Model error space and data assimilation in the Mediterranean Sea and nested models

Luc Vandenbulcke
GeoHydrodynamics and Environment Research laboratory
Departement of Astrophysics, Geophysics and Oceanography
Université de Liège

Thesis submitted for the degree of
Docteur en Sciences Appliquées

April 3, 2007

Thesis Committee

Prof. Jean-Marie Beckers (Université de Liège)
Prof. Eric Delhez (Université de Liège)
Prof. André Lejeune (Université de Liège)
Prof. Louis Wehenkel (Université de Liège)
Prof. Nadia Pinardi (Università di Bologna)
Dr. Alberto Alvarez (IMEDEA)
Dr. Michel Rixen (Saclant NATO Undersea Research Center)
Dr. Alexander Barth (University of South Florida)

Acknowledgments

At the end of my PhD, I would like to express my sincere gratitude to everyone who has collaborated to this accomplishment, whether it be by their insight in the matters treated, by their continuous support, or by both.

I wish to sincerely thank Professor J.-M. Beckers for the perfect blend of guidance and support, and huge freedom regarding the research topics. From my first steps discovering the ocean science, to the last days, his door has always been open for all kinds of questions concerning oceans as well as numerics or informatics, and advice has always been following. Moreover, he made sure that the period of my PhD study was constantly enriched by a multitude of travels and contacts with other people, such as the DEA stage in France, the participation to various national and international conferences, numerous meetings with the scientists of the European MFSTEP project, a very interesting stage in Mallorca (which has led to the last chapter of this thesis), or an exchange program with the Hanoi University of Vietnam.

I would also like to address my gratitude to Michel Rixen, whose office I shared at the beginning of this work, and who flooded me with very interesting literature, ensuring a swift start to my work. He also helped a great deal with the revision of the present text.

My immediate predecessors set very high standards for the following PhD students. I sincerely thank Alexander Barth and Aida Alvera Azcarate for the uncountable amount of time spent helping me with various problems. They were present from the conception of the model grids to the latest review of these pages. And without the functions and programs Alexander added to the GHER model code, I probably would still be working on chapter 2 of this thesis. It was also a pleasure to work with Zied Ben Bouallegue, also in the frame of the MFSTEP project. I am indebted to all three of them.

This leads me to thank all members of the MFSTEP project who provided many ideas to improve my work. In particular, I wish to mention Pierre Demey and his Toulousian team, and Nadia Pinardi, for their immense experience. Their comments were very encouraging time after time.

Alberto Alvarez very kindly welcomed me at the IMEDEA center in Esporles, Mallorca. He spent a lot of time sharing his knowledge of statistical predictors, among others. Therefore I am also very indebted to him.

More recently, the next PhD students at the GHER laboratory already brought very good ideas into my work. I sincerely thank Fabian Lenartz and Charles Troupin for this. I am also grateful to my colleagues Damien Sirjacobs and Nikos Skliris.

I was kindly offered free access to the Pepito data-mining software by the collaborators at Pepite, a spin-off off the University of Liège. The software offers very powerful features and state-of-the-art procedures, which are all only a click away. Therefore, I am very grateful to them.

I wish to acknowledge the members of the SEGI IT department of the University of Liège for their efforts in keeping the Silicon Graphics supercomputer up-and-running despite all difficulties, and for the difficult task of setting up a new cluster of PCs. Only thanks to the latter, the heavy computations presented below were possible.

I also would like to thank all the members of my thesis committee, who kindly accepted to participate to this PhD.

My family, parents and sister not only provided unceasing encouragements during the last 4 years, but from the very beginning of my studies. I owe them more than I could ever write down here.

Finally, I thank Gaëlle. She started supporting the work of her boyfriend, and ended up supporting her husband. All this time, she patiently listened to day-to-day accounts of marginal progress. For these years of happiness, I want to thank her very much.

Abstract

In this work, we implemented the GHER hydrodynamic model in the Gulf of Lions (resolution 1/100°). This model is nested interactively in another model covering the North-Western basin of the Mediterranean Sea (resolution 1/20°), itself nested in a model covering the whole basin (1/4°). A data assimilation filter, called the SEEK filter, is used to test in which of those grids observations taken in the Gulf of Lions are best assimilated. Therefore, twin experiments are used: a reference run is considered as the truth, and another run, starting from different initial conditions, assimilates pseudo-observations coming from the reference run. It appeared that, in order to best constrain the coastal model, available data should be assimilated in that model. The most efficient setup, however, is to group all the state vectors from the 3 grids into a single vector, so that the 3 domains are coherently modified at once during assimilation cycles.

Operational forecasting with nested models often only uses so-called passive nesting: no data feedback occurs from the regional models to the global model. A new idea is proposed: to use data assimilation as a substitute for the feedback. Using again twin experiments, it is shown that assimilating outputs from the regional model in the global model, has beneficial impacts for the subsequent forecasts in the regional model.

The data assimilation method used in those experiments corrects errors in the models using only some privileged directions in the state space. Furthermore, these directions are selected from a previous model run. This is a weakness of the method when real observations are available. We tried to build new directions of the state space using an ensemble run, this time covering only the Mediterranean basin (without grid nesting). This led to a quantitative characterization of the forecast errors we might expect when various parameters and external forcings are affected by uncertainties. We looked in detail at both the spatial and temporal evolution of these forecast errors.

Finally, using these new directions, we tried to build a statistical model supposed to simulate the hydrodynamical model using only a fraction of the computer resources needed by the latter. To achieve this goal, we tried out linear regressions, artificial neural networks, nearest-neighbors and regression trees. The inputs of the algorithms were the weights of the EOF decomposition of the ocean state vector and of some atmospheric forcing fields; the weights after some time constituted the outputs. This study constitutes only the first step towards an innovative statistical model, as in its present form, only a few degrees of freedom are considered and the primitive equation model is still required to build the automatic learning method. We tried forecasting at 2 different time horizons: one day and one week. We tried out both cases where only past states were available in the database, or when some future states were available. In the second case, we showed that automatic learning algorithms performed quite well, at both timescales. In the first case however, satisfying predictions of the future day ocean state could be obtained only with artificial neural networks (and to a lesser extent regression trees). No usable predictions of the next-week ocean state could be obtained at all, indicating that a stable relationship between inputs and outputs simply does not exist within our database.

Résumé

Le modèle hydrodynamique du GHER est implémenté dans le Golfe du Lion (résolution 1/100°). Ce modèle est emboîté interactivement dans un autre modèle couvrant la Méditerrannée du Nord-Ouest (résolution 1/20°), lui-même emboîté dans un modèle couvrant toute la mer (résolution 1/4°). Un algorithme d'assimilation de données, appellé filtre SEEK, est utilisé pour déterminer dans laquelle de ces 3 grilles il vaut mieux assimiler les données qui seraient observées dans le Golfe du Lion. Pour ce faire, nous avons utilisé des expériences jumelles: une simulation de référence est considérée comme l'état réel de la mer, et une autre simulation, partant de condition initiales différentes, assimile des pseudo-observations venant de la simulation de référence. Il en ressort que l'impact des observations sur le modèle côtier est le plus appréciable lorsqu'elles sont assimilées directement dans ce dernier. Cependant, l'implémentation la plus efficace est obtenue lorsque les vecteurs d'état venant des 3 grilles sont rassemblés en un vecteur unique. Lors de l'assimilation de données, les 3 domaines sont alors mis à jour de façon cohérente automatiquement.

Les centres de prévisions opérationnelles utilisent de façon routinière les modèles gigognes tels que celui décrit ci-dessus. Toutefois, seul l'emboîtement passif est généralement utilisé, sans retour de la grille emboîtée vers la grille-parent. Nous proposons dès lors d'utiliser l'assimilation de données pour remplacer le retour d'information. Les prévisions du modèle emboîté sont assimilées dans le modèle global. En utilisant à nouveau des expériences jumelles, nous montrons l'impact positif de cette méthode sur les prévisions suivantes du modèle régional.

Le schéma d'assimilation de données utilisé dans les expériences mentionnées ne corrige les erreurs que si elles suivent certaines directions privilégiées de l'espace d'état. De plus, ces directions sont déterminées par une simulation antérieure. Ceci est un désavantage certain de la méthode lorsque des observations réelles sont disponibles. Nous avons construit un nouveau sous-espace d'erreur au moyen d'une simulation d'ensemble (couvrant uniquement la Mer Méditerrannée sans grilles emboîtées). Ainsi, nous avons pu caractériser quantitativement l'erreur attendue du modèle lorsque différents forçages présentent des incertitudes données. Nous avons examiné l'évolution de cette erreur tant dans l'espace que dans le temps.

Enfin, en utilisant ces nouvelles directions dans l'espace d'erreur, nous avons construit un modèle statistique pour simuler le modèle d'équations primitives, en utilisant seulement une fraction des ressources nécessaires pour ce dernier. Pour ce faire, nous avons utilisé des régressions linéaires, des reseaux de neurones, la méthode des K-plus-proches-voisins et des arbres de régression. Les entrées des algorithmes sont données par les poids d'une décomposition EOF du vecteur d'état de l'océan et des forçages atmosphériques; les sorties sont les poids associés à l'état futur de l'océan. Cependant, cette étude constitue seulement le premier pas vers une méthode statistique réellement innovante. En effet, dans son état actuel, seuls quelques degrés de liberté sont considérés, et les sorties du modèle hydrodynamique sont requises pour la création du modèle statistique. Nous avons essayé de prédire l'état de l'océan après 1 jour et après 1 semaine. Nous avons considéré 2 cas. Dans le premier, seuls des états

passés sont disponibles dans la base de données, tandis que dans le deuxième, quelques états futurs sont également présents. Dans ce dernier cas, nous avons montré que les méthodes d'apprentissage statistique arrivent à prédire relativement correctement l'évolution des états manquants, tant après un jour qu'après une semaine. Par contre, dans le premier cas, des prévisions satisfaisantes de l'état futur après un jour n'ont pu être obtenues qu'avec les réseaux de neurones (et dans une moindre mesure, les arbres de régression). De plus, la prédiction après une semaine n'a été possible avec aucune méthode, ce qui indique qu'une relation (à l'échelle de la semaine), qui soit suffisament stable pour être utilisée dans le futur, n'existe pas entre les entrées et sorties dans notre base de données.

Contents

Chapter 1

Introduction

I never see what has been done;
I only see what remains to be done
Marie Curie

Historically, oceans have first been described following the experimental approach. Later, theoretical progress in physics and mathematics led to the development of the first models supposed to reproduce the oceanic circulation. Following the huge developments of computers realized in the XX^{th} century, numerical models started to be used in prognostic mode, leading to relatively reliable descriptions and predictions of the ocean state. However, hydrodynamic models are generally not able to represent observations completely accurately, for various reasons which have a cumulative effect (see e.g. Lermusiaux et al., 2006).

Structural indetermination is caused by the fact that all physical processes, present in the ocean, are not represented by the equations, or are parameterized in order to simplify the problem. Thus, models will not exactly reproduce observations, which naturally result from all the present phenomena. Among common simplifications, let's cite the very widely used Boussinesq approximation (excluding acoustic waves), the hydrostatic approximation (excluding e.g. dense water formation), or the very limiting geostrophic equilibrium approximation (excluding vertical motion). Another example concerns the parameterization of geophysical turbulence. Models use a chosen spatial resolution (determined by taking limited computer resources into account), and all processes with smaller characteristic lengths cannot be represented, leading to the problem of sub-gridscale parameterization and turbulent closure schemes.

Statistical indetermination is due to the fact that various terms are not known exactly. The initial condition is usually the output of a previous model run, an interpolation of climatological data or the result of an inverse model applied to measurements. The ocean is not an isolated system, but is surrounded by other complex systems known only with their own limited accuracy: other oceans, the atmosphere, rivers etc. Therefore, boundary conditions of the models are not known perfectly either.

The desire to reduce model errors by increasing resolution has naturally led to the use of high-resolution regional models, which are nested in larger models using a coarser resolution. Their usage started as early as in the 1970s for atmospheric models. In oceanography, the last 2 decades saw a renewal of interest in coastal circulation due to the increasing population along coasts, the potentially accompanying rise in pollution, and the parallel growth in environmental concerns. Coastal areas and their terrestrial inputs can favor phytoplankton activity, and hence affect global budgets of biogeochemical elements, such as some greenhouse gases. The relative contribution of coastal seas is much larger than their relative area in the global ocean.

An important theoretical and practical problem associated with nested grids is to impose adequate "open-sea" boundary conditions to the regional model. However the choice and amount of boundary conditions depend on the type of equations, and the general hydrodynamic primitive equations are not easy to classify. In a review of the lateral boundary conditions used in hydrodynamic models, McDonald [1997] pointed out that there is a feeling that we can "overspecify slightly the lateral boundary conditions and not do very much damage". In general, a satisfactory boundary condition scheme is one that (a) transmits

incoming waves through the boundary of the high-resolution model without appreciable change of phase or amplitude, and (b) does not let reflected waves reenter the high-resolution model at its outflow boundaries with appreciable amplitude [Davies, 1983]. In practice, boundary conditions are usually chosen pragmatically and tested numerically to check their appropriateness. A review of the open sea boundary conditions problem is given in Blayo and Debreu [2006]. The particular implementation used in this work is described in section 2.2.

Grid nesting can be applied successively several times, allowing to reach very high resolutions in the smallest domain without an abrupt change in resolution at the open sea boundary.

The nesting procedure involves data exchanges between successive grids. In the so-called one-way nesting, the coarse grid model is interpolated on the fine grid to provide boundary conditions. The coarse grid does not use any information from the fine grid, thus, it can be run standalone first, the fine grid model being run afterward. This is useful for operational systems where the nested systems use different models, or are not run at the same place. One-way nesting was first applied by Spall and Robinson [1989], and is now widespread. It can be found in e.g. Pinardi et al. [2003], Korres and Lascaratos [2003], Echevin et al. [2003b] or Zavatarelli and Pinardi [2003] for the Mediterranean Forecasting System Pilot Project (MFSPP). However, one-way nesting leads to the following disadvantage. If the simulation is run for a long period, discrepancies can appear between solutions of the grids, making the application of boundary conditions delicate, and possibly leading to instabilities in the fine grid model. If the fine grid model is reinitialized regularly (e.g. the Eastern Mediterranean basin subgrid is initialized every week in the Mediterranean Forecasting System (MFS) (N. Skliris, private communication)), its small-scale features are lost. So-called variational initialization methods try to overcome this difficulty and are now used operationally, see e.g. Auclair et al. [2000, 2001]. Another possibility is to reinitialize the fine model a few days before the actual forecast, so as to let the small-scale features develop, e.g. the North-Western basin subgrid in the MFS system performs an 8-day hindcast each week (C. Estournel and M. Lux, private communication).
If both the coarse and fine models are run together, the fine model results can be averaged over each coarse grid cell in the overlapping area, where they replace the coarse model outputs. This yields the two-way or interactive nesting. The technique was tested in idealised experiments (e.g. Spall and Holland [1991], Fox and Maskell [1995], Ginis et al. [1998], Blayo and Debreu [1999]) for different resolution ratios between the fine grid and the coarse grid model. Realistic ocean systems were also studied, e.g. Oey and Chen [1992], Fox and Maskell [1996], Ginis et al. [1998]. Although the dynamic is different in the nested grids, inconsistencies are very less likely to appear between them. It has been shown that two-way nesting yields a more realistic representation of the mesoscale features inside the fine grid domain, and also has positive effects outside the domain [Barth et al., 2005]. Let's note that incoherences could still appear due to the fact that lateral open-sea boundary conditions for the high resolution model are obtained by interpolation of the coarse resolution model outputs, as shown in Auclair et al. [2001].

Grid nesting is not the only way to achieve high resolution in specific areas. Unstructured grids allow to enhance the resolution locally, without the need for different, embedded models (see e.g. Pietrzak et al., 2005, Hanert et al., 2004); they are often used with finite elements methods, but are not common in oceanography yet. Other grids, such as the orthogonal grid used in the ROMS model (see http://www.myroms.org) also allow to obtain a high resolution in specific areas.

In parallel with the development of models, the experimental side did also evolve, with the development of various instruments and sensors, both *in situ* and remote. In the former category, let's cite CTD (Conductivity-Temperature-Depth) sensors and XBT (eXpendable BathyThermographs) instruments, which collect a profile of temperature and salinity measurements at one point at a time and require considerable logistics. Their measurements have been the oldest observation system of the oceans internal structure, and make up the largest historical data set. Let's also cite another, more recent (its deployment begun in 2000) *in situ* instrument: the Argo float. Those floats are drifting independently (requiring less logistic efforts than XBTs) at a given "parking depth", but regularly descend to about 2000 meters and then ascend to the sea surface to transmit measurements (see Fig. 1.1). About 3000 Argo floats are now continuously taking temperature, salinity, pressure and velocity measures over the whole global ocean, which averages to a grid with one Argo float every 3°. The results are available in near-real time. In fact, the Argo results have tend to become reference measurements in many areas, as they provide a number of advantages compared with XBTs (more floats, deeper profiles, year-round operation, not limited to ship routes). If *in situ* observations give a good representation of the vertical structure of the water columns, their horizontal resolution remains sparse. Moreover, they cannot provide (quasi-)synoptic images of the ocean.

In the latter category, different instruments are used. Advanced Very High Resolution Radiometer (AVHRR) instruments onboard satellites lead to sea surface temperature (SST) measurements. Infrared and visible imagery provide so-called ocean color data, microwave and radar imagery leads to surface wind and wave fields and altimetric data. The advantage of satellite observations is certainly that they provide synoptic and global images. However, only the surface layer is observed. Fig. 1.2 gives an overview of most satellites, active or yet to be launched, observing the ocean. Other specific satellites will also be launched, such as the SMOS mission aiming (with a relatively low resolution) at measurements of the sea surface salinity and also soil moisture. Contrary to what one may think when looking at Fig. 1.2, even though many Earth observation missions exist or are in preparation, the future or successor of some satellites ending their life cycle is not assured; fears exist in the scientific community that observations may become sparser in the future than they are now.

Both *in situ* and remote data cover different aspects of the ocean, and are developed to be complementary. For example, combining satellite altimetric measurements with T-S profiles allows to (try to) characterize both the barotropic

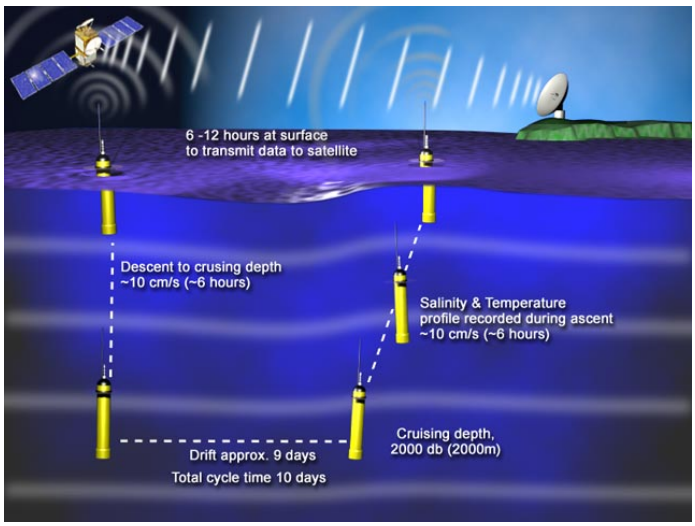

Figure 1.1: Argo float operation cycle: the float descends to cruising depth, drifts for several days, ascends while taking salinity and temperature profiles, and then transmits data to satellites. From http://www.argo.ucsd.edu/FrHow_Argo_floats.html.

and the baroclinic components[1]. Finally, it is also worth to mention that in the last years, huge efforts have been made to quality-control, organize and store the tremendous amount of data coming from all sources mentioned above.

The coexistence of renewed experimental and theoretical approaches has led to new tools to combine informations acquired experimentally and from models. Indeed, on the one hand, experimental measurements never cover the whole tridimensional domain covered by the oceans, and can not lead to predictions, while models suffer from uncertainties as mentioned before. Assimilation of available observations, *in situ* and remote sensed, should thus be considered as a tool to optimally control the model evolution and to reduce the uncertainty affecting it. Data assimilation is a generic name, meaning in the ocean modeling community "to use all available information to determine as accurately as possible the state of the atmospheric or oceanic flow" [Talagrand, 1997]. Uncertainty is most easily described by probability distributions, and this leads to consider data assimilation as a problem in bayesian estimation, *i.e.* to look for the conditional probability distribution of the ocean state, given the available information (model and observations) and the probability distribution of the associated uncertainty. In that respect, assimilation of observations can be seen as one of the many inverse problems that are encountered in many fields of science and technology. In spite of the variety of applications, most (if not all) inverse problems are solved using the same basic mathematical methods. Difficulties that are specific to meteorological and oceanographical applications

[1] As an anectode, we could mention that precisely this fact led to the choice of the name Argos for the measuring devices described above, its T-S profiles being complementary to the altimetric data provided by the Jason satellite. In the greek mythology, Jason sailed a ship called the "Argo" to capture a golden fleece.

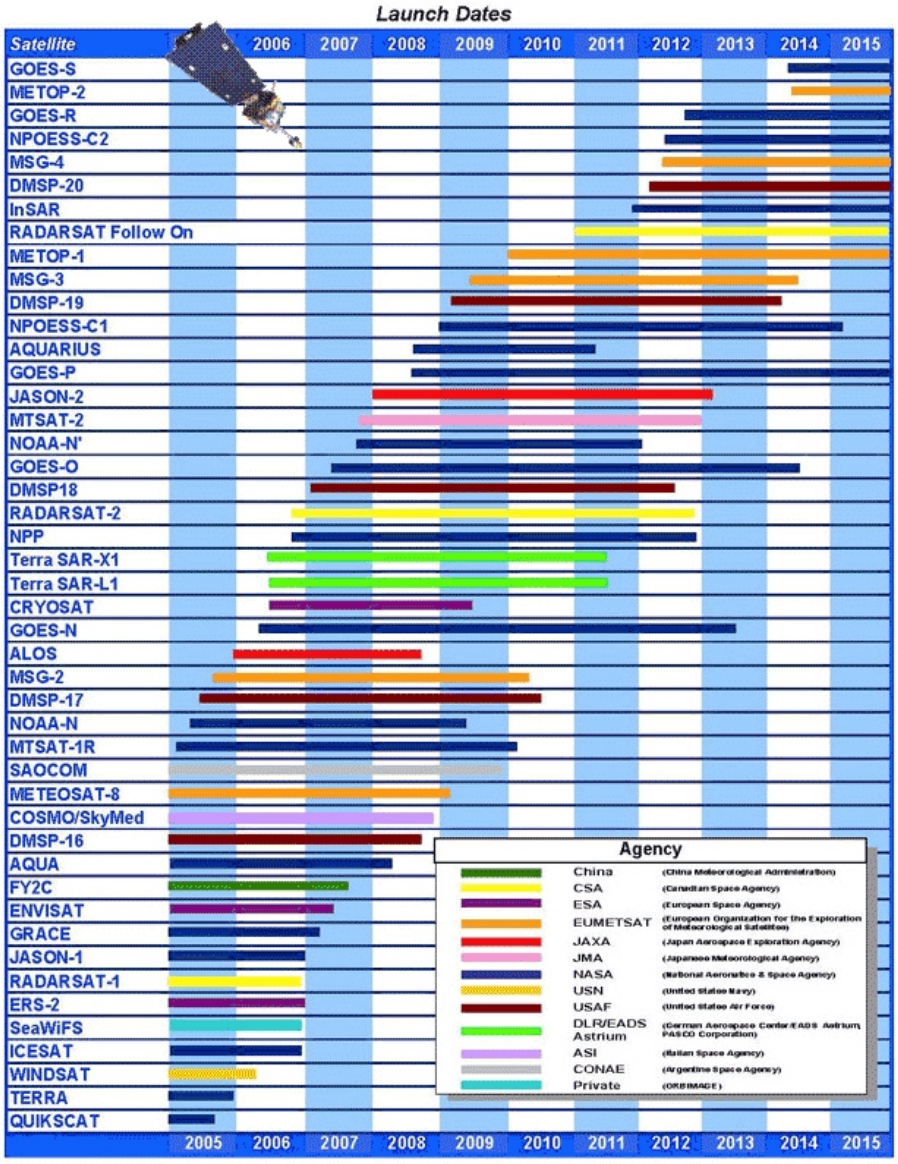

Figure 1.2: Existing and planned ocean observation satellite missions. From http://www.orbit.nesdis.noaa.gov.

are the large numerical dimensions of the problems to be solved (the number of elements required to represent a typical domain is in the range $10^6 - 10^7$), and the nonlinear and chaotic nature of the underlying dynamics. Thus, the general form of bayesian estimation cannot be implemented in practice, and one must restrict his ambitions to much more modest goals. Talagrand identifies two basic approaches:

Statistical linear estimation, which looks for the estimated state as a linear combination of the model state vector and the observations vector, minimizing the statistical variance of the corresponding estimation error. Thus, this method is strictly valid only for linear problems. However, if the estimated state vector is close enough to the real state vector, the tangent linear approximation of the model might be close enough to the real model, so that the estimation problem can be linearized in terms of deviations from that prior estimate. Its solution is called the *Best Linear Unbiased Estimate (BLUE)* of the state of the flow. It requires *a priori* knowledge of the model and observations vectors, and the corresponding covariance matrices. If the data errors are Gaussian, the *BLUE* achieves bayesian estimation in the sense that it entirely defines the conditional probability distribution of the state of the flow, given the data. More details can be found in Talagrand [1997]. Numerous variants of the *BLUE* have been proposed, differing by the specification of the expectation and covariance matrix of the data, and by the algorithm implementation.

When the estimate of the ocean state is updated every time new observations are available, the method is called "sequential". Its exact *BLUE* form is the so-called *Kalman filtering (KF)* [Kalman, 1960] which, because of the need for explicitly computing the temporal evolution of the covariance matrix of the estimation error, goes well beyond available numerical resources. Variants of the exact Kalman filter, such as for example the *Extended Kalman Filter (EKF)*, the square-root filter *(SQRT)* and its reduced rank approximation called *Reduced-Rank Square Root Filter (RRSQRT)* [Verlaan and Heemink, 1997], the *Singular Evolutive Extended Kalman Filter (SEEK)* [Pham et al., 1998b], and the *Error Subspace Statistical Estimation (ESSE)* [Lermusiaux, 1997] will be described in section 2.3.

When the model estimate is adjusted periodically to all observations (past and future) distributed over a period of time, the method is called "variational". It becomes numerically feasible through the use of the *adjoint* of the assimilation model. One advantage of variational assimilation is that it provides a very efficient way to carry information both forward and backward in time. In addition, and contrary to Kalman filter, whose optimality requires that the errors affecting the data must be uncorrelated in time, it can relatively easily cope with time-correlated errors. On the other hand, and contrary to Kalman filter, it does not provide an explicit estimate of the estimation error. This class of methods includes e.g. the 3D-Var and representer methods. The *Kalman smoother* [Gelb, 1974] is an extension of the KF providing the best estimate using both past and future observations.

Ensemble assimilation produces an ensemble of possible ocean states, whose distribution is meant to represent the desired conditional probability distribution. It is of particular interest in situations where a tangent linear approximation is not valid, making it a natural extension of the Kalman Filter to nonlinear dynamics. It was first introduced in geophysical fluid models by Evensen [1994], and called the *Ensemble Kalman Filter (EnKF)*. It is usually sequential in na-

ture. Every member of the ensemble of possible states is evolved in time with the model, and constantly updated as new observations become available. A drawback of the method, as for any Monte Carlo method, is that convergence is slow and a high number of members is generally needed (see Evensen, 2003). Another group of methods is called *particle filters*. An ensemble of members is created, and the most probable ocean state is the average of the members. When observations become available, the members are not modified, but their relative weights are updated.
Following the developments in both statistical and ensemble assimilation methods, equivalences were shown between both groups, and improvements in one lead to improvements and better understanding in the other. For example, a variant of the SEEK filter mentioned above, inspired by the EnKF, is called the *Singular Evolutive Interpolated Kalman filter (SEIK)* [Pham, 2001]; a variant of the RRSQRT filter is called the *Partially Orthogonal Ensemble Kalman Filter(POEnKF)* [Heemink et al., 2001]. Non-sequential versions of the EnKF were also developed, such as the *Ensemble Kalman Smoother* [Evensen and van Leeuwen, 2000].

All assimilation algorithms require an *a priori* estimate of, at least, part of the probability distribution of the errors affecting the (model and observational) data. In sequential methods, that information is used to build the error matrices; in ensemble assimilation methods, it is used to build the ensemble itself through varying initial conditions, model parameters etc. This information cannot be obtained from the data themselves, even through appropriate statistical processing, but entirely depends on independent hypotheses that cannot be objectively validated from the data themselves (Talagrand). Furthermore, as we will see in this work, the results of data assimilation is highly sensitive to the quality of the model error space.

In the last years, progress in assimilation schemes led to their operational use. As a result, modelers now have to choose where to apply the increasing computer power: should the model resolution be increased, or should a more complex assimilation scheme be implemented? For example, is it more useful to increase the model horizontal resolution so that the required computer power is multiplied by a factor 10, or should the model error space comprise 10 times more modes? In that aspect, an example in the MFS system mentioned above is informative. It happens that the root mean square error between model and observations is larger in the regional model covering the North-Western Mediterranean Sea, than in the corresponding part of the OGCM itself, although the resolution of the latter is lower... Indeed, to correctly represent eddies and local features is a very complex task. Therefore, if appropriate observations are available too, more progress could be achieved by dedicating computer resources to the estimation of the errors made in the models, rather than increasing their resolution even more. A nice by-product would then be (more) accurate estimations of the forecast error.

Whenever local observations are available in a system of nested grids, the question arises to know in which grid they should be assimilated to optimally control the model. Indeed, they could be assimilated in the coarse grid model, in the fine grid model, or in both. This question is analysed in detail in chapter 3

using twin experiments. A simulation with two-way nesting is used as a reference. Different setups using one-way nesting are then compared to the reference simulation. A novel approach introduced in Barth et al. [2006], where all the variables from the different grids are assembled in a single state vector, is also tested.

When two-way or interactive nesting is not possible, the high-resolution information generated by the nested model is "lost" for the parent models. This is unfortunate, as it has been shown multiple times that two-way nesting leads to more accurate simulations of the flow, both inside the nested domain and in its vicinity (through advection of the features). However, two-way nesting requires to run both models together, because neither one can continue without the outputs from the other one. This is often difficult in operational setups, particularly when different regional models are nested into the Ocean General Circulation Model (OGCM), or when coastal models are in turn nested in the regional models. It is also impossible when the regional models are run at different physical locations. Thus, we sought an alternative to classic feedback in order to still take some benefits from the regional models in order to improve the OGCM simulation, but that would not require to run both parent and nested model simultaneously. In our method, forecasts from the regional model are assimilated as pseudo-observations in the global model. This constitutes, to our knowledge, a novel approach; it is explained in chapter 4.

All data assimilation experiments in nested grids described in chapters 3 and 4 led us to realize (together with many authors) that it is crucial in assimilation filters to correctly represent, or approximate, the model error space. In particular, some widely used approximations are often not valid. Building the error space from model states taken at different time steps, as suggested by Cane et al. [1996], Pham et al. [1998b], supposes that the model variability in time is a good approximation of its error. This is not always true, and may lead to difficulties in subsequent assimilation cycles, particularly if one uses a filter not updating the error space in time. In chapter 5, we try as carefully as possible to build an error space for a high-resolution implementation of the GHER model covering the Mediterranean Sea, based on the ensemble approach. To our knowledge, no previous systematic computation and study of a Mediterranean model error space has been realized, particularly at the same resolution as modern operational systems (1/16°). Besides being useful to correctly assimilate (real or synthetic) observations in the model, this error space is interesting in itself, as we can now quantify the spatial and temporal variations of expected errors on model predictions. Additionally, it can be used to assess the impact of the uncertainty affecting physical variables on coupled models such as biological models, oil spill models, drift models etc.

The hydrodynamic model considered in chapter 5 is a non-linear function, whose inputs are 5 prognostic variables defined at regular intervals on a grid, the Atlantic Ocean state (for the boundary condition) and 5 atmospheric variables defined also on a regular 2-D grid. All these data are used by a non-linear function to compute the outputs, i.e. the 5 variables over the whole grid after a small time increment. This way of proceeding is the most rigorous one, as it is deduced from basic physical laws and numerical considerations. However, there

is a large spatial and temporal redundancy in the data, and simplified methods might be used to simulate the model. For instance, principal component analysis (PCA), also called Empirical Orthogonal Functions analysis (EOF) in oceanography, might compress the data (see Preisendorfer, 1988, Kantha and Clayson, 2000). The technique can be applied to the model variables as well as to the forcing fields. The size of the state vector representing the sea could be reduced $\mathcal{O}(10)$, or $\mathcal{O}(10^2)$ at most, rather than $\mathcal{O}(10^6 - 10^7)$. The equations of the model must of course be replaced by other equations, and the whole difficulty is to find the most appropriate ones.
In chapter 6, we will use the EOFs computed from the ensemble run in chapter 5 to try and find a link between the ocean state at different instants, based on past model states. We will apply the theory of supervised learning. In particular, we will examine the popular neural networks, but also decision trees and nearest neighbor methods. Here the objective is clearly defined in terms of modeling the underlying function between (some of) the inputs and outputs, as opposed to unsupervised learning algorithms which are not oriented toward a particular prediction task. Rather, they try to find out by themselves the existing relations among states characterized by a set of attributes.
Among the inputs of the statistical function that we try to define, we will also include the coefficients of some atmospheric fields EOFs, as the future ocean state is also determined by the lateral forcings. It should be said that, at best, our method will be able to simulate the hydrodynamic model with some added error due to the huge simplification. Its interest lies in the fact that its computation time will be extremely low compared with the full hydrodynamic model. It is also straightforward to generalize the method to use more inputs, e.g. variables coming from different hydrodynamic models. Here, no general physical development can lead to a function between inputs and outputs, hence inductive learning is not used to simulate a physical model, but rather to infer a possible relation.
It is well-know that statistical methods are only valid in the region of the state space where they have been trained to work. Thus, when radically new ocean states or atmospheric forcings appear, the method will likely fail to provide good results. In fact, the statistical function is simply extrapolating outside of its domain. Unfortunately, this is not the only cause of bad results. A recurring problem with inductive learning algorithms is called "over-fitting", and happens when the obtained functions is modeling the noisy outputs as well as the underlying trend. Results of this function applied on other inputs will then be poor. A difficult aspect is thus to find when the training of the function should be stopped.

Chapter 2

The GHER Model

By three methods we may learn wisdom:
First, by reflection, which is noblest;
Second, by imitation, which is easiest;
and third by experience, which is the bitterest.
Confucius

2.1 Hydrodynamic model

To realize our study, we used the GHER hydrodynamic model developed at the University of Liège. Nihoul et al. [1989] and Beckers [1991] provide a detailed description of the model, including its mathematical and numerical formulations. It is a hydrostatic free-surface primitive equation model solving the prognostic variables of temperature, salinity, sea surface elevation, horizontal components of the velocity and turbulent kinetic energy:

$$\nabla \cdot \mathbf{v} = 0 \tag{2.1}$$

$$\frac{\partial \mathbf{u}}{\partial t} + \mathbf{v} \cdot \nabla \mathbf{u} + f\mathbf{e_z} \wedge \mathbf{u} = -\nabla_h q + \frac{\partial}{\partial z}\left(\tilde{\nu}\frac{\partial \mathbf{u}}{\partial z}\right) \tag{2.2}$$

$$\frac{\partial T}{\partial t} + \mathbf{v} \cdot \nabla T = \frac{\partial}{\partial z}\left(\tilde{\lambda}\frac{\partial T}{\partial z}\right) + \frac{1}{c_p \rho_0}\frac{\partial I}{\partial z} \tag{2.3}$$

$$\frac{\partial S}{\partial t} + \mathbf{v} \cdot \nabla S = \frac{\partial}{\partial z}\left(\tilde{\lambda}\frac{\partial S}{\partial z}\right) \tag{2.4}$$

$$\frac{\partial k}{\partial t} + \mathbf{v} \cdot \nabla k = \tilde{\nu}\left\|\frac{\partial \mathbf{u}}{\partial z}\right\|^2 - \tilde{\lambda}\frac{\partial b}{\partial z} - \epsilon + \frac{\partial}{\partial z}\left(\tilde{\nu}\frac{\partial k}{\partial z}\right) \tag{2.5}$$

Here, f represents the Coriolis frequency, $\mathbf{v}$ is the velocity vector, whose horizontal part is noted $\mathbf{u}$, T is the temperature, S the salinity, k the turbulent kinetic energy and ϵ its dissipation. c_p is the heat capacity of the water, ρ_0 the reference density, I is the insolation (see below) and $\tilde{\nu}$ and $\tilde{\lambda}$ are the eddy viscosity and diffusivity. q is the generalized pressure defined by

$$q = \frac{p}{\rho_0} + gz \tag{2.6}$$

with g the gravitational acceleration, b is the buoyancy given by the state equation

$$b = -\frac{\rho(T,S) - \rho_0}{\rho_0} g \tag{2.7}$$

and linked to the generalized pressure by

$$\frac{\partial q}{\partial z} = b \tag{2.8}$$

The model uses the β-plane and Boussinesq approximations. The vertical turbulence uses a k turbulent kinetic energy closure scheme described in Nihoul et al. [1989]. The evolution equation for ϵ is considered to not be well known, and thus replaced by an algebraic equation. It requires to specify the mixing length l_n, which is assumed to depend only on the geometry of the basin [Beckers, 1991].

Horizontally, the GHER model uses Cartesian coordinates and is discretized using a conventional Arakawa C-grid, where velocities are defined at the grid cell boundaries, and the scalar variables are defined at the center of each grid cell. This immediately yields the advection and the pressure gradient terms, while requiring an averaging operation for the Coriolis term, and is thus well adapted to models with a relative high resolution. In the vertical, the model uses a double sigma coordinate, splitting the domain into two superposed regions, the limit between them being placed at the average shelf break depth

(170 m in our simulations). In each of those regions, a normal σ transform is applied. The double σ transform allows to represent abrupt bathymetry breaks if they are located at the prescribed depth.
The model has a semi-implicit integration scheme which is conservative for tracers. Furthermore, it uses mode splitting, a now common practice: for computational efficiency, the barotropic time step, used to solve the vertically integrated equations, is much smaller than the baroclinic one. This approach is justified by the fact that the external mode describes essentially fast processes, while the baroclinic modes are more representative of slow processes.

Atmospheric forcings interact with the GHER model *via* bulk formula following Kondo [1975], for the momentum flux and the heat flux (sum of the surface shortwave radiation, net longwave radiation, latent and sensible heat fluxes). The implementation is described in Barth [2004]. The precipitation and evaporation terms are neglected in the short model runs presented in the following chapters.

Momentum The winds at the sea surface drag the surface water along its directions. The wind stress τ is parameterized by:

$$\tau = C_D \rho_a \left\| \mathbf{u}_a \right\| \mathbf{u}_a \tag{2.9}$$

with ρ_a the air density, and $\mathbf{u}_a$ the wind velocity at the reference level. The drag coefficient C_D is parameterized by the scheme of Kondo [1975].

Surface shortwave radiation The insolation represents the main energy input into the ocean. It is modeled as a volume heat source in the water column; the radiation flux as a function of depth is given by

$$I(z) = |Q_s| \left(A \exp(g_1 z) + (1 - A) exp(g_2 z) \right) \tag{2.10}$$

with $z = 0$ at the surface and negative in the water column, A=0.58 is the fraction long-wave solar energy and g_1 (0.35 m^{-1}) and g_2 (23.0 m^{-1}) are absorption coefficients in the short-wave and long-wave solar energy respectively. Q_s is the solar energy at the sea surface, and is composed of direct and diffused radiation, which are calculated as a function of the zenith angle of the sun and the fractional cloud coverage.

Net longwave radiation This contribution to the heat flux is the sum of upward longwave radiation of the ocean and downward longwave radiation of the atmosphere. The former is generally larger than the latter, leading to a net loss for the ocean. The net longwave radiation is calculated following the scheme of Clark et al. [1974]:

$$Q_b = \epsilon \sigma T_s^4 (1 - 0.8 C^2)(0.39 - 0.05\sqrt{e_a}) + 4\epsilon\sigma T_s^3 (T_s - T_a) \tag{2.11}$$

where ϵ is the emissivity of the ocean, σ the Stefan-Boltzman constant, e_a the atmospheric vapor pressure in hPa and T_a the air temperature at a reference level.

Latent and sensible heat fluxes The former flux is due to the difference in water vapour content beween the air at the ocean surface and at the reference level, inducing evaporation or condensation. To this mass transfer corresponds a heat exchange equal to the rate of evaporation times the latent heat of evaporation. The sensible heat flux is due to the temperature difference between the air at the ocean surface and the air at the reference level. The heat exchanged by conduction is proportional to the temperature gradient. Both latent and sensible heat flux are parameterized by classical bulk turbulent transfer formula, following Rosati and Miyakoda [1988] and Castellari et al. [1998].

The GHER hydrodynamic model has been applied successfully in different areas, at different resolutions. In particular, its implementation in the Western Mediterranean Sea has been described in Beckers [1991], Beckers et al. [1997], showing that a realistic circulation could be reproduced. During the MEDMEX experiment [Beckers et al., 2002], its results in the Mediterranean were also compared to other models: the POM model [Zavatarelli and Mellor, 1995], the OPA model [Herbaut et al., 1996, 1998] and the MOM model [Álvarez et al., 1994]. It appeared that when forced by monthly mean atmospheric forcings, all models exhibit a similar seasonal cycle and climatic drift. When forced by daily fields, no model clearly outperformed the others; the correct setup and calibration of the models appeared to be the critical step, particularly concerning vertical diffusion, and, to a lesser extent, horizontal diffusion.

2.2 Passive and interactive nesting

As mentioned in chapter 1, the desire to study coastal (or other) areas of interest with high-resolution naturally led to the concept of nested grids. This has been implemented in the GHER model, and is described in Barth et al. [2005]. We summarize it here. The nesting procedure can be used one-way or both-way. In the former case, it simply skips the feedback procedure. The grid refinement ratio r is supposed to be odd and equal in both horizontal directions. Thus, in the overlapping region, each grid value of the coarse grid coincides with a value of the fine grid (see Fig. 2.1). In the present implementation, no refinement is used in the vertical.

At the boundary of successive domains, the bathymetry is kept constant over r fine grid cells (as illustrated by the small "blocks" which can be seen along the open sea boundaries in Figs. 3.1 and 3.2 for the particular implementation described in chapter 3). Thus, over the boundary band, to each vertical σ layer of the coarse grid corresponds one layer in the fine grid at the same depth. Over the band, the land-sea mask is also identical for both grids, which simplifies the interpolation of the boundary conditions and the feedback procedure. The remaining land mask is chosen freely, to follow the real coastline and bathymetry as closely as possible.

The nesting procedure between the coarse and fine grid models can be summarized as follows:

1. interpolation of the fine grid boundary conditions for that variable from the coarse grid model

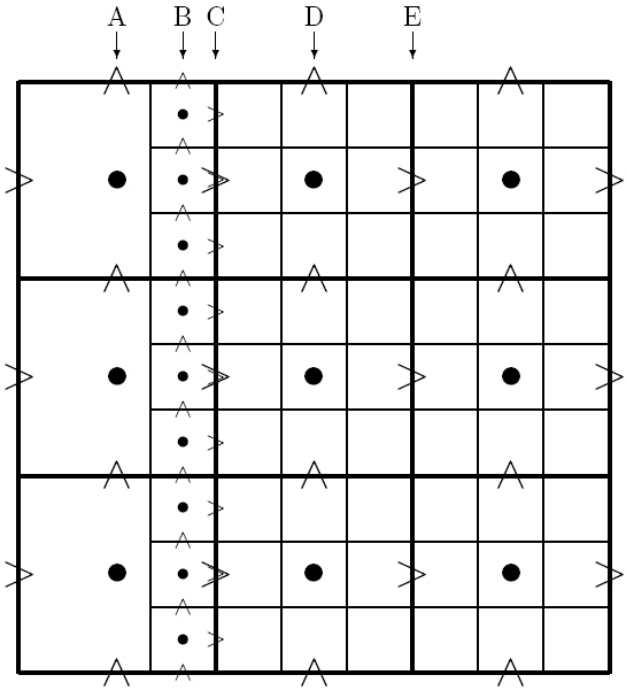

Figure 2.1: The relative position of the coarse (heavy lines) and fine (fine lines) grid. The dots show the position of scalar variables, the arrows the velocity components on the Arakawa C-grid. Large symbols correspond to the coarse grid, small symbols to the fine grid. For clarity, only the position of the variables imposed by boundary conditions are showed for the fine grid. The boundary conditions of the scalars and the tangent (to the nesting boundary) velocities interpolated from columns A et D are imposed in column B. The normal velocity component is imposed on column C. When using interactive nesting, the average values of the scalars and the tangent velocities are injected in the coarse grid, starting with column D. For the normal velocity, the feedback begins with column E. From Barth [2004].

2. integration of both grid models for one time step
3. averaging the values of the fine grid model lying on and inside the "feedback interface" and replacement of the corresponding values in the coarse grid model

Let's note that in principle, when feedback is used, the interior of the overlapping region does not need to be integrated by the coarse model (the values will be replaced anyway). Furthermore, it is not necessary to transfer the entire content of the fine model during the feedback step. Only those values in the immediate neighborhood of the interface are needed to integrate the coarse model at the following timestep. However, the computational burden of both these operations is relatively small, and it allows to have "complete" model outputs in both grids at any time step.

Open sea boundary conditions

We also mentioned in chapter 1 that an important challenge of grid nesting is to implement the open sea boundary conditions.
Relaxation methods let the fine model solution tend toward the external data on the boundary. The most brutal way to do this is to impose the external solution on the boundary, i.e. to use a Dirichlet condition. This condition is often used, but a major drawback is that the outflowing information is totally determined by the external data, and does not depend at all on the interior solution. Therefore, part of the outgoing information will be reflected into the domain as soon as the external data is not perfectly consistent with the internal dynamics. Thus, such a boundary condition should be avoided. It is frequent in practical applications to use a more progressive method, called flow relaxation scheme. This can also be interpreted as adding a nudging term to the original model equations. A so-called *sponge layer* is implemented, where the solution is replaced by a linear combination of the internal and exterior solutions; the model viscosity is also decreased toward the boundary. Other progressive methods are possible. For example, Onken et al. [2005] used a data assimilation filter in order to implement the open sea boundary conditions in a nested system using very different hydrodynamic models in the different grids, also leading to some combination of interior and exterior solution around the open sea grid border.
Radiation methods are based on the transport of the variables through the boundary, and are very popular. It has however been shown recently [Blayo and Debreu, 2005] that these conditions are justified only in the context of wave equations with constant phase velocities; and it can therefore not be justified in ocean modeling.
Model adapted methods, as opposed to the previous boundary conditions, take into account the model equations. However, due to their complicated implementation, they are seldom used and restricted to simplified 1D or 2D models.

The choice of open sea boundary conditions thus is a delicate task. In the GHER model, the integration of the fine grid model requires boundary conditions for horizontal velocity, temperature, salinity and turbulent kinetic energy; they are obtained by interpolation, which is performed for each vertical level independently.

Normal velocity This component of the velocity has the greatest impact on the flow inside the fine grid model. Due to the choice of the grid staggering, only an interpolation tangent to the boundary is required, in column C of figure 2.1. The interpolation is performed in a way so that (a) the volume is conserved, and (b) abrupt variations are penalised. The interpolation coefficients are computed once and stored.

Tangent velocity The velocity tangent to the boundary only plays a role in the horizontal mixing and advection of momentum; these terms are generally small compared to the other forces. Thus, a simple bilinear interpolation is implemented. No volume conservation constraints must be taken into account for this component.

Scalars The interpolation of the 3 scalar variables is performed in two steps. First, the variables are linearly interpolated normally to the boundary; then they are interpolated tangently to the boundary using the same reasoning as for the normal velocity. However, this procedure is not conservative for the scalars. The design of a stable and flux conservative interpolation scheme is not trivial, since a smooth flux boundary condition does not guarantee a resulting smooth scalar field.

Sponge layer

At the nesting boundary, over a distance of two coarse grid cells, the diffusion in the fine model is linearly raised to reach the diffusion of the coarse model. This is justified by the following reasons: (a) small scale features moving outward the fine grid cannot properly be resolved in the coarse grid – the sponge layer allows to damp them, (b) the sponge layer regularizes the reaction of the fine grid model to imperfect boundary conditions coming from the coarse model (which necessarily exhibits a different behavior).

Nested grids initial conditions

Yet another issue raised by grid nesting procedures concerns the initialization of nested grids. This problem is particularly important in operational systems where the nested grids are reinitialized regularly (e.g. weekly). In most cases, the fine grid initial condition is obtained by interpolation of the coarse grid solution in the overlapping domain. However, this has been shown to raise problems, particularly if the successive hydrodynamic models represent different phenomena and use very different algorithms (e.g. a free surface model nested in a rigid-lid model, a model using the σ coordinate nested in a model with the z coordinate, or 2 models using different bathymetries on the boundary, leading to the extrapolation of fields when the fine model bathymetry is deeper). In many cases, the fundamental balance of the fields is destroyed, and spurious waves are generated [Auclair et al., 2001]. Even when this is not considered too harmful, (re-)initializing the nested model with interpolated fields simply leads to the loss of all small scale structures developed previously. In the MFS system mentioned before, some regional models are reinitialized weekly (this is called

slave mode), but perform a hindcast in order to regenerate the small-scale features. Some regional models are simply reinitialized on the first day of the new simulation, without hindcast, while some models are never reinitialized (*offline mode*) and rely on the parent model only for the boundary conditions.
A relatively new technique of variational initialization is now used operationally. It consists of finding the (initial) field that minimizes a cost function with 2 terms. The first one penalizes deviations from a first guess (usually the interpolated coarse-grid field). The second term penalizes fields that do not correspond to "observations", which may be constraints like tendencies of the (high-resolution) tangent linear equations, or some particular local or global observations. The scheme is, in fact, equivalent to a variational data assimilation scheme. Let's mention that it requires considerable computer resources. Tests performed with this initialization scheme showed its beneficial effect on both the short and longer time scale circulations. In particular, it was shown that it does not suffice to crudely interpolate the coarse resolution fields, and then remove numerical high frequency transients, since those also have consequences on the pressure field [Auclair et al., 2000, 2001].
In the present implementation, no such initialization scheme was used. The problem is less severe in our setup, because (a) a single numerical code is used (only the horizontal grid resolution and the diffusion coefficients are modified between grids, but the vertical coordinates and bathymetry are identical), and (b) the model is initialized only once, and spun up (i.e. no reinitializations).

2.3 Data Assimilation

The state vector describing the ocean will always be affected by some uncertainty. Therefore, with the availability of numerous observations, a data assimilation scheme was implemented by Barth et al. [2006] in the GHER hydrodynamic model. Before briefly describing it, it is useful to examine different present day algorithms. We will limit ourselves to sequential methods in the statistical framework of data assimilation, and ensemble filters.

2.3.1 Nudging

The simplest data assimilation method is called nudging; it is an empirical method, its principle consisting in adding a source term in the hydrodynamic model equation governing the observed variable. Whenever an observation is available, a correction, proportional to the difference between the model variable and the observed variable, is applied. Thus, the method is implemented in the model itself, and no separate "assimilation step" is required.
With the nudging method, only model variables can be observed; moreover, when a variable is modified, the other ones cannot be modified accordingly. The method has been applied to assimilate altimetric measurements [Blayo et al., 1996], SST and S-T profiles [Rixen et al., 2001]; it is also often used with climatologic data instead of observations, thus preventing the model to drift too far from the climatology.

2.3.2 Least-squares and Optimal interpolation

Optimal interpolation [Eliassen, 1954, Gandin, 1965] is usually introduced using the following simple example from Talagrand [1997]. An unknown quantity x^t has to be determined from two measurements (y_1 and y_2); the estimation will be called x. Supposing that both measurements are unbiased, uncorrelated and have variances of σ_1^2 and σ_2^2 respectively, we search a solution of the form

$$x = a_1 y_1 + a_2 y_2 \tag{2.12}$$

We are looking for an unbiased estimate x, i.e. $E(x - x^t) = 0$, where E is the expectation operation. This is achieved if $a_1 + a_2 = 1$. The vector x which minimizes the statistical variance $\sigma^2 = E[(x - x^t)^2]$ corresponds to weights a_1 and a_2 which are inversely proportional to the variances of the corresponding observation errors, i.e.

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \qquad a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \tag{2.13}$$

In addition, the error variance corresponding to this solution is given by:

$$\frac{1}{\sigma^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \tag{2.14}$$

This solution can also be found through a different, variational approach: the optimal estimation of x^t, called x, should minimize the "distance" to the observations. Thus, the solution is given by the minimum of a cost-function,

$$J(x) \equiv \frac{(x - y_1)^2}{\sigma_1^2} + \frac{(x - y_2)^2}{\sigma_2^2} \tag{2.15}$$

and corresponds to the solution found *via* the statistical approach. The generalization of the above variational principle leads to a family of data assimilation schemes comprising the 3D-Var and 4D-Var methods. However, we will focus only on the statistical approach.

We now introduce the notation that we will use in the following sections. It corresponds to the unified notation conventions proposed by Ide et al. [1997]. All the values representing different variables of the true ocean state on (part of) the entire 2-D or 3-D model grid are assembled into a vector $\mathbf{x^t}$, whose length n is usually $\mathcal{O}(10^6 - 10^7)$. The model is a function M that propagates the state from time t_{i-1} to time t_i, while adding also an unknown dynamic error η_i because the model is imperfect (in particular, it has been discretized, so that subgrid processes are not included). We will first suppose that the model is linear, i.e. $M = \mathbf{M}$. Then,

$$\mathbf{x^t}(t_i) = \mathbf{M_{t_i}}(\mathbf{x^t}(t_{i-1})) + \eta_\mathbf{i} \tag{2.16}$$

We also suppose that the model is unbiased[1], and that the dynamic error covariance matrix is known:

$$E(\eta_\mathbf{i}) = 0 \tag{2.17}$$

$$E(\eta_\mathbf{i}\eta_\mathbf{i}^\mathbf{T}) = \mathbf{Q_i} \tag{2.18}$$

[1] This might seem a strong limitation, as models usually *are* biased, hence the interest for the multi-model approach that averages out biases. However, it is possible to adapt some methods exposed in this chapter to biased models, see e.g. Dee and Silva [1998].

Of course, we do not know the true ocean state at any time, and have to content ourselves with an estimate $\mathbf{x^f}$, which we again suppose unbiased, and of known covariance:

$$\mathbf{x_i^f} = \mathbf{x_i^t} + \epsilon_\mathbf{i} \tag{2.19}$$

$$E(\epsilon_\mathbf{i}) = 0 \tag{2.20}$$
$$E(\epsilon_\mathbf{i}\epsilon_\mathbf{i}^\mathbf{T}) = \mathbf{P_i} \tag{2.21}$$

When observations are available (and do or do not correspond immediately to model variables), they are assembled in the vector y^o of length m, usually (much) smaller than n. It is necessary to build the so-called observation operator H performing the transformation from model variables to observation space. In ocean science, observations usually correspond to model variables, but are observed at locations outside model grid points; H is a simple linear interpolation operator $\mathbf{H}$. Otherwise, it could be a non-linear operator and include physical laws when the observed variables are not even present in the model. The following equation then links the observations with the true ocean state:

$$\mathbf{y_i^o} = \mathbf{H}\mathbf{x_i^t} + \eta_\mathbf{i}^\mathbf{o} \tag{2.22}$$

$$E(\eta_\mathbf{i}^\mathbf{o}) = 0 \tag{2.23}$$
$$E(\eta_\mathbf{i}^\mathbf{o}\eta_\mathbf{i}^{\mathbf{o}^\mathbf{T}}) = \mathbf{R_i} \tag{2.24}$$

We again supposed that the observation errors represented by η^o are unbiased, and of known covariance. They are of course due to a limited instrumental precision, but also because observations naturally depend on every phenomenon present in the sea at the time of measurement, including those that can not be represented on the chosen model grid. A first estimation of this representativity error can be obtained by projecting the observations on the model space, projecting the result back in the observation space, and then examining the difference between this vector and the original observation vector. Errors due to approximations in $\mathbf{H}$ may also be included. Thus, we write that[2]:

$$\mathbf{R} = \mathbf{R_{instr}} + \mathbf{R_{repr}} + \mathbf{R_H} \tag{2.25}$$

Furthermore, we suppose that observation and model errors are uncorrelated:

$$E(\epsilon_\mathbf{i}\eta_\mathbf{i}) = 0 \tag{2.26}$$

We have supposed that all random variables mentioned above are unbiased. In principle, if this were not true, we could correct the bias before proceeding. Dee

[2]The above equations also implicitly suppose that the analyzed ocean state is computed exactly each time when observations are available. However, this is usually not true, and the following approximation is used. Every time observations are available, the misfit or innovation vector $\mathbf{x^f} - \mathbf{Hy^o}$ is calculated, but no analysis is performed. Only at fixed instants (e.g. once a day), the optimal state is calculated using all previous misfits.
This approximation still supposes that observations are instantaneous, and relate to a given instant of the model simulation. However, observations often relate to time-averaged quantities. A new research topic investigates the possibility to group those instants all together in a new state vector, and relate it to the observations through a new $\mathbf{H}$ operator representing time interpolation as well as spatial interpolation (A. Barth, private communication).

and Silva [1998] show how the model bias can actually be estimated as part of the analysis cycle.

In any case, the following linear equation then yields the optimal, analyzed ocean state $\mathbf{x^a}$,

$$\mathbf{x^a} = \mathbf{x^f} + \mathbf{W}[\mathbf{y^o} - \mathbf{H}(\mathbf{x^f})] \tag{2.27}$$

taking into account both the model output and the observations. The $\mathbf{W}$ matrix indicates the relative importance given to the innovation vector with respect to the model forecast; it is also noted $\mathbf{K}$, the Kalman gain matrix.

The generalization of the solution deduced above for 2 scalar observations yields the classical optimal interpolation method, given by 2.27 with the following expression for the gain matrix:

$$\mathbf{W} = \mathbf{P^f}\mathbf{H^T}(\mathbf{R} + \mathbf{H}\mathbf{P^f}\mathbf{H^T})^{-1} \tag{2.28}$$

where $\mathbf{P^f}$ represents the *a priori* (before the assimilation) model state error covariance matrix. The solution of equations 2.27-2.28 is called the Best Linear Unbiased Estimate, or BLUE, of x^f from y^o, as it represents the linear combination of model and observations with the least error variance. Furthermore, the error covariance of the new estimate, called the *a posteriori* error covariance matrix, is given by:

$$\mathbf{P^a} = \mathbf{P^f} - \mathbf{W}\mathbf{H}\mathbf{P^f} \tag{2.29}$$

It should be noted that it does not depend on the observations themselves, but only on $\mathbf{H}$ and the *a priori* forecast and observations error covariances. This may seem a very demanding requirement, but at the same time, it is obvious that a prerequisite for a rational use of a set of observations is to know what has been observed, and with what accuracy. It is therefore a good thing that this requirement comes out of the mathematics of estimation theory.
The theory of statistical linear estimation can be used for evaluating the performance of an hypothetical observing system defined by what that system would observe and with which accuracy, but independently of any actual observations. The main difficulty in optimal interpolation is thus the specification of the *a priori* error covariance matrix. It could been obtained from the difference, at a given instant, between different forecasts started at different past moments. Alternatively, it could be obtained by considering the difference between model outputs and observations, but this usually leads to poor covariance matrices, as the observations are much sparser than the model grid. Another possibility is suggested by Cane et al. [1996] or Pham et al. [1998b]: to use the model variability in time as a proxy for its error. In any way, as the required information is never explicitly available, one will have to compensate for it by as reasonable as possible hypotheses on the covariance matrices. One advantage of studying assimilation in the perspective of general estimation theory is that it forces us to explicitly formulate hypotheses which are necessarily made in one way or another.
The fact that only the first and second order moments of the error vector are required is due to the fact that the estimate $\mathbf{x^a}$ has *a priori* been sought under the linear form 2.27. The determination of the most general least-variance estimate would require the knowledge of the entire probability distribution function

(pdf). However, in the case when the error vector is Gaussian, the associated conditional pdf is also Gaussian, and the BLUE therefore entirely solves the problem of determining the *a posteriori* conditional pdf.
If one is only interested in the *a posteriori* state, it is sufficient to know the matrix $\mathbf{HP^f}$, representing the covariance between the complete state vector $\mathbf{x^f}$ and its observed part $\mathbf{Hx^f}$:

$$\mathbf{HP^f} = E\left\{(\mathbf{Hx^f} - \mathbf{Hx^t})(\mathbf{x^f} - \mathbf{x^t})^T\right\} \tag{2.30}$$

It tells us how an error, detected in the observed part of the state vector, should affect the remainder of that vector.

Let's note that no assumption had to be made about the probability distribution of the estimates of the ocean state. The same solution of eqs. 2.27-2.28 would be obtained by searching for the estimate of the ocean with the highest probability, in a bayesian approach. Such an estimation is very inefficient in practice in high-dimensional problems, unless the assumption of a Gaussian distribution is made. Then, application of the maximum likelihood criterion is straightforward, and immediately leads to the following equation, whose minimum is the solution we are looking for:

$$J(\mathbf{x}) \equiv (\mathbf{x} - \mathbf{x^f})^T(\mathbf{P^f})^{-1}(\mathbf{x} - \mathbf{x^f}) + (\mathbf{Hx} - \mathbf{y^o})^T\mathbf{R}^{-1}(\mathbf{Hx} - \mathbf{y^o}) \tag{2.31}$$

Thus, the BLUE result obtained above can also be obtained from a variational principle. If the theoretical result is the same, the procedure to obtain it is different, as 2.31 can be solved iteratively, using the adjoint model, at a lower cost than the exact solution of the linear system 2.27-2.28. Therefore, it is often this variant which is implemented. However, when the assumption of gaussianity can not be verified, the most-likely solution (resulting from the bayesian approach) and the most variance-diminishing solution do not coincide anymore. An example of such situation is given by currents that exhibit two possible regimes. For example, the Northern current in the Western Mediterranean Sea can either flow along the shelf break of the Gulf of Lions, or (partly) penetrate over it. The most likely solution is one of both these situations, but the most variance diminishing solution might be something in between.
Optimal interpolation, or close variants, is still widespread in operational centers, and leads to quite satisfactory results. However, experience shows that the fields produced by optimal interpolation are often contaminated by unrealistic noise, which must be filtered out through additional appropriate procedures [Talagrand, 1997].

2.3.3 The Kalman Filter (KF)

The Kalman filter can be seen as an extension of optimal interpolation, when observations are available at different instants [Talagrand, 1997, Kalnay, 2003]. In between those instants, the state vector is updated by the linear model according to equation 2.16, and the corresponding error covariance is also updated, according to the following equation, called the Lyapunov equation:

$$\mathbf{P^f_{i+1}} \equiv \mathbf{MP^a_i M^T} + \mathbf{Q} \tag{2.32}$$

Thus, at any stage, the Kalman filter produces the BLUE of the state of the system under observation, using all observations up to estimation time. It also produces the covariance matrix of the corresponding estimation error. Experiments performed with various linear systems have produced convincing results as to the capability of the method for effectively extracting the information contained in the observations and the model.

2.3.4 The Extended Kalman Filter (EKF)

The equations governing the oceanic flow are strongly nonlinear. Moreover, this nonlinear character is at the origin of one of the most important properties of the flow, namely its chaotic character. This imposes stringent limits on the predictability of the flow, and one can legitimately wonder whether a linear hypothesis is legitimate in the context of assimilation. Let us assume that the model M is non-linear. The operator linking model and observation space can also be non-linear; we will note it H. If the difference $\mathbf{x^a} - \mathbf{x^t}$ is small enough, the quantity $M\mathbf{x^a} - M\mathbf{x^t}$ can be approximated by $M'\mathbf{x^a} - M'\mathbf{x^t}$, where M' is the jacobian matrix of the operator M, taken at point x^a. The Lyapunov equation 2.32 accordingly becomes

$$\mathbf{P^f_{i+1}} = M'\mathbf{P^a_i}M'^T + \mathbf{Q} \tag{2.33}$$

Thus, the mean of the Gaussian pdf (i.e. $\mathbf{x^f}$) is predicted by the full nonlinear model, and the error covariance by the tangent linear model.
Similarly, if the forecast vector $\mathbf{x^f}$ is close enough to the true state vector $\mathbf{x^t}$, $\mathbf{y^o_{i+1}} - H\mathbf{x^f_{i+1}} = H\mathbf{x^t_{i+1}} - H\mathbf{x^f_{i+1}} + \epsilon_{\mathbf{i+1}}$ can be approximated by $H'(\mathbf{x^t_{i+1}} - \mathbf{x^f_{i+1}}) + \epsilon_{\mathbf{i+1}}$. Eqs. 2.27 and 2.29 of the assimilation procedure can then be respectively replaced by:

$$\mathbf{x^a_{i+1}} = \mathbf{x^f_{i+1}} + \mathbf{P^f_{i+1}}H'^T(H'\mathbf{P^f_{i+1}}H'^T + \mathbf{R})^{-1}(\mathbf{y^o_{i+1}} - H\mathbf{x^f_{i+1}}) \tag{2.34}$$

$$\mathbf{P^a_{k+1}} = \mathbf{P^f_{k+1}} - \mathbf{P^f_{k+1}}H'^T(H'\mathbf{P^f_{i+1}}H'^T + \mathbf{R})^{-1}H'\mathbf{P^f_{k+1}} \tag{2.35}$$

In these equations, H has been replaced by the jacobian H', except in the expression for the innovation vector. The algorithm defined by the 3 preceding equations is called the Extended Kalman Filter (EKF), see e.g. Jazwinski [1970]. It is valid whenever the differences between the real and estimated states of the system are small enough to allow local linearizations as just described. This hypothesis can fail either if the model is too nonlinear or if the errors are too large. Although the EKF has been reported to produce satisfactory results, there is strong evidence that the tangent linear approximation may not always be valid (Evensen, 1994, Miller et al., 1994). Thus, the need arises for methods to evolve the full error pdf in time, rather than its first and second order moments[3].

The main difficulty of the (Extended) Kalman filter, which is optimal in the linear case with Gaussian errors, resides in the large dimension of the ocean

[3]Other methods exist, that do propagate a few more moments [Leith, 1971, Fleming, 1971a,b, Leith and Kraichnan, 1972, Leith, 1974]. However, their application to large ocean models has been questioned, as already the specification of the second order moment (the covariance) is very difficult, requires huge amounts of computer memory, and can thus only be done in an approximate way.

state, requiring unrealistic computer memory for the storage of covariance matrices and unrealistic computation time to update them. Indeed, the size of matrix $\mathbf{P}$ is n^2, and equation 2.32 requires $2n$ integrations of the model. A simple alternative to the solution of eq. 2.32 is to multiply $\mathbf{P^f}$ with a constant factor at fixed intervals, but experiments suggest that this substantially deteriorates the quality of subsequent assimilations. Three different approaches have been developed to handle these problems:

Updating the error space with a simplified model This reduces the cost associated to the error forecast of the Kalman filter [Dee et al., 1985, Dee, 1990, Daley, 1992], but it is not obvious if the simplified model represents the same error growth as the complete one. It is also possible to propagate only the variance of the error matrix, the non-diagonal elements being specified based on physical assumptions [Daley, 1991]. Another possible simplification of the model is to update the error matrix in a model space of lower resolution than the state vector [Fukumori and Malanotte-Rizzoli, 1995]. Small-scale errors, not present in the error matrix, will thus not be corrected. This might be inefficient in oceanography since the physical models often produce small-scale errors, as for example the misplacement of some eddies or of ocean meanders, and the error covariance needs small-scale structures.

A low-rank approximation of the state covariance matrix The full non-linear model can then be used to update it (see 2.3.5 and following)[4].

Ensemble runs The third approach, which is popular thanks to its conceptual and implementation simplicity, uses an ensemble of likely model states to represent the error statistics given in the EKF by the state estimate and covariance matrix[5]. Several variants have also been developed, which can be interpreted in one or the other of those approaches. Because all these recent filter developments approximate the covariance matrix by a matrix of low rank, their analysis step operates in a low-dimensional subspace of the true error space. Despite different forecast schemes, the analysis scheme of all filters is provided by some variant of the analysis equation of the EKF given above. We will now present some of these methods.

2.3.5 Reduced Order Optimal Interpolation (ROOI)

The ROOI filter uses exactly the same equations as its full rank counterpart. Only the model error covariance matrix has a lower rank than n, the dimension of the state vector $\mathbf{x}$. Different approaches are possible in order to obtain the covariance. For example, one could write the error covariance matrix as the product of a horizontal and a vertical covariance. Other rank reductions are similar to the ones exposed below for the SEEK and RRSQRT filters.
In the ROOI scheme, this reduced-rank matrix is not modified in time. Experiments show that the first analysis cycle induces a consequent correction,

[4] Chapters 3 and 4 are based on this approximation
[5] This approach will be used in chapter 5

essentially bringing the model state on the most probable trajectory that can be reached, from the forecast, by the reduced rank covariance matrix. However, an error still remains uncorrected, which is orthogonal to the space spanned by the covariance matrix. The subsequent assimilation cycles then induce much smaller corrections, as the orthogonal part of the error now constitutes its largest part.
A nice implementation of the ROOI filter is provided in the SOFA software, see De Mey and Benkiran [2002].

2.3.6 The Singular Evolutive Extended Kalman (SEEK) filter

The SEEK filter described in Pham et al. [1998b], Brasseur et al. [1999] is based on the idea that the specification of the matrices $\mathbf{P}$, $\mathbf{Q}$ and in a lesser extent $\mathbf{R}$ is quite difficult to realize. One usually has only an approximate idea of the initial conditions (obtained from a climatologic atlas, a previous model simulation...), and even less is known about the uncertainty affecting it. Specifying the dynamic error matrix would require the accurate knowledge of the statistical behavior of the state process $\mathbf{x^t}(t)$, but how could this be acquired without actually observing it? As a matter of fact, it is even questionable whether the knowledge of such a vast amount of numbers ($\mathcal{O}(10^{12})$) is even possible or useful [Cane et al., 1996]. Thus, in practice, any KF depending on these matrices is somewhat suboptimal. The idea of the SEEK is then based on 2 key elements:
(a) at initial time, $\mathbf{P}$ is represented by a limited number of EOFs, describing the dominant modes of the system's variability. Ideally, the EOFs would be multi-variate and covering the whole grid, so that the covariance will allow to propagate future corrections to every element of the state vector (e.g. correct deeper layers using surface observations, propagate satellite observations between tracks, or correct salinity with temperature observations).
(b) During assimilation cycles, the rank of the error covariance is conserved while the error statistics propagate with time according to the model dynamics. Thus, we assume that the analysis error covariance is of rank r at time t_{i-1}, so that it can be factorized as

$$\mathbf{P^a}(t_{i-1}) = \mathbf{S^T}(t_{i-1})\mathbf{\Lambda}(t_{i-1})\mathbf{S}(t_{i-1}) \tag{2.36}$$

in which the r columns of $\mathbf{S}$ are orthonormal modes and the matrix $\mathbf{\Lambda}$ is diagonal,

$$\lambda(t_{i-1}) = diag\left\{\lambda_1(t_{i-1}), ..., \lambda_r(t_{i-1})\right\} \tag{2.37}$$

This matrix is then updated not following the classical Lyapunov equation 2.32 but is computed in 2 steps, using its decomposition, for $j = 1, ..., r$,

$$\mathbf{s_j}(t_i) = \left\{\mathbf{S^T}(t_i)\right\}_j = \lambda_j^{-1/2}\left\{M_{i-1}\left[\mathbf{x^a}(t_{i-1}) + \lambda_j^{1/2}\mathbf{s_j}(t_{i-1})\right] - \mathbf{x^f}(t_i)\right\} \tag{2.38}$$

$$\mathbf{P^{f^-}}(t_i) = \mathbf{S^T}(t_i)\mathbf{\Lambda}(t_{i-1})\mathbf{S}(t_i) \tag{2.39}$$

followed by

$$\mathbf{P^f}(t_i) = \rho^{-1}\mathbf{P^{f^-}}(t_i) \tag{2.40}$$

In the first step, the full model may be used. Alternatively, its linear tangent could be applied to each column. The second step is a compensation technique

to account for the system noise, and $\rho \in [0,1]$ is the compensation factor, also called forgetting factor. Thus, the compensation operation represents the increment in uncertainty during the model integration from t_{i-1} to t_i due to the imperfect model. Usually, a value of 0.5 is chosen, but a more precise calibration of the forgetting factor might be achieved in practice by trial-and-error. The rank of $\mathbf{P^f}(t_i)$ is identical to the rank of $\mathbf{P^a}(t_{i-1})$.
By replacing eq. 2.40 in the equation of the Kalman gain matrix 2.28, we obtain its particular expression for the SEEK filter:

$$\mathbf{K_i} = \mathbf{S^T}(t_i)\left[\rho\mathbf{\Lambda^{-1}}(t_{i-1}) + (\mathbf{H_i S^T})^T\mathbf{R_i^{-1}}\mathbf{H_i S^T}\right]^{-1}(\mathbf{H_i S^T})^T\mathbf{R_i^{-1}} \tag{2.41}$$

Writing

$$\mathbf{\Lambda}(t_i) = \left[\rho\mathbf{\Lambda^{-1}}(t_{i-1}) + (\mathbf{H_i S^T})^T\mathbf{R_i^{-1}}\mathbf{H_i S^T}\right]^{-1} \tag{2.42}$$

we thus have

$$\mathbf{K_i} = \mathbf{S^T}(t_i)\mathbf{\Lambda}(t_i)(\mathbf{H_i S^T})^T\mathbf{R_i^{-1}} \tag{2.43}$$

Similarly, eq. 2.29 giving the analysis of the error covariance matrix becomes

$$\mathbf{P^a}(t_i) = \mathbf{S^T}(t_i)\mathbf{\Lambda}(t_i)\mathbf{S}(t_i) \tag{2.44}$$

$\mathbf{\Lambda}(t_i)$ may not be diagonal anymore, and the orthogonality of the columns of $\mathbf{S}$ has been destroyed during their forecast by the model. However, the last equation shows that the rank of the analysis error covariance is left unchanged; and it is still possible to re-orthogonalize the projection operator. The most expensive step in the SEEK filter is given by equation 2.38, requiring r supplementary model integrations. However, the SEEK scheme is numerically feasible in realistic ocean models. In particular, it should be noted that the covariance matrices $\mathbf{P^f}$ and $\mathbf{P^a}$ must never be formed and stored in memory explicitly; only the square root matrices $\mathbf{S^f}$ and $\mathbf{S^a}$ are involved.
The main problem remains to choose an appropriate initial error covariance; in practice, the authors of the SEEK filter recommend to build it from an EOF decomposition of a previous model run achieved without assimilation. However, it has been observed multiple times that covariances obtained from historical runs present unphysical long range correlations due to the seasonal variability; artificial methods, such as limiting each correction to a certain area, must then be implemented to suppress the long range corrections.

Brasseur et al. [1999] notes that, similarly to the situation in the ROOI filter, with the evolutive SEEK filter, if the EOF basis is consistent with the actual model-observation misfits, the largest correction is brought during the first assimilation cycle. If the basis is not statistically representative of the misfit, the assimilation fails even when the covariance matrix is updated according to model dynamics. Therefore, he proposes a new variant where the error space is enriched based on the residual errors after each assimilation cycle, using "statistical learning". It relies upon a relevant inversion of the innovation vector computed after the analysis step,

$$\mathbf{d^a} = \mathbf{y^o} - \mathbf{H x^a} \tag{2.45}$$

containing the piece of information which has not been assimilated into the analysis. Asymptotically, it should contain only the observation noise, $\|\mathbf{d^a}\| =$

$[trace(\mathbf{R})]^{1/2}$. However, in practice, the observation error is often much smaller than the observation error resulting from the omitted error. Thus, a smooth inversion of $\mathbf{d^a}$ formally noted

$$\mathbf{s}(t_i) = \mathbf{H_i^{-1}}\,[\mathbf{d^a}] \tag{2.46}$$

is expected to represent the truncation error that is precisely missing in the sub-space. This inversion is mathematically ill-posed, but can be realized in particular cases by including supplementary hypotheses (an example is given immediately below). The strategy to include the inverted innovation in the algorithm is the following: the error mode which has the smallest contribution to the Kalman gain is determined after each analysis cycle, and replaced by eq. 2.46. This new mode enriches the reduced space, offering a suitable direction of correction for the next analysis.
Brasseur et al. [1999] implemented this variant of the SEEK filter in a twin-experiment using the MICOM model [Bleck, 1978, Bleck and Boudra, 1986], assimilating surface pressure, and using $r = 108$ modes in the covariance matrix. In eqn. 2.46, the inverse of the $\mathbf{H}$ operator is applied to a vector in the observation space, resulting is a vector in the model space. In this particular case, the application of the inverse operator is replaced by the 2 following steps. First, the surface pressure information is propagated in the vertical using vertical correlation coefficients deduced from the first (vertical) pressure EOF, which accounts for 97% of the variability in this particular case. The pressure anomaly is then propagated to the other variable in the state vector (the horizontal velocity variables) using the geostrophic relationship. Results show that this new variant yields better results than the "normal" evolutive SEEK filter, itself better than a non-evolutive version where $\mathbf{P}$ is kept constant in time. However, we note that the inversion of the observation operator is not easy to obtain in the general case.

Another interesting variant of the SEEK filter, called the Semi-Evolutive Partially Local Filter (SEPLEX), is described in Hoteit et al. [2001]. It aims at further reducing the computational cost of the SEEK filter, as $r + 1$ model integration might still be too costly for operational oceanography. Therefore, it introduces the concept of local EOFs covering only a small region, while vanishing elsewhere. This limits the correlation length of the ocean variables, which is consistent with the idea that such correlation should vanish for far away spatial locations. Further, it allows choosing a different amount of EOFs in each sub-domain in order to maximize representativity. There is however a difficulty in the above "local basis" approach: it cannot evolve with the model without destroying its locality property. Therefore, the local basis is kept fixed in time, and augmented by a few global basis vectors which evolve. The resulting filter is much less costly than the SEEK, yet in their experiments, Hoteit et al. [2001] found that the results of this filter are better.

2.3.7 The Reduced Rank SQuare Root (RRSQRT) filter

Verlaan and Heemink [1997] introduce another reduced rank filter called RRSQRT. Its principle is the same as the SEEK filter, and it can be shown that the analysis is mathematically equivalent to the SEEK filter; the model error covariance

matrix is decomposed as:

$$\mathbf{P} = \mathbf{S}\mathbf{S}^{\mathbf{T}} \tag{2.47}$$

and again, $\mathbf{S}$ may be estimated by a $n \times r$ matrix rather than $n \times n$. A difference with the SEEK filter lies in the update of the error covariance. Pham et al. [1998b] do not use any model error and instead amplifies the previous error estimates by a forgetting factor, thus relying entirely on the initial error. In the RRSQRT KF, the dynamic model error is considered, and columns are added accordingly to the $\mathbf{S}$ matrix. In order to keep its rank constant, it is then necessary to "reduce" the matrix by computing its eigenvalues, and retaining only the columns corresponding to the largest eigenvalues. Verlaan and Heemink [1997] propose an efficient algorithm for this operation.
If the columns of $\mathbf{S}$ are updated in time with the tangent linear model, Verlaan and Heemink [1997] note the following problem. If we replace $\mathbf{S}_j$ with its opposite -$\mathbf{S}_j$, a different solution will be obtained compared with the full model. This is solved by gathering in $\mathbf{S}$ the dominant eigenvectors with both signs. The obtained filter, which is more robust, multiplies the computational cost of the forecast step by two.

2.3.8 The Ensemble Kalman filter (EnKF)

The desire to retain more than just the two first moments of the error pdf in data assimilation led to the development of the EnKF, first presented in oceanography by Evensen [1994], and in atmospheric modeling by Houtekamer and Mitchell [1998].
An ensemble of N equally likely ocean states contains the information of all statistical moments of the pdf. Let's note that this pdf needs not to be Gaussian. The analysis of the EnKF then extracts the ensemble mean $\overline{\mathbf{x^f}}$ and covariance $\mathbf{P_e^f}$ as

$$\overline{\mathbf{x^f}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{x^{f(k)}} \tag{2.48}$$

$$\mathbf{P_e^f} = \frac{1}{N-1}\sum_{k=1}^{N}(\mathbf{x^{f(k)}} - \overline{\mathbf{x^f}})(\mathbf{x^{f(k)}} - \overline{\mathbf{x^f}})^T \tag{2.49}$$

In his original paper, Evensen [1994] proposed to assimilate observations into each of these members using the classical Kalman equations, the Kalman gain being obtained from the statistical matrices. Thus, for the analysis step, Gaussianity must be assumed; the mean and covariances needed to describe the Gaussian function are extracted from the ensemble.
About the important question of the choice of N, one should notice that convergence is, in principle, slow, as the convergence rate is proportional to $1/\sqrt{N}$. However, simulated case studies did not show the interest of using more than a hundred members [Evensen, 1994], a few hundred members at most [Evensen, 2002]. Of course, this amount depends on many factors, including the model resolution.
Burgers et al. [1998] showed that in order not to artificially diminish the ensemble variance, the uncertainty affecting observations, given by $\mathbf{R}$, should be taken into account by perturbing the observations accordingly :

$$\mathbf{y^{o(k)}} = \mathbf{y^o} + \epsilon^{(\mathbf{k})} \tag{2.50}$$

and the observation error covariance $\mathbf{R}$ should be close to the observation ensemble covariance $\mathbf{R_e}$

$$\mathbf{R_e} = \frac{1}{N-1} \sum_{k=1}^{N} \epsilon^{(\mathbf{k})} \epsilon^{(\mathbf{k})^{\mathbf{T}}} \tag{2.51}$$

The latter should be used in the Kalman gain matrix computation rather than the original matrix $\mathbf{R}$. If the original matrix $\mathbf{R}$ were to be used, the ensemble spread would be diminished, and the *a posteriori* model error covariance would be smaller than prescribed by equation 2.29. The need to degrade the observations is an undesirable feature of the EnKF, and several variants exist to avoid this: the Ensemble Square Root Filter (EnSQF) [Whitaker and Hamill, 2002], the Ensemble Adjustment Filter (EAKF) [Anderson, 2001], and the Ensemble Transform Kalman Filter (ETKF) [Bishop et al., 2001].
Houtekamer and Mitchell [1998] speak of "inbreeding", since the gain computed from the covariance matrices $\mathbf{P^e}$ and $\mathbf{R^e}$ is built from the ensemble itself. This leads to too small ensemble variance when using small ensembles [van Leeuwen, 1999]. Hence, Houtekamer and Mitchell [1998] propose to use a double ensemble, one of which is used only to obtain the Kalman gain. However, one might argue that these inbreeding problems only occur when the ensemble size is small; in this case, the EnKF technique should be used with great care anyway. In any case, even if the ensemble based covariances are unbiased covariances of the true error covariances (which should be the case by construction for $\mathbf{R^e}$, but is less straightforward for $\mathbf{P^e}$), the *a posteriori* ensemble might be negatively biased as it is the result of a non-linear transformation. Moreover, van Leeuwen [1999] showed that the sampling error of $\mathbf{P^e}$ leads to a systematic underestimation of the *a posteriori* error covariance. When using a double ensemble, there will be a systematic overestimation.
Thus, using $\mathbf{R_e}$ still leads to inaccuracies and large sampling errors if the observations are weakly or not correlated. The use of the original $\mathbf{R}$ matrix solves this problem of sampling error.

Hamill et al. [2001] show that, exactly like the SEEK filter, the EnKF is also subject to unphysical long range correlations: since the correlation between far-away locations should be smaller, the signal/noise ratio becomes larger. Therefore, Hamill et al. [2001] has implemented a distance-dependent reduction of the error covariance in the EnKF, and tested the new filter with a two-layer atmospheric model. This yielded better results than the unmodified EnKF. The authors also showed that the optimal correlation length scale in the filter function depends on the ensemble size; larger correlation lengths are preferable for larger ensembles.

Despite the problems mentioned, the EnKF has been applied to a broad range of physical and coupled biological models (as well as completely different problems), and has proved better adequation for highly nonlinear systems than the various suboptimal schemes of the EKF. Recent examples of a comparison of the RRSQRT and EnKF schemes are given in Canizares [1999] for a 2D model and in Bertino et al. [2002] for an ecological model; computational burden, stability issues and filter performance are compared. Bertino et al. [2002] rightly observes that the filter efficiency depends very much on the correct knowledge of the *a priori* errors, more than on the particular method used.

When one wishes to practically implement an ensemble forecast, an "optimal" member, corresponding to one's best knowledge of the system, serves as a basis for the creation of N members, and this ensemble is then fed to the model in order to yield the forecast. When different models are run, e.g. by different people or in different facilities, each of them corresponds to some optimal estimate of the real state, as each has particular strengths and knowledge. The so-called "super-ensemble" and "hyper-ensemble" techniques function in the same way as regular ensemble techniques, but using different models and data, representing the same and different physical processes respectively. Examples are given in Kalnay and Ham [1989], Fritsch et al. [2000], Krishnamurti et al. [1999, 2000], Rixen and Ferreira-Coelho [2006].

2.3.9 The Singular Evolutive Interpolated Kalman (SEIK) filter

The SEIK filter was introduced by Pham et al. [1998a], starting again from the idea that the EKF and its low-rank approximations such as the SEEK are limited by the assumption of linearity. Nonlinear optimal filters cannot be implemented in high-dimensional cases, and Monte-Carlo methods are used to approximate them, leading to the EnKF. With the SEIK filter, the assumption of linearity is not alleviated, and the goal is restricted to deal with nonlinearities *better* than the EKF. The method can be viewed as an extension of the SEEK filter by using interpolation in place of linearization.
The SEIK filter is based on the same idea as the EnKF, but rather than using random perturbations in order to generate the ensemble members, they are created using the error covariance matrix. Pham et al. [1998a] calls the method *exact second order sampling*, which hopefully leads to faster convergence, i.e. a smaller amount of members will be needed. The obtained members are updated with the full non-linear model, such as in the EnKF. Experiments indeed show better results than those obtained with the EnKF (see Pham et al., 1998a). One should however keep in mind that the EnKF is valid for all non-linear models, including highly non-linear ones.
Nerger et al. [2004a,b] compared the SEEK, EnKF and SEIK schemes using a non-linear shallow-water model. The EnKF was shown to have a more costly analysis scheme, which also introduces noise in the ensemble because of the requirement of an ensemble of observation vectors. The EnKF also generally requires more members than the SEIK filter to achieve the same results. When using large ensembles, both filters act similarly. Due to its direct forecast of the error modes, the SEEK filter produces results which can be strongly distinct from those predicted by the SEIK filter; for very small ensemble sizes, its results are superior. It is well suited to filter rather coarse structures in which non-linearity is not pronounced.
Following this, Evensen [2004] proposed a new implementation of the EnKF, where the initial ensemble is built similarly to the SEIK initial ensemble; the analysis update equation is also modified in order to avoid the necessity of perturbing observations, based on the 3 variants proposed by Whitaker and Hamill [2002], Anderson [2001], Bishop et al. [2001].

2.3.10 The Error Subspace Statistical Estimation (ESSE) filters

Lermusiaux [1997], Lermusiaux and Robinson [1999] introduced 2 analysis schemes, the stochastic and deterministic ESSE filters, which we will only briefly mention here. The stochastic filter differs from the previous filters in the way the *a posteriori* $\mathbf{S^a}$ matrix is computed. Indeed, updating it according to eq. 2.29 rewritten as

$$\mathbf{S^a} = (\mathbf{I} - \mathbf{KH})\mathbf{S^f} \tag{2.52}$$

misses the variance increase due to uncertainties on the observations. Therefore, the following update equation is proposed:

$$\mathbf{S^a} = (\mathbf{I} - \mathbf{KH})\mathbf{S^f} + \mathbf{KE} \tag{2.53}$$

The columns of E are random perturbations drawn from a Gaussian pdf with zero mean and covariance given by $\mathbf{R}$ multiplied by $1/\sqrt{N-1}$, as in the EnKF. The full observation error covariance matrix is used for the Kalman gain computation. Moreover, the scheme does not work directly with the ensemble perturbation but rather with their singular value decomposition (see Lermusiaux, 1997, Barth, 2004).

The deterministic ESSE filter starts from the same decomposition of the model error matrix as in the SEEK filter (eq. 2.36). The dynamic error is modeled by random linear combinations of the dominant eigenvectors of $\mathbf{Q}$. The observations however are not perturbed.

2.3.11 Statistical Importance Resampling filters

Motivation

In our review of sequential data assimilation methods, we started from the filter with the most severe limitations, the KF which is optimal only for a linear model and Gaussian model errors; the latest filters (EnKF and following) allow to use non-linear models, but are still updating the estimates through the same analysis equations based on means and covariances with higher-order moments discarded; the error pdf is thus still supposed to be Gaussian. It has been shown on multiple occasions that error pdfs might not be Gaussian, or might even be multimodal (e.g. currents switching between different regimes). For instance, the EnKF allows to estimate the third-order moment (skewness). If non-zero values are found, there is no way to use this extra information about the pdf during the assimilation cycle. The updated ensemble will have a skewness that is deformed by the update process, and it is unclear what its meaning is. It might be completely wrong, deviating the estimated pdf from the true one.
The Importance Filters relax even this gaussianity requirement. The general principle of these methods is to forecast an ensemble of N states as in the EnKF, starting from an ensemble of equiprobable members which all cary an equal weight $w_k = 1/N$. When observations become available, the members are not modified, but their respective weights are updated from the Bayes's theorem, by the so-called importance resampling algorithm. Possibly some members are replaced with new ones. As a remarkable consequence, the members are always physically balanced, a feature which could not be guaranteed by any of the

previous schemes.
These methods have become increasingly popular over the last years in various fields of science; links to many theoretical developments and applications are given on the following website: http://www-sigproc.eng.cam.ac.uk/smc/. They were first applied to large-scale geophysical fluid problems by van Leeuwen [2003]. They are also investigated for operational use, see e.g. Skachko et al. [2006].

Implementation

At the heart of data assimilation methods lies the notion of combining the probability densities of model and observations. In Bayesian statistics, the unknown model evolution $\mathbf{x}$ is viewed as the value of a random variable $\underline{\mathbf{x}}$. By expressing the problem in terms of the general pdf rather than a Gaussian pdf defined by its mean and covariance, the density function $f_m(\mathbf{x})$ of $\mathbf{x}$ is obtained from the model somehow and is called the prior probability density. Using the definition of a conditional probability density, we derive the new, or posterior, probability density given the observations $\mathbf{y^o}$:

$$f_m(\mathbf{x}|\mathbf{y^o}) = \frac{f_d(\mathbf{y^o}|\mathbf{x}) f_m(\mathbf{x})}{\int f_d(\mathbf{y^o}|\mathbf{x}) f_m(\mathbf{x}) d\mathbf{x}} \tag{2.54}$$

The variance-minimizing model evolution is equal to the mean of the posterior probability density:

$$\overline{\mathbf{x}} = \int \mathbf{x} f_m(\mathbf{x}|\mathbf{y^o}) d\mathbf{x} \tag{2.55}$$

The discrete interpretation of this equation leads to:

$$\overline{\mathbf{x}} = \sum_{i=1}^{N} \mathbf{x_i} f_m(\mathbf{x}|\mathbf{y^o}) \tag{2.56}$$

or equivalently,

$$\overline{\mathbf{x}} = \frac{\sum_{i=1}^{N} \mathbf{x_i} f_m(\mathbf{y^o}|\mathbf{x})}{\sum_{i=1}^{N} f_m(\mathbf{y^o}|\mathbf{x})} \tag{2.57}$$

This means that each ensemble member (or particle) is weighted by its "distance" to the observations, with the weights given by:

$$w_i = \frac{f_d(\mathbf{y^o}|\mathbf{x})}{\sum_{i=1}^{N} f_d(\mathbf{y^o}|\mathbf{x})} \tag{2.58}$$

This "distance" is found from the probability density of the observations. For Gaussian observational errors, which we suppose uncorrelated for simplicity, with standard deviation σ, the weights are found as

$$w_i = \frac{1}{A} \exp\left[-\frac{(\mathbf{y^o} - H\mathbf{x_i})^2}{2\sigma}\right] \tag{2.59}$$

in which H is the (non-linear) observation operator, and the normalization constant A is given by

$$A = \prod \exp\left[-\frac{(\mathbf{y^o} - H\mathbf{x_i})^2}{2\sigma}\right] \tag{2.60}$$

Let's note that no matrix inversions have to be performed in the above equations; the filter is thus not an inversion problem. Data assimilation *can* be written as an inversion problem for convenience, but it does not have to. The fact that no matrix inversions are required might be a huge advantage of the particle methods when important amounts of data are available. With other methods, one might have to consider a trade-off between computational cost (few observations) and statistical relevance (as many observations as possible).
The moments (of any order) of the ensemble are also easy to obtain as soon as the weights are available:

$$\overline{g(\mathbf{x})} = \sum_{i=1}^{N} w_i g(\mathbf{x_i}) \tag{2.61}$$

There is however one important problem with the method explained here. When the weights are updated, only a few members obtain relatively large weights, while the others have such low weights that they make no contribution in the posterior density to the first two moments. A possible solution is to resample the posterior density after some time to create a new ensemble in which all members have an equal weight again. This can be done in a variety of ways; an overview is given in Doucet et al. [2001].
When the weights of each ensemble member are calculated, a new pdf is generated. A random sample of size N is then chosen in this pdf. Each of these numbers then corresponds to a new member. More numbers will be drawn at high weights than at low weights; and when different numbers are sampled at the same weight, as many identical new members are created. On the other hand, when a weight is so low that no numbers are sampled, the corresponding member is simply discarded. Thus, a new ensemble is created in which each member again has a weight $1/N$. It is resampled from the density defined by the previously obtained weights, and therefore by the relative closeness of the observations to the members. The construction of the new ensemble is illustrated in Fig. 2.2.

If the dynamic error cannot be considered, it is pointless to run multiple copies of the same member. In this case, some "jitter" can be applied to obtain a little more spread; however, it is not clear how this will affect the ensemble pdf. Nevertheless, interesting results have been obtained with this method applied to the Lorenz model [Anderson and Anderson, 1999].
van Leeuwen [2003, 2007] proposes to modify the resampling algorithm when using relatively small ensembles which are often used for computational reasons. Instead of choosing randomly from the posterior pdf determined by the weights, members with large weights are chosen directly from the distribution in the following way. First, the density is multiplied by N. For each so-obtained weight larger than 1, the integer part of that weight determines the number of identical copies in the new ensemble. The remaining (fractional) parts form a new pdf, in which the rest of the ensemble members are drawn according to the rules of the importance resampling algorithm described above[6]. The SIR filter was applied to a real-size problem ($n = 2\ 10^5$) with different ensemble sizes;

[6] The remaining members could be drawn from the new pdf by multiplying it again, and apply deterministic sampling again, and then repeating the procedure over again until all N members are chosen. The stochastical sampling is chosen in the fractional pdf for simplicity reasons.

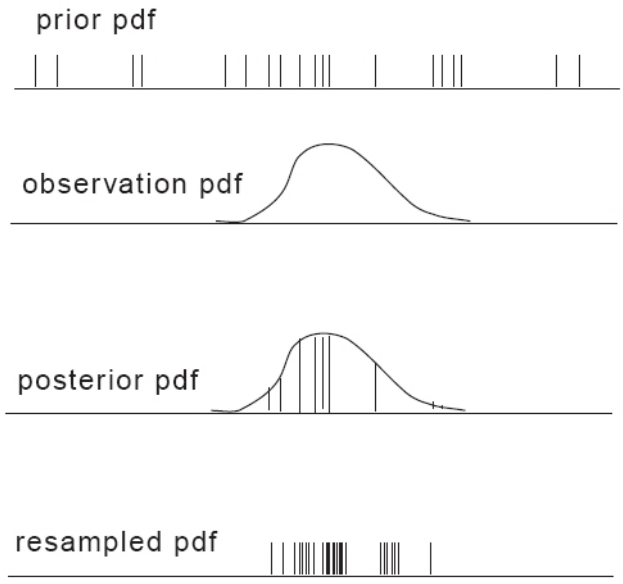

Figure 2.2: Sequential Importance Resampling: the prior pdf, represented by an ensemble, is multiplied with the observation pdf (not necessarily Gaussian) to obtain the posterior pdf as represented by the ensemble. The posterior pdf is resampled to give each member equal weight again. From van Leeuwen [2007].

good results were obtained with $N = 495$. The results showed the uncommon behavior of the SIR filter compared to KF methods. The latter might be drawn too close to observations because of the implicit Gaussian assumption of their prior density during the analysis step. Furthermore, when using the SIR filter, the posterior variance might actually be larger than the prior one, which is impossible with KF-derived filters. The experiments also showed that the method might recover from situations where only a few members conserve some weight (ensemble collapse) provided the dynamic noise is large enough.
In the SEEK and SEIK filters, the model error covariance matrices could also become reduced by the analysis cycle; a forgetting-factor artificially increases the covariance again. However, by doing so, the covariance of observations is relatively diminished, and the model is pulled too close to observations. Thus, the filter partly forgets the prior model state, preventing filter divergence. Indeed, much smaller ensembles can be used without ensemble collapse. It should be noted also that pulling the model very close to observations is not what data assimilation is about. To put it to an extreme, the model should not be used to interpolate between available observations; it contains important information in itself.
The optimal estimate at any given timestep is given by the ensemble (weighted) mean. A drawback we noticed in the theory of the SIR method lies in the fact that this mean might not have much physical sense. Although each particle is physically balanced by the model at all times, there is no guarantee that the mean also exhibits this property. In fact, the mean is a linear combination of the particles, just as the analyzed ocean state in KF variants. Although this is an unpleasant property of the SIR filter, it does not yield stability problems as in other data assimilation schemes (the mean is only used as a diagnostic, not for any computation).

The super-ensemble and hyper-ensemble techniques mentioned above can be interpreted as particular cases of the SIR filter, without the resampling step, and with an *a priori* known uniform observation pdf. The models are run in hindcast mode; when observations are available, a combination of the different models is sought, which minimizes the distance to the observations. The members are then used in forecast mode; the optimal estimate is given by the same combination of the forecasts as during the hindcast. When hyper-ensembles are used, different models concerning physical processes are considered; but it is not straightforward how they should be optimally combined. Rixen and Ferreira-Coelho [2006] use non-linear combinations of the models to find this combination, with tools like neural networks and genetic algorithms. An independent term is also added to account for biases in the models.

The Guided SIR (GSIR) algorithm is a variant of the SIR algorithm, introduced in van Leeuwen [2007], which aims at reducing N for operational implementations. It is explained in Fig. 2.3. When the SIR is applied at the previous observation time (t1 in Fig. 2.3), one waits until the next observations come in. Then the ensemble is integrated forward in time for a few time steps; and then it is assumed that the observations are obtained at this time (t2') rather than at the real instant t2. A SIR-step is then performed. Obviously, one makes an error here, because the observations are not to be used yet. However, this allows to know already at this stage which members are going in the right direction, and

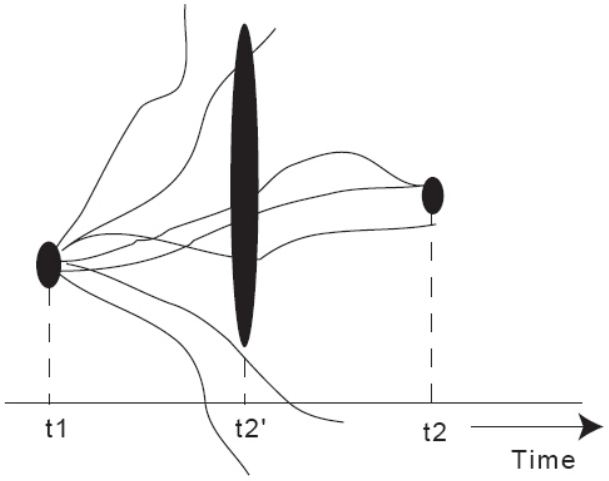

Figure 2.3: Guided Sequential Importance Resampling: a SIR is performed at t2' before the actual measurement time t2 to guide the ensemble towards the observations at time t2. The ellipses denote the observations with their standard deviation. From van Leeuwen [2007].

which members are shooting off to remote areas in state space not supported by observations at all. To avoid being too strict, the measurement error is increased by a large factor, say 100. The system is so non-linear, or the dynamic error so large, that a number of ensemble members is unable to get a reasonable weight at this stage. By using the SIR, they are just abandoned already. The "guiding" step might be repeated one or multiple times between t2' and t2, with an error multiplied by smaller amounts, e.g. 10. Then, the ensemble is integrated up to the true measurement time. Again the SIR is performed, but because the particles are already relatively close to the observations (guided to them), the weights will not differ too much. This procedure allows one to greatly reduce the number of members. In his realistic-sized experiments van Leeuwen [2007] was able to reduce the amount of members needed by a factor 10, making the GSIR filter suitable for operational use. Preliminary experiments also showed that the results of the GSIR filter are not very sensitive to the choice of t2' and the error inflation factor, provided they are in a range along the values suggested above.

Finally, model parameter estimation can be implemented easily within the SIR family of filters. When such a parameter is badly known, it suffices to create different members with different parameter values, and at the end of the simulation, the members with higher weights correspond to more likely parameter values. Other assimilation filters also allow parameter estimation, but never in such a straightforward way. Once again, when non-Gaussian variables are involved, the SIR also presents superior results Kivman [2003].

2.3.12 Data assimilation in the GHER model

The scheme implemented in the GHER model is a fixed SEEK assimilation filter. The equations are those given in section 2.3.6. However, the columns of **S** are not updated by the model in our implementation, due to the computational burden. Thus, eq. 2.38 is not implemented. Furthermore, the "statistical learning" add-on to the filter is not implemented either, because in our general case where temperature, salinity and sea surface elevation could be assimilated at any location of their respective grids, there is no immediate method to invert the observation operator.
The properties of the filter have been described above. In synthesis, only the first- and second-order moments of the error statistics are retained. Using the SEEK filter implies that we suppose that the error processes can be considered quasi-linear, and the model error is approximately Gaussian. This is obviously not always the case, as shown e.g. in Auclair et al. [2003]. However, for relatively short time forecasts, we still will use this widespread approach.

The assimilation procedure implemented in the model allows to correct different variables at once, in a single state vector. Temperature, salinity and surface elevation are correlated through the model covariance matrix, and an observation of any of those variables automatically brings a correction to all of them. The velocity field is not corrected directly by the assimilation procedure, but rather in the following way. A density correction is computed based on the previous variables corrections. The geostrophic velocity correction is then calculated and applied to the model horizontal velocity variables U and V. Thus, we will not be able to correct the ageostrophic part of the error.

The implementation allows global or local assimilation. In the former, observations have an impact on the whole domain, *via* the covariance matrix. In principle, with realistic error covariances, no unphysical long-range correlations should exist. However, many authors notice them using all kinds of filters from Optimal Interpolation to EnKF, and try to tackle the problem and suppress these correlations. Thus, in the latter method, the correction brought by each observation is multiplied with a radial Gaussian function centered on the observation.

The temperature, salinity and surface elevation corrections yielded by observations are linear combinations of the model error modes (in our case, EOFs of model outputs). However, nothing guarantees that these linear combinations still represent a physically coherent ocean state. The velocity field, based on the geostrophic velocity update corresponding to the T, S and elevation update, might also be unbalanced. In order to minimize the chance of yielding unphysical, unbalanced ocean states by the data assimilation procedure, we implemented the following defensive procedure. Before and after assimilation, the Brunt-Väisälä frequency N^2 is calculated for each gridpoint. The spatial mean of its square values is also computed, separately for positive (stable) and negative (unstable) values of N^2. For a point where N^2 is positive, if N^2 after the assimilation is smaller than 3 times the square root of the mean square value of all positive values in the grid, the assimilation correction is completely applied; if N^2 is larger than 10 times the square root of the mean square value, the

correction is not applied. For values of N^2 between 3 and 10 times the root of the mean square value, the correction decreases linearly. The same procedure is applied for points with a negative N^2 frequency. Although this procedure is applied everywhere without *a priori* distinctions, we will notice during our simulations that only in particular places, such as river plumes and continental shelves, the correction will be diminished or totally suppressed.

Chapter 3

Downscaling

La perfection est atteinte,
non pas lorsqu'il n'y a plus rien à ajouter,
mais lorsqu'il n'y a plus rien à retirer.
Antoine de Saint-Exupéry

3.1 Introduction

For various reasons, such as insufficiently known initial conditions, model parameters and atmospheric forcings, or reasons linked to inherent limits to predictability, numerical ocean models progressively drift away from the true state of the ocean state. With the availability of numerous, often real-time or almost real-time observations, data assimilation techniques have proved to be an essential component of operational forecasting systems, as they allow to find a compromise (in some optimal way) between model forecasts and observations. Moreover, end users often ask for high resolution forecasts with adequate physics in coastal zones. A common technique to achieve this is to use nested grids, yielding high resolution results only in a limited region and thus avoiding the computational cost of high resolution in the rest of the basin. Both the data assimilation schemes and the nesting procedure have been introduced in chapter 1, and their implementation in the GHER model has been discussed in chapter 2.
However, after choosing a specific data assimilation scheme, and a specific implementation of the nesting procedure, the combination of both those techniques can be realized in different ways. In particular, the available observations could be assimilated in the coarse, global model, or in the high resolution, local model. When assimilated in the global model, the information they bring would be transferred to the local model through boundary conditions, and possibly through initial conditions is the regional model is reinitialized. When observations are assimilated in the local model, they immediately have an impact on this model's output; but the coarse resolution model, which is not corrected, might then feed the local model with inappropriate boundary conditions. In this chapter, we present series of twin experiments in order to address those questions, under the hypothesis that we are interested mainly in the output of the high resolution model. The results below have also been published in Vandenbulcke et al. [2006].

3.2 The Gulf of Lions

Our study area is the Gulf of Lions (GoL) (bathymetry shown in Figs. 3.1), a large continental margin in the Northwestern Mediterranean Sea, where many small scale processes such as gyres and meanders along the canyons take place. The first internal Rossby radius is 7 to 11 km [Grilli and Pinardi, 1998], with local minima as small as a few km. The presence of coasts and the relatively shallow depths allow the GoL to be influenced by different intense forcings.

1. Due to the surrounding orography, continental winds, when they occur, are mainly channeled in two directions. The northern Mistral and northwestern Tramontane are cold and dry continental winds, which can occur together, the Mistral on the east side of the GoL and the Tramontane on the west side. Those wind fields are known to be key factors for the condition of the sea, leading to sea-surface cooling and the so-called Dense-Water Formation (see e.g. Lacombe and Tchernia, 1974, Millot, 1990, or more recently Estournel et al., 2003). Their spatial variation over the GoL partly determines the circulation in the area. E.g. Estournel et al. [2003] showed that changing a homogeneous Tramontane wind field by a

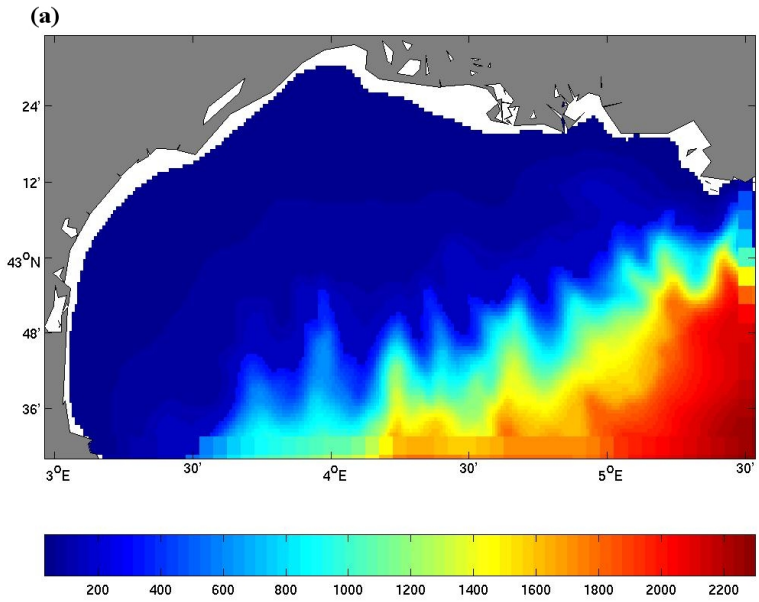

Figure 3.1: Bathymetry of the 3 successive grids, in meters: Gulf of Lions (GoL). The corresponding grid resolution is 0.01°.

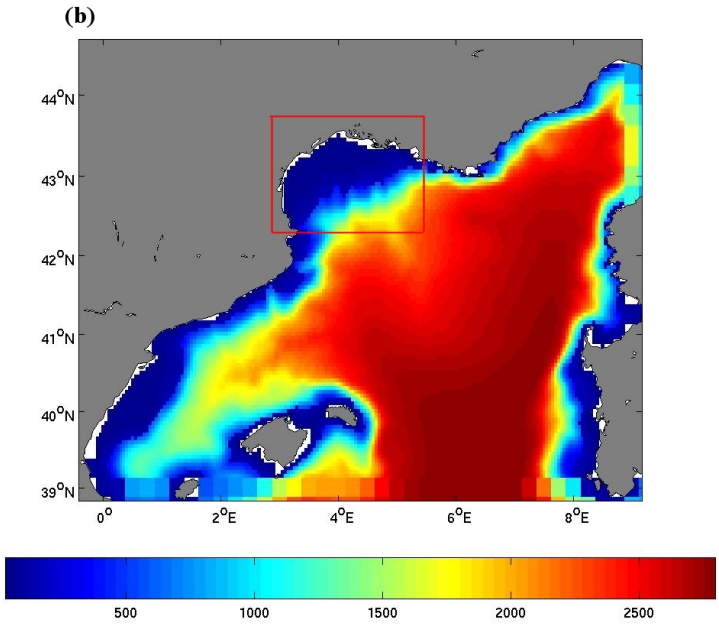

Figure 3.2: Bathymetry of the 3 successive grids, in meters: North-Western Mediterranean (intermediate grid). The corresponding grid resolution is 0.05°. The red box indicates the position of the GoL grid.

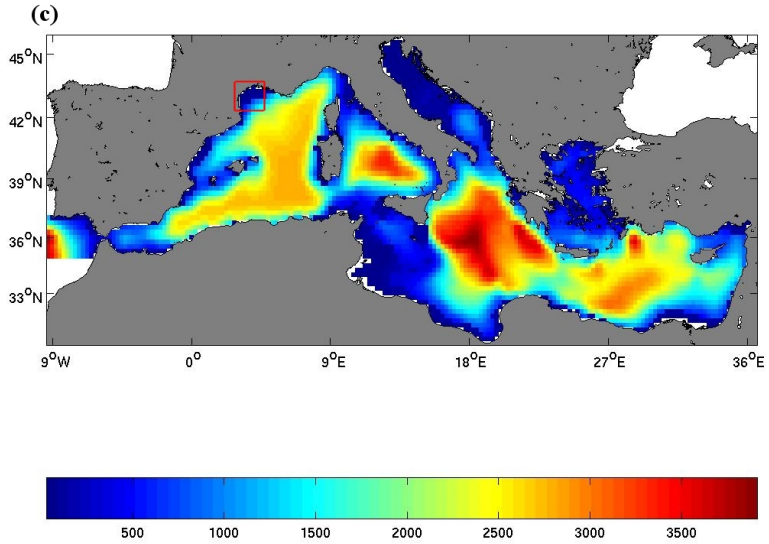

Figure 3.3: Bathymetry of the 3 successive grids, in meters: Mediterranean Sea (coarse grid). The corresponding grid resolution is 0.25°.

spatially varying wind, reverses the shelf circulation! The wind can also cause upwellings [Millot, 1982] and important inertial oscillations, as they become very strong and disappear very quickly [Millot, 1990].

2. Several rivers end in the GoL; the Rhône river is the most important one, and accounts for 80 to over 90% of the total discharge [Fieux, 1974]. Its average flux is about 1800 $m^3\ s^{-1}$. Depending on atmospheric forcings and the surrounding circulation, its plume can reach over tens of km (up to 100 km) in the GoL, and leads to a reduction of the coastal water density.

3. The southern and eastern limits of the domain are open sea boundaries, with a major current flowing through: the Liguro-Provençal-Catalan (LPC) current. This current forms north of Corsica where the Eastern Corsican Current and Western Corsican Current join, and moves westward along the French coast. The LPC is formed of Modified Atlantic Water (down to 300–400 m), and Levantine Intermediate Water (from 500 to 800 m). It can be seen as the northern branch of the general cyclonic gyre of the north-western Mediterranean Sea, and is also called Northern Current. Its importance in the Ligurian Sea has been calculated to be in the range 0.9-1.8 Sv [Alberola et al., 1995, Sammari et al., 1995]. Values up to 2.7 Sv were obtained in January 1982 [Bethoux et al., 1988]. The extreme values obtained in the GoL are 0.5 Sv in June 1992 and 1.8 Sv in April 1992 [Conan and Millot, 1995]. In the winter and spring, the LPC has be found to be deep (about 450 m), narrow (20-30 km), quick (0.5-1 $m.s^{-1}$) and close to the coast, while it is the contrary in summer [Conan and Millot, 1995].

4. The large-scale flow is constrained to flow along the GoL shelf break, but instabilities can make it split in two branches, one of which penetrates the coastal region (particularly at its eastern edge), partly controlling the shelf circulation. Echevin et al. [2003a] writes that when the stratification is strong, a branch of the LPC can flow on the shelf; while when the stratification is weak, the shelf break can act as a barrier. However, Petrenko [2003] found experimentally that a branch of LPC flowed on the eastern side of the GoL even with an almost fully mixed water column. The water "budget" on the gulf could also influence the LPC intrusions: a deficit of water mass in the gulf could favor the intrusions [Estournel et al., 2003]. Thus, the LPC is largely responsible for the exchanges between the open sea and the shelf. Finally, let's note that in the winter, the LPC has been shown to be baroclinically unstable, with rapid variations in its position: the periods are of a few days to 20 days [Crépon et al., 1982]. More references about the LPC current include e.g. Millot [1990], Petrenko et al. [2005] and review paper Millot [1999].

Furthermore, the GoL exhibits a complex bathymetry, with many canyons. When the water column is homogeneous, it has been shown that the influence of the bathymetry can be felt on currents almost up to the surface [Petrenko, 2003]. In any case, the deep portions of the LPC are strongly influenced by the canyons. It should be mentioned that non-hydrostatic models might yield better results in areas with such canyons.

The GoL is also known to be the siege of an extremely important process in the Mediterranean Sea, the formation of dense water. This phenomenon constitutes a widespread process over continental shelves, with tens of locations listed all over the world, mainly in Antarctic and Arctic regions, but also at midlatitude. During the winter, the Mistral and Tramontane winds blow regularly over the GoL, bringing cold and dry atmospheric conditions. When they are associated with strong negative heat fluxes, they generate dense water called Winter Mediterranean Intermediate Water (WIW) above the continental shelf [Fieux, 1974]. This process can conveniently be discussed in three phases. In the *preconditioning* phase, due to cooling and evaporation, an increase in density appears on the surface waters, and a convective movement starts. A cyclonic vortex of about 100 km diameter appears, inscribed in the regional cyclonic circulation. This in turn results in turbulent vertical mixing. The contrast between surface and intermediate layers is very much reduced (even though the 3-layer structure keeps existing), as the surface layer is cooled and has some saline water from the intermediate layer mixed up into it. During the next phase, called *violent mixing*, strong winds blow. Meanders appear along the cyclonic vortex, and smaller eddies (of a few km) are created. Even smaller plumes (a few hundred meters) are also created. The surface layer density reaches that of the layers beneath, mixes rapidly downwards in the small plumes, with vertical speeds about 10 cm/s. It then completely destroys the intermediate layer and incorporates deep water. The few-km eddies do not contribute to the sinking of the new water, as they rather tend to re-stratify the region. It should be noted that hydrodynamic models using monthly atmospheric fluxes are generally not able to represent deep water formation correctly, in contrast to models using high-frequency fluxes (e.g. 6-hourly). This shows the intermittent and often violent nature of the phenomenon, which is linked to a series of specific storms rather than to a gradual cooling [Robinson et al., 2001]. Finally, during the *sinking and spreading* phase, the columns of dense water formed in the violent mixing phase sink and spread laterally in the deeper layers, where their density matches that of the surrounding waters. The most dense waters will reach the bottom, and will thus be more prevented from mixing than the less dense waters, which will be more easily mixed with waters of similar density. There is also a marked variability at annual scale, since meteorological conditions are clearly different from year to year.

The WIW differs from the Winter Mediterranean Dense Water (WMDW) which is formed in the open ocean of the GoL under the same strong atmospheric conditions. WIW's temperature is lower than 12°C, and its salinity is lower than 37.9, whereas WMDW has a temperature of 12.8°C and a salinity of 38.4 [Dufau-Julliand et al., 2004]. WIW is also formed on the continental shelves in the Balearic sea, and is commonly found there below MAW (see e.g. Salat and Font, 1987). The formation of dense water was first observed in 1969 by the MEDOC Group [1970]. The western part of the GoL seems to be more favorable to dense water formation, since the Rhône river fresh water may inhibit this in the eastern part. WIW was later observed over the whole shelf during the 1971 winter, when severe conditions (strong winds, intense negative heat fluxes) occurred [Person, 1974]. Using atmospheric observations, he calculated the average temperatures, wind velocities and evaporation rates and showed that they were very severe. Once formed, the WIW plumes move on to the bottom of the shelf. At the (southwestern part of the) shelf limit, different reasons allow the

dense water to *cascade* over the shelf break. These reasons include friction effects and Ekman drainage: velocity is reduced in the Ekman bottom layer, in turn reducing the Coriolis force responsible for the geostrophic equilibrium and following, the along-slope current. They also include local effects such as the complex bottom topography. Following Hill [1998], the canyon sidewalls can eliminate the effect of the Coriolis force and enhance the downslope motion due to the buoyancy forces.

All the processes detailed above imply that a correct modeling of the GoL requires to take into account all scales, from small to large, in a full 3-D model.

3.3 Model implementation and results

Further details of the implementation used for our study are given below.

As explained before, the Gulf of Lions is the siege of relatively small-scale processes that cannot be resolved by coarse resolution grids, but yet are (also) dependent on the large scales. For the simulations described below, and according to the discussion above, a resolution of 1/100° (approximately 1 km) was desired. To resolve the open boundary problem, a system of nested grids was implemented. In order to avoid a high factor between the resolutions of the coarse and fine grids, an intermediate grid was also implemented, yielding two successive refinement factors of 5. The coarse resolution model covers the whole Mediterranean Sea with a resolution of 1/4°. The intermediate grid covers the area of the North-Western Mediterranean Sea, its eastern boundary being the Corsica and Sardinia islands, with a resolution of 1/20°. Finally, the third grid covers the area of the GoL with a resolution of 1/100°. The three grids are shown in Figs. 3.1 to 3.3.
The original bathymetry is the Smith and Sandwell [1997] bathymetry. It is also smoothed more in the coarse grid than in the GoL grid.
A model covering the whole Mediterranean Sea is started from MODB climatological initial conditions, and spun up for 10 years. Interpolation and averaging then yield the initial conditions of the 2 other grids. The same timestep is used in all 3 grids: 3 s (barotropic mode) and 3 min (baroclinic mode). We use a relaxation term towards the MODB/MEDAR4 climatology in all three grids. Starting on 1 January 1998, the 3 models are spun up one month using two-way nesting. We use climatological Rhône river discharges [Tusseau and Mouchel, 1994]. Atmospheric data (used to calculate fluxes at the air-sea interface) are the 6-hourly European Centre for Medium-Range Weather Forecasts (ECMWF) reanalysis fields with a spatial resolution of half a degree. The wind speed and direction, averaged over the GoL grid, is shown in Figs. 3.4a and b. It can be seen that 3 Mistral/Tramontane wind events take place around 14, 16–17 and 20 January. Other, similar or smaller events take place in February. It should be noted however that these average wind values do not correspond to a wind field homogeneous over the GoL. Finally, on 30 January, the twin experiment runs begin.

Although the purpose of this study is not to examine the hydrodynamics of the GoL itself, but rather to examine data assimilation in nested grids, we

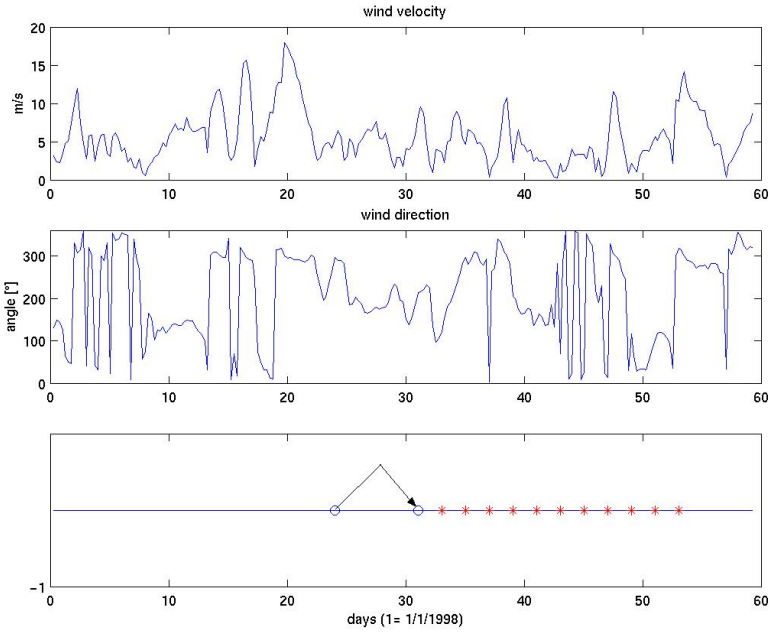

Figure 3.4: **(a)** Average ECMWF-reanalysis wind velocity over the Gulf of Lions **(b)** wind direction, 0° corresponds to east, and 90° to north **(c)** data assimilation cycles on the same time-scale. The blue circles represent the false and correct initial condition used to start the twin experiment on 30 January, while red stars represent assimilation cycles.

still present a brief summary. The results in the intermediate-resolution grid (Fig. 3.5a) clearly show the cyclonic gyre of the Western Mediterranean Sea. Its northern part, the LPC, shows a width of 30 to 40 km, in good agreement with the literature. Its current velocities are about 30 cm.s^{-1} and the associated transport is about 1.3 Sv, which is slightly less than the maximum values usually found for the winter in the literature, of 1.5 to 2 Sv (see e.g. Millot [1999]). The LPC is of course also visible in the lower part of the high-resolution grid (Fig. 3.5b), as well as its strong meanders at various places. The first internal Rossby radius in the GoL is only of a few km. Figure 3.6 shows the salinity along the A'-A line in Fig. 3.5b, just next the Rhône river mouth. The salinity in the GoL is strongly influenced by the river plume, with values as low as 35.5 psu just at the river mouth. Along the shelf break, a vein of Levantine Intermediate Water (LIW) shows good agreement with the literature [e.g. Sparnocchia et al., 1995, Millot, 1999]. It has a high salinity (38.45 psu in our simulation, compared to 38.6 psu in the literature), and its depth ranges from 300 to 700 m, when it can go from 200 up to 1000 m in the literature. Below the LIW, the Western Mediterranean Deep Water (WMDW) has a lower salinity (38.3 psu compared to 38.4 in the literature). The temperature field shows the same water masses, also in good agreement with the literature. Finally, let's note that further away from the Rhône river mouth, the GoL waters are well mixed (not shown).

3.4 Assimilation implementation

As seen before, the data assimilation scheme implemented in the GHER model uses a reduced-rank model error space (dimension rxN instead of N^2, where N is the model state vector dimension). Here, we take r=20 as it is a reasonable trade-off, allowing to represent the errors fairly well while keeping computational cost low enough. As explained in Sect. 3.2, the processes involved in the GoL cannot be described otherwise than fully 3-D. Therefore, multivariate 3-D EOFs are calculated from the daily model states, from 2 model runs, one covering January and February 1997 and one covering January 1998. They are computed at the same time, in a single operation, over 3 variables (T, S and η) and over the 3 grids, after the temporal mean has been removed from all the fields. In the state vector and EOFs, each point uses a norm equal to the product of the corresponding grid cell volume and the variables variance. Some parts of the first EOF are shown in Fig. 3.7.

It has been shown recently [Auclair et al., 2003] that a better data assimilation scheme, or at least a better errorspace, would be obtained by EOFs built from ensemble runs obtained by perturbing variables of the simulation, such as the atmospheric forcings, the forcing field along the open boundary, the initial state (and in particular the position and intensity of the LPC), model parameters and the bathymetry. These considerations will be further developed in chapter 5. It should also be noted that the limited amount of "directions" that we use to build the error space, will probably not be able to correct all the errors encountered. In particular, the high variability in the Rhône river plume will probably not be resolved by our errorspace basis.

As mentioned before, we will not update $\mathbf{P}$ during the simulation. Because

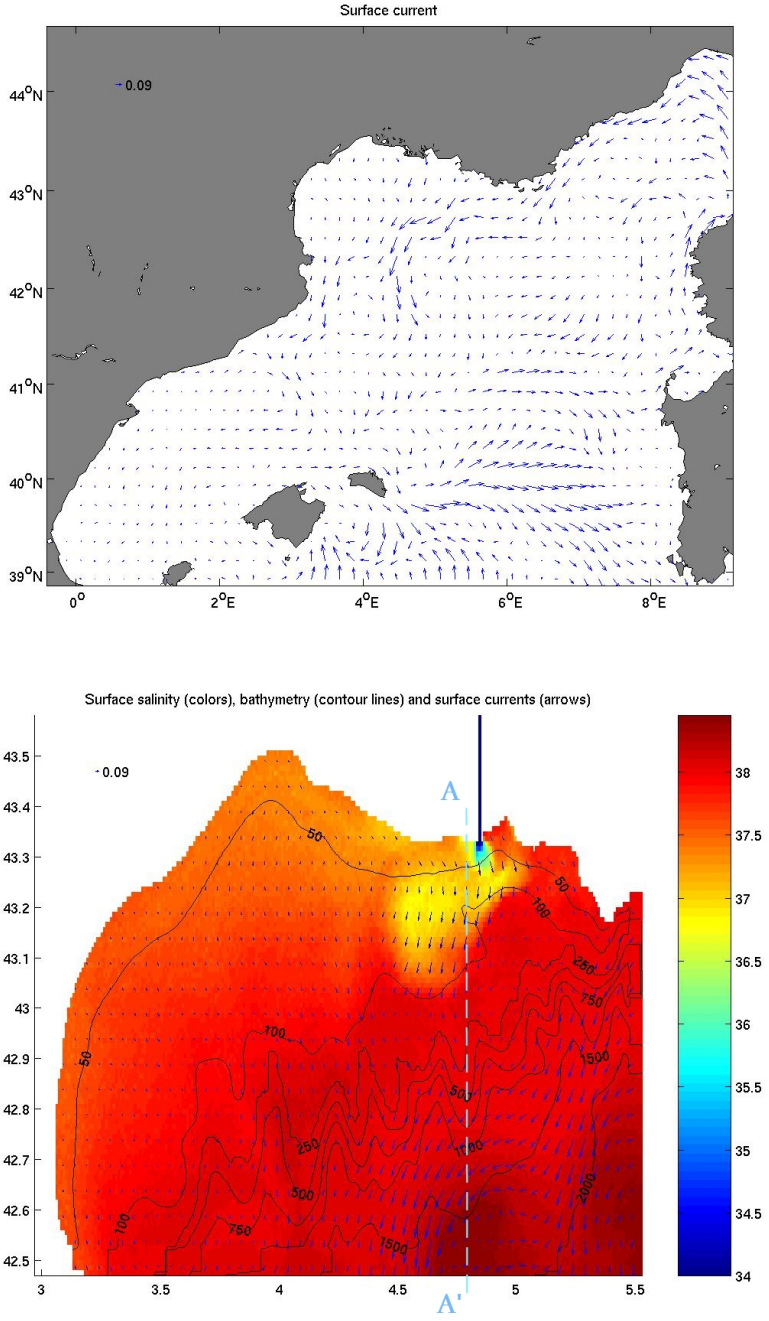

Figure 3.5: Example of model results: plots from 17 January 1998. **(a)** Surface current velocity [m.s^{-1}] in the intermediate grid. The cyclonic gyre and LPC current are clearly visible. **(b)** The arrows represent surface currents [m.s^{-1}], colors represent surface salinity [psu] and the contour lines represent isobaths [m]. The LPC follows the shelf break. Following Tramontane/Mistral wind bursts on 14 and 16 January, an intense current moves surface waters (and the Rhône plume in particular) away from the coastline.

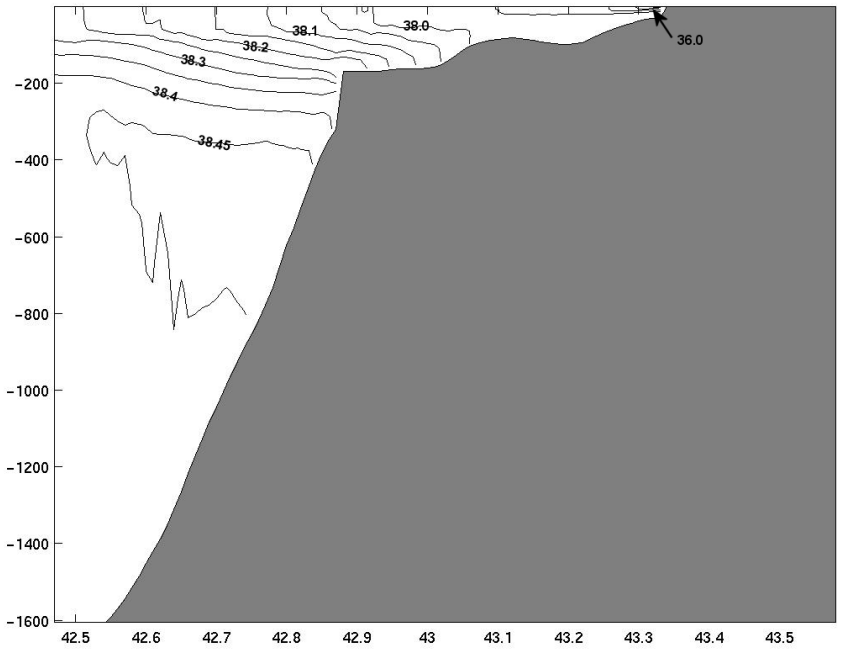

Figure 3.6: Salinity along the A′–A line in Fig. 3.5b. The salinity in the GoL is influenced by the Rhône. A LIW vein is clearly visible along the shelf break.

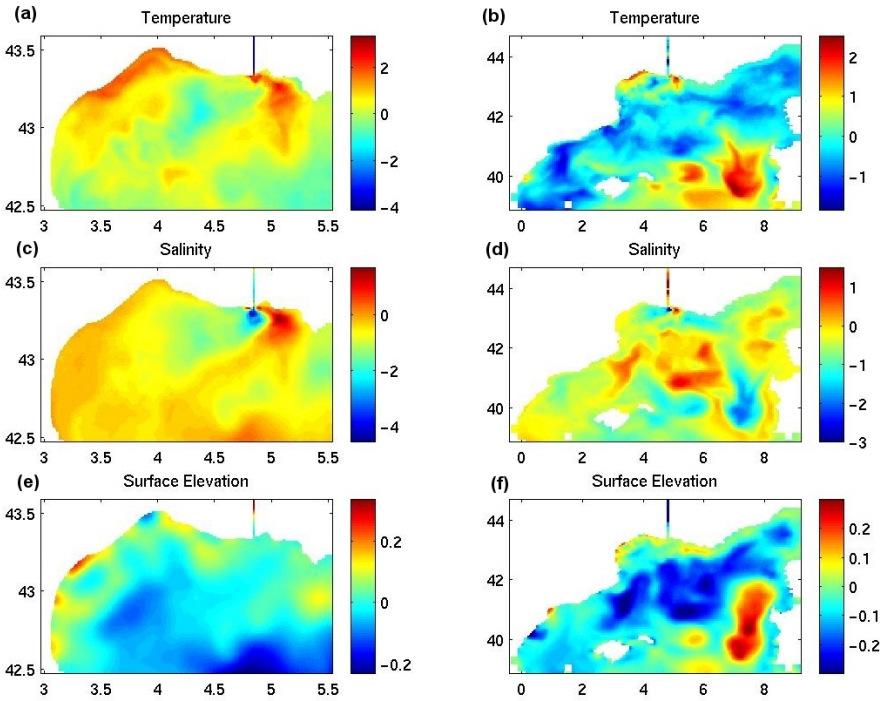

Figure 3.7: Surface plot of the temperature in °C **(a, b)**, surface salinity in psu **(c, d)** and elevation in m **(e, f)** parts, in the GoL (a, c, e) and in the Intermediate grid (b, d, f), of the first multigrid multivariate 3-D EOF, calculated after removing the temporal mean. The 1st EOF shows relatively large-scale structures, when EOFs of higher order represent structures with a smaller scale (not shown).

of the procedure followed to obtain it, **P** is representative for the period covering January and February. However, in the complete RRSQRT or SEEK filters, the error covariance is updated by the model (generally increasing it) and by the assimilation scheme (decreasing it). Thus, the ratio of the eigenvalues of the projection of **P** in the observation space, and the eigenvalues of **R** (the observations error covariance matrix), is modified during the simulation; this is not the case in our experiment as both **P** and **R** are constant in time.

Our model state vector contains temperature (T) and salinity (S) at each 3-D gridpoint, as well as sea surface elevation (η). Thus, whenever data is assimilated, this leads to corrections on the T, S, and η variables. These corrections are multiplied by a radial Gaussian function centered on the corresponding observation in order to limit the spatial extent of the correction that an observation can yield. In the present case, the Gaussians extent σ is put to 100 km. As mentioned in section 2.3, we notice that if **P** would accurately represent the model error covariance, this step would not be necessary; unphysical long-range correlations would not be present in the computed statistics.
Finally, the geostrophic velocity correction (corresponding to the T, S and η corrections) is computed and applied to the model horizontal velocity variables U and V. Indeed, our experience showed that it is often difficult to obtain the correct covariances between T, S and η fields on the one hand, and the velocity fields on the other hand. Therefore, we did not trust the assimilation procedure to update the velocity fields based on T and η observations. However, we realize that our procedure will not be able to correct the ageostrophic part of the error.

3.5 Twin experiment

In the present section, a twin experiment is set up. The model run shown in the previous section is used as the reference run. Another run, called the free run hereafter, will start from different initial conditions, being the ocean state from 23 January instead of 30 January 1998 (see Fig. 3.4c). This choice ensures that the initial conditions are physically balanced. The grids are also coherent with each other, since the reference run uses interactive nesting.

In the free run, some atmospheric forcings (wind velocity, air temperature and cloud coverage) are also modified in the following way. The real fields are decomposed as weighted sums of EOFs, which are obtained from a year time-series. It appears that 30 EOFs are needed to accurately describe the cloud coverage field (95% of variance), 10 to 20 for air temperature and about 80 for wind velocity (see Fig. 3.8). For all 4 fields however, we used 100 principal components, the computational cost being low. The modified fields are then obtained by multiplying the real weights by a random factor included in $[1-\epsilon, 1+\epsilon]$. This random factor is kept constant at all times in order to avoid large, unphysical variations in the forcing fields. Adequate fields were obtained by using ϵ equal to 0.4, 0.5 and 0.35, respectively for the wind velocity, cloud coverage and air temperature fields.

Every 2 days starting on 31 January, pseudo-observations occur and are assimilated in a perturbed run in order to bring it as close as possible to "reality"

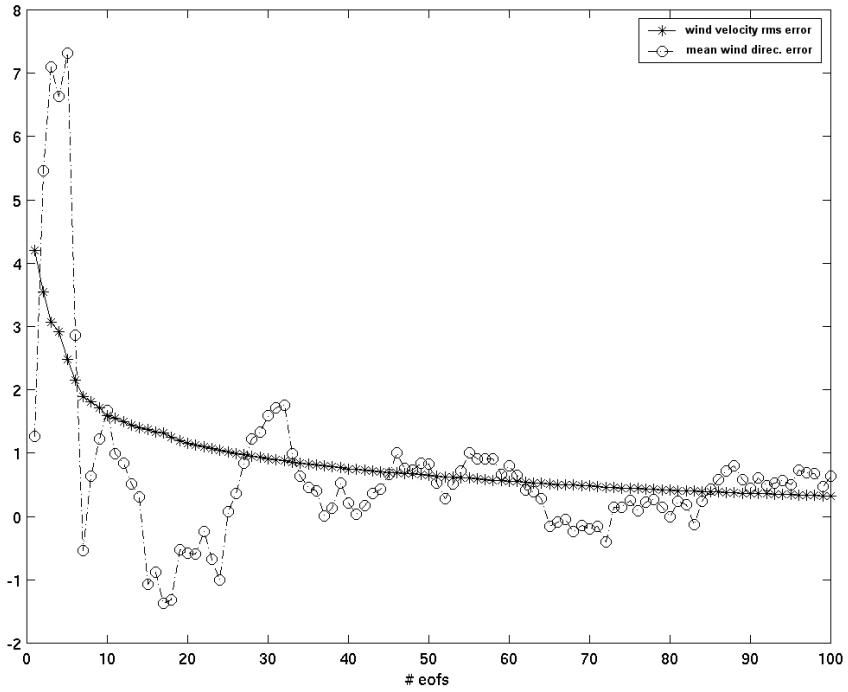

Figure 3.8: Error between the actual wind field, and its recomposition using the N first EOFs, as a function of N: rms error of the velocity [m/s] (stars), mean direction error [°] (circles). Means are calculated spatially over the whole Mediterranean grid.

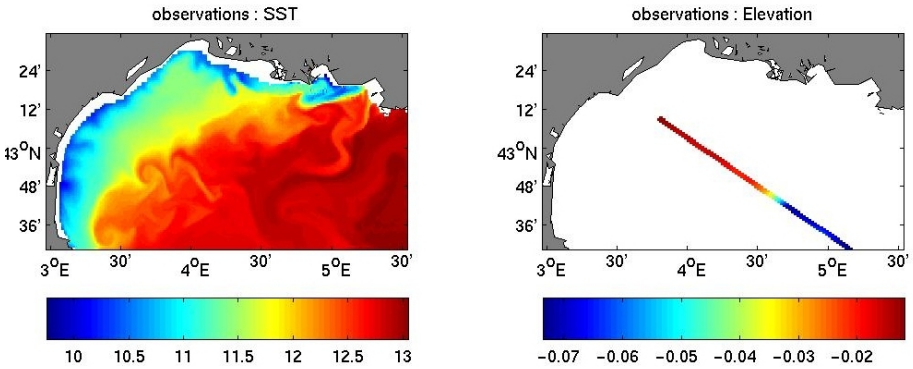

Figure 3.9: Pseudo-observations coming from the reference run: **(a)** Sea Surface Temperature [°C], **(b)** Sea surface elevation corresponding to a typical satellite track [m].

(assimilation cycles are represented by red stars in Fig. 3.4c). As pseudo-data, we assimilate Sea Surface Temperature (SST) over the whole GoL, and Sea Surface Height (SSH) over a pseudo-track in the Gulf of Lions. As an example, the observations to use during the first assimilation cycle (on 31 January) are shown in Fig. 3.9. The error covariance corresponding to these observations is represented by a diagonal matrix $\mathbf{R}$, with values equal to the square of 1°C (for temperature) and 0.4 m (for surface elevation).
In our simulations, we did not add random errors to observations, consistent with $\mathbf{R}$. It is known that this leads to a too strong reduction in variance. Adding noise to the observations would have added a supplementary stochastic variance to model SST and SLA, and thus to the resulting RMS curves. The RMS error measure is a filter, thus it already filters (part of) this stochastic signal. The study area (Gulf of Lions) being relatively small, there is a risk that the RMS still contains a part of the random signal, which could further be eliminated by an "ensemble" of twin experiments (prohibitive for us, due to numerical cost).
Furthermore, we represented the observations error matrix by a diagonal matrix. It is known that this matrix contains different contributions, at least some of them being not diagonal (i.e. the so-called representativity error). Thus, adding noise consistent with a diagonal matrix $\mathbf{R}$ would still be an approximation - even less crude than not adding noise at all. In any case, not adding noise to the observations is an error which we think affects all of the study cases (1 to 5) detailed in section 3.6 in a similar way, and does not change the qualitative comparison between those different cases.

3.6 Comparison of different setups

If observations are available in the area covered by the fine grid (the GoL), they could be assimilated in the coarse (and intermediate) grid, or in the fine grid, or both. Our purpose is to examine which setup is the most efficient in the context of an operational system where only one-way nesting is currently used.

Therefore, we define the following 4 study cases:

- case 1: the nesting is two-way, and data is assimilated in the fine grid. The **P** matrix is built from the part of the multigrid EOFs, corresponding to the fine grid. The corrections are automatically fed back to the other grids via the two-way nesting. As mentioned before, two-way has many advantages in- and outside the coastal domain; but it is not feasible in most operational configurations such as the one used in the "Mediterranean Forecasting System: Toward Environmental Predictions" (MFSTEP) project.

- case 2: will use one-way nesting, observations are still assimilated in the fine grid only.

- case 3: also uses one-way nesting, but data is assimilated in all the grids at the same time.

- case 4: one-way nesting still is used; data is assimilated in the coarse grid only. Let's note that since the observations assimilated in the coarse grid are still physically located in the Gulf of Lions, information should be transported to the intermediate-resolution model through the boundary conditions, and hereafter transported to the high-resolution model.

- case 5: moreover, we will define a fifth case, based on the work of Barth et al. [2006]. In this case, all 3 state vectors from the 3 grids are assembled in a unique state vector. Since the EOFs have been calculated over the 3 grids together too, perfect correlation is assured between data, located at the same physical points in different grids. Hence, an observation automatically yields coherent corrections in the 3 grids. If an evolutive assimilation scheme were to be used (i.e. the model would be used to update **P** in time), the model error covariance update equations would also yield "errorspace" feedback. Case 5 uses 2-way nesting.
  Vandenbulcke [2003] showed that in a 1 dimensional, linear case, and using a full rank evolutive Kalman filter, this approach leads to superior results. In fact, the "feedback" provided by the assimilation scheme and its full "multi-grid" error matrices, proved to be sufficient to adequately correct the different grids all at once, even when no explicit model feedback is used.

Let's note that in cases 2, 3 and 4, one-way nesting is used, so that discrepancies could appear between grids, ultimately leading to instabilities in the coastal model. Other instabilities may appear, as mentioned before, in the Rhône river plume after each assimilation cycle. Therefore, the defensive procedure based on the change in Brunt-Väisälä frequency, described in chapter 2, was implemented. In our study, this procedure yields masks where the corrections in the Rhône plume are almost systematically put to zero, confirming our suspicion that the errorbase would not be able to describe the plume variability (Fig. 3.10).

As an example, the correction to the surface temperature in the GoL, after the first assimilation cycle in case 5, is shown in Fig. 3.11. The corresponding assimilated data are shown in Fig. 3.9.

Figure 3.12 shows plots of the rms error between the reference run, and the

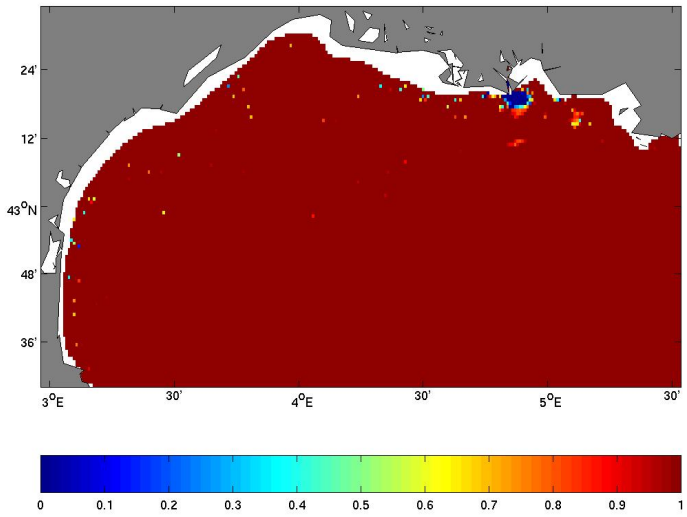

Figure 3.10: An example of a mask used to multiply the correction, and computed from the N^2 field after assimilation. The shown mask is calculated for the first assimilation cycle.

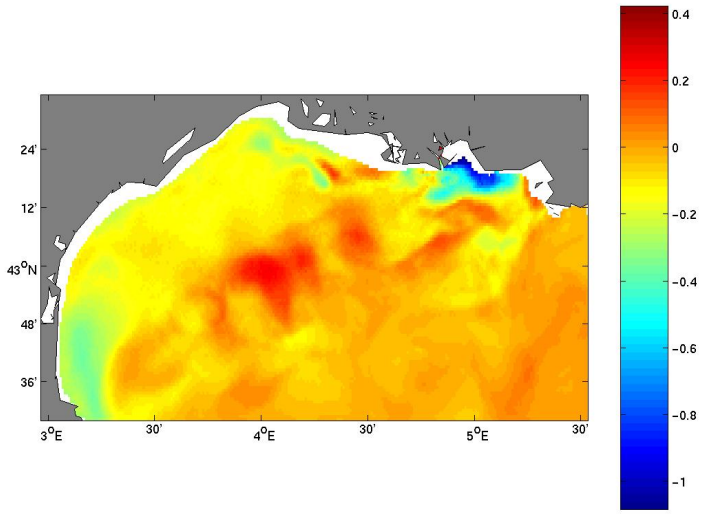

Figure 3.11: Correction yielded by the first assimilation cycle on the Sea Surface Temperature [°C].

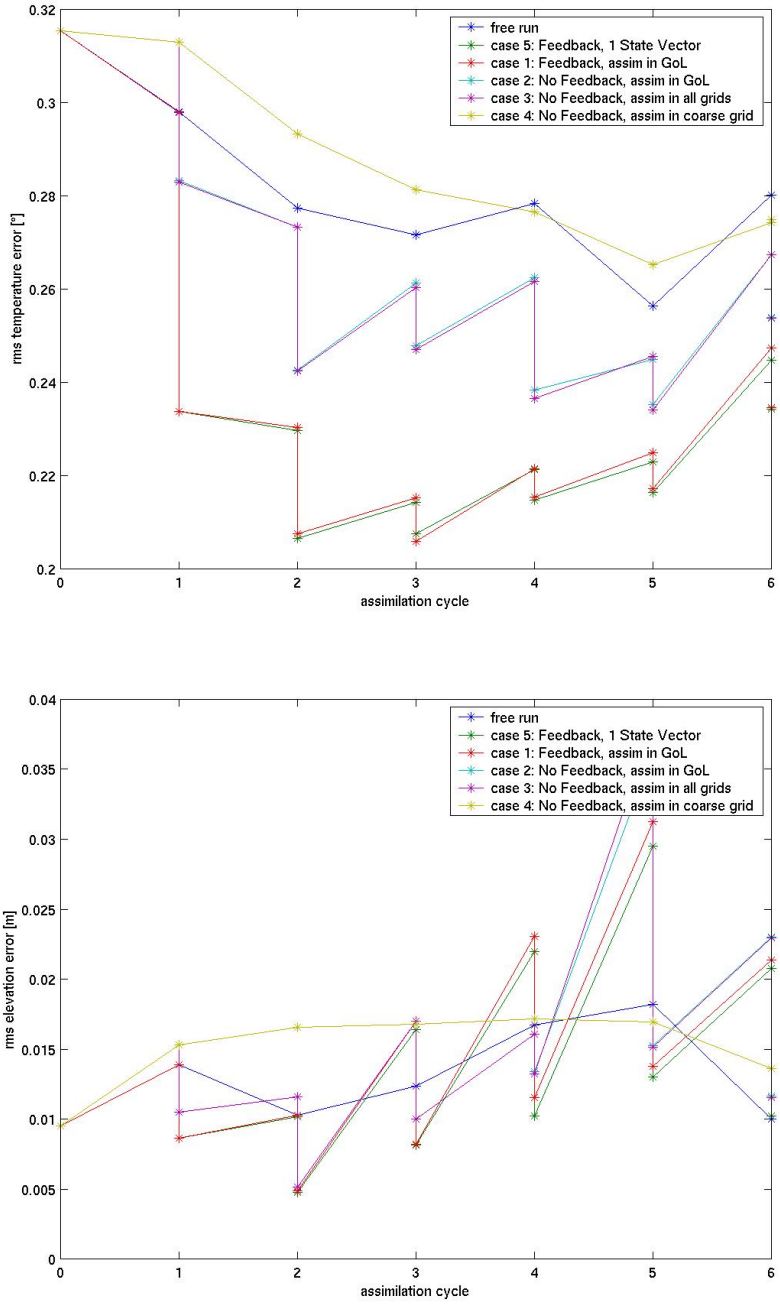

Figure 3.12: Evolution of the rms error in time, between the reference run and the perturbed runs, the latter being the free run (blue curve), case 1 (red curve), case 2 (turquoise curve), case 3 (purple curve), case 4 (yellow curve), and case 5 (green curve), showing **(a)** SST **(b)** Surface elevation. The stars represent assimilation cycles.

free run as well as the runs corresponding to cases 1 to 5. The averages are calculated over the grid points where observations are available. All the runs start from the same rms error, which is the rms between the "right" and alternative initial conditions. We observe that the SST rms error between the free run and the reference run shows a natural tendency to decrease with time. The reason for this is that the SST tends to an equilibrium with the atmosphere. It seems that perturbing the air temperature and cloud coverage leads to similar equilibrium states as the control run. Only during the last part of the simulation does the free run SST again departs from the control run SST. In the corresponding graphic for sea surface elevation, the rms error between free run and control run is increasing most of the time. This is because the principal variability source for SLA is the LPC current, with processes such as baroclinic instabilities and formation and propagation of eddies; they have a more chaotic nature.

It can be seen that case 1 (the simulation with interactive nesting) presents rms errors very close to case 5 (the simulation with feedback, but where corrections are automatically copied to the 3 grids, since all 3 grids are comprised in the state vector). However in our experiment, we kept the model error covariance matrix $\mathbf{P}$ constant. If it would be modified by the assimilation procedure, those modifications would still be consistent, in case 5, for covariances between identical physical points located in different grids. In the other cases, $\mathbf{P}$ covers only one grid, or three different $\mathbf{P}$ matrices are used in the three grids, with no assurance of consistent evolution between them. Hence, the assimilation of observations in subsequent assimilation cycles would probably be more accurate. Cases 1 and 5 both show much smaller errors than the 3 cases using one-way nesting. This indicates that the errors, which were not corrected in the global and intermediate model, are advected back to the local model through the boundary conditions. It can also be seen that the differences between cases 2 and 3 (assimilate in the GoL only, or in all 3 grids), are very small; indicating that the corrections made in the (intermediate and) global model are not sufficient to improve the boundary conditions, and hence, the fields in the high-resolution grid, significantly. In any way, their effect on the rms errors is much smaller than the assimilation of data in the local grid. Small corrections in the coarse grid are due to the fact that, in our twin experiment, these fields are already close to the observations, with respect to the error covariances. Unless the correction in the coarse grids becomes much more important (i.e. the fields in the coarse grid part further away from "reality"), it is thus not very useful to assimilate observations in the coarse grid. Finally, case 4 shows the largest rms errors of all. Corrections are only brought to the coarse grid; they need to propagate via the intermediate model, to the local model. There is no immediate change in the rms errors of the local model at assimilation times. But since the rms error of in case 4 is ultimately smaller than the one in the free run, it seems the information (assimilated in the global model) slowly arrives in the local model anyway. It can also be noted that the error on surface elevation is small (a few centimeters) at all times, even in the "free" run. During assimilation cycles, it is reduced by a factor close to 2, but in between each assimilation cycle, the model causes new errors of the same order as the correction, leading to an oscillating rms error curve. However, the maximum error remains of the order of 3 cm.

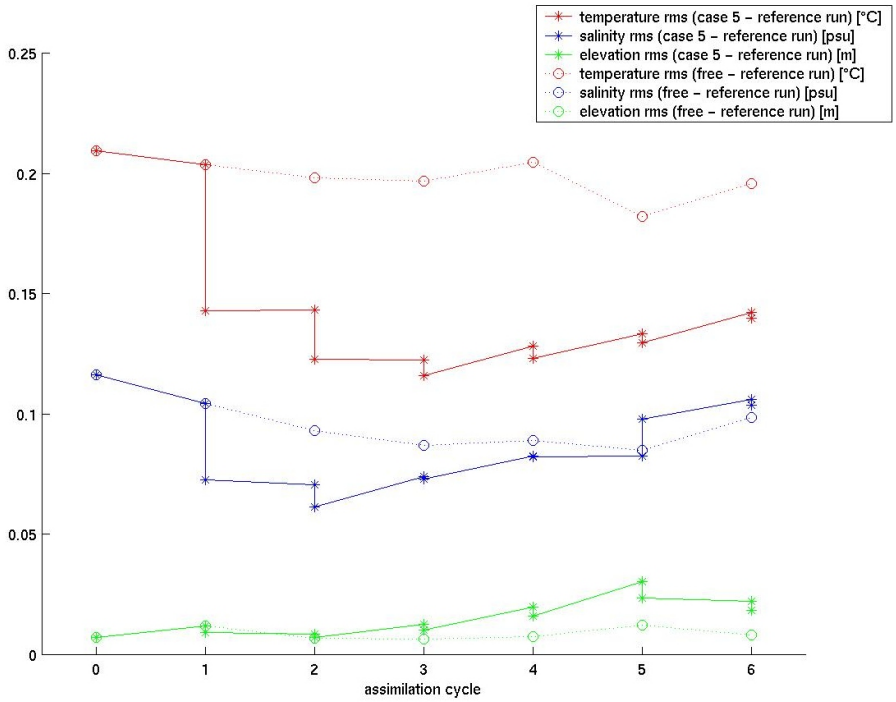

Figure 3.13: Evolution of the rms error in time, between the reference run and case 5 (full lines), and between the reference run and the free run (dotted lines). The rms error is calculated over the entire 3-D field of temperature (blue curves), the entire 3-D field of salinity (green curves), the whole surface elevation field (red curves). The stars represent assimilation cycles.

It is interesting to examine the rms errors between the reference run and the other cases, with the averages calculated over other location than those where observations are available (of course, this is only possible in the framework of a twin experiment). In particular, we show the rms curve calculated over the whole 3-D grid in Fig. 3.13. As example, we used case 5. As can be seen, the data assimilation cycles reduce the overall rms error, even in points where no data was assimilated. In fact, the procedure also reduces rms errors on the salinity variable, although no salinity observations are available. This shows that the errorspace correctly represents the statistical covariance between the salinity and the observed variables. Only during the fifth assimilation cycle (on 8 February 1998 at midnight), the correction on SST and surface elevation along a satellite track, statistically yields a salinity correction which actually increases the salinity rms error. It can be seen that the surface elevation mismatch is, for that assimilation cycle, very important (almost 3 times larger than during the first assimilation cycle). Finally, it can also be observed that the temperature and salinity rms errors are lower than in the free run. However, the rms error corresponding to sea surface elevation is higher in case 5 than in the free run. This indicates that we have some bad estimations in the part of the covariance matrix related to the surface elevation variable, due to the approximations listed

in Sect. 2.3.

3.7 Conclusion

We have studied the impact of data assimilation in a nested hydrodynamic model, assessing the question whether the available observations should be assimilated in the high-resolution, local grid, or in the coarse-resolution grid. Using twin-experiments with different test-cases, we obtained some general conclusions:

- two-way nesting moves the model toward "reality" faster than one-way nesting. Indeed, when new information is added in any grid, it is transmitted to all the other grids; in the one-way nesting paradigm, information only goes from low-resolution to high-resolution grids.
- the assimilation filter that we used, yields satisfactory results, even though we chose a low errorspace dimension (20). Only in places with very high variability (such as a river plume or some points along the coastlines), the filter could not capture the model variability and hence, we artificially diminished the correction in those points, because it was expected to be incorrect.
- using multi-variable state-vectors, we have corrected all variables by observing only some of them.
- using a full 3-D matrix, we have also corrected variables over the whole grid, although only surface variables were available.

Supposing that the global model is approximately correct (i.e. it feeds the local model with boundary conditions "not too far" from reality), we showed that the high-resolution model is better corrected when available observations are assimilated immediately in that grid, rather than to assimilate them in the coarse-resolution grid and transport the information in the local model via boundary conditions. And if the data is assimilated in the high-resolution model, it is then of little use to also have assimilated it in the global model (this is even less useful when using interactive nesting).

Perspectives to improve the nested assimilation system mainly concern the error covariance matrices. Indeed, a number of approximations were used in our assimilation procedure. In particular, it should be beneficial to replace the static model error covariance matrix built with EOFs obtained from a historical run, with a matrix built from an ensemble run, which would also be evolved in time with the model. Updating the errorspace allows for it to represent the "errors of the day", and hopefully leads to better corrections in subsequent assimilation cycles. This should particularly improve the sea surface elevation variable. Even when updating the errorspace, we mentioned above that it is possible, using interactive nesting, to maintain complete consistency between the grids in the state vector as well as in the model errorspace.

Chapter 4

Upscaling

What goes up, must come down

4.1 Rationale

Recent ocean forecast systems do provide end users with predictions at various scales, ranging from global and basin wide estimates to regional or even coastal fine-scaled predictions. The widespread technique used to reach all these scales is model nesting. In many operational cases however, and due to practical considerations, while the coarse-resolution model provides initial conditions and/or boundary conditions to the fine-resolution model, the latter cannot provide the former with any feedback; only so-called one-way nesting is applied. Indeed, two-way nesting, where the nested model returns values to the coarse model, supposes that the parent and child models are sufficiently similar in their implementation to allow data to be exchanged. It also requires the nested models to be run simultaneously, with large data transfer between them. This practical consideration can be especially limitative when multiple models, covering different areas, are nested into a coarse model. A typical example of this configuration is the one used in the MFSTEP project covering the Mediterranean Sea (see Pinardi et al. [2003], `http://www.bo.ingv.it/mfstep`, or `http://www.moon-oceanforecasting.eu` for the operational MOON system). An oceanic general circulation model (OGCM) covering the Mediterranean Sea is run at INGV (Bologna, Italy). Its forecasts are used to provide initial conditions and boundary conditions to four regional models, covering the North-Western Mediterranean Sea region, the Sicily Strait region, the Adriatic Sea and the Eastern Mediterranean Sea. Those models are run in France, Malta, Italy and Greece respectively, and use different numerical codes and different forcings. It is thus impossible, in this configuration, to use two-way nesting.

However, it has been shown that nested models who do provide feedback (two-way nesting) result in more realistic predictions, even outside of the area covered by the fine grid (see e.g. Barth et al. [2005] for a recent example). Feedback also minimizes the chance that discrepancies appear at the open boundary of the high-resolution grid. Those discrepancies could ultimately lead to instabilities in the high-resolution model. Furthermore, it has been shown that observations are best assimilated in the regional models rather than in the OGCM; their impact will improve the OGCM if two-way nesting is applied (see chapter 3 or Vandenbulcke et al., 2006). Thus, it is pitiful not to use available high-resolution information (nested models or observations) to its best availabilities in order improve the OGCM.

The technique of data assimilation is not limited to the use of (real) observations of physical variables into a model. For example, Onken et al. [2005] used assimilation as a substitute for one-way nesting in a cascade of nested models. Another example is shown in Álvarez et al. [2000]; a statistical model is used to predict sea surface temperature, and these forecasts are then used as pseudo-observations and assimilated in a hydrodynamic model.

In the present study, we use data assimilation as an alternative method to classic model feedback. Our aim is to bypass the need to run models simultaneously, and the large data transfers at each time step. To achieve this, outputs from high-resolution, local models are assimilated as pseudo-observations in a basin-wide model. Usually, in recent forecasting systems, a data assimilation

scheme is already present in the code, and the pseudo-observations are readily assimilated. The only practical requirement of our method is thus that some outputs from the high-resolution model must be transferred to the coarse model. The use of data assimilation as a substitute for the feedback in nested grids is, to our knowledge, new; we will refer to it as "upscaling".

Upscaling can be seen, from another point of view, as the use of nested models (hopefully more accurate, in their domain, than the OGCM) as a substitute for ever too sparse real observations. For example, in the North Atlantic, Guinehut et al. [2002, 2004] showed that a coverage of the sea with a 3°-resolution grid of Argo floats allows to effectively represent the large scales. Using a 5° array reduces the precision of the estimated fields two times, while a 1° array increases the precision of the estimated fields 4 times. Combining these data with satellite observations yields good estimates of the ocean state, both large scale and instantaneous. However, at present time, the coverage of Argo floats in the Mediterranean is sparse, and hence, pseudo-observations from high-resolution, regional models, could be used as a next-to-best complement to real observations.

Finally, upscaling can be understood as a complement to variational initialisation techniques such as the one presented in Auclair et al. [2000, 2001] and called VIFOP. The point of these methods is that small-scale structures present in coastal models are lost whenever the model is crudely (re)-initialized by fields interpolated from the OGCM. Worse, the interpolated fields are physically unbalanced with respect to the coastal physics. VIFOP builds a new initial condition as a combination of a background field (the regional field interpolated on the high resolution grid) and some forcing vectors; it is optimal with respect to specified error covariance matrices, and the new initial condition is physically balanced in the coastal model. However, the calculation of the initial field might be a lengthy process. For example, in the MFSTEP project, initializing the "ALERMO" nested model covering the Eastern Mediterranean Sea takes approximately as long as the forecast for a week (Nikos Skliris, private communication). If upscaling is used to improve the OGCM fields and accord them with the coastal model, the need for a variational initialization technique is diminished.

In the following section, we will present our model set-up. Sections 4.3 and 4.4 then show the result of some twin experiments where respectively surface fields and profiles are assimilated. A tentative of real experiment is presented in section 4.5. Some conclusions are given in section 4.6.

4.2 Study area

The hydrodynamic model, data assimilation scheme and nesting procedure have been described in chapter 2. The bathymetry, forcing fields and initial conditions are also the same as those used in chapter 3. A coarse-resolution (1/4 degree) grid covers the whole Mediterranean Sea. Herein, a model is nested which covers the part of the basin, to the North-West of Corsica and Sardinia, with a resolution of 1/20 degree. The implementation of these 2 grids is the

Simulation	*Perturbation*
Simulation 1	Reference simulation using two-way nesting
Simulation 2	Simulation using only one-way nesting
Simulation 3	Simulation using only one-way nesting, but with assimilation in the parent model of outputs from the nested model

Table 4.1: Synthesis of the 3 simulations in the twin experiments. In twin experiment 1, sea surface temperature and elevation is assimilated; in twin experiment 2, 26 T-S profiles are assimilated.

same as in the previous chapter (where a third grid was also implemented). They are shown again in Fig. 4.1.
In the one-way nesting approach, the coarse grid model is interpolated on the boundary of the fine grid to provide boundary conditions. In the two-way nesting approach, the fine model results are averaged over each coarse grid cell in the overlapping area, where they replace the coarse model outputs.
The results of the free model run in those grids have been presented earlier, as well as in Vandenbulcke et al. [2006].

4.3 Twin experiment 1

We will perform 3 simulations, all starting from the same initial conditions, and using the same forcings. In the first one, which will be used as reference, both models are run simultaneously and two-way nesting is applied: at each timestep, in the area covered by both models, the values of the coarse model are replaced with averages of the corresponding gridpoints of the high-resolution model. A sponge-layer is used. As mentioned before, it has been shown that this set-up provides more accurate predictions, in the high-resolution area as well as in its neighborhood.
In the second simulation, single-way nesting is used; nothing is implemented, that could provide feedback. Unfortunately, this is the case often found in operational nested forecast systems.
In the third simulation, only one-way nesting is applied, but outputs from the high-resolution model are assimilated in the coarse-resolution model every day, except the first day. Indeed, some time must be given to the one-way nested model to part from the two-way nested model. The three simulations are summarized in Table 4.1. A nice feature of assimilating data back from the fine model to the coarse model, is that we can freely choose which variables we assimilate. All model variables could, in principle, be used, at every 3D gridpoint, although such a massive amount of pseudo-observations is not necessarily useful. In this section, we limited ourselves to assimilate the complete sea surface temperature and elevation fields. Let's note that in the context of this twin experiment, both the OGCM and the regional model represent the same phenomena, so that sea elevations could be assimilated without prior treatment. The observations error covariance matrix $\mathbf{R}$ is approximated by a diagonal matrix, with an appropriate choice for the standard deviation as 0.1°C (for the temperature part) and 5 cm (for the sea surface height part).

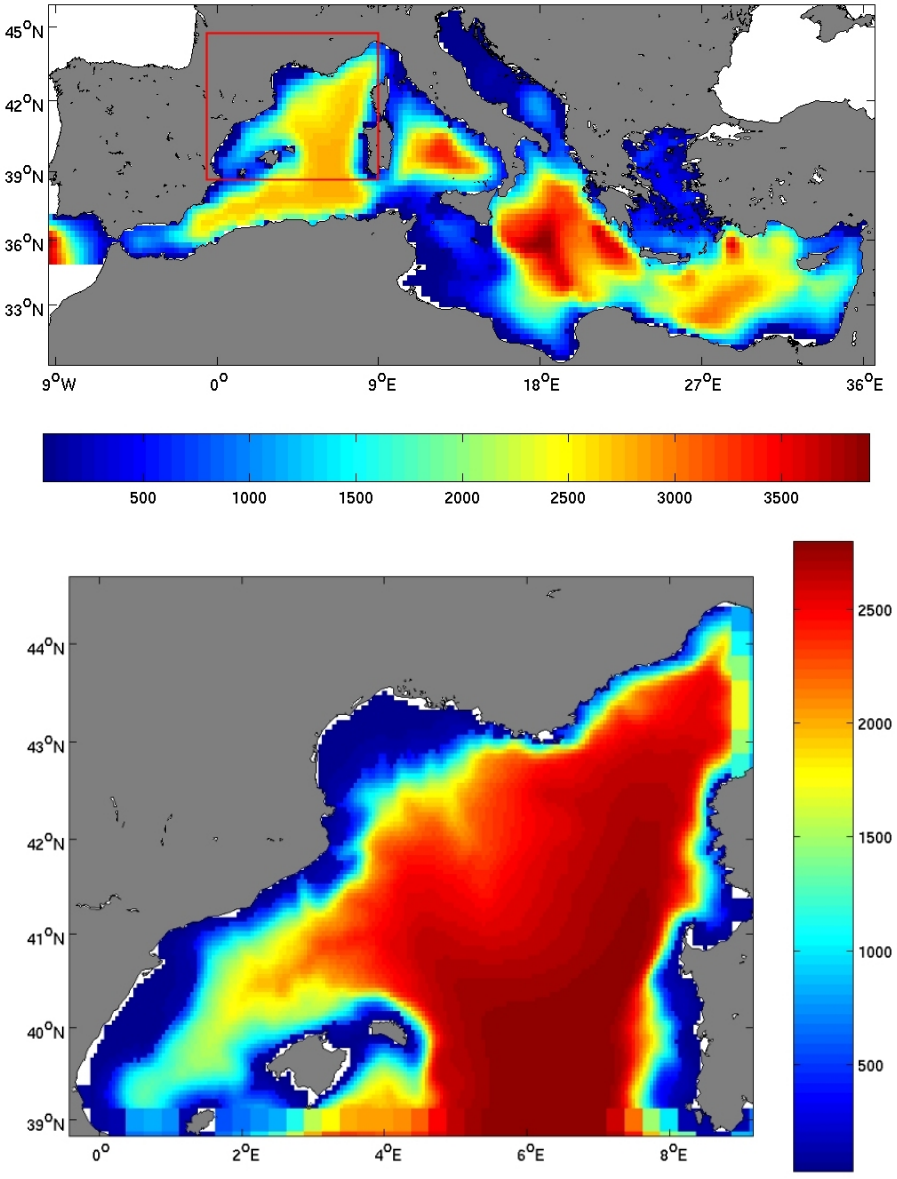

Figure 4.1: (a) The coarse-resolution grid, with the high-resolution grids location indicated by a red rectangle. (b) the high-resolution grid covering the North-Western part of the Mediterranean Sea.

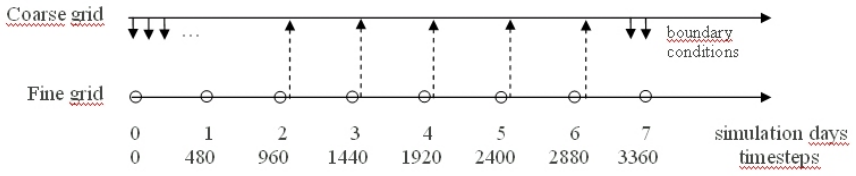

Figure 4.2: Time sequence of the experiment. The upper time-line represents the coarse-resolution model, the lower one the high-resolution model. The small arrows represent the boundary conditions provided to the fine model at every timestep. In the reference simulation, they also represent the feedback. Dashed arrows represent the assimilation cycles in simulation 3.

The time sequence of model runs and data assimilation cycles is summarized in Fig. 4.2. In an operational case such as the MFSTEP program presented above, this would imply the following sequence. First, the OGCM is ran until day 2, followed by the regional model (also until day 2), using the OGCM forecast for the boundary conditions. Then, the OGCM assimilates the forecast of the regional model, and continues its forecast. Finally, the regional model also continues from day 2, with updated boundary conditions. In a more realistic setup (requiring only one weekly data transfer from regional model to OGCM), the OGCM would forecast a whole week, followed by the regional model. Afterwards, the OGCM could be rerun, assimilating forecasts from the regional model; and finally, the regional model would also be rerun, using the new boundary conditions. This implies to run both models twice.

In order to assess the utility of our method, we will compare the outputs of simulations 2 and 3 versus the output of the reference run (simulation 1). Rms errors are calculated every day, just before the fine-resolution data is assimilated in the coarse model. Hence, at these times, the data can be considered independent. We will calculate rms errors for temperature over the whole depth of the basin (even though only surface data is assimilated), and rms errors of sea surface height.

After 2 days of simulation, the models with (reference run) and without (simulations 2 and 3) feedback have departed. Fig. 4.3a shows the difference in sea surface temperature between them. The errors are mainly located in a patch, where an error of approximately 0.4°C appears, and along the coast. It must now be emphasized that the chosen assimilation method (fixed SEEK filter) only corrects the model in the direction of the *a priori* chosen EOFs, which cover the whole Mediterranean Sea. As noted before, only r EOFs were used, and it is well known that the first EOFs, which represent the highest variability in the basin, usually represent large-scale features. Thus, if we must correct such a small feature as the patch shown in Fig. 4.3a, a fair amount of EOFs will probably be needed. It appeared that using 40 EOFs would almost not correct the model; while 60 EOFs brought a very good correction, as shown in Fig. 4.3b. The very small-scale errors along the coast and around the islands could not be

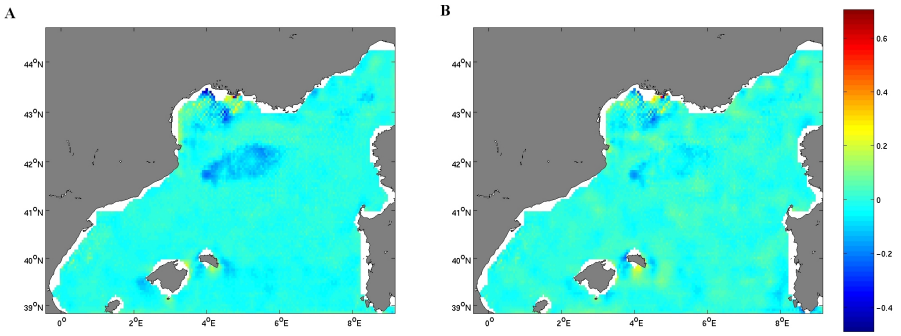

Figure 4.3: (a) Difference in SST between the simulation without and with model feedback, and (b) this difference, shown on the fine-resolution grid, after data assimilation in the coarse-resolution grid, using r=60.

corrected at all, even with 60 EOFs.

Fig. 4.4 shows the daily rms errors between run 3 and the reference run, for sea surface elevation, calculated each day just before the assimilation cycle. The rms errors between run 2 and the reference run, calculated at the same time, are also shown. A lower rms error in the "run 3" curve, means that the previous-day assimilation cycle had a beneficial effect. The figure clearly shows the positive impact of the assimilation. At the end of the simulation, approximately half of the error, due to not using model feedback between nested models, is corrected.

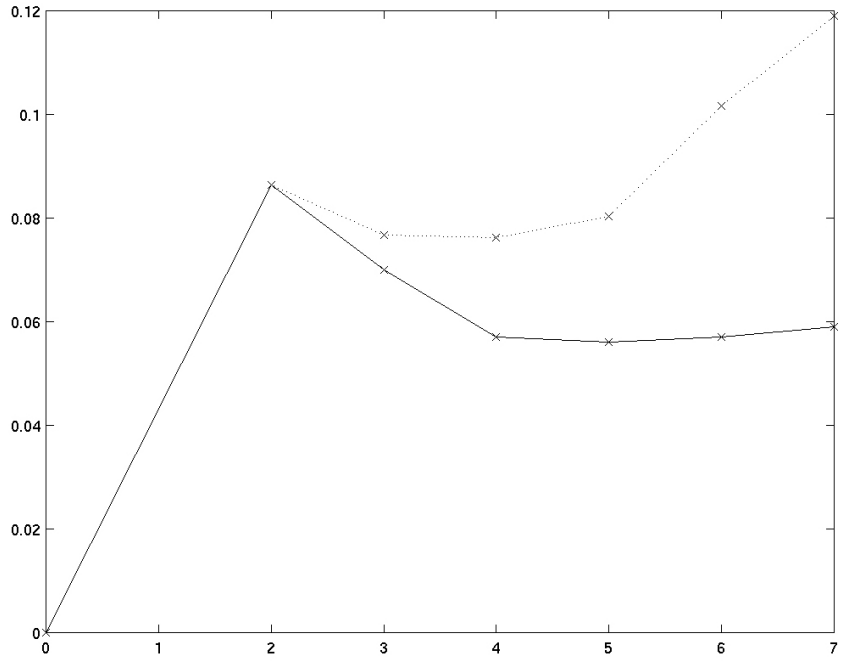

Figure 4.4: Daily RMS errors of sea surface height, in meters, between simulation 2 and the reference run (dotted line), and simulation 3 and the reference run (solid line)

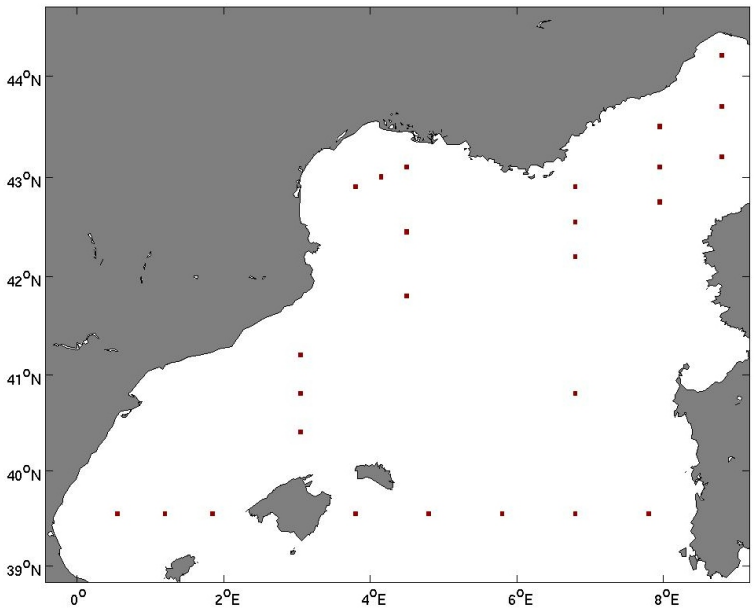

Figure 4.5: Location of the pseudo-profiles which are extracted from the regional model and assimilated in the basin-scale model.

4.4 Twin experiment 2

As explained before, any synthetic observation could in principle be extracted from the regional model to be assimilated in the basin-scale model. In this section, we will perform a new twin experiment using 3 simulations (summarized in Table 4.1). Instead of assimilating surface fields, we will now use 26 pseudo-profiles of temperature and salinity. Their positions are shown in Fig. 4.5; they are chosen in order to constrain the incoming currents, mainly the Northern Current (see chapters 3 and 5). Sampling tools exist to further optimize these positions (see e.g. Rixen et al., 2003aa, Rixen et al., 2003bb, Yilmaz et al., 2006). Moreover, the following causes of differences between simulations 1 and 2 (or 3) are also implemented. We will use different atmospheric forcings in the 2 grids: the air temperature, cloud coverage and wind velocity fields are perturbed in the same way as explained in chapter 3. The relaxation towards the climatology is also suppressed in order for the differences to grow faster. Finally, we will wait long enough before starting to assimilate: only after 20 days, the first pseudo-profiles is extracted and used. The model error covariance matrix used in the data assimilation scheme is again built with 60 EOFs.

The resulting rms errors are shown in Fig. 4.6 for the error calculated in the high-resolution grid, on the profiles themselves, and in Figs. 4.7 to 4.9 for the error calculated over the whole basin. As the mean errors are calculated just before the assimilation cycles in the OGCM grid, they reflect the effect of previous days assimilation cycles, whose effect have been transported to the regional model via the (updated) boundary conditions.

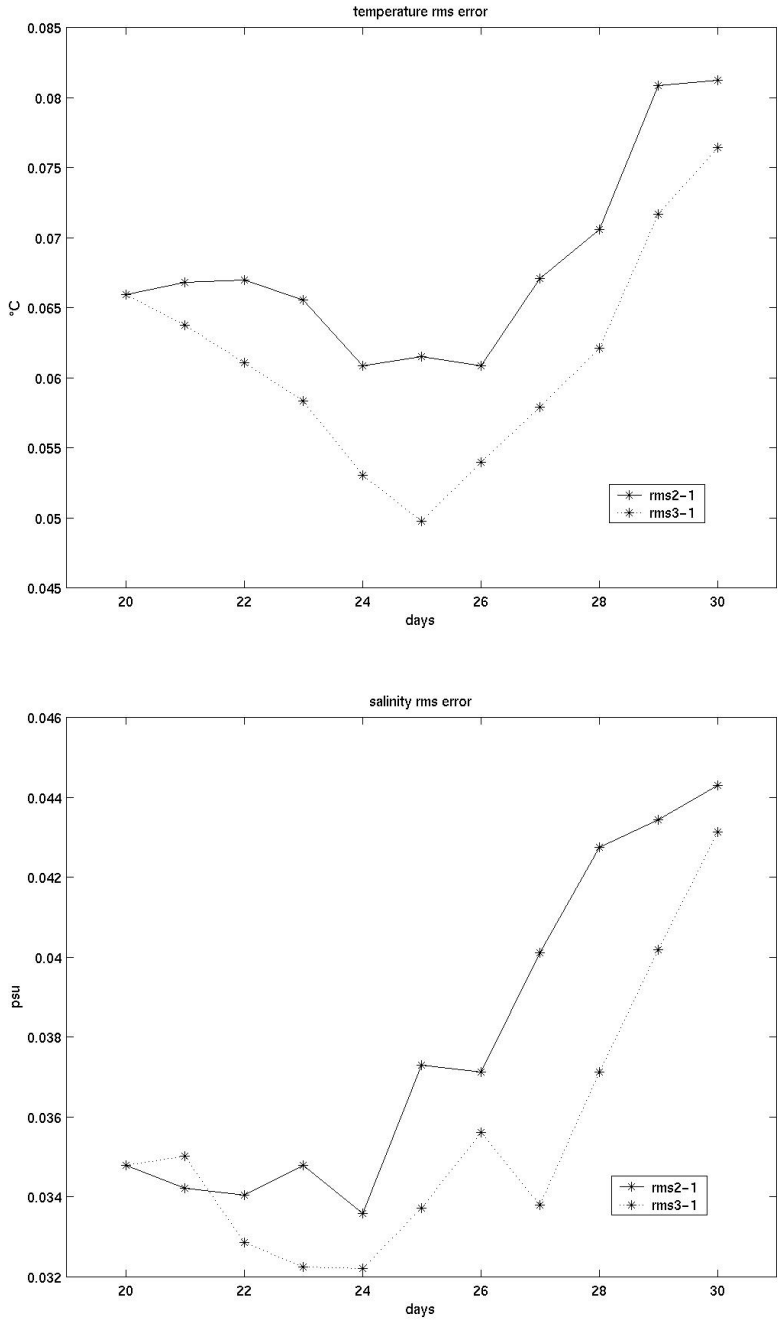

Figure 4.6: Rms error calculated in the regional grid on the pseudo-profiles, just before the assimilation of pseudo-profile data in the OGCM grid: (a) temperature [°C], (b) salinity.

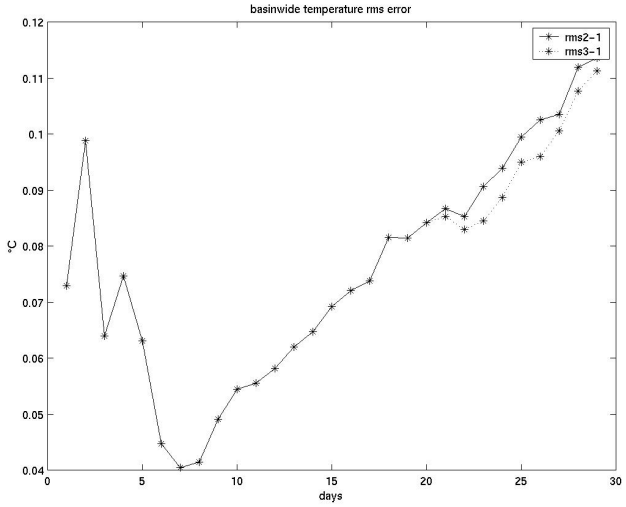

Figure 4.7: Temperature rms error [°C] calculated in the regional grid on the whole grid, just before the assimilation of pseudo-profile data in the OGCM grid. The rms error is represented for simulation 2, without assimilation (plain line) and simulation 3, with "upscaling" (dotted line).

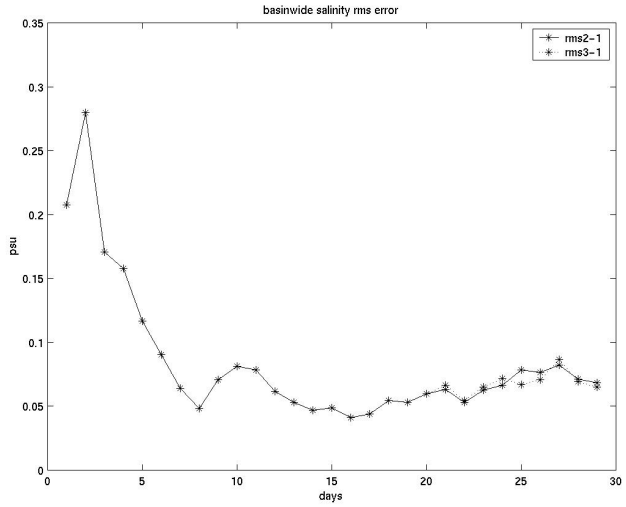

Figure 4.8: Idem, for the salinity [psu]

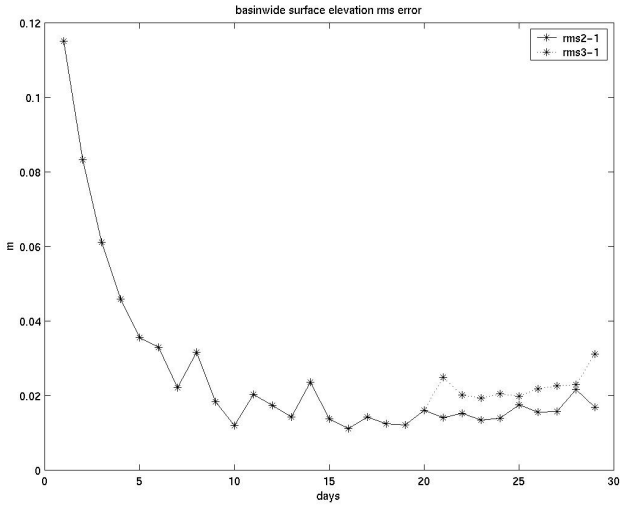

Figure 4.9: Idem, for the surface elevation [m].

We observe that, with the chosen set-up, the upscaling procedure is able to correct both the observed variables of temperature and salinity at the given locations, and even a day after the assimilation, this still has a positive impact on these variables at these locations, albeit this impact is relatively small. When looking at the errors computed over the whole regional basin however, the impact of the 26 profiles is rather poor. The sea temperature is, on average, slightly corrected, but there is almost no impact on salinity, and worse, the surface elevation presents larger errors than without upscaling. Very similar results were obtained with different error bases, using 40 or 60 EOFs, built on instant ocean scales (at midnight) or daily-averaged ocean states.
These results may be explained as follows. The regional model and the OGCM hopefully both represent accurately the large-scale circulation, which is precisely the one that profile-type observations best corrects. The benefit of regional models comes from the better representation of meso-scale activities, which are known to be best corrected with numerous quasi-synoptic observations, not by isolated profiles. Indeed, the differences between the simulations without and with nesting feedback are computed in the regional model, and have small sizes (Fig. 4.10). The corresponding correction brought in the OGCM is shown in Fig. 4.11. Given the resolution of the OGCM, these corrections present approximatively the smallest possible size, particularly in the northern half of the Ligurian Sea. It should be noted *en passant* that the correction naturally spreads out a little further than the overlapping domain.

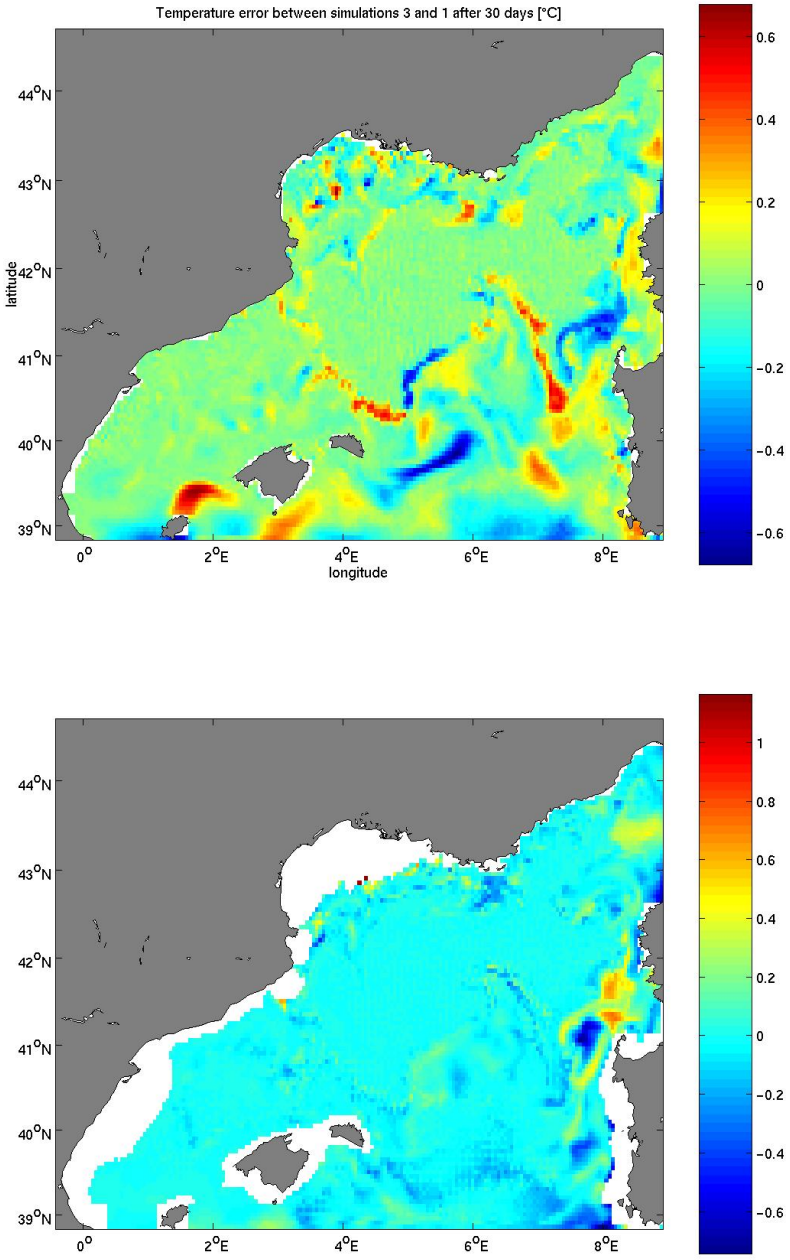

Figure 4.10: Temperature difference between simulation 3 and simulation 1 after 30 days, (a) on the surface layer, (b) in a deep layer, k=10 (k goes from 1 close to the bottom to 31 close to the surface).

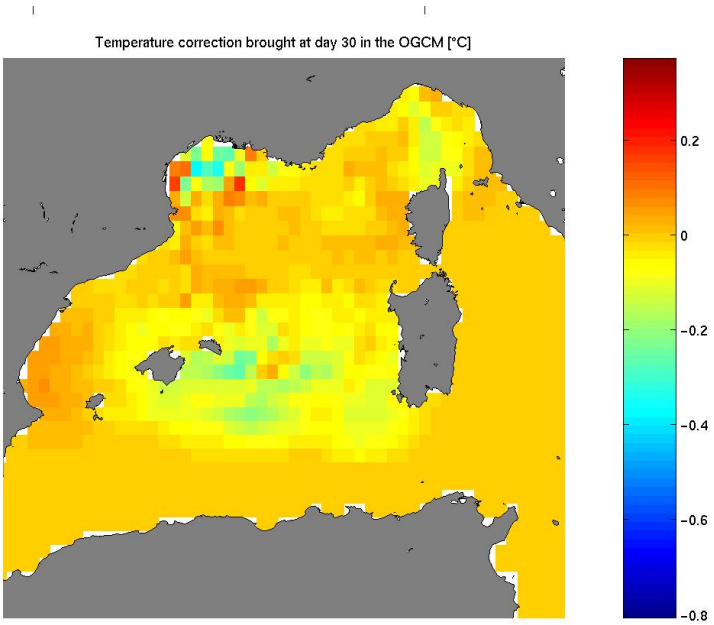

Figure 4.11: Temperature correction brought by the assimilation in the OGCM after 30 days (zoom).

4.5 Real experiment

After performing the successful twin experiments described in the previous sections, we wanted to try out our method in a real experiment, in the MFSTEP system (grids see Fig. 4.12). The test simulations would concern october 2005. The MFS OGCM is ran at the *Istituto Nazionale di Geofisica e Vulcanologia* of Bologna, and performs a forecast for the week from 4 to 10 october. Three sub-domains (North-Western Mediterranean (NWM), Alermo (for the Levantine basin) and Adricosm (for the Adriatic)) then also perform a forecast for the same week, using the MFS outputs for their boundary conditions. We mentioned before that those 3 models are set up differently. The first one, NWM, uses the Symphonie 3D hydrodynamic model with a horizontal resolution of 3 km (about $1/30°$) and 40 vertical σ layers. It is forced with Aladin atmospheric fields from Meteo-France. It is restarted with a 10-days hindcast every week; the initial condition is built with VIFOP from previous outputs and MFS outputs. Alermo uses the POM hydrodynamic model, and Skiron meteorological forcings from the University of Athens. The horizontal resolution is $1/20°$; there are 25 σ layers. It is also initialized with VIFOP from previous outputs and MFS fields, but does not perform any hindcast. Adricosm finally also uses the POM model, with 5 km horizontal resolution and 21 σ layers, but it is never reinitialized. Let's note that a fourth regional model covering the Sicily basin, as well as different local models embedded in the 4 regional models, were not considered for our experiment.

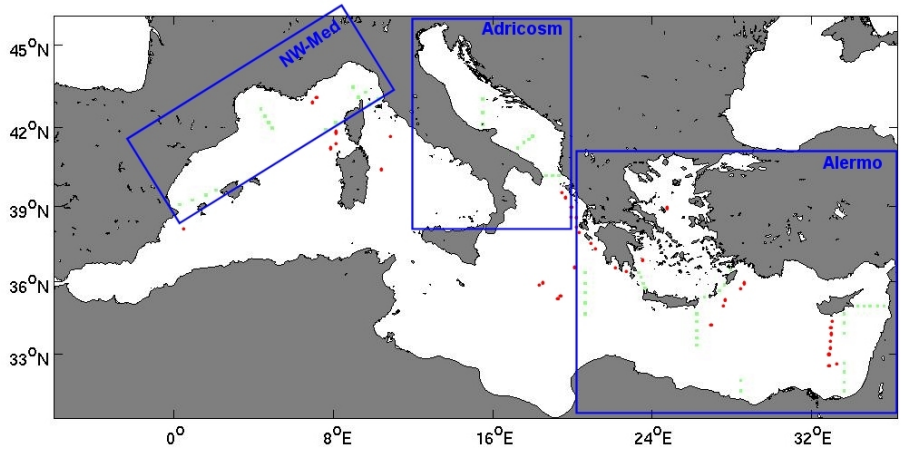

Figure 4.12: Data assimilated in the OGCM during second week of the hindcast from 27 september to 10 october 2005. The red dots indicate the location of real data available; the green dots represent synthetic data extracted from the previous regional model forecasts.

We collected the forecasts of the 3 regional models, and, in collaboration with INGV, decided to extract some pseudo-profiles. Their locations are shown in Fig. 4.12. The following week, when the OGCM starts its next forecast for the week from 11 to 18 october, it first performs a 2-week hindcast, during which real profiles (red dots in the figure) as well as these pseudo-profiles (green dots) would be assimilated. The observational error would be different for real and synthetic data. In the diagonal error covariance matrix $\mathbf{R}$, real temperature observations have an estimated error of 0.1°, salinity errors are taken as 0.03. Synthetic data would have larger expected errors, such as 0.5° and 0.1 respectively. Of course, these parameters, as well as the profiles location, could be fine-tuned later, as experience is acquired. Then, the procedure would be repeated for the next week, and the performance of this new system could be assessed by comparing independent (real) observations with the new model forecasts and with the ones obtained without "upscaling". Unfortunately, the OGCM could not be re-run for our test by lack of manpower, and hence the test was aborted.

4.6 Conclusion

Traditional feedback in nested models, from regional models to an OGCM, is often not implemented for practical reasons (large data transfers and the need for running parent and nested model simultaneously). In this chapter, we tried to replace it by extracting pseudo-observations from the regional model, and assimilating these in the OGCM as if they were observed data. Depending on the set-up and the data chosen as pseudo-observations, we showed that this upscaling procedure might indeed make the passively nested system look more like an interactively nested system, without needing large data transfers.

As the large-scale circulation is (hopefully) relatively well represented in the OGCM, the main improvements brought by the regional model are the meso-scale features of the circulation as well as meso-scale corrections to the main circulation. These are thus the corrections that could be transferred from the regional model into the OGCM by the upscaling procedure. However, assimilating meso-scale data in a large-scale model is not an easy task. In particular, we showed that when using a reduced-rank assimilation scheme, an important amount of directions are needed in the error space. If this would not be the case, a simpler assimilation method, e.g. where the model error matrix is the product of a vertical and horizontal correlation, might yield better results than full 3D error modes.

In the case where the dimension of the error covariance is high enough, we showed that assimilating surface fields can reduce by a factor 2 the "error" caused by the absence of nesting feedback. We could not obtain the same results when assimilating pseudo-profiles. Indeed, the latter are known to efficiently constrain the large scales in a model, while satellite observations are better suited to correct mesoscale features. Thus, we recommend to extract synoptic 2D fields rather than profiles to use as pseudo-observations. As errors are also present in deeper layers, it is certainly useful to add some profiles too, or to include deeper 2D fields.

Chapter 5

High-resolution model error analysis in the Mediterranean Sea

Anyone who attempts to generate random
numbers by deterministic means is,
of course, living in a state of sin
John von Neumann

The present chapter is devoted to the analysis, in a realistic case, of the errors that are inherent to hydrodynamic models. The errors of the GHER model in the Mediterranean Sea have already been studied by comparisons with other primitive equation models [Beckers et al., 2002], or with observations and with the climatology, using usual statistical methods and also wavelet decompositions [Azcàrate, 2004]. In this chapter, we rather study the sensitivity of the model to various variables using an ensemble of models. For this purpose, we will simplify our task by using "only" models covering the Mediterranean Sea, without nested grids. We will however choose a relatively high resolution, $1/16°$, corresponding to the resolution now used in operational OGCMs covering the Mediterranean, such as the MFS system (`http://www.bo.ingv.it/mfs/`). Before this, it is useful to describe the general patterns and characteristics of the water masses in the Mediterranean. Indeed, in the previous chapters, we already performed model runs in the Mediterranean, but in fact we were only interested in the open sea boundary conditions it provided to nested models covering the northwestern basin. Hence, we did not devote time to the description of the oceanography of the whole Mediterranean Sea. Next, we will explain how we generated an ensemble of model simulations, where various more-or-less well known inputs are allowed to vary according to the respective uncertainty. Statistics calculated on this ensemble are, in fact, the response of the non-linear hydrodynamic system to errors on the forcing terms. When those statistics are calculated at a certain timestep, they provide a spatial analysis of the model error; statistics calculated over the time dimension will show whether errors are intensified by the system, or rather disappear.

The model error is interesting as such. However, it can also be used for different purposes. Following Talagrand, we have explained in chapter 1 that all statistical data assimilation methods require at least some *a priori* knowledge of the error associated with the model state vector. In Kalman filter methods (and its variants such as the SEEK filter [Pham et al., 1998b]), (a subspace of) the model error covariance matrix must be specified *a priori*. A natural extension is the family of Ensemble Kalman filters. Here, *a priori* assumptions must be made to generate the ensemble of possible ocean states. Contrary to what Pham et al. [1998b] suggests, we do not have to suppose that the model error is well approximated by the model time variability. This is indeed not always the case. In particular, if the error covariance is approximated by $\mathbf{P} = \mathbf{S}\mathbf{S}^{\mathbf{T}}$, and the columns of S are the first EOFs of a historical run, assimilation experiments with the EOFs calculated on instant outputs of the model yield different results than those were the EOFs are calculated over day-averaged outputs. This shows that the method is not robust with respect to the choice of the model error space specification. Also, the period of time to consider for the EOF computation is not easily determined. For all these reasons, some authors do not even bother to calculate EOFs anymore, and simply put daily (or weekly) model states in the columns of S, without even removing the average state (C. Testut, private communication). An error space built from an ensemble of members perturbed according to our physical intuition of uncertainties on the data, does not require any further assumptions.

Most biogeochemical models, drift models, oil spill models *etc* are coupled with hydrodynamic models. Furthermore, most of these models are quite sensitive

to the provided hydrodynamic boundary conditions. Analyzing the effect of the hydrodynamic model uncertainty on coupled models is a major goal for all coupled models. However, we did not analyze this aspect further in our work. The impact of data assimilation in the hydrodynamic model on coupled models is also very interesting. In particular, data assimilation might generate unbalanced ocean states, yielding spurious and unphysical inertia-gravity waves which strongly (and incorrectly) influence the coupled models (A. Barth, private communication).

By calculating the EOFs of the ensemble of model states over a certain time-span, we obtain a very accurate basis to project each state on. The weights at day $n+1$ can then be calculated as a (non-linear) function of the weights at the n previous days, as well as some other inputs. If this function could be determined, we could emulate the hydrodynamic model, probably at a much lower cost. The drawbacks are, of course, (a) that the function is only valid in the part of the weight-space where it has been built, and (b) that the resulting ocean state cannot escape the space of linear combination of EOFs, i.e. no real innovations are possible. This question will be examined in detail in chapter 6.

5.1 Oceanography of the Mediterranean Sea

The Mediterranean constitutes an almost isolated system, connected only to the Atlantic ocean by the Strait of Gibraltar and to the Black Sea by the Dardanelles. Most processes of the global ocean, fundamental in oceanography, can also be found in the Mediterranean, either identically or analogously. Furthermore, Rixen et al. [2005] showed that the Mediterranean temperature increase over the last years is a proxy to that of the global ocean. Thus, the Mediterranean Sea constitutes a very interesting "test lab" for the global ocean, both theoretically and practically.

The geography of the entire Mediterranean is shown in Fig. 5.1. It is composed of two basins, connected by the Strait of Sicily. The general circulation in the Mediterranean Sea is cyclonic, composed of three predominant and interacting spatial scales: basin scale (including the thermohaline circulation), sub-basin scale, and mesoscale. The circulation patterns are very complex, due to the multiple driving forces, to strong topographic and coastal influences, and to internal dynamical processes. For example, we described in section 3.2 how the formation of deep water in the Gulf of Lions involves mesoscale eddies, yet it also influences the largest scales present in the sea.
Following Robinson et al. [2001], we will now describe the circulation in the Mediterranean following the typical length scales: basin scale, sub-basin scale and mesoscale. Other valuable information is provided by Millot [1999] for the Western basin, and by Malanotte-Rizzoli et al. [1999] for the Eastern basin. A summary is given in Azcàrate [2004].

Large-scale circulation

Processes relevant for the large-scale circulation include the thermohaline circulation, water mass formation and transformation, dispersion and mixing. These

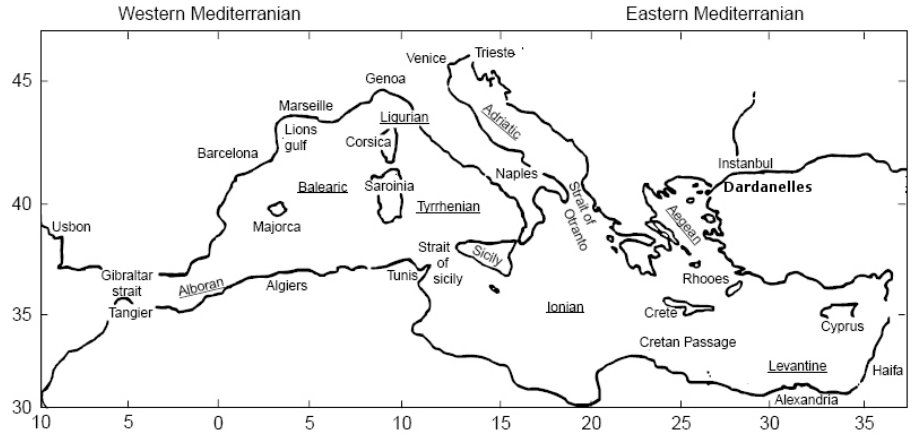

Figure 5.1: The Mediterranean Sea geography and nomenclature of the major sub-basins and straits. From Robinson et al. [2001].

processes are schematically shown in Fig. 5.2. The Mediterranean basins are evaporation basins, with freshwater flux from the Atlantic through the Gibraltar Strait and into the eastern basin through the Sicily Straits. Relatively fresh water of Atlantic origin (AW) circulating in the Mediterranean forms a 100-200 m thick surface layer. It increases in density because evaporation exceeds precipitation (salinity is about 36.5 at Gibraltar and 38.0-38.3 in the North of the Western basin), gradually becoming Modified Atlantic Water (MAW) which partly recirculates in the western basin following a cyclonic circulation, and partly flows to the easter basin via the Strait of Sicily. In the latter basin, it keeps flowing eastwards, eventually reaching the eastern boundary of the Sea. MAW forms new water masses via convection events driven by intense local cooling from winter storms. Dense water is produced, for the western basin, in the Gulf of Lions and the Ligurian Sea (it is called Western Mediterranean Deep Water, WMDW, see Fig. 5.2a, and the mechanism was described in 3.2), and for the eastern basin, in the Adriatic (it is called Eastern Mediterranean Deep Water, EMDW, see Fig. 5.2b), and sinks down through the Strait of Otranto (see Cushman-Roisin et al., 2002). In the eastern basin, intermediate water called Levantine Intermediate Water (LIW) is formed in the whole basin, but preferably in the Rhodes Gyre [Azcàrate, 2004] and in the north [Malanotte-Rizzoli et al., 1999], probably due to meteorological forcings. This is an important water mass which both recirculates in the eastern basin, and flows out to the western basin. There, it follows the general cyclonic circulation on its eastward travel. It was shown recently that its entrainment across the Algerian basin is due to mesoscale eddies rather than a permanent westward flow [Millot and Taupier-Letage, 2005]. Finally, the LIW contributes predominantly to the outflow from Gibraltar to the Atlantic, mixed with WMDW together with some EMDW, and flowing out below the AW inflow. It constitutes the main source of salty Mediterranean water in the Northern Atlantic [Marullo et al., 1999]. The Adriatic/Aegean Intermediate Water (AIW) is a water mass that has been detected along the Tunisian slope and at its bottom. It flows to the Western

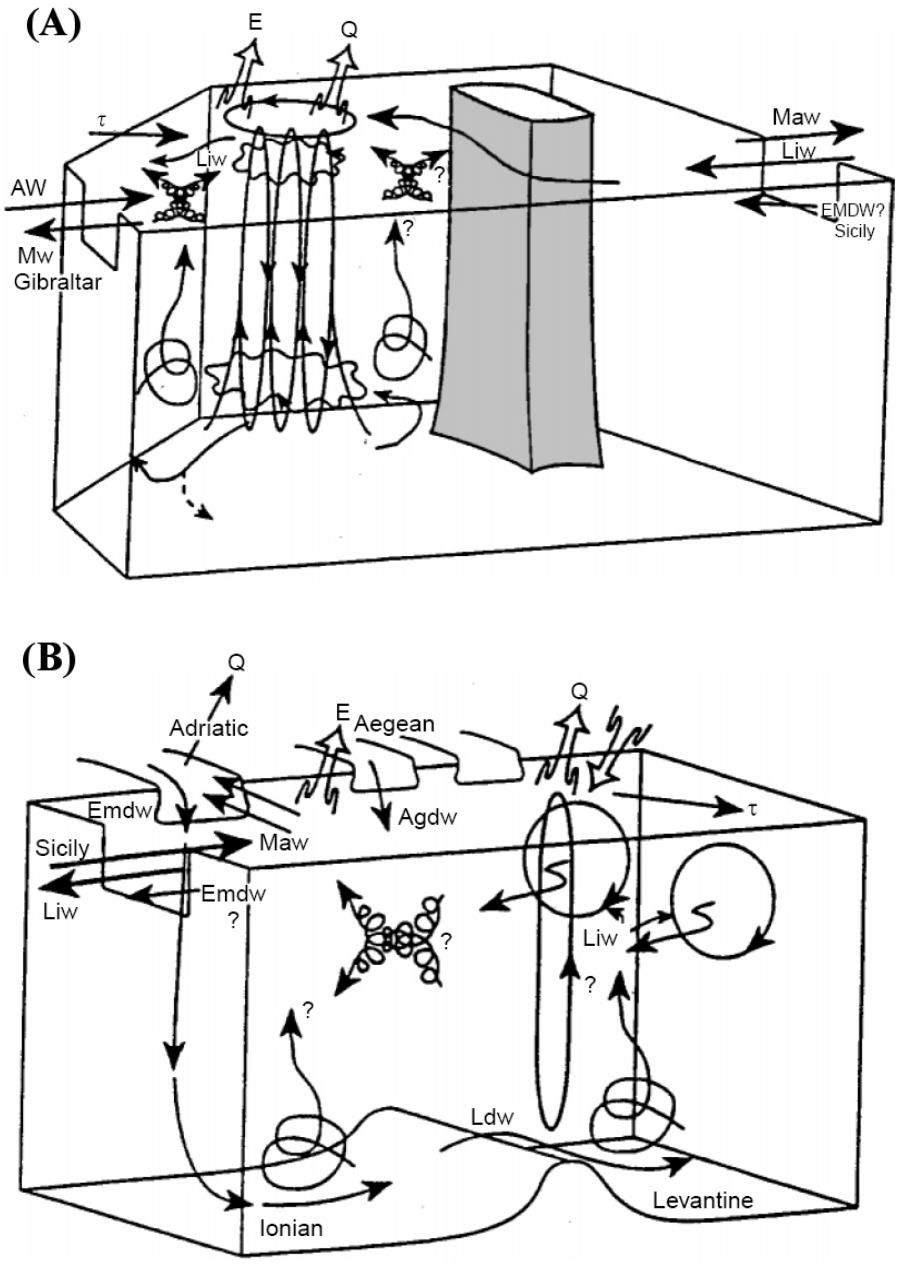

Figure 5.2: Processes of air-sea interaction, water mass formation, dispersion and transformation in (a) the Western Mediterranean, (b) the Eastern Mediterranean or Levantine basin. From Robinson et al. [2001].

basin below the LIW, through the Strait of Sicily. It constitutes a very dense water mass. EMDW is also formed during exceptionally cold winters in the north-eastern part of the eastern basin. Intermediate and deep waters (but not bottom water) are formed also in the Aegean basin (AGDW) and provided to the eastern basin through the Aegean Straits. In recent studies, it would rather be called Cretan Intermediate Water (CIW) and Cretan Deep Water (CDW); research about its formation is still in progress.

Measurements performed during cruises in the late 1980s indicate that the deep layer in the western basin was 0.12°C warmer and about 0.33 more saline than in 1959 [Robinson et al., 2001]; the heat and salt increase is continuing further during the last years. Based on the consideration of the heat and water budget in the Mediterranean, the deep-water temperature trend was originally speculated to be the result of greenhouse gases. A more recent argument also considers the anthropomorphic reduction of river water fluxes into the eastern basin to be the main cause of this warming trend (in the western basin), particularly the Nile river damming at Assouan [Skliris and Lascaratos, 2004]. Modeling both western and eastern basins together is thus increasingly important.
Since the beginning of the XX^{th} century, up to the mid-1980s, both the deep and intermediate conveyor belts in the eastern basin presented rather constant characteristics. After 1987, the most important changes in the thermohaline circulation and water properties basin-wide ever detected occurred. The Aegean, which had only been a minor contributor to deep waters, became more effective than the Adriatic as a new source of deep and bottom waters in the Levantine basin: its production of dense water is 3 times greater than in the Adriatic. The formed water was also warmer and more saline than the previously existing EMDW of Adriatic origin. After 1990, less dense CIW appeared to exit the Aegean basin mainly through the western Cretan Straits, and spread in the intermediate layers in major parts of the Ionian Sea, blocking the westward route of the LIW. The changes in the internal and open conveyor belts of the eastern basin have been named the Eastern Mediterranean Transient (EMT), and are explained by several hypotheses, including (1) internal redistribution of salt, (2) changes in the atmospheric forcing combined with long term salinity change, (3) changes in circulation patterns leading to blocking situations concerning the MAW and LIW, and (4) variations in the fresher water input of Black Sea origin, through the Strait of Dardanelles. A recent study of the changes is given in Millot et al. [2005]. Whether the present thermohaline regime will eventually return to its previous state or reach a new equilibrium is still an open question. Anyway, the signal of the change has spread from the Eastern to the Western Mediterranean.

Sub-basin scale circulation

In the western basin, the paths of the three main water masses (MAW, LIW and WMDW) and their variants are shown in Fig. 5.3. The AW enters the Mediterranean Sea through the Strait of Gibraltar, forming a jet of about 30 km width. It then flows anticyclonically in the western portion of the Alboran Sea, while a more variable pattern occurs in the eastern portion. A vein is also flowing from Spain to Algeria, called the Almeria-Oran jet. Further east, the MAW is

transported by the Algerian Current, with a transport of approximately 1.7 Sv [Benzohra and Millot, 1995]. It is relatively narrow and deep in the west, but then becomes wider and thinner until it reaches the Channel of Sardinia.
In the Tyrrhenian Sea, both the current along Sicily and the Italian peninsula, and the mesoscale activity, are the dominant features. Both the eastern and western coasts of Corsica present northward currents. The Western Corsican Current (WCC) varies seasonally, reaching its maximum transport values at the beginning of spring. Then it decreases progressively to reach a minimum in autumn. These seasonal changes are mainly due to the dense water formation processes occurring in the winter in the Liguro-Provençal basin. Its transport of 1.15 Sv presents no annual variability [Azcàrate, 2004]. The Eastern Corsican Current (ECC) is driven by thermohaline conditions, with a maximum in early winter. High transport values persist all the cold season; they reach a minimum in summer and autumn. The signal can be observed both in the MAW and LIW. The volume has been established at about 0.65 Sv. As described in chapter 3, the flows of MAW east and west of Corsica join and form the Liguro-Provenco-Catalan (LPC) current or Northern current (NC) [Robinson et al., 2001], which flows cyclonically along the French and Spanish coasts. Even though the ECC transport is less than the WCC one, the ECC influences the Northern Current to a major extent. In this latter current, mesoscale activity is more intense in winter. The dominant feature in the Gulf of Lions and the Ligurian Sea is the formation of deep water (WIW). In the Balearic basin, there is also an intense mesoscale activity, presenting a strong seasonal variability.
In the Channel of Sicily, the dominant features are the large mesoscale variability and the water exchanges between both basins. As seen before, the LIW and WMDW follow the general cyclonic circulation and exit the Mediterranean at Gibraltar. The WMDW also recirculates, and water which has accumulated at depths greater than 2000 m in the Algero-Provençal basin flows to the deep Tyrrhenian Sea (about 3900 m). The amount of WMDW there is also controlled by the density of the cascading flow through the Channel of Sicily and thus by the deep water formation in the Levantine basin.
Finally, in the Balearic basin, the islands form an arc that perturbs the flow passing through. The LPC current is weakened while flowing southward through the Channel of Ibiza, resulting in the formation of mesoscale eddies called "weddies" as they are formed by WIW. They cause a partial recirculation of the LPC current in the Arc of Balearic Islands [Pinot et al., 1994]. The remaining water continues to flow southward, and as seen before, the intermediate water exits the Mediterranean through the Strait of Gibraltar while the MAW again joins the fresher MAW entering the Strait in its travel to the east.

In the Eastern Mediterranean, energetic sub-basin scale features influence the basin-wide circulation. Important variabilities exist and include: (1) shape, position and strength of permanent gyres and their unstable lobes, meanders etc., (2) meander pattern, bifurcation structure, and strength of permanent jets, and (3) occurrence of transient and aperiodic eddies, jets and filaments. Therefore, new, object-oriented methods to compare and determine the position and strength of those gyres and currents in models and observations are currently investigater [Ben Bouallègue et al., 2005].
In the Ionian basin, the MAW entering through the Sicily Strait meanders due to the complex local topography. The current is called Atlantic-Ionian Stream

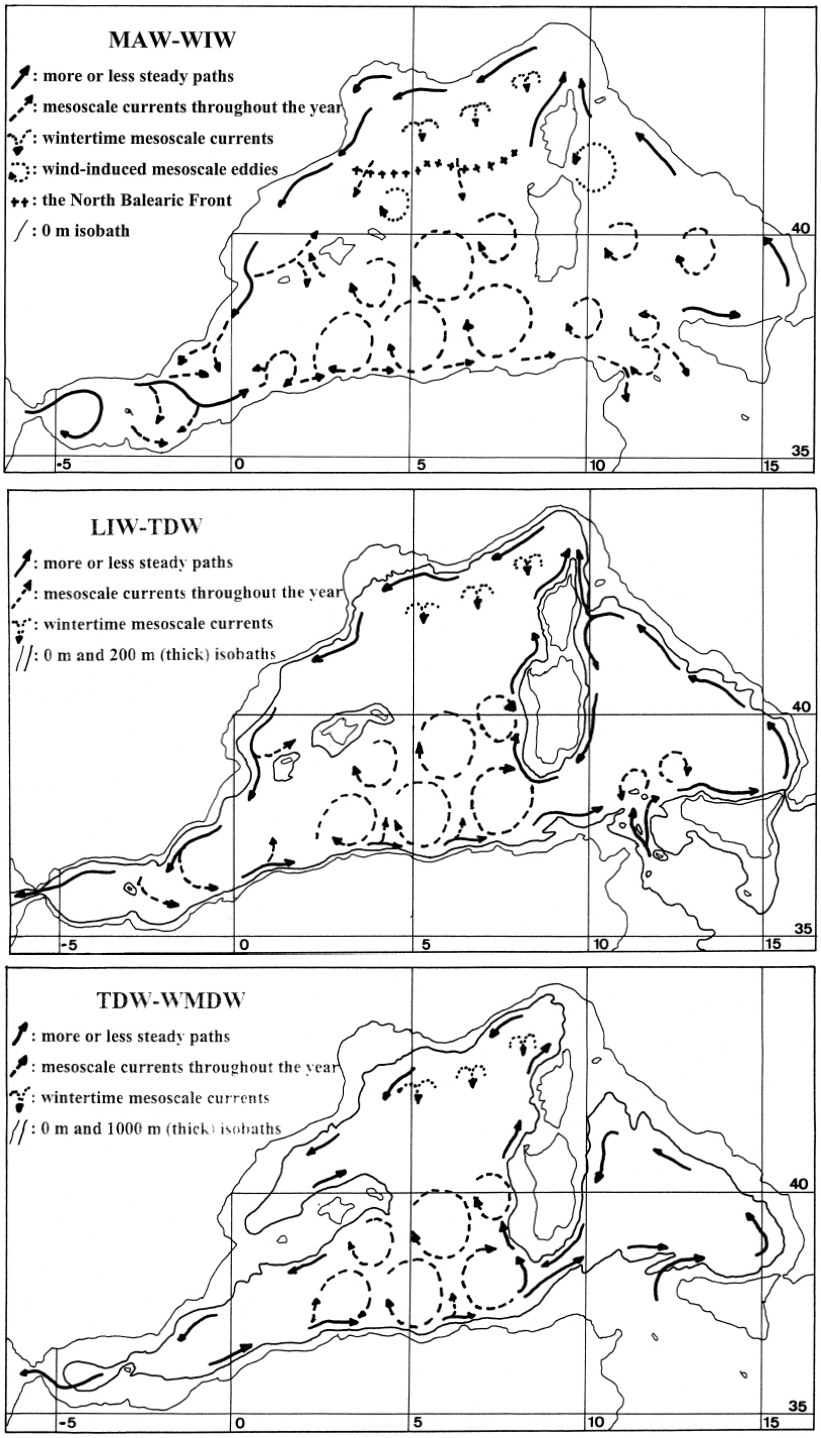

Figure 5.3: Schematics of the circulation of water masses in the Western basin (a) MAW and WIW, (b) LIW, (c) WMDM. From Millot [1999].

(AIS). After passing the Strait of Sicily, it bifurcates into two branches. One turns to the south, describing an anticyclonic pattern to the Tunisian coast. The other one continues eastward, crosses the basin and turns south, and reaches the Levantine basin. It eventually turns north-east and its signal gets weaker at about 27°E, where it contributes to the deep-water formation process.
In the Aegean basin, the very irregular topography, with many channels and islands, as well as the colder and less salty water from the Black Sea, yield a very variable circulation. In general, the surface circulation presents a cyclonic gyre in winter and a southerly movement in summer, caused mainly by the wind.
The Adriatic Sea has very shallow parts in the north and the center (maximum depth of 270 m), is deeper in the south (1200 m), and again less deep at the Strait of Otranto (780 m). The circulation is mainly induced by winds: the south-east wind called Sirocco and the north-east one called Bora. The circulation is cyclonic, with water coming from the Ionian basin entering via the east of the southern boundary of the Adriatic Sea, forming the East Adriatic Current (EAC). It is a weak and wide current, bringing modified LIW water to the north. The Western Adriatic Current (WAC) closes the cyclonic circulation. Three cyclonic gyres are also mostly present.
This image of the eastern basin circulation, with little information given on the southern part of the basin, was modified or completed recently by Hamad et al. [2003], based on the observation of composite daily and weekly infrared images of the period 1996-2000. In particular, the importance of eddies generated by the instabilities of the MAW flow, or by the wind, must not be underestimated, particularly when explaining how the alongslope flow is spread towards the open basin. In the southern Ionian, large eddies are generated as soon as the bathymetry is sufficiently deep (a few hundred meters), and they drift either alongslope or seaward. On average, the MAW does not cross the Ionian in its central or northern parts, but ultimately concentrates in the southern Ionian along the western Libyan slope as an unstable anticlockwise flow, generating eddies. The Libyan eddies then propagate downstream and eventually interact with the Ierapetra gyre, increasing the interannual variability of the latter. When entering the Levantine basin, the eddies tend to follow the deep isobaths and thus detach from their parent current, and the Ierapetra gyre as well. Therefore, and contrary to what was previously believed, the area known as Mersa-Matruh is occupied not by a permanent or recurrent feature, but by slowly propagating and merging anticyclonic eddies originated elsewhere. The northwestern edges of such mesoscale eddies must have been confused with the Mid-Mediterranean Jet. In any case, this whole discussion once more proves the tight interactions between the different scales, and the need to resolve mesoscale features in models.

Fig. 5.4 shows a conceptual model following the path of a jet of MAW entering the eastern basin through the Strait of Sicily, meandering through the Ionian Sea, which is believed to feed the Mid-Mediterranean Jet, and continues to flow through the central Levantine all the way to the Israelian shores. This Mid-Mediterranean Jet bifurcates, one branch flowing to Cyprus and a second branch flows eastward then southward. The most important sub-basin features include the Rhodes gyre, the Mersa-Matruh gyre, the Ierapetra gyre, and the southeastern Levantine system of anticyclonic eddies (see e.g. Larnicol et al. [2002]).

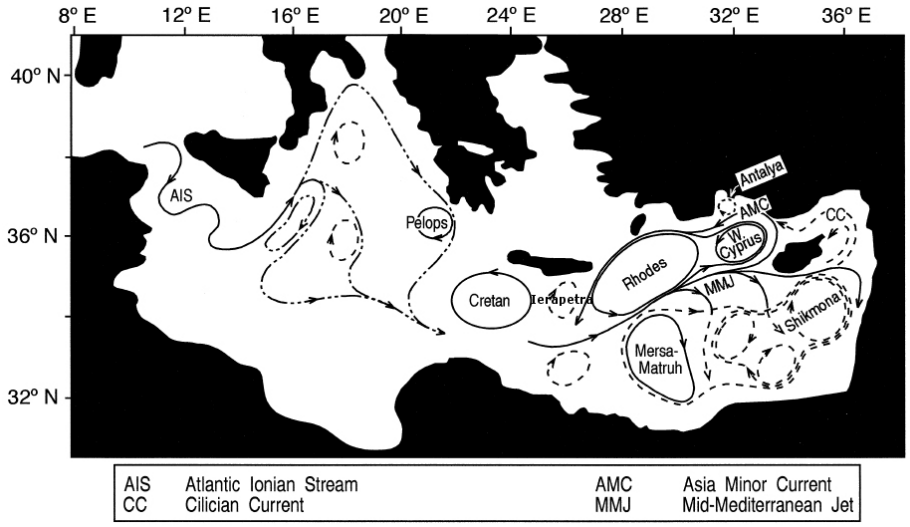

Figure 5.4: Schematics of the circulation of water masses in the Eastern basin. From Malanotte-Rizzoli et al. [1999].

As for the large-scale circulation, the sub-basin scale circulation was also modified during the last decade. In the south Aegean Sea, during the period following the EMT, the usual seasonal variability of the circulation has been replaced by a rather constant pattern. In the Cretan Sea, eddies were in a seasonal evolution in the 1980s, while in the 1990s there was a constant succession of three main eddies (one cyclone in the west, one anticyclone in the central region, and again one cyclone in the east. During the first half of the 1990s, a large three-lobe anticyclonic feature also developed in the south-western Levantine, blocking the free westward LIW flow from the Levantine to the Ionian, and causing a recirculation of the LIW within the west Levantine basin. The data collected in late 1998 and early 1999 indicated that this circulation pattern has been reversed to cyclonic, confirming the transient nature of these eddies. Consequently, the Atlantic Ionian stream was not flowing from Sicily towards the northern Ionian, but directly eastwards crossing the central Ionian towards the Cretan Passage.

Mesoscale circulation

The horizontal scale of mesoscale eddies is related to the internal Rossby radius of deformation, defined as the length at which buoyancy effects become of the same order as the Coriolis effect. In the Mediterranean, it has typical values $\mathcal{O}$(10-14 km) [Robinson et al., 2001], thus requiring very fine sampling resolutions to be observed. For this reason, only recently different mesoscale features were found in both basins; typically with satellite observations.
In the western basin, intense mesoscale activity has been detected as instabilities along the coastal currents, leading to the formation of mesoscale eddies which can eventually move across the basin or interact with the current itself. Along the Algerian current, for example, meanders of few tens of kilometers

are generated due to the unstable character of the current, and both cyclonic and anticyclonic eddies develop. The cyclonic eddies are relatively superficial and short-lived, while the anticyclonic ones last for weeks or months, or according to the recent study of Puillat et al. [2002], up to three years. The quasi permanence of these eddies increases the mixing between the resident and newly-entered surface waters. The anticyclonic eddies generally detach from the coast, evolving eastward (a few km per day). Nevertheless, not all of them have a sufficiently large vertical extent to markedly modify the circulation of all water masses present. The hypothesis that coastal eddies could extract more and more energy from the main current, so that they would increase in size, be advected eastward more slowly and grow deeper, has to be rejected since their depth and kinetic energy fluctuate without any special tendency [Millot et al., 1997]. Only exceptionally, these eddies induce significant currents at depths of 100 m and more. In fact, coastal eddies are now considered to be the upper part of an event also comprising another, deeper, anticyclonic eddy, with non-coinciding axis at least at the beginning of the event. The anticyclonic eddies drifting eastward are associated with large elevations of the sea surface (10-20 cm). They are eventually blocked by the topography of the Sardinia Channel, that does not allow deep structures to progress eastward. Their path is inflected to the north and the coasts of Sardinia, where they are able to pull fragments of LIW seaward, then to the west in the middle of the Algerian Basin. Old eddies extend deep in the water column and have long life-times. This implies that the generation rate is low, of the order of one per several months. Finally, the eddies sometimes enter coastal regions and interact again with the Algerian Current. Furthermore, the current and its meanders can be disturbed by the "open sea eddies", with an energy transfer from the current to the eddies. All this indicates that eddies can modify the circulation over a relatively wide area and for relatively long periods of time.

Mesoscale eddies have also been detected and investigated in the Tyrrhenian Sea, along the Corsican coasts and in the Ligurian Sea. They are present in the meanders of the LPC current, showing a large seasonal variability. In the middle of the Balearic Sea, mesoscale structures have generally been linked to instabilities of the alongslope circulation due to bathymetric features, but they might also be due to interactions with recent MAW entering through the Balearic channels.

Mesoscale currents have also been found; they are characterized by a permanent occurrence and by a baroclinic structure with relatively large amplitude at the surface, moderate at the intermediate level and still noticeable at depth, thus indicating large vertical shear of the horizontal currents.

In the Levantine basin, dedicated high-resolution sampling led to the discovery of open ocean mesoscale energetic eddies, as well as jets and filaments. Mesoscale eddies dynamically interacting with the general circulation occur with diameters of the order of 40-80 kilometers; their relation with the subbasin scale circulation in the Levantine was discussed above.

5.2 Implementation of an Ensemble Run

The generation of an ensemble of members is based on the estimation we make of the uncertainty affecting the model inputs. In his reference paper, Evensen [1994] introduced the Ensemble Kalman filter, later he also formulated a practical algorithm to perturb forcing fields with random, but correlated fields [Evensen, 2003]. By specifying a correlation length (or 2 or 3 according to the domain dimension), generating random fields and taking the (inverse) Fourier transform, one may generate spatially coherent perturbations, and add them (with a specified amplitude) to any initial or forcing field. However, depending on the variable that we wish to perturb, other techniques may lead to more realistic results. In our work, we modified the model initial conditions, its bathymetry, internal parameters (diffusion coefficients) and some atmospheric forcings. Auclair et al. [2003] realized a similar, very careful study, covering only the Gulf of Lions. There, perturbations affected the initial density field, the position and importance of the LPC current in the initial and boundary conditions, the wind stress, and the Rhône river runoff.

Bathymetry perturbation

The bathymetry is a key factor for hydrodynamic simulations. Its importance has been shown e.g. in She et al. [2007]. In chapter 3 of the present work, we noted that the LPC current is constrained to flow along the shelf break, except at the canyons where it can penetrate over the shelf. In any case, the bathymetry plays a crucial role in the fluctuations of the LPC current and its meanders [Petrenko et al., 2005]. Of course, we also expect the bathymetry to influence other currents at other locations.
Since the mesh has a limited resolution (1/16°), the original bathymetry from Smith and Sandwell [1997] is interpolated and then smoothed in order to avoid abrupt variations to cause instabilities. We decided to generate 5 supplementary bathymetries, where the smoothing algorithm has not been applied at all, applied but yielding 2 times less smoothing, or applied to yield 2 fields even more smoothed. This allowed us to create sub-ensemble 1, SE_BATHY, which we believe to be representative of the bathymetries currently used in models. In the original (only interpolated) bathymetry, the deepest point reaches 4910 m, while it is limited to 4333 m in the most smoothed one. The largest differences are found along the Algerian, Greek and Cretean coasts, in the Tyrrhenian Sea, and along the edges of the deep part of the Ionian Sea. As an example, the difference between the unperturbed bathymetry, and the most smoothed bathymetry is shown in Fig. 5.5. SE_BATHY is much smaller than the other sub-ensembles, but we found it pointless to generate 60 members with bathymetries which would necessarily be only slightly different.

Atmospheric forcings

The wind, relative cloud coverage and air temperature, are all perturbed separately following the procedure explained in section 3.5, generating 60 new members each. It consists of decomposing the fields as sums of EOFs, and then randomly (but coherently in time) perturb the weights before summing back the weighted EOFs to obtain the perturbed field. The weights are multiplied

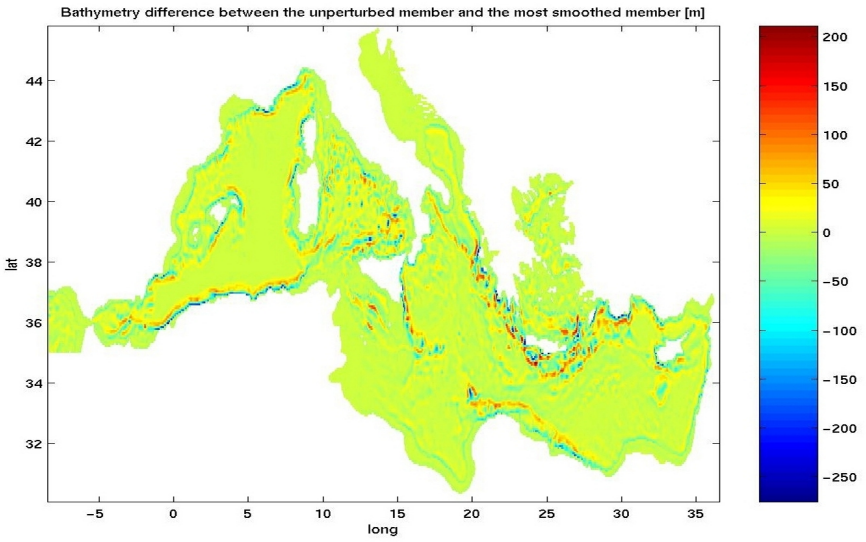

Figure 5.5: Difference between the reference bathymetry and the smoothest bathymetry.

by a random coefficient in the range $[1-\epsilon; 1+\epsilon]$, with ϵ fixed at 0.4 (wind), 0.5 (cloud coverage), 0.35 (air temperature). The corresponding rms perturbation was then found to be about 1 m s^{-1}, 12% and 2°C respectively, which are realistic error values. The cloud coverage is of course limited back to the interval [0%; 100%]. By this procedure, we obtained sub-ensembles 2 (SE_WIND), 3 (SE_CLOUD) and 4 (SE_AIRT).

Diffusion coefficients

In the reference (unperturbed) member, the real horizontal diffusion coefficient and the horizontal supplementary pseudo-diffusion coefficient introduced for the velocity are defined in MKS units as 125 and 1500 respectively. In 4 supplementary members, we modified those coefficients respectively as 25 and 300, 75 and 900, 375 and 4500, and 500 and 10000, thus largely covering the entire range of realistic values for the coefficients. This forms sub-ensemble 5 (SE_DIFFU). As SE_BATHY, it has only few members, but it is pointless to generate 60 members with only small differences in diffusion coefficients between them.

Initial conditions

We generated 60 new members in sub-ensemble 6 (SE_IC), whose initial conditions are perturbed following the method detailed in Barth [2004], Barth et al. [2006]. The initial temperature and salinity fields are perturbed by a pseudo-random field with a horizontal correlation length of 100 km and a vertical correlation length of 20 m. Those fields are generated as explained in Evensen [2003]. At the surface, the perturbations of temperature and salinity are 0.5°C and 0.1 psu respectively. This surface perturbation is then applied to the water

column as a function of depth, decreasing with z as a Gaussian function with an inflection point at the depth of h_{IP}=170 m (the transition between the two sigma domains, and the average depth of shelf breaks):

$$\sigma(z) = \sigma(0)\exp(-\frac{z^2}{h_{IP}^2}) \tag{5.1}$$

Barth [2004] compared CTD profiles measured in the Ligurian Sea and an implementation of the GHER model, showing that below 170 m, differences are inferior to 0.1°C.

The temperature and salinity field perturbations are associated to elevation and velocity field perturbations, that we calculated in the following way. At depth h_0 of 700 m, horizontal variations of the hydrostatic pressure are small. This hypothesis was validated on several model results at different times [Barth, 2004]. It implies a balance between temperature, salinity and surface elevation. If the hydrostatic pressure does not change at depth h_0, the elevation perturbation $\Delta\eta$ can therefore be computed from the temperature and salinity perturbations ΔT and ΔS by the following linearized relation [Haines, 2002]:

$$\Delta\eta = \int_{-h_0}^{0} \alpha\Delta T - \beta\Delta S dz \tag{5.2}$$

The parameters α and β are the mean thermal expansion coefficient and saline contraction coefficient respectively,

$$\alpha = -\frac{1}{\rho}\left(\frac{\partial\rho}{\partial T}\right)_{p,S} \tag{5.3}$$

$$\beta = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial S}\right)_{p,T} \tag{5.4}$$

The temperature, salinity and surface elevation perturbations in turn allow us to compute the hydrostatic pressure perturbation using a linearized state equation.

$$\Delta p_h(z) = g\rho_0\Delta\eta - g\rho_0\int_z^0 \alpha\Delta T - \beta\Delta S dz \tag{5.5}$$

The perturbation of the horizontal velocity is supposed to be in geostrophic balance with the hydrostatic pressure perturbation. While the geostrophic balance is well respected in the open sea, the relation between pressure and velocity often comprises a large ageostrophic part near the coast. For instance, the geometry of the coast, nonlinear and non-stationary effects cannot be neglected. Therefore, at the coast, the velocity field is not perturbed. This also avoids that large velocity perturbations near the coast, inconsistent with the coastline geometry, produce an "adjustment" shock. Thus, if $\mathbf{x}$ is the distance from a point to the nearest coast, a coefficient $c(\mathbf{x})$ ranging from 0 at the coast to 1 at 50 km of the coast ensures a smooth transition.

$$\Delta\mathbf{u} = \frac{c(\mathbf{x})}{f}\nabla(\Delta p_h) \wedge \mathbf{e_z} \tag{5.6}$$

The turbulent kinetic energy is not perturbed. Indeed, it is not clear how a consistent perturbation could be calculated. Furthermore, it adjusts itself very

Sub-ensemble	*Name*	*Perturbation*
1	SE_BATHY	Bathymetry
2	SE_WIND	Wind field
3	SE_CLOUD	Cloud coverage
4	SE_AIRT	Air temperature field
5	SE_DIFFU	Horizontal diffusion parameters
6	SE_IC	Initial Conditions (T,S,η,$\mathbf{v}$)

Table 5.1: Synthesis of the 6 sub-ensembles

rapidly to the density structure and the velocity profile of the model.

We have thus generated 6 sub-ensembles, with perturbations of respectively (1) the bathymetry, (2) the wind forcing, (3) the cloud coverage, (4) the air temperature, (5) the horizontal diffusion coefficients and (6) the initial conditions, summarized in table 5.1. The total count of members in our ensemble is 250. Having different amounts of members in the different sub-ensembles is not a problem, as in this chapter we analyze sub-ensembles separately.
All the members are integrated by the model during 1 month, with daily outputs of the prognostic variables. In the following sections, we will analyze the model response to the perturbations. Strictly speaking, all we are calculating are differences between modified members and an unperturbed simulation, where all the parameters and forcings correspond to our best guess. Of course, we don't know which member best represents the real ocean state. However, following the usual terminology used in twin experiments, we will still consider that the unperturbed, central member (reference member) represents the truth, and hence we will freely mix the terms "difference" (between members and the central member) and "error" in the remainder of this chapter.

5.3 Temporal analysis of the model error

Because of the very high computational power required to obtain the results in this section, only a part of each sub-ensemble was used rather than the full sub-ensembles (except sub-ensembles SE_BATHY and SE_DIFFU which are small). In order to assess the evolution of the error during our one-month simulation, we will first examine the daily evolution of a well-known statistical indicator, the root mean square difference between the members of the different sub-ensembles and the central (unperturbed) member. We calculated those rms differences (a) on the entire 3D fields, (b) on the upper σ region (from the surface to 170 m depth), and (c) on the surface layer, with an average thickness of 1.9 m. The results are shown, for the SE_WIND as an example, in Fig. 5.6, and, for the different sub-ensembles but only for the upper σ region, in Figs. 5.7 to 5.9, for temperature, salinity and surface elevation respectively.

These graphics allow us to conclude that

- modifying atmospheric forcings generally results in an increase (over time) of the rms difference on the temperature and salinity fields. It is natural that the surface layer is perturbed more than the whole basin. How-

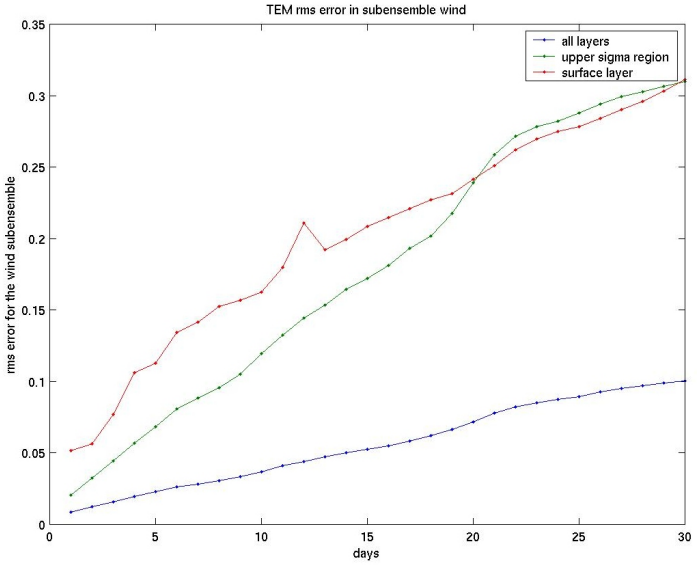

Figure 5.6: Mean (over the members) root mean square (over space) temperature difference in SE_WIND, as a function of time given in days from 1 January. The error is given in °C. The 3 curves represent the rms difference calculated respectively over the whole 3D basin, the upper σ layer, and the surface layer.

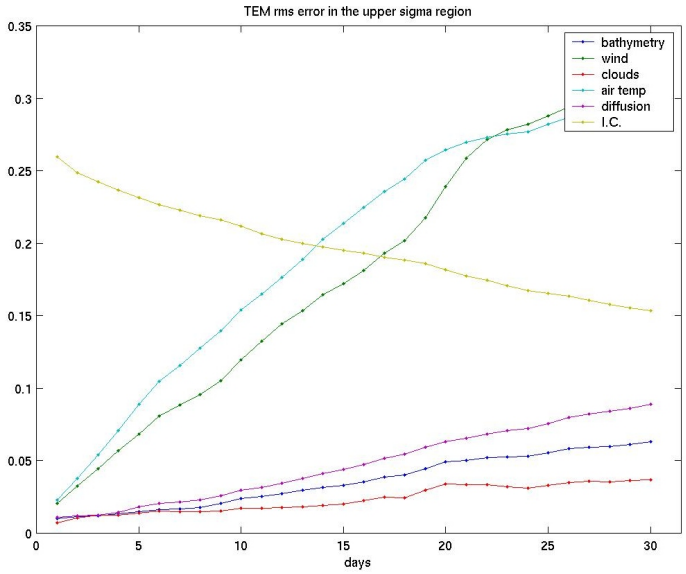

Figure 5.7: Mean (over the members) root mean square (over space, in the upper σ region) temperature difference in the different sub-ensembles, as a function of time [m].

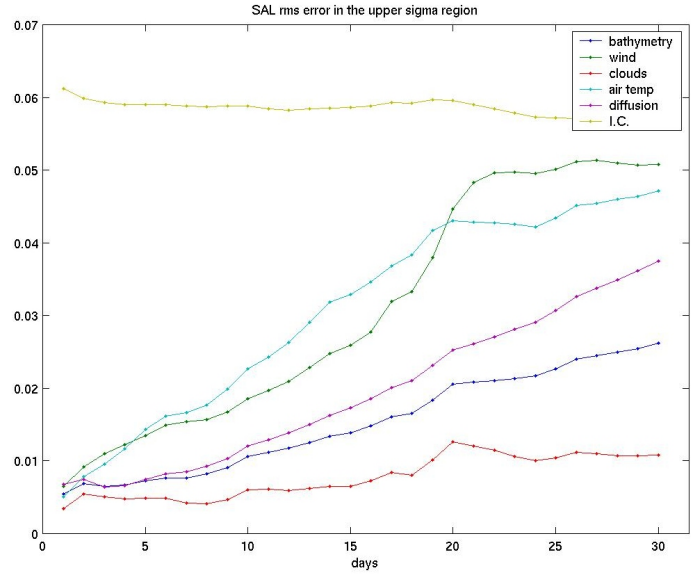

Figure 5.8: Mean rms salinity difference in the different sub-ensembles, as a function of time.

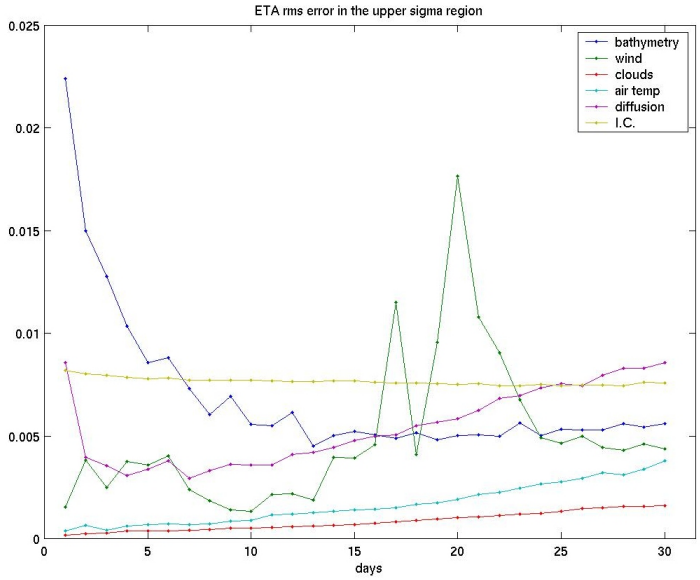

Figure 5.9: Mean rms elevation difference [m] in the different sub-ensembles, as a function of time.

ever, it seems that the perturbations are transmitted downwards relatively fast, as the results in the surface layer and the upper σ region are always similar. This can be explained by remembering the simulations are performed in winter time, and a mixed layer takes place. An example is shown for SE_WIND in Fig. 5.6, but similar graphics are obtained for the SE_CLOUD and SE_AIRT.

- with the chosen perturbation intensities, the air temperature and wind perturbation have the largest impact on the model temperature; the salinity is affected even more by the initial conditions perturbation (Figs. 5.7 and 5.8).

- the corresponding modifications induced on the sea surface elevation are more subtle to describe. The increase of error due to air temperature and cloud coverage perturbations is small. Sudden large errors on the wind field result in peaks in the sea elevation error. Figure 5.10 shows the root mean square of the wind velocity over the basin; a large wind peak centered around 20 January is visible. Due to the method followed to build the perturbations, they are larger when the wind velocity itself is large. Therefore, the bigger error on sea surface around 20 January corresponds to our intuition. However, another relatively large error appears already around 17 January, resulting from other large (random) errors in the wind field. In both cases, after a transition period (1 to 3 days depending on the importance of the perturbation), the error is reduced approximately to its initial values.

- the more or less strong smoothing of the bathymetry is responsible for relatively small errors on the temperature and salinity fields. The errors however slowly increase with time. Concerning the surface elevation, a large error is immediately generated, probably because the model initial conditions are not balanced with respect to the perturbed bathymetry. After approximatively a week, this error is reduced and a plateau is attained.

- smaller or larger horizontal diffusion also leads to a constant increase of error, which is small for the temperature field, but larger for the salinity and even more for the sea surface elevation. When analyzing the errors in detail for each member of the sub-ensemble, it appears that rms errors are larger when diffusion coefficients are too big, than when they are too small. Thus, over-smoothing the output fields is probably worse than the contrary (as long as the model is stable, of course).

- finally, the model tries to adapt itself to modifications to the initial conditions. These initial conditions are already balanced, by construction. Hence, no startup error peaks appear, and for temperature, salinity and surface elevation, the error slowly decreases with time.

To check whether this error is approximately stationary in space, and only changes in intensity, or rather if the error is constantly changing, we can run the following test. Every day, we calculate the central empirical orthogonal functions (EOFs), i.e. the EOFs of the anomaly between the members and the unperturbed reference (central) simulation. The EOFs are calculated over

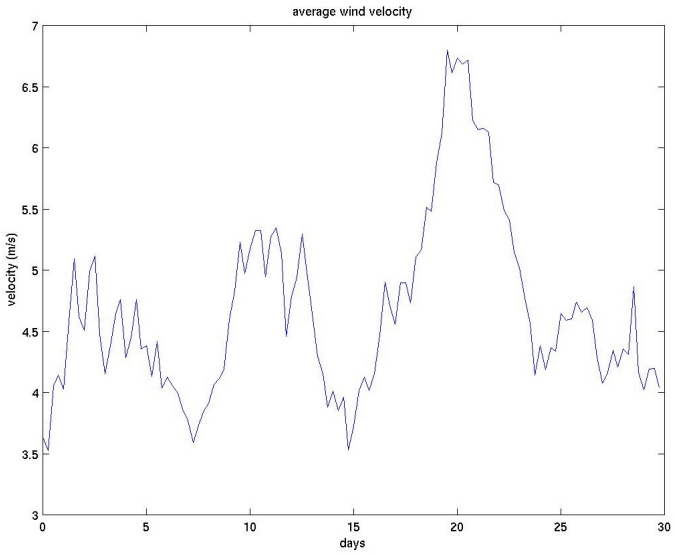

Figure 5.10: Root mean square wind velocity over the basin, as a function of days.

the whole basin, for each variable (T,S,η) separately. We do this in each sub-ensemble separately. The anomalies of the next day are then projected on the previous-day EOFs, and the total variance that can hence be explained is calculated. It is represented in Figs. 5.11 to 5.13. If the variance is close to 100%, it means that the errors are quite stationary in space.

It should be noted first that some oscillations are visible (particularly in Fig. 5.12) that are common to all 6 curves. This can be explained by the variations of one or multiple external parameters. Indeed, if the model was linear, the effect of the variation of model parameters (i.e. atmospheric forcings...) would be the same in the reference run and the ensemble members, hence it would not be visible in the anomalies, and neither in the anomaly EOFs. However, with non-linear models, variating parameters act differently in the reference run and ensemble members, causing visible effects of common oscillations in the graphics. These oscillation are important as such, but not if one's aim is to study the ensemble spatial stability in time.
Let us now analyze the evolution of the graphics, apart from this. It appears that the error structures (on temperature, salinity and elevation), due to perturbations of the initial conditions, and whose intensity slowly decreases in time, are very stationary in space, with over 90% of a day's anomalies corresponding to the previous day's anomalies. The wind and air temperature variations also induce temperature and salinity error patterns which are relatively stable in time, at least after the first few days. In fact, we might hope that the first temperature and salinity anomaly EOFs, calculated from the different members anomalies, represent the response of the model to the air temperature and wind field EOFs, which were used to build the perturbations. The sea surface eleva-

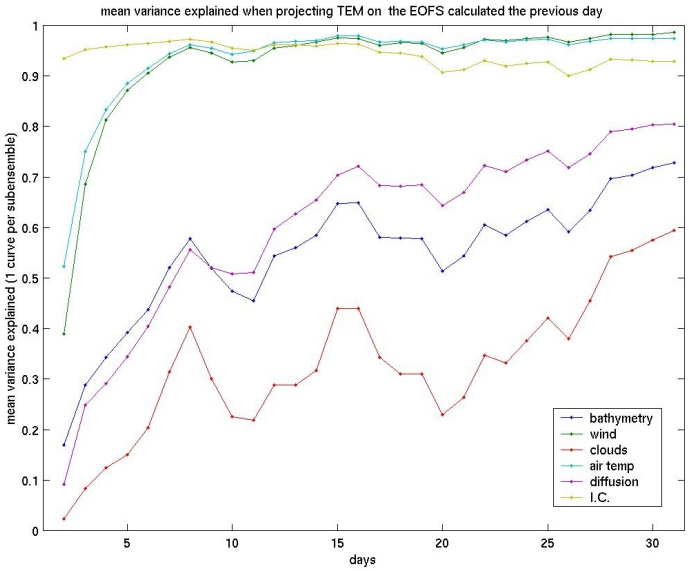

Figure 5.11: Proportion of model temperature anomaly variance, on a certain day, that can be explained by the anomalies of the previous day (in the same sub-ensemble).

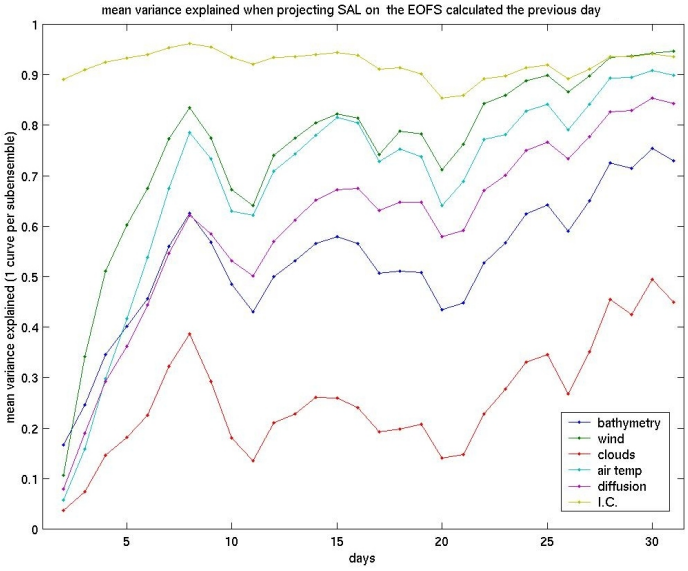

Figure 5.12: Idem, for the salinity.

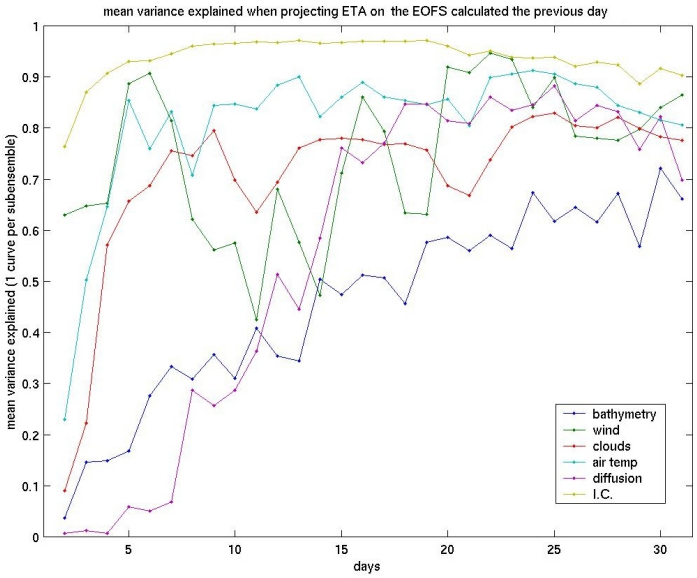

Figure 5.13: Idem, for the surface elevation.

tion error structure, due to air temperature perturbation, is also quite stationary after a few days. However, the sea surface elevation error is, as we expect, more sensitive to wind errors, given the direct reaction of the sea surface to wind stress. This explains the lower values observed after day 17 in Fig. 5.13 for the wind-induced errors; a new anomaly, generated by a large new wind perturbation, is moving rapidly in space; it is not (less) projectable on the previous days anomalies. After the wind error peak, the curve more or less resumes its slow ascension, but new errors cause new slight decreases after day 20.
The model response to perturbations applied to the bathymetry and the diffusion coefficients is more complex; the salinity and temperature error patterns are becoming more stationary with time without ever becoming completely stationary (around 70% of the variance is explained by the previous day's variance). The sea surface elevation spatial error structure due to bathymetry is also increasingly stationary, while the diffusion parameter error causes error structures which are constant in space for a few days, then locate elsewhere.
Finally, perturbations on the clouds field seem to generate ever-changing errors on temperature and salinity fields, with less than half of a day's error explained by the previous days anomalies; the shape of the error on sea surface elevation is somewhat more constant. This is, again, understandable; clouds immediately influence the temperature of the upper layers of the sea; but they affect the surface elevation field only via the model adjusting itself to the new density field.

5.4 Spatial analysis of the model error

In the previous section, we analyzed the time-evolution of some particular numbers representing the whole basin. Let us now analyze the spatial distribution of the impact of the various perturbations at fixed timesteps, e.g. after 2 and 4 weeks. Indeed, the differences between perturbed and unperturbed simulations (which we abusively called "errors") are not distributed uniformly over the basin, and are worth to be looked into. We will first examine the mean and standard deviation of these differences in order to indicate whether some locations are more sensitive to perturbations. We will also look at the third and fourth order statistical moments, mainly to examine whether the response of the model to the applied perturbations can be considered as a Gaussian random variable. Finally, we will compute the central EOFs of the anomalies in order to find the directions of the model space where differences are most likely to appear when perturbations are applied.

As mentioned in section 2.3, ensemble techniques retain all moments of the error probability density function (pdf); they can be calculated from the members at any given timestep. We will first analyze the first order moment. The difference between this mean of members and the unperturbed member would vanish as the ensemble size increases, if (a) the model was linear, (b) the added perturbations had a zero mean. Obviously none of those hypotheses are true here; nevertheless the computation of that difference might be an interesting indicator. The second-order moment, or variance, of the different sub-ensembles obviously depends on the intensity of the perturbations. In order to compare the importance of the effect of different perturbations, as suggested in Auclair et al. [2001], the variance could be divided by some norm, e.g. the energy contained in the perturbation. However, we feel that the perturbations applied in the sub-ensembles represent our knowledge of the uncertainty on the associated forcing, as explained in section 5.2; it is interesting to examine the variance yielded by realistic perturbations rather than compare variances of equally energetic perturbations.
The third and fourth order moments, known as skewness and kurtosis, represent respectively the asymmetry and the "peakedness" of the error pdf. The skewness is defined, for a ensemble of size N, as:

$$s = \frac{\sqrt{N}\sum (x_i - \overline{x})^3}{(\sum (x_i - \overline{x})^2)^{3/2}} \tag{5.7}$$

The definition above is a biased estimator if a sample of size n is used rather than the entire ensemble (size N); the following estimator should then be used:

$$S = \frac{\sqrt{n(n-1)}}{n-2} s \tag{5.8}$$

A positive skewness indicates that the right tail of the pdf is more important, negative skewness is caused by a fatter or longer left tail.
The kurtosis is the fourth order moment; for an ensemble of size N it is estimated by

$$k = \frac{N\sum (x_i - \overline{x})^4}{(\sum (x_i - \overline{x})^2)^2} - 3 \tag{5.9}$$

Again, when only a sample of size n is available in a larger ensemble, the estimator above is biased, and the following equation is usually used:

$$K = \frac{(n+1)n}{(n-1)(n-2)(n-3)} \frac{\sum (x_i - \overline{x})^4}{\sigma^2} - 3\frac{(n-1)^2}{(n-2)(n-3)} \tag{5.10}$$

where σ^2 is an unbiased estimator of the larger ensemble variance. Unfortunately, K is still generally biased (it is unbiased for normal populations). Higher kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to negative kurtosis with frequent modestly-sized deviations. Thus, the former have an acute peak and "fat" tails, while the latter have flatter peaks and thin tails. An extreme example of negative kurtosis is the uniform distribution. Normal populations have a zero kurtosis.
The analysis of the skewness and kurtosis is interesting because it allows to examine whether the processes are Gaussian or not (a desirable property, that is required to use the EKF data assimilation scheme and its widely used variants like e.g. SEEK). Indeed, large absolute values of either the skewness or kurtosis will indicate that the process certainly can not be considered as Gaussian. In principle, the model variables could then be transformed until a Gaussian error pdf is obtained, using so-called anamorphosis methods [Chiles and Delfiner, 1999, Bertino et al., 2003].

The difference between the sub-ensemble means and the unperturbed, reference member, is very small for all 6 sub-ensembles. Mean values over the whole basin are of $\mathcal{O}(10^{-3}°\text{C})$ for the temperature (except in SE_AIRT, where the mean difference is slightly over 0.1°C), $\mathcal{O}(10^{-4})$ for the salinity and $\mathcal{O}(10^{-6}\text{m})$ for the surface elevation. As mentioned before, this is no proof of well built ensembles, but still gives us some confidence. Nevertheless, the difference between means and reference member is not spatially uniform. The surface elevation mean in SE_BATHY is approximately the same as the reference run everywhere, except mainly in the Strait of Sicily, along the Tunisian coast, and in the Aegean Sea, where differences of up to 3 cm are present (Fig. 5.14). Let's note that the results, west of Gibraltar, are not to be fully trusted since they are heavily influenced by the Atlantic Ocean open boundary. The mean temperature in SE_BATHY is also almost identical to the reference temperature, except along the coastlines, with errors of 0.2°C.

In SE_WIND, the wind perturbation induces small surface elevation perturbations, which, after 2 weeks, are more pronounced in the shallow Northern Adriatic, and in the Ionian (Fig. 5.15a). After 1 month, the largest errors are now present along the Libyan and Egyptian coastlines (Fig. 5.15b). The center of Western basin is (very) slightly too low, which indicates that the general cyclonic circulation is a little bit reinforced by the wind. Preliminary tests (not shown here) with only 8 members in the sub-ensemble showed the opposite effect, with larger absolute values, indicating that the wind perturbation caused the cyclonic circulation to slow down. As the sub-ensemble size increases, the effect of random wind perturbations is thus averaged out.
In SE_WIND, the temperature is very slightly too warm in the Ionian (not shown).

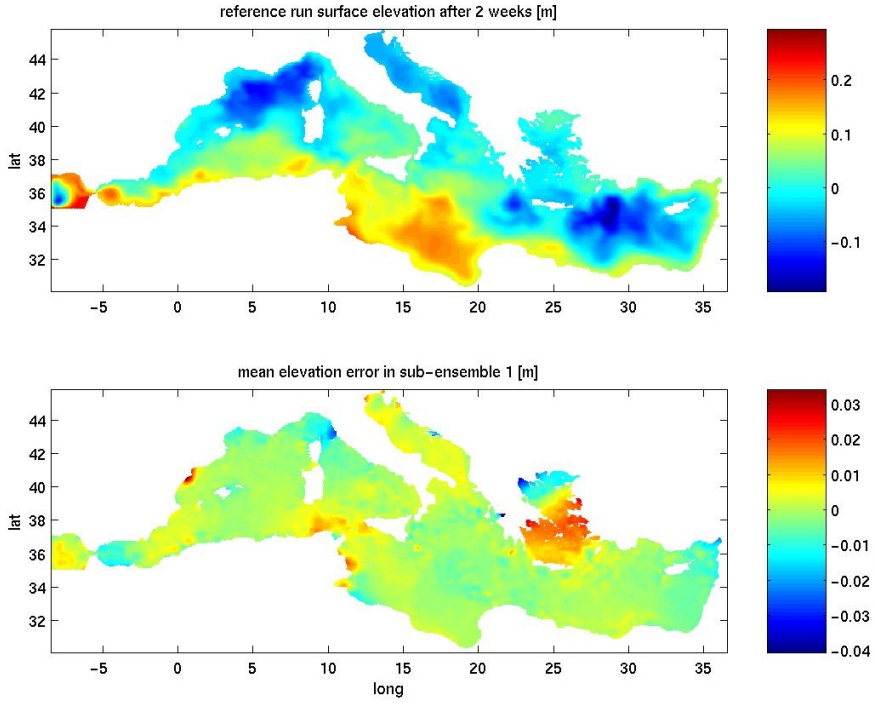

Figure 5.14: (a) Model output after 2 weeks of simulation: reference member sea surface elevation, (b) mean sea surface elevation difference between the members of SE_BATHY and the reference run, after 2 weeks of simulation.

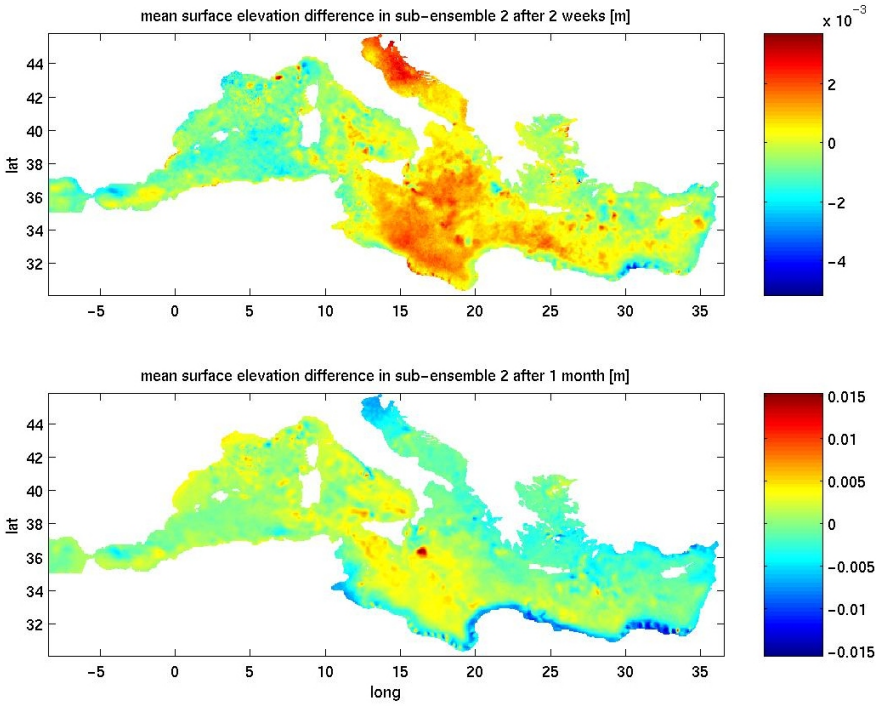

Figure 5.15: (a) Mean sea surface elevation difference between the members of SE_WIND and the reference run, after 2 weeks of simulation, (b) the same figure after 1 month of simulation,

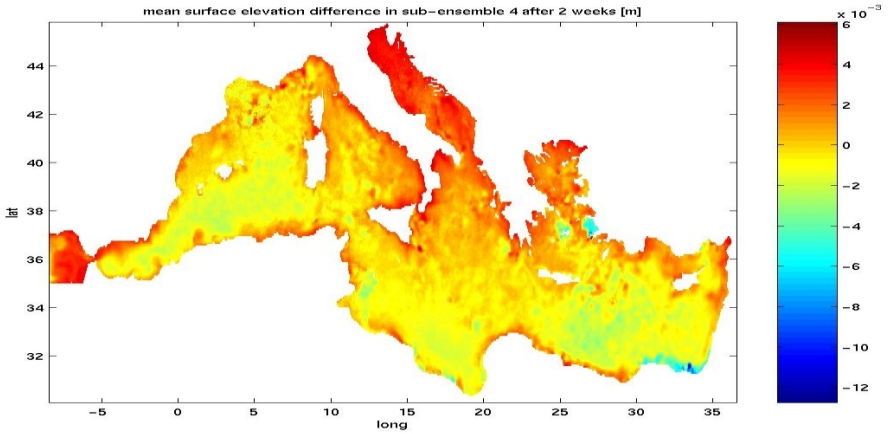

Figure 5.16: Mean sea surface elevation difference between the members in SE_AIRT and the reference run after 2 weeks of simulation

Perturbations on the cloud field or the diffusion parameters do not yield any distinct spatial structure in the surface elevation, temperature or salinity error. Errors caused by perturbations of the initial conditions also do not have preferred location, but correspond to the evolution of the initial errors. The air temperature perturbations have very small impacts on the sub-ensemble mean; surface elevation errors are more pronounced in shallow and coastal zones than in interior seas (see Fig. 5.16). The temperature errors are important along the Tunisian coast, the Ligurian Sea and the Aegean Sea (see Fig. 5.17).

By comparing the difference between the mean and the reference run, for each sub-ensemble, after 2 and 4 weeks, we confirm conclusions already stated in the previous section. The error (on T,S and elevation) caused by initial conditions, locates in areas of sizes corresponding to the evolution of the random error patches in the initial conditions; they are relatively stationary in space. The wind perturbation also causes stable T and S errors, although the elevation error is less constant. Errors appear in relative large areas (of the size of the wind perturbations) while accumulating along the African coast only after 4 weeks (see Fig. 5.15a,b).
The air temperature error causes error structures relatively stationary in space; the bathymetry perturbation leads to error patterns which acquire small scales with time, and are present around areas of large bathymetry modifications. The modification of diffusion coefficients causes errors concentrated in small zones, which relocate during the simulation, probably following specific features that diffusion affects the most. Finally, cloud coverage perturbation also leads to errors concentrated in small, moving zones, corresponding to perturbations on the cloud fields.

The variance plots, computed in each sub-ensemble, indicate the locations with the largest dynamical response to the corresponding perturbations. It appears that diffusion, initial condition, and to a lesser extent air temperature perturba-

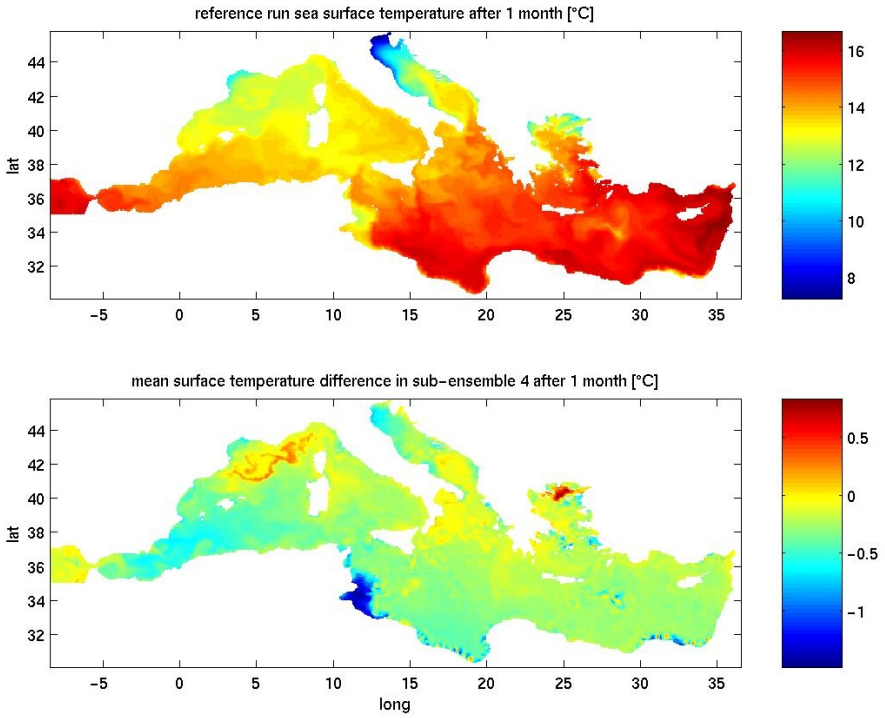

Figure 5.17: (a) Model output after 1 month of simulation: reference member sea surface temperature, (b) mean surface temperature difference between SE_AIRT members and the reference member

tions lead to patches of high variability in the surface elevation response (see e.g. Fig 5.18b). The higher variabilities in specific areas (Egyptian coast, Gibraltar, Aegean Sea and various patches) are probably due to intrinsic higher responses (shallow water etc); they are not due to higher mean random perturbations in these areas, as the temperature and salinity variance shows no such patches. Cloud coverage perturbations yield an almost uniform elevation variance map. The temperature variance is almost uniform for cloud coverage, air temperature, initial conditions, and to a lesser extent diffusion coefficient perturbations.
The surface elevation response to bathymetry perturbations is characterized by a higher variability in regions with large topographic gradients, where the bathymetry has been most modified (Fig. 5.18), see also above. Very clear examples of high variability visible in this figure are the patches south of Greece, around the Rhodes Island, North of Tunisia, in the Sicily Strait, between Italy and Sicily, and along the Algerian shelf break, all places with strong topography gradients. Variability is also high in the Atlantic Ocean box, but this is not reliable in our model. The corresponding temperature variability is shown in Fig. 5.19a. It exhibits a similar behavior, but high variability can now also be seen around all coastlines.
Modifying the wind forcing impacts the temperature most in some specific areas (see Fig. 5.19b), as well as in general in the Southern half of the Western Basin. Of course it mainly impacts the upper σ domain (Fig. 5.20a), although some perturbations reach the lower layers too. In fact, all sub-ensembles present a higher variance in the upper region; Fig. 5.20b shows this for SE_IC. In this sub-ensemble, the higher variability in the upper part is due to the fact that perturbations in the initial conditions are maximal at the surface, and decrease exponentially with depth, as explained in section 5.2.
Finally, let's note that all the above observations for temperature are also valid for the salinity, and that results in the small Atlantic Ocean box are not to be considered.

The skewness graphics are similar for temperature, salinity and surface elevation. The former two also present no clear structure in the vertical dimension; the whole 3D fields seem to behave similarly. In all sub-ensembles except SE_BATHY and SE_IC, the skewness also approximately has a zero mean value, which indicates that globally, the pdf distributions are symmetric. In SE_WIND, the entire skewness field is actually close to zero, with some (apparently random) pixels in the graphic having small non-zero values. The cloud coverage, air temperature and diffusion perturbations lead to pixelised graphics, as shown in Fig. 5.21b for SE_DIFFU.
Perturbing the initial conditions (SE_IC) leads to spatial structures corresponding to the initial perturbations, with a positive skewness, indicating a non-symmetric error pdf.
The bathymetry perturbation (SE_BATHY) leads to a different conclusion: if most of the field is 'pixelised" as for e.g. SE_DIFFU, the areas with largely modified bathymetry present systematic negative deviances (Fig. 5.21a), due to a more important left tail in the pdf. This clearly indicates non-Gaussianity of the error pdf in these zones.

Finally, all 6 kurtosis plots (not shown) present negative values over the entire domain, which are all relatively uniform over the whole fields (for all the vari-

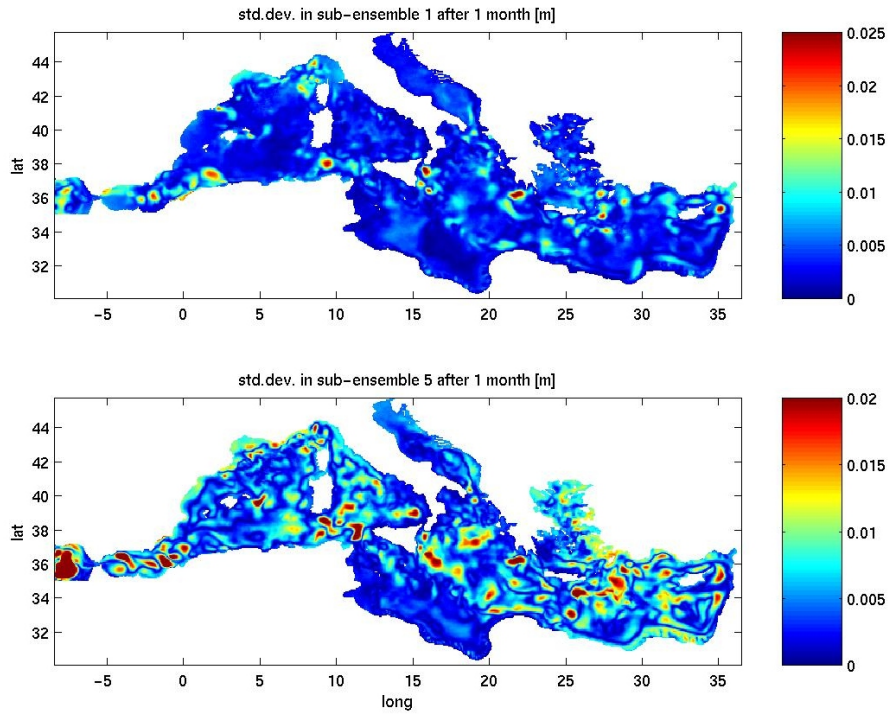

Figure 5.18: (a) Sea surface elevation standard deviation in SE_BATHY after 1 month (b) idem in SE_DIFFU [m]

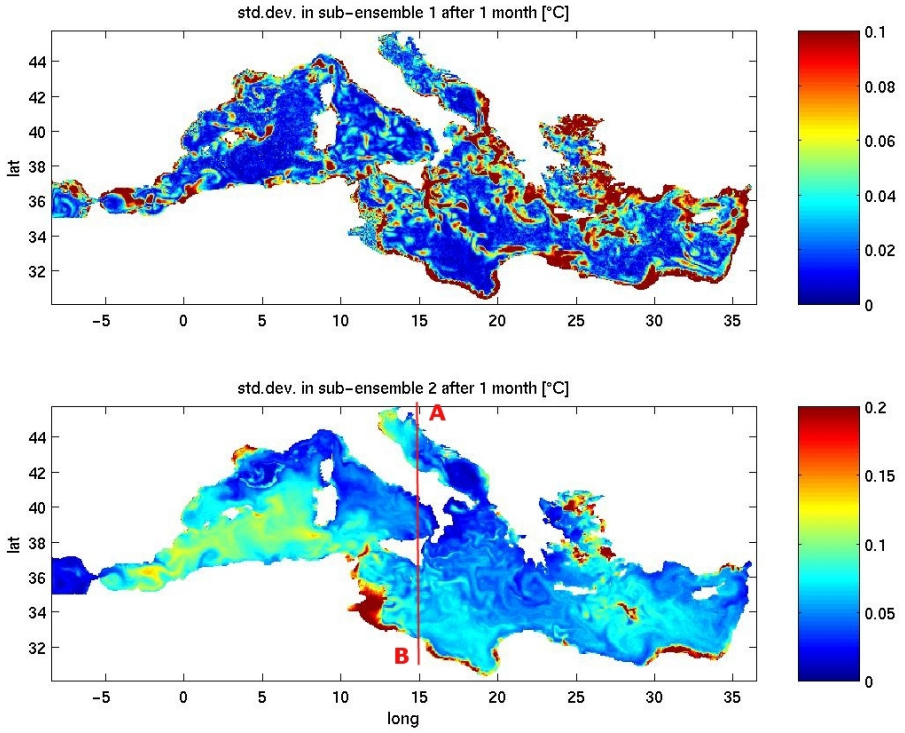

Figure 5.19: (a) Sea surface temperature standard deviation in SE_BATHY after 1 month [°C]. The colorbar is saturated at 0.1°C, even though some extreme values up to 0.5°C are present; (b) idem in SE_WIND, with the colorbar saturated at 0.2°C.

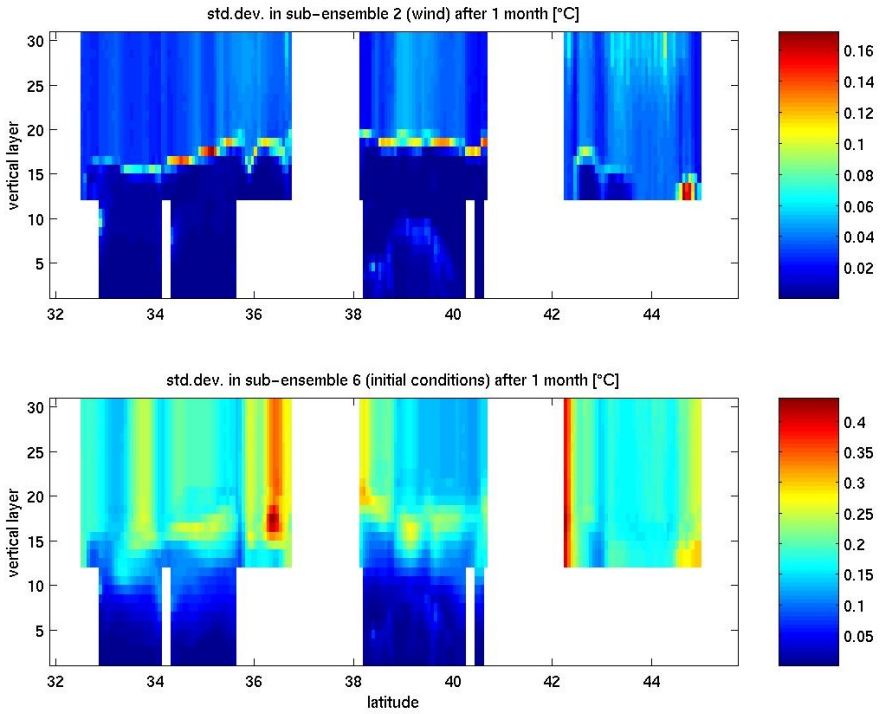

Figure 5.20: (a) Temperature standard deviation in SE_WIND after two weeks, along the horizontal line A-B indicated in figure 5.19 (b) idem in SE_IC. The graphic is drawn using the 31 vertical σ layers.

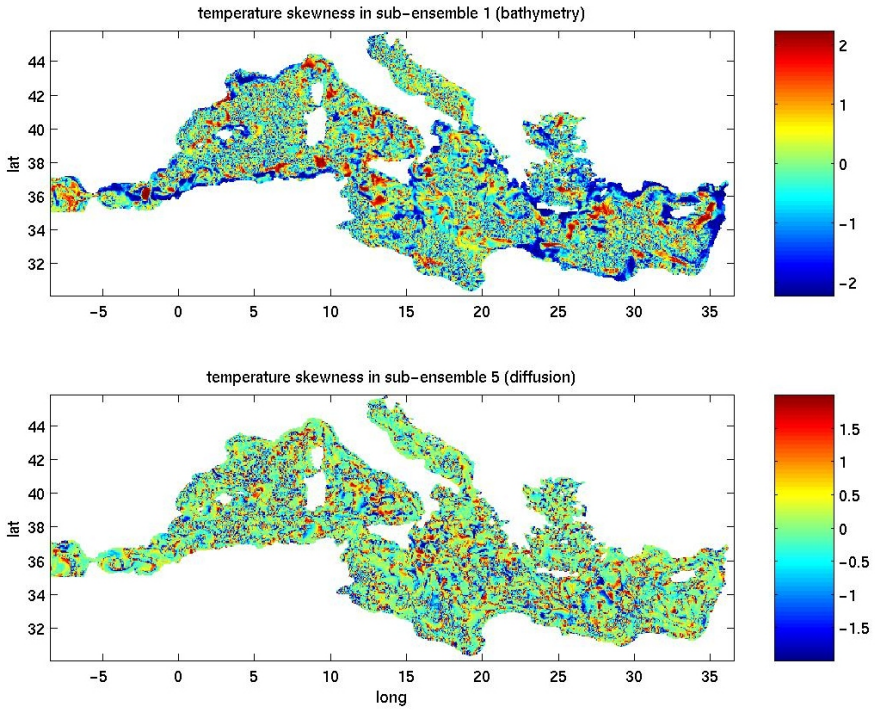

Figure 5.21: (a) Temperature skewness in SE_BATHY, (b) in SE_DIFFU [°C]

Sub-ensemble	EOF1	EOF2	EOF3	EOF4	EOF5
SE_BATHY	74%	11%	10%	3%	2%
SE_WIND	74%	20%	5%	<1%	<1%
SE_CLOUD	71%	3%	2%	1%	1%
SE_AIRT	99%	<1%	<1%	<1%	<1%
SE_DIFFU	68%	28%	3%	1%	N/A
SE_IC	10%	5%	4%	4%	4%

Table 5.2: Relative variance explained by the first sea temperature anomaly EOFs in each sub-ensemble

ables); there is no noticeable difference between the sub-ensembles. In principle, these negative values indicate that the error pdf is "less peaked" than the normal distribution, i.e. that the variance is due to more members (there are less "peak" members close to the sub-ensemble mean, and less members with large variance) than in the Gaussian distribution. However, it is not clear just how far the kurtosis is situated compared to a normal distribution kurtosis; since when the pdf is not exactly normal, its estimator is not unbiased. A reassuring property of the graphics (not shown) is anyway that all the variables present approximately uniform kurtosis values in the whole fields.

Anomaly EOFs

In order to further assess the spatial structures of the error, we will again compute central EOFs of each sub-ensemble, and examine if some physical process can be associated with them. Table 5.2 indicates the variability explained by each sea temperature anomaly EOF in the different sub-ensembles; the values for surface elevation and salinity are very similar.

In SE_BATHY, the first EOFs are shown in Fig. 5.22 for the temperature. It corresponds to variations in all the areas where the bathymetry is most modified (the coastlines around the Strait of Gibraltar, around Sicily, in the Aegean Sea, in the Rhodes area and in Egypt), which is coherent with conclusions given before. As this is still true after 1 month of simulation, we can furthermore affirm that areas with errors in the bathymetry will always be the subject of differences on the outputs: the spatial structure of the differences remains relatively stationary, as already stated before. These differences do not seem to spread out very much to other areas and to the rest of the basins; they rather represent the new equilibrium ocean state corresponding to the modified bathymetry. The remaining EOFs only contribute to the overall variance for a fraction of the first EOFs variance, and contain non-zero values only at some isolated places with abrupt bathymetry changes, mainly in the Aegean Sea or along the African coast.

In SE_WIND, the wind perturbation on the members also only leads to 2 significant central EOFs (for all model variables). This is probably due to the relative large scales present in the historical wind EOFs used to build the wind perturbation. The first surface elevation EOF, while its surface presents features already described in the previous discussions: its largest variations can be observed in the North-Western basin and in Adriatic. The second EOF has

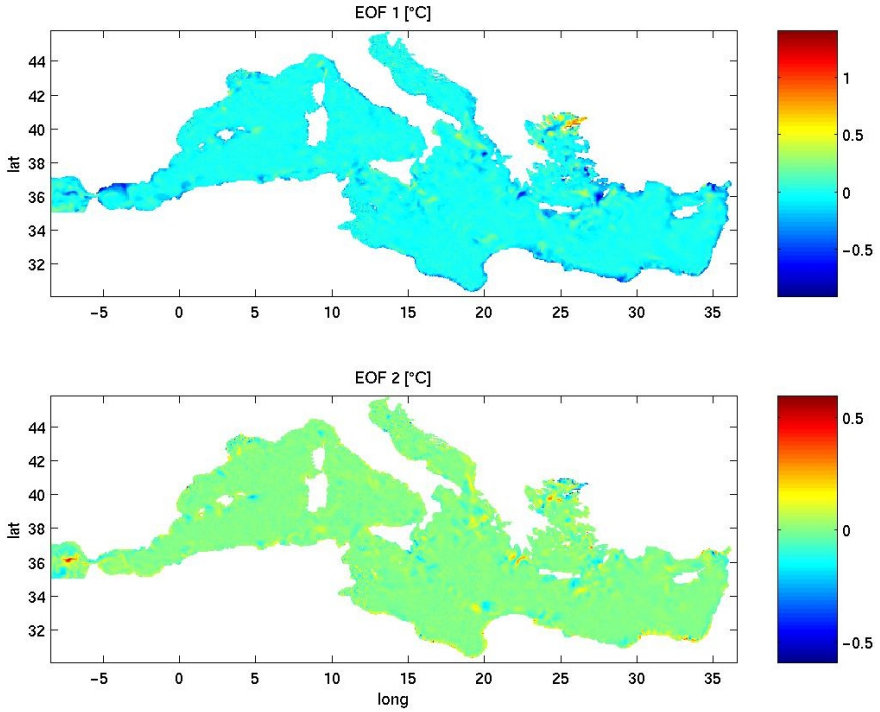

Figure 5.22: (a) First temperature EOF in SE_BATHY, computed on the sub-ensemble members after 1 month, (b) idem, EOF 2.

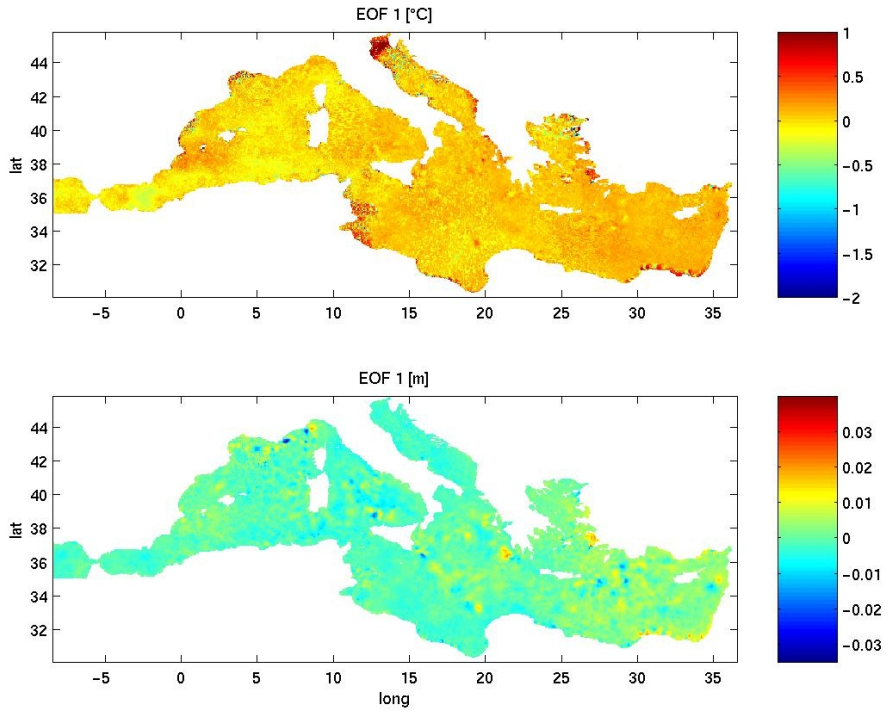

Figure 5.23: (a) First temperature EOF in SE_CLOUD, (b) idem for the sea surface elevation

high values along the coasts. The temperature and salinity EOFs vary the most in the Aegean. The EOFs are also qualitatively the same after 2 and 4 weeks.

The temperature in SE_CLOUD also varies mainly in the subspace spanned by the first EOF. The latter is characterized by the "pixelisation" of all the areas, and intense variations in the shallow Northern Adriatic, the Gulf of Lions, the Tunisian coast, and some other coastal areas (Fig. 5.23a). Its surface elevation counterpart varies following small-sized patches, indicating the creation of small gyres (Fig. 5.23b). The following EOFs, both for temperature and surface elevation, do not present much structure, but are rather generally pixelised.

SE_AIRT yields the same kind of 1st temperature central EOF as SE_WIND, i.e relative large variations in some well-defined areas (Northern Adriatic, Tunesian shelf, Egyptian coastline), and little variation in the remainder of the sea. EOF 2 and the following ones only add small, local corrections. The surface elevation EOF contrasts the shallow and the deep regions, with higher variation in the former than in the latter. The following EOFs again only contain small patches representing little structures. Their absolute value is smaller than the large-scale ones, indicating that they are probably present only in one or two members each.

Most of the variance in SE_DIFFU, for each variable, is again contained in

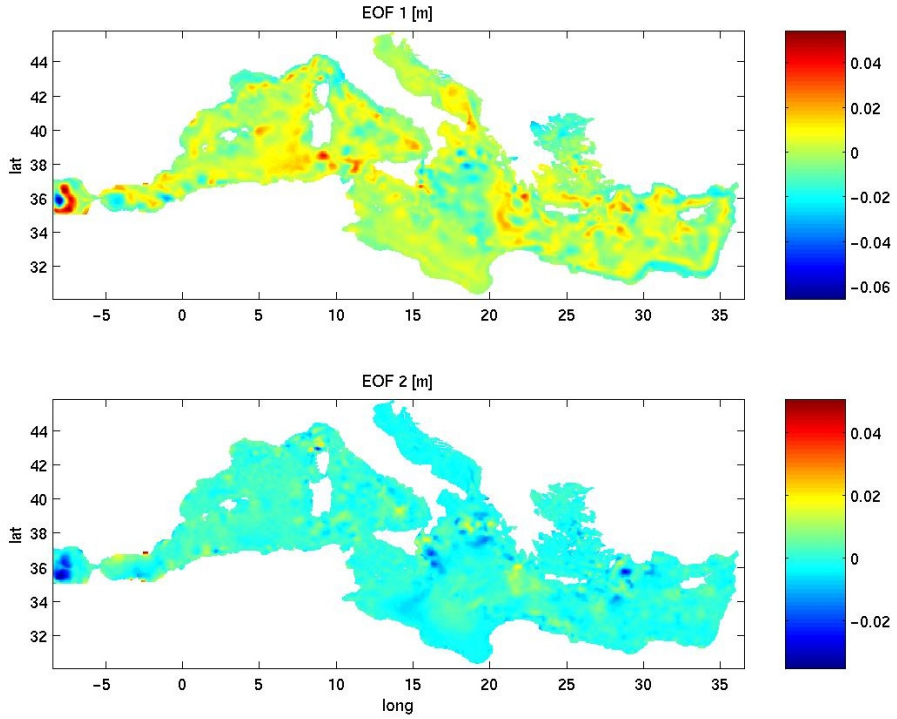

Figure 5.24: (a) First surface elevation EOF in SE_DIFFU, (b) idem, EOF 2.

the first central EOF, and to some extent in the second one. It contains a lot of small zones of relative small variation, corresponding to the increased (or decreased) diffusion of the structures present there. The first two surface elevation EOFs, accounting together for 95% of the variations in the sub-ensemble, are shown in Fig. 5.24.

Finally, the EOFs in SE_IC largely differ from the previous ones. The different random errors added to the initial condition in each member lead to different members at each time step. Therefore, no EOF can capture the variability; and each EOF has approximately the same importance (except the first one which captures about 10% of the variability). During the simulation, the present features seem to become somewhat smaller in space, and their variability is somewhat lower, indicating a progressive attenuation of the initial perturbations (Fig. 5.25). This is consistent with our observations in section 5.3.

5.5 Conclusions

The bathymetry perturbations, whether they come from badly known bathymetries or from voluntary smoothing for stability reasons, lead to important variations in the sea state in those areas where uncertainties or smoothing are the largest. The modified state corresponds to a new equilibrium, or trajectory

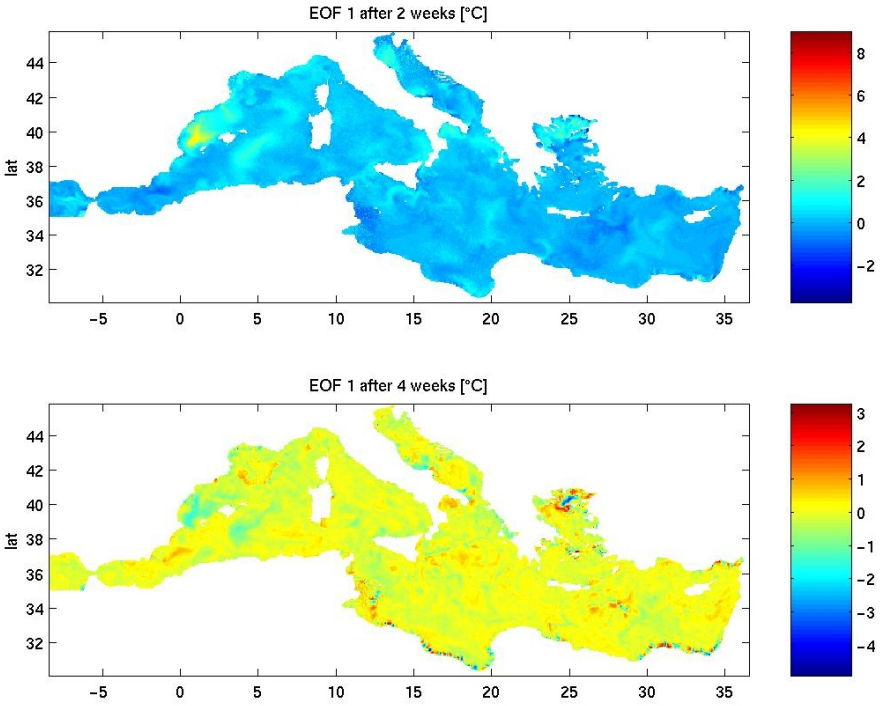

Figure 5.25: (a) First temperature EOF in SE_IC after 2 weeks, (b) idem, after 1 month.

in the state space, corresponding to the new bathymetry. It is important to notice that the system's response to bathymetry modifications does not yield a Gaussian pdf. This implies that data assimilation in these regions must be applied with great care. It does not come as a surprise; areas with canyons, shelf breaks etc. are known to be difficult to model correctly, hence the development of nesting procedures and specialised models in these areas.
Atmospheric perturbations lead to relatively large errors in the upper layers, particularly in shallow areas. Their impact is the largest on variables immediately linked (air temperature impacts sea surface temperature; wind speed most strongly impacts sea elevation); the other variables are modified only in a smaller extent, through the model dynamics. We mentioned that uncertainties on the wind might lead to attenuated or on the contrary to stronger gyres, even for basin-scale gyres. This is only an example of ocean state modifications. It is important to note that only with a relatively large amount of members (60) in the ensemble, this particular effect would (almost) vanish by averaging. Poorly known atmospheric fields yield the largest errors in the upper layers of the sea; and shallow areas are the most affected. In our simulations, the very large shelf east of Tunisia often presented very large uncertainties. This is a real a problem because very few observations are available in that area.
A careful estimation of internal model parameters, such as diffusion coefficients, is very important to obtain quantitatively correct outputs. Their influence is most critical in areas with fronts, gyres etc, since the largest gradients (which could more or less be diffused) are present there.
In agreement with numerous publications, it appears that a correct estimation of the initial conditions is essential; random errors are not superposable, do not disappear easily and are known to be very difficult to correct with data assimilation schemes, particularly reduced rank filters (including Ensemble filters), as the probability to represent the correct errors in the reduced error space is very small.
We obtained conclusions only at sub-basin scale even though we used a resolution of $1/16°$. This is due to the use of averaging statistics, smoothing out local effects and retaining only general trends. However, the moment fields and EOFs shown in the last part of this chapter also exhibit meso-scale properties, linked with local features, that should be studied carefully in order to fully benefit from the model resolution ($1/16°$).

Chapter 6

Statistical Predictions

I'm Popeye the Sailor Man
I'm strong to the finich
Cause I eats me spinach
Elzie Segar

In this chapter, we try to approximate the hydrodynamic model with a statistical model based on applied inductive learning methods, which aim at extracting a model of a complex system from the sole observation (or the simulation) of this system in some situations. These methods are also known under the fancy name of "knowledge discovery in databases" (KDD) methods. Of course it makes no sense to replace the hydrodynamic equations, based on physical and numerical considerations, by statistically-deduced relations. However, there are no such equations that link EOF weights between different time steps. We will focus on the latter objective, i.e. predicting the ocean state using a system of very low dimension. As the ocean state depends on the past states, but also on various forcings, we will include the most important one, the atmosphere, in our statistical laws. Thus, we will try to find a statistic relation between a vector containing n EOF weights at time t as well as m atmospheric EOF weights typical for the time between t and $t+1$, and the vector containing some EOF weights at time $t+1$. This relation can be expressed in various forms, i.e. a "black-box" equation obtained as a neural network, an ensemble of "if-then" rules obtained as a regression tree, or from the analysis of the most similar states in the past, a method called K-nearest-neighbours (K-NN). All 3 algorithms are introduced in the following section. We were given access to the Pepito software (see `http://www.pepite.be`), which contains these 3 methods, as well as a few other tools (principal component analysis, data clustering ...).

Tangang [1997] used a neural network model to seasonally forecast the tropical Pacific sea surface temperature anomalies (SSTA) in the Niño 3.4 region. The inputs to the neural networks were the first seven wind stress EOF amplitudes of the tropical Pacific for four seasons and the Niño 3.4 SSTA itself for the final season. At 6-month lead time, neural networks attained forecast skills comparable to the other El Niño-Southern Oscillation (ENSO) models. The SOFT project aimed at forecasting sea surface temperature using non-linear statistical techniques [Álvarez et al., 2000, Álvarez, 2003], and eventually assimilating these forecasts as pseudo-data in hydrodynamic models.

We will use a similar idea to forecast the ocean state: we decompose the present state as a weighted sum of EOFs (built in the previous chapter) and try to predict the weights of a future state. We also add the coefficients of the wind and air temperature EOFs. We did not consider the cloud coverage EOFs, nor the impact of uncertainty on diffusion coefficients or the bathymetry. We also do not consider other past ocean states. We will examine whether the ocean state can be predicted after 1 day, and after 1 week.

The constitution of the database of states is explained in section 6.2, and results of Pepito on the database are shown in section 6.3. Ideas for possible improvements are given in section 6.4.

6.1 Machine Learning

Wehenkel [2000], and the references herein, provides a description of the machine learning methods summarized below. *Knowledge discovery in databases* (KDD) is the non-trivial process of identifying valid, novel, potentially useful,

and ultimately understandable patterns in data. This data is usually organized in a database (DB), which is a collection of objects (in our case, ocean states) described by a certain number of attributes (in our case, the weights of the EOF decomposition of these ocean states, and also the weights of the EOF decomposition of the atmospheric fields typical for the time preceding the considered ocean state). Furthermore, we will define the learning set as the part of the database used to build the model, and the test set as another part of the database, used to test the obtained model. If no test-set is available, cross-validation techniques might be used to create one. *Data mining* is a step in the KDD process consisting of using an automatic learning (AL) method on a well defined problem in order to build patterns or models. In general, it comprises the subtasks of (a) representation (choosing attributes to characterize the objects), (b) selection (dismissing the attributes which are not believed to carry important information), (c) model selection (the choice of an *ad hoc* model and optimization technique), (d) interpretation and validation (the model is tested on a set of unseen objects, the test set), and (e) model use.

Automatic learning methods are subject to the general problem of over-fitting. It appears when the model is too complex with respect to the information provided in the learning set. For example, the complexity of a decision tree (see below) is proportional to the number of its terminal nodes, or the complexity of an artificial neural network (see below) is proportional to its number of weights. A model which over-fits the learning set will be suboptimal in terms of generalization capabilities. A familiar example is given by polynomial interpolation problems. Using too high a polynomial order will generally result in over-fitting, and the resulting function oscillates strongly.
The over-fitting problem can be interpreted in the following way. The estimation error by a model might be decomposed as the sum of a bias term, a variance term and a residual term. The last term does not depend on the particular automatic learning method used nor on the training set size; it however puts a lower bound on the error rate which may be obtained. Using an automatic learning method to derive successive models from successive learning sets (of same size) will produce different models, leading to different decision rules. This phenomenon is called model variance in the statistical literature: the smaller the learning set, the larger the variance. Furthermore, it is true that if we increase the size of the hypothesis space allowing the AL algorithm to search among more models, variance will increase to some extent. Normally, the variance vanishes for very large training set sizes.
In addition to variance, most models derived by AL algorithms present also some bias for finite learning set sizes. Many methods remain biased even when the learning set becomes very large! For example, using a linear decision rule to approximate a nonlinear problem leads to bias. Using a decision tree with a fixed number of nodes to approximate a complex relationship also leads to bias. Increasing the space of the models in general allows to decrease bias. In other words, if the learning algorithm is well designed, bias will decrease when the model complexity increases; but in the same time, the model will depend in a stronger fashion of the random nature of the learning set, thus its variance will increase. As both bias and variance lead to generalization errors, it is necessary to define a tradeoff (see Fig. 6.1). Thus, the design of "good" learning algorithms consists of two complementary aspects: (i) find model spaces

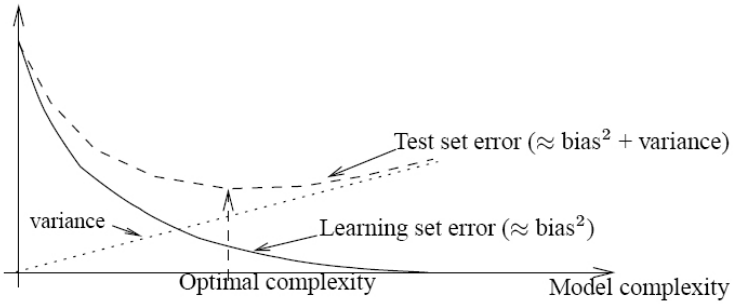

Figure 6.1: Bias/variance tradeoff via cross-validation (neglecting the residual error). From Wehenkel [2000].

which allow to reduce bias significantly without increasing variance too much, (ii) define the optimal model complexity (leading to an optimal bias/variance trade-off) for a given learning problem and a given learning set size. Geman et al. [1992] applied this study in the case of artificial neural networks.

In practice, for a given automatic learning method, bias is an increasing function of the physical problem complexity, not of the learning set size. Variance on the other hand is a decreasing function of the learning set size, thus, the optimal model complexity will be an increasing function of both problem complexity and learning set size. An important consequence is that if we know that a problem is simple (e.g. linear), we should decrease the model space size since by doing this we can reduce variance without increasing bias. Nonlinear techniques such as neural networks should be used only if we have reasons to believe that our problem is indeed nonlinear, and if we have enough data to ensure low variance. The cross-validation technique consists of learning a sequence of models of growing complexity, then using an independent test set to evaluate their generalization capabilities and finaly selecting the best one. The drawback is the computational overhead. Inversely, cross-validation may also be used in order to prune models of large complexity. The pruning starts with a very complex model, which is supposed to be over-parameterized. Then, a sequence of models of decreasing complexity is obtained by progressively simplifying this model; cross-validation selects the best one.
Besides cross-validation, other techniques exist for choosing the optimal model complexity. Regularization consists of modifying the square error criterion in order to penalize models which are not smooth enough, e.g. by adding a term proportional to the model curvature. Model averaging consists of building several models and aggregating their outputs in the form of an average, thus reducing variance. Other particular techniques exist for particular methods, such as the very efficient stop-splitting criterion used in decision trees (see below).
The methods that we briefly introduce below are classified as "supervised learning" methods: they have the general objective of finding a relation between objects. Unsupervised learning, which is not considered any further here, consists of finding relations between objects, without *a priori* specified objectives.

6.1.1 K-NN

The nearest neighbor is a very intuitive method, similar to the human approach where outputs are determined by similarity with past situations. There is no learning phase in the algorithm. When an output has to be computed for a new object, the object which is the closest to the new object is selected in the learning set, and its output y is taken as the solution. Of course, this requires to define a distance in the attribute space. If the i attributes a of the objects $\mathbf{o}$ are normalized, this distance could be the euclidian norm, e.g. between objects m and n:

$$\Delta = \sum_i (a_i(\mathbf{o_m}) - a_i(\mathbf{o_n}))^2 \tag{6.1}$$

Asymptotically, when the learning set size $N \to \infty$, the nearest neighbor converges towards the new object, and the output has the expected asymptotic probability of being correct. However, in real-life problems with finite learning sets (even large ones), the nearest neighbor rule may be rather suboptimal. The first approach to solve this problem consists of reducing the locality of the nearest neighbor information by using more than one neighbor. The leads to the K-NN rule. It consists of searching for the K nearest neighbors; the output $\mathbf{r}$ may then be estimated e.g. by

$$\mathbf{r}(\mathbf{o}) = \frac{\sum_{\mathbf{o}' \in KNN(LS)} y(\mathbf{o}')\Delta^{-1}(\mathbf{o}', \mathbf{o}')}{\sum_{\mathbf{o} \in KNN(LS)} \Delta^{-1}(\mathbf{o}', \mathbf{o})} \tag{6.2}$$

where we have denoted by $KNN(LS)$ the set of the K nearest neighbors of $\mathbf{o}$ in the learning set.

Further improvements to the nearest neighbor technique consists of editing the learning set by removing the learning states which are surrounded by many states (all with a similar output), that have a strongly different output compared with the considered learning state. Condensing algorithms may be used to dramatically reduce the size of the required database, by removing the states which do not contribute to the computed output. However, these techniques also reduce the locality of the nearest neighbor method, which is a desirable practical feature of the 1-NN method.

The nearest neighbor rule is very simple and easy to implement. Its main disadvantage is that it requires a very large number of learning states in order to become robust, particularly in the case of a high dimensional attribute space. Thus, it might be interesting to use another method to preselect a limited amount of attributes (i.e. putting a zero weight on the other attributes in the definition of the distance).
Although there is no learning phase (except the creation of the database containing the learning set objects), the complexity when using the method is directly proportional to the product of the number N of learning objects and the dimension of the attribute space. This may be several orders of magnitude higher that the time required by competing techniques and only rather sophisticated search algorithms can allow us to reduce the CPU time.
Further details about the KNN method can be found in Devijver and Kittler [1982].

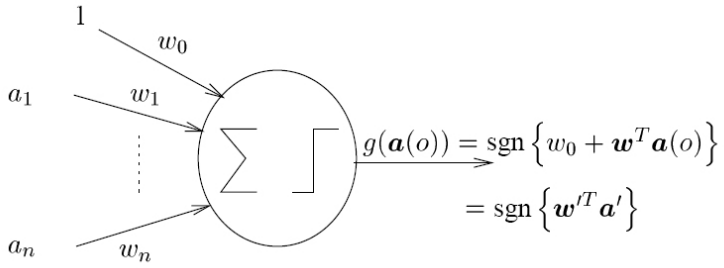

Figure 6.2: Basic linear threshold unit. From Wehenkel [2000].

6.1.2 Artificial neural networks

The learning systems based on artificial neural networks (ANN), which are essentially a nonlinear regression technique, became popular only in the early eighties; they have a much longer research history and some of these methods have evolved towards quite mature techniques. The early work even dates back to the 1940's, and aimed at reproducing the human learning abilities by developing a low-level model reflecting the biological structure of the brain. A second wave of research reached its peak in the early sixties with the perceptron learning theorem of Rosenblatt [1963], which proved that the simple on-line error-correcting learning algorithm proposed would converge in a finite number of steps if the patterns are linearly separable. Negative results however brought a stop to the enthusiasm. Finally, the present wave has started from the conjunction of the rapid increase of available computational resources, theoretical work advances and the improvements of multilayer perceptrons culminating with the republication of the back-propagation algorithm by Rumelhart et al. [1986]. Cybenko [1989] showed that an ANN with a single hidden layer of neurons with large enough a number of neurons can arbitrary well approximate any continuous function. A vast bibliography exists on the subject. An overview is e.g. given in Haykin [1994].

The most simple version of the ANN is constituted by a single linear threshold unit, represented in Fig. 6.2. It is a hyper-plane model, separating the space in two regions with two corresponding different possible outputs, defined by the discrimination function

$$g(\mathbf{a}(\mathbf{o}) = \mathrm{sign}(w_0 + \mathbf{w}^\mathbf{T}\mathbf{a}(\mathbf{o})) \tag{6.3}$$

which assigns a value of -1 or +1 depending on which side of the hyper-plane the attribute vector is located. Of course this model can only be used for so-called classification problems, where only a single boolean output is needed.
The perceptron learning algorithm works iteratively, with successive passes through the learning states, where the weights are adjusted at each step so as to improve the output of the current object.

1. Consider the objects of the learning set in a cyclic or random sequence

2. Let $\mathbf{o}$ be the current object, $y(\mathbf{o})$ its target output and $\mathbf{a}(\mathbf{o})$ its attribute vector

3. Adjust the weight by using the following correction rule:

$$\mathbf{w'^{new}} = \mathbf{w'^{old}} + \eta(\mathbf{y}(\mathbf{o}) - \mathbf{g}(\mathbf{a}(\mathbf{o})))\mathbf{a'} \tag{6.4}$$

If the output is already correct, eq. 6.4 shows that the correction is zero. The prime symbol is used to indicate that we consider the extended vector $\mathbf{w'} = [\mathbf{w_0} \quad \mathbf{w}]$ including an independent term. η denotes the learning rate of the algorithm. It can be shown that if the learning set is separable, then the above rule converges to a solution in a finite number of steps, but the speed of convergence may depend on the precise value of η. If the learning set is not separable, the algorithm will never stop changing the weight vector. Thus, one of the techniques used to ensure convergence consists of using a decreasing sequence of learning rate values.

An immediate generalization of the linear perceptron is a single layer of perceptrons, which is able to learn a more complicated output than a single boolean. Other, continuous and smoother transition functions are considered instead of the *sign* function, such as the sigmoid and hyperbolic tangent functions

$$\text{sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{6.5}$$

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{6.6}$$

Considering output values varying continuously between -1 and +1, the learning objective of finding the weight vector $\mathbf{w'}$ is to minimize the mean square error SE,

$$SE(\mathbf{w'}) = N^{-1} \sum_{o \in LS} |g(\mathbf{a}(\mathbf{o})) - y(\mathbf{o})|^2 \tag{6.7}$$

The gradient of SE with respect to the weight vector is given by

$$\nabla_{\mathbf{w'}} SE = 2N^{-1} \sum_{o \in LS} (g(\mathbf{a}(\mathbf{o})) - y(\mathbf{o}))\mathbf{f'}(\mathbf{w'^T}\mathbf{a'}(\mathbf{o}))\mathbf{a'}(\mathbf{o}) \tag{6.8}$$

where $\mathbf{f'}(.)$ is the derivative of the activation function and $\nabla_{\mathbf{w'}}$ is the gradient in the weight space. Thus, using a fixed step gradient descent approach for minimizing the mean square error, in a sequential object by object correction setting, would consist of using the following weight update rule

$$\mathbf{w'^{new}} = \mathbf{w'^{old}} - \eta \nabla_{\mathbf{w'}} SE(\mathbf{o}) \tag{6.9}$$

$$= \mathbf{w'^{old}} - \eta[y(\mathbf{o}) - g(\mathbf{a}(\mathbf{o}))]\mathbf{f'}(\mathbf{w'^T}\mathbf{a'}(\mathbf{o}))\mathbf{a'}(\mathbf{o}) \tag{6.10}$$

where $SE(\mathbf{o})$ is the contribution of object $\mathbf{o}$ in eqn. 6.7. Thus, it is analog to the learning rule of a single perceptron, where the learning rate η is adapted proportionally to the derivative $\mathbf{f'}$ of the activation function.
An alternative, batch learning approach consists of computing the full gradient of the SE with respect to the complete learning set before correcting the weight

vector. A further improvement would then consist of using a variable step gradient descent method, for example the steepest descent approach, leading at each time step to the maximal decrease of the SE criterion. Other more sophisticated numerical optimization techniques may be thought of; all leading at each iteration to a decrease of the error function until a (local) minimum of the error function is reached. The particular local minima of the error criterion (which usually presents many) to which the search technique will converge mainly depends on the initial weight values. Thus, in order to increase the probability of finding a global minimum, the procedure should be repeated with various randomized initial weight values. Other initialization techniques also exist.

Multilayer perceptrons (MLP) are composed of several layers of soft threshold unites, interconnected so as to enable the approximation of arbitrary nonlinear relationships between inputs and outputs. The corresponding extension of the gradient descent learning algorithm is the back-propagation algorithm; its name stems from the fact that the gradient (with respect to the weights) of the MLP output error is computed by propagating information backwards through the MLP. It is an efficient algorithm, with a complexity linear in the number of weights. It is listed in e.g. Wehenkel [2000]. It sometimes uses a supplementary regularization term in the mean square error cost function, allowing to avoid high frequency components in the input/output mapping so as to reduce over-fitting problems. This approach is also called weight decay, because in practice it penalizes large weight values. The slow back-propagation algorithm can also be replaced by more efficient iterative Newton-like methods for finding the minimum of the error function, or algorithms based on the second derivate of the function. Famous examples are the Broyden-Fletcher-Goldfarb-Shanno (BFGS) and Levenberg-Marquadt algorithms. The latter is also implemented in the Pepito software.

6.1.3 Regression trees

A regression tree consists of an ensemble of "if/then" rules defined on the space of possible attributes, which lead to a partition of the database. Indeed, each interior node carries a test defined on the possible values of an attribute, with 2 possible outcomes. The terminal nodes then each correspond to a sub-ensemble of the database, and all together they constitute a partitioning of it.
During the learning phase, the tests on the nodes are usually chosen from all possible tests of the form

$$a_i(\mathbf{o}) \leq t \tag{6.11}$$

in order to minimize the variance in both subsets. t is a threshold to be determined. More sophisticated tests can also be used, involving e.g. linear combinations of attributes; but they require a lot more computations during the learning phase. The tree growing phase stops when the nodes are considered "pure" enough (the output variable is constant enough in the terminal nodes). However, this may lead to over-fitting, e.g. when the subset of the learning set corresponding to the nodes becomes too small, with sometimes only one or two objects per node. The statistical information obtained from the tree is not reliable any more. There thus exist a trade-off between the two sources of error: bias which results from insufficient splitting and variance which is a

consequence of too much splitting. In the most recent approach, the trees are first fully grown, then the over-fitting problem is considered in a post-processing stage. Tree pruning is a form of "smoothing" obtained by removing the over-specified parts of the tree in a bottom up fashion. It requires to define the quality measure of a tree T, in the form of

$$Q_\beta(T, \mathbf{LS}) = R(T, \mathbf{LS}) - \beta C(T) \tag{6.12}$$

with R a reliability measure (e.g. the amount of objects with the correct output, or the amount of variance reduction), and C a measure of the complexity of the tree, e.g. the amount of terminal nodes. β is a parameter that gives the relative importance one attaches to complexity vs. reliability. In particular, for $\beta = 0$ the pruned tree will be the full tree, for $\beta \to \infty$ the pruned tree shrinks to a single node. With this quality measure, for a fixed β we can extract the optimally pruned tree such that the quality is maximal. Moreover, with β growing from 0 to ∞, the sequence of optimally pruned trees can be constructed efficiently as each tree in contained in the previous one. Each of the obtained optimally pruned trees should then be tested with an independent "pruning set" (not containing the learning set nor the test set) in order to select the final tree in the sequence.
During the test phase, the tree is then simply used by directing the test object towards the appropriate terminal node. At each test node a particular attribute value is tested and the walk through the tree stops as soon as a terminal node is reached. This will correspond to the elementary subset in the attribute space comprising the object and the information store there (i.e. expected value of the output variable) is extrapolated to the current object.
Sophisticated generalizations of these trees have been proposed, such as so-called fuzzy trees, equivalent to constructing different trees and then aggregating them. As an example of fuzzy trees, two thresholds might be defined for a single test. For values of the considered attribute in between the two thresholds, no decision is taken, and the objects follows both paths down to the terminal nodes. Its final output might then be an average of the obtained outputs in the different terminal nodes. It has been shown that fuzzy trees might at the same time be more accurate than their "crisp" counterpart, and also more stable with respect to random fluctuations of their learning set.

An important advantage of the tree methods is their intuitive interpretation, as the results are expressed in a similar fashion to the knowledge which can be formulated by a human expert. This allows to decide which attributes are the most important in the final decision. Therefore, the method is often used in combination with other methods. For example, a tree might used to select the most important attributes, which then serve as inputs to an ANN.

6.2 Inputs and Outputs

The simulation outputs of the ensemble run described in the previous chapter are used to test the automatic learning methods presented above. As mentioned previously, we try to forecast the coefficients of ocean state EOFs based on the previous coefficients as well as on atmospheric EOF coefficients.

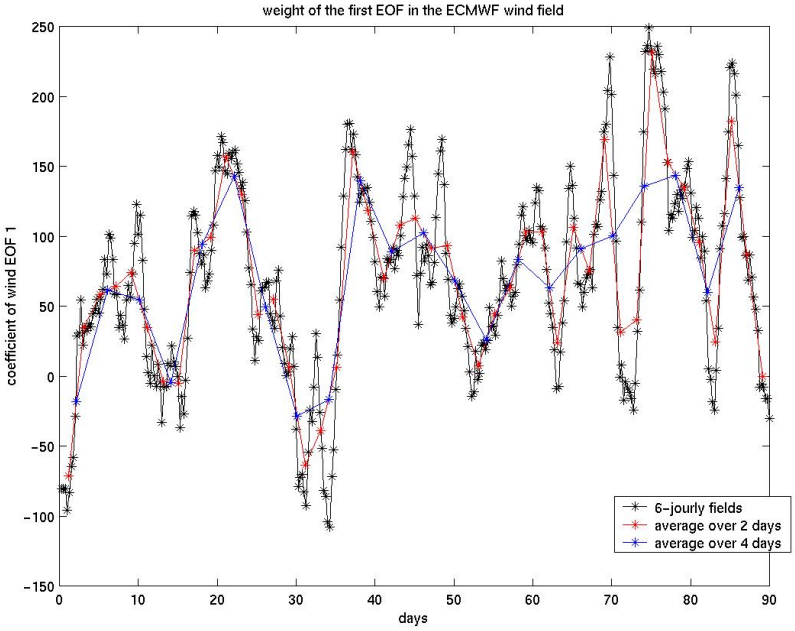

Figure 6.3: Coefficients of the 1st wind EOF in the ECMWF 6-hourly fields: instant fields, averaged over 2 days, and averaged over 4 days.

The coefficients of the 6-hourly ECMWF atmospheric fields EOFs were readily obtained; the EOFs having already been computed before to create the modified atmospheric forcings (see chapters 3 and 5). We noted that with 10 EOFs, we could explain about 80% of the variability in the wind field at each timestep (see also Fig. 3.8 page 63). The first EOF already sufficed to explain over 90% of the air temperature variability. Furthermore, as we will try to forecast the ocean state after 1 day and 1 week, 6-hourly fields (4 per day) would yield 4(10+1) and 28(10+1) numbers respectively, which represent a lot of inputs. Although the AL algorithms should in principle extract the needed information from all these inputs, a considerable amount of learning states might be needed. Thus, we examined if it is possible to use averaged atmospheric fields rather than instanteneous ones; we thus need to know a typical variation-time of the atmospheric fields. Fig. 6.3 shows the "instant" coefficients of the first wind EOF, as well as their average over 2 days and 4 days respectively. The loss of information seems acceptable when averaging over 2 days, but this is about the maximum allowed: with averages over 4 days, the signal is largely missed. Thus, we could average the wind field over the day (when predicting the ocean state after 1 day) or use 3 coefficients for each EOF (when predicting after 1 week).

When forecasting the ocean state, there is no reason anymore to distinguish between the different sub-ensemble EOFs, where different perturbations have been applied. The EOFs should rather be computed from all the members of

the ensemble together. We also decided to compute multivariate EOFs comprising temperature, salinity and surface elevation. These variables are normalized by the corresponding grid cell size and by their own standard deviation; which is 18.57 times larger for temperature as for surface elevation, and 2.04 times larger for temperature as for salinity. The weight of the surface elevation part in the state vector is then artificially increased by a factor 5 in order to give more importance to this variable. The weight is also taken close to zero in the Atlantic Box.
However, the memory space required for the computation is rather huge. If N_{sea} is the amount of grid points in the state vector, N_{time} the amount of instants at which data is available, and N_{eof} the desired amount of EOFs, the computer memory should be equal to $N_{sea} \cdot N_{time} + N_{sea} \cdot N_{eof} + 3 \cdot N_{sea}$, as well as some additional temporary memory space. In order to allow computation of principal components from larger matrices (less temporary memory space is required), and to decrease the needed computation time, we used the Lanczos algorithm proposed by Toumazou and Cretaux [2001] to compute the EOFs, rather than the classic SVD algorithm. Some temporary variables, such as the counter summing the whole original matrix and which is used to compute the variability explained by each EOF, also had to be declared in double precision.
In our case, there are 1.8 million grid points in the 3D grid, and about 64 thousand in the 2D field (used for surface elevation). The state vector comprising temperature, salinity and surface elevation thus requires about 15 MB of memory in simple precision. All by all, when trying to calculate the 50 first EOFs from 3 weekly outputs of 250 members, 11.6 GB of memory would be required. We decided to reduce our effort to computing the EOFs from only 125 members. The members of the smaller sub-ensembles (where the bathymetry or diffusion coefficients are modified) are all considered; the other ones are chosen randomly from the remainder of the complete ensemble. This still leads to 6.3 GB of required memory, which is feasible on the new supercomputer from the SEGI / ULg, some of its nodes having 32 GB of memory.

The EOFs computed are not anomalies with respect to a central reference member as in chapter 5. Instead, we considered all 125 members and removed their common temporal mean. The resulting average surface salinity and elevation fields are shown in Fig. 6.4 as examples.
The effect of the basin-scale circulation features described in section 5.1 can be seen in the figure. For example, the current along the Algerian coast generates 3 large anticyclonic eddies. We have stated before that only the anticyclonic eddies are long-lived, while cyclonic eddies disappear rather rapidly and thus are not present on the average map. In the Ligurian Sea, the general cyclonic circulation leads to a lowering of the center of the basin. The same is visible in the Eastern Basin, while a large anticyclonic zone is present off the Libyan coast. Apart from the basin-scale circulation, features are also visible at smaller scales and with smaller intensities (e.g. in the Aegean Sea), although they tend to be averaged out. The temperature (not shown) and salinity averages also clearly show the impact of the large-scale circulation. Relatively fresh water flows in through Gibraltar, which gets saltier all the way to the Eastern Basin. The Ligurian Sea is also saltier than the waters along the Algerian coast, while the Gulf of Lions is less salty than the offshore waters. The temperature is warmer in the Eastern Basin than in the Western Basin, and in the latter, it is warmer

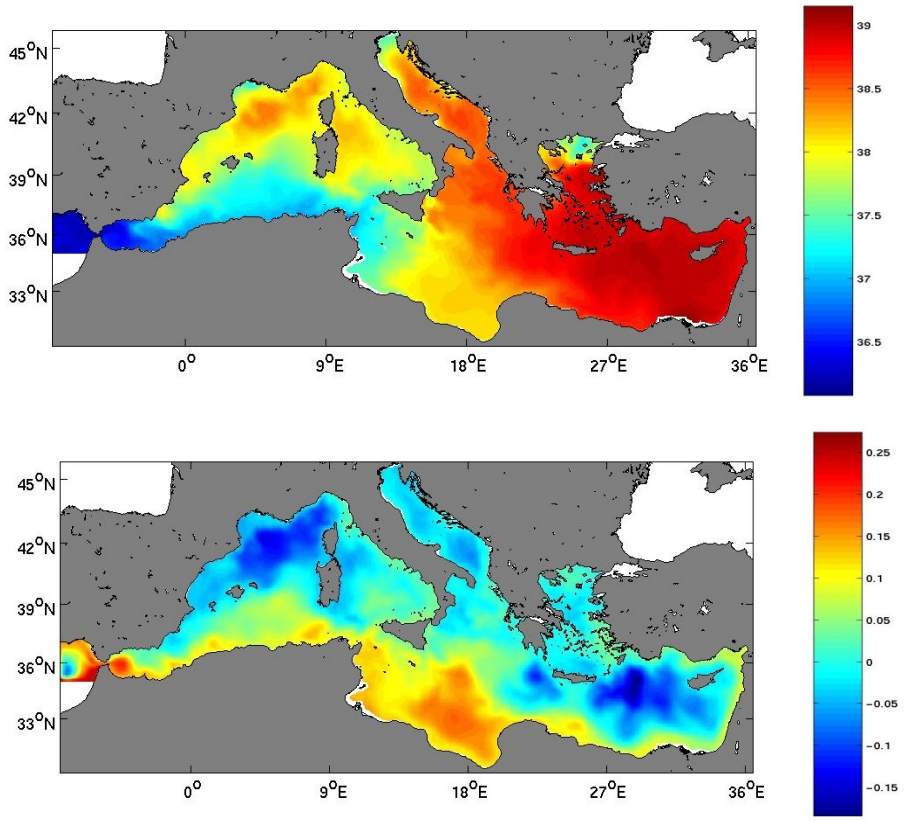

Figure 6.4: Mean ocean state in the ensemble: (a) surface salinity [psu] (b) surface elevation [m]

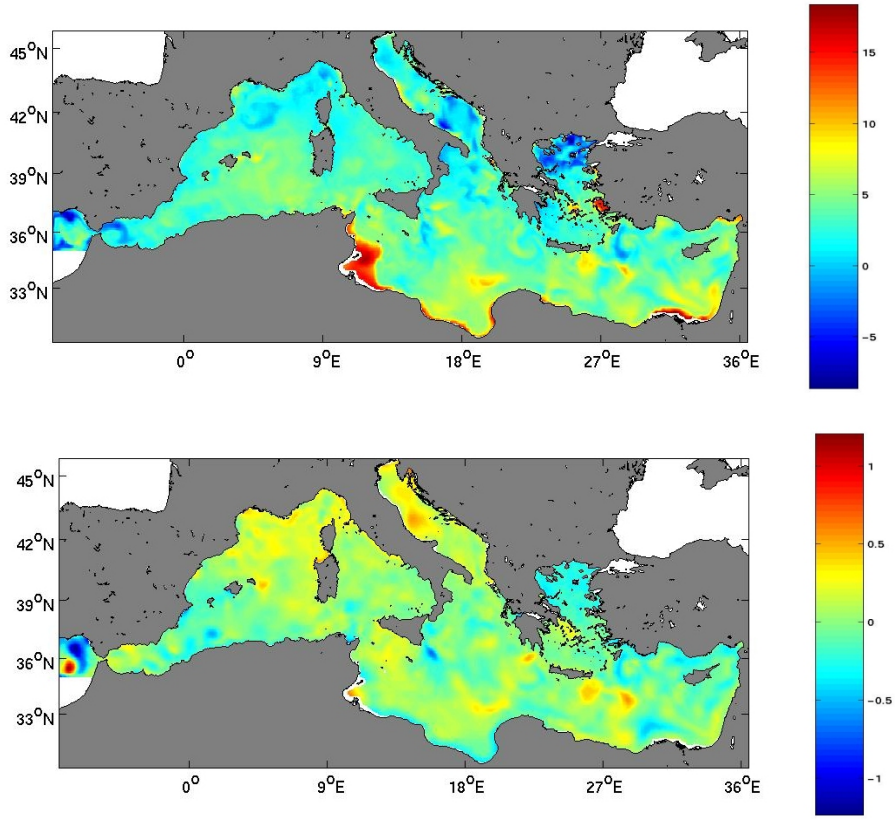

Figure 6.5: First multivariate EOF computed over the ensemble, (a) surface temperature [°C], (b) surface elevation [m]

in the South than in the North. The coldest waters are located in the Northern Adriatic, which is a very shallow area subject to cold winds during the winter.

The first multivariate EOF accounts for 67% of the total variability. Its part concerning surface temperature and surface elevation are shown respectively in Fig. 6.5 a and b. This EOF is not associated with a single type of perturbations. If anything, it is probably most determined by the time evolution of all the members. For the temperature part, this results in an average warming or cooling of the whole basin, except the areas around Gibraltar, the North of the Ligurian Sea, the Adriatic Sea and its neighborhood in the Ionian, and the Aegean Sea, which all get the inverse sign on average. The largest values, on the Tunisian shelf and along the Egyptian coast, have similar shapes to the "errors" observed when the air temperature is perturbed. The patterns in the Ligurian Sea and the Aegean Sea also look similar to the ones in the EOFs computed only in that sub-ensemble. However, perturbing the wind field also leads to large variability along the Tunisian and Egyptian coasts, as well as a similar high-variability area along the Turkish coast in the Aegean Sea (see 5.19). Finally, bathymetry perturbations also lead to higher variability in areas with the

most perturbed topography, including the areas mentioned here. However, since other areas with large topographic gradients are not so present in the first EOF, bathymetric uncertainties certainly are not the major cause. The salinity part (not shown) of this first EOF is very similar to the temperature part, except that the Tunisian shelf area is less variable. The patterns are also different in the area East of Gibraltar. The surface elevation is mainly characterized by the intensification or weakening of some important gyres, such as the Ierapetra and Mersa Matruh in the Eastern Basin, and thus corresponds more to the "normal" time evolution than to any specific perturbation. We already conclude that the prediction of the coefficient corresponding to the first EOF will be particularly important, since it stands for the time evolution of the system.

The second EOF, which accounts for only 7.6% of the total variability, is again difficult to associate with any given type of perturbation. The largest contribution in the surface elevation part of this EOF is brought by the Mersa Matruh, the southern part of the Aegean, and the Libyan coastline. Other gyres also are represented. The Western Basin is globally slightly lifted, but local details are visible. The salinity and temperature parts show smaller scale structures than in the first EOF, but their origin is difficult to assess.
The third EOF still represents 3.2% of the variability (the 4th is 1.4%, and the remaining EOFs account for less than 1% each, with EOF 44 and following accounting for less than 0.01%). The effects of time-evolution and the various perturbations all sum up and are not distinguishable in the EOFs.

The ocean states of all the (250) members at various instants were projected on the EOFs. Then, a database was built, where each line starts with the member number and the start time t (those 2 numbers are used only for easy selection of lines in the database, but not by the AL methods). Each line also comprises (a) the 25 coefficients of the ocean state at time t (b) 11 or 33 averaged coefficients of the wind and temperature atmospheric fields in between times t and $t+1$ and (c) the 25 coefficients at time $t+1$. The delay between t and $t+1$ is 1 day or 1 week (with 11 or 33 atmospheric coefficients respectively).

As observed in chapter 5, when considering only members with perturbed initial conditions, it is impossible to build significant EOFs. Thus, when all kinds of perturbations are mixed together in a database, there will be no representation of the (evolution of) initial condition modifications in the obtained EOFs. This implies that, in our database, the lines corresponding to members with perturbed initial conditions rather add noise to the database than add information. They essentially duplicate the lines corresponding to the unperturbed member, with some added noise. If this noise is realistic, this is a desirable feature that will help us not to over-fit the other data.

Fig. 6.6 represents the ocean states of all 250 members at initial time, and after 1, 2 and 3 weeks, in the space of the first 3 EOFs. Each member is represented by a dot, with a color according to the sub-ensemble is comes from. We notice that the perturbation of initial conditions yields a spread in the initial conditions, corresponding to the cloud of green points noted "I.C." in the figure (all other members having the same initial conditions, they are represented by a single bold black dot in the middle of the "I.C." cloud). The extent of this cloud

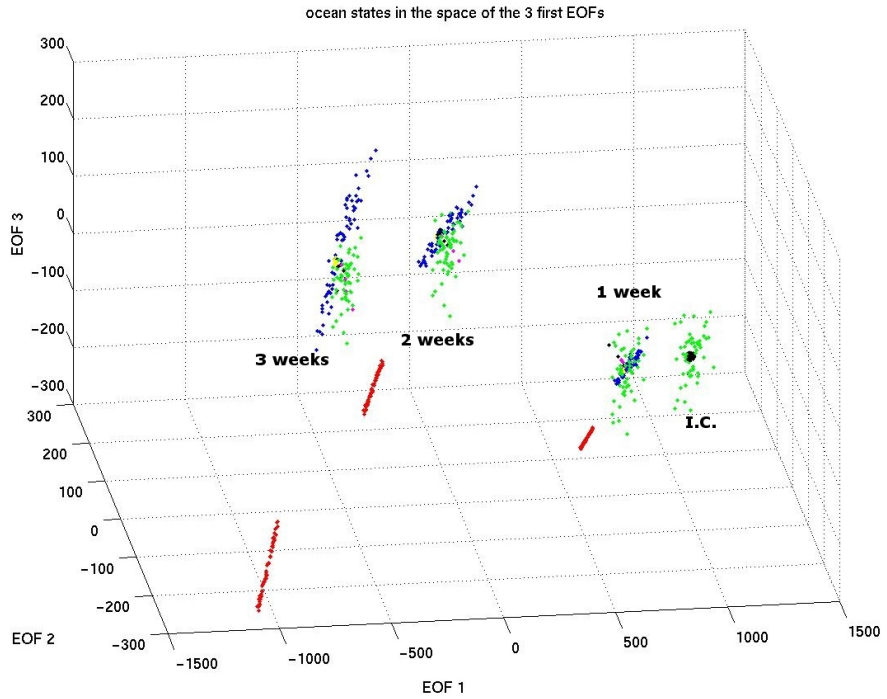

Figure 6.6: Ocean states represented in the space of the 3 first EOFs. The color of each dot corresponds to the sub-ensemble of the member: black for SE_BATHY, blue for SE_WIND, yellow for SE_CLOUD, red for SE_AIRT, magenta for SE_DIFFU and green for SE_IC.

stays more or less constant as it is displaced by the model. It can be seen also that the weekly evolution is larger than the intra-ensemble spread[1]. The blue dots correspond to the members affected by perturbations of the wind field; a relative large spread is also visible. The red dots, representing members with modified air temperature field, have a surprising position in this space: they seem to reside on a straight line, which get further and further away from the other members in time. The members of the other sub-ensembles (bathymetry, cloud coverage, diffusion coefficients) do not spread out very much, and stay close to the reference member.

6.3 Results and conclusions

6.3.1 Daily forecasts

In this section, we try to forecast at the daily time-scale. We thus applied the hydrodynamic model on the 250 members during one week, and projected the outputs on the 25 EOFs. We built a database containing 1750 lines (one per day and per member). Each line contains the member number and starting day (for reference only), the 25 coefficients of the initial state (called $coef01$ to $coef25$), 10 coefficients for the (daily average of the) wind ($wind01$ to $wind10$), 1 coefficient for the air temperature ($tem01$), and 25 coefficients for the final state ($out01$ to $out25$). In this database, we then selected a random learning set ($LS1$) of 1500 lines, the remaining lines forming the test set ($TS1$). We also selected another learning set ($LS2$) containing all the states except the ones starting on the last day, and then a test set containing the states starting on the seventh day ($TS2$).

We first tried a simple linear regression of the 36 inputs to yield $out01$. This simple method is very fast (almost instantaneous on a modern PC) and already gives very accurate results, and it is doubtful that complicated non-linear methods will be able to outperform them. Indeed, the comparison of the real output $out01$ with the one predicted with the linear regression led to a maximum absolute error of 29.6; the error mean is 0.116 and its standard deviation 5.4. These errors should all be related to the value of the coefficients, which is in the range (450,1000); this range is not centered around 0, because the EOFs were obtained from a month of simulation, when now only the first week is considered.
The regression tree obtained from $LS1$ to forecast $out01$, is shown in Fig. 6.7. It comprises 611 nodes and the total variance reduction on the learning set is 99.825%. The attributes used in the tests, most reducing the variance, are the following: $coef01$ (86%), $coef04$ (13%), $wind02$ (0.6%), $coef03$ (0.2%), $tem01$ (0.1%), $coef21$ (0.1%). Then, when the tree is applied to $TS1$, the comparison of the real output $out01$ and the output predicted by the tree, yields a maximum absolute error of about 25, an average error of -0.75 and a std. dev. around this average of about 6. A scatter plot of the real output vs. the predicted one is shown in Fig. 6.8. It is typical for trees that some horizontal groups of points appear in the plot (see e.g. around the values 600, 750 and 900), resulting from different members who reached the same terminal node in the tree, and

[1] The fact that the EOF weights evolve in time, underlines, if necessary, the need to update the model error space in data assimilation methods.

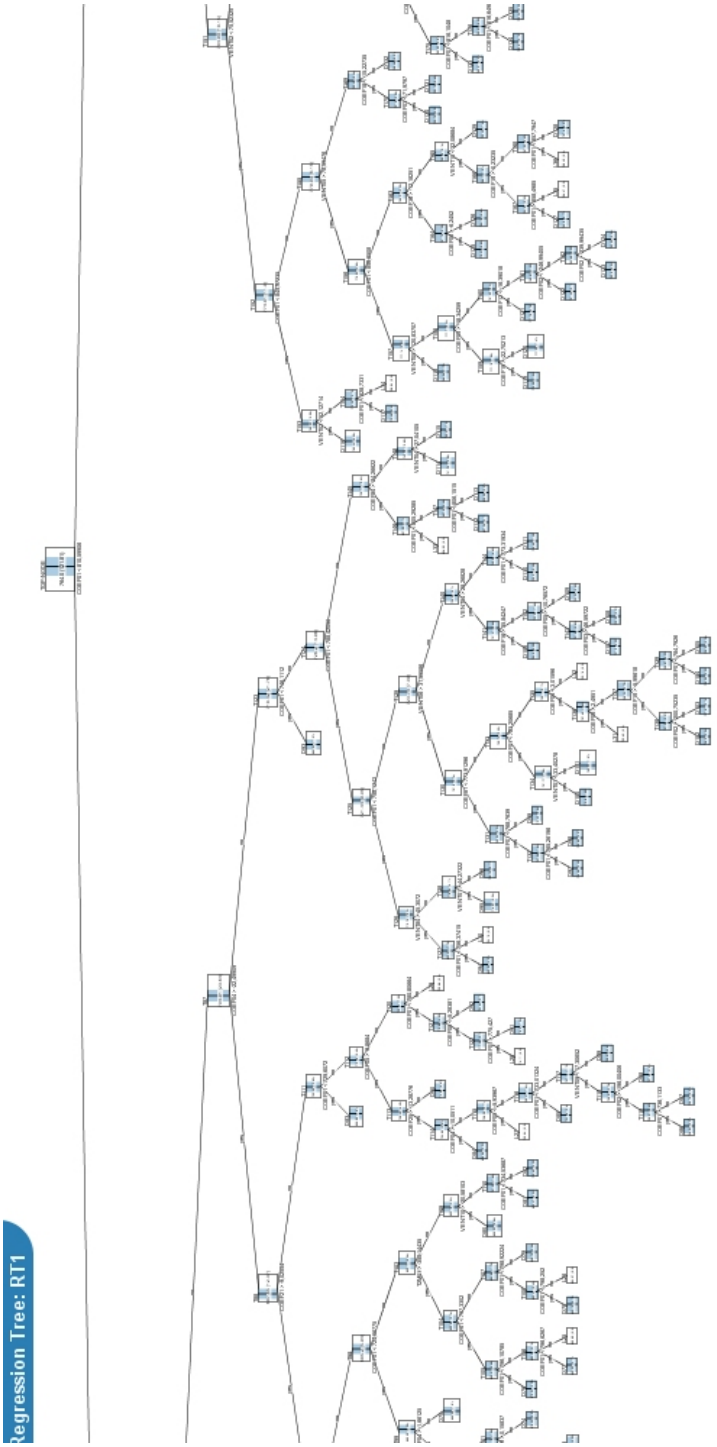

Figure 6.7: Zoom on the center of the regression tree yielding $out01$ after 1 day. The top-node test is: $coef01 < 810.599$

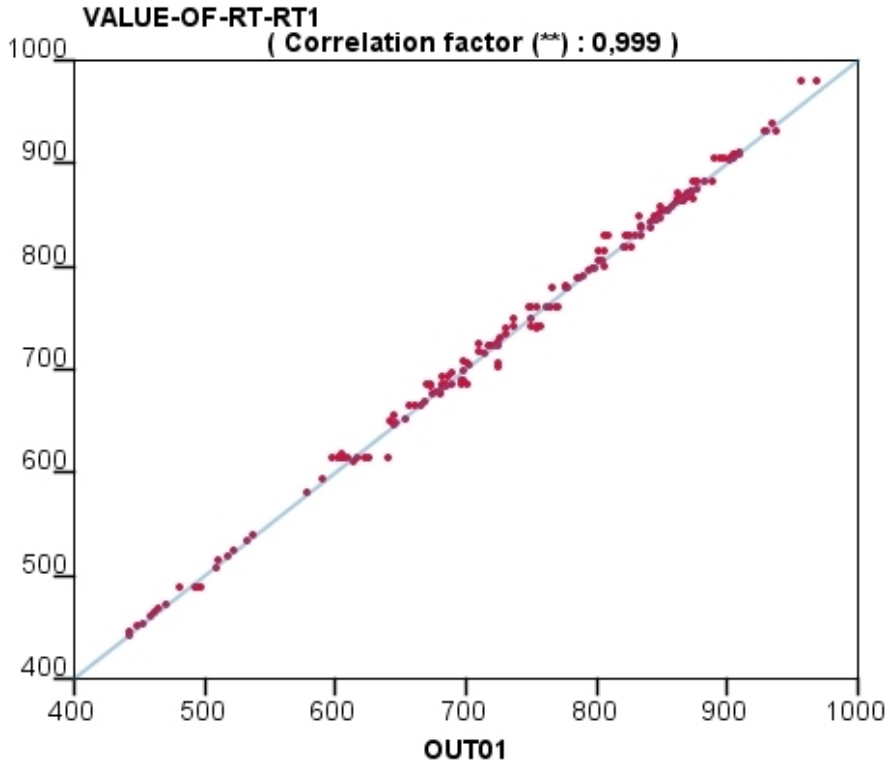

Figure 6.8: Scatter plot of $out01$ versus its value predicted by the regression tree ($TS1$).

were thus given the same output. The computational load to build the tree and use it is higher than the linear regression, but remains very low anyway (a few seconds).
The results obtained with the nearest neighbor method are deceiving (compared with the other methods), with average errors of 3, and std. dev of about 20, depending on K, the amount of neighbors. Pepito allows to automatically search for the optimal K by cross-validation. The computational load is then a little bit heavier (a few minuts), but the results seem almost independent of K.
A neural network with a single hidden layer containing 10 neurons, using a hyperbolic-tangent activation function, leads to very good results, comparable to the linear regression: the prediction of $out01$ leads to a maximum error of 23.6, a mean error of 0.34, and the std. dev. of the error is of 4.7. The computational load is also low (a few tens of seconds).

The following outputs were also predicted with the different AL methods; results are summarized in Table 6.1. It appears clearly that a linear regression is able to predict very well the coefficients in $TS1$. In particular, the bias of the prediction is almost zero for every output coefficient. The standard deviation of the error is also very small compared to the standard deviation of the variable

Output	σ	Lin.regression		Tree		K-NN		ANN	
		mean	σ	mean	σ	mean	σ	mean	σ
$out01$	130.45	0.116	5.4	-0.75	6.3	-2.60	21.6	0.35	4.7
$out02$	76.73	0.095	8.1	0.054	9.3	-1.87	13.3	0.02	4.7
$out03$	54.60	0.168	4.4	0.36	6.2	-1.23	12.2	0.27	4.1
$out04$	62.29	0.247	7.4	-0.015	6.1	-7.26	16.7	-0.17	7.4
$out05$	41.70	0.132	2.8	0.17	5.6	-4.54	23.2	0.29	3.0
$out06$	40.69	0.007	4.2	-0.16	4.3	-6.67	17.6	0.25	4.4
$out07$	52.17	0.084	3.3	-0.75	6.1	-3.14	20.3	0.06	2.9
$out08$	42.79	0.165	2.8	-0.08	4.7	-3.16	20.9	0.31	2.9
$out09$	37.96	-0.044	1.6	-0.16	4.2	-3.58	23.2	-0.42	4.3
$out10$	47.12	-0.042	3.2	-0.87	8.0	0.65	15.0	-0.03	2.8
$out11$	38.65	0.065	3.5	0.02	7.6	2.20	17.5	0.25	4.0
$out12$	35.99	0.047	1.6	0.01	4.5	1.55	16.7	-0.22	3.0
$out13$	40.28	0.232	1.8	0.26	4.6	0.67	16.7	-0.17	4.7
$out14$	37.58	0.043	1.4	-0.27	6.4	1.71	16.0	-0.57	2.1
$out15$	34.64	-0.095	1.6	-0.14	3.2	0.27	16.7	-0.04	1.7
$out16$	34.94	-0.182	1.5	-0.41	3.6	-0.21	17.0	-0.17	2.4
$out17$	34.41	0.138	1.7	-0.35	5.3	-0.03	16.8	-0.05	2.3
$out18$	34.27	-0.083	1.1	-0.75	5.0	-1.38	19.4	-0.05	6.4
$out19$	41.56	0.141	2.4	0.88	8.7	-0.03	14.8	0.08	1.9
$out20$	37.00	0.186	2.1	0.19	6.4	-1.85	18.3	0.03	2.1
$out21$	47.04	0.168	3.2	0.40	6.4	0.52	8.1	-0.75	6.9
$out22$	39.90	0.012	2.2	-0.13	4.1	0.03	10.6	-0.17	2.2
$out23$	34.73	-0.032	1.4	-0.15	4.8	1.18	7.5	-0.08	1.6
$out24$	32.27	0.148	2.2	0.34	7.8	2.05	8.6	0.20	1.9
$out25$	30.63	-0.059	1.7	-0.16	4.3	-0.37	7.6	-0.07	2.0

Table 6.1: Summary of the performance of the various AL methods in $TS1$. For each output, the standard deviation (as indication), the mean error and the standard deviation of the error are given.

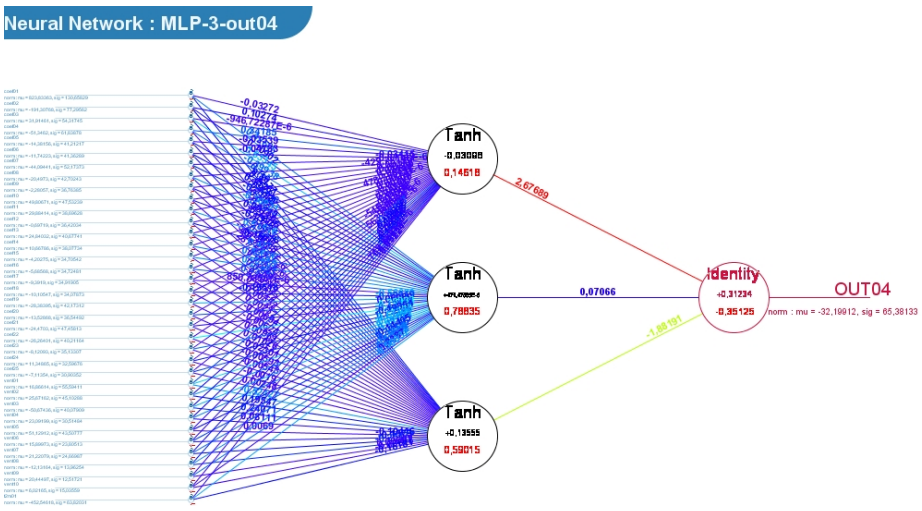

Figure 6.9: The trained ANN leading to an accurate prediction of *out*04, comprising only a single hidden layer with 3 neurons.

itself. It should be noted that the coefficients of the linear regression are not approximately equal to 1 (i.e. the regression is not simply yielding persistance). Non-linear methods (regression trees, neural networks) also perform well. Regression trees do somewhat less good; simple ANNs (1 hidden layer with only 3 neurons, i.e. about 110 parameters to determine, see Fig. 6.9) do as well, or slightly better than the linear method.

Indeed, when a simple relation exists between inputs and outputs, it is well known that a simple method will detect it better than a complex one (this is known as "Occam's Razor", see Greenwood and Oxspring [2001]). Apparently, the relation here is approximately linear, with small non-linear corrections. The nearest neighbor method leads to bad results for all coefficients.
It is important to note that all coefficients lead to similar error rates. If the latest EOFs merely represented noise, their coefficients would be random, and unpredictable; but here, it appears every EOF contains a predictable signal.
Thus, when one seeks to evolve a very large ensemble of members in time, it should be possible to forecast only a smaller ensemble with the hydrodynamic model, and afterwards, train a statistical predictor in order to forecast the remaining members.

Let us now redo the experiments using $LS2$ to train the methods, and $TS2$ to test them. This will indicate whether the data also can be extrapolated in the future. We should note from the start that this might be a difficult task, since only 6 previous days are present to train the method (for each member). The results are shown in Table 6.2. Now it appears that the linear regression yields a relatively important bias on the predictions of output coefficients 1, 2, 4, 10, 11, 19, 21. This indicates that the method misses the time-evolution function linking the input coefficients with these outputs (Fig. 6.10). Moreover, *out*02 also presents a large standard deviation, making the prediction inaccurate.

Output	σ	Lin.regression		Tree		K-NN		ANN	
		mean	σ	mean	σ	mean	σ	mean	σ
$out01$	130.45	18.85	4.32	-9.46	20.73	-57.67	30.53	6.38	5.60
$out02$	76.73	39.69	8.59	39.80	34.38	25.97	12.23	29.81	12.92
$out03$	54.60	-2.78	3.37	37.30	24.49	-4.43	14.55	-2.81	5.52
$out04$	62.29	-26.55	2.55	-14.09	15.59	3.31	17.67	-2.82	5.52
$out05$	41.70	8.61	1.94	1.77	5.05	0.194	23.94	-3.93	14.86
$out06$	40.69	4.21	2.36	6.88	9.86	1.89	23.99	-6.05	7.13
$out07$	52.17	-5.11	2.56	-4.39	5.42	4.89	23.58	-0.48	6.15
$out08$	42.79	-2.66	2.13	-3.68	2.99	1.83	24.12	1.25	5.37
$out09$	37.96	3.77	1.21	-0.76	7.18	0.150	22.97	5.18	5.54
$out10$	47.12	27.82	1.16	6.59	4.66	-7.19	20.58	-3.97	12.12
$out11$	38.65	19.2	1.46	2.22	9.11	-2.52	22.54	3.23	8.59
$out12$	35.99	12.46	0.98	3.53	6.30	4.20	21.02	10.62	6.72
$out13$	40.28	13.65	1.94	4.22	5.33	-3.99	23.28	8.25	5.80
$out14$	37.58	6.92	0.90	2.33	7.76	-2.45	22.82	-1.26	4.64
$out15$	34.64	5.52	1.20	0.244	3.63	0.85	21.85	3.11	1.81
$out16$	34.94	-2.91	0.70	-0.10	4.32	1.60	21.68	-3.37	5.28
$out17$	34.41	0.95	1.33	-4.76	3.23	-0.50	21.30	1.58	10.57
$out18$	34.27	2.16	0.89	2.01	3.85	3.82	21.01	-2.15	4.49
$out19$	41.56	-11.53	2.53	-2.31	8.58	3.39	21.09	-4.58	3.67
$out20$	37.00	-2.49	1.96	-2.58	8.72	2.70	20.46	-0.80	3.38
$out21$	47.04	-14.76	3.75	1.43	7.86	4.17	21.38	-6.01	9.34
$out22$	39.90	-10.71	1.84	-2.39	7.83	7.50	21.20	-4.08	6.00
$out23$	34.73	-8.06	1.04	-7.96	9.63	-2.77	20.51	-5.92	1.71
$out24$	32.27	-2.48	1.44	-5.84	6.58	-6.48	18.21	0.92	6.13
$out25$	30.63	-5.81	1.08	5.35	3.74	5.08	19.52	0.77	9.14

Table 6.2: Summary of the performance of the various AL methods in $TS2$. For each output, the standard deviation (as indication), the mean error and the standard deviation off the error are given.

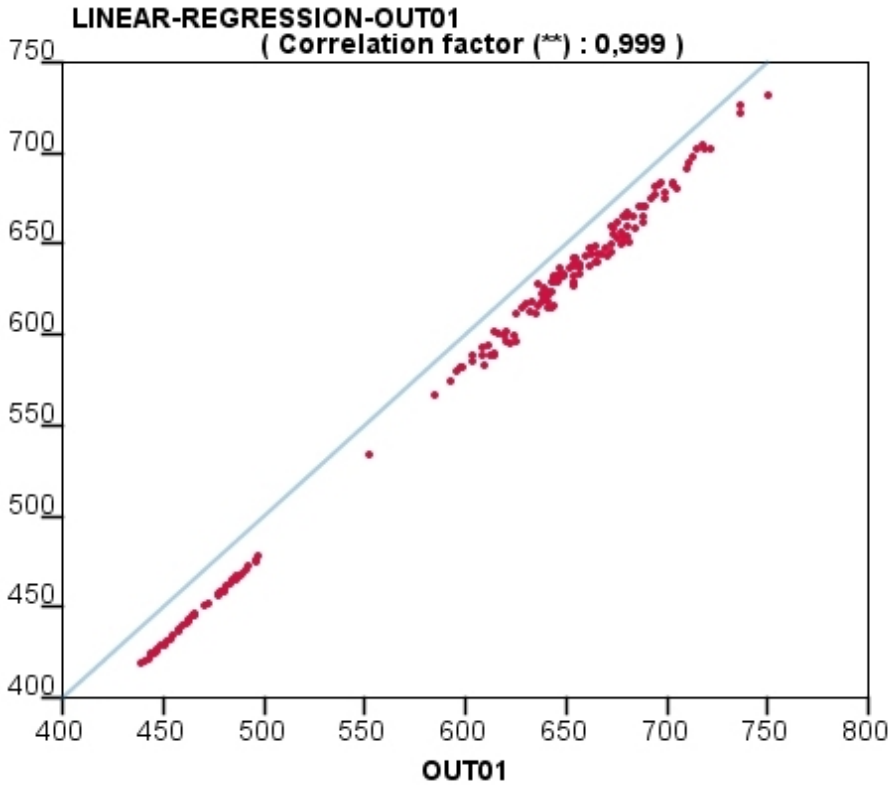

Figure 6.10: Scatter plot of $out01$ versus its value predicted by a linear regression ($TS2$).

The ANNs however lead to better results. It seems that the method is able to extract the time-evolution part of the function, and extrapolate the information to the next day. The bias is quite low, and the error variance is usually acceptable too, even if it is higher than the variance yielded by the linear regression. These results are illustrated in the scatter plot of the real $out01$ vs the one predicted by the ANN (Fig. 6.11a). For some outputs, the error remains relatively large (e.g. $out12$, and particularly $out02$, see Fig. 6.11b). Also, the performance of the ANN varies from one try to another, depending on which local minimum of the cost function is reached, which depends on the (random) initial values of the weights in the network. Thus, when one tries to forecast EOF weights, experiments should be redone many times, and validated with a test set or cross-validation, before drawing conclusions. The outputs also depend on the architecture of the neural net, although we found it quite stable as long as about 10 to 20 neurons were used (distributed in 1 or 2 layers). A more complex network architecture usually leads to a lower bias, but higher variance. For example, the error mean and std. dev. induced for output 09 by a network with a single layer of 10 neurons are given in Table 6.2 : mean=5.18 and std. dev.=5.54. A net with 25 neurons led to a bias of only -0.19, but the std. dev. is 14.25.
A typical learning curve (rms error at each cycle through the learning set during the training phase of the network) is shown in Fig. 6.12. The error in the learning set decreases constantly, while the test set error increases when overfitting occurs.
Regression trees also perform better than linear regression; they are often doing only slightly worse than the ANNs, and sometimes somewhat better (see e.g. outputs 05, 09, 15, 16 and 21 in Table 6.2). However, the very most important coefficient to forecast is $out01$, and there, the ANN performs significantly better. As expected, for every output, the most determining input for the variance reduction in the tree is the corresponding one ($coef02$ for $out02$ etc.). The following important coefficients usually comprise a couple of wind coefficients, the temperature coefficient and a few other input state coefficients.

Finally, the K-NN method is outperformed by regression trees and neural networks, and also usually by linear regressions, whatever K is considered. Thus, this method is not recommended for the present application.

6.3.2 Weekly forecasts

We will now try and predict the ocean state using the EOF coefficients from the previous week. We have run the hydrodynamic model during 1 month, leading to 4 instants (weeks starting with the initial conditions, or weeks starting on one of the three following weeks)[2]. Each line of the database contains the 25 ocean state EOF coefficients of the previous week, 3 average coefficients for each wind and air temperature EOF (i.e. 33 coefficients in total), and the 25 ocean state EOF coefficients at the end of the week. We will again test the capabilities

[2]In principle, with a one-month simulation, we could use a "moving window" of a week, and hence create about 24 input-output relations rather than 4. However, due to limited hard disk space, we could not save daily model outputs, and only 4 outputs were saved, separated by a week each.

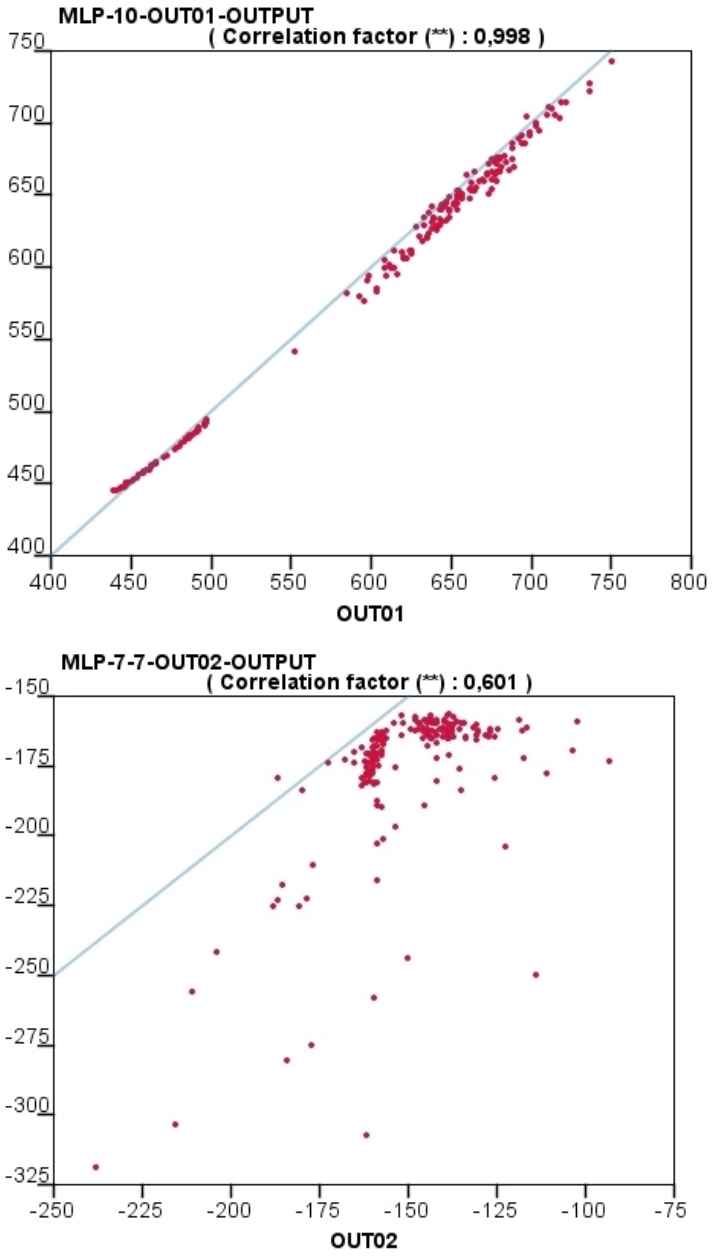

Figure 6.11: Scatter plot of the real output versus its value predicted by the ANN ($TS2$), for (a) $out01$, (b) $out02$.

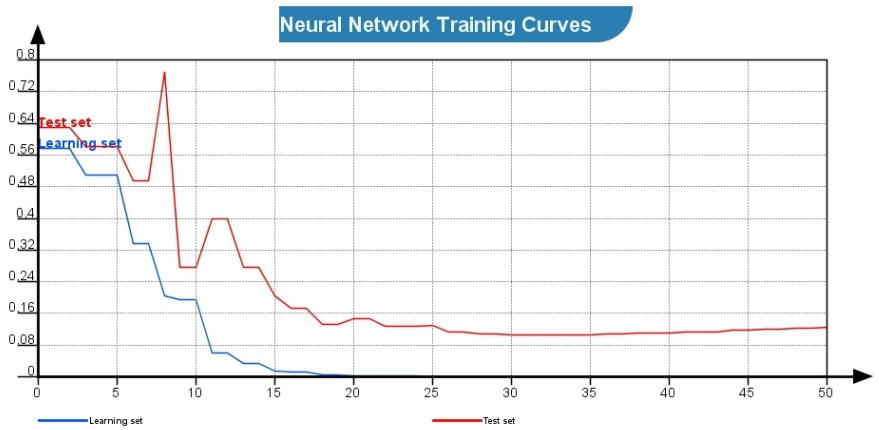

Figure 6.12: Typical ANN learning curve (here, an ANN with 2 layers of 8 neurons each, yielding $out12$.

of the AL algorithms both when some random members are removed from the ensemble, which in this way is split into the learning set $LS1$ and test set $TS1$, and when the members are removed which start on week 3 and end at week 4, forming $LS2$ and $TS2$. We again expect the predictions in $TS2$ to be more difficult to realize. The situation is even more critical then in the previous section, as now only 3 different instants are available to find a relation and extrapolate it in the future. Moreover, this relation is more complicated, as there are more (atmospheric) inputs and the underlying hydrodynamic model has been running longer.

We start by training the AL methods with $LS1$ and testing them using $TS1$. The linear regression leads to good results, e.g. $out01$ is only slightly biased (-0.052) and the error has a std. dev. of 8.8, to be compared with the std. dev. of $out01$ itself (664.3). The following outputs are also predicted quite accurately. Regression trees yield forecasts somewhat less accurate than the linear regression; in particular, the error standard deviation is higher. The trees indicate again, for each output, which input yields the largest variance reduction. For $out01$, the corresponding input is the main contributor (about 99% of the variance reduction), as well as some atmospheric coefficients and some other ocean state input coefficients. The precise choice of atmospheric coefficients seems somewhat random, as there is for example no reason to prefer a certain wind coefficient averaged over the first 2 days of the week, or over the following 2 days, unless by chance critical wind events would preferably happen at the beginning of each week. Thus, the particular choice seems somewhat arbitrary. The prediction of some other outputs leads to surprising rules. For example, in the tree for $out02$, 77% of the variance reduction is obtained with tests on the first wind EOF coefficient averaged over days 3 and 4; 13% is due to input coefficient 03. Another wind EOF accounts for 5%, and the input coefficient 02 finally accounts only for 2%. The results in the test set show that this tree is rather good; it is approximately unbiased, with an error standard deviation of

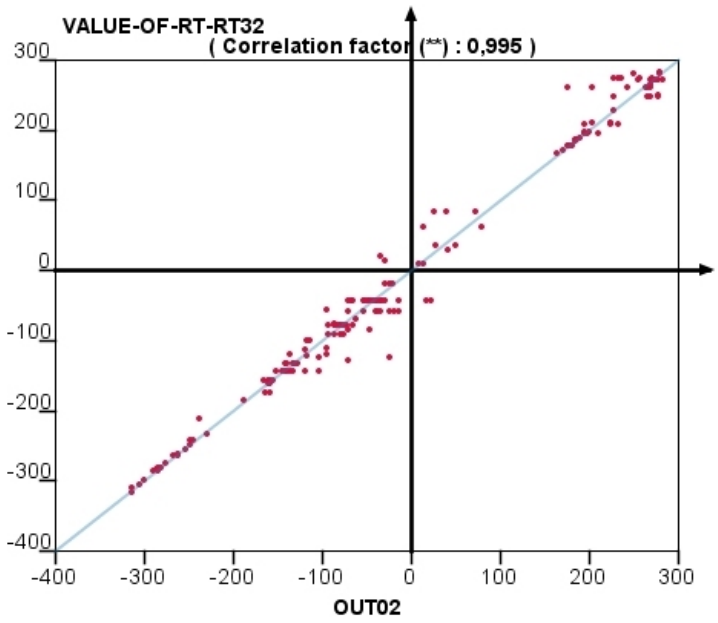

Figure 6.13: Scatter plot of the real output $out02$ versus its value predicted by the tree ($TS1$).

18 (the standard deviation of the output itself is 174). The scatter plot is shown in Fig. 6.13. Other outputs also have unexpected contributors to the decision rules.

Exactly as in the previous section, the nearest neighbor method does not provide interesting results. Neural networks however lead to results comparable to, but not better than, linear regression. Thus, contrary to the preceding section, the non-linear methods are usually not able to extract more information than the linear relation from the database; they rather slightly and erroneously over-fit the data.

The forecasting of the ocean state after week 4, from the state at the beginning of this week, could not be realized correctly by any of the AL methods that we tested. As example, the scatter plot of the linear regression forecast of $out01$ is shown in Fig. 6.14: the forecast is completely wrong. Unfortunately, regression trees (pruned or not) and ANNs (with various architectures) did not lead to conclusive results either. Thus, as no method was able to find a relation between inputs and outputs that could be extrapolated in time, we suspect such a stable relation simply does not exist. Possibly, if more past weeks would be considered (remember that only 3 past weeks are present in our database), something useful could appear from the AL algorithms.

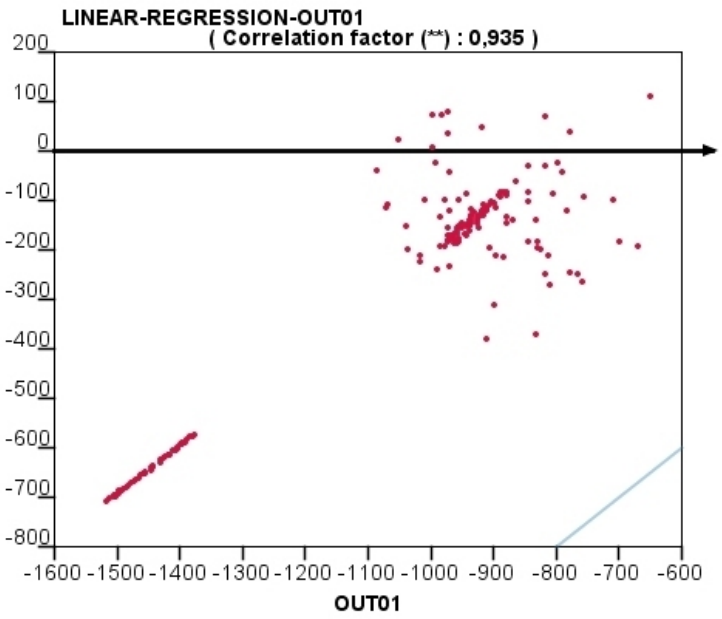

Figure 6.14: Scatter plot of the real output $out01$ versus its value predicted by linear regression ($TS2$).

6.3.3 Conclusions

When we try to forecast an ocean state based on the previous day state, and the wind and air temperature in between, it appeared that the relationship between inputs and outputs is relatively easy to find. Indeed, even linear regression could forecast the outputs; the non-linear techniques of course also managed to find the outputs, except the K-NN method whose performance was lower. Thus, the AL technique could be applied to build larger ensembles from smaller ones, at a marginal computational cost (compared to the full hydrodynamic model). The only limitation of the method is that the forecasted ocean states will be linear combinations of past and other present states. If the EOFs are built from a sufficiently large ensemble, the results should still be interesting.

When using an ensemble of members covering the past days to predict an unseen ocean state for the next day (i.e. there is absolutely no hydrodynamic prediction of any member for the next day), it seems that the linear regression method is not able to extract the information from the database that allows to extrapolate in the future. This indicates that the relationship found above led to such good performances, because it has (also) been trained on the present-day forecast, albeit with different members. A pure extrapolation in time is much harder to implement. However, we found out that ANNs were still able to predict the outputs relatively well. Trees have more difficulties, particularly with the first output coefficient (which is by far the most important one). However, they perform relatively well for some other coefficients.

The exact performance of the methods of course depends on the importance of the changes happening during the day in the test set. If a totally different wind would appear, for example, all algorithms would most probably completely fail. We mentioned from the start of this chapter that this is an inevitable property of statistical methods; only a very large database, covering many past days with every possible situation, could possibly yield a robust statistical predictor.
To further improve the statistical forecasting, the main effort should thus be directed at better sampling of the space formed by the multiple input coefficients. More past states should be considered, not only in the recent past, but maybe also during the same season in previous years. Hence, situations similar to the ones we try to test would already have been encountered, and forecasting performances are expected to be better.

When forecasting the ocean state based on the previous week state, and some atmospheric coefficients, the best results were provided by the simple linear regression. Non-linear methods could not extract any supplementary information from the database. However, the accuracy of (all) the methods is still sufficient to use them to enlarge ensembles. When trying to forecast future states, without having explored that future by other members, only useless results could be obtained; no method was able to retrieve a consistent relation between the inputs and the outputs, which could be used to extrapolate in time. Again, a learning set covering a longer past time period might yield better results.

6.4 Object-oriented methods

As announced in the introduction chapter, we showed above that we are more-or-less able to simulate the complete hydrodynamic model with statistical methods. However, to train these methods, a database is needed which contains enough ocean states to characterize the phase states space, and which contains all the necessary parameters required to compute the evolution from time t to time $t+1$. In our case, the database was built from the outputs of a primitive equation model, but we could imagine building it from observations too, using an inverse model to build the complete ocean states [Brasseur, 1991, Brasseur and Haus, 1991]. A well-built automatic learning (AL) method should then be able to forecast the future ocean state, provided the past ocean state is located in the same region of the phase space. As we have explained, we could then increase ensemble sizes considerably at very low computational costs.

The method also is limited by the fact that only linear combinations of EOFs are possible to represent the forecasts. This is known to be a particular problem when meso-scale and small-scale features are considered. A particular, small gyre, for example, could be advected together with a larger scale current. The successive positions of the gyre should all be represented by a particular EOF, which is unrealistic. Thus, if the smaller scales are to be represented, another method for representing the ocean state is necessary.

In our opinion, the most interesting way of achieving this level of predictions with AL algorithms should be based on human-like reasoning. This leads to object-oriented methods. Thus, the ocean should be decomposed as a back-

ground field, and a multitude of "objects" which are each defined by a few typical numbers. For example, we could define gyres (defined e.g. by the position of their center, their depth, average radii and velocity, temperature, salinity, and surface elevation difference between the middle and the edge of the gyre), currents (defined e.g. by their position, depth, width, velocity, ...), but also particular objects such as upwellings. The creation, destruction and time-evolution of the attributes of these objects could then be forecasted by an AL method, trained with data from a hydrodynamic model or an inverse model. This would be more general than the work showed above, as the forecast of the object attributes would be valid at all places and all times, as long as the parameter space has been sufficiently sampled during the training. It is the definition (and automatic detection in the ocean) of the objects and their attributes that constitutes a very difficult problem in itself. We mentioned before that Ben Bouallègue et al. [2005] presented a method to modify objects (such as gyres) by displacing and intensifying them, but these object were manually selected square zones rather than automatically detected particular features in the ocean. Methods exist however to automatically select structured objects in spatio-temporal systems, and some of them have been applied to oceanography. Thonet et al. [1995] developed a technique to identify eddies in remote sensed AVHRR images, using multiscale analysis of isotherm curvature, and characterized these eddies through the so-called "phase portraits method". Herlin et al. [2004] developed a scheme where eddies are recognized based on contour recognition: a sufficiently regular curve delimiting the eddy (because the eddy is approximately circular) should lie in the vicinity of locations of maximal contrast. Segond et al. [2006] uses genetic algorithms to automatically detect gyres (including characterization of their size) in current velocity fields. After a few tens of passes, the algorithm is able to correctly detect about 80% of the gyres. Guindos-Rojas et al. [2004] developped a system aiming at classifying different objects such as eddies and upwellings, based both on ANNs and tree methods. Finally, Shalizi et al. [2005] presents two such methods were structures are detected based on their different evolution in time, compared with the background field (however, they did not applied these general methods to oceans). Of course, the application of these methods to the characterization, and ultimately, the forecasting of ocean states is a very complex problem which requires much more research.

Chapter 7

Conclusions

Not all who wander are lost
JRR Tolkien

In order to use high-resolution models of particular regions, such as coastal areas, the most common technique is to use nested grids. Furthermore, it has been shown that it is beneficial to send the (averaged) information from the nested grid back to its parent grid. We implemented a doubly nested model, whose 3 grids cover the Mediterranean Sea, its North-Western part, and the Gulf of Lions.
When observations are available in the nested area, we examined in chapter 3 where they should best be assimilated. Therefore, we implemented twin experiments. A reference run, using two-way nesting, is considered the truth, and pseudo-observations are extracted. Then, other simulations start from wrong initial conditions; we try to correct them using data assimilation (DA). We used a fixed SEEK filter, with a few specificities. For example, each observation is multiplied by a radial Gaussian function centered on the corresponding observation, in order to yield a correction only in a limited area. Also, to limit the risk of creating instabilities, corrections leading to strong modifications of the Brunt-Väisälä frequency are attenuated or simply put to zero. An estimation of the model error, required by the assimilation procedure, is obtained from a previous model run; it is kept constant at all times. The way of obtaining and using the model error covariance matrix, constitutes the most limiting aspect of the assimilation procedure in our simulations. Furthermore, the SEEK filter supposes that the error probability function is Gaussian; this is not always the case, particularly in coastal zones.
Both two-way and one-way simulations are tested; and the pseudo-observations could be assimilated in any of the grids. A last possibility is to group all information from the 3 grids in a single vector, and coherently modify it during DA cycles. We found that when two-way nesting is unavailable, the observations best constrain the model when they are assimilated in the nested model rather than in its parent. However, the most efficient setup is the one where a single state vector contains all information.

Thus, it is quite a pity that most often, two-way nesting is not applied in operational forecasts. Indeed, feedback requires permanent data transfer between parent and child models, but those aren't always run together or at the same oceanographic center. Rather, to benefit from the expertise of local oceanographic teams, nested models are often run at different institutes. Hence, we sought to replace the data feedback from child to parent with a less demanding DA procedure. Forecasts are extracted from the nested model, and sent back to the parent model where they are assimilated as pseudo-observations. Usually, operational models include a DA procedure anyway, and the supplementary data transfers required by our upscaling procedure are very limited. The major requirement is that the models should be run twice.
We have shown in chapter 4 that the whole procedure is beneficial for the subsequent forecasts in the nested region, as the corrections to the parent model flow back to its child model via the boundary conditions (and during possible reinitializations as well). Depending on the DA scheme, the chosen error matrices, and the pseudo-data extracted from the nested model (any data could in principle be used), the effect of not using nesting feedback can be more or less diminished. When using a fully 3D DA scheme, we showed that it is best to assimilate 2D fields (e.g. surface temperature, salinity and elevation) in order to capture the small-scale corrections to the large scale flow. The latter

is hopefully already more-or-less well represented in the parent model. In the particular case implemented in this work, we could then remove about half of the error in the child model, which is due to the lack of nesting feedback.

The two problems considered, downscaling and upscaling, and the assimilation experiments, led us to realize that the most critical point in popular DA filters such as the SEEK filter, is to correctly represent the expected model error. Many authors simply take the model variability in time as a proxy for its error, and correct errors only along the most important of these "directions" of the model error space. This is also the approach that we followed in our twin experiments. However, this approximation might often be inadequate. The superiority of filters derived in the "ensemble" approach, where an ensemble of members is evolved in time by the hydrodynamic model, has been shown in many applications. Thus, we decided to try and build a new model error space based on such an ensemble. In order to tackle one problem at a time, we used only a single grid covering the Mediterranean Sea, but implemented it with a relatively high resolution ($1/16°$). We then perturbed the bathymetry, the horizontal diffusion coefficients of the model, the initial conditions and the various atmospheric forcing fields (wind, cloud coverage, air temperature) used interactively by the model through the bulk formula, leading to an ensemble of 250 members. Other perturbations would still be possible, by modifying e.g. the river inflows, lateral open sea boundary conditions (the Atlantic Ocean) or the position and intensity of some particular currents. Then, we analyzed how these modifications were forwarded in time by the model. In particular, we examined if they were stable in space, and if their intensity increased or decreased during the simulation. We could thus quantitatively assess the impact of uncertainties affecting various model parameters and forcing fields. As one might expect, we found that the perturbation of lateral boundary conditions or model parameters led to rms errors increasing in time, while the differences between reference member and members with perturbed initial conditions were slowly attenuated during the simulations. Of course, these general conclusions must be nuanced for each perturbation. For example the members with modified bathymetry presented larger errors in the areas where the bathymetry was most perturbed. As we modified the bathymetry by more or less smoothing it, this corresponds to the areas with the largest bathymetry gradients. Another example concerns the members with modified wind forcings. As we perturbed the wind field by decomposing it in EOFs, and then by multiplying the EOF weights with random factors, the wind perturbations were larger when the wind field itself is larger; hence, the ocean state error generated at (or just after) these moments was also larger. Yet another interesting conclusion concerned the members with modified initial conditions. We found that the propagation in time of these initial perturbations were not superposable between the various members; it would thus be difficult to correct these errors with data assimilation schemes.
We further examined the first to fourth-order moments of the perturbations, and their Empirical Orthogonal Function decomposition. We found that most error processes can be considered approximately Gaussian (a critical hypothesis in many DA filters), but not all of them. In coastal zones, for instance, the "error" usually had no Gaussian probability density function.

The ensemble has a wealth of applications. Not only does it provides an es-

timation of the forecasts accuracy, it also can be used efficiently in DA schemes such as the Ensemble Kalman Filter or the Sequential Importance Resampling filter. Furthermore, coupled models (biology, drift models ...) rely heavily on the ocean physics, and therefore it is very important to assess the impact of uncertainties affecting the latter, on the former. Finally, the ensemble could also be used to generate EOFs (characterizing the perturbations), which can in turn be used to simulate the hydrodynamic model with a statistical model.

In chapter 6, we tried various automatic learning (AL) methods for this purpose: multi-linear regressions, but also artificial neural networks, regression trees and the k-nearest-neighbors method. We tried to forecast the ocean state at the day and week timescales. As inputs, we used the weights of the 25 first EOFs in a decomposition of the initial ocean state, and the weights of the 10 first EOFs of the wind decomposition and of the first EOF of the air temperature field, averaged between the input time and output time. The outputs were the weights of the 25 first EOFs in the decomposition of the forecasted ocean state. When the ensemble covers various past instants as well as (for some members) the future instant, we could relatively easily forecast the future instant ocean state for the other members, even with the linear regression method, both after one day and one week. This indicates that the AL methods are able to find the relationship between inputs and outputs, and use it on unseen ocean states as well. However, when the ensemble does not cover the final day at all, the relation between the ocean states must be derived and extrapolated in the future. Only the non-linear methods, and artificial neural networks in particular, could achieve this when forecasting the ocean state after one day, indicating that the relation is non-linear, but relatively stable in time, at least between the days covered by the database, and the day where the prediction is required. Of course, if a significant new event would take place, the method would fail whatsoever.
When we tried to predict the ocean state after a week (without any member available at the end of the week in question), no method at all could provide a decent forecast. Thus, no relation exists between the EOF weights at the beginning and the end of a certain week, that is sufficiently stable in time, in order still to be used the following week.
All in all, the experiments conducted offer two interesting perspectives: we could augment the size of ensembles, by adding new members at a marginal computing cost (all methods described here require at most a few minutes of computing time, thus about a factor 1000 less time than the hydrodynamic model). Given the past few days ocean states, we could also provide a new forecast of the following day ocean state much faster than a hydrodynamic model. The method could thus be used when an extremely fast forecast is required. The main limitation, in both applications, is of course that the new estimation could only ever be a linear combination of the *a priori* given EOFs.

List of Figures

List of Tables

Bibliography

C. Alberola, C. Millot, and J Font. On the seasonal and mesoscale variabilities of the Northern Current during the PRIMO-0 experiment in the western Mediterranean Sea. *Oceanol. Acta*, 18:163–192, 1995.

A. Álvarez. Performance of Satellite-Based Ocean Forecasting (SOFT) Systems: A study in the Adriatic Sea. *Journal of Atmospheric and Oceanic Technology*, 20:717–729, 2003.

A. Álvarez, J. Tintoré, G. Holloway, M. Elby, and J.M Beckers. Effect of topographic stress on circulation in the western Mediterranean. *Journal of Geophysical Research*, 99:16053–16064, 1994.

A. Álvarez, C. Lopez, M. Riera, E. Hernandez-Garcia, and J. Tintoré. Forecasting the SST space-time variability of the alboran sea with genetic algorithms. *Geophysical Research Letters*, 27(17):2709–2712, 2000.

J. L. Anderson. An Ensemble Adjustment Filter for Data Assimilation. *Monthly Weather Review*, 129:2884–2903, 2001.

J. L. Anderson and S. L. Anderson. A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127:2741–2758, 1999.

F. Auclair, S Casitas, and P. Marsaleix. Application of an inverse method to coastal modeling. *J. Atmos Oceanic Technol.*, 17:1368–1391, 2000.

F. Auclair, P. Marsaleix, and C. Estournel. The penetration of the Northern Current over the Gulf of Lions (Mediterranean) as a downscaling problem. *Oceanologica Acta*, 24:529–544, 2001.

F. Auclair, P. Marsaleix, and P. De Mey. Space-time structure and dynamics of the forecast error in a coastal circulation model of the Gulf of Lions. *Dynamics of Atmospheres and Oceans*, 36:309–346, 2003.

A. Alvera Azcàrate. *Forecast verification of a 3D model of the Mediterranean Sea. Analysis of model results and observations using wavelets and Empirical Orthogonal Functions.* PhD thesis, Université de Liege, 2004. 263 pp.

A. Barth. *Assimilation of sea surface temperature and sea surface height in a two-way nested primitive equation model of the Ligurian Sea.* PhD thesis, Université de Liege, 2004. 202 pp.

A. Barth, A. Alvera-Azcàrate, M Rixen, and J.-M. Beckers. Two-way nested model of mesoscale circulation features in the Ligurian Sea. *Progress In Oceanography*, 66:171–189, 2005.

A. Barth, A. Alvera-Azcárate, J.-M. Beckers, M. Rixen, and L. Vandenbulcke. Multigrid state vector for data assimilation in a two-way nested model of the Ligurian Sea. *Journal of Marine Systems*, 2006.

J.-M. Beckers, M. Rixen, P. Brasseur, J.-M. Brankart, A. Elmoussaoui, M. Crépon, Ch. Herbaut, F. Martel, F. Van den Berghe, L. Mortier, A. Lascaratos, P. Drakopoulos, P. Korres, N. Pinardi, E. Masetti, S. Castellari, P. Carini, J. Tintore, A. Alvarez, S. Monserrat, D. Parrilla, R. Vautard, and S. Speich. Model intercomparison in the Mediterranean. The MedMEx simulations of the seasonal cycle. *Journal of Marine Systems*, 33–34:215–251, 2002.

J.M. Beckers. Application of a 3D model to the Western Mediterranean. *Journal of Marine Systems*, 1:315–332, 1991.

J.M. Beckers, P. Brasseur, and J.C.J. Nihoul. Circulation of the western Mediterranean: from global to regional scales. *Deep-Sea Research*, 44(3–4): 531–549, 1997.

Z. Ben Bouallègue, C. Fratianni, L. Vandenbulcke, M. Rixen, and J.-M. Beckers. A structure-oriented method for forecast assessment: analysis of the forcing winds impact on the Mediterranean Sea structures representation. In *Geophysical Research Abstracts*, Austria, April 2005. EGU General Assembly.

M. Benzohra and C. Millot. Characteristics and circulation of the surface and intermediate water masses off Algeria. *Deep-Sea Research*, 42:1803–1830, 1995.

L. Bertino, G. Evensen, and H. Wackernagel. Combining geostatistics and Kalman Filtering for data assimilation in an estuarine system. *Inverse Methods*, 18:1–23, 2002.

L. Bertino, G. Evensen, and H. Wackernagel. Sequential data assimilation techniques in oceanography. *International statistical review*, 17:223–241, 2003.

J.P. Bethoux, L. Prieur, and J.H. Bong. Le courant Ligure au large de nice. *Ocean. Acta*, 9:56–67 (special issue), 1988.

C. H. Bishop, B. Etherton, and S. J. Majumdar. Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. *Monthly Weather Review*, 129:420–436, 2001.

E. Blayo and L. Debreu. Revisiting open boundary conditions from the point of view of characteristic variables. *Ocean Modelling*, 9:231–252, 2005.

E. Blayo and L. Debreu. *Ocean Weather Forecasting, edited by E. Chassignet and J. Verron*, chapter Nesting Ocean Models. Springer, 2006.

E. Blayo and L. Debreu. Adaptive mesh refinement for finite difference ocean model: some first experiments. *J. Phys. Ocean.*, 29(6):1239–1250, 1999.

E. Blayo, J. Verron, J.M. Molines, and L. Testard. Monitoring of the gulf stream path using geosat and topex/poseidon altimetric data assimilated into a model of ocean circulation. *Journal of Marine Systems*, 8:73–89, 1996.

R. Bleck. Simulation of coastal upwelling frontogenesis with an isopycnic coordinate model. *J. Geophys. Res.*, 83:6163–6172, 1978.

R. Bleck and D. Boudra. Wind-driven spin-up in eddy-resolving ocean models formulated in isopycnic and isobaric coordinates. *J. Geophys. Res.*, 91:7611–7621, 1986.

P. Brasseur, J. Ballabrera, and J. Verron. Assimilation of altimetric data in the mid-latitude oceans using the Singular Evolutive Extended Kalman filter with an eddy-resolving, primitive equation model. *Journal of Marine Systems*, 22 (4):269–294, 1999.

Pierre Brasseur. A Variational Inverse Method for the reconstruction of general circulation fields in the Northern Bering Sea. *Journal of Geophysical Research*, 96(C3):4891–4907, 1991.

Pierre Brasseur and Jacques Haus. Application of a 3–D variational inverse model to the analysis of ecohydrodynamic data in the Northern Bering and Southern Chuckchi seas. *Journal of Marine Systems*, 1:383–401, may 1991.

G. Burgers, P. J. van Leeuwen, and G. Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly Weather Review*, 126:1719–1724, 1998.

M.A. Cane, A. Kaplan, R.N. Miller, B. Tang, E.C. Hackert, and A.J. Busalacchi. Mapping tropical Pacific sea level: Data assimilation via a reduced state space Kalman filter. *J. Geophys. Res.*, 101(C10):22599–22617, 1996.

R. Canizares. *On the application of data assimilation in regional coastal models.* PhD thesis, TU Delft, 1999.

S. Castellari, N. Pinardi, and K. Leaman. A model study of air-sea interactions in the Mediterranean Sea. *Journal of Marine Systems*, 18:89–114, 1998.

J. P. Chiles and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty, Wiley.* 1999.

N.E. Clark, R.M. Eber, J.A. Renner, and J.F.T. Saur. Heat exchange between ocean and atmosphere in the eastern north pacific for 1961-71. Technical report, NOAA, U.S.Dept.Commerce, Wash.,D.C., 1974. FIXME title.

P. Conan and C. Millot. Variability of the northern current off Marseilles, western Mediterranean Sea, from February to June 1992. *Oceanol. Acta*, 18: 193–205, 1995.

M. Crépon, L. Wald, and J.M. Monget. Low-frequency waves in the Ligurian Sea during December. *J. Phys. Oceanogr.*, 87:595–600, 1982.

B. Cushman-Roisin, M. Gacic, P.-M. Poumain, and A. Artegiani. *Physical Oceanography of the Adriatic Sea. Past, Present and Future.* Springer, 2002. 320 pp.

G. Cybenko. Approximations by superpositions of a sigmoidal function. *Matt. Contro. Signals Syst.*, 2:303–314, 1989.

R. Daley. *Atmospheric Data Analysis.* Cambridge University Press, New York, 1991. 457pp.

R. Daley. Forecast-error statistics for homogeneous and inhomogeneous observation networks. *Monthly Weather Review*, 120:627–643, 1992.

H.C. Davies. Limitations on some lateral boundary schemes used in regional nwp models. *Monthly Weather Review*, 111:1002–1012, 1983.

P. De Mey and M. Benkiran. *Ocean Forecasting, Conceptual basis and Applications*, chapter A multivariate reduced-order optimal interpolation method and its application to the Mediterranean basin-scale circulation, page 472. Springer-Verlag, Berlin Heidelberg New York, 2002.

D. Dee. *Realization and Modelling in System Theory*, chapter Simplified adaptative Kalman filtering for large-scale geophysical models. Kaashoek, M.A. and van Schuppen, J.H. andRan, A.C.M., eds, Birkhauser, volume 1 of *Proceedings of the International Symposium MTNS-89*, 567–574, 1990.

D. Dee and A. Silva. Data assimilation in the presence of forecast bias. *Q. J. R. Meteorol. Soc.*, 124:269–295, 1998.

D. Dee, S. Cohn, A. Dalcher, and M. Ghil. An efficient algorithm for estimating noise covariances in distributed systems. *IEEE Trans. Autom. Control*, 30: 1057–1065, 1985.

P.A. Devijver and J. Kittler. *Pattern recognition: A statistical approach.* Prentice-Hall International, 1982.

A. Doucet, N. de Freitas, and N. J. Gordon, editors. *Sequential Monte Carlo Methods in Practice.* Springer, 2001. 581 pp.

C. Dufau-Julliand, P. Marsaleix, A. Petrenko, and I. Dekeyser. Three-dimensional modeling of the Gulf of Lion's hydrodynamics (northwest Mediterranean during January 1999 (MOOGLI3 Experiment) and late winter 1999: Western Mediterranean Intermediate Water's (WIW's) formation and its cascading over the shelf break. *Journal of Geophysical Research*, 1098 (C11002), 2004.

V. Echevin, M. Crépon, and L. Mortier. Interaction of a coastal current with a gulf: application to the shelf circulation of the Gulf of Lion in the Mediterranean Sea. *J. Phys. Ocean.*, 33:188–206, 2003a.

V. Echevin, M. Crépon, and L. Mortier. Simulation and analysis of the mesoscale circulation in the northwestern Mediterranean Sea. *Annales Geophysicae*, 21: 281–297, 2003b.

A. Eliassen. Provisional report on calculation of spatial covariance and autocorrelation of the pressure field. *Report no 5, Videnslaps-Akademiets Institutt for Vaer-Og Klimasforskning*, page 12 pp, 1954.

C. Estournel, X. Durrieu de Madron, P. Marseleix, F. Auclair, C. Julliand, and R. Vehil. Observation and modeling of the winter coastal oceanic circulation in the Gulf of Lion under wind conditions influenced by the continental orography (FETCH experiment). *Journal of Geophysical Research*, 108(C3): 8059–8076, 2003.

G. Evensen. *Sequential Data Assimilation for Nonlinear Dynamics: The Ensemble Kalman Filter*. 2002.

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

G. Evensen. Sampling strategies and square root analysis schemes for the EnKF. *Ocean Dynamics*, 2004.

G. Evensen. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research*, 99, NO. C5:10143–10162, 1994.

G. Evensen and P.J. van Leeuwen. An ensemble kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128:1852–1867, 2000.

M. Fieux. Formation d'eau dense sur le plateau continental du Golfe du Lion. *Colloques Internationaux du CNRS. La Formation des Eaux Océaniques Profondes.*, 1974.

R.J. Fleming. On stochastic dynamic prediction. Part I: The energetics of uncertainty and the question of closure. *Monthly Weather Review*, 99:851–872, 1971a.

R.J. Fleming. On stochastic dynamic prediction. Part II: Predictability and utility. *Monthly Weather Review*, 99:927–938, 1971b.

A. D. Fox and S. J. Maskell. Two-Way Interactive Nesting of Primitive Equation Ocean Models with Topography. *Journal of Physical Oceanography*, 25:2977–2996, 1995.

A. D. Fox and S. J. Maskell. A nested primitive equation model of the iceland-faeroe front. *Journal of Geophysical Research*, 101:18259–18278, 1996.

J.M. Fritsch, J. Hilliker, J Ross, and R.L. Vislocky. Model consensus. *Weather Forecasting*, 15:571–582, 2000.

I. Fukumori and P. Malanotte-Rizzoli. An approximate kalman filter for ocean data assimilation: An example with one idealized Gulf Stream model. *Journal of Geophysical Research*, 100:1831–1855, 1995.

L. S. Gandin. *Objective analysis of meteorological fields*. Israel Program for Scientific Translation, Jerusalem, 1965. 242 pp.

A. Gelb. *Applied optimal estimation*. MIT Press, Cambridge, MA, 1974. 374 pp.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/vairance dilemma. *Neural Computation*, 4:1–58, 1992.

I. Ginis, R. A. Richardson, and L. M. Rothstein. Design of a multiply nested primitive equation ocean model. *Monthly Weather Review*, 126:1054–1079, 1998.

Mark Greenwood and Rob Oxspring. The Applicability of Occam's Razor to Neural Network Architecture. Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK, 2001.

F. Grilli and N. Pinardi. The computation of rossby radii of deformation for the mediterranean sea. *MTP News No. 6*, March, 1998.

F. Guindos-Rojas, M. Cantón-Garbín, J.A. Torres-Arriaza, M. Peralta-López, J.A. Piedra-Fernández, and A. Molina-Martínez. Automatic recognition of ocean structures from satellite images by means of neural nets and expert systems. In *Proceedings of ESA-EUSC 2004 - Theory and Applications of Knowledge-Driven Image Information Mining with Focus on Earth Observation (ESA SP-553). 17-18 March 2004, Madrid, Spain. Compiled by: H. Lacoste, L. Ouwehand. Published on CDROM., p.11.1*, 2004.

S. Guinehut, G. Larnicol, and P.Y. Le Traon. Design of an array of profiling floats in the north atlantic from model simulations. *Journal of Marine Systems*, 35:1–9, 2002.

S. Guinehut, P.Y. Le Traon, G. Larnicol, and S. Philipps. Combining argo and remote-sensing data to estimate the ocean three-dimensional temperature fields-a first approach based on simulated observations. *Journal of Marine Systems*, 46:85–98, 2004.

K Haines. *Ocean Forecasting, Conceptual basis and Applications, edited by N. Pinardi and J.D. Woods*, chapter Assimilation of Satellite Altimetry in Ocean Models, page 472. Springer-Verlag, Berlin Heidelberg New York, 2002.

N. Hamad, C. Millot, and I. Taupier-Letage. The surface circulation in the Eastern Basin of the Mediterranean Sea. In *Proceedings of the European Geophysical Union*, Austria, April 2003. EGU.

T. M Hamill, J.S. Whitaker, and C. Snyder. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Monthy Weather Review*, 129:2776–2790, 2001.

E. Hanert, D.Y. Le Roux, V. Legat, and Deleersnijder E. Advection schemes for unstructured grid ocean modelling. *Ocean Modelling*, 7:39–58, 2004.

S. Haykin. *Neural networks. A comprehensive foundation*. IEEE Press, 1994.

A. W. Heemink, M. Verlaan, and A. J. Segers. Variance Reduced Ensemble Kalman Filtering. *Monthly Weather Review*, 129:1718–1728, 2001.

C. Herbaut, L. Mortier, and M. Crepon. A sensitivity study of the general circulation of the Western Mediterranean Sea: Part I. The response to density forcing through the straits. *Journal of Physical Oceanography*, 26:65–84, 1996.

C. Herbaut, F. Cordron, and M. Crepon. Separation of a coastal current at a strait level: case of the strait of Sicily. *Journal of Physical Oceanography*, 28: 1346–1362, 1998.

I. Herlin, F.-X. Le Dimet, E. Huot, and J.-P. Berroir. Coupling models and data: which possibilities for remotely-sensed images? In P. Prastacos, U. Cortés, J.-L. Diaz De Leon, and M. Murillo, editors, *e-Environment: Progress and Challenge*, volume 11 of *Research on Computing Science*, pages 365–383. Instituto Politécnico Nacional, November 2004.

A. E. Hill. *The Sea*, chapter Buoyancy effect in coastal and shelf seas. John Wiley, 1998.

I. Hoteit, D. Pham, and J. Blum. A Semi-Evolutive Partially Local Filter for Data Assimilation. *Marine Pollution Bulletin*, 43:164–174, 2001.

P. L. Houtekamer and H. L. Mitchell. Data assimilation using ensemble kalman filter technique. *Monthly Weather Review*, 126:796–811, 1998.

K. Ide, P. Bennett, M. Courtier, M. Ghil, and A.C. Lorenc. Unified notation for data assimilation: Operational, sequential and variational. *Journal of Meteorological Society of Japan*, 75(1B):181–189, 1997.

A. H. Jazwinski. *Stochastic Processes and Filtering Theory*. Academic, San Diego, Calif., 1970.

R.E. Kalman. A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82(D):35–45, 1960.

E. Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, Cambridge, 2003.

E. Kalnay and M. Ham. Forecasting forecast skill in the southern hemisphere. In *Preprints of the 3rd Internation Conference on Southern Hemisphere Meteorology and Oceanography, Buenos Aires, 13-17 Nov. 1989. Boston, MA: Amer. Meteor. Soc.*, 1989.

Lakshmi H. Kantha and Carol Anne Clayson. *Numerical models of Oceans and Oceanic processes*, volume 66 of *International geophysics series*. Academic press, 2000. 940 pp.

G. A. Kivman. Sequential parameter estimation for stochastic systems. *Nonlinear Processes in Geophysics*, 10:253–259, 2003.

J. Kondo. Air-sea bulk transfer coefficients in diabatic conditions. *Boundary-Layer Meteorol.*, pages 91–112, 1975.

G. Korres and A. Lascaratos. A one-way nested eddy resolving model of the Aegean and Levantine basins: implementation and climatological runs. *Annales Geophysicae*, 21:205–220, 2003.

T. N. Krishnamurti, C. M. Kishtawal, T. La Row, D. Bachiochi, Z Zhang, E. Williford, S. Gadgil, and S. Surendran. Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433):1548–1550, 1999.

T. N. Krishnamurti, C. M. Kishtawal, Z Zhang, T. La Row, D. Bachiochi, and E. Williford. Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, 13:4196–4216, 2000.

H. Lacombe and P. Tchernia. Les zones de formation d'eau profonde océanique: Caractères–processus–problèmes, in Processus de formation des eaux océaniques profondes. *Colloq. Int. Centre Natl. Rech. Sci.*, 215:249–262, 1974.

G. Larnicol, N. Ayoub, and P.Y. Le Traon. Major changes in Mediterranean Sea level variability from seven years of TOPEX/Poseidon and ERS-1/2 data. *Journal of Marine Systems*, 33–34:63–89, 2002.

C.E. Leith. Atmospheric predictability and two-dimensional turbulence. *Journal of Atmospheric Sciences*, 28:145–161, 1971.

C.E. Leith. Theoretical skill of monte carlo forecasts. *Monthly Weather Review*, 102:409–418, 1974.

C.E. Leith and R.H. Kraichnan. Predictability of turbulent flows. *Journal of Atmospheric Sciences*, 29:1041–1048, 1972.

P. F. J. Lermusiaux. *Error Subspace Data Assimilation Methods for Ocean Field Estimation: Theory, Validation and Appliciation*. PhD thesis, Harvard University, Cambridge, Mass., 1997.

P. F. J. Lermusiaux and A. R. Robinson. Data assimilation via error subspace statistical estimation. Part 1: Theory and schemes. *Monthly Weather Review*, 127:1385–1407, 1999.

P.F.J. Lermusiaux, C.-S. Chiu, G.G. Gawarkiewicz, P. Abbot, A.R. Robinson, R.N. Miller, P.J. Haley, W.G. Leslie, S.J. Majumdar, A. Pang, and F. Lekien. Quantifying uncertainties in ocean predictions. *Oceanography*, 49:80–88, 2006.

P. Malanotte-Rizzoli, B. Manca, M. Ribera d'Alcala, A. Theocharis, S. Brenner, G. Budillon, and E. Ozsoy. The Eastern Mediterranean in the 80s and in the 90s: the big transition in the intermediate and deep circulations. *Dynamics of Atmospheres and Oceans*, 29:365–395, 1999.

S. Marullo, R. Santoleri, P. Malanote-Rizzoli, and A. Bergamasco. The sea surface temperature field in the Eastern Mediterranean from advanced very high resolution radiometer (AVHRR) data. Part I. Seasonal variability. *Journal of Marine Systems*, 20:63–81, 1999.

A. McDonald. Lateral boundary conditions for operational regional forecast models: a review. *Hirlam Technical Report*, 32, 1997.

MEDOC Group. Observation of formation of deep water in the Mediterranean Sea. *Nature*, 227:1037–1040, 1970.

R.N. Miller, M. Ghil, and F. Gauthiez. Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of Atmospheric Sciences*, 51:1037–1056, 1994.

C. Millot. *Hydrodynamics of Semi-enclosed Seas, edited by Nihoul*, chapter Analysis of upwelling in the Gulf of Lion. 1982.

C. Millot. The Gulf of Lions' hydrodynamics. *Cont. Shelf Res.*, 10:885–894, 1990.

C. Millot. Circulation in the western mediterranean sea. *Journal of Marine Systems*, 20:423–442, 1999.

C. Millot and I. Taupier-Letage. Additional evidence of LIW entrainment across the Algerian subbasin by mesoscale eddies and not by a permanent westward flow. *Progress In Oceanography*, 66:231–250, 2005.

C. Millot, M. Benzhora, and I. Taupier-Letage. Circulation in the Algerian Basin inferred from the MEDIPROD-5 current meters. *Deep-Sea Research*, 44:1467–1495, 1997.

C. Millot, J.-L. Fuda, J. Candela, and Y. Tber. Large warming and salting of the Mediterranean outflow due to changes in its composition. In *Proceedings of the European Geophysical Union*, Austria, 2005. EGU. submitted to Nature.

L. Nerger, W. Hiller, and J. Schröter. A comparison of error subspace Kalman filters, part 1: Filter algorithms. *Monthly Weather Review*, 2004a. submitted.

L. Nerger, W. Hiller, and J. Schröter. A comparison of error subspace Kalman filters, part 2: Evaluation of numerical experiments. *Monthly Weather Review*, 2004b. submitted.

Jacques C. J. Nihoul, Eric Deleersnijder, and Salim Djenidi. Modelling the general circulation of shelf seas by 3D $k - \epsilon$ models. *Earth-Science Reviews*, 26:163–189, 1989.

L.-Y. Oey and P. Chen. A Nested-Grid Ocean Model: With Application to the Simulation of Meanders and Eddies in the Norwegian Coastal Current. *Journal of Geophysical Research*, 97(C12):20063–20086, December 1992.

Reiner Onken, Allan R. Robinson, Lakshmi Kantha, Carlos J. Lozano, Patrick J. Haley, and Sandro Carniel. A rapid response nowcast/forecast system using multiply nested ocean models and distributed data systems. *Journal of Marine Systems*, 56:45–66, 2005.

R. Person. Un exemple de descente des eaux superficielles du plateau continental dans un canyon du Golfe du Lion. *in La Formation des Eaux Oceaniques Profondes, edited by Colloques Internationaux du CNRS*, 1974.

A. Petrenko. Circulation features in the Gulf of Lions, NW Mediterranean Sea; importance of inertial currents. *Oceanologica Acta*, 26:323–338, 2003.

A. Petrenko, Y. Leredde, and P. Marsaleix. Circulations in a stratified and wind-forced Gulf of Lions, NW Mediterranean Sea: in situ and modeling data. *Continental Shelf Research*, 25:7–27, 2005.

D. T. Pham. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly Weather Review*, 129:1194–1207, 2001.

D. T. Pham, J. Verron, and L. Gourdeau. Singular evolutive Kalman filters for data assimilation in oceanography. *C. R. Acad. Sci. Ser. II*, 326(4):255–260, 1998a.

D. T. Pham, J. Verron, and M. C. Roubaud. A singular evolutive extended Kalman filter for data assimilation in oceanography. *Journal of Marine Systems*, 16(3–4):323–340, October 1998b.

J. Pietrzak, E. Deleersnijder, and J. Schroeter. The second international workshop on unstructured mesh numerical modelling of coastal, shelf and ocean flows (delft, the netherlands, september 23-25, 2003). *Ocean Modelling (special issue)*, 10, 2005.

N. Pinardi, I. Allen, E. Demirov, P. De Mey, A. Lascaratos, P.-Y. Le Traon, C. Maillard, G. Manzella, and C. Tziavos. The mediterranean ocean forecasting system: first phase of implementation (1998-2001). *Annales Geophysicae*, 21:3–20, 2003.

J.-M. Pinot, J. Tintoré, and D. Gomis. Quasi-synoptic mesoscale variability in the Balearic sea. *Deep-Sea Research*, 41:897–914, 1994.

R.W. Preisendorfer. *Principal Component Analysis in Meteorology and Oceanography*. Elsevier, 1988.

I. Puillat, I. Taupier-Letage, and C. Millot. Algerian Eddies lifetime can near 3 years. *Journal of Marine Systems*, 31:245–259, 2002.

M. Rixen and E. Ferreira-Coelho. Operational surface drift prediction using linear and non-linear hyper-ensemble statistics on atmospheric and ocean models. *Journal of Marine Systems*, 2006.

M. Rixen, J. T. Allen, and J.-M. Beckers. Diagnosing vertical velocities using the QG omega equation: a relocation method to obtain pseudo-synoptic data sets. *Deep-Sea Research*, 48(6):1347–1373, march 2001.

M. Rixen, S. Alderson, J.T. Allen, A. Barth, J.-M. Beckers, V. Cornell, N. Crisp, S. Fielding, A.T. Mustard, R.T. Pollard, E.E. Popova, D.A. Smeed, and M.A. Srokosz. Along or across front survey strategy? An operational example at an unstable oceanic front. *Geophysical Research Letters*, 30(1):1017–1020, 2003a.

M. Rixen, J.T. Allen, R. Pollard, and J.-M. Beckers. Along or across front survey strategy? The estimation of quasi-geostrophic vertical velocities and temperature fluxes. *Geophysical Research Letters*, 30(5):1264–1267, 2003b.

M. Rixen, J.M. Beckers, S. Levitus, J. Antonov, T. Boyer, C. Maillard, M. Fichaut, E. Balopoulos, S. Iona, H. Dooley, M.J. Garcia, B. Manca, A. Giorgetti, G. Manzella, N. Mikhailov, N. Pinardi, and M. Zavatarelli. The Western Mediterranean Deep Water: A proxy for climate change. *Geophysical Research Letters*, 32, 2005.

A.R. Robinson, G.L. Wayne, A. Theocharis, and A. Lascaratos. *Encyclopedia of Ocean Sciences, edited by J. Steele, K. Turekian and S. Thorpe*, chapter Mediterranean Sea Circulation. Academic Press, 2001.

A. Rosati and K. Miyakoda. A general circulation model for upper ocean simulation. *J. Phys. Oceanogr.*, 18(11):1601–1626, 1988.

F. Rosenblatt. *Principles of neurodynamics*. Spartan, 1963.

D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

J. Salat and J. Font. Water mass structure near and offshore the Catalan coast during the winters of 1982 and 1983. *Annales Geophysicae*, 1B:49–54, 1987.

S. Sammari, C. Millot, and L Prieur. Aspects of the seasonal and mesoscale variabilities of the Northern Current in the western Mediterranean Sea inferred from PROLIG-2 and PROS-6 experiments. *Deep Sea Research I*, 42: 893–917, 1995.

Segond, Robilliard, and Fonlupt. Iterative filter generation using genetic programming. In *EuroGP 2006*, 2006.

C. Shalizi, R. Haslinger, J.B. Rouquier, K. Klinkner, and C. Moore. Automatic filters for the detection of coherent structure in spatiotemporal systems. *Physical Review E*, 73, 2005.

J. She, P. Berg, and J. Berg. Bathymetry impacts on water exchange modelling through the Danish Straits. *Journal of Marine Systems*, 65:450–459, 2007.

S. Skachko, L. Berline, L. Bertino, J.M. Brankart, P. Brasseur, Y. Ourmières, J. Schroter, P.J. van Leeuwen, and J. Verron. Recent advances in data assimilation in the mersea project. In *15 years of progress in radar altimetry*, pages 4891–4907, 2006.

N. Skliris and A. Lascaratos. Impacts of the Nile River damming on the thermohaline circulation and water mass characteristics of the Mediterranean Sea. *Journal of Marine Systems*, 52:121–143, 2004.

W. H. F. Smith and D. T. Sandwell. Global sea floor topography from satellite altimetry and ship depth soundings. *Science*, 277:1956–1962, 1997.

M. A. Spall and W. R. Holland. A nested primitive equation model of oceanic application. *Journal of Physical Oceanography*, 21:205–220, 1991.

M. A. Spall and A. R. Robinson. A new open-ocean hybrid coordinate primitive equation model. *Math. Comput.*, 31:241–269, 1989.

S. Sparnocchia, P. Picco, G. M. R. Manzella, A. Ribotti, S. Copello, and P. Brasey. Intermediate water formation in the Ligurian Sea. *Oceanologica Acta*, 18:151–162, 1995.

O. Talagrand. Assimilation of observations, an introduction. *J. Meteor. Soc. Japan*, 75:191–209, 1997.

O. Talagrand. Assimilation of observations into numerical models.

F.T. Tangang. Forecasting the equatorial Pacific sea surface temperatures by neural network models. *Climate Dynamics*, 13(2):135–147, 1997.

H. Thonet, B. Lemonnier, and R. Delmas. Automatic segmentation of oceanic eddies on AVHRR thermal infrared sea surface images. In *OCEANS '95. MTS/IEEE. 'Challenges of Our Changing Global Environment'. Conference Proceedings.*, volume 2, pages 1122–1127, 1995.

V. Toumazou and J.F. Cretaux. Using a Lanczos Eigensolver in the Computation of Empirical Orthogonal Functions. *Monthly Weather Review*, 129: 1243–1250, 2001.

M.H. Tusseau and J.M. Mouchel. Nitrogen inputs to the gulf of lions via the rhône river. In J.M. Martin and H. Barth, editors, *Water Pollution Research Reports, Proceedings of the EROS 2000 Workshop, Hamburg*, March 1994. Commission of the European Communities.

P.J. van Leeuwen. A Variance-Minimizing Filter for Large-Scale Applications. *Monthly Weather Review*, 131:2071–2084, 2003.

P.J. van Leeuwen. Ensemble Kalman filters, Sequential Importance Resampling and beyond. 131:2071–2084, 2007. In preparation.

P.J. van Leeuwen. Comment on "Data assimilation using an ensemble Kalman filter technique". *Monthly Weather Review*, 127:1374–1377, 1999.

L. Vandenbulcke. Assimilation de données dans les modèles gigognes. Master's thesis, University of Liège, DEA Européen en modelisation de l'environnement marin, june 2003. 61 pp.

L. Vandenbulcke, A. Barth, M Rixen, A. Alvera-Azcarate, Z. Ben Bouallegue, and J.M. Beckers. Study of the combined effects of data assimilation and grid nesting in ocean models - application to the Gulf of Lions. *Ocean Science*, 2: 213–222, 2006.

M. Verlaan and A. W. Heemink. Tidal flow forecasting using reduced rank square filters. *Stochastic Hydrology and Hydraulics*, 11:349–368, 1997.

L. Wehenkel. *Applied Inductive Learning. Course notes: october 2000.* Institut Montefiore, ULG, Liège, 2000.

J. S. Whitaker and T. M. Hamill. Ensemble data assimilation without perturbed obsevations. *Monthly Weather Review*, 130:1913–1924, 2002.

N.K. Yilmaz, C. Evangelinos, N.M. Patrikalakis, P.F.J. Lermusiaux, P.J. Haley, W.G. Leslie, A.R. Robinson, Wang D., and H. Schmidt. Path planning methods for adaptive sampling of environmental and acoustical ocean fields. In *Proceedings of IEEE/MTS Oceans 06 Conference, Boston, MA, September 18-21*, page 6pp, 2006.

M. Zavatarelli and G.L. Mellor. A numerical study of the Mediterranean Sea. *Journal of Physical Oceanography*, 25(6), 1995.

M. Zavatarelli and N. Pinardi. The Adriatic Sea modelling system: a nested approach. *Annales Geophysicae*, 21(1), 2003.

Index