left(\sum_{j=L+1}^{\infty}r^{-1}\exp\left(-(2\pi/r)^2\sigma^2j\right)\right)^{1/2}\\ &= r^{-1/2}\left(\frac{\left(\exp\left(-(2\pi\sigma/r)^2\right)\right)^{L+1}}{1+\exp\left(-(2\pi\sigma/r)^2\right)}\right)^{1/2}.\end{aligned}$$

Following similar steps, the L_2 norm of the truncation error for the stable density ap-

proximation is bounded by

$$\|R\|_2 \leq r^{-1/2} \left(\frac{\left(\exp\left(-(2\pi c/r)^\alpha\right)\right)^{L+1}}{1+\exp\left(-(2\pi c/r)^\alpha\right)} \right)^{1/2}.$$

In both cases, we note that the ratio of the parameter controlling the spread of the density and the width of the region where the approximation is made is the main determinant of how large the bound for the approximation error is. As the spread parameter gets smaller, a larger L is required to maintain a similar approximation error.

For the asymmetric Laplace distribution, the Fourier coefficients (given in Appendix B) squared and summed in pairs simplify to $\{b_{2j-1}(s,\boldsymbol{\theta})\}^2 + \{b_{2j}(s,\boldsymbol{\theta})\}^2 = 2/\{(1+.5\rho_j^2\sigma^2)^2 + (\rho_j\mu)^2\}$. The L_2 norm is then bounded above by

$$\|R\|_2 \leq \sum_{j=L+1}^{\infty} \frac{2}{(1+.5(2\pi j\sigma/r)^2)^2 + (2\pi j\mu/r)^2}.$$

While this does not simplify to a closed form expression, we note that the same relationship holds in that the ratio of the spread parameter to the width of the approximation region is important in determining how many basis functions are needed to approximate the density well. Also, note that for the Gaussian and stable densities, the bound is exponentially decaying with L, whereas the bound for the asymmetric Laplace approximation has a polynomial decay with L, which supports the empirical observation that the Gaussian and stable density approximation requires fewer basis functions than the asymmetric Laplace.

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.

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

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, New York: John Wiley & Sons.

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*, 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.

Gneiting, T. (2002), "Nonseparable, stationary covariance functions for space–time data," *Journal of the American Statistical Association*, 97, 590–600.

Hamilton, J. D. (1994), *Time Series Analysis*, vol. 2, Cambridge University Press.

Heck, W. W., Cure, W. W., Rawlings, J. O., Zaragoza, L. J., Heagle, A. S., Heggestad, H. E., Kohut, R. J., Kress, L. W., and Temple, P. J. (1984), "Assessing impacts of ozone on agricultural crops: II. Crop yield functions and alternative exposure statistics," *Journal of the Air Pollution Control Association*, 34, 810–817.

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

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.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001), *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering, and Finance*, Boston: Birkhäuser.

Matheron, G. (1963), "Principles of geostatistics," *Economic Geology*, 58, 1246–1266.

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.

Papoulis, A. and Pillai, S. U. (2002), *Probability, random variables, and stochastic processes*, Tata McGraw-Hill Education.

Prado, R. and West, M. (2010), *Time Series: Modeling, Computation, and Inference*, Florida: CRC Press.

Robeson, S. and Steyn, D. (1990), "Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations," *Atmospheric Environment. Part B. Urban Atmosphere*, 24, 303–312.

Samorodnitsky, G. and Taqqu, M. S. (1997), "Stable Non-Gaussian Random Processes," *Econometric Theory*, 13, 133–142.

Shumway, R. H. and Stoffer, D. S. (2011), *Time Series Analysis and its Applications, with R Examples*, New York: Springer.

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

Stein, M. L. (2005), "Space–time covariance functions," *Journal of the American Statistical Association*, 100, 310–321.

Steutel, F. W. and Harn, K. V. (2003), *Infinite Divisibility of Probability Distributions on the Real Line*, Florida: CRC Press.

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

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.

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.

Zheng, Y. and Aukema, B. H. (2010), "Hierarchical dynamic modeling of outbreaks of mountain pine beetle using partial differential equations," *Environmetrics*, 21, 801–816.