DISTRIBUTION STATEMENT A: Distribution approved for public release; distribution is unlimited.

Bayesian Hierarchical Model Characterization of Model Error in Ocean Data Assimilation and Forecasts

Ralph F. Milliff
Colorado Research Associates Division, NWRA
3380 Mitchell Lane
Boulder, CO 80301
phone: (303) 415-9701 fax: (303) 415-9702 email: milliff@cora.nwra.com

Christopher K. Wikle
Department of Statistics, University of Missouri
146 Middlebush
Columbia, MO 65211
phone: (573) 882-9659 fax: (573) 884-5524 email: wikle@stat.missouri.edu

L. Mark Berliner and Radu Herbei
Department of Statistics, The Ohio State University
1958 Neil Ave.
Columbus, OH 43210
phone: (614) 292-0291 fax: (614) 292-2096 email: mb@stat.osu.edu

Award Number: N00014-10-C-0354

LONG-TERM GOALS

Quantitative uncertainty management attributes of the Bayesian Hierarchical Model (BHM) methodology are applied to the identification, characterization, and modelling of irreducible model error in ocean data assimilation and forecast systems.

OBJECTIVES

We describe 4 objectives addressed in the fiscal year October 2011 - September 2012.

First, the modifications of a control vector variable in the Regional Ocean Model System (ROMS) four-dimensional variational (4dvar) data assimilation system (Moore et al., 2011a,b,c) are compared with posterior distributions of a surface wind BHM adapted to the California Current System (CCS) from its original implementation in the Mediterranean Sea (Milliff et al., 2011).

Second, we investigate how Feynman-Kac representations of solutions to advection-diffusion

models help characterize model error resulting from mapping observations to a pre-determined physical model grid.

Third, we develop an efficient way to represent time-varying error covariances so that they can easily accommodate covariate information. Given that error covariance matrices in complicated systems are expected to be nonstationary, a flexible framework should allow for changes due to internal variability as well as variations in external conditions. The main challenges in incorporating these components into a model for time-varying covariances are the dimensionality of the state process and the requirement that any model for a covariance must be positive definite.

Fourth, we account for dynamical model uncertainty by the use of statistical emulators of deterministic models within the context of data assimilation. The statistical emulators are efficiently placed within a Bayesian hierarchical framework to account for uncertainty in parameters as well as different potential models. This project is tied to work on a related project funded by the NSF US Globec Program (see "Related Projects" below).

APPROACH

Comparing 4dvar Increments with a Surface Wind BHM: The surface wind BHM due to Milliff et al. (2011) is adapted to the CCS, with data stage inputs from the QuikSCAT scatterometer data record and analyses of the Coastal Ocean Atmosphere Mesoscale Prediction System (COAMPS). The posterior distribution includes daily surface vector winds at 0.33° resolution in the CCS. These surface wind distributions are compared, over 10-day analysis-forecast cycle, with increments in the surface stress control vector for the ROMS 4dvar system. As a temporary measure, surface stress is related to surface wind through a sequence of regressions, modeling wind speed as a function of stress and the conversion factor $\rho_a c_d |u|$ as a function of wind speed (where ρ_a is atmospheric density, c_d is a drag coefficient and $|u|$ is the wind speed amplitude).

In regions where the increments of the 4dvar procedure are driven far from the modes of the BHM posterior distributions, we can suspect model error is driving increments and not errors in surface stress.

Model Error Arising from a Discrete Grid: We derive a general probabilistic representation for a solution to an advection-diffusion equation with Dirichlet boundary conditions. Such models are used to link tracer concentration observations to ocean circulation (velocities, diffusion coefficients). In this framework, one can approximate the solution of the advection-diffusion equation at any location of interest via a Monte Carlo simulation. This lifts the requirement of selecting a *model grid* and consequently mapping the data to this grid. The output of the Monte Carlo simulation is used in a BHM framework to explore the posterior distribution for velocities and diffusion coefficients, conditional on observed tracer concentrations. Although we are using Monte Carlo approximations for the forward model (advection-diffusion equation), our approach is *exact* from a statistical perspective.

Time-Varying Error Covariance Models: Our approach is to construct a new covariate based parameterization of a spatio-temporal error covariance process via the so-called "cepstral"

formulation. This approach makes use of the duality between spectral space and covariance space through modeling the log-spectrum of a spectral density $f(\cdot)$ with frequency $\lambda \in [-\pi, \pi]$,

$$\log f(\lambda) = \sum_{k=0}^{p} \theta_k \cos(k\lambda).$$

This framework conveniently allows for the parameterization of the covariance function based on the cepstral coefficients θ_k, through the autocovariance γ_h at lag h, through an infinite order moving-average model, MA(∞), representation where,

$$\gamma_h = 2\pi \exp \theta_0 \sum_{j=h}^{\infty} \phi_j \phi_{j-h}, \quad \phi_j = \frac{1}{2j} \sum_{k=1}^{j} k\theta_k \phi_{j-k}.$$

This autocovariance can be suitably estimated and computed efficiently through a Fast Fourier Transform. Using this formulation, a valid covariance matrix can be constructed using only the cepstral coefficients on unbounded support of θ_k. Furthermore, due to the exponential decay of ϕ_j, a covariance can be constructed using relatively few cepstral parameters.

For a vector of cepstral coefficients $\theta_t = (\theta_{1t}, \ldots, \theta_{pt})$ at time t, propagator matrix M, covariate vector x_t, and covariate coefficient matrix B, the time-evolving spatial covariance can be modeled

$$\theta_t = M\theta_{t-1} + Bx_t + \eta_t,$$

where η_t is a correlated error process. This framework is flexible in accounting for internal time-variation (through the autoregressive component) and external (covariate-based) variation. For example, setting $M \equiv 0$ allows only the covariates to influence the covariance, while setting $B \equiv 0$ gives a pure autoregressive dynamic structure.

The time-varying covariance model is considered within a BHM framework to allow for better uncertainty quantification. Namely, for data δ_t, error process e_t, and cepstral coefficients $\xi_t \equiv \{\theta_0, \theta_{1t}, \ldots, \theta_{pt}\}$,

$$\begin{aligned}
\delta_t &= H_t e_t + \varepsilon_t \quad \varepsilon_t \sim \mathrm{Gau}(0, \sigma_\varepsilon^2 I), && (1)\\
e_t|\xi_t &\sim \mathrm{Gau}\left(0, \Sigma_e(f_{\xi_t})\right), && (2)\\
\theta_t &= M\theta_{t-1} + Bx_t + \eta_t, \\
\theta_0 &\sim \mathrm{Gau}(0, \Sigma_0), && (3)\\
\mathrm{vec}(B) &\equiv \beta \sim \mathrm{Gau}(0, \sigma_\beta^2 I), \\
\sigma_\varepsilon^2 &\sim \mathrm{IG}(q, r), \\
\Sigma_\eta^{-1} &\sim \mathrm{Wishart}\left((\nu Q)^{-1}, \nu\right), \\
\mathrm{vec}(M) &\equiv m \sim \mathrm{Gau}_{p^2}\left(0, \Sigma_m\right),
\end{aligned}$$

where hyperparameters $\Sigma_0, \sigma_\beta^2, q, r, \nu, Q, \Sigma_m$ are specified, and the notation $\mathrm{vec}(B)$ means stacking the columns of the matrix B on top of each other to create a long vector.

Emulator Assisted Data Assimilation: Most data assimilation problems can be characterized by a state-space model consisting of an observation equation, e.g.,

$$\mathbf{Z}_t = \mathbf{H}(\theta_h)\mathbf{Y}_t + \varepsilon_t, \tag{4}$$

and state equation, e.g.,

$$\mathbf{Y}_t = \mathscr{M}(\mathbf{Y}_{t-1}; \theta_m) + \eta_t, \tag{5}$$

$\mathbf{Z}_t$ is an $m_t \times 1$ observation vector, $\mathbf{Y}_t$ is an $n \times 1$ state vector, $\mathbf{H}(\theta_h)$ is a mapping/observation function with parameters θ_h, $\mathscr{M}(\cdot)$ is a state transition function that depends on parameters θ_m, and the additive error processes $\{\varepsilon_t\}$ and $\{\eta_t\}$ are mutually independent, independent in time, zero mean, and with variance-covariance matrices $\mathbf{R}_t$ and $\mathbf{Q}_t$, respectively. Although more complicated error processes are possible (e.g., Cressie and Wikle, 2011), this structure suffices to illustrate the methodology. The role of the matrix $\mathbf{H}(\theta_h)$ is to bring the observations to the state process, and to accommodate change of support, alignment, nonlinear transformation, and/or missing observations. In addition, this function may serve the role of projecting from observation space into a lower-dimensional manifold that accommodates the dynamical evolution. The role of $\mathscr{M}(\cdot)$ is then to evolve the process forward in time dynamically, accommodating interactions (linear or nonlinear) in the state-process. Typically, $\mathscr{M}(\cdot)$ is a deterministic model such as a numerical solution to differential equations (e.g., in the case of fluid dynamical models of dispersion). One of the biggest challenges in data assimilation for high-dimensional systems is the expense of running such a deterministic model.

In recent years, the use of statistical emulators (or surrogates) for complicated computer models have proven an effective way to perform parameter inference and calibration to remediate the cost of running the deterministic computer models (e.g., Sacks et al. 1989; Kennedy and O'Hagan, 2001). There has been a substantial growth in the literature in this area incorporating dimension reduction (Higdon et al., 2008; Hooten et al. 2011), dynamical models (Drignei, 2008; Conti et al., 2009), multivariate output (Rougier, 2008; Conti et al. 2010) and hierarchical formulations (Hooten et al. 2011). Most emulation approaches in statistics have been concerned with modeling response surfaces through second-order (covariance) properties analogous to spatial modeling in geostatistics. There has also been an interest in modeling so-called "first-order" emulators for dynamical processes; i.e., in the conditional mean (e.g., van der Merwe et al. 2007; Frolov et al. 2009; Hooten et al. 2011). Such models are closer to the essence of the dynamical process evolution and thus are likely more appropriate for data assimilation applications.

Higdon et al. (2008) describe an emulator that is based on a singular value decomposition (SVD) of multiple mechanistic model runs, given various inputs. In particular, the modeling focus is on the right-singular vectors from this SVD. For example, say we have K computer model outputs in vector form, denoted, $\mathbf{y}_k, k = 1, \ldots, K$, where the dimension of $\mathbf{y}_k$ is $n \times 1$ and each of these computer model output vectors is associated with an input vector θ_k. Then, consider the SVD $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_K) = \mathbf{UDV}'$. In this case, the right singular vectors in $\mathbf{V}$ are the projection of the computer model output onto the structures associated with the vectors of the matrix $\mathbf{UD}$. Hooten et al. (2011) describe a first-order emulator that models the relationship between these right-singular vectors and the inputs, θ, directly in the conditional mean, making the case that this

is simple and, for dynamical emulators, more true to the underlying process. Thus, in the first-order setting, one builds models for $\mathbf{v}(\theta)$ in the expression, $\mathbf{y} = \mathbf{UDv}(\theta) + \varepsilon$, where $\mathbf{y}$ is observed data, $\mathbf{U}$ and $\mathbf{D}$ are fixed, as obtained from the SVD of the computer model output, and one builds a stochastic model for $\mathbf{v}(\theta)$ given a prior distribution on the input vector θ. Hooten et al. (2011) describe both linear and nonlinear emulators of this type, which depend on parameters that must be estimated based on the computer model output.

To build a dynamical one-step ahead emulator in this framework, let $\mathbf{y}_k$ correspond to the computer model output at the t-th time, say $\mathbf{y}_t$, and the "input" θ_k correspond to the previous time, say $\mathbf{y}_{t-1}$ (note, however, that we can also include other inputs in such a dynamical emulator). In this case, the modeling effort is still directed at the right singular vectors, which are now assumed to be a function of the past value of the computer model, i.e., $\mathbf{v}(\mathbf{y}_{t-1})$. This framework also allows a natural dimension reduction in terms of the leading structures in $\mathbf{UD}$. That is, we can write

$$\mathbf{y}_t = \mathbf{U}^{(1)}\mathbf{D}^{(1)}\mathbf{v}^{(1)} + \mathbf{U}^{(2)}\mathbf{D}^{(2)}\mathbf{v}^{(2)},$$

where $\mathbf{U}^{(1)}\mathbf{D}^{(1)}$ correspond to the first q columns of $\mathbf{UD}$ from the SVD above and $\mathbf{U}^{(2)}\mathbf{D}^{(2)}$ correspond to the remaining columns. The choice of the number of columns q in $\mathbf{U}^{(1)}\mathbf{D}^{(1)}$ should be made to account for a significant portion of the variation in the computer model output. In the case where the vectors $\mathbf{y}_t$ correspond to spatial fields, the columns of $\mathbf{UD}$ are the empirical orthogonal functions (EOFs) of the computer model output, which account for decreasing amounts of variation in the output with increasing order, and where the modes corresponding to largest variation also typically correspond to the largest scale spatial structures (i.e., see the overview in Cressie and Wikle, 2011, Chap. 5). It is thus a reasonable assumption to allow the right singular vectors associated with these largest modes of variability, $\mathbf{v}^{(1)}$, to depend dynamically on the past values of the right singular vectors (recall $\mathbf{v}_t = (\mathbf{U}^{(1)}\mathbf{D}^{(1)})'\mathbf{y}_t$), while the smaller scale components associated with $\mathbf{v}^{(2)}$ need not evolve dynamically. That is, we seek to build a statistical model for $\mathbf{v}(\mathbf{v}_{t-1})$, which will depend on unknown parameters that can be estimated from the computer model output. A powerful component of this framework is that $\mathbf{v}_t$ is of dimension $q << n$, and this rank reduction reduces the number of parameters that are necessary to describe the dynamical evolution. In the context of data assimilation, such an emulator provides a reduced-rank approach to blending mechanistic models and observations.

WORK COMPLETED

Comparing 4dvar Increments with a Surface Wind BHM: Posterior distributions for four-times daily surface vector winds were obtained from the surface wind BHM in the CCS for the calendar year 2003. This coincides with a period for which ROMS 4dvar forecasts and analyses have been run in the CCS by collaborators (Smith, Moore, Edwards) at University of California, Santa Cruz. Comparisons of the surface stress component of the 4dvar control vector and the surface wind posterior distributions were made for forecast cycles in March 2003. Regions of large discrepancies in control vector vs. BHM were noted and animated over the course of the month.

Future work will adapt to the CCS a surface stress summary of the surface wind BHM posterior distribution for the Mediterranean Sea. This removes the need for regressions to relate surface

wind to surface stress. Developing BHMs for other components of the 4dvar control vector (e.g. surface heat and fresh water fluxes) will allow us to extend the methodology demonstrated here for surface stress. Increments in the 4dvar procedure that drive vectors and amplitudes to extremes of posterior distributions for all components of the control vector are most likely due to irreducible model error. The dynamical evolution of model error can be studied using space-time inferences of these kinds.

Model Error Arising from a Discrete Grid: We implemented the general probabilistic representation of a solution to the advection-diffusion equation using tracer observations collected during the WOCE experiment in the South Atlantic Ocean. We use the physical model described in McKeague et al. (2005) and rather than using a FORTRAN solver for the advection-diffusion equation on a pre-specified grid, we use the Feynman-Kac representation. We use a parallel computing environment, consisting of several Graphics Processing Units (GPUs). We also developed the associated statistical methodology which allows us to sample the correct posterior distribution, eliminating the errors associated with the Monte Carlo simulation.

Time-Varying Error Covariance Models: The theoretical derivations and simulation testing of the cepstral model have been completed. We have preliminary results for the sea-surface temperature (SST) long-lead prediction example and are testing various model combinations. Testing should be completed by the first week of October, 2012. A complete draft manuscript is written and, pending the final results, will be submitted in October 2012. We will next move, simultaneously, to time-varying models for the observation error and model error covariances.

Emulator Assisted Data Assimilation: This work was written-up and submitted to the journal *Statistical Methodology*. We have received positive comments and were encouraged to revise the manuscipt. We are in the final stages of implementing some of the changes suggested by the reviewers and will resubmit the manuscript by the first week in October. We intend to investigate potential benefits from allowing switching between different process models in this setting. This will be greatly facilitated by the emulator approach as it will allow us to consider multiple models fairly inexpensively.

Relevant Meetings and Presentations:

(Wikle) Nonlinear dynamic spatio-temporal statistical models.
– *International Invited Talk*, Norwegian Computing Center, University of Oslo, Oslo, Norway; September 13, 2011.
– *Invited Talk*, University of California, Santa Cruz, Department of Applied Mathematics and Statistics, Santa Cruz, CA; October 17, 2011.
– *Invited Talk*, University of Illinois, Department of Statistics, Champaign-Urbana, IL; November 29, 2011.

(Wikle) A hierarchical Bayesian statistical perspective on spatio-temporal dynamics. *Invited Talk*, Courant Institute of Mathematical Sciences, New York University, New York, NY; October 5, 2011.

(Milliff) ONR Code 32 Review Presentation, Denver, CO; November 2011.

(Herbei, Milliff, Moore) Presentations at ONR Model Error Project Meeting, Courant Institute of Mathematical Sciences, New York University, New York, NY; November, 2011.

(Milliff) Visit to INGV for work on MFS-Error-BHM, Bologna, Italy; February 2012.

(Edwards, Milliff, Moore) Experimental design considerations, on-site visit with collaborators at Univ. California, Santa Cruz; April, 2012.

(Milliff) Presentation at Rotating, Stratified Turbulence Workshop, NCAR Geophysical Turbulence Program, Univ. Colorado, Boulder, CO; May 2012.

(Wikle) Statistical methods for nonlinear dynamic spatio-temporal models.
– *International Invited Talk*, French Statistical Society Meeting (JdS'2012), Brussels, Belgium; May 25, 2012.
– *International Invited Talk*, ASC 2012, Conference of the Statistical Society of Australia, Adelaide, Australia; July 9, 2012.

(Milliff) Model Error presentation at International Ocean Vector Winds Science Team Meeting, Utrecht, Netherlands; June 2012.

(Herbei, Berliner) Estimating ocean-circulation: a likelihood-free approach via a Bernoulli factory. Presented at :
– International Society for Bayesian Analysis World Meeting, Kyoto, JP; June, 2012.
– Seminar : Ohio State University, Statistics Department; Sept, 2012
– Seminar: Purdue University, Statistics Department; Oct 5th, 2012.

(Wikle) Data assimilation, data fusion, and emulators: A gentle introduction. *International Invited Talk*, CSIRO Great Barrier Reef Pollutant Load Workshop, Brisbane, Australia; July 3, 2012.

(Milliff) External Review Panel, NRL SSC Oceanography Program, Bay St. Louis, MI; July 2012.

(Herbei, Milliff, Moore, Wikle) Informal presentations and discussions at the annual "All-Hands" project meeting (Confab) at CIRES/Univ. Colorado, Boulder, CO; August 2012.

(Milliff) Presentation at Coupled Data Assimilation Workshop, NCAR, Boulder, CO; August 2012.

RESULTS

Comparing 4dvar Increments with a Surface Wind BHM: Figure 1 depicts a snapshot on 23 March 2003 comparing the initial and final values of the surface stress control vector (the 4dvar prior is the green vector and the 4dvar posterior is the blue vector) with a sample from the BHM posterior for surface winds (BHM samples from the posterior in red and the posterior mean wind is the black vector). Surface stress has been converted to surface wind for this comparison. A region off Northern California is identified wherein the surface stress control vector does not change very much from initial to final increment (i.e. blue overlies green), while the control

vector realizations both lie outside the BHM ensemble. We reason that these are locations where information from the scatterometer data stage inputs are correcting the COAMPS forcing. The ROMS 4dvar forecast calculations do not involve scatterometer winds. More importantly, the identified region centered on about $128°W$ and $34°N$ is an example of possible model error. Here the inital vector in the surface stress component of the 4dvar control vector lies within the sample from the BHM posterior distribution. But, 4dvar increments are driving the final vector outside the BHM posterior, hinting at an error source other than errors in the surface wind forcing.

Model Error Arising from a Discrete Grid: We developed a novel statistical MCMC approach which does not require an exact evaluation of the likelihood function. Our method is based on the recent developments by Flegal and Herbei (2012) who devise a fast Bernoulli Factory algorithm. This allows us to simulate a Metropolis-Hastings sampler which only requires unbiased estimates of the likelihood function. The decision whether to accept or reject a proposed state in the Metropolis-Hastings algorithm is made using the Bernoulli Factory.

We apply this method to tracer observations in a deep layer of the South Atlantic Ocean. In Figure 2 we compare a Fortran solver (right panel) to a Feynman-Kac solver (left panel). The two methods are nearly indistinguishable; however, the Feynman-Kac approach has the advantage that the solution can be obtained at any location in the domain. In Figure 3 we show our estimated posterior distributions for the zonal and meridional diffusion coefficients. Compared to previous estimates from grid-based analyses, we notice an increase in the meridional diffusion: $300m^2s^{-1}$ vs. $\sim 100m^2s^{-1}$ (McKeague et al., 2005).

Time-Varying Error Covariance Models: We demonstrated the power of the cepstral model through two simulated examples and through an SST forecasting example. We consider the difference between the Zebiak-Cane El Niño-Southern Oscillation 6-month lead forecast and observed SST anomalies along the equator using the Southern Oscillation Index (SOI) and Pacific Decadal Oscillation (PDO) index as covariates for the period January 1972 through December 1994. Three different error process models were tested: the full proposed model from (2) (hereafter called ARCOV); a model setting $B \equiv 0$ (AR model); and a model with $M \equiv 0$ (COV model). Also, the model was tested for various lengths p for the vector θ_t (i.e., $p = 2, \ldots, 8$). Utilizing the Deviance Information Criterion for model selection, preliminary results suggest that the ARCOV($p = 6$) model is the best model, with COV($p = 6$) also a good model. We note that the posterior distributions of the parameters associated with the two covariates in the ARCOV($p = 6$) model cover 0. In contrast, the COV model suggests that the PDO is significant in informing the time-evolving covariance while the SOI is not. Figure 4 shows the results from the ARCOV($p = 6$) model.

Emulator Assisted Data Assimilation: We focus on lower-trophic level marine ecosystem dynamics and an associated coupled physical-biological model that mimics this system. The dynamical ocean circulation component of the model is an implementation of ROMS (Haidvogel et al. 2008), and the ocean ecosystem component is a six-compartment Nutrient-Phytoplankton-Zooplankton-Detritus (NPZD) model (Powell et al. 2006), with iron limitation (NPZDFe; Fiechter et al. 2009). A description of the coupled model can be found in Fiechter et al. (2011). The lower-trophic level ecosystem model is coupled to the ocean

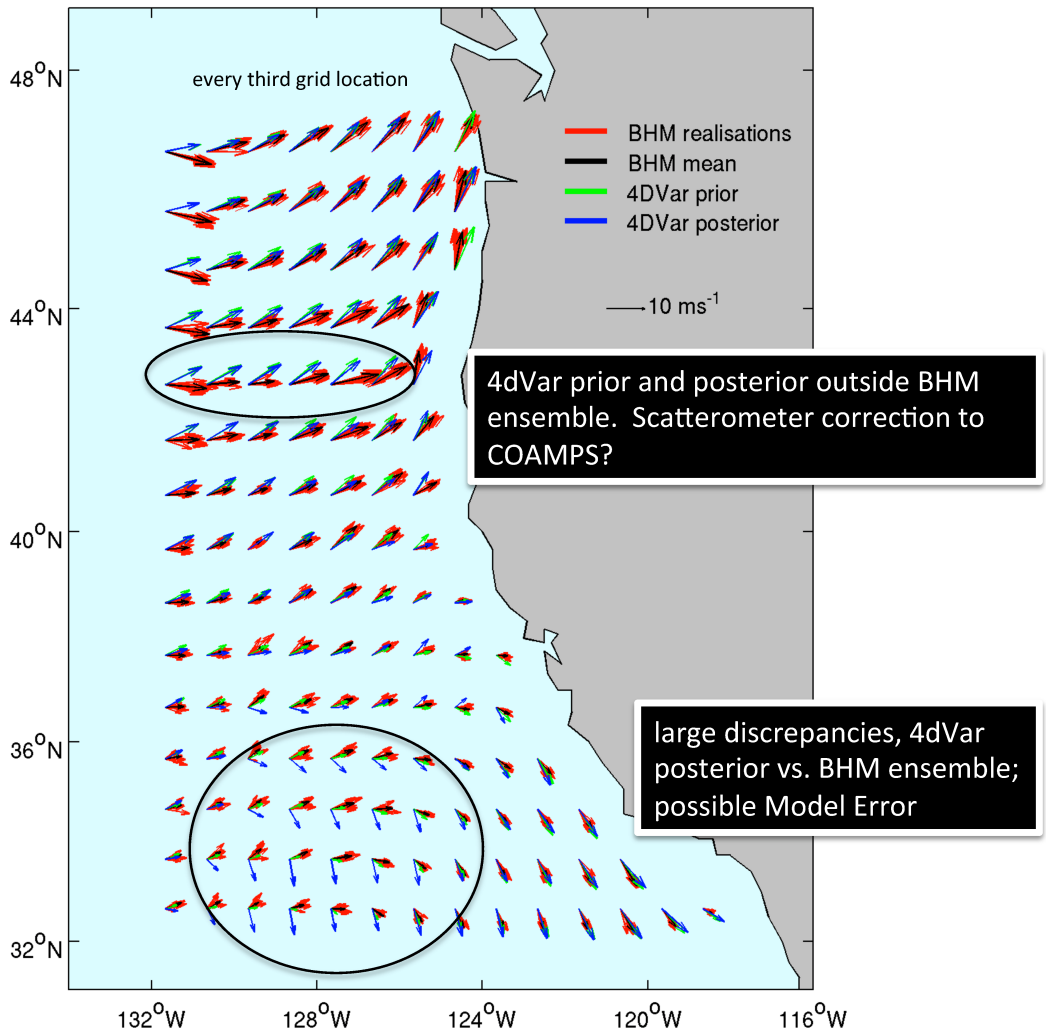

Figure 1: ***Snapshot of comparison between initial (green) and final (blue) vectors from the surface wind implied by the ROMS 4dvar control vector and samples from the posterior distribution for surface wind in the California Current System from a Bayesian Hierarchical Model. The BHM ensemble from the posterior distribution is shown in red vectors and the posterior mean vector is black. Comparisons are shown for every third grid location in the CCS domain for a snapshot on 23 March 2003. Regions are identified wherein surface winds from COAMPS are being corrected by scatterometer data stage inputs in the BHM (off N. California) and regions where the increments in the 4dvar are driving the initial vector (originally within the BHM ensemble) outside the range suggested by the BHM (see discussion in text).***

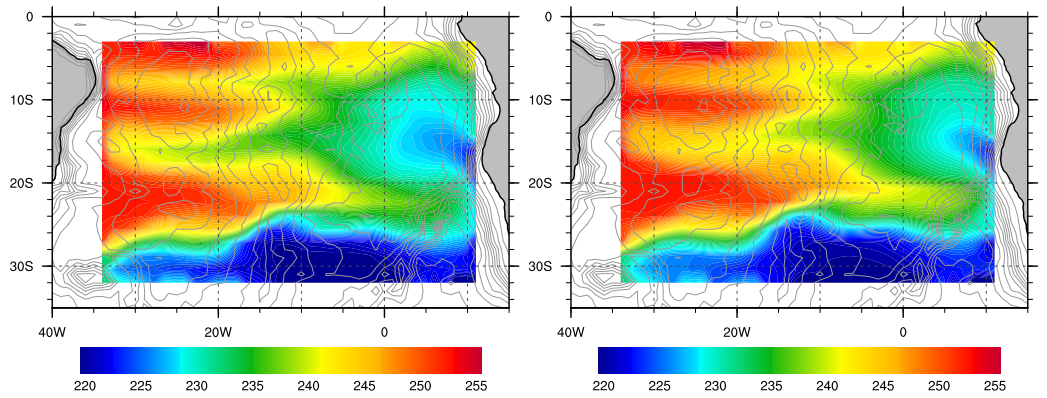

Figure 2: A comparison between a Feynman-kac solver (left) and FORTRAN solver (right)

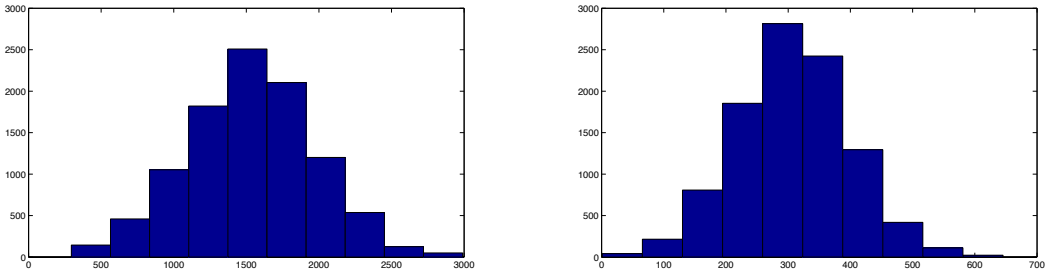

Figure 3: A comparison between a Feynman-kac solver (left) and FORTRAN solver (right)

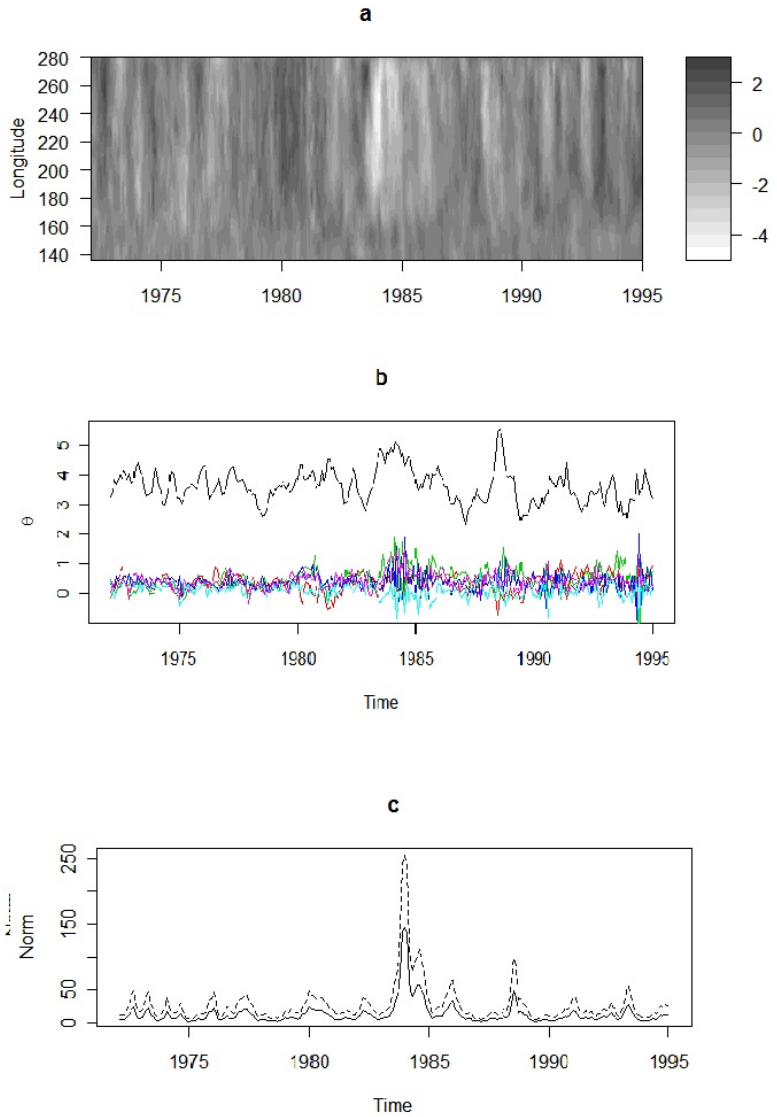

Figure 4: Results of ARCOV $p = 6$ model for ENSO. a) Top plot shows the posterior mean of process e_t, January 1972 - December 1994, for longitudes $136°E$ to $80°W$. b) Middle plot shows evolution of components of θ_t, January 1972 - December 1994 (black is θ_{1t}, colored lines $\theta_{it}, i = 2, \ldots, 6$). c) Bottom plot shows the Frobenius norm (solid) and trace (dashed) of the resulting spatially referenced covariance from January 1972 through December 1994. A pronounced model error is indicated around 1983 as well as a smaller pronounced model error covariance in middle 1988.

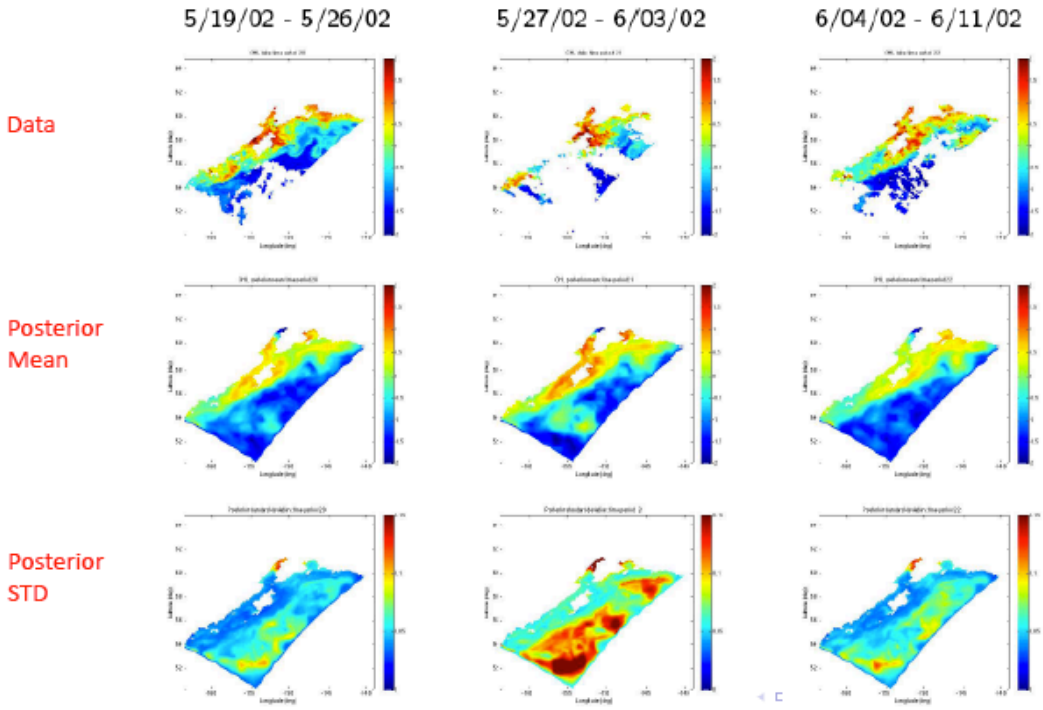

Figure 5: The top panel of each column shows the log of the SeaWiFS ocean color data. The middle panels provide the posterior mean estimate over the entire study area. The bottom panels display the posterior standard deviation. Three consecutive 8-day periods are shown for late May and early June, 2002.

circulation model by solving a transport equation in ROMS for each of the NPZDFe model compartments at each time step. A vertical sinking velocity is also specified for detritus in this coupled framework. Each realization was generated from 1998 through 2002 using the ROMS-NPZDFe coupled physical-biological model for the northwestern Coastal Gulf of Alaska (CGOA; from ca. $50^\circ - 62^\circ N$ and $140^\circ - 164^\circ W$).

For the data assimilation experiments, we trained the emulator on 8-day averages of sea surface height (SSH), SST, and phytoplankton (chlorophyll) data from 1998, 1999, 2000, and 2001. We then used remotely sensed SeaWiFS ocean color data for 2002 as our data. Preliminary results are shown in Figure 5.

IMPACT/APPLICATIONS

Our research thus far, demonstrates the wide scope of applicability of the BHM methodology in characterizing, identifiing and modelling irreducible model error in ocean forecast systems. Our work is leading to operationally useful estimations of the space-time properties of uncertainties in these systems.

TRANSITIONS

Presentations and discussions at the Bayesian Confab meeting in Boulder this August focused on Irreducible Model Error issues (see www.cora.nwra.com ; password available on request).

Milliff served as one of 5 reviewers on an External Review Panel for the 6.1 and 6.2 oceanography programs in the Battlespace Environments focus area of the Naval Research Laboratory, Stennis Space Center, Bay St. Louis, MI (29 July - 2 August 2012).

RELATED PROJECTS

"Bayesian Hierarchical Models to Augment the Mediterranean Forecast System", ONR Physical Oceanography Program, May 2009 - May 2011.

"Estimating Ecosystem Model Uncertainties in Pan-Regional Syntheses and Climate Change Impacts on Coastal Domains of the North Pacific Ocean", NSF US Globec Program, October 2009 - September 2012.

"Quantifying the Amplitude, Structure and Influence of Model Error during Ocean Analysis and Forecast Cycles", ONR Physical Oceanography Program, A. Moore (PI).

"Ocean Surface Vector Winds in Multi-Platform Bayesian Hierarchical Model Applications", International Ocean Vector Winds Science Team, NASA Physical Oceanography Program, R. Milliff (PI).

REFERENCES

Conti S, J. Gosling, J. Oakley and A. O'Hagan, 2009: "Gaussian process emulation of dynamic computer codes", *Biometrika*, **96**(3), 663-676.

Conti S and A. O'Hagan, 2010: "Bayesian emulation of complex multi-output and dynamic computer models", *J. Stat. Plan. Infer.*, **140**(3), 640-651.

Cressie, N. and C.K. Wikle, 2011: **Statistics for Spatio-Temporal Data**, John Wiley and Sons Inc., 588 pgs.

Drignei D., 2008: "Fast statistical surrogates for dynamical 3-D computer models of brain tumors", *J. Comput. Graph. Stat.*, **17**(4), 844-859.

Fiechter, J., and A.M. Moore, 2009: "Interannual Spring Bloom Variability and Ekman Pumping in the Coastal Gulf of Alaska", *Journal Geophysical Research*, **114**, C06004.

Fiechter, J., A.M. Moore, C.A. Edwards, K.W. Bruland, E. Di Lorenzo, C.V.W. Lewis, T.M. Powell, E.N. Curchitser and K. Hedstrom, 2009: "Modeling Iron Limitation of Primary

Production in the Coastal Gulf of Alaska", *Deep Sea Research II*, **56**, 25032519.

Flegal, J. and R. Herbei, 2012: "Exact sampling for intractable probability distributions via a Bernoulli factory", *Electronic Journal of Statistics*, **6**, 10-37.

Frolov S., A. Baptista, T. Leen, Z. Lu and R. van der Merwe, 2009: "Fast data assimilation using a nonlinear Kalman Filter and a model surrogate: An application to the Columbia River estuary", *Dynam. Atmos. Oceans*, **48**(1-3).

Haidvogel, D. B., H. Arango, W.P. Budgell, B.D. Cornuelle, E.N. Curchitser, E. Di Lorenzo, K. Fennel, W.R. Geyer, A.J. Hermann, L. Lanerolle, J. Levin, J.C. McWilliams, A.J. Miller, A.M. Moore, T.M. Powell, A.F. Shchepetkin, C.R. Sherwood, R.P. Signell, J.C. Warner and J. Wilkin, 2008:, "Ocean Forecasting in Terrain-Following Coordinates: Formulation and Skill Assessment of the Regional Ocean Modeling System", *Journal of Computational Physics*, **227**, 3595–3624.

Herbei, R. and L.M. Berliner, 2012: "Estimating ocean circulation: a likelihood-free MCMC approach via a Bernoulli factory" *Journal of American Statistical Association – Applications and Case Studies*, submitted.

Higdon D., J. Gattiker, B. Williams and M. Rightley, 2008: "Computer model calibration using high-dimensional output", *J. Am. Stat. Assoc.*, **103**(482), 570-583.

Hooten M., W. Leeds, J. Fiechter and C.K. Wikle, 2011: "Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models", *J. Agr., Biol., Envir. Stat.*, **16**(4), 475-494.

Kennedy, M.C. and A. O'Hagan, 2001: "Bayesian calibration of computer models", *Journal of the Royal Statistical Society, Series B*, **63**, 425-464.

McKeague, I.W., G. Nicholls, K. Speer, and R. Herbei, 2005: "Statistical inversion of South Atlantic circulation in an abyssal neutral density layer", *J. Mar. Research*, **63**, 683-704.

Milliff, R.F., A. Bonazzi, C.K. Wikle, N. Pinardi and L.M. Berliner, 2011: "Ocean Ensemble Forecasting, Part I: Mediterranean Winds from a Bayesian Hierarchical Model", *Quarterly Journal of the Royal Meteorological Society*, **137**, 858-878.

Moore, A.M., H.G. Arango, G. Broquet, B.S. Powell, J. Zavala-Garay and A.T. Weaver, 2011a: "The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems. Part I: System overview and formulation", *Progress in Oceanography*, **91**, 34-49.

Moore, A.M., H.G. Arango, G. Broquet, C. Edwards, M. Veneziani, B. Powell. D. Foley, J. Doyle, D. Costa and P. Robinson, 2011b: "The Regional Ocean Modeling System (ROMS)

4-dimensional variational data assimilation systems. Part II: Performance and application to the California Current System", *Progress in Oceanography*, **91**, 50-73.

Moore, A.M., H.G. Arango, G. Broquet, C. Edwards, M. Veneziani, B. Powell. D. Foley, J. Doyle, D. Costa and P. Robinson, 2011c: "The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems. Part III: Observation impact and observation sensitivity in the California Current System", *Progress Oceanography*, **91**, 74-94.

Powell, T. M., C.V.W. Lewis, E.N. Curchitser, D.B. Haidvogel, A.J. Hermann and E.L. Dobbins, 2006: "Results from a Three-Dimensional, Nested, Biological-Physical Model of the California Current System and Comparisons with Statistics from Satellite Imagery", *Journal of Geophysical Research*, **111**(C0), 7018.

Rougier J., 2008: "Efficient Emulators for Multivariate Deterministic Functions", *J. Comput. Graph. Stat.*, **17**(4), 827-843.

Sacks, J., W.J. Welch, T.J. Mitchell, and H.P. Wynn, 1989: "Design and analysis of computer experiments", *Statistical Science*, **4**, 409-435.

van der Merwe, R., T. Leen, Z. Lu, S. Frolov and A. Baptista, 2007: "Fast neural network surrogates for very high dimensional physical-based models in computational oceanography", *Neural Networks*, **20**(4), 462-478.

PUBLICATIONS

Flegal, J. and R. Herbei, 2012: "Exact sampling for intractable probability distributions via a Bernoulli factory", *Electronic Journal of Statistics*, **6**, 10-37.

Gladish, D.W., C.K. Wikle and S.H. Holan, 2012: "Covariate-based cepstral parameterizations for time-varying spatial error covariances", draft.

Herbei, R. and L.M. Berliner, 2012: "Estimating ocean circulation: a likelihood-free MCMC approach via a Bernoulli factory" *Journal of American Statistical Association – Applications and Case Studies*, submitted.

Hooten, M.B., W.B. Leeds, J. Fiechter and C.K. Wikle, 2011: "Assessing first-order emulator inference for physical parameters in nonlinear mechanistic models", *Journal of Agricultural, Biological, and Ecological Statistics*, **16**, 475-494.

Leeds, W.B., C.K. Wikle and J. Fiechter, 2012: "Emulator-assisted reduced-rank ecological data assimilation for multivariate dynamical spatio-temporal processes", in revision.