Hierarchical Models for Uncertainty Quantification: An Overview

7

Christopher K. Wikle

Abstract

Analyses of complex processes should account for the uncertainty in the data, the processes that generated the data, and the models that are used to represent the processes and data. Accounting for these uncertainties can be daunting in traditional statistical analyses. In recent years, hierarchical statistical models have provided a coherent probabilistic framework that can accommodate these multiple sources of quantifiable uncertainty. This overview describes a science-based hierarchical statistical modeling approach and the associated Bayesian inference. In addition, given that many complex processes involve the dynamical evolution of spatial processes, an overview of hierarchical dynamical spatio-temporal models is also presented. The hierarchical and spatio-temporal modeling frameworks are illustrated with a problem concerned with assimilating ocean vector wind observations from satellite and weather center analyses.

Keywords

Bayesian • Basis functions • BHM • Integro-difference equations • Latent process • Quadratic nonlinearity • MCMC • Multivariate • Ocean • Reduced-rank representation • Spatio-temporal • Wind

Contents

C.K. Wikle (✉)
Department of Statistics, University of Missouri, Columbia, MO, USA
e-mail: wiklec@missouri.edu

R. Ghanem et al. (eds.), *Handbook of Uncertainty Quantification*,
DOI 10.1007/978-3-319-12385-1_4

1 Introduction

Scientists and engineers are increasingly aware of the importance of accurately characterizing various sources of uncertainty when trying to understand complex systems. When performing statistical modeling on complex phenomena, the goal is typically either inference, prediction, or forecasting. To accomplish these goals through modeling, one must synthesize information. This information can come from a variety of sources, including direct (*in situ*) observations, indirect (remotely sensed) observations, surrogate observations, previous empirical results, expert opinion, and scientific principles. In order to make inferential or predictive decisions with a statistical model, one must consider these sources of information in a coherent manner that accounts adequately for the various sources of uncertainty that are present. That is, there may be measurement error, model representativeness error, error associated with differing levels of support between observations and process, parameterization error, and parameter uncertainty. Over the last 20 years or so, one of the most useful statistical paradigms in which to consider complex models in the presence of uncertainty is *hierarchical modeling* (HM). The purpose of this overview is to outline the general principles of science-based statistical HM and its utility to a wide class of processes.

Hierarchical modeling is, at its core, just a system of coherently linked probability relationships. In this sense, it is certainly not a new idea, and from a modeling perspective, such ideas have been at the core of fundamental statistical methods such as mixed models, structural equation models, spatial models, directed acyclic graph models, among others. This class of models might be referred to as "little h" hierarchical models. That is, one is either focused on a data model (i.e., "likelihood") and parameters, with the process considered a nuisance, or a data model and process model, with the parameters considered a nuisance. The perspective presented in this overview follows more closely the perspective originally outlined by Mark Berliner [4] in a somewhat obscure conference proceedings paper written while he was the director of the Geophysical Statistics Project at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado, USA. In this seminal paper, Berliner presents a simple, yet fundamentally important, way to think about partitioning uncertainty associated with data, processes, and parameters in complex systems. As described below, the basic tenet of this modeling paradigm is to characterize uncertainty in the joint model of data, process, and parameters in terms of component conditional and marginal distributions, which is often facilitated by the inclusion of scientific information. The advent of this formulation coincided with the so-called computational Bayesian revolution, specifically in terms of Markov chain Monte Carlo (MCMC) methods that were facilitated by the classic paper of [10]. This understanding provided the practical tools necessary to implement these models in the Bayesian context. One of the key components of thinking about models from this perspective is that one deliberately pushes complexity into the conditional mean, in which case subprocesses and parameters are often modeled with fairly complex dependence structures. This [4] hierarchical modeling paradigm might be referred to as a "big H" hierarchical model (HM), to emphasize that the conditional structure and parameter models are fundamental to the HM, not just a nuisance, and that scientific/mechanistic information is included in the various components of the HM.

The chapter begins with a general overview of hierarchical modeling and its Bayesian implementation. This is then followed by an overview of discrete-time spatio-temporal dynamical processes, given their importance as component models in many complex hierarchical modeling applications. A discussion of process and parameter space reduction is included in this overview of spatio-temporal processes. A simple illustrative example based on blending different sources of information for ocean surface vector winds is presented to highlight some of the important components of hierarchical modeling. Finally, a brief conclusion is presented that outlines the trade-offs one has to consider when building complex BHMs.

2 Hierarchical Modeling in the Presence of Uncertainty

This section presents a broad overview of statistical hierarchical modeling. The focus of this presentation is on the role of conditional models and, specifically, the separation of the joint model into coherently linked models for data, process, and parameters. This discussion follows similar discussions in [7, 8, 38], and [43].

To motivate the discussion of HMs, consider a problem in which one has observations of near-surface winds over the ocean from a satellite scatterometer and wishes to "predict" the distribution of complete spatial fields of the true wind across time. That is, there are satellite observations of east-west and north-south wind components that occur at a fairly fine spatial resolution, but are incomplete spatially due to the polar orbit of the satellite, and the goal is to interpolate the observations to form complete spatial fields for a regular sequence of times. In this case, the "process" corresponds to the wind components and, potentially, other relevant atmospheric state variables (e.g., sea-level pressure). In addition to the satellite observations, there is additional information from weather center "analysis" fields (i.e., model output from operational data assimilation systems that combine worldwide weather observations and deterministic weather forecasting models).

It is reasonable to assume that the satellite-derived scatterometer observations have not been used to produce the weather center data assimilation products.

For purposes of this general exposition, let the wind observations (data) be denoted by D. One possible approach to solving the aforementioned interpolation problem is to apply some deterministic curve-fitting interpolation algorithm $\hat{D} = f(D)$ (e.g., linear, polynomial, or spline interpolation). However, such approaches do not account for the uncertainty associated with the observations and, more importantly, do not utilize scientific knowledge to help fill in the data gaps in a physically plausible manner.

A more traditional statistical modeling alternative to this curve-fitting interpolation approach might consider a distribution for the data conditioned on some parameters, say θ_o, which is denoted by $[D \mid \theta_o]$. Note the use of a bracket notation for distribution, "[ ]," which is common in the hierarchal modeling literature, where the vertical bar, "|," denotes conditioning, $[A, B]$ represents the joint distribution of A and B, and $[A \mid B]$ represents the conditional distribution of A "given" B. In the traditional statistical model, one would seek the parameters, θ_o, that maximize the likelihood of observing the data D. Of course, this assumes that the distributional assumption adequately captures all of the variability (spatial, temporal, multivariate, etc.) in the data, subject to the correct specification of the parameters. Although this is very much a reasonable paradigm in many traditional statistical modeling problems, it is extremely tenuous in the example considered here (and, indeed, most complex physical, biological, or engineering problems) because it is typically not possible to adequately represent the complexity in the data via a single distributional assumption. In particular, this approach does not consider the fact that much of the complexity of the data arises from the scientific *process* (e.g., the atmospheric state variables in the wind example).

2.1 Basic Hierarchical Structure

A scientific modeling approach considers a model for the process of interest, say W here for "wind." Recognizing the fact that one's understanding of such scientific processes is always limited, this uncertainty is accounted for via a stochastic representation, denoted by the distribution $[W \mid \theta_W]$, where θ_W are parameters. The traditional statistical approach described above does not explicitly account for this uncertainty nor the uncertainty about the relationship between D and W. To see this more clearly, one might decompose the joint distribution of the data and the process given the associated parameters as

$$[D, W \mid \theta_D, \theta_W] = [D \mid W, \theta_D][W \mid \theta_W], \tag{7.1}$$

where the parameters in the conditional distribution of D given the process W are denoted by θ_D, which are different than the parameters θ_o for the marginal distribution of the data described above. That is, integrating out the process, W, from (7.1), gives $[D \mid \theta_o = \{\theta_D, \theta_W\}]$, which implies that the complexity associated with the process W is present in the marginal form of this data distribution and the associated parameters. Typically, this integration cannot be done analytically, and so one does not know the actual form of this marginal data likelihood nor could one generally account for the complicated multivariate spatio-temporal dependence in such a parameterization for real-world processes (e.g., nonlinearity, nonstationarity, non-Gaussianity). Even in the rare situation where one can do the integration analytically (e.g., Gaussian data and process models), the marginal dependence structure is typically more complicated than can be captured by traditional spatio-temporal parameterizations, that is, the dependence is some complicated function of the parameters θ_D and θ_W. Perhaps more importantly, in the motivating application considered here, the interest is with W, so one does not want to integrate it out of (7.1). Indeed, one typically wants to predict this process distribution. This separation between the data model conditional on the process and the process model is exactly the paradigm in traditional state-space models in engineering and time-series applications [e.g., 29]. More generally, the trade-off between considering a statistical model from the marginal perspective, in which the random process (parameters) are integrated out, and the conditional perspective, in which complicated dependence structures must be parameterized, is just the well-known trade-off that occurs in traditional mixed-model analysis in statistics [e.g., 31].

The decomposition given in (7.1) above is powerful in the sense that it separates the uncertainty associated with the process and the uncertainty associated with the observation of the process. However, it does not factor in the uncertainty associated with the parameters themselves. Utilizing basic probability, one can always decompose a joint distribution into a sequence of conditional and marginal distributions. For example, $[A, B, C] = [A \mid B, C][B \mid C][C]$. Thus, the hierarchical decomposition can be written as

$$[D, W, \theta] = [D \mid W, \theta][W \mid \theta][\theta], \tag{7.2}$$

where $\theta = \{\theta_D, \theta_W\}$. This hierarchical decomposition is not unique, e.g., it is equally valid probabilistically to write $[A, B, C] = [C \mid B, A][A \mid B][B]$, but the decomposition in (7.2) is meaningful scientifically as it implies causality in the sense that the parameters drive the process and the process generates the data, etc. In addition, note that the distributions on the right-hand side (RHS) of (7.2) could be simplified such that $[D \mid W, \theta] = [D \mid W, \theta_D]$ and $[W \mid \theta_W]$, that is, it might be reasonable to assume conditional independence in the parameter decomposition. This is a modeling choice, but it is reasonable in this case based on how the individual data and process distributions were specified above. More generally, it is helpful to consider the following schematic representation of [4] when partitioning uncertainty in hierarchical decompositions as it provides a framework for building probabilistically consistent models:

$$\begin{aligned}[\text{data, process, parameters}] &= [\text{data}|\text{process, parameters}] \\ &\quad \times [\text{process}|\text{parameters}] \times [\text{parameters}].\end{aligned} \tag{7.3}$$

2.2 Data Models

Each of the stages of the hierarchy given in (7.3) can be decomposed into products of distributions or submodels. For example, say there are three datasets for the near-ocean surface wind process (W) denoted by $D^{(1)}$, $D^{(2)}$, and $D^{(3)}$. These might correspond to the satellite scatterometer data mentioned previously, ocean buoy data, and the weather center analysis data product. These observations need not be coincident nor even of the same spatial or temporal support as the other data nor the process. In this case, the data model might be represented as

$$[D^{(1)}, D^{(2)}, D^{(3)} \mid W, \theta_D] = [D^{(1)} \mid W, \theta_D^{(1)}][D^{(2)} \mid W, \theta_D^{(2)}][D^{(3)} \mid W, \theta_D^{(3)}], \qquad (7.4)$$

where the parameters for each submodel are given by $\theta_D = \{\theta_D^{(1)}, \theta_D^{(2)}, \theta_D^{(3)}\}$. The RHS of (7.4) makes use of the assumption that the three datasets are all conditionally independent given the true process. This is not to say that the data are independent marginally, as they surely are not. Yet, the assumption of conditional independence is a powerful simplifying modeling assumption that is often reasonable in complex systems, but must be justified in practice. It is important to emphasize that the specific forms of the component distributions on the RHS of (7.4) can be quite different from each other, accounting for the differing support and measurement properties associated with the specific dataset. For example, satellite scatterometer wind observations have fairly well-known measurement-error properties and are associated with fairly small areal “footprints” (depending on the specific instrument), but wind observations from an ocean buoy are best considered point-level support with well-calibrated measurement-error properties.

2.3 Process Models

Typically, the process model in the hierarchical decomposition can also be further decomposed into component distributions. For example, in the case of the wind example described here, the wind process is a vector composed of two components, speed and direction or, equivalently, north-south and east-west components that depend on pressure. That is, one might write

$$[W^{(1)}, W^{(2)}, W^{(3)} \mid \theta_W] = [W^{(1)}, W^{(2)} \mid W^{(3)}, \theta_W^{(1,2)}][W^{(3)} \mid \theta_W^{(3)}], \qquad (7.5)$$

where $W^{(1)}$ and $W^{(2)}$ correspond to the east-west and north-south wind components (typically denoted by u and v, respectively) and $W^{(3)}$ corresponds to the near-surface atmospheric pressure (typically denoted P). The decomposition in (7.5) is not unique, but is sensible in this case because there is strong scientific justification for conditioning the wind on the pressure [e.g., 14]. The process parameters are again decomposed into those components associated with each distribution, $\theta_W = \{\theta_W^{(1,2)}, \theta_W^{(3)}\}$. The decomposition in (7.5) simplifies the joint dependence structure between the various process components by utilizing simplifying assumptions based on scientific input. It is important to recognize that these components are still distributions, so that the uncertainties in the relationships (say, between wind and pressure) can be accommodated through appropriate modeling components (e.g., bias and error terms).

Other types of joint interactions in the process can also be simplified through such conditional probability relationships. For example, given that the wind process is time varying, one might be able to make Markov assumptions in time. For example, if W_t corresponds to the wind process at time t for $t = 0, \ldots, T$, then

$$[W_0, W_1, \ldots, W_T \mid \theta_W] = \prod_{t=1}^{T} [W_t \mid W_{t-1}, \theta_W][W_0], \tag{7.6}$$

represents a first-order Markov assumption, that is, the process is independent of the past if conditioned on the most recent past. This is a significant simplifying assumption, and must be justified in practice, but such assumptions are often very realistic for real-world time-varying processes. Similar sorts of conditioning arguments can be made for networks, spatial processes (e.g., Markov random fields), and spatio-temporal processes (e.g., spatio-temporal dynamical models) as described in [7].

2.4 Parameter Models

An important consequence of the hierarchical modeling paradigm described above is the recognition that additional complexity can be accommodated by allowing the parameters to be random and endowing them with dependence structures (e.g., multivariate, spatial, temporal, etc.). That is, the parameter models can themselves be quite complex and can incorporate additional information, whether that be through exogenous data sources (e.g., a sea-surface temperature index corresponding to the El Niño/La Niña phenomenon) or scientific knowledge (e.g., spatial turbulent scaling relationships). For example, one might write $[\theta_W | X, \theta_X]$, where X is some exogenous covariate and θ_X are parameters. It can be very difficult, if not impossible, to account for such complex parameter dependence structures in the classical modeling approach discussed above.

Now, one must decide how to account for the uncertainty in X and θ_X, which often leads to yet another data or parameter level of the model hierarchy. Typically, at some point, there is no more information that can assist the specification of these distributions, and one either assigns some sort of non-informative distribution to the parameters or, in some cases, estimates them through some other means.

It is apparent that the distinction between "process" and "parameter" may not always be precise. This can be the case in some applications, but the strength of the hierarchal paradigm is that it is the complete sequence of the hierarchical decomposition that is important, *not* what one calls "process" or "parameter."

This suggests that one requires a flexible inferential paradigm that allows one to perform inference and prediction on both process and parameters and even their joint interaction.

2.5 Bayesian Formulation

The Bayesian paradigm fits naturally with hierarchical models because the posterior distribution is proportional to the product distributions in the hierarchical decomposition. For example, in the schematic representation of [4] given in (7.3), the posterior distribution can be written via Bayes' rule as

$$\begin{aligned}[\text{process, parameters} \mid \text{data}] \;\propto\; & [\text{data}|\text{process, parameters}] \\ & \times [\text{process}|\text{parameters}] \times [\text{parameters}], \qquad (7.7)\end{aligned}$$

where the normalizing constant is the integral (in the case of continuous distributions) of (7.3) with respect to the process and parameters (i.e., the marginal distribution of the data). In the context of the wind example, the posterior distribution can be written

$$[W, \theta_W, \theta_D \mid D] \;\propto\; [D \mid W, \theta_D][W \mid \theta_W][\theta_D, \theta_W]. \qquad (7.8)$$

In practice, it is not typically possible to calculate the normalizing constant ($1/[D]$) analytically. With the understanding that Markov chain Monte Carlo (MCMC) methods could be used generally for such purposes (i.e., after the seminal paper of [10]), this has not been a serious limitation.

MCMC methods seek to draw simulation samples from a distribution that coincides with the posterior distribution of interest. In particular, a Markov chain is constructed algorithmically such that samples from the stationary distribution of the Markov chain correspond to samples from the desired posterior distribution. Details of the implementation of such algorithms are beyond the scope of this overview, but they can be found in references such as [25] and [6]. Alternatively, approximate solutions can sometimes be found with less computational burden, such as with variational methods, approximate Bayesian computation (ABC), and integrated nested Laplace approximations (INLA) [e.g., 21, 27, 30]. In general, one must find trade-offs between model complexity and computational complexity when building complex statistical models in the presence of uncertainty (see the Conclusion of this chapter).

In some simpler modeling situations (e.g., state-space models), one might be content with assuming the parameters are fixed but unknown rather than assign them distributions. In that case, one could write (7.8) as

$$[W \mid D, \theta_W, \theta_D] \;\propto\; [D \mid W, \theta_D][W \mid \theta_W]. \qquad (7.9)$$

In applications where the component models are not too complex, these parameters can be estimated using classical statistical approaches, and then the parameters are used in a "plug-in" fashion in the model. For example, in state-space modeling, one might estimate the parameters through an E-M algorithm and then evaluate the process distributions through a Kalman filter/smoother [e.g., 29]. Such an approach is sometimes called "empirical Bayes" or, in the context of hierarchical models, empirical hierarchical modeling (EHM) [e.g., 7]. A potential concern using such an approach is accounting for the uncertainty in the parameter estimation. In some cases, this uncertainty can be accounted for by Taylor approximations or bootstrap resampling methods [e.g., 29]. Typically, in complex models, the BHM framework provides a more sensible approach to uncertainty quantification than EHM approaches.

3 Dynamical Spatio-temporal Process Models

The motivating wind example discussed above can be thought of as a data assimilation (DA) problem. [33] characterize DA as a set of methods that blend observations with prior system knowledge in an optimal way in order to obtain a distributional summary of a process of interest. In this context, "system knowledge" can correspond to deterministic models, scientific/mechanistic relationships, model output, and expert opinion. As summarized in [33], there is a large literature in the physical sciences dedicated to various methods for DA. In many ways, this is just a type of inverse modeling, and many different solution approaches are possible. However, if DA is considered from a BHM perspective, then one can gain a more comprehensive characterization of the uncertainty associated with the data, process, and parameters. From a statistical perspective, these methods typically require a dynamical spatio-temporal model (DSTM) of some sort. Hence, this section gives a brief overview of hierarchical DSTMs. More complete details can be found in [7] and [40]. This overview considers only DSTMs from a discrete-time perspective for the sake of brevity. However, it should be noted that many science-oriented process models are specified from a continuous time perspective (e.g., differential equations) and these can be used either to motivate HMs or can be implemented directly within the HM framework (e.g., [4]).

The data model in a general DSTM can be written

$$Z_t(\cdot) = \mathcal{H}(Y_t(\cdot), \boldsymbol{\theta}_d(t), \epsilon_t(\cdot)), \quad t = 1, \ldots, T,$$

where $Z_t(\cdot)$ corresponds to the data at time t and $Y_t(\cdot)$ is the corresponding latent process of interest, with a linear or nonlinear mapping function, $\mathcal{H}(\cdot)$, that relates the data to the latent process. The data model error is given by $\epsilon_t(\cdot)$, and data model parameters are represented by $\boldsymbol{\theta}_d(t)$. These parameters may vary spatially and/or temporally in general. As discussed more generally above, an important assumption that is present here, and in many hierarchical representations of DSTMs, is that the data $Z_t(\cdot)$ are independent in time when conditioned on the true process, $Y_t(\cdot)$

and parameters $\boldsymbol{\theta}_d(t)$. Thus, the observations conditioned on the true process and parameters can be represented

$$\prod_{t=1}^{T}[Z_t(\cdot) \mid Y_t(\cdot), \boldsymbol{\theta}_d(t)].$$

The key component of the DSTM is the dynamical process model. As discussed above, one can simplify this by making use of conditional independence through Markov assumptions. For example, a first-order Markov process can be written as

$$[Y_t(\cdot)|Y_{t-1}(\cdot), \ldots, Y_0(\cdot), \{\boldsymbol{\theta}_p(t), t = 0, \ldots, T\}] = [Y_t(\cdot)|Y_{t-1}(\cdot), \boldsymbol{\theta}_p(t)],$$

for $t = 1, 2, \ldots$ so that

$$\begin{aligned}[Y_0(\cdot), Y_1(\cdot), \ldots, Y_T(\cdot)|\{\boldsymbol{\theta}_p(t), t = 0, \ldots, T\}] = &\prod_{t=1}^{T}[Y_t(\cdot)|Y_{t-1}(\cdot), \boldsymbol{\theta}_p(t)] \\ &\times\ [Y_0(\cdot)|\boldsymbol{\theta}_p(0)]. \qquad (7.10)\end{aligned}$$

Higher-order Markov assumptions could be considered if warranted by the specific problem of interest. Such relationships are critical for real-world spatio-temporal processes because they follow the etiology of process development.

Now, the modeling focus is on the component Markov models in (7.10). For example, a first-order process can be written generally as

$$Y_t(\cdot) = \mathcal{M}(Y_{t-1}(\cdot), \boldsymbol{\theta}_p(t), \eta_t(\cdot)), \quad t = 1, 2, \ldots, \qquad (7.11)$$

where $\mathcal{M}(\cdot)$ is the evolution operator (linear or nonlinear), $\eta_t(\cdot)$ is the noise (error) process, and $\boldsymbol{\theta}_p(t)$ are process model parameters that may vary with time and/or space. Typically, one would also specify a distribution for the initial state, $[Y_0(\cdot)|\boldsymbol{\theta}_p(0)]$.

The hierarchical model then requires distributions to be assigned to the parameters $\{\boldsymbol{\theta}_d(t), \boldsymbol{\theta}_p(t), t = 0, \ldots, T\}$. Specific distributional forms for the parameters (e.g., spatially or temporally varying, dependence on auxiliary covariate information, etc.) depend strongly on the problem of interest. Indeed, as mentioned above, one of the most critical aspects of complex hierarchical modeling is the specification of these distributions. This is illustrated below with regard to linear and nonlinear DSTMs.

3.1 Linear DSTM Process Models

In the case where one has a discrete set of spatial locations $D_s = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ of interest (e.g., a lattice or grid), the first-order evolution process model (7.11) can be written as

$$Y_t(\mathbf{s}_i) = \sum_{j=1}^{n} m_{ij}(\boldsymbol{\theta}_m) Y_{t-1}(\mathbf{s}_j) + \eta_t(\mathbf{s}_i), \tag{7.12}$$

for $t = 1, 2, \ldots$, with redistribution (transition) components $m_{ij}(\boldsymbol{\theta}_m)$ that depend on parameters $\boldsymbol{\theta}_m$. If interest is in continuous space and discrete time, one can also write this in terms of an *integro-difference equation (IDE)*

$$Y_t(\mathbf{s}) = \int_{D_s} m(\mathbf{s}, \mathbf{x}; \boldsymbol{\theta}_m) Y_{t-1}(\mathbf{x}) d\mathbf{x} + \eta_t(\mathbf{s}), \quad \mathbf{s}, \mathbf{x} \in D_s, \tag{7.13}$$

for $t = 1, 2, \ldots$, where $m(\mathbf{s}, \mathbf{x}; \boldsymbol{\theta}_m)$ is a *transition kernel* that gives redistribution weights for process at the previous time and $\eta_t(\mathbf{s})$ is a time-varying (continuous) spatial error process. Analogous stochastic partial differential equation models could be specified for continuous time and space.

Now, denoting the process vector $\mathbf{Y}_t \equiv (Y_t(\mathbf{s}_1), \ldots, Y_t(\mathbf{s}_n))'$, (7.12) can be written in vector/matrix form as a first-order vector autoregression (VAR(1)) DSTM

$$\mathbf{Y}_t = \mathbf{M}\mathbf{Y}_{t-1} + \boldsymbol{\eta}_t, \tag{7.14}$$

where the $n \times n$ transition matrix is given by $\mathbf{M}$ with elements $\{m_{ij}\}$ with the associated time-varying spatial error process given by $\boldsymbol{\eta}_t \equiv (\eta_t(\mathbf{s}_1), \ldots, \eta_t(\mathbf{s}_n))'$, which is typically specified to be zero mean and Gaussian, with spatial variance-covariance matrix $\mathbf{C}_\eta$. Usually, $\mathbf{M}$ and $\mathbf{C}_\eta$ are assumed to depend on parameters $\boldsymbol{\theta}_m$ and $\boldsymbol{\theta}_\eta$, respectively, to mitigate the curse of dimensionality that often occurs in spatio-temporal modeling. As discussed below, the parameterization of these matrices is one way that additional mechanistic information can be incorporated into the HM framework.

3.2 Nonlinear DSTM Process Models

Many mechanistic processes are best modeled nonlinearly, at least at some spatial and temporal scales of variability. A class of nonlinear statistical DSTMs can be specified to accommodate such processes with quadratic interactions. Such a *general quadratic nonlinear (GQN)* DSTM [35] can be written as

$$Y_t(\mathbf{s}_i) = \sum_{j=1}^{n} m_{ij} Y_{t-1}(\mathbf{s}_j) + \sum_{k=1}^{n} \sum_{\ell=1}^{n} b_{i,k\ell} Y_{t-1}(\mathbf{s}_k) g(Y_{t-1}(\mathbf{s}_\ell); \boldsymbol{\theta}_g) + \eta_t(\mathbf{s}_i), \tag{7.15}$$

where m_{ij} are the linear transition coefficients seen previously and quadratic interaction transition coefficients are denoted by $b_{i,k\ell}$. A transformation of one of the components of the quadratic interaction is specified through the function $g(\cdot)$, which can depend on parameters $\boldsymbol{\theta}_g$. This function $g(\cdot)$ is responsible for the "general" in GQN, and such transformations are critical for many processes such as density-dependent growth that one may see in an epidemic or invasive species population process. The spatio-temporal error process is again typically assumed to be independent in time and Gaussian with mean zero and a spatial covariance matrix. Note that the conditional GQN model is Gaussian, but the marginal model will not in general be Gaussian because of the nonlinear interactions.

3.3 Multivariate DSTM Process Models

There are three primary approaches to modeling multivariate spatio-temporal dynamical processes in statistics. An obvious approach is to simply *augment* the process vector (e.g., concatenating the process vectors for a given time) and then using one of the univariate models (such as described above) to model the evolution of the process. That is, if there are J processes given by $\{\mathbf{Y}_t^{(j)}\}, j = 1, \ldots, J$, then for time t one could write $\mathbf{W}_t \equiv (\mathbf{Y}_t^{(1)\prime}, \ldots, \mathbf{Y}_t^{(J)\prime})'$ and then evolve $\mathbf{W}_t$ as above. The simplicity of this approach is appealing, but it is often more difficult to incorporate scientific information into the process evolution. Perhaps more critically, this often leads to very high-dimensional process vectors, which compounds the curse of dimensionality issue that is endemic in spatio-temporal statistical modeling.

As discussed generally above, multivariate processes can be modeled hierarchically by using the law of total probability and applying some conditional independence assumptions. As a simple example, consider $J = 3$ processes for the component conditional distribution for time t given time $t - 1$ might be written as

$$\begin{aligned}[\mathbf{Y}_t^{(1)}, \mathbf{Y}_t^{(2)}, \mathbf{Y}_t^{3} | \mathbf{Y}_{t-1}^{(1)}, \mathbf{Y}_{t-1}^{(2)}, \mathbf{Y}_{t-1}^{3}] &= [\mathbf{Y}_t^{(1)} | \mathbf{Y}_t^{(3)}, \mathbf{Y}_{t-1}^{(1)}, \mathbf{Y}_{t-1}^{(2)}] \\ &\quad \times\ [\mathbf{Y}_t^{(2)} | \mathbf{Y}_t^{(3)}, \mathbf{Y}_{t-1}^{(1)}, \mathbf{Y}_{t-1}^{(2)}][\mathbf{Y}_t^{(3)} | \mathbf{Y}_{t-1}^{(3)}].\end{aligned}$$

That is, processes 1 and 2 are conditionally independent at time t given process 3 at time t and previous values of processes 1 and 2 at time $t - 1$, and process 3 at time t is conditionally independent of the others given its previous values. Such a model formulation has the advantage of being able to match up to mechanistic knowledge about the processes and their interactions. However, if such knowledge is not available, this conditional formulation is arbitrary (or there is no basis for the conditional independence assumptions), and such an approach is not recommended.

The third primary approach for modeling multivariate dynamical spatio-temporal processes is to condition the J processes on one or more latent processes, much like what is done in multivariate factor analysis. For a set of $K \leq J$ common latent dynamical processes, $\{\alpha_{\ell,t}^{(k)}\}$, which may or may not be spatially referenced, consider

$$Y_t^{(j)}(\mathbf{s}_i) = \sum_{\ell=1}^{n_\alpha} \sum_{k=1}^{K} h_{i,\ell}^{(jk)} \, \alpha_{\ell,t}^{(k)} + \eta_t^{(j)}(\mathbf{s}_i), \tag{7.16}$$

for $i = 1, \ldots, n$, $j = 1, \ldots, J$, where $h_{i,\ell}^{(jk)}$ are interaction coefficients that account for how the ℓth element of the kth latent process influences the jth process at location i. This is a powerful modeling framework, but the curse of dimensionality in parameter space can easily make this impracticable. In addition, care must be taken when modeling the latent processes, which is typically done at the next level of the model hierarchy, as there are identifiability problems between the h parameters at this level and potential dynamical evolution parameters for the α processes at the next level [see 7, Section 7.4.2, for more discussion].

3.4 Process and Parameter Reduction

As mentioned above, one of the greatest challenges when considering DSTMs in hierarchical settings is the curse of dimensionality associated with the process and parameter space. For the fairly common situation where the number of spatial locations (n) is much larger than the number of time replicates (T), even the fairly simple linear VAR(1) model (7.14) is problematic as there are on order n^2 parameters to estimate. This is compounded for the GQN model (7.15), which has on order n^3 free parameters and similarly for the multivariate model. To proceed, one must reduce the number of free parameters to be estimated in the model and/or reduce the dimension of the dynamical process. These two approaches are discussed briefly below.

3.4.1 Parameter Reduction

Very seldom would one estimate the full variance/covariance matrix ($\mathbf{C}_\eta$) in the DSTM. Rather, given that these are spatial covariance matrices, one would either use one of the common spatial covariance function representations (e.g., Matérn, conditional autoregressive, etc.; see Cressie and Wikle [7, Chapter 4]) or a spatial random effect representation (see the "Process Reduction" section below). Generally, the transition parameters in the DSTM require the most care. For example, in the case of the simple VAR model (7.14), one could parameterize the transition matrix $\mathbf{M}$ simply as a random walk (i.e., $\mathbf{M} = \mathbf{I}$), a spatially homogeneous autoregressive process (i.e., $\mathbf{M} = \theta\mathbf{I}$), or a spatially varying autoregressive process ($\mathbf{M} = \text{diag}(\boldsymbol{\theta}_m)$). The first two parameterizations are somewhat unrealistic for most real-world dynamical processes, and the latter, although able to accommodate non-separable spatio-temporal dependence, does not account for interactions dynamically across space and time. Although in the context of evolving a spectral latent process (see below), such models can be very effective.

More mechanistically realistic dynamical parameterizations in the context of physical space representations recognize that spatio-temporal interactions are crucial for dynamical propagation. For example, in the linear case, the asymmetry and rate of decay of the transition parameters relative to a location (say, $\mathbf{s}_i$) control propagation (linear advection) and spread (diffusion). This suggests that a simple *lagged-nearest-neighbor* parameterization can be quite effective. For example,

$$Y_t(\mathbf{s}_i) = \sum_{j \in \mathcal{N}_i} m_{ij} Y_{t-1}(\mathbf{s}_j) + \eta_t(\mathbf{s}_i), \tag{7.17}$$

where $\mathcal{N}_i$ corresponds to a prespecified neighborhood of location $\mathbf{s}_i, i = 1, \ldots, n$ and $m_{ij} = 0$ for all $\mathbf{s}_j \notin \mathcal{N}_i$. Such a parameterization reduces the number of parameters from $O(n^2)$ to $O(n)$. It can be shown that such a parameterization can be motivated by many mechanistic models, such as those suggested by standard discretization of differential equations (e.g., finite difference, Galerkin, spectral) [e.g., see 7, 35]. In these cases, the m_{ij} parameters in (7.17) can be parameterized in terms of other mechanistically motivated knowledge, such as spatially varying diffusion or advection coefficients [e.g., 16,17,32,37,45]. Mechanistically motivated parameterizations can also be applied to nonlinear and multivariate processes [35].

3.4.2 Process Rank Reduction

Useful process reductions can be formulated with the realization that the essential dynamics for spatio-temporal processes typically exist on a relatively low-dimensional manifold [e.g., 41]. This is helpful because instead of having to model the evolution of the n-dimensional process $\{\mathbf{Y}_t\}$, one can model the evolution of a much lower-dimensional (n_α) process $\{\boldsymbol{\alpha}_t\}$, where $n_\alpha << n$. Thus, consider a decomposition of $\mathbf{Y}_t$ [36] such that

$$\mathbf{Y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Phi}\boldsymbol{\alpha}_t + \boldsymbol{\Psi}\boldsymbol{\xi}_t + \boldsymbol{\nu}_t, \tag{7.18}$$

where $\boldsymbol{\mu}_t$ is an n-dimensional time-varying (potentially) spatial mean corresponding to large-scale non-dynamic features and/or covariate effects; $\boldsymbol{\Phi}$ is an $n \times n_\alpha$ matrix of basis vectors corresponding to the latent dynamical expansion coefficient process, $\{\boldsymbol{\alpha}_t\}$; and $\boldsymbol{\Psi}$ can either be an $n \times n_\xi$ basis function matrix corresponding to the latent process, $\{\boldsymbol{\xi}_t\}$, which typically is assumed to have different dynamical characteristics than $\{\boldsymbol{\alpha}_t\}$ or this component might account for non-dynamical spatial variability. The error process $\{\boldsymbol{\nu}_t\}$ is typically Gaussian and assumed to be mean zero with simple dependence structure. Note that a continuous space representation of this decomposition can be expressed in terms of IDEs [e.g., see 7, Section 7.1.3].

The evolution of the latent $\boldsymbol{\alpha}_t$ process can proceed according to the basic linear or nonlinear models described above. Even in this low-dimensional context, parameter space reduction may still be necessary, particularly the case in nonlinear models (e.g., there are on the order of n_α^3 free parameters to estimate in the GQN model). Mechanistic knowledge can again be used to motivate such parameterizations in some cases [11, 36], and/or model selection approaches can be used to reduce the parameter space, such as stochastic search variable selection [e.g., 34].

There are many choices for the basis vectors that make up $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$. It has become quite common in recent years to represent spatial processes in terms of basis decompositions, and there are many choices for these, such as orthogonal polynomials, empirical spectral decompositions (i.e., empirical orthogonal functions (EOFs)), stochastic optimals, balanced truncations, wavelets, splines, bisquare bases, Wendland bases, Moran's I bases, kernel convolution and "predictive process" bases, and dynamic factor bases [e.g., see the discussion in 7,39]. Each has its proponents, although it does not seem to matter too much in the spatial context which basis is used, so long as it can accommodate the appropriate variability. In the context of DSTMs, it can make more of a difference because it is important that interactions across spatial scales be allowed [e.g., 40]. This is more difficult to do in the standard “knot-based” representations (e.g., splines, kernel convolutions, predictive processes) in which case the $\boldsymbol{\alpha}_t$ coefficients are spatially referenced, but not necessarily multi-resolution. Most of the other basis representations are in some sense multi-scale, and the associated expansion coefficients $\boldsymbol{\alpha}_t$ are not indexed in space. However, the product of the basis times the expansion coefficients is spatially referenced, and more importantly, dynamical evolution in the DSTM can accommodate scale interactions. Note that the coefficients $\boldsymbol{\xi}_t$ associated with the matrix $\boldsymbol{\Psi}$ are typically specified to have much-simpler dynamical structure (if at all) as the assumption is that the controlling dynamics are associated with $\boldsymbol{\alpha}_t$. In general, the $\boldsymbol{\xi}_t$ coefficient portion of the expansion is used to accommodate extra-dynamical spatial variability and/or exogenous effects.

The projection of the process $\{\mathbf{Y}_t\}$ to the lower-dimensional manifold need not be linear as shown in (7.18). There are a bewildering number of choices for nonlinear dimension reduction, and some of them could potentially represent the dynamics more realistically (e.g., Laplacian eigenmaps [2], kernel principal components [e.g., 44], etc.). However, these methods are somewhat limited by a lack of uniqueness in the back projection of the expansion coefficients into physical space, which requires either some sort of ad hoc procedure or an additional modeling component in the HM.

In some cases, the process is so complicated that it might be very difficult to specify an adequate process model. If deterministic simulation models are available, it can sometimes be easier to incorporate the mechanistic information through a surrogate model or statistical emulator. That is, much like the design and analysis of computer modeling experiment literature [e.g., 12, 18, 28], one builds a statistical model for the fairly rich simulation output (in terms of spatio-temporal behavior) and uses that either as a black box [e.g., 15, 22] or to inform prior distributions for simpler mechanistically motivated DSTMs [e.g., 19]. In other cases, one can build simpler lower-dimensional emulators and link them together hierarchically to represent the dynamical process [e.g., 20]. It is important to note that emulators in the context of dynamical spatio-temporal processes typically are built from what [15] call the “first-order” perspective. That is, process evolution is accounted for explicitly in the conditional mean structure, following the etiology of the real-world process. This is unlike the design and analysis of computer modeling literature, which typically considers so-called “second-order” emulators, in which the focus is on the covariance structure. Such an approach is well suited for the model calibration problem.

4 Example: Near-Surface Winds Over the Ocean

To illustrate many of the HM concepts described above, consider the motivating example of prediction of complete near-surface wind fields from a blend of weather center analysis winds and satellite scatterometer winds (so-called *surface vector* *winds* (SVW)). Some relevant background on the problem is presented, followed by a fairly simple BHM illustration applied to this problem.

4.1 Surface Vector Wind Background

Near-surface ocean winds are a critical component of the atmosphere/ocean interface as they are directly responsible for the transfer momentum to the ocean and the wind speed modulates the exchanges of heat and freshwater to and from the upper ocean. The advent of spaceborne scatterometer instruments in the 1990s provided the first global wind fields, on daily timescales, from observations. Prior to these scatterometer instruments, ocean winds were largely inferred from global weather forecast models (so-called *analyses*). These analyses depend on sparse global network of in situ wind observations from buoys and ships of opportunity and blend them with a mechanistic model of the atmosphere. The practical spatial resolution of such winds is limited to the relatively large spatial and temporal scales of variability.

Scatterometer SVW observations are not direct measures of the wind [see 23, for a more detailed description]. The winds are derived from complicated ("geophysical model function") relationships concerning the roughness imparted on the ocean by surface capillary waves in response to the shear stress vector at the air-sea interface. Depending on the specific sensor, SVW estimates from scatterometers are accurate to within at least 2 ms^{-1} in speed and 30° in direction, and resolutions are on the order of 12.5–50 km for up to 90% global coverage on daily timescales. The SVW retrievals occur in swaths along the polar-orbiting satellite ground track, with varying swath widths depending on the instrument system. For the purposes of predicting complete spatial fields, it is important to note that because of the polar orbit (approximately 14 polar orbits per day), the swaths overlap at high latitudes and are separated by gaps in coverage at low latitudes. So, although there are gaps over a day, areas in which there are SVWs exhibit much finer spatial resolution of atmospheric wind features than the analysis wind products from the same period, which are complete in space but have much lower effective spatial feature resolution in general (i.e., an unrealistic kinetic energy spatial spectrum). The goal of a statistical data assimilation is then to blend the complete, but energy-deficient, weather center analyses with the incomplete, yet energy-realistic, SVW in order to provide spatially complete wind fields at sub-daily intervals while managing the uncertainties associated with the different data sources and the blending procedure.

Uncertainty management via BHM with process dynamics motivated by mechanistic models (i.e., leading order terms and/or approximations of the primitive equations) has been shown to be a very effective approach for this wind data assimilation problem [e.g., see the sequence of papers: 5, 13, 23, 24, 26, 36, 42]. In particular, these methods have been shown to be quite useful in the context of providing inputs to ocean forecasting systems such as the Mediterranean Forecast System (MFS) [23, 24].

The MFS produces 10-day forecasts for upper ocean fields every day. This forecast model resolves medium-scale (in time and space) features (e.g., *synoptic scale*) in the upper ocean, and the most uncertain parts of the forecast fields correspond to so-called *mesoscales* (i.e., hourly and 10–50 km scales). These are the primary scales of the upper ocean hydrodynamic instabilities driven by the surface wind. Thus, modeling uncertainty in the surface wind field can be an important means of quantifying uncertainty in the MFS ocean forecasts on the scales that are most important to daily users.

4.2 Ocean SVW BHM

[23] describe the details of a SVW BHM for the MFS, and [24] discuss the impacts of the resulting BHM SVW fields in an ensemble forecast methodology built around realizations from the posterior distribution for SVW from the BHM. The process model in [23] involves the leading-order terms in a Rayleigh friction equation (RFE) approximation at synoptic scales, with extra-spatial variability added to account for turbulent scaling relationships in the wind field. A critical component of the [23] model is that it is multivariate in terms of modeling the east-west (u) and north-south (v) wind components and surface pressure (all of which are spatio-temporal processes) such that the wind components are independently conditioned on the pressure, which is a reasonable and justifiable assumption to first order. However, higher-order interactions of wind components are most likely important even after conditioning on the pressure field, so the model presented here considers a multivariate low-rank representation of the residual wind components after accounting for potential pressure gradient effects as suggested by the RFE. The data, process, and parameter models are described below.

4.2.1 Data Models

Two sources of wind data are considered, along with sea-level pressure data. In particular, there are satellite wind observations from the QuikSCAT scatterometer and surface winds and pressures from an analysis by the European Centre for Medium-Range Weather Forecasts (ECMWF). In this simple illustrative application, the pressure will be considered "known," and only the wind components are modeled as a process, i.e., the pressure is used as an exogenous variable here. The wind data models are then:

$$
\begin{aligned}
\mathbf{d}_t^{Q_u} | \mathbf{u}_t, \sigma_Q^2 &\sim ind.\, Gau(\mathbf{H}_t^Q \mathbf{u}_t, \sigma_Q^2 \mathbf{I}), \\
\mathbf{d}_t^{Q_v} | \mathbf{v}_t, \sigma_Q^2 &\sim ind.\, Gau(\mathbf{H}_t^Q \mathbf{v}_t, \sigma_Q^2 \mathbf{I}), \\
\mathbf{d}_t^{E_u} | \mathbf{u}_t, \sigma_E^2 &\sim ind.\, Gau(\mathbf{H}_t^E \mathbf{u}_t, \sigma_E^2 \mathbf{I}), \\
\mathbf{d}_t^{E_v} | \mathbf{v}_t, \sigma_E^2 &\sim ind.\, Gau(\mathbf{H}_t^E \mathbf{v}_t, \sigma_E^2 \mathbf{I}),
\end{aligned}
$$

where $\mathbf{d}_t^{Q_u}$ and $\mathbf{d}_t^{Q_v}$ are m_t-dimensional vectors of scatterometer u-wind and v-wind observations, respectively, within a specified time window indexed by t; and $\mathbf{d}_t^{E_u}$ and $\mathbf{d}_t^{E_v}$ are ECMWF u-wind and v-wind component observations, respectively, within the same window. The spatially vectorized true wind process components are given by the n-dimensional vectors $\mathbf{u}_t$, $\mathbf{v}_t$. The mapping matrices for the scatterometer and ECMWF observations are given by $\mathbf{H}_t^Q$ and $\mathbf{H}_t^E$, respectively. In this case, these are just incidence matrices that map the observations to the nearest process grid location [see 7, Chapter 7 for details]. The measurement errors are assumed to have Gaussian distributions that are independent in space and time, conditioned upon the true process values. The measurement-error variances, σ_Q^2 and σ_E^2, correspond to scatterometer and ECMWF wind components, respectively. The conditional independence of these data models follows from the more general discussion above concerning the relative ease of incorporating multiple data sources in the BHM framework.

The wind data for February 2, 2005, are shown in Figs. 7.1 and 7.2. These plots show the QuikSCAT scatterometer and ECMWF analysis observations available within a window of ±3 h of t = 00:00, 06:00, 12:00, and 18:00 UTC ("Coordinated Universal Time"). The ECMWF analysis winds and pressures are specified on a $0.5^\circ \times 0.5^\circ$ spatial grid, and they are available at each time period for all locations. This grid is also used for the process vectors, $\mathbf{u}_t$ and $\mathbf{v}_t$. As described above, the QuikSCAT observations are available intermittently in space due to the polar orbit of the satellite, but at much higher spatial resolution (25 km) when they are available. Thus, the mapping matrices for the scatterometer data, $\mathbf{H}_t^Q$, are defined as incidence matrices such that all scatterometer observations within 0.25° of a process grid point, and within 3 h of time t, are associated with the wind process at that grid point and time.

4.2.2 Process Model

The wind component process models are specified as

$$\mathbf{u}_t \equiv \mathbf{D}_x \mathbf{p}_t \theta_{ux} + \mathbf{D}_y \mathbf{p}_t \theta_{uy} + \boldsymbol{\Phi}_u \boldsymbol{\alpha}_t \tag{7.19}$$

$$\mathbf{v}_t \equiv \mathbf{D}_x \mathbf{p}_t \theta_{vx} + \mathbf{D}_y \mathbf{p}_t \theta_{vy} + \boldsymbol{\Phi}_v \boldsymbol{\alpha}_t, \tag{7.20}$$

where $\mathbf{D}_x$ and $\mathbf{D}_y$ are matrix operators that give the x-direction- and y-direction-centered differences of the spatial field vector on which they operate, respectively, and $\mathbf{p}_t$ is the vectorized gridded ECMWF pressure data (assumed known here). In the context of the process rank reduction decomposition given in (7.18), $\boldsymbol{\Phi}_u$ and $\boldsymbol{\Phi}_v$ correspond to $n \times n_\alpha$ matrices of basis functions for the u- and v-wind components, respectively, with the common random reduced-rank latent process expansion coefficients, $\boldsymbol{\alpha}_t$. In addition, relative to (7.18), let $\boldsymbol{\mu}_{u,t} \equiv \mathbf{D}_x \mathbf{p}_t \theta_{ux} + \mathbf{D}_y \mathbf{p}_t \theta_{uy}$ and $\boldsymbol{\mu}_{v,t} \equiv \mathbf{D}_x \mathbf{p}_t \theta_{vx} + \mathbf{D}_y \mathbf{p}_t \theta_{vy}$, where these terms represent the importance of winds on the gradient of pressure, as controlled by the parameters $\boldsymbol{\theta} \equiv \{\theta_{ux}, \theta_{uy}, \theta_{vx}, \theta_{vy}\}$. This particular formulation does not include a separate small-scale spatial variability component (e.g., $\boldsymbol{\Psi}\boldsymbol{\beta}_t$ in (7.18)) for simplicity. [36] and [23] include such a term

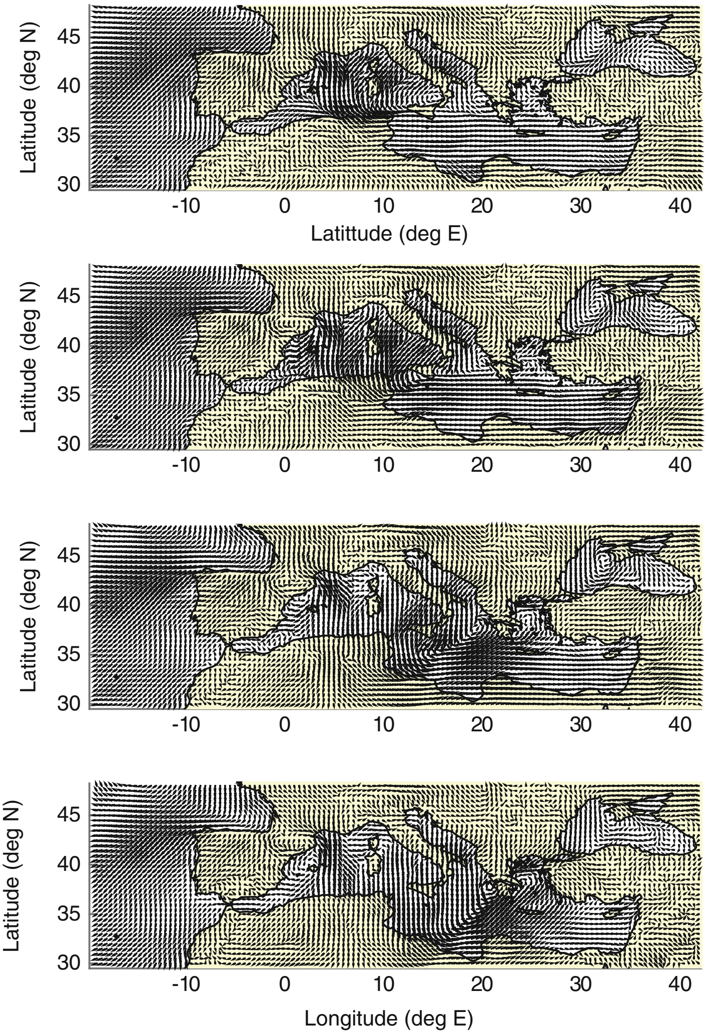

Fig. 7.1 Wind observations from February 2, 2005. From *top* to *bottom*, the panels correspond to the available data at 00:00, 06:00, 12:00, and 18:00 UTC (Universal Coordinated Time). The panels correspond to the ECMWF analysis winds on a 0.5° × 0.5° grid. The length of the wind quiver (arrow) corresponds to speed, where the smallest is 0.06 m/s and the largest is 17.7 m/s.

and parameterize it in terms of two-dimensional spatial wavelet basis functions to account for the turbulent scaling relationships that are inherent in the SVW.

Note that the basis function matrices, $\boldsymbol{\Phi}_u$ and $\boldsymbol{\Phi}_v$, are constructed from multivariate empirical orthogonal functions (EOFs) of the joint ECMWF u- and v-wind components [see 7, for an overview of EOF basis functions]. The advantage of such bases in this context is that they are constructed multivariately, so that the joint dependence of the wind components is considered in their construction. In

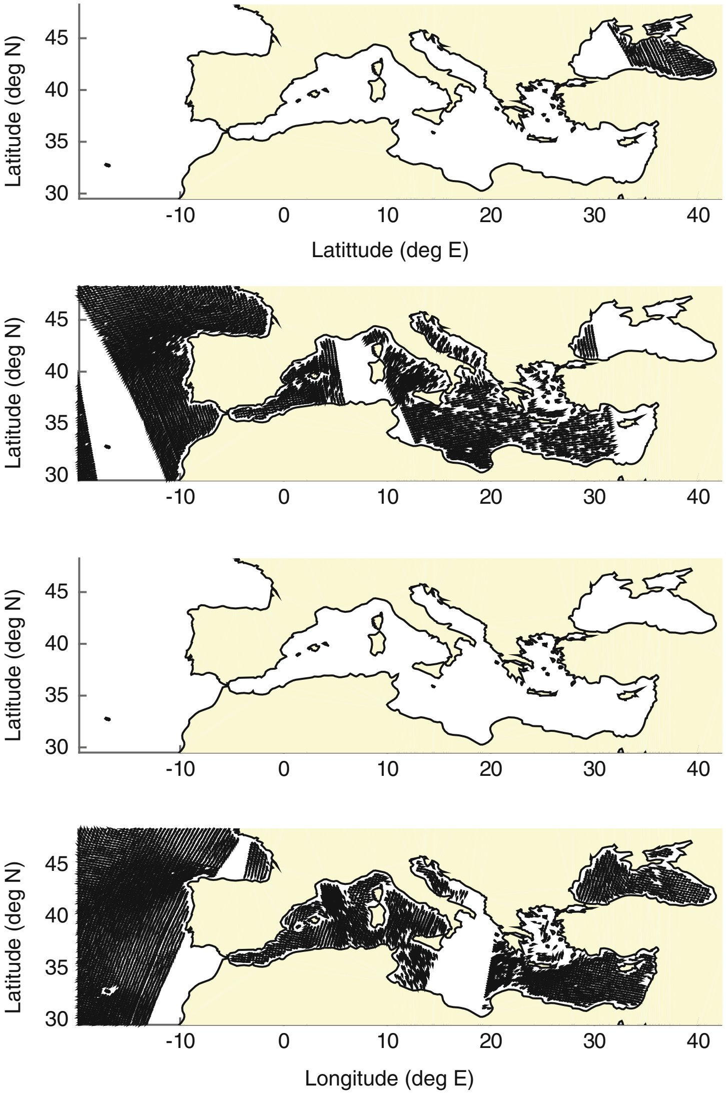

Fig. 7.2 Wind observations from February 2, 2005. From *top* to *bottom*, the panels correspond to the available data at 00:00, 06:00, 12:00, and 18:00 UTC (Universal Coordinated Time). The panels correspond to the high-resolution (25 km), but spatially intermittent, QuickSCAT scatterometer wind retrievals. The length of the wind quiver (arrow) corresponds to speed, where the smallest is 0.2 m/s and the largest is 21.3 m/s.

addition, EOFs generally are useful for dynamical reduced-rank modeling because the dimension reduction is quite significant (in the case here, $n = 4096$ and $n_\alpha = 32$, which accounts for approximately 98% of the variability in the ECMWF wind data). Although they can be quite useful for DSTM rank reduction, given that EOFs are essentially spatial principal component loadings, they are optimal for variance reduction but not typically for dynamical propagation. Regarding the notions of inclusion of mechanistic information in BHMs, note that by conditioning the wind components on common processes $\mathbf{p}_t$ and $\boldsymbol{\alpha}_t$, the process decomposition in (7.19) and (7.20) allows a reasonable mechanistic-based approach for building in the conditional independence between the wind components.

The dynamical evolution of the common latent process coefficients is specified fairly simply in this illustrative example as

$$\boldsymbol{\alpha}_t = \text{diag}(\mathbf{m}_\alpha)\, \boldsymbol{\alpha}_{t-1} + \boldsymbol{\eta}_t, \quad \boldsymbol{\eta}_t \sim \; Gau(\mathbf{0}, \mathbf{C}_\eta), \tag{7.21}$$

for $t = 1, \ldots, T$, where $\text{diag}(\mathbf{m}_\alpha)$ corresponds to an n_α-dimensional diagonal matrix with $\mathbf{m}_\alpha$ on the main diagonal and zeros elsewhere. The initial condition is specified as $\boldsymbol{\alpha}_0 \sim \; Gau(\mathbf{0}, \mathbf{C}_0)$. Note that this fairly simple dynamical structure is motivated by the components of the RFE described in [23] that do not depend on pressure. Marginal dependence between the elements that make up $\boldsymbol{\alpha}_t$ is accommodated by a non-diagonal variance-covariance matrix, $\mathbf{C}_\eta$.

4.2.3 Parameter Models

To facilitate computation for this simple illustrative example, the parameters in the previous stages are given conjugate prior distributions. In particular, specify $\theta_k \sim N(\mu_i, \sigma_i^2)$ for $i = \{ux, uy, vx, vy\}$, $\mathbf{m}_\alpha \sim N(\boldsymbol{\mu}_\alpha, \mathbf{C}_\alpha)$, and $\mathbf{C}_\eta^{-1} \sim W((d_\eta \mathbf{S}_\eta)^{-1}, d_\eta)$, where $W(\,)$ corresponds to a Wishart distribution. The remaining parameters and hyperparameters are fixed at scientifically plausible values (e.g., σ_Q^2, σ_E^2, μ_i, and $\boldsymbol{\mu}_\alpha$ as described in [23]) or given values to suggest vague (non-informative) priors (e.g., $\mathbf{C}_0$, $\mathbf{S}_\eta$, d_η, σ_i^2, $\mathbf{C}_\alpha$).

4.3 Implementation

The posterior distribution for the random components of the model is given by

$$\begin{aligned}
[\{\boldsymbol{\alpha}_t\}_{t=0}^T, \boldsymbol{\theta}, \mathbf{m}_\alpha, \mathbf{C}_\eta \mid \{\mathbf{d}_t^{Q_u}\}_{t=1}^T, \{\mathbf{d}_t^{Q_v}\}_{t=1}^T, \{\mathbf{d}_t^{E_u}\}_{t=1}^T, \{\mathbf{d}_t^{E_u}\}_{t=1}^T] \propto& \\
\prod_{t=1}^T [\mathbf{d}_t^{Q_u} \mid \boldsymbol{\alpha}_t, \boldsymbol{\theta}] \prod_{t=1}^T [\mathbf{d}_t^{Q_v} \mid \boldsymbol{\alpha}_t, \boldsymbol{\theta}]& \\
\times \prod_{t=1}^T [\mathbf{d}_t^{E_u} \mid \boldsymbol{\alpha}_t, \boldsymbol{\theta}] \prod_{t=1}^T [\mathbf{d}_t^{E_v} \mid \boldsymbol{\alpha}_t, \boldsymbol{\theta}]& \\
\times \prod_{t=1}^T [\boldsymbol{\alpha}_t | \boldsymbol{\alpha}_{t-1}, \mathbf{m}_\alpha, \mathbf{C}_\eta][\boldsymbol{\alpha}_0][\mathbf{m}_\alpha][\mathbf{C}_\eta][\boldsymbol{\theta}].&
\end{aligned}$$

Although an analytical posterior distribution is not available in this case, given the conjugate distributional forms, the required full-conditional distributions for a Gibbs sampler can all be derived analytically [e.g., see 7, for the details of a similar DSTM example]. In the example given here, the spatial grid sizes are $n = 4096$ and $T = 57$ times and were considered corresponding to the period from 12:00 UTC January 25, 2005, through 12:00 UTC February 8, 2005, at 6-h intervals. The reduced-rank vectors were of dimension $n_\alpha = 32$. The MCMC was run for 100,000 iterations after a 20,000-iteration burn-in period. The algorithm is quite efficient given the number of prediction locations 4096×57 and large amount of data (e.g., the MCMC was run in less than 4 h on a standard 2014 vintage laptop computer).

4.4 Results

The posterior mean wind fields for 12:00 UTC on February 2, 2005, are shown in Fig. 7.3. In addition, Fig. 7.4 shows a portion of the prediction domain, in which ten samples widely separated in the MCMC chain are plotted. One can see from this plot that the uncertainty associated with the spatial prediction of the wind fields is not homogeneous in the domain. For example, the strong flow off the south coast of France into the Gulf of Lion (so-called mistral winds) shows more variability in wind direction than areas over land. Note that this area of increased posterior variability is over a fairly small spatial region, which is important when one is using winds to force an ocean model such as with the MFS. That is, the small-scale variations in the wind forcing can lead to similar-scale uncertainties in the ocean state variables, which can make a substantial difference in ocean forecasts [see 24, for an in-depth discussion].

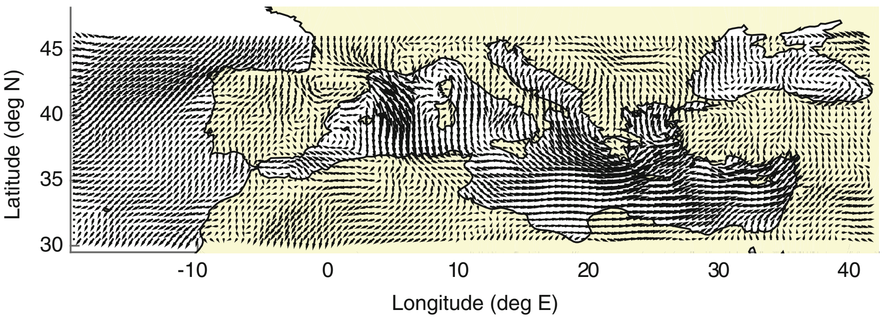

Fig. 7.3 Posterior mean wind vectors for 12:00 UTC on February 2, 2005, on the prediction grid. Wind speed is proportional to the length of the vectors, with the direction of wind toward the "arrowhead" on the wind quiver. The length of the wind quiver corresponds to speed, where the smallest is 0.0 m/s and the largest is 18.0 m/s.

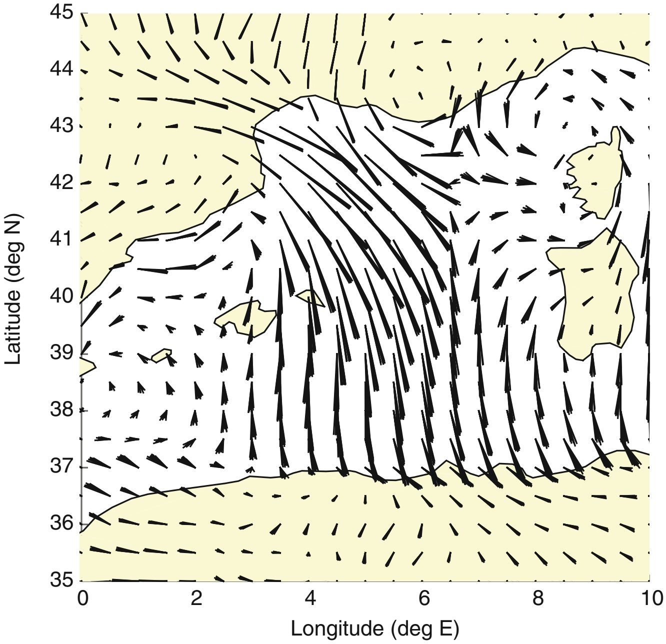

Fig. 7.4 Ten samples of wind vectors taken from the posterior distribution for 12:00 UTC on February 2, 2005, over the western Mediterranean basins. Wind speed is proportional to the length of the vectors. The direction of the vector is away from the vertex at the center of its grid cell. In this case, the "arrowhead" on the wind quivers is suppressed so that the variability in the wind corresponds to the width of the downwind portion of the vector.

5 Conclusion

This chapter has presented a brief summary of hierarchical modeling in the context of complex processes, typically those with mechanistically motivated spatio-temporal dependence. When modeling the complex processes one sees in many science and engineering applications, hierarchical modeling is a coherent approach to accommodate uncertainty in the observations (measurement error and sampling error), in the process specification, and in the knowledge of the parameters and potential additional forcings. The approach is very flexible, but with that flexibility comes potentially significant challenges and compromises when it comes to implementation.

Consider what [7] call the "data/model compromise." Even for complicated spatio-temporal processes, if one has *enough* (whatever that may be) high-quality observations, then the model can be fairly simple since the complex dependence structure is already contained in the observations and, presumably, can be "learned" by the statistical model. In many respects, this was the case with the SVW example presented above. There are a lot of observations from two different observation sources in this example, so the wind component process model is actually fairly simple relative to a mechanistic model that would be used with little or no data. In other SVW implementations, more sophisticated process models may have to be used depending on the data coverage and complexity of the dynamic environment [e.g., 36]. On the other hand, if one specifies a very complex mechanistic process model, but has very little data, there may not be enough information in the data to inform the posterior distributions associated with the parameters and process [e.g., 9]. That is, when the data are not rich enough to learn about the process and parameters, then one effectively has a practical lack of identifiability that may inhibit fitting the BHM. In practice, one tries to strike a balance between these two competing data/model trade-offs.

Perhaps the greatest challenge with implementing complex BHMs is recognizing the need to trade the complexity of the model for computational simplicity or what [7] call the "computing/model compromise." Despite the ever-advancing state of statistical computation for HMs, the algorithms can still be difficult to implement, both in terms of time required to code and the effort required to tune the algorithm. Software packages to implement BHMs are increasing in number and quality, but it is still often difficult to implement very complex BHMs with these packages. Thus, one is often faced with the dilemma of either simplifying the model and sacrificing some realism or utilizing an approximate estimation/inference approach (e.g., ABC, INLA, variational Bayes, etc.) and either limiting the sorts of inference that can be accomplished or accepting some inaccuracy relative to the true posterior distribution of interest. Thus, when implementing a complex BHM, one must always consider the difference between what one *wants* to do and what one *can* do and whether it is best for the particular problem at hand to sacrifice model complexity or computational efficiency. Regardless, the BHM paradigm still remains one of the most powerful frameworks in which to quantify uncertainty.

This chapter is concerned with science-based hierarchical modeling, in which one has mechanistic information available to inform the model components (either data, model, or parameters). In recent years, alternative hierarchical modeling approaches have been developed from the statistical learning perspective [e.g., see the review outlined in 3], which typically do not make use of scientific/mechanistic information, but seek to build multilayer models (e.g., "deep learning") through nonparametric approaches. These approaches can sometimes be useful in situations where subject-matter knowledge is not readily available yet can overfit in situations with complex spatial and temporal dependencies. In both the science-based and statistical learning-based HM approaches, much more work remains to be done on the theoretical properties of the estimators and predictors under various amounts of uncertainty in observations, process models and parameter structure, as well as data volume. Promising approaches are being developed [e.g., 1], but to date, these approaches have not been able to speak to the multilevel-dependent parameter structures common in the science-based BHM setting.

References

1. Agapiou, S., Stuart, A., Zhang, Y.X.: Bayesian posterior contraction rates for linear severely ill-posed inverse problems. J. Inverse Ill-Posed Probl. **22**, 297–321 (2014). doi:10.1515/jip-2012-0071
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neur. Comput. **15**(6), 1373–1396 (2003)
3. Bengio, S., Deng, L., Larochelle, H., Lee, H., Salakhutdinov, R.: Guest editors' introduction: special section on learning deep architectures. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1795–1797 (2013). doi:10.1109/TPAMI.2013.118
4. Berliner, L.: Hierarchical Bayesian time-series models. Fund. Theor. Phys. **79**, 15–22 (1996)
5. Berliner, L.M., Milliff, R.F., Wikle, C.K.: Bayesian hierarchical modeling of air-sea interaction. J. Geophys. Res. Oceans (1978–2012) **108**(C4), 2156–2202 (2003)
6. Brooks, S., Gelman, A., Jones, G., Meng, X.L.: Handbook of Markov Chain Monte Carlo. CRC Press, Boca Raton (2011)
7. Cressie, N., Wikle, C.: Statistics for Spatio-Temporal Data, vol. 465. Wiley, Hoboken (2011)
8. Cressie, N., Calder, C., Clark, J., Hoef, J., Wikle, C.: Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. Ecol. Appl. **19**(3), 553–570 (2009)
9. Fiechter, J., Herbei, R., Leeds, W., Brown, J., Milliff, R., Wikle, C., Moore, A., Powell, T.: A Bayesian parameter estimation method applied to a marine ecosystem model for the coastal gulf of Alaska. Ecol. Model. **258**, 122–133 (2013)
10. Gelfand, A.E., Smith, A.F.: Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**(410), 398–409 (1990)
11. Gladish, D., Wikle, C.: Physically motivated scale interaction parameterization in reduced rank quadratic nonlinear dynamic spatio-temporal models. Environmetrics **25**(4), 230–244 (2014)
12. Higdon, D., Gattiker, J., Williams, B., Rightley, M.: Computer model calibration using high-dimensional output. J. Am. Stat. Assoc. **103**(482), 570–583 (2008)
13. Hoar, T.J., Milliff, R.F., Nychka, D., Wikle, C.K., Berliner, L.M.: Winds from a Bayesian hierarchical model: computation for atmosphere-ocean research. J. Comput. Graph. Stat. **12**(4), 781–807 (2003)
14. Holton, J.: Dynamic Meteorology. Elsevier, Burlington (2004)
15. Hooten, M., Leeds, W., Fiechter, J., Wikle, C.: Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. J. Agric. Biolog. Environ. Stat. **16**(4), 475–494 (2011)
16. Hooten, M.B., Wikle, C.K.: A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove. Environ. Ecol. Stat. **15**(1), 59–70 (2008)
17. Hooten, M.B., Wikle, C.K., Dorazio, R.M., Royle, J.A.: Hierarchical spatiotemporal matrix models for characterizing invasions. Biometrics **63**(2), 558–567 (2007)
18. Kennedy, M.C., O'Hagan, A.: Bayesian calibration of computer models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **63**(3), 425–464 (2001)
19. Leeds, W., Wikle, C., Fiechter, J.: Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. Stat. Methodol. (2013). doi:10.1016/j.stamet.2012.11.004
20. Leeds, W., Wikle, C., Fiechter, J., Brown, J., Milliff, R.: Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators. Environmetrics **24**, 1–12 (2013). doi:10.1002/env.2187
21. Marin, J.M., Pudlo, P., Robert, C.P., Ryder, R.J.: Approximate Bayesian computational methods. Stat. Comput. **22**(6), 1167–1180 (2012)

22. van der Merwe, R., Leen, T., Lu, Z., Frolov, S., Baptista, A.: Fast neural network surrogates for very high dimensional physics-based models in computational oceanography. Neur. Netw. **20**(4), 462–478 (2007)
23. Milliff, R., Bonazzi, A., Wikle, C., Pinardi, N., Berliner, L.: Ocean ensemble forecasting. Part I: ensemble mediterranean winds from a Bayesian hierarchical model. Q. J. R. Meteorol. Soc. **137**(657), 858–878 (2011)
24. Pinardi, N., Bonazzi, A., Dobricic, S., Milliff, R., Wikle, C., Berliner, L.: Ocean ensemble forecasting. Part II: mediterranean forecast system response. Q. J. R. Meteorol. Soc. **137**(657), 879–893 (2011)
25. Robert, C., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer New York (2004)
26. Royle, J., Berliner, L., Wikle, C., Milliff, R.: A Hierarchical Spatial Model for Constructing Wind Fields from Scatterometer Data in the Labrador Sea. Lecture Notes in Statistics, pp. 367–382. Springer, New York (1999)
27. Rue, H., Martino, S., Chopin, N.: Approximate Bayesian inference for latent gaussian models by using integrated nested laplace approximations. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **71**(2), 319–392 (2009)
28. Sacks, J., Welch, W., Mitchell, T., Wynn, H.: Design and analysis of computer experiments. Stat. Sci. **4**(4), 409–423 (1989)
29. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and its Applications: With R Examples. Springer, New York (2010)
30. Šmídl, V., Quinn, A.: The Variational Bayes Method in Signal Processing. Springer, Berlin/New York (2006)
31. Verbeke, G., Molenberghs, G.: Linear Mixed Models for Longitudinal Data. Springer, New York (2009)
32. Wikle, C.: Hierarchical Bayesian models for predicting the spread of ecological processes. Ecol. **84**(6), 1382–1394 (2003)
33. Wikle, C., Berliner, L.: A Bayesian tutorial for data assimilation. Phys. D: Nonlinear Phenom. **230**(1), 1–16 (2007)
34. Wikle, C., Holan, S.: Polynomial nonlinear spatio-temporal integro-difference equation models. J. Time Ser. Anal. **32**(4), 339–350 (2011)
35. Wikle, C., Hooten, M.: A general science-based framework for dynamical spatio-temporal models. Test **19**(3), 417–451 (2010)
36. Wikle, C., Milliff, R., Nychka, D., Berliner, L.: Spatiotemporal hierarchical Bayesian modeling tropical ocean surface winds. J. Am. Stat. Assoc. **96**(454), 382–397 (2001)
37. Wikle, C.K.: A kernel-based spectral model for non-gaussian spatio-temporal processes. Stat. Model. **2**(4), 299–314 (2002)
38. Wikle, C.K.: Hierarchical models in environmental science. Int. Stat. Rev. **71**(2), 181–199 (2003)
39. Wikle, C.K.: Low-Rank Representations for Spatial Processes. Handbook of Spatial Statistics, pp. 107–118. CRC, Boca Raton (2010)
40. Wikle, C.K.: Modern perspectives on statistics for spatio-temporal data. Wiley Interdiscip. Rev. Comput. Stat. **7**(1), 86–98 (2015)
41. Wikle, C.K., Cressie, N.: A dimension-reduced approach to space-time Kalman filtering. Biometrika **86**(4), 815–829 (1999)
42. Wikle, C.K., Berliner, L.M., Milliff, R.F.: Hierarchical Bayesian approach to boundary value problems with stochastic boundary conditions. Mon. Weather Rev. **131**(6), 1051–1062 (2003)
43. Wikle, C.K., Milliff, R.F., Herbei, R., Leeds, W.B.: Modern statistical methods in oceanography: a hierarchical perspective. Stat. Sci. **28**(4), 466–486 (2013). doi:10.1214/13-STS436, http://dx.doi.org/10.1214/13-STS436
44. Wu, G., Holan, S.H., Wikle, C.K.: Hierarchical Bayesian spatio-temporal Conway–Maxwell poisson models with dynamic dispersion. Jo. Agri. Biol. Environ. Stat. **18**(3), 335–356 (2013)
45. Xu, K., Wikle, C.K., Fox, N.I.: A kernel-based spatio-temporal dynamical model for nowcasting weather radar reflectivities. J. Am. Stat. Assoc. **100**(472), 1133–1144 (2005)