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

Ralph F. Milliff
Cooperative Institute for Research in Environmental Science
University of Colorado
UCB 216
Boudler, CO 80309-0216
phone: (303) 492-3013 fax: (303) 492-1149 email: milliff@colorado.edu

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: N00014-13-1-0622

LONG-TERM GOALS

The Bayesian Hierarchical Model (BHM) methodology is exploited to identify, characterize, and model the irreducible model error in ocean data assimilation and forecast systems.

OBJECTIVES

We describe 4 objectives addressed in the fiscal year October 2012 - September 2013.

First, we seek to extend the proof-of-concept results comparing a BHM surface wind ensemble with the increments in the surface momentum flux control vector in a four-dimensional variational (4dvar) assimilation system. The current objective is to convert BHM surface wind realizations to create an ensemble of surface stress vectors.

Second, continuing the effort to understand irreducible model error induced by representing the ocean state vector on a discrete grid, the current objective is to estimate the Hellinger distance between posterior distributions described in the next section.

Third, we have extended the hierarchical models for stochastic time-varying error-covariance matrices associated with data assimilation to include the case where both the observation and background error covariances are updated, yet dependent upon each other.

Fourth, we have extended the emulator-assisted data assimilation methodology by extending the parameterization of the spectral quadratic nonlinear spatio-temporal models to accommodate the inclusion of nonlinear interactions from small scales to inform the evolution of large scale modes.

APPROACH

Converting Surface Wind Realizations to a Surface Stress Ensemble: Sea-level pressure (*SLP*) and surface air and dew point temperature fields (T_a and T_d, respectively) for the Mediterranean Sea were obtained from collaborators (Prof. Nadia Pinardi) at Istituto Nazionale di Geofisica e Vulcanologia (INGV) in Bologna. These fields are used with ensemble winds from the BHM due to Milliff et al. (2011) to derive ensemble surface stress realizations as outlined in the flowchart shown in Figure 1.

Model Error Arising from a Discrete Grid: Let the parameter θ denote the ocean state (velocities and diffusion coefficients) in a deep layer in the South Atlantic, as shown in Figure 2. Define two posterior distributions as follows:

- $p(\theta|\tilde{Y})$ –which we call model $\tilde{M}$. Here the data $\tilde{Y}$ are tracer concentration measurements which have been *averaged to the nearest site on a regular spatial grid* (Fig 2); and
- $p(\theta|Y)$ – which we call model M. Here the data Y are tracer concentration measurements available at the original spatial locations (Fig 2).

The Hellinger distance between the two posterior distributions is defined as

$$H = \sqrt{\frac{1}{2}\int \left(\sqrt{p(\theta|\tilde{Y})} - \sqrt{p(\theta|Y)}\right)^2 d\theta}$$

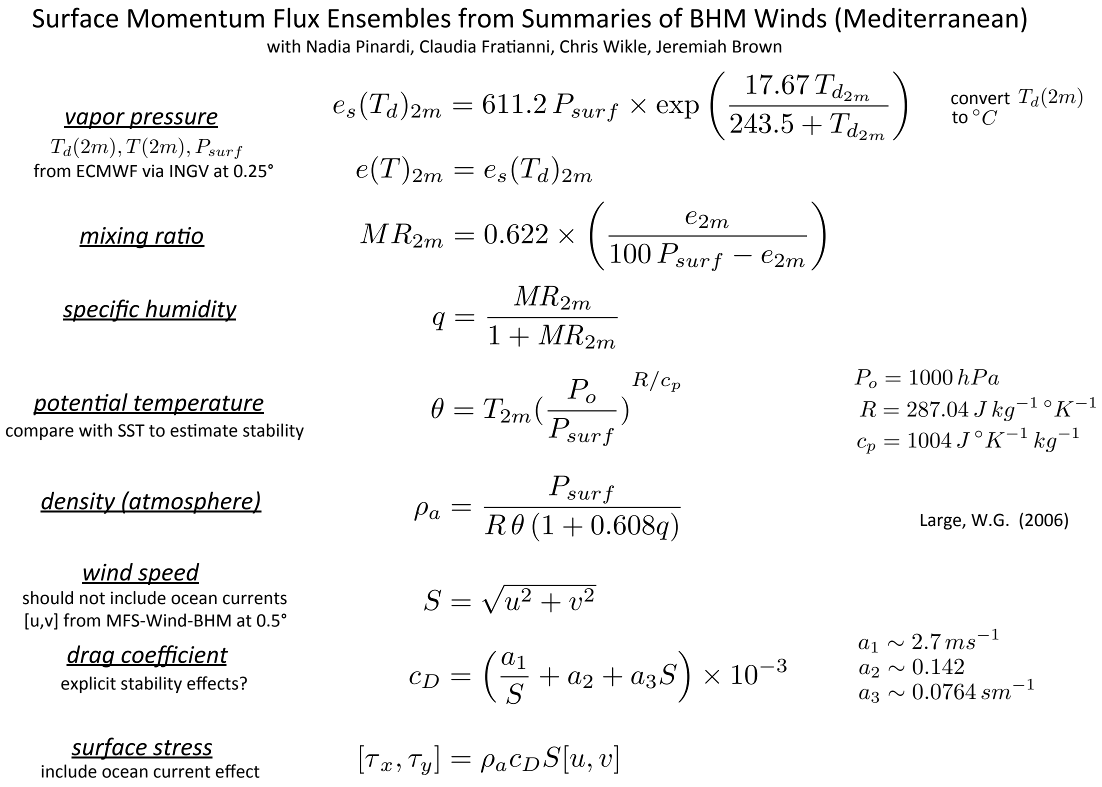
Obtaining an estimate of H gives us a measure of discrepancy between the two models $\tilde{M}$ and M.

We work under the assumption that model $\tilde{M}$ can be explored efficiently via Markov-Chain Monte Carlo (MCMC) methods and that model M can be explored via parallel computing. We derive a Monte Carlo estimate of H which requires the following components:

(i) A MCMC exploration of model $\tilde{M}$; Assume $\theta_1, \theta_2, \ldots \theta_B$ represents a Markov chain which explores the target $p(\theta|\tilde{Y})$.

(ii) An evaluation of the un-normalized posteriors $p(\theta_i|Y)$, $i = 1, \ldots, B$ which can be obtained efficiently via parallel computing;

(iii) An estimate of $p(\theta^*|\tilde{Y})$ and $p(\theta^*|Y)$ at *one single* user selected value θ^*.



Surface Momentum Flux Ensembles from Summaries of BHM Winds (Mediterranean)

with Nadia Pinardi, Claudia Fratianni, Chris Wikle, Jeremiah Brown

vapor pressure

$T_d(2m), T(2m), P_{surf}$

from ECMWF via INGV at 0.25°

$$e_s(T_d)_{2m} = 611.2\, P_{surf} \times \exp\left(\frac{17.67\, T_{d_{2m}}}{243.5 + T_{d_{2m}}}\right)$$

convert $T_d(2m)$ to $^\circ C$

$$e(T)_{2m} = e_s(T_d)_{2m}$$

mixing ratio

$$MR_{2m} = 0.622 \times \left(\frac{e_{2m}}{100\, P_{surf} - e_{2m}}\right)$$

specific humidity

$$q = \frac{MR_{2m}}{1 + MR_{2m}}$$

potential temperature

compare with SST to estimate stability

$$\theta = T_{2m}\left(\frac{P_o}{P_{surf}}\right)^{R/c_p}$$

$P_o = 1000\, hPa$

$R = 287.04\, J\, kg^{-1}\, {}^\circ K^{-1}$

$c_p = 1004\, J\, {}^\circ K^{-1}\, kg^{-1}$

density (atmosphere)

$$\rho_a = \frac{P_{surf}}{R\, \theta\, (1 + 0.608q)}$$

Large, W.G. (2006)

wind speed

should not include ocean currents

[u,v] from MFS-Wind-BHM at 0.5°

$$S = \sqrt{u^2 + v^2}$$

drag coefficient

explicit stability effects?

$$c_D = \left(\frac{a_1}{S} + a_2 + a_3 S\right) \times 10^{-3}$$

$a_1 \sim 2.7\, ms^{-1}$

$a_2 \sim 0.142$

$a_3 \sim 0.0764\, sm^{-1}$

surface stress

include ocean current effect

$$[\tau_x, \tau_y] = \rho_a c_D S[u, v]$$

Figure 1: ***Flow chart demonstrating path from input analysis fields to specific humidity and atmospheric density approximations. These are combined with estimates for drag coefficient and surface wind speed given ensemble winds from a Bayesian Hierarchical Model to provide surface momentum flux ensembles.***

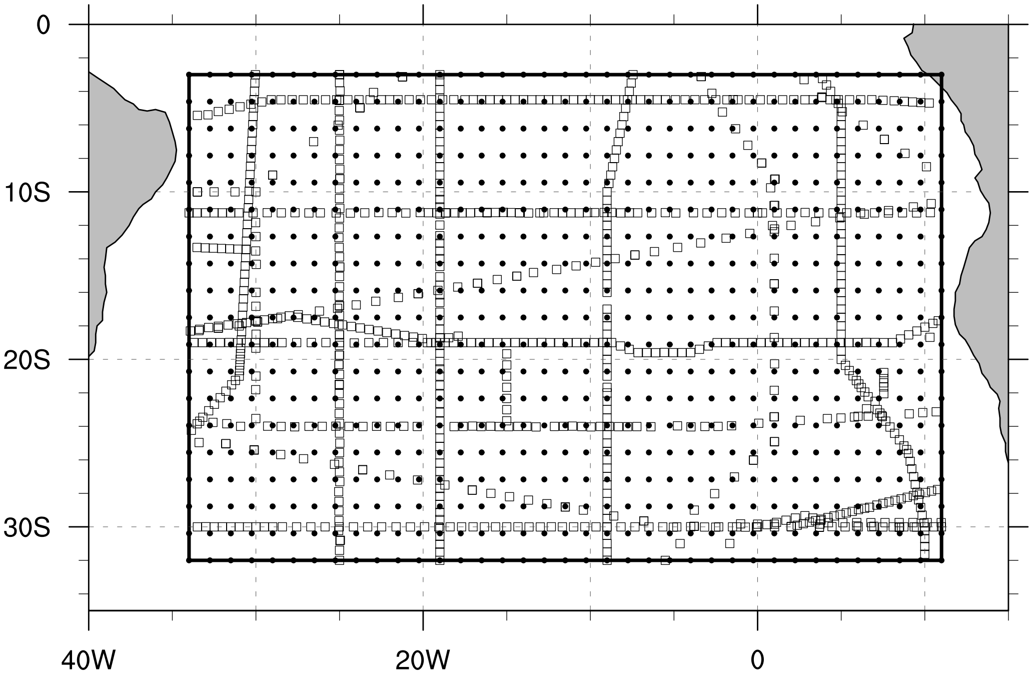


Figure 2: Domain of interest : squares indicate spatial locations where tracer concentration measurements are available; circles indicate a regular 19×37 ***spatial grid.***

<u>*Time-Varying Error Covariance Models:*</u> Extending the time-varying covariance methodology described in Dobricic et al. (2013), we consider modeling the simultaneous evolution of the background and observation error covariance matrices of a spatial field for use in data assimilation. In this context, there are two sources of data, one corresponding to historical model misfits (related to the background error covariance) and one associated with *in situ* observations, such as those from ARGO floats in the ocean. Critically, both covariance matrices are conditioned on a common reduced-dimensional covariance structure whose elements are allowed to evolve in a Markovian fashion through a Cholesky decomposition formulation. Although both the background and observation error covariance matrices are conditioned on this common time-varying low-rank matrix, the overall structure is somewhat different between them because the low-rank matrix gets transformed differently. This allows common structure but allows one to capture the scale variations between the two types of error processes.

<u>*Emulator Assisted Data Assimilation:*</u> We are interested in reduced-rank parametric emulators for non-linear spatio-temporal error processes. We allow the spatial process at time t (say, $\mathbf{Y}_t$) to be decomposed in terms of two basis expansions:

$$\mathbf{Y}_t = \Phi^{(1)}\alpha_t + \Phi^{(2)}\beta_t + \nu_t,$$

where $\Phi^{(i)}$, $i = 1,2$ correspond to $n \times p$ and $n \times q$ matrices containing large-scale and small-scale basis functions, respectively. Then, α_t and β_t correspond to vectors of large-scale and small-scale expansion coefficients of lengths p and q, respectively, and ν_t an error process assumed to be independent of α_t and β_t. Our primary interest is in the evolution of the large-scale coefficients given by α_t, as it is often the case in real-world processes wherein the important dynamics exist on a lower-dimensional manifold. Critical to our model development is allowing the propagation of α_t to $\alpha_{t+\tau}$ (where $\tau \geq 1$) to be influenced by the small-scale coefficients β_t but not allowing α_t to influence $\beta_{t+\tau}$ directly in the dynamical formulation. This provides a physically realistic way in which to reduce the parameter space in the rank-reduced general quadratic nonlinearity (GQN) formulation (e.g., Wikle and Hooten, 2010; Leeds et al. 2013). Specifically, we consider the following model for the evolution of α_t

$$\alpha_{t+\tau} = \mathbf{M}_\alpha \alpha_t + (\mathbf{I}_p \otimes \mathscr{G}(\alpha_t)')\mathbf{M}_{\alpha,Q}\alpha_t + \mathbf{M}_{\beta,L}\beta_t + (\mathbf{I}_p \otimes \mathscr{G}(\beta_t)')\mathbf{M}_{\beta,Q}\beta_t + \eta_{t+\tau}, \qquad (1)$$

for $t = 1, \ldots, T$ and some appropriate time increment τ, where $\eta_t \sim \text{Gau}(\mathbf{0}, \mathbf{Q}_\alpha)$, $\mathbf{M}_\alpha$ corresponds to the linear evolution of coefficients for the α_t process, $\mathbf{M}_{\alpha,Q}$ corresponds to the nonlinear evolution coefficients for the α_t process, $\mathbf{M}_{\beta,L}$ corresponds to the linear interactions between β_t and $\alpha_{t+\tau}$, $\mathbf{M}_{\beta,Q}$ corresponds to the nonlinear interactions between β_t and their impact on $\alpha_{t+\tau}$, and $\mathbf{Q}_\alpha$ is a $p \times p$ covariance matrix. Note that $\mathbf{M}_\alpha$ and $\mathbf{M}_{\beta,L}$ are $p \times p$ and $p \times q$ matrices while $\mathbf{M}_{\alpha,Q}$ and $\mathbf{M}_{\beta,Q}$ are $p^2 \times p$ and $pq \times p$ matrices. Though we may consider a variety of transformation functions $\mathscr{G}(\cdot)$, for the applications of interest in our DA examples, it is reasonable to specify this function be the identity and hence define $\mathscr{G}(\alpha_t) \equiv \alpha_t$ (similarly for β_t). A major challenge in the implementation of this methodology is the development of efficient sampling algorithms.

WORK COMPLETED

Converting Surface Wind Realizations to a Surface Stress Ensemble: Codes have been developed to follow the progression toward a surface momentum flux (e.g. surface stress vector) ensemble following Fig 1 using the analysis fields provided by INGV. Figure 3 depicts the intermediate output stages up to the surface density (ρ_a) calculation.

Following Large (2006) we approximate the surface drag coefficient as noted in Fig 1. The coefficients a_i; $i = 1, 2, 3$ can be treated as random variables to be estimated in the BHM. This enhancement is left for later work. For now, we use the ensemble winds from the BHM due to Milliff et al. (2011) and convert each surface vector wind realization to a surface momentum stress realization according to the flowchart in Fig 1.

Model Error Arising from a Discrete Grid: We have implemented and tested this approach on a series of examples where the Hellinger distance can be computed analytically. We are currently implementing this approach for a synthetic example in which the forward model is approximated via a first order emulator (see Hooten et al., 2011). We are also implementing this approach for the oceanographic tracer inversion problem described in Herbei and Berliner, (2012).

Time-Varying Error Covariance Models: The time-varying error covariance methodology for simultaneous estimation of background and observation error covariance matrices has been implemented in several simulated examples and with an application associated with data assimilation in the Mediterranean Sea (e.g., see Dobricic et al. 2013). This has been written-up as

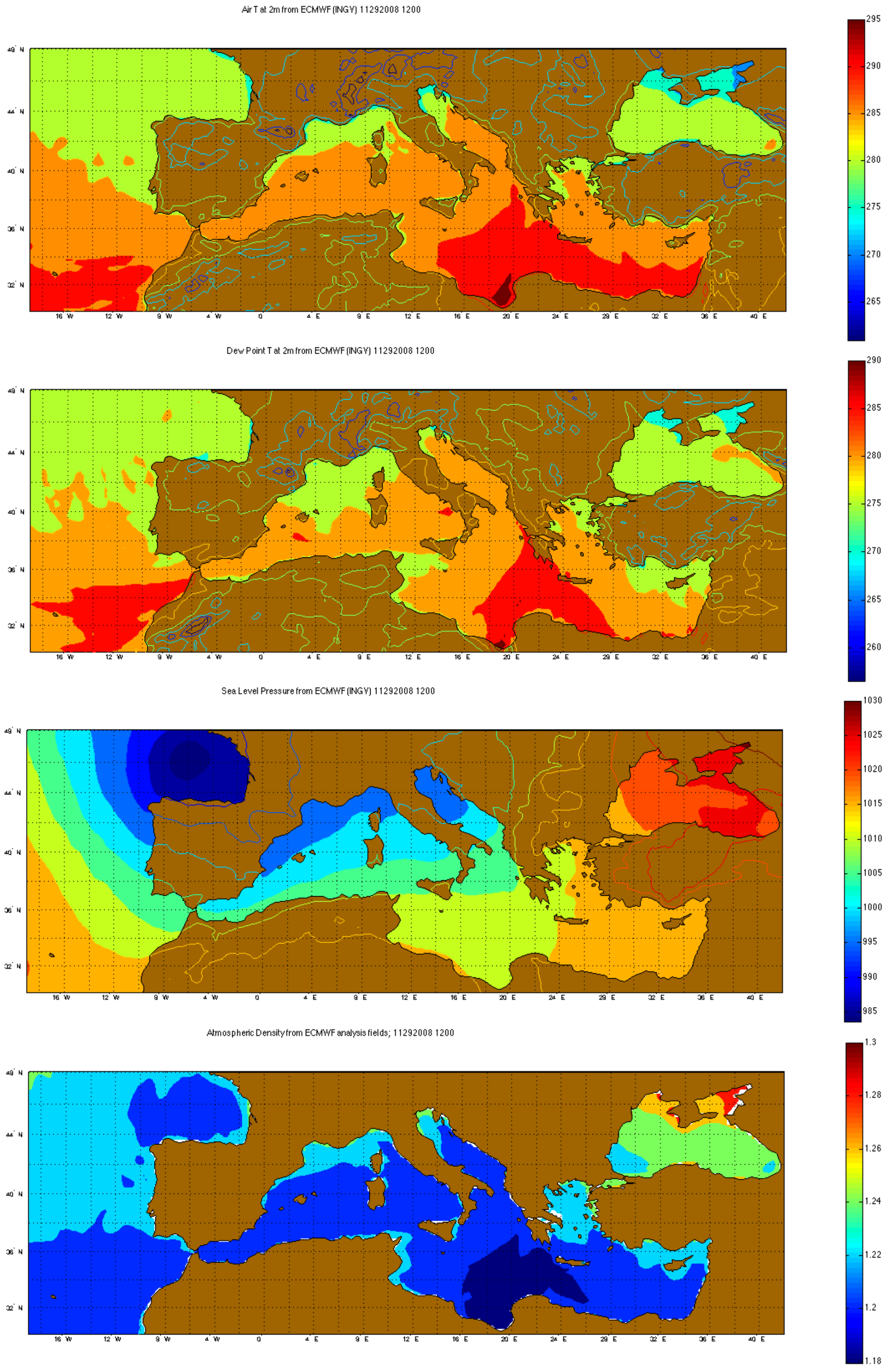

Air T at 2m from ECMWF (INGV) 11292008 1200
Dew Point T at 2m from ECMWF (INGV) 11292008 1200
Sea Level Pressure from ECMWF (INGV) 11292008 1200
Atmospheric Density from ECMWF analysis fields; 11292008 1200
48° N
44° N
40° N
36° N
32° N
16 W
12 W
8 W
4 W
0
4 E
8 E
12 E
16 E
20 E
24 E
28 E
32 E
36 E
40 E
295
290
285
280
275
270
265
290
285
280
275
270
265
260
1030
1025
1020
1015
1010
1005
1000
995
990
985
1.3
1.28
1.26
1.24
1.22
1.2
1.18

Figure 3: ***Input analysis fields from ECMWF via INGV for 29 November 2008 at 1200 UTC. The top 2 panels are atmospheric temperature and dew point temperature at $2m$, in $^{\circ}K$. The third panel is sea level pressure in hPa. These fields are combined to estimate atmospheric density (bottom panel) in $kg\,m^{-3}$.***

a chapter in the dissertation of Dan Gladish (U. Missouri) and is in the process of being turned into a paper for publication.

Emulator Assisted Data Assimilation: The development of the scale-interaction GQN model is outlined in a chapter in the dissertation of Dan Gladish (U. Missouri) and will be submitted in early October to a special issue of the journal *Environmetrics* that is focused on physical-statistical modeling in the environmental sciences.

Relevant Meetings and Presentations:

(Wikle) *Invited;*, Hierarchical general quadratic nonlinear models for spatio-temporal dynamics. Red Raider Conference, Texas Tech University, Lubbock, TX, October 2012.
(Wikle) *Invited;* Efficient time-frequency representations in high-dimensional spatial and spatio-temporal models. Invited Talk, ASA ENVR Workshop on Environmetrics, North Carolina State University, October 2012.
(Herbei, Berliner) *Poster;* Estimating ocean-circulation: a likelihood-free approach via a Bernoulli factory. Institute for Mathematics and its Applications, Workshop on Stochastic Modeling of the Ocean and Atmosphere; U. Minnesota, March 2013.
(Milliff) *Invited;* Uncertainty in Ensemble Ocean Forecasts; Deducing Ocean Model Error with Ensemble Winds from a Bayesian Hierarchical Model, Institute for Mathematics and its Applications, Workshop on Stochastic Modeling of the Ocean and Atmosphere; U. Minnesota, March 2013.
(Wikle) *Invited;* Using quadratic nonlinear statistical emulators to facilitate ocean biogeochemical data assimilation, Institute for Mathematics and its Applications, Workshop on Stochastic Modeling of the Ocean and Atmosphere; U. Minnesota, March 2013.
(Wikle) *Invited;* Statistics and the environment: Overview and challenges. 41st Annual Meeting of the Statistical Society of Canada, Edmonton, Alberta, Canada; May 2013.
(Wikle) *Invited;* Nonlinear Dynamic Spatio-Temporal Statistical Models. Southern Regional Council on Statistics Summer Research Conference; June 2013.
(Milliff) *Invited;* Bayesian Hierarchical Model Applications in Ocean Forecasting, Society for Industrial and Applied Mathematics, Annual Meeting Minisymposium on Uncertainty Quantification in Climate Modeling and Prediction; San Diego, CA, July 2013.
(Milliff) *Invited;* A Tale of Two Bayesian Hierarchical Models, IMAGe 2013 Climate Analytics Theme-of-the-Year Workshop; Next Generation Climate Data Products; July 2013.
(Herbei) Exact MCMC Using Approximations, Joint Statistical Meetings, Montreal, CANADA; August, 2013.
(Wikle) *Invited Keynote Lecture;*, Nonlinear dynamic spatio-temporal statistical models. Third Workshop on Bayesian Inference and Latent Gaussian Models with Applications, Reykjavik, Iceland, September 2013.

RESULTS

Converting Surface Wind Realizations to a Surface Stress Ensemble: Sample surface momentum flux ensembles are shown for a snapshot in the western Mediterranean Sea in Figure 4. As noted

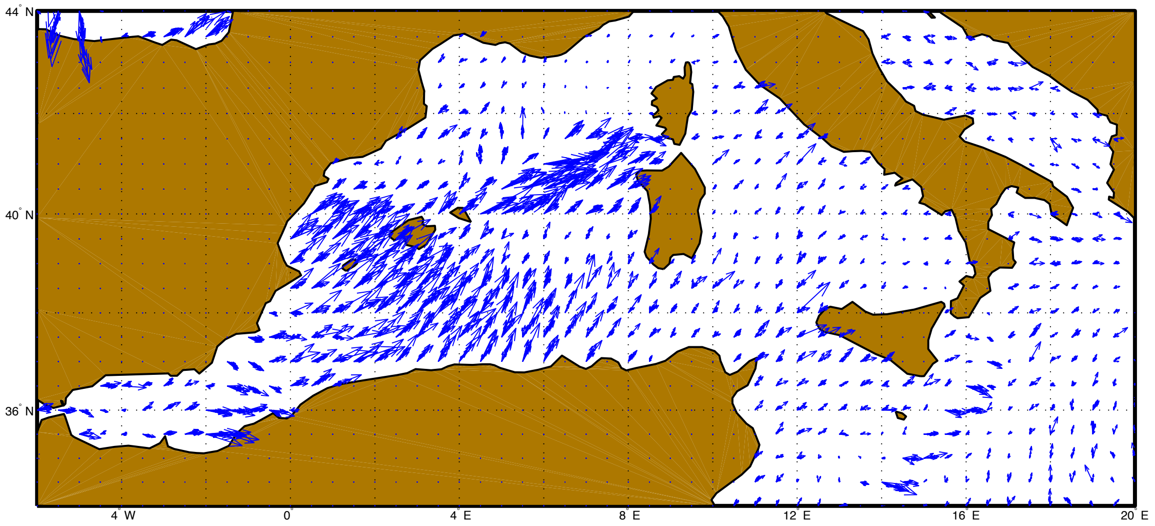


Figure 4:** **Sample surface momentum flux vectors at 1200 UTC on 29 Nov 2008. Ensemble $\vec{\tau}$ are obtained as summaries of ensemble winds from a BHM (Milliff et al., 2011), given input surface and $2m$ analysis fields.

in the progress report from last year, ensemble surface wind stress estimates from the BHM can be used to diagnose loci of potential model error in 4dvar ocean data assimilation systems. If the iterations in the strong-constraint 4dvar between forward and adjoint models shifts the surface flux control vector outside the probabilistic ensemble estimated in the BHM, there is the possibility that the control vector variable is being used to correct for model error rather than forcing error. This was demonstrated for the California Current System ROMS 4dvar forecast system due to Prof. Andrew M. Moore and colleagues.

Our next focus will be on similar BHM ensemble developments for the other variables comprising the forcing part of the control vector in ROMS 4dvar; e.g. surface heat and fresh-water fluxes.

Model Error Arising from a Discrete Grid: Our initial tests show that the estimate we propose worked well in all cases. We are currently running simulations for the oceanographic tracer concentration inversion.

Time-Varying Error Covariance Models: We have preliminary results for an application in the Mediterranean Sea, which uses the data described in Dobricic et al. (2013). This is currently in a draft chapter of Dan Gladish's dissertation (U. Missouri), the final version of which will be submitted in November, 2013. At that time, we expect to finish converting the chapter to a paper for submission.

Emulator Assisted Data Assimilation: We have preliminary results for an application to long-lead sea surface temperature prediction in the tropical Pacific ocean, as well as 6-24 hour forecasts of SLP over the midwest USA. These currently reside in a draft chapter of Dan Gladish's dissertation (U. Missouri), the final version of which will be submitted in November,

2013. We are in the process of converting this chapter to paper for submission to a special issue of *Environmetrics* in early October 2013.

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, 31 July - 2 August 2013, focused on Irreducible Model Error issues.

RELATED PROJECTS

"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).

"Bayesian Hierarchical Climate Prediction", NSF, April 2011 - March 2014, C.K. Wikle and L.M. Berliner (PIs)

REFERENCES

Dobricic, S., C.K. Wikle, R.F. Milliff, N. Pinardi, and L.M. Berliner, 2013: Assimilation of oceanographic observations with estimates of vertical background error covariances by a Bayesian hierarchical model, submitted.

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

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., 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.

Large, W.G., 2006: "Surface fluxes for practitioners of global ocean data assimilation", Chapt. 9 in **Ocean Weather Forecasting**, E.P. Chassignet and J. Verron (eds.), Springer, pgs. 229-270.

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.

PUBLICATIONS

Dobricic, S., C.K. Wikle, R.F. Milliff, N. Pinardi, and L.M. Berliner, 2013: "Assimilation of oceanographic observations with estimates of vertical background error covariances by a Bayesian hierarchical model", submitted.

Gladish, D.W., C.K. Wikle and S.H. Holan, 2013: "Covariate-based cepstral parameterizations for time-varying spatial error covariances", *Environmetrics*, accepted pending minor revision.

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*, in revision.

Leeds, W.B., C.K. Wikle and J. Fiechter, 2013: "Emulator-assisted reduced-rank ecological data assimilation for multivariate dynamical spatio-temporal processes", in press, doi:10.1016/j.statmet.2012.11.004.

White, S. and R Herbei, 2013: "Quantifying model error in posterior distribution – a Monte Carlo approach" in preparation.

Wikle, C.K., R.F. Milliff, R. Herbei and W.B. Leeds, 2013: "Modern Statistical Methods in Oceanography: A Hierarchical Perspective", *Statistical Science*, in press.

HONORS/AWARDS/PRIZES

N. Cressie and C.K. Wikle, PROSE Award, For excellence in the Mathematics Category from the Association of American Publishers, for the 2011 Wiley book, *Statistics for Spatio-Temporal Data*.