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

We seek to focus quantitative uncertainty management attributes of the Bayesian Hierarchical Model (BHM) methodology on the identification, characterization, and evolution of irreducible model error in ocean data assimilation and forecast systems.

OBJECTIVES

A sequence of project objectives build upon experience gained under prior Office of Naval Research (ONR) support. First, we will extend time- and space-dependent error covariance BHM from the Mediterranean Forecast System (MFS) to Regional Ocean Model System (ROMS) applications in the California Current System (CCS). Second, reduced-dimension error process models will be developed from ensembles of ROMS analyses and forecasts wherein selected model parameterizations (e.g. diffusion) are treated as random. Monte Carlo sampling algorithms will be developed to obtain posterior distributions for prescribed error models (e.g. additive, multiplicative, etc.). Third, based on the experience gained in the first and second sets of objectives, we will develop an ocean forecast model error process BHM to evolve distributions for model error.

Funding for this research arrived at the cooperating institutions in the latter half of the fiscal year (NWRA/CoRA funding in place as of late May 2010, University of Missouri funding arrived as late as August 2010). In this report, we elaborate plans and progress in pursuit of the first set of objectives.

APPROACH

Time-Varying Error Covariance Models

Consider a vector of spatially distributed, time dependent errors, denoted by e_t. Let the error processes can associated with differences between a deterministic model and its long term averages and/or differences between the deterministic model and observations for multiple state variables. The goal is to obtain the time-varying error covariance matrix, defined to be Σ_t, for the error process. In traditional linear Kalman filter approaches to data assimilation, one estimates the error covariance matrix through the Kalman recursions, updating the estimates as new data become available. In nonlinear or non-Gaussian systems, analytical forms for the estimated covariance are not available. Furthermore, in high dimensional settings, sequential importance sampling approaches that can give estimates for nonlinear and non-Gaussian systems, are not efficient and rely on potentially unrealistic approximations. These difficulties demonstrate the need for new approaches. In our research we develop a hierarchical approach to model these covariances directly, given observations of the model errors.

A critical component of our approach relies on the use of basis function expansions. Specifically, we write the $n \times n$ error covariance matrix as

$$\Sigma_t = \Phi B_t \Phi',$$

where Φ is an $n \times p$ matrix of EOFs and B_t is a $p \times p$ positive definite matrix. The important idea here is that there are a set of EOF modes that are thought to be important, yet their relative importance through time varies. This then implies that B_t is not diagonal (as it would be for the stationary EOF decomposition of the error covariance matrix). One statistical challenge is to develop an efficient model for B_t. Note that the dimension reduction (from n to p, where $n >> p$) is crucial, as it allows us to focus on models for time-varying error covariance matrices through the treatment of comparatively few parameters contained in B_t rather than the full Σ_t.

The error covariance BHM development is an extension of a BHM application in the MFS project that is in its final stages. In that application, the model for error vectors e_t is given by

$$e_t = \Phi \beta_t + \eta_t \tag{1}$$

where Φ are vertical EOF bases, β_t are time-dependent amplitudes, and $\eta_t \sim Gau(0, \sigma_\eta^2 I)$ account for additional uncertainty, such as that arising from the dimension reduction. Critically, we assume that $\beta_t \sim Gau(0, B_t)$, where, as discussed above, B_t is the time-dependent contribution to Σ_t. We write B_t in terms of its modified Cholesky decomposition (Chen and Dunson, 2003),

$$B_t = \Lambda_t \Gamma_t \Gamma_t' \Lambda_t,$$

where Λ_t is a diagonal matrix with elements proportional to the standard deviations of the elements of β_t and Γ_t is a lower triangular matrix associated with the correlations among the β_t. The hierarchical Bayesian specification allows the non-zero elements of Λ_t and Γ_t to be expressed as regression coefficients in a linear model (Chen and Dunson, 2003). In our time-varying context, these "regression"

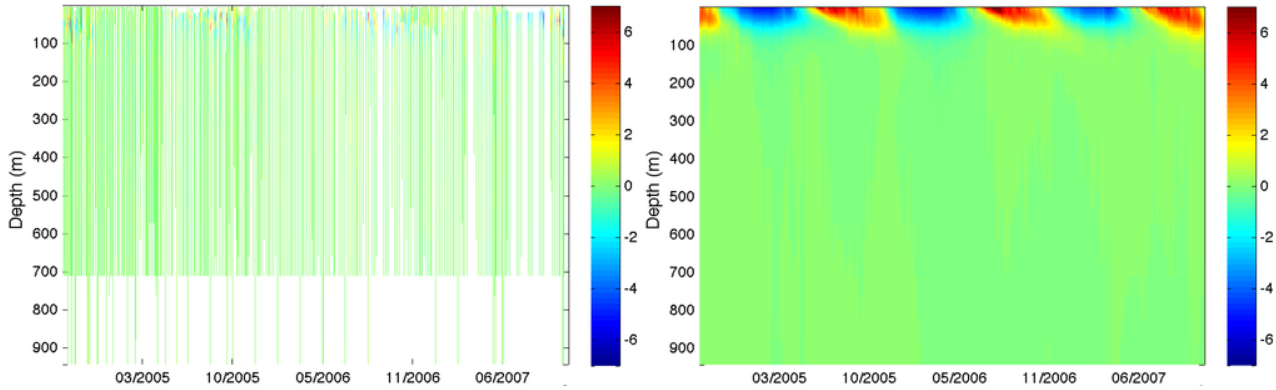

***Figure 1:** Misfit (left panel) and anomaly (right panel) data stage inputs for a time-dependent error covariance matrix BHM application in the Gulf of Lions region of the Western Mediterranean Sea. Similar datasets and BHM applications will be constructed for the CCS as part of research focus 1 in the proposed research.*

coefficients are modeled as autoregressive time series, with parameters modeled probabilistically in the BHM.

The data stage inputs to our BHM are *model misfits* d_t and *anomalies* q_t. The model misfits are forecast differences with respect to *in-situ* observations. The anomalies are departures from the model "year minus day" climatologies. These vectors can be written

$$\begin{aligned} d_t &= H_t(X_{t|t-1}) - x_t^{obs} \\ q_t &= X_{t|t-1} - \bar{x} \end{aligned} \tag{2}$$

where H_t is the operator that moves the forecast $X_{t|t-1}$ to the observation x_t^{obs} locations for comparison, and $\bar{X}_t$ is the climatology value for the model state variable X.

Figure 1a and b depict the d_t and q_t for the Gulf of Lions region of the western Mediterranean Sea for the period October 2004 through October 2007. The misfits are with respect to Argo data in the Gulf of Lions during this period.

Milliff will coordinate with Prof. Andrew M. Moore of the ROMS 4DVAR project to obtain d_t and q_t data sets from ROMS applications during interesting oceanographic events (e.g. upwelling, offshore streamer development, etc.) in the CCS.

Additional Statistical Model Development

While the MFS implementation of the modified Cholesky BHM is showing promise (see below), there are additional covariance modeling methodologies that might prove beneficial for the CCS domain. In particular, we are exploring the possibility of using so-called "mixture models" to account for rapid regime-shifts in the error covariance model. For example, consider the time-varying matrix B_t defined above. In this case, assume that B_t is controlled by parameters, say θ_t, that are time varying. The current version of the MFS BHM assumes these parameters evolve in time by a multivariat autoregressive process (i.e. a "random walk"). Alternatively, in the mixture approach, we assume that the distribution of these parameters in time corresponds to a mixture of possible distributions at each time. That is,

$$[\theta_t] = \sum_{i=1}^{q} \pi_{i,t} [\theta_t(i) | \eta_t]$$

where the bracket notation "[ ]" refers to probability distribution, $\pi_{i,t}$ corresponds to mixture probabilities, where $\pi_{i,t}$ is the probability of the distribution associated with $\theta_t(i)$ is appropriate at time t. In this case, the distribution of the possible parameters is controlled by other parameters η_t. Note that the power of the hierarchical approach is that we can then focus our modeling attention on the mixture probabilities $\pi_{i,t}$ and the controlling parameters η_t. The advantage is that scientifically meaningful covariates can be included in these lower levels of the hierarchy to suggest scientifically meaningful regimes that are likely to exhibit different error covariance properties.

Another path that is currently being explored by the graduate student at U. Missouri is a statistical time-varying covariance model that does not rely on the EOF expansion, but can still be represented in terms of small numbers of parameters. This work is in its very early stages.

Stochastic Diffusion Based MCMC

As in all of our analyses in these research projects, the computations for assessing the posterior distributions rely on Markov chain Monte Carlo (MCMC) methods. However, MCMC is severely tested in settings involving nonlinear, non-Gaussian models; particularly in high dimensions. The nature of the physical models used in our work limit the efficiency of common MCMC algorithms such as Gibbs Sampling, Metropolis-Hastings methods, and "Metropolis-within-Gibbs" hybrids. Berliner and Herbei are developing practical implementations, and identifying properties of an alternative MCMC, known as *diffusion* (or *Langevin*) MCMC. In this approach, we formulate a model for a diffusion process that is a solution to a stochastic differential equation (SDE). By choosing the drift and diffusion function of the SDE appropriately, we can insure that the stationary distribution of the diffusion process coincides with our posterior distribution. The method uses the Fokker-Planck equation and its stationary solution.

This approach may be very useful in our work in that there is no computation or simulation of the probability distributions used in Gibbs Sampling. Neither are there any direct needs for Metropolis steps. However, efficient simulation of complicated, diffusion processes in high-dimensions is stil not easy in general. We are currently developing algorithms.

WORK COMPLETED

Time-Varying Error Covariance Models

In initial experiments we found substantial disagreements between assimilation results using the MFS operational system error covariance and the BHM time-varying error covariance. Sensitivity studies suggested that the differences were due to variations in how seasonality was removed in the operational system versus the BHM, as well as how vertical level-thickness information was included in the EOF decomposition (e.g. North et al., 1982). Recent test simulations produced with BHM EOFs calculated in a fashion similar to that used in the MFS system gave much closer agreement to the MFS operational results. The latest BHM results were based on a run with just the anomaly data (i.e., q_t), so as to compare with the MFS assimilation. The time period considered was the six month span from January - May 2007.

Figure 2 shows four of the associated temperature-salinity error covariance estimates (i.e., the posterior mean) for the Gulf of Lions region during the data stage period covered in Fig. 1. Large amplitude temperature error covariances at the surface and in the upper ocean vary over the 15-day period spanned by the matrix evolution depicted in Fig. 2. Due to the inherent differences in variability in the salinity

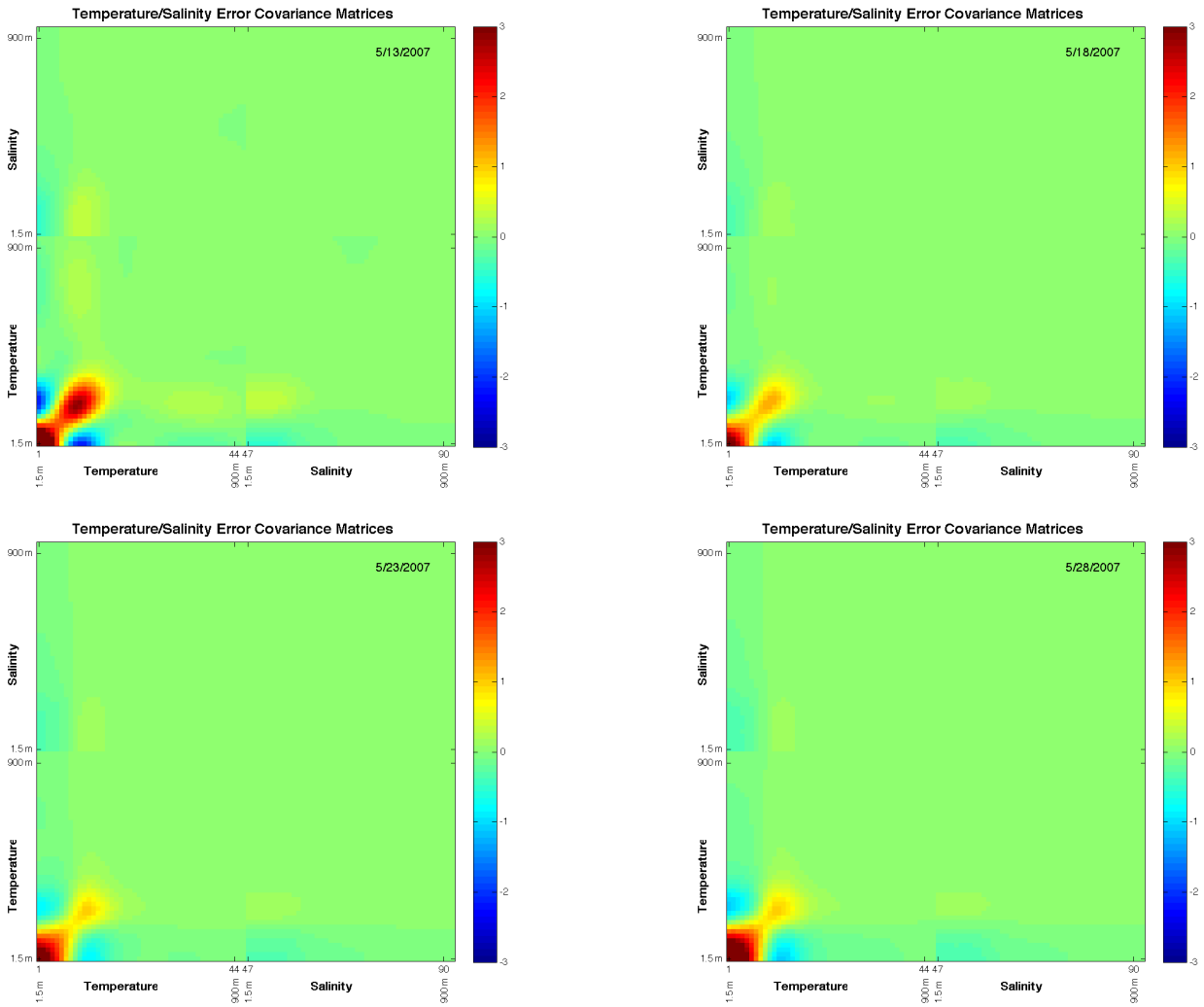

***Figure 2:** Multi-variate (T,S) error covariance matrix evolution, every 5 days from 13 May 2007 (upper left) to 28 May 2007 (lower right) from error covariance BHM in (1) given data from (3). Sub-regional error covariance characterization is planned for focus 1 of the proposed research in the CCS.*

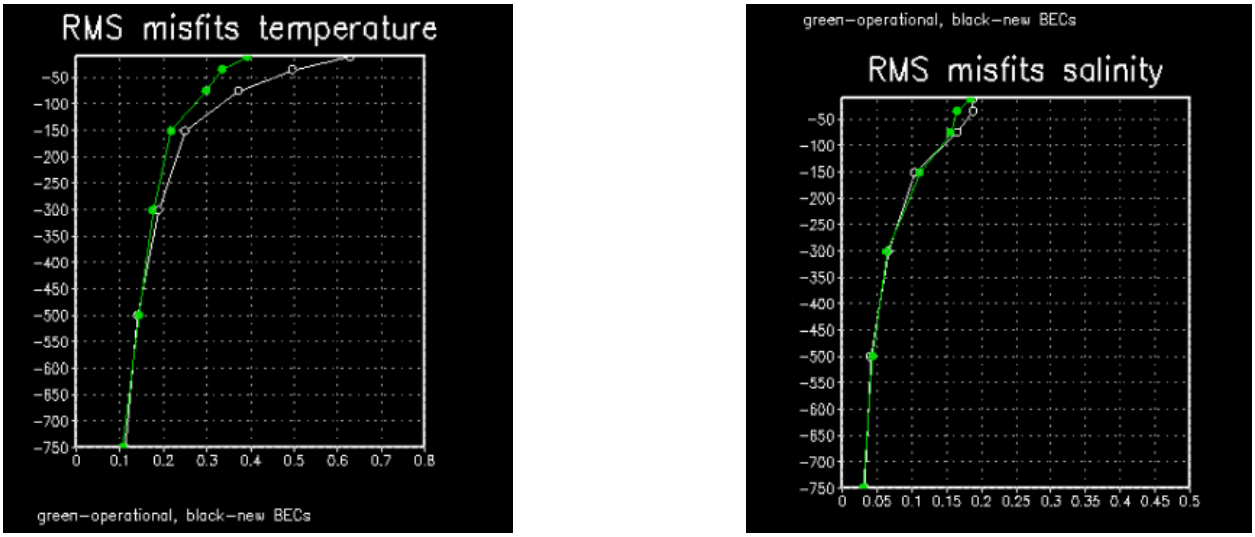

Figure 3: RMS misfits for temperature and salinity. Green lines/circles corresponds to the operational MFS assimilation and the white lines/circles correspond to the MFS system using the BHM covariances for a 6 month experiment.

anomalies, the associated variances and covariances do not stand out in these images. However, note that the covariances are modeled on a scale that does allow for the temperature-salinity cross-covariances to play a role. The figures shown here are rescaled back to the original observation space.

Comparison of the RMS misfits for the operational MFS system and the MFS system with the BHM covariances are shown for temperature, salinity and sea level anomalies (SLA) in Figures 3 and 4, respectively. These figures show that the RMS for salinity and SLA are comparable between the BHM and the operational system and, other than the temperature at upper levels, the temperature is reasonably close as well. These results are encouraging in that there was no attempt to optimize the BHM results for this time period and the actual misfit data were not used. Given the favorable comparisons of the BHM time-varying error covariances and the MFS seasonally-varying error covariances, we will finish up experiments to:

(i) consider the effect of using a trivariate EOF, adding the surface height anomaly;

(ii) add the d_t misfit data; and

(iii) contrast seasonally-varying EOFs and the effects of horizontal averaging of the anomaly data.

It is important to note that the BHM methodology is now mature and further development of this particular model from a statistical perspective is not likely to be necessary.

The MFS Med results suggest that it is useful to apply a similar methodology to e_t in the CCS domain (e.g. in the CalCOFI and Globec regions of the domain). The error covariance structures that are products of this research focus will be provided to the ROMS 4DVAR project, for application to their cost function estimation.

Relevant Presentations

(Berliner, Herbei, Milliff, Wikle) Informal presentations and discussions at the annual “All-Hands”

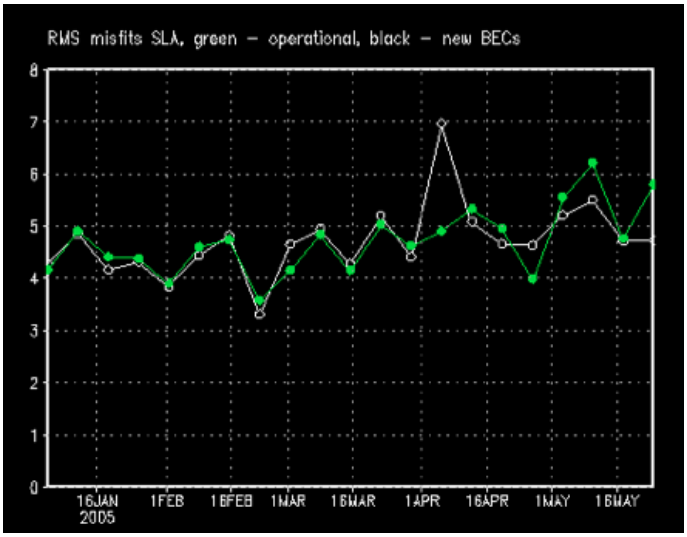

Figure 4: ***RMS misfits for SLA. Green lines/circles corresponds to the operational MFS assimilation and the white lines/circles correspond to the MFS system using the BHM covariances for a 6 month experiment.***

project meeting at NWRA/CoRA, August, 2010.

(Milliff, Wikle; session co-conveners) Probabilistic Models in Ocean Sciences: Applications in Data Assimilation, Coupled Ecosystem Models and Air-Sea Interaction Studies, American Geophysical Union, Ocean Sciences Meeting, Portland, OR, February, 2010.

(Berliner) Combining Models and Data: The Bayesian Approach to Modeling and Prediction, Invited Talk, AGU Ocean Sciences Meeting, Portland, OR, February, 2010.

(Milliff, Pinardi, Wikle, Berliner, Bonazzi) Process model considerations for a surface wind Bayesian hierarchical model. Poster, AGU Ocean Sciences Meeting, Portland, OR, February 2010.

(Milliff) Estimating semivariograms to build covariance matrices for J. Workshop on the ROMS 4D-Var Data Assimilation Systems for Advanced ROMS Users, University of California, Santa Cruz, July 2010.

(Wikle) A hierarchical approach to motivate spatio-temporal statistical models. Institute for Pure and Applied Mathematics (IPAM), UCLA, Los Angeles, CA, May 25, 2010.

(Wikle) Bayesian hierarchical models to augment the Mediterranean forecast system. Invited talk. Iowa State University. Ames, IA, October 15, 2009.

(Wikle) Don't forget the process! Using scientific process knowledge to motivate spatio-temporal models. Invited talk. SAMSI Program on Space-Time Analysis for Environmental Mapping, Epidemiology and Climate Change, Opening Workshop, RTP, North Carolina, September 14, 2009.

(Wikle) A class of nonlinear spatio-temporal dynamic models. Invited Talk, Joint Statistics Meetings, Washington, DC, August 4, 2009.

RESULTS

Time-Varying Error Covariance Models

Embedded scales in the error covariance estimations of ocean forecast systems act to rescale the error covariance magnitudes. This will impact the cost function estimation in the CCS implementations of ROMS 4DVAR. Anomaly data stage inputs are probably not sufficient to represent abrupt regime shifts in the ocean state. Experiments adding misfit data stage inputs and using mixture models will be useful in modelling error covariance response to ocean regime shifts in the CCS.

IMPACT/APPLICATIONS

The research overlapping the ONR project to use BHM to augment MFS, with the initial few months of the ONR model error project demonstrates practical methods to add time- and space-dependence to error covariance representations in operational (MFS) and near-operational (ROMS) ocean forecast systems. Refining estimates of the time-dependent changes in forecast uncertainty across regime shifts adds value to ocean forecast system output.

TRANSITIONS

Informal communications with scientists in the Ocean Modelling branch of the Naval Research Laboratory, Bay St. Louis, MI have carried over from the ONR MFS project.

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 2008 - October 2011.

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

REFERENCES

Chen, Z. and Dunson, D. B., 2003: "Random effects selection in linear mixed models", *Biometrics*, **59**, 762–769.

North, G.R., T.L. Bell, R.F. Cahalan and F.J. Moeng, 1982: "Sampling errors in the estimation of empirical orthogonal functions", *Mon. Wea. Rev.*, **110**, 699–706.

PUBLICATIONS

Wikle, C.K. and M.B. Hooten, 2010: A general science-based framework for spatio-temporal dynamical models. Invited discussion paper for TEST, Official Journal of the Spanish Society of Statistics and Operations Research, In press.