Q. J. R. Meteorol. Soc. (2005), **131**, pp. 3397–3408 doi: 10.1256/qj.05.99

Multi-model fusion and error parameter estimation

By O. G. LOGUTOV* and A. R. ROBINSON
Division of Engineering and Applied Sciences, Harvard University, Cambridge, USA

(Received 23 May 2005; revised 3 January 2006)

SUMMARY

A robust and practical methodology for multi-model ocean forecast fusion has been sought. Present regional ocean forecasting systems adapt and evolve in response to modelled processes. This makes it imperative that a forecast combination methodology be adaptive and capable to operate with a small sample of past validating events. To this end, we consider an extension of maximum-likelihood error parameter estimation to multi-model predictive systems, and utilize the resulting methodology for adaptive Bayesian model fusion. The methodology consists of the following three general steps: (a) parametrization of forecast uncertainties through either a suitable parametric family (e.g. covariance models) or through a low-rank approximation (e.g. flow-dependent error subspaces); (b) update of uncertainty parameters via maximum likelihood; and (c) combining model forecasts based on their uncertainty parameters via maximum likelihood. In order to implement step (b), we have extended the maximum-likelihood error parameter estimation methodology to multi-model forecasting systems using the expectation-maximization technique, with the true state-space vector at observation locations treated as missing data. With only one forecasting model, the procedure reduces to the standard maximum-likelihood error parameter estimation. The proposed multi-model fusion methodology neglects cross-model error correlations in order to gain the capability to work with a small sample of past events. We illustrate the methodology with a two-model forecasting example (HOPS, ROMS) within the framework of the real-time forecasting experiment held in Monterey Bay during 2003.

KEYWORDS: Adaptive methods Data assimilation Ocean and atmospheric forecasting

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.* 1999; Doblas-Reyes *et al.* 2000; Kharin and Zwiers 2002) and require a substantial training dataset. On the other hand, adaptive modelling has been moving steadily into the mainstream of many real-world forecasting applications, particularly in the area of regional ocean forecasting (Robinson 1999; Robinson *et al*. 2002). On regional scales, dominant ocean processes and conditions can undergo significant changes over time and adaptiveness is increasingly recognized as the key determinant of a successful modelling system (Lermusiaux 1999a; Robinson and Lermusiaux 2004). Because such forecasting systems must adapt and evolve in response to modelled processes, 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, which presents a difficulty in the use of multi-models in regional ocean applications. In this paper, we work around this limitation by treating the optimal model combination as a non-stationary problem that calls for a methodology that operates with a small sample of past validating events.

In regional ocean forecasting, observational data typically come in batches, with a certain spatial coverage, and each batch of data is treated as a single validating event. Within new regional oceanographic applications, as soon as several validating events become available, a change to a forecasting system is often made to correct for deficiencies exemplified in the validating data (e.g. change of model parameters, parametrization schemes, model grids, etc.). As a result, only a few (often one or two) batches

* Corresponding author: Pierce Hall, 29 Oxford Street, Cambridge, MA 02138, USA.
e-mail: oleg@pacific.harvard.edu

of spatially distributed measurements are available as training data for the purposes of model combination. In our own practice with the Harvard Ocean Prediction System (HOPS) for regional synoptic forecasting including oceanic (sub)mesoscale variability (Robinson 1999; Robinson *et al*. 2002), a training dataset typically has sufficient spatial coverage to infer the spatial parameters of forecast errors, but insufficient to infer the cross-model structure of forecast-error covariances. On the other hand, a high variability across multiple scales is distinctive of the regional ocean applications (Lermusiaux 1999b). One kind of synoptic phenomenon is often followed by a very different kind of dynamical process in the time span between validation surveys. The error patterns in forecasts are, therefore, also highly variable. Under these settings, although the forecast errors from different models might indeed be correlated, accounting for cross-model error correlations is risky and generally counterproductive because of over-fitting and due to the fact that different dynamical events generally result in different error patterns in forecasts. In this paper, we advocate a Bayesian model fusion approach that treats forecast errors from different models as uncorrelated in order to gain its capability to work with a small sample of past validating events, while still accounting for spatial structure in forecast-error covariances.

The proposed methodology consists of the following three general steps: (a) parametrization of forecast uncertainties through either a suitable parametric family (covariance models) or through a low-rank approximation (dominant error subspaces); (b) update of forecast uncertainty parameters via maximum likelihood (ML) (section 2); and (c) combining model forecasts based on their uncertainty parameters via ML (section 3). The methodology is amenable to an ensemble-based representation of error covariances in individual models. The forecasting models are allowed to be defined on different grids.

The rest of the paper is organized as follows. In section 2, we describe the multi-model error parameter estimation based on ML. The core part of the construction is the expectation–maximization technique. In section 3, we apply this developed methodology for multi-model Bayesian fusion. Finally, we illustrate it on the example of two-model real-time ocean forecasting exercises held in the Monterey Bay/California Current system in August 2003.

2. Multi-model error parameter estimation

(*a*) *Set-up, notation, motivation*

Suppose a multi-model ocean/atmospheric predictive system consists of m models that produce independent forecasts, $\{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_m^k\}_{k=1}^K$, valid at times $\{t_k\}_{k=1}^K$, with the corresponding forecast errors $\{\boldsymbol{\epsilon}_1^k, \boldsymbol{\epsilon}_2^k, \ldots, \boldsymbol{\epsilon}_m^k\}_{k=1}^K$. Suppose also that validating measurements, $\{\mathbf{y}^k\}_{k=1}^K$, with error $\{\boldsymbol{\epsilon}_o^k\}_{k=1}^K$, become available. Each $\mathbf{x}_i^k$ in the above designates a grid-space vector of forecast values, $\mathbf{x}_i^k \in \mathbb{R}^{n_i}$, corresponding to model i, and index k designates a validating event. Each vector $\mathbf{y}^k \in \mathbb{R}^{p_k}$ denotes a batch of spatially distributed validating measurements taken at time t_k. Measurements $\mathbf{y}^k$ and $\mathbf{y}^l$ might correspond to different locations, and forecasts $\mathbf{x}_i$ and $\mathbf{x}_j$ are allowed to be defined on different grids. The total number of validating events, K, is assumed to be small, possibly one, as explained in the introduction. We pose ourselves the problem of finding the optimal strategy for combining model forecasts, $\mathbf{x}_1^{K+1}, \mathbf{x}_2^{K+1}, \ldots, \mathbf{x}_m^{K+1}$, to generate the central forecast, $\mathbf{x}^{K+1} \in \mathbb{R}^n$, the next time a prediction is being made.

Let $\mathbf{x} \in \mathbb{R}^n$ denote the grid-space vector of forecast values on a central forecast grid. A central forecast grid can be chosen as one of the model grids or defined separately.

The ML fusion of $\mathbf{x}_1^{K+1}, \mathbf{x}_2^{K+1}, \ldots, \mathbf{x}_m^{K+1}$ corresponds to maximizing the posterior probability,

$$\widehat{\mathbf{x}}^{K+1} = \arg\max_{\mathbf{x}} \{p(\mathbf{x} \mid \mathbf{x}_1^{K+1}, \mathbf{x}_2^{K+1}, \ldots, \mathbf{x}_m^{K+1}, \mathcal{D})\}, \tag{1}$$

where $\mathcal{D} = \{\mathbf{x}_1^k, \mathbf{x}_2^k, \ldots, \mathbf{x}_m^k, \mathbf{y}^k\}_{k=1}^K$ denotes a set of all past validating data within the time window K. The posterior probability (1) typically involves a set of unknown hyper-parameters, $\boldsymbol{\Theta}$, such as forecast-error covariance parameters. We simplify (1) by making a Markov process assumption and finding the hyper-parameters, $\boldsymbol{\Theta}$, from the validating data $\mathcal{D}$ based on ML,

$$\widehat{\mathbf{x}}^{K+1} = \arg\max_{\mathbf{x}} \{p(\mathbf{x} \mid \mathbf{x}_1^{K+1}, \mathbf{x}_2^{K+1}, \ldots, \mathbf{x}_m^{K+1}, \widehat{\boldsymbol{\Theta}}^*)\}, \tag{2}$$

$$\text{with } \widehat{\boldsymbol{\Theta}}^* = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta} \mid \mathcal{D}). \tag{3}$$

We assume unbiased Gaussian forecast errors, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{B}_i)$, $\mathbf{B}_i = \langle \boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^{\mathrm{T}} \rangle$. The parametrization of forecast uncertainties is needed to reduce the number of degrees of freedom in the system in order to ensure robust parameter estimation from the data available within the time window K. We assume that the forecast-error covariances, $\{\mathbf{B}_i\}_{i=1}^m$, are parametrized in some sensible way, $\mathbf{B}_i \approx \widehat{\mathbf{B}}_i(\boldsymbol{\alpha}_i)$, and denote the vector of all error covariances parameters of model i as $\boldsymbol{\alpha}_i \in \mathbb{R}^{q_i}$. The collection of the error parameters in all m models we formally denote as $\boldsymbol{\Theta} \equiv \{\boldsymbol{\alpha}_i\}_{i=1}^m$.

The exact way of parametrizing the $\mathbf{B}_i$ remains at the discretion of a researcher or is determined by the legacy of a data assimilation system. One popular approach to parametrization is the use of covariance models (e.g. Gaspari and Cohn 1999). In this case $\boldsymbol{\alpha}_i$ could involve decorrelation length-scales and variances, among other parameters. An example of $\boldsymbol{\alpha}_i$ is given in section 2(d). A possible alternative to covariance models is the use of a low-rank approximation (e.g. Lermusiaux 1999a), $\mathbf{B}_i \approx \mathbf{U}_i \boldsymbol{\Lambda}_i(\boldsymbol{\alpha}_i)\mathbf{U}_i^{\mathrm{T}}$, where $\mathbf{U}_i$ is a low-rank orthogonal basis of the dominant forecast-error subspace. The error basis $\mathbf{U}_i$ can be found from a forecast ensemble spread using the methodology described in Lermusiaux and Robinson (1999). The diagonal matrix of singular values, $\boldsymbol{\Lambda}_i(\boldsymbol{\alpha}_i)$, can be parametrized, with parameters $\boldsymbol{\alpha}_i \in \mathbb{R}^{q_i}$. The simplest choice of $\boldsymbol{\alpha}_i$ are the singular values themselves treated as free parameters. Regardless of the form of forecast-error covariance parametrization in model i, we refer to it as $\widehat{\mathbf{B}}_i(\boldsymbol{\alpha}_i)$ hereafter.

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

The ML error parameters are found through maximizing the probability of observing data $\mathcal{D}$ given error parameters, $p(\mathcal{D} \mid \boldsymbol{\Theta})$. Since $\mathcal{D}$ is given and hence fixed, the probability $p(\mathcal{D} \mid \boldsymbol{\Theta})$ is a deterministic function of parameters $\boldsymbol{\Theta}$, called the likelihood function, $\mathcal{L}$. ML error parameter estimation amounts to maximizing the likelihood function, $\mathcal{L}$,

$$\widehat{\boldsymbol{\Theta}}^* = \arg\max_{\boldsymbol{\Theta}} p(\mathcal{D} \mid \boldsymbol{\Theta}) = \arg\max_{\boldsymbol{\Theta}} \mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}). \tag{4}$$

We use $\boldsymbol{\Theta}$, as before, to formally denote the collection of all error parameters in all m models, $\boldsymbol{\Theta} \equiv \{\boldsymbol{\alpha}_i\}_{i=1}^m$. With only one forecasting model, all the relevant information contained in the validating data $\mathcal{D} = \{\mathbf{x}_1^k, \mathbf{y}^k\}_{k=1}^K$ can be fully expressed via model–data misfits, $\mathbf{d}_k = \mathbf{y}^k - \mathbf{H}_1 \mathbf{x}_1^k$. Gaussian error assumption leads to the standard ML error parameter estimation from model–data misfits (Dee 1995; Dee and da Silva 1999; Purser and Parrish 2003). Because the sum of Gaussian errors, $\mathbf{H}_1 \boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_o$ (where $\mathbf{H}_1$ is linear mapping from model state-space onto the observational space), is also

distributed as a Gaussian with covariance $\langle \mathbf{H}_1 \boldsymbol{\epsilon}_1 \boldsymbol{\epsilon}_1^{\mathrm{T}} \mathbf{H}_1^{\mathrm{T}} \rangle + \langle \boldsymbol{\epsilon}_o \boldsymbol{\epsilon}_o^{\mathrm{T}} \rangle$, the model–data misfits, $\mathbf{d}_k = \mathbf{y}^k - \mathbf{H}_1 \mathbf{x}_1^k$, are random samples from the normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{Q}_1(\boldsymbol{\alpha}_1))$, where $\mathbf{Q}_1(\boldsymbol{\alpha}_1)$ is the sum of the forecast- and the observation-error covariances, $\mathbf{Q}_1(\boldsymbol{\alpha}_1) = \mathbf{H}_1 \mathbf{B}_1(\boldsymbol{\alpha}_1) \mathbf{H}_1^{\mathrm{T}} + \mathbf{R}$. Therefore, the likelihood function $\mathcal{L}(\boldsymbol{\alpha} \mid \mathcal{D})$ in (4) is readily expressed as

$$\log \mathcal{L}(\boldsymbol{\alpha} \mid \mathcal{D}) \propto (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{\alpha} - \boldsymbol{\alpha}_0) + K \log \det \mathbf{Q}(\boldsymbol{\alpha}) + \sum_{k=1}^{K} \mathbf{d}_k^{\mathrm{T}} \mathbf{Q}^{-1}(\boldsymbol{\alpha}) \mathbf{d}_k. \quad (5)$$

ML parameter estimation in a one-model system amounts to minimizing (5) with respect to $\boldsymbol{\alpha}$. The first term in (5) describes prior information on parameter values, $\mathcal{N}(\boldsymbol{\alpha}_0, \boldsymbol{\Sigma})$, and can be omitted if no such prior information is available (Dee 1995; Purser and Parrish 2003).

In a multi-model system, the likelihood function $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D})$ is not easily expressed because the relevant information contained in $\mathcal{D}$ cannot be fully represented via model–data misfits. Model–model differences are another source of information about forecast errors that complement 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 differences and results in suboptimal error parameter estimation. The amount of information contained in the model–model differences depends on the accuracy of the models, as well as on the degree of independence of their errors. Synthetic data tests presented later in this section demonstrate that ignoring the information contained in model–model differences leads to significant degradation of error parameter estimates (Table 1). In the next subsection we present a method that solves the multi-model error parameter estimation problem in a way that is both practical and easy to implement.

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

Our method of solving (4) in the multi-model settings rests on the observation that the problem is amenable to a missing data interpretation. The likelihood function $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D})$ is not easily expressed in the multi-model case. However, the multi-model error parameter estimation becomes straightforward if we know the values of the true state of the ocean/atmosphere at observation locations. In order to take advantage of that fact, we augment the data, $\mathcal{Y} = \{\mathcal{D}, \mathcal{Z}\}$, where $\mathcal{Z} = \{\mathbf{z}^k\}_{k=1}^K$ is the true ocean/atmosphere state at observation locations ($\mathbf{z}^k \in \mathbb{R}^{p_k}$), and employ the *expectation–maximization* (EM) technique to replace $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D})$ with $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}, \mathcal{Z})$ for the purposes of error parameter estimation. We call $\mathcal{Y}$ the *complete data*. As opposed to $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D})$, the complete-data likelihood $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}, \mathcal{Z})$ is readily expressed. However, $\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}, \mathcal{Z})$ is not a deterministic function of parameters $\boldsymbol{\Theta}$, since $\mathcal{Z}$ is unknown and therefore must be treated as random, with some underlying probability distribution, $p(\mathcal{Z} \mid \mathcal{D}, \boldsymbol{\Theta})$. The expectation, $\mathcal{E}_{\mathcal{Z}|\mathcal{D},\boldsymbol{\Theta}}\{\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}, \mathcal{Z})\}$, of the complete-data likelihood with respect to missing data $\mathcal{Z}$, on the other hand, is a deterministic function of parameters $\boldsymbol{\Theta}$ ($\mathcal{E}_{\mathcal{Z}|\mathcal{D},\boldsymbol{\Theta}}\{\cdot\} \equiv \int_{\mathcal{Z}} \{\cdot\} p(\mathcal{Z} \mid \mathcal{D}, \boldsymbol{\Theta})\, \mathrm{d}\mathcal{Z}$). The EM method amounts to replacing the maximization of $\log \mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D})$ by maximization of $\mathcal{E}_{\mathcal{Z}|\mathcal{D},\boldsymbol{\Theta}}\{\log \mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}, \mathcal{Z})\}$ as this substitution leads to the same parameter values (Dempster *et al.* 1977; Redner and Walker 1984).

The parameter values, $\boldsymbol{\Theta}$, are unknown for the purposes of evaluating $p(\mathcal{Z} \mid \mathcal{D}, \boldsymbol{\Theta})$ when taking the expectation $\mathcal{E}_{\mathcal{Z}|\mathcal{D},\boldsymbol{\Theta}}\{\cdot\}$. The EM methodology overcomes that difficulty through an iterative succession of *expectation* (E) and *maximization* (M) steps.

At an E-step, the expectation of the complete-data log-likelihood is computed based on the current estimate of parameter values, $Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}) \equiv \mathcal{E}_{\mathcal{Z} \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}}\{\log \mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{D}, \mathcal{Z})\}$. At an M-step, $Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}})$ is maximized as a function of $\boldsymbol{\Theta}$ to improve upon $\widehat{\boldsymbol{\Theta}}^{\text{iter}}$.

$$\left.\begin{aligned} \text{E}: \quad & Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}) = \int_{\mathcal{Z}} \log\{\mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{Z}, \mathcal{D})\} p(\mathcal{Z} \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}})\, \mathrm{d}\mathcal{Z} \\ \text{M}: \quad & \widehat{\boldsymbol{\Theta}}^{\text{iter}+1} = \arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}). \end{aligned}\right\} \tag{6}$$

The EM algorithm has global convergence, with linear or better rate (Dempster *et al.* 1977). Under the Gaussian error assumption, the complete-data log-likelihood is expressed as

$$\log \mathcal{L}(\boldsymbol{\Theta} \mid \mathcal{Z}, \mathcal{D}) \propto K \sum_{i=1}^{m} \log\det \widetilde{\mathbf{B}}_i(\boldsymbol{\alpha}_i) + \sum_{k=1}^{K} \sum_{i=1}^{m} (\mathbf{z}^k - \mathbf{H}_i \mathbf{x}_i^k)^{\mathrm{T}} \widetilde{\mathbf{B}}_i^{-1}(\boldsymbol{\alpha}_i)(\mathbf{z}^k - \mathbf{H}_i \mathbf{x}_i^k), \tag{7}$$

where $\mathbf{H}_i$ denotes linear mapping from the ith model state-space onto the observational space, and $\widetilde{\mathbf{B}}_i = \mathbf{H}_i \mathbf{B}_i \mathbf{H}_i^{\mathrm{T}}$. The prior term $(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)^{\mathrm{T}} \boldsymbol{\Sigma}_{\boldsymbol{\Theta}}^{-1} (\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)$ can also be included in (7) in the same way as in (5) if prior information on $\boldsymbol{\Theta}$ is available. Assuming the independence of validating events in $\mathcal{D}$, the marginal probability density $p(\mathcal{Z} \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}})$ expands as

$$p(\mathcal{Z} \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}) = \prod_{k+1}^{K} p(\mathbf{z}^k \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}), \tag{8}$$

where each probability density function, $p(\mathbf{z}^k \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}})$, is the normal distribution,

$$p(\mathbf{z}^k \mid \mathcal{D}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}) = \mathcal{N}\{{}^{\text{iter}}\mathbf{z}_{\text{a}}^k, \widetilde{\mathbf{B}}_{\text{a}}(\widehat{\boldsymbol{\Theta}}^{\text{iter}})\}. \tag{9}$$

The vector ${}^{\text{iter}}\mathbf{z}_{\text{a}}^k \in \mathbb{R}^{p_k}$ is the analysis at observation locations and $\widetilde{\mathbf{B}}_{\text{a}}(\widehat{\boldsymbol{\Theta}}^{\text{iter}})$ is the analysis-error covariance, both evaluated with parameter values $\widehat{\boldsymbol{\Theta}}^{\text{iter}}$:

$${}^{\text{iter}}\mathbf{z}_{\text{a}}^k = \arg\min_{\mathbf{z}} \sum_{i=1}^{m} (\mathbf{z} - \mathbf{H}_i \mathbf{x}_i^k)^{\mathrm{T}} \widetilde{\mathbf{B}}_i^{-1}(\widehat{\boldsymbol{\alpha}}_i^{\text{iter}})(\mathbf{z} - \mathbf{H}_i \mathbf{x}_i^k) + (\mathbf{z} - \mathbf{y}^k)^{\mathrm{T}} \mathbf{R}^{-1} (\mathbf{z} - \mathbf{y}^k) \tag{10}$$

$$\widetilde{\mathbf{B}}_{\text{a}}^{-1}(\widehat{\boldsymbol{\Theta}}^{\text{iter}}) = \widetilde{\mathbf{B}}_1^{-1}(\widehat{\boldsymbol{\alpha}}_1^{\text{iter}}) + \widetilde{\mathbf{B}}_2^{-1}(\widehat{\boldsymbol{\alpha}}_2^{\text{iter}}) + \cdots + \widetilde{\mathbf{B}}_m^{-1}(\widehat{\boldsymbol{\alpha}}_m^{\text{iter}}) + \mathbf{R}^{-1}. \tag{11}$$

An alternative to (10) is the closed-form expression

$$\mathbf{z}_{\text{a}}^k = \widetilde{\mathbf{B}}_{\text{a}} \mathbf{H}_1^{\mathrm{T}} \mathbf{B}_1^{-1} \mathbf{x}_1^k + \widetilde{\mathbf{B}}_{\text{a}} \mathbf{H}_2^{\mathrm{T}} \mathbf{B}_2^{-1} \mathbf{x}_2^k + \cdots + \widetilde{\mathbf{B}}_{\text{a}} \mathbf{H}_m^{\mathrm{T}} \mathbf{B}_m^{-1} \mathbf{x}_m^k + \widetilde{\mathbf{B}}_{\text{a}} \mathbf{R}^{-1} \mathbf{y}^k. \tag{12}$$

By using the Gaussian integral identity for an expectation of a quadratic form,

$$\int_{\mathbf{z}} (\mathbf{z} - \mathbf{H}_i \mathbf{x}_i^k)^{\mathrm{T}} \widetilde{\mathbf{B}}_i^{-1} (\mathbf{z} - \mathbf{H}_i \mathbf{x}_i^k) \mathcal{N}(\mathbf{z}; \mathbf{z}_{\text{a}}^k, \widetilde{\mathbf{B}}_{\text{a}})\, \mathrm{d}\mathbf{z} = (\mathbf{z}_{\text{a}}^k - \mathbf{H}_i \mathbf{x}_i^k)^{\mathrm{T}} \widetilde{\mathbf{B}}_i^{-1} (\mathbf{z}_{\text{a}}^k - \mathbf{H}_i \mathbf{x}_i^k) + \mathrm{Tr}(\widetilde{\mathbf{B}}_i^{-1} \widetilde{\mathbf{B}}_{\text{a}}),$$

for every term in (7), and keeping the terms dependent on $\boldsymbol{\Theta}$, we obtain a closed-form expression for the expectation of complete-data log-likelihood:

$$Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}) \propto K \sum_{i=1}^{m} \log\det \widetilde{\mathbf{B}}_i(\boldsymbol{\alpha}_i) + \sum_{k=1}^{K}\sum_{i=1}^{m} ({}^{\text{iter}}\widehat{\mathbf{z}}_{\text{a}}^k - \mathbf{H}_i\mathbf{x}_i^k)^{\text{T}}\widetilde{\mathbf{B}}_i^{-1}(\boldsymbol{\alpha}_i)({}^{\text{iter}}\widehat{\mathbf{z}}_{\text{a}}^k - \mathbf{H}_i\mathbf{x}_i^k). \quad (13)$$

The EM procedure simplifies to a sequence of iterative updates to the analysis, ${}^{\text{iter}}\widehat{\mathbf{z}}_{\text{a}}^k$, based on the current error parameter estimates, $\widehat{\boldsymbol{\Theta}}^{\text{iter}}$, and the subsequent updates to the error parameter values from model–analysis misfits:

$${}^{\text{iter}}\widehat{\mathbf{z}}_{\text{a}}^k = \arg\min_{\mathbf{z}} \sum_{i=1}^{m} (\mathbf{z} - \mathbf{H}_i\mathbf{x}_i^k)^{\text{T}}\widetilde{\mathbf{B}}_i^{-1}(\widehat{\boldsymbol{\alpha}}_i^{\text{iter}})(\mathbf{z} - \mathbf{H}_i\mathbf{x}_i^k) + (\mathbf{z} - \mathbf{y}^k)^{\text{T}}\mathbf{R}^{-1}(\mathbf{z} - \mathbf{y}^k) \quad (14a)$$

$$\widehat{\boldsymbol{\Theta}}^{\text{iter}+1} = \arg\min_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}), \quad \text{with } Q(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{\text{iter}}) \text{ given in (13).} \quad (14b)$$

Because we have assumed that forecast errors from different models are uncorrelated, error parameter estimation (14b) can be carried out independently in the individual modelling systems using their own error parameter estimation procedures. Only vectors ${}^{\text{iter}}\widehat{\mathbf{z}}_{\text{a}}^k$ have to be communicated between the models. Thus, the methodology is easy to implement and carries little overhead in terms of new software development.

(*d*) *Synthetic data tests*

The EM-based multi-model error parameter estimation procedure described in the previous subsection improves an estimate of forecast-error parameters based upon model–data misfits by the inclusion of additional information contained in model–model differences. Synthetic data tests in which the 'truth' and the 'true' error parameter values are generated (and hence known) can be carried out to examine the improvement attributed to the method for different uncertainty parameters. Figure 1 illustrates such an experiment, and Table 1 summarizes the statistics. Two 'models' are drawn independently at random from a normal distribution around a synthetic 'truth', with chosen error covariance parameters. An isotropic covariance model, $\mathbf{B}(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2) = \sigma^2\rho(||\,\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\,||)$, with the fifth-order piecewise rational representing function, $\rho(r) = \rho_{\text{c}}(r, L)$, 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 a well-conditioned $\mathbf{B}_i(\boldsymbol{\alpha}_i)$ for the whole range of parameter values. The vector $\boldsymbol{\alpha}_i$ in this case consists of only two parameters, $\boldsymbol{\alpha}_i = (\sigma_i, L_i)^{\text{T}}$. The error length-scale parameters, L_i, in the models have been intentionally chosen to be very different. Synthetic 'observations' have also been drawn at random from a Gaussian distribution, with diagonal covariance $\mathbf{R}$. Figure 1(a) illustrates the data on which the multi-model error parameter estimation procedure is run. Figures 1(b) and (c) show the analyses computed with error parameters obtained from model–data misfits and through the EM procedure described in this paper, respectively. The experiment illustrated in Fig. 1 was repeated 100 times for each uncertainty level attributed to the observations (i.e. different ratio values in Table 1) to collect statistics. Table 1 illustrates that, given a sufficient level of uncertainty in the measurements as compared to uncertainty in the models, the information contained in the model–model differences leads to a significant improvement of the error parameter estimates and of the multi-model analysis.

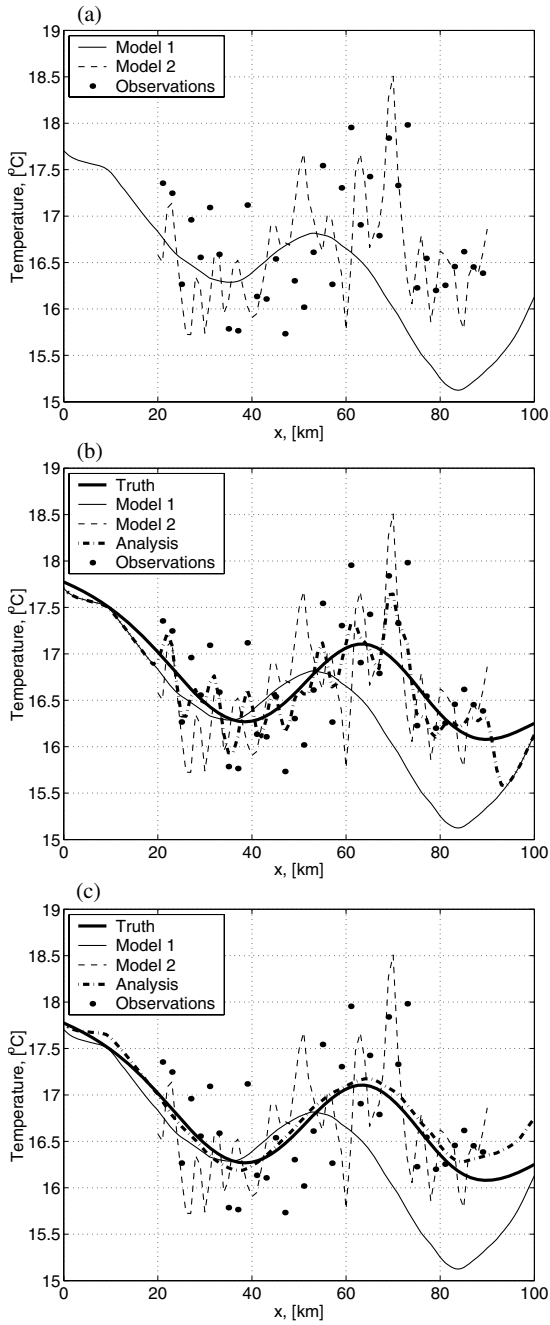

Figure 1. Synthetic data test illustration: (a) two model forecasts and validating observations, (b) analysis based on error parameter estimation from model–data misfits, and (c) analysis based on multi-model error parameter estimation described in this paper.

TABLE 1. NORMALIZED* r.m.s. ERROR OF ESTIMATES $\widehat{\sigma}$, $\widehat{L}$ AND $\widehat{\mathbf{x}}_{\mathrm{a}}^{\dagger}$

	Via EM formalism			From model–data misfits		
Ratio‡	$\widehat{\sigma}$	$\widehat{L}$	$\widehat{\mathbf{x}}_{\mathrm{a}}$	$\widehat{\sigma}$	$\widehat{L}$	$\widehat{\mathbf{x}}_{\mathrm{a}}$
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 $\widehat{\sigma}$ and $\widehat{L}$, and by the r.m.s. of the 'true' analysis (obtained using true error parameters) in case of $\widehat{\mathbf{x}}_{\mathrm{a}}$.
‡ Ratio $\equiv rms_{\mathrm{o}}/(rms_1 + rms_2)$, where rms_{o}, rms_1, and rms_2 are the r.m.s. errors of observations, first and second models, respectively.
† See text for definitions.

3. MAXIMUM-LIKELIHOOD FORECAST FUSION

Once the multi-model error parameters are estimated through the EM procedure (14), we combine the models based on their relative uncertainties via the ML 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 the individual forecast p.d.f.s (we drop the index k hereafter as all the vectors correspond to t_{K+1}),

$$p(\mathbf{x} \mid \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) \propto p(\mathbf{x} \mid \mathbf{x}_1)\, p(\mathbf{x} \mid \mathbf{x}_2) \ldots p(\mathbf{x} \mid \mathbf{x}_m). \tag{15}$$

The independence of multi-model forecasts can be questioned since the individual models have qualitatively similar model errors and typically use the same set of measurements for initialization. We advocate (15) for the reasons explained in the introduction—with only a few validating events, an attempt to account for cross-model correlation structure amounts to over-fitting and is counterproductive.

Under the Gaussian error assumption, the individual p.d.f.s $p(\mathbf{x} \mid \mathbf{x}_i)$ in (15) are normal distributions $\mathcal{N}\{\mathbf{H}_i^{\mathrm{c}}\mathbf{x}_i, \boldsymbol{\mathcal{B}}_i(\widehat{\boldsymbol{\alpha}}_i)\}$, where $\mathbf{H}_i^{\mathrm{c}}$ denotes linear mapping from the ith model state-space onto the central forecast state-space and $\boldsymbol{\mathcal{B}}_i(\widehat{\boldsymbol{\alpha}}_i) = \mathbf{H}_i^{\mathrm{c}}\mathbf{B}_i(\widehat{\boldsymbol{\alpha}}_i)\mathbf{H}_i^{\mathrm{cT}}$. The ML (minimum variance) central forecast, $\mathbf{x}_{\mathrm{c}}$, corresponding to (2) is found as

$$\mathbf{x}_{\mathrm{c}} = \arg\min_{\mathbf{x}} \sum_{i=1}^{m} (\mathbf{x} - \mathbf{H}_i^{\mathrm{c}}\mathbf{x}_i)^{\mathrm{T}} \boldsymbol{\mathcal{B}}_i^{-1}(\widehat{\boldsymbol{\alpha}}_i)(\mathbf{x} - \mathbf{H}_i^{\mathrm{c}}\mathbf{x}_i), \tag{16}$$

or using a closed-form solution,

$$\mathbf{x}_{\mathrm{c}} = \widehat{\mathbf{B}}_{\mathrm{c}}\mathbf{H}_1^{\mathrm{T}}\widehat{\mathbf{B}}_1^{-1}\mathbf{x}_1 + \widehat{\mathbf{B}}_{\mathrm{c}}\mathbf{H}_2^{\mathrm{T}}\widehat{\mathbf{B}}_2^{-1}\mathbf{x}_2 + \cdots + \widehat{\mathbf{B}}_{\mathrm{c}}\mathbf{H}_m^{\mathrm{T}}\widehat{\mathbf{B}}_m^{-1}\mathbf{x}_m, \tag{17}$$

where the central forecast-error covariance, $\widehat{\mathbf{B}}_{\mathrm{c}}$, is expressed as

$$\widehat{\mathbf{B}}_{\mathrm{c}} = (\widehat{\boldsymbol{\mathcal{B}}}_1^{-1} + \widehat{\boldsymbol{\mathcal{B}}}_2^{-1} + \cdots + \widehat{\boldsymbol{\mathcal{B}}}_m^{-1})^{-1}. \tag{18}$$

The above procedure is equivalent to Bayesian Model Averaging (Hoeting *et al.* 1999), where $\mathbf{x}_{\mathrm{c}} = \widehat{\mathbf{C}}_1\mathbf{x}_1 + \widehat{\mathbf{C}}_2\mathbf{x}_2 + \cdots + \widehat{\mathbf{C}}_m\mathbf{x}_m$, with Bayesian weight matrices, $\widehat{\mathbf{C}}_i$, given by $\widehat{\mathbf{B}}_{\mathrm{c}}\mathbf{H}_i^{\mathrm{T}}\widehat{\mathbf{B}}_i^{-1}$. Denote the sum of columns of $\widehat{\mathbf{C}}_i$ as $\widehat{\mathbf{p}}_i$,

$$\widehat{\mathbf{p}}_i = \sum_{n'=1}^{n_i} \widehat{\mathbf{C}}_i(j, n')\ \forall j, \qquad \widehat{\mathbf{C}}_i = \widehat{\mathbf{B}}_c\mathbf{H}_i^{\mathrm{T}}\widehat{\mathbf{B}}_i^{-1}, \tag{19}$$

where j is the central forecast grid point index. Bayesian weights $\mathbf{p}_i$ provide the spatially varying optimal weight of the ith model in the central forecast and satisfy $\mathbf{p}_i(j) > 0\ \forall j$ and $\sum_{i=1}^{m} \mathbf{p}_i = 1$. These weights translate model error parameters in terms of relative model skill. If it is desired that no extra smoothing be introduced through model fusion, the $\mathbf{C}_i(j, :)$ are to be replaced with the $\mathbf{p}_i(j)$ in the overlapping parts of the central and ith model domains, $\mathbf{x}_c^j \in \mathbf{x}_i$, so that the multi-model Bayesian fusion becomes

$$\mathbf{x}_c = \begin{cases} \widehat{\mathbf{p}}_1 \mathbf{x}_1,\ \mathbf{x}_c^j \in \mathbf{x}_1 \\ \widehat{\mathbf{C}}_1 \mathbf{x}_1,\ \mathbf{x}_c^j \notin \mathbf{x}_1 \end{cases} + \begin{cases} \widehat{\mathbf{p}}_2 \mathbf{x}_2,\ \mathbf{x}_c^j \in \mathbf{x}_2 \\ \widehat{\mathbf{C}}_2 \mathbf{x}_2,\ \mathbf{x}_c^j \notin \mathbf{x}_2 \end{cases} + \cdots + \begin{cases} \widehat{\mathbf{p}}_m \mathbf{x}_m,\ \mathbf{x}_c^j \in \mathbf{x}_m \\ \widehat{\mathbf{C}}_m \mathbf{x}_m,\ \mathbf{x}_c^j \notin \mathbf{x}_m. \end{cases} \tag{20}$$

Bayesian model fusion (20) ignores inter-model forecast-error correlations in order to gain its capability to work with a small sample of past validating events in response to the needs discussed in the introduction. In applications with a sufficient amount of validating data for robust estimation of cross-model error covariance parameters, steps could be taken to correct for the inter-model error correlations. One possible approach is given in Julier and Uhlmann (2001) who describe the Covariance Intersection algorithm that seeks to account for the inter-model error correlation structure by introducing and optimizing an additional scalar parameter that specifies a convex combination of the model covariances. If this general approach can be extended to accommodate the high dimensionality of real-world data and models, then it might be useful for applications to atmospheric and oceanographic forecasting systems.

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 2003 and designated as AOSN-2. The Autonomous Ocean Sampling Network (AOSN) project 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, the forecast protocols and technical infrastructure behind an integrated observing and modelling system for regional ocean predictions*. The exercises included two ocean models, the HOPS (Robinson 1999; Robinson and Lermusiaux 2002) and the University of California (Los Angeles) version of the Regional Ocean Modeling System (ROMS; Shchepetkin and McWilliams 2005). These two forecasting models were defined in overlapping but slightly different domains (Fig. 2(a)) and operated independently of each other in real time. The use of two independent models provided the opportunity to consider improving the forecasting by combining the two models.

Our approach to model combination led us to develop and apply the methodology described in this paper. The AOSN-2 program included two conductivity–temperature–depth (CTD) surveys for initialization and assimilation, with nominally similar station locations, the first on 5–7 August, with the stations depicted in Fig. 2(a), and the second on 21–23 August, with the stations depicted in Fig. 2(d). (CTD instruments measure vertical profiles of *in situ* temperature and salinity, and other seawater properties.) 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. The data from the second CTD survey were utilized to evaluate HOPS and ROMS forecast skill as compared to the skill of the combined forecast which we now call the central forecast, CNTR. Figures 2(b) and (c) show

* The full description of the experiment is available at http://www.mbari.org/aosn/

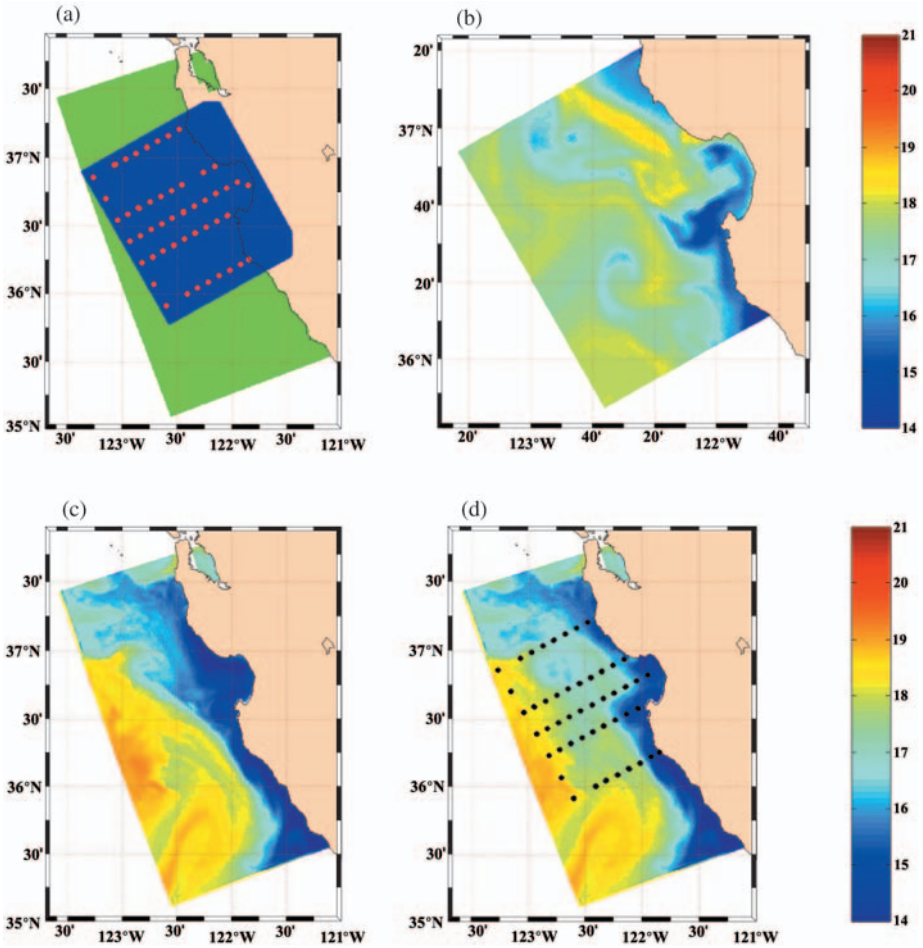

Figure 2. (a) HOPS (blue) and ROMS (green) domains, with dots indicating conductivity–temperature–depth stations (first survey), (b) HOPS and (c) ROMS sea surface temperature (°C) forecasts, and (d) Bayesian fusion of HOPS/ROMS forecasts, with validating CTD stations (second survey).

examples of the individual HOPS and ROMS forecasts. The corresponding CNTR, obtained via Bayesian fusion of HOPS and ROMS, is shown in Fig. 2(d).

Table 2 summarizes some skill metrics, specifically the r.m.s. error and the correlation of forecast with observed SST. These skill metrics have been accumulated for the individual HOPS and ROMS forecasts, and for the central forecasts obtained via Bayesian fusion of HOPS and ROMS using two different error parameter estimation methodologies. The first set of central forecasts, $\mathrm{CNTR}_{\mathrm{EM}}$, was computed based on the EM multi-model error parameter estimation methodology described in this paper. The second set of central forecasts, $\mathrm{CNTR}_{\mathbf{d}}$, was based on the standard ML error parameter estimation from model–data misfits, within each model separately, following the methodology described in Dee (1995) and Purser and Parrish (2003). (**d** represents the innovation vector.) Table 2 shows improvements in both the r.m.s. error statistics and the correlation of forecast with validating measurements for $\mathrm{CNTR}_{\mathrm{EM}}$ compared with $\mathrm{CNTR}_{\mathbf{d}}$. This improvement stems from the inclusion of information contained in the model–model differences.

TABLE 2. SST FORECAST SKILL

	HOPS	ROMS	$CNTR_{EM}$	$CNTR_{\mathbf{d}}$
R.m.s. forecast error (degC)	1.2	1.6	0.9	1.1
Forecast-to-data correlation	0.75	0.68	0.83	0.80

$CNTR_{EM}$: central forecast with the EM error parameter estimation.
$CNTR_{\mathbf{d}}$: central forecast with error parameter estimation as in Dee (1995).

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. Today, in regional ocean forecasting applications, the time-scale for changes to a forecasting system is often shorter than the time it takes to collect a statistically significant sample of validating events. We have outlined a methodology designed to work with a small sample of past validating events. Our method consists of parametrization of forecast uncertainties, multi-model error parameter estimation, and Bayesian fusion of multi-model forecasts based on their estimated uncertainty parameters. The method is not optimal because it does not directly treat the inter-model forecast-error correlations, but it does take into account the information contained in model–model differences. The method is practical and well suited for regional adaptive ocean forecasting and possibly other ocean and atmospheric forecasting applications.

Multi-model error parameter estimation is the essential component of the methodology. Synthetic data tests indicate that model–model differences supply important additional information to forecast-error parameter estimation beyond the information contained in each of the model–data misfits. We have put forward a practical algorithm to account for the information contained in model–model differences via the EM-based scheme. The strength of our algorithm is that it makes full use of the existing error parameter estimation procedures within the individual models and carries little overhead in terms of software development. We have successfully implemented Bayesian multi-model fusion via error parameter estimation in the AOSN-2 real-time forecasting exercises which led to improved forecast quality. In August 2006, HOPS and ROMS will again forecast together in real time in the Monterey Bay area for an adaptive sampling experiment involving Woods Hole and Scripps fleets of gliders and several other observational platforms. The methodology presented here will be utilized in real time.

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 this subject. We are also thankful to Dr Patrick Haley and Mr Wayne Leslie for comments and support with the AOSN-2 experiment data and HOPS forecasting products. The AOSN-2 ROMS forecast products have been provided by Dr Yi Chao and Dr Jei-Kook Choi from JPL. This work was funded through the ONR grants N00014-02-1-0989 and N00014-97-1-0239.

REFERENCES

Dee, D. P. 1995 Online estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.*, **123,** 1128–1145

Dee, D. P. and da Silva, A. M. 1999 Maximum-likelihood estimation of forecast and observation error covariance parameters. I: Methodology. *Mon. Weather Rev.*, **127,** 1822–1834

Dempster, A. P., Laird, N. M. and 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., Déqué, M. and 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. *Q. J. R. Meteorol. Soc.*, **125,** 723–757

Hoeting, J., Madigan, D., Raftery, A. E. and Volinsky, C. 1999 Bayesian model averaging: a tutorial. *Statist. Sci.*, **14**(4)**,** 382–417

Julier, S. J. and Uhlmann, J. K. 2001 Generalized decentralized data fusion using covariance intersections. In *Handbook of multisensor data fusion*. Eds. D. L. Hall and J. Llinas, CRC Press

Kharin, V. V. and Zwiers, F. W. 2002 Climate predictions with multimodel ensembles. *J. 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. 1999a Estimation and study of mesoscale variability in the Strait of Sicily. *Dynam. Atmos. Oceans*, **29,** 255–303

1999b Evolving the subspace of the three-dimensional multiscale ocean variability: Massachusetts Bay. *J. Marine Sys.*, **29,** 385–422

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

Purser, R. J. and Parrish, D. F. 2003 A Bayesian technique for estimating continuously varying statistical parameters of a variational assimilation. *Meteorol. Atmos. Phys.*, **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. 1999 Forecasting and simulating coastal ocean processes and variabilities with the Harvard Ocean Prediction System. Pp. 77–100 in *Coastal Ocean Prediction*, AGU Coastal and Estuarine Studies Series, Ed. C. N. K. Mooers, American Geophysical Union, Washington DC

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.* **12,** Eds. A. R. Robinson, J. J. McCarthy and B. J. Rothschild, John Wiley and Sons, New York

2004 Prediction systems with data assimilation for coupled ocean science and ocean acoustics. Pp. 325–342 in *Theoretical and computational acoustics 2003*, Eds. A. Tolstoy, Y. C. Teng and E. C. Shang, World Scientific

Robinson, A. R., Haley, P. J., Lermusiaux, P. F. J. and Leslie, W. G. 2002 Predictive skill, predictive capability and predictability in ocean forecasting. Pp 787–794 in Proceedings of Oceans 2002, MTS/IEEE conference, Holland Publications

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,** 347–404