

# A Review of Physical-Statistical Bayesian Hierarchical Modeling of Air-Sea Interaction

Suz Tolwinski

April 16, 2009

## Abstract

Bayesian hierarchical models represent an untapped method for combining scientific information and observational data for the reconstruction of geophysical fields. Here I review an example of such a model presented by Berliner et al. in “Bayesian hierarchical modeling of air-sea interaction” [1]. This article provides an informative example of how traditional Bayesian hierarchical modeling can be extended to include scientific first-principles information, and demonstrates the potential of this “physical-statistical” approach as a reconstructive tool. In addition, the shortcomings of the results highlight the fact that model parsimony and computational efficiency both represent trade-offs for an optimal estimation of ones target.

## 1 Introduction

Use of Bayesian statistics combined with first-principles modeling is not new in the geosciences. In fact, developments in *data assimilation* (DA)– the set of techniques for combining observed data with scientific prior knowledge to estimate a given state space– have historically been driven by applications in weather forecasting and hydrology. In this context, one generally combines a statistical *data model* for the distribution of observations  $Y$  given the true process  $X$ , and a physics-based *prior model* of the scientific information one has about the process of interest  $X$ . Most DA techniques for merging data with the information on the process exploit Bayes’ Law in one way or another, which inverts “forward” models to give the desired “inverse” *posterior distribution*:

$$p(x|y) = \frac{p(y|x)p(x)}{\int_{\mathcal{X}} p(y|x)p(x)dx} \quad (\text{Bayes' Law}) \quad (1)$$
$$\text{posterior distribution} = \frac{\text{data model} \cdot \text{prior model}}{\text{marginal distribution}}$$

Many DA approaches are based on a two-step algorithm, where first the prior model is used to provide an estimate of the process of interest, and then the data model “updates” the estimate via Bayes’ Law.

Bayesian hierarchical models constitute a form of data assimilation in which the prior model can be expanded in a chain of dependences between the system variables. This approach allows for an emphasis on forward modeling, as well as considerable freedom in the way this modeling is carried out. In their 2003 paper “Bayesian Hierarchical Modeling of Air-Sea Interaction,” Berliner, Milliff and Cressie demonstrate how an understanding of the physics of atmosphere-ocean interaction can be incorporated into the machinery of a BHM. Their experiment in reconstructing a known oceanic model target provides a benchmark for estimating the

past oceanic states from sparse data, and showcases the general potential of the BHM methodology for geoscientific reconstructions.

## 2 Bayesian Basics

In contrast to the frequentist point of view, in which an unknown parameter  $\theta$  is believed to have one fixed, true value, the Bayesian paradigm treats  $\theta$  as a random variables. The field of Bayesian statistics deals with how to estimate the parameter  $\theta$  given data  $x$  from this point of view.

### 2.1 Bayesian Notions of Risk, Loss, and Estimation

Suppose one would like to estimate the value of a parameter  $\theta$  given some related data  $x$ . Bayesian inference on  $\theta$  begins with the specification of three structures:

1. A parametric statistical model  $f(x|\theta)$  that describes how the data depends on the parameter of interest;
2. A prior distribution  $\pi(\theta)$  that represents any a priori information the modeler has on the parameter  $\theta$ . Together, the parametric statistical model and prior distribution comprise a Bayesian statistical model;
3. A loss function  $L(\theta, \delta)$ , representing the loss (or error) owing to an estimate of  $\theta$  by  $\delta$ .

Given these quantities, the field of Decision Theory deals with how to estimate the quantity  $\theta$  of interest given a set of data  $x$ . From a frequentist perspective, the quantity one would like to minimize with one's decision (estimator)  $\delta(x)$  is the frequentist risk  $R(\theta, \delta)$ , defined as

$$R(\theta, \delta) = E_{\theta} [L(\theta, \delta(x))] = \int_X L(\theta, \delta(x))f(x|\theta)dx \quad (2)$$

Note that the frequentist risk is defined for a fixed value of  $\theta$  (the value that corresponds to the truth, according to the frequentist viewpoint).

In the Bayesian point of view, one is instead concerned with minimizing the posterior expected loss  $\rho(\pi, \delta|x)$ , defined as

$$\rho(\pi, \delta|x) = E_{\pi} [L(\theta, \delta|x)] = \int_{\Theta} L(\theta, \delta)\pi(\theta|x)d\theta \quad (3)$$

Note that the posterior expected loss is an inherently Bayesian quantity; the loss is integrated over all possible values in the range of the random variable  $\theta$  weighted by its posterior distribution.

It turns out that choosing ones estimator  $\delta$  to minimize the posterior expected loss is equivalent to minimizing the integrated (Bayes) risk  $r(\pi, \delta)$ , a quantity given by

$$r(\pi, \delta) = E_{\pi} [R(\theta, \delta)] = \int_{\Theta} \int_X L(\theta, \delta(x))f(x|\theta)dx\pi(\theta)d\theta \quad (4)$$

The proof that the integrated Bayes risk and the posterior expected loss are both minimized for the same estimator  $\delta$  reduces simply to an application of Fubini's theorem.

The estimator  $\delta^{\pi}$  that minimizes the Bayes risk  $r(\pi, \delta)$  for a particular prior distribution  $\pi$  and loss function  $L$  is called a Bayes' estimator for  $\theta$ . In application, the quadratic loss function  $L(\theta, \delta) = (\theta - \delta)^2$  is most frequently used because of its simplicity, and because it provides a second-order Taylor expansion approximation to any more complicated symmetric convex choices. (Indeed, quadratic loss is implicitly assumed in the article reviewed in this paper.)

Under quadratic loss, it is easy to show that the Bayes estimator  $\delta^\pi$  is the posterior mean. In this case, the posterior expected loss can be written

$$\begin{aligned}\rho(\pi, \delta|x) &= \mathbb{E}_\pi [L(\theta, \delta|x)] \\ &= \mathbb{E}_\pi [(\theta - \delta)^2|x] \\ &= \mathbb{E}_\pi[\theta^2|x] - 2\delta\mathbb{E}_\pi[\theta|x] + \mathbb{E}_\pi[\delta^2|x] \\ &= \mathbb{E}_\pi[\theta^2|x] - 2\delta\mathbb{E}_\pi[\theta|x] + \delta^2(x)\end{aligned}$$

which is clearly minimized for the choice  $\delta^\pi = \mathbb{E}_\pi[\theta|x]$ . Thus, because the squared-error loss is the most common loss used in scientific applications, the posterior mean is the most common Bayes estimator in these settings.

## 2.2 Bayesian Hierarchical Modeling

Bayesian hierarchical modeling is a DA approach with the ability to model arbitrarily many variables involved in the process of interest in a *hierarchy* of dependencies. Formally, Robert [2] defines a Bayesian hierarchical model as

...a Bayesian statistical model  $(f(x|\theta), \pi(\theta))$ , where the prior distribution  $\pi(\theta)$  is decomposed in conditional distributions

$$\pi_1(\theta|\theta_1), \pi_2(\theta_1|\theta_2), \dots, \pi_n(\theta_{n-1}|\theta_n) \quad (5)$$

and a marginal distribution  $\pi_{n+1}(\theta_n)$  such that

$$\pi(\theta) = \int_{\Theta_1 \times \Theta_2 \times \dots \times \Theta_n} \pi_1(\theta|\theta_1)\pi_2(\theta_1|\theta_2) \dots \pi_{n+1}(\theta_n)d\theta_1 d\theta_2 \dots d\theta_n \quad (6)$$

The parameters  $\theta_i$  are called hyperparameters of level  $i$  ( $1 \leq i \leq n$ ).

In settings with many complex interrelationships between state variables, this type of model is ideal for separating those relationships into simpler levels of conditional dependence. With continuing advances in computational speed, the geosciences community is beginning to experiment with Bayesian hierarchical models (BHMs). Haslett and colleagues condition a climate reconstruction at Glendalough, Scotland on pollen data using Bayesian hierarchies [3], and Korhola et al. pursue a similar approach to reconstruct temperature in northern Fennoscandia given chironomid data from the region [4]. Tebaldi and Sanso use a Bayesian hierarchical technique to formally evaluate the uncertainty associated with climate projections from global climate models (GCMs) [5], and Hegerl et al. use the methods to estimate the true sensitivity of the earth system to a doubling of CO<sub>2</sub> [6]. All of these models exploit Bayes' Law to relate a *data model*, a *process model*, and a *parameter model* on all the model parameters:

$$\begin{aligned}[\text{process}|\text{data}] &\propto \int_{\Theta_1 \times \Theta_2} [\text{data}|\text{process}, \vec{\theta}_1] \cdot [\text{process}|\vec{\theta}_2] \cdot [\vec{\theta}_1, \vec{\theta}_2] d\vec{\theta}_1 d\vec{\theta}_2 \quad (7) \\ \text{posterior distribution} &\propto \text{data model} \cdot \text{process model} \cdot \text{parameter model}\end{aligned}$$

The Bayesian treatment of the parameters as random variables allows for modeling of their uncertainty in the prior distribution. Although traditional BHMs use parametric statistical models at all levels, the modularity of this approach allows for the use of first-principles scientific information in the development of the process level model. Scientific information on how the data is derived or formed given the process of interest may be used in the data model. Although use of mechanistic knowledge at the data level is found only very rarely in the geosciences literature, no additional mathematics is needed for the theoretical description of a BHM with this added information. In circumstances where the data is not a direct measurement of the field of interest, but is derived through a chemical, biological, or physical process depending on that field, mechanistic modeling at the data level is a natural extension of these methods.

### 3 Observing System Simulation Experiment

#### 3.1 Setting

To test the potential of a BHM for capturing characteristics of oceanic flow given realistically sparse data, Berliner et al. perform an observing system simulation experiment (OSSE). In such an experiment, “data” is simulated by sampling a known dynamical model “truth” and adding noise in a manner consistent with real data sampling. The experimenter then carries out the reconstruction methodology, and can evaluate its performance against the target (eg. the model “truth”).

The simulated reality in this study is provided by a numerical, primitive-equation (PE) shallow-water-equation (SWE) gridded model of the ocean by Milliff and McWilliams [7]. Specifically, Berliner et al. are interested in reconstructing how the model ocean responds to an energetic, transient atmospheric cyclone. The cyclone takes six simulation days to cross the domain, which is given a size, geographic location, and initial conditions that mimic conditions of the Labrador Sea. Paleoclimate records show that the meridional overturning circulation is sensitive to changes in the oceanic flow in this region of the North Atlantic. Because this circulation is crucial to the redistribution of heat from the equator to the poles, the response of this region of the ocean to energetic atmospheric events is important to understand from a climatological perspective.

#### 3.2 Physics-Based Process Models

The first-principles based modeling in this work occurs at the process level in the Bayesian hierarchy. Berliner et al. develop a stochastic model for the oceanic streamfunction and atmospheric winds based on simplifications of the laws of physics governing the system.

For this synoptic-scale ocean basin, a quasi-geostrophic approximation of the oceanic flow is generally valid. That is, the Coriolis (apparent) force is taken to balance the pressure gradient force in the horizontal momentum equations, and vertical advection of momentum is neglected. A  $\beta$ -plane approximation is invoked wherein the meridional variation of the Coriolis parameter is approximated to first order by  $f = f_0 + \beta y$ . In geophysical contexts, the zonal component of flow (flow parallel to lines of constant latitude) is denoted by  $u$ , and the meridional component (parallel to lines of constant longitude) is denoted by  $v$ . Since purely geostrophic flow is non-divergent, it becomes useful to define the *streamfunction*  $\Psi$  such that  $\frac{\partial \Psi}{\partial x} = v$  and  $-\frac{\partial \Psi}{\partial y} = u$ . The flow field can therefore be completely recovered from the scalar streamfunction field. Another important quantity is the vertical component of the fluid’s relative vorticity,  $\xi$ , defined by  $\xi = \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}$ . The stream function and vorticity can thus be related by  $\nabla^2 \Psi = \xi$ .

The quasi-geostrophic vorticity equation may be derived from the horizontal momentum equations and the continuity equation (see, eg., [8]), and is written (in terms of the stream function) as:

$$\left(\nabla^2 - \frac{1}{r^2}\right) \frac{\partial \Psi}{\partial t} = -J(\Psi, \nabla^2 \Psi) - \beta \frac{\partial \Psi}{\partial x} + \frac{1}{\rho H} \text{curl}_z \tau - \gamma \nabla^2 \Psi + a_h \nabla^4 \Psi \quad (8)$$

Here  $r$  is the Rossby radius of deformation, the length scale at which earth’s rotational effects become important. The Jacobian in the first term on the right hand side can be expanded to show that the term represents advection of vorticity. The second term on the right hand side results from variation in the Coriolis force with latitude. The quantity  $\tau$  in the third term represents the wind stress (so  $\text{curl}_z(\tau) = \frac{\partial \langle u'w' \rangle}{\partial y} - \frac{\partial \langle v'w' \rangle}{\partial x}$ , where  $\langle \cdot \rangle$  denotes time average).  $\rho$  in this term is the fluid density, and  $H$  is the fluid depth. The fourth term models the damping effects of bottom friction, and the last term models dissipation of vorticity by eddy processes.

### 3.2.1 Oceanic process model and priors

A discretized approximation to the quasi-geostrophic equation (8) for the evolution of the oceanic streamfunction on the interior gridpoints over one time step is given by

$$\Psi_I^{t+1} = \left( I + \Delta \tilde{G} (-\beta D_x - \gamma G + a_h G^2) \right) \Psi_I^t + \Delta \tilde{G} \left( -J(\Psi_I^t) + \frac{1}{\rho H} C(U^t, V^t) \right) + B \Psi_B^{t+1} \quad (9)$$

Here,  $\tilde{G} = (G - r^{-2}I)^{-1}$ , where  $G$  is a discrete Laplacian operator,  $\Delta$  is the discrete time step,  $D_x$  denotes the zonal derivative,  $J$  the discrete Jacobian,  $C(U^t, V^t)$  stands for the discretized vertical component of the wind curl stress, and  $B \Psi_B^{t+1}$  represents the effects of the boundary flow. To write down an evolutive stochastic process model for the oceanic streamfunction, the authors note that in this simple QG model,  $\Psi_I^{t+1}$  has simple dependence on various operations on  $\Psi_I^t$ , an operation on the wind-stress, and on the boundary values  $\Psi_B^{t+1}$ . The model is composed of a random parameter times each of these terms plus some noise:

$$\begin{aligned} \Psi_I^{t+1} &= \left( l_1 I + l_2 \tilde{G} D_x + l_3 \tilde{G} G^2 \right) \Psi^t + j \tilde{G} J(\Psi_I^t) + c \tilde{G} C(U^t, V^t) + b B \Psi_B^{t+1} + e_{t+1} \quad (10) \\ &\text{where } e_{t+1} \sim N(0, \Sigma_e) \end{aligned}$$

The parameters  $\vec{l} = (l_1, l_2, l_3), j, c, b$  and  $\Sigma_e$  are priors of the oceanic process model, and as such, will be modeled as random variables. Prior information on many of these parameters can be derived from direct comparison of the oceanic process model (8) with the discretized QG model (10). For example,  $l_2$  is the coefficient on contributions from  $\tilde{G} D_x \Psi^t$ , and so one would expect its value to be around  $\beta$  times the time step  $\Delta$ . The authors choose a noninformative uniform distribution centered on this value, with upper and lower boundaries deviating by 4% of this value. Parameters  $l_1, l_3, j, b$ , and  $c$  are modeled following similar intuition.

The other priors are modeled using empirical information. For example, the authors mention they have a rough sense of the variability corresponding to errors between a QG approximation and the PE-SWE ‘‘truth,’’ as well as a sense of the spatial correlation of the errors. The covariance structure of the error is then modeled as  $\Sigma_e = \sigma_e^2 R(\theta)$ . They set  $\sigma_e^2 \sim U(1.1 \times 10^6, 1.6 \times 10^6)$ , representing their prior beliefs about the size of the errors, and  $R(\theta)$  approximates the discretized exponential correlation  $\Sigma_{ij} = \exp(-\theta d_{ij})$ , where  $d_{ij}$  is the distance between gridpoints  $i$  and  $j$ .

Finally, the streamfunction is assumed to be a Markov process; that is, the dependence of the  $t+1$ st value of this field on past values of all other fields only through their  $t$ th value. Given the wind fields  $U$  and  $V$ , and denoting all of the parameters of the model above by  $\eta_\psi$ , the Markov property allows the conditional distribution of  $\Psi$  to be written

$$[\Psi|U, V, \eta_\psi] = [\Psi^1|\eta_\psi] \prod_{t=1}^{T-1} [\Psi^{t+1}|\Psi^t, U^t, V^t, \eta_\psi] \quad (11)$$

### 3.2.2 Atmospheric process model and priors

The atmospheric process model developed by Berliner et al. does not contain explicit dynamical time dependence, but is based on an assumption of approximate geostrophy (eg. that the Coriolis apparent force on the atmospheric flow is balanced approximately by the pressure gradient force). For zonal and meridional wind components  $U(t)$  and  $V(t)$ , geostrophic balance can be written

$$\begin{aligned} f_0 U(t) &= -1/\rho_a D_y P(t) \\ f_0 V(t) &= 1/\rho_a D_x P(t) \end{aligned}$$

Berliner et al. assume normally distributed deviations from geostrophic balance, so that the model for the zonal and meridional components of the wind is written as

$$f_0 U^t | P^t, \sigma_{U|P}^2 \sim N\left(-1/\rho_a D_y P^t, \sigma_{U|P}^2 I\right) \quad (12)$$

$$f_0 V^t | P^t, \sigma_{V|P}^2 \sim N\left(1/\rho_a D_x P^t, \sigma_{V|P}^2 I\right) \quad (13)$$

Since there are no observations of atmospheric pressure, but simulations of the field  $P$  are clearly necessary for this specification of the wind, pressure is referred to as a “hidden process.” The pressure field is simulated by  $P^t \sim N(\mu_p, \Sigma_p)$ .

Strictly speaking, the priors of the atmospheric process model should be  $\sigma_{U|P}^2, \sigma_{V|P}^2$ , the mean pressure field  $\mu_p$  and its variance,  $\Sigma_p$ . The authors mention the conjugate inverse gamma distribution would be a feasible choice for  $\sigma_{U|P}^2$  and  $\sigma_{V|P}^2$ . However, to simplify the experiment, they choose to fix  $\sigma_{U|P}^2 = \sigma_{V|P}^2 = 3\text{m}^2/\text{s}^2$  (note that this means the hierarchical model is not fully Bayesian). To model the pressure field, Berliner et al. mention an observational study of the Labrador Sea that provides an estimate of the mean pressure field  $\mu_P$ , and suggests a simple covariance structure given by  $c(d) = \sigma_p^2 \exp(-\theta_p d)$ , with  $d$  being the distance between two points. They reparameterize the process by  $P^t = \mu_p I + E(\theta_p) \alpha^t$ , where  $E(\theta_p)$  are the first few EOFs of the covariance matrix, and  $\alpha^t \sim N(0, \Lambda)$  are the variances (amplitudes) of the EOFs in time. Here  $\Lambda$  is a diagonal matrix of i.i.d. variances with a conjugate inverse gamma distribution. The parameters of this distribution are set from an EOF decomposition of the empirical covariance pressure matrix across the area. The approach of setting model hyperparameters according to empirical data is referred to as an *empirical Bayes* approach.

### 3.3 Data modeling and priors

Berliner, Milliff and Wikle simulate two kinds of data in their OSSE. The first is scatterometer data, which provides a measurement of wind fields. Scatterometers send radio frequency pulses toward the ocean from a satellite, and the near-surface wind velocities are inferred from the energy of the reflected pulse. The simulated data cover a wide swath of the testbed domain at 0 and 12 hours each simulation day, and provide a measurement at each gridpoint in the swath. This simulated data is comparable to what can be obtained from NASA’s QuickSCAT scatterometer, which has a 1,800 km swath covering roughly 90% of earth’s oceans each day with a 25-km wind vector resolution [9].

Secondly, simulated altimeter data of the ocean pressure field is used in the OSSE. The altimetry used in ocean studies is also satellite-based, and uses microwaves to measure surface topography. The simulated altimeter data crosses the domain on its orbital track at 0 and 6 hours each simulation day, with an along-track resolution of 12 km. The resolution used is about half of that of real modern-day altimeters, but Berliner et al. use twice as many data tracks as can normally be obtained in real systems.

#### 3.3.1 Scatterometer Data Model

The data level model for how scatterometer data should arise from the “real” wind fields is especially simple. The data is given by an incidence matrix  $K_w^t$  on the wind field (which returns the wind field value at each place covered by the data swath at time  $t$ , but zero elsewhere) plus uncorrelated gaussian noise:

$$\begin{aligned} D_U^t &= K_w^t U^t + \epsilon_w^t \\ D_V^t &= K_w^t V^t + \epsilon_w^t \\ &\text{where } \epsilon_w^t \sim N(0, \sigma_{w,\epsilon}^2 I) \end{aligned} \quad (14)$$

The variance  $\sigma_{w,\epsilon}^2$  is fixed at a value determined through studies of the accuracy of scatterometer data.

### 3.3.2 Altimeter Data Model

The specification of the altimeter data model is similarly simple. Note that real-world altimeter data provides a reading of the pressure field, but the model truth in the OSSE simulates the streamfunction rather than surface pressure. The authors invoke the geostrophic assumption and model measurements of the perturbation pressure field data as  $f_0$  times the streamfunction plus noise:

$$D_{p'}^t = K_0^t \Psi^t f_0 + \epsilon_0^t \quad (15)$$

where  $\epsilon_0^t \sim N(0, \sigma_{o,\epsilon}^t I)$

Again,  $K_0^t$  is an incidence matrix mapping to precise location of the altimeter tracks on the domain at time  $t$ . The variance of the noise process is again assumed to be known.

## 4 Numerics

Unless one uses only conjugate priors, the posterior distribution of a BHM generally can not be computed analytically. Even when it can be easily evaluated, sampling from a high-dimensional probability distribution is nontrivial— the naive approach of uniform sampling scales exponentially with the number of dimensions. Numerical methods for sampling probability distributions and evaluating integrals therefore go hand in hand with Bayesian hierarchical modeling, where one is interested in a chain of dependences (and therefore a model of high dimension).

In ‘Bayesian hierarchical modeling of air-sea interaction,’ two such Monte Carlo methods are used: importance sampling, and a Monte Carlo Markov Chain Gibbs sampler. These are presented here after the expositions in [10] and [2].

### 4.1 Importance Sampling

Importance sampling is a technique for estimating moments or expectations of functions of a random variable, rather than for generating a numerical probability distribution. The method can be viewed as a generalization of the naive approach of evaluating the desired function at uniformly distributed points and taking a sum

Suppose one has a Bayesian model  $(f(x|\theta), \pi(\theta))$ , and wants to evaluate

$$\mathbb{E}[g(\theta|x)] = \frac{1}{M} \int_{\theta} g(\theta) f(x|\theta) \pi(\theta) d\theta \quad (16)$$

In importance sample, it is assumed that the distribution according to which one would like to integrate is known to a multiplicative constant. This is the case in most Bayesian analyses, where the posterior  $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$ , but the normalizing constant  $M$ , given by a marginal distribution, is an integral that cannot be evaluated exactly. Suppose that  $h(\theta)$  is a simpler distribution which is also known to within a multiplicative constant, eg.  $h(\theta) = h^*(\theta)/M_h$ , where  $h^*(\theta)$  can be sampled directly.

Now, the right hand side of (16) is proportional to

$$\int_{\theta} \frac{g(\theta) f(x|\theta) \pi(\theta)}{h(\theta)} h(\theta) d\theta = \int_{\theta} \frac{g(\theta) f(x|\theta) \pi(\theta)}{h^*(\theta)} h^*(\theta) d\theta \quad (17)$$

Importance sampling proceeds by generating  $N$  samples  $\{\theta_n\}_{n=1}^N$  from  $h^*(\theta)$ , and evaluating  $g(\theta_n)$  at each point. To account for the fact that these constitute samples from the distribution  $h^*(\theta)$ , instead of  $f(x|\theta)\pi(\theta)$  as desired, one defines weights  $w_n$  by

$$w_n \equiv \frac{f(x|\theta)\pi(\theta)}{h^*(\theta_n)} \quad (18)$$

Then, the sum

$$\frac{1}{N} \sum_{n=1}^N g(\theta_n)w_n = \sum_{n=1}^N \frac{g(\theta_n)f(x|\theta)\pi(\theta)}{h^*(\theta_n)} \quad (19)$$

converges almost surely to  $\int_{\theta} g(\theta)f(x|\theta)\pi(\theta)d\theta$  by the strong law of large numbers, and the expectation (16) can be approximated by

$$\frac{1}{N} \frac{\sum_{n=1}^N g(\theta_n)w_n}{\sum_{n=1}^N w_n} \quad (20)$$

Note that if  $h(\theta)$  is small in region where  $|g(\theta)f(x|\theta)\pi(\theta)|$  is large, then this important region for the integral is only sampled infrequently. For this reason, choosing a distribution  $h(\theta)$  with heavy tails, and such that  $\text{supp}(g(\theta)f(x|\theta)\pi(\theta)) \subseteq \text{supp}(h(\theta))$ , is necessary for quick convergence.

## 4.2 Sampling via Monte Carlo Markov Chain Gibbs Sampler

A Markov chain is a discrete-parameter stochastic process  $X_n$  possessing the *Markov property*, which in its simplest form can be written

$$P(X_{n+1}|X_n, X_{n-1}, \dots, X_1) = P(X_{n+1}|X_n) \quad (21)$$

In words, the Markov property says that the future of the chain depends on its past states only through the present. A Markov chain can be specified by an initial probability distribution  $p^{(0)}(x)$  and a transition probability density  $T(x', x)$ , defined such that

$$P(X_{i+1} \in B|X_i = x') = \int_B T(x', x)dx \quad (22)$$

Then, the probability distribution of the chain at the  $(n + 1)$ st iteration is

$$p^{(n+1)}(x') = \int T(x', x)p^n(x)d^N x \quad (23)$$

A distribution  $\pi(x)$  is called the *invariant distribution* of the Markov process if it satisfies

$$\pi(x') = \int T(x', x)\pi(x)d^N x \quad (24)$$

Three technical properties of a Markov chain are necessary to use it for sampling. First, for a Markov chain  $X_n$ , define  $\tau_y = \inf\{n \geq 1 : X_n = y|X_0 = y\}$ . This random variable  $\tau_y$  is called the “time of first return to  $y$ .” The state  $y$  is said to be *positive recurrent* for the Markov chain if  $P(\tau_y < \infty) = 1$  and  $E[\tau_y] < \infty$ . If all states in the state space of the chain are positive recurrent, then this property is also used to describe the Markov chain.

Secondly, a state  $y$  has *periodicity*  $k$  if any return to  $y$  must occur in a multiple of  $k$  steps. That is, the period  $k$  is the greatest common divisor of  $\{n : P(x_n = y|X_0 = y) > 0\}$ . If  $k = 1$ ,

the state is said to be *aperiodic*, and this property also describes the Markov chain if all states possess this quality.

Finally, one says that “ $x$  communicates with  $y$ ” if the probability transition density  $T^n(x, y)$  is nonzero for some  $n > 0$ . In other words, if  $x$  communicates with  $y$ , then there is some chance of ending up at  $y$  some time in the future if the chain starts at  $x$ . If every state communicates with every other state, then the state space and the Markov chain are said to be *irreducible*. Under these three conditions and the assumption that a stationary distribution  $\pi(x)$  exists, the following nontrivial theorem holds:

$$p^n(x) \rightarrow \pi(x) \text{ in distribution as } t \rightarrow \infty, \text{ for any } p^{(0)}(x) \quad (25)$$

Monte Carlo Markov Chain (MCMC) techniques exploit the fact that if one can find an irreducible, aperiodic, and positive recurrent Markov chain for which the target distribution is the invariant distribution, then the chain can be used to produce samples from the target.

The Gibbs sampling method provides such a Markov chain for drawing samples from the target distribution. The method is based on the assumption that while the target  $P(x)$  is too difficult to sample directly, the full conditional distributions  $P(x_i | \{x_j : j \neq i\})$  of each  $x_i$  can be sampled. This is frequently the case in the hierarchical Bayes’ context for the posterior, which of course is described by simpler, conditional models.

Starting from an arbitrary initial condition, the Gibbs sampling algorithm proceeds by sampling each model variable from its full conditional while holding all the others constant. A round of sampling of each variable constitutes one iteration. At step  $n$ ,

$$\begin{aligned} x_1^{(n+1)} &\sim P(x_1 | x_2^{(n)}, x_3^{(n)}, \dots, x_N^{(n)}) \\ x_2^{(n+1)} &\sim P(x_2 | x_1^{(n+1)}, x_3^{(n)}, \dots, x_N^{(n)}) \\ &\vdots \\ x_N^{(n+1)} &\sim P(x_N | x_1^{(n+1)}, x_2^{(n+1)}, \dots, x_{N-1}^{(n+1)}) \end{aligned} \quad (26)$$

as  $n \rightarrow \infty$ , the distribution of  $x^{(n)}$  converges to the target distribution.

### 4.3 Sampling in the Air-Sea Problem

For brevity, denote all of the wind parameters of the process model described above by  $\theta_w$ , and the hyperparameters by  $\eta_w$ . The ocean parameters and hyperparameters will be denoted the same way but with subscript  $\psi$ . Then, in terms of the data, process, and prior level models described above, the model developed by Berliner, Milliff, and Wikle has posterior given by

$$[U, V, \psi, \theta_w, \theta_\psi, \eta_w, \eta_\psi | D_\psi, D_w] \propto \quad (27)$$

$$[D_\psi | \psi, \theta_\psi] \quad (28)$$

$$[\Psi^1 | U, V, \eta_\psi] \prod [\Psi^{t+1} | \Psi^t, U, V, \eta_\psi] [\eta_\psi, \theta_\psi] \quad (29)$$

$$[D_w | U, V, \theta_w] [U, V, \eta_w] [\eta_w, \theta_w] \quad (30)$$

In order to sample the posterior (27), the authors use a combination of Gibbs and importance sampling. The prior distributions in this case are all familiar forms that are easy to sample. Running a simple Gibbs sampler on the last line (the factor (30)) is therefore not difficult, and produces wind samples from  $[U, V, \theta_w, \eta_w | D_w]$ . Given these samples, it is a simple matter to generate samples of  $\Psi$  by sampling the oceanic prior distributions and process level model in term (29). At this stage, the algorithm has produced samples from the distribution

$$[U, V, \psi, \theta_w, \theta_\psi, \eta_w, \eta_\psi | D_w] \propto [\Psi^1 | U, V, \eta_\psi] \prod [\Psi^{t+1} | \Psi^t, U, V, \eta_\psi] [\eta_\psi, \theta_\psi] [U, V, \theta_w, \eta_w | D_w] \quad (31)$$

In the final stage of computation, this distribution is used as the importance sampler  $h^*$  to include conditioning on the oceanic data  $D_\Psi$  and compute expectations of functions of the posterior (27).

Berliner et al. comment on a practical complexity that arose during the computational stage of their work. While the simple sampling algorithm described above is theoretically valid, running it was inefficient. The standard deviation of the altimetry data model,  $\sigma_{o,e}$ , was inflated by a factor of one hundred to speed the computation. Since the error of the altimeter data level model  $e_o$  is given by  $e_o \sim N(0, \sigma_{o,e}^2 I)$ , this modification amounts to a significant weakening of the influence of the altimeter data on the posterior, and has important implications in the results.

## 5 Results and Evaluation

Berliner et al. examine several aspects of their posterior distribution to assess their modeling efforts. First, they check that the results provide a qualitative match to the model “truth.” By plotting the target and the reconstruction side-by-side, they show that to the eye, the Bayes estimator provides a good match to the general field morphology. Figure (1) here shows the comparison for the 7th day, which is the latest OSSE time for which results are presented. Since the reconstruction is initialized with a coarsened version of the initial target state with only a small amount of noise added, later “snapshots” of the field provide a more stringent evaluation of the reconstruction performance.

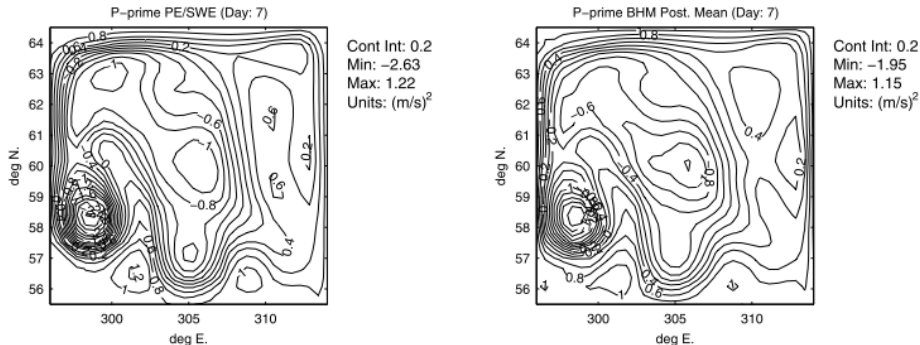


Figure 1: Berliner et al. [1], figure 3. Target (left) and reconstructed (right) perturbation pressure fields on the 7th day of the OSSE.

Next, the difference of the Bayes’ estimator from the target is presented, for reconstructions based on both the scatterometer and altimeter data, and then for the scatterometer data alone. In figure (2), the former field of residuals is shown on the left, and the latter on the right. The scatterometer-only residuals clearly show a global bias. Throughout the domain, the differences are negative, indicating a tendency for the reconstruction to overestimate the oceanic pressure. The model does most poorly in the southwest corner of the domain, where the target has a low pressure system of magnitude around  $-2 \text{ m}^2/\text{s}^2$  in the center. (Note that while the deviations from the mean pressure field are overestimated, this means the *magnitude* of the low pressure system is underestimated.) The presence of the cyclone, a relic of the dynamical “truth” spin-up conditions, represents an unlikely realization from the standpoint of the process level model, small gaussian deviations about a quasi-geostrophic balance. Quasi-geostrophy assumes only small vertical motions, and so the energetic cyclonic activity in the data is inconsistent with

the prior model. Without increasing the amount of data, improvements in reconstructing this feature could be made by weighting the importance of the data more heavily in the BHM. Indeed, some of the difficulty in capturing the magnitude of this feature may have been overcome had the authors not inflated the variance in the altimeter data model for computational speed (see discussion of numerics).

The comparison of the difference fields from reconstructions with and without altimeter data clearly shows an improvement with the greater amount of data. On the seventh day of the OSSE, an altimeter data transect crosses directly through the SW corner low pressure system. By including this data in the reconstruction, the error in estimating the peak of the system is cut in half. Still, the error of  $.6 \text{ m}^2/\text{s}^2$  in the center of the low pressure system represents a significant fraction of the magnitude of the true value ( $-2 \text{ m}^2/\text{s}^2$ ). In addition, despite the reduction of error close to the altimeter tracks, the model still displays a tendency to underestimate the magnitude of deviations from the mean pressure almost everywhere in the domain.

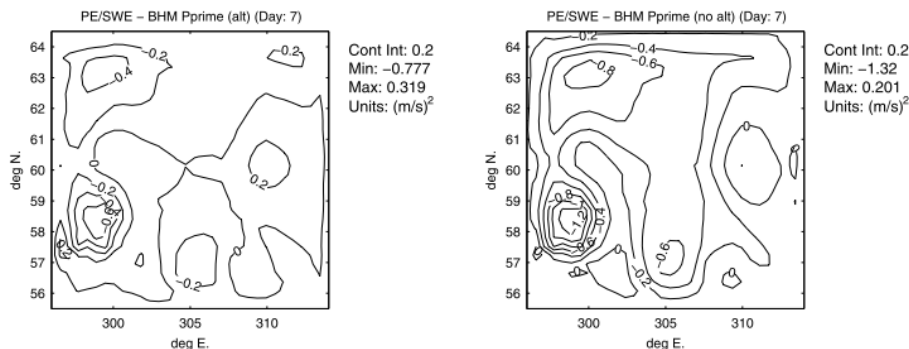


Figure 2: Berliner et al. [1], figure 4. Day-7 residuals of reconstruction conditioned on altimeter and scatterometer data (left) and scatterometer data only (right).

The standard deviation of the posterior estimates of the pressure field provide a measure of the reconstruction uncertainty. The most striking feature in the standard deviation over the domain on OSSE day seven is the sharp peak in the standard deviation in the southwest corner (see figure (3)). The greater degree of uncertainty in this region provides further confirmation that the model has the most trouble reconstructing this cyclone. It is somewhat heartening to note that the target truth is only one standard deviation away from the Bayes' estimator of the pressure field in the center of this feature. On the other hand, the standard deviation field generally takes on values on the same order of magnitude as the truth field over much of the domain. Even though the Bayes' estimator of pressure is morphologically reasonable, then, the estimate comes with a large envelope of uncertainty.

Berliner et al. also look at the kinetic energy (KE) of their reconstruction integrated over the whole basin, and compare it to that of the target. Their plot of the posterior KE over the time interval of the OSSE is shown below in figure (4). The heavy black line represents the target integrated KE, and the dashed black line is the posterior mean integrated KE. A model bias to underestimate the overall kinetic energy is clearly apparent, and worsens in time as well. Physically speaking, this result shows that the model consistently underestimates the magnitude of the flow fields. This bias is probably mainly due to the model's tendency to underestimate the magnitude of the cyclone in the SW-corner of the domain. To test this hypothesis, it would be useful to run the reconstruction algorithm again under spin-up conditions of the dynamical model "truth" that do not generate an underlying cyclonic field.

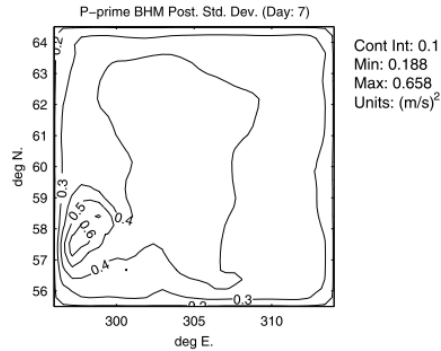


Figure 3: Berliner et al. ([1]), figure 5. Standard deviation of reconstructed perturbation pressure field on the 7th day of the OSSE.

Finally, analysis of the posterior distributions of the model parameters provides yet another evaluation of performance. The posterior of  $a_h$  in particular suggests the modeling could be improved. This parameter was originally intended as the coefficient of dissipation of vorticity by eddy processes. The authors mention briefly that “only by trial and error did we arrive at a uniform prior for  $a_h$  that included both positive and negative values.” In fact, the distribution of  $a_h$  conditional on the data strongly favors negative values (figure (5)), indicating that the last term in the process model is no longer serving to represent diffusion as originally intended. It is possible that with a negative coefficient, this term is acting as “negative diffusion,” acting to pump energy into the system suggested by the data, but unallowed by the QG process model. This discrepancy between the physical interpretation of the process modeling and its realizations suggests that the results could be improved by a reconsideration of which effects from equation (8) are appropriate to include in the coarsened process model.

## 6 Conclusion

Including mechanistic information from science is not typical in Bayes hierarchical modeling. However, Berliner, Milliff and Wikle demonstrate the usefulness of this extension of the classical methodology for data assimilation in the reconstruction of geophysical fields. Their results appear somewhat compromised by measures taken to improve numerical efficiency, but perhaps represent a lower bound for the potential of the methodology to solve such estimation problems given perpetually increasing computational power. In using a “physical-statistical” reconstruction approach, one should evaluate results carefully in light of the mechanistic information used to derive the model. Any inconsistencies that arise point to a need for more general modeling of the phenomena at hand under the fixed availability of computational power. Though the aim of Berliner and colleagues was mainly pedagogical, their results demonstrate that prior knowledge of oceanic flow in situations beyond quasi-geostrophy is probably necessary to better understand the ocean response to energetic atmospheric phenomena through hierarchical modeling.

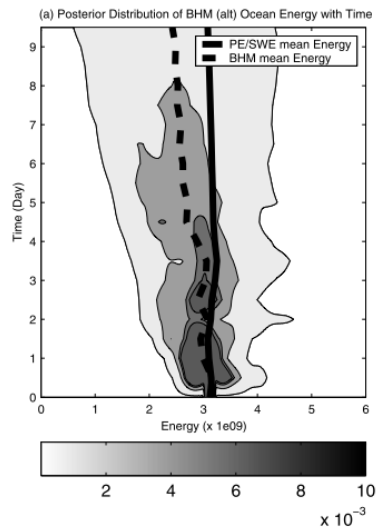


Figure 4: Berliner et al. [1], figure 7a. Basin integrated kinetic energy distribution of the posterior, compared to truth (solid black line).

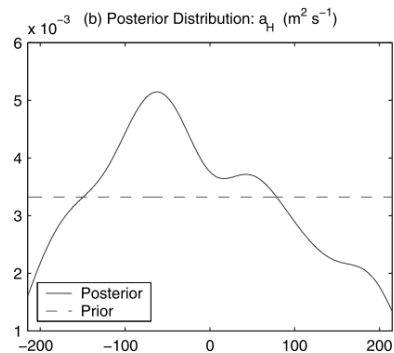


Figure 5: Berliner et al. [1], figure 6b. Posterior distribution of parameter intended as coefficient of eddy diffusion.

## References

- [1] Berliner, L.M, R.F. Milliff, and C.K. Wikle, Bayesian hierarchical modeling of air-sea interaction, *Journal of Geophysical Research*, 2003, doi:10.1029/2002JC001413.
- [2] Robert, C.P., (2007). *The Bayesian Choice*. Paris: Springer Texts in Statistics.
- [3] Haslett J, Whitley M, Bhattacharya S, et al., Bayesian palaeoclimate reconstruction, *J. Royal Stat. Society Series A*, 2006, 169, 395-430
- [4] Korhola A, Vasko K, Toivonen HTT, et al., Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling, *Quat. Sci. Rev.*, 2002, 21, 1841-1860.
- [5] Tebaldi C, Sanso B, Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach, *J. Royal Stat. Society Series A*, 2009, 172, 83-106.
- [6] Hegerl, G.C., T. Crowley, W.T. Hyde and D. Frame, Constraints on climate sensitivity from temperature reconstructions of the past seven centuries, *Nature*, 2006, 440, 1029-1032.
- [7] Milliff, RF. and McWilliams, J.C., The Evolution of Boundary Pressure in Ocean Basins, *J. Phys. Oceanogr.*, 24, 1317-1338, 1994.
- [8] Pedlosky, J. (1987). *Geophysical Fluid Dynamics*. New York: Springer-Verlag.
- [9] NASA Jet Propulsion Laboratory WINDS:Missions:SeaWinds webpage; <http://winds.jpl.nasa.gov/missions/quikscat/index.cfm#measurements>
- [10] Mackay, D.J. , Introduction to Monte Carlo Methods, review paper in the proceedings of an Erice summer school, retrieved from <http://www.inference.phy.cam.ac.uk/mackay/>