Polynomial nonlinear spatio-temporal integro-difference equation models

Christopher K. Wikle* and Scott H. Holan

Spatio-temporal statistical models are ubiquitous in the environmental sciences. Often the underlying spatio-temporal process can be written as a hierarchical state-space process that utilizes prior understanding of the physical process to formulate models for the state transition. In situations with complicated dynamics, such as wave propagation, or population growth and dispersal models, parameterizing the transition function associated with the high-dimensional state process may prove difficult or even impossible. One approach to overcoming this difficulty is through the use of polynomial stochastic integro-difference equation (IDE) models. Here, in the context of discrete time and continuous space, complicated dynamics can be accommodated through redistribution kernels that are allowed to vary with space. To facilitate computation and dimension reduction, we consider a basis-function-expansion representation of this model. Even in this framework there are too many parameters to estimate efficiently from a classical perspective. Therefore, we utilize the hierarchical Bayesian framework in order to implement a stochastic-search variable-selection algorithm that mitigates these estimation issues. Finally, the method is illustrated on the problem of long-lead prediction of equatorial sea surface temperature.

Keywords: Bayesian; hierarchical statistical model; quadratic nonlinearity; sea surface temperature; spectral; stochastic search variable selection.

1. INTRODUCTION

Spatio-temporal processes are the driving mechanism in the environmental sciences. Increasingly, deterministic and stochastic models have been applied to such processes so as to predict, forecast and draw inference. In the statistical sciences, interest in spatio-temporal processes has tended to focus in two primary areas: (i) descriptive, covariance-based joint models, and (ii) dynamic, conditionally-specified models (for a comprehensive overview see Cressie and Wikle, 2011). Our interest is with dynamic models because they tend to be more true to the etiology of the underlying natural process, and because they fit conveniently into a broader hierarchical statistical modelling construct.

Within the class of dynamic spatio-temporal models, most statistical applications have focused on linear or quasi-linear models (e.g. threshold models). However, given that many real-world environmental processes are inherently nonlinear in the sense that they are governed by interactions of random-process components, it is important to develop broader classes of nonlinear statistical models for spatio-temporal processes. Such models provide unique challenges in that they must be parameterized so that they can accommodate the real-world process. In addition, given the potential for a severe ‘curse of dimensionality’, one must take special care with parameter estimation. The focus of this article is to present a broad class of useful nonlinear spatio-temporal dynamic models as well as a coherent approach to deal with the curse of dimensionality.

Dynamic spatio-temporal models (DSTMs) can be difficult to implement, even in the context of linear evolution with Gaussian error processes. The primary reason for this is that the state-process dimension of interest is typically quite large, which suggests that inference on the associated linear propagator (i.e. transition matrix) requires estimation (or specification) of a large number of parameters. As shown in Wikle and Hooten (2010), this can be mitigated by the sensible use of science-based parameterizations such as suggested by theoretical process models (e.g. partial differential equations (PDEs), integro-difference equations (IDEs), and difference equations). Such science-based process models typically allow one to parameterize the dynamic model propagator in terms of a relatively few number of parameters. Furthermore, this approach is even more powerful when placed within a hierarchical modelling framework in which the parameters are endowed with dependencies at a lower stage of the hierarchy.

Wikle and Hooten (2010) also show that the general science-based framework can be used to motivate classes of Markovian nonlinear DSTMs. In particular, they describe how processes that exhibit what they describe as ‘general quadratic nonlinearity’ (GQN; cf. Section 2.1) can accommodate many real-world processes such as the spread of invasive species, and nonlinear advective processes in geophysics. Not surprisingly, the number of parameters associated with these processes is an order of magnitude larger than for linear DSTMs. As is the case for linear formulations, science-motivated parameterizations can dramatically reduce the parameter space. However, there are many instances where specific scientific parameterizations are not known (or are too complicated to implement), yet one has knowledge that certain types of interactions (e.g. quadratic, triad, etc.) are appropriate.
In that case, a sensible way to accommodate the curse of dimensionality is required. Wikle and Hooten (2010) considered a dimension-reduced state-process, based on a basis-function expansion, and then used simple arguments from turbulence theory to reduce the parameter space.

Although the GQN approach of Wikle and Hooten (2010) presents a powerful modelling approach for nonlinear spatio-temporal processes, there are some limitations and possible extensions. Their turbulence-based parameter reduction approach requires a somewhat arbitrary specification as to which basis-function expansion coefficients are allowed to interact quadratically (based on spatial scaling arguments). In addition, the process in that case is assumed to be discrete in space, yet many environmental spatio-temporal processes of interest can best be thought of as continuous spatial processes that evolve with time. Finally, we note that there are real-world processes that include higher-order interactions than quadratic (e.g., see the discussion of triad interactions in Majda et al., 2005). In such cases, the curse of dimensionality increases by an order of magnitude for each additional polynomial interaction.

This article presents a continuous-space, discrete-time (IDE) approach that allows for higher-order polynomial interactions. Basis expansion formulations are considered to simplify the IDE integrals and facilitate state-process dimension reduction and parameter dimension reduction. Additionally, we utilize a stochastic search variable selection (SSVS) mechanism to help deal with the curse of dimensionality in the parameter space. The remainder of this article is organized as follows. Section 2 gives the statistical modelling background for our approach. This is followed by the new methodology in Section 3. Section 4 provides an application of our methodology to the real-world problem of long-lead forecasting Pacific sea surface temperatures (SSTs). We conclude with a discussion in Section 5.

2. STATISTICAL MODELLING BACKGROUND

Here, we introduce notation for spatio-temporal processes as well as the general notion of a DSTM. Then, as motivation for our polynomial nonlinear model, we describe a class of nonlinear spatio-temporal models defined by GQN. This is followed by an introduction to continuous space DSTMs by way of stochastic IDEs. Finally, we provide some background on stochastic search variable selection (SSVS). The new methodology we discuss in Section 3 relies strongly on these three concepts: polynomial nonlinearity, stochastic IDEs, and Bayesian SSVS.

We define the discrete-time spatio-temporal process to be \( \{ Y_t(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^2, t \in T \} \), where \( D \) is the continuous spatial domain of interest (assumed here to be two-dimensional) and \( T = \{1, 2, \ldots \} \) is the discrete-time index set. We are interested in a dynamic model approach in which conditional probability distributions describe how \( Y_t(\mathbf{s}) \) behaves given the behaviour of ‘nearby’ (in space) past values of the process taken from the set \( \{ Y_r(\mathbf{x}) : \mathbf{x} \in D, r < t \} \). In this article, we will only consider first-order Markovian time dependence for this conditional model, in which case we are interested in the distribution \( \{ Y_t(\mathbf{s}) \mid Y_{t-1}(\mathbf{x}) : \mathbf{x} \in D, \theta \} \). However, we note that the methodology presented here could be readily extended to the case of higher-order Markovian dependence. We have also explicitly accounted for a conditional dependence on the parameters \( \theta \), which control the evolution through the propagator, and conditional dependencies in the error process.

2.1. Nonlinear dynamic spatio-temporal models

If interest is on the process at a finite collection of spatial locations, say \( Y_t = \{ Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_n) \} \), then we can write our Markovian model generally as a traditional nonlinear autoregressive model:

\[
Y_t = \mathcal{M}(Y_{t-1}; \eta_t; \theta); \quad t = 1, 2, \ldots,
\]

where \( \mathcal{M}() \) is a nonlinear operator, \( \{ \eta_t \} \) is an error process, and \( \theta \) are parameters (which may vary in time and/or space). Although a conceptually simple model, such a structure is typically too general, and thus we must further specify the form of \( \mathcal{M}(\cdot) \). For example, state-dependent models have been a particularly useful class of models in the time series literature (e.g., see the overviews in Priestley, 1988; Tong, 1990; Fan and Yao, 2005).

Wikle and Hooten (2010) show that a special class of nonlinear autoregressive models for spatio-temporal processes is given by GQN. The class of models they propose arise naturally since quadratic interactions are critical for many geophysical (Kondrashov et al., 2005; Majda et al., 2005) and ecological (Hooten and Wikle, 2008) processes. In particular, they consider the model

\[
Y_t(\mathbf{s}_i) = \sum_{j=1}^n a_{ij} Y_{t-1}(\mathbf{s}_j) + \sum_{k=1}^n \sum_{l=1}^n b_{ikl} Y_{t-1}(\mathbf{s}_k) g(Y_{t-1}(\mathbf{s}_l); \theta_{ijkl}) + \eta_t(\mathbf{s}_i),
\]

for \( i = 1, \ldots, n \), where \( a_{ij} \) and \( b_{ikl} \) are the linear and nonlinear evolution parameters, respectively, and \( \{ \eta_t(\mathbf{s}_i) \} \) is a correlated (in space) Gaussian noise process. The adjective ‘general’ is due to the potentially nonlinear transformation function \( g(\cdot; \theta_{ijkl}) \), which can give the model much more flexibility in accommodating real-world processes (e.g. growth processes). Critically, there are \( O(n^3) \) parameters in model (2), where \( \Phi(x) \) is an \( n \times n \) matrix of basis functions, and the expansion coefficient vector \( \alpha = (\alpha_1, \ldots, \alpha_n) \), and the range of the summations going to \( n \) instead of \( n \). Obviously, one still has \( O(n^2) \) parameters in this model, so direct estimation of parameters can be problematic unless \( n \) is exceptionally small relative to \( n \).
It is important to be clear as to what we mean by 'nonlinearity' in this methodology. In statistics, traditional nonlinearity refers to the parameters in a particular model (e.g. nonlinear regression. In the geophysical sciences, the notions of nonlinearity apply to the process. Specifically, nonlinearity refers to the interaction of some component of a process at the same instant in time, either in physical space or in some spectral representation of physical space. In our view, if the process is random and contains contemporaneous interactions in physical space or in the coefficients of a basis expansion, then we consider the process to be 'nonlinear'. It is possible that there could be parameter nonlinearity as well, as is the case with the GQN framework described above.

2.2. Integro-difference equation models

An alternative to the state-dependent discrete-space representation of the nonlinear AR process eqn (1) is given by the nonlinear stochastic IDE,

\[ Y_t(s) = \int_D k_s(u; \theta_g) g(Y_{t-1}(u); \theta_p) \, du + \eta_t(s); \quad s, u \in D, t \in T, \]

where \( g(\cdot; \theta_g) \) is a nonlinear function in \( \{Y_{t-1}(u) : u \in D\} \) and \( \theta_p \) and \( k_s(u; \theta_g) \) is the kernel that 'redistributes' the nonlinearly transformed lagged state-process forward to the current time. Deterministic IDEs have been used for quite some time in the ecological sciences to describe the change in the spatial distribution of a species over time (Kot, 1992; Kot et al., 1996; Lewis, 2000). For example, in the context of an invasive species, the growth of the population from one time to the next is controlled by \( g(\cdot; \theta_g) \), and the nature of its spread (dispersal or diffusion) is governed by the redistribution kernel \( k_s(\cdot; \cdot) \). The shape and speed of the dispersal depends on the kernel shape (i.e. kernel width and tail behaviour). In addition, non-diffusive propagation is dependent on the skewness of the kernel (e.g. Wikle, 2002). Such knowledge of the kernel shape relative to the propagation dynamics can lead to highly efficient parameterizations in a hierarchical context in which kernel parameters are allowed to vary with space (and/or time), and these spatially explicit parameters are modelled as spatial random fields at the next level of the model hierarchy (e.g. Wikle, 2002; Xu et al., 2005).

In many environmental applications, the effective dimensionality of the manifold on which the dynamics reside is substantially less than \( n \) (e.g. Majda et al., 2006). That is, the dynamics can be effectively characterized by relatively few 'spectral modes', or, as they might be described in statistics, as coefficients in the basis-function expansion. In that case, we might reasonably assume \( Y_t \sim \Phi \mathbf{x}_t \), where the dynamics are controlled by \( \mathbf{x}_t \). Note, additional 'non-dynamic' spatio-temporal variability in the process can be added if warranted (e.g. see Wikle and Cressie, 1999; Cressie et al., 2010), and this is typically accounted for by a spatial process that is independent across time or, as shown in Berliner et al. (2000), as spatial dependence in the conditional covariance in the data model (i.e. in the error process after conditioning the data on the true dynamic process of interest). In the linear case, one can show that the IDE model eqn (3) can be written in terms of the components of \( \mathbf{x}_t \) (e.g. Wikle and Cressie, 1999). Estimation can be carried out using traditional state-space models (e.g. Wikle and Cressie, 1999; Dewar et al., 2009; and Scerri et al., 2009), or in a hierarchical Bayesian framework through parameterization of the kernel (Wikle, 2002; Xu et al., 2005).

2.3. Stochastic search variable selection

In general, Bayesian SSVS algorithms provide an effective means of model selection when interest lies in considering a large number of potential submodels (e.g. see George, 2000, for a detailed overview.) That is, as popularized by George and McCulloch (1993, 1997), SSVS provides a quick and efficient way of performing variable selection within the hierarchical Bayesian framework. For example, assume we have a basic multiple regression problem with \( \mathbf{Y}_t \mid \mathbf{B}, \sigma^2 \sim N(\mathbf{X}_t \mathbf{B}, \sigma^2) \), where \( \mathbf{B} \) is an \( n_p \)-dimensional vector and \( n_p \) is quite large. One standard implementation then assumes the following prior specification for the elements of \( \mathbf{B}_j \):

\[ \beta_j \mid \gamma_j \sim \gamma_j N(0, c_j \tau^2_j) + (1 - \gamma_j) N(0, \tau^2_j), \quad j = 1, \ldots, n_p, \]

where \( \{\gamma_j\} \) are specified at the next level of the hierarchy to have independent Bernoulli distributions with probability parameters \( \{p_j\} \). In this case, \( p_j \) can be viewed as the prior probability that \( \beta_j \) should be included in the model. Furthermore, note that \( \gamma_j = 1 \) indicates that the \( j \)-th variable is included in the model. Typically, \( c_j, \tau_j, p_j \) are fixed hyperparameters. George and McCulloch (1993, 1997) provide various alternatives for the specification of these hyperparameters. They suggest that one would like \( \tau_j \) to be small so that, when \( \gamma_j = 0 \), it is reasonable to specify an effective prior for \( \beta_j \) that is near zero. In addition, one typically wants \( c_j \) to be large (>1) so that, if \( \gamma_j = 1 \), then our prior would favour a non-zero \( \beta_j \).

Similar to eqn (4) one could also impose a mixture prior on \( \beta_j \) composed of a dirac function at zero and a normal distribution; for example,

\[ \beta_j \mid \gamma_j, \sigma^2 \sim (1 - \gamma_j) \delta_0(\beta_j) + \gamma_j N(0, h_j \sigma^2), \quad j = 1, \ldots, n_p, \]

with \( \sigma^2 \) typically assumed to have a conjugate \( IG(\nu/2, \lambda/2) \) prior distribution and the hyperparameters \( \nu, \lambda \), and \( \{h_j : j = 1, \ldots, n_p\} \) are specified. Typically, one assumes hyperparameters for the parameter in eqn (5) such that the respective independent given \( \gamma \) and \( \sigma^2 \). Many alternative hyperparameter specifications have been proposed for specific applications (e.g. see Stingo and Vannucci (2011) and the references therein for an overview).

The Bayesian variable selection literature is extensive, with many variations on these basic methods having been proposed (e.g. see O’Hara and Sillanpää (2009) for a comprehensive review). The diverse approaches reflect the need for high-dimensional variable selection across a broad range of applications. For example, originally motivated by detecting differentially expressed genes in microarray experiments, Ishwaran and Rao (2003, 2005) proposed a 'spike and slab' model similar to eqn (5), but using a rescaled
response variable. Although this method has been shown to be effective, the rescaling assumes independence across the response variable and, thus, is not directly applicable in our setting. Another interesting approach is proposed by Stingo and Vannucci (2011), wherein the authors incorporate scientific information into eqn (5) by putting a Markov random field prior on \( \{ \gamma_i \} \). In principal, similar scientific information could be incorporated into eqn (4); see Section 3.2 for further discussion.

Among the several methods available for conducting Bayesian variable selection, which approach is ‘best’ for a given application? To a large extent, this depends on the goals of the analysis being conducted. For example, in the context of econometrics, George et al. (2008) proposed a Markov chain Monte Carlo (MCMC) algorithm that selects elements of both the VAR regression coefficient and the error variance matrices. In the spatio-temporal nonlinear forecasting problem, the goal is prediction-oriented. Thus, one reasonable strategy for SSVS hyperparameter specification that has proven effective when the main interest resides in predictive ability, is the use of eqn (4), coupled with sensitivity analysis driven by predictive results (e.g. Holan et al., 2010). This is the approach employed below in Section 4.2.

3. METHODOLOGY

Section 2 provided the necessary background on nonlinear DSTMs, IDEs, and SSVS. These three components make up the modelling framework that we develop in this article. First, we note that a natural extension to the GQN methodology, previously discussed, is to include higher-order polynomial interactions. For example, a model having \( p \)th order polynomial interactions can be specified as

\[
Y_t(s_k) = \sum_{j=1}^{n} \alpha_{1j} Y_{t-1}(s_k) + \sum_{j=1}^{n} \sum_{j=1}^{n} \alpha_{1j} Y_{t-1}(s_k) g_2(Y_{t-1}(s_k); \theta_{p_1}) \\
+ \sum_{j=1}^{n} \sum_{j=1}^{n} \sum_{j=1}^{n} \alpha_{1j} Y_{t-1}(s_k) g_3(Y_{t-1}(s_k); \theta_{p_1}) + \cdots \\
+ \sum_{j=1}^{n} \cdots \cdots \cdots \cdots \cdots \alpha_{1j} Y_{t-1}(s_k) g_p(Y_{t-1}(s_k); \theta_{p_1}) + \eta_t(s_k),
\]

for \( i = 1, \ldots, n \) and \( t = 1, 2, \ldots, T \). Although this ‘general polynomial nonlinear’ (GPN) framework provides a flexible and useful extension of GQN, this model does not directly accommodate continuous space. In many spatio-temporal problems that arise in the environmental sciences, one is interested in spatial prediction at any location within a specified spatial domain. As mentioned previously, the continuous-space, discrete-time IDE model is a particularly flexible model on which to parameterize DSTMs in this context. Unfortunately, the nonlinear form of the stochastic IDE shown in eqn (3) is far too general to be of direct use. Thus, our interest is to extend the quadratic nonlinear DSTM given in Wikle and Hooten (2010) to the case of polynomial nonlinearity in \( Y_{t-1}(u) \), for \( u \in \mathcal{D} \), but within the IDE framework. This is facilitated greatly by considering a basis-function-expansion representation of the process.

Consider the following polynomial stochastic IDE:

\[
Y_t(s) = \int_{\mathcal{D}} k_1^{(1)}(u_1) Y_{t-1}(u_1) \, du_1 + \int_{\mathcal{D}} \int_{\mathcal{D}} k_2^{(2)}(u_1, u_2) Y_{t-1}(u_1) Y_{t-1}(u_2) \, du_1 du_2 \\
+ \int_{\mathcal{D}} \int_{\mathcal{D}} \cdots \int_{\mathcal{D}} k_p^{(p)}(u_1, \ldots, u_p) Y_{t-1}(u_1) \cdots Y_{t-1}(u_p) \, du_1 \cdots du_p + \eta_t(s),
\]

for \( s, u_1, \ldots, u_p \in \mathcal{D}, t \in T \), where \( \{ k_i^{(l)}(u_1, \ldots, u_{l-1}) \} \) are \( l \)-dimensional kernel functions. Note, although we assume that each of these kernels could depend on parameters \( \theta_{p_2} \), we suppress the dependence on these parameters for notational simplicity. In addition, as in GQN, we could allow certain polynomial interactions of the \( Y \)-process to have one or more transformed components (i.e. \( Y_{t-1}(u) g(Y_{t-1}(u)) \)). However, such formulations lose the convenient simplification found in the basis-function expansion described above. Additionally, the non-transformed interactions are sufficiently flexible to accommodate many higher-order interaction processes (e.g. Majda et al., 2005). The form of eqn (6) is similar in spirit to the functional polynomial regression model of Yao and Müller (2010). The key difference is that eqn (6) represents a spatio-temporal process, and the nonlinearity is with respect to the past (lagged) values of the process \( Y \), not fixed covariates as in Yao and Müller (2010). That is, the model presented here represents Markovian interactions in a stochastic process.

A convenient way to deal with the integrals in eqn (6) is to consider an expansion of the process and kernels in terms of a complete and orthonormal basis set (e.g. Wikle and Cressie, 1999). Specifically, let

\[
Y_t(u) = \sum_{j=1}^{\infty} \phi_j(u) x_j(t),
\]

where \( \alpha(t) \) are mean-zero time series, and \( \{ \phi_i : i = 1, 2, \ldots \} \) are complete and orthonormal basis functions (i.e. \( \int \phi_i(\mathbf{u})\phi_j(\mathbf{u}) \text{d}\mathbf{u} = \delta_{ij} \) where \( \delta_{ij} = 0 \) for \( i \neq j \) and \( \delta_{ii} = 1 \) for \( i = j \)). Similarly, we expand the kernels,

\[
\begin{align*}
    k^{(1)}_s(\mathbf{u}_1) &= \sum_{j=1}^{\infty} b^{(1)}_{j_1}(\mathbf{s}) \phi_{j_1}(\mathbf{u}_1) \\
    k^{(2)}_s(\mathbf{u}_1, \mathbf{u}_2) &= \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} b^{(2)}_{j_1, j_2}(\mathbf{s}) \phi_{j_1}(\mathbf{u}_1) \phi_{j_2}(\mathbf{u}_2) \\
    &\vdots \\
    k^{(p)}_s(\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p) &= \sum_{j_1=1}^{\infty} \cdots \sum_{j_p=1}^{\infty} b^{(p)}_{j_1, \ldots, j_p}(\mathbf{s}) \phi_{j_1}(\mathbf{u}_1) \phi_{j_2}(\mathbf{u}_2) \cdots \phi_{j_p}(\mathbf{u}_p),
\end{align*}
\]

(8)

where \( b^{(l)}_{j_1, \ldots, j_p}(\mathbf{s}) \) are the expansion coefficients for the \( l \)th kernel with corresponding location \( \mathbf{s} \) and \( \phi \)'s denoting the orthonormal basis functions, as described above. In general, we do not require the same basis functions in the kernel expansion as we do in the process expansion; however, we would require biorthonormality to simplify calculations. It is important to note that without an explicit reason to justify preference toward biorthonormal basis functions for a particular application, we see little reason to deviate from a single set of orthonormal basis functions in this context. Furthermore, although we could consider non-separable multi-dimensional basis functions (e.g. \( \phi_{j_1, j_2}(\mathbf{u}_1, \mathbf{u}_2) \) would correspond to a \( j_1, j_2 \)-th two-dimensional spatial basis function for locations \( \mathbf{u}_1 \) and \( \mathbf{u}_2 \)), for simplicity we choose the product form (e.g. \( \phi_{j_1, j_2}(\mathbf{u}_1, \mathbf{u}_2) = \phi_{j_1}(\mathbf{u}_1)\phi_{j_2}(\mathbf{u}_2) \)). In this case, substituting the process and kernel expansions (eqns (7) and (8), respectively) into the polynomial IDE gives the infinite-basis-function-expansion form for the polynomial IDE model:

\[
Y_t(\mathbf{s}) = \sum_{j_1=1}^{\infty} \sum_{j_2=1}^{\infty} \cdots \sum_{j_p=1}^{\infty} b^{(p)}_{j_1, \ldots, j_p}(\mathbf{s}) \phi_{j_1}(\mathbf{u}_1) \phi_{j_2}(\mathbf{u}_2) \cdots \phi_{j_p}(\mathbf{u}_p)
\]

for \( \mathbf{s} \in D, t \in T \). Importantly, the orthonormality of the basis functions mitigates the integrals in this context.

Similar to applications in functional data analysis, we consider a finite approximation to the infinite summations in eqn (9) and assume that the order of truncation increases with the sample size. A good heuristic motivation for this truncation strategy can be found in Müller and Stadtmüller (2005); see Crainiceanu et al. (2009) for further discussion. In the case of spatio-temporal environmental processes, such a finite truncation is reasonable because the dynamic process of interest typically exists on a relatively low-dimensional manifold (e.g. for a recent discussion, see Majda et al., 2010). For example, we might consider the spatio-temporal process (say, \( Y_t'(\mathbf{s}) \)) to be decomposed into a spatio-temporal dynamical component (say, \( Y_t(\mathbf{s}) \)) and a non-dynamic residual component (say, \( \nu_t(\mathbf{s}) \)), where the dynamic component is considered in terms of the finite basis (spectral) expansion:

\[
Y_t'(\mathbf{s}) = Y_t(\mathbf{s}) + \nu_t(\mathbf{s})
\]

\[
= \sum_{j=1}^{n_t} \phi_j(\mathbf{s}) \alpha_j(t) + \nu_t(\mathbf{s})
\]

(10)

where \( \phi(\mathbf{s}) = (\phi_1(\mathbf{s}), \ldots, \phi_{n_t}(\mathbf{s}))' \) and \( \alpha_t = (\alpha_1(1), \ldots, \alpha_{n_t}(\mathbf{s}))' \). Note that it is often the case in spatio-temporal modeling of geophysical processes that one uses the term ‘spectral’ quite generally to refer to an expansion in terms of deterministic basis functions, similar in spirit to Galerkin approaches in the numerical solution of PDEs (e.g., see Cressie and Wikle, 2011, Chapter 7). As will be shown in eqn (12), one can consider the \( \nu_t(\mathbf{s}) \) portion of the process in the data model and then the process model can be written in terms of the \( \alpha_j(t) \) for \( j = 1, \ldots, n_t \) through the finite truncation of eqn (9). Specifically, the polynomial IDE in ‘finite spectral form’ is written:

\[
Y_t(\mathbf{s}) = \sum_{j_1=1}^{n_1} \sum_{j_2=1}^{n_2} \cdots \sum_{j_p=1}^{n_p} b^{(p)}_{j_1, \ldots, j_p}(\mathbf{s}) \phi_{j_1}(\mathbf{u}_1) \phi_{j_2}(\mathbf{u}_2) \cdots \phi_{j_p}(\mathbf{u}_p)
\]

(11)

for \( \mathbf{s} \in D, t \in T \).
### 3.1. State-space formulation

In general, one has observations \( \{ Z(i) : i = 1, \ldots, m \} \), which are ‘noisy’ and related, either linearly or nonlinearly, to the \( Y \)-process. Conditioned on \( Y \), the observation errors associated with \( Z \) can be Gaussian or non-Gaussian. Such observation models can be quite complex and accommodate biases and change-of-support (in space and/or time), as well as conditional dependencies (see the overview in Cressie and Wikle, 2011, Chapter 7). Since the focus of this article is on the nonlinear process \( Y \), we will assume fairly simple data/observation models. However, we recognize that hierarchical formulations allow for straightforward application (in principle) of substantially more complicated data models. As shown below, a critical component of this observation model is that it accommodates the basis-function expansion and associated dimension reduction, if the dynamics live on a lower-dimensional manifold.

After specification of the observations conditioned on the true process, we must specify a model for this process. It is at this stage that the polynomial nonlinear spatio-temporal IDE model is considered. Typically, one is interested in spatio-temporal prediction at a set of fixed spatial locations \( \{ s_k, \ldots, s_n \} \) at times \( t \in \{ 1, \ldots, T \} \). These locations may, but need not, coincide with the observation locations. Together, the observation model and process model combine to make a state-space model (see Shumway and Stoffer, 2006, for a comprehensive overview).

Let \( Z_t = (Z_t(r_1), \ldots, Z_t(r_m))^\top \) denote the \( m \)-dimensional observation vector at time \( t \). Similarly, define the \( n \)-dimensional process vector, \( Y_t = (Y_t(s_1), \ldots, Y_t(s_n))^\top \). The state-space model is then given as follows.

**Observation (data) model:**

\[
Z_t = a_t + H_t(Y_t + v_t) + \bar{e}_t, \quad \bar{e}_t \sim \text{Gau}(0, \mathbf{R}_t),
\]

\[
Z_t = a_t + H_t \mathbf{X}_t + e_t, \quad e_t \sim \text{Gau}(0, \mathbf{R}_t),
\]

for \( t = 1, 2, \ldots, T \), where \( v_t = (v_t(s_1), \ldots, v_t(s_n))^\top \), \( a_t \) is an \( m \)-dimensional ‘bias’ vector with \( H_t \), an \( m \times n \) observation matrix that maps the process to the (effectively bias-corrected) observations, and

\[
\Phi \equiv \begin{pmatrix}
\phi^t(s_1) \\
\vdots \\
\phi^t(s_n)
\end{pmatrix}
\]

is the matrix of basis functions for the \( n \)-dimensional process locations of interest. In this case, \( e_t = H_t v_t + \bar{e}_t \) is assumed to be a mean-zero Gaussian observation error process that includes measurement error \( (\bar{e}_t) \) and non-dynamic process variability \( (v_t) \), and it has variance-covariance matrix \( \mathbf{R}_t \). As noted above, other error distributions can be considered for the observation model (e.g. see Cressie and Wikle, 2011, Chapter 7), and one may model the \( v_t \) process explicitly (e.g. Wikle and Cressie, 1999; Cressie et al., 2010).

**Process model:**

Note that the finite-expansion-polynomial IDE model eqn (11) can be written more concisely as

\[
Y_t(s) = f(x_{t-1}; \mathbf{B}(s)) + \eta_t(s),
\]

where \( f(\cdot) \) represents the polynomial nonlinear terms on the right-hand side of eqn (11), which are a function of \( x_{t-1} \) and \( \mathbf{B}(s) \) (all of the expansion coefficients that are associated with location \( s \)). Thus, given that we are interested in the dynamic process at locations \( \{s_1, \ldots, s_n\} \), we can write the process model compactly in terms of \( x_t \) using eqns (10) and (15):

\[
\mathbf{X}_t = \mathbf{F}(x_{t-1}; \mathbf{B}) + \eta_t,
\]

where \( \mathbf{F} \) is a matrix operator with \( \mathbf{f}(x_{t-1}; \mathbf{B}(s)) \) defining the \( i \)-th row, and where \( \mathbf{B} = \{ \mathbf{B}(s_1), \ldots, \mathbf{B}(s_n) \} \), and \( \eta_t = (\eta_t(s_1), \ldots, \eta_t(s_n))^\top \). Critically, we once again make use of the orthogonality of the basis functions that make up the \( n \times n_x \) matrix \( \Phi \) to give the equivalent polynomial dynamic model in terms of the spectral process \( x_t \):

\[
x_t = \Phi' \mathbf{F}(x_{t-1}; \mathbf{B}) + \Phi' \eta_t.
\]

Typically, we will assume \( \eta_t \sim \text{Gau}(0, \mathbf{Q}) \) for these models, but note that other error distributions could be considered if desired. Finally, although it is possible to write out \( \mathbf{F} \) explicitly in terms of matrix and vector functions of \( x_{t-1} \) and \( \mathbf{B} \), we do not pursue this general exposition for the sake of brevity.

**Parameter and state-process estimation:**

In principle, standard methods for parameter and state estimation can be considered for the state-space model defined by eqns (13) and (17). These can include sequential importance sampling (SIS) methods (e.g. Doucet et al., 2001), Markov chain Monte Carlo (MCMC) methods (e.g. see the summary in Shumway and Stoffer, 2006, Chapter 6), or hybrid combinations of SIS and MCMC (Andrieu et al., 2010). However, as described in Wikle and Hooten (2010) and Cressie and Wikle (2011, Chapters 7 and 8), the curse of dimensionality in polynomial nonlinear DSTMs precludes efficient estimation for most problems of realistic interest in the environmental sciences. Even with the dimension reduction discussed above in terms of basis-function expansions, \( n_x \) must be extremely small to ensure efficient parameter estimation. As an alternative, in the next section, we describe a hierarchical SSVS approach.
3.2. Hierarchical SSVS parameterization

As described in Section 2.3, SSVS typically places a mixture prior on the parameters of a regression model. In our case, we simply allow each set of interaction parameters given in $B$ to have such priors. Specifically, let

$$b_{i}^{(1)}(s_{i}) | \gamma_{i}^{(1)} \sim N(0, \sigma_{i}^{(1)}(\gamma_{i}^{(1)})^{2}) + (1 - \gamma_{i}^{(1)})N(0, (\gamma_{i}^{(1)})^{2}),$$

$$b_{i}^{(2)}(s_{i}) | \gamma_{i}^{(2)} \sim N(0, \sigma_{i}^{(2)}(\gamma_{i}^{(2)})^{2}) + (1 - \gamma_{i}^{(2)})N(0, (\gamma_{i}^{(2)})^{2}),$$

for $i = 1, \ldots, n$, $j_{i} = 1, \ldots, n_{p}$, $l = 1, \ldots, p$. Then, we specify Bernoulli distributions,

$$\gamma_{i}^{(1)} \sim Bern(p_{i}^{(1)}),$$

$$\gamma_{i}^{(2)} \sim Bern(p_{i}^{(2)}).$$

As previously discussed, we typically assume these distributions of $\gamma$'s are independent and treat the $c$'s, $\tau$'s, and $\pi$'s as fixed hyperparameters, where the $\pi$'s represent the prior probability that a parameter is in the model. Recall that we want the $\tau$'s to be small and the $c$'s to be large so that when $\gamma$ is zero it is reasonable that the associated $b$-parameters should be close to zero, and when $\gamma$ is one, this prior favours non-zero $b$-parameters.

As alluded to in Section 2.3, having each of the $c$, $\tau$ and $\pi$ parameters vary by polynomial order, as well as spatial location $s$, could give the model a great deal of flexibility, so that one could specify subjective priors concerning anticipated dynamic interactions. This would be particularly useful if one considered a polynomial dynamical model that was based on some theoretical construct (e.g. a nonlinear advection-diffusion model) or, if the basis functions were ‘ordered’ (in the sense that $a_{x}^{(1)}(s_{i})$ results in the same values of $a_{x}^{(2)}(s_{i})$, but note that the more general methodology can be implemented fairly easily, but with substantially more tedious calculations.

4. LONG-LEAD FORECAST OF PACIFIC SEA SURFACE TEMPERATURE

The variation of tropical Pacific sea surface temperature (SST) is one of the most important sources of variability in the ocean/atmosphere system on interannual time scales (for background, see Philander, 1990). This variation is a result of a complicated interaction of the atmosphere and ocean across a wide range of spatial and temporal time scales. The most dominant signal in the Pacific SST corresponds to the El Niño-Southern Oscillation (ENSO) phenomenon. This consists of quasi-periodic warming (El Niño) and cooling (La Niña) in the tropical Pacific, with periods roughly from 3 to 5 years. Due to the extreme disruption of global atmospheric circulation patterns caused by these episodic warmings and coolings, there are significant impacts in weather across the globe. In that sense, it is important to be able to forecast the likely development of ENSO conditions several months in advance. Indeed, such ‘long-lead’ forecasts have been shown to have reasonable skill and represent one of the few situations in the atmospheric and oceanic sciences where purely statistical forecast methodologies are of value. If not better, than forecasts based on deterministic models (Barnston et al., 1999; van Oldenborgh et al., 2005).

Although dimension-reduced linear Markov models (i.e. VAR(1) models) have reasonable skill in long-lead forecasting of tropical Pacific SST (Penland and Magorian, 1993), the ENSO phenomenon is best characterized as a nonlinear process (e.g. Hoerling et al., 1997; Burgers and Stephenson, 1999). In that regard, nonlinear statistical models have been developed and typically are better at capturing the magnitude of the predicted El Niño and La Niña events (Tangang et al., 1998; Berliner et al., 2000; Tang et al., 2000; Timmermann et al., 2001; Kondrashov et al., 2005; Wikle and Hooten, 2010).

Kondrashov et al. (2005) demonstrate the effectiveness of a quadratic nonlinear model for long-lead prediction of ENSO from a classical regression perspective. Wikle and Hooten (2010) demonstrate the implementation of such a quadratic nonlinear model but in the context of a Bayesian hierarchical model framework in which the process is ‘hidden’ and parameters are random. Critically, these quadratic nonlinear implementations still suffer from a fairly high curse of dimensionality in the parameters. In this context, it is difficult to know a priori which quadratic interactions are important. To alleviate this curse of dimensionality and to provide a natural framework in which to ‘average’ across the various model specifications, we employ the SSVS methodology described in Section 3, applied to the long-lead SST forecasting problem. We note that this analysis is included simply as an illustration and therefore is not meant to be a complete study of SST long-lead forecasting.
4.1. Data

The data used in this example were anomalies from monthly averaged SSTs as described in Berliner et al. (2000). These anomalies are on a $2 \times 2^\circ$ resolution grid covering 2261 oceanic locations in the tropical Pacific over the region from 29S to 29N latitude and 124E to 70W longitude. The time period of interest includes January 1970–October 1997, with the purpose of producing a long-lead forecast of the development of the extraordinarily intense La Niña that peaked in the northern hemisphere autumn of 1988, and the intense El Niño that peaked in the autumn of 1997. In each case, the forecast is made given observations 7 months before the forecast target month. Because of their unusual intensity, these are classic ENSO cases that are often used to test new long-lead forecasting methodologies (e.g. Barnston et al., 1999; Berliner et al., 2000). The basis functions in this analysis are empirical orthogonal functions (EOFs; which are simply spatio-temporal principal components) calculated from the empirical spatial covariance matrix for the SST anomalies from January 1970 to March 1997 for the El Niño example and for January 1970 to February 1988 for the La Niña example. For detailed discussion on EOFs, see Jolliffe (2002), Wilks (2006), Cressie and Wikle (2011), and the references therein. In our example, we consider $n_s = 10$ basis functions and note that the first 10 EOFs account for about 88% of the variance in these SST anomalies. This choice was based on the previous studies by Berliner et al. (2000) and Wikle and Hooten (2010).

4.2. Model

Given the notation introduced in Sections 2 and 3, the complete hierarchical model for the long-lead SST forecasting example is given as follows:

$$ Z_t = \Phi x_t + e_t, e_t \sim \text{Gau}(\mathbf{0}, R), $$

$$ x_t = \Phi F(x_{t-1}; B) + \eta_t, \eta_t \sim \text{Gau}(\mathbf{0}, \mathbf{Q}), $$

$$ b_{j_k}^{(1)}(i) \sim N(\tilde{b}_{j_k}, \tilde{\sigma}_{b_k}^2), i = 1, \ldots, n_2; \quad j_k = 1, \ldots, n_s, $$

$$ b_{j_k}^{(2)}(i) | \gamma_{j_k}^{(2)} \sim N(0, \mathbf{c}_{j_k}^{(2)} (\gamma_{j_k}^{(2)})^2) + (1 - \gamma_{j_k}^{(2)}) N(0, (\gamma_{j_k}^{(2)})^2), $$

$$ i = 1, \ldots, n_s; j_k = 1, \ldots, n_s, $$

$$ \gamma_{j_k}^{(2)} \sim \text{Bern}(\pi_{j_k}^{(2)}), \quad j_k = 1, \ldots, n_s, $$

$$ \mathbf{Q}^{-1} \sim \text{Wishart}(\langle \mathbf{v} \mathbf{S} \rangle^{-1}, v), $$

and

$$ R = k_0 \left( \mathbf{c}_0 \mathbf{I} + \sum_{k=n_3+1}^{n_2+K} \hat{\lambda}_k \phi_k \phi_k^T \right), \quad k_0 \sim \text{IG}(q, r), $$

where $\Phi$ is given by eqn (14), $F(x_{t-1}; B)$ is given by eqn (16), and the following hyperparameters are used: $\tilde{b}_{j_k} = 0.9, \tilde{\sigma}_{b_k}^2 = 100, \mathbf{S} = \mathbf{I}$, $v = n_s = 10$, $C_0 = 0.005, K = 10$, and $q = 12, r = 9.09 \times 10^{-4}$ (corresponding to a mean of 100 and a variance of 1000). Note, the lag in this case corresponds to $t’ = 7$ months. Furthermore, following the approach of Berliner et al. (2000) and Wikle and Hooten (2010), the data-model covariance includes spatial structures associated with the next $K = 10$ (beyond the first $n_s$) EOFs, where $\hat{\lambda}_k$ is the associated eigenvalue. This formulation implicitly assumes that there is spatial structure in the higher-order EOFs, but that there is little spatio-temporal dynamical structure associated with these EOF components. The constant, $c_0$, is added to ensure that the matrix is not rank deficient and $k_0$ adds flexibility to accommodate the residual variance. Note also that we require the model include all linear interaction terms; consequently we fix $\gamma_{j_k}^{(1)} = 1, j_k = 1, \ldots, n_s$. In general, the results presented here are not overly sensitive to these choices for hyperparameters. That is, the long-lead predictions do not change appreciably when the hyperparameters are varied within reasonable ranges.

Because of the difficulty in knowing a priori the optimal choices for the SSVS parameters, we considered a sensitivity analysis including all possible combinations of $\pi^{(2)}_{j_k} = \pi \in \{0.1, 0.5\}$, $\mathbf{c}_{j_k}^{(2)} = c \in \{10, 100, 1000\}$, and $(\gamma_{j_k}^{(2)})^2 = \tau^2 \in \{0.001, 0.01, 0.1\}$. Specifically, we ran all 18 of these combinations for both the El Niño and La Niña 7-month-ahead forecasting cases. We evaluated the performance based on the model’s ability to forecast the average SST in a region of the Pacific known to be a strong indicator of ENSO activity (i.e. $5^\circ$–$5^\circ$N, $120^\circ$–$170^\circ$W); this region is sometimes referred to as the ‘Niño 3.4 region’. These results suggested $\pi = 0.1$ for both cases, but $c = 1000$ for the El Niño case and $c = 100$ for the La Niña case. Similarly, this procedure selected $\tau^2 = 0.001$ for the El Niño case and $\tau^2 = 0.1$ for the La Niña case.

The model was implemented in a Gibbs sampler. All updates were conjugate Gibbs steps apart from for the $x_t$-process values. These were updated using a random-walk Metropolis-Hastings update (with an acceptance rate of approximately 40%). Note that, for
the purposes of illustration, we assumed that \( \{x_t : t = 1, \ldots, 7\} \) were known and were given by \( \{\Phi Z_t : t = 1, \ldots, 7\} \). The results are based on 10,000 iterations for each case, of which the first 1000 were considered burn-in. Convergence of the MCMC was verified through visual inspection of trace plots of the sample chains.

4.3. Results

The quadratic interaction parameter matrices associated with the process evolution model are of interest here. Figures 1a and 1c show plots of the inclusion index for the quadratic interaction parameters (i.e. \( \gamma_j^{(2)} : j, j = 1, \ldots, n_2 \)) for the El Niño and La Niña cases respectively. In addition, the posterior means for the corresponding interaction parameters (i.e. \( b_j^{(2)} \)) are shown in Figures 1b and 1d, respectively. A couple of interesting features are evident from these plots. First, we note that the posterior likelihood of an interaction coefficient being in the model is very small for the La Niña case (i.e. all of the \( \gamma_j^{(2)} \) coefficients have posterior means very close to zero). In contrast, the posterior means for the \( \gamma_j^{(0)} \) coefficients for the El Niño case indicate that there are several nonlinear interactions that are frequently favoured in the model (i.e. they have posterior means close to 1). As is well known in the geophysical literature, the El Niño dynamical evolution is quite different than the La Niña evolution (e.g. Hoerling et al., 1997). Thus, we expect these interaction matrices to behave differently. Second, we note that in general, as expected, if an interaction is chosen fairly frequently for the model, its associated parameter is relatively large in absolute value as is evident in the El Niño example (Figure 1b). However, importantly, this is not always the case. For example, some interaction coefficients for the El Niño case that are included frequently in the model do not have coefficients with large posterior-mean magnitudes. On closer reflection, this is not unexpected as very small values of a given state variable, or interaction, can be critically important in nonlinear dynamical evolution (e.g. this is a critical component of chaotic systems; see the overview in Cressie and Wikle, 2011, Chapter 3). Furthermore, the posterior means for the La Niña case (Figure 1d) are relatively large for many parameters, even though the interactions all have small probability of being in the model. Note that in this case, most of these parameters have posterior credible intervals that cover zero (not shown).

In the context of the predictions, Figures 2a–2d show the SST anomalies as well as the forecast site-wise means and 2.5 and 97.5 percentiles from the posterior forecast distribution of \( \Phi x_{t+1} \) for October 1997, given observations through March 1997. Similarly, Figures 2e–2h show the analogous plots for the forecast for September 1988, given observations through February 1988. These plots show that the La Niña forecast did a reasonable job of capturing the intensity of the 1988 La Niña, and the forecast intervals include the true intensities. The El Niño case suggests the possibility of a strong El Niño, but it still under-forecasts the intensity, both in the mean and in the forecast uncertainty. We note that there was no attempt to ‘tune’ the model to optimize the forecast uncertainty in this simple illustration beyond the previously mentioned crude SSVS parameter sensitivity analysis. For example, different choices of SSVS prior structure and/or the number of basis-function coefficients considered in the dynamic and error portions of the model could be modified to improve forecast accuracy and/or uncertainty characterization.

El Niño and La Niña activity is often summarized in terms of various time series indices. A common index is the spatial average of SST anomalies over the so-called Niño 3.4 region (defined in Section 4.2). Thus, it is useful to see how our model is able to forecast this index, calculated by averaging the forecast SST anomalies in our data set over this region for each iteration of the MCMC. The results for the two cases are given in Table 1. In the case of October 1997, the forecast definitely suggests a strong El Niño (positive

![Figure 1](image-url)
values of the index), but the truth is not contained within the 2.5 percentile and 97.5 percentile of the posterior forecast distribution. However, the intensity of the October 1998 La Niña event is forecasted quite well, and the truth is within the 2.5 percentile and 97.5 percentile of the posterior forecast distribution. In both cases, the nonlinear model presented produced better forecasts than a purely linear model. These results are corroborated for the 1987 El Niño and 1998 La Niña as well (results not shown).

5. DISCUSSION

Spatio-temporal modelling of environmental processes has become an important research area in statistics over the last 10–15 years. Most of the models used in this context have been linear. However, given that most environmental processes are nonlinear, at least at some spatial and temporal scales of variability, it is critical that classes of nonlinear spatio-temporal models be developed. As discussed in Cressie and Wikle (2011), we are just now entering the development phase of these models in statistics. Thus, the purpose of this article is to illustrate a potentially important class of nonlinear dynamic spatio-temporal models that can be used for a variety of environmental processes.

The key components of the model class presented here are its formulation in continuous space, and its ability to incorporate polynomial interactions, which are known to be important in many environmental processes. At least in the case of geophysical
processes, polynomial interactions in spectral space are an important mechanism for energy transfer, and thus the formulation of polynomial models in this context is scientifically appealing (e.g. Majda et al. 2005, 2006, 2010). In addition, the basis-function representation is convenient both mathematically and computationally in the case of spatially continuous (IDE) dynamic models. Finally, despite the benefits from state-process dimension reduction in terms of finite approximations to infinite basis expansions, these models still suffer from the curse of dimensionality in that there are just too many parameters to estimate reliably using traditional frameworks. Consequently, we employ a stochastic search variable selection (SSVS) approach to these models within the Bayesian hierarchical framework.

The methodology is demonstrated on the important problem of long-lead prediction of Pacific sea surface temperatures. Although this example is not intended to be comprehensive, it does illustrate that there is some selection of important quadratic interactions by way of the SSVS implementation, and that the associated forecast uncertainty is well characterized by this model when it is applied to the classically difficult problem of long-lead prediction (i.e. the onset of the 1997 El Niño and 1998 La Niña events).

Acknowledgements

This work was made possible by NSF grants ATM-0434213, GEOSCI-0814934, and Awards N00014-07-1-0513 and N00014-10-1-0518 from the Physical Oceanography Program of the U.S. Office of Naval Research (ONR). The authors would like to thank Noel Cressie and two anonymous reviewers for providing many helpful suggestions to improve the article.

REFERENCES


11


