Q. J. R. Meteorol. Soc. (2005), Submitted to WMO DA Symposium Special Issue May 20, 2005

Multi-Model Fusion and Error Parameter Estimation

By O. G. LOGUTOV* and A. R. ROBINSON
Harvard University, USA

(Received 20 May 2005)

Summary

A robust and practical methodology for multi-model ocean forecast fusion has been sought. To this end, we consider an extension of Maximum-Likelihood (ML) error parameter estimation to multi-model predictive systems and utilize the resulting methodology for adaptive Bayesian model fusion. Proposed multi-model error parameter estimation is based on the Expectation-Maximization (EM) method, with the true state-space vector treated as missing data to simplify the log-likelihood expression. With only one forecasting model, the method reduces to the standard maximum-likelihood error parameter estimation. Synthetic data tests indicate the importance of the EM-based approach as opposed to simple transfer of the standard methodology developed for a single model system to multi-model systems.

Efficient solution of the pertinent minimization problem is the focus of the second half of the study. Bayesian Multi-Model Fusion represents a computationally intensive task and might be impractical for real-size applications. We discuss a randomized algorithm that alleviates the problem and cuts the computational complexity and storage to practical limits at a controlled expense of optimality. Our method is based on constructing and maintaining "randomized sketches" of the full Bayesian Model Averaging matrices instead of their full high-dimensional counterparts.

We illustrate the methodology on the example of two-model HOPS/ROMS forecasting within the framework of AOSN-2 real-time forecasting experiment held in Monterey Bay in 2003.

Keywords: Data Assimilation Ocean and Atmospheric Forecasting Adaptive Methods

1. Introduction

Various forecasting models have different skill in capturing aspects of reality and therefore forecasting could be improved through model combination. The methodology for ocean/atmospheric multi-model forecasting, however, is at an early stage. Current practices are dominated by the multiple-regression based approaches (*e.g.* Krishnamurti *et al.* 1990, Kharin and Zwiers 2002, Doblas-Reyes *et al.* 2000) and require a substantial training data set. On the other hand, real-world forecasting systems must adapt and evolve in response to modeled processes. Use of multi-models has been hampered by the fact that the time scale for changes to a forecasting system is often shorter than the time it takes to collect a sufficient sample of past events for robust model combination. In this paper, we work around this limitation by treating the optimal model combination as a non-stationary problem that calls for an adaptive version of methodology. We advocate an adaptive Bayesian model fusion that consists of the following three general steps: a. parameterization of forecast uncertainties through either a suitable parametric family (offers computational advantage) or through a low-rank approximation (allows for non-homogeneous dynamically motivated error subspaces); b. update of forecast uncertainty parameters via maximum-likelihood (Section 2); and c. combining model forecasts based on their relative uncertainties via maximum-likelihood (Section 3). In order to implement step b. in the foregoing we have extended the maximum-likelihood Error Parameter Estimation to multi-model forecasting systems through the expectation-maximization technique.

The rest of the paper is organized as follows. In Section 2, we describe the multi-model error parameter estimation based on maximum-likelihood. The core

* Corresponding author: Pierce Hall, 29 Oxford St., Cambridge, MA, USA 02138

part of the construction is the expectation-maximization technique. In Section 3, we apply this developed methodology for multi-model Bayesian fusion and also discuss efficiency issues. Finally, we illustrate the methodology on the example of AOSN-2 HOPS/ROMS real-time forecasting exercises in Monterey Bay in August 2003.

2. Multi-Model Error Parameter Estimation

(*a*) *Setup, Notation, Motivation*

Suppose a multi-model ocean/atmospheric predictive system consists of m models that produce independent forecasts, $\left\{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_m^k\right\}_{k=1}^K$, valid at times $\{t_k\}_{k=1}^K$, with the corresponding forecast error $\left\{\boldsymbol{\epsilon}_1^k, \boldsymbol{\epsilon}_2^k, \ldots, \boldsymbol{\epsilon}_m^k\right\}_{k=1}^K$. Suppose also that validating measurements, $\left\{\mathbf{y}^k\right\}_{k=1}^K$, with error $\left\{\boldsymbol{\epsilon}_o^k\right\}_{k=1}^K$, become available. We pose ourselves with the problem of finding the optimal strategy for combining model forecasts, $\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \ldots, \mathbf{x}_m^{k+1}$, the next time prediction is being made.

Let $\mathbf{x}$ be the true state-space on a central forecast grid (the central forecast domain, for example, being the largest model domain in the forecasting system) and $\mathbf{H}_i$ be linear mapping from the central forecast state-space onto the ith model state-space

$$\left\{\begin{array}{lclcll} \mathbf{x}_i & = & \mathbf{H}_i\mathbf{x} & + & \boldsymbol{\epsilon}_i & i=1,\ldots,m \\ \mathbf{y} & = & \mathbf{H}_o\mathbf{x} & + & \boldsymbol{\epsilon}_o & \end{array}\right. \tag{1}$$

The length of time window K is taken to reflect the time scale of changes to a forecasting system. Denote all past forecasts and validating data within the time window K as $\mathcal{D}$, $\mathcal{D} = \left\{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_m^k, \mathbf{y}^k\right\}_{k=1}^K$. The maximum-likelihood fusion of forecasts $\mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \ldots, \mathbf{x}_m^{k+1}$ is defined as

$$\mathbf{x}^* = \arg\max_{\mathbf{x}} \left\{ p\left(\mathbf{x} \middle| \mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \ldots, \mathbf{x}_m^{k+1}, \mathcal{D}\right) \right\} \tag{2}$$

We simplify (2) by parameterizing forecast uncertainties and combining forecasts via

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} \left\{ p\left(\mathbf{x} \middle| \mathbf{x}_1^{k+1}, \mathbf{x}_2^{k+1}, \ldots, \mathbf{x}_m^{k+1}, \boldsymbol{\Theta}^*\right) \right\} \tag{3}$$

where $\boldsymbol{\Theta}^*$ denotes the estimates of forecast uncertainty parameters in m models found by maximizing the log-likelihood of parameters given data, $\mathcal{D}$

$$\boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\Theta} \middle| \mathcal{D}\right) \qquad \boldsymbol{\Theta} = \{\boldsymbol{\alpha}_i\}_{i=1}^m \tag{4}$$

Parameterization of uncertainties is needed to reduce the number of degrees of freedom in the system to ensure robust parameter estimation from data available within the time window K. We hereby assume unbiased Gaussian forecast errors, $\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \mathbf{B}_i\right)$, $\mathbf{B}_i = \langle \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T \rangle$, and forecast error covariances, $\{\mathbf{B}_i\}_{i=1}^m$, represented either via a suitable parametric family, $\mathbf{B}_i \approx \hat{\mathbf{B}}_i(\boldsymbol{\alpha}_i)$, or via a low-rank approximation, $\mathbf{B}_i \approx \mathbf{V}_p \mathbf{Q}_{(\alpha)} \mathbf{V}_p^T = \sum_{i=1}^p q_i^{(\alpha)} \mathbf{v}_i \mathbf{v}_i^T$. The exact way of parameterizing $\mathbf{B}_i$s remains at a discretion of a researcher. In cases when errors are known to be non-Gaussian, a mixture of Gaussians can be utilized to approximate the probability densities, $p(\boldsymbol{\epsilon}_i) \sim \sum_{l=1}^{L} \mathcal{N}_l\big(\mathbf{0}, \mathbf{B}_{il}\big)$, and the expectation-maximization procedure that solves the mixture of Gaussians estimation problem can be included on top of the methodology described in this paper, with only minor modifications. For the sake of simplicity, we do not discuss this case here and refer a reader to Redner (1984) for a review of mixture of Gaussians problem.

(*b*) *single-model and multi-model error parameter estimation*

With only one forecasting model, $\mathbf{x}_1$, the Gaussian error assumption leads to a standard maximum-likelihood error parameter estimation from model-data misfits (Dee 1995, Dee and da Silva 1999, Purser and Parrish 2003). Since a sum of Gaussians, $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_o$, is also a Gaussian, with covariance $\langle \boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_1^T \rangle + \langle \boldsymbol{\epsilon}_o \boldsymbol{\epsilon}_o^T \rangle$, model-data misfits, $\mathbf{d} = \mathbf{y} - \mathbf{H}\mathbf{x}_1$, are random samples from the normal distribution $\mathcal{N}\big(\mathbf{0}, \mathbf{Q}(\boldsymbol{\alpha})\big)$, where $\mathbf{Q}(\boldsymbol{\alpha})$ is the sum of the forecast and the observation error covariances, $\mathbf{Q}(\boldsymbol{\alpha}) = \mathbf{H}\mathbf{B}(\boldsymbol{\alpha})\mathbf{H}^T + \mathbf{R}$. The Maximum-Likelihood (ML) error parameters are then found through minimizing the log-likelihood expression

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{\alpha}) \tag{5}$$

$$\mathcal{L}(\boldsymbol{\alpha}) = \big(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\big)^T \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}^{-1} \big(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0\big) + \log\det \mathbf{Q}(\boldsymbol{\alpha}) + \frac{1}{K}\sum_{k=1}^{K} \mathbf{d}_k^T \mathbf{Q}^{-1}(\boldsymbol{\alpha}) \mathbf{d}_k$$

where the first term describes prior information on parameter values, $\mathcal{N}\big(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma}_{\boldsymbol{\alpha}}\big)$, and can be omitted if no such prior information is available.

In a multi-model system (1), model-model misfits are another source of information about forecast errors that complements the information contained in model-data misfits. Learning forecast errors from model-data misfits only, within each model separately, is equivalent to ignoring the information contained in model-model misfits and results in suboptimal error parameter estimation. Synthetic data tests presented later in this Section indicate that this leads to significant degradation of error parameter estimates (Table 1). Multi-model maximum-likelihood error parameter estimation corresponds to maximizing the joint posterior probability

$$\{\hat{\mathbf{x}}, \hat{\boldsymbol{\Theta}}\} = \arg\max_{\mathbf{x},\boldsymbol{\Theta}} p(\mathbf{x}, \boldsymbol{\Theta}|\mathcal{D}) \tag{6}$$

which contains all the information about the true state, $\mathbf{x}$, and the true error parameters in m models, $\boldsymbol{\Theta}$, given available data $\mathcal{D}$ within an appropriate time window K.

(*c*) *multi-model error parameter estimation via expectation-maximization*

We discuss one practical way of solving (6). We notice that the problem is amenable to a missing data interpretation. If we knew the true state-space $\mathbf{x}$, error parameter estimation would have become straightforward. We therefore augment the data, $\mathcal{X} = \big\{\mathcal{D}, \mathbf{x}\big\}$, where $\mathbf{x}$ is the true state-space in the observation space, and expand the joint probability density, $p(\mathbf{x}, \boldsymbol{\Theta}|\mathcal{D})$, in terms of the complete-data likelihood, $p(\mathbf{x}, \boldsymbol{\Theta}|\mathcal{D}) \propto p(\boldsymbol{\Theta}|\mathcal{X})p(\mathbf{x}|\mathcal{D})$. The incomplete-data log-likelihood function $\log p(\boldsymbol{\Theta}|\mathcal{D})$ can not be easily expressed. However, the complete-data log-likelihood, $\log p(\boldsymbol{\Theta}|\mathcal{X})$, is readily expressible. The *expectation-maximization* methodology has been designed specifically to handle this type of problems (Dempster *et al.* 1977). The complete-data log-likelihood is expressed as

$$\log p(\boldsymbol{\Theta}|\mathbf{x},\mathcal{D}) \propto \sum_{i=1}^{m} \log\det \boldsymbol{\mathcal{B}}_i(\boldsymbol{\alpha}_i) + \sum_{i=1}^{m} \big(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\big)^T \boldsymbol{\mathcal{B}}_i^{-1}(\boldsymbol{\alpha}_i)\big(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\big) \quad (7)$$

where $\mathbf{H}_i$ denotes linear mapping from the ith model state-space onto the observational space, and $\boldsymbol{\mathcal{B}}_i(\boldsymbol{\alpha}_i)$ is the ith model forecast error covariance in the observational space, $\boldsymbol{\mathcal{B}}_i = \mathbf{H}_i\mathbf{B}_i\mathbf{H}_i^T$. The prior term $\big(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\big)^T\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}^{-1}\big(\boldsymbol{\Theta}-\boldsymbol{\Theta}_0\big)$ again could be included in (7) given $\mathcal{N}\big(\boldsymbol{\Theta}_0,\boldsymbol{\Sigma}_{\boldsymbol{\Theta}}\big)$ prior on $\boldsymbol{\Theta}$. The marginal probability density $p(\mathbf{x}|\mathcal{D})$ is a normal distribution

$$p(\mathbf{x}|\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_m,\mathbf{y}) = \mathcal{N}\Big(\mathbf{x},\mathbf{x}_a,\boldsymbol{\mathcal{B}}_a(\boldsymbol{\Theta})\Big) \quad (8)$$

with the analysis, $\mathbf{x}_a$

$$\mathbf{x}_a = \arg\min_{\boldsymbol{x}} \sum_{i=1}^{m} \big(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\big)^T \boldsymbol{\mathcal{B}}_i^{-1}(\boldsymbol{\alpha}_i)\big(\mathbf{x}_i-\boldsymbol{\mathcal{H}}_i\mathbf{x}\big) + \big(\mathbf{x}-\mathbf{y}\big)^T\mathbf{R}^{-1}\big(\mathbf{x}-\mathbf{y}\big) \quad (9)$$

and the analysis error covariance, $\boldsymbol{\mathcal{B}}_a(\boldsymbol{\Theta})$

$$\boldsymbol{\mathcal{B}}_a^{-1}(\boldsymbol{\Theta}) = \boldsymbol{\mathcal{B}}_1^{-1}(\boldsymbol{\alpha}_1) + \boldsymbol{\mathcal{B}}_2^{-1}(\boldsymbol{\alpha}_2) + \ldots + \boldsymbol{\mathcal{B}}_m^{-1}(\boldsymbol{\alpha}_m) + \mathbf{R}^{-1} \quad (10)$$

An alternative to (9) is a closed form expression

$$\mathbf{x}_a = \boldsymbol{\mathcal{B}}_a\mathbf{H}_1^T\mathbf{B}_1^{-1}\mathbf{x}_1 + \boldsymbol{\mathcal{B}}_a\mathbf{H}_2^T\mathbf{B}_2^{-1}\mathbf{x}_2 + \ldots + \boldsymbol{\mathcal{B}}_a\mathbf{H}_m^T\mathbf{B}_m^{-1}\mathbf{x}_m + \boldsymbol{\mathcal{B}}_a\mathbf{R}^{-1}\mathbf{y} \quad (11)$$

The forecast error parameters, $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}_i\}_{i=1}^m$ are not known in (7)–(11). We proceed by applying the expectation-maximization formalism which consists of maximizing the expectation of the complete-data log-likelihood given previous estimate or guess at parameter values, $Q(\boldsymbol{\Theta},\hat{\boldsymbol{\Theta}}^k)$

$$\begin{cases} \text{E}: \quad Q(\boldsymbol{\Theta},\hat{\boldsymbol{\Theta}}^k) = \mathcal{E}_{\mathbf{x}|\mathcal{D},\hat{\boldsymbol{\Theta}}^k}\Big\{\log p(\mathbf{x},\mathcal{D}|\boldsymbol{\Theta})\Big\} = \int_{\mathbf{x}} \log\Big(\mathcal{L}(\boldsymbol{\Theta}|\mathbf{x},\mathcal{D})\Big)p(\mathbf{x}|\mathcal{D},\hat{\boldsymbol{\Theta}}^k)d\mathbf{x} \\ \text{M}: \quad \hat{\boldsymbol{\Theta}}^{k+1} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta},\hat{\boldsymbol{\Theta}}^k) \end{cases} \quad (12)$$

That is, we use $\hat{\boldsymbol{\Theta}}^k$ for the purposes of estimating the marginal density (8)

$$p(\mathbf{x}|\mathcal{D},\hat{\boldsymbol{\Theta}}^k) = \mathcal{N}\Big(\mathbf{x},\mathbf{x}_a^k,\mathbf{B}_a(\hat{\boldsymbol{\Theta}}^k)\Big) \quad (13)$$

and employ (13) to evaluate the expectation of the complete-data log-likelihood (E step). We next update the current parameter estimates by maximizing the expected value of the complete-data log-likelihood (M step). The procedure is known as *Expectation-Maximization* and is provably convergent (Dempster *et al.* 1977, Redner 1984). By substituting (13) in (12) and using Gaussian integral identity, $\int_{\mathbf{x}} \left(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\right)^T \boldsymbol{\mathcal{B}}_i^{-1}\left(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\right)\mathcal{N}(\mathbf{x},\mathbf{x}_a,\boldsymbol{\mathcal{B}}_a)d\mathbf{x} = \left(\mathbf{x}_a-\mathbf{H}_i\mathbf{x}_i\right)^T\boldsymbol{\mathcal{B}}_i^{-1}\left(\mathbf{x}_a-\mathbf{H}_i\mathbf{x}_i\right) + \mathrm{Tr}\left(\boldsymbol{\mathcal{B}}_i^{-1}\boldsymbol{\mathcal{B}}_a\right)$ for every term in (12), we obtain a closed form expression for the expectation of complete-data log-likelihood

$$Q(\boldsymbol{\Theta},\boldsymbol{\Theta}^k) \propto \sum_{i=1}^{m} \log\det\boldsymbol{\mathcal{B}}_i(\boldsymbol{\alpha}_i) + \sum_{i=1}^{m}\left(\hat{\mathbf{x}}_a^k-\mathbf{H}_i\mathbf{x}_i\right)^T\boldsymbol{\mathcal{B}}_i^{-1}(\boldsymbol{\alpha}_i)\left(\hat{\mathbf{x}}_a^k-\mathbf{H}_i\mathbf{x}_i\right) \tag{14}$$

Hence, the expectation-maximization procedure simplifies to a sequence of iterative updates to the analysis in the observation space, $\hat{\mathbf{x}}_a^k$, based on current error parameter estimates, $\hat{\boldsymbol{\Theta}}^k$, and subsequent updates to the error parameter values from model-analysis misfits

$$\left\{\begin{array}{rl} \hat{\mathbf{x}}_a^k = & \arg\min\limits_{\boldsymbol{x}} \sum_{i=1}^{m}\left(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\right)^T\boldsymbol{\mathcal{B}}_i^{-1}(\boldsymbol{\alpha}_i^{k-1})\left(\mathbf{x}-\mathbf{H}_i\mathbf{x}_i\right) + \left(\mathbf{x}-\mathbf{y}\right)^T\mathbf{R}^{-1}\left(\mathbf{x}-\mathbf{y}\right) \\ \hat{\boldsymbol{\Theta}}^{k+1} = & \arg\min\limits_{\boldsymbol{\Theta}} \sum_{i=1}^{m}\log\det\boldsymbol{\mathcal{B}}_i(\boldsymbol{\alpha}_i) + \sum_{i=1}^{m}\left(\hat{\mathbf{x}}_a^k-\mathbf{H}_i\mathbf{x}_i\right)^T\boldsymbol{\mathcal{B}}_i^{-1}(\boldsymbol{\alpha}_i)\left(\hat{\mathbf{x}}_a^k-\mathbf{H}_i\mathbf{x}_i\right) \end{array}\right. \tag{15}$$

The analyses, $\hat{\mathbf{x}}_a^k$, in (15) could also be computed through (11).

(*d*) *Synthetic Data Tests*

An improvement that the EM-based procedure brings as compared to error parameter estimation from model-data misfits stems from inclusion of information contained in model-model misfits. Synthetic data tests in which the "truth" and the "true" error parameter values are generated and hence known can be carried out to illustrate the improvement for different uncertainty parameters attributed to observations. Figure 1 illustrates an example of such a synthetic data experiment, whereas Table 1 summarizes the statistics. Two "models" (blue and green lines in Figure 1) are drawn independently at random from a normal distribution around a synthetic "truth" (black line in Figure 1), with chosen error covariance parameters. An isotropic covariance model, $\mathbf{B}(\mathbf{x},\mathbf{y}) = \sigma^2\rho(||\mathbf{x}-\mathbf{y}||)$, with the fifth-order piecewise rational representing function, $\rho(r) = \rho_c(r,L)$, given by (4-10) in Gaspari and Cohn (1999), has been utilized. The choice of the covariance model was arbitrary and made solely because it yields well-conditioned $\mathbf{B}_i(\boldsymbol{\alpha}_i)$ for the whole range of parameter values. Error length scale parameters in models have been intentionally chosen very different. Synthetic "observations" have also been drawn at random from a Gaussian distribution, with diagonal covariance $\mathbf{R}$. The upper panel in Figure 1 illustrates the data that the multi-model error parameter estimation procedure is run on. The middle and the bottom panels show the analyses (magenta lines) computed with error parameters obtained from model-data misfits only (middle panel) and through the EM procedure described in this paper (bottom panel). An experiment illustrated in Figure 1 was repeated 100 times for each uncertainty level attributed to observations (`Ratio` in Table 1) to collect statistics. Table 1 illustrates that given a sufficient level of uncertainty in measurements as compared to uncertainty in models, the information contained in model-model misfits leads to a significant improvement of error parameter estimates and of a multi-model analysis.

3. Maximum-Likelihood Forecast Fusion

(a) *Formalism*

Once error parameters have been estimated we can combine models based on their relative uncertainties via the maximum-likelihood principle. Given independent forecasts, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, the conditional probability density of the true state, $\mathbf{x}$, expands via individual forecast pdfs

$$p(\mathbf{x}|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) = p(\mathbf{x}|\mathbf{x}_1)p(\mathbf{x}|\mathbf{x}_2)\ldots p(\mathbf{x}|\mathbf{x}_m) \tag{16}$$

and, under the Gaussian error assumption, individual pdfs read $p(\mathbf{x}|\mathbf{x}_i) = \mathcal{N}\big(\mathbf{H}_i\mathbf{x}, \mathbf{x}_i, \mathbf{B}_i(\hat{\boldsymbol{\alpha}}_i)\big)$, where $\mathbf{H}_i$ maps from the central forecast state-space onto the ith model state-space as in (1). The forecast error parameters, $\hat{\boldsymbol{\alpha}}_i$, are estimated through (15). The maximum-likelihood (minimum variance) central forecast, $\mathbf{x}_c$, corresponding to (3) is found as

$$\mathbf{x}_c = \arg \min_{\boldsymbol{x}} \sum_{i=1}^{m} \big(\mathbf{x}_i - \mathbf{H}_i\mathbf{x}\big)^T \mathbf{B}_i^{-1}(\hat{\boldsymbol{\alpha}}_i)\big(\mathbf{x}_i - \mathbf{H}_i\mathbf{x}\big) \tag{17}$$

or

$$\mathbf{x}_c = \mathbf{B}_c\mathbf{H}_1^T\mathbf{B}_1^{-1}\mathbf{x}_1 + \mathbf{B}_c\mathbf{H}_2^T\mathbf{B}_2^{-1}\mathbf{x}_2 + \ldots + \mathbf{B}_c\mathbf{H}_m^T\mathbf{B}_m^{-1}\mathbf{x}_m \tag{18}$$

with the central forecast error covariance, $\mathbf{B}_c$, given by

$$\mathbf{B}_c = \left(\boldsymbol{\mathcal{B}}_1^{-1} + \boldsymbol{\mathcal{B}}_2^{-1} + \ldots + \boldsymbol{\mathcal{B}}_m^{-1}\right)^{-1} \qquad \boldsymbol{\mathcal{B}}_i^{-1} = \mathbf{H}_i^T\mathbf{B}_i^{-1}\mathbf{H}_i \tag{19}$$

The above procedure is equivalent to Bayesian Model Averaging (BMA), $\mathbf{x}_c = \mathbf{C}_1\mathbf{x}_1 + \mathbf{C}_2\mathbf{x}_2 + \ldots + \mathbf{C}_m\mathbf{x}_m$, with Bayesian weight matrices, $\mathbf{C}_i$s, given by $\mathbf{B}_c\mathbf{H}_i^T\mathbf{B}_i^{-1}$. Denote the sum of columns of a $\mathbf{C}_i$ as $\mathbf{p}_i$

$$\mathbf{p}_i = \big|\mathbf{C}_i\big|_1 = \sum_{n'=1}^{n_i} \mathbf{C}_i(j, n') \ \forall\, j \qquad \mathbf{C}_i = \mathbf{B}_c\mathbf{H}_i^T\mathbf{B}_i^{-1} \tag{20}$$

where j is the central forecast grid point index. Bayesian weights, $\mathbf{p}_i$s, provide the spatial distribution of a fraction of information to be assimilated from the ith model into the central forecast and satisfy $\mathbf{p}_i(j) > 0 \ \forall j$ and $\sum_{i=1}^{m} \mathbf{p}_i = \mathbf{1}$. $\mathbf{p}_i$s indicate how model error parameters translate in terms of relative model skill.

Since it is generally desirable that no extra smoothing be introduced through model fusion, $\mathbf{C}_i(j, :)$s are to be replaced with $\mathbf{p}_i(j)$s in the overlapping parts of the central and ith model domains, $\mathbf{x}_c^j \in \mathbf{x}_i$, for the purposes of Bayesian model fusion

$$\mathbf{x}_c = \left\{\begin{array}{l} \hat{\mathbf{p}}_1\mathbf{x}_1, \ \mathbf{x}_c^j \in \mathbf{x}_1 \\ \hat{\mathbf{C}}_1\mathbf{x}_1, \ \mathbf{x}_c^j \notin \mathbf{x}_1 \end{array}\right. + \left\{\begin{array}{l} \hat{\mathbf{p}}_2\mathbf{x}_2, \ \mathbf{x}_c^j \in \mathbf{x}_2 \\ \hat{\mathbf{C}}_2\mathbf{x}_2, \ \mathbf{x}_c^j \notin \mathbf{x}_2 \end{array}\right. + \ldots + \left\{\begin{array}{l} \hat{\mathbf{p}}_m\mathbf{x}_m, \ \mathbf{x}_c^j \in \mathbf{x}_m \\ \hat{\mathbf{C}}_m\mathbf{x}_m, \ \mathbf{x}_c^j \notin \mathbf{x}_m \end{array}\right. \tag{21}$$

It is worthwhile to point out the differences and similarities of (21) to the multiple-regression based approaches. In the foregoing, we assumed independent model forecasts and therefore uncorrelated errors between models. This is equivalent to enforcing the block-diagonal structure of cross-covariances, $\langle \mathbf{X}^T \mathbf{X} \rangle$, in the multiple-regression based methodology where $\mathbf{X}$ is the design matrix, such that the regression weights are found as $\hat{\mathbf{w}}_{LS} = \langle \mathbf{X}^T \mathbf{X} \rangle^{-1} \langle \mathbf{X}^T \mathbf{y} \rangle$. Bayesian model fusion is more general than the multiple-regression analysis in that the optimal weights, $\hat{\mathbf{p}}_i$s, are inhomogeneous in space. $\hat{\mathbf{p}}_i$s have a direct interpretation as probabilities of an ith model being correct, variable over the central forecast domain. The weights in the multiple-regression analysis do not have such clear interpretation.

(*b*) *Efficiency Issues*

Bayesian model fusion (3), both if implemented via (17) or (18)-(19), represents a computationally intensive task, with $\mathcal{O}(n^3 m)$ complexity and $\mathcal{O}(n^2)$ storage requirement when carried out directly (n being the dimensionality of an individual model state-space, $\mathcal{O}(10^6 - 10^7)$, and m the number of models). This might be impractical for some real-world applications. The difficulty of Bayesian multi-model fusion stems from the fact that variational analysis in a multi-model system is fundamentally computationally more challenging than in a single-model system. In a single-model system, the full background error covariance matrix, $\mathbf{B}$, never has to be stored and it's inversion can be avoided through change of variables, *e.g.* $\mathbf{v} = \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b)$ (Huang 2000), such that each iterative step of cost function minimization with respect to the new state-space $\mathbf{v}$ involves the multiplication by $\mathbf{B}$ rather than its inverse. It is generally not possible to completely avoid the background error covariance inversion in a multi-model system. The matrix identities that express the analysis error covariance, $\mathbf{B}_a = (\mathbf{B}^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H})^{-1} = \mathbf{B} - \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{B}$, and bring the analysis equation to a familiar incremental form, $\mathbf{x}_a = \mathbf{x}_b + \mathbf{B}\mathbf{H}^T(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}_b)$, are not applicable in a multi-model case. On the other hand, if the m forecasts are merged sequentially, one by one, the central forecast error covariance, $\mathbf{B}_c$, has to be updated every time a forecast is merged. This liquidates the initial parametric or low-rank representation of background error covariances and a new parametric or low-rank approximation needs to be developed for $\mathbf{B}_c$ at every sequential step. This is computationally expensive.

Several approaches can potentially alleviate the problem. We have followed one of such approaches that derives its efficiency from *randomization.* The method is based on constructing and maintaining randomized "sketches" of Bayesian weights, $\mathbf{p}_i$, and makes use of synopsis data structures that resemble "count-min sketches" developed by Cormode and Muthukrishnan (2005).

(*c*) *Bayesian weight computation via randomized sketches*

The method conceptually amounts to randomized grid sub-scaling through the use of d hash functions chosen at random and averaging over the results obtained for these different hash functions. Let $\mathcal{I} = \{1, \ldots, n\}$ be indices of elements in a forecast state-space and $\mathcal{J} \subseteq \mathcal{I}$ be elements of $\mathcal{I}$ that has already been sampled. Choose d hash functions $h_1, \ldots, h_d : \{\mathcal{I} \setminus \mathcal{J}\} \to \{1, \ldots, w\}$ uniformly at random from a pairwise-independent family. Denote $s_x[d, w]$ to be a sketch of $\mathbf{x} \in \mathcal{R}^n$, of size (d, w), formed by hashing the elements of $\mathbf{x}$ using $h_1, \ldots, h_d$ into d

w-dimensional vectors (Cormode and Muthukrishnan 2005). The size of a sketch is independent of the size of the central forecast state-space, n, and $(d, w) \ll n$.

Hash the central forecast state-space element indices, $\{\mathcal{I}_n^c \setminus \mathcal{J}_n^c\} \to s_c[1, w]$. Hash the individual forecast state-space indices of the m models into m sketches, each of size (d, w), $\{s_i[d, w]\}_{i=1}^m$. We, next, use the indices hashed in $s_c[1, w]$ and $\{s_i[d, w]\}_{i=1}^m$, to compute Bayesian Model Averaging matrices, $\{\tilde{\mathbf{C}}_i^j\}_{i=1}^m$, for $\forall j : 1 \leqslant j \leqslant d$. Every $\tilde{\mathbf{C}}_i^j$ is a $w \times w$ matrix $\tilde{\mathbf{C}}_i^j = \tilde{\mathbf{B}}_c^j \tilde{\mathbf{H}}_{ij}^T \tilde{\mathbf{B}}_{ij}^{-1}$, where $\tilde{\mathbf{B}}_{ij}$ corresponds to the ith forecast error covariance evaluated for the state-space elements with indices in the jth row of sketch $s_i[d, w]$, i.e. $\tilde{\mathbf{B}}_{ij} = \hat{\mathbf{B}}_{\{s_i(j,:)\}}(\hat{\boldsymbol{\alpha}}_i)$, $\tilde{\mathbf{H}}_{ij}$ are linear mappings from central forecast state-space elements with indices in $s_c[1, w]$ onto ith model state-space elements with indices in $s_i(j, :)$, $\tilde{\mathbf{B}}_c^j$ is computed as in (19), with $\mathbf{B}_i$ and $\mathbf{H}_i$ replaced by $\tilde{\mathbf{B}}_{ij}$ and $\tilde{\mathbf{H}}_{ij}$ correspondingly.

Once $\{\tilde{\mathbf{C}}_i^j\}_{i=1}^m$, $\forall j : 1 \leqslant j \leqslant d$ are computed, fill in the Bayesian weight sketches, $\{\tilde{p}_i[d, w]\}_{i=1}^m$, by summing up the columns of $\tilde{\mathbf{C}}_i$, $\tilde{p}_i(j, :) = \sum_{w'=1}^{w} \tilde{\mathbf{C}}_i(j, w')$, for every model i and every hash function j. Answer the Bayesian weight queries for central forecast indices $\{j : j \in s_c[1, w]\}$ by averaging the sketches of Bayesian weights, $\{\tilde{p}_i[d, w]\}_{i=1}^m$, across the hash functions

$$\hat{\mathbf{p}}_i(j : j \in s_c[1, w]) = \frac{1}{d} \sum_{d'=1}^{d} \tilde{p}_i[d', w] \tag{22}$$

Continue hashing the central forecast state-space element indices, $j \in \{\mathcal{I}_n^c \setminus \mathcal{J}_n^c\}$, into $s_c[1, w]$ and repeat all the above steps in constructing randomized sketches until all elements in $\{\mathcal{I}_n^c \setminus \mathcal{J}_n^c\}$ are exhausted, $\{\mathcal{I}_n^c \setminus \mathcal{J}_n^c\} = \emptyset$, and full $\{\hat{\mathbf{p}}_i\}_{i=1}^m$ are obtained. The computational complexity of obtaining all m $\hat{\mathbf{p}}_i$s is $\mathcal{O}(nw^2md)$ and the storage requirement is $\mathcal{O}(nwd)$, where n is the dimensionality of a model state-space and m the number of models, as compared to the $\mathcal{O}(n^3m)$ complexity and $\mathcal{O}(n^2)$ storage of Bayesian model fusion if carried out directly. By properties of randomized sketches (Cormode and Muthukrishnan 2005), $|\hat{\mathbf{p}}_i - \mathbf{p}_i| \leqslant \epsilon$ with probability $1 - \delta$ if the sketch size parameters are set $w = \lceil e/\epsilon \rceil$ and $d = \lceil \ln 1/\delta \rceil$, independent of n. $\{\hat{\mathbf{p}}_i\}_{i=1}^m$ sketches that satisfy a 1% accuracy level with probability 99% require $w \sim 500$ and $d \sim 6$. By changing the size parameters, (d, w), we set the computational complexity and storage to practical limits at a controlled expense of optimality.

4. Real-time Forecasting in the Monterey Bay/California Current system

We illustrate the methodology based on the real-time ocean forecasting exercises held in the Monterey Bay/California Current system in August of 2003 and designated as AOSN-2. The Autonomous Ocean Sampling Network (AOSN) project (Figure 2) brought together a wide range of measurements from various platforms with the state-of-the-art numerical ocean models and was designed to test and further improve the methods and engineering solutions behind an integrated observing and modeling system for regional ocean predictions. The full description of the experiment is available at `http://www.mbari.org/aosn/`. The exercises included two competing ocean models, the Harvard Ocean Prediction System (Robinson and Lermusiaux 2002, Robinson 1999) and the UCLA version of the Regional Ocean Modeling System (Shchepetkin and McWilliams 2005). These two forecasting models, HOPS and ROMS, were defined in overlapping but slightly different domains (Figure 3a) and operated independently of each other in real-time. A question was raised of a possibility to boost the overall forecasting skill of a system by combining the models.

To address the question, we have applied the methodology described in this paper. AOSN-2 program included two validation CTD surveys, in August 5-7 and in August 21-23 (Figure 3a,d). We have utilized the data from the first CTD survey to estimate the forecast error parameters in HOPS and ROMS. We have then applied these parameters for model fusion, as described in this paper. We have used the data from the second CTD survey to evaluate HOPS and ROMS forecast skill as compared to the skill of the Central forecast computed through model combination. Figure 3 provides an illustration of these tests. Panels b. and c. show an example of HOPS/ROMS individual SST forecasts. The corresponding Central forecast obtained via Bayesian fusion of HOPS and ROMS is shown in panel d. Table 2 summarizes some skill metrics, specifically, the rms error and the forecast SST correlation with the observed SST. These skill metrics have been accumulated for the individual HOPS and ROMS forecasts, and for the Bayesian fusion of HOPS and ROMS based on the multi-model error parameter estimation (CNTR_{EM}) and error parameter estimation from model-data misfits only ($\mathrm{CNTR}_{\mathbf{d}}$). We have found improvements in both the rms error statistics and the forecast correlation with validating measurements in Central forecasts computed through Bayesian model fusion in the course of described exercises. EM error parameter estimation procedure brings an additional improvement as compared to standard error parameter estimation from model-data misfits.

5. Conclusions

Describing the model fusion process within a probabilistic Bayesian framework through the formalism of multi-model error parameter estimation is a sensible venue for multi-model forecasting. An essential attribute of the proposed methodology is adaptiveness which is essential in view of continuous changes made to a forecasting system as part of typically real-world operational practices. We have demonstrated how the multi-model error parameter estimation can be carried out efficiently via the EM-based methodology. Synthetic data tests indicate the importance of the coupled error parameter learning in multi-models as opposed to parameter learning from model-data misfits within each model separately. Multi-model fusion via error parameter estimation has been successfully implemented in AOSN-2 real-time forecasting exercises to improve forecast quality.

Acknowledgements

We would like to express our gratitude to Dr. Pierre F.J. Lermusiaux, Division of Engineering and Applied Sciences of Harvard University, for many useful discussions on the subject. We are also thankful to Dr. Patrick Haley and Dr. Wayne Leslie for comments and support with the AOSN-2 experiment data and HOPS forecasting products. AOSN-2 ROMS forecast products have been provided by Dr. Yi Chao and Dr. Jei-Kook Choi from JPL. The work was funded through the ONR grants N00014-02-1-0989 and N00014-97-1-0239.

References

Cormode, G. and Muthukrishnan, S. 2005 An improved data stream summary: The count-min sketch and its applications. *Journal of Algorithms,* **55,** 58–75

Dee, D. P. 1995 Online Estimation of Error Covariance Parameters for Atmospheric Data Assimilation. *Monthly Weather Rev.,* **123,** 4, 1128–1145

Dee, D. P. and da Silva, A. M. 1999 Maximum-Likelihood Estimation of Forecast and Observation Error Covariance Parameters. I: Methodology. *Monthly Weather Rev.,* **127**, 1822–1834

Dempster, A. P., Laird, N. M., Rubin, D. B. 1977 Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc., B,* **39,** 1–38

Doblas-Reyes, F. J., Deque M., Piedelievre J. P. 2000 Model and multi-model spread and probabilistic seasonal forecasts in PROVOST. *Q. J. R. Meteorol. Soc.,* **126,** 2069–2088

Gaspari, G., and Cohn, S. E 1999 Construction of correlation functions in two and three dimensions. *Quart. J. Roy. Meteor. Soc.,* **125,** 723–757

Huang, X. Y. 2000 Variational analysis using spatial filters. *Mon. Wea. Rev.,* **128,** 2588-2600

Kharin, V. V., Zwiers, F. W. 2002 Climate Predictions with Multimodel Ensembles. *Journal of Climate,* **15,** 793-799

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R.,Zhang, Z., Williford, C. E., Gadgil, S. and Surendran, S. 1999 Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550

Lermusiaux, P. F. J. and Robinson, A. R. 1999 Data assimilation via error subspace statistical estimation. Part I: Theory and schemes. *Monthly Weather Rev.,* **127,** 7, 1385–1407

Purser, R. J. and Parrish, D. F. 2003 A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation. *Meteorology and Atmospheric Physics.,* **82,** 209 – 226

Redner R. A. and Walker H. F. 1984 Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review.,* **26**(2), 195–239

Robinson, A. R. and Lermusiaux, P. F. J. 2002 Data Assimilation for Modeling and Predicting Coupled Physical-Biological Interactions in the Sea. Pp. 475-536 in *THE SEA:Biological-Physical Interactions in the Sea.,* **Vol. 12,** Eds. Robinson, A. R., J. J. McCarthy and B. J. Rothschild, John Wiley and Sons, NY.

Robinson, A. R. 1999 Forecasting and Simulating Coastal Ocean Processes and Variabilities with the Harvard Ocean Prediction System in Coastal Ocean Prediction. Pp. 77-100 in *AGU Coastal and Estuarine Studies Series.* Ed. C. N. K. Mooers, American Geophysical Union.

Shchepetkin, A. F. and McWilliams, J. C. 2005 The Regional Oceanic Modeling System: A split-explicit, free-surface, topography-following-coordinate ocean model. *Ocean Modelling,* **9**(4), 347-404

TABLE 1. Normalized* RMS error

	via EM formalism			from model-data misfits		
Ratio**	σ	L	$\mathbf{x}_a$	σ	L	$\mathbf{x}_a$
0.10	0.10	0.18	1.05	0.17	0.21	1.1
0.25	0.11	0.20	1.10	0.18	0.23	1.2
0.50	0.14	0.23	1.25	0.21	0.27	1.4
0.75	0.16	0.25	1.35	0.25	0.32	1.7
1.00	0.17	0.26	1.40	0.30	0.38	2.1

* by true parameter values in case of σ, L and by the rms of the "true" analysis for $\mathbf{x}_a$. ** Ratio $= \frac{rms_o}{rms_1+rms_2}$

TABLE 2. SST Forecast Skill

	HOPS	ROMS	CNTR$_{EM}$*	CNTR$_{\mathbf{d}}$**
rms	1.2	1.6	0.9	1.1
cor^o	0.75	0.68	0.83	0.80

* Central forecast based on error parameter estimation via EM, ** error parameter estimation from model-data misfits, ^{o} correlation.

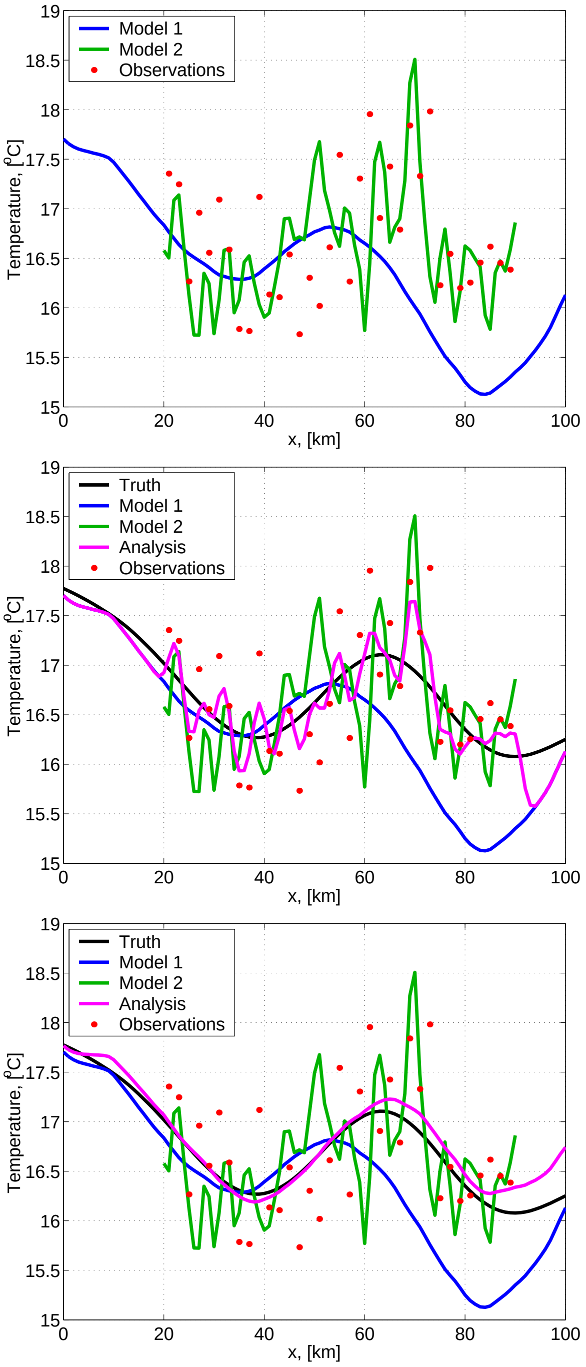

Figure 1. Synthetic data test illustration. (Upper panel): two model forecasts and validating observations; (Middle panel): analysis based on error parameter estimation from model-data misfits (Bottom panel): analysis based on multi-model error parameter estimation described in this paper.

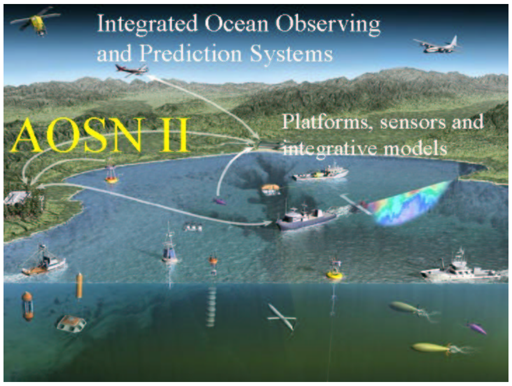

Figure 2. AOSN-2 experiment schematic (courtesy of Dr. James Bellingham, Monterey Bay Aquarium Research Institute).

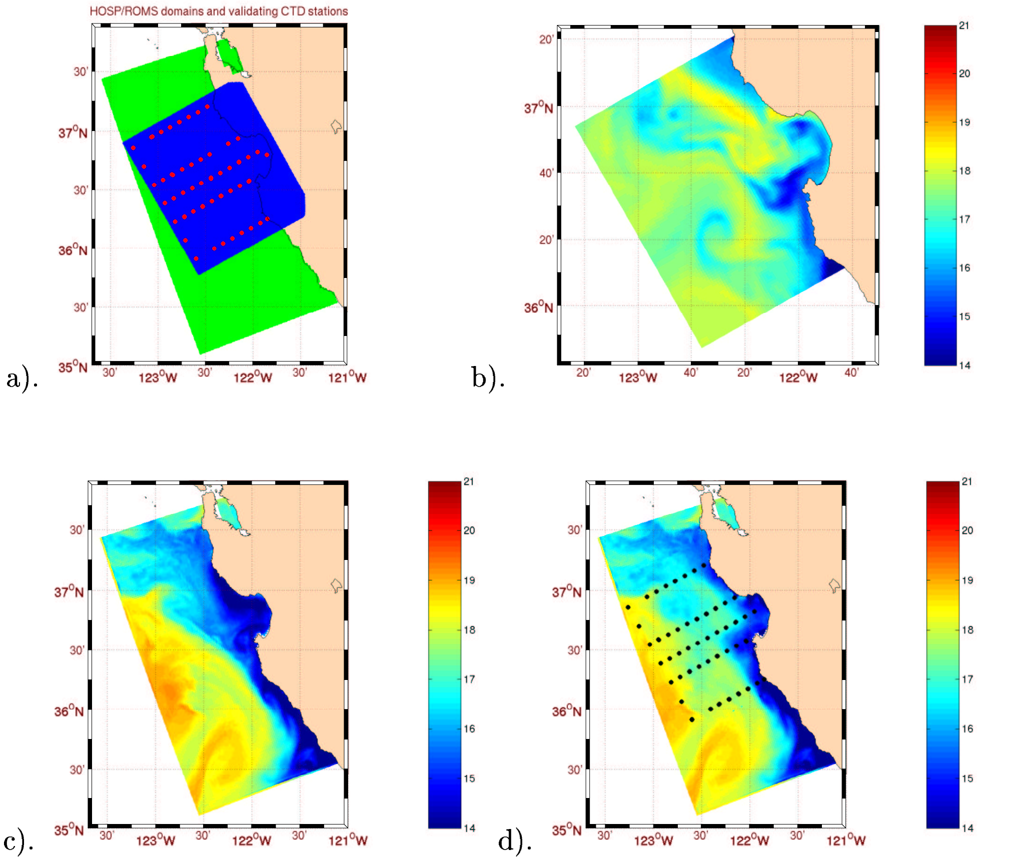

Figure 3. (Panel a.) HOPS (blue) and ROMS (green) domains, and 1st survey CTD stations; (Panel b.) HOPS SST forecast; (Panel c.) ROMS SST forecast; (Panel d.) Bayesian fusion of HOPS/ROMS forecasts, and validating CTD stations (2nd survey).