Ensemble Forecasting in the Mediterranean Sea

by
Alessandro Bonazzi

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Ensemble Forecasting: Atmospheric and Ocean Weather

Despite the fact that operational weather ensemble forecasting is a recent achievement, the first ideas about the necessity of evaluating the forecast uncertainty date back to the very beginning of the Numerical Weather Prediction (NWP) history. A notable review of the origin of ensemble forecasting was recently written by Lewis (2004 [51]). In the early 1950's the first successful experiments of NWP were performed at the Institute of Advance Study, Princeton. Electronic computer were for the first time used to solve a set of dynamical equations to describe the atmospheric flow. In the same years Eady and Charney independently developed the theory of baroclinic instability (Charney 1947 [14], Eady 1949 [28]) that explains the basic instability mechanism by which small perturbations of the atmospheric state might have an exponential growth. Eady first realised the implication of baroclinic instability on weather forecasting: since the initial condition of the atmospheric state is known only up to a certain accuracy, the initial errors will propagate undermining the prediction skills. Eady stated that "long-range forecasting is necessarily a branch of statistical physics" (Eady 1951 [29]). In the late 1950' several practical problems of NWP systems were investigated and solved and in parallel the work of Phil Thompson addressed from a theoretical point of view the problem of predictability. He calculated the derivative of the error at the initial time for a quasi-geostrophic two-level model and extrapolated the error growth concluding that the limit of predictability is in the order of 1 week (Thompson 1957 [89]).

One of the major contribution to the understanding of the chaotic nature of the atmospheric flow was given by Lorenz. In 1963 he showed that the solution of a simple set of differential equations representing a forced dissipative hydrodynamic flow is aperiodic and unstable with respect to small modifications; then small perturbation in the initial condition can evolve in considerably different solutions(Lorenz 1963 [55]). In the conclusion of this seminal paper, Lorenz states that "prediction of the sufficiently distant future is impossible by any method, unless the present conditions are known exactly". In 1965 Lorenz also envisioned ensemble forecasting suggesting to apply a system of dynamical equation to a set of different initial states representing the error and inadequacies in observations; then he remarked that the future state is unpredictable when two states chosen from the ensemble becomes indistinguishable from two randomly chosen atmospheric states.

The existing link between dynamical and probabilistic aspects of weather forecasting was clarified in the late 1960s by Edward Epstein. He proposed a geometrical interpretation of ensemble forecasting where each member is a point is a phase space and the spread of points represents the uncertainty. The dynamical system is then made of random variables that are characterised by stochastic properties. Using a simple but nontrivial system of equations Epstein explicitly derived the full set of stochastic-dynamic (SD) equations for the mean and the variances of the state variables (Epstein 1969 [31]). Epstein also used Monte Carlo simulations as a reference for the numerical experiments performed with the SD model. He showed that the SD prediction were a better representation of the 'true' state then the determinist forecast. However the SD machinery suffers from limitations that prevented a wide application of the method; for instance the analytical derivation of SD equations is only possible if assumptions are made on the shape of ensemble distributions. Furthermore the computational time required to solve the the stochastic-dynamic equation for a system of n components growth with a factor n^2.

Interestingly it was the Monte Carlo method that survived the evolution of the stochastic prediction. In the 1970' Leith continued the work of Epstein and focused on the Monte

Carlo approximation to the SD prediction (Leith 1971 [44], 1974 [45] and Leith and Kraichnan 1972 [46])). He also studied the theoretical skill of Monte Carlo forecasting as a function of ensemble size (Leith 1974 [45]) claiming that a small number of members is sufficient to estimate the ensemble mean; however he provided no indication about the ensemble size that is required for the determination of higher-order moments of the forecast distribution. Leith explained that ensemble mean is a better estimate than a deterministic forecast starting from the best initial condition because averaging together individual members of a Monte Carlo simulation has the effect of filtering out the small scale structures of the prediction for which there is little predictability. He admitted that this point was controversial since the smooth mean field may become an unrealistic representation of the atmospheric state.

Today the state of the art global ensemble prediction systems (EPS) are all based on a slightly modified Monte Carlo approach. In traditional Monte Carlo the initial state probability distribution function (pdf) is assumed to be known and it is randomly sampled, where the operational weather forecasting centres have adopted different strategies to sample the initial state distribution. This choice is made to optimise the usage of computational time reducing the ensemble size to the smallest number of members that are able to capture the larger portion of atmospheric state variability. Again this technique was already being suggest by Lorenz back in 1965 (Lorenz 1965 [53]). The optimal perturbations by which it is possible to initialise an ensemble forecast are generally found as Singular Vectors that identify the perturbation of the initial state that have a maximum linear growth for a given norm (Lacarra and Talagrand 1988 [42], Farrel 1990 [34]). These structure can be qualitatively associated with the unstable baroclinic modes of the basic-state flow (Buizza and Palmer 1995 [9]).

Ocean ensemble forecasting is quite a novel field and has not undergone the growth as in weather forecasting. Ensemble method for the ocean are generally used in data assimilation schemes to provide the error covariances statistics (Evensen 2003 [33]), rarely used to quantify forecast uncertainty in short-term operational prediction systems. The aim of this work is to implement an ensemble forecasting methodology to be applied in the framework of the Mediterranean Forecasting System (MFS, see Pinardi et al. 2003 [73]) .

The increasing computational power available to the oceanographic community is feeding interest on the application of Monte Carlo methods to the ocean prediction problem. The development of large distributed computing network, such the Grid infrastructure (Foster and Kesselman 1999 [35]) makes possible to simultaneously run a very high number of simulations within the time constraint imposed by an operational application. A technical study of the feasibility of operational ensemble forecasting using the Italian Grid was performed applying a Monte Carlo approach based on randomly selected initial perturbation for a large number of members (see Pinardi et al. 2007 [74]). The development of sharing computing network makes available even to small scientific community large resources that were accessible only to few large centres in the world.

The major problem in establishing an ensemble procedure is the identification of the initial perturbation of the system that are required to *i*) represent the error of the system due to insufficient and imperfect observation of the initial state, *ii*) identify an optimal sub-sample of the initial error that is most likely to affect the prediction. For ocean forecasting this problem concerns both the initial and the boundary conditions. A natural implementation of ocean ensemble forecasting can be implemented using the ensemble forcing produced by a weather ensemble prediction system, such as European Centre for Medium-range Weather Forecasting (ECMWF). An ocean ensemble forecast can be produced forcing several ocean simulations with different ECMWF ensemble wind members. This approach is feasible but is assume that the forcing errors affects only the atmospheric prediction phase and not the atmospheric analyses.

The launch of scatterometer satellite missions have led to a new understanding of the wind stress over the sea surface. Since 1999 the QuikSCAT satellite mission has provided high resolution and spatially extensive wind measurements over the global ocean and revealed persistent small-scale features in the curl and divergence of the wind stress that are important kinetic energy input to the ocean (Chelton et al. 2003 [15]). What is relevant for ocean ensemble forecasting is that wind analyses from state of the art NWP such as ECMWF or US Natianal Centre for Environmental Prediction (NCEP) present a KE deficiency that is particularly relevant at the small scales (Chin et al. 1998 [16], Milliff et al. 1996 [65] , 1999 [66] , 2004a [63], 2004b [64]).

The fact that wind analyses present systematic and persistent errors in representing the external forcing over the sea surface constitutes an ideal starting point to address the problem of ocean ensemble forecasting. This is especially true for high resolution ocean modelling. This is by no means the only source of uncertainty in ocean forecasting but it is the most controllable. QuikSCAT wind constitutes a large data-set to compare the ECWMF products and derive a realistic representation of the errors. The availability of wind measurement allows to build an "objective" method for creating the Monte Carlo perturbations. This system is not optimal in the sense the reduces the ensemble size that is necessary to consider, but it is a sound representation of a well-known source of uncertainty for ocean forecast

A large part of the effort of this thesis were dedicated to the development of an "objective" modelling of uncertainty in the wind forcing. Full probabilistic modelling based on the Bayesian paradigm has been rapidly growing in recent years due to the increased computational power and the development of new sampling techniques. Chapter 2 presents a general explanation of the Bayesian modelling and the sampling strategies that are commonly applied. A Bayesian Hierarchical Model (BHM) of winds over the sea-surface is presented in Chapter 3. Chapter 4 describes the BHM ocean ensemble method together with an inter-comparison with other commonly used techniques such random perturbation of initial conditions and the direct application of ECMWF ensemble forcing. Finally Chapter 5 present a study of the relation between ensemble statistics and model error with a particular attention to data assimilation problems.

Chapter 2

Bayesian Analysis

2.1 Basic Bayesian Analysis

The first step in Bayesian inference is the identification of the full probability model (Wikle and Berliner 2007 [92]) . This is the joint probability of all the observable and non - observable components of interest. For instance if we want to estimate the wind field in a given region of the ocean, we should collect all the information we might have about this event. This means collecting all the available wind measurements together with our best knowledge of the dynamical process and with an estimate of non-observable parameters that enters the definition of the physical laws. So we can say that data, physical laws and parameters define our full probability model.

Our aims is to gain a better understanding of the non-observable components using the information that is carried by observations, i.e. we want to solve an inverse problem. To do this, we need to write the conditional probability of the unknown random variables given the observations. The theoretical solution of this problem is given by the Bayes' rule, however we shall note that most of the time we are unable to solve this in practice and then we should rely on some approximation methods.

The last part of our analysis is the evaluation of the model we derived so far. This is a good and common practice for all modelling efforts. At this points we need to check our initial assumtions, all the simplification that we might have done to see how to improve the solution we found.

A review of bayesian analysis with particular attention to data assimilation problems can be found in Wikle and Berliner 2007 [92] and in the classical texts by Berger (1985 [4]) and Bernardo and Smith (1994 [7]).

In the following random variable will be denoted by capital letter, while fixed and observed quantities by low case letter. The will write probability density using $p()$ and low case letter as arguments. Bold quantities will refer to matrix or vector. Finally Greek letters will usually denote parameters.

Let's say the Y represents our data and X is some unobservable random quantities. The full probability model is given by joint probability $p(x, y)$. Is always possible to break down a joint distribution in it's component:

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x)$$

where we use the notation $p(x|y)$ to define the conditional distribution of X given Y. The Bayes's rule simply follows:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{2.1}$$

It's worth to have a closer look of the components of this equation, because we will be commonly referring to them with this nomenclature:

- *Prior distribution*: the distribution $p(x)$ summarises our scientific understanding of the random variable X before adding information through the data. We should observe that the specification of this distribution is subjective. We can use past data or a set of physical laws, or even a mixture of this two. This is the crucial step is designing a Bayesian model. The choices we take here are likely to propogate through all the model, and we should keep in mind this when we will be evaluating the performancies of the Bayesian model we have built.

- *Data model*: the quantification of the random nature of the data Y and their relationships to X are modelled as the probability distribution $p(y|x)$. In literature this distribution takes the name of likelihood function $L(x|y)$ when we consider fixed data for random X; this is the case of the maximum likelihood estimation. Here we will think at the data stage as the distribution of imperfect observation Y around the true (unobservable) value (Wikle and Berliner 2007 [92]).

- *Marginal distribution*: the distribution $p(y) = \int p(y|x)p(x)dx$ enters the denominator of Bayes' rule 2.1 and can be thought as a "normalising constant". This shape of this distribution is generally unknown since it can be analytically computed for a very limited range of cases.

- *Posterior distribution*: $p(x|y)$ is the updated distribution of X having observed Y . The full specification of this distribution, or some of its moments are the final aim our analysis.

2.1.1 The Univariate Case; Normal Prior and Normal Data

Here we review a simple univariate example that it also described in Wikle and Berliner (2007 [92]). Let's say we want to estimate the temperature value on a given location and at a given time in the ocean; then our random variable is a scalar and will be denoted by X. The climatology distribution $X \sim N(\mu, \tau^2)$ constitutes our prior knowledge, where μ is the climatological mean temperature and τ^2 is the variance of the distribution. Then suppose we collect a series of n independent but noisy observations $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_n\}$. Each observation can be described by a normal distribution $Y_i|X = x \sim N(x, \sigma^2)$. The data stage containing all observations is:

$$
\begin{aligned}
p(\mathbf{y}|x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-1/2(y_i - x)^2/\sigma^2\} \\
&\propto \exp\{-1/2 \sum_{i=1}^{n} (y_i - x)^2/\sigma^2\}
\end{aligned}
$$

Applying the Bayes' rule 2.1 we can write the posterior distribution as:

$$
p(x|\mathbf{y}) \propto \exp\{-1/2[\sum_{i=1}^{n} (y_i - x)^2/\sigma^2 + (x - \mu)^2/\tau^2]\} \tag{2.2}
$$

The posterior distribution is given by the product of Gaussian distributions. Solving the squares in equation 2.2 we obtain :

$$p(x|\mathbf{y}) \quad \propto \quad \exp\{-1/2[x^2(n/\sigma^2 + 1/\tau^2) - 2x(\sum_{i=1}^{n} y_i/\sigma^2 + \mu/\tau^2)\} \tag{2.3}$$

where all the terms that do not contain x are ignored. The probability distribution expressed in 2.12 is still Normal and can be write as:

$$X|\mathbf{y} \quad \sim \quad N\big((\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1}(\sum_{i=1}^{n} \frac{y_i}{\sigma^2} + \frac{\mu}{\tau^2}), (\frac{n}{\sigma^2} + \frac{1}{\tau^2})^{-1}\big) \tag{2.4}$$

The expectation and variance of the Normal distribution 2.4 can be rearranged in this form:

$$\begin{aligned} E(X|\mathbf{y}) &= \mu + K(\bar{y} - \mu) \\ var(X|\mathbf{y}) &= (1 - K)\tau^2 \\ K &= n\tau^2(\sigma^2 + n\tau^2)^{-1} \end{aligned}$$

where we introduced the term K that is normally referred as “gain”. Several interesting considerations can be drawn from this simple example. First the posterior estimation of our temperature value is obtained as weighted average of prior and data. Note that the weights are inversely proportional to the variances of the input distributions. For a given σ^2 the posterior weights more the data stage as the number n of observation increases. The second consideration concerns the fact that only the statistics $\bar{y}$ enter the definition of the expectation and variance of $X|\mathbf{y}$. This means that the statistics $T(y) = \bar{y}$ is sufficient to describe the posterior distribution and that the knowledge of the data $\mathbf{Y}$ does not add any information. The sufficient statistics principle is a powerful form of dimension reduction in statistics and we will see an application of this idea in the next sections.

2.2 Monte Carlo Tecqniques

The Bayesian inference relies on the application of the simple theory we showed so far. However for the majority of practical applications we are not able to find the analytical solution. Indeed many problems arise in large dimensional spaces. Given some unknown variables $\mathbf{X} \in \mathbb{R}^m$ and data $\mathbf{Y} \in \mathbb{R}^c$ those are some of the integration problems that we commonly find in the Bayesian inference contest (Andrieu et al. 2003 [2] and Freitas [20]):

- *Normalization.* As was shown in previous section to compute the posterior distribution of $\mathbf{X}$ given $\mathbf{Y}$ we should be able to solve: $p(\mathbf{y}) = \int_{\mathbb{R}^m} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}$.

- *Expectation.* The computing of summary statistics of a distribution is achieve as:

$$E_{p(\mathbf{x}|\mathbf{y})}(f(\mathbf{x})) = \int_{\mathbb{R}^m} f(\mathbf{x})p(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

  where $f(\mathbf{x})$ is some function of interest integrable with respect to $p(\mathbf{x}|\mathbf{y})$. A common example are the mean ($f(\mathbf{x}) = \mathbf{x}$) and the covariance ($f(\mathbf{x}) = \mathbf{x}\mathbf{x}' - E_{p(\mathbf{x}|\mathbf{y})}(\mathbf{x})E'_{p(\mathbf{x}|\mathbf{y})}(\mathbf{x})$).

- *Marginalization.* Given a joint distribution $p(x, z) \in \mathbb{R}^m \times \mathbb{R}^n$ the marginal distribution for $\mathbf{X}$ is given by:

$$p(\mathbf{x}|\mathbf{y}) = \int_{\mathbb{R}^n} p(\mathbf{x}, \mathbf{z}, \mathbf{y})d\mathbf{x}$$

When the analytical computation of the above integrals is not feasible we can try analytical approximations, numerical integration or Monte Carlo simulations. Many assimilation techniques make some analytical approximation, such the Kalman filter that assumes the distributions are normal and the model is linear. This technique proved to be very useful for many application, but they might disregard some salient statistical feature of the processes under consideration (Freitas 1999 [20]). The numerical computation of high-dimension integrals is very expensive and might prove to be of little use in practical application. Monte Carlo approximation might be viewed as an intermediate approach.

The basic idea behind Monte Carlo methods is to draw N sample $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ from the distribution $p(\mathbf{x})$. Then, the samples can be used to describe the target distribution through an empirical point-mass function:

$$p_N(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_{(i)}}(\mathbf{x})$$

where $\delta_{\mathbf{x}_i}$ is the delta-Diract mass function located at $\mathbf{x}_{(i)}$. As a consequence we can approximate expectation of the form:

$$E_{p(\mathbf{x})}(f(\mathbf{x})) = \int_{\mathbb{R}^m} f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

by the following estimate:

$$E_N(f(\mathbf{x})) = \frac{1}{N} \sum_{i=1}^{N} f(\mathbf{x}_i)$$

The estimate $E_N(f(\mathbf{x}))$ is unbiased and it will almost surely converge to $E_{p(\mathbf{x})}(f(\mathbf{x}))$ by the strong law of large number (Robert & Casella 2001 [11]). In many applications it is unfeasible to sample directly from the probability distribution, that might to complicated, or we may know only up to a proportional constant. There are two families of possible solutions of this problem: Monte Carlo Markov Chain and Importance Sampling Monte Carlo. The overall aim is to produce ensemble members that are representative of the true posterior as efficiently as possible. There is no way to say which is the best, the choice should be based on the particular problem under investigation.

2.3 Importance Sampling Monte Carlo

Let's consider the problem of estimating the mean of the posterior distribution of a time depend process X_τ for $\tau = 0, \ldots, t$ conditioned on observation Y_τ available at time step $\tau = 1, \ldots, t$. In the following we define X_τ and Y_τ to be scalar without any loss of generality. The expectation is given by the following integral:

$$E(g_t(x_{0:t})) = \int g_t(x_{0:t}) p(x_{0:t}|y_{1:t}) dx_{0:t}$$

To apply the Monte Carlo approximation we need to draw samples from the posterior distribution $p(x_{0:t}|y_{1:t})$. When it is not possible we can apply an Importance Sampling Monte Carlo (ISMC) scheme. We introduce a easier to sample proposal distribution $q(x_{0:t}|y_{1:t})$ and make the following substitution:

$$\begin{aligned} E(g_t(x_{0:t})) &= \int g_t(x_{0:t}) \frac{p(x_{0:t}|y_{1:t})}{q(x_{0:t}|y_{1:t})} q(x_{0:t}|y_{1:t}) dx_{0:t} \\ &= \int g_t(x_{0:t}) \frac{p(y_{1:t}|x_{0:t}) p(x_{0:t})}{q(y_{1:t}|x_{0:t}) q(x_{0:t})} q(x_{0:t}|y_{1:t}) dx_{0:t} \\ &= \int g_t(x_{0:t}) \frac{w(x_{0:t})}{p(y_{1:t})} q(x_{0:t}|y_{1:t}) dx_{0:t} \end{aligned} \tag{2.5}$$

where $w(x_{0:t})$ are the un-normalized importance weights and are defined as:

$$w(x_{0:t}) = \frac{p(y_{1:t}|x_{0:t}) p(x_{0:t})}{q(x_{o:t}|y_{1:t})} \tag{2.6}$$

To avoid the computation of the normalizing density we substitute in 2.5 $p(y_{1:t})$ with its equivalent marginal integral:

$$\begin{aligned} E(g_t(x_{0:t})) &= \frac{1}{p(y_{1:t})} \int g_t(x_{0:t}) w(x_{0:t}) q(x_{0:t}|y_{1:t}) dx_{0:t} \\ &= \frac{\int g_t(x_{0:t}) w(x_{0:t}) q(x_{0:t}|y_{1:t}) dx_{0:t}}{\int p(y_{1:t}|x_{0:t}) p(x_{0:t}) \frac{q(x_{0:t}|y_{1:t})}{q(x_{0:t}|y_{1:t})} dx_{0:t}} \end{aligned} \tag{2.7}$$

then we substitute 2.6 in 2.7:

$$\begin{aligned} E(g_t(x_{0:t})) &= \frac{\int g_t(x_{0:t}) w(x_{0:t}) q(x_{0:t}|y_{1:t}) dx_{0:t}}{\int w(x_{0:t}) q(x_{0:t}|y_{1:t}) dx_{0:t}} \\ &= \frac{E_{q(\cdot|y_{1:t})}(w_t(x_{0:t}) g_t(x_{0:t}))}{E_{q(\cdot|y_{1:t})}(w_t(x_{0:t}))} \end{aligned}$$

Then by drawing sample from the proposal distribution $q(x_{0:t}|y_{1:t})$ it is possible to approximate the expectation as:

$$\begin{aligned}\overline{E(g_t(x_{0:t}))} &= \frac{\frac{1}{N}\sum_{i=1}^{N} g_t(x_{0:t}^{(i)})w_t(x_{0:t}^{(i)})}{\frac{1}{N}\sum_{i=1}^{N} w_t(x_{0:t}^{(i)})} \\ &= \sum_{i=1}^{N} g_t(x_{0:t}^{(i)})\tilde{w}_t(x_{0:t}^{(i)})\end{aligned}$$

where $\tilde{w}_t^{(i)}$ are the normalised importance weights:

$$\tilde{w}_t^{(i)} = \frac{w_t(x_{0:t}^{(i)})}{\sum_{i=1}^{N} w_t(x_{0:t}^{(i)})}$$

The normalised importance weights are nonnegative values and sum to one. Then it is possible to consider the normalised importance weight as as the probability of the samples obtained by the proposal distribution. A useful review of ISMC can be found in Merwe et al. (2000 [60]).

2.3.1 Particle Filter

For many application we are only interested in describing the filtering density that is the probability of X_t conditional on the previous state only of the variable. In the smoothing case we will be trying to modify the state X_t conditional of the observation we might have at subsequent time steps. The smoothing problem is not well resolved by ISMC, the sequential incorporation of observation prove to be more effective in this framework.

The first assumption we make in filtering is that the current state do not depend on future observation, and then we won't be correcting previously simulated states using new observations. In this case we can use proposal distribution of this shape:

$$q(x_{0:t}|y_{1:t}) = q(x_{0:t-1}|y_{1:t-1})q(x_t|x_{t-1}, y_{1:t}) \tag{2.8}$$

Now we assume that states correspond to a Markov process and that the observations are conditionally independent give the state, i.e:

$$p(x_{0:t}) = p(x_0)\prod_{j=1}^{t} p(x_j|x_{j-1}) \qquad \text{and} \tag{2.9}$$

$$p(y_{1:t}) = \prod_{j=1}^{t} p(y_j|x_j) \tag{2.10}$$

If we substitute 2.13 and 2.9 in equation 2.6 we obtain:

$$\begin{aligned} w_t &= \frac{p(y_{1:t}|x_{0:t})p(x_{0:t})}{q(x_{0:t-1}|y_{1:t-1})q(x_t|x_{t-1}, y_{1:t})} \\ &= \frac{p(y_{1:t-1}|x_{0:t-1})p(x_{0:t-1})}{q(x_{0:t-1}|y_{1:t-1})} \\ &\times \frac{p(y_{1:t}|x_{0:t})p(x_{0:t})}{p(y_{1:t-1}|x_{0:t-1})p(x_{0:t-1})}\frac{1}{q(x_t|x_{t-1}, y_{1:t})} \\ &= w_{t-1}\frac{p(y_{1:t}|x_{0:t})p(x_{0:t})}{p(y_{1:t-1}|x_{0:t-1})p(x_{0:t-1})}\frac{1}{q(x_t|x_{t-1}, y_{1:t})} \\ &= w_{t-1}\frac{p(y_t|x_t)p(x_t|x_{t-1})}{q(x_t|x_{t-1}, y_{1:t})} \end{aligned} \tag{2.11}$$

Now we have a method to sequentially update the weights given their previous state and a proposal distribution $q(x_t|x_{t-1}, y_{1:t})$. We can write a generic particle filter algorithm in two main steps, known as forecast and analysis (or update). An idealised algorithm can be written as:

---

1. Initialisation. Sample $x_0^{(i)}$ from the prior $p(x_0)$ for $i = 1, \ldots, N$
2. For $t = 1, \ldots, T$
   - Forecast: sample $x_t^{(i)} \sim q(x_t|x_{t-1}^{(i)}, y_{1:t})$ for $i = 1, \ldots, N$
   - Analysis: compute the importance weights $w_t^{(i)}$ using equation 2.11 for $i = 1, \ldots, N$ and then normalised weigths $\tilde{w}_t^{(i)}$
3. Compute the target distribution as:

$$p(x_t|y_{0:t}) \approx \hat{p}(x_t|y_{1:t}) = \sum_{i=1}^{N} \tilde{w}_t^{(i)} \delta_{x_t^{(i)}}(dx_t)$$

This generic algorithm is subject to many drawback as we shall see. A major problem is know as degeneracy. After some time steps the weight are likely to concentrate on a small portion of the ensemble, this is enough to prevent the scheme from describing a correct target posterior distribution. Several variants of particle filter have been proposed. A comparison between different schemes can be found in Arulampalam et al. (2002 [3])

In the next paragraphs we will focus on the choice of the proposal distribution and on some technique that might be used to mitigate the degeneracy of the particle filter.

2.4 Monte Carlo Markov Chain

Monte Carlo Markov Chain is a method to draw sample $x^{(i)}$ from the probability distribution $p(x)$ using a Markov chain mechanism (Andrieu et al. 2003 [2]). It is easier to describe MCMC in finite state space, i.e. where $x^{(i)}$ can only take a series s of discrete values $x^{(i)} = \{x_1, x_2, \ldots, x_s\}$. We say that the stochastic process $x^{(i)}$ is a Markov chain if the transitional probabilities between different values in the state spaces depends only on the stochastic variable current state, i.e.:

$$p(x^{(i)}|x^{(i-1)}, \ldots, x^{(1)}) = T(x^{(i)}|x^{(i-1)})$$

where $T(x^{(i)}|x^{(i-1)})$ is defined as the transition probability or kernel of the Markov chain. We say that the chain is homogenous is the kernel remains invariant for all i and the sum of the transitional probability over all $x^(i)$ is equal to 1. In this case the evolution of the random variable depends only on the current state and a fixed transition matrix. For every starting point the chain will converge to the invariant distribution $p(x)$ provided that the kernel is irreducible and aperiodic. The irreducibility condition requires that transitional kernel to explore all possible states of the random variable x. The aperiodicity condition requires that the chain to do not get trapped in cycles. The detailed balance provides a sufficient, but not necessary condition, to ensure that the a particular $p(x)$ is the invariant distribution:

$$p(x^{(i)})T(x^{(i-i)}|x^{(i)}) = p(x^{(i-1)})T(x^{(i)}|x^{(i-1)})$$

Summing both sides over $x^{(i-1)}$ gives:

$$p(x^{(i)}) = \sum_{x^(i-1)=x_1}^{x_s} p(x^{(i-1)})T(x^{(i)}|x^{(i-1)})$$

In continuous spaces the transitional matrix T becomes an integral kernel K:

$$p(x^{(i)}) = \int p(x^{(i-1)})K(x^{(i)}|x^{(i-1)})dx^{(i-1)}$$

MCMC samplers are irreducible and aperiodic Markov chain that have the target distribution as invariant distribution. The following description of Metropolis-Hasting and Gibbs sampler follow the work of Andrieu et al. (2003 [2]).

2.4.1 Metropolis-Hasting

The Metropolis-Hasting (MH) sampler is the most used MCMC algorithm (Metropolis and Tweedie 1949 [62] Metropolis et al. 1953 [61] and Hasting 1979 [38]). The algorithm provides a practical way to implement the transition kernel required by the Markov chain.

A step of the MH algorithm requires the sampling of a candidate value x^* from a proposal distribution $q(x^*|x)$. The Markov chain moves to the candidate x^* with an acceptance probability that is given by:

$$\mathcal{A}(x, x^*) = \min\Big(1, \frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}\Big)$$

The transition kernel for the MH algorithm can be written as:

$$K_{MH}(x^{(i+1)}|x^i) = q(x^{(i+1)}|x^i)\mathcal{A}(x^{(i)}|x^{i+1}) + \delta_{x^{(i+1)}}(x^{i+1})r(x^{(i)})$$

where the term $r(x^{(i)})$ represents the rejection probability:

$$r(x^{(i)}) = \int q(x^*|x^{(i)})(1 - \mathcal{A}(x^{(i)}|x^*))dx^*$$

Since by construction the MH kernel satisfies to detailed balance condition:

$$p(x^{(i)})K_{MH}(x^{(i+1)}|x^i) = p(x^{(i+1)})K_{MH}(x^{(i)}|x^{i+1})$$

the MH algorithm admits $p(x)$ as invariant distribution. Moreover the algorithm is aperiodic since the it always allow for rejection of the candidate value and it is irreducible if we ensure that the support of the proposal distribution covers the support of the targeted $p(x)$. The last condition means that the candidate distribution is required to have positive probability for all the possible outcome of the invariant distribution $p(x)$. If the MCMC is aperiodic and irreducible then the chain asymptotically converges (Tierney 1994 [90]).

We shall now present a particular case of the MH algorithm that prove to be numerically convenient for application that require the evaluation of high-dimesional problems.

2.4.2 Gibbs sampler

The Gibbs sampler is a special case of the MH algorithm that was introduced by Geman and Geman in 1984 ([37]) in the contest of digital image processing. The main idea is that we only consider univariate conditional distribution (or full conditional), i.e. the distribution when all random variables are assigned fixed values except one. Such conditional distribution are normally easy to sample often being normal , gamma or other common prior distributions. However we shall note that it is not always possible to analytically derive the expression of full conditionals, hence the applications of Gibbs sampler are limited.

Suppose that we are in a finite space and the state vector can take values $\{x_1, x_2, \ldots, x_n\}$ and that we know all the full conditional distributions
$p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, x_n)$. In such case we can use as proposal distribution for $j = 1, \ldots, n$:

$$q(x^*|x^{(i)}) = \begin{cases} p(x^*_j|x^{(i)}_{-j}) & \text{if } x^*_{-j} = x^{(i)}_{-j} \\ 0 & \text{Otherwise.} \end{cases}$$

The corresponding acceptance probability is:

$$\begin{aligned} \mathcal{A}(x, x^*) &= \min\Big(1, \frac{p(x^*)q(x^{(i)}|x^*)}{p(x)q(x^*|x^{(i)})}\Big) \\ & \min\Big(1, \frac{p(x^*)q(x^{(i)}_j|x^{(i)}_{-j})}{p(x)q(x^*_j|x^*_{-j})}\Big) \\ & \min\Big(1, \frac{p(x^*_{-j})}{p(x^{(i)}_{-j})}\Big) = 1 \end{aligned}$$

The acceptance probability for the Gibbs sampler is always 1, making this algorithm extremely efficient.

Hence a generic Gibbs sampler can be written as:

```
1. Initialise s $x_{0,1:n}$
2. For $i = 0$ to N -1
    • Sample $x_1^{(i+1)} \sim p(x_1|x_2, x_3, \ldots, x_n)$
    • Sample $x_2^{(i+1)} \sim p(x_2|x_1, x_3, \ldots, x_n)$
    • ...
    • Sample $x_j^{(i+1)} \sim p(x_j|x_1, \ldots, x_{j-1}, x_{j+1}, x_n)$
    • ...
    • Sample $x_n^{(i+1)} \sim p(x_j|x_1, x_2 \ldots, x_{n-1})$
```

2.5 Bayesian Hierarchical Model

It is often convenient to apply a hierarchical approach to solve complex statistical problem. The basic idea is that any joint distribution $p(x, \theta, \phi)$ can be decomposed as:

$$p(x, \theta, \phi) = p(x|\theta, \phi)p(\theta|\phi)p(\phi)$$

where we think at x as an output of a model, and θ, ϕ as parameters. When x depend on ϕ only through the parameter θ the problem reduces to:

$$p(x, \theta, \phi) = p(x|\theta)p(\theta|\phi)p(\phi)$$

The combination between hierarchical strategy and bayesian modelling is at basis of the statistical tool that we are aimed to develop. In the following section we present the Bayesian Hierarchical Model (BHM) implemented for an idealised problem. The sampling from the posterior distribution is achieved through the application of Particle Filter and Gibbs Sampling.

2.5.1 A Toy BHM: an application for wind at sea surface

So far we have discussed some basic concept of the bayesian analysis. Here we present an example of statistical modelling applied to a 2-dimensional wind field. This problem has some inherent properties that make the application of a bayesian model easy and successful. Namely it exist a strong physical balance that allows to describe the wind as a function of the pressure field, i.e. the geostrophic equilibrium. Geostrophy is both easy to model and effective to diagnostic large scale circulation. This physical property of the wind fits with the hierarchical approach that we have just described. We model the zonal and meridional components of the wind as function of the pressure field, without any need to explicit any correlation structure between the components of the wind vector. A real world application is presented in Chapter 3, here we discussed an idealised toy problem.

We consider a uniformly spaced grid of 25 × 25 points located at mid-latidute on the northern hemisphere. Suppose that we have measurement $\mathbf{d}_u = \{\mathbf{d}_\mathbf{u}, \mathbf{d}_\mathbf{v}\}$ of the wind $\mathbf{u} = \{\mathbf{u}_x, \mathbf{u}_y\}$ that covers only a portion of the basin and that we have more accurate observations $\mathbf{d}_p$ of the pressure field $\mathbf{p}$ over all the domain (figure 2.1). The wind observations are simulated applying the geostrophic relation to the pressure field:

$$u = -\frac{1}{f}\frac{\partial p}{\partial x} \quad \text{and} \quad v = \frac{1}{f}\frac{\partial p}{\partial y}$$

where we used $1/f = 11000\text{s}^{-1}$. The assimilation of the observed quantity (wind and pressure) should provide information about the unobserved parameter θ that links to two fields, i.e. $1/f$.

The full probability problem is :

$$p(\mathbf{u}, \mathbf{p}, \theta|\mathbf{d}_u) \quad \propto \quad p(\mathbf{d}_u, \mathbf{d}_p|\mathbf{u}, \mathbf{p}, \theta)p(\mathbf{u}, \mathbf{p}, \theta)$$

Assuming that the data stages $\mathbf{d}_u$ and $\mathbf{d}_p$ are independents we can write an hierarchy to represent the posterior distribution as:

$$\begin{aligned} p(\mathbf{u}, \mathbf{p}, \theta|\mathbf{d}_u, \mathbf{d}_p) \quad \propto \quad & p(\mathbf{d}_u|\mathbf{u}) \\ & p(\mathbf{u}|\theta, \mathbf{p}) \\ & p(\theta)p(\mathbf{p}|\mathbf{d}_p) \end{aligned} \tag{2.12}$$

where we substituted the $p(\mathbf{d}_p|\mathbf{p})p(\mathbf{p})$ with its posterior $p(\mathbf{p}|\mathbf{d}_p)$ via Bayes' rule.

2.5.2 Particle Filter

The posterior distribution can be found via Particle Filter following this procedure:

---

1. randomly sample an ensemble of size $m, \{\mathbf{p}^i, w_p^i\}$ to represent the $p(\mathbf{p}|\mathbf{d}_p)$. The normalised weight w_p^i can be found as:

$$w_p^i \quad \propto \quad \frac{\exp[-\frac{1}{2}(\mathbf{d}_p - H_p\mathbf{p}^i)'R^{-1}(\mathbf{d}_p - H_p\mathbf{p}^i)]}{\sum_{i=1}^m \exp[-\frac{1}{2}(\mathbf{d}_p - H_p\mathbf{p}^i)'R^{-1}(\mathbf{d}_p - H_p\mathbf{p}^i)]}$$

where R is the diagonal measurement error covariance matrix and H is an observational operator that moves the sample $\mathbf{p}^i$ to the location of data $\mathbf{d}_p$.

2. sample an ensemble of size $m, \{\theta^i, w_\theta^i\}$ from the independent distribution $p(\theta)$. Since θ are unobserved parameters impose $w_\theta^i = 1/m$ for all i.

3. for each sample $\mathbf{p}^i$ and θ^i generate $\mathbf{u}^i$ from $p(\mathbf{u}|\theta^i, \mathbf{p}^i)$. Then the weights:

$$w^i = p(\mathbf{d}_u|\mathbf{u}^i)w_p^i$$

can be used to make inference under the full posterior distribution 2.12.

However we shall note that this procedure is likely to be ineffective even if we consider a very large number m. This issue arises since we are approximating a continuos distribution as a finite ensemble of limited size. The presence of unobserved variable θs is a further complication and make this approach unfeasible unless we introduce some approximations that are aimed to constrain the filter to the most relevant portion of the probability space. The recent work of Berliner and Wikle (2007 [6]) presented an Approximate Partilce Filter (APF) that can be used to reduce the degeneracy of the sampler in high-dimensional space, where the re-sampling strategy are generally not sufficient.

Before presenting the modified algorithm we consider this re-arrangment of the posterior distribution. Integrating both sides of 2.12 with respect to $\mathbf{u}$ yields:

$$p(\mathbf{p}, \theta|\mathbf{d}_u, \mathbf{d}_p) \quad \propto \quad \tilde{p}(\mathbf{d}_u, |\theta, \mathbf{p})p(\mathbf{p}|\mathbf{d}_p) \tag{2.13}$$

where,

$$\tilde{p}(\mathbf{d}_u|\theta, \mathbf{p}) = \int p(\mathbf{d}_u|\mathbf{u}, V)p(\mathbf{u}|\mathbf{p}, \theta)d\mathbf{u} \tag{2.14}$$

Let's say that the marginal distribution 2.14 is computable, then we further assume that for fixed $\mathbf{p}$ the distribution $\tilde{p}$ admits a sufficient statistics $T(\mathbf{d}_u; \mathbf{p})$ for θ such that:

$$\tilde{p}(\mathbf{d}_u|\theta, \mathbf{p}) = G(T(\mathbf{d}_u; \mathbf{p})|\theta)h(\mathbf{d}_u; \mathbf{p}) \tag{2.15}$$

where the distribution $h(\mathbf{d}_u; \mathbf{p})$ do not depend on θ. Then we can rewrite equation 2.13 as:

$$\tilde{p}(\mathbf{p}, \theta|\mathbf{d}_u, \mathbf{d}_p) \quad \propto \quad G(T(\mathbf{d}_u; \mathbf{p})|\theta)h(\mathbf{d}_u; \mathbf{p})p(\theta)p(\mathbf{p}|\mathbf{d}_p) \tag{2.16}$$

The introduction of the sufficient statistics is the crucial point of the work of Berliner and Wikle (2007 [6]) and need not to be overlooked. The marginal 2.14 links the distribution of the unobservable parameter θ to the observable quantity $\mathbf{d}_u$ recognising that given a pressure field there is a relation between the velocity field and the geostrophic parameters. However this relation explains only up to a certain degree the behaviour of the observed wind field. For instance the atmospheric flow over the sea surface can be induced by convective turbulent processes that are totally pressure independent. As we shall see in the following the proposed assumption relies an the idea that for weighting the θ priors we can assume that the wind observations are in geostrophic balance with pressure field. This approximation is aimed to reduce the probability space rejecting all wind samples that are too far away from geostrophy without assuming that the wind field are in perfect geostrophic balance.

Now we can write the APF algorithm:

---

1. draw an ensemble as size m, $\{\mathbf{p}^i, w_p^i\}$, where the weights are defined according to 2.13.
2. draw an ensemble m, $\{\theta^i, 1/m\}$ from the prior distribution $p(\theta)$.
3. the resulting ensemble of Step 1 and 2 represent $\tilde{p}(\mathbf{d}_u|\theta, \mathbf{p})$ (2.13) if they are weighted according to:

$$w_1^i \quad \propto \quad w_p^i G(T(\mathbf{d}_u; \mathbf{p})|\theta^i)h(\mathbf{d}_u; \mathbf{p}^i)$$

We approximate the weigths w_1^i retaining only:

$$\tilde{w}_1^i \quad \propto \quad w_p^i G(T(\mathbf{d}_u;\mathbf{p})|\theta^i)$$

4. the posterior distribution 2.12 can be re-written as:

$$\tilde{p}(\mathbf{u},\mathbf{p},\theta|\mathbf{d}_u,\mathbf{d}_p) \quad = \quad \tilde{p}(\mathbf{u}|\mathbf{p},\theta,\mathbf{d}_u)\tilde{p}(\mathbf{p},\theta|\mathbf{d}_u,\mathbf{d}_p)$$

   If it is possible to simulate $\mathbf{u}^i$ from $\tilde{p}(\mathbf{u}|\mathbf{p},\theta,\mathbf{d}_u)$ proceed with inference under the full posterior distribution with weights w_1^i.

To increase the efficiency of the scheme Berliner and Wikle (2007 [6]) proposed to replace $G(T(\mathbf{d}_u;\mathbf{p})|\theta^i)p(\theta)$
with $G(T(\mathbf{d}_u;\mathbf{p}))p(\theta|T(\mathbf{d}_u;\mathbf{p})$ in equation 2.16. The resulting algorithm will be more computationally expensive but the APF will sample only physically consistent values for the θ parameter. The modified APF2 algorithm can be written as:

---

1. same of step 1 of APF.
2. draw an ensemble of size m, $\{\theta^i, 1/m\}$ from the posterior distribution
   $p(\theta|T(\mathbf{d}_u;\mathbf{p}))$.
3. the ensemble $\{\mathbf{p}^i,\theta^i\}$ represents $\tilde{p}(\mathbf{d}_u|\theta,\mathbf{p})$ (2.13)
   if weighted by:

$$w_2^i \quad \propto \quad w_p^i G(T(\mathbf{d}_u;\mathbf{p}))h(\mathbf{d}_u;\mathbf{p}^i)$$

   We approximate the weigths w_2^i retaining only:

$$\tilde{w}_2^i \quad \propto \quad w_p^i G(T(\mathbf{d}_u;\mathbf{p}))$$

4. same as APF replacing $\tilde{w}_1^i$ with $\tilde{w}_2^i$

2.5.3 Gibbs Sampling

If the distributions are conforms then a Gibbs sampler can be written to describe the posterior distribution using a sequence of full conditional. A long iteration of the chain will produces a set of samples that will be distributed according to the targeted distribution. In our case the Gibbs sampling chain can be written as:

```
1. Initialise $\mathbf{u}^0, \mathbf{p}^0$ and $\theta^0$ and set $k = 0$
2. Iteration k
    • Sample $\mathbf{u}^{k+1} \sim p(\mathbf{u}^{k+1}|\mathbf{p}^{k-1}, \theta^{k-1}, \ldots)$
    • Sample $\mathbf{p}^{k+1} \sim p(\mathbf{p}^{k+1}|\mathbf{u}^{k+1}, \theta^{k-1}, \ldots)$
    • Sample $\theta^{k+1} \sim p(\theta^{k+1}|\mathbf{u}^{k+1}, \mathbf{p}^{k+1}, \ldots)$
3. set $t = t + 1$ and go to 2
```

The first step of the chain require the initialisation of the random variable, this initial choice do not influence the behaviour of the MCMC. The second step of the chain requires the sampling from the 'full conditionals' that represent the probability distribution of a variable when all the other random variable are kept fixed. When the prior distributions belongs to the family of standard distributions (Gamma, Guassian, Exponential) we can derive the analytical expression for the 'full conditional' and then the writing of the sampling algorithm become strait-forward.

2.6 Results

The Gibbs sampling and the particle filter were tested on the same problem. The posterior distribution for the geostrophy parameter θ is used here to compare the performances of these techniques. The number of samples m and iteration k was set to 10000 for the two algorithm.

Figure 2.2 shows the results of the APF scheme..The first panel shows the sampling cloud of the geostrophic parameter θ. Given a vague prior of the $p(\theta)$ only a small portion of the sampled values are given weights different from zero (see panel b); out of 10000 iteration, only 5 or 6 sample fall in the relevant portion of the probability space for the geostrophic parameter. Consequently the posterior distribution of θ is not well represented by the sampling (see panel c).

Figure 2.3 shows the result of modified APF2 where we sample θ from its posterior distribution given the significant statistics $T(\mathbf{d}_u; \mathbf{p})$. Panel a shows that the θ samples span a limited portion of the probability space, consequently the associated weights are almost equally distributed between the θ ensemble (see panel b). The posterior distribution is well defined (see panel c)and it is centered around the analytical value of θ that was used to generate the synthetic wind data.

The Gibbs sampling proves to be the most efficient scheme in solving the posterior distribution for the toy model. Panel a of figure 2.4 show the θ for the 10000 iteration of the Markov chain. After few hundreds iteration the chain converges to the true value of θ. Panel b show the Markov chain after the burn-in period, the sample of θ are confined to a very space even if the prior for the θ was vague and the initial value was 0. The posterior distribution (panel c) is centered around the true value of θ with a smaller variance then the APF schemes.

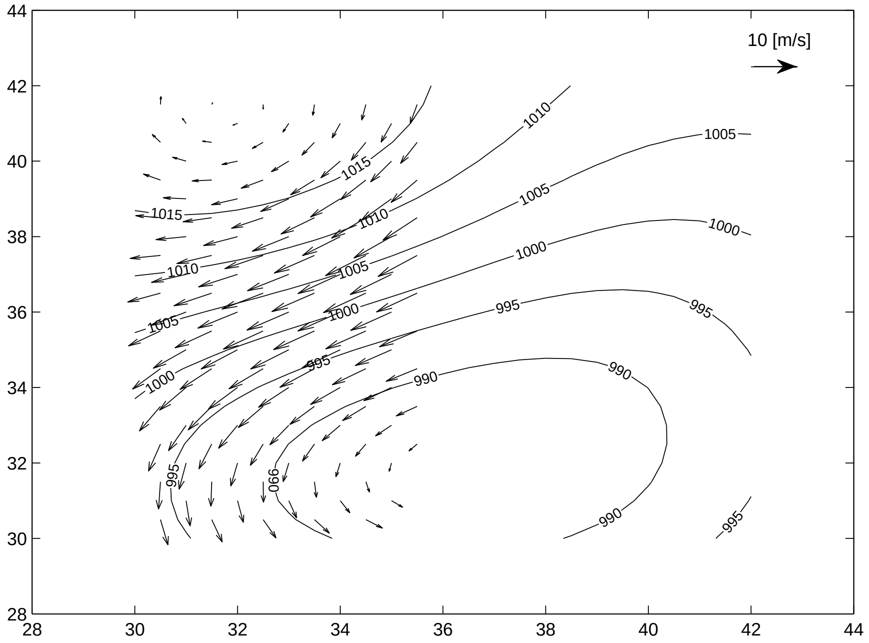


Figure 2.1: Toy model domain and data. The contour lines represent pressure data [mbar]. Arrows are wind synthetic observations derived from the pressure field.

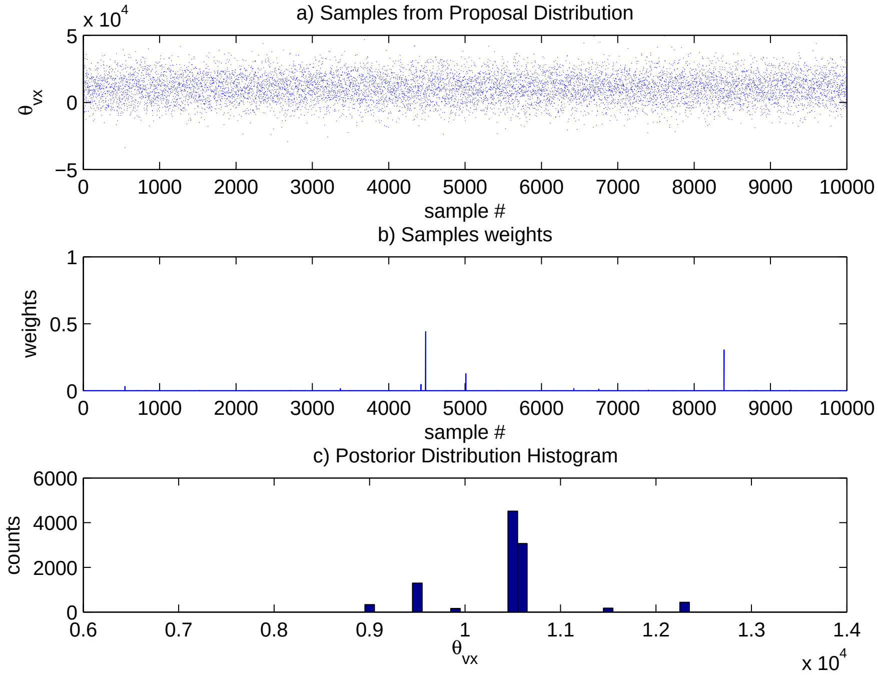


Figure 2.2: a) Sampling cloud of the geostrophic parameter θ for APF from the vague prior distribution $p(\theta)$. b) sampling weight for each of the 10000 sampled values of θ. c) posterior distribution for θ after resampling.

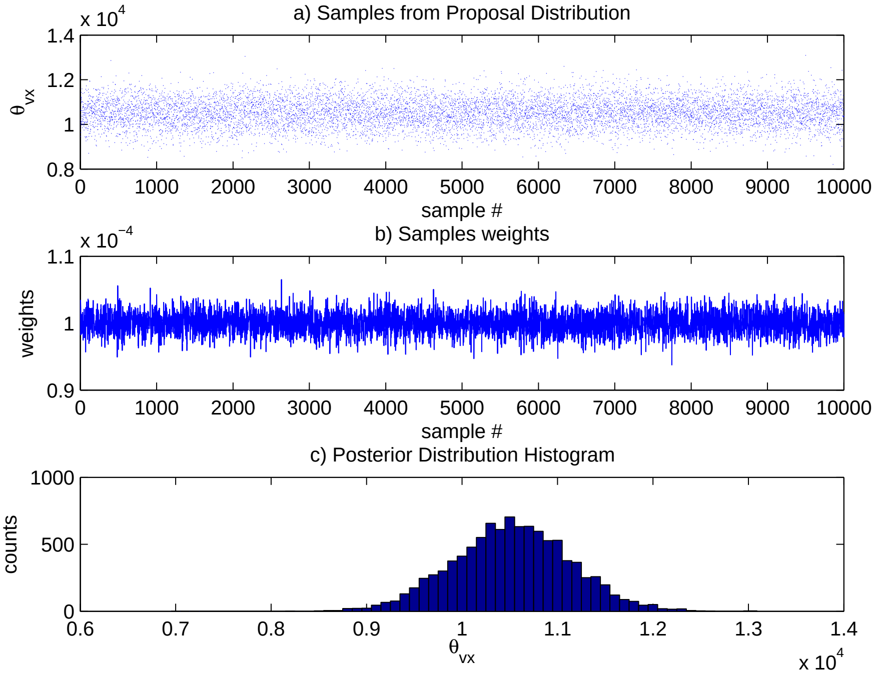


Figure 2.3: a) Sampling cloud of the geostrophic parameter θ for APF from the conditional distribution $p(\theta|T(\mathbf{d}_u; \mathbf{p}))$. b) sampling weight for each of the 10000 sampled values of θ. c) posterior distribution for θ after resampling.

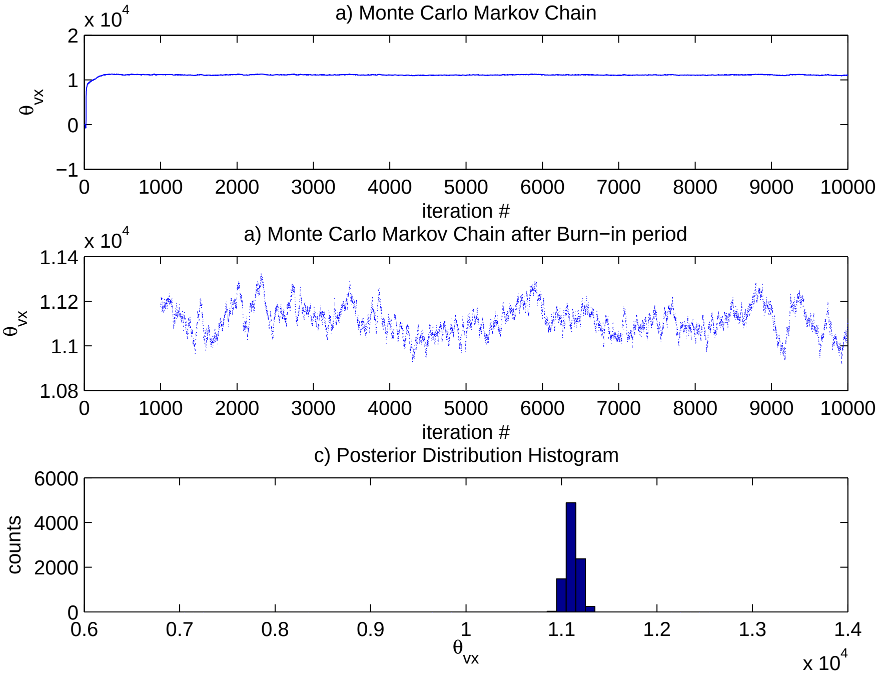


Figure 2.4: a) Markov chain for the geostrophic parameter θ for Gibbs sampling. b) same as a) but skipping the first 1000 iteration, i.e. burn-in period. c) posterior distribution for θ

Chapter 3

The Bayesian Hierarchical Model for surface winds

The study of the air-sea interactions for ocean modelling relies on the atmospheric surface fields provided by Numerical Weather Predicition (NWP) systems such as the European Centre for Medium-range Weather Forecasting (ECMWF). The introduction of back-scattering wind measurement from satellites in the recent years has uncovered a new picture of the winds over the ocean (Chelton 2003 [15]) . In particular since the launch of the QuikSCAT mission in 1999, the scientific community benefits from the availability of high-resolution and high-frequency measurements of the wind at sea surface.

In this work we aim to implement a statistical model of the wind at sea surface melding ECMWF products with QuikSCAT data. Following the approach that is described in Chapter 2, we propose to use a Bayesian Hierarchical Model (BHM).

The BHM model will be presented in its components that are the data stages, the process models and the priors. The algorithm used to solve the posterior distribution is the Gibbs Sampling. A theoretical discussion of this method can be found in Chapter 2.

3.1 ECMWF and QuikSCAT Sea Surface Winds

ECMWF provides data for wind, pressure, cloud cover, relative humidity, dew point and air temperature. This data set is available to the Mediterranean Forecasting System community at a 6 hour frequency (at 0, 6, 12, and 18UTC) and at a space resolution of half a degree. The ECMWF forecast is released daily to the MFS group since September 1998. In this work the forecast is assumed to be available only once a week, as it was in the configuration of the MFS between 1998 and 2004. The forecast window has a length of 10 days and it starts at 18:00 p.m. of every Tuesday. For the purpose of the wind BHM only wind and pressure data are used in the data stage.

The QuikSCAT satellite carries a microwave radar that measures near-surface wind speed and direction under all weather and cloud conditions over the world oceans. QuikSCAT data are available at $25 \times 25 km$ resolution in continuous 1,800 km swaths which cover about 90% of the ice-free ocean every 24 hours (see the data product user's manual [70]) For the Mediterranean Sea the QuikSCAT data were collected following the time frequency of the EMCWF products. The QuikSCAT orbits were searched for observation during a 6 hour period centred at 0, 6, 12 and 18 UTC. The maximum time lag between EMCWF analyses and QuikSCAT data is 3 hours. The algorithm used to extract wind speed and direction from the microwave measurement is the 'Direction Interval Retrieval with Thresholded Nudging' developed by Styles (1999). Further post-processing guarantees the elimination of rain-flagged data for wind speed grater then 15 m/s and of the 3 outermost wind vector cells of each swath (see Milliff and Morzel 2004 [63]).

In the Mediterranean Sea the average wind speed difference between ECMWF analyses and QuikSCAT data is 0.8 ± 0.5 $[m/s]$ for the period between September 2004 and March 2005. The difference is mainly due to a positive bias suggesting that the ECMWF usually underestimates wind speed (Milliff 2004 [64]); positive misfit up to 10 m/s are usually observed during strong wind events. There is not a preferential differential direction between the two data set, the mean difference being $-1.1 \pm 6.9°$.

Recent works have demonstrated a Kinetic Energy (KE) deficiency in global ocean surface wind provided by Numerical Weather Prediction (NWP) systems with respect to coincident surface wind retrieved by scatterometer data (see Chin et al. 1998 [16] and Milliff et al. 1996 [65], 1999 [66], 2004 [63]). An analysis of the differences in the Mediterranean Sea between ECMWF and QuikSCAT (see Milliff 2004 [64]) demonstrate that a KE deficiency also characterises the analyses and forecast currently used to force the MFS ocean model. The KE spectra for QuikSCAT data is consistent with a power-law relation with the form $PSD(k) \sim k^p$ where k is the wavenumber and p is the exponent of the power-law. Freilich and Chelton (1986 [36]) highlighted the relation between the theoretical power-law decay of isotropic two-dimensional turbulence with QuikSCAT spectra. This idea was further investigated by Wikle et al. (1999 [93]) that analysed the TOGA COARE IOP region in the Pacific Ocean cross-comparing satellite, in-situ and high resolution doppler-radar measurements from aircrafts. Milliff (2004 [64]) showed that ECMWF analyses depart from the theoretical power-law relation at about 400 km. At the spatial scale of 100 km the KE amplitude differences are grater then two order of magnitude. These differences are evident in all the ECMWF analyses in a time period ranging from 2000 to 2004.

The energy lack at small scales of ECMWF products might have a strong impact on the mesoscale ocean circulation and it is regarded as a source of uncertainty in an ocean prediction effort.

Figure 3.1 show the error growth of the ECMWF forecast. The statistic is computed as the RMS difference between forecast and analysis on the same day. For each one of the 10 days of the prediction window a set of 216 weeks of forecast, ranging from January 2000 to December 2004, was analysed to compute the mean and variance of the error. The results are divided by year.

Forecast error grows linearly with leading time reaching a mean value of 3 m/s at the end of the forecast.

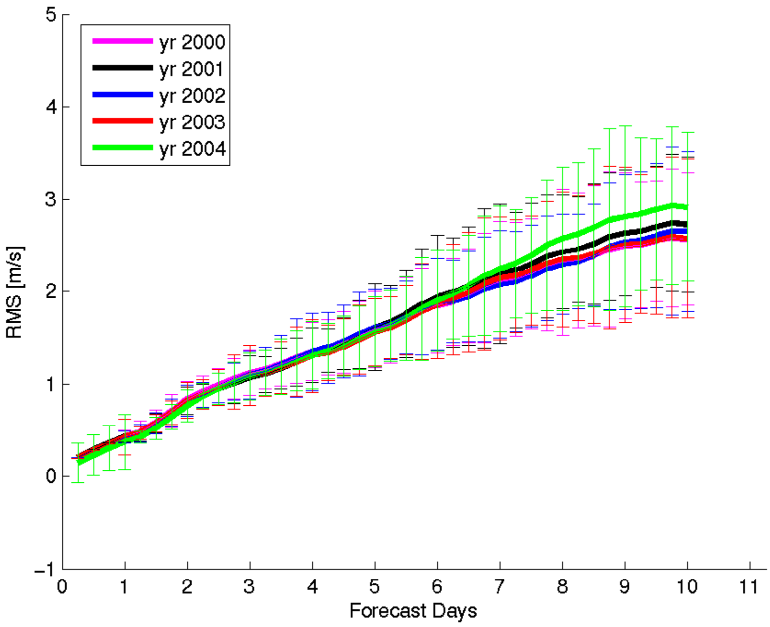


Figure 3.1: RMS difference between ECMWF forecast and analysis as a function of leading time for the years 2000-2004 ($[m/s]$).

3.2 The Data Stage

The BHM data stages includes the ECMWF analyses and forecast and the QuikSCAT wind data. Two assumptions have been made:

- The data stages are treated as independent. This is a major approximation but has it was shown the ECMWF in not effective in retaining the information content of the QuikSCAT data, at least for what concerns the small scales of motion.

- White-noise error. The spatial structure of the data stage error is assumed to behaves has a white-noise process. This is a common simplification in data assimilation and it is normally reasonable to represent observational error as un-correlated in space and time. However it could be objected that the ECMWF data stage should be defined differently since ECMWF forecasts are a modelling product. The choice was made to simplify the solution of the problem, and it may be revisited in the future.

The data stages are represented by multi-variate Normal distribution centred around the ‘true’ value of the zonal and meridional wind components. The covariance of these distributions is entirely defined by the a scal measurement error. Hence the QuikSCAT and ECWMF data stages for the wind component U and V are written as:

$$\begin{aligned}\begin{pmatrix} S_u^t|U^t \\ S_v^t|V^t \end{pmatrix}, \sigma_s^2 &\sim N\left(\begin{pmatrix} K_s U^t \\ K_s V^t \end{pmatrix}, \sigma_s^2 I\right) \\ \begin{pmatrix} A_u^t|U^t \\ A_v^t|V^t \end{pmatrix}, \sigma_a^2 &\sim N\left(\begin{pmatrix} K_a U^t \\ K_a U^t \end{pmatrix}, \sigma_a^2 I\right)\end{aligned}$$

where S_u, S_v are the QuikSCAT zonal and meridional wind components and A_u, A_v are the ECMWF homologues. The error variances σ_s^2 and σ_a^2 are defined as known and constants. The matrices K_s and K_a are the observational operator for QuikSCAT and ECMWF data that move the ’true’ gridded value to the location of the observations.; I is the identity matrix.

The QuikSCAT data are assumed to have an measurement variance of 1 $[m^2/s^2]$ A reference for this number can be found in Ebuchi et al. (2004 [30]). The choice of the ECMWF variance was based on a preliminary study of the error properties of the ECMWF products. The wind analyses were compared with the QuikSCAT data, while the forecasts were tested against the analyses. It should be noticed that this parameter has a strong impact on the behaviour of the BHM. It is one of the few fixed parameters and a small value for the ECMWF variance will constraint the solution to the input data. Especially because this data set it is ubiquitous with respect to the model grid. The variance of the ECMWF wind distribution was set to 10 $[m^2/s^2]$.

The third data stage is the distribution of the pressure data. Again a multi-variate Normal distribution is chosen:

$$A_p^t|P^t, \sigma_{ap}^2 \sim N(K_p P^t, \sigma_{ap}^2)$$

where A_p is the ECMWF pressure field and K_p is its observational operator. The variance σ_p^2 is not considered constant and we only define the initial value to be 1 mbar^2.

3.3 The APBL Process Model

The process model expresses the relation between the random variables in the BHM. This relation can be represented in a statistical sense or applying physical balances. Given the nature of the problem treated here, is is strait-forward to apply the geostrophic balance to prescribe to relation between wind components and the pressure field. Royle et al. (1999 [83]) first proposed to model the spatial correlation of the wind field imposing a geostrophic constraint in the formulation of the process model. Here we expand this approach deriving two formulation of the process model as approximations of an Atmospheric Planetary Boundary Layer (APBL, see Anderson and Gill 1974 [1]).

The linear momentum equation for the APBL are:

$$\frac{\partial u}{\partial t} = +fv - \frac{1}{\rho_0}\frac{\partial p}{\partial y} - \gamma u \tag{3.1}$$

$$\frac{\partial v}{\partial t} = -fu - \frac{1}{\rho_0}\frac{\partial p}{\partial x} - \gamma v \tag{3.2}$$

where γ is a "bottom" friction parameter, f is the Coriolis parameter and ρ_0 is a constant reference atmospheric boundary layer densisty. The dependent variable (u, v) are the winds components and p is the pressure field.

We solve for u and v in the equations 3.1 and 3.2, to get:

$$u = -\frac{1}{f\rho_0}\frac{\partial p}{\partial y} - \frac{\gamma}{f}v - \frac{1}{f}\frac{\partial v}{\partial t} \tag{3.3}$$

$$-v = -\frac{1}{f\rho_0}\frac{\partial p}{\partial x} - \frac{\gamma}{f}u - \frac{1}{f}\frac{\partial u}{\partial t} \tag{3.4}$$

Then substituting 3.3 into 3.4 for u and 3.4 into 3.3 for v we derive:

$$\frac{1}{f}\left[\frac{\partial^2}{\partial t^2} + (f^2+\gamma^2)\right]v + 2\frac{\gamma}{f}\frac{\partial v}{\partial t} = \frac{1}{\rho_0}\frac{\partial p}{\partial x} - \frac{1}{f\rho_0}\frac{\partial^2 p}{\partial y \partial t} - \frac{\gamma}{f\rho_0}\frac{\partial p}{\partial y} \tag{3.5}$$

$$\frac{1}{f}\left[\frac{\partial^2}{\partial t^2} + (f^2+\gamma^2)\right]u + 2\frac{\gamma}{f}\frac{\partial u}{\partial t} = -\frac{1}{\rho_0}\frac{\partial p}{\partial y} - \frac{1}{f\rho_0}\frac{\partial^2 p}{\partial x \partial t} - \frac{\gamma}{f\rho_0}\frac{\partial p}{\partial x} \tag{3.6}$$

Two approximations of these equations are used to defined different versions of the process model.

3.3.1 APBL Process Model 1

Here we ignore all the time derivative in equations 3.6 and 3.5. This leads to a simple set of equations:

$$\begin{aligned}\rho_0 v &= \frac{f}{\gamma^2+f^2}\frac{\partial p}{\partial x}+\frac{\gamma}{\gamma^2+f^2}\frac{\partial p}{\partial y}\\ \rho_0 u &= -\frac{f}{\gamma^2+f^2}\frac{\partial p}{\partial y}+\frac{\gamma}{\gamma^2+f^2}\frac{\partial p}{\partial x}\end{aligned}$$

both the pressure derivative in x and y are used model the u and v fields.

A stochastic analogue of this set of equation can be written on a discrete grid as:

$$V_t = \theta_{vx}D_xP_t+\theta_{vy}D_yP_t+\epsilon \tag{3.7}$$

$$U_t = \theta_{uy}D_yP_t+\theta_{vx}D_xP_t+\epsilon \tag{3.8}$$

where V, U, P are random variables defined on a discrete grid and D_x, D_y are discrete differential operators. The equation error ϵ and θ are stochastic parameters. For θ and ϵ we are required to specify their prior distribution. Details about the specification of these distributions can be found in next paragraphs. This set of stochastic equation was already used in the work of Royle et al. (1999 [83]) that however do not explicitly derived 3.7 and 3.8 from the APBL equations 3.5 and eq:gillU.

A multivariate Normal distribution is used to represent the process model stage:

$$[U_t|P_t,\theta_{ux},\theta_{uy},\sigma_u^2] \sim N(\theta_{uy}D_yP_t+\theta_{vx}D_xP_t,\Sigma_u) \tag{3.9}$$

$$[V_t|P_t,\theta_{vx},\theta_{vy},\sigma_v^2] \sim N(\theta_{vy}D_yP_t+\theta_{vx}D_xP_t,\Sigma_v) \tag{3.10}$$

where Σ_u and Σ_v are the covariance matrices of the two distribution. This is a crucial point in designing the Bayesian model and it will discussed in detail in the following.

3.3.2 APBL Process Model 2

Ignoring the second order time derivative in u and v in equations 3.6 and 3.5, the APBL equations reduce to:

$$
\begin{aligned}
2\frac{\gamma}{f}\frac{\partial v}{\partial t} + \frac{1}{f}(f^2+\gamma^2)v &= \frac{1}{\rho_0}\frac{\partial p}{\partial x} - \frac{1}{f\rho_0}\frac{\partial^2 p}{\partial y \partial t} - \frac{\gamma}{f\rho_0}\frac{\partial p}{\partial y} \\
2\frac{\gamma}{f}\frac{\partial u}{\partial t} + \frac{1}{f}(f^2+\gamma^2)u &= -\frac{1}{\rho_0}\frac{\partial p}{\partial y} - \frac{1}{f\rho_0}\frac{\partial^2 p}{\partial x \partial t} - \frac{\gamma}{f\rho_0}\frac{\partial p}{\partial x}
\end{aligned}
$$

Then using central finite difference we derive the descretized equations:

$$
\begin{aligned}
V_t &= \left[(f^2+\gamma^2) - \frac{2\gamma}{\Delta}\right]^{-1} \\
&\left(-\frac{2\gamma}{\Delta}V_{t-1} + \frac{f}{\rho_0}D_xP_t + \frac{1}{\rho_0}(\gamma - \frac{1}{\Delta})D_yP_t + \frac{1}{\Delta\rho_0}D_yP_{t-1}\right) \\
U_t &= \left[(f^2+\gamma^2) - \frac{2\gamma}{\Delta}\right]^{-1} \\
&\left(-\frac{2\gamma}{\Delta}U_{t-1} - \frac{f}{\rho_0}D_yP_t + \frac{1}{\rho_0}(\gamma - \frac{1}{\Delta})D_xP_t + \frac{1}{\Delta\rho_0}D_xP_{t-1}\right)
\end{aligned}
$$

where Δ is the time step and capital letters refer to gridded variables.

A stachastic analogue of these equations is:

$$
\begin{aligned}
V_t &= +\theta_{v(1)}V_{t-1} + \theta_{vx}D_xP_t + \theta_{vy}D_yP_t \qquad (3.11) \\
&\quad +\theta_{vy(1)}D_yP_{t-1} + \epsilon \\
U_t &= +\theta_{u(1)}U_{t-1} + \theta_{uy}D_yP_t + \theta_{ux}D_xP_t \qquad (3.12) \\
&\quad +\theta_{ux(1)}D_xP_{t-1} + \epsilon
\end{aligned}
$$

where again we ignore the constant air density term. The term θ and ϵ are assumed to be random variables.

Similarly to the previous case, a multivariate Normal distribution is used to represent the process model:

$$
\begin{aligned}
&[U_t|U_{t-1}, P_t, P_{t-1}, \theta_{u(1)}, \theta_{ux}, \theta_{uy}, \theta_{ux(1)}, \Sigma_u] \\
&\sim N(\theta_{u(1)}U_{t-1} + \theta_{uy}D_yP_t + \theta_{ux}D_xP_t + \theta_{ux(1)}D_xP_{t-1}, \Sigma_u) \\
&[V_t|V_{t-1}, P_t, P_{t-1}, \theta_{v(1)}, \theta_{vx}, \theta_{vy}, \theta_{vy(1)}, \Sigma_v] \\
&\sim N(\theta_{v(1)}V_{t-1} + \theta_{vy}D_yP_t + \theta_{vx}D_xP_t + \theta_{vy(1)}D_yP_{t-1}, \Sigma_v)
\end{aligned}
$$

3.4 The prior and hyper-prior distributions

To complete the design of the BHM it is necessary to specify a prior distribution for each one of the stochastic variables that were introduced so far.

The random variables for which a prior distribution is needed are the pressure field, the equation parameter $\theta_{ux}, \theta_{uy}, \theta_{vx}, \ldots$ and the covariance $\Sigma_u, \Sigma_v, \ldots$ of the process models. The specification of these distributions will introduce other parameters that are normally referred as hyper-priors. At this level of the hierarchy the distributions are chosen to be vague and a particular choice for the quantities at the hyper-prior level should not affect the behaviour of the model. Table 3.1 presents the prior distribution and hyper-prior values for all the parameters of the BHM.

3.4.1 The Pressure Field

The wind components are modelled using either 3.7-3.8 or 3.11-3.12 that are mainly based on the pressure field for which a prior distribution is required. To specification of this quantity is performed in a reduced space. The pressure is decomposed in a set of m basis $\Phi_1, \ldots, \Phi_m$ as:

$$P^t \equiv \mu + \Phi\alpha^t$$

where μ is a basin mean pressure value and α_t is a vector coefficient of dimension m. Since the set of basis function Φ is kept constant, it suffices to write the α coefficient prior:

$$\begin{aligned} [\alpha_t|\Lambda] &\sim N(0, diag(\Lambda)) \\ [\lambda_i|\xi_i, r_i] &\sim IG(\xi_i, r_i) \end{aligned}$$

where $diag(\Lambda) = \lambda_1, \ldots, \lambda_m$: i.e. there is no correlation between the coefficient α. The Inverse Gamma (IG) distribution parameter ξ, r are constant, see table 3.1. Several experiments were conduced to identify the best set of basis function Φ. Two possible choices are a geometrical set of eigenvectors, computed from a Gaussian covariance matrix, or a set of Empirical Orthogonal Functions (EOF) of the pressure field itself. The choice of the EOFs proved to be optimal in reducing the difference between the original and the reconstructed P field (not shown here).

3.4.2 Equation Parameters and Errors

The priors of the equation parameters θ are defined as Normal distributions.

$$[\theta|\mu_\theta, \sigma_\theta^2] \sim N(\mu_\theta, \sigma_\theta^2)$$

The mean and variance of the parameter prior are shown in table 3.1. It should be noticed that having very large variances allows the BHM to find the optimal value for the θ variables. As it was shown in Chapter 2, the prior assumption made at this level doesn't influence the behaviour of the Markov Chain. This is one of most desirable property of a well constructed Monte Carlo Markov Chain (MCMC) and make this techniques preferable to the Importance Sampling Monte Carlo (ISMC). On the other hand, posterior θ values that are not physically consistent are indicator of errors in the MCMC coding, making this an important check on the numerical implementation of the Bayesian model.

The last part of the BHM design concerns the definition of the equation error form for the process model. Three approaches have been tested: a scalar white noise process and a structured field based on EOF or wavelet.

1. White-noise process $\epsilon = \tau I$

   In the first case the process model error is supposed to be a scalar quantity, then the covariance is a diagonal matrix $\Sigma \equiv \sigma_u^2 I$, where I is the identity matrix and σ_u^2 is a scalar variance for to the u. The prior distribution for the positive-definite variance is chosen the be an Inverse Gamma, i.e. $\sigma_u^2 \sim IG(\xi_\sigma, r_\sigma)$ the initial value for the ξ_σ and r_σ are reported in table 3.1. A similar relation applies for the process model of the v component of the wind field.

2. EOF representation of process model error:

   $$\epsilon = \Psi \mathbf{b}_u + \tau I \tag{3.13}$$

   the process model error for is wrote as a combination of a scalar τ plus a term that is defined by a set of k EOFs Ψ and their amplitude coefficient vector $\mathbf{b}_u$. If n is the size of state vector then $k << n$. The term $\Psi \mathbf{b}_u$ does not have a unique interpretation. It can be view as an additional term in the physical process model equations or as a nuisance term in an hierarchical model of a full covariance matrix. To show this point consider this simple process model:

   $$\begin{aligned} X &= 0 + \Psi\ \mathbf{b} + \tau I \\ \tau &\sim N(0, \sigma^2) \\ \mathbf{b} &\sim N(0, C) \end{aligned}$$

   where X is a generic random variable, and C is the covariance matrix of the $\mathbf{b}$ coefficients. This system is equivalent to write:

   $$\begin{aligned} X &= 0 + \mathbf{f} \\ \mathbf{f} &\sim N\ (0, \Psi\ C\ \Psi^T + \sigma^2\ I) \end{aligned}$$

The two formulations are identical, in the sense that the marginal distribution of $\int [X|\Psi\mathbf{b}, \sigma^2][\mathbf{b}|C]db = [X|\Psi C\Psi' + \sigma^2]$. The major advantage of using the first system instead of the second is that the dimension of the covariance matrix reduces to $[k \times k]$, making trivial the numerical solution of the inverse problems.

3. Wavelet representation of process model error:

$$\epsilon_t = \Xi \mathbf{d_t} + \tau$$

where $\mathbf{d_t}$ is a vector of size n of temporally evolving random coefficients and $\boldsymbol{\Xi}$ is a matrix of size $n \times n$ of Daubechies wavelet basis functions of order two that are defined on the prediction grid (Daubechies and Paul 1987 [19]); a modification is introduced to take into account the close domain (see Cohenamd et al. 1993 [17]). The coefficient vector $\mathbf{d_t}$ is assumed to follow a first order Markov vector autoregression, hence its prior distribution is set to:

$$\begin{aligned} \mathbf{d_t}|\mathbf{H_d}, \mathbf{d_{t-1}}, \boldsymbol{\Sigma}_{\eta_\mathbf{d}} &\sim N(H_d \mathbf{d_{t-1}}, \boldsymbol{\Sigma}_{\eta_\mathbf{d}}) \\ \Sigma_{\eta_d} &\equiv diag(\sigma^2_{\eta_b}(1), \ldots, \sigma^2_{\eta_b}(n)) \end{aligned}$$

This model it is initialised at time 0 as $\mathbf{d_0} \sim \mathbf{N}(\mu_\mathbf{d}, \boldsymbol{\Sigma}_{\eta_\mathbf{d}})$. The covariance matrix Σ_{η_d} is diagonal and is defined accordingly to a multi-resolution scaling. This constraint follows from the consideration that the sea surface wind as it is observed by QuikSCAT has an energy spectrum that follow a power-law relation. Wornell (1993 [95]) and Chin et al. (1998 [16]) derived the variance for such processes in one and two dimensions. Wikle et al. (2001 [94]) combined the previous results with the innovation variance of an auto-regressive process. They show that the variance of this process can be written as:

$$\sigma^2_{\eta_b} \propto [1 - h_b^2(l)][2^{-l(1+d)-1}]$$

where $h_b(l)$ is the marginal variance of the auto-regressive process and $2^{-l(1+d)-1}$

is the variance of a two dimensional wavelet coefficient, where l is the level of the multi-resolution decomposition and d is the slope of the power law relation. This relationship was used to find the prior of an Inverse Gamma distribution.

3.5 Posterior Distribution Estimates

The posterior distribution of the wind model can now be written. The time window over which the problem will be solved is divided in an analysis $[1 : T]$ and a forecast $[T + 1 : T + L]$. The QuikSCAT data stage is available only on the analysis while during the forecast only ECMWF prediction are available. The time integration appears as a product of distribution since each step is supposed to be independent from the others. Then the posterior distribution of the wind model defined by the first approximation and the first error type is:

$$\begin{aligned}
&\prod_{t=1}^{t=T+L} [U^t, V^t, P^t, \ldots | S_u^t, S_v^t, A_u^t, A_v^t, A_p^t] \\
&\qquad \propto \quad \prod_{t=1}^{t=T+L} ([S_u^t | U^t, \sigma_s^2][S_v^t | V^t, \sigma_s^2] \\
&\qquad\qquad [A_u^t | U^t, \sigma_a^2][A_v^t | V^t, \sigma_a^2][A_p^t | P^t, \sigma_p^2] \\
&\qquad\qquad [U^t | P^t, \theta_{ux}, \theta_{uy}, \Sigma_u][V^t | P^t, \theta_{vx}, \theta_{vy}, \Sigma_v] \\
&\qquad\qquad [\alpha^t | \Lambda])[\Lambda][\ldots]
\end{aligned}$$

The posterior distributions for the other process models are similarly found. As it was discussed in Chapter 2, the solution of this problem can be obtained numerically following different approaches. The most interesting being the Metropolis-Hasting and the Importance Sampling Monte Carlo. Given the nature of the probability distributions considered here, and the linearity of the process model, it is convenient to choose the Gibbs sampler, that is a very efficient variation of the Metropolis-Hasting algorithm. This scheme works as a smoother, so the solution of the posterior distribution will be found in a time window, where past and future observations are used to constraint the solution at a certain time

Parameter	Prior Distribution	Hyper-prior Spec.	Initial Values	Version	Comment
$\theta_{ux}, \theta_{uy}, \theta_{vx}$ $\theta_{vy}, \theta_{ux(1)}, \theta_{vx(1)}$	$\sim N(\mu_\theta, \sigma_\theta^2)$	$\mu_\theta = 0\ [\frac{m^3 s}{kg}]$ $\sigma_\theta^2 = 10^4\ [\frac{m^6 s^2}{kg^2}]$	$\theta_{\cdot} = 0$	All	same prior for all θs
$\theta_{u(1)}, \theta_{v(1)}$	$\sim N(\mu_\theta, \sigma_\theta^2)$	$\mu_\theta = 0\ [\frac{m^3 s}{kg}]$ $\sigma_\theta^2 = 10^2\ [\frac{m^6 s^2}{kg^2}]$	$\theta_{u(1)}, \theta_{v(1)} = 0$	A2-E1, A2-E2, A2-E3	
λ_i	$\sim IG(\xi_i, r_i)$	$\xi_i = 100$, $r_i = 100$	$\lambda_i = 0$	All	same prior for all $i = 1, \ldots, 20$
σ_u^2, σ_v^2	$\sim IG(\xi_\sigma, r_\sigma)$	$\xi_\sigma = 2$ $r_\sigma = 2$	$\sigma_u^2, \sigma_v^2 = 10[m^2/s^2]$	A1-E1, A2-E1	
σ_u^2, σ_v^2	known	n.a.	$0.5\ [m^2/s^2]$	A1-E2, A1-E3 A2-E2, A2-E3	constant diagonal term
σ_{ap}^2	$\sim IG(\xi_{ap}, r_{ap})$	$\xi_\sigma = 3$ $r_\sigma = 3$	$\sigma_{ap}^2 = 10\ [mbar^2]$	A1-E1, A1-E2, A2-E1	
σ_{ap}^2	known	n.a.	$\sigma_{ap}^2 = 1\ [mbar^2]$	A1-E3, A2-E2, A2-E3	
b_i	$\sim N(0, c_i)$	$c_i \sim IG(\xi_{ci}, r_{ci})$ $\xi_{ci} = 100, r_{ci} = 100$	$b_i = 0$	A1-E2, A2-E2	for all b_i in $\mathbf{b} = \{b_1, \ldots, b_{20}\}$
$\mathbf{d}_t$	$\sim N(H_b \mathbf{d}_{t-1}, \Sigma_{\eta d})$	$\sigma_{\eta di} \sim IG(q_{\eta d}, r_{\eta d})$ $\mathbf{d}_0 = N(\mu_d, \sigma_d^2)$ $H_b \sim N(\mu_{Hb}, \sigma_{Hb})$	$\mathbf{d}_t = 0$ $\mathbf{d}_0 = 0$	A1-E3, A1-E3	$\Sigma_{\eta d}$ is diagonal with $\sigma_{\eta di}^2$ elements for i in $1, \ldots, n$

Table 3.1: Prior and hyper-prior distributions for different Process Models and Equation Errors.

step.

The derivation of the full conditional distribution for this simple model implementation, required by the Gibbs sampler can be found on Appendix A.

3.6 Results

A set of experiments was performed to assess the sensitivity of the BHM solution to the process model and the error covariance structure. The selection of a process model influences ability of the Bayesian model to statistically interpolate between the data stages. It is reasonable to expect that adding complexity to the process model we should improve the representation of the wind field operated by the BHM. However there is a point over which the model is not able to isolate the effect of different terms in eqautions and compensating effects might become important. In this case the solution loose a physical consistency, while still holding a statistical sense.

The results shown here refers to a time window defined around a forecast start day February 8th 2005. The analysis part starts 2 weeks before, while the forecast leading time is 10 days. As it will presented in the next chapter this choice in made to be conform with the operational chain adopted by the Mediterranean Forecasting System.

APBL Process Model 1, error type 1 (A1-E1)

The process model of version A1-E1 is based on the balance between geostrophic and friction terms as expressed in equation 3.7-3.8. The equation error covariance matrix that appears in 3.9 (and 3.10) is modelled as a diagonal matrix with uniform components, i.e. the random variable that represents the error structure is a scalar.

The geostrophy terms dominates process model. The posterior distributions for the term θ_{uy} and θ_{vx} have an expectation value of -5500 and 5100 $[\frac{m^3}{kgs^{-1}}]$ respectively , while friction terms θ_{ux} and θ_{vy} have an expectation value of 1500 and 1900 $[\frac{m^3}{kgs^{-1}}]$ (see figure 3.2). We remind that in a friction-less case the geostrophic parameters $-\theta_{uy}$ and θ_{vx} would have value $1/f \sim 1^4[\frac{m^3}{kgs^{-1}}]$. The equation for U exhibits a stronger geostrophy balance then the V equation. Since strongest winds over the Mediterranean region are mostly meridional, as it is the case for Mistral or Bora, the BHM represents a non-linear process, such as friction, with different posterior distributions for the equation parameters of the meridional and zonal wind components.

The error term σ_u^2 and σ_v^2 have expectation values of 5.5 and 5.8 $[m^2/s^2]$ respectively (see figure 3.3). The error terms are a quantification of the difference between the posterior field for U and V that includes the data stages and the wind field inferred by the process model: the error terms reflects the ability of the process model to mimic the behaviour of the wind data. The variance σ_{ap}^2 represents the difference between ECMWF and BHM pressure field. A variance of 2000 $[Pa^2]$, that corresponds to a standard deviation of 0.4 $[mbar]$, reflects a substantial agreement between the pressure field estimate by the Bayesian model and the data stage.

APBL Process Model 2, error type 1 (A2-E1)

In version A2-E1, beside geostrophy and friction, two more terms enter the process model in equation 3.11 and 3.12 . The $\theta_{u(1)}$ (or $\theta_{v(1)}$) relates the U (or V) wind component between two time steps. From a statistical point of view this is an auto-regressive parameter of a AR-1 process, while physically it comes from the time-derivative of u (or v). One interpretation doesn't exclude the other. Even if the derivation of equations 3.11 and 3.12 was rigourous, the BHM do not perform a time integration with CFL constraints then the term $\theta_{u(1)}$ (or $\theta_{v(1)}$) do not properly represent a time step process. The terms $\theta_{ux(1)}$ and $\theta_{vy(1)}$ are significant for motion at time scale grater then the inertial time scale and where first described by Anderson and Gill (1974 [1]) to be responsible for Rossby wave propagation in a boundary layer. The results shown in figure 3.4 and 3.5 suggest that the AR(1) terms dominate the process model and are responsible for a significant reduction of the expectation value for geostrophy and friction terms respect to version A1-E1. The ratio between geostrophy friction terms is still in favour of geostrophy but to a minor extent to what has been observed for the previous process model. The planetary waves term $\theta_{ux(1)}$ and $\theta_{vy(1)}$ have a small impact on version A2-E1 and account for half the value of the friction. The pressure field variance σ^2_{ap} is about 2500 $[Pa^2]$. The model is able to reproduce the data stage for pressure with an error that is less then 1 $[mbar]$ (see figure 3.6).

The expectation values for equation error variance σ^2_u and σ^2_v are 5.1 and 4.4 $[m^2/s^2]$ respectively. Overall they are smaller then what observed in version 5 but the difference between the two process model is not large.

The fact that the error covariance matrix if specified only through a scalar value implies that a portion of un-correlated white noise is introduced at each sampling from the distribution of U or V. To avoid this problem it is introduced a more structured form of the error covariance matrix for the process model equation.

APBL Process Model 1, error type 2 (A1-E2)

A set of 20 EOFs of U and V were used to build the covariance prior as defined in 3.13. The introduction of the EOF model for the process model equations change the posterior distribution of the θs parameters with respect to version A1-E1. The major changes affect the geostrophy terms, whose expectation value is reduced to 3000 $[m^3skg^1]$ (see figure 3.2).

The EOFs would represent a natural basis to build a purely statistical wind model and have the potential to represent the dynamical patterns of the wind. Therefore they compete in the BHM with the other equation terms in fitting the data stages. In the work of Berliner at al.[5] the error covariance was defined by a set of k EOFs that do not appeared in a statistical process model that was already modelled through EOFs. The separation between process and error was introduced by splitting the series of Empirical Orthogonal Function in two halves; then no direct competition between the two parts is possible since they represent different aspects of targeted field. However in this application it is not strait-forward to ensure that the EOFs do not compete with the dynamical terms of the process model since the EOFs are not scale-selective.

In version A1-E2 the variances σ^2_u and σ^2_v, see figure 3.3, representing the diagonal component of the process model covariances models are fixed values.

APBL Process Model 2, error type 2 (A2-E2)

A drastic change affects the posterior distribution of the equation parameters, see figure 3.4 and 3.5 . The interaction between the U and V error EOFs changes the balance between the terms of equations 3.11 and 3.12; the AR(1) coefficients $\theta_{v(1)}$ and $\theta_{u(1)}$ invert sign and become negatives. The geostrophic terms θ_{uy} and θ_{vx} augment their value in order to compensate the AR(1) coefficients. The model does converge to some equilibrium but the drastic alteration of process model parameters indicates that the physical consistency is lost.

Note that all the variances shown in figure 3.6 for version A2-E2 are constant values.

APBL Process Model 1, error type 3 (A1-E3)

A more explicit scale separation between process model and error modelling can be introduced using multi-scale wavelet. Here the most critical prior assumption concerns the variance distribution at different scale. As it was earlier in this Chapter discussed a reasonable assumption is to prescribe the variance of the wavelet coefficient accordingly to the slope of a power-law relation.

The θs parameter expectation values are similar to version A1-E1, see figure 3.2. There is no interference between the error term and the rest of the process model. The wavelet implementation is able to capture only spatial scale that are not resolved by the geostrophic balance and friction. We observe an increase in the variance of the pressure data stages that do not alter the representation of the major features of the pressure field as it is described by the ECMWF analyses and forecasts. Note that all the variances shown in figure 3.3 for version A1-E3 have been set to fixed values.

APBL Process Model 2, error type 3 (A2-E3)

The wavelet term has an impact on the posterior distribution of the θs parameters. Interestingly the AR(1) terms almost vanish making this formulation very similar to version A1-E3, see figure 3.4 and 3.5. The effect of the scale selective error term suggest that the version A1-E3 formulation suffices to explain the m large scale motion of the problem considered here.

The diagonal terms of the process model covariances σ_u^2 and σ_v^2 and the variance σ_{ap}^2 of the pressure data stage are kept constants to make the sampler converge, see figure 3.6.

Conclusive remark

The parametrization of equation error through EOF reduces the expectation values of the large scale parameter θ_{uy} and θ_{vx} in version A1-E2 compared to version A1-E1. We also note an inversion of the sign of the AR(1) parameters $\theta_{u(1)}$ and $\theta_{v(1)}$ in version A2-E2 compared to version A2-E1. This is due to the fact that the error EOFs (error type 2) and the equation parameters compete for the representation of the same dynamical processes. The competition between the error terms and the equation parameter causes the Gibbs sampler to converge to a solution that is unphysical; then version A1-E2 and A2-E2 loose their analogies with the APBL equations. For this reason in the following we focus only on the physically consistent version A1-E1, A2-E1, A1-E3 and A2-E3.

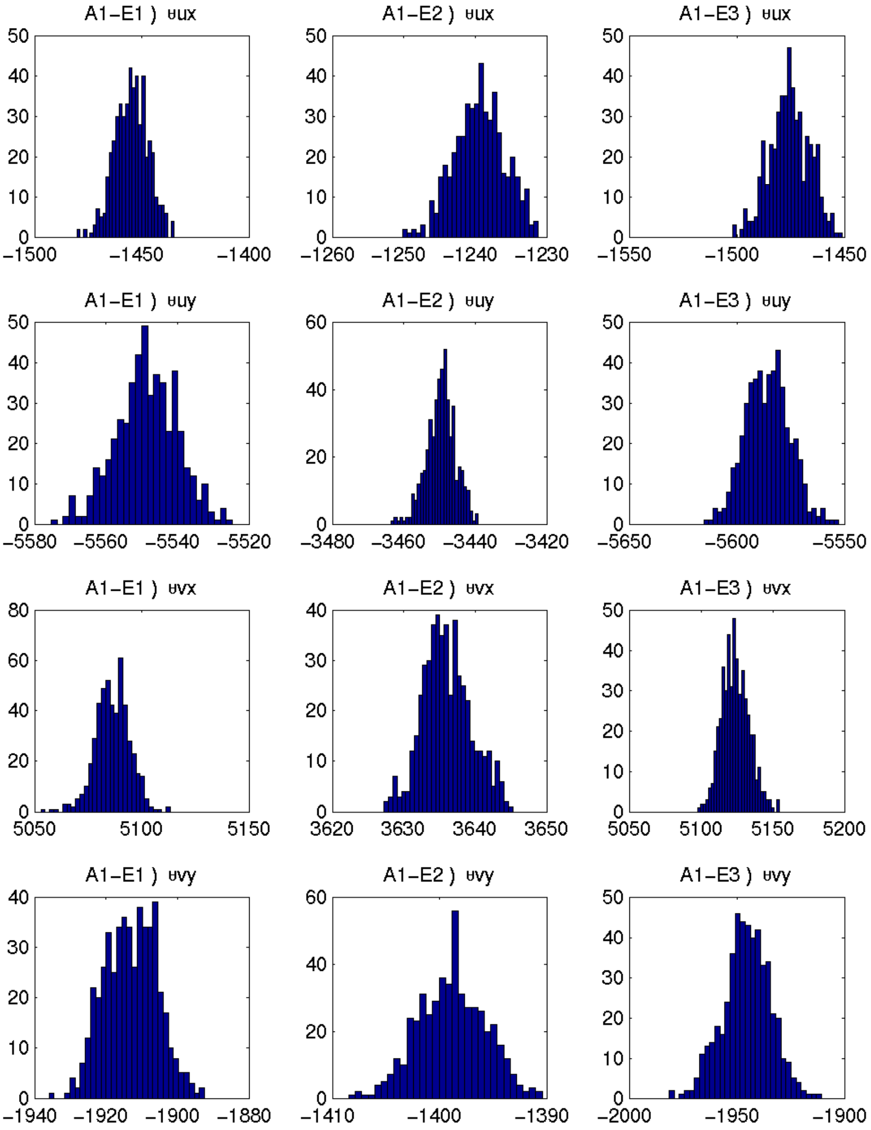


Figure 3.2: Posterior distribution of $\theta_{ux}[\frac{m^3 s}{kg}], \theta_{uy}[\frac{m^3 s}{kg}], \theta_{vx}[\frac{m^3 s}{kg}]$ and $\theta_{vy}[\frac{m^3 s}{kg}]$ for version A1-E1, A1-E2 and A1-E3.

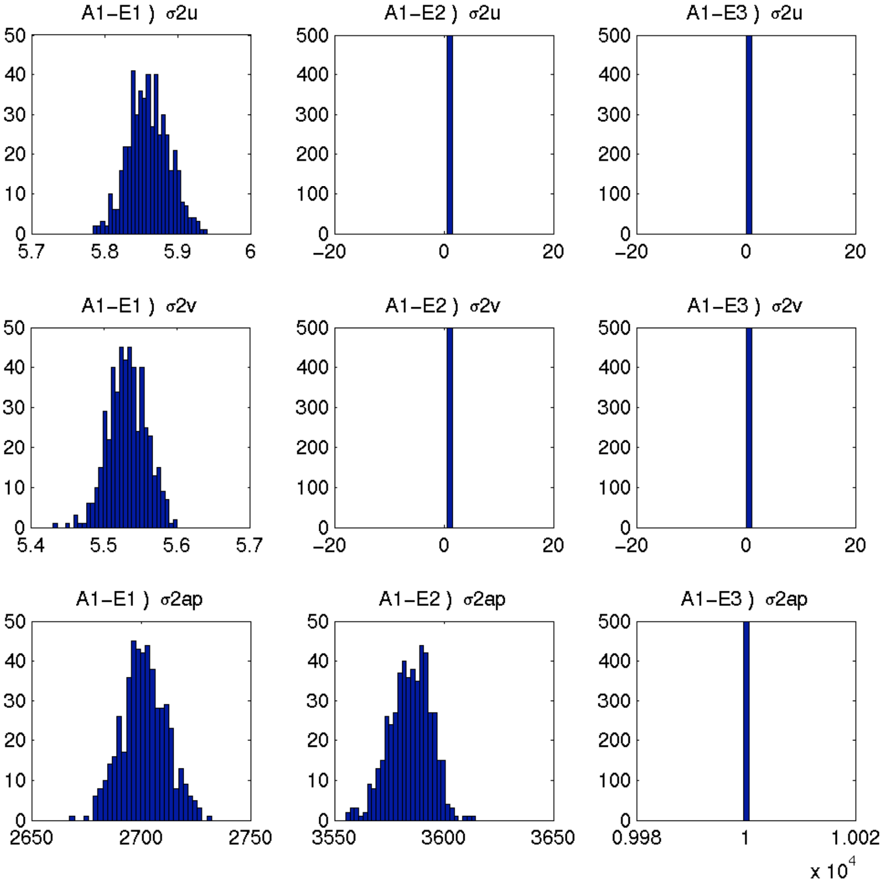


Figure 3.3: Posterior distribution of $\sigma_u^2[\frac{m^2}{s^2}], \sigma_v^2[\frac{m^2}{s^2}]$ and $\sigma_{ap}^2[\text{mbar}^2]$ for version A1-E1, A1-E2 and A1-E3. Note that in version A1-E2 σ_u^2 and σ_v^2 are fixed values; in version A1-E3 σ_u^2, σ_v^2 and σ_{ap}^2 are fixed values.

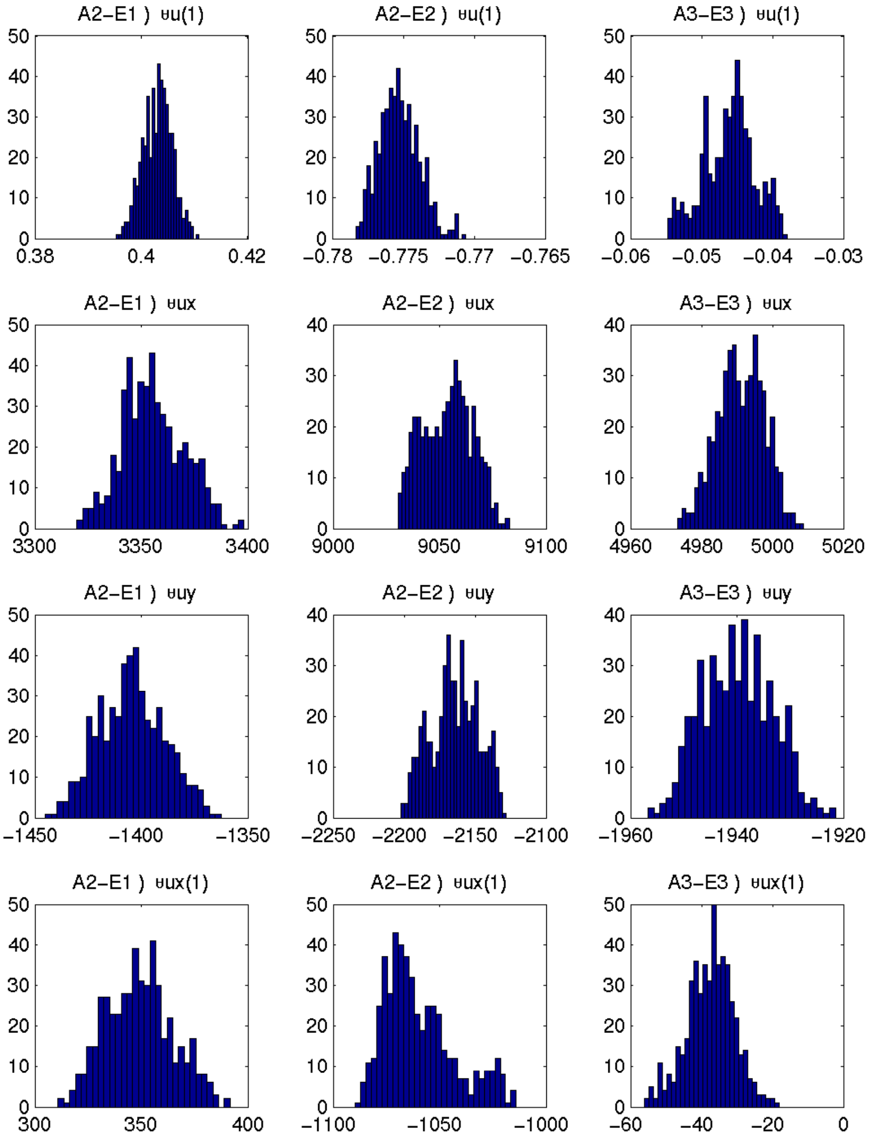


Figure 3.4: Posterior distribution of $\theta_{u(1)}, \theta_{ux}[\frac{m^3 s}{kg}], \theta_{uy}[\frac{m^3 s}{kg}]$ and $\theta_{ux(1)}[\frac{m^3 s}{kg}]$ for version A2-E1, A2-E2 and A2-E3.

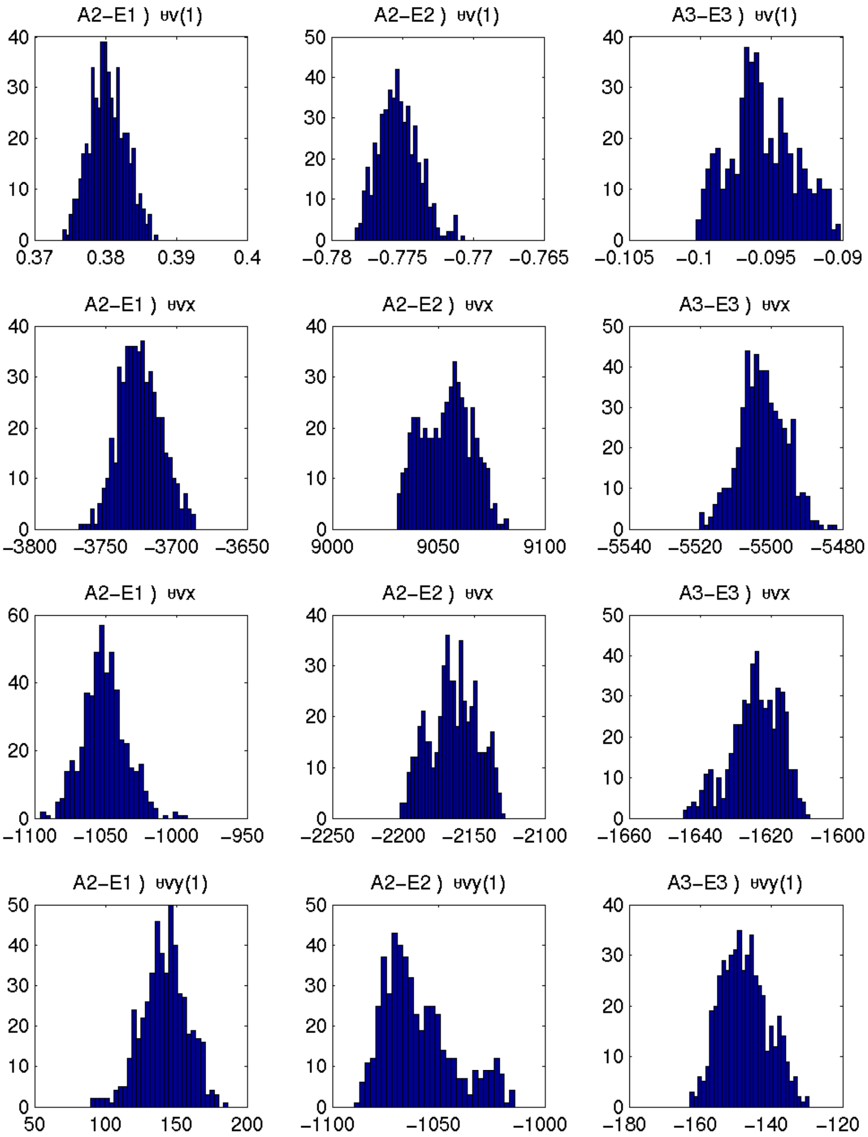


Figure 3.5: Posterior distribution of $\theta_{v(1)}, \theta_{vx}[\frac{m^3s}{kg}], \theta_{vy}[\frac{m^3s}{kg}]$ and $\theta_{vy(1)}[\frac{m^3s}{kg}]$ for version A2-E1, A2-E2 and A2-E3.

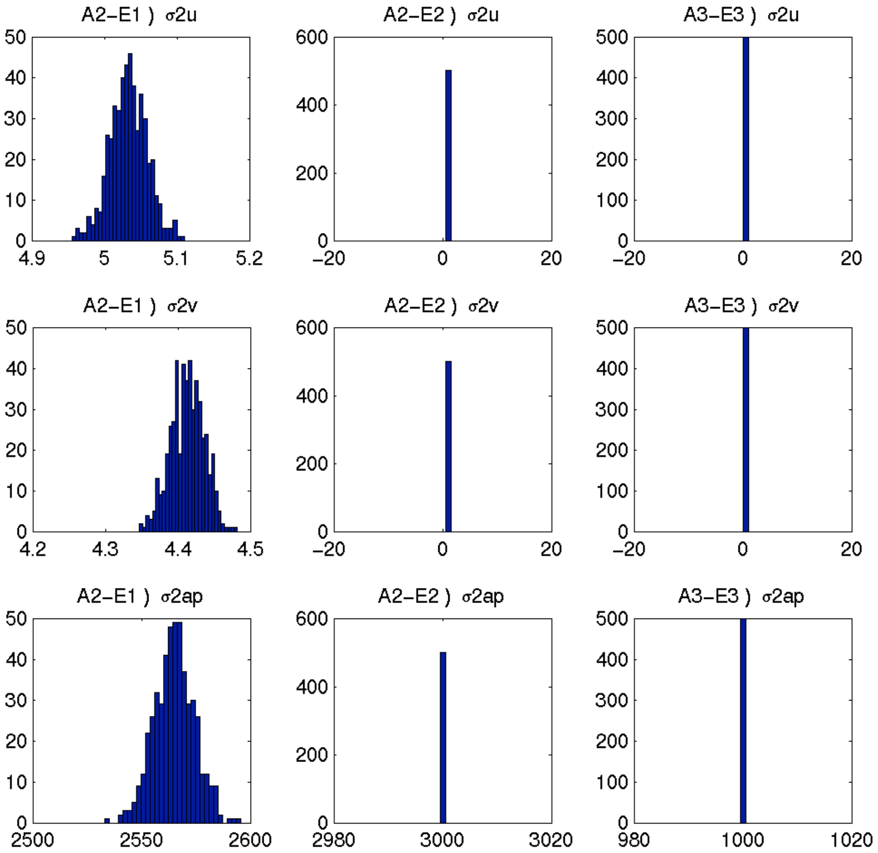


Figure 3.6: Posterior distribution of $\sigma_u^2[\frac{m^2}{s^2}], \sigma_v^2[\frac{m^2}{s^2}]$ and $\sigma_{ap}^2[\text{mbar}^2]$ for version A1-E1, A1-E2 and A1-E3. Note that in version A1-E2 and A1-E3 σ_u^2, σ_v^2 and σ_{ap}^2 are fixed values.

3.6.1 The wind posterior distributions

The posterior distribution of the wind and the pressure will be considered here in terms of the sensitivity to QuikSCAT data insertion and Kinetic Energy (KE) spectra.

Sensitivity to QuikSCAT data insertion

The time window of the BHM span 14 days in the past and 10 days in the future. During the fist phase, QuikSCAT data and ECMWF analysis are inserted as data stages, while in the next 10 days only ECMWF forecast products are used. The QuikSCAT data cover portions of the Mediterranean Sea three times a day leaving the time step at 12 am always empty. Since the insertion of scatterometer data affect the posterior distribution of all BHM implementations very similarly, here we show the results only for the A1-E3 version.

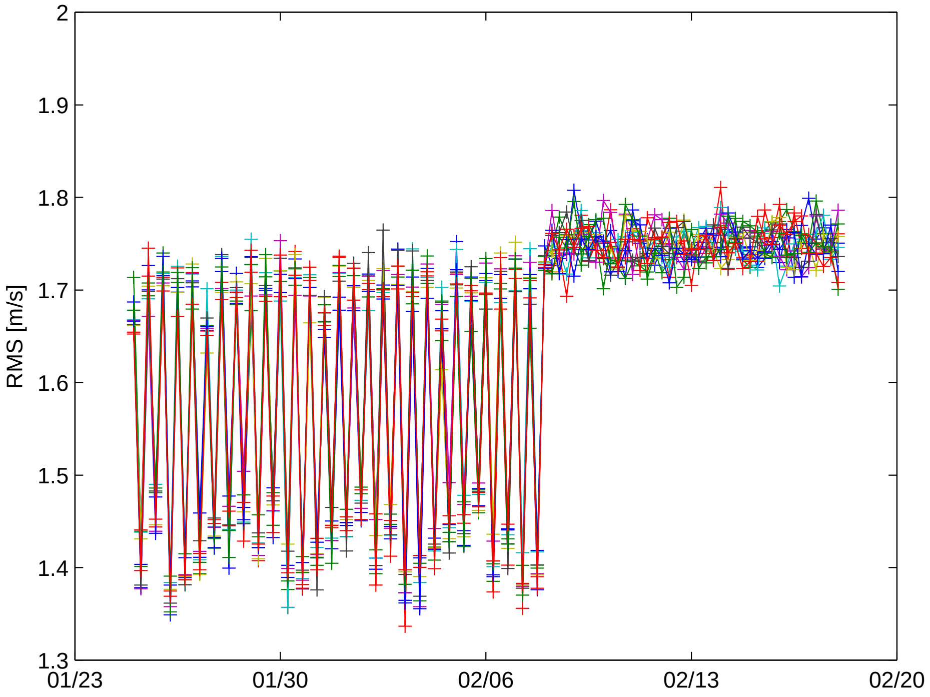


Figure 3.7: Root Mean Square difference between posterior mean and 10 realisations of the BHM for the meridional component of the wind field for the period from 25th January to 18th February 2005; each line in the plot refers to a single realisation of version A1-E3.

This feature is evident in the time evolution of the overall variability of the wind realisations. Figure 3.7 shows the standard deviation of the 10 members around their posterior mean for version A1-E3. Two different periods are evident on the plot. During the first 14 days the posterior spread oscillates between 1.4 and 1.9 [m/s], then it remains almost constant at 1.9 [m/s]. The oscillation is the footprint of QuikSCAT data insertion. The BHM wind uncertainty doesn't increase with the leading time during the forecast period. This feature do not mimic the pattern that was observed in the comparison between EMCWF analysis and forecast (see figure 3.1) and constitute a direction of further development of the BHM. However this simple implementation of the BHM forecast do not require any ad hoc assumption about the behaviour of the wind forecast error and constitutes a reasonable initial choice.

Figures 3.9 and 3.10 show two snapshots of the surface wind for February 3rd 2005 at 12.00 and 18.00 UTC as it is seen by the ECMWF analysis, the QuikSCAT data, the BHM A1-E3 posterior mean and 10 realisations. A mistral event is active in the Gulf of Lion region with winds stronger then 20 m/s. At 12.00 there are no QuikSCAT data available, while at 18.00 a portion of the Gulf of Lion region is covered with satellite data. The spread of the 10 realisations drawn from the posterior distribution is larger where QuikSCAT data are not present.

The insertion of the QuikSCAT data in the BHM is a constraint for the wind posterior distribution. Two simulations of the BHM model were performed with and without scatterometer data. Figure 3.8 show the scatterplot of the BHM A1-E3 posterior distribution against the QuikSCAT data for the wind components U and V for the two experiments: the upper panels show the results for a run that was made without using scatterometer data, while the lower one refers an experiment for which the QuikSCAT data were inserted. The cloud of the wind distribution collapse when the scatterometer data are included in the analysis.

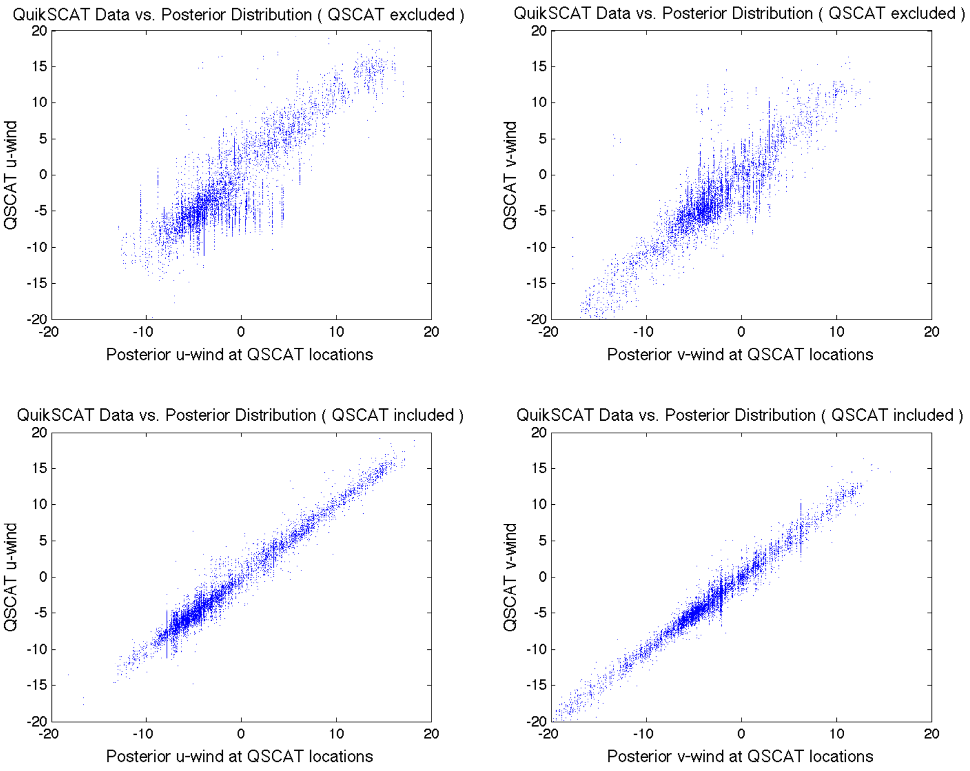


Figure 3.8: QuikSCAT wind at 18.00 UTC 3 February 2005 versus the posterior BHM A1-E3 at QuikSCAT location; (a) Data versus BHM A1-E3 posterior mean for u-wind when QuikSCAT data are not included in the data stage; (b) Data versus BHM A1-E3 posterior mean for u-wind when QuikSCAT data are included; (c) same as (a) for v component; (d) same as (b) for v component. All data are in $[m/s]$.

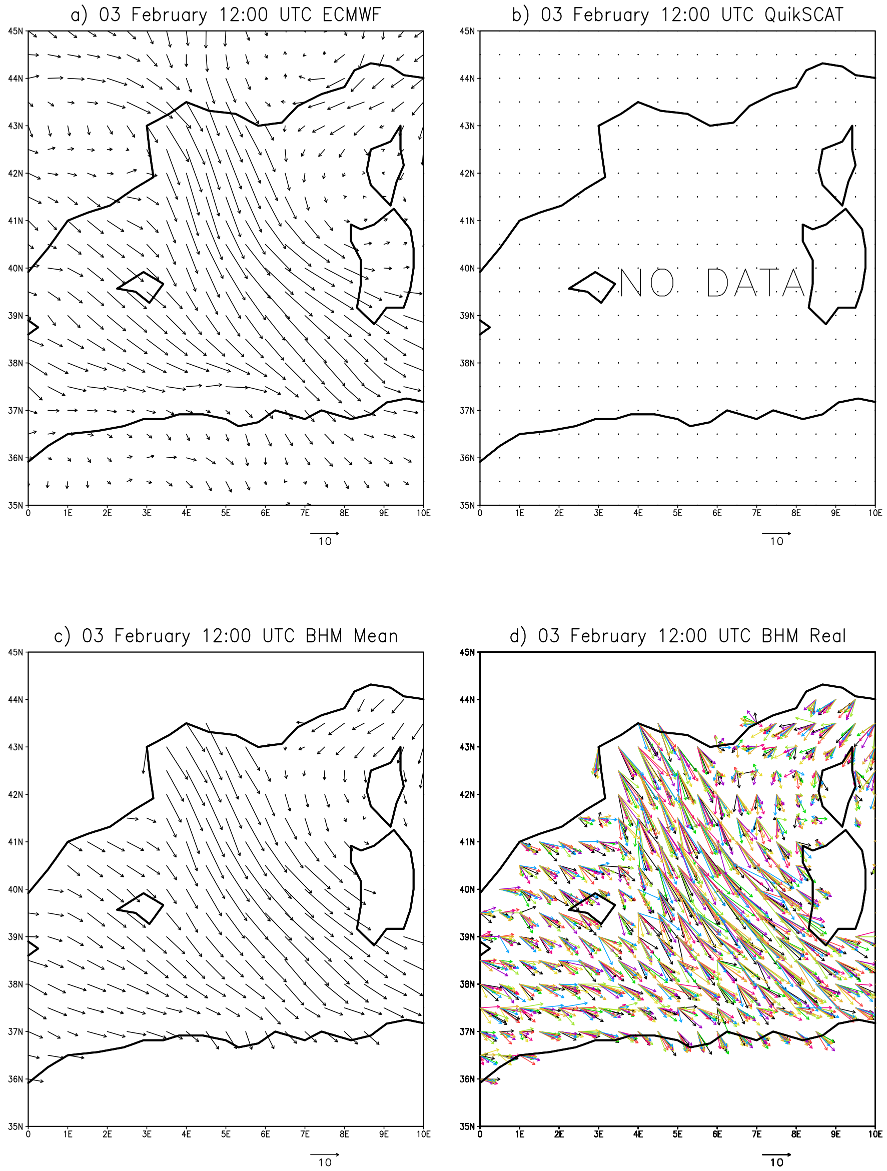
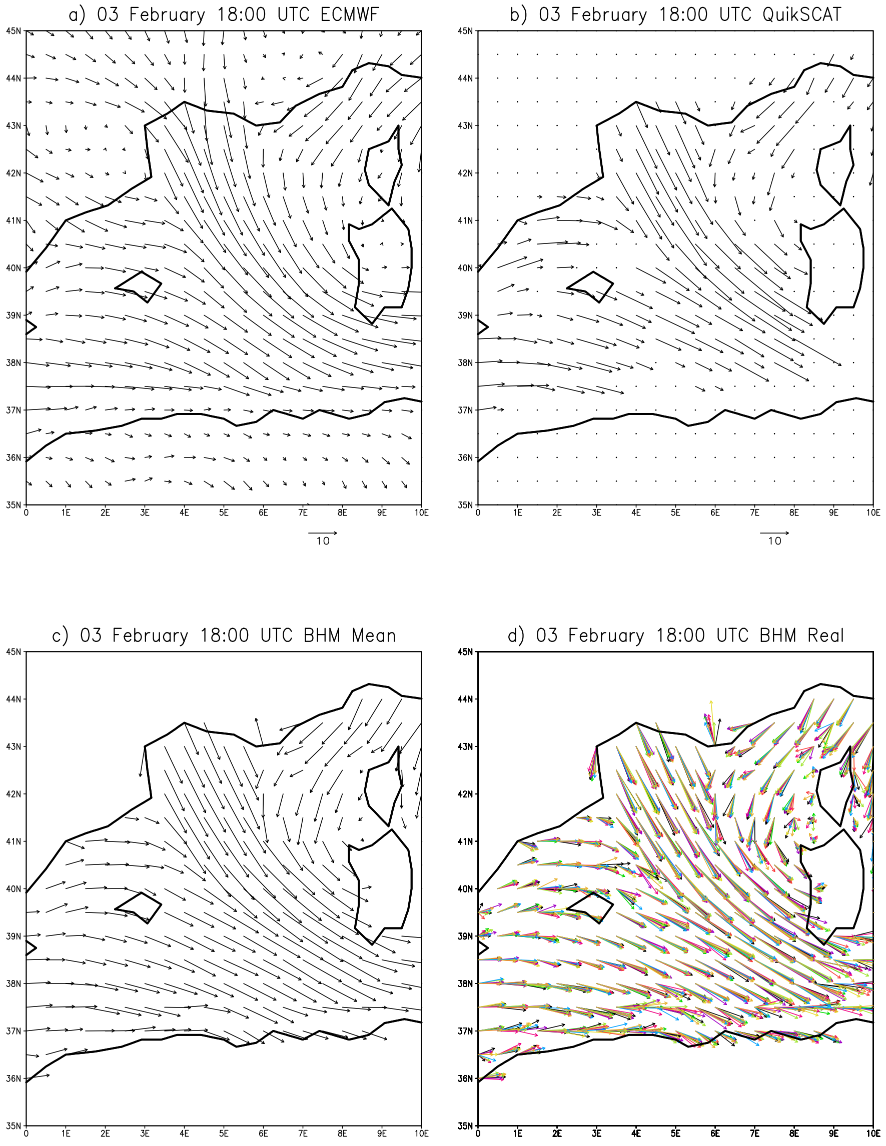

a) 03 February 12:00 UTC ECMWF
b) 03 February 12:00 UTC QuikSCAT
NO DATA
c) 03 February 12:00 UTC BHM Mean
d) 03 February 12:00 UTC BHM Real
45N
44N
43N
42N
41N
40N
39N
38N
37N
36N
35N
0
1E
2E
3E
4E
5E
6E
7E
8E
9E
10E
10

Figure 3.9: Wind snapshot 03/02/2005 12:00. Up-left panel: ECMWF analysis. Up-right panel: QuikSCAT data (empty). Low-left panel: BHM A1-E3 posterior mean. Low-right panel: 10 BHM A1-E3 realisations.


a) 03 February 18:00 UTC ECMWF

b) 03 February 18:00 UTC QuikSCAT

c) 03 February 18:00 UTC BHM Mean

d) 03 February 18:00 UTC BHM Real

Figure 3.10: Wind snapshot 03/02/2005 18:00. Up-left panel: ECMWF analysis. Up-right panel: QuikSCAT data. Low-left panel: BHM A1-E3 posterior mean. Low-right panel: 10 BHM A1-E3 realisations.

The BHM Energy Spectra

The energy spectra for QuikSCAT wind are computed along 40 consecutive wind retrieval lines of 1000 km length (with a resolution of 25 km). These lines are equally distributed in the south part of the western and eastern Mediterranean Sea (note that the northern part of the basin is not wide enough to extract sufficiently long lines of scatterometer data).

The ECMWF and BHM winds are not interpolated on the location of QuikSCAT data. The KE spectra are computed along a series of 20 longitudinal lines spaced by 0.5 ° that cover the same regions of the Mediterranean Sea that are used for the scatterometer data analysis.

Each line is transformed in the Fourier space after removing the along-track mean and then it is tapered and treated with a Hanning window (Press et al. 1986 [79]) to reduce the spectral leakage. The smallest spatial scale, i.e. Nysquist scale, that is resolved is 50 km for the QuikSCAT data and 90 km for the ECMWF and BHM winds. The largest spatial scale is about 900 km for all cases. An explanation of the method used to computed the KE spectra can be found in Milliff 2004 [64].

The KE spectrum for the QuikSCAT data (solid green line in figures 3.14 and 3.13) shows an approximate power-law relation with exponent $p = -2$. Beyond 100 km we note the QuikSCAT spectrum flattens suggesting a possible concentration of energy at the smallest resolvable scales. However caution must be taken in evaluating this result since the high wave-number tail of the spectra might affected by spurious signals due to spectral windowing and tampering.

The ECMWF analyses (dashed green line in figures 3.14 and 3.13) departs from the theoretical power-law relation at a spatial scale equivalent to 500 km. The energy decay that affect the ECMWF is severe; the KE content at 100 km is two order of magnitude less then what observed for the QuikSCAT winds.

The energy spectra of BHM version A1-E1, A1-E2 is shown in figure 3.11 and 3.12. Even in neither of the BHM versions is able to reproduce the exact power-slope of the QuikSCAT wind data, the BHM winds clearly represent an improvement with respect to the ECMWF analyses, at least for what concern the posterior mean. The wind realisations of version A1-E1 and A2-E1 show a flat spectra of high KE below the spatial scale of 200 km that is a clear indication of a large white-noise input in the BHM winds. An improvement of the kinetic energy spectra for the wind realisations is observed in version A1-E3 and A1-E2 (see figure 3.13 and 3.14).

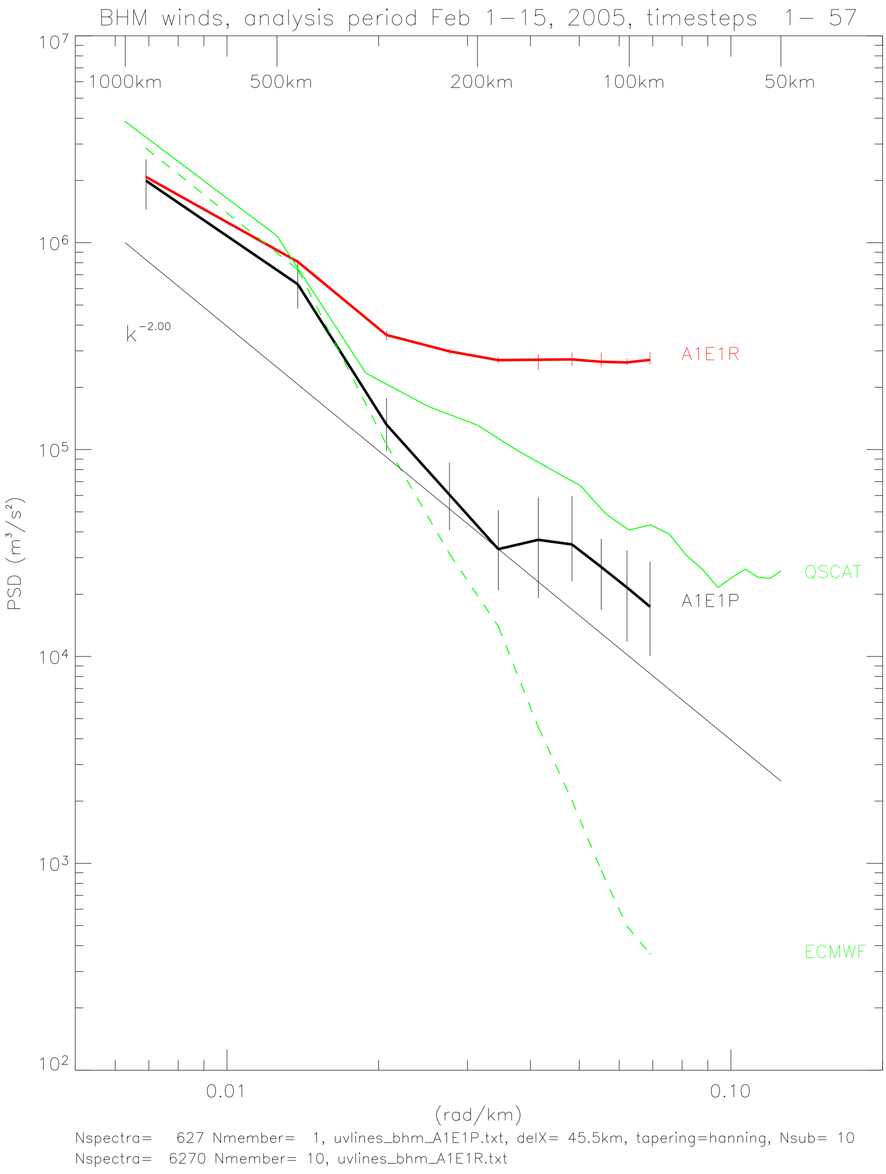


Figure 3.11: Average kinetic energy vs. wavenumber spectra for QuikSCAT (green solid line), ECMWF (green dashed line) and BHM A1-E1 expectation mean (red solid line) and BHM A1-E1 realisations (black solid line) for period 25 January to 7 February 2005.. Power spectral density (PSD, the ordinate) plotted against spatial wavenumber (k, the abscissa) . The spatial scales corresponding to selected wavenumbers are noted on the horizontal axis at top. A reference slope for a power law relation $PSD \sim k^{-2}$ is also shown in dashed black line. Confidence intervals are shown for BHM spectra.

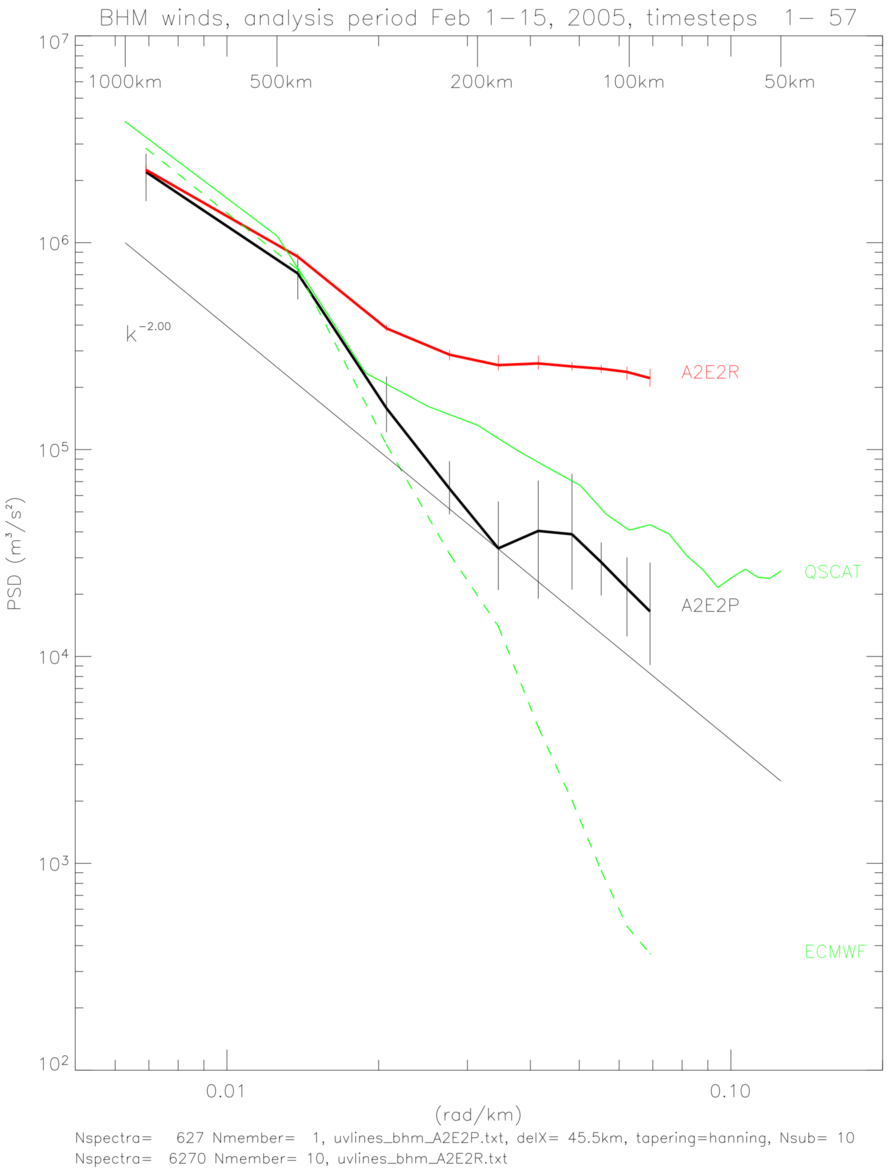


Figure 3.12: Average kinetic energy vs. wavenumber spectra for QuikSCAT (green solid line), ECMWF (green dashed line) and BHM A2-E1 expectation mean (red solid line) and BHM A2-E1 realisations (black solid line) for period 25 January to 7 February 2005.. Power spectral density (PSD, the ordinate) plotted against spatial wavenumber (k, the abscissa) . The spatial scales corresponding to selected wavenumbers are noted on the horizontal axis at top. A reference slope for a power law relation $PSD \sim k^{-2}$ is also shown in dashed black line. Confidence intervals are shown for BHM spectra.

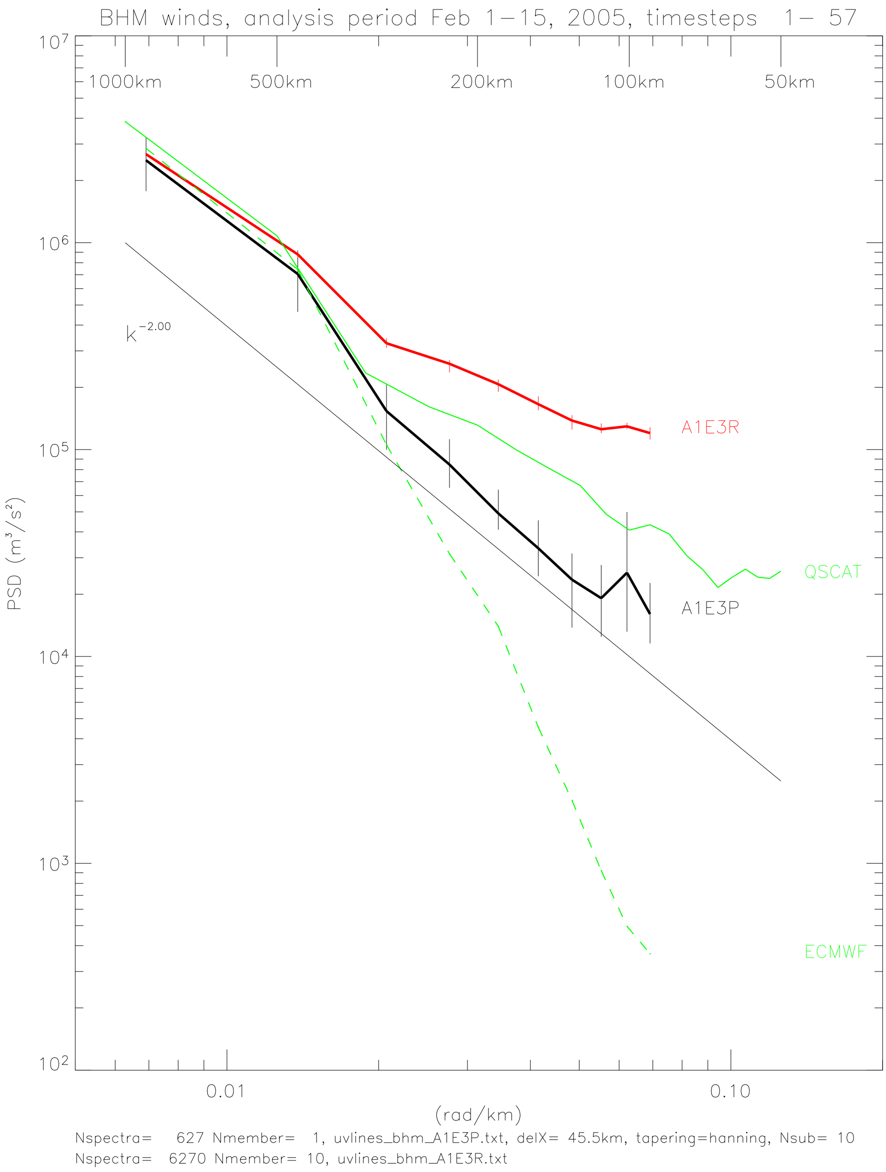


Figure 3.13: Average kinetic energy vs. wavenumber spectra for QuikSCAT (green solid line), ECMWF (green dashed line) and BHM A2-E3 expectation mean (red solid line) and BHM A2-E3 realisations (black solid line) for period 25 January to 7 February 2005.. Power spectral density (PSD, the ordinate) plotted against spatial wavenumber (k, the abscissa) . The spatial scales corresponding to selected wavenumbers are noted on the horizontal axis at top. A reference slope for a power law relation $PSD \sim k^{-2}$ is also shown in dashed black line. Confidence intervals are shown for BHM spectra.

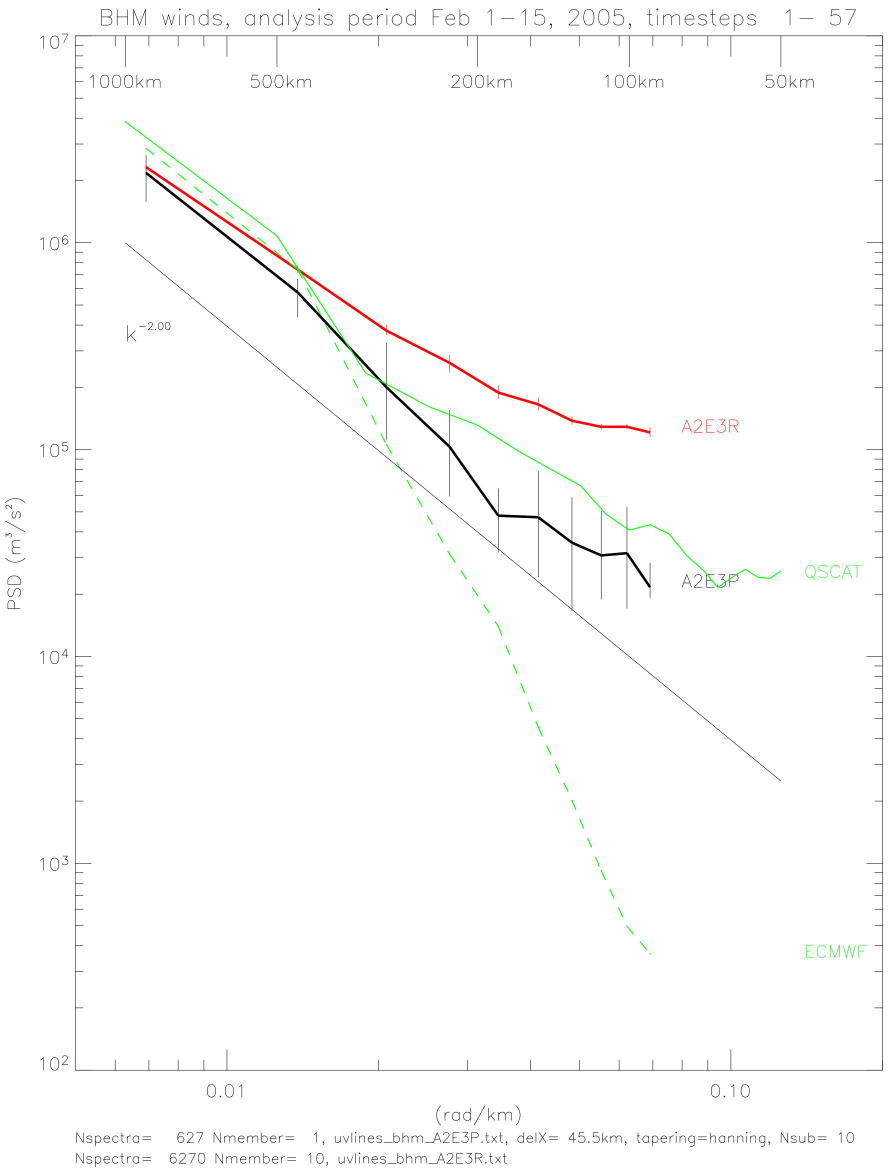


Figure 3.14: Average kinetic energy vs. wavenumber spectra for QuikSCAT (green solid line), ECMWF (green dashed line) and BHM A1-E3 expectation mean (red solid line) and BHM A1-E3 realisations (black solid line) for period 25 January to 7 February 2005. Power spectral density (PSD, the ordinate) plotted against spatial wavenumber (k, the abscissa) . The spatial scales corresponding to selected wavenumbers are noted on the horizontal axis at top. A reference slope for a power law relation $PSD \sim k^{-2}$ is also shown in dashed black line. Confidence intervals are shown for BHM spectra.

3.7 Summary

A Bayesian Hierarchical Model was developed to produces estimates of winds as a combination of ECMWF and scatterometer winds. The BHM has been built with different process models and representation of model equation errors. The approximation of the physical laws or the error representation has a strong impact on the final results.

Version A1-E1 and A2-E1 are the easiest implementation of the BHM that were tested for what concern the parametrization of model equation errors. The results for the wind posterior mean show that both models are able to retain most of the energy content that is missing in the ECMWF analyses. However the simple parametrization of model error introduces a large white-noise signal that affect the wind realisations; the flat energy spectra for the wind realisation of version A1-E1 and A2-E1 is typical of a random process.

The wavelet implementation of model error allow to significantly reduces the white-noise level that affect the wind realisation in version A1-E3 and A2-E3. The usage of wavelets makes possible a proper scale separation between the large scale, prevalently geostrophic atmospheric flow, and the small scale of the error. Furthermore the posterior distribution for the equation parameters of version A1-E3 and A2-E3 mimics the distribution observed for the simplest model A1-E1.

Based on this consideration we identify in version A1-E3 and A2-E3 the best implementation of wind BHM. Since the result of these two BHM are very similar for what concerns the posterior distribution of the equation parameters and the energy spectra of the wind realisations, we select the easiest implementation A1-E3 as the BHM set-up to be applied in the contest of ocean ensemble forecasting.

However we note that the presence ofa major drawback that still need to be addressed. The spread of the BHM winds do not growth with the forecast time but remains constant at a relatively small value of 1.8 $[m/s]$. The BHM developed so far allows to represent the forcing error in the past, but does not include any sensible information about the error evolution in the prediction phase. A direction of development is constituted by the integration of the stochastic forcing produced by NWP system, such the ensemble forecast of ECMWF, in the forecast wind data stage.

Chapter 4

Ocean Ensemble Forecasting Part I: the methodology

A new methodology is being devised for ensemble ocean forecasting using distributions of the surface wind field derived from a Bayesian Hierarchical Model (BHM). The ocean members are forced with samples from the posterior distribution of the wind during the assimilation of satellite and in-situ ocean data. The initial condition perturbations are then consistent with the best available knowledge of the ocean state at the beginning of the forecast.

An experimental array is being devised to test the BHM-ocean ensemble against traditional techniques that include initial condition random perturbation and the direct application of an ensemble of winds produced by the European Centre for Medium-range Weather Forecasting (ECWMF). Two implementation of the MFS ocean model were used to perform the ensemble experiments; the first is a high resolution eddy resolving set up, the latter is a low resolution model.

The study period is February 2005 when the maximum number of observation were available in the Mediterranean Sea including XBT, Argo floats and satellite measurements. Strong Mistral events were recorded during the months of January and February making this period particularly suitable to test the BHM-Wind ensemble method.

4.1 Method Description

4.1.1 BHM-wind Ocean Ensemble Method

The stochastic forcing produced by the BHM described in Chapter 3 is used here to generate an ensemble of ocean forecasts. A set of wind realisations sampled from the BHM posterior distribution is used to force an ensemble of ocean runs. This approch is strait-forward and relies on the important role that the wind plays on the circulation of the Mediterranean Sea. There is a large literature on this topic. Back in the 1974 Moskalenko [69] showed that the wind stress and its curl are responsible for the formation of basin-scale gyres. Malanotte-Rizzoli and Bergamasco (1989) [57] reproduced the eastern basin circulation using a model forced by realistic winds and heat fluxes. Pinardi and Navarra (1993) described the correlation between the sub-basin structures and the curl of the wind forcing. Zavatarelli and Mellor (1995) [96] and Roussenov et al. (1995) [82] demonstrated that the major control on the Mediterranean circulation is provided by the combined effect of wind and thermohaline forcing. Recently Molcard et al. (2002) [68] demonstrated that it is possible to reproduce the major patterns of the Mediterranean circulation using the wind alone in a flat topography model. Here we want to demonstrate that the uncertainty in the wind forcing has an impact on short term ocean forecast.

The working hypothesis is that all the ocean model uncertainty is due to a misrepresentation of the wind forcing both in the analysis and in the forecast period.The uncertanty in the wind forecast is a well know problem in literature (Epstein 1963 [31], Lorenz 1993 [54], Buizza and Palmer 2003 [8]) and is widely recognised that after 3 or 4 days the wind forecast looses most of its skill. Thus building a set of statistically consistent wind estimates from ECMWF and scatterometer winds allows us to have a physically consistent representation of the wind error in the forcing fields of an ocean model.

The BHM ocean ensemble method is tailored on the MFS operational scheme (Tonani et al. 2007 [91] and Dobricic et al. 2007 [27]). Each ocean member replicates a full cycle of assimilation and forecast.

For a forecast cycle starting at day j an ensemble of M members is initialised from a single restart file at day $j-14$. During the first 14 days each ocean member is forced with a BHM realisation and it experiences the assimilation of temperature, salinity and SLA data. Then the M ocean states at day j are consistent with all the available observations. The ensemble forecast is obtained continuing to force the M ocean run with BHM prediction winds.

The initial condition perturbation is defined to be the spread of the ocean ensemble at the day j, i.e. the uncertainty of the forecast initial condition. We will be referring to the ensemble spread as the standard deviation of the ensemble member around the ensemble mean.

4.1.2 ECMWF wind ocean ensemble method

The European Centre for Medium-range Weather Forecasting (ECMWF) runs a global Ensemble Prediction System (EPS). The EPS is one of the most successful prediction system and proved to be extremely useful in a wide range of applications (Buizza 2006 [8]). The capability of this product in generating an ensemble of ocean forecast is tested here.

The key feature of the EPS systems is the usage of singular vector to produce the initial condition perturbations . Singular vector identifies the perturbations with maximum growth rate for an energy norm in the first 48 hours of forecast (Lacarra and Talagrand 1988 [42], Farrel 1990 [34]). Small errors in the initial condition along this direction would amplify most rapidly and affect the forecast accuracy [9]. Since 2004, the operational implementation of the EPS includes 50 perturbed members plus a control forecast run at a horizontal resolution of approximately 100 km. The data are interpolated on a Gaussian grid at 0.5 degrees resolution.

Figure 4.1 shows the first 3 EOFs of the ECMWF surface wind ensemble for the time period considered in this study. The first EOF (panel a) shows a large scale wind pattern that modulates the intensity of a Mistral event occurring between February 10th - February 13th. The second EOF has the shape of a cyclone-anticyclone centred over central Mediterranean. The third EOF (panel c of figure 4.1) exhibits a topography modulation of the winds that interests the Gibraltar Strait and Sicily Channel regions.

Figure 4.2 shows the growth in amplitude of the standard deviation of the ECMWF 51 members around the ensemble mean. The ensemble spread perfectly mimics the behaviour the forecast errors as depicted by figure 1 of Chapter 3. We remind that the BHM wind spread is constant all along the 10 days of forecast (see figure 8 of Chapter 3).

There is no doubt that the ECMWF ensemble contains more information about the forecast error in winds then what can be extracted by statistical models based on past observations. However the ability of the ECMWF stochastic forcing to generate a large ocean response need to be verified.

The ocean ensemble is constructed forcing several ocean forecasts with different members of the ECMWF wind ensemble. Each ocean forecast starts from a single un-perturbed initial condition.

4.1.3 Initial Condition Random Perturbation Method

A "standard" initial condition perturbation technique is presented here. In the seventies Leith [45] conceived ensemble forecasting as a Monte Carlo approach based on the integration of numerical models starting from randomly perturbed initial conditions. In this application we perturb the prognostic field temperature and salinity. The T and S perturbations are:

$$T_p(x,y,z,t_0) = T(x,y,z,t_0) + p(x,y)\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij} f_{ij}(z) \tag{4.1}$$

$$S_p(x,y,z,t_0) = S(x,y,z,t_0) + p(x,y)\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} e_{ij} g_{ij}(z) \tag{4.2}$$

where T_p and S_p are the perturbed temperature and salinity and $p(x,y)$ is a 2D random field that is extrapolated in the vertical dimension using a set of precalculated vertical statistical modes. These modes, $f_{ij}(z)$ and $g_{ij}(z)$, are Empirical Orthogonal Functions (EOFs) and represent the largest vertical variability of temperature and salinity computed

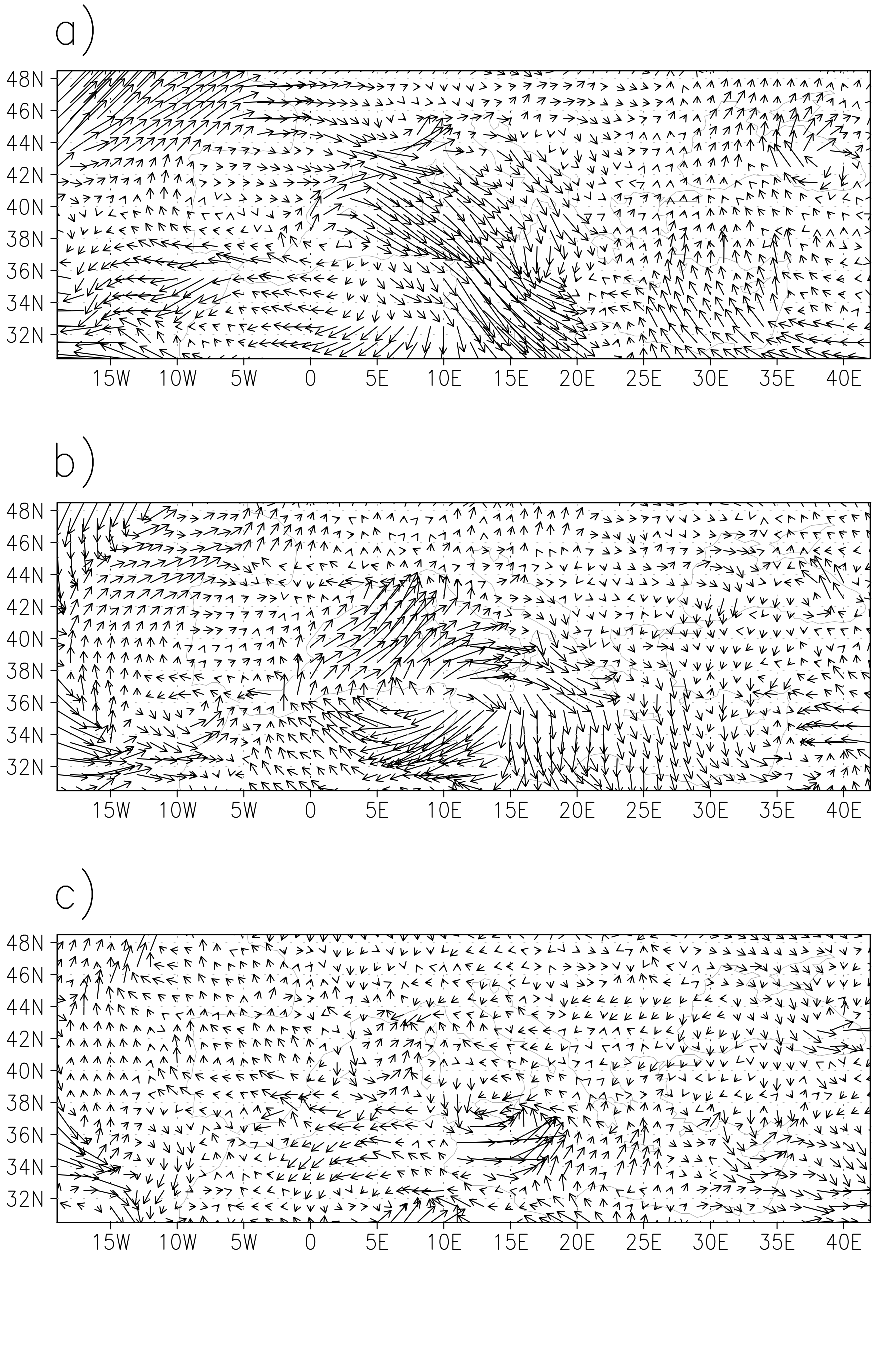


Figure 4.1: EOFs of the ECMWF ensemble winds computed over the 10 days of forecast from 8 to 18 February 2005; a) First EOF, Percentage of Explained Variance (PEV): 12%. b) Second EOF, PEV 7%. c) Third EOF, PEV 4%

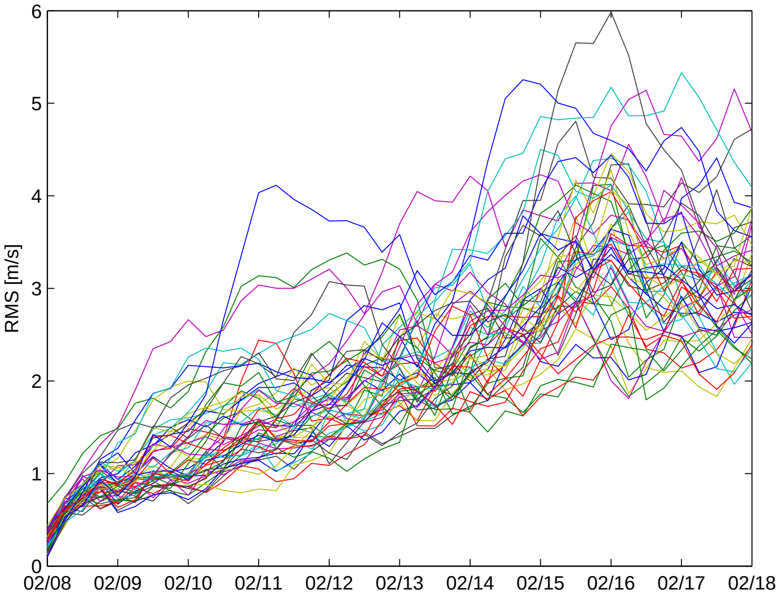


Figure 4.2: The growth in amplitude of the standard deviation (std) for the 10 days forecast of the 51 members of ECMWF ensemble. Each line represent a member of the ensemble. Units are in [m/s].

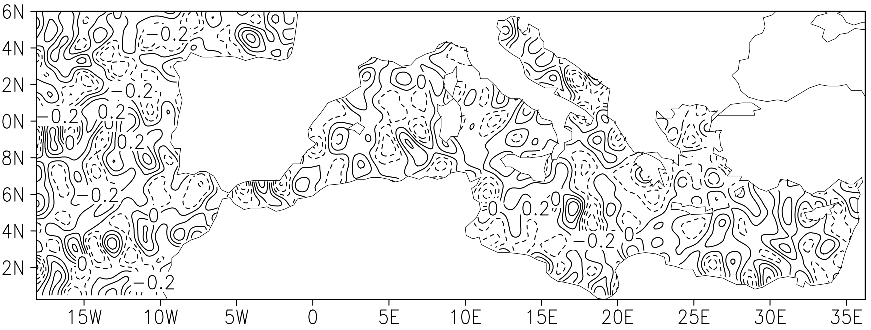


Figure 4.3: Example of a 2-dimensional random field $p(x, y)$ with a correlation radius of 30 km produced following the procedure describe in Evensen 2003 [32]. The contour interval is arbitrary.

over a 7 years-long simulation of the model itself (see Dobricic et al. [26] and Appendix B) . The first 20 modes of T and S variability are used (M=20) and each mode is weighted using the normalized percentage of explained variance e_{ij}. In the operation system, the EOFs were computed separately for 13 subdomains of the Mediterranean Basin (N=13), but for this experiment a single mean EOF is considered over the entire domain.

The pseudorandom field $p(x, y)$ is generated following the procedure proposed by Evensen (2003 [32]) . The mean of $p(x, y)$ is zero and the covariance, which determines the smoothness of the field, is specified a priori. A generator of pseudorandom numbers is used in the algorithm and it insures that different output will be produced given different input seeds. Figure 4.3 shows an example of the p field.

The initial perturbation of temperature and salinity was calibrated to simulate a standard deviation of 2 cm in the Sea Surface Height field. The random ensemble scheme is not constrained by any observation.

The ocean ensemble is produced running several ocean forecasts from a perturbed initial condition. Each ocean run is forced by the deterministic ECMWF wind forecast.

4.1.4 Experimental Design

The three perturbation methods are tested on the same forecast cycle that spans between February 8th to February 18th 2005. A control run was obtained replicating the operational forecast; the control run starts from the previous day analysis and it is forced by ECMWF deterministic fields. Each method was tested with high (MFS1671) and low (MFS471) resolution models described in appendix B.

Then the experiment array is composed by 6 ensembles that we will be compared on the basis of their spread, mean error and predictable patterns (see table 4.1)

Table 4.1: Experiment array. See text for a description of perturbation methods.

	Ocean model	Initial Condition perturbation	Wind Forcing perturbation	Ensemble size
BHM-MFS1671	high resolution	yes (section 1.1.1)	yes (section 1.1.1)	10
BHM-MFS471	low resolution	yes (section 1.1.1)	yes (section 1.1.1)	100
EC-MFS1671	high resolution	no	yes (section 1.1.2)	10
EC-MFS471	low resolution	no	yes (section 1.1.2)	51
ICRP-MFS1671	high resolution	yes (section 1.1.3)	no	10
ICRP-MFS471	low resolution	yes (section 1.1.3)	no	100

4.2 High resolution results

4.2.1 Ensemble Initial Conditions

The initial condition perturbations are shown for the 3 different perturbation method in figure 4.4. The initial condition spread of the BHM ensemble is concentrated in following circulation features (see Robinson et al. 1992 and Pinardi et al. 2005 [76] for a detail description of Mediterranean circulation):

- Algerian Current (AC); the AC is a boundary intensified jet-like current that transports Atlantic Water (AW) following the northern African continental slope. Large anticyclonic eddies of diameter of about 100 km detach from the mean current and have a lifetime that range between few months to several years (Millot and Taupier-Letage 2005 [67] and Testor 2005 [88]). During the period of this study a large anticyclonic eddy is located at $6°E$ and it has already separated from the Northern African coast. The wind perturbation effectively displaces the border of the eddy generating a large ensemble variance along the sharp density front.

- Atlantic-Ionian (AI) Current; the AC separates in two branches at the western tip of Sicily. The north branch enters the cyclonic circulation of the Thyrrenian Sea, while the second branch heads south in the Sicily Channel and becomes the Ionian-Atlantic stream. The AI current leaves the Sicily shelf plateau and undergoes a meandering process in the deep Ionian basin. The separation of the AC and the meandering of the AI stream are highlighted in the ensemble SSH spread. A barotropic signal in the esemble SSH standard deviation it is evident over the shallow African continental shelf.

- North African Current; in the Levantine basin the AI current takes the name of North African Current (NAC). The NAC flows along the African coast before bifurcating in the Mid Mediterranean Jet (MMJ). The ensemble SSH spread is concentrated in the boundary intensified portion of the NAC to the east of the Mersa-Matruh gyre.

The EC-ensemble is obtained without any initial condition perturbation but at the end of the first forecast day the ensemble standard deviation of SSH has reached an amplitude comparable with the BHM ensemble (panel c of figure 4.4). The spread concentration in the northern Adriatic Sea is key feature of the EC ensemble and it suggests that only the barotropic ocean circulation is affected by the ECMWF wind perturbation.

The initial condition of the ICRP ensemble reflects only the perturbation method and it is not informative about the ocean dynamics and its uncertainty (panel d of figure 4.4).

4.2.2 Ensemble spread during forecast time

The IC perturbations of the BHM-MFS1671 ensemble grows during the forecast time. We observe well defined structures of ensemble spread that interest the mesoscale circulation of the Mediterranean Sea (panel b of figure 4.5). The maximum spread in Sea Surface Height reaches the value of 6 cm that is comparable in amplitude with a typical model error that in a mean sense oscillates between 4-6 cm (Dobricic et al. 2005 [27]). There is a clear continuity between the ensemble spread at the beginning and end of the forecast (panel a of figures 4.5 and 4.4), the main difference being the disappearance of any swallow water signal in the SSH ensemble spread at the forecast end time.

The EC ensemble presents a reversed scenario. Most of the ocean spread is found on shelf areas, with maximum SSH standard deviation of 12 cm in the Northern Adriatic (panel c of figure 4.5). In the Algerian Current and in the Ionian Sea we observe weak ensemble variability that concerns the mesoscale circulation.

The ICRP ensemble shows a large ocean spread, up to 10 cm, that is not well organised around circulation structures, suggesting that 10 members are not sufficient to highlight the most relevant instability regions (panel d of figure 4.5). However three maxima of ensemble variability for the SSH can be identified in the Algerian current in the Ionian Sea and in the Levantine basin around the Mersa-Matru gyre. It should be noticed that both the initial perturbation and the final variability of the ensemble members in the costal areas are low. The applied perturbation scheme is sensitive to the depth of the water column. The perturbation of temperature and salinity results in volume anomalies that are integrated between 1000 m and the surface to obtain the perturbation on SSH. In shallow water ($z < 1000$) this integral is smaller being computed over a shorter interval.

In figure 4.6 we show the ensemble density standard deviation along a section at 5E. The ensemble variance vertical structure reveals interesting similarities between the ensembles, especially between the BHM and ICRP experiments (panel b and d). The EC ensemble presents a small vertical penetration of the wind perturbation (see panel c of the same figure) confirming the prevalent barotropic structure of the ocean response.

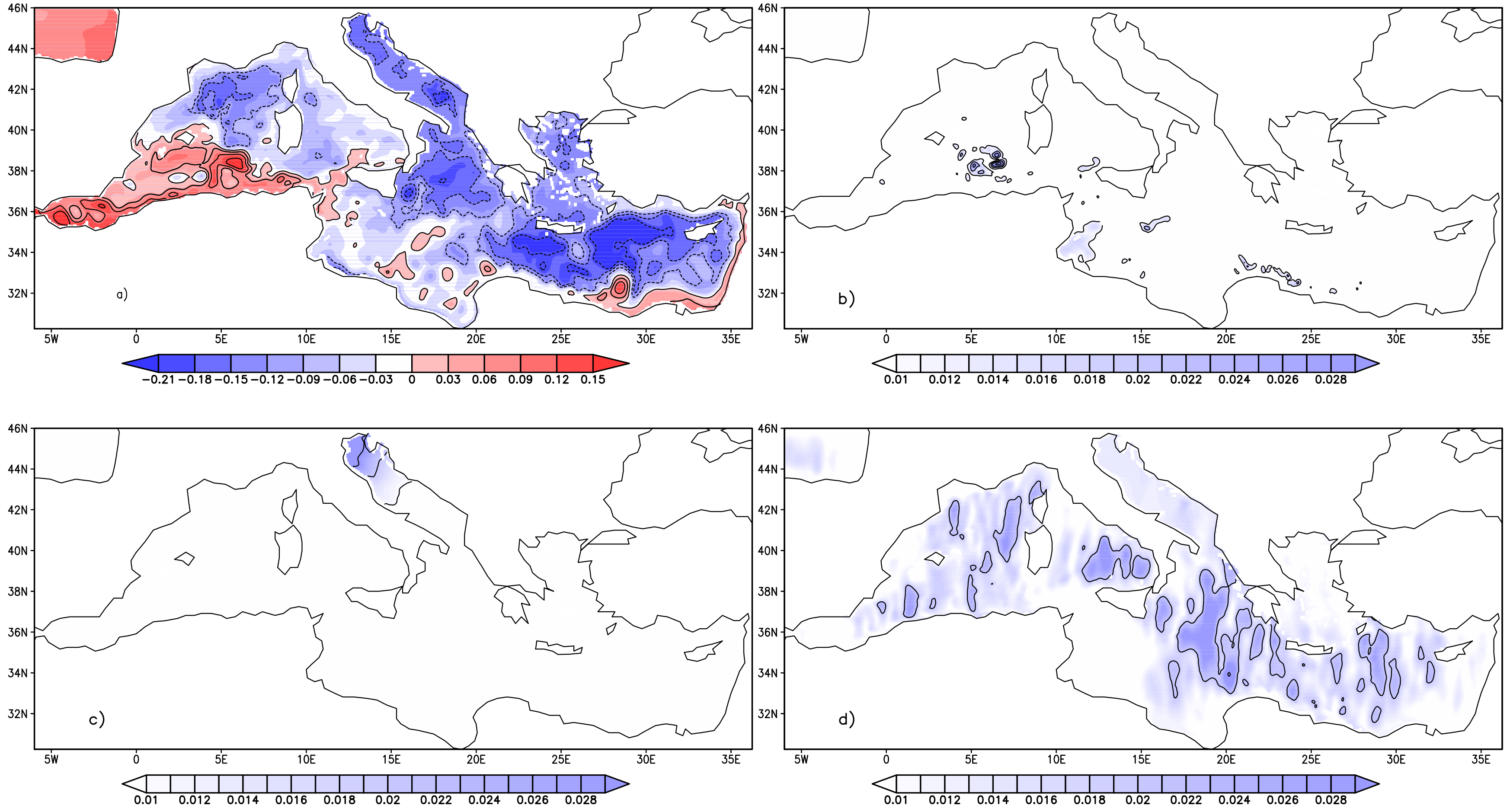


Figure 4.4: a) MFS1671 control run SSH [m] at February 8th 2005 (forecast initial condition, here called F0). The contour interval is 0.05 m. b) Standard Deviation [m] of BHM-MFS1671 ensemble at F0 . c) Standard deviation [m] of EC-MFS1671 ensemble at the end of the fist day of forecast (F0 + 1 day) . d) Standard deviation [m] of ICRP-MFS1671 ensemble at F0. The contour interval in b), c) and d) is 0.01 m.

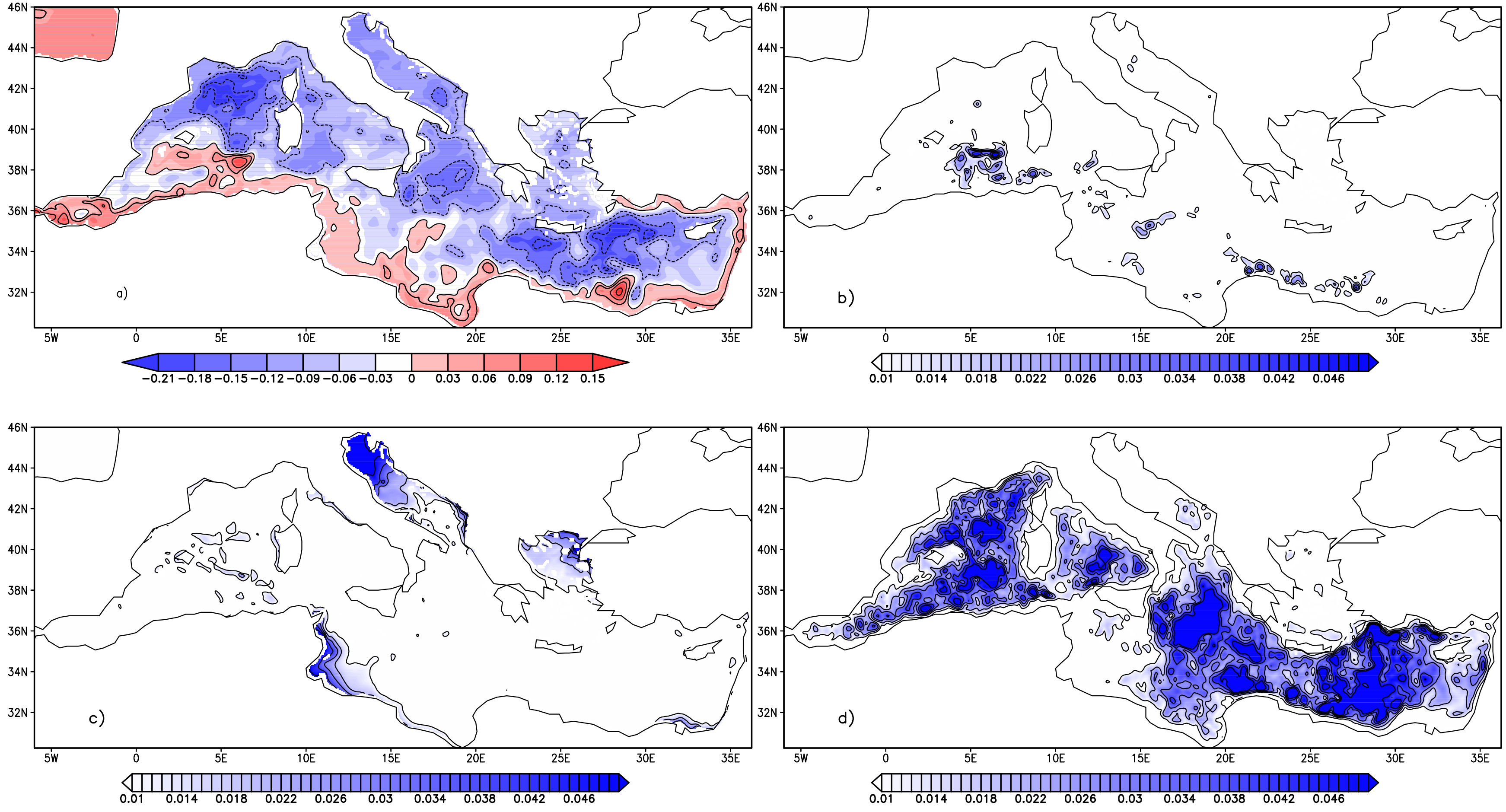


Figure 4.5: a) MFS1671 control run SSH [m] at February 18th 2005 (last day of forecast, here called F10). The contour interval is 0.05 m. b) Standard Deviation [m] of BHM-MFS1671 ensemble at F10 . c) Standard deviation [m] of EC-MFS1671 ensemble at F10 . d) Standard deviation [m] of ICRP-MFS1671 ensemble at F10. The contour interval in b), c) and d) is 0.01 m.

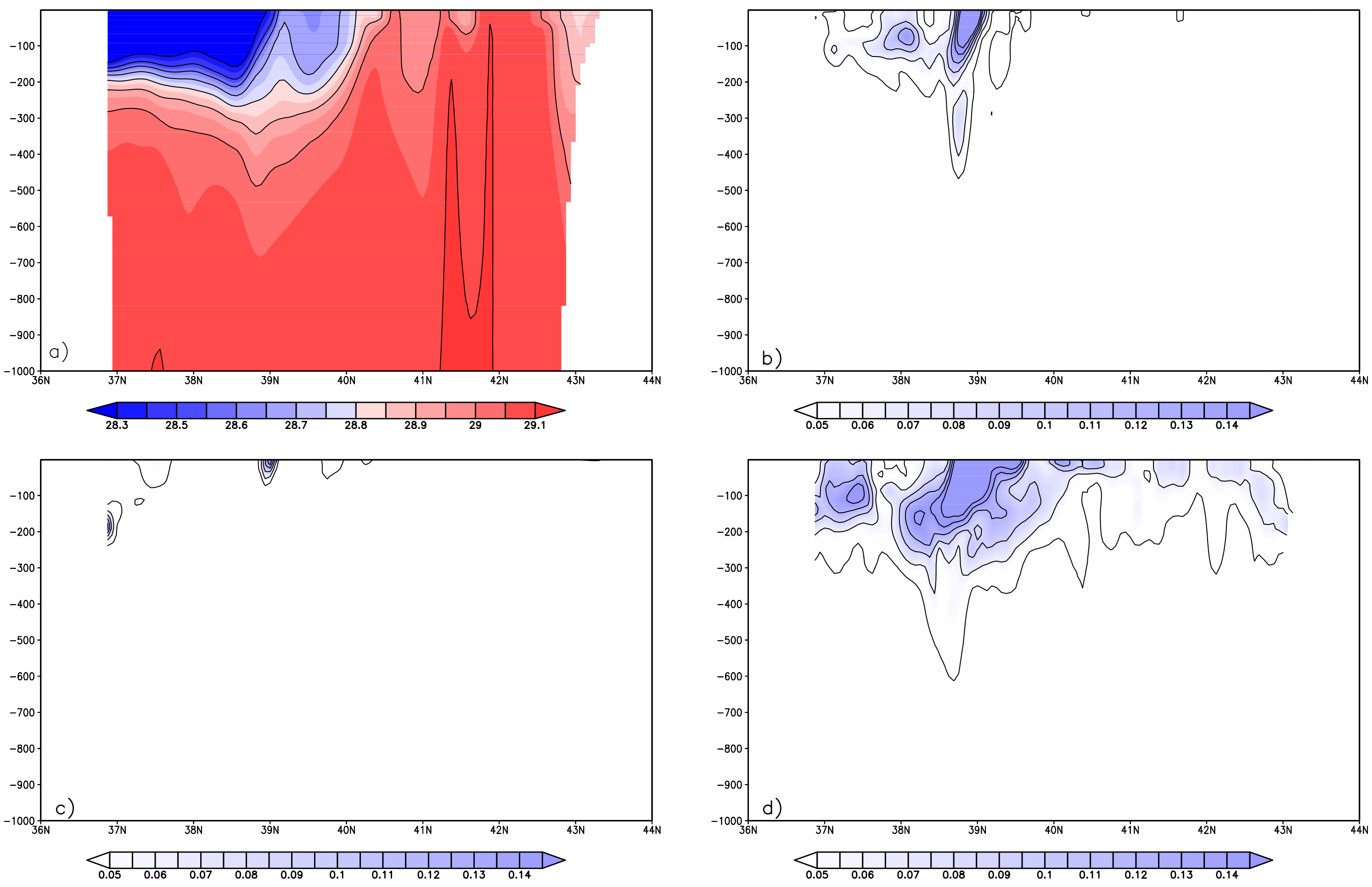


Figure 4.6: a) MFS1671 control run density $\sigma = \rho - 1000$ $[kg^3/m^3]$ at February 18th 2005 (last day of forecast, here called F10). The contour interval is 1 σ. b) σ standard deviation of BHM-MFS1671 ensemble at F10 . c) σ standard deviation of EC-MFS1671 ensemble at F10 . d) σ standard deviation of ICRP-MFS1671 ensemble at F10. The contour interval in b), c) and d) is 1 σ.

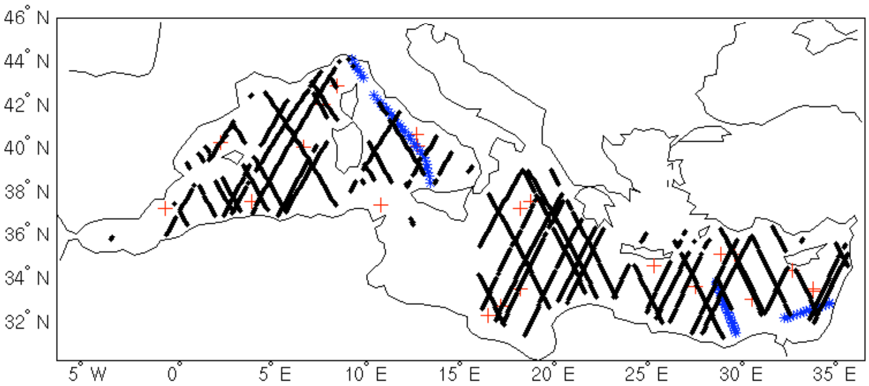


Figure 4.7: Observation locations from 8 to 18 February 2005 in the Mediterranean Sea. Blue points are XBT temperature profiles, red crosses refer to ARGO temperature and salinity profiles. Black points refer to Sea Level Anomaly measurements from satellite altimetry.

4.2.3 Comparison with data

So far we have being looking at the properties of the ensemble variance, trying to understand the ocean response to a set of perturbations of different nature. But the question of how effective the ensembles are to reduce to forecast error has being left aside. The presence of buoys and satellite data makes it possible to check the forecast error reduction. It is believed that if the ensemble method is valid it should reduce forecast error with respect to the deterministic single forecast (Leith 1971 [44]). Here we evaluate this comparing the ensemble mean state with observations.

In the 10 days that are covered by this hind-cast experiment, 22 ARGO profiles of temperature and salinity and 77 XBT temperature profiles were taken in various areas of the Mediterranean Sea. A set of 55 Sea Level Anomaly tracks are also available for this period for a total of 1761 points. Figure 4.7 shows the space distribution of the data during the 10 days of forecast.

Table 4.2 presents the 10 day averaged forecast error for the control run, forced with ECMWF winds, and the three high resolution ensemble experiments. The results are presented for temperature, salinity and model error. Temperature and salinity skill scores are further divided considering three groups; the first 30 m, the first 100 m and the whole water column.

- BHM: the ensemble mean presents similar skill score of the control run. This is probably due to the fact that the ensemble spread is very limited in small regions.
- EC: the ensemble mean presents a slightly smaller RMS on temperature but show a 0.4 cm increment in the Sea Surface Height RMS error.
- ICRP: the ensemble mean is worse then the control run for all the observed variables. The temperature ensemble mean error is 20% worse then the control run and the salinity ensemble mean is 10% worse then the reference run.

The BHM method is not very effective in reducing the forecast error but the EC method worsen the scores for SLA. The worst results are observed for the random perturbation ensemble that also proved to be the most successful method in exiting a large ocean response.

Table 4.2: MFS1671 ensemble and control rms errors. Units are [$^\circ C$] for temperature, [psu] for salinity and [m] for SLA.

	Temperature			Salinity			SLA
	All	30 m	100 m	All	30 m	100 m	
Control	0.37	0.41	0.4	0.18	0.28	0.22	0.04
BHM Ensemble Mean	0.36	0.41	0.34	0.18	0.28	0.23	0.04
EC Ensemble Mean	0.36	0.39	0.39	0.18	0.28	0.22	0.05
ICRP Ensemble Mean	0.43	0.50	0.48	0.20	0.30	0.25	0.05

4.3 Low resolution results

The high resolution of the MFS ocean model makes a Monte Carlo approach computationally prohibitive unless the ensembles size is limited. The possibility to enlarge the number of ensemble members arises if we consider a low resolution version of the ocean model. A 1/4 of a degree set-up of MFS model was implemented reducing by a factor 16 the dimension of the state vector. The MFS471 ocean model fits a single CPU of a standard PC, making possible to run multiple experiments on cluster machines without the need of any interface between nodes. This approach is optimal for the usage of High Throughout Computing systems, such the Grid. A technical efficiency and reliability investigation was conducted on the Italian Grid.it system, see Pinardi et al. (2007 [74]) and Appendix C.

Generally we observe that the MFS471 EC and ICRP ensembles are qualitatively similar to their high resolution homologues (figure 4.5 and figure 4.9):

- The BHM ensemble method is highly sensitive to the resolution of the ocean model since the BHM winds prevalently act on the meso-scale circulation that is poorly represented by the low resolution model. The spread maxima both in the initial condition and at the forecast end are localised similarly in the two BHM ensembles but the signal amplitude is weakened in the MFS471 case (panel b of figure 4.8 and 4.9). The vertical section at 5E (figure 4.10) shows that the BHM forcing perturbation propagates less in the depth then the high resolution experiment.

- The EC ensemble spread after the first day of forecast (panel b of figure 4.8) presents a clear variance maximum on the Northern Adriatic and little if any impact on the rest of the Mediterranean basin. At the end of the forecast a small portion of the initial variance has been transferred to the meso-scale circulation, but the most evident signal is still concentrated in shallow regions (panel c of figure 4.9).The density spread along section 5E, see figure 4.10, is more baroclinic but weak as usual.

- The ensemble size is a crucial factor for the random perturbation scheme. The low resolution model allows us to easily run 100 members making possible to fully explore the probability distribution that simulates the uncertainty of the initial state. Panel d of figure 4.8 show the initial condition spread for the SSH field. We clearly recognise that the initial perturbation is homogenous in region of similar depth and this is very unrealistic. At the end of the forecast the SSH spread is generally localised in region of high dynamical circulation, see panel c of figure 4.9 with maxima in the Algerian current, along the Atlantic current in the Ionian Sea and around the Mersa-Matruh gyre in the Levantine basin. The section at 5E (see panel d of figure 4.10) reveals a vertical structure of the density spread that is similar to the high resolution ensemble.

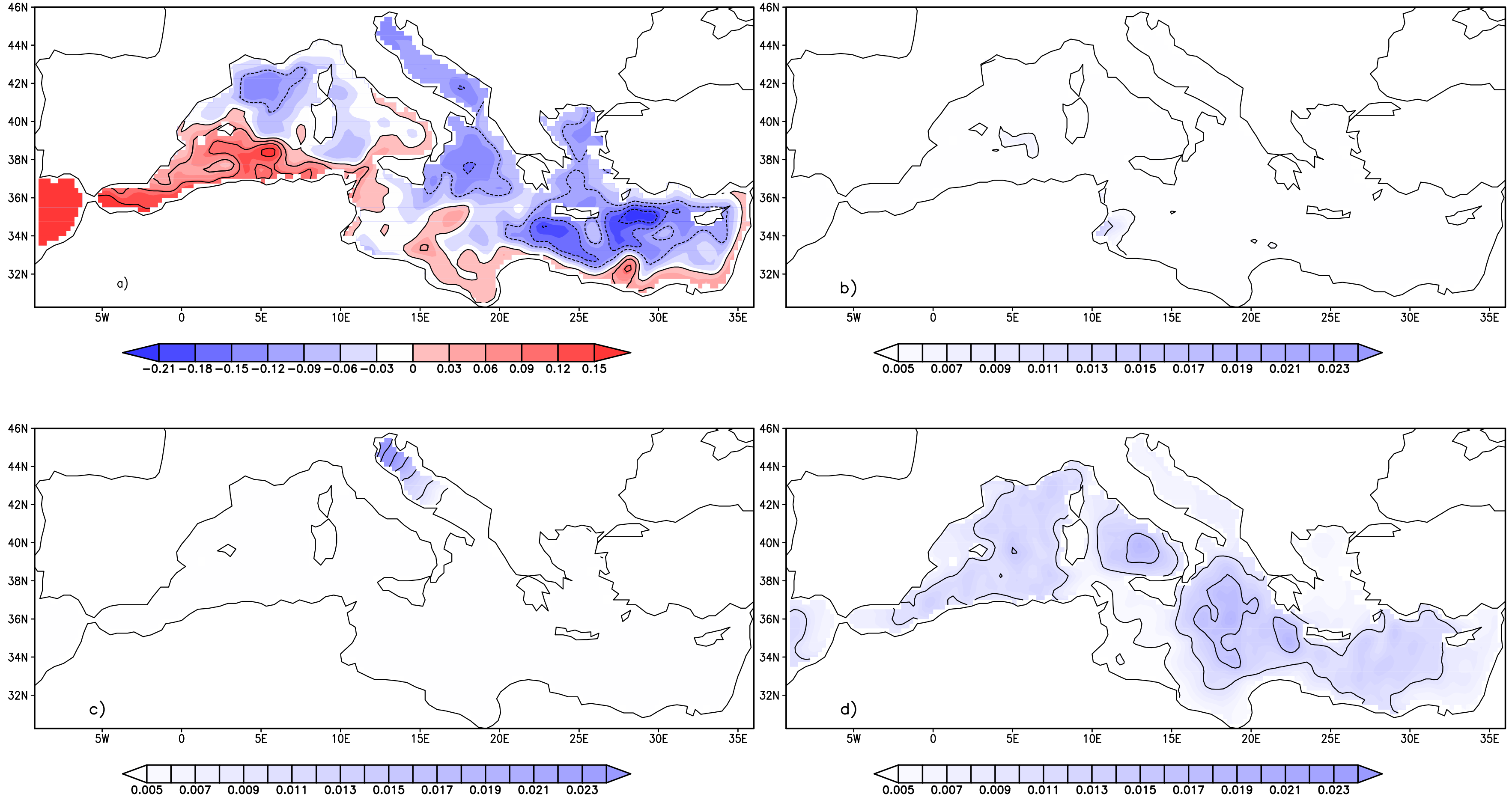


Figure 4.8: a) MFS471 control run SSH [m] at February 8th 2005 (forecast initial condition, here called F0). The contour interval is 0.05 m. b) Standard Deviation [m] of BHM-MFS471 ensemble at F0 . c) Standard deviation [m] of EC-MFS471 ensemble at the end of the fist day of forecast (F0 + 1 day) . d) Standard deviation [m] of ICRP-MFS471 ensemble at F0. The contour interval in b), c) and d) is 0.01 m.

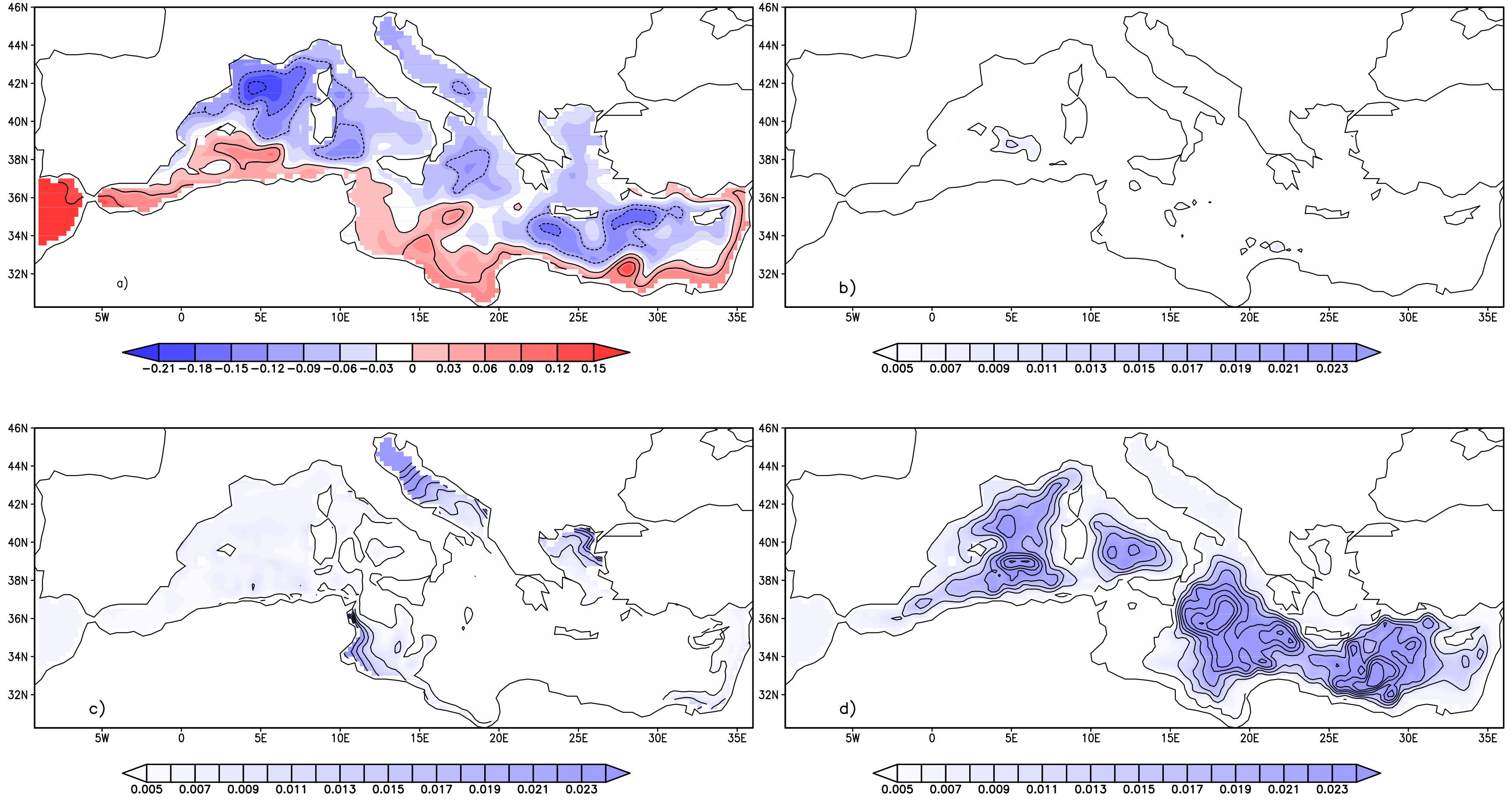


Figure 4.9: a) MFS471 control run SSH [m] at February 18th 2005 (last day of forecast, here called F10). The contour interval is 0.05 m. b) Standard Deviation [m] of BHM-MFS471 ensemble at F10 . c) Standard deviation [m] of EC-MFS471 ensemble at F10 . d) Standard deviation [m] of ICRP-MFS471 ensemble at F10. The contour interval in b), c) and d) is 0.01 m.

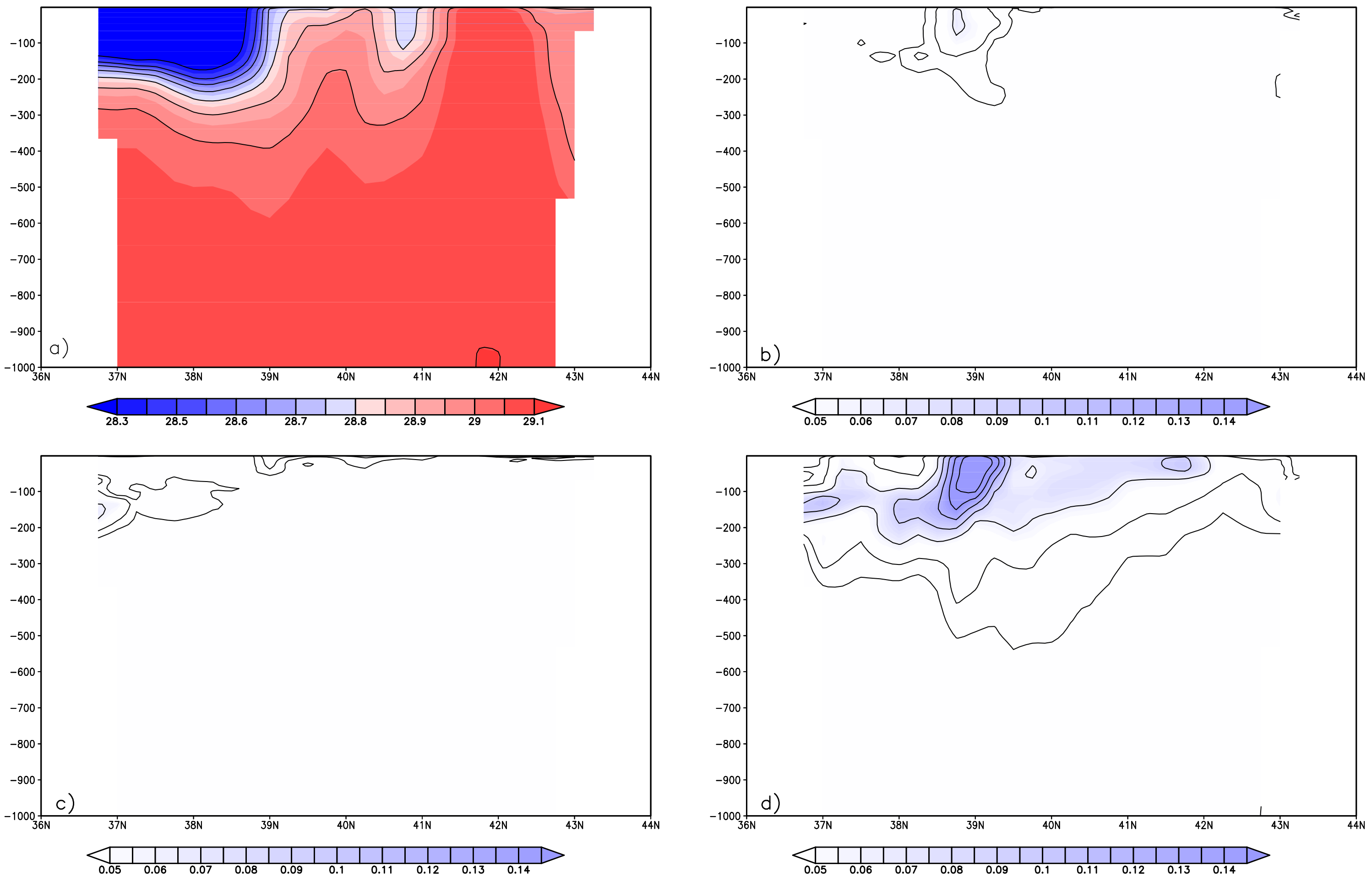


Figure 4.10: a) MFS471 control run density $\sigma = \rho - 1000$ $[kg^3/m^3]$ at February 18th 2005 (last day of forecast, here called F10). The contour interval is 1 σ. b) σ standard deviation of BHM-MFS471 ensemble at F10 . c) σ standard deviation of EC-MFS471 ensemble at F10 . d) σ standard deviation of ICRP-MFS471 ensemble at F10. The contour interval in b), c) and d) is 1 σ.

4.3.1 Comparison with observations

Table 4.3.1 presents the general conclusion on the skill scores of the coarse resolution ensemble experiments. We note that:

- The low resolution implementation of the BHM ensemble method does better then the control run for SLA. The ensemble mean error is 1 cm less the the reference run. The temperature and salinity comparison show an opposite situation; the control run is better then the ensemble mean. The fact that we do not observe consistent forecast skills for temperature, salinity and SLA suggests that the number of available observations in not sufficient to properly sample the model error in a coherent way.
- The EC ensemble mean do not show significant variation of the error skills in comparison with the control run for salinity and SLA. However we register a reduction of the temperature error.
- The ICRP ensemble seems to benefit from the increment of the ensemble size; the skill scores of the ensemble mean are closer to the control run then in the high resolution case.

Table 4.3: MFS471 ensemble and control rms errors. Units are $[°C]$ for temperature, $[psu]$ for salinity and $[m]$ for SLA.

	Temperature			Salinity			SLA
	All	30 m	100 m	All	30 m	100 m	
Control	0.38	0.42	0.40	0.17	0.26	0.22	0.05
BHM Ensemble Mean	0.45	0.56	0.50	0.20	0.28	0.25	0.04
EC Ensemble Mean	0.35	0.36	0.36	0.18	0.26	0.22	0.05
ICRP Ensemble Mean	0.39	0.44	0.42	0.18	0.26	0.22	0.05

4.4 Predictability

We measure the information contained in a probability distribution as its entropy (Shannon 1948 [85]) that is defined as:

$$H(X) = -k \int p(x)\ log\ p(x)\ dx \tag{4.3}$$

where k is a normalizing factor and $h(x) = -log\ p(x)$ is the information content of a single event $X = x$.

If the entropy is a measure of the amount of information contained in a distribution function, it is strait-forward to think at the predictability as the difference between the entropies of two probability distributions (DelSole 2004 [21] and [22]):

1. The climatological probability distribution: this distribution represent the prior knowledge that we have about a certain process. We assume that this distribution is stationary:
$$p(\mathbf{x}_{t+\tau}) = p(\mathbf{x}_t)$$
where we define that $p(\mathbf{x}_t)\ dx$ is the probability that the ocean N-dimensional state X at time t lies in the range $\mathbf{x}_t$ and $\mathbf{x}_t + d\mathbf{x}_t$; τ is the forecast lead time. The expectation of this distribution is the climatology mean that is simply called climatology.

2. The forecast probability distribution: also known as posterior distribution represents the best knowledge that we have of the ocean state X at time $t+\tau$ having observed all available data $\mathbf{o}_t$ up to time t. The chain rule allows to write the forecast distribution as:
$$p(\mathbf{x}_{t+\tau}) = \int p(\mathbf{x}_{t+\tau}|\mathbf{x}_t) p(\mathbf{x}_t|\mathbf{o}_t) d\mathbf{x}$$
where $p(\mathbf{x}_t|\mathbf{o}_t)$ is the analysis distribution and here its mean is obtained using the Reduced Order Optimal Interpolation scheme (Dobricic 2007 [27]). The transitional distribution $p(\mathbf{x}_{t+\tau}|\mathbf{x_t})$ is computed from a dynamical or stochastic model of ocean circulation. We recognise that the ensemble forecasts are a discrete approximation

of the forecast probability distribution.

Thus the predictive information can be defined as (Schneider and Griffies 1999 [84] and DelSole 2004 [21]):

$$R_v = H_1 - H_2 \tag{4.4}$$

where H_1 and H_2 are the entropy of the climatology and forecast distributions.

A major simplification arises if we consider all the distributions as Gaussian. Then the climatology distribution can be written as:

$$p_1(\mathbf{x}) \quad = \quad \frac{(2\pi)^{-m/2}}{(det\boldsymbol{\Sigma})^{\mathbf{1/2}}}\text{exp}\left[-\frac{1}{2}\Big(\mathbf{x} - \langle\mathbf{X}\rangle\Big)^{\mathbf{T}}\boldsymbol{\Sigma}^{\mathbf{-1}}\Big(\mathbf{x} - \langle\mathbf{X}\rangle\Big)\right] \tag{4.5}$$

where $\mathbf{x}$ is the m-dimensional state vector, $\langle\mathbf{X}\rangle$ is the expectation value and $\boldsymbol{\Sigma}$ is the climatology covariance matrix.

Similarly we can write the forecast distribution as:

$$p_2(\mathbf{x}_t) \quad = \quad \frac{(2\pi)^{-m/2}}{(det\mathbf{C_t})^{\mathbf{1/2}}}\text{exp}\left[-\frac{1}{2}\Big(\mathbf{x}_t - \langle\tilde{\mathbf{X}}_\mathbf{t}\rangle\Big)^{\mathbf{T}}\mathbf{C}_\mathbf{t}^{-\mathbf{1}}\Big(\mathbf{x_t} - \langle\tilde{\mathbf{X}}_\mathbf{t}\rangle\Big)\right] \tag{4.6}$$

where $\langle\tilde{\mathbf{X}}_\mathbf{t}\rangle$ is the forecast expectation and $\mathbf{C_t}$ is the covariance of the forecast distribution.

Inserting 4.5 in 4.3 we find that the entropy for the climatological distribution of dimension m is:

$$H_1 = \frac{k}{2}(m + mlog2\pi + logdet\boldsymbol{\Sigma}) \tag{4.7}$$

and similarly the entropy of the forecast distribution is:

$$H_2 = \frac{k}{2}(m + mlog2\pi + logdet\mathbf{C_t}) \tag{4.8}$$

Substituting 4.7 and 4.8 in 4.4 it follows that the predictive information is:

$$R_v = -\frac{k}{2}log\left(\frac{det\mathbf{C}_t}{det\boldsymbol{\Sigma}}\right) \tag{4.9}$$

Finally the Predictive Power (PP) can be defined as:

$$\alpha = 1 - (det\ \mathbf{\Gamma}_t)^{1/(2m)},$$

where we set the normalising constant k equals to $1/m$ and we applied the product theorem of determinants. The matrix $\mathbf{\Gamma}_t = \mathbf{C}_t\ \mathbf{\Sigma^{-1}}$ is call the predictive information matrix. For the univariate case the Predictive Power reduces to 1 minus the ratio of root mean square errors, i.e the standard deviations, of the climatology and forecast distributions.

The predictive information matrix can be decomposed the ocean state in a set of subspaces that are ordered by their predictive power. Schneider and Griffies (1999 [84]) show that the Predictable Patterns can be found as the solutions of an eigenvalue problem. In particular the Least Predictable Pattern (LPP) satisfy the condition:

$$\mathbf{\Gamma_t v_1} = \gamma_\mathbf{1} \mathbf{v_1} \tag{4.10}$$

where γ_1 is the biggest eigenvalue of the Information Matrix $\mathbf{\Gamma}$ to which correspond the least predictable pattern of the ensemble at time t.

4.4.1 Climatological and forecast covariance matrices

Assuming that the forecast and climatological distribution are Normal, the predictable patterns can be found using only the covariance matrices of the two distribution. We compute the climatology covariance matrix as:

$$\mathbf{\Sigma} = \frac{1}{T-1}\sum_{t=1}^{T}(\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T$$

where the t index refers to a time series of the model state vector $\mathbf{x}$ and the bar denotes the time mean. If $\mathbf{X} = \{\mathbf{x}_1 - \bar{\mathbf{x}}, \ldots, \mathbf{x}_T - \bar{\mathbf{x}}\}$ we can write the covariance matrix as

$$\mathbf{\Sigma} = \frac{1}{T-1}\mathbf{X}\mathbf{X}^T = \frac{1}{T-1}\mathbf{W\Lambda^2W^T}$$

where $\mathbf{W}$ and $\mathbf{\Lambda}$ are the EOFs and diagonal eigenvalue matrix derived from a singular value decomposition of $\mathbf{X}$. Equation 4.9 requires non-singular $\mathbf{\Sigma}$; this is often not the case since the dimension m of the state vector $\mathbf{x}$ is usually greater then the degree of freedom T over which the covariance is computed. The similarity between predictable pattern and principal component analysis is used to solve the rank deficiency problem the analysis is performed in the reduced space of the first $k << m$ eigenvectors. It is clear that the definition of the climatological covariance matrix is subjective and requires physical consideration about the properties of the predictability that we are interested in.

The forecast covariance matrix $\mathbf{C}_t$ is easily computed as:

$$\mathbf{C}_t = \frac{1}{M-1}\sum_{i=1}^{M}(\mathbf{x}_t^i - \bar{\mathbf{x}}_t)(\mathbf{x}_t^i - \bar{\mathbf{x}}_t)^T$$

where $\mathbf{x}_t^i$ is the state vector of the ith ensemble member and $\bar{\mathbf{x}}_t$ is the ensemble mean at time t. Note that also this matrix can be written using a singular value decomposition of the ensemble anomaly field. In order to compare all the experiments with the same climatological covariance matrix the low resolution fields where interpolated on the finer MFS1671 grid.

4.4.2 Results: Density predictability

We apply the PP formalism to study the predictability of density from the model ensemble forecasts. In this experiment the climatological matrix is defined using the single deterministic evolution of the density field from January 1 to February 8, 2005 along a section at 5°E. Several test were made using different periods for the definition of $\mathbf{\Sigma}$ (not shown here) and it was found that using a relatively short period for building $\mathbf{\Sigma}$ allows to catch the ensemble variability that is associated with the eddy field. Since the climatological matrix is singular, the analysis was performed in a truncated space, retaining the first $r = 14$ eigenvectors This truncation allow to capture 88 %of the total variability of the density field.

We focus our attention on the LPP (equation 4.10) of BHM, ICRP and EC ensembles.

For the BHM ensemble we consider both the analysis and forecast periods since both of them are associated to a probability distribution that we can compare with the same climatological covariance matrix; for the ICRP and EC experiments we have a probability distribution only for the 10 days of forecast.

Figure 4.11 and 4.12 shows the LPP for the high and low resolution ensembles at February 18th 2005, i.e. the last day of forecast. There is a substantial agreement in all the results and especially between high and low resolution counterparts.

The BHM and ICRP ensembles (see panel a and c of figure 4.11 and 4.12) highlight an isolated thick structure at 38.5° E that extends down in the water column to 150 m. The least predictable pattern for these experiments is the location and shape of the strong anticyclone in the Algerian Current. The predictable power that is associated to this pattern varies drastically between the experiments, see figure 4.13. The ICRP high resolution experiment show the a rapid decrease of PP during the very first days of forecast, reaching values close to zero after 6 days. The low resolution ICPR presents a similar behaviour suggesting that the initial condition random perturbation is capable of producing a strong response from both model set-ups. The high resolution BHM ensemble presents a less drastic decrease of predictable power, but at the forecast end the uncertainty associated with the least predictable pattern is about 50% of the dynamical variance of the density field along 36 days of simulation. The predictability power analysis confirm that the BHM wind perturbation is less effective for the low resolution model.

The least predictable pattern of the ECMWF ensemble (see panel b of figures 4.11 and 4.12) presents a slightly different picture. The maximum unpredictability is almost equally divided between three structure centred at 38, 40 and 43°; the anticyclonic gyre in the Algerian Current is not preferentially exited the by large scale wind perturbation. The predictable power associated to this structure clearly shows that the stochastic EC wind forcing is the least effective perturbation of the vertical stratification of the model density.

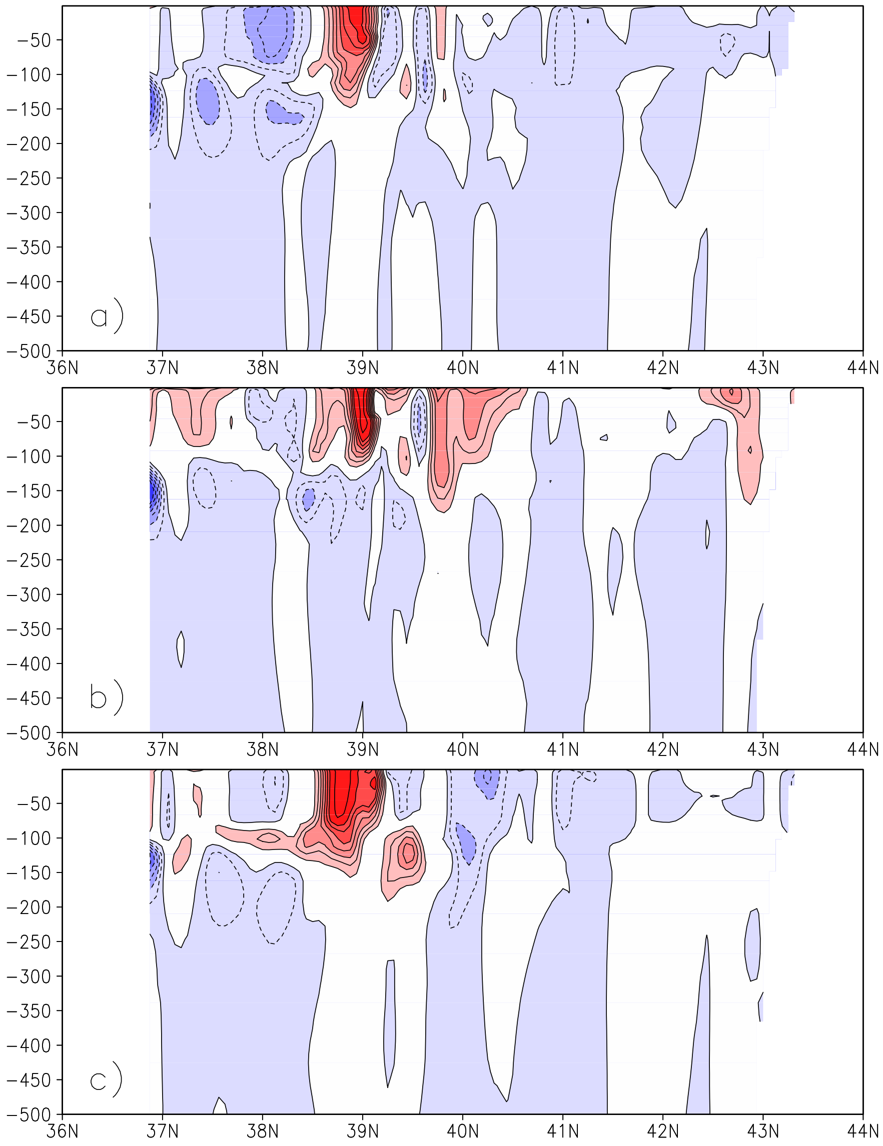


Figure 4.11: Density Least Predictable Pattern (LPP) along a section at 5◦E at day February 18th 2005 (last day of forecast); contour interval is arbitrary. a) BHM-MFS1671 ensemble, b) EC-MFS1671 ensemble, c) ICRP-MFS1671 ensemble. The control run density field is depicted in panel a of figure 4.6.

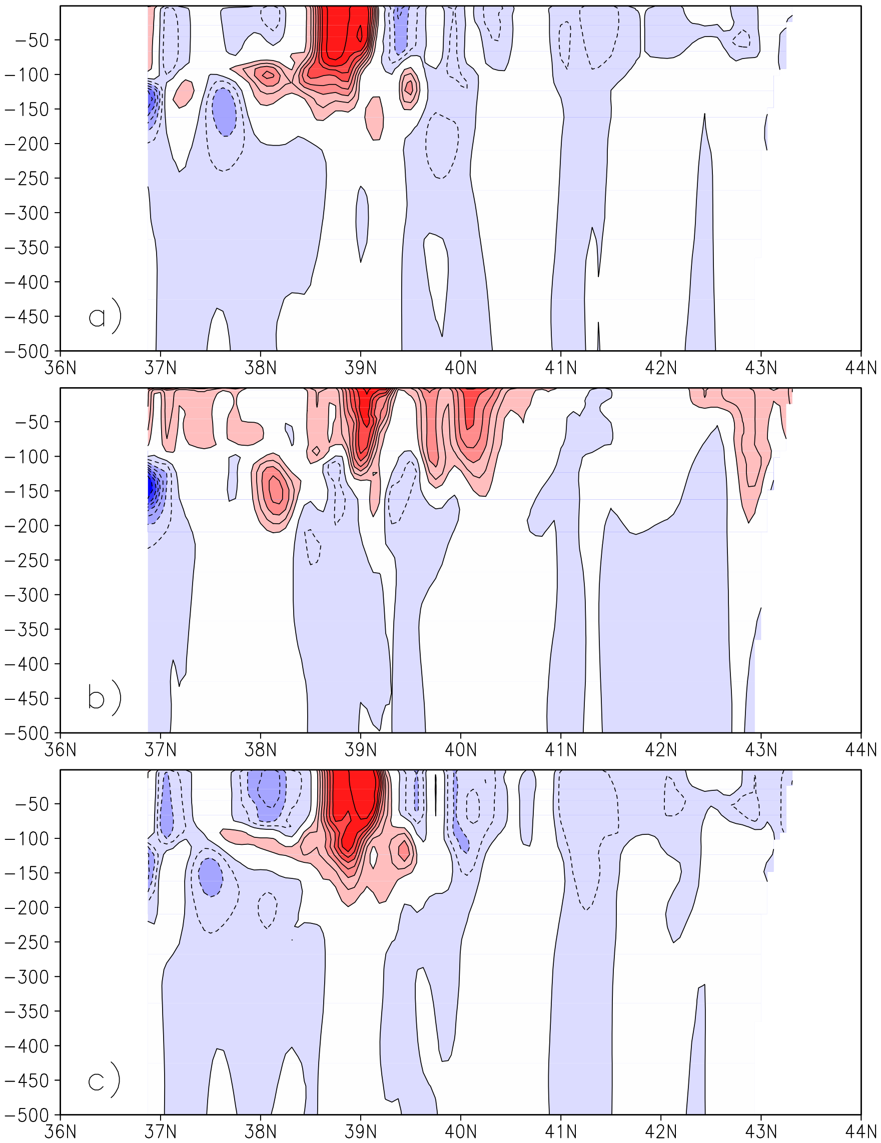


Figure 4.12: Density Least Predictable Pattern (LPP) along a section at 5○E at day February 18th 2005 (last day of forecast); contour interval is arbitrary. a) BHM-MFS471 ensemble, b) EC-MFS471 ensemble, c) ICRP-MFS471 ensemble. The control run density field is depicted in panel a of figure 4.9.

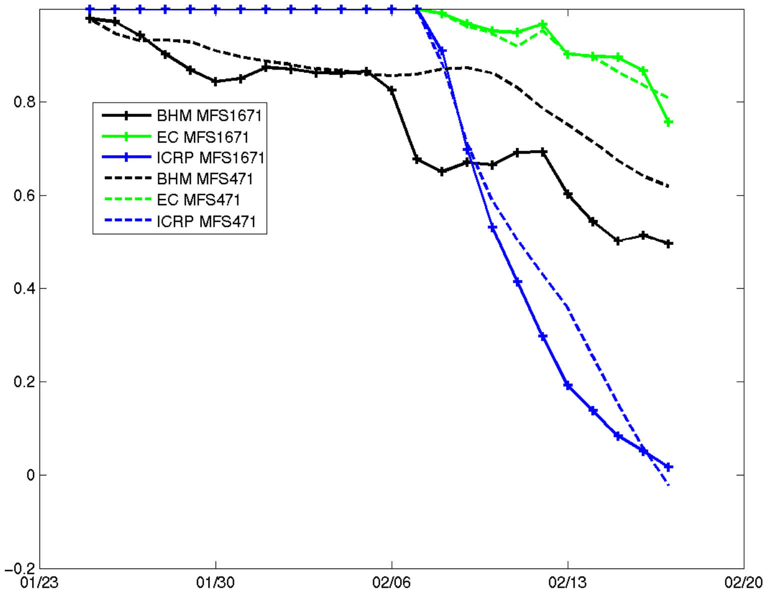


Figure 4.13: Predictive Power associated to the density Least Predictable Pattern for 6 ensemble experiments. The analysis period extend from January 25th to February 8th, following days are in forecast period. EC and ICRP have predictable power equal to 1 before February 8th, since there is no ensemble during the analysis period.

4.4.3 Results: SSH predictability

The climatological covariance matrix is defined using the dynamical evolution of the SSH field is a box region that extends from 2°E to 8°E in the Western basin for the same 38 days period used for the study of density predictability. The covariance matrix is truncated using the first $k = 14$ EOFs that explain the 81% of the variability of the SSH field.

The least predictable SSH pattern of the ensemble distributions for the three experiments at high and low resolution is presented in figures 4.14 and 4.15 for the day February 8th 2005. The SSH mean field for the high and low resolution ensemble is shown in panel a of figures 4.14 and 4.15. The complex dynamics of the SSH is only poorly solved by the MFS471 implementation of the ocean model that do not solve the eddy field.

The least predictable pattern for all experiment highlights the north-eastern border of the anti-cyclonic gyre that is detaching from the core. Fontanet et al. (2004 [40]) explained the dynamics of similar anti-cyclonic eddies of the Algerian Current in terms of two-dimensional turbulence. In particular they show that the circulation cell that encircle the eddy core is characterised by strain dominated processes that are responsible for abrupt changes in the behaviour of the Lagrangian motion of drifters. Then the forecast uncertainty might affect our ability to predict the dispersion of particles crossing these regions and need to be taken into consideration for end-user applications of ocean ensemble forecast.

Panel b of figures 4.14 and 4.15 shows the least predictable pattern for the BHM ensemble for high and low resolution model set-up. The shape of LLPs is similar but the associated predictability power (see figure 4.16) is almost 1, i.e. total predictability, for the MFS471 indicating that the BHM winds are a less efficient perturbation for coarse resolution model.

The least predicable pattern of the ICRP and EC ensemble are shown in panel c and d of figure 4.14 and 4.15 respectively . Again the shape of the LPP is similar between all the ensemble regardless to the kind of perturbation and resolution, the only significant difference being the PP the is associated to each pattern.

The results indicate that the eigenvectors of the Information Matrix $\mathbf{\Gamma}_t$ are found along the directions of maximum variability of the climatological covariance matrix $\mathbf{\Sigma}$ that is a unique reference for the 6 ensemble experiments. Then also for the low resolution ensemble experiments the least predictable pattern highlights meso-scale feature that are not fully resolved the coarse MFS471 ocean model.

We observe that generally the low resolution experiments presents higher Predictable Power then their high resolution counterparts; it is useful to clarify this point. What we are estimating here is only the potential predictability of the system given by accessible forecast distribution that is based on our imperfect model representation of the dynamical evolution of the ocean state vector (Del Sole 2004, [22]). DelSole extends the problem in order to find which is the predictability of the "true" state given the accessible forecast distribution. This conditional distribution will be different for the various experiments that we presented here in order to relate them to the unique predictability of the "true" system.

4.5 Summary

In this chapter we presented a comparison between 3 methodologies of ocean ensemble forecasting applied to the Mediterranean Forecast System. The application of Bayesian Hierarchical Model winds (that are described in Chapter 3 of this thesis) during 14 days of ocean analysis proves to be an effective method in perturbing the forecast initial condition given realistic estimated of uncertainties in the surface wind forcing; the parallel assimilation of in-situ and satellite data provide a reasonable constraint to the initial condition perturbations reducing the uncertainty in the ocean field according to the best available knowledge of the system. The BHM winds are also applied in the forecast phase to sustain the perturbations growth. The ocean spread at the end of the forecast is mainly concentrated in the meso-scale of the Mediterranean circulation suggesting the eddy field is the most unpredictable component of the ocean state.

The Initial Condition Random Perturbation built by vertical extrapolation of a random 2-dimesional field through EOF generate a large ocean ensemble spread, as it was already

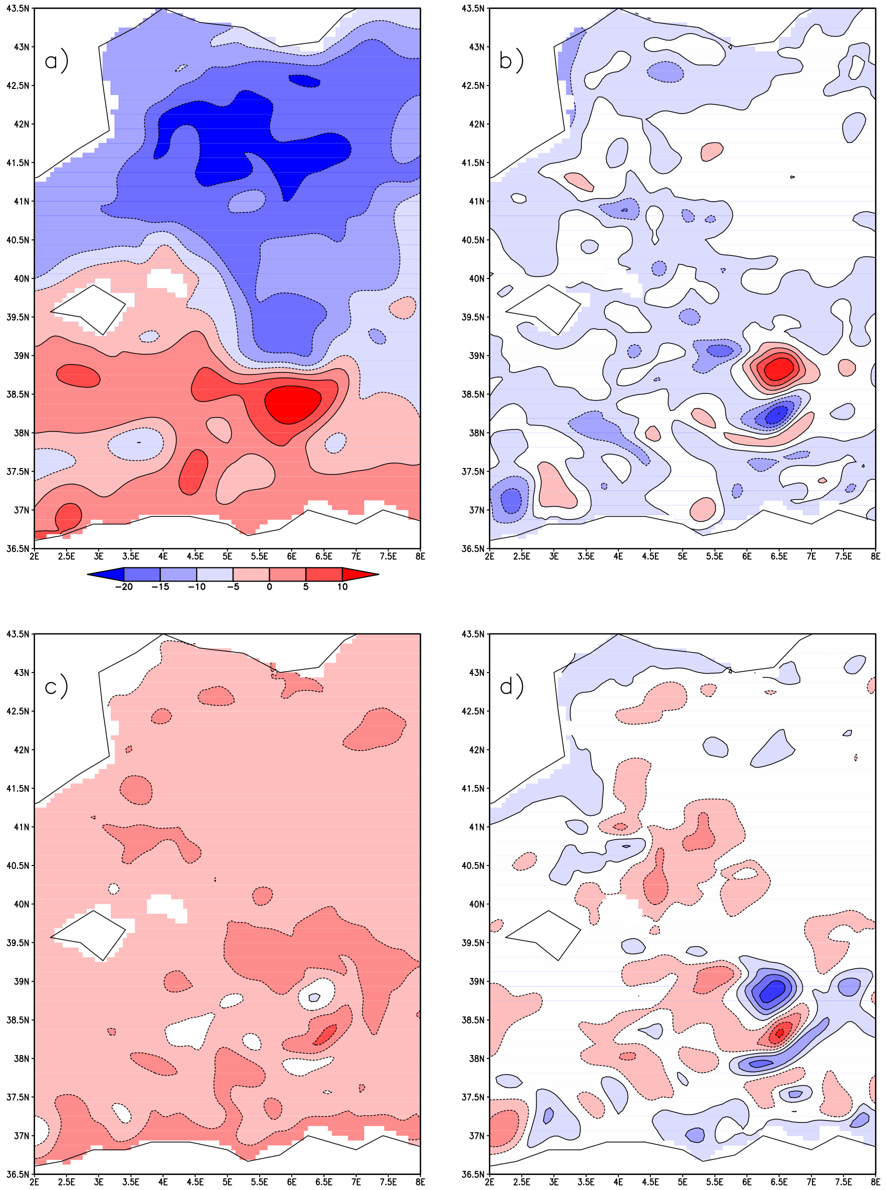


Figure 4.14: a) MFS1671 control run SSH [m] at February 18th 2005 in a box region in the Western Mediterranea. The contour interval is 5 cm. b) SSH Least Predictable Pattern (LPP) of BHM-MFS1671, c) EC-MFS1671 and d) ICRP-MFS1671 ensembles. Contour interval for panel b, c and d is arbitrary.

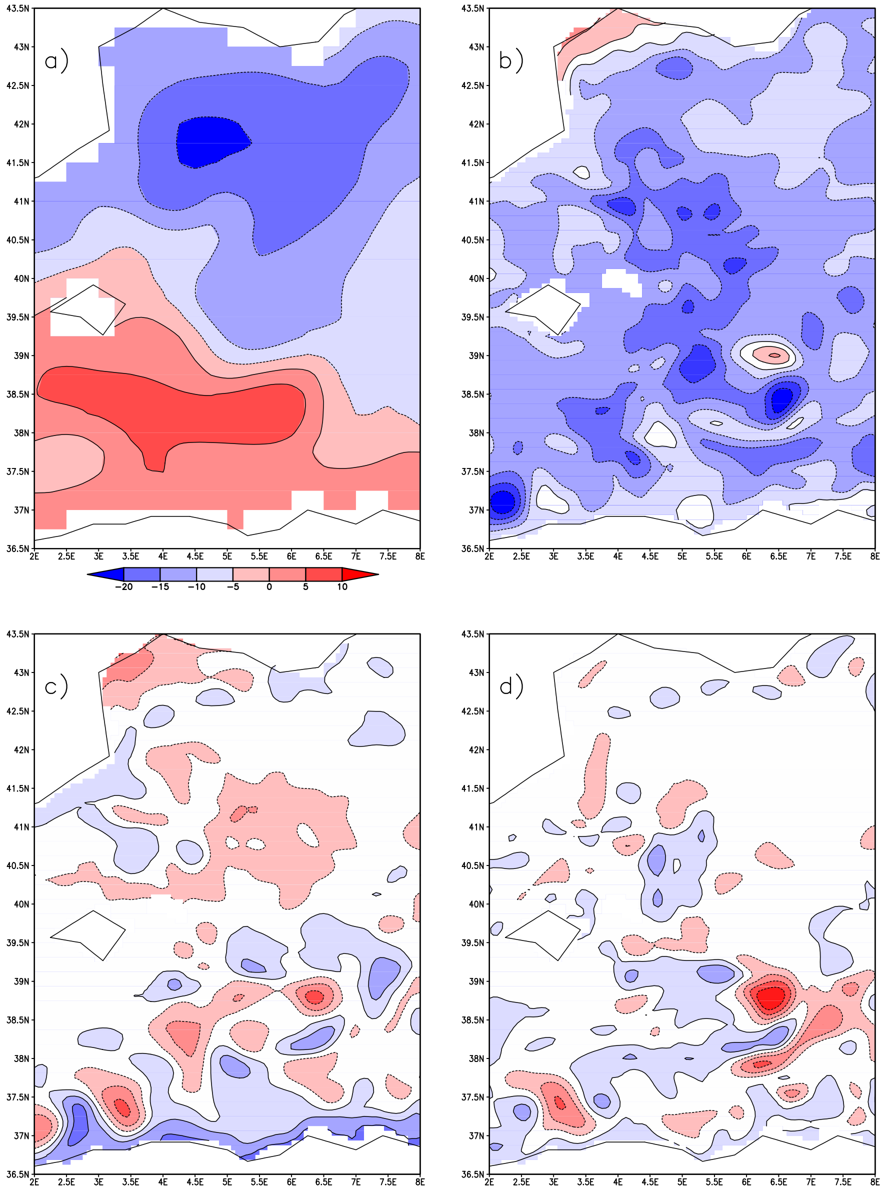


Figure 4.15: a) MFS471 control run SSH [m] at February 18th 2005 in a box region in the Western Mediterranea. The contour interval is 5 cm. b) SSH Least Predictable Pattern (LPP) of BHM-MFS471,c) EC-MFS471 and d) ICRP-MFS471 ensembles. Contour interval for panel b, c and d is arbitrary.

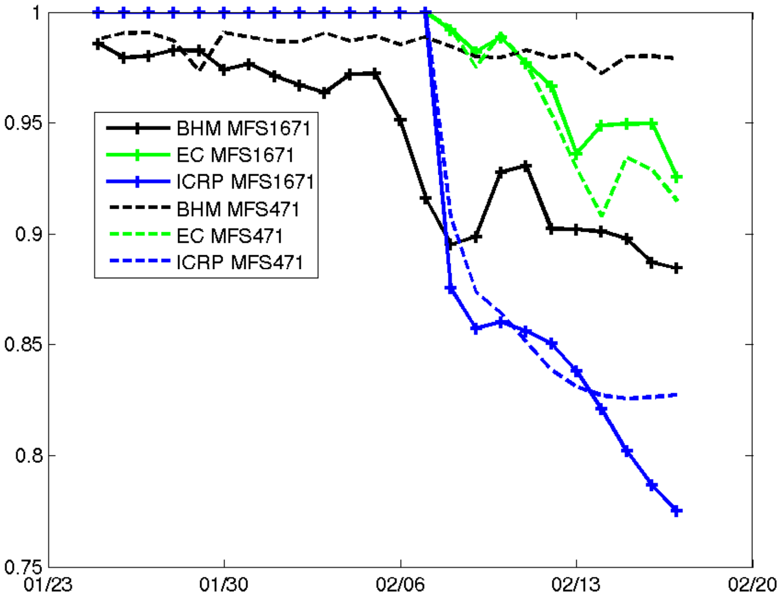


Figure 4.16: Predictive Power associated to the SSH Least Predictable Pattern for 6 ensemble experiments. The analysis period extend from January 25th to February 8th, following days are in forecast period. EC and ICRP have predictable power equal to 1 before February 8th, since there is no ensemble during the analysis period.

demostrated in a recent work of Pinardi et al. (2007 [74]) applied to a similar set-up. However the random nature of the perturbation to not permit to integrate the best available knowledge of the ocean state; the data comparison shows that ICRP ensemble mean has lower skills then the control forecast.

The ensemble generated by the ECMWF ensemble winds prove to be centred around the control forecast, presenting only slightly deteriorate skills. This method generates mostly a large-scale signal in ocean spread, suggesting that the EC winds are not an effective perturbation of the meso-scale circulation.

The experiments performed with a low resolution ocean model set-up show strong analogies with their high resolution counterparts; however BHM winds appear to be an ineffective perturbations of a coarse resolution model most of their energy being confined in scale not properly resolved by the low-resolution model. On the other hand the ICRP scheme benefits from the increased numbers of members that can be computed for the less-demanding low resolution implementation of the ocean model.

The differences observed in the responses of the ocean model to BHM, ICRP and EC collapse when we isolated the least predictable pattern of the forecast. The ocean model present a preferential direction of variability that is similarly found in all ocean ensembles.

We identify two directions of further development of the present work. The first one is the investigation of the singular vectors of the ocean model that will help the understanding of the observed similarities between the ocean ensemble forecast experiments. The second direction of development is the investigation of the relation between forecast probability distribution and forecast errors. In the next Chapter we focus on this second issue; we present a study of the ensemble properties viewed as a proxy for model error. The first issue is left to a future evolution of this work.

Chapter 5

Ocean Ensemble Forecasting Part II: error covariance estimates

5.1 Experiment Description

An eight month long experiment from April to December 2005 of weekly BHM ensemble forecasts was performed in the framework of the operational Mediterranean Forecasting System. The statistical properties of the ensemble are compared with model errors throughout the seasonal cycle proving the existence of a strong relationship between forecast uncertainties due to atmospheric forcing and the seasonal cycle of model errors.

The new ensemble forecast method based on the BHM winds described in Chapter 4 has being applied in a long re-forecast experiment for the year 2006.

The Mediterranean Sea is an ideal testbed since it has been the site of an important modelling and real time forecasting effort; operational prediction of the ocean state are released regularly by the MFS group since 1998 [73]. The operational modelling activities are complemented by a large observational network. Vertical profiles of temperature and salinity were regularly taken by a ship of opportunity program (Manzella et a. 2003 [58]) and a drifting profiling program that deployed ARGO floats in the Mediterranean Sea (see Poulain at al. 2007 [77]). Satellite measurements of Sea Level Anomaly (SLA) data and Sea Surface Temperature were available in near real time for the operation activities of MFS (Le Traon et al. 2003 [49] and Buongiorno Nardelli et al. 2003 [10]).

A set of 36 ensemble forecast experiments was performed for the period April to December 2006. Each week an ensemble cycle is conducted: it is composed of 14 days of analysis and 10 days of forecast repeated for 10 ocean members forced by different realisations of the stochastic forcing produced by the BHM. The ensemble is initialised at every cycle with a control run forced with deterministic ECWMF winds; hence there is no propagation of ensemble uncertainties between successive forecast cycles. The experiment period spans between spring to late autumn and allows to study the ensemble properties all along the seasonal variability of the Mediterranean Sea.

The ocean model used here is the high resolution version of the MFS model described in Tonani et al. (2007 [91]), so-called MFS1671. The assimilation is carried out with an Optimal Interpolation scheme that was originally developed by De Mey and Benkiran

(2002 [23]) and which last MFS implementation is described in Dobricic et al. (2005 [26]). A brief review of the ocean model and of the assimilation scheme can be found in Appendix B.

The BHM produces small scale, high frequency wind perturbations that do not change the mean property of the atmospheric flow as it is depicted by the ECMWF analysis and forecast. Figure 5.1 show a 10 days moving average of the ensemble mean and the ensemble standard deviation of BHM wind stress over the Mediterranean Sea. We observe that the variability of wind stress is constant throughout the seasons, while the standard deviation of the wind stress curl is generally larger during spring and fall than summer. In summer in the Mediterranean Sea we achieve the lowest wind speeds and wind curl values; our method to perturb the winds by BHM is not capable to create speed in the wind stress curl when the amplitude is low.

The surface boundary conditions for the ocean model include a set of bulk formula to account for the surface heat budget (Castellari et al. 1999 [12] and 2000 [13]) then the perturbation of the input winds indirectly affects the heat and water fluxes. The time series of the net heat flux is shown in figure 5.2. The ocean warms up until August then it starts to loose heat to the atmosphere reaching highly negative values of heat flux that a typical of the winter season in the Mediterranean Sea. The standard deviation of net heat flux slowly increases during the experiments reaching maximum values of 10 W/m^2 at the end of November.

The Mediterranean Sea has negative balance between evaporation and precipitation that is compensated by a net water inflow through the Strait of Gibraltar. The scarcity of precipitation data over the ocean makes difficult to model this important process directly and it has led to alternative parametrization of the water fluexes. The MFS models the air-sea water exchanges as a salt flux, where the driving force is the difference between model and climatological surface salinity, see Chapter 1 and Tonani et al. (2007 [91]). The wind perturbation acts on the modelled water balances not through evaporation, as it would be physically correct, but altering the dynamics of surface salinity advection and mixing. Figure 5.2 shows the 10 days moving average ensemble mean and standard

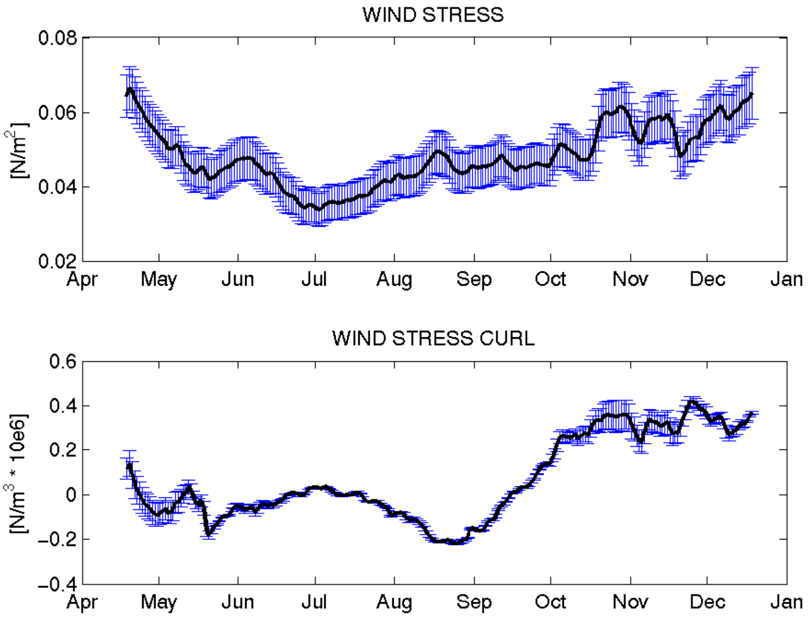


Figure 5.1: Basin mean wind stress $[N/m^2]$ and wind-stress curl $[N/m^3]$ over the Mediterranean Sea. Dark solid line refer to ensemble mean values, vertical bars denote ensemble standard deviations.

deviation of the net upward water flux. The standard deviation is large compared to the signal suggesting that the water fluxes scheme is sensible to small variation of the surface salinity field. This is probably non completely consistent with the heat flux perturbations and simulations should be carried out in the future with new water flux formulation of the MFS model.

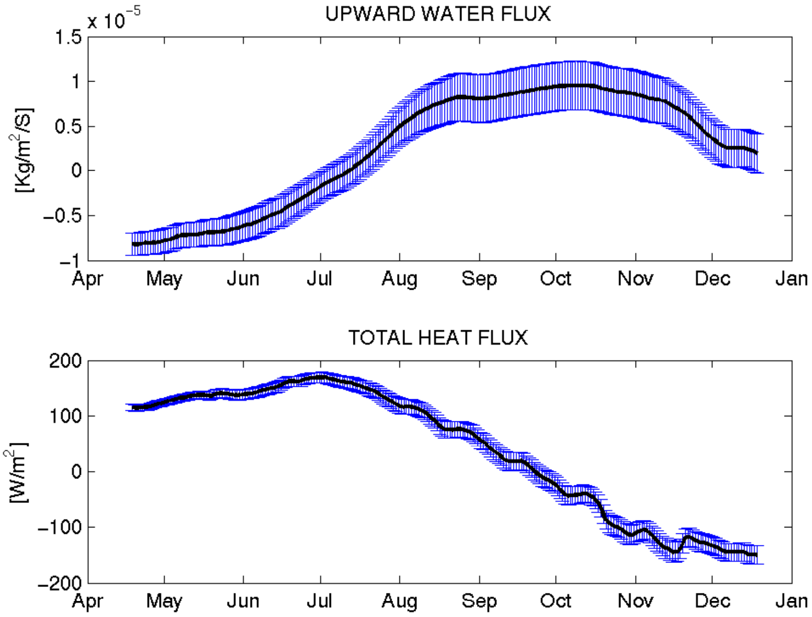


Figure 5.2: Basin mean upward water flux $[kg/m^2s]$ and total heat flux $[W/m^2]$. Dark solid line refers to ensemble mean values, vertical bars denote ensemble standard deviation.

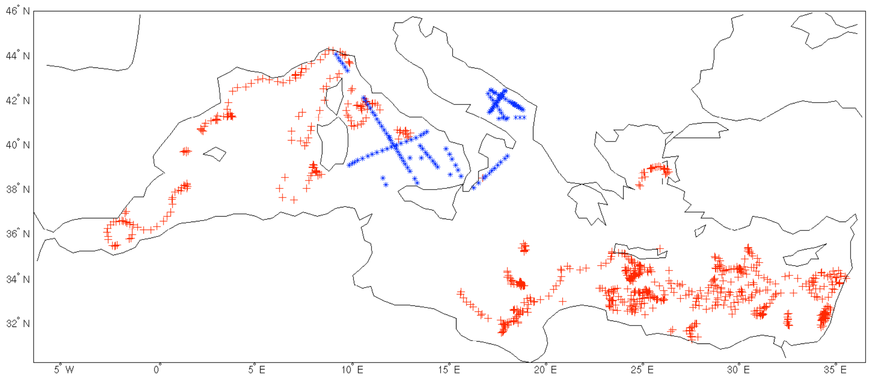


Figure 5.3: Location of ARGO (red crosses) and XBT (blue points) profile from April 18th to December 14th 2006 in the Mediterranean Sea.

5.2 Temperature and Salinity forecast errors

The aim of this work is two-folded; first we want to understand if the perturbation of the air-sea fluxes produced by the BHM winds is able to generate a stochastic ocean response that is informative of the model error; second we want to use the ocean ensemble variance produced by the BHM ensemble to calculate the the background error covariances to be eventually used by the assimilation scheme.

To understand the MFS forecast skills we compare the data with the model forecast fields interpolated on the position and time of the XBT and ARGO profiles computed for the 36 ensemble forecast in 2006.

A relatively large data-set of ARGO and XBT vertical profiles was collected during the time period covered by this study; figure 5.3 shows the spatial distribution of temperature and salinity profiles. The availability of a large data set makes feasible to analyse the ensemble properties against the forecast errors for a time period that covers a complete cycle of formation and destruction of the thermocline in the Mediterranean Sea.

Figure 5.4 show the results of this comparison for the Eastern Mediterranean Sea for the whole period of the ensemble forecasting experiment. The model-data intercomparison is carried out only in the vertical direction since all the data and interpolated model fields were horizontally averaged.

The observed temperature (panel a of figure 5.4) presents a sharp vertical gradient that appears to be smoothed in the model temperature vertical profile (panel b of the same figure). The model is generally colder than observations, in particular it is not able to properly reproduce the temperature maximum that is recorded by XBT and ARGO profiles in September.

Panel c of figure 5.4 shows the salinity field as viewed by ARGO floats. We recognise in the observations the high surface salinity signal that is controlled by summer evaporation. Below the surface salty waters there is a layer of relatively fresh water that is produced by the horizontal advection of Modified Atlantic Water from Gibraltar. The deep high salinity layer is characterised by Levantine Intermediate Water that is formed in the Rhode gyre and advected in all the Mediterranean Sea (Demirov and Pinardi 2002 [24]) . The model underestimates the high salinity observed values (see panel d of figure 5.4) suggesting that the crude parametrization of the water fluxes is affecting the quality of the reconstructed salinity even if observations are assimilated.

The ability of the MFS forecasts to reproduce the observations depends on the performance of the ocean model and the assimilation scheme. The work of Tonani et al. (2007 [91]) shows that during summer the MFS ocean model lacks the high evaporative components of the heat fluxes and that the parametrization of the upper mixed layer physics does not reproduce well the relative deep, saline and hot mixed layer. On the other hand the assimilation does not fully correct the model fields.

The parametrization of the background error covariance matrix plays a key role on the performances of on Optimal Interpolation scheme such SOFA that is used by MFS. The work of Dobricic et al. [26] proved that a better representation of the background error improved the skills of the MFS analyses. Here we want investigate the possibility to extract valuable information from the ensemble variance that is generally interpreted as a proxy for model error.

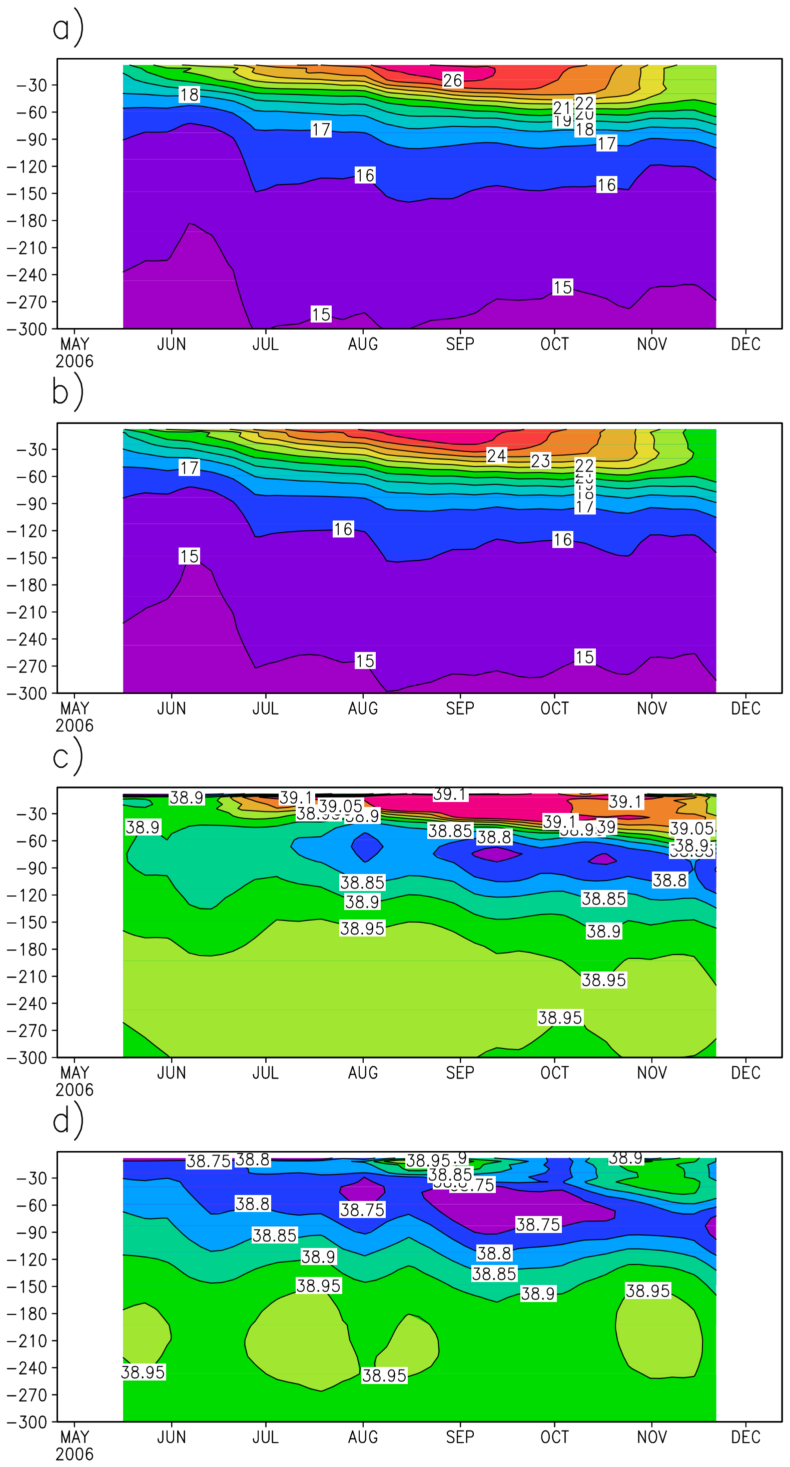


Figure 5.4: Mean temperature and salinity profile for the period May to December 2006. a) ARGO and XBT temperature [$^\circ C$]. b) Ensemble Mean forecast temperature [$^\circ C$]. c) ARGO salinity [psu]. d) Ensemble mean forecast salinity [psu].

5.2.1 Estimates of Temperature and Salinity vertical errors covariances.

The RMS of the model error is compared with the standard deviation of the ocean ensemble and with a statistics derived from the vertical EOFs used in the assimilation scheme.

We then define three different representations of the model vertical error:

- RMS misfit; for each forecast cycle we compute the RMS of the misfits between the control forecast and data. The control run is the deterministic forecast produced with ECMWF winds. The measurements are XBT and ARGO profiles of temperature and salinity. We average the RMS along the vertical direction.

- BHM ensemble spread; this quantity is the standard deviation of the ensemble members around the ensemble mean. We evaluate this quantity at the observation locations and then we average for all the data that fall into the forecast window grouping by vertical levels.

- EOF spread; the MFS assimilation scheme, described in Appendix B, uses vertical EOFs to represent the vertical model errors. For each temperature and salinity profile we compute the standard deviation of an ensemble drawn from the multivariate normal distribution prescribed by the EOFs

  $$e \sim N(0, S_{jt}\Lambda_{jt}S_{jt}^T)$$

  where e is background model error, the matrix S contains 20 vertical temperature and salinity EOFs and Λ is the eigenvalue matrix. The index $j = 1, \ldots, 13$ and $t = 1, \ldots, 4$ refers to the region and the season to which the profile belongs. The profile of the ensemble standard deviation are interpolated at the depth of the observations and then averaged for all the data that fall into the 10 day forecast cycle grouping by vertical levels.

In figure 5.5 we intercompare the different model vertical errors for the temperature. A clear maximum is present at the depth of the upper thermocline. As the summer season advances the mixed layer depth and the location of maximum temperature error deepens.

Overall we observed a difference of 30 m in the position of the error maxima between June and October.

The spread of the BHM and EOF underestimate the amplitude of the background model error by an order of magnitude.

The BHM ensemble spread a structure of vertical error that mimics the behaviour of the background model error. The EOFs spread presents a limited time variability since the EOFs are only updated with a seasonal frequency and amplitude are underestimated by a factor 20. The true variability of the EOF ensemble spread seems also shifted in the with respect to the temperature misfit. During summer the RMS of misfits present an intensification of the surface signal that does not appear in neither the BHM or EOF spread.

The maximum amplitude of salinity misfit errors are recorded between September and October. The salinity RMS of misfit has a constant maximum at sea surface. During late summer and autumn we observe a vertical penetration of salinity error down to 50 m. Generally we observe salinity errors of 0.1 PSU down to 200 m. The BHM ensemble spread amplitude is one order of magnitude less than the misfit error and the the highest salinity spread is recorded in mid-September, in good agreement with the timing of the maximum model error.The salinity EOF spread generated by the vertical EOFs has a smaller amplitude than BHM but shows a vertical penetration that is comparable with the salinity model error. The EOF spread does not show any surface intensified signal.

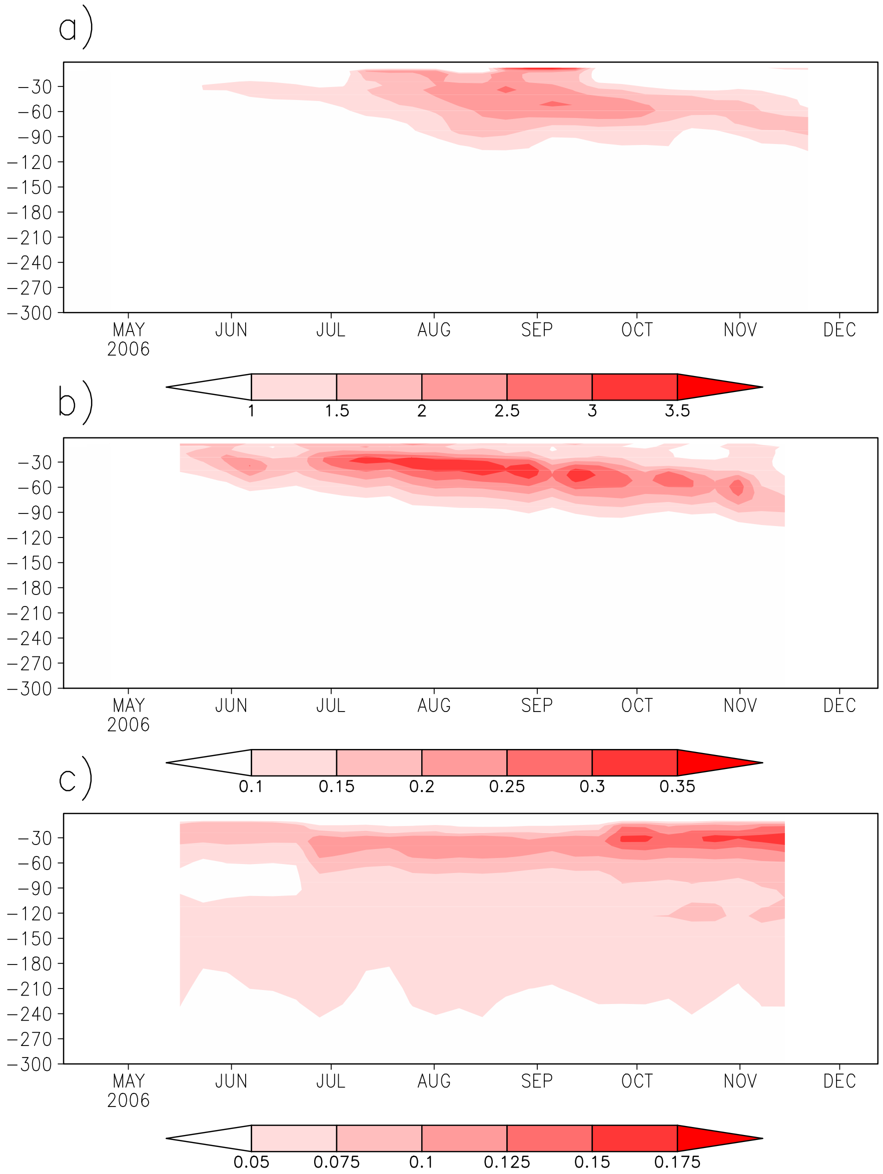


Figure 5.5: a) RMS Temperature misfit $[^\circ C]$ for the period May to December 2006. b) Temperature BHM ensemble spread $[^\circ C]$ for the same time period. c) Temperature EOF ensemble spread $[^\circ C]$. See text for details.

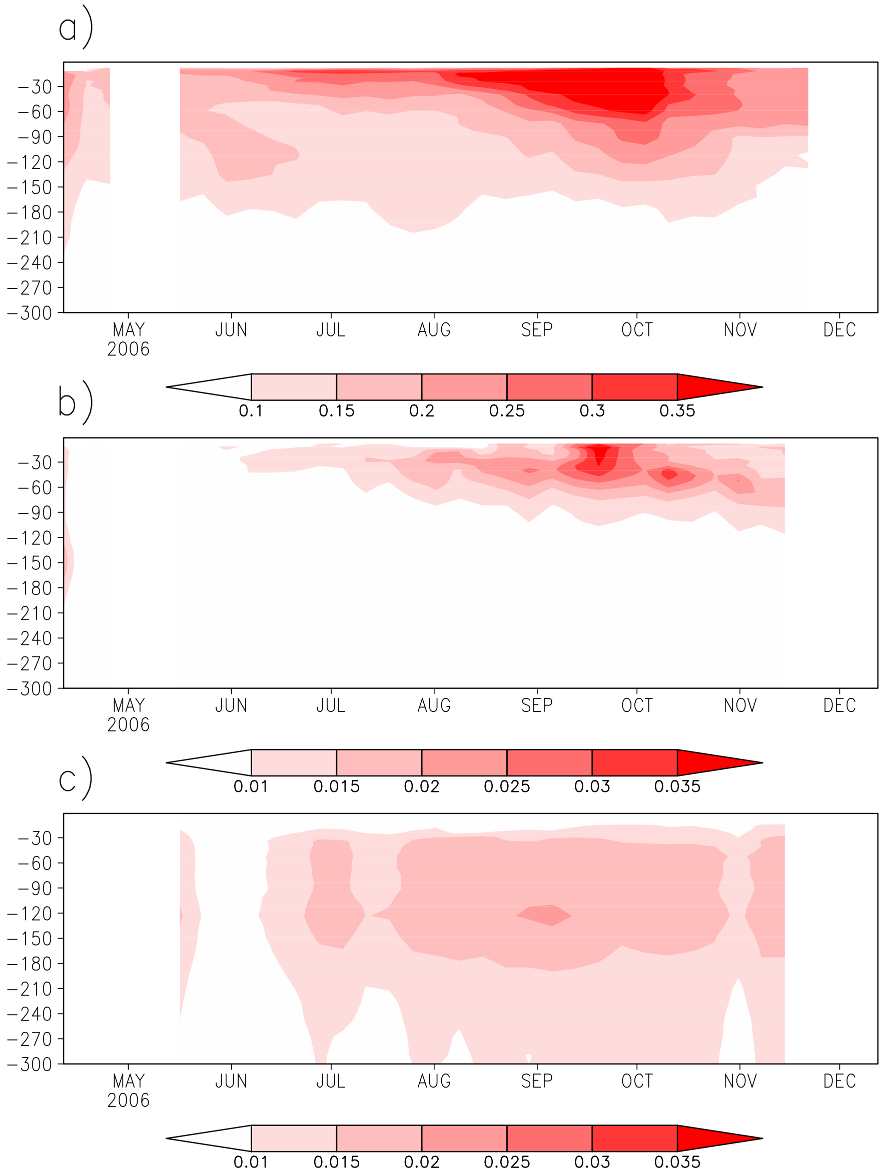


Figure 5.6: a) RMS Salinity misfit [psu] for the period May to December 2006. b) Salinity BHM ensemble spread [psu] for the same time period. c) Salinity EOF ensemble spread [psu]. See text for details.

5.2.2 Likelihood estimates of vertical error covariances

Ensemble techniques are commonly used as a step in data assimilation (see Ensemble Kalman Filter [33]) to approximate the time varying model error covariance matrix. In this section we estimate the ability of the BHM ensemble to mimic the background model error in theoretical framework. We compute the likelihood of the observed temperature and salinity misfits for two prediction models that are different on the specification of the error covariance matrix.

The likelihood model has a long history in statistics and it is mathematically optimal in the sense that the estimates of parameters of calibration model that are fitted by maximising likelihood are the most accurate possible (Casella and Berger, 2002 [11]). The likelihood estimates are an intuitive way to compare stochastic predictions. In our case we assume that the likelihood functions are Gaussian, then we write:

$$L \quad = \quad \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{Hx})^{\mathbf{T}}(\mathbf{H\Sigma H^T})^{-\mathbf{1}}(\mathbf{y}-\mathbf{Hx})} \tag{5.1}$$

where Σ is the covariance matrix, $||$ is the determinant operator, $\mathbf{y}$ is the observed profile of temperature or salinity or both, $\mathbf{x}$ is the temperature and salinity model vector and H is the observational operator with dimension $m \times n$ where m is length of y and n the dimension of the model vertical profile. The covariance matrix in the likelihood has been projected in the observation space using the operator H.

Two covariance model were tested here:

- BHM Ensemble Covariance; given a set of m realisation of the model state vector $X = \{X_1, X_2, \ldots, X_m\}$ we define the anomaly vector as $X' = X - \overline{X}$, where $\overline{X}$ is the ensemble mean. Then we compute the covariance matrix as:

$$\Sigma_{BHM} \quad = \quad \frac{1}{m}\tilde{X}_k' \tilde{X}_k'^T + \sigma I \tag{5.2}$$

where $\tilde{X}_k$ is a reconstruction of the anomaly state vector with a truncated number of EOFs, we retain $m-2$ singular vectors. To ensure that the explained the covariance matrix is full rank we add a diagonal term that is scaled as:

$$\sigma \quad = \quad Tr(X'X^{'T} - X_k X_k^{'T})/m \tag{5.3}$$

where Tr is the trace. Since the MFS assimilation scheme implements a separation of vertical and horizontal components of the error covariance matrix, we evaluate a vertical covariance matrix locally at each point on the horizontal plane.

- Covariance; given the set of EOF U used by the operational assimilation scheme we write:

  $$\Sigma_{EOF} \quad = \quad U\Lambda^2 U^T + \tau I \tag{5.4}$$

  where U are the vertical temperature and salinity EOFs, Λ is the eigenvalue matrix and τ is a scaling factor that represents the unresolved variance and ensures the matrix is full rank. For each season and each region there is a set of 20 EOFs. Then the EOF covariance model is build according to the position and time of the observed data.

The validation experiment worked in the following way. For each ensemble cycle we collected all the temperature and salinity profiles $y = \{y_1, y_2, \ldots y_N\}$ and their model counterparts $x = \{x_1, x_2, \ldots, x_N\}$ at the nearest grid point. The total likelihood is defined as:

$$L_{total} = \prod_{i=1}^{T} \prod_{j=1}^{N_i} L_i \tag{5.5}$$

where L_i was defined in equation 5.1 and the index i spans the $N = 36$ ensemble forecast cycles and j the number of the available profile of temperature and salinity. For practical reasons, instead of maximising the product of the likelihood, we will minimise the sum of negative likelihood logarithm.

As we discussed in the previous section both the ensemble spread and the EOFs underestimate the expected model error, see figure 5.5 and 5.6. Then we can isolate two separate problem; what is the best scaling factor for the covariances matrices and which covariance model better reproduces the structure of the background errors.

Figure 5.7 the negative log likelihood for BHM and EOF covariance models as a function of a scaling factor α . The minimum of the likelihood functions is reached between 15 and 20 confirming the substantial underestimation of the model error that affect the two estimates.

The probability gain can be defined as the difference between the likelihood of the stochastic estimates:

$$PG \quad = \quad L_{BHM} - L_{EOF} \tag{5.6}$$

Figure 5.8 shows the sign of the probability gain for the 36 forecast cycles for each of the 13 regions of the Mediterranean Sea. The probability gain is computed comparing L_{BHM} and L_{EOF} at the local minimum of the likelihood function with respect to the scaling factor α. The error covariance matrix derived from the ocean ensemble is generally a better estimate of the model error than the EOFs. From mid August to mid October in almost all regions the BHM covariance performs better, while during late spring is the EOF model that wins in the western regions (add region map).

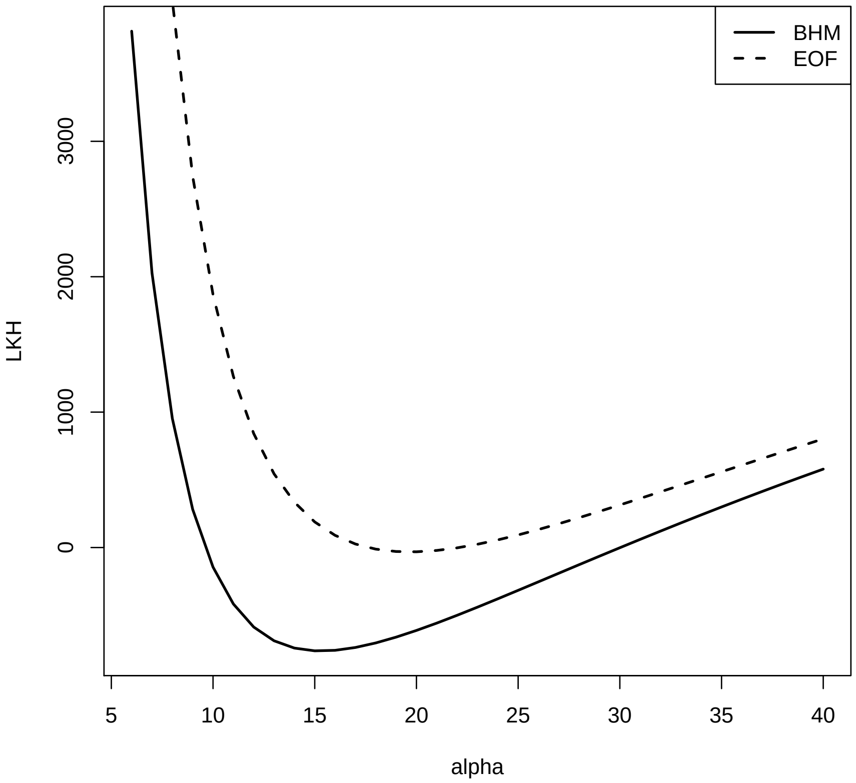


Figure 5.7: Negative Log Likelihood as a function of the α scaling coefficient for BHM and EOF covariance matrices computed for the period August to December 2006. The minimum of the function indicates the optimal value for α.

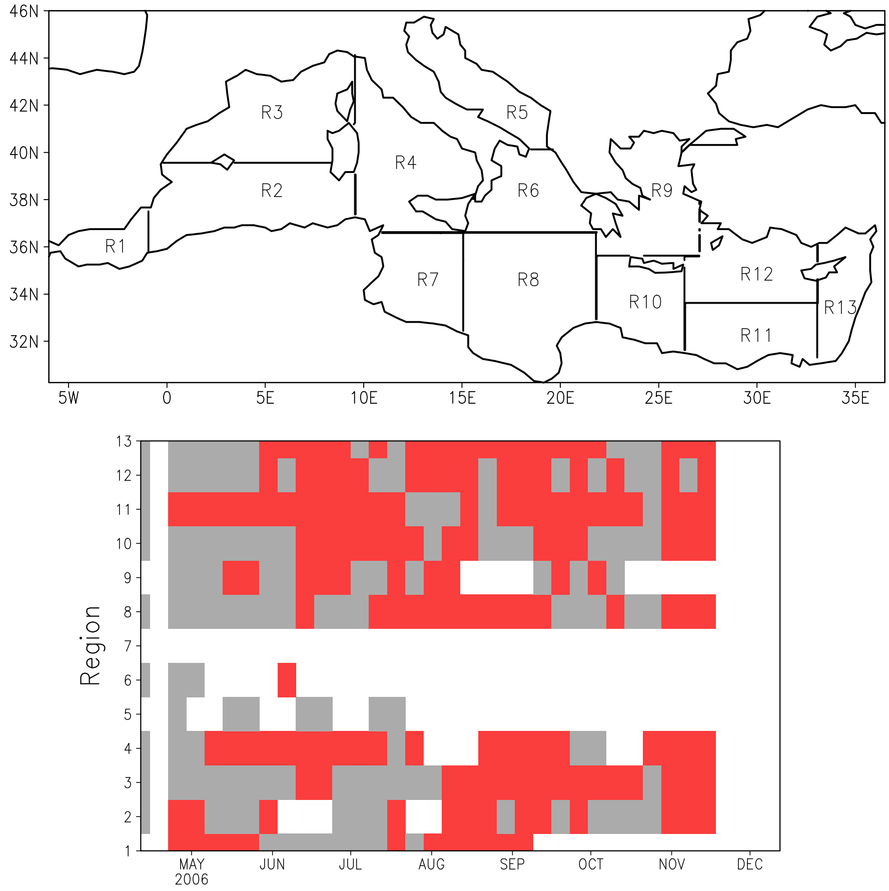


Figure 5.8: Upper panel: the 13 Mediterranean regions. Lower panel: Probability gain divided in the 13 region of the Mediterranean Sea. Red boxes BHM wins, grey boxes EOF wins.

5.3 Sea Level Anomaly forecast errors

The SLA data used by MFS are along track measurement from two satellite missions: Jason 1 and Geosat Follow On (GFO). The SLA data are post-processed in order to provide a homogenous and inter-calibrated data set; the global crossover calibration is described by La Traon 1995 ([47]) and La Traon et Ogero (1998 [50]). The data set is also corrected to account for inverse barotropic and tidal effects and for the signal due to tropospheric and ionospheric contamination (Le Traon and Gauzelin [48]). The along track data then are re-sampled at 7 km resolution and a 7 year SSH mean (corresponding to the period 1993-1997) is removed to provide Sea Level Anomaly (SLA) data.

The SLA model values are computed with reference to a mean dynamic topography level that was calculated by Rio et al. (2007 [81]) on the basis of a ocean general circulation model simulation for the period 1993-1997 corrected by surface velocity observations by drifter and SLA satellite measurements.

The normalised time series of basin averaged misfit errors and BHM ensemble spread are shown in figure 5.9. The mean value of misfit errors is 5.4 ± 0.7133 $[cm]$, while the mean ensemble spread is one order of magnitude less being 0.3672 ± 0.0983 $[cm]$. The time series of misfit errors and ensemble spread have a correlation coefficient of 0.37. The maximum of the ensemble spread time series is reached between August and September, while the maximum misfit error is observed between November and December.

5.3.1 Estimates of horizontal error covariances

In order to evaluate the spatial patterns of the ensemble spread and misfit errors we divided the re-forecast experiment in three parts; April to June, July to September and October to December. For each of the period we computed the average SSH model field, the mean ensemble spread, the mean Kinetic Energy derived from geostrophic velocities and the mean misfit between model and satellite SLA.

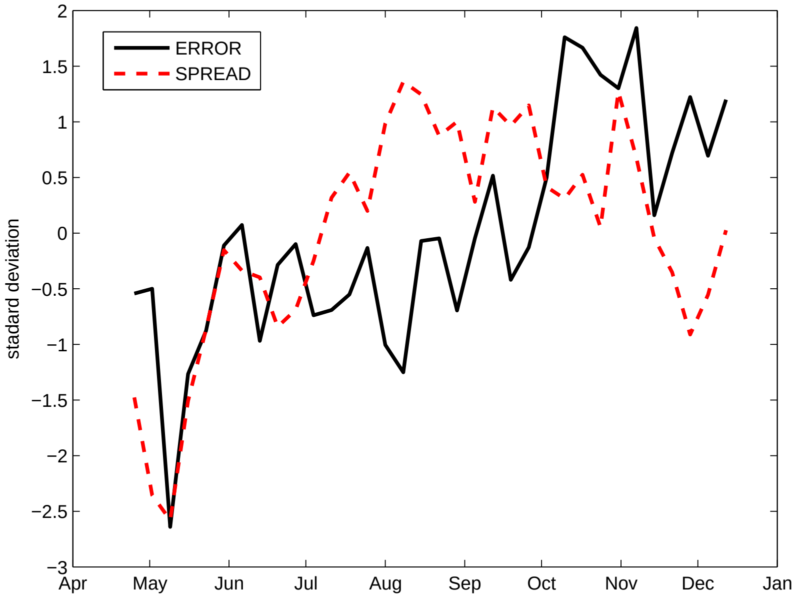


Figure 5.9: Time Series of normalised background model error (dark solid line) and ensemble spread (red dashed line) for SLA. The normalisation of the time series is computed subtracting the mean and dividing by standard deviation; mean spread $0.3672 \pm 0.0983[cm]$, mean error $5.4 \pm 0.7133[cm]$

The averages are performed on the prediction fields using the last 7 days of the forecast window. We skipped the first 3 days of forecast to avoid double-counting of the overlapping days between two consecutive forecast cycles.

In figure 5.10, 5.11 and 5.12 we show that the SSH ensemble spread is mostly concentrated in the Western basin, Ionian and Aegean Sea. The Tyrrhenian, the Adriatic Sea and large portion of the Levantine basin show a low ensemble variability. The maximum amplitude in the ensemble spread is reached in summer between July and September where maximum spread was also observed in temperature and salinity.

From April to June the maximum ensemble spread is found in the Alborean Sea and along the Algerian Current that are well-known regions of meso-scale activity; panel c of figure 5.10 shows that the regions of maximum ensemble spread are also interested by high values of KE. However the correlation between spread and KE do not extend to the Eastern basin.

The along track misfit errors are plotted in panel d of figure 5.10; we consider the subsample of satellite data that were also used by the assimilation scheme in order to have a reference for the amplitude of model error. We observe up to 11 cm of rms between model and satellite SLA; we consider significant any misfits grater then 2-3 cm that is taken as the reference measurement error. Maximum errors are found in the Western basin along circulation structures that also show larger ensemble variance, while the Levantine basin presents generally lower error and ensemble variance.

During the summer the maximum KE patterns are associated with jet-currents. The most energetic structures include the Liguro-Provencal and the Algerian Current, a large KE signal is observed in the gyres of the Alborean Sea and along the western coast of Sardinia, see panel a and c of figure 5.11. The maximum ensemble spread is found in the northern part of the Gulf of Lion gyre, where the Liguro-Provencal current detaches from the coast. However the high ensemble spread region covers almost uniformly the northern part of the Western basin suggesting that the ocean spread is generated by a differential breaking of summer stratification produced by stochastic BHM winds. The satellite data indicate that the regions of maximum model errors are in good agreement with the patterns of high ensemble spread; the major difference is found in the eddy field of the Algerian Current that shows large model error but weak ensemble spread. In the Levantine basin we observe an intensification of jet currents like the Mid-Mediterranean Jet and the Asia Minor Current flowing along the southern coast of Turkey; also the Ierapetra Gyre appears to be intensified with respect to the previous months (see panel a and c of figure 5.11). None of this structures are found in the ensemble spread that appears to be weak in all the Eastern basin. The model error appears to be un-correlated to the ensemble spread.

The third period of the analysis present similar SSH and KE patterns to the previous months, see panel a and c of figure 5.12. Again the Eastern Mediterranean present a weak signal in the ensemble spread suggesting the BHM winds are not an effective perturbation of the Levantine basin dynamics. High values of model errors are found within the Ierapetra Gyre that is totally absent in the ensemble spread map.

The results of the comparison between background model error and ensemble spread are less intuitive than what the vertical errors for temperature and salinity. In the previous section we showed that the vertical structure of the ensemble spread is a good representation of misfit model errors, showing better skills than EOFs in a likelihood sense. For the SLA spread we can not draw a similar conclusion; in fact a further complication is present due to the uncertainty on the mean dynamic topography (Dobricic et al 2005 [26]).

5.4 Summary

In this Chapter we presented an analysis of the relation between misfit errors and ensemble spread. A large set of observation of temperature, salinity and SLA was used to evaluate the misfit errors and their seasonal variations.

The analysis of temperature and salinity profiles showed interesting similarity between the BHM vertical ensemble spread and the misfit error structure. In particular the BHM spread reproduces the deepening of maximum temperature error that is related to the depth of the thermocline. In the contest of an Optimal Interpolation scheme we suggest

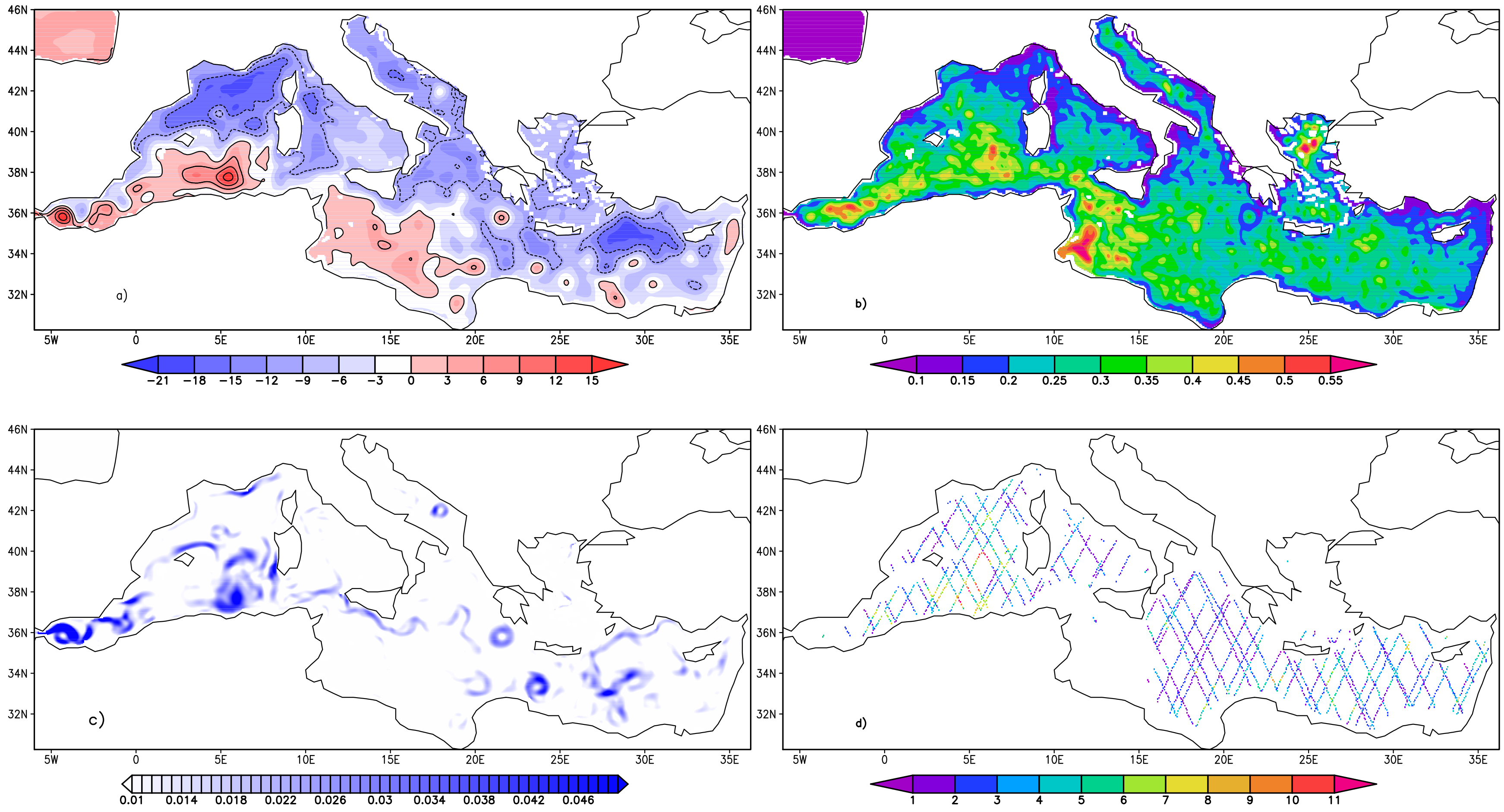


Figure 5.10: a) mean SSH for period April to June 2006 $[cm]$. Contour interval is 5 cm. b) Mean SSH ensemble forecast spread for the same period $[cm]$. c) Mean Kinetic Energy $[m^2/s^2]$. d) SLA mean misfit $[cm]$.

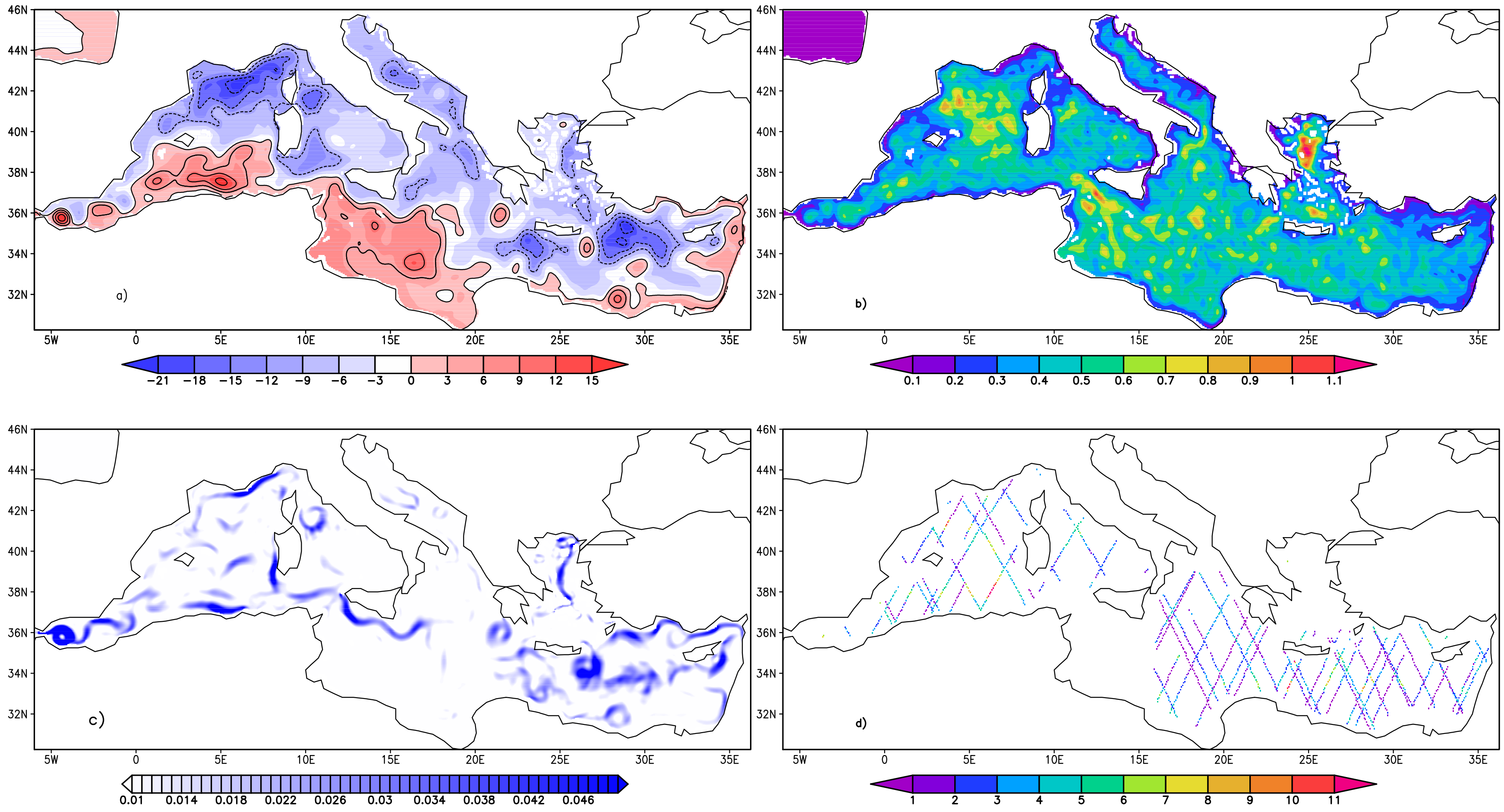


Figure 5.11: a) mean SSH for period July to September 2006 $[cm]$. Contour interval is 5 cm. b) Mean SSH ensemble forecast spread for the same period $[cm]$. c) Mean Kinetic Energy $[m^2/s^2]$. d) SLA mean misfit $[cm]$.

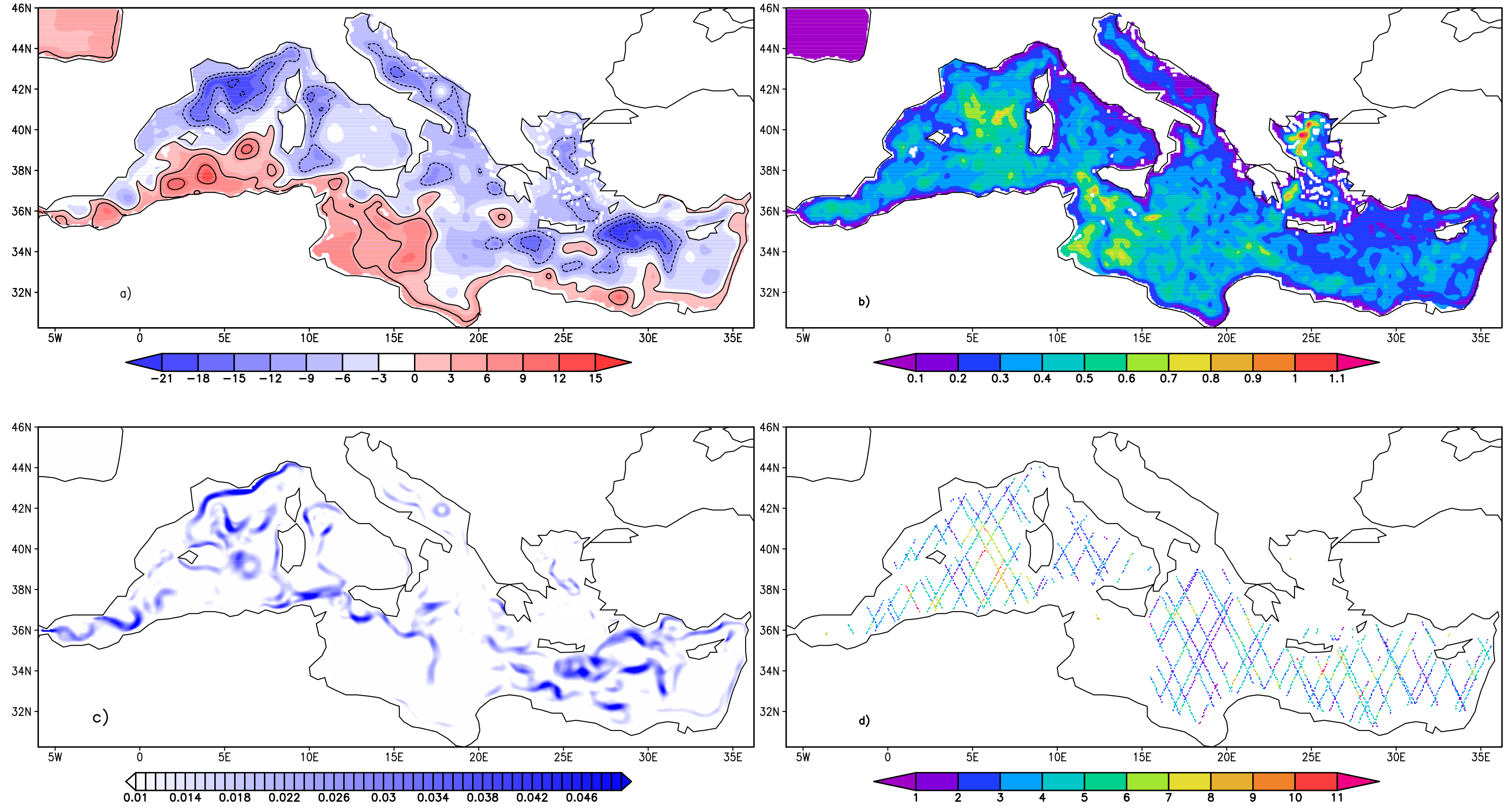


Figure 5.12: a) mean SSH for period October to December 2006 $[cm]$. Contour interval is 5 cm. b) Mean SSH ensemble forecast spread for the same period $[cm]$. c) Mean Kinetic Energy $[m^2/s^2]$. d) SLA mean misfit $[cm]$.

that the ocean spread generated by BHM winds can be used to prescribe the error covariance matrix and its temporal evolution; this idea was tested on a theoretical framework computing the likelihood of the observed misfit using two error covariance matrices; the covariance matrix derived from the 10 BHM ocean members and the covariance matrix built using the EOF currently used in the MFS assimilation scheme. Overall the BHM ensemble provides a better representation of background model error than the EOF.

The results of the SLA comparison shows that the BHM ensemble spread generally fails to represent the areas of maximum misfit errors in particular in the Levantine basin. The presence of strong biases in the estimation of background error due to the uncertainty in mean dynamic topography field only partially explain the observed results. The horizontal covariance structure of the model error seems not be easily represented. It is useful to remind that the MFS assimilation scheme uses Gaussian homogeneous functions that depends only on the distance between points. The ensemble spread seems at this point not capable to give a better estimates of the horizontal model errors.

Chapter 6

Conclusions and Future Directions

In this work we presented a new methodology for ocean ensemble forecasting. The novelty consists in the derivation of an "objective" method to represent the uncertainty of the winds at the sea surface. Using a Bayesian approach in a similar contest to the work of Royle et al. (1999 [83]) and Berliner and Wikle 2007 ([6]) we derive a Bayesian Hierarchical Model (BHM) to represent the full posterior distribution of the wind forcing given the available data (see Chapter 3). The BHM proved to be able to effectively meld the ECWMF winds with high-resolution scatterometer observations. The final BHM wind product is *i*) well anchored to the mean atmospheric flow that is depicted by ECMWF and *ii*) represents the uncertainty of the wind forcing at the small spatial scales that are unresolved by state of the art Numerical Weather Prediction systems.

The posterior distribution of the wind field derived from a BHM has been sampled to force an ensemble of ocean forecast (Chapter 3). The BHM winds generate an ocean initial condition perturbation that is constrained by the assimilation of in-situ ocean data and that concentrates the ocean uncertainty in the eddy field. The differential Ekman pumping that is produced by the BHM winds produces modification in the vertical stratification of the water column. Marshall et al. (2002 [59]) suggested that the restoring force that is exited by vertical advection is the baroclinic instability . Then BHM wind perturbation, acting on the vertical stratification of the fluid, is able to generate a large ocean ensemble variability even in the short time scale of the forecast.

The relation between ensemble variance and forecast error has been studied in the last part of this work (Chapter 5). The results of the analysis of temperature and salinity data are encouraging and indicate that the forecast error is likely to be larger where the ocean ensemble variability is stronger. We tested this hypothesis in a likelihood sense and we concluded that the BHM ensemble provides a better representation of model background error than the statistics that are currently used in the assimilation scheme of the Mediterranean Forecasting System. However a similar conclusion can not be drawn for the horizontal structure of background model error.

Numerous directions of development are left open. In particular we identify in the integration between data assimilation and ensemble forecast the added value of this work that need to be further expanded. The Bayesian sequential filtering reduces the data assimilation step to the identification of the weights to be associated with the members of an ensemble forecast. This technique is extremely power since it does not require any assumption of normality and linearity but it is prohibitive in high-dimensional problems. However the introduction of new sampling schemes (Merwe et al. 2000[60], Andrieu et al. 2003 [2] and Berliner and Wikle 2007 [6]) and the increased computational power that is provided by sharing computation network such the Grid provide the theoretical and technical framework to develop new high-dimensional Monte Carlo applications.

Appendix A

Full Conditional A1-E1

In this appendix we present the formulation of the full conditional distributions that enter the Gibbs sampler for the BHM version A1-E1. The Gibbs sampler can be implemented following the pseudo-code presented in section 2.5.3. See Chapter 3 for a description of all the random variables that enter the full conditional distributions.

---

1. Full conditional for U

   $$\begin{aligned} [U^t|\cdot] &\propto [S^t|U^t, \sigma_s^2][A^t|U^t, \sigma_a^2] \\ & \quad [U^t|P^t, \theta_{ux}, \theta_{uy}, \sigma_u^2] \end{aligned}$$

   the full conditional for U is a normal distribution:

   $$[U^t|\cdot] \sim N(A^{-1}B, A^{-1})$$

   where:

   $$\begin{aligned} A &= K_s' K_s/\sigma_s^2 + K_a' K_a/\sigma_a^2 + I/\sigma_u^2 \\ B &= S_u^{t'} K_s/\sigma_s^2 + A_u^{t'} K_a/\sigma_a^2 \\ & \quad +(\theta_{ux} D_x P^t + \theta_{uy} D_y P^t)'/\sigma_u^2 \end{aligned}$$

2. Full conditional for V

   $$\begin{aligned} [V^t|\cdot] &\propto [S^t|V^t, \sigma_s^2][A^t|V^t, \sigma_a^2] \\ & \quad [V^t|P^t, \theta_{vx}, \theta_{vy}, \sigma_v^2] \end{aligned}$$

   the full conditional for V is a normal distribution:

   $$[V^t|\cdot] \sim N(A^{-1}B, A^{-1})$$

   where:

   $$\begin{aligned} A &= K_s' K_s/\sigma_s^2 + K_a' K_a/\sigma_a^2 + I/\sigma_v^2 \\ B &= S_v^{t'} K_s/\sigma_s^2 + A_v^{t'} K_a/\sigma_a^2 \\ & \quad +(\theta_{vx} D_x P^t + \theta_{vy} D_y P^t D_y)'/\sigma_v^2 \end{aligned}$$

3. Full Conditional for α

$$\begin{aligned}
[\alpha^t|\cdot] \quad \propto \quad & [A_p^t|K_p(\mu+\Phi\alpha^t),\sigma_{ap}^2] \\
& [U^t|\mu+\Phi\alpha^t,\theta_{ux},\theta_{uy},\sigma_u^2] \\
& [V^t|\mu+\Phi\alpha^t,\theta_{vx},\theta_{vy},\sigma_v^2] \\
& [\alpha^t|\lambda]
\end{aligned}$$

the full conditional for α is a normal distribution:

$$[\alpha|\cdot] \sim N(A^{-1}B, A^{-1})$$

where:

$$\begin{aligned}
A &= \Phi' K_p' K_p \Phi/\sigma_{ap}^2 + \Phi'(\theta_{ux}D_x+\theta_{uy}D_y)'(\theta_{ux}D_x+\theta_{uy}D_y)\Phi/\sigma_u^2 + \\
& \quad \Phi'(\theta_{vx}D_x+\theta_{vy}D_y)'(\theta_{vx}D_x+\theta_{vy}D_y)\Phi/\sigma_v^2 + \Lambda^{-1} \\
b &= (A_p^t - K_p\mu)' K_p \Phi/\sigma_{ap}^2 + \\
& \quad (U^t - (\theta_{ux}D_x+\theta_{uy}D_y)\mu)'(\theta_{ux}D_x+\theta_{uy}D_y)\Phi/\sigma_u^2 + \\
& \quad (V^t - (\theta_{vx}D_x+\theta_{vy}D_y)\mu)'(\theta_{vx}D_x+\theta_{vy}D_y)\Phi/\sigma_v^2
\end{aligned}$$

4. Full conditioanl for λ_i

$$[\lambda_i|\cdot] \quad \propto \quad \prod_{t=1}^{T}([\alpha^t|\Lambda])[\lambda_i]$$

the full conditional for λ_i is an Inverse Gamma:

$$[\lambda_i|\cdot] \sim IG(\xi, r)$$

where:

$$\begin{aligned}
\xi &= \frac{T}{2} + \xi_{prior} \\
r &= \frac{1}{r_{prior}} + \frac{1}{2}\sum_{t=1}^{T}(\alpha_i^t)^2
\end{aligned}$$

5. Full Conditional for σ_u^2

$$[\sigma_u^2|\cdot] \quad \propto \quad \prod_{t=1}^{T}([U^t|P^t,\theta_{ux},\theta_{uy},\sigma_u^2])[\sigma_u^2]$$

the full conditional for σ_u^2 is an Inverse Gamma:

$$[\sigma_u^2|\cdot] \sim IG(\xi, r)$$

$$\begin{aligned}
\xi &= \frac{kT+L}{2} + \xi_{prior} \\
r &= \frac{1}{r_{prior}} + \sum_{t=1}^{T+L} (U^t - (\theta_{ux} D_x P^t + \theta_{uy} D_y P^t)' \\
&\quad (U^t - (\theta_{ux} D_x P^t + \theta_{uy} D_y P^t)
\end{aligned}$$

6. Full Conditional for σ_v^2

$$[\sigma_v^2|\cdot] \quad \propto \quad \prod_{t=1}^{T}([V^t|P^t, \theta_{vx}, \theta_{vy}, \sigma_v^2])[\sigma_v^2]$$

the full conditional for σ_u^2 is an Inverse Gamma:

$$[\sigma_v^2|\cdot] \sim IG(\xi, r)$$

$$\begin{aligned}
\xi &= \frac{kT+L}{2} + \xi_{prior} \\
r &= \frac{1}{r_{prior}} + \sum_{t=1}^{T+L} (V^t - (\theta_{vx} D_x P^t + \theta_{vy} D_y P^t)' \\
&\quad (V^t - (\theta_{vx} D_x P^t + \theta_{vy} D_y P^t)
\end{aligned}$$

7. Full Conditional for σ_{ap}^2

$$[\sigma_{ap}^2|\cdot] \sim IG(\xi, r)$$

the full conditional for σ_u^2 is an Inverse Gamma:

$$[\sigma_{ap}^2|\cdot] \sim IG(\xi, r)$$

where:

$$\begin{aligned}
\xi &= \frac{kT+L}{2} + \xi_{prior} \\
r &= \frac{1}{r_{prior}} + \sum_{t=1}^{T+L} (A_p^t - K_p P^t)
\end{aligned}$$

8. Full Conditional for θ_{ux}

$$[\theta_u x|\cdot \quad \propto \quad \prod_{t=1}^{T}([U^t|P^t, \theta_{ux}, \theta_{uy}, \sigma_u^2])[\theta_{ux}|\mu_{\theta_{ux}}, \sigma_{\theta_{ux}}^2]$$

the full conditional for θ_{ux} is a Normal Distribution:

$$[\theta_{ux}|\cdot] \sim N(A^{-1}B, A^{-1})$$

where:

$$\begin{aligned} A &= \sum_{t=1}^{T}(P^{t'}D_x'D_xP^t)/\sigma_u^2 + 1/\sigma_{\theta_{ux}}^2 \\ b &= \sum_{t=1}^{T}((U^t - \theta_{uy}D_y'P^t)'D_xP^t)/\sigma_u^2 + \mu_{\theta_{ux}}/\sigma_{\theta_{ux}}^2 \end{aligned}$$

9. Full Conditional for θ_{uy}

$$[\theta_{uy}|\cdot \quad \propto \quad \prod_{t=1}^{T}([U^t|P^t, \theta_{ux}, \theta_{uy}, \sigma_u^2])[\theta_{uy}|\mu_{\theta_{uy}}, \sigma_{\theta_{uy}}^2]$$

the full conditional for θ_{uy} is a Normal Distribution:

$$[\theta_{uy}|\cdot] \sim N(A^{-1}B, A^{-1})$$

where:

$$\begin{aligned} A &= \sum_{t=1}^{T}(P^{t'}D_y'D_yP^t)/\sigma_u^2 + 1/\sigma_{\theta_{uy}}^2 \\ b &= \sum_{t=1}^{T}((U^t - \theta_{ux}D_x'P^t)'D_yP^t)/\sigma_u^2 + \mu_{\theta_{uy}}/\sigma_{\theta_{uy}}^2 \end{aligned}$$

10. Full Conditional for θ_{vx}

$$[\theta_v x|\cdot \quad \propto \quad \prod_{t=1}^{T}([V^t|P^t, \theta_{vx}, \theta_{vy}, \sigma_v^2])[\theta_{vx}|\mu_{\theta_{vx}}, \sigma_{\theta_{vx}}^2]$$

the full conditional for θ_{vx} is a Normal Distribution:

$$[\theta_v x|\cdot] \sim N(A^{-1}B, A^{-1})$$

where:

$$\begin{aligned} A &= \sum_{t=1}^{T}(P^{t'}D_x'D_xP^t)/\sigma_v^2 + 1/\sigma_{\theta_{vx}}^2 \\ b &= \sum_{t=1}^{T}((V^t - \theta_{vy}D_y'P^t)'D_xP^t)/\sigma_v^2 + \mu_{\theta_{vx}}/\sigma_{\theta_{vx}}^2 \end{aligned}$$

11. Full Conditional for θ_{vy}

$$[\theta_v y|\cdot \quad \propto \quad \prod_{t=1}^{T}([V^t|P^t, \theta_{vx}, \theta_{vy}, \sigma_v^2])[\theta_{vy}|\mu_{\theta_{vy}}, \sigma_{\theta_{vy}}^2]$$

the full conditional for θ_{vy} is a Normal Distribution:

$$[\theta_{vy}|\cdot] \sim N(A^{-1}B, A^{-1})$$

where:

$$\begin{aligned} A &= \sum_{t=1}^{T}(P^{t'} D'_y D_y P^t)/\sigma_v^2 + 1/\sigma_{\theta_{vy}}^2 \\ b &= \sum_{t=1}^{T}((V^t - \theta_{vx} D'_x P^t)' D_y P^t)/\sigma_v^2 + \mu_{\theta_{vy}}/\sigma_{\theta_{vy}}^2 \end{aligned}$$

Appendix B

Ocean model set-ups and data assimilation scheme

B.1 Model Equation

The Navier Stokes equations and a non-linear equation of state are used to describe the ocean dynamics. In order to make the problem treatable we adopt the following assumption:

- Buissinesq approximation: density variation are considered only for their contribution to the buoyancy force
- hydrostatic hypothesis: the vertical velocity equation is reduced to a balance between buoyancy force and pressure gradient
- the geopotential surfaces are assumed to be spheres and the gravity field is assumed parallel to the earth's radius, then the equation are written in spherical coordinate (λ, ϕ, z) where λ is longitude, ϕ latitude and z is depth
- thin-shell approximation: the ocean depth is neglectable compared to the earth radius
- turbolent closure approximation: the small scale processes like turbolent fluxes are expressed in terms of the large-scale dynamics.
- incompressibility approximation: the divergence of the velocity vector is assumed to be zero.

The momentum equations for the zonal and meridional velocity components u and v, the hydrostatic equation for pressure p, the continuity equation, the conservation equation for potential temperature T and salinity S and the state equation for density ρ define a set of 7 equations that together with their boundary conditions describe the ocean dynamic. Those equation are analytically written as:

$$\begin{aligned}\frac{\partial u}{\partial t} &= (\zeta + f)v - w\frac{\partial u}{\partial z} - \frac{1}{2\ a\ cos\phi}\frac{\partial}{\partial \lambda}(u^2 + v^2) \\ &- \frac{1}{\rho_0\ a\ cos\phi}\frac{\partial p}{\partial \lambda} - A^{lm}\nabla^4 u + \frac{\partial}{\partial z}(A^{vm}\frac{\partial u}{\partial z})\end{aligned} \tag{B.1}$$

$$\begin{aligned}\frac{\partial v}{\partial t} &= -(\zeta+f)u - w\frac{\partial v}{\partial z} - \frac{1}{2\ a}\frac{\partial}{\partial \phi}(u^2+v^2) \\ &- \frac{1}{\rho_0\ a}\frac{\partial p}{\partial \phi} - A^{lm}\nabla^4 v + \frac{\partial}{\partial z}(A^{vm}\frac{\partial v}{\partial z})\end{aligned} \tag{B.2}$$

$$\frac{\partial v}{\partial t} \quad = \quad -\rho g \tag{B.3}$$

$$\frac{1}{a\ cos\phi}\Big(\frac{\partial u}{\partial \lambda} + \frac{\partial}{\partial \phi}[cos\ \phi v]\Big) + \frac{\partial w}{\partial z} \quad = \quad 0 \tag{B.4}$$

$$\begin{aligned}\frac{\partial T}{\partial t} &= -\frac{1}{a\ cos\phi}\Big[\frac{\partial}{\partial \lambda}(Tu) + \frac{\partial}{\partial \phi}(cos\ \phi T v)\Big] - \frac{\partial}{\partial z}(Tw) \\ &- A^{lT}\nabla^4 T + A^{vT}\frac{\partial^2 T}{\partial z^2}\end{aligned} \tag{B.5}$$

$$\begin{aligned}\frac{\partial S}{\partial t} &= -\frac{1}{a\ cos\phi}\Big[\frac{\partial}{\partial \lambda}(Su) + \frac{\partial}{\partial \phi}(cos\ \phi S v)\Big] - \frac{\partial}{\partial z}(Sw) \\ &- A^{lS}\nabla^4 S + A^{vS}\frac{\partial^2 S}{\partial z^2}\end{aligned} \tag{B.6}$$

$$\rho \quad = \quad \rho(T,S,p) \tag{B.7}$$

where a is the earth radius; $f = 2\ \Omega\ sin\phi$ is the Coriolis term with constant earth radius rate Ω; ρ_0 is the reference density. In equation B.1 and B.2, that are here written in their vorticity form (see Pedlosky 1987 [72]), the terms A^{lm} and A^{vm} are the horizontal and vertical eddy viscosity and ζ is the vorticity and it is defined as:

$$\zeta \quad = \quad \frac{1}{a\ cos\phi}\Big(\frac{\partial v}{\partial \lambda} - \frac{\partial}{\partial \phi}(u\ cos\phi)\Big)$$

In equation B.5 and B.6 the terms A^{vT}, A^{vS} and A^{lT}, A^{lS} are the vertical and horizontal diffusivity for the temperature and salinity tracers.

The ocean model OPA version 8.2 described in Madec et al. (1998 [56]) discretizes equation B.1 to B.7. Here we use the free-surface implementation; in the following η stands for sea surface height and it is a prognostic variable.

B.2 Boundary and Initial Conditions

The depth of the ocean bottom $z = -H(\lambda, \phi)$, the sea surface height $z = \eta(\lambda, \phi, t)$ and the coastlines define the ocean boundaries. Since this work deals with wind perturbations we will concentrate on the description of the air-sea interaction, namely the boundary conditions at $z = \eta$.

The surface boundary condition for momentum is:

$$A^{vm}\frac{\partial \mathbf{u_h}}{\partial z}|_{z=\eta} \quad = \quad \frac{\tau}{\rho_0}$$

where $\tau = (\tau_{\mathbf{u}}, \tau_{\mathbf{v}})$ represents the zonal and meridional wind stress components and $\mathbf{u_h} = (\mathbf{u}, \mathbf{v})$. The wind stresses are derived from wind at 10 m height $\mathbf{U_{10}} = (\mathbf{U_{10}}, \mathbf{V_{10}})$ following the bulk aerodynamic formula suggested by Smith (1980 [86]):

$$\tau \quad = \quad \rho_a C_D |\mathbf{U_{10}}| \mathbf{U_{10}}$$

where ρ_a is the air density and C_D is the drag coefficient that is derived using the Hellermann and Rosenstein approximation [39] that relates the drag coefficient to the wind speed and the air-sea temperature difference.

The boundary condition for heat flux is:

$$A^{vT}\frac{\partial T}{\partial z}|_{z=0} \quad = \quad \frac{Q}{\rho_0 C_p}$$

where C_p $[J/kg/^{\circ}K]$ is the ocean heat capacity constant and Q $[W/m^2]$ is the heat budget and consists of the solar radiation flux Q_S minus the net long-wave radiation flux Q_B, the latent heat flux Q_E and the sensible heat flux Q_H. The details of the heat fluxes parameterization are described in Castellari et al. (1998 [12]). The heat fluxes are affected by wind perturbation through two processes; wind speed is an input in the parameterization scheme of sensible and latent heat; variations of sea surface temperature due to wind driven circulation affect the parameterization schemes of sensible and latent heat and the net outgoing long-wave flux.

The water flux boundary condition states that a particle of water can enter or escape the sea surface only through precipitation or evaporation:

$$w \quad = \quad \frac{D\eta}{Dt} - (E - P)$$

P and E are precipitation and evaporation; D is the total derivative $\frac{D}{Dt} = \frac{\partial}{\partial t} + u_h \cdot \nabla$.

The salinity boundary condition is coupled to the water flux by the $E - P$ term:

$$\rho_0 A^{vS}\frac{\partial S}{\partial z}|_{z=0} \quad = \quad (E - P) S_{z=0} \rho_0$$

where $S_{z=0}$ is the model surface salinity.

The unreliability of precipitation data over the oceans has led to alternative formulation of the surface fluxes dependent on evaporation and precipitation . A Newtonian relaxation to monthly mean salinity S^* has been used to replace $E - P$:

$$E - P \quad = \quad \frac{1}{\rho_0 \gamma}\frac{S - S^*}{S}$$

where γ $[m^2 s/kg]$ is the salinity relaxation term that defines the strength of the constraint.

At the ocean bottom there is an insulating condition for heat and salt. The vertical boundary condition for momentum is a function of a drag coefficient and it takes in consideration the effect of eddy kinetic energy due to tides. The bottom flow is required to obey the no-normal flow condition. The lateral boundary condition is no-slip with insulating walls for heat and salt.

The initial condition for the ocean simulation are taken from existing analysis field produced by MFS.

B.3 Sub grid scale parametrization

The vertical mixing coefficient is parametrized as a function of the Richardson number using the Pakanowsky and Philander-PP (1981 [71]) scheme:

$$\begin{aligned} A^{vT} &= \frac{10^{-2}}{(1+5(N^2/(\partial\mathbf{U}/\partial\mathbf{z})^\mathbf{2}))^\mathbf{2}} + (1.5\times10^{-4}) \\ A^{vm} &= \frac{A^{vT}}{(1+5(N^2/(\partial\mathbf{U}/\partial\mathbf{z})^\mathbf{2}))^\mathbf{2}} + (3\times10^{-4}) \end{aligned}$$

The horizontal eddy viscosity A^{lm} and diffusivities A^{lT}, A^{lS} are set constants.

B.4 Model Implementations

Two different set ups of OPA Ocean General Circulation Model (OGCM) were used in this study:

- MFS1671: it is the operational model used within MFS project. It has 72 vertical levels and a horizontal resolution of 1/16 X 1/16 degrees. The model domain covers the Mediterranean Sea and a portion of the Atlantic ocean, see panel a of figure B.1. The coastlines resolves 49 islands and the bathymetry was derived from Digital Bathymetric Data Base. The lateral boundary of the model in the Atlantic box are closed and the model fields in proximity of the boundary are relaxed toward climatology at all depths (see Tonani et al. [91] for details on the sponge layer). A climatological wind forcing is applied in the Atlantic box region. The horizontal eddy viscosity and diffusivity are set to $5\times10^9[m^4/s]$ and $3\times10^9[m^4s]$ respectively.

- MFS471: it is a low resolution set up expecially developed to decrease the amount of computational resources needed for a single integration. It has 72 vertical levels and an horizontal resulution of 1/4 X 1/4 degrees. The geographical domain covered by the MFS471 implementation includes a smaller Atlantic Box that extends to 9.5W, see panel b of figure B.1. In the Atlandic box the model fields are relaxed toward climatology and forcing is switched off. The coastlines resolves 6 islands and the bathymetry has been interpolated from the MFS1671 bathymetry. The horizontal eddy viscosity and diffusivity are set to $5\times10^{10}[m^4/s]$ and $3\times10^{10}[m^4/s]$ respectively.

Both models were coupled with the MFS operational assimilation scheme.

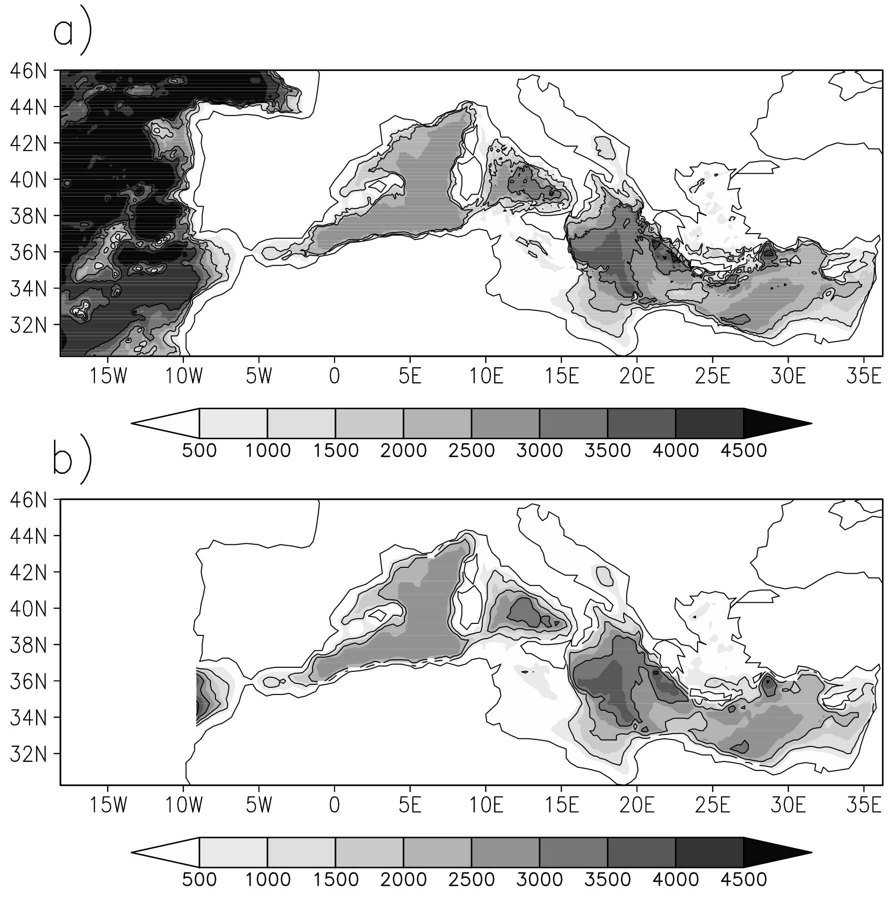


Figure B.1: a) MFS1671 bathymetry and domain MFS1671 [m]. b) same for MFS471.

B.5 Assimilation Scheme

The MFS applies a multivariate optimal interpolation scheme that is based on the System for Ocean Forecasting and Analysis (SOFA) that was originally developed by De Mey and Benkiran (2002 [23]). The initial set up for the Mediterranean is described in Demirov et al. (2003 [25]), while further development where introduced by the work of Dobricic et al. (2005 [26]).

The optimal interpolation scheme is an approximation of the Kalman filter (Kalman 1960 [41]) where the time evolving error covariance matrix is replaced by a background error covariance matrix. The analysis are estimated correcting the background model fields $\mathbf{x}_b$ with increments that depends on available observations $\mathbf{y}$. This can be written as:

$$\begin{aligned} \mathbf{x}_a &= \mathbf{x}_b + K[\mathbf{y} - \mathcal{H}(\mathbf{x}_b)] \\ K &= BH^T(HBH^T + R)^-1 \end{aligned}$$

where $\mathbf{x}_a$ is the analysis field and $\mathcal{H}$ is a non-linear observational operator that move the model state vector to the observations. The matrix K is known as Kalman Gain and it is defined by a linear combination of the background error covarinace B, the linearized observational operator H and the observational error covariance R. The insertion of an observation $\mathbf{y}$ will impact the state vector $\mathbf{x}_b$ according to the distance between observation and the model grid points and the multivariate correlations. The information necessary to perform these two tasks are contained in the error covariance matrix. To explain this we look at the increment part of the assimilation scheme:

$$\delta\mathbf{w} = BH^T(HBH^T + R)^{-1}\,\mathbf{d}$$

all the terms $(HBH^T + R)^{-1}\mathbf{d}$ are defined in the observation space; the misfit $\mathbf{d} = (\mathbf{y} - H(\mathbf{x}_b))$ is weighed by the measurement error and by the model covariance reduced by the observational operator H and its adjoint H^T to the observation location and kind. Assume that d is a scalar temperature misfit, then the H operator and its adjoint H^T reduces HBH^T to a scalar that retains only the variance of model temperature on the required position. The correction $c = (\sigma_b^2 + \sigma_r^2)^{-1}\; d$ is then defined by the misfit and the amplitude of model and observation variance. The term H^T move the correction back to the model grid points that are directly interested by the observation and then the matrix B expand the increments in space and on the other variables of the model state vector. This is the only mechanism by which information can be transferred from observed to the unobserved variables (Ricci et al. 2004 [80]).

B.6 The background error covariance matrix

In SOFA the error covariance matrix is seperated in horizontal and vertical components:

$$B = S^T B_r S$$

where S contains the vertical multivariate error covariances that are represented by Empirical Orthogonal Functions (EOFs) , and B_r is defined as:

$$B_r \;=\; \Lambda^{1/2} C \Lambda^{1/2}$$

where C are the horizontal covariances modelled as Gaussian function of distance and $\mathbf{\Lambda}$ are the eigenvalues of the vertical EOFs. The Empirical Orthogonal Functions analysis, also known as the Principal Components analysis, is a widely used tool in atmospheric sciences and oceanography (Lorentz 1956 [52], 1998 Preisendorfer [78],Sparnocchia et al. 2003 [87] and many others) . This method is commonly used to reduce the dimension of the problem and to transform interdependent coordinates into significant and independent patterns (De Mey and Benkiran 2002 [23]).

B.7 Vertical EOF

The fundamental assumption underneath the computation of the vertical structure of the background error covariance matrix is that the variability of the model field in a long simulation is a statistical representation of the model error. This approach is alternative to the usage of misfit between in-situ data and model fields that was at basis of the work of Sparnocchia et al. (2003 [87]). The usage of model simulation presents the advantage that all the cross-covariance between variable of the state vector can be estimated, where in-situ data allow only the estimation of temperature and salinity covariances in a limited set of locations in the basin where CTD or ARGO profile are available.

Given the separation between vertical and horizontal scales it would be sufficient to specify a single vertical structure of the background covariance matrix and apply it to whole basin. However given the high variability of the water masses of the Mediterranean Sea this approach prove to be ineffective. Separate sets of EOFs were computed for 13 regions of the Mediterranean Sea on the basis of a model simulation ranging from 1993 to 1999. Each EOF is multivariate and was build from a singular value decomposition of the matrix $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ whose column are the multivariate vertical state vector for each point of the model grid belonging to a certain region. For instance the column k is:

$$\mathbf{x}_k = \left[\frac{\delta\eta}{\sigma_\eta}, \frac{\delta\psi}{\sigma_\psi}, \frac{\delta T_1}{\sigma_T}, \ldots, \frac{\delta T_n}{\sigma_T}, \frac{\delta S_1}{\sigma_S}, \ldots, \frac{\delta S_n}{\sigma_S}\right]^T$$

were η is the sea surface height, ψ is the stream function, $T_1, \ldots, T_n$ are temperature values along the vertical profiles and similarly for salinity S; δ indicates the difference between the daily averaged field and the seasonal mean the anomaly and the standard deviation σ is used here to non-dimensionalize the state vector. An additional scaling factor was applied to accounts for geometric considerations; for a detailed description of the method see Dobricic et al. 2005 [26]. The stream function ϕ was diagnostically computed from the barotropic velocity.

The OI scheme relies on the specification of a background covariance matrix that is constant in time, i.e the assimilation scheme do not evolve the error covariance according to the insertion of data and the integration of the dynamical operator. Anyway the seasonal variability was captured grouping the EOFs in blocks that accounts for the 4 seasons.

Appendix C

A Grid application to ocean ensemble forecasting

Here I present a review of the work I have co-authored 'Very Large ensemble ocean forecasting experiment using the Grid computing infrastructure' [74]. Some of the results that were omitted in the paper for synthesis sake are presented here.

C.1 The ocean model and ensemble method

The model used in this experiment is a copy of the MFS operational system from July 2005. It is a MOM1.1 ocean model (1984 [18]) implemented in the Mediterranean at $1/8° \times 1/8°$ resolution and 31 levels in vertical. Results from this model implementation are described in Pinardi and Masetti (2000 [75]) and Demirov et al. (2003 [25]). The model equations are the rigid-lid approximation to the primitive equations for oceanic fluids. The ocean state variable that are forecasted are temperature, salinity, density, pressure, three velocity components and sea level. The size of the problem in terms of dimension of the state vector is somewhat like 10^7 and it is comparable in size to a coarse resolution global ocean model. The atmospheric forcing is provided by the European Centre for Medium-range Weather Forecast (ECMWF) at a resolution of half a degree.

A random perturbation scheme of the initial condition is used here to generate an ensemble of ocean forecast. A detailed description of the scheme can be found in Chapter 3.

C.2 The Italian Grid Infrastructure

The ensemble experiment was carried out on a distributed computing network, the so-called Grid (Foster and Kesselman 1999 [35]). This system is characterised by the implementation of a Grid Production Framework that is oriented on large-scale resource sharing and high performances applications, making this approach different from conventional distribuited computing.

The institutions that participate in creating and maintaining the Grid infrastructure are dived in abstract entities called Virtual Organisations (VO); each VO groups into the same administrative domain users and resources.

The Italian Grid infrastructure is made of 30 sites that are equipped with Computing Elements (CE) formed by 10 to hundreds of nodes that are called Worker Node (WN). Each CE also provide a disk-based storage service that is organised in Storage Elements which capacity range from hundreds of gigabytes up to hundred of terabytes. The CEs are the entry points of queues that are managed by Local Resources Management System (LRMS). The jobs are submitted to the CEs by a Resource Broker (RB) to which the user can connect through a User Interface (UI). The experiment described in this section were run on a maximum number of 15 different sites.

C.3 Ensemble Forecast Experiment

A total of 67 ensemble forecast experiments have been carried out at different hours and week days over a period of 20 days in order to test the Grid efficiency through its normal workload cycle. No special arrangement has been made to the Grid configuration and operation policies for this experiment. Each ensemble forecast experiment is designed to launch 1000 jobs within a total time of five hours, after which the jobs are deleted without paying any attention to their status. The ensemble forecast experiments were done following a 3 phases procedure:

- Phase 1:

  the model input files and the executable code are uploaded to the closest SE belonging to the INFN-CNAF CE in Bologna. Then the SE replicates these file to 15 sites that are distributed over the national territory. The input file replication avoid the bottleneck due to multiple and simultaneous request from hundreds of WNs to a single SE. This is a crucial point because we observed a significant percentage of failures (up to 10%) that are due to unsuccessful copies of input files. The size of files that are transferred is in the order 100 Mb.

- Phase 2:

  the jobs are submitted to a INFN-CNAF RB that looks for the best available CE to execute the jobs. The RB interrogates an Information Service that provide the status of computational and storage resources and the File Catalogue that provide information on the location of the required data. A quasi-parallel submission of 1000 jobs on the WNs is handled by the LRMS. The jobs are submitted only to the CEs that belong to the same farm of the SE were the input file are located.

- Phase 3:

  each time a job finished, a procedure for the downloading the model output files is activated. Only a limited portion (1MB) of the output file produced by the ocean simulation is recovered. All the jobs that are still pending after 5 hours are cancelled.

The wall-clock-time for 67 ensemble forecast experiments in shown in Figure C.1. The results indicate that a minimum number of 200 jobs were finished in 2 hours and at least 450 in five hours. The user can specify to the RB the requirements that need be address in order to submit a job to a certain CE. In this experiment we imposed very

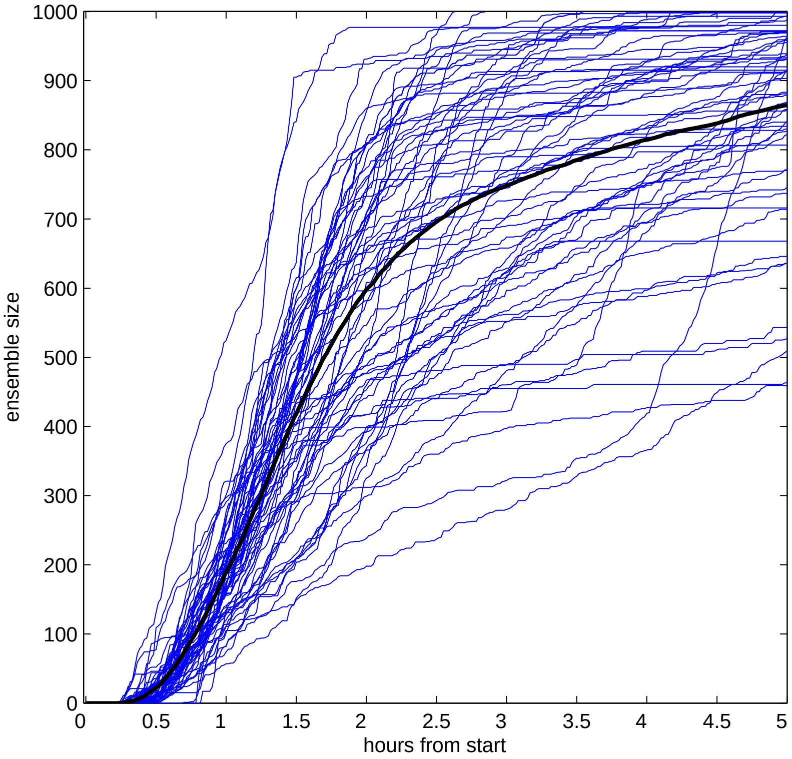


Figure C.1: Numeber of member jobs successfully carried out from 67 ensemble experiments launched on the Grid as a function of time. Each experiment is set to last maximum 5 hours and should run as much as 1000 jobs. The central black curve is the average.

generic constraint to ensure the largest usage of Grid CE and reasonable efficienty. We requirement imposed were that at least a CPU is free on the CE and the time limit of the queues is greater then 80 minutes. Figure C.2 shows how the jobs were distributed on the Grid network. The large farmer T1 INFN sustained more then the 40 % of the overall work load, but nevertheless the importance of all the other smaller CEs overcome in importance the single contribution of the T1 farm.

The ensemble experiments never used more the 20% of the Grid computing resources.

C.4 Results

We present the results for one single experiment of the 67 produced, the forecast from 16 to 25 November 2005. In this analysis we concentrate on the Sea Surface Height (SSH) field that is two-dimensional field (small in term of data retrieving) but it is highly informative about ocean circulation since the horizontal gradients of SSH are in balance with the geostrophic velocities. These consideration justified to working choice to retrieve only the SSH field and discard the rest of the 3-dimensional model output.

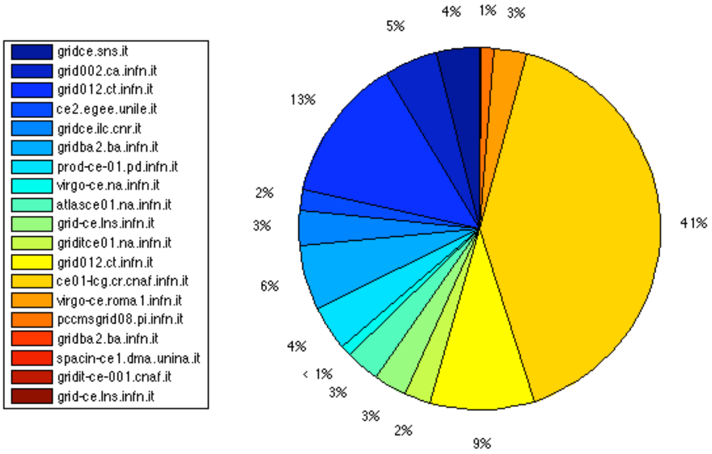


Figure C.2: Overall job distribution on the 20 Computing Elements of Italian Grid that were used throughout the 67 experiments.

Since this experiment mainly concerns the technical aspects of using the GRID as a tool for ensemble forecasting we presents only a limited analyses of the oceanographic results; in particular we will be dealing with the ensemble spread that is defined as the standard deviation of the ensemble members around their mean.

The initial perturbation is in the order of few tens of centimetres and it is homogeneously distributed in the deep part of the basin, see panel a of figure C.4. After ten day of forecast the ensemble spread reaches values of several centimetres (up to 4 cm , see panel b of figure C.4) that are in agreement with the analysis error standard deviation that was estimated to be 5 cm for a three year period (Dobrici et al. 2005 [26]). The spatial distribution of ensemble standard deviation at the end of the 10 days of forecast is limited in extent and it is concentrated along strong current jets, frontal structures and eddy borders. While the initial perturbation do not present a connection to the dynamical structure of circulation, the final ensemble spread is clearly connected to regions of strong ocean dynamics. The areas where we observe a maximum growth of the initial perturbations are thought to be the least predictable regions of the flow field.

Figure C.3 shows the growth of the ensemble standard deviation for areas having different ensemble spread on the last day of the forecast. The growth of initial condition perturbation is almost linear in all cases after the 2 day. Within the first day we observe a fast growth of the ensemble standard deviation that is probably due to geostrophic adjustment. The region with maximum perturbation growth account for less of 2% of the total area of the Mediterranean Sea, where in the largest part of the domain we actually observe a decrease of the amplitude of the initial signal.

The large computing power provided by the Grid allowed us to test the sensitivity of the ensemble spread along with the size of the ensemble. Figure C.5 show the ensemble standard deviation at the forecast end (for the same experiment) for three ensemble sizes; 10, 50 and 200 members. The fact that the experiment with 100 members is qualitatively equal to the 1000 members case presented in figure C.4 indicates a clear saturation of the ensemble variance after a few hundred members. Such number is clearly a function of the perturbation scheme and the ocean model but recent work on storm surge indicates a that a few hundreds members would saturate the ensemble standard deviation (Lamourou et at. 2006 [43]).

C.5 Discussion

In this experiment we have shown a large ensemble experiment carried out with a state of the art operational forecasting system. The forecast was performed with a primitive equation, eddy permitting model. We showed that the ensemble variance saturates at roughly 200-300 members and that it concentrates in region of high dynamics of the Mediterranean circulation that account for a small portion of the basin. All the consideration about the relevance of this information in terms of model error and predictability are left to Chapter 3 and 4 of this thesis.

The major aim of this work was to demonstrated that a very large ocean ensemble activity can be sustained by the Italian Grid in normal work-load conditions. Approximately 500 members can be executed in operational wall clock time, i.e. within 5 hours after the submission of the first job.

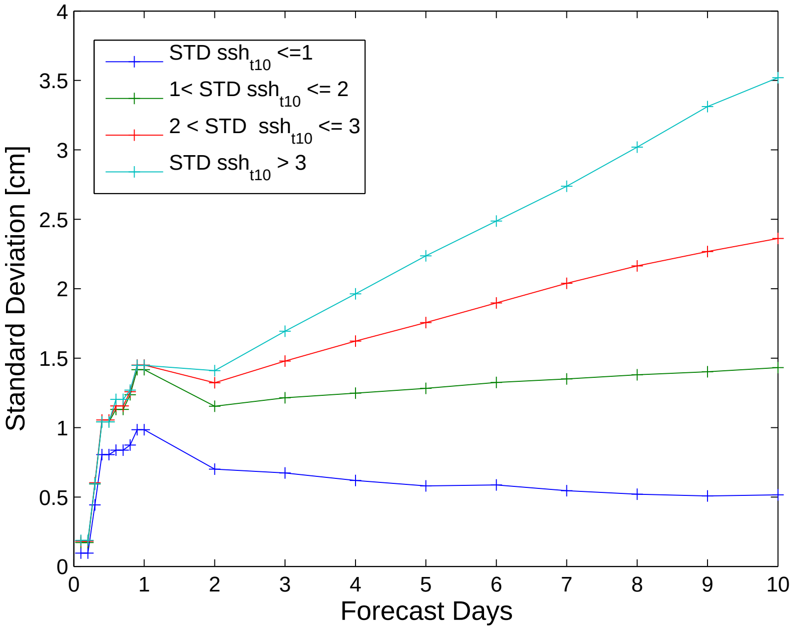


Figure C.3: The growth in amplitude of the standard deviation (std) for the 10 days ensemble forecast experiment with 500 members. Different curves are averages done in regions of figure C.4 with different std at day 10.

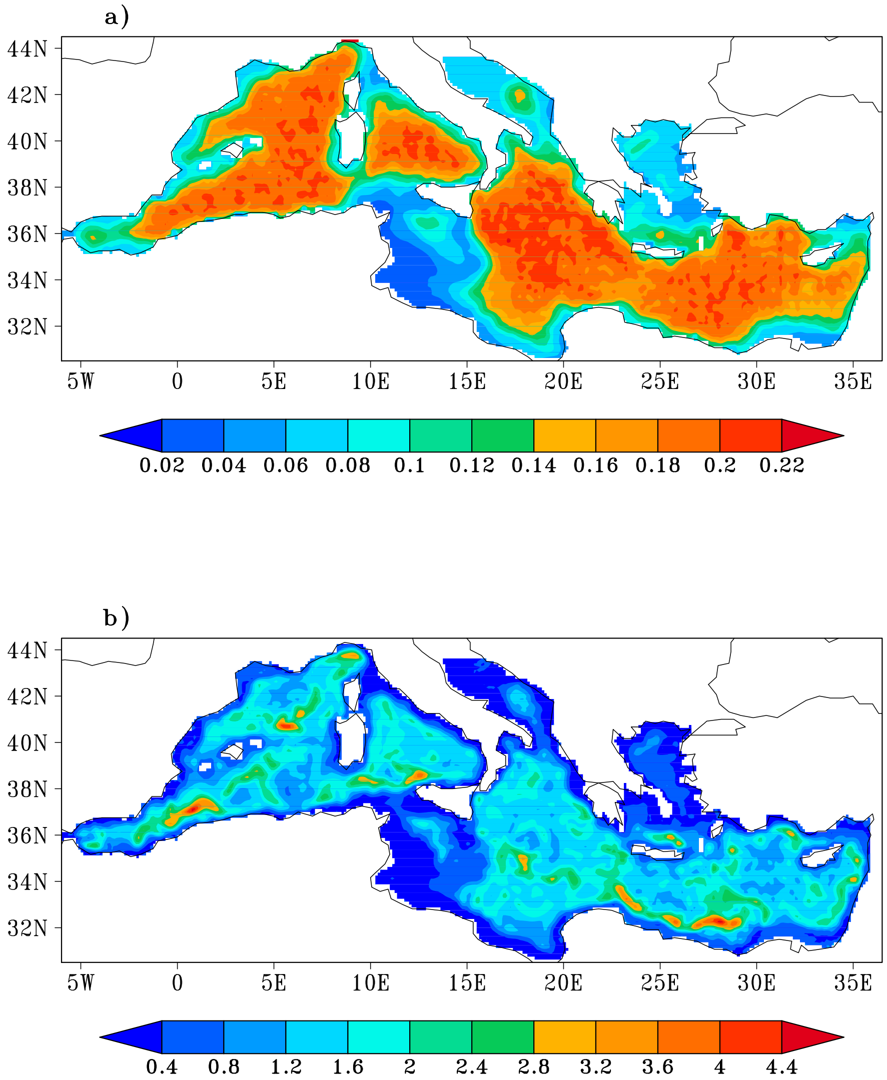


Figure C.4: The amplitude and structure of the standard deviation at forecast day 1 (a) and 10 (b) for sea surface height. The 500 members ensemble mean has been subtracted and the units are cm.

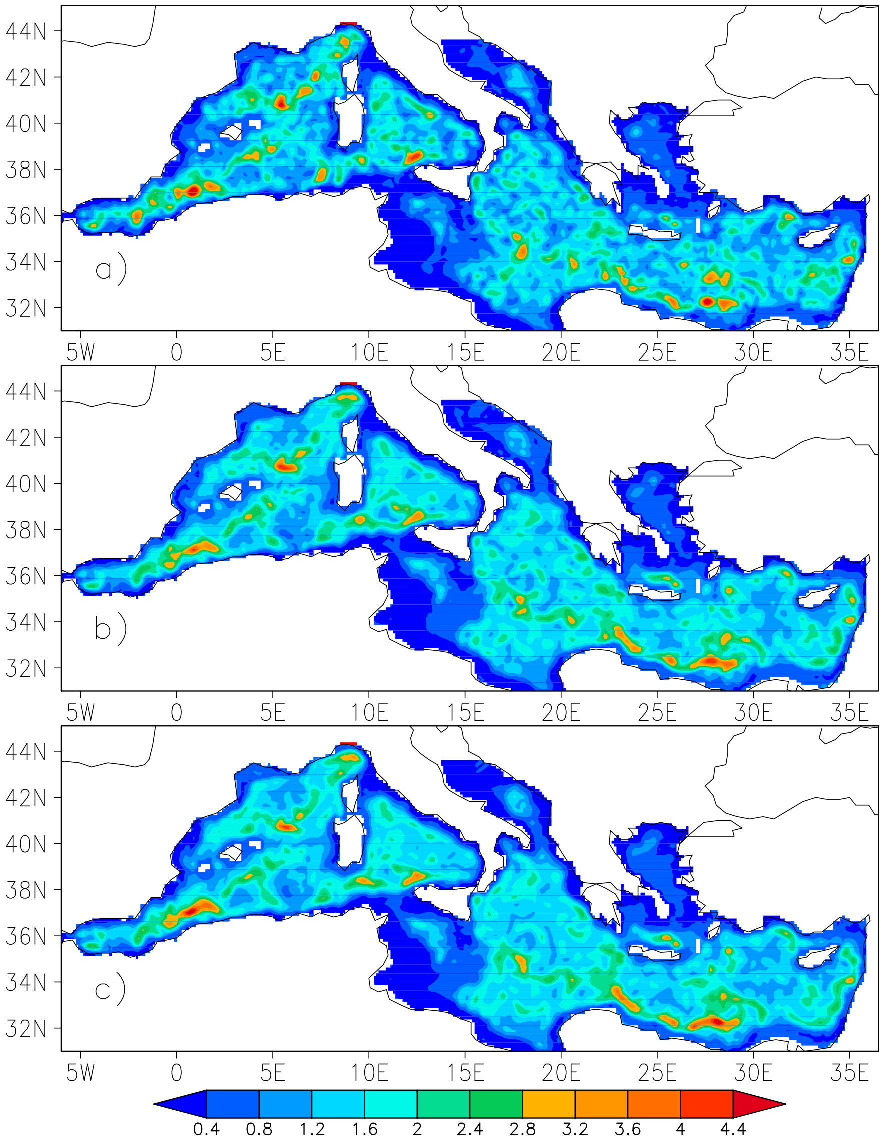


Figure C.5: The amplitude and structure of the standard deviation at forecast day 10 for sea surface height computed with 10 members (a), 100 members (b) and 200 members (c).

Bibliography

[1] Anderson, D. L. T., and Gill, A. E. Spin-up of a stratified ocean, with application to upwelling. *Deep-Sea Research 22* (1974), 583–596.

[2] Andrieu, C., Freitas, N. D., Doucet, A., and Jordan, M. I. An introduction to "mcmc" for machine learning. *Machine Learning 50* (2003), 5–43.

[3] Arulampalam, M. S., Maskell, S., Gordon, N., and Clapp, T. A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEE Transactions on Signal Processing 50*, 2 (2002), 174–188.

[4] Berger, J. O. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, 1985.

[5] Berliner, L. M., Wikle, C. K., and Cressie, N. Long-lead prediction of Pacific SSTs via bayesian dynamic modelling. *Journal of Climate 13* (2000), 3953–3968.

[6] Berliner, M., and Wikle, C. K. Approximate importance sampling monte carlo for data assimilation. *Physica D 230* (2007), 37–49.

[7] Bernardo, J. M., and Smith, A. F. M. *Bayesian Theory*. John Wiley and Sons, Inc., New York, 1994.

[8] Buizza, R. The ECMWF Ensemble Prediction System. In *Predictability of Weather Climate*, T. Palmer and R. Hagedorn, Eds. Cambridge University Press, 2006.

[9] Buizza, R., and Palmer, T. N. The singular-vector structure of the atmospheric general circulation. *Journal of Atmopheric Sciences 52*, 9 (1995), 1434–1456.

[10] Buongiorno-Nardelli, B., Larnicol, G., D'Acunzo, G., Santoleri, R., Marullo, S., and LeTraon, P.-Y. Near Real Time SLA and SST products during 2-years of MFS pilot project: processing, analysis of the variability and of the coupled patterns. *Annales Geophysicae 21* (2003), 103–121.

[11] Casella, G., and Berger, R. G. *Statistical Inference*. Duxbury, 2001.

[12] Castellari, S., Pinardi, N., and Leaman, K. A model of air-sea interaction in the Mediterranean Sea. *Journal of Marine Systems 18* (1998), 89–114.

[13] Castellari, S., Pinardi, N., and Leaman, K. Simulation of water mass formation processes in the Mediterranean Sea: Influence of the time frequency of the atmospheric forcing. *Journal of Geophysical Research 105*, C10 (2000), 24,157–24,181.

[14] Charney, J. The dynamcs of long waves in baroclinic westerly currents. *Journal of Meteorology 4* (1947), 135–162.

[15] Chelton, D. B., Schlax, M. G., Freilich, M. H., and Milliff, R. F. Satellite measurements reveal persistent small-scale features in ocean winds. *Science 303* (2003), 978–983.

[16] Chin, T. M., Milliff, R. F., and Large, W. G. Basin-scale, high wavenumber sea surface wind fields from a multiresolution analysis of scatterometer data. *Journal of Atmospheric and Ocean Technology 15* (1998), 741–763.

[17] Cohenamd, A., Daubechies, I., and Vital, P. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis 1* (1993), 58–81.

[18] Cox, M. D. A primitive equation, 3-dimensional model of the ocean. Tech. Rep. 1, GFDL Ocean Group Tech. Rep., 1984.

[19] Daubechies, I., and Paul, T. Wavelet - some applications. In *Proceedings of the International Conference on Mathematical Physics* (1987), M. Mebkhout and R. Seneor, Eds., pp. 675–686.

[20] de Freitas, J. F. G. *Bayesian Methods for Neural Networks.* PhD thesis, Trinity College and University of Cambridge, 1999.

[21] DelSole, T. Predictability and Information Theory. part 1: Measures of Predictability. *Journal of Climate 61* (2004), 2425–2440.

[22] DelSole, T. Predictability and Information Theory. part II: Imperfects Forecasts. *Journal of Atmopheric Sciences 62*, 3368-3381 (2004).

[23] DeMey, P., and Benkiran, M. A multivariate reduced-order optimal interpolation method and its application to the mediterranean basin-scale circulation. In *Ocean Forecasting*, S. Verlag, Ed. N.Pinardi and J. Woods, 2002, pp. 281–306.

[24] Demirov, E., and Pinardi, N. The simulation of the Mediterranean Sea circulation from 1979 to 1993. part i: the interannual variability. *Journal of Marine Systems 33-34* (2002), 23–50.

[25] Demirov, E., Pinardi, N., Fratianni, C., Tonani, C., Giacomelli, L., and de Mey, P. Assimilation scheme in the Mediterranean forecasting system: operational implementaion. *Annales Geophysicae 21* (2003), 189–204.

[26] Dobricic, S., Pinardi, N., Adani, M., Bonazzi, A., Fratianni, C., and Tonani, M. An improved assimilation scheme for sea-level anomaly and its validation. *Q. J. R. Meteorol. Soc. 131* (2005), 3627–3642.

[27] Dobricic, S., Pinardi, N., Adani, M., Tonani, M., Fratianni, C., Bonazzi, A., and Fernandez, V. Daily oceanographic analyses by Mediterranean Forecasting System at basin scale. *Ocean Sciences 3* (2007), 149–157.

[28] Eady, E. Long waves and cyclone waves. *Tellus* (1), 33–52.

[29] Eady, E. The quantitative theory of cyclone development. In *Compendium of Meteorology*, T. Malone, Ed. American Meteorological Society, 1951.

[30] Ebuchi, N., Graber, H. C., and Caruso, M. J. Evaluation of wind vectors observed by QuikSCAT/SeaWinds using ocean buoy data. *Journal of Atmospheric and Ocean Technology 19* (2002), 2049–2062.

[31] Epstein, E. Stochasitc dynamic prediction. *Tellus 21* (1969), 739–759.

[32] Evensen, G. Inverse methods and data assimilation in nonlinear ocean models. *Physica D 77* (1994), 108–129.

[33] Evensen, G. The Ensemble Kalman Filter: theoretical formulation and practical implementation. *Ocean Dynamics 53* (2003), 343–367.

[34] Farrell, B. F. Small error dynamics and the predictability of atmospheric flow. *Journal of Atmopheric Sciences 47* (1990), 2191–2199.

[35] Foster, I., and Kesselman, C. *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, 1999.

[36] Freilich, M. H., and Chelton, D. B. Wavenumber spectra of Pacific winds measurements by the Seasat scatterometer. *Journal of Physical Oceanography 16* (1986), 741–757.

[37] Geman, S., and Geman, D. Stochastic relaxation, Gibbs distribution and Bayesian restoration of images. *IEE Transactions on Pattern Analysis and Machine Intelligence 6* (1984), 721–741.

[38] Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika 57* (1970), 97–109.

[39] Hellermann, S., and Rosenstein, M. Normal monthly wind stress over the world ocean with error estimates. *Journal of Physical Oceanography 13* (1983), 1093–1104.

[40] Isern-Fontanet, J., Font, J., Garcia-Ladona, E., Emelianov, M., Millot, C., and Taupier-Letage, I. Spatial structure of anticyclonic eddies in the Algerian basin (Mediterranean Sea) analysed using the Okubo-Weiss parameter. *Deep-Sea Research 51* (2004), 3009–3028.

[41] Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering 82*, Seried D (1960), 35–45.

[42] Lacarra, J., and Talagrand, O. Short-range evolution of small perturbations in a barotropic model. *Tellus 40A* (1988), 81–95.

[43] Lamourou, J., DeMey, P., Lyard, F., and Jensou, E. Control of a barotropic model of the Bay of Biscay in presence of atmosperic forcing errors. *Journal of Geophysical Research In press* (2006).

[44] Leith, C. E. Atmospheric predictability and two-dimensional turbulence. *Journal of Atmopheric Sciences 28* (1971), 145–161.

[45] Leith, C. E. Theoretical skills of monte carlo forecasts. *Monthly Weather Review 102* (1974), 409–418.

[46] Leith, C. E., and Kraigman, R. Predictability od turbulent flows. *Journal of Atmopheric Sciences 29* (1972), 1041–1058.

[47] LeTraon, P.-Y. Bains-scale oceanic circulation from satellitealtimetry. In *Oceanographic applications of remote sensign*, M. Ikeda and F. Dobson, Eds. CRC Press Inc., 1995.

[48] LeTraon, P.-Y., and Gauzelin, P. Response of the Mediterranean mean sea level to atmospheric pressure forcing. *Journal of Geophysical Research 102* (1997), 973–984.

[49] LeTraon, P.-Y., Nadal, F., and Ducet, N. An improved mapping method of multisatellite altimeter data. *Journal of Atmospheric and Ocean Technology*, 15 (2003), 522–533.

[50] LeTraon, P.-Y., and Ogor, F. ERS-1/2 orbit improvement using TOPEX/POSEIDON: the 2cm challenge. *Journal of Geophysical Research 103* (1998), 8045–8057.

[51] Lewis, J. M. Roots of ensemble forecasting. *Monthly Weather Review 133* (2005), 1865–1885.

[52] Lorentz, E. N. Empirical Orthogonal Functions and statistical weather prediction. Scientific Report 1, Massachussetts Institute of Technology, Dept. of Meteorology, 1956.

[53] Lorenz, E. A study of the predictability of a 28-variable atmospheric model. *Tellus 17* (1965), 321–333.

[54] Lorenz, E. *The Essence of Chaos.* University of Washington Press, 1993.

[55] Lorenz, E. N. Deterministic Nonperiodic Flow. *Journal of Atmopheric Sciences 20* (1963), 130–141.

[56] Madec, G., Delecluse, P., Imbard, M., and Levy, C. *OPA 8.1 Ocean General Circulation Model Reference Manual.* Institut Pierre Simone Laplace des Sciences de l'Environment Global, 1998.

[57] Malanotte-Rizzoli, P., and Bergamasco, A. The wind and thermally driven circulation of the eastern Mediterranean. part i. *Oceanologica Acta 12*, 4 (1989), 335–371.

[58] Manzella, G., Scoccimarro, E., Pinardi, N., and Tonani, M. Improved near real time data management procedures for the Mediterranean ocean Forecasting System-Voluntary Observing Ship program. *Annales Geophysicae 21* (2003), 3–20.

[59] Marshall, J., Jones, H., Karsten, R., and Wardle, R. Can Eddies Set Ocean Stratification? *Journal of Physical Oceanography 32* (2002), 26–38.

[60] Merwe, R., Doucet, A., deFreitas, N., and Wan, E. The unscented particle filter. Technical Report CUED/F-INFENG/TR 380, Cambridge University Engeneering Department, 2000.

[61] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teler, A. H., and Teller, E. Equations of state calculations by fast computing machines. *Journal of Chemical Physics* (1953), 1087–1091.

[62] Metropolis, N., and Tweedie, R. L. The Monte Carlo method. *Journal of the American Statistical Association 44*, 227 (1949), 335–341.

[63] Milliff, R., and Morzel, J. Wind stress curl and wind stress divergence biases from rain effects on QSCAT Surface Wind Retrivals. *Journal of Atmospheric and Ocean Technology 21*, 8 (2004), 1216–1231.

[64] Milliff, R. F. Comparing the Kinetic Energy vs. Wavenumber in Surface Wind fields from ECMWF Analyses and the NASA QuikSCAT Scatterometer. Tech. rep., Instituto Nazionale Geofisica e Vulcanologia, 2004.

[65] Milliff, R. F., Large, W. G., Holland, W. R., and McWilliams, J. C. The general circulation reponses of high-resolution north atlantic ocean models to synthetic scatterometer winds. *Journal of Physical Oceanography 26* (1996), 1747–1768.

[66] Milliff, R. F., Large, W. G., Marzel, J., Danabasoglu, G., and Chin, T. M. Ocean general circulation model sensitivity to forcing from scatterometer winds. *Journal of Geophysical Research 104* (1999), 11337–11358.

[67] Millot, C., and Taupier-Letage, I. Circulation in the Mediterranean Sea. In *Handbook of Environmental Chemistry*. Springer-Verlag, 2005.

[68] Molcard, A., Pinardi, N., Iskandarami, M., and Haidvogel, D. B. Wind driver general circulation of the Mediterranean Sea simulated with a Spectral Element Ocean Model. *Dynamics of Atmosphere and Oceans 17* (2002), 687–700.

[69] Moskalenko, L. V. Steady state wind driven currents in the eastern half of the Mediterranean Sea. *Okianologia 4* (14), 491–494.

[70] NASA. QuikSCAT Science Data Product User's Manual, Overview and Geophysical Data Products, version 2.2. Tech. rep., Jet Propulsion Laboratory, Pasadena, CA, 2001.

[71] Pacanovsky, R. C., and Philander, S. G. H. Parametrization of the vertical mixing in the numerical models of tropicals oceans. *Journal of Physical Oceanography 11* (1981), 1443–1451.

[72] Pedlosky, J. *Geophysical Fluid Dynamics*. Springer-Verlag, 1987.

[73] Pinardi, N., Allen, I., Demirov, E., DeMey, P., Korres, G., Lascaratos, A., LeTraon, P.-Y., Maillard, C., Manzella, G., and Tziavos, C. The Mediterranean ocean Forecasting System: first phase of implementation (1998-2001). *Annales Geophysicae*, 21 (2003), 49–62.

[74] Pinardi, N., Bonazzi, A., Scoccimarro, E., Dobricic, S., Navarra, A., Ghiselli, A., and Veronesi, P. Very large ensemble ocean forecasting experiment using the Grid computing infrastructure. *BAMS in press* (2008).

[75] Pinardi, N., and Masetti, E. Variability of the large scale circulation of the Mediterranean Sea from observation and modelling: a review. *Palaeogeography, Palaeoclimatology, Palaeoecology 158* (2000), 153–173.

[76] Pinardi, N., Zavatarelli, M., Arneri, E., Crise, A., and Ravioli, M. The physical, sedimentary and ecological structure and variability of shelf areas in the Mediterranean Sea. In *The Sea*. Allan R. Robinson adn Kenneth H. Brink, 2005, ch. 32.

[77] Poulain, P. M., Barbanti, R., Font, J., Cruzado, A., Millot, C., Gertman, I., Griffa, A., Molcard, A., Rupolo, V., Bras, S. L., and de-la Villeon, L. P. Medargo: a drifting profiler program in the Mediterranean Sea. *Ocean Sciences 3* (2007), 379–395.

[78] Preisendorfer, R. W. *Principal Components Analysis in Meteorology and Oceanography.* (Ed) Mobley, C.D., Elsevier, Amsterdam, 1988.

[79] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. *Numerical Recipes.* Cambridge University Press, 1986.

[80] Ricci, S., Weaver, A. T., Vialard, J., and Rogel, P. Incorporating state-dependent Temperature and Salinity constraints in the background error covariance of variational ocean data assimilation. Technical Memorandum 441, ECMWF, 2004.

[81] Rio, M. E., Poulain, P. M., Mauri, A., Larnicol, E., and Santoleri, L. A mean dynamic topography of the Mediterranean Sea computed from the altimetric data and in situ measurements. *Journal of Marine Systems* (65), 484–508.

[82] Roussenov, V., Stanev, E., Artale, V., and Pinardi, N. A seasonal model of the Mediterranean Sea general circulation. *Journal of Geophysical Research 100* (1995), 13515–13538.

[83] Royle, J. A., Berliner, L. M., Wikle, C. K., and Milliff, R. M. *A hierarchical spatial model for constructing wind fields from scatterometer data in the Labrador Sea.* Springer-Verlag, 1998, ch. Case studies in Bayesian Statistics, pp. 367–382.

[84] Schneider, T., and Griffies, S. M. A conceptual framework for predicatibility studies. *Journal of Climate 12* (1999), 3133–3155.

[85] Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J. 27* (1948), 370–423.

[86] Smith, S. D. Wind stress and heat flux over the ocean in Gale force winds. *Journal of Physical Oceanography 10* (1980), 709–725.

[87] Sparnocchia, S., Pinardi, N., and Demirov, E. Multivariate Empirical Orthogonal FUNCTION analysis of the upper thermocline structure of the Mediterranean

Sea from observations and model simulations. *Annales Geophysicae 21* (2003), 167–187.

[88] Testor, P., and Gascard, J. C. Large scale flow separation and mesoscale eddy formation in the Algerian basin. *Prog. Oceanogr. 66* (2005), 211–230.

[89] Thompson, P. D. Uncertainty of initial state as a factor in predictability of large-scale atmospheric flow patterns. *Tellus 9* (1957), 225–295.

[90] Tierney, L. Markov chain for exploring posterior distributions. *The Annals of Statistics 22*, 4 (1994), 1701–1762.

[91] Tonani, M., Pinardi, N., Dobricic, S., Pujol, I., and Fratianni, C. A high-resolution free-surface model for the mediterranean sea. *Ocean Sciences 4* (2008), 1–14.

[92] Wikle, C. K., and Berliner, L. M. A Bayesian Tutorial for Data Assimilation. *Physica D 230*, 1-2 (2007), 1–16.

[93] Wikle, C. K., Milliff, R. F., and Large, W. G. Surface wind variability on spatial scales form 1 to 1000 km observed during TOGA COARE. *Journal of Atmopheric Sciences 56* (1999), 2222–2231.

[94] Wikle, C. K., Milliff, R. F., Nychka, D., and Berliner, L. M. Spatio-temporal hierarchical bayesian modelling: Tropical ocean surface winds. *Journal of American Statistical Association*, 97 (2001), 382–397.

[95] Wornell, G. W. Wavelet-based representations for the 1/f family of fractal processes. *Proceeding of the IEEE*, 81 (1993), 1482–1450.

[96] Zavatarelli, M., and Mellor, M. A numerical study of the Mediterranean Circulation. *Journal of Physical Oceanography 46* (1995), 680–688.