Statistical Science
2013, Vol. 28, No. 4, 466–486
DOI: 10.1214/13-STS436
© Institute of Mathematical Statistics, 2013

Modern Statistical Methods in Oceanography: A Hierarchical Perspective

Christopher K. Wikle, Ralph F. Milliff, Radu Herbei and William B. Leeds

Abstract. Processes in ocean physics, air–sea interaction and ocean biogeochemistry span enormous ranges in spatial and temporal scales, that is, from molecular to planetary and from seconds to millennia. Identifying and implementing sustainable human practices depend critically on our understandings of key aspects of ocean physics and ecology within these scale ranges. The set of all ocean data is distorted such that three- and four-dimensional (i.e., time-dependent) in situ data are very sparse, while observations of surface and upper ocean properties from space-borne platforms have become abundant in the past few decades. Precisions in observations of all types vary as well. In the face of these challenges, the interface between Statistics and Oceanography has proven to be a fruitful area for research and the development of useful models. With the recognition of the key importance of identifying, quantifying and managing uncertainty in data and models of ocean processes, a hierarchical perspective has become increasingly productive. As examples, we review a heterogeneous mix of studies from our own work demonstrating Bayesian hierarchical model applications in ocean physics, air–sea interaction, ocean forecasting and ocean ecosystem models. This review is by no means exhaustive and we have endeavored to identify hierarchical modeling work reported by others across the broad range of ocean-related topics reported in the statistical literature. We conclude by noting relevant ocean-statistics problems on the immediate research horizon, and some technical challenges they pose, for example, in terms of nonlinearity, dimensionality and computing.

Key words and phrases: Bayesian, biogeochemical, ecosystem, ocean vector winds, quadratic nonlinearity, spatio-temporal, state–space, sea surface temperature.

1. INTRODUCTION

The global ocean dominates the iconic image of Earth viewed from space, leading to the now famous "blue marble" descriptor for our planet. The ocean covers more than 70% of the planetary surface and ocean processes are critical to life-sustaining (and life-challenging) events and processes occurring across broad ranges of temporal and spatial scales. Understanding issues of ocean resource consumption (e.g., fisheries, coastal pollution, etc.) lead to foci on ocean ecosystem dynamics and their coupling to physical processes (e.g., mixing, transports, upwelling, ice dynamics, etc.). Understanding the ocean role in climate dynamics (e.g., sequestration of atmospheric CO_2, ab-

Christopher K. Wikle is Professor, Department of Statistics, University of Missouri,146 Middlebush Hall, Columbia, Missouri 65203, USA (e-mail: wiklec@missouri.edu). Ralph F. Milliff is Senior Research Associate, CIRES, University of Colorado, Boulder, Colorado, USA (e-mail: Ralph.Milliff@colorado.edu). Radu Herbei is Assistant Professor, Department of Statistics, The Ohio State University, Columbus, Ohio, USA (e-mail: herbei@stat.osu.edu). William B. Leeds is Postdoctoral Researcher, Department of Statistics and Department of Geophysical Sciences, University of Chicago, Chicago, Illinois, USA (e-mail: leedsw@uchicago.edu).

sorption of atmospheric heat, impacts on the hydrologic cycle, teleconnections, etc.) lead to foci on massive and complex simulations and forecasts based on equations of geophysical fluid dynamics. In all instances, the broad range of scales, the energetic exchanges across them, and the associated uncertainties drive innovations that involve methods of modern statistical modeling.

Oceanography has historically been a "data poor" science and the need to use advanced statistical methodology to perform inference and prediction has been paramount throughout its development. Although it is the case that there are too few in situ observations of the ocean to characterize its evolution and its interaction with marine ecosystems, in an ironic twist, the discipline also suffers from having an abundance of particular data types when one factors in the satellite observations that have become available in the last couple of decades. The need for statistical collaboration in Oceanography results from both the situation of not having enough observations in some parts of the system and having huge amounts of data in other parts of the system.

1.1 The Physical Ocean

The physical ocean is governed by basic laws of physics (see, e.g., Vallis, 2006). The primitive equations consist of the following: three equations corresponding to the conservation of momentum (for the two horizontal and one vertical components of velocity), a continuity equation representing the conservation of mass, an equation of state (relating density, pressure, temperature and salinity), and equations corresponding to the conservation of temperature and salinity. There are seven state variables (three velocity components, density, pressure, temperature and salinity). This system of equations is nonlinear and exhibits a huge range of spatial and temporal scales of variability. Given that many of these scales of variability are not resolved in data or in deterministic or "forward" ocean models, the equations are typically simplified by scale analysis arguments and the small scale (turbulent) structures are parameterized. These parameterizations involve relationships between the mean of the state variables and their gradients. In this way the eddy viscosity and diffusivity terms serve as so-called "sub-grid scale" parameterizations, representing the unresolved processes that are sinks of momentum and heat at the grid scale of a given model, for example, $O(10)$ km in global ocean models and $O(1)$ km in regional ocean models.

The ocean system is nonlinearly coupled to the atmosphere across a broad range of scales. At the largest scales, air–sea fluxes of heat and fresh water drive mostly vertical or "thermohaline" circulations while the surface shear stress and wind stress curl drive mostly horizontal or wind-driven gyre circulations (e.g., Pedlosky, 1998). At smaller scales, the vertical-horizontal separation breaks down and the ocean response to external forcing and internal instabilities results in a broadband spectrum of vigorous eddy circulations (e.g., McWilliams, 2006). On all scales, the circulation provides the context for ocean processes affecting other components of the ocean system such as those related to ocean biology and chemistry. There are significant nonlinear interactions between the ocean chemistry, biology and its physical state.

1.2 Ocean Biogeochemistry

Ocean biogeochemistry is concerned with the interaction of the biology, chemistry and geology of the ocean (e.g., Miller, 2004). This is a very complex system that contains many interactions across a variety of scales. The system can be simply illustrated by thinking about the interactions of broad classes of its components. For example, the presence of nutrients near the ocean surface, where there is light, allows for the growth of phytoplankton, which deplete the nutrients as their population expands. The increased abundance of phytoplankton then provides a food source for zooplankton, which leads to growth in the zooplankton population. The consumption of phytoplankton leads to waste products from the zooplankton that settles as detritus to the ocean floor. As the zooplankton deplete the phytoplankton, the zooplankton population decreases due to the lack of a sufficient food source. Eventually, the detritus at depth is transferred to the surface through upwelling and mixing, providing the nutrients that lead to another bloom in phytoplankton, etc.

This simple four component system is a vast oversimplification, as there are many different species interacting at any one time. More critically, this lower trophic ecosystem is also coupled to higher levels of the food web, for example, with foraging fish predating the zooplankton, which are in turn predated by higher trophic fish, marine mammals, commercial fishing, etc. The chemical component of the cycle is critical in several respects. It provides a way for carbon to be removed from the atmosphere, as the phytoplankton remove carbon from the ocean water and are consumed by the zooplankton. Some of that carbon is contained in the detritus that sinks to the ocean floor, becoming buried in the sediment and leading to a (temporary) carbon sink in the global carbon cycle, which is very important in the context of sequestration of carbon relative to potential climate change from greenhouse gases. In addition, the biological cycle in the ocean is closely tied to the distribution of dissolved oxygen in the water and also influences the distribution of other chemicals, such as silicon, nitrates and phosphates, for example, in the shells of diatoms.

1.3 Uncertainty

Given the complexities in the ocean system, it is not surprising that there are numerous sources of uncertainty. First, although the large-scale equations of motion are in some sense deterministic, the scale issues that lead to eddy viscosity/diffusivity parameterizations are inherently uncertain. Furthermore, the forms of the linkages between system components (e.g., wind stress, heat and moisture fluxes between atmosphere and ocean) are not known with certainty. The components of traditional biogeochemical models are even more uncertain, both in terms of the functional forms and parameters.

The process and parameter uncertainty is compounded by the inherent data issues in the ocean system. In situ observations of the ocean are quite limited in terms of spatial and temporal resolution, and in terms of the variables measured. For example, it is a painstaking process to measure zooplankton abundance, often requiring a scientist or technician to literally count critters through a microscope. Fortunately, many surface variables can be observed remotely, particularly through satellite proxies. In some cases, for example, near surface winds from scatterometers, sea surface height from altimeters and sea surface temperature (SST) from radiometers, the satellite observations are typically quite precise, albeit with gaps corresponding to orbital geometries, swath widths and fields of view. In other cases, for example, ocean color as a proxy for phytoplankton, the observational representation of the process is more uncertain, at least on fairly short time scales. Thus, a key issue in state prediction, parameter estimation and inference is to deal with incomplete observations that vary in precision and in spatial and temporal support. This is particularly important when one considers ocean "data assimilation," that is, the blending of prior information (e.g., a numerical solution of the deterministic representation of the ocean state) with observations.

Another traditionally important component of uncertainty in ocean process modeling corresponds to the selection of reduced-dimensional representations of the process. Given the assumption that much of the larger scale processes in the ocean can be represented in a lower dimensional manifold, with smaller scales corresponding to turbulent scales (that certainly may interact with the larger scale modes, or at least suggest the form of parameterizations or stochastic noise terms), there has been considerable attention given to different approaches to obtain the reduced-dimension basis functions. The choices vary depending on the part of the system being considered as well as whether one is looking at the system diagnostically or predictively.

In the context of statistical models used to describe or predict portions of the ocean system, the nature of the error structures is important. Given the nonlinearity that is inherent in the system, many process distributions are not well represented by Gaussian errors. In addition, in some cases (e.g., biological abundance variables) the distributions can only have positive support.

1.4 Statistical Methods

The ocean and atmospheric sciences have benefitted from a strong tradition in applying fairly complex statistical methods to deal with many of the uncertainty issues described above. In particular, general monographs such as Emery and Thomson (2001), Von Storch and Zwiers (2002), and Wilks (2011) provide comprehensive descriptions of traditional methods used to analyze such data. In addition to overviews of basic statistical concepts, these books describe multivariate methods (e.g., principal components—empirical orthogonal functions, canonical correlation analysis, discriminant analysis, etc.), spectral methods (e.g., cross-spectral analysis) and dynamically-based reduction methods (e.g., principal oscillation patterns) to facilitate analysis of high-dimensional data that has inherent dependence in time and space. This is in addition to more focused monographs such as Preisendorfer and Mobley (1988), which gives a comprehensive overview of eigen-decomposition methods, and several books and review papers devoted to various aspects of data assimilation (see Section 3.1). Statistical presentations of many of these methods can be found in Jolliffe (2002) and Cressie and Wikle (2011).

Recognizing the challenges related to uncertainty in the ocean system and the need to foster more collaborative research between oceanographers and statisticians, the U.S. National Research Council (NRC) commissioned a panel to write a report on "Statistics and

Physical Oceanography" (NRC, 1994); see also the accompanying article by Chelton (1994) and published comments. This report contains a very nice review of physical oceanography for nonoceanographers and outlined the need for research in several key areas, including the change of support problem and the indirect nature of satellite observations, non-Gaussian random fields, the incorporation of Lagrangian and Eulerian data, data assimilation, inverse modeling, model/data comparison and feature identification, to name some of the most prominent. The report focused on the physical component of the ocean and did not address biogeochemistry nor many issues of current interest, such as climate change and reduced-dimensional representations.

1.5 Paper Outline

Our goal with this review is to provide an overview of some of the advancements that have occurred at the interface of Statistics and Oceanography since the NRC (1994) report. In particular, we believe strongly that the hierarchical statistical perspective has played a significant role in this development and will focus our review from that perspective. Section 2 presents a brief discussion of hierarchical modeling, both empirical and Bayesian, with some discussion of the need for computational tools. We note that although this paper is in a Statistics journal, we hope that it will generate interest from both statisticians and oceanographers. For the same reason that we gave a brief and general overview of Oceanography above, we will also give a brief and general overview of hierarchical modeling for those readers with little exposure to these ideas. In Section 3 we focus in more depth on examples related to data assimilation and inverse modeling, long lead forecasting and uncertainty quantification in biogeochemical models. We will follow this review with a brief discussion of current and future challenges in Section 4.

2. THE HIERARCHICAL MODELING PARADIGM

The idea of hierarchical modeling of scientific processes arose largely out of Berliner (1996), when Mark Berliner was the director of the Geophysical Statistics Project at the National Center for Atmospheric Research. The idea, although fundamentally quite simple, was revolutionary in that it provided a probabilistically consistent way to partition uncertainty in systems with complicated data, process and parameter relationships, and coincided with the development and popularization of Markov Chain Monte Carlo (MCMC) methods in Bayesian statistics. As described below, the key idea is to consider the joint model of data, process and parameters as three general linked model components, that is, the data conditioned on the process and parameters, the process conditioned on parameters, and the parameters. These ideas quickly spread into Statistics and subject matter journals in Climatology, Meteorology, Oceanography and Ecology, to name a few. This particular perspective on hierarchical modeling is summarized in several books, including Clark (2007), Royle and Dorazio (2008) and Cressie and Wikle (2011). More traditional Bayesian presentations of hierarchical models can be found in many books on Bayesian statistics (e.g., Gelman et al., 2004; Banerjee, Carlin and Gelfand, 2003).

Statistical modeling and analysis are about the synthesis of information. This information may come from expert opinion, physical laws, previous empirical results or various observations—both direct and indirect. Consider the case where we have a scientific process of interest, denoted by Y. As an example, say that this process corresponds to the near-surface north/south and east/west wind components over a portion of the ocean (i.e., a multivariate spatio-temporal process). We also have observed data associated with this process, say Z, which might come from a satellite-based scatterometer (i.e., wind component observations derived from speed and direction that are incomplete in space and time). We assume that we have parameters associated with the measurement process, say θ_Z, that might represent differences in support and representativeness between the satellite observations and the true wind process at the resolution of interest. In addition, we assume that there are some parameters, say θ_Y, that describe the underlying wind process dynamics (i.e., the evolution operator and innovation covariances that propagate the joint spatial fields of the wind components through time). Thus, using the total law of probability, it is natural to write the decomposition of the joint distribution of the data and process conditioned on the parameters as

$$(1) \qquad [Z, Y|\theta_Z, \theta_Y] = [Z|Y, \theta_Z][Y|\theta_Y],$$

where $[Z|Y, \theta_Z]$ is the "data distribution" (or "data model") and $[Y|\theta_Y]$ is the "process distribution" (or "process model"), and we have assumed conditional independence of the parameters on the distributions of the right-hand side (RHS) of (1). Note, we are using brackets "[ ]" to refer to a distribution and the vertical bar "|" to denote "conditioned upon." Clearly, we could also consider an alternative decomposition in which

Y is conditioned on Z followed by the marginal distribution of Z. However, such a decomposition is less scientific as described below.

In traditional statistics, one might think about the data Z given some specified distributional form and some associated parameters, θ (e.g., corresponding to a spatio-temporal mean and associated variances and covariances, or their parameterization). In the context of (1), such a distribution arises from integrating out the random Y process, yielding the distribution $[Z|\theta \equiv \{\theta_Z, \theta_Y\}]$. We are then typically interested in estimating these parameters given the data. Such estimation (e.g., maximum likelihood estimation) does not include an explicit representation of a model for the underlying dynamical wind process, Y, but rather includes it implicitly through the first and second moments (as a consequence of the integration). In addition, this distribution accounts for the uncertainty that is due to sampling and measurement.

The question is then why might we be interested in Y? First, in many such applications, one is actually interested in predicting the true, but unobserved process, Y, rather than just accounting for its (co)variability. Second, given the complexity of most ocean and atmospheric processes, the multivariate spatiotemporal dependence structures associated with Y can be very complicated (e.g., nonlinear in time, nonstationary in space and/or time) and potentially very high-dimensional. This seriously complicates the likelihood-based inference, as it puts much of the modeling burden on the realistic specification of complicated dependence structures. Rather, by focusing attention on the process Y directly, one can incorporate scientific insight (e.g., Markovian approximations to mathematical representations of the process, spatially or time-varying parameters, etc.) and, critically, disentangle the measurement uncertainty (which can also be quite complicated) and the process (co)variability and uncertainty. That is, marginal means and covariances contain potentially complicated functions of θ_Z and θ_Y, which can be difficult to specify without explicitly doing the integration of the random process. Thus, one is effectively trading the complexity of specifying very complicated marginal dependence structures with a more scientific specification of the conditional mean as a random process at the next level of the hierarchy. This also allows one to focus effort on modeling the conditional error structure in the data stage, without having to try to disentangle the measurement uncertainty with the process (co)variability. Statisticians will recognize this as just a manifestation of the issue that one faces in traditional mixed-model analysis where one has a choice of considering a marginal model, whereby the random effects are integrated out, or a conditional model, whereby the random effects are predicted and the conditional covariance structure in the data model is simpler (e.g., Verbeke and Molenberghs, 2009).

Applying Bayes' rule to the hierarchical decomposition in (1) gives

$$[Y|Z, \theta_Z, \theta_Y] \propto [Z|Y, \theta_Z][Y|\theta_Y], \tag{2}$$

with the normalizing constant obtained by the integral of (1) with respect to the process Y. Note that this implies that we are updating our knowledge of the process of interest given the data we have observed. This is the goal of prediction and confirms that the hierarchical decomposition on the RHS of (2) is the more plausible scientific decomposition of the joint distribution of data and process described above in (1). Clearly, (2) assumes that the parameters are known without uncertainty. This is seldom the case in reality, although one may have estimates of these parameters from some source and be comfortable with substituting them into (2), leading to what Cressie and Wikle (2011) refer to as an empirical hierarchical model (EHM). Alternatively, and more realistically for most ocean processes, at least some of the parameters are typically not known, and we are interested in learning more about them or, at least, would like to account for their uncertainty. In this case, we consider the fully-hierarchical or Bayesian hierarchical model (BHM):

$$[Y, \theta_Y, \theta_Z|Z] \propto [Z|Y, \theta_Z][Y|\theta_Y][\theta_Z, \theta_Y], \tag{3}$$

where one must specify a prior distribution for the parameters $[\theta_Z, \theta_Y]$ and we note that the normalizing constant integrates over the parameters in addition to the process. It is often the case that we have information to inform these parameter distributions. For example, in the case of the wind process described above, we have knowledge about the quality of scatterometer observations of wind components, and recognize that certain turbulent scaling laws must be followed, which can be incorporated through restrictions and informative prior distributions (e.g., Wikle et al., 2001).

Schematically, it is helpful to think about the RHS of (3) by using the following representation of Berliner (1996):

$$\begin{aligned} &[\text{data, process, parameters}] \\ &\quad = [\text{data}|\text{process, parameters}] \\ &\qquad \times [\text{process}|\text{parameters}] \times [\text{parameters}]. \end{aligned} \tag{4}$$

The strength of this hierarchical representation is that it provides a probabilistically consistent way to think about modeling complex systems while quantifying uncertainty. Critically, each of the stages on the RHS of (3) or (4) can be split into sub-models. For example, multiple data sets with differing supports can be accommodated with a model such as

$$\begin{aligned}(5)\qquad & [Z^{(1)}, Z^{(2)} | Y, \theta_{Z^{(1)}}, \theta_{Z^{(2)}}] \\ & \quad = [Z^{(1)} | Y, \theta_{Z^{(1)}}][Z^{(2)} | Y, \theta_{Z^{(2)}}],\end{aligned}$$

where $Z^{(1)}$ and $Z^{(2)}$ correspond to data sets (1) and (2), respectively, with associated parameters $\theta_{Z^{(1)}}$ and $\theta_{Z^{(2)}}$. We note that (5) makes the assumption that, conditioned on the true process, both data sets are independent. This is often a reasonable simplifying assumption and greatly facilitates the combination of differing data sets (although this assumption should be verified in specific applications). The parameters and distributional form for the two distributions on the RHS of (5) can be quite different, perhaps accommodating differing types of spatial or temporal support, and/or measurement and sampling errors. Examples of this in the context of the wind example discussed previously are data from satellite scatterometers as well as data from ocean buoys or even weather center analysis winds (e.g., Wikle et al., 2001).

It is also important to note that the process model component on the RHS of (3) or (4) can also be decomposed into subcomponents. This could correspond to a hierarchical Markov decomposition in time, for example, $[Y] = \prod_{t=1}^{T} [Y_t | Y_{t-1}][Y_0]$, where $Y = \{Y_0, Y_1, \dots, Y_T\}$. Or, it could correspond to a multivariate decomposition, $[Y] = [Y^{(2)} | Y^{(1)}][Y^{(1)}]$, where $Y = \{Y^{(1)}, Y^{(2)}\}$. In both cases, there would also be process model parameters. Clearly, combinations of these types of distributions and other types of dependencies (e.g., spatial, spatiotemporal, etc.) could be considered. In the context of the wind example, it would be natural for the wind components to be represented by $Y^{(2)}$ and these could be conditioned on near surface atmospheric pressure fields, say $Y^{(1)}$, both of which would be spatiotemporal processes with further conditioning possible.

Last, we recognize that a huge advantage of hierarchical models in complicated systems is that the parameters are themselves endowed with potentially quite complex distributions. That is, they exhibit multivariate dependence between parameters or in terms of space and time, or may themselves be dependent on exogenous information. Such complex dependencies in parameters are very difficult to accommodate in the classical paradigm. As discussed above, in the wind example, it is critical to specify parameter distributions that adhere to known empirical and theoretical turbulent scaling properties (e.g., Wikle et al., 2001).

There is no free lunch! Although the hierarchical modeling paradigm is extremely powerful, it often comes at a substantial computational cost. In particular, the normalizing constant in (3) involves the integration over all random quantities in the model, which can correspond to a very high-dimensional integration in many applications. Given that analytical solutions to these integrals are almost never available in complex models, one has to resort to numerical methods. In the fully Bayesian context this is typically some type of Markov Chain Monte Carlo algorithm (e.g., Robert and Casella, 2004). In the EHM case represented schematically in (2), one may be able to work out alternative computational solutions [e.g., expectation-maximization (E–M) or numerical maximum likelihood or method-of-moments estimation; e.g., see Chapter 7 of Cressie and Wikle (2011)]. It is critical to note that the complexity of these calculations often leads to modifications in model structure to facilitate practical computation.

3. HIERARCHICAL MODELING AND OCEANOGRAPHY

There have been many examples of hierarchical modeling in the ocean sciences since the late 1990s. In this section we briefly review some of that work at the interface of Statistics and Oceanography. We focus most attention on two data assimilation examples (ocean vector winds and ocean tracer state estimation), long lead forecasting of SST, and uncertainty quantification and assimilation of biogeochemical models. For these topics, we present vignettes to illustrate the power of hierarchical modeling from problems we have worked on. We also briefly describe some of the work at the interface related to other important oceanographic problems.

3.1 Data Assimilation and Inverse Modeling

Wikle and Berliner (2007) summarize data assimilation (DA) as an approach for optimally blending observations with prior information concerning the system (i.e., mathematical representations of mechanistic relationships, model output, etc.) to obtain a distributional characterization of the true state of the system. Relevant to this overview paper, these concepts originated in the atmospheric and ocean sciences and there is an extensive literature describing various methodologies (Daley, 1991; Ghil and Malanotte-Rizzoli, 1991; Bennett, 2002; Kalnay, 2003; Lorenc, 1986; Tarantola, 1987; Wikle and Berliner, 2007). In essence, DA can be thought of as an inverse problem, and one can derive algorithms from numerous perspectives, including optimal estimation theory, variational analysis and Bayesian statistics. The Bayesian approach to DA (e.g., Lorenc, 1986; Tarantola, 1987) is helpful because the problem is inherently hierarchical and, thus, it provides a coherent probabilistic approach that can be used to describe the various approaches. In addition, if one is willing to consider the BHM perspective, one can often obtain a more realistic characterization of uncertainty in various components of the data and mechanistic models (e.g., Wikle and Berliner, 2007).

Most oceanographic processes of interest in the DA context concern spatial fields that vary with time. A general dynamical spatiotemporal model (DSTM) can be written in the BHM paradigm (e.g., Cressie and Wikle, 2011). The data model is given by

$$\mathbf{Z}_t(\cdot) = \mathcal{H}\big(\mathbf{Y}_t(\cdot), \boldsymbol{\theta}_d(t), \boldsymbol{\varepsilon}_t\big), \quad t = 1, \ldots, T,$$

where $\mathbf{Z}_t(\cdot)$ represents the data at time t, $\mathbf{Y}_t(\cdot)$ the associated process, where the mapping function, $\mathcal{H}$, may be linear or nonlinear, the error process, $\boldsymbol{\varepsilon}_t$, may be additive or multiplicative, and the model depends on parameters given by $\boldsymbol{\theta}_d(t)$ that may be spatial or time-varying. The process model is given by

$$\mathbf{Y}_t(\cdot) = \mathcal{M}\big(\mathbf{Y}_{t-1}(\cdot), \boldsymbol{\theta}_p(t), \boldsymbol{\eta}_t\big), \quad t = 1, 2, \ldots,$$

where the evolution operator $\mathcal{M}$ may be linear or nonlinear for the oceanographic process of interest, the error process, $\boldsymbol{\eta}_t$, may be additive or multiplicative, and the parameters, $\boldsymbol{\theta}_p(t)$, may be spatial or time-varying. Note that this model is assumed to be valid beyond the maximum data time-period (T), so that forecasting is appropriate. Finally, the model is completed by the specification of distributions for the parameters in the previous stages, $[\boldsymbol{\theta}_d(t)][\boldsymbol{\theta}_p(t)]$, where we have assumed that the parameters from the two stages are independent for convenience. Note, this hierarchical framework would also include a specification of the initial process distribution, $[\mathbf{Y}_0|\boldsymbol{\theta}_0]$. The sequential modeling perspective is a natural framework in which to consider the DA problem, as one can interpret the prior distribution from above as a forecast distribution, which gets updated given new observations (e.g., see Cressie and Wikle, 2011).

3.1.1 *Ocean surface vector winds.* Surface winds directly transfer momentum to the ocean and surface wind speed modulates the exchanges of heat and fresh water to and from the upper ocean as modeled by bulk transfer formulae (e.g., Large, 2006). The advent of space-borne scatterometer instruments in the 1990s provided the first global wind fields, on daily timescales, from observations. Prior to the scatterometer era, ocean winds were inferred from global weather forecast models and reanalyses (e.g., Hellerman and Rosenstein, 1983) that depended upon a very sparse global network of in situ wind observations from buoys and ships of opportunity. Reliable resolutions in the pre-scatterometer wind fields were limited to ocean basin scale features (e.g., associated with large-scale high and low pressure systems) and monthly averages.

The wind fields retrieved from scatterometer observations are not direct measures of the wind, but rather the observations are of the roughness imparted on the ocean by surface capillary waves in response to the shear stress vector at the air–sea interface. The amplitudes and orientations of surface capillary waves are in equilibrium with the surface shear stress, and these amplitudes and orientations are retrievable from measures of how the capillary waves scatter impinging radar pulses of known frequency and polarization, that is, so-called microwave backscatter cross-sections; for details, see Freilich (1996). Radar backscatter cross-sections are spatially averaged over wind-vector cells and related to a surface vector wind (SVW) via a geophysical model function. The SVW retrievals from scatterometers are accurate to within at least 2 ms^{-1} in speed and 30° in direction. Depending on the sensor system and agency providing the data, 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-orbitting satellite ground track. Swath widths vary by system from 600–1800 km, that is, between about 20 and 100 wind-vector cells across a given swath depending on resolution. Because of the polar orbit (about 14 polar orbits per day) the swaths overlap at high latitudes and are separated by gaps in coverage at low latitudes, with the largest swath gaps occurring at the equator. Again, depending on the system, the swath gaps at the equator are on the order of 10 degrees longitude and can take up to three days to fill. The SVW from scatterometers resolve features of the synoptic storms (e.g., fronts, convergences in rain bands, closed circulations, etc.) that form, propagate and dissipate every day, over the world ocean.

To fill the gaps in the scatterometer winds, one could make use of the complete (yet lower resolution) wind fields from the operational meteorological centers. However, these wind fields have different properties than the scatterometer observations. Differences in the true resolutions of weather-center analyses and scatterometer winds are efficiently described in terms of surface wind kinetic energy spectra, that is, kinetic energy as a function of spatial wavenumber. Wikle, Milliff and Large (1999) showed a power-law relation for surface winds from multiple data sources in the tropical Pacific and the power law behavior in surface winds from scatterometry has been documented for the globe (Milliff et al., 1999). Milliff et al. (2011) compare kinetic energy between weather center analyses and SVW retrievals and find that the kinetic energy drops off unrealistically in weather center analyses at smaller scales. For example, at spatial scales on the order of 10^2 km, the weather center kinetic energy content is more than an order of magnitude weaker than the SVW observations. 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.

Wikle et al. (2001) implemented a spatiotemporal BHM for tropical winds in a region of the equatorial western Pacific. As noted above, the inter-swath gaps in scatterometer coverage on daily timescales are largest in the tropics. A BHM formulation to blend weather-center analyses and scatterometer winds requires space–time properties that account for the greater intermittency in the scatterometer data. This is achieved in the process model stage in Wikle et al. (2001), which is posed in terms of three scientifically-motivated components as

$$\mathbf{Y}_t = \boldsymbol{\mu}^u + \boldsymbol{\Phi}^{(1)}\boldsymbol{\alpha}_t^u + \boldsymbol{\Phi}^{(2)}\boldsymbol{\beta}_t^u, \tag{6}$$

where $\mathbf{Y}_t$ is a vector of spatially-indexed zonal (east–west) wind velocity components (typically denoted by u). Analogous terms apply for the meridional velocity (v). Here, $\boldsymbol{\mu}^u$ is the mean zonal wind process that accounts for the prevailing wind and its variance in the tropical domain, $\boldsymbol{\Phi}^{(1)}$ is a basis function matrix corresponding to the large-scale modes of the equatorial beta-plane approximation (i.e., a linear, thin fluid approximation) to the momentum equations, and $\boldsymbol{\Phi}^{(2)}$ are nested wavelet basis functions used to model fine-scale winds according to observed kinetic energy-wavenumber properties for the region. The use of scientifically-motivated basis functions is crucial here, as we have information from the thin-fluid analytical approximation to the wind field that is appropriate over the tropics. In particular, the leading equatorial normal modes [$\boldsymbol{\Phi}^{(1)}$] are large-scale wave signals in the tropics, that is, westward propagating Rossby waves, eastward Kelvin waves and mixed Rossby–Gravity waves (e.g., Matsuno, 1966). Based on data analyses (e.g., Wheeler and Kiladis, 1999), the equatorial normal mode basis functions were limited to the leading modes defined by two zonal wavenumbers and four wavenumbers in the meridional direction (Wikle et al., 2001). These are sufficient to support the important basin-scale Kelvin, Rossby and mixed Rossby–Gravity waves that can be used to describe much of the tropical dynamics within the Pacific basin. The amplitude coefficients, $\boldsymbol{\alpha}_t^u$ and $\boldsymbol{\beta}_t^u$, are treated as random variables in the BHM, with first-order Markov models. The priors on the parameters associated with these models correspond to time series of the theoretical amplitudes for each mode for the $\boldsymbol{\alpha}_t^u$ coefficients and the slopes of the kinetic energy spectrum for the $\boldsymbol{\beta}_t^u$ coefficients as obtained from Wikle, Milliff and Large (1999). Data stage inputs were obtained from NSCAT SVW retrievals and from sea level pressure and surface winds taken from reanalyses of the National Centers for Environmental Protection (NCEP).

The posterior mean surface wind and surface convergence/divergence exhibited variability on scales not achievable in the NCEP reanalyses. This is particularly evident in a comparison of the NCEP reanalysis wind field with the posterior mean winds from the tropical wind BHM for the period when tropical cyclone Dale crossed the study area (Figure 1). The posterior mean winds are organized into strong convergence regions that spiral into the center of the tropical cyclone. These are consistent with rain band features of tropical cyclones and this was confirmed by comparing the posterior mean winds with cloud-top temperatures from an independent satellite observation nearly coincident with the snapshot from the posterior distribution of the BHM (see Wikle et al., 2001).

Uncertainty management properties of SVW BHM based on mechanisitic models (i.e., leading order terms and/or approximations of the primitive equations) are particularly relevant in ocean forecast settings. The Mediterranean Forecast System (MFS) is an operational system producing 10-day forecasts for upper ocean fields every day. The MFS ocean forecast models resolve synoptic variability in the upper ocean.

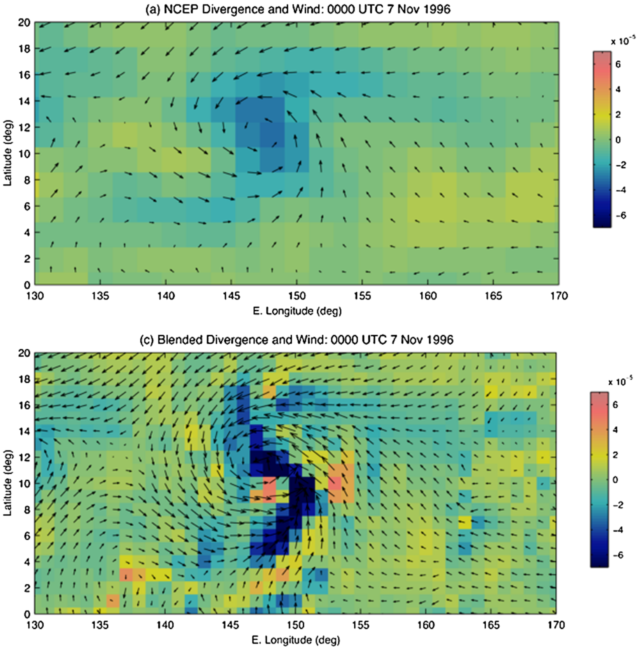

FIG. 1. *Tropical surface wind BHM after Wikle et al.* (2001). *Top panel is the NCEP reanalysis at the time tropical cyclone Dale occupied the BHM domain. Color contours denote convergence* (*blues*) *and divergence* (*reds*) *in the surface wind field. Bottom panel is the posterior mean wind field* (*vectors*) *and convergence/divergence* (*colors*) *map for tropical cyclone Dale from the tropical wind BHM. Strong convergences in rainband structures spiraling into the tropical cyclone center coincide with coldest cloud-top temperatures from independent satellite observations.*

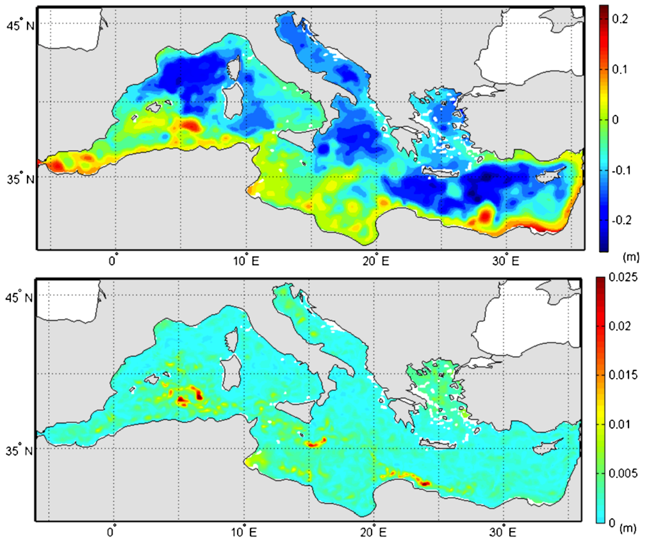

FIG. 2. *Forecast initial condition mean* (*top*) *and spread* (*bottom*) *in the sea-surface height field computed from* 10 *ensemble members, each driven by a realization from the posterior distribution for the SVW BHM* (*Milliff et al.*, 2011). *Ensemble spread is usefully localized in the most uncertain regions of the domain where wind forcing is driving hydrodynamic instabilities and a local pulse in the mesoscale eddy field* (*Pinardi et al.*, 2011).

The uncertain parts of the forecast fields are at ocean mesoscales, that is, hourly and 10–50 km scales. These are the scales of upper ocean hydrodynamic instabilities driven by the surface wind. So, modeling uncertainty in the surface wind field is a useful means of quantifying uncertainty in the MFS ocean forecasts on the scales that are most important to daily users.

Milliff et al. (2011) describe the details of the SVW BHM for MFS, and Pinardi et al. (2011) review the impacts of BHM SVW fields in an ensemble forecast methodology built around realizations from the posterior distribution for SVW from the BHM. The mechanistic process model in Milliff et al. (2011) involves the leading-order terms in a Rayleigh Friction equation approximation at synoptic scales and, again, a nested wavelet basis model to represent turbulent closure at the finest spatial scales. Thus, again it is critical to incorporate scientific information into the specification of the process rather than try to model such behavior through the marginal covariance structure. This then also allows specification of the measurement uncertainties associated with the data conditioned upon the true process. Specifically, data stage inputs are obtained from the QuikSCAT scatterometer SVW retrievals and from weather center sea level pressure and surface wind fields.

Ten realizations of the posterior distribution for SVW were used to drive ensemble data assimilation and ensemble forecasts in the MFS. Figure 2 depicts the data assimilation system response in sea-surface height (SSH) at the forecast initial condition time. The mean SSH initial condition (Figure 2, top) reflects the accurate spatial scales for MFS forecasts where synoptic variability overlies the general circulation patterns for the Mediterranean Sea. Sub-basin scale cyclonic gyres are represented by blue (depressed) SSH and anticyclones by reds. The gradients between blue and red signals are proportional to surface current speeds. The spread in the SSH initial condition (Figure 2, bottom) demonstrates the value of managing uncertainty in the SVW. The largest amplitude signals in the spread are localized in a few places only—where surface wind is driving hydrodynamic instabilities and pulsing the energy in ocean mesoscale response.

3.1.2 *Inversion of oceanographic tracer data*. Inferring the structure of the circulation in the world's oceans is a significant part of our quest for understanding the climate. Direct measurements of the circulation

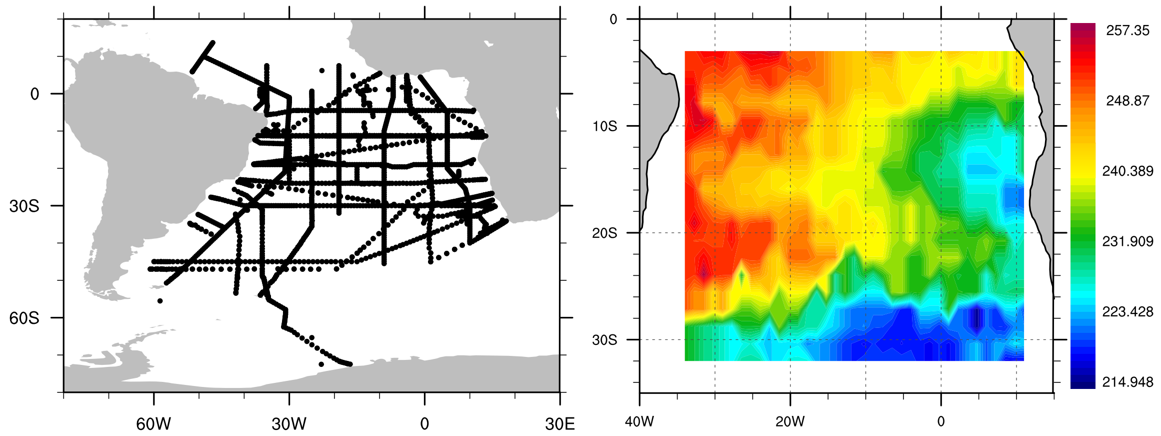

FIG. 3. *Left*: *World Ocean Circulation Experiment era cruise tracks in the South Atlantic Ocean. Right*: *Interpolated oxygen measurements in a deep neutral layer of the South Atlantic Ocean.*

are difficult to obtain; however, in the past decades, scientists have collected large amounts of hydrographic data (hydrostatic pressure, temperature, salinity, oxygen concentration, etc.). Such data have been used to produce climatological maps that exhibit the large scale structure (Lozier, Owens and Curry, 1995). Based on these maps and data, one can infer some constituents of the ocean circulation, such as boundary currents, wind-driven gyres and abyssal interior flow. Yet, hydrographic data are typically available at very sparse locations in space and time. For example, the World Ocean Circulation Experiment lasted from 1990 until 1998. In Figure 3 we illustrate these data that were available for the South Atlantic Ocean. Consequently, statistical smoothing methods are required to provide estimates at space–time locations of interest. To that end, one would typically combine data sets collected at different times. This seems reasonable only under the assumption that the distribution of dynamically-passive tracers is representative of a mean circulation and that such tracers are sufficiently stable on a fixed neutral density layer.

A classical approach to inverting hydrographic data into circulation structure is the *box inverse method* (Wunsch, 1996). Starting with the thermal wind equations, the problem reduces to estimation of a reference velocity field. This is achieved via a large system of linear equations expressing conservation of mass within a collection of connected boxes. Due to the size of the problem, a high-dimensional regression approach is subsequently employed. Traditionally, ridge regression estimators have proved useful, however, modern statistical/machine learning methods can provide more effective alternatives. Other approaches are based on more complex inverse modeling and nonlinear optimization (Paillet and Mercier, 1997; Zika, McDougall and Sloyan, 2010; Zhang and Hogg, 1992; Wunsch, 1994). The inverse problem described here is ill-posed and some type of regularization is required (Kirsch, 1996). We describe the Bayesian inversion approach of McKeague et al. (2005) (see also Herbei, McKeague and Speer, 2008), which connects the data (tracer concentrations) to parameters (ocean circulation) using a system of partial differential equations. In this case, regularization is imposed through a prior distribution over the parameters. The *solution* is the posterior distribution. It can be used to select "representative" values and the associated uncertainty for any aspect of the circulation.

Formally, let $C = C(x, y)$ denote the concentration of a tracer of interest at a location (x, y) in a rectangular domain $\Omega \subset \mathbb{R}^2$ representing the layer of the ocean being studied. The problem is to estimate the (horizontal) water velocities and diffusion coefficients based on noisy measurements of C at a sparse set of locations in Ω. In the right panel of Figure 3 we display an interpolated map of oxygen concentration for a 2000 m deep layer in the South Atlantic Ocean. The connection between the tracer concentration C and the velocities and diffusion coefficients is modeled by the steady-state advection-diffusion equation

$$(7)\qquad \mathbf{u} \cdot \nabla C = \nabla \cdot (K \nabla C) + Q_C, \quad (x, y) \in \Omega,$$

where $\mathbf{u} = (u, v)$ is the horizontal water velocity, the diagonal diffusivity matrix $K = \operatorname{diag}(\kappa^{(x)}, \kappa^{(y)})$ does not vary with location, and the sink term $Q_C = -\lambda C$ is present only when C represents oxygen concentration. Equation (7) is augmented with Dirichlet boundary conditions $C = C_{\partial\Omega}$ when $(x, y) \in \partial\Omega$, where $\partial\Omega$ denotes the boundary of Ω.

The statistical model assumes additive observational error

$$C_j^{\text{obs}}(x_i, y_i) = C_j(x_i, y_i) + \varepsilon_j(x_i, y_i).$$

Here, $j = 1, \dots, n_C$ indexes a particular tracer and $i = 1, \dots, N_j$ indexes a spatial location where data for tracer j are available. The underlying tracer concentration $C(x, y)$ is obtained as a solution of the advection diffusion equation (7). As this solution is not available in closed form, one uses a numerical (grid-based) approximation. We collect all quantities of interest (velocities, diffusion coefficients, boundary values) in a high-dimensional vector $\boldsymbol{\gamma}$. Under the assumption that the measurement errors $\varepsilon(\cdot)$ are unbiased Gaussian variables with constant (yet tracer-dependent) variance, the posterior distribution is written as

$$\begin{aligned} &\pi(\boldsymbol{\gamma}|C^{\text{obs}}) \\ (8)\quad &\propto \prod_{j=1}^{n_C}\prod_{i=1}^{N_j} \exp\left\{-\frac{(C(\boldsymbol{\gamma}; x_i, y_i) - C^{\text{obs}}(x_i, y_i))^2}{2\sigma_j^2}\right\} \\ &\qquad \cdot \pi(\boldsymbol{\gamma}), \end{aligned}$$

where $\pi(\boldsymbol{\gamma})$ represents the selected prior distribution. The posterior probability model (8) is explored via Monte Carlo methods (e.g., Robert and Casella, 2004).

The ill-posedness mentioned above is resolved by specifying a proper prior distribution $\pi(\boldsymbol{\gamma})$. This implies that the posterior distribution is proper. However, an efficient MCMC approach is still required. In addition, one can design specific sampling strategies as described in McKeague et al. (2005) and Herbei, McKeague and Speer (2008). It is important to understand that each component described above (physical model, prior distribution, likelihood function, data) plays a crucial role in determining the solution. Under the Bayesian approach, a sensitivity analysis, although possible, is hampered by the immense computational cost associated with the MCMC sampler. In addition, the under-determination problem is present in this case. There are (roughly) 300 data points, while there are thousands of parameters to estimate (velocities, diffusions, boundary values). In the results given in McKeague et al. (2005) and Herbei, McKeague and Speer (2008), not all estimated velocities are significantly different from zero, which is the prior mean. However, the data are informative about the large-scale features (zonal jets, gyres). The posterior mean velocities are compared with velocities determined from float data (Hogg and Owens, 1999), showing a comforting consistency (McKeague et al., 2005 and Herbei, McKeague and Speer, 2008).

3.2 Long Lead Forecasting: SST

Tropical Pacific SST exhibits some of the most important variability on inter-annual time scales for the ocean/atmosphere system (e.g., see the overview in Philander, 1990). This variation arises from complicated interactions, across a large range of spatiotemporal scales, between the ocean and the atmosphere. The most prominent signal on these time scales is the El Niño-Southern Oscillation (ENSO) phenomenon. This is the well-known quasi-periodic (3–5 year period) warming (El Niño) and cooling (La Niña) in the tropical Pacific. These warming and cooling events lead to dramatic effects in weather across the globe due to teleconnections with the global atmospheric circulation. Because of these significant weather related impacts (e.g., droughts, floods, etc.), it is critical to be able to forecast several months in advance the possible development and transition of these events. Increasingly, such "long lead" forecasts have shown useful skill and correspond to one of the few situations in ocean science where a purely statistical forecast methodology is competitive with, and in many cases better than, equivalent deterministic model forecasts (Barnston, He and Glantz, 1999; van Oldenborgh et al., 2005).

At typical spatial resolutions, there can easily be several thousand gridded spatial locations corresponding to the tropical Pacific region of forecasting interest. Complicated spatiotemporal statistical models are difficult or impossible to implement at these dimensions. For this reason, statistically-based models for tropical SST have been "spectral" in the sense that they are based on coefficients associated with a projection of the SST fields on spatial basis functions. The associated projection coefficients that are evolved are typically a reduced set, usually corresponding to larger modes. That is, let $\mathbf{Y}_t$ correspond to an n-dimensional vector of the true SST process at n spatial locations and time t and consider the decomposition of this process vector

$$(9)\qquad \mathbf{Y}_t = \boldsymbol{\Phi}^{(1)}\boldsymbol{\alpha}_t + \boldsymbol{\Phi}^{(2)}\boldsymbol{\beta}_t,$$

where $\boldsymbol{\Phi}^{(i)}$, $i = 1, 2$, is an $n \times p_i$ matrix of spatial basis functions, and $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are the associated expansion coefficients. A first-order linear Markov assumption on the evolution of the coefficients $\boldsymbol{\alpha}_t$, for example, $\boldsymbol{\alpha}_{t+\tau} = \mathbf{M}\boldsymbol{\alpha}_t + \boldsymbol{\eta}_{t+\tau}$, for appropriate time increment, τ, and with Gaussian errors, $\boldsymbol{\eta}_t \sim \text{Gau}(\mathbf{0}, \mathbf{Q})$, was shown in the early 1990s to be a model with reasonable skill at long-lead forecasting (e.g., Penland and Magorian, 1993). As with the tropical wind case presented previously, the specification of the process in terms of two basis sets is critical from a scientific perspective. In particular, it is thought that the active dynamical process is driven by a lower dimensional manifold (corresponding to the first set of basis functions) and, yet, the residual spatially-dependent structures captured by the second set of basis functions remain an important source of variability.

Although dimension-reduced linear Markov models showed reasonable skill, the ENSO phenomenon is better characterized as a nonlinear process (e.g., Hoerling, Kumar and Zhong, 1997; Burgers and Stephenson, 1999). Nonlinear statistical models typically are better at capturing the magnitude of the predicted El Niño and La Niña events (Tangang et al., 1998; Berliner, Wikle and Cressie, 2000; Tang et al., 2000; Timmermann, Voss and Pasmanter, 2001; Kondrashov et al., 2005). Most nonlinear stochastic methods have not successfully characterized the uncertainty in the forecasts. An exception to this is the model of Berliner, Wikle and Cressie (2000), who use a dimension-reduced threshold Markov model with certain components governed by the onset of intra-seasonal oscillations (so-called "westerly wind bursts").

There are several key components of the model in Berliner, Wikle and Cressie (2000) that illustrate the power of hierarchical modeling. First, the data model is constructed

$$\text{(10)} \qquad \mathbf{Z}_t = \boldsymbol{\Phi}^{(1)}\boldsymbol{\alpha}_t + \boldsymbol{\varepsilon}_t, \quad \boldsymbol{\varepsilon}_t \sim \text{Gau}(\mathbf{0}, \mathbf{R}),$$

where the data vector, $\mathbf{Z}_t$, is very high-dimensional (say, $n \times 1$), and $\boldsymbol{\Phi}^{(1)}$ is a reduced-dimension set of spatial basis functions as described above (an $n \times p_1$ matrix). It is important to note that the remaining basis functions (e.g., $\boldsymbol{\Phi}^{(2)}$ from above) associated with the basis expansion are used to parameterize the observational spatial covariance matrix, $\mathbf{R}$, which now accounts for both observation error and the truncation. The assumption is that the active dynamics exist on the lower dimensional manifold represented by $\boldsymbol{\Phi}^{(1)}$ and, like in typical turbulence parameterizations, the small scale structures accounted for by the $\boldsymbol{\Phi}^{(2)}$ portion of the expansion are associated with nonpredictive covariability.

Critically, the evolution of the dynamical process is nonlinear, but conditionally linear. Specifically,

$$\text{(11)} \quad \boldsymbol{\alpha}_{t+\tau} = \boldsymbol{\mu}_t + \mathbf{M}_t\boldsymbol{\alpha}_t + \boldsymbol{\eta}_{t+\tau}, \quad \boldsymbol{\eta}_t \sim \text{Gau}(\mathbf{0}, \mathbf{Q}),$$

where prior distributions are also specified for $\{\boldsymbol{\alpha}_t : t = 1, \ldots, \tau\}$. The important modeling assumption is that $\mathbf{M}_t$ and $\boldsymbol{\mu}_t$ depend on both the current (time t) and future (time $t+\tau$) climate "regimes," that is, $\mathbf{M}_t = \mathbf{M}(I_t, J_t)$ and $\boldsymbol{\mu}_t = \boldsymbol{\mu}(I_t, J_t)$, where I_t classifies the current SST regime as "cool," "normal" or "warm," and J_t anticipates one of these three regimes at time $t+\tau$. Specifically, I_t is based on the Southern Oscillation Index and J_t is based on a latent variable that is modeled, hierarchically, in terms of the east-west component of the near surface wind in the Western Pacific ocean (for details, see Berliner, Wikle and Cressie, 2000). The point is that this leads to nine potential mean states and nine potential dynamical operators, accommodating the nonlinear transition between ENSO phases.

Because this model was developed in a Bayesian hierarchical framework, it was able to capture the uncertainty characterized by the data, process and parameters. However, despite the fact that this model used physical notions (i.e., westerly wind bursts) in the lower stages of the model hierarchy, it was unable to directly account for quadratic interactions across spatial scales. As mentioned previously, nonlinearity is important in most atmosphere and ocean processes. 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 a quadratic nonlinear model in the context of a Bayesian hierarchical framework in which the random nonlinear process is "hidden" and parameters are random as well. They illustrate that simple arguments from turbulence scaling support the form of this model and further suggest a dimension reduction strategy.

As defined in Wikle and Hooten (2010), general quadratic nonlinearity (GQN) with respect to the $\boldsymbol{\alpha}_t$ process can be written

$$\begin{aligned} &\alpha_{t+\tau}(i) \\ \text{(12)} \quad &= \sum_{j=1}^{p} m_{ij}^{L}\alpha_t(j) \\ &\quad + \sum_{k=1}^{p}\sum_{l=1}^{p} m_{i,kl}^{Q}\alpha_t(k)g(\alpha_t(l); \boldsymbol{\theta}_g) + \eta_{t+\tau}(i) \end{aligned}$$

for $i = 1, \ldots, p$, where $g(\cdot)$ is some transformation of $\boldsymbol{\alpha}_t$ that depends on parameters $\boldsymbol{\theta}_g$ and gives the process more generality than the simple dyadic interactions alone. This framework is exceptionally flexible in accommodating many real-world mechanistic processes, but it comes at the cost of a curse of dimensionality in the parameter space; that is, there are $O(p^3)$ parameters corresponding to the nonlinear coefficients ($m^Q_{i,kl}$), in addition to the linear coefficients (m^L_{ij}) and $\boldsymbol{\theta}_g$.

One can use scale analysis to help with the dimensionality concerns, and also to motivate components of the hierarchical model. Specifically, as before, assume we can decompose the spectral coefficients into large-scale components ($\boldsymbol{\alpha}_t$) and small-scale coefficients ($\boldsymbol{\beta}_t$) corresponding to the spatial basis functions $\boldsymbol{\Phi}^{(1)}$ and $\boldsymbol{\Phi}^{(2)}$ discussed above. Consider all of the possible dyadic interactions of the elements of this vector—that is, there are small-scale/small-scale, small-scale/large-scale and large-scale/large-scale interactions. Loosely motivated by "Reynolds averaging" in turbulence theory (e.g., Holton, 2004), Wikle and Hooten (2010) suggest considering the large-scale/large-scale pairwise interactions explicitly, with the small-scale/small-scale interactions contributing to a correlated dependence structure in the additive error, and the small-scale/large-scale interactions corresponding to the linear term in the large-scale coefficients with random parameters (i.e., the small-scale coefficients play the role of random coefficients in the interaction). Such an argument leads to a quadratic nonlinear model on the large-scale coefficients, which can be written in matrix form as

$$\begin{aligned}\boldsymbol{\alpha}_{t+\tau} &= \mathbf{M}_L\boldsymbol{\alpha}_t + (\mathbf{I}_{p_1} \otimes \boldsymbol{\alpha}'_t)\mathbf{M}_Q\boldsymbol{\alpha}_t + \boldsymbol{\eta}_{t+\tau},\\ &\qquad \boldsymbol{\eta}_t \sim \mathrm{Gau}(\mathbf{0}, \mathbf{Q}).\end{aligned} \tag{13}$$

Prior distributions are then given to the parameters in $\mathbf{M}_L$ and $\mathbf{M}_Q$ as well as the covariance matrix $\mathbf{Q}$ (e.g., Wikle and Hooten, 2010). This approach was shown to quite reasonably account for the uncertainties so that the prediction error bounds covered the extreme ENSO events, even when the forecasts did not adequately capture the true magnitude.

Critically, these quadratic nonlinear implementations still suffer from a fairly high curse of dimensionality in the parameters. In this sense, various model reduction approaches have to be implemented that necessarily fail to account for as much model uncertainty as is likely present. 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, Wikle and Holan (2011) employed a stochastic search variable selection methodology (e.g., George and McCulloch, 1993, 1997).

In general, although the GQN statistical models are very flexible in representing oceanographic processes, these models can easily experience finite-time blow up (i.e., explosive growth) when fit to data (Majda and Yuan, 2012). This is seldom an issue when one is using these models for data assimilation, given the presence of observations to act as a control (e.g., Leeds et al., 2013), nor is it typically a problem when one is only forecasting out one time step (e.g., as in the SST example). It can be a problem when multiple time steps are forecast that require some form of constraint, either statistical or physical (Majda and Harlim, 2013).

3.3 Biogeochemical Models

Analysis of marine ecosystem dynamics involves various sources of uncertainty in the observations, the underlying scientific process and the parameters that describe the process dynamics. Critically, it is often the case that the observation errors are non-Gaussian and that the process being modeled is nonlinear, that is, there is an explicit system of coupled nonlinear differential equations that describe the complex ecosystem dynamics. As a result of these uncertainties and complexities, the BHM framework is natural, but can be difficult to implement.

Soon after the introduction of MCMC methods into Bayesian computation, the BHM approach was used in data assimilation for marine ecosystem models. Anticipating the forthcoming increase in remotely-sensed ocean color observations, Harmon and Challenor (1997) implemented an MCMC-based sampling protocol that explored the ability to recover various model parameters in a seven-compartment marine ecosystem model with and without added model noise. They noted the computational difficulties required to adequately account for correlation in these parameters. While identifying correlation in only ten parameters may no longer be a computational issue, there are still issues related to the sampling methodology that adequately generates reasonable block proposals for the entire parameter vector. Adaptive Metropolis–Hastings algorithms, which update the covariance of the proposal distribution, may be useful when trying to generate adequate proposals for a nonlinear marine ecosystem model (see, e.g., Haario, Saksman and Tamminen, 2001). Current research has taken into account advances in sampling methodology, including the following: sequential importance resampling, particle filters, ensemble Kalman filters and state-augmentation approaches (Evensen, 1994, 2009; Dowd, 2006, 2007, 2011; Parslow et al., 2013; Stroud et al., 2010).

An innovative practice in statistical modeling of ocean ecosystems is the inclusion of information from deterministic, mechanistic models into the statistical framework, typically in the process stage of a BHM. However, the necessary estimation procedures often require iteratively running the mechanistic model, which poses a problem when the model is computationally expensive and can only be run a very limited number of times. In certain situations where the computer model is too computationally expensive to run a sufficient number of times for the desired analysis, statistical surrogates (i.e., emulators) are used. In its simplest form, an emulator is simply the resulting estimated statistical model when computer model output is used as data. Then, this model is used to predict the output of the computer model under untried input settings (e.g., initial conditions, parameter values, forcings).

Traditionally, so-called "computer model" emulation has been done using Gaussian Process (GP) models (e.g., Sacks et al., 1989; Currin et al., 1991). In our case, these "computer models" are deterministic ocean forward models. GP emulators are related to spatial GPs, which use a correlation function such that the model output is more highly correlated for inputs that are "nearer" to one another in a given sense than those that are "farther apart." However, because ocean ecosystem models are nonlinear, a GP emulator of the joint distribution of the output may be inappropriate (as a nonlinear process cannot be specified by only two moments). In this case, one could consider the use of a dynamic GP emulator, which considers the value of the process at the current time step (i.e., the initial conditions) as an input to the ocean forward model (and GP emulator). This offers several benefits over the traditional GP emulation approach (e.g., Conti et al., 2009; Liu and West, 2009).

Rather than modeling the dynamics through a covariance function in a GP, it may be more appropriate to model the output using first-order characteristics (as the computer models themselves are written). The use of so-called "first-order emulators" has appeared in several applications, for example, van der Merwe et al. (2007) and Frolov et al. (2009) use neural networks and Hooten et al. (2011) use random forests, a machine learning algorithm. There is also the potential to use other nonlinear statistical models, polynomial chaos expansions (Mattern, Fennel and Dowd, 2012) or GP models in a dynamical context (Margvelashvili and Campbell, 2012). These approaches use dimension-reduction techniques to overcome the curse of dimensionality, with the emulator describing the dynamics of the reduced-dimensional process. Most of these implementations use nonparametric approaches. Alternatively, Leeds, Wikle and Fiechter (2013) use an emulator based on a reduced-rank multivariate quadratic nonlinear statistical model as described above.

Related to the use of emulators for dynamic models, there is also a growing body of work considering the use of emulators for spatial or spatiotemporal forward model output. Marine ecosystem models can vary from nonspatial, 0-D models, to models that have 3-D spatial structure. Leeds et al. (2013) used a 1-D (in the vertical) four-component model that included nutrients, zooplankton, phytoplankton and detritus (e.g., a "1-D NZPD" model) with iron limitation (i.e., "1-D NPZDFe") to model a 3-D process by creating a forest of 1-D models (more specifically, a forest of emulators of the 1-D model). These 1-D models resolve vertical processes, but do not account for horizontal diffusion and advection. Variability resulting from horizontal advection and diffusion is accommodated by putting a spatial GP model on the parameters (inputs) to the 1-D NPZDFe model.

In certain circumstances, the forward model emulator is developed simply to learn about the behavior of the forward model itself. However, in other situations, the forward model emulator may be useful in a data assimilation context. In those circumstances, the emulator may be used in a two-stage approach, where the emulator is fit "off-line," and then is used inside a Bayesian hierarchical model in the place of the forward model itself. Calder et al. (2011) and Leeds, Wikle and Fiechter (2013) developed BHMs using forward model output as data.

Leeds, Wikle and Fiechter (2013) performed data assimilation using SeaWiFS ocean color observations, as well as sea surface height and SST output from the Regional Ocean Model System (ROMS) forward model coupled with the NPZDFe model (Fiechter and Moore, 2012). They consider a reduced-dimension quadratic nonlinear process model similar to (9) and (13), but where $\mathbf{Y}_t = (\mathbf{Y}'_{1,t}, \mathbf{Y}'_{2,t}, \mathbf{Y}'_{3,t})'$, representing the three state variables of interest. Basis functions $\mathbf{\Phi}^{(1)}$ and $\mathbf{\Phi}^{(2)}$ were based on output from the ROMS-NPZDFe model for a different time period using a singular value decomposition and the priors for $\mathbf{M}_L$ and $\mathbf{M}_Q$ were developed using the remaining right singular vectors. Leeds, Wikle and Fiechter (2013) show that this approach was sufficient to accommodate nonlinear dynamics even in the absence of observations over a substantial portion of the domain as shown in Figure 4.

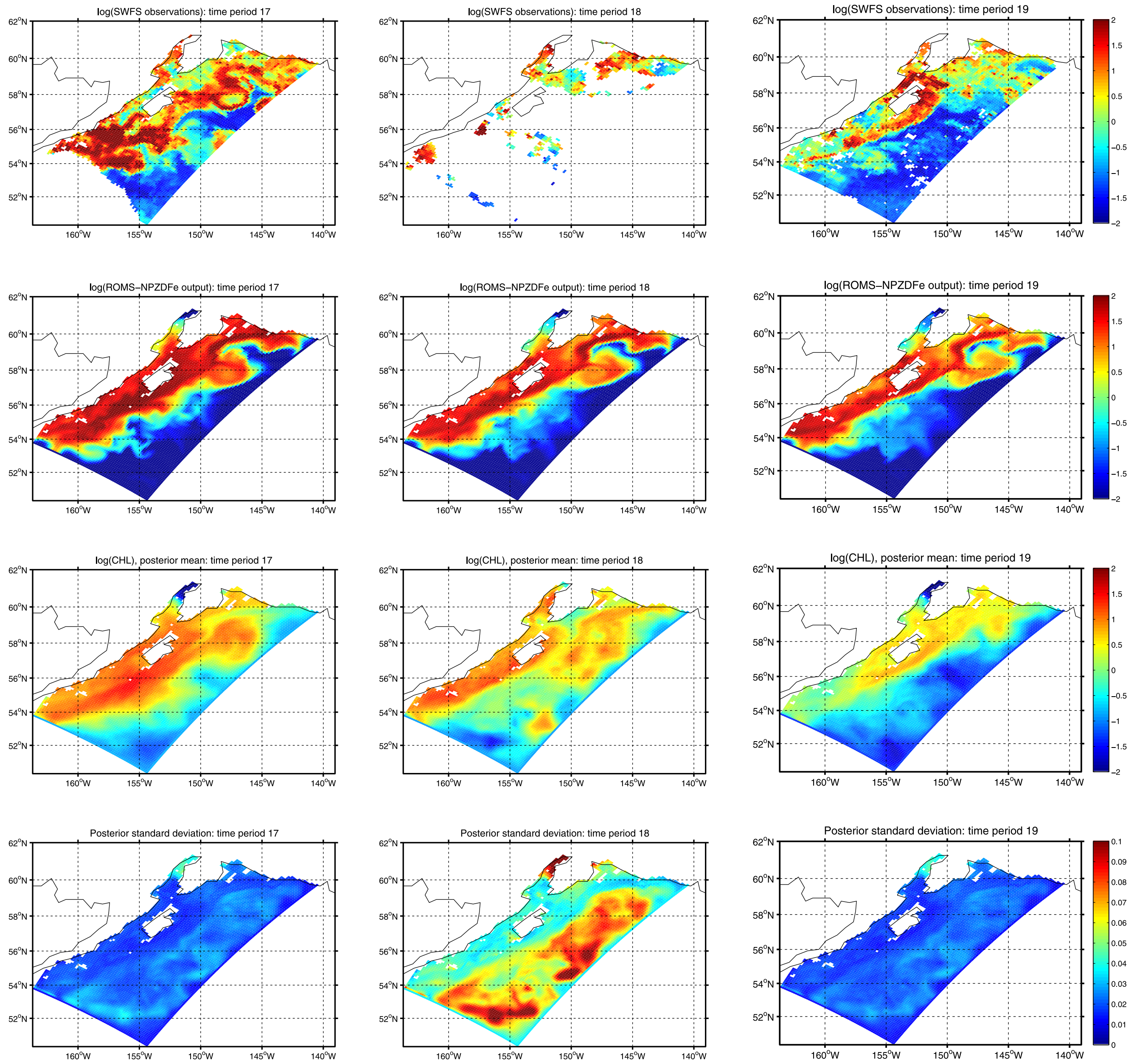
log(SWFS observations): time period 17
log(SWFS observations): time period 18
log(SWFS observations): time period 19
log(ROMS–NPZDFe output): time period 17
log(ROMS–NPZDFe output): time period 18
log(ROMS–NPZDFe output): time period 19
log(CHL), posterior mean: time period 17
log(CHL), posterior mean: time period 18
log(CHL), posterior mean: time period 19
Posterior standard deviation: time period 17
Posterior standard deviation: time period 18
Posterior standard deviation: time period 19
62°N
60°N
58°N
56°N
54°N
52°N
160°W
155°W
150°W
145°W
140°W
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0

FIG. 4. *A plot of SeaWiFS ocean color observations* (*top row*), *ROMS-NPZDFe output for phytoplankton* (*second row*), *the posterior mean for phytoplankton* (*third row*) *and the posterior standard deviation* (*bottom row*), *for three consecutive eight-day time periods, corresponding to May* 16, *May* 24 *and June* 1, 2002, *respectively. Note that observations were log-transformed and that the MCMC was run using these log-transformed observations.*

3.4 Other Important Problems

Space limitations prevent us from giving complete overviews of all hierarchical statistical models in oceanography. However, we mention here some notable and important papers that have been published in recent years in areas outside of those mentioned above. By necessity, this list is not exhaustive, neither in terms of topics nor citations within topics, but it does give some sense of the breadth and depth of work being done in the area.

As mentioned above, the amount of in situ data for ocean state variables is fairly limited and there is a need to construct complete spatial fields representing climatological features of the ocean state, as well as the uncertainty in those states using spatial and spatiotemporal models. There are many notable examples from a BHM perspective (e.g., Higdon, 1998; Lavine and

Lozier, 1999; Lemos and Sansó, 2009, 2012; Lemos, Sansó and Santos, 2009).

It has long been known that there are strong teleconnections between the ocean and the atmosphere and this has led to useful statistical models for prediction (e.g., Davis, 1976; Barnett, 1981; Barnett and Preisendorfer, 1987). There is an increasingly large number of papers that have used the BHM perspective to model teleconnections as well as their impacts. For example, Wikle and Anderson (2003) used such a model when evaluating the impacts of SST on tornado report counts in the eastern two-thirds of the US; Elsner, Murnane and Jagger (2006) (and references therein) have shown great success in modeling hurricane activity and forecasts (see also Flay and Nott, 2007); Lima and Lall (2009) use a BHM to model precipitation considering ocean conditions. In addition, there are recent studies showing linkages between the physical ocean and ecological impacts (e.g., Cloern et al., 2010; Ruiz et al., 2009).

Bayesian hierarchical models have long been used to model the higher trophic levels of the marine ecosystem. For example, stock-recruitment models have been an important tool in marine fisheries management (e.g., Thompson, 1992; Hilborn, Pikitch and McAllister, 1994; McAllister and Kirkwood, 1998; Dorn, 2002; Michielsens and McAllister, 2004; Hirst et al., 2005). In addition, BHMs have been used extensively in recent years to model distributions and movement of marine mammals (e.g., Jonsen, Myers and James, 2007; Ver Hoef and Jansen, 2007; Johnson et al., 2008; Cressie et al., 2009; Hanks et al., 2011; Moore and Barlow, 2011; Conn et al., 2012; Hiruki-Raring et al., 2012; McClintock et al., 2012).

Clearly, a critical problem of vast societal interest concerns climate change. Given the role of the ocean in the climate system and its direct connections to weather events and ecological impacts, oceanic climate change and its uncertainty characterization are extremely important. This is a vast research area with many papers that have taken a hierarchical Bayesian approach. This topic is beyond the scope of this review, but a few notable examples include Tebaldi et al. (2005), Furrer et al. (2007), Tebaldi and Sansó (2009), Aldrin et al. (2012) and Satterthwaite et al. (2012).

4. CURRENT AND FUTURE CHALLENGES

As demonstrated throughout this review, the ocean system is quite complex with numerous interacting subcomponents and external processes, and with substantial uncertainty in terms of knowledge and data. Statistical methods have been critically important to improve our understanding of this system and to characterize uncertainty. In recent years, the hierarchical statistical modeling paradigm has proven to be an exceptionally useful tool to manage the various sources of uncertainty. We have only just scratched the surface in terms of our presentation of the work that has been done in this area, but it is clear that the use of hierarchical modeling in oceanography has blossomed. That being said, in addition to continued development in the areas mentioned above, there are still many important problems at the interface of Statistics and Oceanography to be considered and methodologies to be explored.

One of the biggest challenges for statisticians working in Oceanography is to develop statistical models that can effectively parameterize the complex nonlinear interactions that are associated with the ocean system. In particular, models must be developed to account for potentially high-dimensional state-processes, and yet effectively manage uncertainty with varying data quality and non-Gaussian errors. In general, to be effective in this context, dimension reduction should not be independent of the physical and biological environment of the problem under consideration.

One component of nonlinearity that is important concerns the coupling of subsystems, whether that be the atmosphere/ocean interface, the ocean/ice interface, the physical/biological interface or interactions between trophic levels in the ocean. Many of these couplings require parameterizations for fluxes across boundaries, and there is often only limited observational data available to help inform the process. This presents an extreme challenge and one in which we must use the large amounts of remotely sensed observations along with the sparse in situ data to build these nonlinear relationships.

Not unrelated is the notion of accounting for and describing model error. The precise characterization of uncertainty that is fundamental in BHM can be used to help identify and characterize model error in deterministic models of the ocean system, for example, ocean forecast models. As with all model abstractions, deterministic approaches accept trade-offs in resolution and approximations of the ocean variability to gain affordability in simulations and forecasts. Because the ocean system and the models are inherently nonlinear, the errors introduced by acceptable approximations can grow and lead to model error properties that are difficult to diagnose. Parameters of posterior distributions from independent BHM analyses for ocean model response or control variables can be used as a standard against which deterministic model error can be quantified. For example, if the incremental adjustment of the surface momentum flux control vector in a variational data assimilation procedure pushes the momentum flux at a point outside the reasonable bounds of a posterior distribution for momentum flux from a BHM, the variational procedure is probably compensating for forecast model error as opposed to uncertainty in the control vector.

To gain understanding of the impact of physical processes on biota as well as the interaction of biological organisms, there is increasing reliance on individual (or agent-based) models. These models have become more prevalent in the ecological realm and are starting to be considered from the BHM perspective (e.g., Hooten and Wikle, 2010). The use of these models in a statistical context across the spectrum of scales and processes involved in oceanography is a growing area of interest. For example, Megrey et al. (2007) link physical ocean models to individual-based bioenergetic models for fish.

The issues described above will almost certainly require new computational strategies, particularly in the Bayesian paradigm. As the models get more and more complex and larger data sets become available, standard MCMC methods begin to fail to provide meaningful results. Among the many challenges associated with Bayesian computing, two stand out: (1) there are very high-dimensional distributions to be explored, and (2) the models are complex, which leads to inexact and sometimes impossible likelihood evaluations. Novel MCMC methodology will address these issues, while maintaining feasibility. For example, particle MCMC (PMCMC) methods are designed to address the first issue. For high-dimensional posterior distributions (thousands of state variables and parameters), it is nearly impossible to design good Metropolis–Hastings proposal distributions and in this case, even adaptive MCMC may fail. Andrieu, Doucet and Holenstein (2010) propose to use sequential Monte Carlo (SMC) methods combined with importance sampling to design near optimal proposal distributions. The resulting algorithm, which has similar features to a particle filter, will update the entire collection of state variables at once and can be extremely useful for space–time models. Parslow et al. (2013) use PMCMC for state and parameter estimation as well as state forecasting in the context of a marine biogeochemical model. In addition, Hamiltonian MCMC methodologies may provide an attractive alternative for situations for which samples from complicated high-dimensional processes and parameter distributions are required (e.g., Beskos, Kalogeropoulos and Pazos, 2013).

Approximate Bayes Computing (ABC) methods are a new and emerging class of likelihood-free MCMC algorithms. They address the second issue above—that is, cases when evaluation of the likelihood function involves integrals over very large spaces that are impossible to calculate (e.g., Beaumont, Zhang and Balding, 2002). ABC relies on the ability to simulate from the selected model without much computing effort. Consequently, the user is forced to select a relevant (multivariate) statistic and "posterior samples" are defined as parameter values that result in a statistic similar to the one observed. This algorithm, in fact, explores the distribution of the variables of interest *conditional on the selected statistic*, not the data. Although this may raise some concern regarding the interpretation of the results, Fearnhead and Prangle (2012) describe a semi-automatic way of selecting good summary statistics in a general setting.

In general, it seems likely that for inference and prediction of complicated oceanographic processes in the foreseeable future, one will have to continue to make compromises between model complexity and computational feasibility. Indeed, one must be willing to accept some reasonable lack of "optimality" in the solution in order to make headway on many of these problems. This is true regardless of whether one takes a Bayesian or frequentist perspective. The hierarchical paradigm helps to some extent as it allows one to consider trade-offs between approximate computational strategies, incorporation of scientific information and model specification for the different model components (data, process and parameters) separately.

In conclusion, there is a long history of activity at the interface of Statistics and Oceanography. In recent years, the hierarchical statistical paradigm has proven to be very helpful for managing the uncertainty associated with data, process and parameters in modeling the ocean and its related systems. There are increasingly more oceanographers with advanced statistical training and more statisticians with oceanographic backgrounds and this is sure to lead to even more innovative and exciting methodological developments in years to come.

ACKNOWLEDGMENTS

The authors wish to thank the Editor, AE and reviewer who provided substantial suggestions to improve the manuscript. In addition, Wikle acknowledges the support of National Science Foundation (NSF) Grants DMS-10-49093, OCE-0814934 and Office of Naval Research (ONR) Grant ONR-N00014-10-0518. Herbei acknowledges support from NSF Grant DMS-12-09142 and ONR Grant N00014-10-1-0488. Milliff acknowledges early support for BHM research from the Physical Oceanography Program, ONR; from the NASA, International Ocean Vector Winds Science Team, and recent support from US Globec under Grant number NSF OCE-0815030. It is a pleasure to acknowledge our BHM mentor L. Mark Berliner and the inspirations he provides dating back to his days as Director, Geophysical Statistics Project at NCAR through the present. We are also pleased to acknowledge our colleagues and collaborators who participate in vigorous cross-disciplinary discussions at summer Confabs in Boulder, CO. In recent years, these include, Professors Berliner, Mevin B. Hooten, Andrew M. Moore, Nadia Pinardi and Thomas M. Powell, and Doctors Jeremiah Brown, Manny Fiadeiro and Jerome Fiechter.

REFERENCES

ALDRIN, M., HOLDEN, M., GUTTORP, P., SKEIE, R. B., MYHRE, G. and BERNTSEN, T. K. (2012). Bayesian estimation of climate sensitivity based on a simple climate model fitted to observations of hemispheric temperatures and global ocean heat content. *Environmetrics* **23** 253–271. MR2914207

ANDRIEU, C., DOUCET, A. and HOLENSTEIN, R. (2010). Particle Markov chain Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **72** 269–342. MR2758115

BANERJEE, S., CARLIN, B. and GELFAND, A. (2003). *Hierarchical Modeling and Analysis for Spatial Data* **101**. Chapman & Hall/CRC, Boca Raton.

BARNETT, T. (1981). Statistical prediction of North American air temperatures from Pacific predictors. *Monthly Weather Review* **109** 1021–1041.

BARNETT, T. and PREISENDORFER, R. (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review* **115** 1825–1850.

BARNSTON, A., HE, Y. and GLANTZ, M. (1999). Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997–98 El Niño episode and the 1998 La Niña onset. *Bulletin of the American Meteorological Society* **80** 217–243.

BEAUMONT, M. A., ZHANG, W. and BALDING, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics* **162** 2025–2035.

BENNETT, A. F. (2002). *Inverse Modeling of the Ocean and Atmosphere*. Cambridge Univ. Press, Cambridge. MR1920432

BERLINER, L. M. (1996). Hierarchical Bayesian time series models. In *Maximum Entropy and Bayesian Methods* (*Santa Fe, NM*, 1995). *Fund. Theories Phys.* **79** 15–22. Kluwer Academic, Dordrecht. MR1446713

BERLINER, L., WIKLE, C. and CRESSIE, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *Journal of Climate* **13** 3953–3968.

BESKOS, A., KALOGEROPOULOS, K. and PAZOS, E. (2013). Advanced MCMC methods for sampling on diffusion pathspace. *Stochastic Process. Appl.* **123** 1415–1453. MR3016228

BURGERS, G. and STEPHENSON, D. (1999). The "normality" of El Nino. *Geophys. Res. Lett.* **26** 1027–1030.

CALDER, C., BERRETT, C., SHI, T., XIAO, N. and MUNROE, D. (2011). Modeling space–time dynamics of aerosols using satellite data and atmospheric transport model output. *J. Agric. Biol. Environ. Stat.* **16** 495–512.

CHELTON, D. (1994). Physical oceanography: A brief overview for statisticians. *Statist. Sci.* **9** 150–166.

CLARK, J. S. (2007). *Models for Ecological Data*: *An Introduction*. Princeton Univ. Press, Princeton, NJ. MR2292764

CLOERN, J., HIEB, K., JACOBSON, T., SANSÓ, B., DI LORENZO, E., STACEY, M., LARGIER, J., MEIRING, W., PETERSON, W. and POWELL, T. et al. (2010). Biological communities in San Francisco Bay track large-scale climate forcing over the North Pacific. *Geophys. Res. Lett.* **37** L21602.

CONN, P., JOHNSON, D., LONDON, J. and BOVENG, P. (2012). Accounting for missing data when assessing availability in animal population surveys: An application to ice-associated seals in the Bering Sea. *Methods in Ecology and Evolution* **3** 1039–1046.

CONTI, S., GOSLING, J. P., OAKLEY, J. E. and O'HAGAN, A. (2009). Gaussian process emulation of dynamic computer codes. *Biometrika* **96** 663–676. MR2538764

CRESSIE, N. and WIKLE, C. K. (2011). *Statistics for Spatio-temporal Data*. Wiley, Hoboken, NJ. MR2848400

CRESSIE, N., CALDER, C., CLARK, J., HOEF, J. and WIKLE, C. (2009). Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. *Ecological Applications* **19** 553–570.

CURRIN, C., MITCHELL, T., MORRIS, M. and YLVISAKER, D. (1991). Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. *J. Amer. Statist. Assoc.* **86** 953–963. MR1146343

DALEY, R. (1991). *Atmospheric Data Analysis*. *Cambridge Atmospheric and Space Science Series* **4** 57. Cambridge Univ. Press, Cambridge.

DAVIS, R. (1976). Predictability of sea surface temperature and sea level pressure anomalies over the North Pacific Ocean. *Journal of Physical Oceanography* **6** 249–266.

DORN, M. (2002). Advice on West Coast rockfish harvest rates from Bayesian meta-analysis of stock-recruit relationships. *North American Journal of Fisheries Management* **22** 280–300.

DOWD, M. (2006). A sequential Monte Carlo approach for marine ecological prediction. *Environmetrics* **17** 435–455. MR2240936

DOWD, M. (2007). Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. *Journal of Marine Systems* **68** 439–456.

DOWD, M. (2011). Estimating parameters for a stochastic dynamic marine ecological system. *Environmetrics* **22** 501–515. MR2843404

ELSNER, J., MURNANE, R. and JAGGER, T. (2006). Forecasting US hurricanes 6 months in advance. *Geophys. Res. Lett.* **33** L10704.

EMERY, W. and THOMSON, R. (2001). *Data Analysis Methods in Physical Oceanography*. Elsevier, Amsterdam.

EVENSEN, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research* **99** 10.

EVENSEN, G. (2009). *Data Assimilation*: *The Ensemble Kalman Filter*, 2nd ed. Springer, Berlin. MR2555209

FEARNHEAD, P. and PRANGLE, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **74** 419–474. MR2925370

FIECHTER, J. and MOORE, A. M. (2012). Iron limitation impact on eddy–induced ecosystem variability in the coastal Gulf of Alaska. *Journal of Marine Systems* **92** 1–15.

FLAY, S. and NOTT, J. (2007). Effect of ENSO on Queensland seasonal landfalling tropical cyclone activity. *International Journal of Climatology* **27** 1327–1334.

FREILICH, M. (1996). Sea winds algorithm theoretical basis document. Jet Propulsion Laboratory, Pasadena, CA.

FROLOV, S., BAPTISTA, A., LEEN, T., LU, Z. and VAN DER MERWE, R. (2009). Fast data assimilation using a nonlinear Kalman filter and a model surrogate: An application to the Columbia River estuary. *Dynamics of Atmospheres and Oceans* **48** 16–45.

FURRER, R., SAIN, S. R., NYCHKA, D. and MEEHL, G. A. (2007). Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Environ. Ecol. Stat.* **14** 249–266. MR2405329

GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2004). *Bayesian Data Analysis*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR1385925

GEORGE, E. and MCCULLOCH, R. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.* **88** 881–889.

GEORGE, E. and MCCULLOCH, R. (1997). Approaches for Bayesian variable selection. *Statist. Sinica* **7** 339–374.

GHIL, M. and MALANOTTE-RIZZOLI, P. (1991). Data assimilation in meteorology and oceanography. *Adv. Geophys* **33** 141–266.

HAARIO, H., SAKSMAN, E. and TAMMINEN, J. (2001). An adaptive Metropolis algorithm. *Bernoulli* **7** 223–242. MR1828504

HANKS, E., HOOTEN, M., JOHNSON, D. and STERLING, J. (2011). Velocity-based movement modeling for individual and population level inference. *PloS One* **6** e22795.

HARMON, R. and CHALLENOR, P. (1997). A Markov chain Monte Carlo method for estimation and assimilation into models. *Ecological Modelling* **101** 41–59.

HELLERMAN, S. and ROSENSTEIN, M. (1983). Normal monthly wind stress over the world ocean with error estimates. *Journal of Physical Oceanography* **13** 1093–1104.

HERBEI, R., MCKEAGUE, I. W. and SPEER, K. G (2008). Gyres and jets: Inversion of tracer data for ocean circulation structure. *J. Phys. Ocean.* **38** 1180–1202.

HIGDON, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.

HILBORN, R., PIKITCH, E. and MCALLISTER, M. (1994). A Bayesian estimation and decision analysis for an age-structured model using biomass survey data. *Fisheries Research* **19** 17–30.

HIRST, D., STORVIK, G., ALDRIN, M., AANES, S. and HUSEBY, R. (2005). Estimating catch-at-age by combining data from different sources. *Canadian Journal of Fisheries and Aquatic Sciences* **62** 1377–1385.

HIRUKI-RARING, L., HOEF, J., BOVENG, P. and BENGTSON, J. (2012). A Bayesian hierarchical model of Antarctic fur seal foraging and pup growth related to sea ice and prey abundance. *Ecological Applications* **22** 668–684.

HOERLING, M., KUMAR, A. and ZHONG, M. (1997). El Niño, La Niña, and the nonlinearity of their teleconnections. *Journal of Climate* **10** 1769–1786.

HOGG, N. G. and OWENS, W. B. (1999). Direct measurement of the deep circulation within the Brazil Basin. *Deep-Sea Res.* **46** 335–353.

HOLTON, J. (2004). *An Introduction to Dynamic Meteorology*, 4th ed. Elsevier Academic Press, Burlington, MA.

HOOTEN, M. B. and WIKLE, C. K. (2010). Statistical agent-based models for discrete spatio-temporal systems. *J. Amer. Statist. Assoc.* **105** 236–248. MR2757201

HOOTEN, M. B., LEEDS, W. B., FIECHTER, J. and WIKLE, C. K. (2011). Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models. *J. Agric. Biol. Environ. Stat.* **16** 475–494. MR2862294

JOHNSON, D. S., LONDON, J. M., LEA, M.-A. and DURBAN, J. W. (2008). Continuous-time correlated random walk model for animal telemetry data. *Ecology* **89** 1208–1215.

JOLLIFFE, I. (2002). *Principal Component Analysis*, 2nd ed. Springer, New York.

JONSEN, I., MYERS, R. and JAMES, M. (2007). Identifying leatherback turtle foraging behaviour from satellite telemetry using a switching state-space model. *Marine Ecology Progress Series* **337** 255–264.

KALNAY, E. (2003). *Atmospheric Modeling*, *Data Assimilation and Predictability*. Cambridge Univ. Press, Cambridge.

KIRSCH, A. (1996). *An Introduction to the Mathematical Theory of Inverse Problems*. *Applied Mathematical Sciences* **120**. Springer, New York. MR1479408

KONDRASHOV, D., KRAVTSOV, S., ROBERTSON, A. and GHIL, M. (2005). A hierarchy of data-based ENSO models. *Journal of Climate* **18** 4425–4444.

LARGE, W. (2006). Surface fluxes for practitioners of global data assimilation. In *Ocean Weather Forecasting* (E. Chassignet and J. Verron, eds.) 229–270. Springer, Dordrecht.

LAVINE, M. and LOZIER, S. (1999). A Markov random field spatio-temporal analysis of ocean temperature. *Environ. Ecol. Stat.* **6** 249–273.

LEEDS, W., WIKLE, C. and FIECHTER, J. (2013). Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. *Stat. Methodol.* To appear. DOI:10.1016/j.statmet.2012.11.004.

LEEDS, W., WIKLE, C., FIECHTER, J., BROWN, J. and MILLIFF, R. (2013). Modeling 3-D spatio-temporal biogeochemical processes with a forest of 1-D statistical emulators. *Environmetrics* **24** 1–12.

LEMOS, R. T. and SANSÓ, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of North Atlantic sea surface temperature. *J. Amer. Statist. Assoc.* **104** 5–18. MR2662306

LEMOS, R., SANSÓ, B. and SANTOS, F. (2009). Hierarchical Bayesian modelling of wind and sea surface temperature from the Portuguese coast. *International Journal of Climatology* **30** 1423–1430.

LEMOS, R. T. and SANSÓ, B. (2012). Conditionally linear models for non-homogeneous spatial random fields. *Stat. Methodol.* **9** 275–284. MR2863614

LIMA, C. and LALL, U. (2009). Hierarchical Bayesian modeling of multisite daily rainfall occurrence: Rainy season onset, peak, and end. *Water Resources Research* **45** W07422.

LIU, F. and WEST, M. (2009). A dynamic modelling strategy for Bayesian computer model emulation. *Bayesian Anal.* **4** 393–411. MR2507369

LORENC, A. (1986). Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society* **112** 1177–1194.

LOZIER, M. S., OWENS, W. and CURRY, R. (1995). The climatology of the North Atlantic. *Prog. Oceanogr.* **36** 1–44.

MAJDA, A. J. and HARLIM, J. (2013). Physics constrained nonlinear regression models for time series. *Nonlinearity* **26** 201–217. MR3001768

MAJDA, A. J. and YUAN, Y. (2012). Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems. *Discrete Contin. Dyn. Syst.* **4** 1333–1363.

MARGVELASHVILI, N. and CAMPBELL, E. (2012). Sequential data assimilation in fine-resolution models using error-subspace emulators: Theory and preliminary evaluation. *Journal of Marine Systems* **90** 13–22.

MATSUNO, T. (1966). Quasi-geostrophic motions in the equatorial area. *J. Meteor. Soc. Japan* **44** 25–43.

MATTERN, J., FENNEL, K. and DOWD, M. (2012). Estimating time-dependent parameters for a biological ocean model using an emulator approach. *Journal of Marine Systems* **96** 32–47.

MCALLISTER, M. and KIRKWOOD, G. (1998). Bayesian stock assessment: A review and example application using the logistic model. *ICES Journal of Marine Science*: *Journal du Conseil* **55** 1031–1060.

MCCLINTOCK, B., KING, R., THOMAS, L., MATTHIOPOULOS, J., MCCONNELL, B. and MORALES, J. (2012). A general discrete-time modeling framework for animal movement using multistate random walks. *Ecological Monographs* **82** 335–349.

MCKEAGUE, I. W., NICHOLLS, G. K., SPEER, K. G. and HERBEI, R. (2005). Statistical inversion of South Atlantic circulation in an abyssal neutral density layer. *J. Mar. Res.* **63** 683–704.

MCWILLIAMS, J. (2006). *Fundamentals of Geophysical Fluid Dynamics*. Cambridge Univ. Press, Cambridge.

MEGREY, B., ROSE, K., KLUMB, R., HAY, D., WERNER, F., ESLINGER, D. and SMITH, S. (2007). A bioenergetics-based population dynamics model of Pacific herring (*Clupea harengus pallasi*) coupled to a lower trophic level nutrient–phytoplankton–zooplankton model: Description, calibration, and sensitivity analysis. *Ecological Modelling* **202** 144–164.

MICHIELSENS, C. and MCALLISTER, M. (2004). A Bayesian hierarchical analysis of stock recruit data: Quantifying structural and parameter uncertainties. *Canadian Journal of Fisheries and Aquatic Sciences* **61** 1032–1047.

MILLER, C. (2004). *Biological Oceanography*. Blackwell, Oxford.

MILLIFF, R., LARGE, W., MORZEL, J., DANABASOGLU, G. and CHIN, T. (1999). Ocean general circulation model sensitivity to forcing from scatterometer winds. *Journal of Geophysical Research* **104** 11337–11411.

MILLIFF, R., BONAZZI, A., WIKLE, C., PINARDI, N. and BERLINER, L. (2011). Ocean ensemble forecasting. Part I: Ensemble Mediterranean winds from a Bayesian hierarchical model. *Quarterly Journal of the Royal Meteorological Society* **137** 858–878.

MOORE, J. and BARLOW, J. (2011). Bayesian state-space model of fin whale abundance trends from a 1991–2008 time series of line-transect surveys in the California Current. *Journal of Applied Ecology* **48** 1195–1205.

NRC (1994). Report on statistics and physical oceanography. *Statist. Sci.* **9** 167–201.

PAILLET, J. and MERCIER, H. (1997). An inverse model of the eastern North Atlantic general circulation and thermocline ventilation. *Deep-Sea Res.* **44** 1293–1328.

PARSLOW, J., CRESSIE, N., CAMPBELL, E., JONES, E. and MURRAY, L. (2013). Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications* **23** 679–698.

PEDLOSKY, J. (1998). *Ocean Circulation Theory*. Springer, Berlin.

PENLAND, C. and MAGORIAN, T. (1993). Prediction of Niño 3 sea surface temperatures using linear inverse modeling. *Journal of Climate* **6** 1067–1076.

PHILANDER, S. (1990). *El Niño*, *La Niña*, *and the Southern Oscillation* **46**. Academic Press, San Diego, CA.

PINARDI, N., BONAZZI, A., DOBRICIC, S., MILLIFF, R., WIKLE, C. and BERLINER, L. (2011). Ocean ensemble forecasting. Part II: Mediterranean Forecast System response. *Quarterly Journal of the Royal Meteorological Society* **137** 879–893.

PREISENDORFER, R. and MOBLEY, C. (1988). *Principal Component Analysis in Meteorology and Oceanography* **425**. Elsevier, New York.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York. MR2080278

ROYLE, J. and DORAZIO, R. (2008). *Hierarchical Modeling and Inference in Ecology*: *The Analysis of Data from Populations*, *Metapopulations and Communities*. Academic Press, San Diego, CA.

RUIZ, J., GONZÁLEZ-QUIRÓS, R., PRIETO, L. and NAVARRO, G. (2009). A Bayesian model for anchovy (Engraulis encrasicolus): The combined forcing of man and environment. *Fisheries Oceanography* **18** 62–76.

SACKS, J., WELCH, W. J., MITCHELL, T. J. and WYNN, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.* **4** 409–435. MR1041765

SATTERTHWAITE, W., MOHR, M., O'FARRELL, M., WELLS, B. and WALTERS, C. (2012). A Bayesian hierarchical model of size-at-age in ocean-harvested stocks—quantifying effects of climate and temporal variability. *Canadian Journal of Fisheries and Aquatic Sciences* **69** 942–954.

STROUD, J. R., STEIN, M. L., LESHT, B. M., SCHWAB, D. J. and BELETSKY, D. (2010). An ensemble Kalman filter and smoother for satellite data assimilation. *J. Amer. Statist. Assoc.* **105** 978–990. MR2752594

TANG, B., HSIEH, W., MONAHAN, A. and TANGANG, F. (2000). Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial Pacific sea surface temperatures. *Journal of Climate* **13** 287–293.

TANGANG, F., TANG, B., MONAHAN, A. and HSIEH, W. (1998). Forecasting ENSO events: A neural network-extended EOF approach. *Journal of Climate* **11** 29–41.

TARANTOLA, A. (1987). *Inverse Problem Theory*: *Methods for Data Fitting and Model Parameter Estimation*. Elsevier Science Publishers B.V., Amsterdam. MR0930881

TEBALDI, C. and SANSÓ, B. (2009). Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *J. Roy. Statist. Soc. Ser. A* **172** 83–106. MR2655606

TEBALDI, C., SMITH, R., NYCHKA, D. and MEARNS, L. (2005). Quantifying uncertainty in projections of regional climate change: A Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate* **18** 1524–1540.

THOMPSON, G. (1992). A Bayesian approach to management advice when stock-recruitment parameters are uncertain. *Fishery Bulletin* **90** 561–573.

TIMMERMANN, A., VOSS, H. and PASMANTER, R. (2001). Empirical dynamical system modeling of ENSO using nonlinear inverse techniques. *Journal of Physical Oceanography* **31** 1579–1598.

VALLIS, G. (2006). *Atmospheric and Oceanic Fluid Dynamics*: *Fundamentals and Large-Scale Circulation*. Cambridge Univ. Press, Cambridge.

VAN DER MERWE, R., LEEN, T. K., LU, Z., FROLOV, S. and BAPTISTA, A. M. (2007). Fast neural network surrogates for very high dimensional physics-based models in computational oceanography. *Neural Netw.* **20** 462–478.

VAN OLDENBORGH, G. J., BALMASEDA, M., FERRANTI, L., STOCKDALE, T. and ANDERSON, D. (2005). Did the ECMWF seasonal forecast model outperform statistical ENSO forecast models over the last 15 years? *Journal of Climate* **18** 3240–3249.

VER HOEF, J. M. and JANSEN, J. K. (2007). Space–time zero-inflated count models of harbor seals. *Environmetrics* **18** 697–712. MR2408939

VERBEKE, G. and MOLENBERGHS, G. (2009). *Linear Mixed Models for Longitudinal Data*. Springer, New York. MR2723365

VON STORCH, H. and ZWIERS, F. (2002). *Statistical Analysis in Climate Research*. Cambridge Univ. Press, Cambridge.

WHEELER, M. and KILADIS, G. N. (1999). Convectively coupled equatorial waves: Analysis of clouds and temperature in the wavenumber-frequency domain. *J. Atmospheric Sci*. **56** 374–399.

WIKLE, C. and ANDERSON, C. (2003). Climatological analysis of tornado report counts using a hierarchical Bayesian spatiotemporal model. *J. Geophys. Res* **108** 9005.

WIKLE, C. K. and BERLINER, L. M. (2007). A Bayesian tutorial for data assimilation. *Phys. D* **230** 1–16. MR2345198

WIKLE, C. K. and HOLAN, S. H. (2011). Polynomial nonlinear spatio-temporal integro-difference equation models. *J. Time Series Anal*. **32** 339–350. MR2841788

WIKLE, C. K. and HOOTEN, M. B. (2010). A general science-based framework for dynamical spatio-temporal models. *TEST* **19** 417–451. MR2745992

WIKLE, C., MILLIFF, R. and LARGE, W. (1999). Surface wind variability on spatial scales from 1 to 1000 km observed during TOGA COARE. *J. Atmospheric Sci*. **56** 2222–2231.

WIKLE, C. K., MILLIFF, R. F., NYCHKA, D. and BERLINER, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *J. Amer. Statist. Assoc*. **96** 382–397. MR1939342

WILKS, D. (2011). *Statistical Methods in the Atmospheric Sciences* **100**. Academic Press, San Diego, CA.

WUNSCH, C. (1994). Dynamically consistent hydrography and absolute velocity in the eastern North Atlantic Ocean. *J. Geophys. Res*. **99** 14071–14090.

WUNSCH, C. (1996). *The Ocean Circulation Inverse Problem*. Cambridge Univ. Press, Cambridge. MR1410266

ZHANG, H. M. and HOGG, N. (1992). Circulation and water mass balance in the Brazil Basin. *J. Mar. Res*. **50** 385–420.

ZIKA, J., MCDOUGALL, T. and SLOYAN, B. (2010). A tracer-contour inverse method for estimating ocean circulation and mixing. *J. Phys. Ocean*. **40** 26–47.