Constraining model biases in a global general circulation model with ensemble data assimilation methods

Martin Canter

GHER, GeoHydrodynamics and Environment Research

Department of Astrophysics, Geophysics and Oceanography

University of Liege

A thesis submitted for the degree of

Docteur en Sciences

March 2017

Committee members

Prof. Jean-Paul Donnay, President (University of Liège)
Prof. Jean-Marie Beckers (University of Liège)
Prof. Hugues Goosse (SST/ELI/ELIC, Université Catholique de Louvain)
Directeur de recherche CNRS Pierre Brasseur (IGE, Université Grenoble Alpes)
Dr. Charles-Emmanuel Testut (Mercator-Ocean)
Dr. François Counillon (Nansen Environmental and Remote Sensing Center)
Dr. Alexander Barth (University of Liège)

Acknowledgements

I would like to thank my thesis supervisor, Dr. Alexander Barth. He initially gave me the opportunity to discover and awaken my interest for numerical modelling and oceanography during my master thesis. During the four years of my PhD, his support, expertise and advices guided me through the multiple difficulties of this research. He was always available and never failed to provide a response to any of my question.

I would like to thank Prof. Jean-Marie Beckers the insightful review of my work, and the multiple ideas to be explored, especially during the very pleasant DIVA workshops.

I would like to thank the rest of my thesis committee, Prof. Jean-Paul Donnay, Prof. Hugues Goosse (SST/ELI/ELIC, Catholic University of Louvain), Directeur de recherche CNRS Pierre Brasseur, Dr. Charles-Emmanuel Testut, and Dr. François Counillon for the careful review of this manuscript.

A special thank to Arthur Capet, who helped me through the first months of my PhD, and shared his office with me.

I would also like to thank my colleagues from the GHER, Sylvain Watelet, Aïda Alvera Azcarate, Thi Hong Ngu Huynh, Charles Troupin, Stephane Lesoinne, Igor Tomazic, Subekti Mujiasih, Gaelle Parard and Svetlana Karimova for the Friday afternoon coffees and the ensuing discussions and presentations who made me discover other aspects of oceanography.

To my upstairs and downstairs neighbours and colleagues, Maxime Hubert, Sebastien Mawet, and Christophe Becco, I want to express my thanks for your help, encouragements, and questions which provided me with a different point of view on my work.

I am also thankful to the two wonderful persons who organise all the events around the GHER, Charlotte Peelen and Cécile Pregaldien.

A special thanks to Manon Mathieu, without whose support throughout those four years this thesis would not have been possible.

I would like to thank two friends, Romain Van der Keilen, and Werrick Deneu, whose companionship and help to solve programming issues have been immensely helpful.

A special thank to my family, whose encouragements helped me through all my levels of educations.

This work was funded by the project PREDANTAR (SD/CA/04A) from the federal Belgian Science policy (http://www.climate.be/PREDANTAR) and the Sangoma FP7-SPACE-2011 project (grant 283580) (http://www.data-assimilation.net/). Alexander Barth is an F.R.S. - FNRS Research Associate. Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11. MDT_CNES-CLS09 was produced by CLS Space Oceanography Division and distributed by Aviso, with support from CNES (http://www.aviso.altimetry.fr/).

Abstract

A new method of bias correction using an ensemble transform Kalman filter as data assimilation scheme is developed. The objective is to create a stochastic forcing term which will partially remove the bias from numerical models. The forcing term is considered as a parameter to be estimated through state vector augmentation and the assimilation of observations.

The theoretical formulation of this method is introduced in the general context of numerical modelling. A specially designed and modified Lorenz '96 model is studied, and provides a testing environment for this new bias correction method. Several different aspects are considered through both single and iterative assimilation in a twin experiment.

The method is then implemented on the global general circulation model of the ocean NEMO-LIM2. The forcing term generation is detailed to respect particular physical constraints. Again, a twin experiment allows to assess the efficiency of the method on a realistic model. The assimilation of sea surface height observations is performed, with sea surface salinity and temperature as control variable. Subsequently, a multivariate assimilation shows further improvement of the bias correction.

Finally, the method is confronted to real sea surface height observations from the CNES-CLS09 global mean dynamic topography. A thorough study of the NEMO-LIM2 model response to the bias correction forcing term is proposed, and specific results are highlighted. An iterative assimilation concludes the method investigation. Possible ideas and future developments are suggested.

Abstract

Une nouvelle méthode de correction de biais utilisant un filtre d'ensemble de Kalman transformé est développée. L'objectif est de construire un terme de forçage stochastique afin de réduire le biais d'un modèle numérique. Ce terme de forçace est considéré comme un paramètre à estimer via l'augmentation du vecteur d'état lors de l'assimilation d'observations.

La formulation théorique de cette méthode est présentée dans le contexte de la modélisation numérique. Le modèle de Lorenz '96, spécifiquement modifié dans le cadre de cette étude, permet de disposer d'un environnement controlé pour éprouver la méthode de correction du biais. Une expérience jumelle est utilisée pour expérimenter ses différents aspects au travers d'assimilation successivement simple et itérative.

Cette méthode est ensuite implémentée sur le modèle global et général de circulation de l'ocean NEMO-LIM2. La génération du terme de forçage est détaillée afin de respecter différentes contraintes physiques et numériques. Une expérience jumelle permet d'évaluer l'efficacité de la méthode sur un modèle réaliste. La hauteur de la surface de la mer est considérée comme donnée d'observation et assimilée, la température et la salinité de surface de la mer servant de variable de contrôle. Enfin, ces deux premières variables sont assimilées simultanéement, permettant la comparaison avec l'assimilation simple.

Pour terminer, la méthode est confrontée à un cas opérationnel, avec des données de topographie dynamique moyenne provenant de CNES-CLS09. Une étude approfondie du modèle NEMO-LIM2 lors de l'application de la méthode de correction du biais est présentée. L'assimilation itérative de ces mêmes données cloture les expériences menées autour de cette méthode. Différentes idées de développements futurs sont proposées.

Contents

Chapter 1

Introduction

Contents

1.1 Historical perspective

Historically, in 1922, Richardson published the first attempt at a numerical forecast of weather (Richardson, 1922; Lynch, 2008). His so called "forecast factory" interpolated available observations to build initial conditions at t_0. From there, by hand, he produced a 6-hour forecast of the atmosphere over two points in central Europe, using a hydrostatic variation of Bjerkness' primitive equations. The results were, unfortunately, completely unrealistic, resulting in a surface pressure over a six-hour period of 145 hPa (Lynch, 2006). The simple interpolation used did not respect the physical constraints, therefore developing unstable modes which eventually conducted to model instability. Moreover, the initial conditions were unnatural. This inevitably led to spurious tendencies and imbalance between the pressure and wind fields. Despite the generally favorable reviews of his work, the impracticality of his method with the available tools deterred others to follow his path for a couple of decades.

It is only with the arrival of computers that weather prediction and climate modelling would again be conceivable. In the early 1950's, a team was able to produce 24-hour forecasts, which needed 24 hours of computation. It was the first time that

numerical weather prediction was able to keep up with time. Improvements in numerical analysis, with new and stable algorithms, the development of data retrieval tools, such as satellites, and the introduction of data exchange, provided a solid and fertile ground for further research in numerical modelling of the earth.

The close relationship between the atmosphere and the ocean, whether it be the physical properties and transfers between both, or the similarities regarding the numerical particularities of those systems, led to ocean modelling to follow the path of weather prediction. Two major operational differences remain between the two geofluids. The first resides in the interest of producing daily forecasts of those systems: the need for daily atmospheric weather forecasts did not originally find a similar interest in ocean modelling. It is only recently that the need for seasonal and inter-annual oceanic forecasts arose. In particular, the influence of the ocean on global climate, and the prediction of large-scale phenomena linked to the ocean, such as El-Niño, required a knowledge of oceanic processes to be available beforehand. There is also an interest in short-range forecasts (on an oceanic scale) for e.g. world navies, fisheries, off-shore drilling, search and rescue operations, or oil spill forecasts. The second difference relates to the available data sets, which are considerably smaller for the ocean. The ocean, compared to the atmosphere, is a harsh environment which strains data collection systems. Whereas, on the continents, observations stations have been monitoring the weather for a long time, similar tools could not be used in the ocean. For a long time, due to the necessity of having a human performing the measurements, oceanic measures only relied on boats sailing through the ocean. It is only since the arrival of electronics that automatic drifters and buoys have provided regular data sets. The rise of satellites also greatly helped to provide large coverage for the size of the oceans. Consequently, the available measures of the ocean still lack sufficient coverage for extended time periods, and for remote regions.

Nonetheless, numerical modelling of geophysical systems brings a certain number of problems which can not be ignored. Examples are the limitation that comes from the finite computational power available which results in limited resolution, imperfect specification of boundary conditions, and poor representation of sub-grid physical processes. The initialization of a model also requires either a climatology based on another model or retrieved through observations, or an excellent knowledge of the system, in order to avoid early imbalances in the model. All those issues require a particular treatment with specifically developed methods.

1.2 Data assimilation

One of those methods is data assimilation. It is a technique developed to handle the difficulties that arise from numerical modelling. It describes a method which aims at determining the state of a model by combining heterogeneous and imperfect observations in an optimal way. In particular, its purpose is to correct numerical models representing an dynamical system for which observations are available. A perfect representation of a real dynamical system is not possible. As such, compromises have to be made, and lead to poor processes parametrisation and approximate initial states. Errors are therefore inevitable, and need to be coped with, in order to obtain a more accurate estimation of the state of the system.

Data assimilation can be summarised as the combination of mathematical algorithms which incorporates observations into the model state of a numerical model, using prior error and statistical information about both the observations and the model state, and the mathematical equations describing and governing the model. Those available information are all different in nature, quantity and quality. Commonly, the state of the model prior to assimilation is called the forecast, and the result of the assimilation is then an optimal representation of the model state which is called the analysis.

A concrete example can easily be detailed for the weather forecast. After taking the temperature in two cities during a couple of days, one wishes to forecast the daily temperature in the forthcoming days, for both cities, and the land that separates them. A relatively straightforward method is to extrapolate the evolution of the temperature, assuming that the measurements curve is smooth enough. One could also rely on a statistical database of previous evolution of temperature. With the help of numerical modelling, one could also aim to forecast the temperature through simulations. On the next day, the forecast can be compared with the real weather. By taking new measurements in both cities, the model forecast can be corrected. However, supposing that no data is available for the space separating both cities, the model would still also need correction at those points. Finally, errors on the measurements need also be to taken into account. Data assimilation techniques try to optimise those corrections by using the available information about the model and the measured data.

Originally developed for meteorology and numerical weather prediction, data assimilation dates back to the 1950's (Panofsky, 1949) , and has a long history of applications in many different fields such as oceanography, glaciology, seismology, nuclear fusion, medicine, agronomy, ... The recurring common factor of data assimilation applications is inaccurate numerical models for which observations are available. In particular in geoscience, with high dimensional systems, the number of observations and the huge number of variables in the system, respectively up to 10^7-10^9 nowadays, are a real issue and impose constraints due to the limited computational power available.

At first, empirical methods were developed. The successive corrections method (SCM) dates back to 1955 (Bergthörsson and Döös, 1955), and is an iterative process which starts with a background field or first guess. A linear weighing scheme is then used to adjust the initial guess and subsequent estimates to fit the grid point values to the observations, taking into account the ratio of the observation error variance to the background error variance (Cressman, 1959). It was improved later by Barnes (1964) for analyses where no available background field could be provided. Bratseth (1986) showed that SCM can be made to converge to a optimal interpolation provided that adequate weights are chosen.

Data assimilation methods can currently be divided into different categories, depending on their inherent hypotheses, and their approaches. In essence, they can be either sequential or nonsequential. In sequential assimilation, the model is integrated over a period of time with no available observations. When the model reaches a point in time where observations are present, it is stopped. The model is then used as a first guess or background estimate which needs to be corrected or updated with the observations. The statistical information about the model and the observations error are used to obtain the best corrected state, or analysis. The model is then restarted from that point in time. This process is repeated until all observations are used. One can thus look at sequential data assimilation as a succession of integrations and corrections over a period of time where observations are distributed and available. The particularity of sequential data assimilation is that the observations are only used to correct the model forward in time. There is no backwards correction of the previous or initial state from later observations.

Nudging is an empirical and sequential method, also called Newtonian relax-

ation. In this method, an additional source term, or sink term, is added to the prognostic equations of the observed variables. This will then nudge the solution towards the observation (interpolated to the model grid) during the integration of the model equations. Hence, nudging has a continuous effect on the model, at every time step. A relaxation time scale is chosen based on empirical considerations. A too short relaxation time and the solution will converge too fast, causing imbalanced model states. A too long relaxation time causes model errors to grow too large. Example of this method application are altimetry assimilation in the North Atlantic (Blayo et al., 1996), or a 15 year reanalysis from the ECMWF (European Centre for Medium-Range Weather Forecasts) (Kaas et al., 1999). In practice, nudging is easy to implement, and the computational cost is nearly negligible. However, it is not applicable when using indirect observations, and unobserved variables have to adjust themselves through the model dynamics. Still, nudging is still used for specific applications, such as the assimilation of gridded data from a reanalysis. More recently, Auroux and Blum (2005) proposed a back and forth nudging, using an inverse source term and integrating the model backwards in time to the initial state.

Optimal interpolation is also a sequential assimilation method (Gandin and Hardin, 1965). Unlike nudging, optimal interpolation uses physical assumptions and error statistics. For instance, one can connect sea surface temperature and altimetry to general circulation models, such as will be used later in this work. Optimal interpolation refers to methods originating from linear estimation theory. It is based on a background error covariance matrix which is assumed constant in time. More recently, optimal interpolation schemes using a different parametrisation for the error covariance matrix and its numerical representation have been developed, such as the multivariate optimal interpolation scheme (Daley, 1991), or the ensemble optimal interpolation scheme (Oke et al., 2002), used in Counillon and Bertino (2009) to forecast the Loop Current in the Gulf of Mexico.

3D-Var is another sequential assimilation method (Courtier et al., 1998). The idea is to provide the model state as input, which is then adjusted in such a way that the model output is as close as possible to the observations and to the background field. 3D-Var can be seen as a variational method in the sense that it requires the development of the adjoint of the observation operator. Therefore, the model dynamics is not involved, and the relationship between the model state and the observations is performed through the adjoint. If this observational model is linear, 3D-Var becomes equivalent to optimal interpolation.

The Kalman filter, and its numerous derivations, are also sequential methods (Kalman, 1960). Essentially the background and its error covariance matrix are computed and updated at each assimilation cycle. It assumes that the errors are additive, unbiased, uncorrelated and Gaussian distributed for the model and the observations, and that the model and observation operator are linear. The extended Kalman filter proposes to linearise the model dynamics and the nonlinear observation operator (Jazwinski, 1970). The Kalman filter and the extended version can not be used directly on realistic models, where the model dynamics is nonlinear, causing the error covariance matrix to become difficult to compute. An interesting derivation of the Kalman filter is its ensemble formulation, known as the ensemble Kalman filter (Evensen, 1994). It was introduced to overcome the the limitations of the first-order approximation of the extended Kalman filter through a Monte Carlo approach. The full derivation of the Kalman filter, the extended Kalman filter and the ensemble Kalman filter, is detailed in chapter 3. Because of their efficiency, robustness, and the ease of implementation, the Kalman filter and all the existing modifications are very popular in geophysical applications (Edwards et al., 2015).

The particle filter eliminates the need for the Gaussian assumption required by the Kalman filter (DEL MORAL et al., 1995; Van Leeuwen, 2009). Therefore, it handles the nonlinearities much better than the Kalman filter. It is directly drawn from the Bayes theorem, without any additional assumption. Each ensemble member (similarly to the ensemble Kalman filter) is called a particle, and is integrated through the nonlinear model. The objective is to represent the posterior probability function without any prior assumption on its distribution. When observations become available, the information contained in the observations is incorporated into the particles. Probabilities based on an estimated likelihood function resample then the ensemble, and provide a new analysed ensemble. The particle filter is also a sequential assimilation method. It is however difficult to apply to realistic cases due to the curse of dimensionality of high dimensional data and state spaces, causing a large number of particles to be required.

In nonsequential assimilation methods, the information is propagated both forward and backwards. This allows the estimation of a past model state based on posterior observations. For example, the Kalman smoother (Gelb, 1974) is an extension of the Kalman filter which uses past and future observations by integrating the time dimension into the state vector. In Cosme et al. (2010), a square-root

smoother algorithm is presented as an extension from the singular evolutive extended Kalman filter (Pham et al., 1998). The same generalisation can be applied to the particle filter, which is then called a particle smoother (Tanizaki, 2001).

Another class of nonsequential assimilation methods are adjoint methods. They are powerful tools that provide and estimate of the sensitivity of a model output with respect to an input. In particular, in data assimilation, an optimal analysis is one that fits the observations best, using assumptions and available information about the error characteristics of the data used. This is where adjoint methods perform efficiently, allowing the optimization problem to be solved in a reasonable time for application to real-time forecasting (Errico, 1997). Basically, adjoint methods make use of adjoint of the model whose solution is being examined (hence, "adjoint"), where the adjoint is formally defined as the transpose of the tangent linear model. A control solution and a measure of the forecast are considered. The gradient of this forecast at the control solution is evaluated with respect to perturbations of each component of the model output. This gradient can be interpreted as the sensitivity of the forecast to the perturbations. Those perturbations can be on the initial conditions, boundary conditions, or even model parameters.

However, adjoints methods also show limitations. The model equations have to be differentiable, which is not always the case (Zupanski et al., 2008). One also needs to store either the complete trajectory, or to be able to partially recompute it to evaluate the adjoint. This involves sophisticated check-point techniques to efficiently solve the problem. The derivation of the adjoint (and the tangent linear) model can prove to be difficult and tedious. Some improvements in automatic adjoint compilers have been performed, but the procedure is still challenging (Heimbach et al., 2005). Additionally, the cost-function might have a local minimum, causing the minimisation to converge only to the local minimum, and not the global one. In particular, for 4D-Var, if the probability density function has multiple secondary modes, finding the global one can be a challenge.

4D-Var is a nonsequential data assimilation method which makes use of the adjoint of the model (Dimet and Talagrand, 1986; Talagrand and Courtier, 1987). It is able to take all observations into account. However, all error sources must be control variables of the optimisation process. Since one can not take into account the error introduced at every time step, the model is assumed to be perfect. This

is imposed as a strong constraint. One can interpret the 4D-Var scheme as an time extension of the 3D-var. 4D-Var is still commonly applied to various practical cases in oceanography (Ngodock and Carrier, 2014; da Rocha Fragoso et al., 2016).

In the representer method (Bennett, 1992), the model error is accounted for, unlike the 4D-Var method. However, this can be prohibitive if a lot of information are available, such as satellite data coverage.

1.3 State of the art

Bias is commonly defined as a systematic error with a nonzero mean. In a more formal formulation, any kind of component of error which is systematic, with regard to the notion of the average of a model or estimator, can be considered as bias (Dee, 2005). The effects of bias can significantly deteriorate the model solution. Bias can take multiple different forms, spatially variable, seasonal, or even depend on specific situations. In numerical modelling, current limitation comes among others from the finite computational power available, which, in ocean models, results in limited resolution. With inaccurate surface forcings, those are examples of model associated bias. The limited knowledge of the system also leads to imperfect specification of boundary conditions, and poor representation of subgrid physical processes (Baek et al., 2009). Those differences between the numerical model and the dynamics of the real ocean induce systematic errors in the numerical forecasts. Daytime high-altitude radiosonde temperatures can be biased due to solar radiation effects, and radar altimetry affected by electromagnetic bias originating from the smaller reflectivity of wave crests than troughs, are examples of observation related bias (Ghavidel et al., 2016). Finally, bias can even come from the assimilation itself, when unbiased observations are assimilated into a biased model. The model drift towards its biased state causes bias in the assimilation, and can lead to apparent changes in characteristics of the observing system (Santer et al., 2004).

When used for prediction, or long term simulations with a limited number of available observations, those systematic errors cause the model to exhibit significant differences in climatologies when compared to the reality. In some circumstances, they can even be comparable or larger than random or nonsystematic error of the model solution. While the random part of the model error has been reduced thanks to several advances in numerical modelling, it has become increasingly necessary to address the systematic model error (Keppenne et al., 2005). The bias in climatic

modelling can be so large, that only variation and anomalies are studied rather that the absolute model results (Zunz et al., 2013).

In the context of oceanography, state of the art ocean models exhibit significant differences in the climatological mean state when compared to observations from the real ocean (Flato et al., 2013). For instance, eddy-mean processes can be poorly represented, which causes western boundary currents to be responsible for large sea surface temperature bias, such as for the Gulf Stream and the Kuroshio currents (Large and Danabasoglu, 2006).

To reduce the error of the model, one can make use of data assimilation schemes. However, a critical assumption for analysis schemes is that the mean of the background error is zero. This hypothesis is clearly violated in the presence of bias. Data assimilation schemes that are designed to use nonbiased observations to correct random errors with zero mean in a model background estimate, are called bias-blind. In presence of bias, those analysis schemes are suboptimal, and can generate spurious corrections and undesired trends in the analysis (Dee and Uppala, 2009). Most data assimilation schemes are designed to handle small, random errors and make small adjustments to the background fields that are consistent with the spatial structure of random errors (Dee, 2005). Unfortunately, due to the systematic character of model errors, their representation as random errors, or noise, is rather poor. In some cases, such as satellite observations, bias can even be larger that the actual, useful signal present in the observations (Cucurull et al., 2014).

Bias-aware data assimilation scheme are designed to simultaneously estimate the model state variables and parameters that are set to represent systematic errors in the system. However, assumptions need to be made about the error covariance of the bias and its attribution to a particular source. It also needs to be represented and expressed in a set of well-defined parameters.

Model bias estimation was first introduced by Friedland (1969), and more deeply described by Jazwinski (1970); Gelb (1974). He suggested a scheme in which the model state vector should be augmented with a decoupled bias component that can be isolated from the other state vector variables. This allows the estimation of the bias prior to the estimation of the model.

Bias correction approaches can be separated into different approaches (Keppenne

et al., 2005; Chepurin et al., 2005). In offline methods, bias is estimated from the model mean and the climatology, using a preliminary model run. Offline methods are simple to implement and have a small computational cost. In online methods, bias is updated during the data assimilation step, resulting in an analysed bias.

The most known and referred to algorithm for online estimation and correction of the bias in sequential data assimilation is presented in Dee and Da Silva (1998). Bias is estimated during the assimilation by adding an extra and separate assimilation step. It was successfully applied in Dee and Todling (2000) to the global assimilation of humidity observations in the Goddard Earth Observing System. A simplified version of this algorithm using a single assimilation step (where Dee and Da Silva (1998) needed two) was applied by Radakovich et al. (2001) to land-surface temperature assimilation, and by Bell et al. (2004) for the online estimation of subsurface temperature bias in tropical oceans. It was also used for model bias estimation by Baek et al. (2006), and observation bias correction in Fertig et al. (2009). In Carton et al. (2000b), a 46-year global retrospective analysis of the upper-ocean temperature, salinity and currents, was performed, with bias originating both from model limitations and poor surface forcings. In Keppenne et al. (2005), bias between the climatology of the model and the data was problematic for the use of satellite altimeter data from TOPEX/Poseidon. In Chepurin et al. (2005), the effect of bias on a 31-year long historical analysis of the physical state of the ocean is studied, with a focus on the mixed layer and thermocline depth in the tropical Pacific Ocean, and in Nerger and Gregg (2008), a singular evolutive interpolated Kalman filter was extended with an online bias correction scheme.

However, a critical requirement is that most methods of bias correction need a reference data set which is defined as bias free, from which a bias estimation can be provided. In practice, it can be difficult to find such a data set. The bias also needs to be charaterised in terms of some well-defined set of parameters. While this is obvious for bias estimation, it is a critical condition when attempting bias correction. The attribution of bias to an erroneous source will force the assimilation to be consistent with a biased source. In some cases, the bias correction would even deteriorate the assimilation procedure, and perform worse than a classic, bias-blind assimilation (Nakamura et al., 2013; Massari et al., 2015).

Hence, the effect of bias on the model climatology can not be neglected. The necessity of removing, or at least, reducing the effects of bias on the model has driven

to the development of methods allowing to force the model towards a nonbiased climatology. Addressing systematic model errors, such as oceanographic biases, is even more tricky, since a representation of the bias itself, or the generation mechanism, is needed. The bias in the background field can be directly modelled by assuming some kind of persistence (Dee and Da Silva, 1998; Chepurin et al., 2005). Background errors (defined as the nonzero mean residuals) being observable, it is relatively straightforward to formulate a consistent bias-estimation scheme. Suppressing the bias generation during the integration of the model would even be preferable.

For example, in Derber and Rosati (1989), a variational continuous assimilation technique is applied. In the same way as nudging for data assimilation, it is a modification of the adjoint techniques, where a correction term is added to the equations, in order to correct the bias. It aimed at optimally fitting the data throughout the assimilation period, rather than relaxing the solution towards the values at observation times. It has been applied to radiative transfer model in Derber and Wu (1998).

Another example is in Radakovich et al. (2004), the model is so heavily affected by bias that a classic bias aware assimilation scheme (Dee and Da Silva, 1998) is not sufficient enough. The bias correction term is only applied during the assimilation scheme, but due to the model characteristics, it quickly slips back to its biased state and dissipates the correction term. In that study, an adapted incremental bias correction term was applied, during the model run, proportional to the initial state and the time separating two analysis steps (Radakovich et al., 2004). In some cases, the bias is handled through a post integration bias correction (Stockdale, 1997).

Recently, Vossepoel et al. (2004) evaluated the possibility of reconstructing wind stress forcing fields with both a random and constant error part, with a 4D-Var assimilation scheme on a twin experiment. Leeuwenburgh (2008) performed the estimation of surface wind-stress through an ensemble Kalman filter and corrected the boundary conditions of the model, effectively reducing the model bias. A very similar study to the work presented here is Ngodock et al. (2016), where an extra term is introduced in the tidal forcing, to correct errors in the tide model due to imperfect topography and damping terms.

1.4 Objectives

In this work, the problematic of model bias correction is tackled by developing a new method which combines stochastic forcing and data assimilation. While most previously developed and existing methods correct bias in the model results, the objective here is to come closer to the origin of the bias, and correct it by applying a stochastic forcing into the model equations. Data assimilation, and in particular the Ensemble Transform Kalman Filter (ETKF) is used, in a similar way to parameter estimation, to tune and find an optimal forcing term which is directly injected into the modified model equation. The aim is to provide a continuous bias correction by forcing the model.

The initial motivation to develop a new bias correction method arose in the context of the PredAntar project (Goosse et al., 2015), which consisted in the study of the Antarctic sea-ice coverage during the period 1980-2009 through the use of the coupled sea ice ocean NEMO-LIM2 model. Considering the long integration period of the project, compromises were made to respect the multiple limitations inherent to the project, such as a coarse model resolution. These caused the model to suffer from bias. Reanalysis throughout the model run provided adequate corrections, but highlighted the effects of bias, and the current bias correction methods limitations.

This novel approach is detailed in a general Kalman filter theoretical framework to prove its theoretical consistency. The successive steps are carefully detailed in the context of data assimilation, so that it can easily be transposed from the current oceanic application to any biased numerical model.

For the first application of the new bias correction method, the classic Lorenz '96 mathematical model (Lorenz, 1996; Lorenz and Emanuel, 1998) is chosen for its chaotic characteristics. Necessary modifications are applied to adjust the model to the specific needs of this study. Hence, the modified Lorenz '96 model characteristics are investigated to show particular connections to realistic ocean models. A classic twin experiment on a Lorenz '96 model is then implemented to test the efficiency and adaptability of the new bias correction method. Results are presented and studied in the context of a single assimilation procedure. Since applying a forcing in a nonlinear model exposes one to a nonlinear response of the model, an iterative assimilation is performed and compared to the single assimilation experiment.

The encouraging results of the Lorenz '96 twin experiment lead to the application of the bias correction method to the coupled sea ice ocean NEMO-LIM2 model. This model is presented in the context of its recent use in a research project, with a comparison with similar models from the CMIP5 framework, and bias affecting the model results is highlighted. The new bias correction method is implemented to the NEMO-LIM2 model, respecting particular constraints. Again, a twin experiment is used to test the stability of the forcing term generation and the efficiency of the bias correction. A single and monovariate assimilation is performed, showing model response to the forcing. A second multivariate assimilation shows the improvement obtained when more observations are used to estimate the bias correction term.

Finally, the method is confronted to a realistic case using real observations from the CNES-CLS09 global mean dynamic topography. A single assimilation experiment shows the effect of the different choices of the bias correction ensemble generation. The model response to the forcing term is interpreted in relation to the general circulation of the ocean. An iterative assimilation is also performed, and indicates the nonlinear model response to the forcing term.

Chapter 2

Basic concepts

Contents

In this chapter, the basic concepts required for the comprehension of the work presented in this thesis will be listed and detailed. Most of them are assumed to be known to the data assimilation community in numerical weather prediction and oceanography through general use and practice. With the intention to keep this work as clear as possible, the notation used will respect the unified notation for data assimilation (Ide et al., 1997) where relevant.

2.1 The model

In numerical modelling, the aim is to describe and reproduce a system through a set of discrete equations. This set of equations is called a numerical model. When applied to oceanography (or other geosciences for instance), models are partial, simplified and sometimes inadequate representations of the real world. It is clear that a model can never describe the whole complexity of the ocean. Choices, assumptions and hypotheses have to be made in order for the model to be viable in practice. Depending on the objective of the model, constraints are applied on the resolution, the

number and type of processes represented, etc. For instance, biological processes are not necessary for a physical ocean model of the long term Antarctic sea ice coverage. Still, numerical modelling is an essential and powerful tool for scientific inquiry. Controlled experiments (Lyard et al., 2006), the influence of the variation of a parameter (Baker et al., 2013), or "what if" scenarios (Dufresne et al., 2013), can help to understand which processes are important, and which assumptions are valid.

A numerical model forwards its model state in time using prognostic variables. Prognostic variables such as temperature, velocity or salinity, are necessary for the model calculations. They are regrouped into the state vector which uniquely describes the state of the system at a particular point in time. The model forwards prognostic variables from the model state at the previous time step. They can be compared to diagnostic variables, which help to interpret the model state, but can always be reconstructed from prognostic variables. An example of prognostic variable is the horizontal velocity, which depends on the horizontal velocity of the previous time step. However, the vertical velocity does not require the previous time step, but can be directly derived from the model state at the required time, making it a diagnostic variable. One must be able to distinguish between a complex reality, which can not be truthfully represented by a set of numbers, and the best way to represent reality as a state vector of a numerical model.

Formally, based on the state vector $\mathbf{x}_{m-1}$ at a time t_{m-1}, with the subscript $m = 1, ..., m_{max}$ being the time index, the model allows to compute the state vector $\mathbf{x}_m$ at the following time step t_m. With M being the forward model operator, one can write that

$$\mathbf{x}_m = M(\mathbf{x}_{m-1}). \tag{2.1.1}$$

The successive model states defined by equation (2.1.1) can be referred to as the model trajectory. The model trajectory describes the path of the different model variables, hence the model state, during the time period over which the model is run. The trajectory can be written as follow (van Leeuwen, 2001; Hunt et al., 2004)

$$\mathbf{x}' = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{m_{\max}} \end{bmatrix}. \tag{2.1.2}$$

2.2 State vector augmentation

A common procedure in numerical modelling is the state vector augmentation. Indeed, whereas the prognostic variables of the model are sufficient to fully describe the state of the model, it can be necessary to include other variables in the state vector, such as additional forcing (which will be extensively used in this thesis), scalar parameters, or diagnostic variables. In practice, one can algorithmically extend the state vector by appending the additional variables to the state vector. One can rewrite equation (2.1.2) by augmenting the state vector with one or multiple variables $\mathbf{e}$ as

$$\mathbf{x}' = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_{m_{\max}} \\ \mathbf{e} \end{bmatrix}. \tag{2.2.1}$$

State vector augmentation is commonly used in parameter estimation (Barth et al., 2010; Sakov et al., 2010), for the parameter to be estimated needs to be present in the description of the model through the state vector. A key advantage, in particular for Kalman filters, is that the incremental cost of an augmented-state vector is relatively small compared to the cost of the state vector alone (Kondrashov et al., 2008).

2.3 Parameter estimation

In numerical modelling, one needs to configure the numerical model through a series of well-defined parameters. Those parameters can contain errors, and contribute to some part of the model error. Parameter estimation inherently considers that the

parameters should be treated as variables of the model. One can include parameters in the state vector through state vector augmentation (equation (2.2.1)). Those parameters can be either fixed, or have a spatial and/or temporal evolution. For example, Bryan (1987) carries out a series of experiments on a low resolution, primitive equation ocean general circulation model to study the processes controlling important aspects of the circulation, and in particular the sensitivity of the model to the magnitude of vertical diffusivity. In Bergman and Hendon (2015), monthly radiative fluxes and heating rates are determined from monthly observations of cloud properties from the International Satellite Cloud Climatology Project and temperature and humidity from ECMWF analysis.

It is common for parameters to be strictly positive or constrained, such as the albedo of the ocean. Parameters are often also difficult, if not impossible, to measure directly (Losa et al., 2004). Optimally estimated parameter can attain nonphysical values due to either overfitting of data, or lack of identifiability with the available data. The complex and often nonlinear feedback between parameters is a particular issue if one wants to increase the number of parameters estimated at the same time, hence the dimension of the estimation problem (Navon, 1998). Parameters can also be updated locally, independently of their properties. In particular, if a global parameter is updated differently in a local assimilation scheme, one can use the different analysed values of said parameter to estimate the optimal global value.

One of the earliest applications of direct parameter estimation in oceanography using contemporary techniques was the estimation of the Cressman term of a barotropic model used for the parametrisation of the divergence associated with long waves (Rinne and Järvinen, 1993). Good parameter tuning is crucial in numerical modelling. In Zhu and Navon (1999) for instance, a complex global spectral model with its adjoint were used to tune both parameters and initial conditions. They concluded that even though the initial conditions dominated in the early stages of assimilation, the optimality of model parameters had a greater importance and persisted much longer than optimal initial conditions. Different approaches for parameter estimation have been tested and explored in the past literature.

For instance, adjoint methods borrowed from the optimal control theory can adjust model parameters through the use of available data. However, the model is supposed to be perfectly known and without error, which is never the case for practical applications in oceanography. One can choose to account for the errors of

the parameter estimation problem, but this greatly increases the dimension of the control space (Ten Brummelhuis et al., 1993). In some cases, extending the dynamical model equation (equation (2.1.1)) with the parameter to be estimated can cause the system to become nonlinear, even if the original system were linear (Kivman, 2003). Gong et al. (1998) is another example of adjoint method application, where a simple linear model was used as equivalent to a barotropic vorticity equation for the stream function on a latitude circle. It concluded that physical parameters to which the analysis is sensitive can be tuned along with one or two weighting parameters, and a smoothing parameter. Nevertheless, adjoint models are not always available, and demand considerable efforts to be developed.

Batch calibration techniques assume the time-invariance of the parameters and rely on statistical measures to minimise the long-term prediction error over some period of calibration and validation data (Kuczera, 1983; Vrugt et al., 2003). They require a set of historical data to be stored and processed, which can be a computational burden. Sequential data assimilation techniques have also been used for parameter estimation in oceanic systems. They present the advantage of overcoming this drawback and being able to explicitly take into account both the uncertainties on the model parameter, and the uncertainties of the model structure, its input, and its output (Moradkhani et al., 2005). The Kalman filter is a classic example of recursive data-processing algorithm, but it is limited to linear dynamic models with Gaussian error statistics. The extension to nonlinear systems with the EKF using first order linearisation can be used, but leads to instabilities when the nonlinearities are too strong (Miller et al., 1994).

Ensemble Kalman filters are also used for parameters estimation. In the EnKF framework, those parameters can be an unabridged part of the analysis and be updated along other model variables (Annan et al., 2005a). One of the first uses of the EnKF for parameter estimation occurs in Anderson (2001), where a demonstration is performed on a Lorenz '96 model by developing an ensemble adjustment Kalman filter. As suggested by Derber and Rosati (1989), the state vector is augmented with the parameters to be estimated. The analysis then contains both the updated conventional state variables, and the newly estimated parameters. In Aksoy et al. (2006), the performance of an EnKF is investigated through a simultaneous state and parameter estimation, where the source of the model error is contained in the uncertainty of the model parameters. The large scale applicability of the EnKF has also been highlighted by Annan et al. (2005a) in an earth model of intermediate

complexity. In Kondrashov et al. (2008), a coupled ocean-atmosphere system is investigated and shows that the simultaneous estimation of two erroneous parameters and the model state allows the improvement of the model state and of unobserved variables.

Kivman (2003) highlights a severe drawback of any Kalman filtering scheme: due to utilizing only first two statistical moments in the analysis step, it is unable to deal with probability density functions that are badly approximated by the normal distribution. In that study, an extension of the sequential importance resampling filter is proposed in order to deal with strongly non-Gaussian distributions. It also highlighted the benefits of specifically developed nonlinear methods like particle filters for non-Gaussian framework. Particle filters can be seen as variance minimizing schemes for any probability function (Simon and Bertino, 2012), which allows them to much better handle nonlinear parameters estimation. However, one must be careful when using them for large scale systems, as the size of the ensemble required is too large for realistic applications.

For large scale applications, the EnKF remains a practical and high-performance choice for parameter estimation. It scales much better than variational methods to large models. Additionally, it does not require the use of a linear tangent and adjoint model, making it straightforward to implement. Bertino et al. (2003) suggested an extended framework in which a nonlinear change of variables is applied in order to solve the obstacles posed by the non-Gaussian distribution of the variables. This procedure is called anamorphosis and allows the analysis step to be performed with transformed Gaussian distributed variables. Simon and Bertino (2009) demonstrated the feasibility of this technique in realistic configurations. In Simon and Bertino (2012), an improvement of the deterministic EnKF is proposed through such a Gaussian anamorphosis extension and solves in particular the inability of the EnKF to estimate negative parameters. This technique is also confirmed in Doron et al. (2013), where a combined state-parameter estimation in a twin experiment of a 3D ocean-coupled physical and biogeochemical model is successfully performed.

2.4 The observation operator

The model state at time m can be described by a vector $\mathbf{x}_m$. Each element of $\mathbf{x}_m$ is attributed a value on a grid, which can represent temperatures, winds compo-

nents, coefficients, etc. One can also make use of observations $\mathbf{y}_m$, which represent a specific measurement of a quantity at time m. The difference between the model state $\mathbf{x}_m$ and the observations $\mathbf{y}_m$ resides in the fact that there is not necessarily a one-to-one correspondence between the two quantities, and of course, only a small fraction of the state vector is observed (directly or indirectly). The observations are, in fact, rarely located exactly on the model grid points. Some interpolation might be necessary to obtain observation values on the model grid. Additionally, observations are not always exact measurements of any variable of the model. An example is the sea surface temperature, where most recent SST global fields are not direct measurements of the ocean with a thermometer, but rather reconstructed satellite measurements of the ocean radiation.

To relate the observations with the model, one can determine a model equivalency of the observations through the observation operator H (Lorenc, 1986). Hence, one can express that

$$\mathbf{y}_m = H_m(\mathbf{x}_m) + \epsilon_m, \tag{2.4.1}$$

where ϵ_m is the observational error, whose covariance matrix is $\mathbf{R}$, and which consists of instrumental and representativeness errors with respective covariance matrices $\mathbf{O}_i$ and $\mathbf{O}_r$. One thus has that $\mathbf{R} = \mathbf{O}_i + \mathbf{O}_r$. The instrumental error is rather straightforward, and represents the error coming from the instrument making the measurement of the observed quantity. On the other hand, the representativeness error is more complex. It represents unresolved processes by the model and is also called the unresolved scales error. It does not correspond to a problem with the observations, but is inherent to inadequacies in the dynamical model.

An example of the representativeness error for the sea surface temperature is the difference between the quantity measured by satellites, which is the skin temperature, and the model surface temperature. Whereas the skin temperature represents the thin layer at the surface of the observed fluid, whose thickness is less than 500μ m, the actual model surface temperature is much thicker, of the order of the meter (5 m for the NEMO-LIM2 model, section 6.1). The representativeness error is generally much larger than the instrumental error.

The observation operator can be nonlinear, and can contain an explicit time dependence in addition to the implicit dependence via the state vector (Ide et al.,

1997). It contains the approaches required to make the correspondence between an observed quantity and the model-equivalent variable, through interpolation, complex transformation of model variables, integration, etc. In equation (2.4.1), one can interpret the observation operator as a sequence of separated operators transforming the control variable $\mathbf{x}_m$ into the equivalent of each observation $\mathbf{y}_m$, at the location of the observations.

2.5 The observations

Observations are at the core of data assimilation. In this work, three different variables will be used as observations for the purpose of assimilation, interpretation of the model results, and validation. Those three variables are sea surface height (SSH), sea surface salinity (SSS) and sea surface temperature (SST). However, SSS will not be further detailed, as it is only used as control variable in a twin experiment, and not in the realistic scenario.

As presented with the observation operator, one must keep in mind that the use of observations is not straightforward. It is necessary to specify how those observations are initially defined, measured, represented on a grid and used for a particular objective.

2.5.1 Sea surface height

To define a height or a distance, one must first set a reference from which one can measure said distance. The difficulty to measure the SSH lies in the difficulty to determine those references.

To approximate the global shape of the earth (or other planetary bodies), geodesy defines a mathematical surface called the ellipsoid of reference. The level surface which corresponds to the surface of the ocean, when at rest, is called the geoid. It is close to the ellipsoid of reference which corresponds to the surface of a fluid under idealised homogeneous and rotating hypotheses, in a solid-body rotation and with no internal flow of the fluid (Stewart, 2008). However, the geoid differs from the ellipsoid due to local variations of the gravitational field. Those differences range up to 100 m (Lemoine et al., 1998). For example, seamounts are typically three times more dense than water, and increase the local gravity which causes a plumb line at the surface of the ocean. On the other hand, trenches in the oceanic floor tend

to create a deficiency of mass, thus a downward bulge. Finally, one must keep into account that the ocean is never at rest. Heat content of the water, tides, Rossby and Kelvin waves, eddies and ocean currents also affect the sea surface height, with ranges of up to 1 m. The deviation of the sea level from the geoid is defined as the dynamic topography. Formally, one defines the height of the geoid to the ellipsoid of reference as N, the sea level above the reference ellipsoid as η, and the sea level above the geoid as h. One has that: $h = \eta - N$ (Rio and Hernandez, 2004).

In practice, altimetry measurements are performed by radar on board of a satellite using high frequencies. The signal is reflected by the surface of the ocean. The time difference between the emission of the pulse and the reception of the echo allows the satellite to measure its distance to the sea surface. Even though the altitude of the satellite is determined by its orbit around the earth, which depends on the gravitational field of the earth, a precise knowledge of the geoid is not available. Hence, the SSH measurement provided by satellites are made with respect to the ellipsoid of reference. Parameters affecting those measurements, such as the speed of light in the atmosphere and interference from the ionosphere have to be taken into account. To provide a global coverage, the orbit of the satellite is chosen in order for the ground tracks to form a closed circuit after a predetermined number of cycles. Many altimetric satellites have flown in space to observe the ellipsoid of reference and measure the SSH. The most well-known space missions are Seasat (1978), geosat (1985–1988), ers–1 (1991–1996), ers–2 (1995–2011), Topex/Poseidon (1992–2006), Jason 1 (2002–2013), Envisat(2002), and Jason 2(2008-) (figure 2.1). In particular, Topex/Poseidon and Jason 1/2 were designed to provide a new level of precision with an accuracy of ± 0.05 m. In a close future (2021), a joint mission of US NASA and French CNES (Centre national d'études spatiales), called the Surface Water and Ocean Topography (SWOT), will be launched with a radar interferometer for making high-resolution measurements of the SSH. It will provide an increased spatial resolution to study ocean surface processes and circulation (Durand et al., 2010). Other ways to measure the SSH exist, in particular *in situ* measurements such as drifters or buoys velocities. The type of coverage required determines which measurement method is the most adapted.

Ocean models work with a grid which does not take the real geoid into account. Instead, they use a model geoid for which the sea level has a depth equal to zero when the ocean is at rest, in a constant gravitational field. The SSH of the model corresponds therefore to the dynamic topography. Hence, when one compares the

model SSH with the SSH obtained from altimetry measurements, one can actually only consider the anomalies to the zero level surface by subtracting the average SSH level from the data. In essence, the zero level surface is a measure of the total volume of the ocean. This way, the difficulty of defining the exact geoid is removed.

Ellipsoid, geoid and dynamic topography schematic

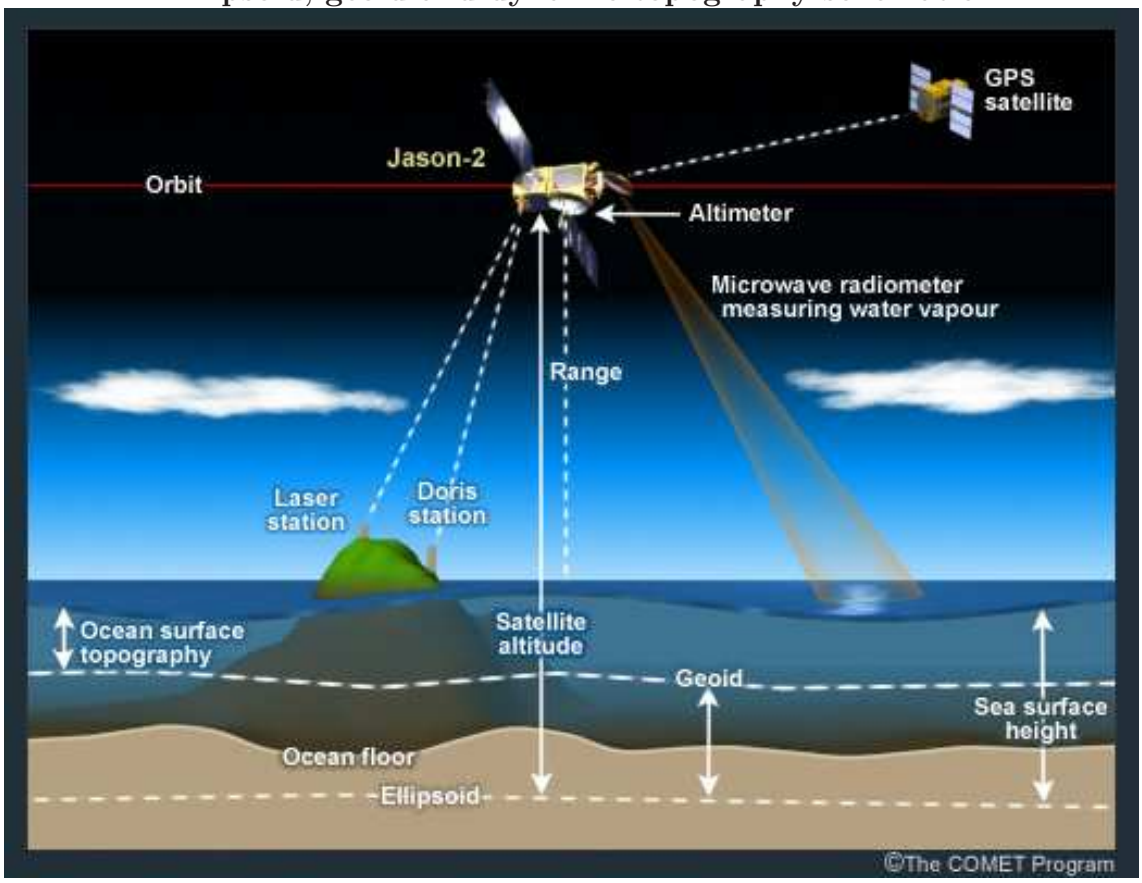


Figure 2.1: Schematic of the Jason-2 mission, with the ellipsoid of reference, the geoid, the dynamic topography (here ocean surface topography) and the sea surface height. Adapted from https://www.eumetsat.int/jason/print.htm

The data set used in this work comes from the CNES-CLS09 global mean dynamic topography (MDT) (Rio et al., 2011). It is a combination of GRACE data covering 4.5 years of measurements, altimetric measurements, and oceanographic *in situ* data. It uses an optimal filtering method to compute the large-scale MDT first guess. The altimetric data was computed at the CLS (Collecte, localisation, satellite), for the 1993-2009 period. The *in situ* measurements comprise drifting buoys velocities covering the 1993-2008 period, for which an Ekman model allows the extraction of the geostrophic velocity component. Wind stress data needed to

estimate the Ekman currents come from the ERA INTERIM reanalysis covering the 1989-2009 period (Simmons et al., 2007). Hydrological profiles, in particular temperature and salinity, were measured by Argo floats from 2002 to 2008. The final product is a global mean dynamic topography of the ocean with a resolution of 1/4° combining multiple different data sources, in particular satellites and *in situ* measurements.

2.5.2 Sea surface temperature

The importance of the ocean and its role in heat transport around the globe have been, in the last decades, the subject of major studies due to their relation to climate change (Wiens, 2016). The mechanisms with which exchanges take place between the atmosphere and the ocean are quite complex and include heat, momentum, moisture and gases. One can consider the SST as a global thermometer coupling the ocean and the atmosphere, which constrains the upper-ocean circulation and thermal structure. Similarly to the SSH, the most accurate SST products are provided by the combination of multiple sources of satellite data, *in situ* data and the underlying processes.

The water column extends from the surface to the ocean floor. Its vertical structure is both complex and variable. For global general circulation models and long term simulation, the vertical resolution is rather poor and the SST is considered as the temperature of the first layer of the ocean, with an order or magnitude of 10 m. One can easily realise the difficulty of measuring the SST by simply plunging one's arm into the sea, detecting the surface temperature gradient of the water. In practice, one can classify the vertical structure of the SST from the surface to the depth as follow (Donlon, 2002):

1. SST_{int}, the interface SST between the atmosphere and the ocean. It represents the infinitely thin layer where the ocean and the atmosphere are in contact, at the top of the SST_{skin} layer. It can not be measured using current technology.

2. SST_{skin}, the skin SST is a thin layer of $\sim 500\mu$ m of water, corresponding to the maximum penetration of infrared waves. It contains the waterside air-sea interface where the conductive and diffusive heat transfer processes dominate. Depending on the magnitude of the heat flux, a strong temperature gradient can be maintained in this thin layer. Radiometers are typically used to measure the SST_{skin}. However, since different wavelengths of the emitted radiation have

different penetration depths, the measured temperature varies depending on the measured wavelength (figure 2.2).

3. SST_{sub}, the subskin SST corresponds to the bottom on the SST_{skin}, at a depth of ~ 1 mm. The molecular and viscous heat transfers dominate. It is measured by low-frequency microwave radiometers and has a typical timescale variation of minutes.

4. SST_{bulk}, the bulk SST or subsurface SST. This is the region beneath the SST_{sub} where turbulent heat transfer processes dominate. It varies with depth, over timescale of hours and should also be noted with a reference to its depth: SST_{5m}. Buoys are used to perform *in situ* measurements of the SST_{bulk}.

The existence of the surface skin layer has been demonstrated both in theory (Hinzpeter, 1967) and practice (Schluessel et al., 1990). Its existence is required to regulate long wave radiation and turbulent heat fluxes across the sea surface. Indeed, turbulent eddy heat fluxes cannot transport heat across the ocean surface itself. The processes responsible for the heat transportation are molecular, hence the relatively thin size of the surface skin layer. Strong winds are able to destroy the skin layer through waves, but it is rapidly re-established when the waves dissipate.

The vertical temperature profile of the SST is shown on figure 2.3. One can easily note the difference between the day and night profiles, due to the presence or absence of the solar radiation.

2.6 Model skill

To assess the forecast skill of the model, one can compare the accuracy of the model trajectory and the degree of association to observations, expected or estimated values of the model, persistence forecast (values of the predictand in the previous time period), or another model on which improvement is expected. This forecast skill, or just skill, is used both qualitatively and quantitatively. It can relate to localised or overall forecast performance according to metrics. It is commonly represented in terms of correlation, root mean square error (RMSE), mean absolute errors, Brier score, or bias, among others (Landman, 2015). For the skill study to be statistically robust, the skill score calculations should be made over a large enough sample. The study of the model output through its skill is in fact the primary target of numerical modelling. The model skill under all its forms allows the interpretation of the model

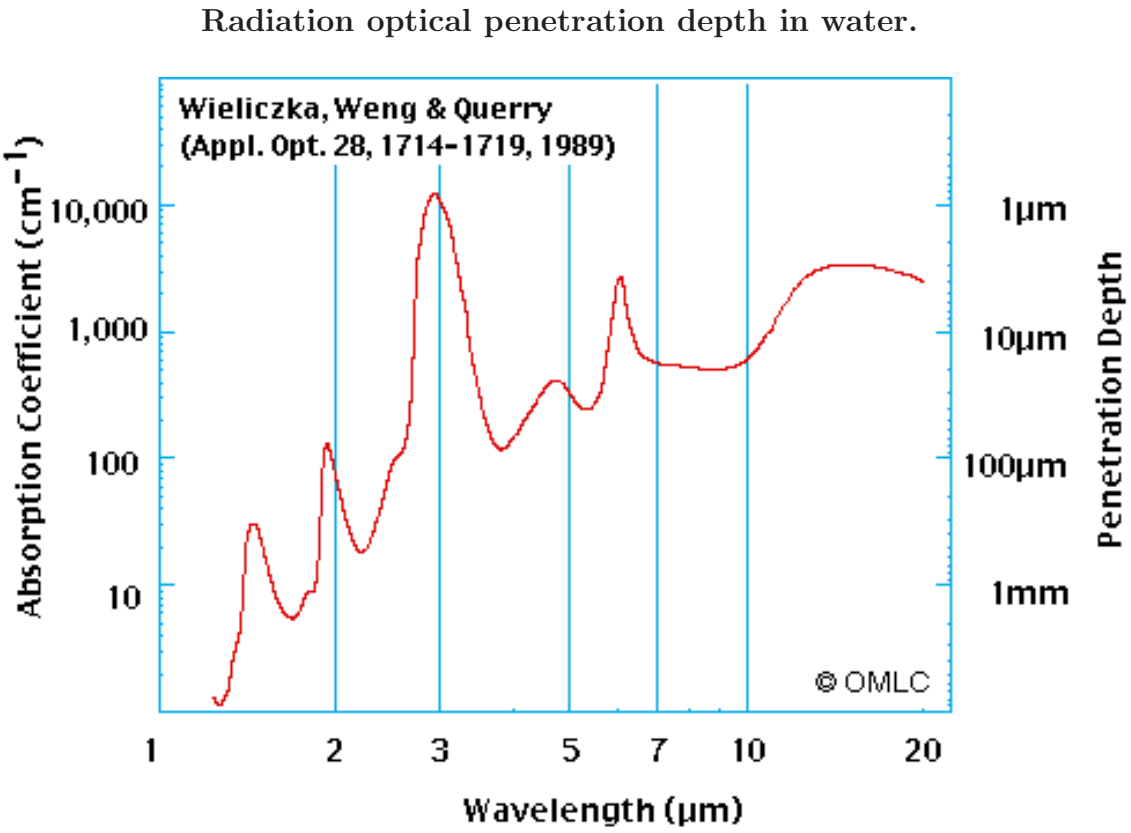


Figure 2.2: Radiation optical penetration depth in water. Adapted from Wieliczka et al. (1989).

results, and conclusions to be drawn.

Usually, forecast skill is presented as a percentage which is interpreted as a skill score and improvement over a reference, or a batch of observations. Formally, it is characterised by a measure of accuracy A with respect to a reference A_{ref}. With A_{perf} being the value of the accuracy measure achieved by a perfect forecast, one can represent the model skill as follow (Wilks, 2011)

$$SS_{ref} = \frac{A - A_{ref}}{A_{perf} - A_{ref}} \cdot 100\%. \tag{2.6.1}$$

If $A = A_{perf}$, the maximum value for the skill score SS_{ref} attains 100%. $A = A_{ref}$ indicates no changes compared to the reference accuracy, with a skill score of 0%. A skill score between 0 and 100% implies an improvement over the reference, while a negative skill $SS_{ref} < 0\%$ score denotes a deterioration. Equation (2.6.1) can eas-

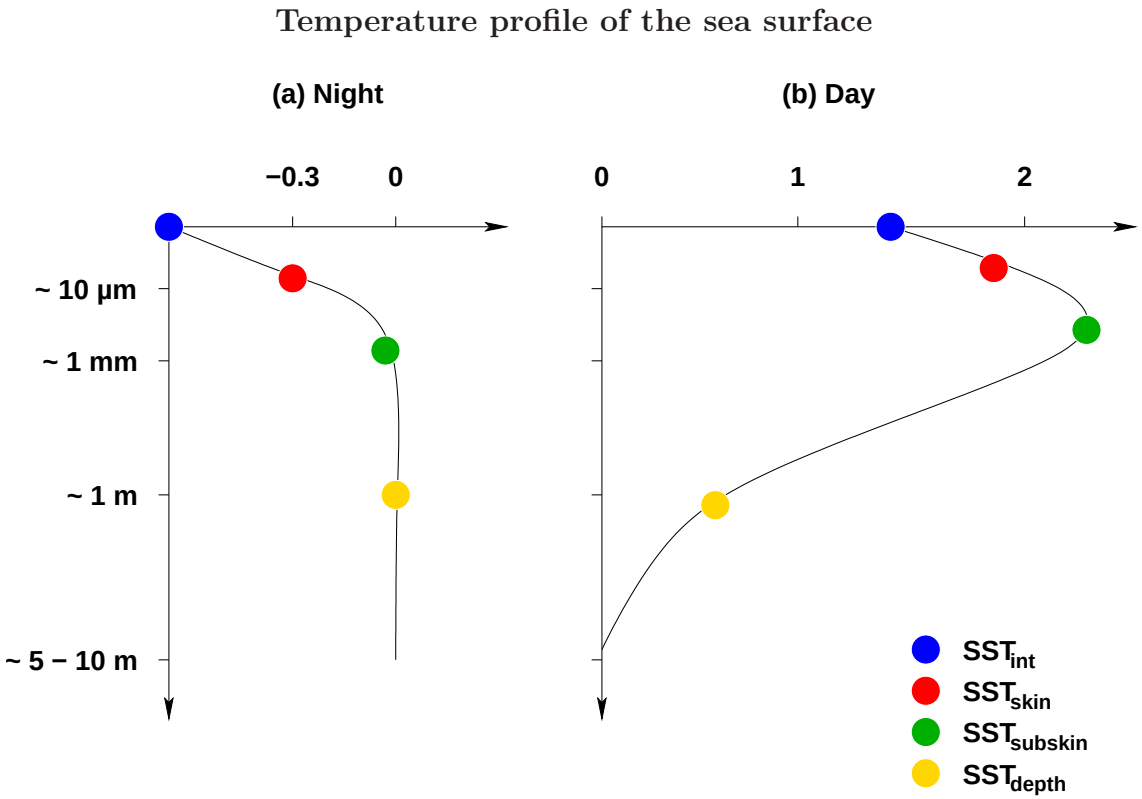


Figure 2.3: Left hand side: Day profile. Right hand side: Night profile. Adapted from Donlon (2002).

ily be constructed by using the root mean square error as the underlying accuracy statistics, and one obtains

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\mathbf{x}} - \mathbf{x}_i)^2}, \tag{2.6.2}$$

where $\hat{\mathbf{x}}$ is the estimator of the estimated variable $\mathbf{x}$ from which the RMSE is calculated.

In the next chapters, the skill score will not be explicitly used, but will be in some sense represented by the comparison of the RMSE values obtained from the different experiments.

Chapter 3

Theoretical framework

Contents

3.1 Kalman Filter

As presented in the introduction, numerical modelling and data assimilation relate to a long history of developments and advances following practical needs and constraints. One popular method is the Kalman Filter (Kalman, 1960). Literature related to Kalman Filtering is large, and dates back to its original expression, named after Rudolf Emil Kalman (who recently passed away, on the 2nd of July 2016). The first known application of the Kalman filter is to the nonlinear problem of space trajectory estimation for the Apollo program (Grewal and Andrews, 2010). The Kalman filter was incorporated to the Apollo navigation computer and used

during the mission. Indeed, estimating the position of an object in space was (and still is to some extent) a very difficult task. The combination of on-board position measurements through velocity and time, with the estimated and expected trajectory of the spacecraft with classic celestial mechanics, was the ideal first application for the Kalman filter.

For applications to oceanography, the Kalman filter was confronted with difficulties related to the size of the encountered systems or nonlinearities of the system. 4D variational techniques avoid some of those implementation problems while, in practice, providing satisfactory results. In the last decade of the 20th century, several variants of the Kalman filter were proposed, in particular the ensemble Kalman filter (Evensen, 1994; Houtekamer and Mitchell, 1998).

The following sections presenting the Kalman filter and the ensemble Kalman filter closely follow the development of C. Snyder (Blayo et al., 2014).

3.1.1 Bayesian formulation

Data assimilation can be formulated through a Bayesian perspective of the problem. One can rewrite equation (2.1.1) by adding the model error η_m, and formulate observations as

$$\mathbf{x}_m = M(\mathbf{x}_{m-1}) + \eta_m, \tag{3.1.1}$$

$$\mathbf{y}_m = H(\mathbf{x}_m) + \epsilon_m, \tag{3.1.2}$$

where $\mathbf{y}_m$ are the observations taken by the observation operator H, with observational error ϵ_m. No assumption is currently made about the errors, except that they are random variables. $\mathbf{x}_m$ and $\mathbf{y}_m$ are also supposed to be random variables. Their evolution in time contains the model and observation errors, and can thus not be expected to be perfectly known.

The classic way to represent random variables is by using a probability density function. Formally, it represents the relative likelihood of a random variable to have a given value. The probability density function expression for everything which is effectively known about the model state at time m is $p(\mathbf{x}_m|\mathbf{y}_0, \ldots, \mathbf{y}_m)$, where $(\mathbf{y}_0, \ldots, \mathbf{y}_m)$ is a sequence of observations.

Using the well known Bayes theorem (Bayes and Price, 1763), one can express the probability density function of $p(\mathbf{x}_m|\mathbf{y}_0,\ldots,\mathbf{y}_m)$ as

$$p(\mathbf{x}_m|\mathbf{y}_0,\ldots,\mathbf{y}_m) = \frac{p(\mathbf{x}_m|\mathbf{y}_0,\ldots,\mathbf{y}_{m-1})\ p(\mathbf{y}_m|\mathbf{x}_m,\mathbf{y}_0,\ldots,\mathbf{y}_{m-1})}{p(\mathbf{y}_m|\mathbf{y}_0,\ldots,\mathbf{y}_{m-1})}. \tag{3.1.3}$$

For the sake of readability, lets note $(\mathbf{y}_0,\ldots,\mathbf{y}_m) = (\mathbf{y}_{0,\ldots,m})$. Hence, one can update the probability density function $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m-1})$ with new observations $\mathbf{y}_m$ when they become available. To express the likelihood function, one starts from equation (3.1.2), assume that the observational error has no time correlation, thus that for $i \neq k$, $\epsilon_i \perp\!\!\!\perp \epsilon_k$ (where $\perp\!\!\!\perp$ stands for independent), and that it is also independent from the model error, hence $\epsilon_i \perp\!\!\!\perp \eta_j$ for all j. If the distribution of the observational error is known, for example a Gaussian distribution with zero mean: $\epsilon_k \sim \mathcal{N}(0,\mathbf{R})$, one can write from equation (3.1.2) that

$$p(\mathbf{y}_m|\mathbf{x}_m) \propto \exp\{[\mathbf{y}_m - h(\mathbf{x}_m)]^T\mathbf{R}^{-1}[\mathbf{y}_m - h(\mathbf{x}_m)]\}. \tag{3.1.4}$$

One needs also to know the probability density function of the model state at time m prior to the observations. This is given by the propagation equation

$$p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m-1}) = \int_{-\infty}^{\infty} p(\mathbf{x}_m|\mathbf{x}_{m-1},\mathbf{y}_{0,\ldots,m-1})\ p(\mathbf{x}_{m-1}|\mathbf{y}_{0,\ldots,m-1})\ d\mathbf{x}_{m-1}. \tag{3.1.5}$$

One can, using the probability density function of the model state at time $m-1$, propagate the observations available at time $m-1$ to the time step m. The link between the propagation rule and the model dynamics is implicit, for that equation (2.1.1) is equivalent to $p(\mathbf{x}_m|\mathbf{x}_{m-1},\mathbf{y}_{0,\ldots,m-1})$. If one assumes that the model error has no time correlation, thus that for $i \neq k$, $\eta_i \perp\!\!\!\perp \eta_k$, and that it is also independent from the observational error, hence $\eta_i \perp\!\!\!\perp \epsilon_j$ for all j, and one knows the distribution of the model error, for example a Gaussian distribution with zero mean: $\eta_k \sim \mathcal{N}(0,\mathbf{Q})$, one can write from equation (3.1.1) that $\mathbf{x}_m|\mathbf{x}_{m-1} \sim \mathcal{N}(f(\mathbf{x}_{m-1}),\mathbf{Q})$.

Those two relationships, the representation of the probability density function through the Bayes theorem (equation 3.1.3)) and propagation rule (equation (3.1.5)), allow the formulation of a sequential and recursive algorithm in which the model is propagated forward in time, from time $m-1$ to m. Then, the model state is updated

by using newly available observations at time m through the Bayes theorem. All is needed to start is an initial estimate of the model state, or initial conditions, and the knowledge about the model and observational errors η_m and ϵ_m.

3.1.2 Gaussian distribution

The original equations for the Kalman filter can be derived from the Bayesian framework (Blayo et al., 2014). Though, while in the Bayesian formulation, no hypotheses were made concerning the model and observational errors, one needs to assume that the system is Gaussian and linear. Since a mean and a covariance are sufficient to completely determine a Gaussian distribution, one can write the probability density function of a random Gaussian vector $\mathbf{x}$ as

$$p(\mathbf{x}) = (2\pi)^{-N_x/2}|\mathbf{P}|^{-1/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\overline{\mathbf{x}})^T\mathbf{P}^{-1}(\mathbf{x}-\overline{\mathbf{x}})\right], \tag{3.1.6}$$

where $\overline{\mathbf{x}} = E[\mathbf{x}]$ is the mean of the random Gaussian vector $\mathbf{x}$, E the expectation operator, and its covariance matrix $\mathbf{P} = E[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^T]$. Hence, one can write that $\mathbf{x} \sim \mathcal{N}(\overline{\mathbf{x}}, \mathbf{P})$. Like all covariance matrices, $\mathbf{P}$ is symmetric and positive definite.

Now, lets assume that the model dynamics (and operator), and the observations operator, are both linear. Lets also assume that the model and observational errors are Gaussian, with zero mean and respective covariance matrices $\mathbf{Q}$ and $\mathbf{R}$. This means, as for most assimilation schemes, that the model and observations are assumed to be unbiased. One can apply those hypotheses to equations (3.1.1) and (3.1.2), and one obtains

$$\mathbf{x}_m = \mathbf{M}_m\mathbf{x}_{m-1} + \eta_m, \tag{3.1.7}$$

$$\mathbf{y}_m = \mathbf{H}_m\mathbf{x}_m + \epsilon_m. \tag{3.1.8}$$

With the Gaussian, hence unbiased, hypothesis, the model and observation errors are completely defined by their respective covariance matrix

$$\mathbf{Q}_m = [\eta_m\eta_m^T], \tag{3.1.9}$$

$$\mathbf{R}_m = [\epsilon_m\epsilon_m^T]. \tag{3.1.10}$$

The covariance matrix is a keystone of the Kalman filter equations. During the update step of the sequential algorithm, they spread information from the observations to the unobserved state variables, if and when such a covariance is present.

3.1.3 Original Kalman Filter

A commonly adopted notation is to use the superscripts "a" and "f" for the analysis and forecast respectively. The forecast meaning that the variable is propagated through time to, say, the time step m, but without using the observations available at said time step. The analysis means that the variable has been updated through the assimilation of observations from time step m.

In addition, for readability reasons, the temporal subscript m will be dropped for the observation operator $\mathbf{H}_m$, the Kalman gain $\mathbf{K}_m$, the model operator $\mathbf{M}_m$ and the covariance matrix of the observational error $\mathbf{R}_m$. One must keep in mind that they retain a time dependence. For example, the observation operator depends on the location of the available observations at time m. If they are differently located than at time $m-1$, $\mathbf{H}_m$ will be different from $\mathbf{H}_{m-1}$.

One can write respectively the mean and variance of the forecast $\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1}$ and analysis $\mathbf{x}_m|\mathbf{y}_{0,\dots,m}$ as follow

$$\mathbf{x}_m^f \equiv E(\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1}), \tag{3.1.11}$$
$$\mathbf{P}_m^f \equiv cov(\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1}), \tag{3.1.12}$$
$$\mathbf{x}_m^a \equiv E(\mathbf{x}_m|\mathbf{y}_{0,\dots,m}), \tag{3.1.13}$$
$$\mathbf{P}_m^a \equiv cov(\mathbf{x}_m|\mathbf{y}_{0,\dots,m}). \tag{3.1.14}$$

Lets suppose that the model state at time m is known: $\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1} \sim \mathcal{N}(\mathbf{x}_m^f, \mathbf{P}_m^f)$ and that the initial conditions or initial state of the model also follows a Gaussian distribution $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_0^f, \mathbf{P}_0^f)$. Using new observations $\mathbf{y}_m$ for the analysis, with c_i being constants independent from $\mathbf{x}_m$, one can rewrite the right hand side of equation (3.1.3) by using equation (3.1.6). $p(\mathbf{y}_m|\mathbf{x}_m, \mathbf{y}_{0,\dots,m-1})$ and $p(\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1})$ become respectively

$$p(\mathbf{y}_m|\mathbf{x}_m, \mathbf{y}_{0,\ldots,m-1}) = c_1 \exp\,[-\frac{1}{2}(\mathbf{y}_m - \mathbf{H}\mathbf{x}_m)^T\mathbf{R}^{-1}(\mathbf{y}_m - \mathbf{H}\mathbf{x}_m)], \tag{3.1.15}$$

$$p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m-1}) = c_2 \exp\,[-\frac{1}{2}(\mathbf{x}_m - \mathbf{x}_m^f)^T(\mathbf{P}_m^f)^{-1}(\mathbf{x}_m - \mathbf{x}_m^f)]. \tag{3.1.16}$$

One can combine equation (3.1.15) and (3.1.16) to obtain the expression for the left hand side term of equation (3.1.3)

$$\begin{aligned} p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m}) &= c_3 \exp[-\frac{1}{2}(\mathbf{y}_m - \mathbf{H}\mathbf{x}_m)^T\mathbf{R}^{-1}(\mathbf{y}_m - \mathbf{H}\mathbf{x}_m) \\ &\quad - \frac{1}{2}(\mathbf{x}_m - \mathbf{x}_m^f)^T(\mathbf{P}_m^f)^{-1}(\mathbf{x}_m - \mathbf{x}_m^f)] \qquad (3.1.17) \\ &= c_3 \exp[-\frac{1}{2}J(\mathbf{x}_m)]. \qquad (3.1.18) \end{aligned}$$

One can explicitly rewrite $J(\mathbf{x}_m)$ and regroup terms together as

$$J(\mathbf{x}_m) = (\mathbf{y}_m - \mathbf{H}\mathbf{x}_m)^T\mathbf{R}^{-1}(\mathbf{y}_m - \mathbf{H}\mathbf{x}_m) - (\mathbf{x}_m - \mathbf{x}_m^f)^T(\mathbf{P}_m^f)^{-1}(\mathbf{x}_m - \mathbf{x}_m^f) \tag{3.1.19}$$

$$\begin{aligned} &= \mathbf{y}_m^T\mathbf{R}^{-1}\mathbf{y}_m - \mathbf{y}_m^T\mathbf{R}^{-1}\mathbf{H}\mathbf{x}_m - \mathbf{H}\mathbf{x}_m^T\mathbf{R}^{-1}\mathbf{y}_m + \mathbf{H}\mathbf{x}_m^T\mathbf{R}^{-1}\mathbf{H}\mathbf{x}_m \\ &\quad - \mathbf{x}_m(\mathbf{P}_m^f)^{-1}\mathbf{x}_m + \mathbf{x}_m(\mathbf{P}_m^f)^{-1}\mathbf{x}_m^f - (\mathbf{x}_m^f)^T(\mathbf{P}_m^f)^{-1}\mathbf{x}_m + (\mathbf{x}_m^f)^T(\mathbf{P}_m^f)^{-1}\mathbf{x}_m^f \end{aligned} \tag{3.1.20}$$

$$= \mathbf{x}_m^T\left[(\mathbf{P}_m^f)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right]\mathbf{x}_m - 2\left[\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}_m + (\mathbf{P}_m^f)^{-1}\mathbf{x}_m^f\right]^T\mathbf{x}_m + c_4, \tag{3.1.21}$$

where c_4 is independent from $\mathbf{x}_m$, containing, amongst other, the terms quadratic in $\mathbf{y}_m$ and $\mathbf{x}_m^f$. As long as the observation operator $\mathbf{H}$ is linear, $J(\mathbf{x}_m)$ is a quadratic form in $\mathbf{x}_m$. Hence, the probability density function $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m-1})$ is also a Gaussian density. One can see that the minimum of this function $J(\mathbf{x}_m)$ is reached when $\mathbf{x}_m = \mathbf{x}_m^a$ by rewriting equation (3.1.21) as

$$J(\mathbf{x}_m) = (\mathbf{x}_m - \mathbf{x}_m^a)^T(\mathbf{P}_m^a)^{-1}(\mathbf{x}_m - \mathbf{x}_m^a) + c_5, \tag{3.1.22}$$

where, in equation (3.1.21), one replaces

$$\mathbf{P}^a_m = \left[(\mathbf{P}^f_m)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right]^{-1} \tag{3.1.23}$$
$$= \mathbf{P}^f_m - \mathbf{P}^f_m\mathbf{H}^T(\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^f_m, \tag{3.1.24}$$
$$\mathbf{x}^a_m = \mathbf{P}^a_m\left[\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}_m + (\mathbf{P}^f_m)^{-1}\mathbf{x}^f_m\right]. \tag{3.1.25}$$

The Sherman-Morrison-Woodbury identity was used from equation (3.1.23) to (3.1.24) (see appendix section 11.1 for demonstration). Those results provide the update for the forecast mean $\mathbf{x}^f_m$ and covariance $\mathbf{P}^f_m$ with respect to the observations $\mathbf{y}_m$. Finally, using equation (3.1.18), one obtains the expression of the probability density function after assimilating the observations $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m})$

$$p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m}) = c_3\exp[-\frac{1}{2}(\mathbf{x}_m - \mathbf{x}^a_m)^T(\mathbf{P}^a_m)^{-1}(\mathbf{x}_m - \mathbf{x}^a_m)]. \tag{3.1.26}$$

The mean and covariance of this Gaussian probability density function are respectively $\mathbf{x}^a_m$ and $\mathbf{P}^a_m$.

An other interesting form of equation (3.1.25) allows to explicitly bring out the difference between the observations and the forecast, or "innovation", $\mathbf{y}_m - \mathbf{H}\mathbf{x}^f_m$. The matrix multiplying this difference is known as the "Kalman gain" $\mathbf{K}$, obtained from

$$\mathbf{x}^a_m = \mathbf{P}^a_m\left[\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}_m - \mathbf{H}\mathbf{x}^f_m + \mathbf{H}\mathbf{x}^f_m) + (\mathbf{P}^f_m)^{-1}\mathbf{x}^f_m\right] \tag{3.1.27}$$
$$= \mathbf{x}^f_m + \mathbf{P}^a_m\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{y}_m - \mathbf{H}\mathbf{x}^f_m) \tag{3.1.28}$$
$$= \mathbf{x}^f_m + \mathbf{K}(\mathbf{y}_m - \mathbf{H}\mathbf{x}^f_m). \tag{3.1.29}$$

The Kalman gain can also be expressed fully in terms of the forecast covariance as

$$\mathbf{K} = \mathbf{P}^a_m\mathbf{H}^T\mathbf{R}^{-1} \tag{3.1.30}$$
$$= \mathbf{P}^a_m\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{R})(\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{R})^{-1}$$
$$= \mathbf{P}^a_m\left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{H}^T\right)(\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{R})^{-1}$$
$$= \left[(\mathbf{P}^f_m)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right]^{-1}\left[\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + (\mathbf{P}^f_m)^{-1}\right]\mathbf{P}^f_m\mathbf{H}^T(\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{R})^{-1}$$
$$= \mathbf{P}^f_m\mathbf{H}^T(\mathbf{H}\mathbf{P}^f_m\mathbf{H}^T + \mathbf{R})^{-1}. \tag{3.1.31}$$

This allows to rewrite equation (3.1.24) as

$$\mathbf{P}_m^a = (\mathbf{I} - \mathbf{KH})\mathbf{P}_m^f. \tag{3.1.32}$$

The last step necessary to obtain the original Kalman filter equations is to propagate the probability density function after assimilation with the model. The propagation equation (equation (3.1.5)) requires both integrand factors $p(\mathbf{x}_m|\mathbf{x}_{m-1}, \mathbf{y}_{0,\dots,m-1})$ and $p(\mathbf{x}_{m-1}|\mathbf{y}_{0,\dots,m-1})$. One can suppose that $p(\mathbf{x}_{m-1}|\mathbf{y}_{0,\dots,m-1}) \sim \mathcal{N}(\mathbf{x}_{m-1}^a, \mathbf{P}_{m-1}^a)$. Knowing that the model and observation errors are assumed to be independent at each time step m, one can remove the dependence on the observations from the previous time step: $p(\mathbf{x}_m|\mathbf{x}_{m-1}, \mathbf{y}_{0,\dots,m-1}) = p(\mathbf{x}_m|\mathbf{x}_{m-1})$. One can finally imply from equation (3.1.7) that

$$\mathbf{x}_m|\mathbf{x}_{m-1} \sim \mathcal{N}(\mathbf{M}\mathbf{x}_{m-1}, \mathbf{Q}), \tag{3.1.33}$$

where $\mathbf{Q}$ is the model error covariance from equation (3.1.9). Consequently, the propagation equation (equation (3.1.5)) is the product of two Gaussian probability density functions, and the product of the two integrand becomes

$$\begin{aligned} p(\mathbf{x}_m|\mathbf{x}_{m-1})\ p(\mathbf{x}_{m-1}|\mathbf{y}_{0,\dots,m-1}) &= c\exp[-(\mathbf{x}_m - \mathbf{M}\mathbf{x}_{m-1})^T\mathbf{Q}^{-1}(\mathbf{x}_m - \mathbf{M}\mathbf{x}_{m-1})/2 \\ &\quad - (\mathbf{x}_m - \mathbf{x}_{m-1}^a)^T(\mathbf{P}_{m-1}^a)^{-1}(\mathbf{x}_m - \mathbf{x}_{m-1}^a)/2] \quad (3.1.34) \\ &= c_6\exp[-\frac{1}{2}J(\mathbf{x}_m, \mathbf{x}_{m-1})], \quad (3.1.35) \end{aligned}$$

where c_6 is a constant independent from $\mathbf{x}_m, \mathbf{x}_{m-1}$. With a classic addition - subtraction term with $\mathbf{M}\mathbf{x}_{m-1}^a$, one can rewrite $J(\mathbf{x}_m, \mathbf{x}_{m-1})$ as

$$\mathbf{x}_m - \mathbf{M}\mathbf{x}_{m-1} = (\mathbf{x}_m - \mathbf{M}\mathbf{x}_{m-1}^a) - \mathbf{M}(\mathbf{x}_m - \mathbf{x}_{m-1}^a), \tag{3.1.36}$$

$$J(\mathbf{x}_m, \mathbf{x}_{m-1}) = \begin{pmatrix} \mathbf{x}_{m-1} - \mathbf{x}_{m-1}^a \\ \mathbf{x}_m - \mathbf{M}\mathbf{x}_{m-1}^a \end{pmatrix}^T \mathbf{S} \begin{pmatrix} \mathbf{x}_{m-1} - \mathbf{x}_{m-1}^a \\ \mathbf{x}_m - \mathbf{M}\mathbf{x}_{m-1}^a \end{pmatrix}, \tag{3.1.37}$$

where

$$\mathbf{S} = \begin{pmatrix} (\mathbf{P}_{m-1}^a)^{-1} + \mathbf{M}^T\mathbf{Q}^{-1}\mathbf{M} & -\mathbf{M}^T\mathbf{Q}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{M} & \mathbf{Q}^{-1} \end{pmatrix}. \tag{3.1.38}$$

One can conclude that the joint distribution $(\mathbf{x}_m, \mathbf{x}_{m-1})|\mathbf{y}_{0,\dots,m-1}$ is Gaussian. Its mean is $(\mathbf{x}^a_{m-1}, \mathbf{M}\mathbf{x}^a_{m-1})$, and its covariance is $\mathbf{S}^{-1}$ and can be written as

$$\mathbf{S}^{-1} = \begin{pmatrix} \mathbf{P}^a_{m-1} & \mathbf{P}^a_{m-1}\mathbf{M}^T \\ \mathbf{M}\mathbf{P}^a_{m-1} & \mathbf{M}\mathbf{P}^a_{m-1}\mathbf{M}^T + \mathbf{Q} \end{pmatrix}, \tag{3.1.39}$$

where the demonstration for inversion $\mathbf{S}^{-1}$ of the block matrix $\mathbf{S}$ can be found in the appendix section 11.1.

The integrand of (equation (3.1.5)) should be integrated over $\mathbf{x}_{m-1}$ to obtain the distribution of $\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1}$. However, one can avoid this integration by noting that the mean and covariance corresponding to $\mathbf{x}_m$ are sufficient, and one has that

$$\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1} \sim \mathcal{N}(\mathbf{M}\mathbf{x}^a_{m-1}, \mathbf{M}\mathbf{P}^a_{m-1}\mathbf{M}^T + \mathbf{Q}). \tag{3.1.40}$$

The forecast step of the original Kalman filter equation are formally expressed by combining the Gaussian distribution of equation (3.1.40) with the definitions of the forecast mean and covariance from equations (3.1.11) and (3.1.12), as

$$\mathbf{x}^f_m = \mathbf{M}\mathbf{x}^a_{m-1}, \tag{3.1.41}$$

$$\mathbf{P}^f_m = \mathbf{M}\mathbf{P}^a_{m-1}\mathbf{M}^T_m + \mathbf{Q}. \tag{3.1.42}$$

To conclude, the Gaussian distribution hypothesis is at the core of the Bayesian framework derivation of the original Kalman filter. The forecast distribution $p(\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1})$ and the filtering distribution $p(\mathbf{x}_m|\mathbf{y}_{0,\dots,m})$ are both Gaussian, since the initial system (equation (3.1.9),(3.1.10)) is Gaussian. Hence, the Kalman filter is optimal for linear Gaussian systems.

The Kalman filter provides a set of recursive equations, allowing a sequential procedure for propagating (equation (3.1.41), (3.1.42)) and updating the system mean (equation (3.1.25)) and covariance (equation (3.1.23), (3.1.24)) through time using available observations.

3.1.4 Best unbiased linear estimator

Another approach to derive the Kalman filter is the best linear unbiased estimator (hereafter BLUE) framework. In this approach, there are no assumptions made

about the probability density functions of the initial state and the noise of the model. Instead, linear estimators are used, the objective being to find the one which minimises the expected mean squared error. The update step can be interpreted as a linear combination of the model state and the observations, with a weighting depending on the respective errors. Basically, estimators are a rule that relates a quantity of interest, here the model state, with its results, here the observations. In practice, both approaches provide the same results and are equivalent in the linear Gaussian case.

3.2 Extended Kalman filter

The Kalman filter equations are optimal and designed to work with linear models and Gaussian distributed random variables. The Gaussianity of the considered systems is at the source of the Bayesian framework derivation, since the Kalman filter provides the conditional mean $E(\mathbf{x}_m|\mathbf{y}_{0,\dots,m})$ and covariance $cov(\mathbf{x}_m|\mathbf{y}_{0,\dots,m})$ of the model state. The conditional probability density function $p(\mathbf{x}_m|\mathbf{y}_{0,\dots,m})$ is completely determined by its mean and covariance.

However, in practice, nonlinearities and non-Gaussianity often affect the model and the observations. Unknown model and observation errors can also affect the performance of the Kalman filter. The underlying error statistics is therefore also non-Gaussian, leading to a non-Gaussian conditional probability density function. Hence, the Kalman filter becomes suboptimal.

An important difficulty in implementing the Kalman filter is filter divergence (Miller et al., 1994). Essentially, if the model forecast contains unknown errors, the analysis error increases through the assimilation cycles and becomes larger than the filter estimation of the analysis spread. Too little weight is given to the observations at each assimilation step. The filter slowly diverges from the observations by trusting its own forecast over the observations. In the most extreme cases, the filter divergence leads to a blow-up of the solution, known as catastrophic filter divergence(Gottwald and Majda, 2013). A possibility to counter filter divergence is to add noise during the forecast step, to artificially increase the forecast variance.

In the Kalman filter update equations (3.1.25), the conditional mean $\mathbf{x}_m^a$ shows a linear dependence to the observations $\mathbf{y}_m$ and the prior mean $\mathbf{x}_m^f$. Moreover, the

updated covariance $\mathbf{P}^a_m$ in equation (3.1.23) is independent of the observations, for one only needs to know which observations are available to know $\mathbf{H}$ and $\mathbf{R}$. $\mathbf{P}^a_m$ can be computed before knowing $\mathbf{y}_m$, which is useful to target specific observations to improve a particular aspect of the analysis, or subsequent forecast. Using equation (3.1.24), one can also show that the analysis variance $tr(\mathbf{P}^a_m)$ is smaller than the forecast variance $tr(\mathbf{P}^f_m)$. Indeed, the second term on the right hand side is the symmetric product of symmetric and positive definite matrices. It is thus also symmetric and positive-definite, and has a positive trace. Hence, the trace of the error covariance matrix is reduced: $tr(\mathbf{P}^a_m) < tr(\mathbf{P}^f_m)$. In case that the state vector contains elements with different units such as temperature in degrees Celcius, and sea surface height in m, one can define a normalisation matrix $\mathbf{W}$ and still have that $tr(\mathbf{W}\mathbf{P}^a_m) < tr(\mathbf{W}\mathbf{P}^f_m)$.

For non-Gaussian distributions, one must make a distinction between the conditional mean, $E(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m})$, and the conditional mode, which is basically the $\mathbf{x}_m$ that maximises $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m})$. In the Gaussian case, those were equivalent. This is no longer true in the non-Gaussian case. A choice needs to be made between both (mean and mode) when one wants to find the optimal estimate $\hat{\mathbf{x}}_m$.

This choice is of utmost importance, since it provides an optimal $\hat{\mathbf{x}}_m$ with specific characteristics. For instance, in the conditional mean case, the expected square error $E(tr((\hat{\mathbf{x}}_m - \mathbf{x}_m|\mathbf{y}_{0,\ldots,m})(\hat{\mathbf{x}}_m - \mathbf{x}_m|\mathbf{y}_{0,\ldots,m})^T)$ is minimal. However, this does not automatically guarantees that the estimate $\hat{\mathbf{x}}$ will correspond to the most likely state. In case of a bimodal probability density function $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m})$, the conditional mean could lie between to peaks, and the estimate $\hat{\mathbf{x}}$ would be unlikely.

3.2.1 Nonlinear and non-Gaussian correction

One way to overcome those difficulties is to extend the Kalman filter to the nonlinear, non-Gaussian case (hence, the extended Kalman filter, or EKF). One can generalise equations (3.1.2) and (3.1.1) to the nonlinear case by considering nonlinear model and observation operators, respectively M and H. One can suppose that the model and observational errors η_m and ϵ_m, as well as $\mathbf{x}_{m-1}$ are approximately Gaussian, where $\mathbf{x}_{m-1} \sim \mathcal{N}(\mathbf{x}^a_{m-1}, \mathbf{P}^a_{m-1})$. One can suppose $\tilde{\mathbf{M}}_m$ to be the linearisation of M at $\mathbf{x}_{m-1}$, defined as follow

$$\tilde{\mathbf{M}}_m = (\frac{\partial M}{\partial \mathbf{x}})|_{\mathbf{x}^a_{m-1}}. \tag{3.2.1}$$

One can expand $\mathbf{x}_m = M(\mathbf{x}_{m-1})$ in a Taylor series using $\tilde{\mathbf{M}}_m$, and inject it back into equation (3.1.1) to obtain

$$M(\mathbf{x}_{m-1}) = M(\mathbf{x}^a_{m-1}) + \tilde{\mathbf{M}}_m(\mathbf{x}_{m-1} - \mathbf{x}^a_{m-1}) + O(|\mathbf{x}_{m-1} - \mathbf{x}^a_{m-1}|^2), \tag{3.2.2}$$

$$\mathbf{x}^m \approx M(\mathbf{x}^a_{m-1}) + \tilde{\mathbf{M}}_m(\mathbf{x}_{m-1} - \mathbf{x}^a_{m-1}) + \eta_m. \tag{3.2.3}$$

With the hypothesis that the probability density function of $\mathbf{x}_{m-1}$ can be approximated to a Gaussian distribution, one can then conclude that $\mathbf{x}_m \sim \mathcal{N}(\mathbf{x}^f_m, \mathbf{P}^f_m)$, where

$$\mathbf{x}^f_m = M(\mathbf{x}^a_{m-1}), \tag{3.2.4}$$

$$\mathbf{P}^f_m = \tilde{\mathbf{M}}_m \mathbf{P}^a_{m-1} \tilde{\mathbf{M}}^T_m + \mathbf{Q}. \tag{3.2.5}$$

The forecast step, equations (3.2.4) and (3.2.5), is thus similar to the linear Kalman filter of equations (3.1.41) and (3.1.42). However, in this linear approximation of the nonlinear case, the mean is propagated using the nonlinear model dynamics M, while the covariances are propagated using the linear approximation $\tilde{\mathbf{M}}_m$, which depends on the analysis at the previous time step $\mathbf{x}^a_{m-1}$.

For the linear approximation of M to be valid, one also needs to consider $O(|\mathbf{x}_{m-1} - \mathbf{x}^a_{m-1}|^2)$ in equation (3.2.2) to be negligible. This term contains and depends on $E(|\mathbf{x}_{m-1} - \mathbf{x}^a_{m-1}|^2) = tr(\mathbf{P}^a_{m-1})$ and on the second derivative of M. Hence, one needs to make the assumption that $\mathbf{P}^a_{m-1}$ must be small in the sense of its magnitude, when supposing that $\mathbf{x}_{m-1}$ is approximately Gaussian (for equation (3.2.1)).

One can apply the same procedure to the observation operator H. Its linearisation at $\mathbf{x}^f_m$, and the $\mathbf{y}_m$ approximation with the expansion with a Taylor series can be written as

$$\tilde{\mathbf{H}}_m = (\frac{\partial H}{\partial \mathbf{x}})|_{\mathbf{x}^f_m}, \tag{3.2.6}$$

$$H(\mathbf{x}_m) = H(\mathbf{x}^f_m) + \tilde{\mathbf{H}}_m(\mathbf{x}_m - \mathbf{x}^f_m) + O(|\mathbf{x}_m - \mathbf{x}^f_m|^2). \tag{3.2.7}$$

Using the linear approximation of $H(\mathbf{x}_m)$ into equation (3.1.2), one obtains

$$\mathbf{y}^m \approx H(\mathbf{x}^f_m) + \tilde{\mathbf{H}}_m(\mathbf{x}_m - \mathbf{x}^f_m) + \epsilon_m, \tag{3.2.8}$$

$$\mathbf{y}^m - H(\mathbf{x}^f_m) = \tilde{\mathbf{H}}_m(\mathbf{x}_m - \mathbf{x}^f_m) + \epsilon_m. \tag{3.2.9}$$

When applying the Kalman filter update from equation (3.1.29), with the observations $\mathbf{y}_m$, to $\mathbf{x'}_m = (\mathbf{x}_m - \mathbf{x}^f_m) \sim \mathcal{N}(0, \mathbf{P}^f_m)$, one obtains the linear approximation of the Kalman filter update and the Kalman gain with

$$\mathbf{x}^a_m = \mathbf{x}^f_m + \mathbf{K}(\mathbf{y}_m - H(\mathbf{x}^f_m)), \tag{3.2.10}$$

$$\mathbf{P}^a_m = (\mathbf{I} - \mathbf{K}\tilde{\mathbf{H}}_m)\mathbf{P}^f_m, \tag{3.2.11}$$

$$\mathbf{K} = \mathbf{P}^f_m\tilde{\mathbf{H}}^T_m(\tilde{\mathbf{H}}_m\mathbf{P}^f_m\tilde{\mathbf{H}}^T_m + \mathbf{R})^{-1}. \tag{3.2.12}$$

Similarly to the previous reasoning for M, the linear approximation of H respectively in equations (3.2.7) requires $O(|\mathbf{x}_{m-1} - \mathbf{x}^f_{m-1}|^2)$ to be neglected. This terms contains and depends on $E(|\mathbf{x}_m - \mathbf{x}^f_m|^2) = tr(\mathbf{P}^f_m)$, and on the second derivative of M. Using equation (3.2.5), H also depends on $\mathbf{P}^a_{m-1}$, as does M. Hence, this confirms the need for $\mathbf{P}^a_{m-1}$ to be sufficiently small for the Taylor series expansion of H and M to be accurate.

This set of three equations (3.2.10, 3.2.11, 3.2.12), can be compared to the original Kalman filter update step from equations (3.1.25), (3.1.29) and (3.1.32). It calculates the innovation from a nonlinear observation operator H applied to the forecast mean $\mathbf{x}^f_m$. Its linear approximation $\tilde{\mathbf{H}}_m$ is then used, in the same way as for M, to compute the gain and covariance update.

To summarise, the extended Kalman filter provides a generalisation of the original Kalman filter to nonlinear systems through a linear approximation. It is suboptimal in the nonlinear case, and rejoins the original Kalman filter in the linear case.

However, both the Bayesian and BLUE framework approaches to derive the origi-

nal and extended Kalman filter suffer greatly in high-dimensional systems, which are common in geophysics and, in particular, oceanography. The computational requirement for the Kalman filter prohibit direct calculations. For instance, the covariance matrices used in equation (3.1.42) for the propagation, and (3.1.32) for the update, are typically of the size N_x^2 and N_y^2, where it is common to have $N_x > N_y > 10^6$. Such matrices are impossible to store, yet be used in calculations.

In practice, the EKF assumes that everything is known. While this is a reasonable assumption about the model and observation operators M and H, the complete knowledge of the covariance matrices $\mathbf{P}^a$, $\mathbf{Q}$ and $\mathbf{R}$ is too optimistic, especially with the large size of those matrices. Finally, another critical assumption for the EKF is the linearisation of the model operator in equation (3.2.1). This linearisation is not straightforward and represents a major difficulty for the EKF implementation.

Other filters derived from the EKF, such as the singular evolutive extended Kalman filter (SEEK) (Pham et al., 1998) or singular evolutive interpolated extended Kalman filter (SEIK) (Pham, 1996) offer alternatives and solutions to those practical problems. In particular, the SEEK reduces the computational cost to an acceptable level by using a low rank approximation of the error covariance matrix. The evolution of the error covariance matrix is described by a reduced size basis of statistical functions which evolve in time.

3.3 Ensemble Kalman filter

The ensemble Kalman filter (EnKF) is another Kalman filter update scheme developed to handle specific issues of the Kalman filter. This algorithm is particular in the sense that the update step provides, either deterministically or stochastically, an ensemble of analyses, and that the forecast step uses this ensemble as initial conditions. When viewed from a Bayesian perspective, those random ensembles can be seen as a way to represent a probability density function (PDF) using a sample of said PDF. Algorithms that are used to generate and manipulate such random ensembles and samples are commonly referred to as Monte Carlo algorithms (Metropolis and Ulam, 1949; Doucet et al., 2001).

Formally, if $\mathbf{x}$ is a random variable with density $p(\mathbf{x})$, one can draw a series of $\mathbf{x}^{(i)}$ (where $i = 1, \ldots, N_e$), and call this series a sample of $\mathbf{x}$ as long as each member is drawn randomly and independently from each other. This random sample is then

equivalent to an ensemble.

In practice, ensemble members are randomly picked with a given distribution. Information about this given distribution is crucial to assure the representativity of the ensemble in systems where multiple modes are present. For example, suppose that a model climatology is strongly bipolar, producing two main modes. One could design an ensemble distribution representing those two main modes, instead of taking a Gaussian distribution.

The sample mean and covariance can then be written as

$$\hat{\mathbf{x}} = (N_e)^{-1} \sum_{i=1}^{N_e} \mathbf{x}^{(i)}, \tag{3.3.1}$$

$$\hat{\mathbf{P}} = (N_e - 1)^{-1} \sum_{i=1}^{N_e} (\mathbf{x}^{(i)} - \hat{\mathbf{x}})(\mathbf{x}^{(i)} - \hat{\mathbf{x}})^T, \tag{3.3.2}$$

where the "hat" symbol $\hat{}$ stands for a sample estimate. In this sense, $\hat{\mathbf{x}}$ and $\hat{\mathbf{P}}$ are both estimates of the mean $E(\mathbf{x})$ and covariance $\mathbf{P} = cov(\mathbf{x})$. The sample covariance $\hat{\mathbf{P}}$ can be written in terms of a square root. If $\mathbf{X}$ is the $N_x \times N_e$ matrix, which contains the ensemble perturbations in its columns, $(\mathbf{x}^{(i)} - \hat{\mathbf{x}})$, with a scaling factor $(N_e - 1)^{-1/2}$, then the sample estimator of the covariance from equation (3.3.2) becomes

$$\hat{\mathbf{P}} = \mathbf{X}\mathbf{X}^T. \tag{3.3.3}$$

As mentioned before, the ensemble Kalman filter aims at correcting specific issues of the Kalman filter. In particular, the full and continuous probability density function of the error covariance is first approximated in the original KF with a Gaussian distribution, through its mean and covariance. In the ensemble KF, this probability density function is represented through a sample of its distribution. In some sense, the ensemble Kalman filter is an approximation of the Kalman filter.

Beginning at time $m-1$ from a sample from the distribution $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m-1})$, and given at time m the observations $\mathbf{y}_m$, the sample can be updated to approximate a sample from the distribution $p(\mathbf{x}_m|\mathbf{y}_{0,\ldots,m})$. The aim of the next developments will be to find an algorithm producing such an update. With the produced updated

sample, one then wants to propagate said sample in time, to $m+1$, approximating the forecast distribution $p(\mathbf{x}_{m+1}|\mathbf{y}_{0,\dots,m})$. One must keep in mind that, using the Bayes theorem from equation (3.1.3), one can write that

$$p(\mathbf{x}_{m+1}, \mathbf{x}_m|\mathbf{y}_{0,\dots,m}) = p(\mathbf{x}_{m+1}|\mathbf{x}_m)p(\mathbf{x}_m|\mathbf{y}_{0,\dots,m}). \tag{3.3.4}$$

Hence, by starting from a sample $\mathbf{x}_m^{(i)}$ from the distribution $p(\mathbf{x}_m|\mathbf{y}_{0,\dots,m})$, one can draw, for each $\mathbf{x}_m^{(i)}$, a corresponding $\mathbf{x}_{m+1}^{(i)}$ from the distribution $p(\mathbf{x}_{m+1}|\mathbf{x}_m^{(i)})$, obtaining a sample $(\mathbf{x}_m^{(i)}, \mathbf{x}_{m+1}^{(i)})$ from the joint distribution $p(\mathbf{x}_{m+1}, \mathbf{x}_m|\mathbf{y}_{0,\dots,m})$. Discarding then every $\mathbf{x}_m^{(i)}$ from said sample produces in the end the sought sample $\mathbf{x}_{m+1}^{(i)}$ from the distribution $p(\mathbf{x}_{m+1}|\mathbf{y}_{0,\dots,m})$.

To draw from $p(\mathbf{x}_{m+1}|\mathbf{x}_m^{(i)})$ for a system as equation (3.1.1), one only needs the forecast model, and one can compute it with

$$\mathbf{x}_{m+1}^{(i)} = f(\mathbf{x}_m^{(i)}) + \eta_{m+1}^{(i)}, \tag{3.3.5}$$

where $\eta_{m+1}^{(i)}$ is drawn randomly from the assumed distribution of η. In fact, a forecast of $\mathbf{x}_m^{(i)}$ simply produces a random draw from $p(\mathbf{x}_{m+1}|\mathbf{x}_m^{(i)})$, resulting in the forecast step to be the propagation of the ensemble in time, thus an ensemble forecast. No approximations are required in this approach.

In the EnKF, no linearisation is required for the model forward operator, as opposed to the EKF. Hence, the EnKF is less affected by errors due to the nonlinearities. Indeed, when errors develop, the nonlinear model will fall into another regime which can contain said error. This cannot happen with a linearised model operator, such as in the EKF.

3.3.1 The stochastic Ensemble Kalman filter

To derive the stochastic ensemble Kalman filter, one can start from the description of the system and observations in equations (3.1.7) and (3.1.8). Suppose that, at time m, one know the distribution $\mathbf{x}_m|\mathbf{y}_{0,\dots,m-1} \sim \mathcal{N}(\mathbf{x}_m^f, \mathbf{P}_m^f)$, and one has the observations $\mathbf{y}_m$ available. Using the Kalman gain described in equation (3.1.31), one can consider the following random variable

$$\xi = \mathbf{x} + \mathbf{K}[\mathbf{y}^o - (\mathbf{H}\mathbf{x} + \epsilon)]. \tag{3.3.6}$$

Here, ξ is a linear function of Gaussian random variables, and is therefore a Gaussian random variable too. Its mean is then given by

$$E(\xi) = \mathbf{x}^a = \mathbf{x}^f + \mathbf{K}[\mathbf{y}^o - \mathbf{H}\mathbf{x}^f]. \tag{3.3.7}$$

Its covariance, with $\xi' = \xi - \overline{\xi}$ and $\mathbf{x}' = \mathbf{x} - \overline{\mathbf{x}}$, becomes

$$\begin{aligned} cov(\xi) &= E(\xi'\xi'^T) && (3.3.8)\\ &= E([\mathbf{x}' - \mathbf{K}(\mathbf{H}\mathbf{x}' + \epsilon)][\mathbf{x}' - \mathbf{K}(\mathbf{H}\mathbf{x}' + \epsilon)]^T) \\ &= \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^T\mathbf{K}^T - \mathbf{K}\mathbf{H}\mathbf{P}^f + \mathbf{K}(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})\mathbf{K}^T \\ &= \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^f \\ &= \mathbf{P}^a. && (3.3.9) \end{aligned}$$

Hence, the Gaussian random variable $\xi \sim \mathcal{N}(\mathbf{x}^a, \mathbf{P}^a)$ has its mean $\mathbf{x}^a$ and covariance $\mathbf{P}^a$ given by the Kalman filter update equations, and has the same distribution as $\mathbf{x}|\mathbf{y}_m$. To draw from the analysis distribution, one can start with $\mathbf{x}^{(i)}$ drawn from the forecast distribution $\mathcal{N}(\mathbf{x}^f, \mathbf{P}^f)$, and $\epsilon^{(i)}$ from $\mathcal{N}(0, \mathbf{R})$, and update each member of the sample with

$$\xi^{(i)} = \mathbf{x}^{(i)} + \mathbf{K}[\mathbf{y}^o - (\mathbf{H}\mathbf{x}^{(i)} + \epsilon^{(i)})]. \tag{3.3.10}$$

In equation (3.3.5), the system provides a way to build a sample from $\mathbf{x}_m|\mathbf{x}_{m-1}$ when starting with a sample from the distribution of $\mathbf{x}_{m-1}$. Similarly, equation (3.3.6) shows the relation between ξ and $\mathbf{x}$, and provides the basis for sampling from $\mathbf{x}|\mathbf{y}_m$ starting with a sample $\mathbf{x}^{(i)}$ from the distribution of $\mathbf{x}$.

The next step in the ensemble Kalman filter is to modify the update equations, which are required in the sampling, in order to avoid forming explicitly the covariances in the Kalman gain $\mathbf{K}$, and replace them by sample estimates from the forecast ensemble $\mathbf{x}^{(i)}$. With the sample mean definition from equation (3.3.1), one can define the columns $\mathbf{X}^{(i)}$ and $\mathbf{Y}^{(i)}$ of the matrices $\mathbf{X}$ and $\mathbf{Y}$ respectively as

$$\mathbf{X}^{(i)} = (N_e - 1)^{-1/2}(\mathbf{x}^{(i)} - \hat{\mathbf{x}}^f), \tag{3.3.11}$$

$$\mathbf{Y}^{(i)} = (N_e - 1)^{-1/2}(\mathbf{H}\mathbf{x}^{(i)} + \epsilon^{(i)} - \mathbf{H}\hat{\mathbf{x}}^f - \hat{\epsilon}), \tag{3.3.12}$$

where ϵ is the observation error, where

$$\hat{\epsilon} = (N_e)^{-1} \sum_{i=1}^{N_e} \epsilon^{(i)}, \tag{3.3.13}$$

and $\hat{\epsilon} \to 0$ for $i \to \infty$. Using the expression of the covariance from equation (3.3.3), one has that $\mathbf{X}\mathbf{Y}^T$ is the sample estimate for $\mathbf{P}^f\mathbf{H}^T$, and $\mathbf{Y}\mathbf{Y}^T$ is the sample estimate from $\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}$. One can then modify the Kalman gain from equation (3.1.31), and obtain the stochastic form of the ensemble Kalman filter

$$\hat{\mathbf{K}} = \mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}, \tag{3.3.14}$$

$$\xi^{(i)} = \mathbf{x}^{(i)} + \hat{\mathbf{K}}[\mathbf{y}^o - (\mathbf{H}\mathbf{x}^{(i)} + \epsilon^{(i)})], \tag{3.3.15}$$

$$= \mathbf{x}^{(i)} + \hat{\mathbf{K}}[\mathbf{y}^{(i)} - (\mathbf{H}\mathbf{x}^{(i)})]. \tag{3.3.16}$$

However, in equation (3.3.14), $\mathbf{Y}\mathbf{Y}^T$ is not always invertible. Indeed, to invert $\mathbf{Y}\mathbf{Y}^T$, one must decompose it in eigenvectors and eigenvalues so that

$$\mathbf{Y}\mathbf{Y}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T, \tag{3.3.17}$$

$$(\mathbf{Y}\mathbf{Y}^T)^{-1} = \mathbf{U}\mathbf{\Sigma}^{-1}\mathbf{U}^T. \tag{3.3.18}$$

Here $\mathbf{U}$ is the matrix whose ith column is the eigenvector $\mathbf{u}_i$, $\mathbf{U}\mathbf{U}^T = \mathbf{I}$, and $\mathbf{\Sigma}$ the diagonal matrix with the i eigenvalues. In oceanography, there are usually more observations available than ensemble members. This causes some eigenvalues to be zero and $\mathbf{Y}\mathbf{Y}^T$ not to be invertible, in which case one uses the pseudo-inverse of the matrix $\mathbf{Y}\mathbf{Y}^T$, written as $(\mathbf{Y}\mathbf{Y}^T)^+$. This way, one avoids the inverse of eigenvalues equal to zero. One can write that

$$(\mathbf{Y}\mathbf{Y}^T)^+ = \mathbf{U}\mathbf{\Sigma}^+\mathbf{U}^T. \tag{3.3.19}$$

It is common in literature to use perturbed observations $\mathbf{y}^{(i)} = \mathbf{y} + \epsilon^{(i)}$, by

replacing $\epsilon^{(i)}$ with $-\epsilon^{(i)}$ in equations (3.3.15) and (3.3.16). Those two equations are equivalent for the first and second moments of the analysis ensemble when the observation error ϵ distribution is symmetric, since the updated sample mean and covariances of the $\mathbf{x}^{(i)}$ ensemble remain the same in both cases.

Additionally, by adjusting the observation perturbations such as to have a zero ensemble mean, one can make sure that the ensemble mean is the same as for unperturbed observations, and reduce the sampling error.

Each member of the sample must undergo a separate analysis using the estimated Kalman gain $\hat{\mathbf{K}}$, with a realization of the observation error ϵ. This update step produces an ensemble of analyses which approximates a random sample from the distribution $p(\mathbf{x}_m|\mathbf{y}_m)$. The sample mean corresponds to the form of equation (3.1.25) of the original Kalman filter, where sample means replace the expected values, and sample covariances are used to approximate the gain as follow

$$\hat{\mathbf{x}}^a \equiv \hat{\xi} = \hat{\mathbf{x}}^f + \hat{\mathbf{K}}[\mathbf{y}^o - (\mathbf{H}\hat{\mathbf{x}}^f + \hat{\epsilon})]. \tag{3.3.20}$$

Again, if one defines the deviations from the sample mean update as the column matrix $\mathbf{X}^a$ as $(N_e - 1)^{-1/2}(\xi^{(i)} - \hat{\mathbf{x}}^a)$, and replaces covariance matrices with sample estimates from the ensemble, one can rewrite the sample covariance to have a similar expression as equation (3.1.24)

$$\mathbf{X}^a = \mathbf{X} - \hat{\mathbf{K}}\mathbf{Y}, \tag{3.3.21}$$

$$\begin{aligned}\hat{\mathbf{P}}^a &= \mathbf{X}^a(\mathbf{X}^a)^T \\ &= \mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{Y}^T\hat{\mathbf{K}}^T - \hat{\mathbf{K}}\mathbf{Y}\mathbf{X}^T + \hat{\mathbf{K}}\mathbf{Y}\mathbf{Y}^T\hat{\mathbf{K}}^T \\ &= \mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^+\mathbf{Y}\mathbf{X}^T. \end{aligned} \tag{3.3.22}$$

When the number of ensemble members tends to infinity, the sample estimates converges to the correct values. The mean and covariance of the ensemble Kalman filter will also converge to the Kalman filter as the ensemble size grows larger. The practical gain of the ensemble Kalman filter is, in the end, to reduce the computational requirements while having approximated values that tend to the expected values. The objective is attained for the covariance matrices, since $\mathbf{P}^f$, $\mathbf{P}^f\mathbf{H}^T$ and $\mathbf{H}\mathbf{P}^f\mathbf{H}^T$ are never explicitly stored or used in calculations. Instead, the EnKF algorithm uses the column matrices $\mathbf{X}$ and $\mathbf{Y}$, which are square roots of the co-

variance matrices. Their size are equivalent to the number of ensemble members N_e for the ensemble perturbations and the predicted observation vectors for those perturbations. However, the approximated Kalman gain $\hat{\mathbf{K}}$ however still uses the pseudo-inverse $(\mathbf{Y}\mathbf{Y}^T)^+$ in equation (3.3.22), which size is $N_e \times N_e$.

Finally, another interesting notion is the analysis increment $\mathbf{a}$. It is defined for each member, from equation (3.3.15), as

$$\begin{aligned} \xi^{(i)} - \mathbf{x}^{(i)} &= \hat{\mathbf{K}}[\mathbf{y}^o - (\mathbf{H}\mathbf{x}^{(i)} + \epsilon^{(i)})], && (3.3.23) \\ &= \mathbf{X}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^+[\mathbf{y}^o - (\mathbf{H}\mathbf{x}^{(i)} + \epsilon^{(i)})] \\ &= \mathbf{X}\mathbf{a}. && (3.3.24) \end{aligned}$$

$\mathbf{a}$ has a size of N_e, and represents the coefficients of the ensemble perturbations $\mathbf{x}^{(i)} - \hat{\mathbf{x}}^f$ when one considers $\xi^{(i)} - \mathbf{x}^{(i)}$ as a linear combination of those ensemble perturbations. This interpretation of the analysis increment in the context of the ensemble Kalman filter is crucial to understand how this algorithm diminishes the computational cost.

The major improvement of the EnKF compared to the EKF is the nonlinearisation of the model. In the EKF, this linearisation is necessary, but arises difficulties in particular circumstances. The EnKF directly uses the nonlinear model to propagate the ensemble states and represents therefore a significant improvement over the EKF. However, the main obstacle for the EnKF resides in the necessary objective to retain a small ensemble size, which can be a problem in cases of high dimensional systems. Indeed, a too small ensemble does not allow to fully reconstruct, via linear combination, the distribution $p(\mathbf{x}_m|\mathbf{y}_m)$.

3.3.2 Local assimilation

Techniques exist to overcome the issue of having a too small ensemble, such as local assimilation. Basically, with a small ensemble with N_e members, the global analysis will only allow adjustments and correction in the sub-space described by those N_e members. In the common case of chaotic systems, this ensemble is not sufficient to fully describe the background covariance matrix (Hunt et al., 2007). However, if one allows the analysis to perform locally different linear combinations of the ensemble members in different regions, one effectively allows the analysis to explore a much larger sub-space than previously. Two main form of localisation approaches

exist. They are classified into domain and covariance localisation (Janjić et al., 2011; Nerger et al., 2012a).

For domain localisation, one can reduce the global assimilation to localised regions, where the system is driven by the dynamics of the neighbouring regions. Subdomains (e.g., single grid point or vertical column) are created from the state vector, allowing independent assimilation to be performed. This is particularly interesting, as one can decompose the entire assimilation algorithm into smaller processes, for parallel computing (Nerger and Hiller, 2013). However, discontinuities can appear in the analysis field. One avoids them by combining domain localisation with observation localisation, by decreasing the weight of distant observations through the increase of the error variance $\mathbf{R}$ (Kalnay and Toth, 1994; Brankart et al., 2003).

In the case of covariance localisation, the algorithm sequentially assimilates every observations through a localisation function used as filter. Whereas the observation localisation acts on $\mathbf{R}$, the covariance localisation operates on the error covariance matrix $\mathbf{P}$. The sequential algorithm however prevents parallelisation, which can be a significant hinder for large models.

3.3.3 The deterministic Ensemble Kalman filter

The stochastic EnKF aims at generating an ensemble of members that approximate a particular PDF through a random sample of that same PDF. On the contrary, the deterministic EnKF aims at producing an analysis with a mean and covariance that correspond to the KF update for the observations, and with a mean and covariance from the forecast ensemble. If the system is linear and Gaussian, the deterministic EnKF also converges to the classic KF if the ensemble becomes large enough. The term "deterministic" is used in contrast to the random sampling of the stochastic EnKF.

Similarly to the previous section, suppose that a forecast ensemble $x^{(i)}, i = 1, \ldots, N_e$ is available. One can use the definition of the sample mean and covariance from equation (3.3.1,3.3.3) respectively. Then, the sample mean $\hat{\mathbf{x}} = \mathbf{x}^f$ can be updated with the Kalman filter by construction is given by

$$\hat{\mathbf{x}}^a = \hat{\mathbf{x}} + \hat{\mathbf{K}}(\mathbf{y}^o - \mathbf{H}\hat{\mathbf{x}}). \tag{3.3.25}$$

With $\mathbf{X}^a$ being the $N_x \times N_e$ matrix that contains the ensemble perturbations $(\mathbf{x}^{(i)} - \hat{\mathbf{x}})$ in its columns, with a scaling factor $(N_e - 1)^{-1/2}$, the sample covariance $\mathbf{P}^f = \mathbf{X}\mathbf{X}^T$ can be updated with

$$(N_e - 1)^{-1} \sum_{i=1}^{N_e} (\xi^{(i)} - \hat{\mathbf{x}}^a)(\xi^{(i)} - \hat{\mathbf{x}}^a)^T = \mathbf{X}^a(\mathbf{X}^a)^T, \tag{3.3.26}$$

where $\xi^{(i)}$ are the members of the analysis ensemble. One can rewrite the analysis covariance of the Kalman filter with the updated sample covariance, and obtains

$$\begin{aligned}
\mathbf{X}^a(\mathbf{X}^a)^T &= \mathbf{P}^f - \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{P}^f && (3.3.27)\\
&= \mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{X}^T\mathbf{H}^T(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{X}\mathbf{X}^T && (3.3.28)\\
&= \mathbf{X}[\mathbf{I} - \mathbf{X}^T\mathbf{H}^T(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{X}]\mathbf{X}^T && (3.3.29)\\
&= \mathbf{X}\mathbf{T}\mathbf{T}^T\mathbf{X}^T. && (3.3.30)
\end{aligned}$$

Equation (3.3.29) is satisfied if $\mathbf{X}^a = \mathbf{X}\mathbf{T}$, where

$$\begin{aligned}
\mathbf{T}\mathbf{T}^T &= \mathbf{I} - \mathbf{X}^T\mathbf{H}^T(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{X} && (3.3.31)\\
&= \mathbf{I} - \mathbf{S}^T(\mathbf{S}\mathbf{S}^T + \mathbf{R})^{-1}\mathbf{S} && (3.3.32)\\
&= \mathbf{I} - \mathbf{S}^T(\mathbf{F})^{-1}\mathbf{S}. && (3.3.33)
\end{aligned}$$

Here $\mathbf{S} = \mathbf{H}\mathbf{X}$ is the ensemble observation matrix and $\mathbf{F}$ is the innovation covariance. The columns of the matrix $\mathbf{T}$ are the coefficients of the linear combination of the forecast ensemble perturbations. One can see the resemblance with equation (3.3.24), where the $\mathbf{a}$ are the coefficients for the linear combination. $\mathbf{T}$ is called the transform matrix that transforms $\mathbf{X}$ into $\mathbf{X}^a$.

Different ways exist to compute it for large-dimensional systems, each corresponding to a particular derivation of the ensemble Kalman filter. Examples are the ensemble square root filter (EnSRF) (Tippett et al., 2003), the ensemble adjustment Kalman filter (EAKF) (Anderson, 2001), the singular evolutive interpolated Kalman filter (SEIK) (Pham, 1996), the error-subspace transform Kalman filter (ESTKF) (Nerger et al., 2012b) and the ensemble transform Kalman filter (ETKF) (Bishop et al., 2001).

3.3.4 The Ensemble Transform Kalman Filter

For the ETKF, one must start from equation (3.3.28) and apply the Sherman-Morrison-Woodbury formula to compute the second term of the right hand side (see appendix section 11.1 for demonstration) (Golub and Van Loan, 1996) to obtain that

$$\begin{aligned}
\mathbf{X}^a(\mathbf{X}^a)^T &= \mathbf{X}\mathbf{X}^T - \mathbf{X}\mathbf{X}^T\mathbf{H}^T(\mathbf{H}\mathbf{X}\mathbf{X}^T\mathbf{H}^T + \mathbf{R})^{-1}\mathbf{H}\mathbf{X}\mathbf{X}^T && (3.3.34)\\
&= \mathbf{X}\mathbf{X}^T - \mathbf{X}(\mathbf{H}\mathbf{X})^T[\mathbf{H}\mathbf{X}(\mathbf{H}\mathbf{X})^T + \mathbf{R}]^{-1}\mathbf{H}\mathbf{X}\mathbf{X}^T && (3.3.35)\\
&= \mathbf{X}\mathbf{X}^T - \mathbf{X}[\mathbf{I} + (\mathbf{H}\mathbf{X})^T\mathbf{R}^{-1}\mathbf{H}\mathbf{X}]^{-1}(\mathbf{H}\mathbf{X})^T\mathbf{R}^{-1}\mathbf{H}\mathbf{X}\mathbf{X}^T && (3.3.36)\\
&= \mathbf{X}[\mathbf{I} - (\mathbf{I} + \underbrace{(\mathbf{H}\mathbf{X})^T\mathbf{R}^{-1}\mathbf{H}\mathbf{X}}_{A})^{-1}(\mathbf{H}\mathbf{X})^T\mathbf{R}^{-1}\mathbf{H}\mathbf{X}]\mathbf{X}^T. && (3.3.37)
\end{aligned}$$

The aim is to find an expression for the analysis covariance $\mathbf{P}^a = \mathbf{X}^a(\mathbf{X}^a)^T$. The inverse of the observation error covariance matrix $\mathbf{R}^{-1}$ is supposed to be known. Then, the bracket A (present twice) in equation (3.3.37) can undergo an eigenvalue decomposition

$$(\mathbf{H}\mathbf{X})^T\mathbf{R}^{-1}\mathbf{H}\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \tag{3.3.38}$$

where $\mathbf{U}\mathbf{U}^T = \mathbf{I}$ and $\mathbf{\Lambda}$ is diagonal and both are $N_e \times N_e$ matrices. Equation (3.3.37) then becomes

$$\begin{aligned}
\mathbf{X}^a(\mathbf{X}^a)^T &= \mathbf{X}[\mathbf{I} - (\mathbf{I} + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T]\mathbf{X}^T && (3.3.39)\\
&= \mathbf{X}[\mathbf{I} - (\mathbf{I} + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T + \mathbf{I} - \mathbf{I})]\mathbf{X}^T\\
&= \mathbf{X}[\mathbf{I} - (\mathbf{I} + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T + \mathbf{I}) + (\mathbf{I} + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}]\mathbf{X}^T\\
&= \mathbf{X}[\mathbf{I} - \mathbf{I} + (\mathbf{I} + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}]\mathbf{X}^T\\
&= \mathbf{X}(\mathbf{I} + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}\mathbf{X}^T\\
&= \mathbf{X}(\mathbf{U}\mathbf{U}^T + \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)^{-1}\mathbf{X}^T\\
&= \mathbf{X}\mathbf{U}(\mathbf{I} + \mathbf{\Lambda})\mathbf{U}^T\mathbf{X}^T\\
&= \mathbf{X}\mathbf{U}(\mathbf{I} + \mathbf{\Lambda})^{-1/2}(\mathbf{I} + \mathbf{\Lambda})^{-1/2}\mathbf{U}^T\mathbf{X}^T\\
&= \mathbf{X}\mathbf{U}(\mathbf{I} + \mathbf{\Lambda})^{-1/2}\mathbf{U}^T\mathbf{U}(\mathbf{I} + \mathbf{\Lambda})^{-1/2}\mathbf{U}^T\mathbf{X}^T. && (3.3.40)
\end{aligned}$$

We obtain a square root decomposition of $\mathbf{P}^a$ in terms of $\mathbf{X}\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}\mathbf{U}^T$. The addition of the product with $\mathbf{U}^T$ allows the ensemble composed of the columns of

$\mathbf{X}^a = \mathbf{X}\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}\mathbf{U}^T$ to have a mean set to zero, without modifying $\mathbf{P}^a$ (Sakov and Oke, 2008).

The sum of all columns of $\mathbf{HX}$ is zero, if $\mathbf{H}$ is a linear observation operator, and

$$\mathbf{HX} = \mathbf{1}_{N_e\times 1}, \tag{3.3.41}$$

$$(\mathbf{HX})^T\mathbf{R}^{-1}\mathbf{HX}\mathbf{1}_{N_e\times 1} = 0\mathbf{1}_{N_e\times 1}. \tag{3.3.42}$$

Hence, $\mathbf{1}_{N_e\times 1}$ is an unnormalised eigenvector of $(\mathbf{HX})^T\mathbf{R}^{-1}\mathbf{HX}$ with eigenvalue 0. If one sorts all the eigenvalues in Λ, then $\mathbf{1}_{N_e}$ is the smallest and last of the N eigenvalue, since they all must be positive. With $\mathbf{e}_{N_e}$ being a zero vector with only the last element equal to one, one has

$$\mathbf{U}\mathbf{e}_{N_e} = \frac{1}{\sqrt{N}}\mathbf{1}_{N_e\times 1}, \tag{3.3.43}$$

$$\Lambda_{N,N} = 0, \tag{3.3.44}$$

$$\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-1/2}\mathbf{U}^T\mathbf{1}_{N_e\times 1} = \mathbf{1}_{N_e\times 1}. \tag{3.3.45}$$

Hence, the mean of all the columns of $\mathbf{X}^a$ is zero. One can also rewrite the Kalman gain $\mathbf{K}$ using the decomposition from equation (3.3.38), where $\mathbf{K}$ is comprised in the right-hand-side term of equation (3.3.36), as

$$\mathbf{K} = \mathbf{X}[\mathbf{I}+(\mathbf{HX})^T\mathbf{R}^{-1}\mathbf{HX}]^{-1}(\mathbf{HX})^T\mathbf{R}^{-1} \tag{3.3.46}$$

$$= \mathbf{X}[\mathbf{U}\mathbf{U}^T+\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T]^{-1}(\mathbf{HX})^T\mathbf{R}^{-1} \tag{3.3.47}$$

$$= \mathbf{X}\mathbf{U}(\mathbf{I}+\mathbf{\Lambda})^{-1}\mathbf{U}^T(\mathbf{HX})^T\mathbf{R}^{-1}. \tag{3.3.48}$$

Finally, the analysed ensemble mean is given by

$$\overline{\mathbf{x}}^a = \overline{\mathbf{x}}^f + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\overline{\mathbf{x}}^f). \tag{3.3.49}$$

One can reconstruct the ensemble from the columns of $\mathbf{X}^a$ and the ensemble mean $\overline{\mathbf{x}}^a$ to obtain

$$\mathbf{x}^{a(i)} = \overline{\mathbf{x}}^a + \sqrt{N_e-1}\mathbf{X}^{a(i)}. \tag{3.3.50}$$

3.4 Conclusion

In this chapter, the theoretical framework of the Kalman filter, from the original formulation to the ensemble transform Kalman filter, has been detailed. The assumptions on which each successive formulation relies, and the different hypotheses and constraints, have been presented. The different drawbacks and specific issues have been discussed, with the possible solutions when applicable. In the next part of this thesis, the ensemble transform Kalman filter is the data assimilation scheme which is used.

Chapter 4

Bias correction

Contents

4.1 Theoretical formulation

The aim is to develop a new method of bias correction for numerical modelling using an ensemble method. While most previously developed and existing methods correct bias in the model results, the objective here is to come closer to the origin of the bias and correct it by applying a stochastic forcing into the model equation.

4.1.1 Bias definition

Bias is a term which regroups numerous different definitions coming from different domains. In social sciences one can be confronted to confirmation bias, cultural bias, media bias or publication bias. Bias can also be defined in mathematics and engineering. However, the general idea of bias remains identical. It represents an inclination, predisposition or preference, towards a particular result, opinion, or tendency.

Bias as it is referred to in numerical modelling is more commonly known as statistical bias. Given a random variable corresponding to observed data x, with a probability distribution $p(x|\theta)$, where θ is a parameter indexing that probability distribution, one can define bias in the Bayesian framework for a statistic $\hat{\theta}$ which serves as an estimator for θ (Lehmann, 1951; Noorbaloochi and Meeden, 2000). This estimator is thus a function of the observed data: $\hat{\theta}(x)$. It is said to be biased when the expected value of the estimator is different from the real value of that parameter. In other words, with E being the expected value, one has that

$$E[\hat{\theta}(x)] - \theta = E[\hat{\theta}(x) - \theta] \neq 0. \tag{4.1.1}$$

Hence, one must keep in mind that the definition of bias depends of the notion of expected value. Indeed, the expectation value is to be determined over a period of time. In particular, in numerical modelling, this period of time has to be related in some sense to the time scale of the model, whether it represents a couple of hours, days, or even years.

4.1.2 Numerical model bias

Consider the following nonlinear stochastic discrete-time dynamical system, such as described by equation (2.1.1)

$$\mathbf{x}_m = M_m \left(\mathbf{x}_{m-1} \right), \tag{4.1.2}$$

where $m = 1, ..., m_{max}$ is the time index, $\mathbf{x}_m$ the n dimensional model state and M_m the forward model operator. One can assume that the model error is additive, as presented in Evensen (2007) and describe the real dynamical system by

$$\mathbf{x}_m^t = M_m^t \left(\mathbf{x}_{m-1}^t \right) + \boldsymbol{\beta}_m, \tag{4.1.3}$$

where $\mathbf{x}_m^t$ is the n dimensional true state, M_m^t the true model forward operator, and $\boldsymbol{\beta}_m$ the stochastic error. This model error can be split into two parts, namely a random part which average is zero: $< \widetilde{\boldsymbol{\beta}}_m > = 0$, and a systematic error, or bias, $\mathbf{b}$ (Dee, 2005), as

$$\boldsymbol{\beta}_m = \widetilde{\boldsymbol{\beta}}_m + \mathbf{b}. \tag{4.1.4}$$

One considers here the bias $\mathbf{b}$ to be constant in time. This assumption on the properties of the bias can be removed, to handle time-varying bias such as seasonal biases. Although finding an adequate correction would prove more difficult and computationally more costly, the principle of the method would remain identical.

4.1.3 Bias estimation and correction

The method for bias correction presented in this work relies on data assimilation. The principle is rather straightforward. One aims at estimating a stochastic forcing term $\widehat{\mathbf{b}}$ which will be used to modify the model forward operator M_m from equation (4.1.2) and correct the model bias. To provide said estimation, an ensemble transform Kalman filter is used (presented in section 3.3.4).

Using the definition of the model trajectory (equation (2.1.2)) and state vector augmentation (equation (2.2.1)), one can augment the state vector with an estimator of the bias correction term $\widehat{\mathbf{b}}^{(i)}$, where $\widehat{\mathbf{b}}^{(i)}$ can be seen as a parameter to be estimated (Barth et al., 2010; Sakov et al., 2010), by writing that

$$\mathbf{x}'^{(i)} = \begin{bmatrix} \mathbf{x}_1^{(i)} \\ \mathbf{x}_2^{(i)} \\ \vdots \\ \mathbf{x}_{m_{\max}}^{(i)} \\ \widehat{\mathbf{b}}^{(i)} \end{bmatrix}, \tag{4.1.5}$$

where $\mathbf{x}'^{(i)}$ is a member of the ensemble of trajectories with $i = 1, \ldots, N_e$ and N_e is the size of the ensemble. To each ensemble member $\mathbf{x}'^{(i)}$ corresponds a bias correction estimation $\widehat{\mathbf{b}}^{(i)}$. The average of the ensemble is then written as

$$\mathbf{x}' = \frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{x}'^{(i)}. \tag{4.1.6}$$

Hence, the ensemble after assimilation with the ensemble transform Kalman filter (equation (3.3.25)) is provided by

$$\mathbf{K}' = \mathbf{P}'^{f}\mathbf{H}'^{T}(\mathbf{H}'\mathbf{P}'^{f}\mathbf{H}'^{T} + \mathbf{R})^{-1}, \tag{4.1.7}$$

$$\mathbf{x}'^{a} = \mathbf{x}'^{f} + \mathbf{K}'\left(\mathbf{y}^{o} - \mathbf{H}'\mathbf{x}'^{f}\right), \tag{4.1.8}$$

where $\mathbf{y}^o$ are the time average of the observations. The observation operator $\mathbf{H}'$ applied on the trajectory $\mathbf{x}'$ also includes a time average and an extraction operator $\mathbf{H}$ of the observed part of the model state, with

$$\mathbf{H}'\mathbf{x}' = \frac{1}{m_{max}}\sum_{m=1}^{m_{max}}\mathbf{H}\mathbf{x}_m = \mathbf{H}\overline{\mathbf{x}}, \tag{4.1.9}$$

$$\overline{\mathbf{x}} = \frac{1}{m_{max}}\sum_{m=1}^{m_{max}}\mathbf{x}_m, \tag{4.1.10}$$

where $\overline{\mathbf{x}}$ is the time average of the model state vector. This method assumes that the bias represents a large part of the model error. Taking the time average of the model allows to concentrate on the bias, or the systematic error, with regards to the time scale of the model, and not on the random error contained in $\boldsymbol{\beta}_m$.

In addition, since one only requires the climatology of the model and the bias correction, the complete model trajectory is not needed. The average state of the model is sufficient, computationally much more interesting and justified by the definition of bias. One can use the state vector including only the model mean state and the bias correction term with

$$\mathbf{x}'' = \begin{bmatrix}\overline{\mathbf{x}} \\ \widehat{\mathbf{b}}\end{bmatrix}, \tag{4.1.11}$$

and an observation operator defined as

$$\mathbf{H}''\mathbf{x}'' = \mathbf{H}\overline{\mathbf{x}}. \tag{4.1.12}$$

One can show that the analysis using the average model state (equation (4.1.13)) provides the same analysed bias correction $\widehat{\mathbf{b}^a}$ as when the full trajectory is included in the estimation vector (equation (4.1.8)) with

$$\mathbf{x}''^{a} = \mathbf{x}''^{f} + \mathbf{K}'' \left(\mathbf{y}^{o} - \mathbf{H}''\mathbf{x}''^{f} \right). \tag{4.1.13}$$

The mathematical proof is shown in the appendix section 11.2. In practice, the assimilation of observations on the climatology of the model $\overline{\mathbf{x}}$ allows the update and tuning of the estimator $\widehat{\mathbf{b}^{a}}$ through the Kalman equations. The model is then rerun with the bias correction term, providing one with a bias corrected model run $\mathbf{x}^{r}_{m}$ as follow

$$\mathbf{x}^{r}_{m} = M_{m}\left(\mathbf{x}^{r}_{m-1}\right) - \widehat{\mathbf{b}}^{a}. \tag{4.1.14}$$

The interest of this method is that when the model is rerun, it provides a new model trajectory $\mathbf{x}^{r}_{m}$. This new trajectory, hence its average $\overline{\mathbf{x}^{r}}$ are different from the analysis $\mathbf{x}''^{a}$, since the model is nonlinear. Indeed, the former results from a new run by the model corrected equation (equation (4.1.14)), whereas the latter results directly from the analysis (equation (4.1.13)).

If the model were linear, the rerun would be equivalent to the assimilation result, and one would have $\overline{\mathbf{x}^{r}} = \mathbf{x}''^{a}$. In addition, the rerun allows to perform a first validation of the estimation of the bias correction term $\widehat{\mathbf{b}}^{a}$.

4.1.4 Discussion

It is common for bias correction methods to estimate the bias during the model run (whether online or offline) with a dynamic model for the bias. The bias estimation can then either be subtracted before the data assimilation, or be taken into account during the assimilation. However, it is not the case here. In this method, there is no dynamical model for the bias, and it is never directly estimated. Instead, the focus is turned towards the statistical estimation of a bias correction term, similarly to parameter estimation techniques (section 2.3).

The bias correction estimation performed by the analysis uses all available information, and not only the past information. Hence, one can consider this method as a smoother, a term commonly used in data assimilation literature to refer to methods propagating information both forward and backwards in time. However, here observations and forcing terms are chosen to be time averages. Thus, the notion of

propagating information in time is no longer relevant. This method provides one with a bias correction term $\widehat{\mathbf{b}^a}$ aimed at correcting the model. It can be used to run a corrected model, either in forecast or reanalysis mode.

In practice, the computational cost of this bias correction method is nearly identical to a classic ETKF scheme. One needs to construct an ensemble of runs, assimilate the observations, and only perform one additional model run with the bias correction. However, a major difference with the classic use of ensembles in data assimilation is that it is not necessary to keep the full trajectory of the analysis ensemble. The point of the assimilation is only to provide an updated bias correction forcing term, and not to provide an ensemble of analysed model states to rerun. As a consequence, the method is not affected by covariance inflation and filter divergence. However, one could consider the use of this method in a dual state-parameter estimation by making the adapted adjustments.

4.2 Practical formulation

A practical description of the different steps can greatly help to clarify the full procedure and its application in practice. The description is generic and valid for any numerical model. The different and successive steps are as follow:

1. Produce an initial model run $\mathbf{x}_m$.

2. Estimate the model bias $\mathbf{b}$ and its possible source through study of the model results. In particular, one can look at the model climatology to determine the bias.

3. Identify in the model equations this error source and determine how it can be modelled in statistical terms.

4. Modify the model equation to add an additional forcing term. This forcing needs to act on the bias source to counteract its effect, and dissipate the bias.

5. Create an ensemble of random parametrised bias correction estimates $\widehat{\mathbf{b}}^{(i)}$. The characteristics of those estimates can be specified. In particular, the shape, magnitude or correlation length can be estimated. The distribution of the ensemble can also be determined through assumptions on the bias source.

6. Run the ensemble of parametrised bias estimates to produce an ensemble of corresponding model trajectories $\mathbf{x}^{(i)}_{1,\ldots,m_{max}}$.

7. Extend the state vector of the model trajectories with the corresponding bias estimates. To diminish the computational cost, one can take the time average of the model $\mathbf{x}''^f$, instead of the full trajectories $\mathbf{x}'^f$, as the observations depend only on the time averaged model solution.

8. Assimilate adequate observations $\mathbf{y}^o$ in order to produce an analysed ensemble of state vectors $\mathbf{x}''^{a(i)}$.

9. Extract the ensemble mean of the analysed bias after assimilation $\widehat{\mathbf{b}}^a$.

10. Rerun the model with the bias correction term to obtain a bias corrected model run $\mathbf{x}_m^r$.

11. Validate the bias correction term with external and independent data, by comparing the uncorrected model run $\mathbf{x}_m$ with the corrected model run $\mathbf{x}_m^r$.

This procedure is summarised as a schematic on figure 4.1.

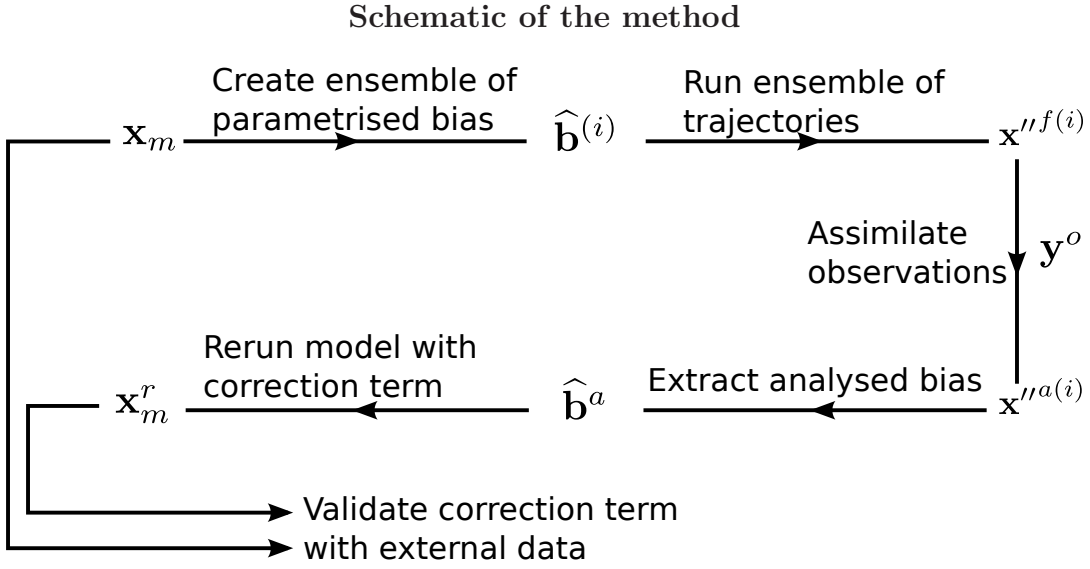


Figure 4.1: General schematic of the bias correction method, from the initial model run $\mathbf{x}_m$ to the corrected model run $\mathbf{x}_m^r$.

4.3 Experiment set up

In the next chapters of this work, the terminology used must be defined and linked to the notation introduced in this chapter. The reference run or true run corresponds to $\mathbf{x}^t$ (equation (4.1.3)), and the free run to the trajectory of $\mathbf{x}_m$ (equation (4.1.2)).

The ensemble run, when run with an ensemble of guessed estimators $\widehat{\mathbf{b}}^{(i)}$ before assimilation will be $\mathbf{x}''^f$, and after assimilation $\mathbf{x}''^a$ (equation (4.1.13)). Finally, the corrected run, or rerun will correspond to the trajectory of $\mathbf{x}_m^r$ (equation (4.1.14)) with the bias correction $\widehat{\mathbf{b}}^a$ provided by the analysis (equation (4.1.13)).

Chapter 5

Lorenz '96 Model

Contents

5.1 Model description

The first test for the new method developed in chapter 4.1 is its application on a fully controlled mathematical model. In 1963, Edward Lorenz developed a simplified mathematical model aimed at reproducing atmospheric convection. It is notable for having chaotic solutions for certain parameter values and initial conditions. Originally, it consists of a system of three differential equations (Lorenz, 1963). In 1996, it was updated in its 40-variables form, known as the Lorenz '96 model (Lorenz, 1996; Lorenz and Emanuel, 1998). It consists of a circular closed boundaries system with advection and diffusion properties.

This model has been widely used to test and improve data assimilation methods, ensemble filters or parameter estimation. In Li et al. (2009), a method for estimating the inflation factor and observational error simultaneously with an EnKF is presented, and first investigated on a low-order Lorenz 96' model. Terasaki and Miyoshi (2014) investigates the impact of observation error correlations and nonorthogonal observation operators on analysis accuracy using, again, the Lorenz 96' chaotic dynamical model. In Yang et al. (2012a), an outer loop is proposed for an EnKF to solve its weakness for Gaussian and linear models, and improve its ability to handle nonlinear dynamics. In particular, a 3-variable version of the Lorenz model is used.

Indeed, developing new methodologies relies on multiple specific procedures which need to be tested. This preparation work is better done beforehand on a small model which, even if it does not stand comparison with the complexity of realistic models, still enables to address the multiple issues one will be facing later on. Even if the Lorenz '96 model is not particularly complex, it still shows similarities with the ocean, in particular, the chaotic behaviour which makes forecasting a real issue.

In this work, the model is used in a different way than the one originally intended. Many of the previous works based on this model concentrate on the value of each variable during the model run. In particular for data assimilation benchmarks, the ability of the assimilation scheme to effectively catch and pull the model towards the observations is evaluated. The chaotic behaviour of the model renders this objective difficult to attain, since a small error in the variables inevitably grows over time. However, since the aim here is not to correct the specific value of the variables, but rather correct the bias that affects those variables, the focus is directed at the mean value of those variables over a period of time. This choice is motivated by the fact that, in some sense, bias is defined as a systematic error over a period of time. The next sections will cover more thoroughly the characteristics of the Lorenz '96 model and its responsiveness to some specific modifications.

5.2 Model characteristics

The system is described by the set of the $K = 40$ following equations

$$\frac{dX_k}{dt} = -X_{k-2}X_{k-1} + X_{k-1}X_{k+1} - X_k + F, \tag{5.2.1}$$

where $k = 1, \ldots, K$, and F is a constant variable independent from k. In this particular form, one can look at the periodic boundary conditions of the one-dimensional system as a circle around the Earth, in which the variables represent some atmospheric quantity in K different sectors of a latitude circle. Hence, $X_{k-K} = X_k = X_{k+K}$. External forcing and internal dissipation of the system, such as mechanical or thermal damping, are represented by the constant and linear terms, while the quadratic terms simulate advection. The coefficients before the quadratic and linear terms have been reduced and are equal to unity, through variable scaling (Lorenz, 1996). Together, they conserve the system total energy E (not strictly, but on the long term), where

$$E = \frac{1}{2}\sum_{k=1}^{K} X_k^2 \tag{5.2.2}$$

$$= \frac{K}{2}s^2. \tag{5.2.3}$$

If one notes the averages of the linear X_k and quadratic X_k^2 terms respectively as r and s^2, one has that total energy becomes $s^2/2$ (Lorenz, 2005). The evolution in time of the total energy is written as

$$\frac{d(s^2/2)}{dt} = \frac{d}{dt}\left(\frac{1}{2K}\sum_{k=1}^{K} X_k^2\right) \tag{5.2.4}$$

$$= \frac{1}{2K}\sum_{k=1}^{K} \frac{dX_k^2}{dt} \tag{5.2.5}$$

$$= \frac{1}{K}\sum_{k=1}^{K} X_k\frac{dX_k}{dt} \tag{5.2.6}$$

$$= \frac{1}{K}\sum_{k=1}^{K} X_k(-X_{k-2}X_{k-1} + X_{k-1}X_{k+1} - X_k + F) \tag{5.2.7}$$

$$= \frac{1}{K}\sum_{k=1}^{K} (-X_k^2 + X_kF - X_{k-2}X_{k-1}X_k + X_{k-1}X_kX_{k+1}) \tag{5.2.8}$$

$$= -s^2 + Fr. \tag{5.2.9}$$

The quadratic terms of equation (5.2.1) cancel out in equation (5.2.8) due to the periodic boundary conditions of the system, and do not affect the total energy balance of the system, similarly to advection in a realistic system. equation (5.2.9) also constrains the values of the system. One can show that the right-hand-side of this equation is negative for values of X_k that make $s > F$, for that $s \geq r$ using the Cauchy–Schwarz inequality (proof in the appendix section 11.3), causing $s^2 > Fr$. In other words, s^2 will decrease until $r < s < F$, hence s and r are always pulled back to values smaller than F.

One can also look at the long-term behaviour of this model. If R and S^2 are respectively the long term averages of r and s^2, in the sense that time derivatives become negligible, then one can write the variance of the model as

$$\sigma^2 = S^2 - R^2. \tag{5.2.10}$$

With the left-hand-side of equation (5.2.9) equal to zero for long-term averages, one has that $S^2 = FR$, hence $\sigma^2 = R(F - R)$. Variances are always positive, which causes the model average to be constrained by $0 < R < F$ (Lorenz, 2005). In particular, for very small values of F, the model solutions all decay to a steady solution, where $X_k = F = R = S$ and $\sigma = 0$. Lorenz and Emanuel (1998) already noted that if $F < 4$, the waves can extract energy fast enough to offset the effect of the external forcing. When $F > 4$, the model becomes completely chaotic over time and shows spatially irregular patterns. Even more, when $F > 15$, the model becomes totally unstable and diverges.

5.3 Model modification

For the intent of this work and as will be justified later, the Lorenz '96 model used here also needs to be tweaked. The method presented in chapter 4.1 needs to be confronted to a spatially variable bias. Hence, the forcing parameter F from equation (5.2.1) is modified to have a spatial structure depending on k as follow

$$\frac{dX_k}{dt} = -X_{k-2}X_{k-1} + X_{k-1}X_{k+1} - X_k + F_k. \tag{5.3.1}$$

5.3.1 Model average

Specific inquiries are necessary to obtain further information about the model behaviour after the modification presented in equation (5.3.1). As detailed in chapter 4.1, the bias correction method developed here relies on the assimilation of observations representing an average value, in order to highlight the potential bias hidden in the model state. Hence, one needs to look at the effect of the X_k parameter on the model average X_k values.

One notes that there is a significant relationship between the variables mean over time and the forcing parameter F_k. Parameters are set to $k = 1, .., 40$, and a time step of 0.05. 30 evenly distributed values are chosen for $0 < F_k < 10$. The model is then run with 450 different initial conditions for each F_k. The 200 first time steps are discarded, as spin-up time for the model to adjust itself to its parameters. The mean of the model variables is taken for the last 800 time steps and averaged over the 40 variables to obtain the model mean state.

Two cases are studied: in the first, the F_k are constant relatively to k for all the variables: $F_k = F$ (figure 5.1a). In the second, a random spatially correlated noise is added on the forcing parameter in order to obtain a different F_k for each k (figure 5.1b). That new forcing parameter is described by

$$\mathbf{F} = F\mathbf{1} + \mathbf{S_P}\mathbf{z}, \tag{5.3.2}$$

$$P_{i,j} = 0.3\, e^{\frac{-(i-j)^2}{15}}. \tag{5.3.3}$$

Here, $\mathbf{1}$ is a vector of size 40 with all values equal to one, $\mathbf{S_P}$ is the Cholesky decomposition of the covariance matrix $\mathbf{P}$ ($\mathbf{P} = \mathbf{S_P}\mathbf{S_P}^T$, $i, j = 1, \ldots, 40$ are rows and columns indices), and $\mathbf{z}$ is a random vector of 40 variables with a normal distribution $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$.

One can clearly see from figures 5.1a and 5.1b that there is a monotonic relationship between the system mean and the forcing parameter, whether the later one is or is not constant. This encourages the working hypothesis that even a fully nonlinear system in each of its variable can, under some conditions, be expected to show a global simple behaviour, as long as the system does not include a regime shift. This also confirms that even though the model state at a specific point in time depends on the initial conditions, the time average of the model over the last 800 time steps only has a minimal dependence on the initial conditions. This is important since the

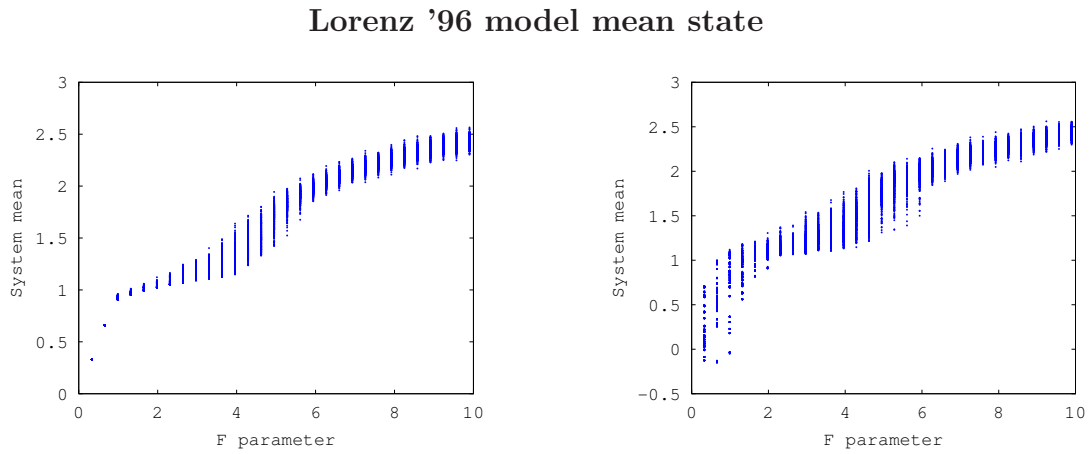


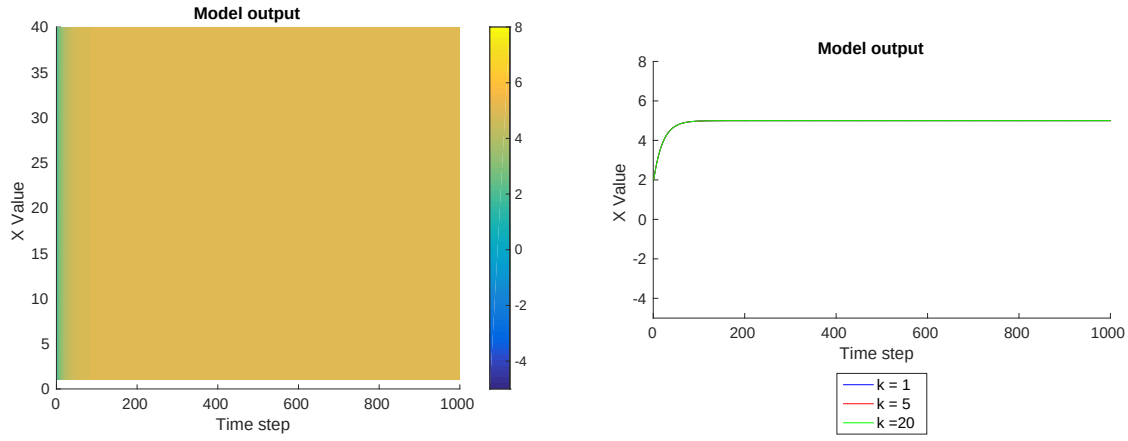
Figure 5.1: Lorenz '96 model mean state compared to: Panel (a) a constant forcing parameter F, Panel (b) a function of the average of the spatially variable forcing parameter F_k as defined by equation (5.3.1). The X-axis represents the 30 different $0 < \overline{F} < 10$ tested. For panel (b), only the mean part corresponding to $\overline{F}$ is plotted for more readability. The Y-axis represents the model mean state for the 450 initial conditions as a function of $\overline{F}$

aim is not to predict the exact value of the system at a given point in time. Instead, the aim is only to correct the model forcing parameter and the bias it causes on the model mean state.

5.3.2 Spatial average

Different set-ups are tested to show the influence of the different parameters on the behaviour of the model described by equation (5.2.1). The model is initially run for every set-up for 10000 time-step. Parameters are set to $k = 1, .., 40$, and a time step of 0.05, which corresponds to about 6 hours in the atmosphere (Lorenz and Emanuel, 1998). Results are shown on figure 5.2a to 5.5d. Panels (a) are a 2D plot of the value of each variable X_k for the first 1000 time-steps, where $k = 1, \ldots, 40$. Panels (b) are a time plot of the variables X_k, $k = 1, 5, 20$, to visualise their particular evolution through time for the first 1000 time-steps. Panels (c) are the values of F_k, in order to see their influence on the model mean. Panels (d) are the time average over the first 1000 time steps of X_k for one initial condition (in red), the time average over for 10000 time-steps for one initial condition (in blue), and for the first 1000 time-steps for 15 initial conditions (in green), to see how the average of each variable is influenced by the model forcing parameters F_k, the integration timescale of the model, and the number of initial conditions.

Figure 5.2a and (5.2b) show the model output for uniform initial conditions $X_k = 2, F_k = 5$, for all k. It is clear that the model rapidly tends to a steady solution where $X_k = F_k = 5$. Panels (c) and (d) are absent from this case, since the values are constant for all k. Figure 5.3a and 5.3b show the influence of uniform initial conditions on $F_k = 5$, for all k, with an initial condition $X_k = \mathcal{N}(2, 1)$. One can observe the particular advective pattern on figure 5.3a, where the signal is slowly propagated towards greater values of k, and the periodic boundary conditions of the system. Moreover, the system needs some time to adapt from the initial conditions towards a globally stable state, in the sense of the long-term average of the global energy, with $S^2 = FR$. This spin-up takes quickly place, and the X_k variables already attain their long-term variability after the first 100 time steps of the model. Figure 5.3d also shows that with random initial conditions and a chaotic behaviour on X_k, the variability of the time average of X_k is one order or magnitude smaller than the variability of X_k. In fact, those fluctuation are only statistical. For a sufficiently long integration period, the time average of X_k would tend to a constant. This can be seen by comparing the model average output of the first 1000 time-steps with the full 10000 time-steps. Moreover, the average over the 10000 time-steps provide the same lower variance as the average over the first 1000 time-steps and over 15 initial conditions.



(a) (b)

Figure 5.2: Lorenz '96 model evolution for 1000 time steps, for uniform $F_k = 5$, for all k parameter and uniform $X_k = 2$, for all k initial conditions. Panel (a) is a 2D plot of X_k over time and spatial index. Panel (b) is the temporal evolution of X_k for $k = 1, 5, 20$.

One can see on figures 5.4a to 5.5d that the spatially variable parameter F_k, as

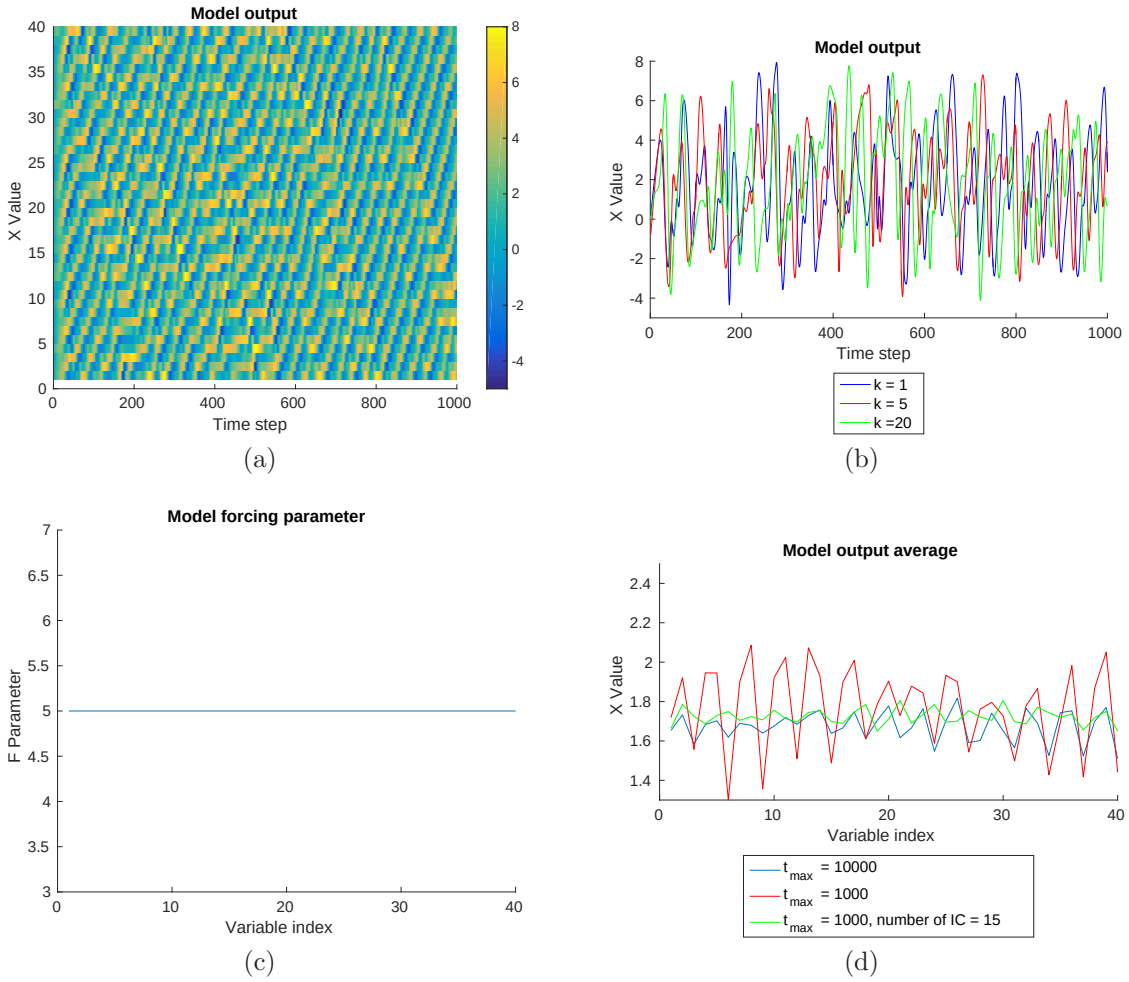


Figure 5.3: Lorenz '96 model evolution for 1000 time steps, for uniform $F_k = 5$, for all k parameter and random X_k initial conditions with average $\overline{\mathbf{X}} = 2$. Panel (a) is a 2D plot of X_k over time and spatial index. Panel (b) is the temporal evolution of X_k for $k = 1, 5, 20$. Panel (c) is the values of F_k. Panel (d) are different time averages of the model state X_k.

defined by equation (5.3.2), adds more variability and instability to the model. In particular, the time average of X_k (figures 5.4d and 5.5d), for the same period of time, show a larger variance, than when F_k is constant. The pseudo-physical processes, such as advection and diffusion, are still present. Figures 5.4a and 5.4b show that with constant initial conditions on X_k, variability is introduced by the forcing parameter F_k. The spin-up time necessary for perturbations to be introduced to the system is less than 100 time steps, but is longer than the constant F_k and variable X_k case.

Finally, one can clearly observe a correlation between figures 5.5c and 5.5d.

The constraints resulting from equation (5.2.10) are also valid on a local scale of the model. In particular, the model average is locally constrained by $0 < R < F$. A higher average value on X_k is obtained in the regions of high F_k values, when $k = 15, \ldots, 30$, and this value drops when F_k becomes smaller, at $k = 1, \ldots, 15, 30, \ldots, 40$. One can this expect to be able to recover the local values of F_k is one possesses long term observations on X_k.

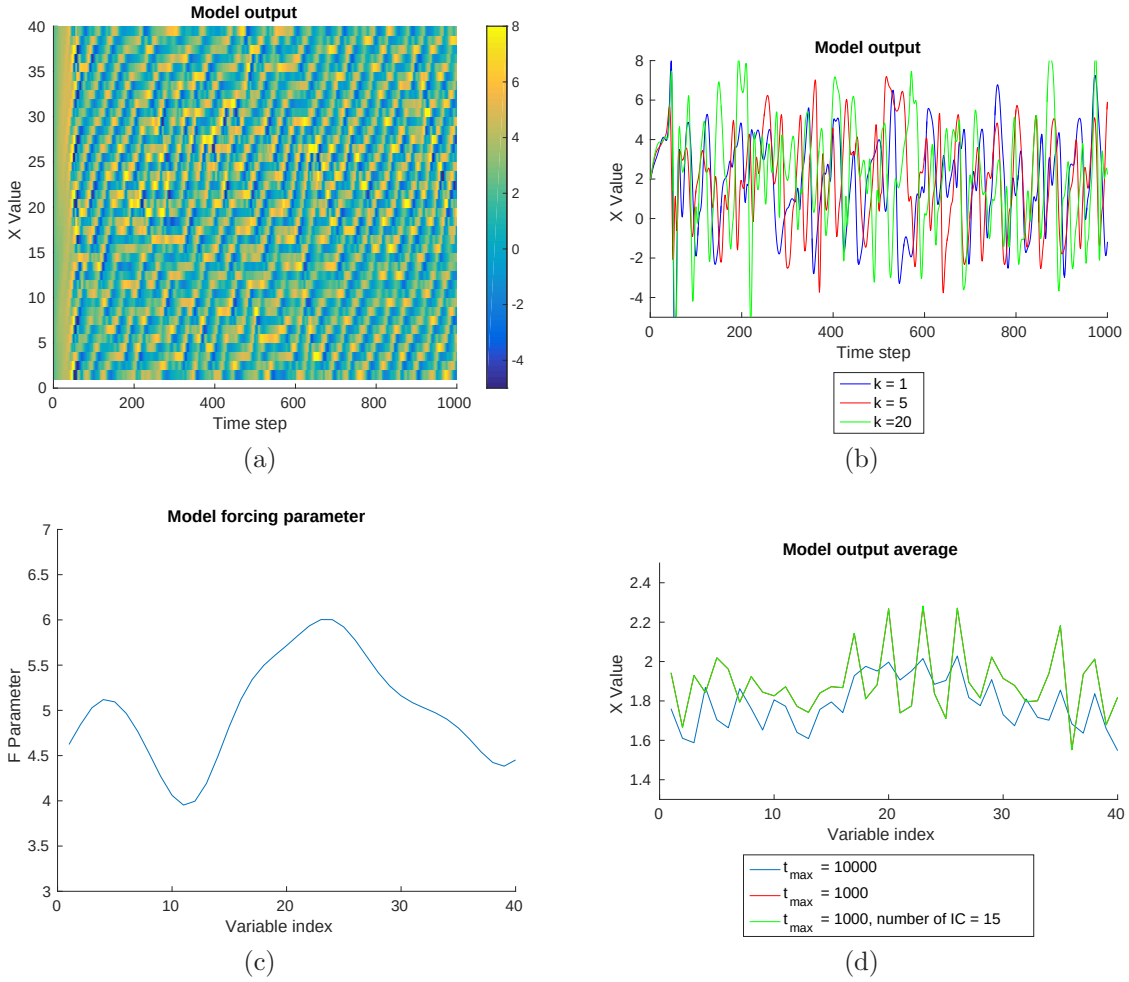


Figure 5.4: Lorenz '96 model evolution for 1000 time steps, for variable F_k parameter with average $\overline{\mathbf{F}} = 5$, and uniform $X_k = 2$, for all k initial conditions. Panel (a) is a 2D plot of X_k over time and spatial index. Panel (b) is the temporal evolution of X_k for $k = 1, 5, 20$. Panel (c) is the values of F_k. Panel (d) are different time averages of the model state X_k.

Those particular cases have not been considered by Lorenz in his extensive study of his model in Lorenz (2005). Lorenz first focused on the perturbations introduced by the variability in the initial conditions on X_k, and in particular the wave length

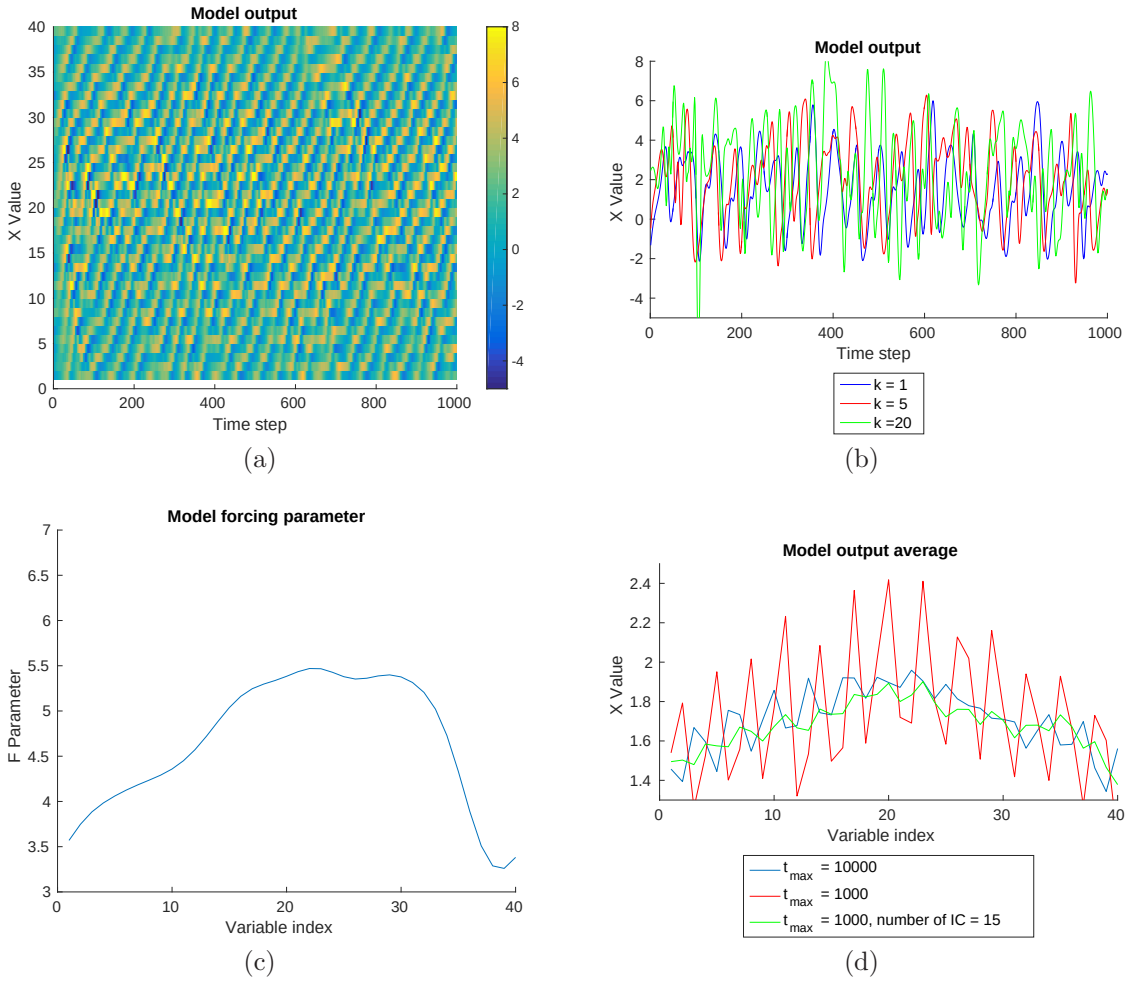


Figure 5.5: Lorenz '96 model evolution for 1000 time steps, for variable F_k parameter with average $\overline{\mathbf{F}} = 5$, and random X_k initial conditions with average $\overline{\mathbf{X}} = 2$. Panel (a) is a 2D plot of X_k over time and spatial index. Panel (b) is the temporal evolution of X_k for $k = 1, 5, 20$. Panel (c) is the values of F_k. Panel (d) are different time averages of the model state X_k.

of the variability of X_k. Lorenz also investigated a modification of his model by transmuting the factors in his equation into random numbers i and j as follow

$$\frac{dX_k}{dt} = -X_{k-i}X_{k-j} + X_{k-i+j}X_{k+j} - X_k + F. \tag{5.3.4}$$

He intended to solve the too abrupt variations in X_k that occurred in the first form of his model.

Lorenz finally concluded that his model, while not being appropriate for certain studies, still offered a potential ground for improvement and modifications to suit

one's needs. In particular, once chaos is installed when looking at the short-term evolution of X_k, the long-term variability remains insignificant. This could be an issue to the investigation of problems where regime shifts over long period of time are observed. One could produce long-term variability by modifying F in time. This could be compared to a parameter of a realistic model varying over long period of time, such as a seasonal variation.

This is to some extent comparable to what has been done here. Instead of having a variation over time, one can look at the spatial variations of F_k as a parameter that has a particular spatial structure. Different regimes can appear in different regions of the model.

5.4 Single assimilation

The bias correction method is implemented and tested with a Lorenz '96 model twin experiment. As shown before, the forcing parameter F_k can be considered to be directly linked with the model mean over a period of time. First, a random, but spatially correlated F_k^t parameter is created following equation (5.3.1), with a mean $\overline{F^t} = 4$. The model is then run once over $m_{max} = 1000$ time steps, with $l_{max} = 15$ different initial conditions. It is then averaged over the initial conditions and over time while ignoring the first 200 time steps to avoid the initial conditions to strongly influence the model mean. This provides the reference (or true) solution X_k^t, obtained from the full model trajectory $X_{k,l,m}^t$ as follow:

$$X_k^t = \frac{1}{l_{max}} \sum_{l=1}^{l_{max}} \frac{1}{m_{max} - 199} \sum_{m=200}^{m_{max}} X_{k,l,m}^t. \tag{5.4.1}$$

The exact same procedure is used to generate an ensemble of $N_e = 100$ different $F_{k,N}^f$. Each one is also run over 1000 time steps, with 15 initial conditions, and averaged without the first 200 time steps, producing an ensemble of model solutions noted $X_{k,N}^f$.

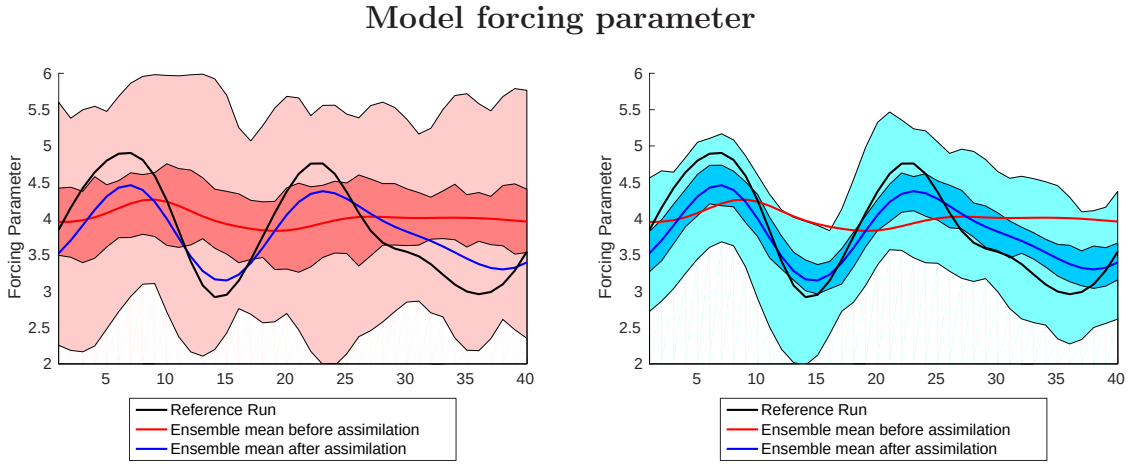
In the context of a classic twin experiment, one wants to assimilate observations y_k^o from the reference run mean X_k^t. In order to reproduce the behaviour and difficulties of a realistic experiment, noise is added to the reference run mean X_k^t and observations are created following

$$y_k^o = X_k^t + \beta s_{X_k^t} \mathbf{z}. \tag{5.4.2}$$

Here $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is a random vector, $s_{X_k^t}$ is the standard deviation of X_k^t, and $\beta = 0.1$.

5.4.1 Bias correction results

Using an ETKF scheme, the state vector, which consists of the ensemble model mean $X_{k,N}^f$, is extended with the ensemble $F_{k,N}^f$ (equation 4.1.11). After the analysis step, one obtains a new and updated vector of forcing parameter $F_{k,N}^a$, and the analysed ensemble model mean $X_{k,N}^a$. The model is rerun with those updated forcings, and one expects the ensemble model mean reruns $X_{k,N}^r$ to improve and come closer to the reference run. The results of this procedure are shown in figures 5.6a, 5.6b, 5.7a and 5.7b.



(a) (b)

Figure 5.6: Lorenz '96 model F_k value (Y-axis) for each $k = 1, .., 40$ (X-axis). The reference run is shown in black: F_k^t. The ensemble mean before assimilation, representing 100 members, is shown in red: F_k^f. The ensemble mean after assimilation is presented in blue: F_k^a. The light and darker areas represent then 25% and 50% percentile of the corresponding colored ensemble before assimilation (a) and after assimilation (b).

In this experiment, the whole ensemble with assimilated forcings is used for the final run. Figures 5.6a and 5.6b show the forcing ensemble envelope before (F_k^f) and after (F_k^a) assimilation respectively. Figures 5.7a and 5.7b show the model mean

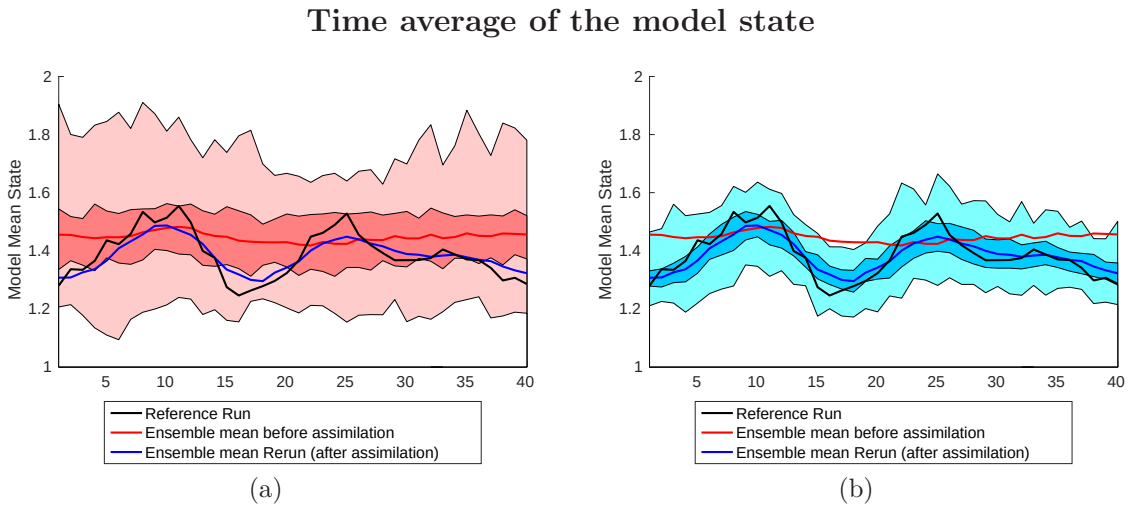


Figure 5.7: Lorenz '96 model X_k model mean state (Y-axis) for each $k = 1, .., 40$ (X-axis). The reference run is shown in black: X_k^t. The ensemble mean before assimilation, representing 100 members, is shown in red: X_k^f. The ensemble mean rerun after assimilation is presented in blue: X_k^r. The light and darker red areas represent then 25% and 50% percentile of the corresponding colored ensemble before assimilation (a) and after assimilation (b).

before assimilation (X_k^f) and after rerun (X_k^r) respectively.

The assimilation of observations on the model mean X_k^t allowed the correction of the bias on F_k^f (figure 5.6b). The root mean square error (RMSE) on F_k^f before assimilation was 0.653. After the assimilation, it has been reduced to 0.323 for F_k^a, and it is already able to reproduce the global shape of the reference run. One also needs to look at the model mean (figure 5.7b). The RMSE of the ensemble mean X_k^f is 0.099. However, one can clearly see that the model rerun with the assimilated F_k^a gives much better results. The RMSE of X_k^r is only 0.037, and reproduces much better the shape of the observations. Thus, not only does the assimilation show an improvement on the forcing parameter of the model, but its mean climatology is also improved by effectively correcting the source of its bias.

5.5 Iterative assimilation

The previous classic twin experiment has shown that the method can be applied to a chaotic system and provide a correction for the biased parameter. Kivman (2003) pointed out that the EnKF performs poorly when estimating simultaneously the state of the model and its parameters. However, by taking the average of the model over time and only estimating the biased parameter, the curse of highly non-

Gaussian probability distribution which affects chaotic models, and in particular the Lorenz '96 model, is diminished.

To obtain further improvement of the bias correction for specifically highly nonlinear systems, one can make use of the work of presented in Evensen and van Leeuwen (2000). It shows that, under the assumption of absence of correlation between the observational errors, one can choose to assimilate data sequentially. In particular, one can artificially create batches of data with a corresponding observation error covariance matrix. As long as the sequential assimilation of those batches are equivalent to the original data, one can bypass the curse of nonlinearities by applying smaller adjustments instead of a single and huge correction (Annan et al., 2005b).

This approach is similar to the "running in place" algorithm (RIP) presented in Kalnay and Yang (2010). Basically, the EnKF needs time to adapt to observations during a quick regime change. Examples are a storm developing in a weather forecast model, or a model initialised from a global, lower resolution, needing time to adapt to mesoscale observations. Since this scheme needs to be guided with observations to the optimal analysis, one can consider the EnKF to be blind in regime shifting situations. The idea of running in place is to assimilate the same observation multiple times during the spin up time, in order to extract the maximum amount of initial information. RIP allows a faster spin up, without loss of accuracy after the spin up, or requiring prior information such as a good estimation of the initial background error covariance. RIP uses a no-cost ensemble Kalman smoother (Kalnay et al., 2007), which is then turned off after the spin up.

The major difference with the iterative assimilation however is that the observations are assimilated without changing the observation error covariance matrix Yang et al. (2012b). The assumption is that due to the nonlinearities causing a regime change, the background may be in an unlikely state, and it may be desirable to extract more information from the observations. Similarly, during the initialisation, the ensemble may be started from scratch, and may not be representative of the most likely state.

Formally, for the iterative assimilation, one starts from the Kalman filter equations (equation (3.3.40) to (3.3.49)). Using the Sherman-Morrison-Woodbury formula, one can write the equivalent equations for the covariance matrix

$$(\mathbf{P}^a)^{-1} = (\mathbf{P}^f)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}, \tag{5.5.1}$$

$$(\mathbf{P}^a)^{-1}\mathbf{x}^a = (\mathbf{P}^f)^{-1}\mathbf{x}^f + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o. \tag{5.5.2}$$

As an illustration, one can split the observations $\mathbf{y}^o$ into two different batches, written as

$$\mathbf{y}_2^o = \begin{pmatrix} \mathbf{y}^o \\ \mathbf{y}^o \end{pmatrix}, \tag{5.5.3}$$

$$\mathbf{R}_2 = \begin{pmatrix} 2\mathbf{R} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{R} \end{pmatrix}, \tag{5.5.4}$$

$$\mathbf{H}_2 = \begin{pmatrix} \mathbf{H} \\ \mathbf{H} \end{pmatrix}. \tag{5.5.5}$$

One can show that the analysis provided by the two batches is equivalent to the assimilation of all the data in one single batch, since equation (5.5.1) and (5.5.2) become

$$(\mathbf{P}^a)^{-1} = (\mathbf{P}^f)^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} = (\mathbf{P}^f)^{-1} + \mathbf{H}_2^T\mathbf{R}_2^{-1}\mathbf{H}_2, \tag{5.5.6}$$

$$(\mathbf{P}^a)^{-1}\mathbf{x}^a = (\mathbf{P}^f)^{-1}\mathbf{x}^f + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o = (\mathbf{P}^f)^{-1}\mathbf{x}^f + \mathbf{H}_2^T\mathbf{R}_2^{-1}\mathbf{y}_2^o. \tag{5.5.7}$$

One can duplicate a single data set multiple times, and obtain the more general expression for the observation covariance matrix

$$\mathbf{R}_\gamma = \begin{pmatrix} \gamma_1\mathbf{R} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \gamma_2\mathbf{R} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \gamma_n\mathbf{R} \end{pmatrix}, \tag{5.5.8}$$

as long as the sum of the coefficient before the observation covariance matrix $\mathbf{R}$ remains

$$1 = \tfrac{1}{\gamma_1} + \tfrac{1}{\gamma_2} + \ldots + \tfrac{1}{\gamma_n}. \tag{5.5.9}$$

Equation (5.5.6) and (5.5.7) clearly show that in the absence of correlation be-

tween the different subsets of data, the assimilation in a single batch is equivalent to an iterated assimilation. Hence, one can easily iterate the assimilation cycle of one data set $\mathbf{y}^o$ by adapting the observation covariance matrix following equation (5.5.9). This result also applies in the context of the ETKF, in particular for the covariance matrices $\mathbf{P}^f$ and $\mathbf{P}^a$ described by the ETKF.

In the case of a linear observation operator, there is no advantage to perform an iterative assimilation, for that the analysis will be identical. However, for a nonlinear observation operator, differences will appear, and smaller corrections provided by the analysis will better handle the nonlinearities. It is certainly relevant for the Lorenz '96 model, where the observation operator even involves running the model again between assimilation iterations.

5.5.1 Observations batches creation

For clarity, an important notation is introduced. A difference is made between the total number of iterations for an entire assimilation experiment, noted n_{iter}^{max}, and the corresponding iteration assimilation n_{iter} within that experiment. Hence, n_{iter} takes all the values up to the maximum value n_{iter}^{max}: $n_{iter} = 1, \ldots, n_{iter}^{max}$. The objective is to compare a single assimilation experiment, $n_{iter}^{max} = 1$, with a double iteration, $n_{iter}^{max} = 2$, $n_{iter} = 1, 2$, or even more iterations.

Applied to the Lorenz '96 experiment, one can take the observations on the model mean $\mathbf{y}^o = X_k^t$ which are assimilated with an error covariance matrix $\mathbf{R}$. A set of n_{iter}^{max} artificial observations $\mathbf{y}'^o = \mathbf{y}^o$ is created. Assimilating this set of observation $\mathbf{y}'^o$ is equivalent to assimilating only once $\mathbf{y}^o$ as long as the covariance matrix of $\mathbf{y}'^o$ are set to $n_{iter}^{max}\mathbf{R}$, that equation (5.5.9) is fulfilled with $\sum_{i=1}^{n_{iter}^{max}} \frac{1}{n_{iter}^{max}} = 1$. In practice, the combination of those sets of observations provides the same posterior estimate.

The model can be integrated again between the assimilation of the different batches, resulting in smaller corrections, and reducing the apparition of unbalanced solutions, or even regime shifts. Hence, the nonlinear forward model operator is included in the observation operator $\mathbf{H}$. Performing the successive analyses provides one with a better posterior estimate, when compared to a single assimilation.

In order to avoid the collapse of the ensemble towards a particular solution, hence reduce the PDF described by the ensemble to a single value, an inflation factor is used to preserve the ensemble spread. Were the ensemble to collapse toward

a particular solution, the PDF describing the ensemble would not contain the observation anymore. The ETKF would then be unable produce a linear combination of the ensemble members to reconstruct the observations.

In other words, after each assimilation, the ensemble is artificially inflated with a factor. With $F_{k,N}^{a,n}$ being an ensemble member after the nth assimilation, $F_k^{a,n}$ being the ensemble mean after the nth assimilation, every ensemble member before the $n+1$th assimilation, noted $F_{k,N}^{f,n+1}$, is corrected with an inflation factor $1 \leq \alpha_{\mathbf{P}}$ as follow

$$F_{k,N}^{f,n+1} = F_k^{a,n} + \alpha_{\mathbf{P}}(F_{k,N}^{a,n} - F_k^{a,n}). \tag{5.5.10}$$

Hence, one also needs to make sure that, when generating the ensemble, the PDF described by the ensemble contains the observations. Otherwise, the ETKF would be confronted to the same issue as with a collapsed ensemble.

5.5.2 Experiment set-up

One needs to remember that since the model is rerun between the assimilation of each batch of observation, the computational cost of this method is proportional to n_{iter}^{max}, since one needs to rerun the model n_{iter}^{max} times for the iterative assimilation cycles. In the context of a Lorenz '96 twin experiment, n_{iter}^{max} is arbitrarily set to cover values from $n_{iter}^{max} = 1, \ldots, 4$. The objective is to study how the increase in the number of iterations affects the bias estimation. The inflation factor is first set to $\alpha_{\mathbf{P}} = 1$. The initial ensemble is created with a similar procedure as for the previous twin experiment (section 5.4).

However, for practical reasons and in order to investigate different assimilation parameters, the ensemble size and number of initial conditions have to be reduced. The model is run over $m_{max} = 1000$ time steps with $l_{max} = 10$ different initial conditions instead of $l_{max} = 15$. The ensemble size is reduced from $N_e = 100$ members to $N_e = 50$ members.

Furthermore, aiming at a clearer difference between a single and iterative assimilation cycle, the ensemble background estimate is different from the true run: $F^f = 6$, whereas the true, or reference run, is created with a mean $F^t = 5$. The ensemble spread is however sufficient for the observations to be contained by the

ensemble. The observations of the true run X_k^t, and of the ensemble model solutions $X_{k,N}^f$, are produced as in the twin experiment from section 5.4. Noise is also added to the observations, following the same procedure as for the previous twin experiment (equation (5.4.2)).

To make sure that the comparison between different iterative assimilation cycles is fair, the initial conditions of the ensemble $X_{k,l,m=1}^f$ and the initial parameter estimates $F_{k,N}^f$ are the same for every $1 \leq n_{iter}^{max} \leq 4$. This allows to show the unique influence of the number of assimilation cycles performed, while having all the other initial conditions and parameters to remain identical.

5.5.3 Results

The results of the experiments, for an inflation factor $\alpha_{\mathbf{P}} = 1$, are shown on figures 5.8a to 5.9b. Figure 5.8a represents the forcing parameter of the model for the reference run F_k^t (in black), the ensemble forecast F_k (in red) and the analysed ensemble F_k (blue). Figure 5.8b represents the average model state X_k over the initial conditions and the integration period, with the same color code as the corresponding F_k parameter. Both panels are for a single iteration (equivalent to a single assimilation experiment).

Figure 5.8c represents the forcing parameter F_k of the model for the reference run (in black), the ensemble forecast (in red) and the multiple iterations of the analysed ensemble, with $n_{iter}^{max} = 4$. Figure 5.8d represents the average model state X_k over the initial conditions and the integration period, with the same color code as the corresponding F_k parameter.

One can see the effect of the iterated assimilation from the comparison between Figs. 5.8a and 5.8c. For a single assimilation iteration, one has $n_{iter}^{max} = 1$. The analysis is shown with the blue line on figure 5.8a. With the same background ensemble and observations, an iterated analysis is performed, with $n_{iter}^{max} = 4$, shown on figure 5.8c. One can note that the correction to the forcing parameter for the first iteration ($n_{iter} = 1$) is stronger for the single assimilation than for the iterated assimilation, by comparing the blue lines on both figures. However, after the end of the cycle, with $n_{iter} = 4$, the last iteration provides a better estimate of F_k than the single assimilation. This is represented by the grey line on figure 5.8c. This shows the difference between one large correction for a single assimilation, and smaller

successive corrections for the iterative assimilation. Only the results for $n_{iter}^{max} = 1$ and $n_{iter}^{max} = 4$ are shown.

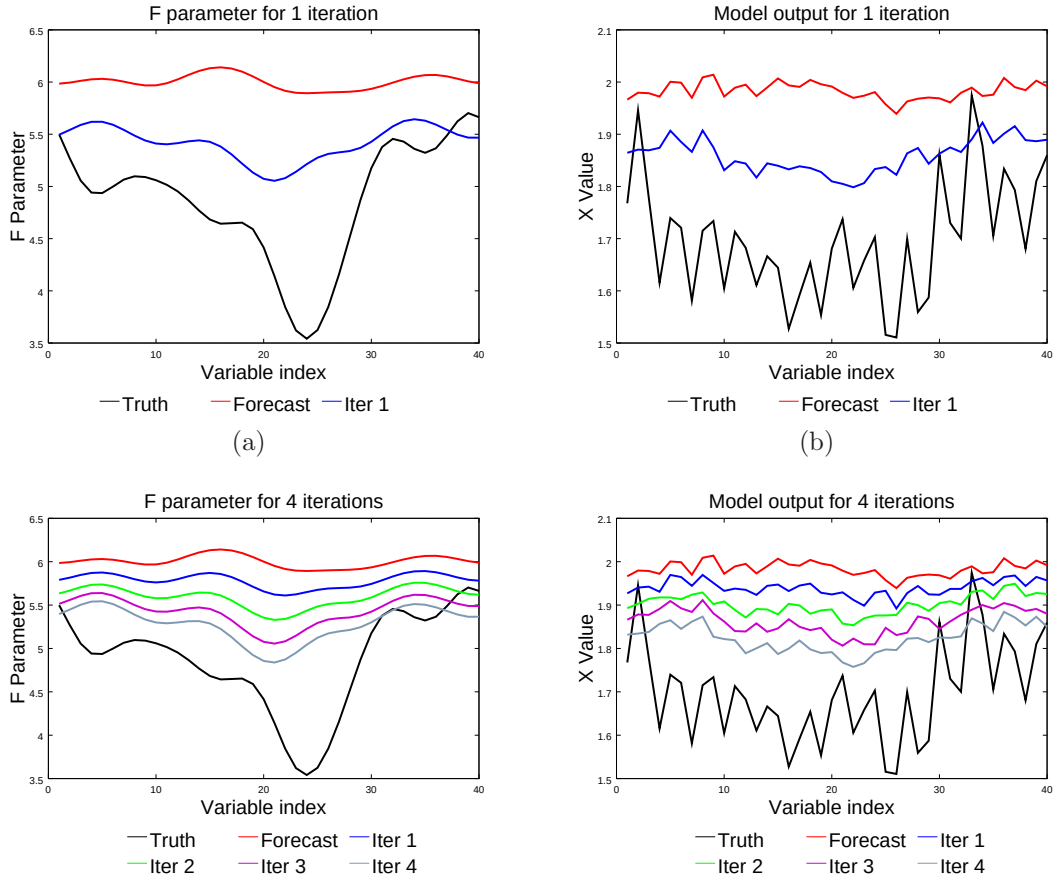
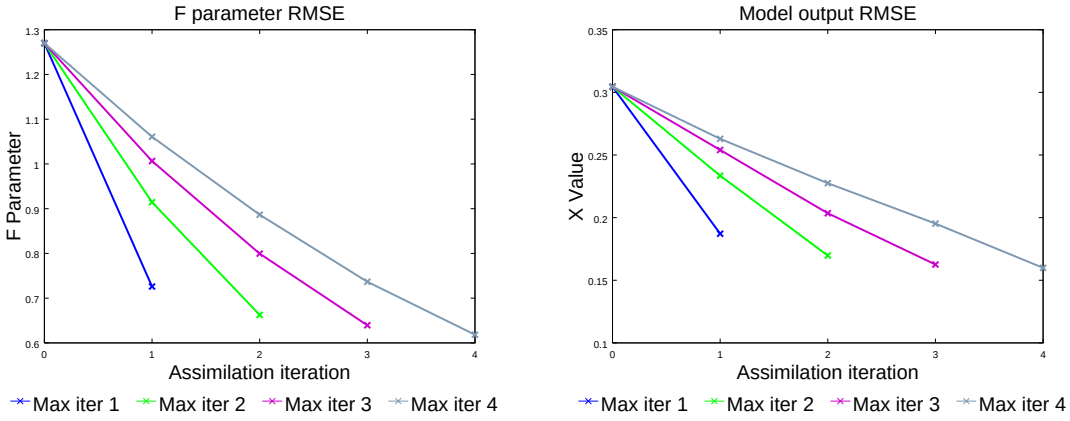


Figure 5.8: Panel (a): F_k for a single assimilation, $n_{iter}^{max} = 1$. Panel (b): Time average of the model state corresponding to F_k parameter from panel (a). Panel (c): F_k for an iterated assimilation, with $n_{iter}^{max} = 4$. Panel (d): Time average of the model state corresponding to F_k parameter from panel (b).

Figure 5.9a shows the root mean square error (RMSE) between the ensemble of estimated and analysed F_k, with the reference F_k^t. Here, every color corresponds to the number of iterations n_{iter}^{max} performed. Each point hence corresponds to the RMSE of the iteration n_{iter} with respect to the total number of iteration n_{iter}^{max} in that particular cycle. One can see that all simulations start from the same RMSE, corresponding to the initial ensemble estimate. For the first assimilation iteration, the RMSE on F_k increases as the maximum number of iterations increases too. This is due to the lower confidence in the assimilated observations: $\mathbf{R} < 2\mathbf{R} < 3\mathbf{R} < 4\mathbf{R}$.

Hence, the assimilation scheme applies a smaller correction to the ensemble. However, the last analysis of every cycle provides a better correction, hence lower RMSE, than the last analysis of cycles with less iterations. The exact values of the plots of the F_k parameter corresponding to iterations with $1 \leq n_{iter}^{max} \leq 4$, $n_{iter} = 1, \ldots, n_{iter}^{max}$ are given on table 5.1.

Figure 5.9b shows the corresponding time average of the model state. One can note that when the RMSE on F_k decreases, so does the RMSE on X_k. The RMSE values of the time average of the model state corresponding to iterations with $1 \leq n_{iter}^{max} \leq 4$, $n_{iter} = 1, \ldots, n_{iter}^{max}$ are given on table 5.2.



(a) (b)

Figure 5.9: Panel (a): RMSE on F_k for $1 \leq n_{iter}^{max} \leq 4$, for every $n_{iter} = 1, \ldots, n_{iter}^{max}$. Panel (b): RMSE on the time average of the model state with F_k corresponding to panel (a).

Background RMSE		Analysed RMSE			
n_{iter}^{max}	Background	$n_{iter} = 1$	$n_{iter} = 2$	$n_{iter} = 3$	$n_{iter} = 4$
1	1.270	0.726			
2	1.270	0.915	0.663		
3	1.270	1.007	0.799	0.639	
4	1.270	1.060	0.887	0.737	0.619

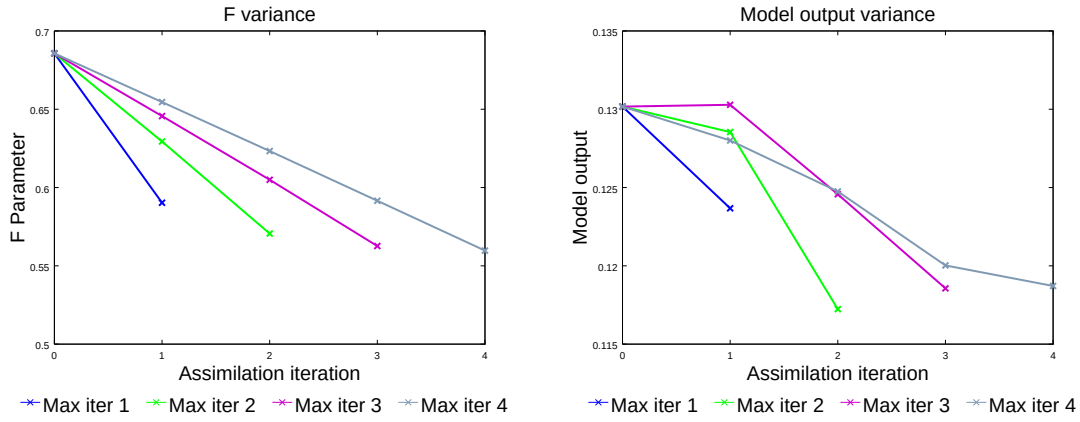
Table 5.1: RMSE on F_k, $\alpha_{\mathbf{P}} = 1$

The variance of the ensembles for every $1 \leq n_{iter}^{max} \leq 4$ are shown on figure 5.10a for F_k, and on figure 5.10b for the time average of the model state. One can see that, with no inflation ($\alpha_{\mathbf{P}} = 1$) of the ensemble between every assimilation iteration, the

Background RMSE		Analysed RMSE			
n_{iter}^{max}	Background	$n_{iter} = 1$	$n_{iter} = 2$	$n_{iter} = 3$	$n_{iter} = 4$
1	0.304	0.187			
2	0.304	0.233	0.170		
3	0.304	0.254	0.203	0.163	
4	0.304	0.263	0.227	0.195	0.160

Table 5.2: RMSE on the time average of the model state, $\alpha_{\mathbf{P}} = 1$

ensemble slowly converges.



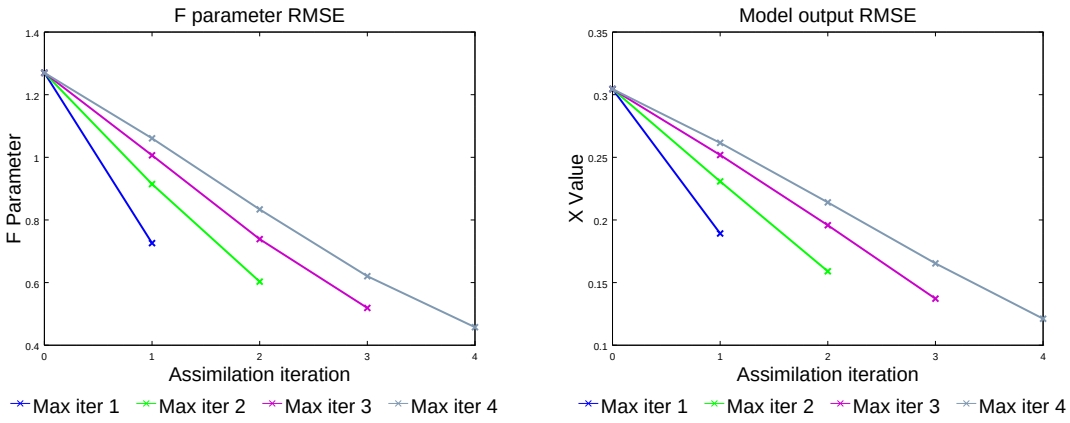
(a) (b)

Figure 5.10: Panel (a): Variance on F_k for $1 \leq n_{iter}^{max} \leq 4$, for every $n_{iter} = 1, \ldots, n_{iter}^{max}$. Panel (b): Variance on the time average of the model state with F_k corresponding to panel (a).

Inflation factor increase

The same experiment, with exactly the same initial conditions, is performed with an inflation factor of $\alpha_{\mathbf{P}} = 1.2$. This helps to counteract the collapse of the ensemble due to the assimilation. The values of the RMSE on the F_k parameter and on the the time average of the model state are plotted on Figs. 5.11a and 5.11b respectively. The results of the corresponding RMSE are shown on tables 5.3 and 5.4.

The reason for the better results obtained with the inflated ensemble lies within the generation of the ensemble compared to true run. For the assimilation to be efficient, the PDF described by the ensemble members must include the true run as much as possible, since the assimilation aims at combining the different ensemble



(a) (b)

Figure 5.11: Experiment with inflation $\alpha_{\mathbf{P}} = 1.2$. Panel (a): RMSE on F_k for $1 \leq n_{iter}^{max} \leq 4$, for every $n_{iter} = 1, \ldots, n_{iter}^{max}$. Panel (b): RMSE on the time average of the model state with F_k corresponding to panel (a).

Background RMSE		Analysed RMSE			
n_{iter}^{max}	Background	$n_{iter} = 1$	$n_{iter} = 2$	$n_{iter} = 3$	$n_{iter} = 4$
1	1.270	0.726			
2	1.270	0.915	0.603		
3	1.270	1.007	0.739	0.519	
4	1.270	1.060	0.834	0.620	0.458

Table 5.3: RMSE on F_k, $\alpha_{\mathbf{P}} = 1.2$

members to reconstruct the true run. If the PDF of the ensemble is too far away from the true run, artificially inflating the ensemble will help to widen the PDF and will provide better results. It is the origin of the improved results with $\alpha_{\mathbf{P}} = 1.2$.

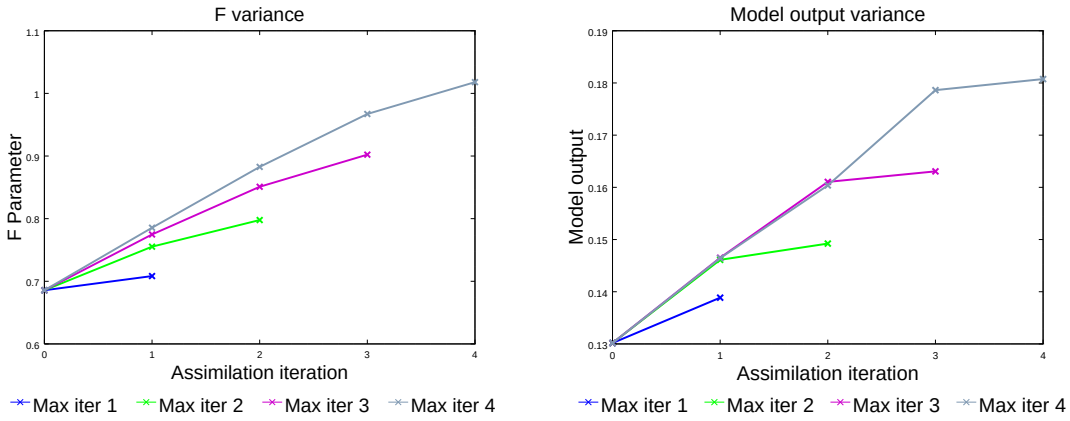
In particular, one can understand the importance of having a correct model for the error on F. Here, the true run has a mean $F^t = 5$, whereas the ensemble is generated with $F^f = 6$. While in this case, the ensemble spread is wide enough to contain the observations, one can conceive how this can be problematic in the context of a more complex model.

5.5.4 Conclusion

In this chapter, the Lorenz '96 model is adapted to correspond to the requirements of the experimentation of the bias correction method. The behaviour of the Lorenz '96 model is thoroughly investigated after its modification. The particular approach of considering the average model state of a nonlinear system in order to estimate

Background RMSE		Analysed RMSE			
n_{iter}^{max}	Background	$n_{iter} = 1$	$n_{iter} = 2$	$n_{iter} = 3$	$n_{iter} = 4$
1	0.304	0.187			
2	0.304	0.233	0.159		
3	0.304	0.254	0.196	0.137	
4	0.304	0.263	0.214	0.165	0.121

Table 5.4: RMSE on the time average of the model state, $\alpha_{\mathbf{P}} = 1.2$



(a) (b)

Figure 5.12: Experiment with inflation $\alpha_{\mathbf{P}} = 1.2$. Panel (a): Variance on F_k for $1 \leq n_{iter}^{max} \leq 4$, for every $n_{iter} = 1, \ldots, n_{iter}^{max}$. Panel (b): Variance on the time average of the model state with F_k corresponding to panel (a).

a biased parameter is detailed. A clear relationship between the forcing parameter and the average model state is highlighted.

The single assimilation twin experiment is presented. The results show that the bias is effectively reduced with a corrected bias parameter estimation. The particular spatial structure of the biased parameter is recovered by the ETKF analysis. The rerun with the corrected parameter clearly shows a significant improvement.

The iteration on the analysis cycle shows that, in the presence of nonlinearities either in the model or observation operator, one can make smaller adjustments to obtain a better result. The procedure is relatively simple to apply when one already has an operational assimilation set-up at hand. The increase in computational cost is substantial, but improvements are already available for a double iterated assimilation. However, one should always care for the collapse of the ensemble, which can be a considerable issue to the efficiency of the assimilation procedure.

Chapter 6

NEMO-LIM2

Contents

6.1 Model presentation

The primitive equations model used in this study is NEMO (Nucleus for European Modelling of the Ocean, Madec (2008)), coupled to the LIM2 (Louvain-la-Neuve Sea Ice Model) sea ice model (Fichefet and Maqueda, 1997; Timmermann et al., 2005; Bouillon et al., 2009).

6.1.1 Primitive equations

The Navier-Stokes equations (Navier, 1823) describing the motion of fluids through viscosity and pressure, can be used to describe the ocean. In addition, one also needs to couple temperature and salinity to velocity through a nonlinear equation of state. However, for a practical application to ocean modelling, additional assumptions need to be established in order to obtain a simplified and usable set of equations to be solved:

- The earth is considered to be a perfect sphere. The geopotential surfaces are assumed to be spherical. Hence, the local vertical vector defined by gravitation is always parallel to the earth radius. The aforementioned definition of sea surface height (section 2.5) is largely simplified.
- The ocean depth is considered to be negligible compared to the earth radius. This is referred to as the thin-shell approximation.
- Small scale processes have an effect on large-scale behaviour of the model. This is represented by turbulent fluxes, expressed in terms of large-scale features. This is referred to as the turbulent closure.
- The Boussinesq hypothesis assumes that one can neglect density variations in the ocean, except in their contribution to the buoyancy force.
- Convective processes are removed from the Navier-Stokes equations and are parametrised instead. The vertical momentum equation is thus reduced to the balance between the buoyancy force and the vertical pressure gradient.
- The ocean is considered to be incompressible. The three dimensional divergence of the velocity vector is assumed to be zero.

The dominant force acting on large-scale motions in the ocean are gravity, the Coriolis acceleration and the pressure gradient. NEMO uses an orthogonal set of unit vectors $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, which are directly linked to gravity. The two vectors $(\mathbf{i}, \mathbf{j})$ are tangent to the geopotential surfaces. The vector $(\mathbf{k})$ is the local upward vector, defined by gravity on a perfect sphere. Hence, $(\mathbf{i}, \mathbf{j})$ are orthogonal to $(\mathbf{k})$ by construction.

One can then define the following set of variables:

- The vector velocity $\mathbf{U} = \mathbf{U}(u, v, w) = \mathbf{U_h} + w\mathbf{k}$, where h is the local horizontal vector defined by $(\mathbf{i}, \mathbf{j})$.

- The potential temperature T.
- The salinity S.
- The *in situ* density ρ.

One can then write the vector invariant form of the primitive equations expressed in the $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ vector system under the form of a set of six equations. Those are respectively the momentum balance (equation (6.1.1)), the hydrostatic equilibrium (equation (6.1.2)), the incompressibility equation (equation (6.1.3)), the heat conservation (equation (6.1.4)), the salinity conservation (equation (6.1.5)), and finally an equation of state (equation (6.1.6)), as follow

$$\frac{\partial \mathbf{U_h}}{\partial t} = -\left[(\nabla \times \mathbf{U}) \times \mathbf{U} + \frac{1}{2}\nabla(\mathbf{U}^2)\right] - f\mathbf{k} - \frac{1}{\rho_0}\nabla_h p + \mathbf{D^U} + \mathbf{F^U}, \tag{6.1.1}$$

$$\frac{\partial \mathbf{p}}{\partial z} = -\rho g, \tag{6.1.2}$$

$$\nabla \cdot \mathbf{U} = 0, \tag{6.1.3}$$

$$\frac{\partial \mathbf{T}}{\partial t} = -\nabla \cdot (T\mathbf{U}) + D^T + F^T, \tag{6.1.4}$$

$$\frac{\partial \mathbf{S}}{\partial t} = -\nabla \cdot (S\mathbf{U}) + D^S + F^S, \tag{6.1.5}$$

$$\rho = \rho(T, S, p). \tag{6.1.6}$$

Here, ∇ is the generalised derivative vector operator in $(\mathbf{i}, \mathbf{j}, \mathbf{k})$, t is the time, z refers to the vertical coordinate, ρ_0 a reference density, p the pressure, f the Coriolis acceleration, and g the gravitational acceleration.

Small-scale physics is parametrised here by $\mathbf{D^U}$, D^T and D^S for the momentum, temperature and salinity respectively. Similarly, $\mathbf{F^U}$, F^T and F^S refer to the surface forcing term of the corresponding quantities.

6.1.2 Boundary conditions

The integration of NEMO-LIM2 over long time periods and on the ORCA2 grid means that the ocean edges will be contoured by complex coastlines on the sides, by a bottom topography at the ocean floor, and by a sea-ice and sea-atmosphere interface at the surface. The depth of the ocean is constant in time and defined at $z = H(x, y)$, with H being the local depth. However, the surface of the ocean $\eta = -H(x, y, t)$ is variable in time. Both H and η are defined with respect to a

given mean surface $z = 0$. In particular, η is used to define the model sea surface height anomaly.

Outwards and inwards exchanges of fluxes of heat, fresh water, salt and momentum of the ocean happen through those interfaces. In NEMO, one can describe the different boundary conditions choices of the ocean made to run the model as follow:

- Land: The major exchange between continental masses and the ocean happens through riven runoff, adding fresh water to the water cycle (evaporation, precipitation, ...)
- Oceanic floor: The exchange of heat and salt with the ocean floor are limited, and can be neglected by the model. However, momentum exchange is crucial, as the normal velocity at the interface is zero. In addition, friction also plays an important role, and must be parametrised in terms of turbulent fluxes.
- Atmosphere: Wind friction with the ocean surface leads to an exchange of horizontal momentum, which is also called wind stress. The atmosphere also transfers important fluxes of heat and fresh water with the ocean.
- Sea-ice: Sea-ice has a salinity of $\sim 4 - 6$ PSU, whereas the average value of the ocean is around ~ 34 PSU. Hence, mass exchanges of water through freezing and melting must be taken into account by the model. In addition, the sea-surface temperature is also constrained to be at the freezing point at the interface.

6.1.3 Subscale processes

The scale of space and time for which the primitive equations describe the ocean are valid for the order of magnitude of the kilometer (10^3 m) in the horizontal dimensions, for the meter (10^0 m) in the vertical dimension, and for a couple of minutes (10^2 s) for time. This strong anisotropy is induced between the vertical and horizontal motions by the dominance of gravitational forces on the system.

In particular, the grid used by NEMO in following experiments is the ORCA2 grid, with grid cells of up to 200 km. By construction, small-scale physical processes can not be explicitly solved, and must be represented in terms of large-scale patterns in the equation. This representation is called parametrisation.

For instance, the small-scale motions effects coming from the advective terms in the Navier-Stokes equations appear in the equations as the divergence of turbulent fluxes associated with the mean correlation of small-scale perturbations. By assuming the turbulent closure hypothesis to establish the primitive equations, one must choose a formulation for these fluxes.

In some cases and for short term integration of the model, the subgrid processes have a weak influence. For example, the river runoff at the land-ocean interface is the main source of exchange of fresh water. Sea surface salinity is directly impacted, in particular in the vicinity of river mouths. This can be neglected for short range integrations, but this process influences the characteristics of water masses which are formed over long time periods, especially at high latitudes. The water cycle of the earth climate system greatly depends on its closure through river runoff in the ocean. Another example is the importance of small scale processes for the balance of surface input of kinetic energy and heat.

Taking into account the anisotropy of the scales in the system, one can split the subgrid-scales physics into vertical and lateral motions. One can decompose $\mathbf{D}^{\mathbf{U}}$, D^T and D^S from equation (6.1.1), (6.1.4) and (6.1.5) respectively into their lateral part $\mathbf{D}^{l\mathbf{U}}$, D^{lT} and D^{lS} and vertical part $\mathbf{D}^{v\mathbf{U}}$, D^{vT} and D^{vS}.

6.1.4 ORCA2 grid

The ORCA2 grid has a peculiar structure, whose purpose is to overcome the North Pole singularity which usually poses problems to ocean models. Indeed, in a traditional longitude-latitude coordinates system, one creates a singular point in the Arctic Ocean, precisely at the North Pole. Meridians converge inside the computational domain, causing a severe restriction on the length of the time step to solve the model equation. The integration of the model over too long time steps causes computational stability issues for finite differences schemes. The ORCA2 grid solves this obstacle by moving the mesh poles to land points, effectively removing the singularities from the ocean (Madec and Imbard, 1996).

In practice, the ORCA2 grid is a 2-degree resolution grid which consists of the combination of two adjacent grid tiles. The first tile, located in the southern part of the globe, is a rectangular grid which extents from latitude 78.190582275 S to latitude 19.605793 N. In the longitude direction, the spacing of 2 degree is constant. However, in the latitude direction, the spacing is variable and is proportional to

$\sim \cos(lat)$. The first grid tile is located at $(i = 1, j = 1) = (78.190582275S, 102W)$. The second tile, situated in the north of the first tile, has a more complex structure aimed at avoiding singularities located in the ocean. Hence, it consists of a dipolar grid in which the grid cells are irregular quadrilaterals. The two poles are located at $(50N, 80E)$ and $(70N, 100W)$. Both poles are connected through a great circle arc which represents the locus of the fold in the grid tile. The upper boundary of the north tile is folded on itself along this line. Those two tiles are produce a global grid of index-space dimensions $i = 180$ and $j = 148$, which are attached at row $j = 92$.

Due to the particular structure of the ORCA2 grid, vector quantities evaluated at the grid midpoints, which are not exactly halfway between the corners. Additionally, cell boundaries are not straightforward curves in the sense that they are neither great circles nor straight lines in the latitude-longitude space. Hence, the grid contains actually four separate places T, U, V, W located respectively at the midpoint, on the vertical and horizontal cell sides, and at the corner of the cell. They allow the different quantities to be represented on the correct part of a grid cell.

6.1.5 Implementation

The global ORCA2 implementation is used, which is based on an orthogonal grid with a horizontal resolution of the order of 2° and 31 z-levels (Mathiot et al., 2011; Massonnet et al., 2013). The hydrodynamic model is configured to filter free-surface gravity waves by including a damping term. The leap-frog scheme uses a time step of 1.6 hours for dynamics and tracers. The model is forced using air temperature and wind from the NCEP/NCAR reanalysis (Kalnay et al., 1996). Relative humidity, cloud cover, and precipitation are based on a monthly climatological mean. The sea surface salinity is relaxed towards climatology with a freshwater flux of -27.7 mm/day.

Because of its low resolution of about 2°, the NEMO-LIM2 model is subject to strong bias due to poorly located currents in the ocean. This leads to a poorly represented heat transport around the globe and causes bias on other variables in the model, such as on the sea surface height and temperature. It is assumed that these bias are constant in time but may have a spatial structure.

6.2 Mixed layer depth

A characteristic of the ocean, or any sufficiently large water mass for instance, is stratification. Water masses in different layers exhibit different properties for salinity (halocline), density (pynocline), temperature (thermocline), etc. The layers act as a barrier to the mixing of the different strata.

One can define the mixed layer depth in the ocean as the surface layer in which the turbulent mixing processes are active and provide an almost vertically uniform profile for temperature, salinity and density (de Boyer Montégut et al., 2004). Fluxes of mass, momentum and energy through the mixed layer, and its thickness, determine the direct interactions between the ocean and the atmosphere.

The mixed layer depth is arbitrarily defined and based on different parameters such as temperature or density gradients. Its spatial variability can range from 20 m to 500 m, and its temporal variability includes diurnal, seasonal and intraseasonal variability (Kara et al., 2003).

In de Boyer Montégut et al. (2004), a 0.2-degree resolution global climatology of the mixed layer depth is constructed based on individual profiles. The selected criterion is a threshold value of the temperature ($\Delta T = 0.2°\text{C}$) or density ($\Delta\sigma_\theta = 0.03$ kg m^{-3}) from a 10 m depth value. The NEMO-LIM2 model adopts this definition of the mixed layer depth.

The use of both temperature and density is encouraged by the appearance of vertically density-compensated layers in the mid- and high-latitude winter hemispheres, effectively creating an isopynical but not mixed layer. A criterion using both quantities allows a more precise determination of the mixed layer depth.

Since the aim is to provide a single constant forcing term, the mixed layer depth used to constrain the forcing has been obtained with a yearly average of a NEMO-LIM2 free run. Clearly, the strong vertical variations of the mixed layer between seasons, in particular due to ice melting and deep water formation in both hemisphere, is not represented. The average value of the yearly mixed layer depth is around 25 m, as shown on figure 6.2.

6.3 Bias in NEMO-LIM2

The NEMO-LIM2 model has been used in the context of the PredAntar project (Goosse et al., 2015), which consisted in the study of the Antarctic sea-ice coverage during the period 1980-2009. Whereas the Arctic sea-ice has been decreasing drastically (Stroeve et al., 2007; Zeng and Delworth, 2015), the sea-ice extent in the Southern hemisphere around the Antarctic has been slightly increasing over the same period. In particular, between November 1978 and December 2012, the increase in ice extent is estimated between 0.13 and 0.2 million km^2 (Vaughan et al., 2013).

The PredAntar project first aimed at expanding the understanding of the complex mechanisms acting on sea-ice in the Southern hemisphere and in particular the processes involved in the Southern Ocean, despite the difficulties imposed by imperfect models and incomplete observations. The improvement of predictions was allowed through the development of post-processing tools providing an assessment of model errors, and through corrections.

Reconstructions of the Antarctic sea ice cover were obtained through data assimilation techniques. They proved to be a valuable compensation for the lack of observations over the considered time period. The correction of the Antarctic Circumpolar Current allowed a better estimate of its position and strength.

It is in this context that the bias correction method presented in chapter 4.1 is an appropriate tool. The following study aims at highlighting the presence of bias of the NEMO-LIM2 model regarding the Antarctic sea-ice coverage. The modification of the NEMO-LIM2 model in this particular case was considered and served as initial motivation, though the project was finished before the development of a stable NEMO-LIM2 bias correction term.

6.3.1 CMIP5

A series of comparisons were performed between the free and analysed runs of NEMO-LIM2 in the PredAntar project context (called PredAntar free and analysed runs), and the model results provided through the 5^{th} Coupled Model Intercomparison Project (CMIP5). Systematic biases in the mean state and in the internal variability of the Antarctic sea ice cover were highlighted (Goosse et al., 2015).

The sea ice coverage data from the PredAntar free and analysed runs are compared with two CMIP5 projects, namely the Centro Euro-Mediterraneo sui Cambiamenti Climatici - Climate Model without a resolved stratosphere (CMCC-CM), and the same model with a resolved stratosphere (CMCC-CMS) (Scoccimarro et al., 2011). Observations are obtained from the Operational SST and Sea Ice Analysis (OSTIA) system. Those particular models were chosen for that they operate on the same ORCA2 grid, similarly to NEMO-LIM2. In addition, the atmospheric part of NEMO-LIM2 consists of a reanalysis of the atmosphere. Hence, it is already influenced by the observations. On the other hand, both CMCC-CM and CMCC-CMS models have a running model for the atmosphere and for the ice coverage.

CMCC-CM(S)

The CMCC-CM(S) models has as ocean component the OPA 8.2 (Madec et al., 1998). It also includes the Louvain-La-Neuve (LIM) model for the dynamics and thermodynamics of the sea ice (Fichefet and Maqueda, 1999). Ocean physics includes a free-surface parametrisation (Roullet and Madec, 2000) and the Gent and McWilliams (1990) scheme for isopynical mixing. The atmospheric model component is ECHAM5 (Roeckner et al., 2003) with a T159 horizontal resolution, corresponding to a Gaussian grid of about 0.75° by 0.75°. A more detailed description of the ECHAM model performance can be found in Roeckner et al. (2006).

The communication between the atmospheric model and the ocean models is carried out with the Ocean Atmosphere Sea Ice Soil version 3 (OASIS3) coupler (Valcke, 2006). Every 160 min (coupling frequency), heat, mass, and momentum fluxes are computed and provided to the ocean model by the atmospheric model. SST and sea surface velocities are provided to the atmospheric model by both ocean models. The global ocean model also provides sea ice cover and thickness to the atmospheric model. The relatively high coupling frequency adopted allows an improved representation of the interaction processes occurring at the air–sea interface. No flux corrections are applied to the coupled model.

OSTIA

Global foundation sea surface temperature from OSTIA (Operational Sea Surface Temperature and Sea Ice Analysis Stark et al. (2007); Roberts-Jones et al. (2012); Donlon et al. (2012)) at an original resolution of 0.05° was reduced to a resolution of 2° by averaging all temperature values within a 2° by 2° grid cell.

Global sea ice fraction from the EUMETSAT Ocean and Sea Ice Satellite application Facility (OSI-SAF Roberts-Jones et al. (2012)) was also reduced to a resolution of 2° and assimilated with an error standard deviation of 0.1. The OSI-SAF sea ice fraction are distributed by MyOcean.

6.3.2 Preparation work

All sea ice coverage data available from the models use the ORCA2 grid. Using the same procedure, they are then all interpolated on the grid from OSTIA observations. This grid is constantly spaced with a 2° resolution, giving a global coverage of 180 by 90 cells. The data sets cover a period of 21 years, from January 1985 up to December 2005. Data from CMCC-CM(S) models are already monthly averages. Consequently, the monthly average for OSTIA observations and data from the PredAntar free and analysed runs are taken. Also, only the southern hemisphere is considered for all the following comparisons since the focus of the PredAntar is the sea ice coverage in the Southern Hemisphere.

Finally, all the sea ice coverage are in fraction of 1. In order to get the sea ice area in m^2, each cell coverage is multiplied by its area

$$A_{i,j} = \Delta\lambda \; \Delta\varphi \; \frac{\pi^2}{180^2} \; R^2 \cos(\varphi), \tag{6.3.1}$$

where $R = 6371000$ m is the mean earth radius, λ and φ are respectively the longitude and the latitude in degrees, and i, j spatial indices.

6.3.3 Seasonal Cycle

First, the seasonal cycle of the models is considered (figure 6.3). To obtain this figure, the monthly sea ice area is calculated from

$$SIA_{p,n} = \sum_i \sum_y A_{i,j} \; SIC_{p,n,i,j}, \tag{6.3.2}$$

where the indices p, n refer to months and years respectively, and i, j to the spatial dimensions. The monthly sea ice area is then obtained by taking the monthly average for all the years included in the 1985 - 2005 period. One can clearly see on figure 6.3 that all models are globally able to reproduce the mean seasonal cycle of the sea ice area (SIA) over the south pole. All models underestimate the SIA during the summer period (December-March). Both the CMCC-CM(S) and the PredAntar free run clearly tend to overestimate the sea ice area during the winter

(July - September). The PredAntar free run also overestimates the SIA during the autumn, starting from April, and performs better than the CMCC-CM(S) models during the winter. One can also note that the CMCC-CMS systematically performs worse than the CMCC-CM model. Finally, the PredAntar analysed run sticks to the OSTIA observations, as expected due to the analysis. Interestingly, it slightly underestimates the SIA throughout the whole year. This might be due to the fact that because of the assimilation, the data is smoothed through the whole domain. This tends to slightly reduce the SIC, hence the SIA, of the PredAntar analysed run.

The Root Mean Square Error of the different models with the OSTIA observations averaged over the 1985-2005 period is shown on figure 6.4. First, the monthly sea ice area averaged over the 1985-2005 period is calculated by

$$SIA_{p,i,j} = A_{i,j} \frac{1}{N} \sum_{n} SIC_{p,n,i,j}. \tag{6.3.3}$$

The RMSE with the OSTIA observation over the domain is obtained with

$$SIA_{p}^{rmse} = \sqrt{\sum_{i} \sum_{j} \left(SIA_{p,i,j} - SIA_{p,i,j}^{obs}\right)^2}. \tag{6.3.4}$$

Figure 6.4 represents the mean monthly RMSE of the models compared to the OSTIA observations. The RMSE of the PredAntar analysed run is much lower than the other models, since it assimilates the data from which the RMSE is calculated. However, the PredAntar free run performs overall similarly to the CMCC-CM(S) models. One can note that for all the models, the RMSE is at its lowest during the summer months, and at its highest during the winter. The main difference between the PredAntar free run and the CMCC-CM(S) models is the period from February to May, where the former has a decreasing RMSE, whereas the later ones have an increasing RMSE. One can note in particular the huge increase in March and April for the CMCC-CMS.

From figure 6.3, one can think that the CMCC-CM(S) models would at least perform better during the summer, since they better reproduce the total SIA. However, this is not the case, and the PredAntar free run has a RMSE similar to the CMCC-CM(S) models throughout the whole year. This difference could come from the fact that the PredAntar free run which, though it does not reproduces the correct total SIA, is able to place the ice at better locations than the CMCC-CM(S) models, thus producing a smaller RMSE with OSTIA observations.

This hypothesis is confirmed when looking at the mean spatial RMSE of sea ice concentration (SIC) of the models with the OSTIA observations. It is obtained by calculating the RMSE of the SIC of the models with OSTIA observations, but not averaging over the domain. The monthly mean state is first calculated by averaging over the whole considered period, then the mean RMSE with OSTIA observations is computed by averaging over a year, as follow

$$SIC_{p,i,j} = \frac{1}{N} \sum_{n} SIC_{p,n,i,j}, \tag{6.3.5}$$

$$SIC^{rmse}_{i,j} = \sqrt{\frac{1}{12} \sum_{p} \left(SIC_{p,i,j} - SIC^{obs}_{p,i,j} \right)^2}. \tag{6.3.6}$$

Both CMCC-CM(S) models produce more localised, but larger errors in the sea ice area. Those errors are the strongest around the Lazarev and Riiser-Larsen seas for CMCC-CM (figure 6.5a), and in the Amundsen sea for CMCC-CMS (figure 6.5b). As expected, the PredAntar analysed run performs very well, and has a nearly uniform RMSE over the whole sea ice domain (figure 6.5c). Finally, the PredAntar free run seems to perform rather well, with errors mainly located in the Somov and D'Urville sea, and along the coast of Graham land (figure 6.5d).

6.3.4 Internal Variability

One can also look at the respective internal variability of all the models. While this is not a direct measurement of the model bias, the behaviour of the model over different years still results from the representation of the processes in the model. In particular, biased currents will affect sea-ice formation over different years, hence the model variability.

First, the mean RMSE of the model is calculated and compared to one particular reference year. This is performed while considering all the years as reference year, and averaged, so that the final result is a monthly mean internal variability of the model, as follow

$$SIA_p^{rmse} = \sqrt{\frac{1}{N}\sum_{n' \neq n}\left[\frac{1}{N}\sum_n\sum_i\sum_j A_{i,j}\left(SIC_{p,n,i,j} - SIC_{p,n',i,j}\right)^2\right]}. \quad (6.3.7)$$

One notes from figure 6.6 that all the models have the same order of magnitude for their respective internal variability. The variability of CMCC-CM(S) models is at its most 20% higher than that of the OSTIA observations. Interestingly, the PredAntar analysed run has a lower internal variability than the observations, but reproduces the exact same shape of the observations. Again, this shows the smoothing of the data through the whole domain due to the assimilation.

One can also look at the spatial internal variability of one month in particular. September is chosen, as is it the month when both the sea ice area, and the RMSE on the sea ice area are at their highest. One can use equation (6.3.7) and remove the spatial sum to obtain

$$SIA_{p,i,j}^{rmse} = \sqrt{\frac{1}{N}\sum_{n'}\left[\frac{1}{N}\sum_n\left(SIC_{p,n,i,j} - SIC_{p,n',i,j}^{ref}\right)^2\right]}. \quad (6.3.8)$$

One notes that, for the CMCC-CM(S) models, the area where the internal variability is the highest tend to correspond with the area where the mean RMSE with OSTIA observations were the largest (figures 6.5a, 6.5b). This is especially true for the Lazarev and Riiser-Larsen seas. The PredAntar free and analysed runs seem to much better reproduce the internal variability of the OSTIA observations.

6.3.5 Conclusion

As expected, the PredAntar analysed run reproduces exactly the behaviour of the OSTIA observation, since it assimilated those data. When comparing the performance of the PredAntar free run and the CMCC-CM(S) models, one can conclude that though the former one has a worse total sea ice area estimation, it has a better localisation for the ice. Its mean spatial RMSE is thus lower than the one of the CMCC-CM(S) models. All models do reproduce the internal variability of the observations quite correctly. However, all models are clearly affected by bias.

This study highlights the issues of model bias on long-term integrations, and the importance of accounting for model bias. Model reanalysis through the use of

independent observations proves to be efficient. Specifically, the PredAntar assimilated run is an example of the general idea of classic methods of data assimilation. Although the model state and output are corrected over the considered time period, the model bias source remains unmodified. Classic reanalysis is only possible for periods of time when observations are available. The estimation of a bias correction term aimed at correcting the intrinsic model bias would allow one to both correct model simulations of the past, and the future. One could benefit both methods simultaneously, by applying the bias correction term during the model run, and performing reanalyses when possible.

The development of an efficient model bias correction method is an essential step into improving numerical modelling. It is in this framework that the correction of the bias on the general circulation in NEMO-LIM2 has been considered.

6.4 Bias correction generation

The aim here is to estimate a forcing term which will correct the oceanic currents of the model. This forcing will be, in practice, a constant acceleration term directly injected into the momentum equations of the ocean-dynamics part of the model. These added constant forces on water masses will create currents correcting the model bias also for other variables. Although the term "forcing" usually refers to external forcings such as atmospheric wind stress, the forcing term here refers thus to an additional source term in the momentum equations. It does not have an external origin, but rather aims at correcting the model error such as those arising from poorly represented physical processes.

However, since the NEMO-LIM2 model is a realistic model, specific constraints need to be imposed to the forcing term in order to maintain a physical and realistic model behaviour.

6.4.1 Horizontal structure

To create a constrained random forcing term, DIVA-ND proves to be a useful tool. It is a Data-Interpolating Variational Analysis in N dimensions (Barth et al., 2009, 2014). This tool allows to generate a random, spatially correlated fields $\Psi(x, y)$.

DIVA-ND defines a cost function $J(\Psi)$, which is expressed as

$$J(\Psi) = \int_{\Omega} L_h^4(\nabla^2\Psi)^2 + 2L_h^2(\nabla\Psi)^2 + \Psi dx, \tag{6.4.1}$$

where $\Psi = \Psi(x, y)$ is the random field and Ω the domain on which it is built. This cost function penalises abrupt variations over a given length-scale L_h, and decouples disconnected areas based on topography. The length-scale L_h will be specified in the experimental set-up with the NEMO-LIM2 model. The Hessian matrix of this discretised cost function is used to create random fields taking the periodicity in the model domain into account, with

$$J(\mathbf{x}_\Psi) = \mathbf{x}_\Psi^T \mathbf{P}_\Psi^{-1} \mathbf{x}_\Psi, \tag{6.4.2}$$

$$\mathbf{x}_\Psi^{(n)} = \mathbf{P}_\Psi^{1/2} \mathbf{z}_{(n)}, \tag{6.4.3}$$

$$\mathbf{z}_{(n)} \sim \mathcal{N}(0, \mathbf{I}). \tag{6.4.4}$$

Here, $\mathbf{x}_\Psi$ is the discretised random field on the model grid, $\mathbf{P}_\Psi^{-1}$ the Hessian matrix, and $\mathbf{z}_{(n)}$ a random vector with a normal distribution $\mathcal{N}(0, \mathbf{I})$. More extensive information can be found in Barth et al. (2009).

6.4.2 Stream function

One can use the DIVAN-ND tool to generate a stream function, which describes the streamlines of a flow under the assumption of incompressibility. One can derive the velocity components from a stream function and obtain a divergence free flow. Applied to an ocean model, meridional and zonal forcing fields for the currents can then be derived from $\Psi(x, y)$.

However, this could produce currents which are perpendicular to the coasts. In order to avoid such physically impossible currents, an additional constraint is applied when generating the random field Ψ. The generated stream function is subjected to the strong constraint $\nabla\Psi \bullet \mathbf{t} = 0$ where $\mathbf{t}$ is the vector tangent to the coast.

Additional spatial filtering on the obtained field Ψ is needed in order to remove very small scale signals when calculating the first derivatives of Ψ. This spatial filtering improves the stability of the NEMO-LIM2 model by a factor of 10^2 when it is forced. The variability of the SSH rises from 0.3 cm up to 30 cm. It is obtained

through a convolution product between Ψ and a 3 by 3 dissipation matrix, effectively smoothing the field.

6.4.3 Vertical extension

Since the aim is to create currents only in the upper layers of the ocean, but avoid modifying the global circulation in depths, the forcing is extended vertically as follow

$$\Psi(x,y,z) = \frac{\Psi(x,y)}{1+\exp(\frac{z-T(x,y)}{L_v})}, \tag{6.4.5}$$

where $T(x,y)$ is defined as the yearly average ocean mixed-layer thickness, and $L_v = 1$ m is a factor to obtain an adimensional exponential. One obtains $T(x,y)$ from an average of a NEMO-LIM2 free run over a one year integration period.

Even though the mixed layer does show a seasonal variability, its definition is arbitrary and based on specific criteria (presented in section 6.2). Additionally, the objective here is not to completely avoid the perturbation of the mixed layer, but rather avoid to modify the deep oceanic circulation. Hence, a yearly average of the mixed layer depth value is sufficient to restrain the ensemble to a specific vertical structure.

The resulting field is used as a stream function from which zonal and meridional divergence-free forces are derived as

$$F_u(x,y,z) = -\frac{\partial \Psi(x,y,z)'}{\partial y}, \tag{6.4.6}$$

$$F_v(x,y,z) = \frac{\partial \Psi(x,y,z)'}{\partial x}. \tag{6.4.7}$$

One can directly add this stochastic forcing terms into the momentum equations of NEMO-LIM2, where $F_u(x,y,z)$ and $F_v(x,y,z)$ are zonal and meridional components respectively. One then has

$$\frac{du}{dt} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + fv + \frac{1}{\rho}\frac{\partial \tau_x}{\partial z} + F_u, \tag{6.4.8}$$

$$\frac{dv}{dt} = -\frac{1}{\rho}\frac{\partial p}{\partial y} - fu + \frac{1}{\rho}\frac{\partial \tau_y}{\partial z} + F_v. \tag{6.4.9}$$

Equation (6.4.8) and (6.4.9) provide a set of bias-corrected ocean-dynamics equa-

tions governing the NEMO-LIM2 model by applying a forcing term on the ocean currents while the model is running. The forcing term is a physically coherent correction that will remove some part of the bias of the model. It has been calibrated such that the variability of the sea surface height (SSH) caused by the forcing is about 28 cm, which can be compared to the root mean square error between the NEMO-LIM2 model and the CNES mean dynamic topography of 20 cm (Rio et al., 2011).

In terms of magnitude, one can compare the forcing against the Coriolis acceleration on ocean currents. Typically, the average velocity in the ocean has a magnitude of 0.1 ms^{-1} (with peaks up to 2 ms^{-1} in the Gulf Stream). With the Coriolis parameter having a magnitude of 10^{-4} rad s^{-1}, the Coriolis acceleration scales around 10^{-5} ms^{-2}. The maximum magnitude of the forcing obtained from this generation mechanism peaks at 10^{-5} ms^{-2}, which is at the most comparable to Coriolis, and which is in average an order of magnitude smaller.

ORCA2 grid.

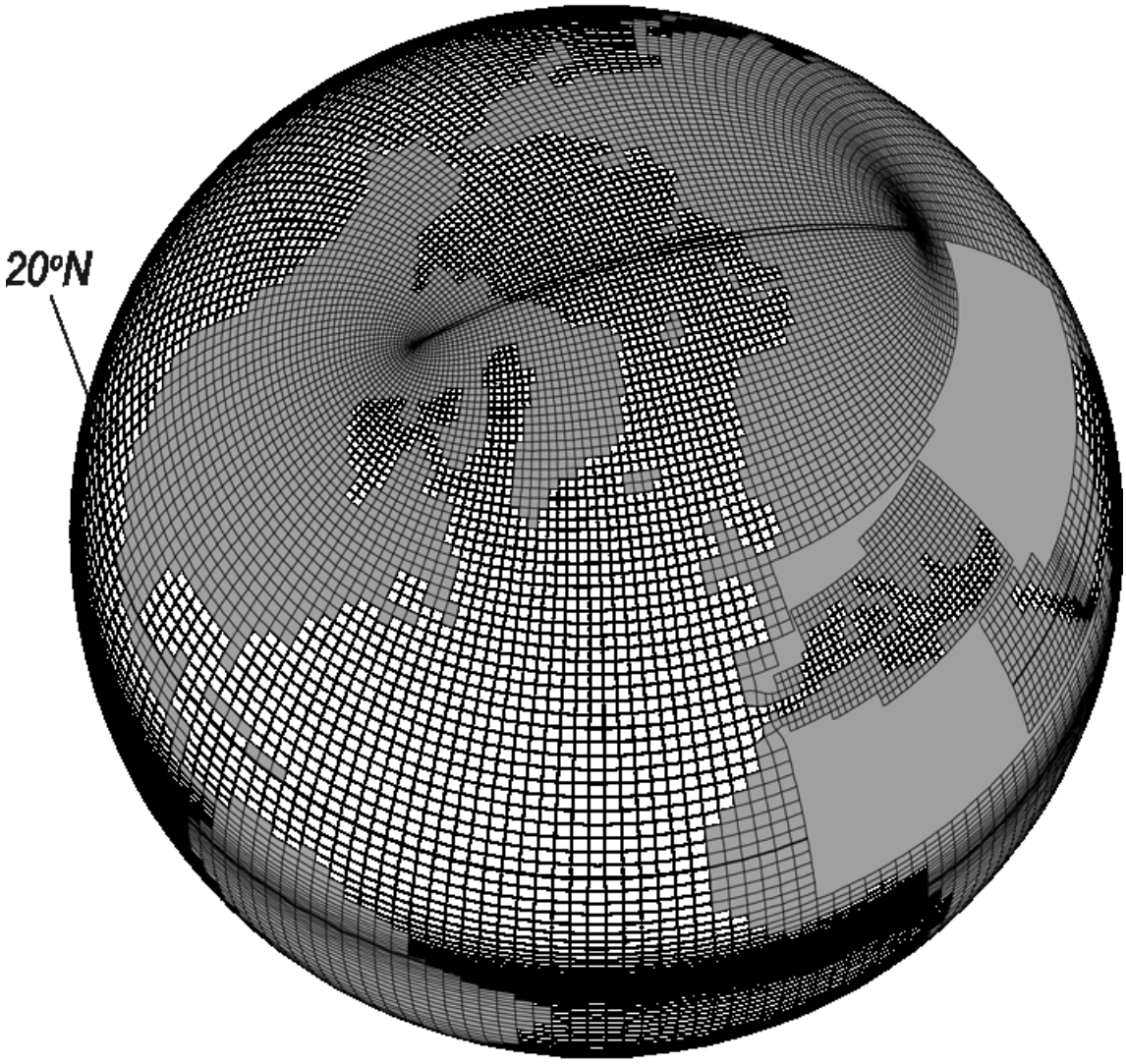


Figure 6.1: ORCA2 grid, adapted from http://www.elic.ucl.ac.be/textbook/glossary_g.xml.

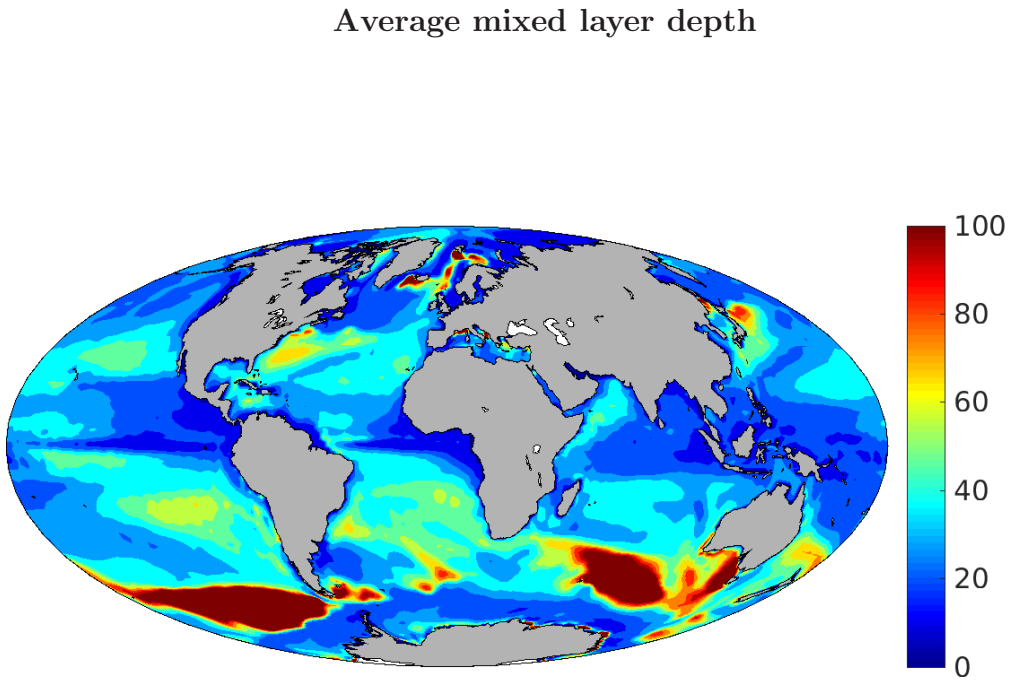


Figure 6.2: Yearly average of the mixed layer depth from a NEMO-LIM2 free run, in m.

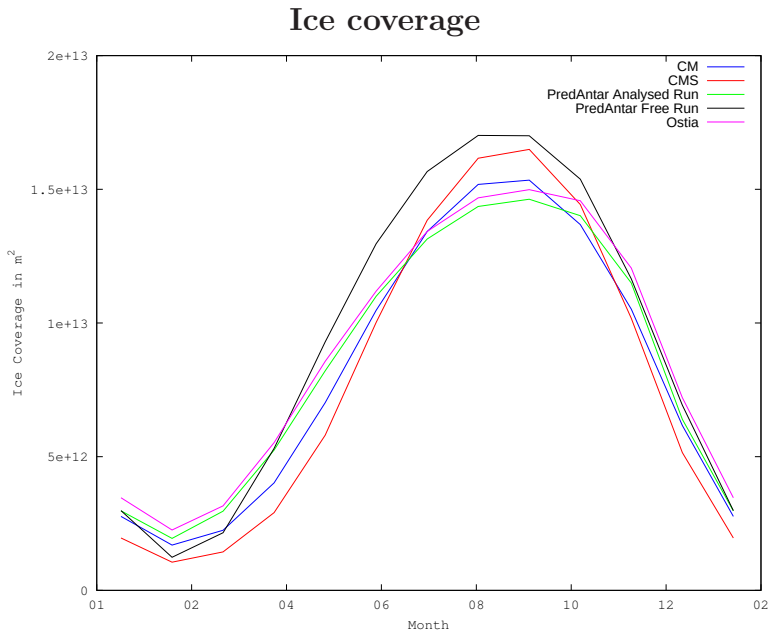


Figure 6.3: Mean monthly seasonal cycle of ice coverage (in m^2) for period 1985-2005.

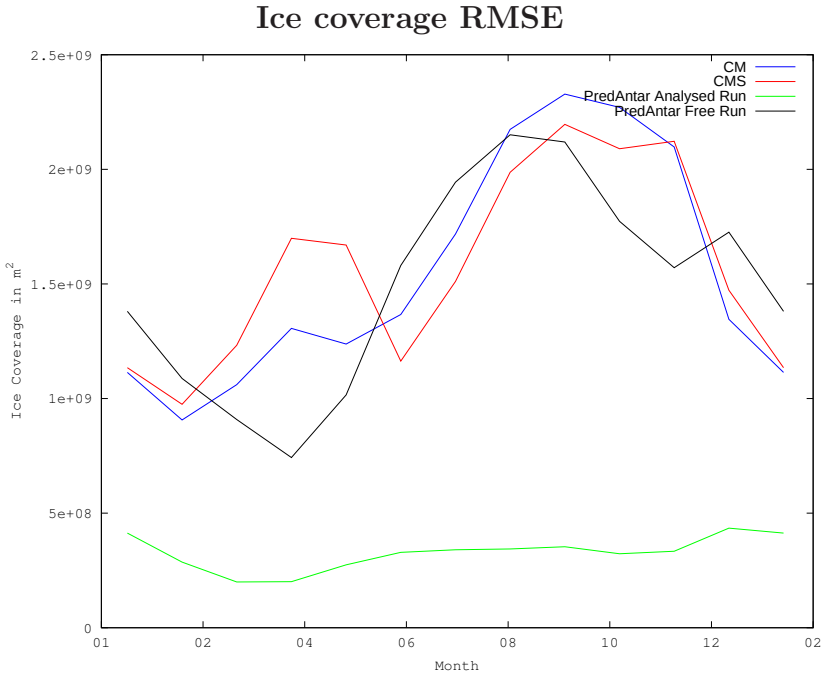


Figure 6.4: Mean monthly RMSE of the ice coverage (in m^2) for period 1985-2005.

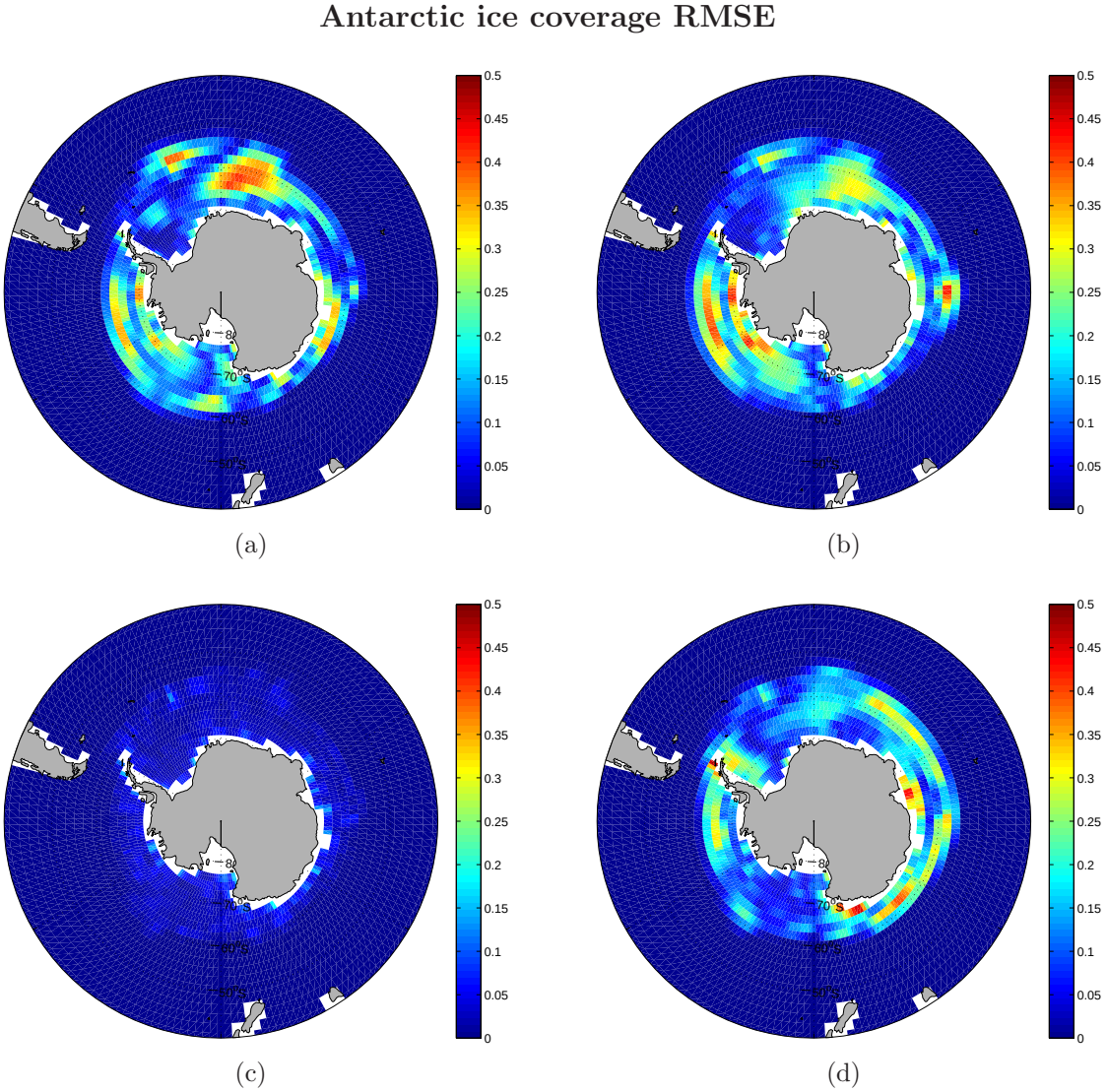


Figure 6.5: Mean RMSE of the ice coverage (in m^2) for period 1985-2005. Panel (a): CMCC-CM. Panel (b): CMCC-CMS. Panel (c): PredAntar analysed run. Panel (d): PredAntar free run.

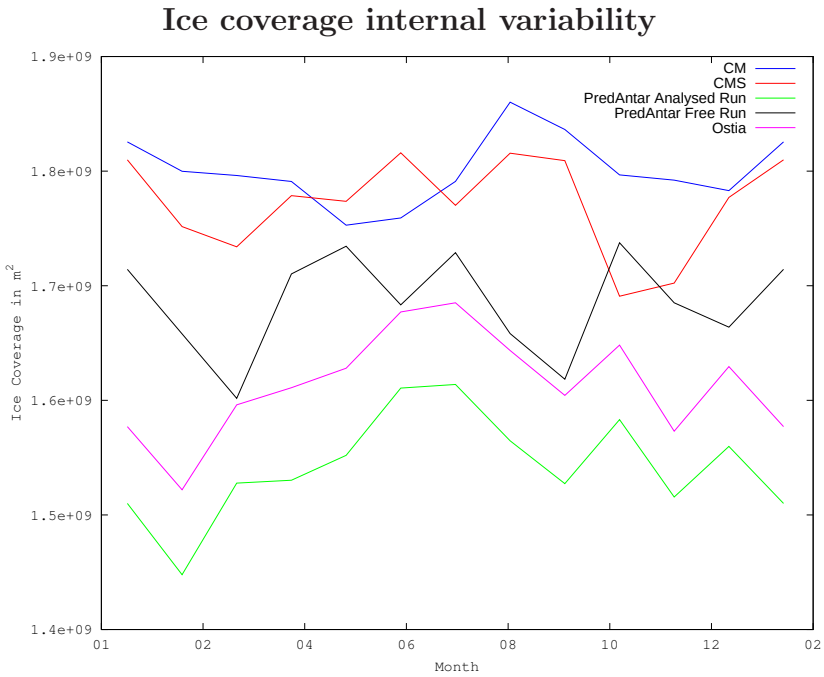


Figure 6.6: Mean monthly internal variability of ice coverage (in m^2) for period 1985-2005.

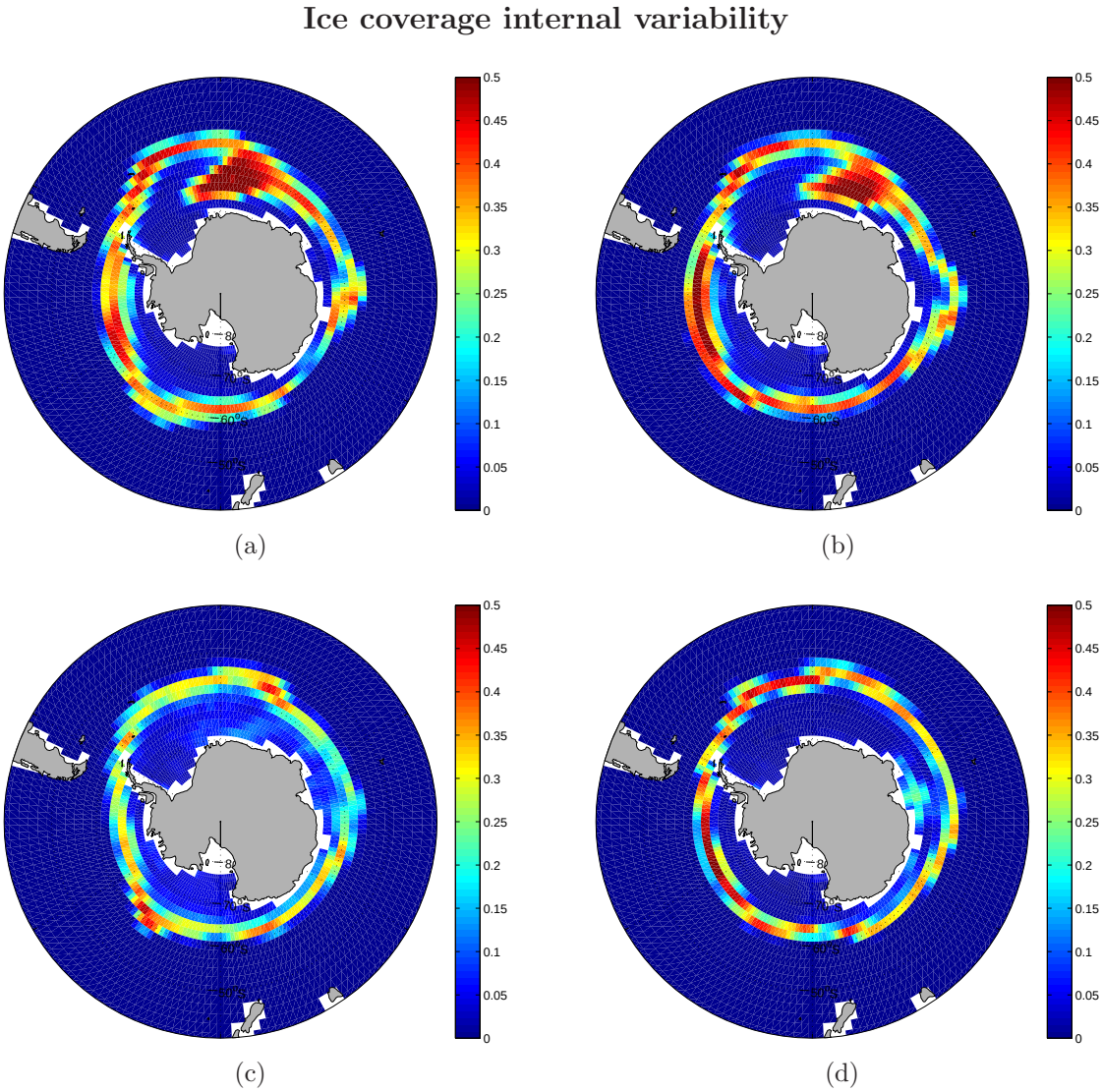


Figure 6.7: Spatial internal variability of ice coverage (in m^2) for September, 1985-2005. Panel (a): CMCC-CM. Panel (b): CMCC-CMS. Panel (c): PredAntar analysed run. Panel (d): PredAntar free run.

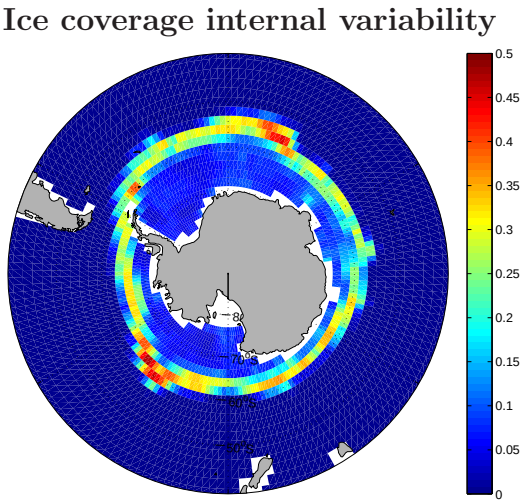


Figure 6.8: Spatial internal variability of ice coverage (in m^2) for September, 1985-2005.

Chapter 7

Twin experiment

Contents

7.1 Monovariate assimilation

The next step to test the efficiency of this method is to apply it to the realistic ocean model NEMO-LIM2. A twin experiment is performed, using a similar procedure to the one presented in the Lorenz '96 chapter.

7.1.1 Model variability

First, a random forcing is generated, with a correlation length of 5000 km. It is afterwards referred to as the truth or the reference forcing. The correlation length is chosen in order to be sufficiently large enough compared to the ORCA2 grid size (about 200 km at the equator). Longer correlation length (up to 10000 km) and shorter (down to 2000 km) were also tested, and provided different forcing structure

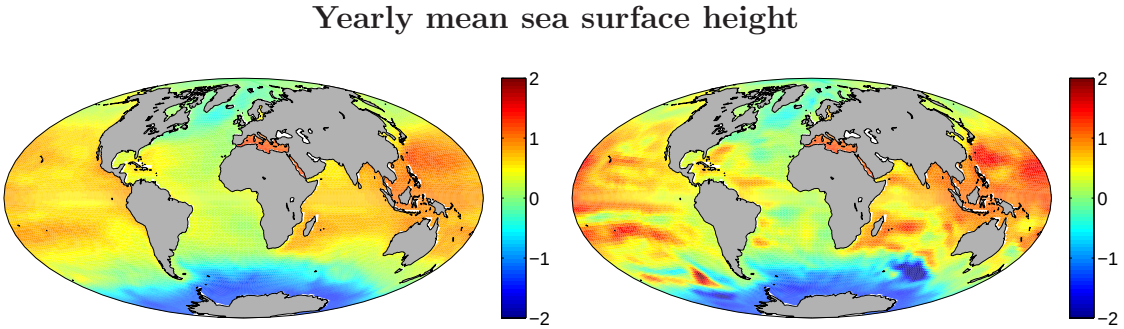
(but are not presented here). This reference forcing is then used with the NEMO-LIM2 model over a one year integration period.

Direct measurements of currents are currently too sparse. Although recently, an anomaly detected in the predicted signal of the median Doppler shift of radar echoes from ENVISAT Advanced Synthetic Aperture Radar (ASAR) have been investigated. Converted to surface Doppler velocity, this anomaly contains high-resolution information on surface currents. The combination of this Doppler signal with sea surface roughness measurements provides high resolution current fields (Chapron et al., 2005). Global climatologies are also available for use, such as Sudre et al. (2013), where sea level and wind stress satellite-derived measurements are combined to provide an estimate of surface currents.

For the realistic case (see chapter 8), real SSH fields representing time averages will be used. Due to the geoid problem, SSH altimetry data is represented as anomalies without any information about the mean state. If one would average SSH altimetry data, one would simply obtain zero (or a quantity close to zero). The mean dynamic topography is thus derived by other means, such as drifter and gravimetric measurements. Hence, the observations already represent an average. One can thus create the observations for the twin experiment by taking the mean SSH of the reference run over one year. When the average the model SSH is taken, the reduction in observational error due to this time averaging is already taken into account, since every ensemble member is averaged in time, causing short time-scale variability to be filtered out.

An ensemble of 100 random forcings is created and each of them is run separately. This produces an ensemble of yearly mean SSH. The Ocean Assimilation Kit (OAK), which contains an implemented ETKF scheme, is used for the analysis step (Barth et al., 2015). A local assimilation scheme option is used with an assimilation length equal to the correlation length of the perturbations (5000 km). The mean SSH from the reference run (figure 7.1b) are taken as the observations. The state vector (equation 4.1.11) consists of the ensemble of mean SSH (figure 7.1a), and is extended with its corresponding forcings

$$\mathbf{x}'' = \begin{bmatrix} \overline{SSH} \\ \widehat{F}_u \\ \widehat{F}_v \end{bmatrix}. \tag{7.1.1}$$



(a) (b)

Figure 7.1: (a) yearly mean sea surface height (SSH) of the ensemble mean runs (in m). The correlation length of the perturbation is 5000 km. (b) yearly mean sea surface height (SSH) of the twin experiment true run (in m).

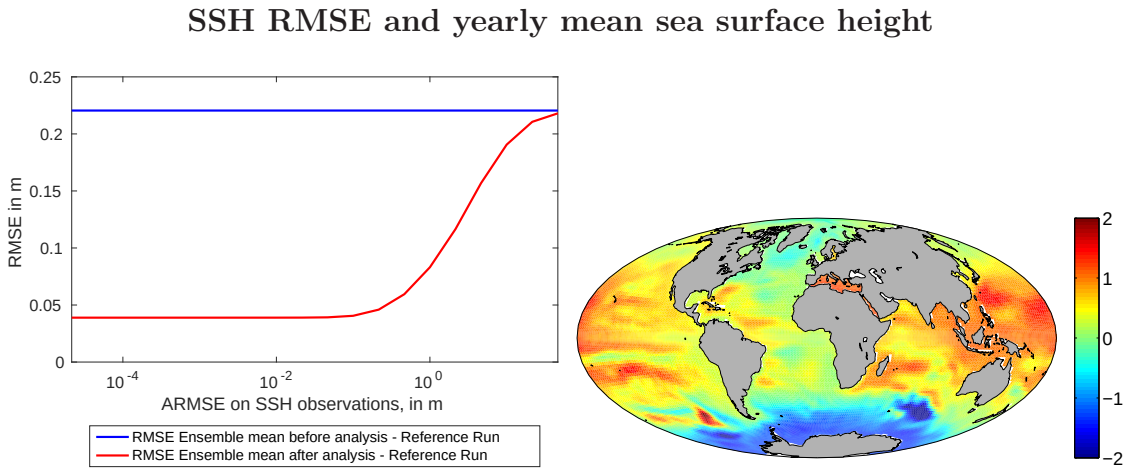
7.1.2 Error adjustment

Similarly to the Lorenz '96 case, one aims at finding the true forcing from the reference run. Noise is added to the observations, with a value representing 10% of the local SSH variability of the ensemble, in order to have strong noise signal in high variability areas, and low noise in low variability area. Aiming at this experiment to be as realistic as possible, the added noise is not taken into account for the observation error covariance matrix $\mathbf{R}$, which is estimated uniform over the domain. The assimilation is expected to provide a satisfying analysis if the relationship between F_u,F_v and the SSH can be captured by a linear covariance. Additionally, the observations used for the assimilation could contain redundancy. This is expressed by a redundancy factor $\alpha = \sqrt{r}$. It can be shown (Barth et al., 2007) that the error variance can be approximated through its multiplication by the number of redundant observations r : $\mathbf{R} = r\mu\mathbf{I}$, where μ is the error variance, and $\mathbf{I}$ the identity matrix. $\alpha RMSE$ is thus the square root of the diagonal of $\mathbf{R}$. Hereafter, $\alpha RMSE$ is refered to as the adjusted $RMSE$ ($ARMSE$). Also, all the model errors are not taken into account, which justifies the increase of the $ARMSE$.

The choice of the value of the error variance is critical. Indeed, in the case of

an underestimated error variance or a too small $ARMSE$, the analysis deteriorates unobserved variables due to the observations overconstraining the analysis. However, if overestimated with a too large $ARMSE$, the information contained in the observations is not sufficiently transferred into the model. This would not allow the assimilation scheme to apply a sufficiently large correction.

Therefore, the assimilation is performed with $ARMSE$ values between 10^{-5}m$< ARMSE < 10^{2}$m, in order to test the sensitivity and efficiency of the assimilation scheme (figure 7.2a). From figure 7.2a, one sees that $ARMSE \leq 4.6$ cm (x-axis) gives the lowest RMSE on the SSH (y-axis) for the assimilation. The corresponding analysed ensemble mean of yearly mean SSH is shown in figure 7.2b. When compared to figure 7.1a, one sees that the analysis is satisfactory and is able to retrieve the pattern of the reference run. The RMSE value of the ensemble mean before assimilation is 0.220 cm, whereas the analysis ranges from 0.039 cm up to 0.218 cm.



(a) (b)

Figure 7.2: (a) RMSE on SSH from Ensemble Mean before and after analysis, with True Run (in m). (b) Sea surface height of the ensemble mean after assimilation (in m).

7.1.3 Forcing field correction

However, this is only the first step of this procedure. The real objective is not the direct analysis of the ensemble SSH, but rather the analysis of the zonal and meridional forcings with which the state vector is augmented. Since one considers not to have any information about the true forcing, the initial background estimate

(or prior guess) of the forcing is zero. The analysis of the zonal and meridional currents are shown respectively in figure 7.3a and figure 7.3c, and must be compared to the true forcing in figure 7.3b and figure 7.3d. One notes that the analysed forcings are convincingly reproducing the structure of the true forcings that are sought to be found.

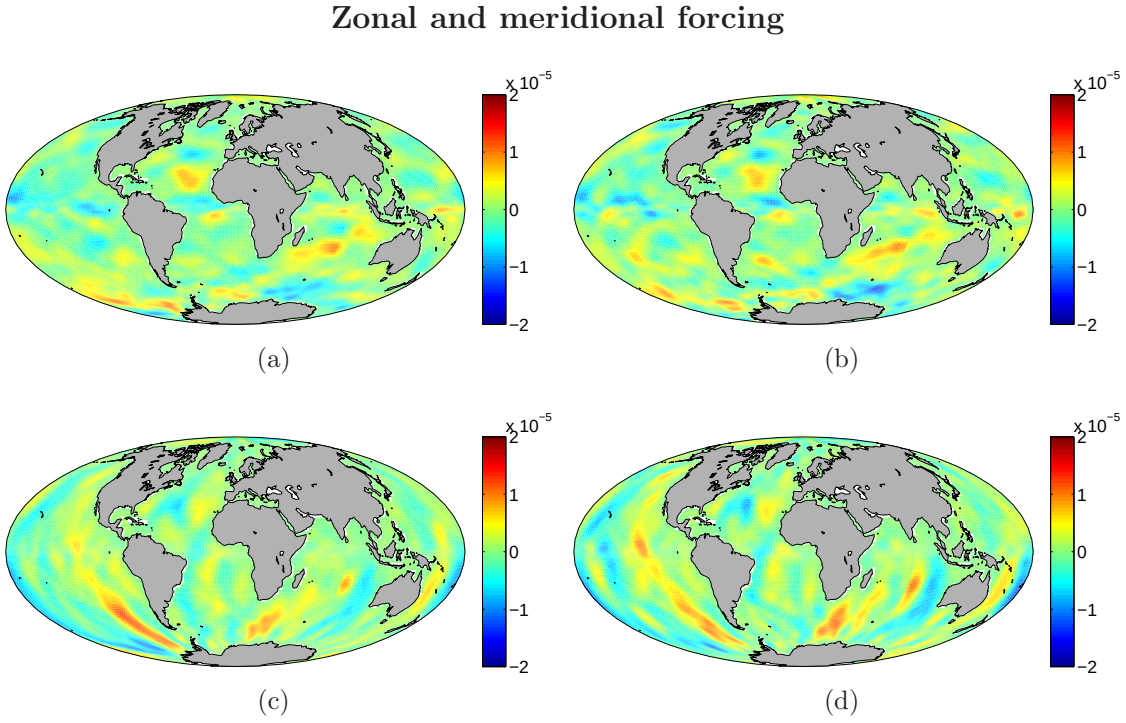
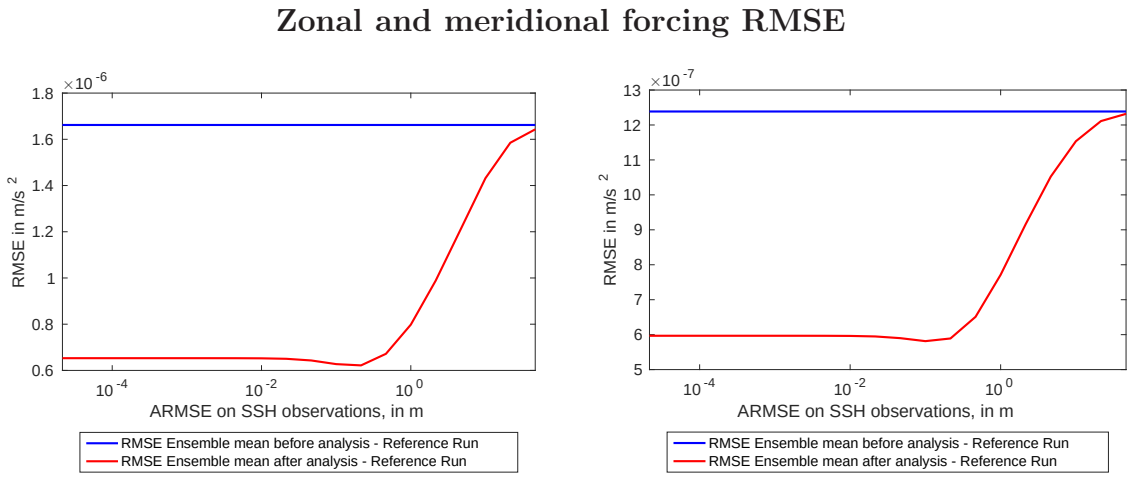


Figure 7.3: (a) Zonal Forcing ensemble mean after analysis (in ms^{-2}). (b) Zonal Forcing from the true run (in ms^{-2}). (c) Meridional Forcing ensemble mean after analysis (in ms^{-2}). (d) Meridional Forcing from the true run (in ms^{-2}).

Using the twin experiment, and the perfect knowledge that one has on the reference run, one can also look at the RMSE between the analysed forcings and the reference run. This is shown in figure 7.4a and figure 7.4b for the zonal and meridional forcings respectively, with different $ARMSE$ on the SSH observations. One can see that the choice of $ARMSE = 4.6$ cm on the observations, which corresponds to the small dent, gives the best possible results. Since this choice is made solely based on the efficiency of the SSH analysis, the relationship between the forcings and the yearly mean SSH of the model can be considered to be strong enough for this experiment. Interestingly, one can interpret the small dent in the RMSE of the the forcings as the overconstraints imposed by the observations on the SSH with values of $ARMSE < 4.6$ cm. The exact lowest RMSE values are 6.27×10^{-7} ms^{-2} and 5.81×10^{-7} ms^{-2} for the zonal and meridional forcings respectively, whereas the

RMSE values with the ensemble forecast are 1.66×10^{-6} ms^{-2} and 1.24×10^{-6} ms^{-2} respectively.



(a) (b)

Figure 7.4: (a) RMSE on Zonal Forcing from Ensemble mean before and after Analysis, with True Run (in ms^{-2}). (b) RMSE on Meridional Forcing from Ensemble mean before and after Analysis, with True Run (in ms^{-2}).

One can also obtain the total analysed forcing by combining the zonal and meridional components into a vector form in figure 7.5. One can the compare this with the geostrophic currents derived from the SSH bias between the twin experiment reference run and the free model run in figure 7.6. Because of the nongeostrophic balance near the equator, where the horizontal Coriolis force tends to zero, a 5° region around the equator has been removed for this comparison. One can see on figure 7.6 that the geostrophic current derived from the SSH bias is not directly linked to the reference forcing from figure 7.5. This stems from the fact that the forcing affects the model globally, whereas the geostrophic current has a more local origin.

7.1.4 Model rerun

The last step to take is to rerun the model with the estimated bias correction term. The forcing from the reference run is considered as the source of the bias acting on the model, and the analysed forcings from the assimilation as the bias correction term to apply to the model. The model is rerun a single time with the analysed ensemble mean forcing, which corresponds to the analysed bias estimator $\widehat{\mathbf{b}}$ from

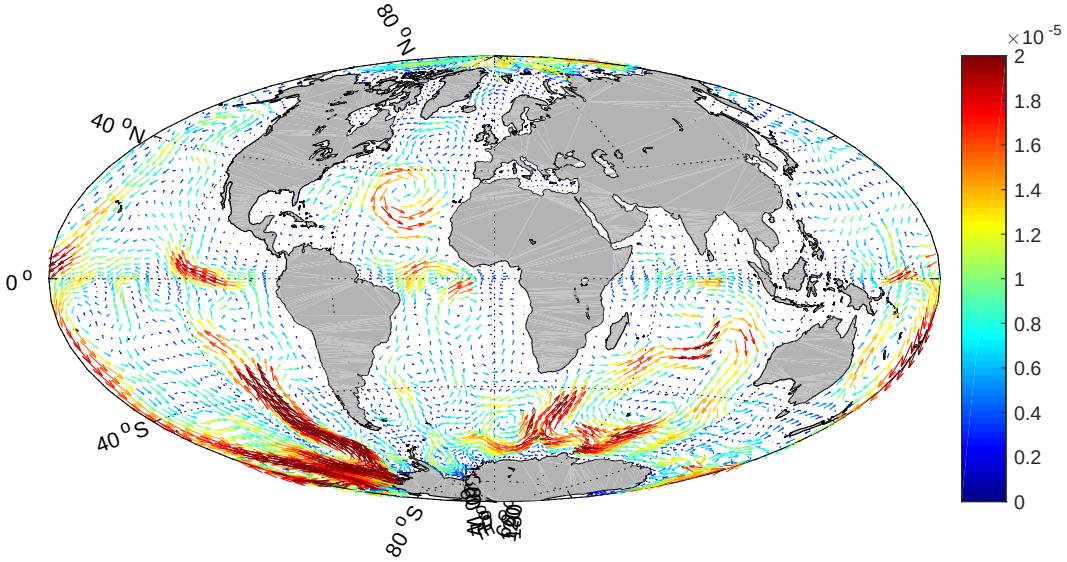


Figure 7.5: Total forcing ensemble mean after analysis (in ms^{-2}).

equation (4.1.14). Without this correction, the model free run without any forcing would be biased. The result of the model rerun with bias correction is shown in figure 7.7a, and can be compared with the true run, displayed in figure 7.1b. Analogously to the Lorenz '96 case, figure 7.7a is not the result of the assimilation of observations from the true run. It is the rerun of the model with the analysed forcing, obtained from the augmented state vector used during the assimilation procedure. The rerun with bias correction is able to reproduce patterns in the SSH that are particular to the reference run, produced by the true forcing. The last validation of the bias correction term forcing the model is shown in figure 7.7b, where the RMSE on the SSH between the rerun of the model and the true run is compared to the initial ensemble mean and the analysis. One can note that a significant part of the model bias has been removed. The lowest RMSE values of the rerun is 0.064 cm.

7.1.5 SST and SSS validation

Further validation of this procedure is done by the comparison of the model forced rerun with the reference run on independent variables. Sea surface temperature

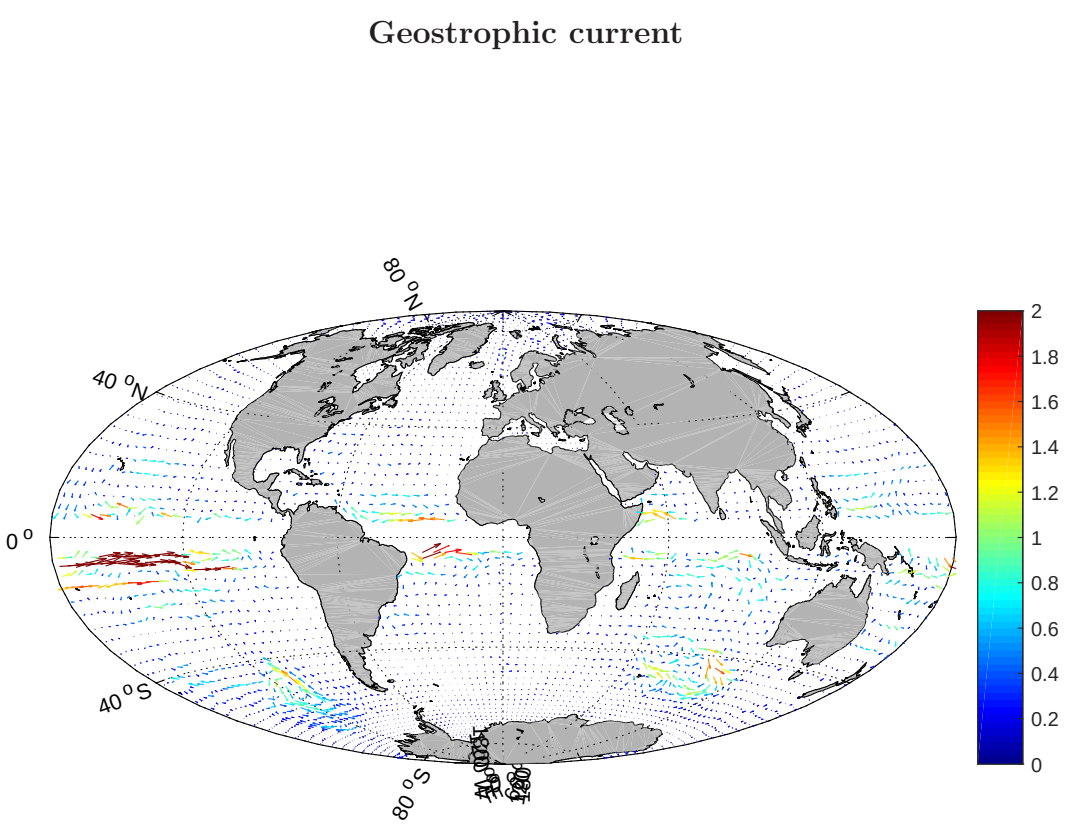
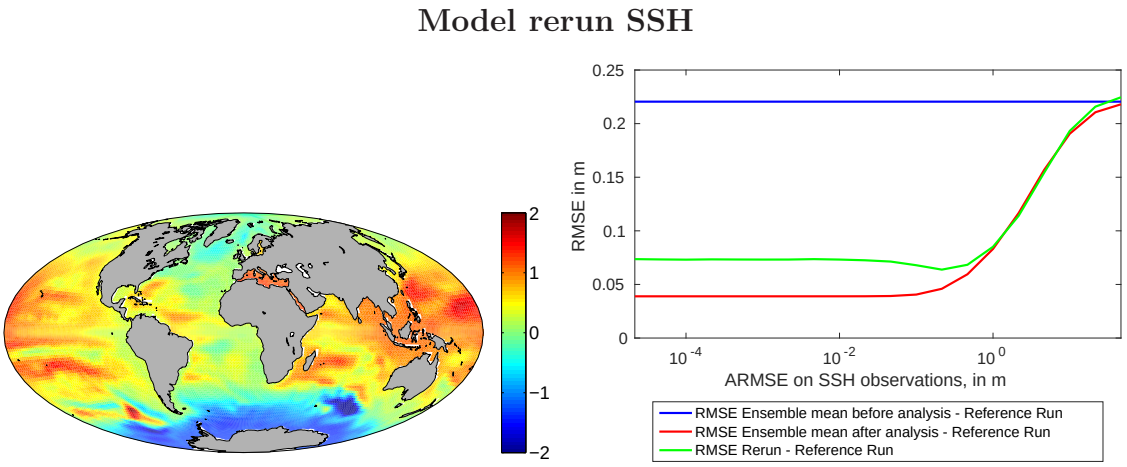


Figure 7.6: Geostrophic current derived from the SSH bias between the twin experiment reference run and the model free run (in ms^{-1}).

(SST) and salinity (SSS) are chosen for their relationship to the currents in the ocean through specific mixing and redistribution of salinity and heat in the ocean. The bias on the currents that this method aims to correct has a direct effect on the SST and SSS. The yearly average SST is shown in figure 7.8a for the ensemble mean, in figure 7.8d for the reference run, and in figure 7.8b for the model rerun with analysed forcing. Figure 7.9a, figure 7.9d and figure 7.9b show the SSS for the same runs respectively.

It is clear that typical structures on the SST and SSS fields from the reference run are reproduced by the rerun, and are completely absent on the ensemble mean. One can also note from figure 7.8c and figure 7.9c that the RMSE on the SST and SSS shows a similar behaviour to the RMSE on SSH from figure 7.7b. However, whereas there is a systematic improvement on the SSH reruns with analysed forcings, the analysed forcings appear to be deteriorating the SST and SSS for a specific set of parameters, in particular when the $ARMSE$ on the SSH is large. The lowest RMSE values for the rerun are 0.537 C° and 0.155 PSU respectively.



(a) (b)

Figure 7.7: (a) Sea surface height (SSH) of the rerun with analysed forcing (in m). (b) RMSE on SSH from Ensemble Mean before and after analysis, and Rerun, with True Run (in m)

7.2 Multivariate assimilation

The monovariate assimilation has proven to be an efficient tool to recover the reference forcing fields in a classic twin experiment. However, one can consider using more observations to see if the results can be improved. In particular, the improvement of the SST of the model can be considered as a biproduct and one can also include observations on the SST to incorporate even more constraints on the assimilation.

Similarly to the previous experiment, the state vector (equation 4.1.11) consists of the ensemble of mean SSH, mean SST, and is extended with its corresponding forcings

$$\mathbf{x}'' = \begin{bmatrix} \overline{SSH} \\ \overline{SST} \\ \widehat{F}_u \\ \widehat{F}_v \end{bmatrix}. \tag{7.2.1}$$

To see if an improvement can be obtained, the optimal parameters of the previous experiment for the ARMSE on the SSH observations are taken ($ARMSE = 4.6$

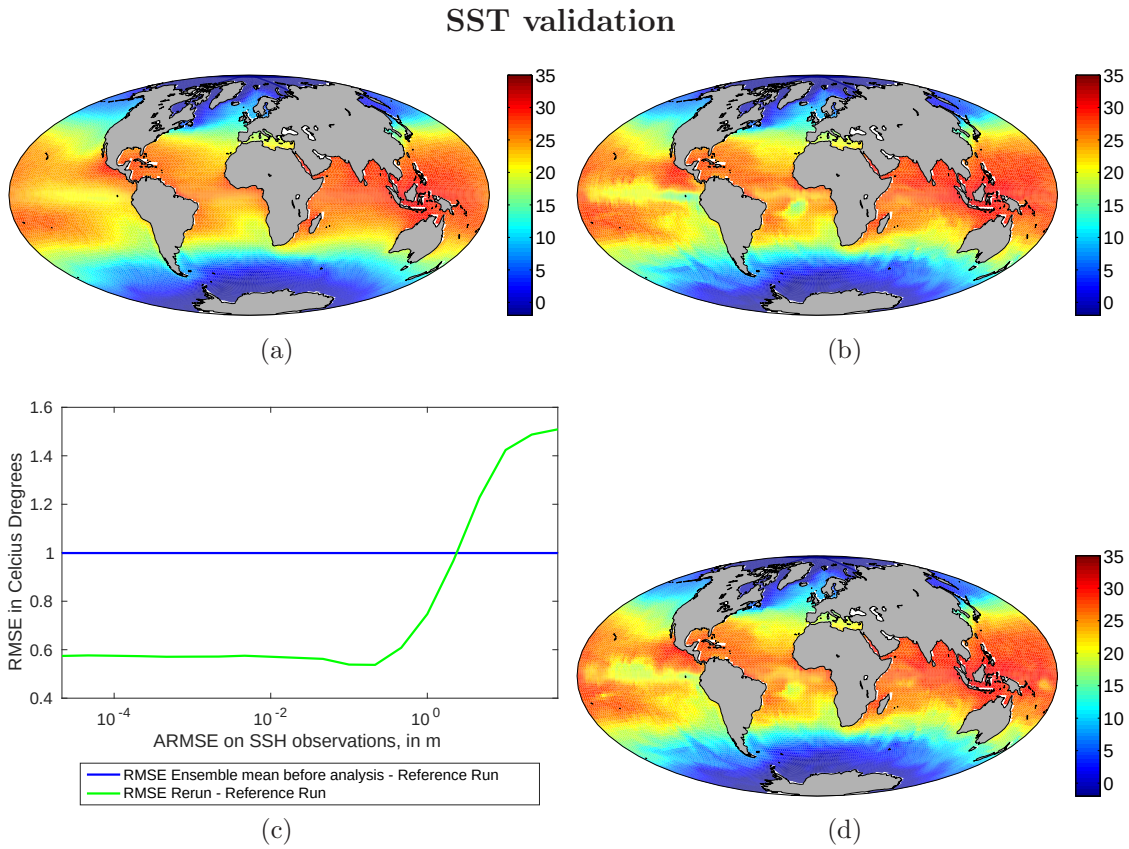


Figure 7.8: (a) Yearly mean sea surface temperature (SST) of the ensemble mean (in C°). (b) Sea surface temperature (SST) of the rerun with analysed forcing (in C°). (c) RMSE on SST from Ensemble Mean after analysis, and Rerun, with True Run (in C°). (d) Yearly mean sea surface temperature (SST) of the twin experiment true run (in C°).

cm). This will provide an objective to attain and exceed. The ARMSE on the SST ranges over 10^{-5} C° $< ARMSE < 10^{2}$ C°. The results of this multivariate assimilation twin experiment are shown on Figs. (7.10a) to (7.10c), and the exact RMSE values are given on table 7.1. Empty values are not relevant to the experiment. For instance, the SSH, SST and SSS values are absent from the monovariate analysis, and only the SSH is analysed in the multivariate experiment.

For the RMSE on the meridional and zonal forcings (Figs. (7.10a) and (7.10b)) respectively, one can note that the analysis results deteriorate with a too low ARMSE on the SST observations, causing the analysis to be overconstrained. However, when the $ARMSE = 1$ C°, the assimilation is able to provide a better estimation of the forcing terms. Finally, when the $ARMSE$ grows larger, the assimilation tends towards the monovariate analysis, as expected.

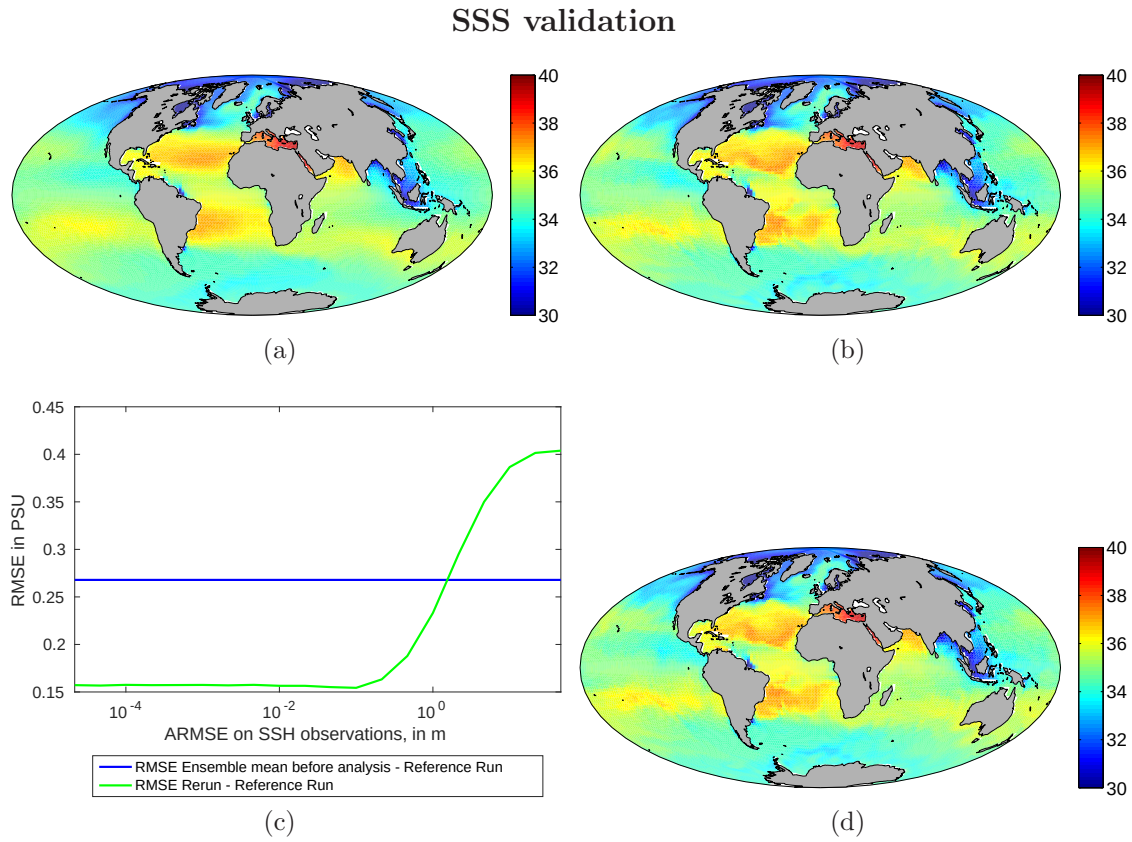


Figure 7.9: (a) Yearly mean sea surface salinity of the ensemble mean (in PSU). (b) Sea surface salinity of the rerun with analysed forcing (in PSU). (c) RMSE on sea surface salinity from Ensemble Mean after analysis, and Rerun, with True Run (in PSU). (d) Yearly mean sea surface salinity of the twin experiment true run (in PSU).

Figure 7.10a represents the RMSE on the SSH observations, before and after analysis for the multivariate assimilation, and the RMSE of the monovariate and multivariate assimilation reruns. It is interesting to note that, similarly to the previous experiment, the rerun shows a slightly larger RMSE than the analysis. However, an improvement is still shown, compared to the monovariate assimilation.

One can also look at the results of this multivariate assimilation experiment on the SSS and SST RMSE. Those are shown on Figs. (7.11a) and (7.11b) respectively. The behaviour of the RMSE on the SST is similar to that of the meridional and zonal forcings. A slight improvement is to be noted on the rerun, when one compares the monovariate assimilation to the multivariate assimilation. However, the SSS case is much more interesting. Whereas the previous improvements only

represented a fraction of the initial correction, the SSS shows an RMSE reduction which is comparable to the initial assimilation correction.

One can conclude that the SST and the meridional and zonal forcings are directly linked through the mixing of surface waters. The addition of SST observations does slightly improves the assimilation results and the model corrected rerun. However, the multivariate assimilation has a much larger impact on the SSS RMSE rerun.

Variable name	Forecast	Monovariate		Multivariate	
		analysis	rerun	analysis	rerun
Zonal forcing in ms^{-2}	1.66×10^{-6}	6.27×10^{-7}		5.96×10^{-7}	
Meridional forcing in ms^{-2}	1.24×10^{-6}	5.81×10^{-7}		5.45×10^{-7}	
SSH in m	0.220		0.068	0.0457	0.061
SST in C°	0.999		0.539		0.509
SSS in PSU	0.268		0.197		0.150

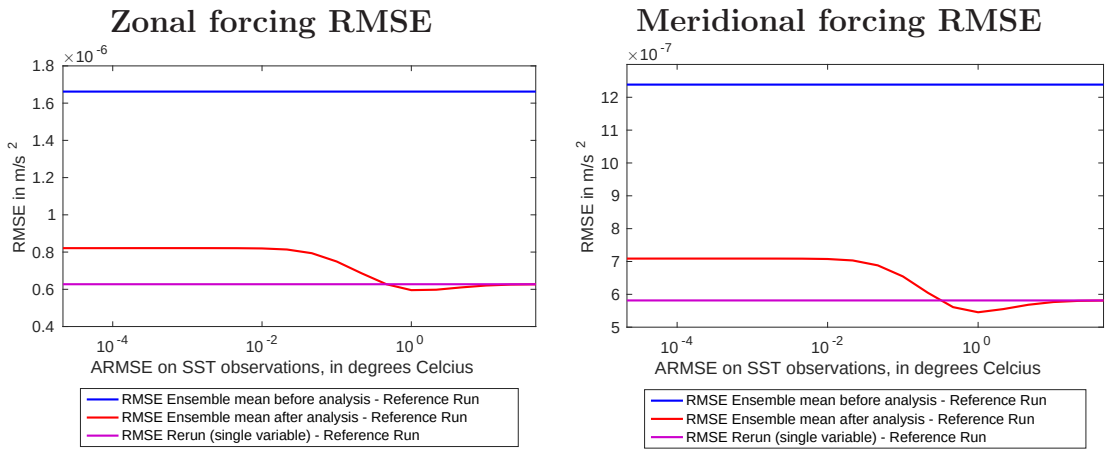
Table 7.1: RMSE values of the multivariate rerun for a $ARMSE = 1$ C° value, compared to the monovariate assimilation. Empty values are not relevant.

7.3 Conclusion

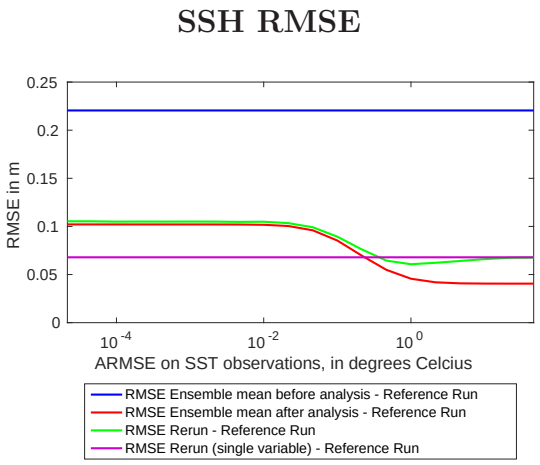
The twin experiments performed in this chapter have allowed to effectively implement and evaluate the effectiveness and feasibility of the bias correction method on a realistic model. The spatial structure of the reference forcing term is successfully estimated by the ETKF scheme. The bias corrected model rerun shows a significant improvement on the average SSH of the model when compared to the reference run. The multivariate assimilation twin experiment shows that further improvements can be obtained when more information are available. Specifically, observations on other variables, such as the SST, allow the estimation of the bias correction term to be more accurate. The constraints imposed by the SST on the estimation also impacts other unobserved variables such as the SSS.

Although choices concerning the set up of this twin experiment aim at performing a realistic trial of the bias correction method, one must keep in mind that twin experiment are generally too optimistic. Clearly, the generation of the bias correction term is exactly the same for the ensemble as for the reference run. From this follows that the PDF described by the ensemble is ensured to contain the reference perturbation. Moreover, one is sure that all physical processes of the reference model

run are accurately represented by the ensemble. Nonrepresented physical processes mentioned as a source of error (hence bias) of numerical models are absent from this twin experiment. One can only expect a real experiment to be less efficient, though the method clearly shows that potential improvements are obtainable.

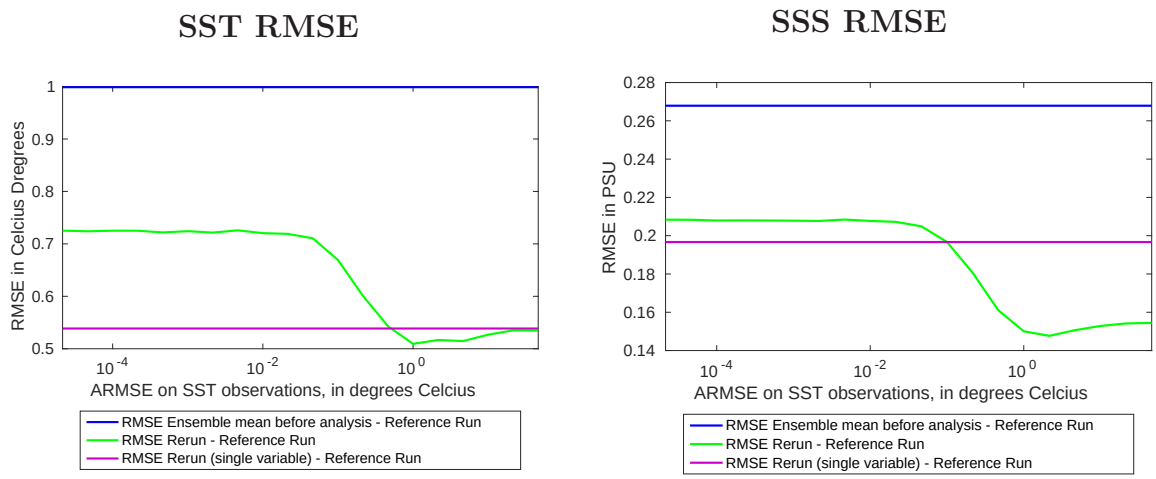


(a) (b)



(c)

Figure 7.10: (a) RMSE on Zonal Forcing from Ensemble mean before and after multivariate analysis, and monovariate analysis, with True Run (in ms^{-2}). (b) RMSE on Meridional Forcing from Ensemble mean before and after multivariate analysis, and monovariate analysis, with True Run (in ms^{-2}). (c) RMSE on SSH from Ensemble Mean before and after multivariate analysis, multivariate analysis rerun , and monovariate analysis rerun, with True Run (in m).



(a) (b)

Figure 7.11: (a) RMSE on SST from Ensemble Mean before multivariate analysis, from multivariate analysis rerun, and monovariate analysis rerun, with True Run (in C°). (b) RMSE on sea surface salinity from Ensemble Mean before multivariate analysis, from multivariate analysis rerun, and monovariate analysis rerun, with True Run (in PSU).

Chapter 8

Realistic case

Contents

8.1 Single assimilation

The efficiency of this bias correction method has been successfully tested on a twin experiment test case in the previous chapter. The following covers the results of this method in a realistic case experiment.

The same setup as the twin experiment is taken for the NEMO model configuration. Observations are however taken from the mean dynamic topography (MDT) of CNES (Centre National d'Etudes Spatiales) (Rio et al., 2011). The SSH provided by the MDT of CNES is interpolated on the ORCA2 grid. Again, an ensemble of forced model runs is created. The observations are assimilated with a range of RMSE fields to find the best compromise between the ensemble and the observations. This procedure provides a forcing which is used to rerun the model. The

same parameters as for the twin experiment are taken: a correlation length of 5000 km and 100 ensemble members.

The different relevant RMSE are shown in figure 8.1. One can notice that the RMSE between the ensemble mean and ensemble members shows a sufficient enough variability on the model to cover the RMSE between the model free run and the CNES-MDT observations. Like in the previous chapter, the RMSE of the analysed SSH field is significantly reduced compared to the RMSE between the ensemble mean before analysis and the CNES-MDT observations. Finally, the rerun of the model with the assimilated forcing shows a significant improvement on the SSH RMSE when compared to the free run. This means that the analysed forcing effectively removes a part of the error of the model on the SSH, through the forcing on the zonal and meridional currents. The lowest RMSE obtained for the rerun is 0.155 cm, for an ARMSE value of $ARMSE \leq 21$ cm (x-axis). The relevant corresponding RMSE values are 0.197 cm for the ensemble mean forecast, 0.098 cm for the analysed ensemble mean, and 0.193 cm for the model free run.

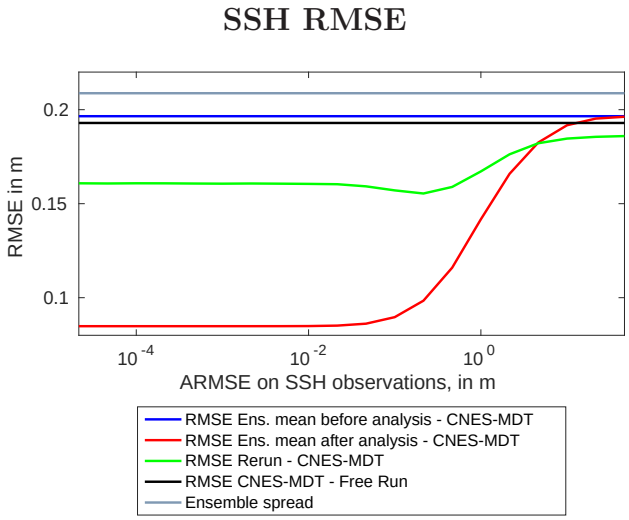


Figure 8.1: RMSE on SSH from the ensemble mean before and after analysis with CNES-MDT observations, from the forced rerun with the observations, from the model free run with the observations, and the internal variability of the ensemble (in m).

8.1.1 Global SSH

More extensive results are shown in the following figures. Figure 8.2a shows the interpolated yearly mean SSH of the CNES-MDT observations on the ORCA2 grid.

Figure 8.2b show the yearly mean SSH of the model free run, for the year 1984-1985. in figure 8.2c, the yearly mean SSH of the ensemble mean is shown. One can notice the differences between the model free run and the ensemble mean of forced runs on the yearly mean SSH. This is due to the fact that, even though the ensemble of zonal and meridional forcings has a close to zero mean, the presence of those forcings do increase the currents in the ocean, producing a nonzero mean SSH modification. Finally, 8.2d shows the yearly mean SSH of the rerun with the analysed forcing.

When comparing figures 8.2a, 8.2b and 8.2d, one can notice the differences on the SSH between the observations, the free model run and the forced rerun. The SSH of the model free run appears to be very smooth and does not show the same variability as the CNES-MDT observations. This property, directly influenced by strong, localised, currents, shows to be improved in the forced rerun. In particular, the SSH variations caused by the Gulf Stream are absent from the free run but present in the forced run. Other similar improvements are present around the Cape of Good Hope and along the coast of Chili.

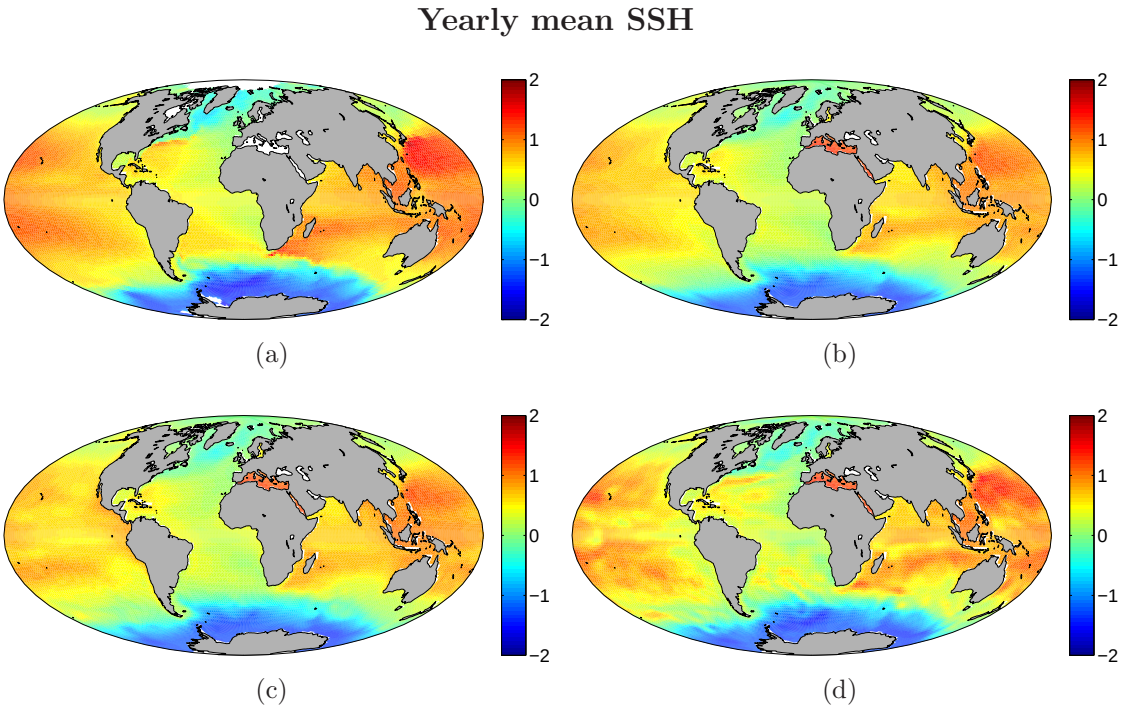


Figure 8.2: Yearly mean SSH (in m) of (a) CNES-MDT observations, (b) model free run, (c) ensemble mean forecast, (d) lowest RMSE model forced rerun.

Average error

Yearly averaged SSH errors relative to the CNES-MDT observations are shown on figures 8.3a to 8.3d. They represent the considered data minus the CNES-MDT observations. The initial model free run SSH average error shown on figure 8.3a shows specific structures following strong oceanic currents, such as the Gulf Stream, the Kuroshio current, or the ACC. The forecast ensemble mean logically exhibits similar patterns on figure 8.3b. The SSH average error of the analysis on figure 8.3c is clearly reduced compared to the previous figures. One must take into account that the analysis is the result of a local assimilation producing a linear combination of ensemble members through the state vector composed by the forcing and yearly averaged SSH. It is thus to be expected that the analysis is able to significantly reduce the global SSH average error. After rerunning the model, the SSH average error of the rerun on figure 8.3d shows smaller average errors structures. This can be compared to the smaller spatial structures shown on figure 8.2d. In particular, improvements can be noted in strong oceanic current regions. This stems from the spatial structure of the forcings, whose correlation length is 5000 km. Hence, the small scale of the correction results in the introduction of small scale structures in the model rerun.

It is worth noting that for the lowest RMSE rerun (figure 8.3d), are the two large errors in the southern hemisphere, located in the south of the Indian and Pacific Ocean. In particular, the model rerun shows a larger average error in those two localised regions, reflecting the deterioration of the SSH due to the analysed forcings. They are caused by the limitations of the yearly averaged mixed layer depth, which is deeper in those two particular locations as shown on figure 6.2. This causes the forcing to produce unrealistic corrections causing the two strong abnormal signals. This pattern was not noted on the twin experiment run, since both the reference run and the ensemble display this particular pattern.

8.1.2 Analysis confidence

To estimate the confidence of the analysis, one can look at the spread of the ensemble before and after assimilation. One can interpret the analysed ensemble spread as a measure of confidence of the analysis. In regions where the spread is low, the higher the confidence in the analysis is. In regions where the spread is large, the lower the confidence is.

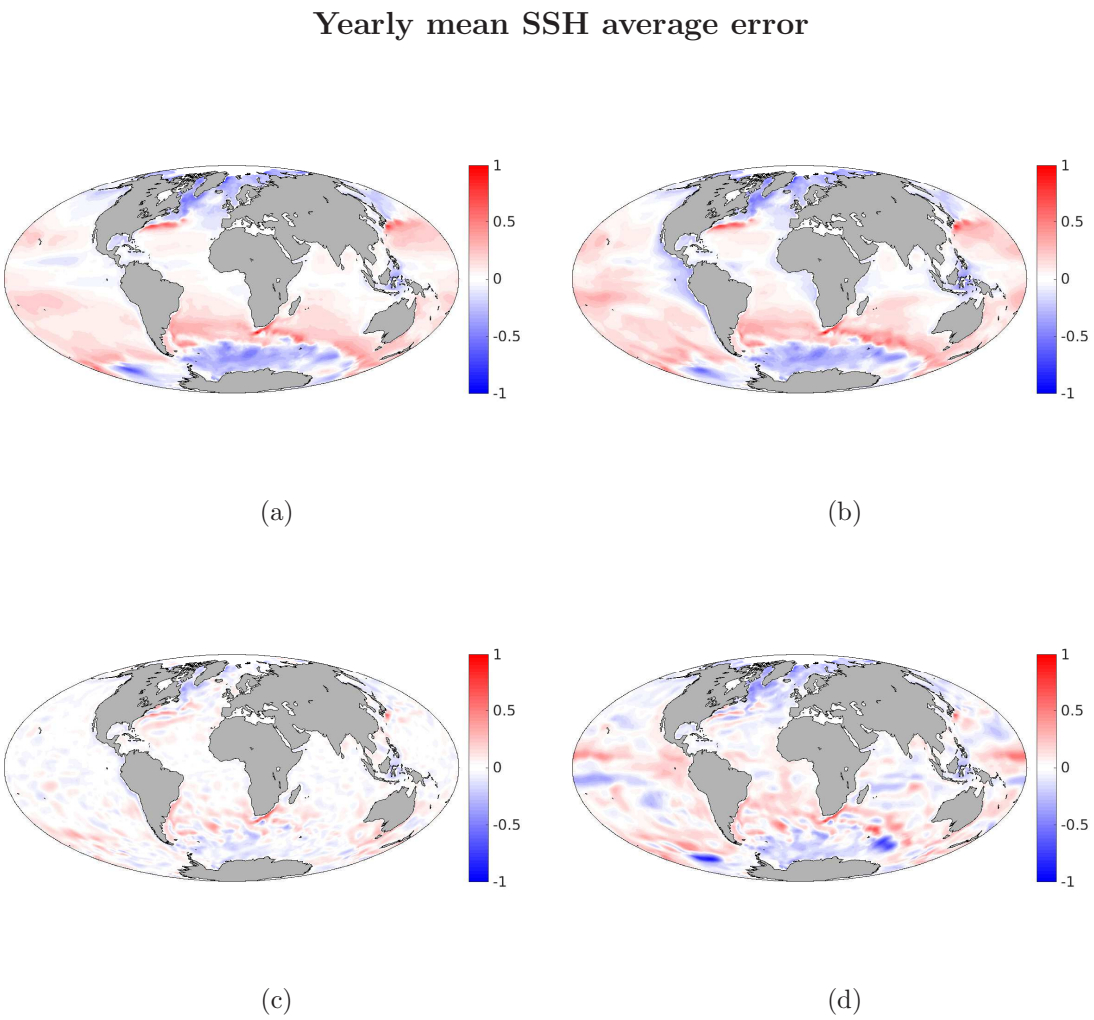


Figure 8.3: Yearly mean SSH average errors with the CNES-MDT observations (in m) of (a) free model run, (b) ensemble mean forecast, (c) ensemble mean after analysis, (d) lowest RMSE model forced rerun .

Figures 8.4a to 8.4d show the ensemble standard deviation of the meridional and zonal forcing fields, before and after the assimilation of the CNES-MDT observations. As explained in section 6.4.3, the initial forcing term is built in order to provide a large enough ensemble spread on the SSH to cover the observations in this realistic experiment. One can note the large ensemble spreads in the forecasts, aimed at producing the adequate SSH spreads to contain the observations. The spread is generally larger in wide oceanic regions, in particular in the Pacific. This is due to the choice of a 5000 km correlation length, which constrains the forcing term generation.

The analysed ensemble spread is also interesting. One can compare figures 8.4b and 8.4d with figure 8.6, which represents the ensemble mean forcing used to rerun the model. The global structure of the zonal and meridional forcing ensemble spread are similar. The analysis clearly reduces the ensemble spread over the whole globe. However the lack of observations on and around the North Pole does not provide the analysis with enough information to reduce the spread to values similar to other regions.

The strong currents created by the analysis and represented on figure 8.6 are mostly located in larger spread areas for both the meridional and zonal forcings (figures 8.4b and 8.4d). In particular, in areas along the equator, the Antarctic Circumpolar current, or the Gulf Stream.

The ensemble SSH spreads before and after analysis are shown on figure 8.5a and 8.5b respectively. One can observe that the spread before assimilation is much larger than after assimilation. In addition, the large spread regions provide information about how the model react to the stochastic forcing. In particular, in the south of the Indian and Pacific Ocean, the ensemble spread rises up to 1.2 m. As mentioned before, this reflects the unrealistic deeper average mixed layer depth, which deteriorates the model in those locations.

After assimilation, the spread decreases drastically, with values up to 4 cm. One can note that, similarly to the zonal and meridional forcings spread, the SSH spread along the equator is notably low. The initial assumption on the forcing generation is to avoid artificial surface gravity waves. For large scale perturbations outside the tropics, the flow is dominated by the geostrophic equilibrium. Hence, the divergence of the flow must be close to zero. However, the geostrophic equilibrium is not valid along the equator. This could lead to other forcing generation mechanisms along the equator.

One can also note a larger spread on the western coasts of land masses. This is clearly visible over both North and South America, along the coast of Madagascar and the south of the Arabian Peninsula, and in the Pacific Ocean. This reflects a general western intensification of boundary currents (Carton et al., 2000a).

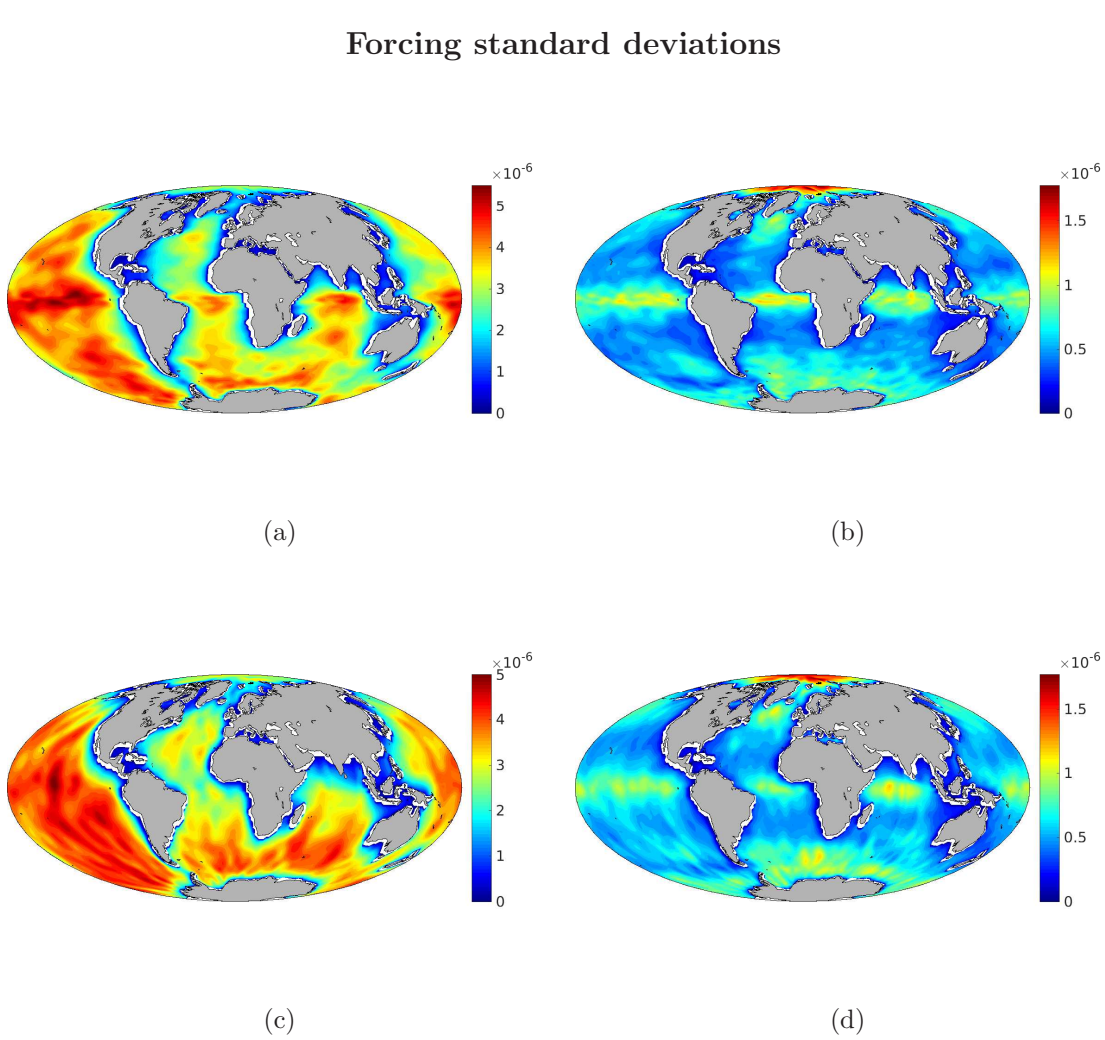
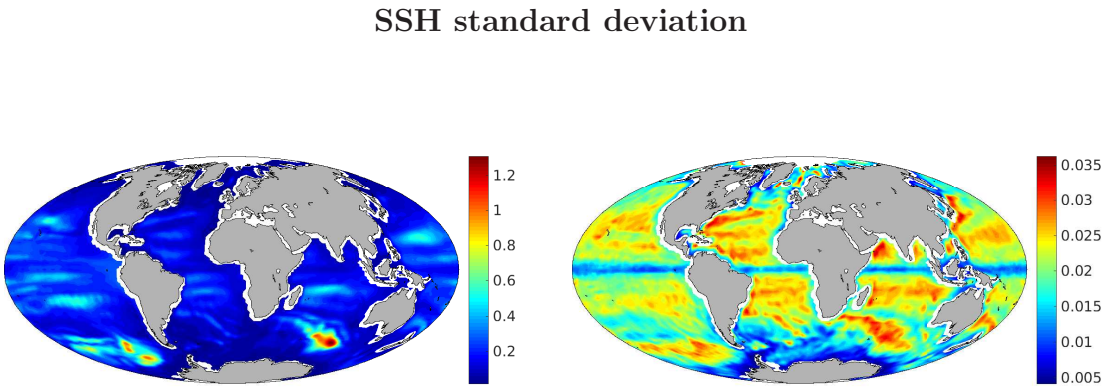


Figure 8.4: Ensemble standard deviations (in ms^{-2}) of: (a) Zonal forcing before assimilation. (b) Zonal forcing after assimilation. (c) Meridional forcing before assimilation. (d) Meridional forcing after assimilation.

8.1.3 Final correction

The final forcing field produced by this procedure is shown in figure 8.6, in vector form. It is the optimal forcing resulting from the analysis with the CNES-MDT SSH observations, applied to the rerun of the NEMO model, a single time, producing the rerun SSH field from figure 8.2d. This can be compared to a global map of the real currents, displayed on figure 8.7. One must remember that even though the initial perturbations did contain some specific physical constraints, especially regarding the currents perpendicular to the coasts, the correlation lengths and the depth of the forcing, no other properties of the oceanic currents was present in the ensemble of forcings. However, figure 8.2a clearly shows some specific real currents, like



(a) (b)

Figure 8.5: Ensemble standard deviations (in m) of: (a) SSH before assimilation. (b) SSH after assimilation.

the Gulf Stream in the North Atlantic Ocean, the Humboldt Current, in the South Pacific Ocean, or the Antarctic Circumpolar current. This result is coherent with the limitations inherent with the low resolution of the NEMO model, which tends to underestimate the strength of those strong currents. The forcing reinforces those currents with a specific correction, effectively accounting for the limitations of the noncorrected model. This forcing, intended to correct current biases in the NEMO model, could thus be used in the future as an additional forcing on the currents to provide a better and more realistic ocean dynamic climatology for NEMO.

8.1.4 SST Validation

In order to validate the final correction field from figure 8.6, the model rerun mean SST is compared against a mean SST climatology (hence observations) from NODC-WOA13V2 data provided by the National Oceanic and Atmospheric Administration (NOAA) (Locarnini et al., 2013), interpolated on the ORCA2 grid. The RMSE of the model free run, the ensemble mean before assimilation, and the model rerun are shown on figure 8.8a.

One can see that the optimal forcing from figure 8.6 does deteriorate the SST. The SST RMSE value corresponding to the final forcing is 1.33 C°, compared to 0.961 C° for the model free run. The origin of this behaviour lies in the origin of the model bias, and the average of the mixed layer depth. In this work, the bias is only

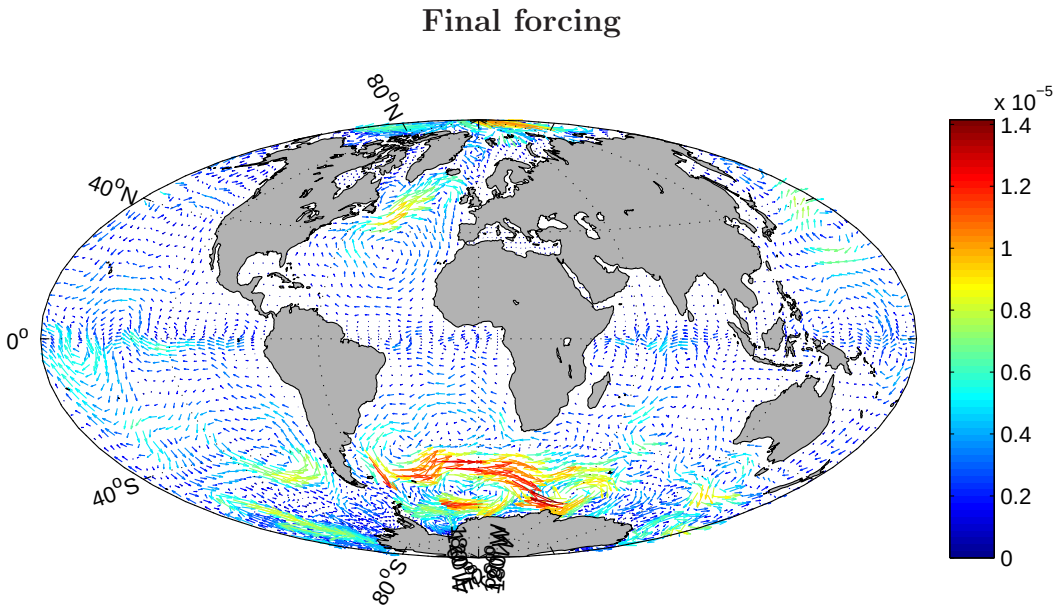


Figure 8.6: Analysed forcing from CNES-MDT observations, used to rerun the model (in ms^{-2}).

corrected for the ocean circulation, whereas in reality multiple other bias sources also affect the model and the SST. Additionally, the average mixed layer depth constraining the forcing does not respect seasonal variations, and is therefore unrealistic. Figure 8.8c is the corresponding spatial SST bias to the final forcing, while 8.8e is the spatial SST bias of the free model run. A clear deterioration is present on the SST around the equator in the Pacific Ocean, where a large scale forcing creates a strong westward current. Other regions do not react better. Interestingly though is that in the Antarctic region, the model reacts to the forcing by cooling the surface waters.

However, with other parameters for the bias correction on the ocean currents, in particular with currents forcing an order of magnitude weaker than the final forcing presented, and a correlation length of 10000 km, the effect on the SST climatology of the model rerun shows a very slight improvements, with RMSE as low as 0.954 C°, as shown on figure 8.8b. Those results show that a slight improvement can be obtained on other nonassimilated variables, but the complicated relations between the different variables and the model bias renders those improvement particularly difficult to obtain. Figure 8.8d shows the spatial SST bias of the lowest RMSE from figure 8.8b.

This shows that it is possible to obtain an amelioration of the SST bias of the model, when an adequate forcing is applied. With the results presented with the

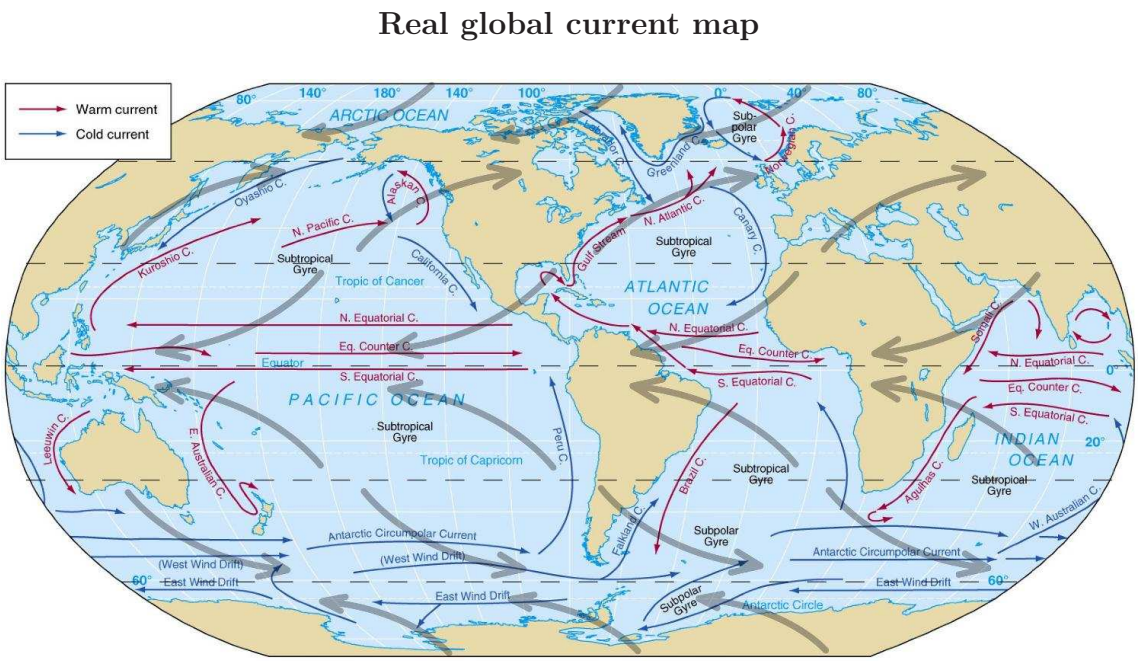


Figure 8.7: Real global average current map of the oceans. Adapted from http://www.georgemaps.com/world-map-ocean-currents-geographic-of-my-nameless-unlabeled-so-far-worldbuilding-warm-current-red-with-circulation/

multivariate assimilation of the twin experiment, one can hope to effectively obtain an optimal correction for both the SSH and the SST, given that optimal parameters are chosen for the forcing generation.

8.2 Iterative analysis

The iterative experiment on the Lorenz '96 model presented in section 5.5 showed a significant improvement on the estimation of the bias correction term and subsequently the bias corrected model output. This iterative assimilation approach is particularly interesting in nonlinear situations, where smaller correction steps are more efficient then a single one.

The NEMO-LIM2 model fulfills the distinct nonlinear behaviour in which the iterative analysis is expected to provide better results.

8.2.1 Experiment set-up

The experiment with the real SSH CNES-MDT observations provided the final correction term presented in section 8.1.3. This experiment aims at obtaining an even

better correction through the iterative assimilation of those same SSH CNES-MDT observations.

The error covariance matrix is adapted following equation (5.5.4), with two equivalently weighed assimilation steps: $\mathbf{R}_1 = \mathbf{R}_2 = 2\mathbf{R}$. The same initial ensemble as for the single assimilation experiment is used, to make sure that the starting points of both the single and iterative assimilation are equivalent. After the first assimilation iteration, the whole analysed ensemble of forcing terms is inflated with an $\alpha = 1.1$ factor, following equation (5.5.10). This analysed ensemble is rerun over one year, and is used as basis for the second assimilation iteration.

The iterative assimilation is computationally heavy to perform. Indeed, one must rerun the whole ensemble for each iteration, and for each ARMSE value on the SSH observations. In practice, one must also remember that due to the complexity and nonlinearity of the model, the mean of the model output from the ensemble rerun is not equivalent to the model output of the ensemble mean rerun. It is the latest which is presented as final forcing in figure 8.6, and which is plotted on the different figures of the model output RMSE on the SSH and SST.

8.2.2 Results

The results of the iterative assimilation experiments are shown on figure 8.9. Because the whole ensemble of forcing has to be rerun through the model for each ARMSE value on the SSH observations, only the most interesting values have been investigated. For clarity, the plotted values on the figure represent the model output of the ensemble mean rerun after the first and second iterations. Furthermore, for the sake of comparing equivalent experiments, the iterative assimilation steps corresponding to the ARMSE value $\mathbf{R}_1 = \mathbf{R}_2 = 2\mathbf{R}$ are plotted, on the X-axis, with the value $\mathbf{R}$ corresponding to the equivalent single assimilation (instead of $2\mathbf{R}$). This way, one can compare on the same vertical line the difference between the single and iterative assimilations. This is visible through the slight horizontal offset between the single assimilation (at value $\mathbf{R}$), and the first iteration (at value $2\mathbf{R}$, plotted at value $\mathbf{R}$).

The first iteration performs similarly to the single assimilation experiment, as visible on figure 8.9. This is due to the fact that a large spectrum of ARMSE values are surveyed, and that the difference between the error covariance matrix $\mathbf{R}_1$

and **R** are small (the x-axis being logarithmic). The second iteration exhibits a systematic improvement over the first iteration, and performs significantly better than the single assimilation step. The second iteration follows the rise in RMSE when the ARMSE values grow, indicating that less information is extracted from the observations, hence less correcting the model.

One can conclude that whereas the initial single assimilation is able to partially correct the bias present in the model through the assimilation of the SSH observations, the iterative assimilation performs better. The possibility for the iterative assimilation scheme to rerun the inflated analysed ensemble and perform a second analysis helps to tune the bias correction term according to the model response to the forcing term. The exact values of figure 8.9 are displayed in table 8.1.

ARMSE on SSH (in m)		RMSE (in m)			
		Background	Single Assim	Iterative Assim	
R	2**R**			iter 1	iter 2
0.0215	0.0431	0.1965	0.1604	0.1604	0.1315
0.0464	0.0928	0.1965	0.1592	0.1579	0.1305
0.1000	0.2000	0.1965	0.1571	0.1554	0.1341
0.2154	0.4308	0.1965	0.1554	0.1574	0.1416
0.4642	0.9284	0.1965	0.1589	0.1640	0.1511

Table 8.1: RMSE on SSH from the ensemble mean before analysis with CNES-MDT observations, from the forced rerun with the observations, and from the first and second successive iterated assimilations.

8.2.3 SSH average error

The yearly mean SSH average errors of the first and second iteration rerun are shown on figure 8.10a and 8.10b. They correspond to the best second iteration rerun, with $2\mathbf{R} = 0.0928$ and a global RMSE on the SSH of 0.1305 m. The SSH average error of the first iteration rerun is very similar to the optimal rerun of the single assimilation experiment, shown on figure 8.3d. However, the SSH average error of the second iteration rerun does show specific improvements. In particular, the two problematic errors in the southern hemisphere, caused by the too deep perturbations, are clearly reduced. Globally, one can conclude that the second iteration rerun allows to attenuate the large errors caused by the initial ensemble.

In a perfect case scenario with infinite computational power, one would perform more iterations, until the improvements on the rerun become negligible, similarly to the "running in place" algorithm mentioned in section 5.5.

8.3 Conclusion

In this chapter, the bias correction method is applied to the NEMO-LIM2 model using realistic observations from CNES-CLS09 global mean dynamic topography. In the first single assimilation experiment, the method showed a significant reduction on the SSH bias compared to the model free run. Specifically, the average SSH affected by strong currents such as the Gulf Stream and the Kuroshio shows an improvement on the model corrected rerun. However, due to the limitations imposed by initial practical choices, the average mixed layer depth limiting the vertical extension in the forcing term generation caused a degradation in two locations around the Antarctic. The final forcing resulting from the single assimilation experiment shows typical realistic currents, though no information about those currents is provided for the ensemble generation. This reflects the ability of the analysis to correct the ensemble and extract useful information from the observations. The intensification of those currents is coherent with the low resolution of the NEMO-LIM2 model, which tends to underestimate said currents. The validation of this final correction with independent SST data however shows a degradation of the global average SST of the ocean. Another set of parameters for the generation of the ensemble, in particular weaker forcing and longer correlation length, showed however a very slight improvement on the SST. This indicates that an improvement is possible, depending on the choices made for the generation of the forcing term.

A second experiment with an iterative assimilation showed that further bias reduction on the SSH is possible. This improvement stems from the nonlinear behaviour of the model. The first rerun of the model allowed the corrected ensemble to provide a better probability density function for the analysis scheme. The second iteration of the assimilation provided a better bias correction, which when rerun by the model, reduced the SSH bias more than any previous obtained reruns.

One can conclude that the bias correction method applied to a realistic model with realistic observations does provide a bias correction term which significantly reduces the bias on the observed quantities. However, one needs to be careful to

the way one generates the ensemble of forcings, for that an incorrect ensemble will not allow the analysis to produce an adequate correction. Unobserved variables are deteriorated by the initial analysis, but a more adequate initial ensemble generation should perform better. A multivariate assimilation should also greatly improve this realistic assimilation by imposing more constraints on the analysis.

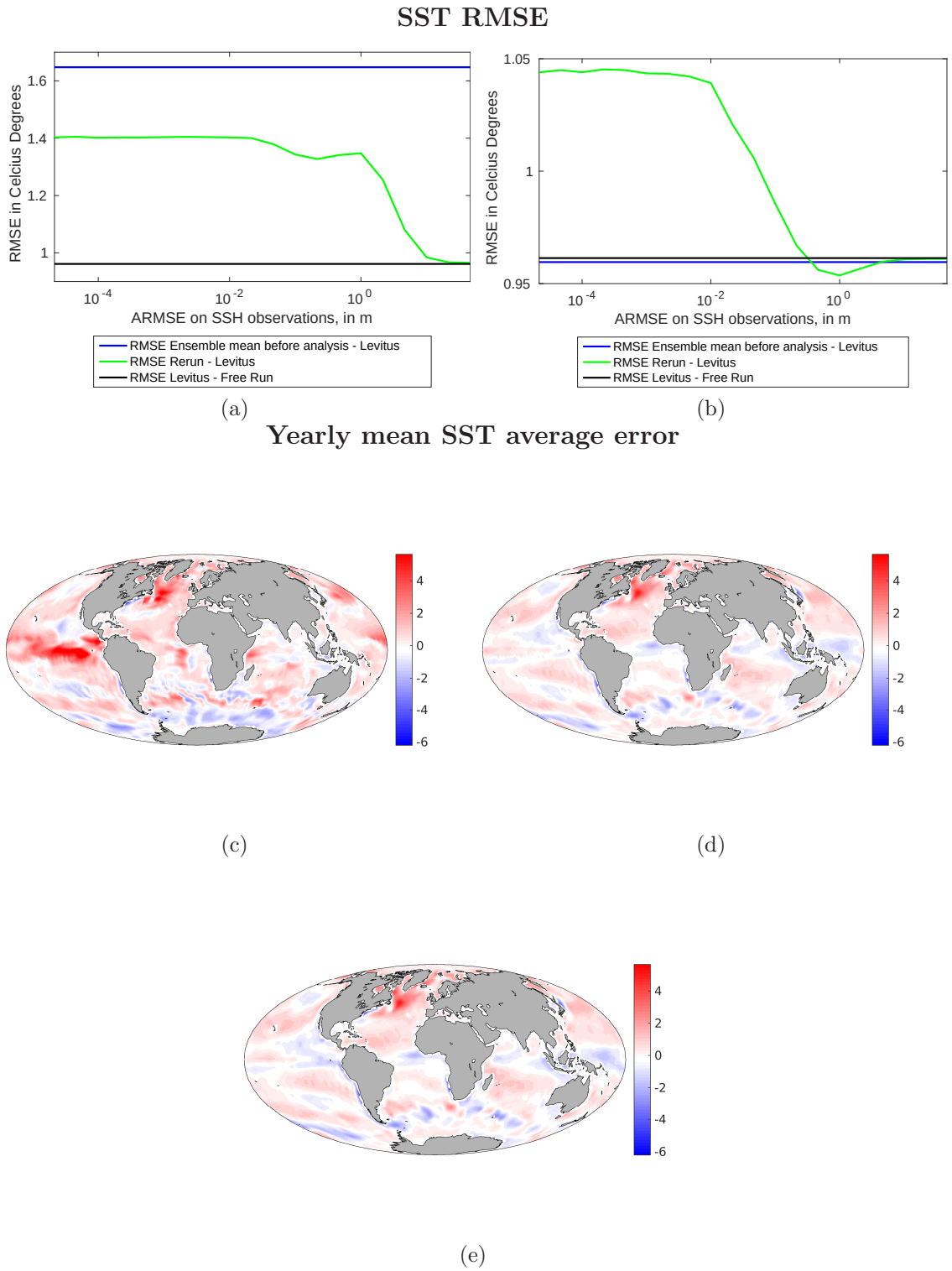


Figure 8.8: Panel (a,b) are RMSE on SST in C° from Ensemble Mean after analysis, Model Free Run, and Rerun, with Levitus observations. Panel (a) are values for the optimal SSH correction from figure 8.6, Panel (b) are the best obtained results for the SST with weaker forcing and a 10000 km correlation length. Panel (c,d,e) are the spatial SST bias in C° corresponding to: Panel (c) the rerun of the final forcing from figure 8.6, Panel (d) the lowest RMSE value from Panel (b) with weaker forcing and 10000 km correlation length, Panel (e) the model free run.

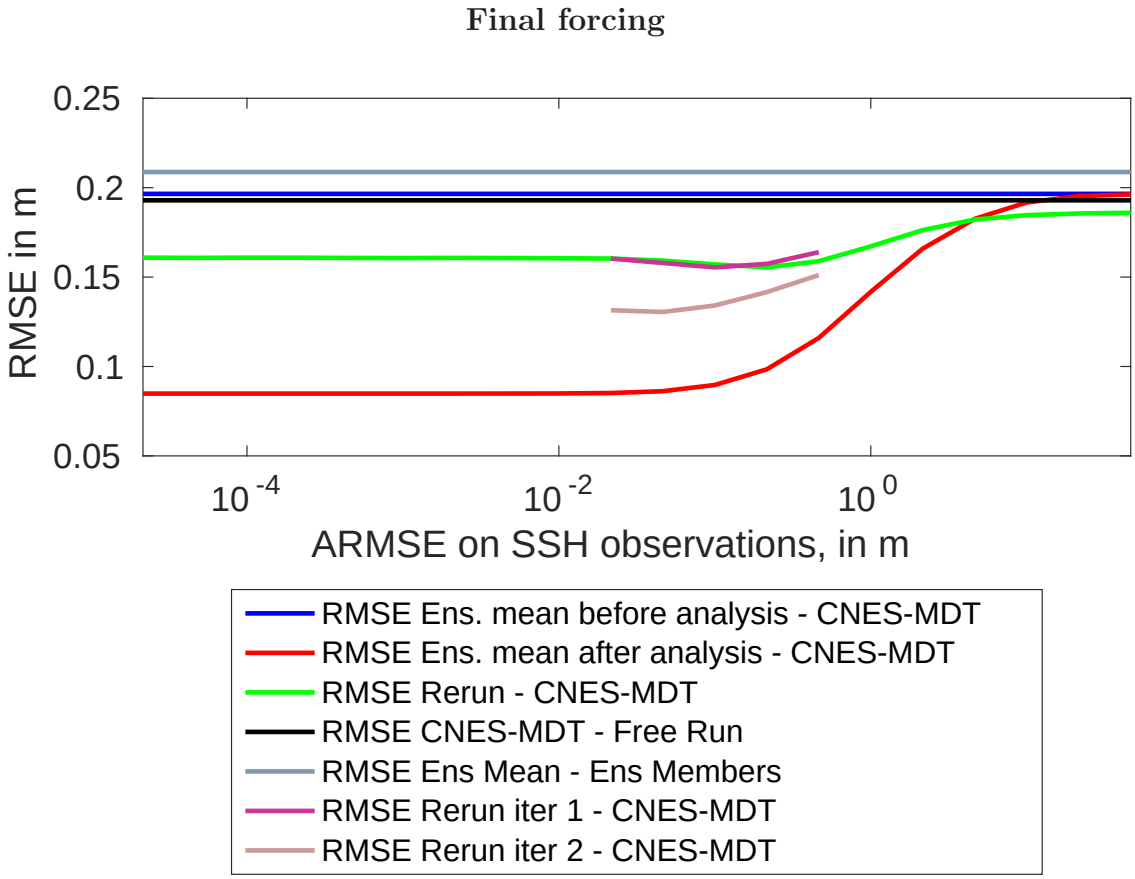
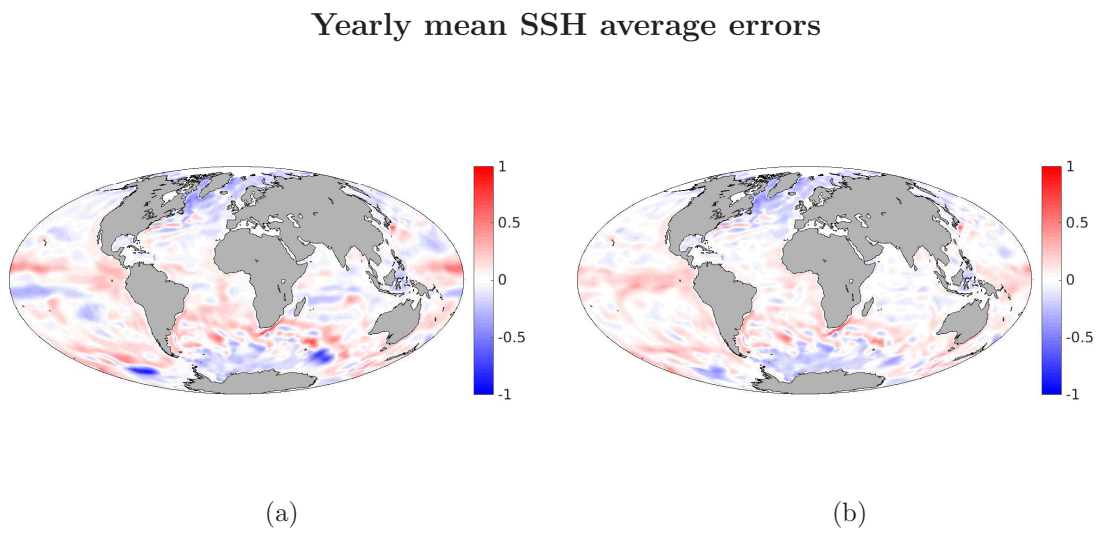


Figure 8.9: RMSE on SSH from the ensemble mean before and after analysis with CNES observations, from the forced rerun with the observations, from the model free run with the observations, from the internal variability of the ensemble, and from the first and second successive iterated assimilations (in m).



(a) (b)

Figure 8.10: Yearly mean SSH average errors with the CNES observations (in m) of (a) rerun of the first iteration, (b) rerun of the second iteration.

Chapter 9

Perspectives

Contents

9.1 Development options

The multiple experiments performed on the NEMO-LIM2 model have pointed out several opportunities to enhance this bias correction method, and showed that significant improvements can be obtained. The twin experiment does show that, in a perfect case scenario, this method is able to estimate and correct the artificially introduced bias. However, in that idealised experiment, one knows exactly how the bias is generated, and how to create a similarly structured bias correction forcing term.

The realistic experiment however indicated the limitations of the method through simultaneous improvement of the global SSH of the model, but deterioration of some other unobserved variables such as the SST. One can only conclude that the model limitations and unresolved processes are more complex that initially assumed. While a part of the SSH bias is effectively corrected, the generated forcing term does not solve all the model bias. This was however to be expected, as the initial assumption

was to focus on the model currents. The forcing term can only correct bias to a certain extent, and only that which can be reproduced by the forcing term. Several improvements of the bias correction method can hence be advised for future works.

9.1.1 3D forcing

The forcing term generation has completed a long journey until its current state. The introduction of the average mixed layer depth and the smoothing of the stream function prior to the zonal and meridional derivatives have improved the stability of the forced runs and obtained a large enough spread of the model. The next step of forcing generation is to replace pseudo-3D vertical extension of the surface forcing term with a real 3D forcing term.

Equation (6.4.5) describes the vertical extension of a single 2D stream function $\Psi(x, y)$. However, if instead of extending the same stream function vertically, one generates a number of stream function corresponding to the number of vertical layers $k = 31$, one can instead obtain a real 3D stream function $\Psi(x, y, z_k)$.

To make sure however that the successive layers do not create contradictory currents, one must ensure a vertical coherence of the applied perturbations $\Psi(x, y, z_k)$ with care. In addition, to keep the variance of the vertical structure of the stream function one can use a coefficient $\alpha = [0, 1)$ to determine the decorrelation of the successive layers (Evensen, 2003). For $\alpha = 0$, they will be fully decorrelated, and $\alpha = 1$ will only propagate the first layer of $\Psi(x, y, z_k)$.

An adequate 3D stream function $\Psi'(x, y, z_k)$ is obtained as follow

$$\Psi'(x, y, z_k) = \alpha\Psi'(x, y, z_{k-1}) + \sqrt{1 - \alpha^2}\Psi(x, y, z_k). \tag{9.1.1}$$

One initialises $\Psi'(x, y, z_1) = \Psi(x, y, z_1)$, and one can then attenuate the real 3D stream function when one reaches the mixed layer depth through

$$\Psi''(x, y, z) = \frac{\Psi'(x, y, z_k)}{1 + \exp(\frac{z - T(x,y)}{L})}. \tag{9.1.2}$$

One can expect a real 3D structure for the forcing term to provide a larger ensemble spread throughout the vertical dimension.

9.1.2 Time dependent forcing

As shown in section 6.3, the bias affecting a model can have both a spatial and a temporal dependence. In the framework of the presented experiments, the temporal dependence has been ignored. The hypothesised bias source was the poor resolution of the model, and no seasonal currents were closely studied. In addition, adding more complexity to the forcing term generation would increase the size of the ensemble, for the number of degrees of freedom would rise proportionally.

For a time-depending forcing, one considers a seasonal bias correction term by taking seasonal averages, instead of yearly averages. One then slowly replaces one forcing term by the next through the use of the coefficient $\alpha = [0, 1)$, and obtains a seasonal correction, for seasonal biases. The observations should then be separated into the corresponding time frames. Seasonal climatologies of the observations are available for use.

The mixed layer depth used as a yearly average would also benefit from a time dependence. Specifically, this would allow the forcing to be active in the polar regions during ice melting periods, solving the two large SSH errors of the real assimilation experiment (figure 8.3d).

9.2 Validation

9.2.1 Method comparison and combination

The different experiments performed in the context of this work cover a wide scope of the problems encountered for the development of the method. They also investigate the efficiency of this method in a realistic case. However, the initial development and implementation from scratch did not allow to effectively compare this method with other bias correction methods.

The most commonly referred to method is described by Dee and Da Silva (1998). It has been adapted to many different cases, in particular by Leeuwenburgh (2008), where it is modified for use with the EnKF to estimate and correct surface wind-stress bias in the Tropical Pacific ocean.

One can always consider combining classic assimilation schemes with a bias correction term. After first estimating an adequate bias correction term aiming at

reducing a part of the systematic model error, one can apply a classic assimilation scheme to further reduce the model error. The combination of both approaches can prove to be even more effective, since they are not mutually exclusive.

9.2.2 Parametrisation

The forcing presented in section 8.1.3 does show some specific physical properties which have been discussed. However, one could also aim at parametrising the estimated correction term. Indeed, the best case scenario would allow the forcing term to lead to a new parametrisation of physical processes in NEMO-LIM2.

In Ferreira and Marshall (2006), a parametrisation of mesoscale eddies in coarse-resolution ocean general circulation models is formulated and implemented using a residual-mean formalism. Residual velocity, defined as the sum of the Eulerian and eddy-induced velocities, advects the mean buoyancy, which is then modified by a residual flux accounting for the diabatic effects of mesoscale eddies. In that approach, the residual velocity is obtained by stepping forward a residual-mean momentum equation in which eddy stresses appear as forcing terms, and are used as control parameters to fit the model to the observations.

Both approaches, from this bias correction method and from Ferreira and Marshall (2006), act on the ocean circulation in a coarse-resolution model. To see if they are equivalent in any sense, one can consider the following expressions from section 6.4.2,

$$\frac{du}{dt} = -\frac{1}{\rho}\frac{\partial p}{\partial x} + fv + \frac{1}{\rho}\frac{\partial \tau_x}{\partial z} - \frac{\partial \Psi'}{\partial y}, \tag{9.2.1}$$

$$\frac{dv}{dt} = -\frac{1}{\rho}\frac{\partial p}{\partial y} - fu + \frac{1}{\rho}\frac{\partial \tau_y}{\partial z} + \frac{\partial \Psi'}{\partial x}, \tag{9.2.2}$$

and compare them to the residual momentum balance appropriate to large scales as describer in Ferreira and Marshall (2006),

$$\frac{\partial \mathbf{v}_{res}}{\partial t} + \mathbf{v}_{res} \bullet \nabla \mathbf{v}_{res} + f\hat{\mathbf{z}} \times \mathbf{v}_{res} = -\frac{1}{\rho_0}\nabla p + \frac{1}{\rho_0}\frac{\partial(\tau^w + \tau^e)}{\partial z} + \nu\nabla^2 \mathbf{v}_{res}. \tag{9.2.3}$$

Here, $\mathbf{v}_{res}$ represents the residual velocity, τ^w the surface wind stress, p the pressure, f the Coriolis parameter, ρ_0 a constant reference density, and τ^e an eddy stress. The objective is to show the equivalency between the forcing term from

the bias correction method and the parametrisation of mesoscale eddies, of in other words, that

$$-\frac{\partial \Psi'}{\partial y} = \frac{1}{\rho_0}\frac{\partial \tau_x^e}{\partial z}, \tag{9.2.4}$$

$$\frac{\partial \Psi'}{\partial x} = \frac{1}{\rho_0}\frac{\partial \tau_y^e}{\partial z}. \tag{9.2.5}$$

Unfortunately, no significant correspondence has been found between the final forcing term of figure 8.6, and the parametrisation of mesoscale eddies from Ferreira and Marshall (2006). Other parametrisations of the final forcing obtained from the real NEMO-LIM2 experiment could prove to be interesting, but do require a deeper knowledge of the model and subscale processes.

9.2.3 Localised corrections

Instead of performing a global correction for the whole model, one can consider to apply the forcing term locally. For instance, the realistic experiment has shown an improvement of the SSH around the Gulf Stream. Typically, coarse resolution climate models struggle with the representation of this surface current, causing the Gulf Stream to exhibit a mean pathway north of the observations (Schoonover et al., 2016). The correction would then be validated by the position of the Gulf Stream, and its total heat transport.

One could imagine to retrieve and adapt the bias correction term of the final forcing locally, similarly to the local assimilation methods to other areas of improvements from the global correction, where the rerun performs better than the free and biased model run.

Chapter 10

Conclusion

In the general context of bias correction in numerical modelling, Dee and Da Silva (1998) is commonly seen as a reference. However, like similar methods, the approach to correct bias in the model output does not remove the source of the bias. Instead, they account for the bias, through offline or online bias estimation, and attempt to reduce its effect on the model. More recently, Leeuwenburgh (2008) performed the estimation of surface wind-stress through an ensemble Kalman filter and corrected the boundary conditions of the model, effectively reducing the model bias.

The bias correction method presented in details in this thesis has as objective to come closer to the source of the model bias. One must though understand that only the bias originating from the source is corrected. If the bias has multiple origins, all must be corrected separately. Its general theoretical formulation allows its transposition and implementation to any numerical model for which an ensemble transform Kalman filter can serve as a practical data assimilation scheme.

The interpretation of the method can be conducted through different angles. In essence though, the model state vector is augmented with a forcing term. This forcing term is estimated through the assimilation of observations with an ensemble transform Kalman filter. The estimated correction is then rerun with the model, providing one with a bias reduced model trajectory. This rerun allows the bias correction to be validated with independent data.

To this end and as suggested by Lorenz, a Lorenz '96 model is modified to suit to the needs of the hypotheses. The Lorenz model (Lorenz, 1963, 1996; Lorenz and Emanuel, 1998), in its 3 or 40 variables form, is usually used for its short term chaotic behaviour, with rapid and unexpected regime changes (Li et al., 2009; Yang

et al., 2012a; Terasaki and Miyoshi, 2014). Bias being considered and described as a systematic error of nonzero mean, one can consider the average of model trajectories instead of full trajectories. While in theory both approaches should provide similar results, in practice, the computational cost is greatly reduced. In addition, a spatially correlated structure is given to the forcing parameter. Hence, the average Lorenz model forcing parameter F_k and output X_k are considered. With an initial biased ensemble of forcing parameters $F_{k,i}$, the ensemble transform Kalman filter is able to reduce the bias. The corrected rerun of the ensemble shows a clear improvement on the value of the forcing parameter.

However, considering the usual complexity of numerical modelling, nonlinearities represent a common obstacle, causing most data assimilation schemes to be sub-optimal. One way to circumvent such difficulties is to perform an iterative assimilation (Evensen and van Leeuwen, 2000; Annan et al., 2005b). Instead of performing a single huge correction, one can assimilate the same observation iteratively by adapting the observation error covariance matrix accordingly. In the context of the bias correction method presented here, this is particularly interesting, since one can rerun the model between each assimilation iteration, obtaining a new ensemble, hence a new probability density function to describe and represent the model solution. It is shown that the iterative assimilation systematically outperforms the single assimilation procedure. The larger the number of iterations, the better the parameter estimation becomes. However, due to the rerun of the ensemble, one is confronted to specific issues such as filter divergence and ensemble collapse. Finally, the computational cost is directly proportional to the number of iterations. While this is not an issue for a simple Lorenz '96 model, this can become challenging for more complex models.

After the successful implementation of the method on a modified Lorenz '96 model twin experiment, a realistic model is considered. Initially, the idea of estimating a bias correction forcing term arose during the PredAntar project with the NEMO-LIM2 model (Goosse et al., 2015). The specific needs of the project required 30 years long ocean simulation of the globe. A comparison of the NEMO-LIM2 results of sea-ice coverage over the considered period with other similar models from CMIP5 shows that all models are subject to bias. The classic approach of observations assimilation provides one with a satisfactory reanalysis when observations are available. However, forecasts are much less accurate due to the absence of observations.

By construction, NEMO-LIM2 (and other similar models) is subject to currents bias due to the coarse resolution of its ORCA2 grid. Surface currents being directly connected to sea surface height, the estimation of a forcing term for the currents can be derived from the assimilation of SSH observations. The average model trajectory is represented by the average SSH. To construct an initial ensemble of forcing, the generation of random fields considered as stream functions is performed with the DIVA tool. Specific constraints are imposed on the current forcing term generation to avoid unrealistic patterns in the bias correction term and its effect on the model. The bias correction method applied to a twin experiment with the NEMO-LIM2 model shows that with an adequate forcing term generation, one can successfully estimate the bias correction to apply. In addition, reducing the bias on the currents also reduced the bias on other variables, such as sea surface temperature and salinity. The same experiment is also performed while assimilating both SSH and SST observation, reducing the bias even further.

Finally, real observations obtained from the CNES mean dynamic topography allow a practical application of this bias correction method. The estimated bias correction term shows specific structures corresponding to realistic current patterns, such as the Gulf Stream in the North Atlantic Ocean, the Humboldt Current, in the South Pacific Ocean, or the Antarctic Circumpolar current. Even though the bias correction term generation is completely random, realistic structures are obtained from the analysis. The ensemble mean corrected rerun shows a reduced RMSE on the average SSH. However, due to physical processes unrepresented in the model, a significant deterioration is caused by the correction term both on the SST and SSS. The origin of this deterioration lies in the hypothesis of a constant forcing term throughout the year, constrained by a constant mixed layer depth. This unrealistic behaviour does not follow the seasonal deep water formation and sea-ice formation cycle in the summer and the winter.

Confronted to a realistic model and observations, and their inherent nonlinear behaviour, the iterative assimilation approach is also experimented. Results clearly show a significant reduction of the SSH bias between the single assimilation rerun, and the second iteration rerun. Furthermore, the unrealistic patterns caused by the initial ensemble generation are partially removed. The computational cost of the rerun procedure is however significant since the search for the optimal bias correction generation and assimilation parameters is performed by trial and error.

This promising bias correction method proved successful, both in fully controlled mathematical models and realistic experiment with real observations. Specific problems are pointed out, such as the importance of having an adequate ensemble, and a good knowledge of the model. Possible developments and perspectives are suggested to provide improvements on the whole bias correction procedure.

Chapter 11

Appendix

11.1 Inverse of block matrix

One can show that

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \\ -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1} & (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1} \end{bmatrix}, \tag{11.1.1}$$

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}^{-1} = \begin{bmatrix} (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & -(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}\mathbf{B}\mathbf{D}^{-1} \end{bmatrix}. \tag{11.1.2}$$

One can derive equation (11.1.1) by imposing that

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{11.1.3}$$

Hence, one has that

$$\mathbf{A}\mathbf{W} + \mathbf{B}\mathbf{Y} = \mathbf{I}, \tag{11.1.4}$$

$$\mathbf{A}\mathbf{X} + \mathbf{B}\mathbf{Z} = \mathbf{0}, \tag{11.1.5}$$

$$\mathbf{C}\mathbf{W} + \mathbf{D}\mathbf{Y} = \mathbf{0}, \tag{11.1.6}$$

$$\mathbf{C}\mathbf{X} + \mathbf{D}\mathbf{Z} = \mathbf{I}. \tag{11.1.7}$$

From equations (11.1.7) and (11.1.5), one shows respectively that

$$\mathbf{X} = -\mathbf{A}^{-1}\mathbf{B}\mathbf{Z}, \tag{11.1.8}$$
$$\mathbf{D}\mathbf{Z} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B}\mathbf{Z} = \mathbf{I}. \tag{11.1.9}$$

From equations (11.1.8) and (11.1.9), one obtains

$$\mathbf{Z} = (\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}, \tag{11.1.10}$$
$$\mathbf{X} = -\mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}, \tag{11.1.11}$$

but also

$$\begin{bmatrix} \mathbf{W} & \mathbf{X} \\ \mathbf{Y} & \mathbf{Z} \end{bmatrix} \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}. \tag{11.1.12}$$

Thus, one has that

$$\mathbf{W}\mathbf{A} + \mathbf{X}\mathbf{C} = \mathbf{I}, \tag{11.1.13}$$
$$\mathbf{Y}\mathbf{A} + \mathbf{Z}\mathbf{C} = \mathbf{0}. \tag{11.1.14}$$
$$\tag{11.1.15}$$

Combining equations (11.1.12), (11.1.13) provides equation (11.1.16), while equation (11.1.17) stems from equation (11.1.14) as follow

$$\mathbf{W} = \mathbf{A}^{-1} - \mathbf{X}\mathbf{C}\mathbf{A}^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}, \tag{11.1.16}$$
$$\mathbf{Y} = -\mathbf{Z}\mathbf{C}\mathbf{A}^{-1} = -(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}. \tag{11.1.17}$$

Following the same reasoning, one can prove equation (11.1.2).

Sherman–Morrison–Woodbury

The Sherman–Morrison–Woodbury is actually a special case of the inverse of block matrices. In fact, from equation (11.1.4) and (11.1.6), one has that

$$\mathbf{Y} = -\mathbf{D}^{-1}\mathbf{C}\mathbf{W}, \tag{11.1.18}$$

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})\mathbf{W} = \mathbf{I}, \tag{11.1.19}$$

$$\mathbf{W} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}. \tag{11.1.20}$$

From the two equations for $\mathbf{W}$, thus equations (11.1.16) and (11.1.20), one obtains the Sherman–Morrison–Woodbury formula, with

$$(\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C}) = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} - \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}. \tag{11.1.21}$$

In particular, for the Kalman filter:

$$\left(\mathbf{R} + \mathbf{H}\mathbf{P}\mathbf{H}^T\right)^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{H}\left(\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}, \tag{11.1.22}$$

or

$$\mathbf{A}\left(\mathbf{I} + \mathbf{A}^T\mathbf{A}\right)^{-1} = \left(\mathbf{I} + \mathbf{A}\mathbf{A}^T\right)^{-1}\mathbf{A}. \tag{11.1.23}$$

From the Sherman-Morrison-Woodbury formula, with the inverse done in the space of matrix $\mathbf{C}$ instead of $\mathbf{A}$, one can show that

$$(\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C}^{-1} = \mathbf{A}\mathbf{B}^T(\mathbf{C} + \mathbf{B}\mathbf{A}\mathbf{B}^T)^{-1}. \tag{11.1.24}$$

11.2 Equivalency of bias estimator

One can show that the analysis using the average model state (equation (4.1.13)) provides the same analysed bias $\widehat{\mathbf{b}^a}$ as when the full trajectory is included in the estimation vector (equation (4.1.8)).

Using $i = 1, \ldots, N$ to refer to the ensemble members, the forecast of the model trajectory can be defined as

$$\mathbf{x}'^f_i = \begin{bmatrix} \mathbf{x}_i^{f(1)} \\ \mathbf{x}_i^{f(2)} \\ \vdots \\ \mathbf{x}_i^{f(m_{\max})} \\ \widehat{\mathbf{b}}_i^f \end{bmatrix}, \qquad \mathbf{x}'^a_i = \begin{bmatrix} \mathbf{x}_i^{a(1)} \\ \mathbf{x}_i^{a(2)} \\ \vdots \\ \mathbf{x}_i^{a(m_{\max})} \\ \widehat{\mathbf{b}}_i^a \end{bmatrix}. \tag{11.2.1}$$

The analysis is provided by

$$\mathbf{x}'^a = \mathbf{x}'^f + \frac{1}{N-1}\mathbf{X}'^f \underbrace{(\mathbf{X}'^f)^T \mathbf{H}'^T (\mathbf{H}'\mathbf{P}'^f\mathbf{H}'^T + R)^{-1}(\mathbf{y}^o - \mathbf{H}'\mathbf{x}'^f)}_{\mathbf{w}'}, \tag{11.2.2}$$

where

$$\mathbf{x}'^f = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}'^f_i, \qquad \mathbf{x}'^a = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}'^a_i, \tag{11.2.3}$$

$$\mathbf{P}'^f = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}'^f_i - \mathbf{x}'^f)(\mathbf{x}'^f_i - \mathbf{x}'^f)^T \tag{11.2.4}$$

$$= \frac{1}{N-1}\mathbf{X}'^f(\mathbf{X}'^f)^T. \tag{11.2.5}$$

The observation operator $\mathbf{H}'$ applied to the trajectory $\mathbf{x}'$ also includes a time average and an extraction operator $\mathbf{H}$ of the observed part of the model state

$$\mathbf{H}'\mathbf{x}' = \frac{1}{m_{max}} \sum_{m=1}^{m_{max}} \mathbf{H}\mathbf{x}^{(m)} = \mathbf{H}\overline{\mathbf{x}}, \tag{11.2.6}$$

$$\overline{\mathbf{x}} = \frac{1}{m_{max}} \sum_{m=1}^{m_{max}} \mathbf{x}^{(m)}. \tag{11.2.7}$$

Hence, the ensemble mean of the analysed bias correction term $\widehat{\mathbf{b}'^a}$ is contained in the analysed model trajectory $\mathbf{x}'^a$. One can also first take the time average of the trajectory, defined as

$$\mathbf{x}''^f_i = \begin{bmatrix} \overline{\mathbf{x}}^f_i \\ \widehat{\mathbf{b}}^f_i \end{bmatrix}, \qquad \mathbf{x}''^a_i = \begin{bmatrix} \overline{\mathbf{x}}^a_i \\ \widehat{\mathbf{b}}^a_i \end{bmatrix}. \tag{11.2.8}$$

The analysis is then given by

$$\mathbf{x}''^a = \mathbf{x}''^f + \frac{1}{N-1}\mathbf{X}''^f \underbrace{(\mathbf{X}''^f)^T \mathbf{H}''^T (\mathbf{H}''\mathbf{P}''^f\mathbf{H}''^T + R)^{-1}(\mathbf{y}^o - \mathbf{H}''\mathbf{x}''^f)}_{\mathbf{w}''}, \tag{11.2.9}$$

where

$$\mathbf{x}''^f = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}''^f_i, \qquad \mathbf{x}''^a = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}''^a_i, \tag{11.2.10}$$

$$\mathbf{P}''^f = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x}''^f_i - \mathbf{x}''^f)(\mathbf{x}''^f_i - \mathbf{x}''^f)^T \tag{11.2.11}$$

$$= \frac{1}{N-1}\mathbf{X}''^f(\mathbf{X}''^f)^T. \tag{11.2.12}$$

The ensemble mean of the analysed bias correction term $\widehat{\mathbf{b}''^a}$ is contained in the analysed mean model state $\mathbf{x}''^a$. Given that

$$\mathbf{H}'\mathbf{x}' = \mathbf{H}''\mathbf{x}'', \tag{11.2.13}$$

it follows that $\mathbf{W}' = \mathbf{W}''$. Hence, $\widehat{\mathbf{b}''^a} = \widehat{\mathbf{b}'^a}$, since they are both constrained by the same linear combination of $\widehat{\mathbf{b}}^f_i$.

11.3 Lorenz long term averages

For **a** and **b** being random vectors, one has that

$$\left(\sum_{k=1}^{n} \mathbf{ab}\right)^2 \leq \left(\sum_{k=1}^{n} \mathbf{a}^2\right)\left(\sum_{k=1}^{n} \mathbf{b}^2\right). \tag{11.3.1}$$

In particular, one can pose $\mathbf{b} = \mathbf{1}$. If one writes expresses the averages of the linear X_k and quadratic X_k^2 terms respectively as r and s^2 as

$$s^2 = \frac{1}{K}\sum_{k=1}^{K} X_k^2, \tag{11.3.2}$$

$$r = \frac{1}{K}\sum_{k=1}^{K} X_k, \tag{11.3.3}$$

$$\tag{11.3.4}$$

One then has that

$$\left(\sum_{k=1}^{K} X_k\right)^2 \leq \left(\sum_{k=1}^{K} X_k^2\right), \tag{11.3.5}$$

$$(r)^2 \leq \left(s^2\right), \tag{11.3.6}$$

$$r \leq |r| \leq s. \tag{11.3.7}$$

Where the $\frac{1}{K}$ factor has been dropped for simplicity.

11.4 List of variables

Variable	Name
$\hat{}$	Hat symbol: Estimate
A	Model accuracy
^{a}	Analysis superscript
$\mathbf{a}$	Analysis increment
α	Observation redundancy factor
$\alpha_{\mathbf{P}}$	Inflation factor
$\mathbf{b}$	Bias
$\boldsymbol{\beta}$	Stochastic error
c_i	Independent integration constant
$\mathbf{D^U}$	Small scale momentum parametrisation in NEMO
D^T	Small scale temperature parametrisation in NEMO
D^S	Small scale salinity parametrisation in NEMO
E	Expectation operator
E	Total energy of Lorenz Model
$\mathbf{e}$	Additional variable in state vector augmentation
ϵ	Observational error
η	Sea level height above the reference ellipsoid
η	Model error
^{f}	Forecast superscript
$\mathbf{F}$	Innovation covariance
F	Forcing parameter of Lorenz 96 model
F_u	Zonal forcing
F_v	Meridional forcing
$\mathbf{F^U}$	Surface momentum forcing in NEMO
F^T	Surface temperature forcing in NEMO
F^S	Surface salinity forcing in NEMO
f	Coriolis acceleration
g	Gravitational acceleration
H	Nonlinear observation operator
$\tilde{\mathbf{H}}$	Tangent linear observation operator (linearisation of H)
$\mathbf{H}$	Linear observation operator
H	Local depth of the ocean
h	Sea level above the geoid
$\mathbf{I}$	Identity matrix
i	Ensemble indice
$(\mathbf{i}, \mathbf{j}, \mathbf{k})$	Orthogonal set of unit vectors in NEMO
J	Cost function
$\mathbf{K}$	Kalman gain
K	Size of Lorenz 96 model
l	Number of initial conditions
L	NEMO forcing generation correlation length

Variable	Name
λ	Longitude
M	Nonlinear forward model operator
$\tilde{\mathbf{M}}$	Tangent linear forward model operator (linearisation of M)
$\mathbf{M}$	Linear forward model operator
m	Time index
N	Height of the geoid to ellipsoid of reference
$\mathcal{N}$	Normal distribution
N_e	Ensemble size
n_{iter}	iteration cycle index
n_{iter}^{max}	Maximum iteration cycle
$O()$	Superior order truncation
$\mathbf{O}_i$	Instrumental error
$\mathbf{O}_r$	Representativeness error
$\mathbf{P}$	Covariance matrix of state vector
p	Pressure
ϕ	Latitude
Ψ	Stream function
$\mathbf{Q}$	Model error covariance matrix
$\mathbf{R}$	Observation error covariance matrix
R	Long term linear average of Lorenz model
R	Mean earth radius
r	Linear average of Lorenz model
ρ	*In situ* density
ρ_0	Reference density
$\mathbf{S}$	$\mathbf{S}^{-1}$ is covariance of the joint distribution $(\mathbf{x}_m, \mathbf{x}_{m-1})\|\mathbf{y}_{0,\ldots,m-1}$
$\mathbf{S}$	Ensemble observation matrix
S	Salinity
$\mathbf{S_P}$	Cholesky decomposition of the covariance matrix
S^2	Long term quadratic average of Lorenz model
s^2	Quadratic average of Lorenz model
σ	Variance of Lorenz Model
$\mathbf{T}$	Transform matrix
T	Average ocean mixed layer depth
T	Potential temperature
^{t}	Superscript for truth
t	Time
$\mathbf{t}$	Tangent vector
tr	Trace
$\mathbf{U}$	Vector velocity
$\mathbf{W}$	Normalisation matrix
$\mathbf{x}$	State vector
$\mathbf{x}'$	Model trajectory
ξ	Linear function of Gaussian random variables
$\mathbf{y}$	Observations
$\mathbf{z}$	Random vector

Abbreviation	Name
$ARMSE$	Adjusted root mean square error
ACC	Antarctic Circumpolar Current
CMCC-CM(S)	Centro Euro-Mediterraneo sui Cambiamenti Climatici Climate Model
CNES	Centre national d'études spatiales
EKF	Extended Kalman Filter
EnKF	Ensemble Kalman Filter
ETKF	Ensemble transform Kalman Filter
KF	Kalman Filter
LIM	Louvain-la-Neuve sea Ice Model
MDT	Mean dynamic topography
NEMO	Nucleus for European Modelling of the Ocean
OAK	Ocean assimilation kit
OSTIA	Operational Sea Surface Temperature and Sea Ice Analysis
PDF	Probability density function
RIP	Running in place
RMSE	Root mean square error
SIA	Sea ice area
SIC	Sea ice concentration
SSH	Sea surface height
SSS	Sea surface salinity
SST	Sea surface temperature

List of Figures

List of Tables

Bibliography

Aksoy, A., Zhang, F., Nielsen-Gammon, J. W., 2006. Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea-breeze model. Monthly Weather Review 134 (10), 2951–2970.

Anderson, J. L., 2001. An Ensemble Adjustment Filter for Data Assimilation. Monthly Weather Review 129 (12), 2884–2903.

Annan, J., Hargreaves, J., Edwards, N., Marsh, R., 2005a. Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. Ocean Modelling 8 (1), 135–154.

Annan, J., Lunt, D., Hargreaves, J., Valdes, P., 2005b. Parameter estimation in an atmospheric GCM using the ensemble Kalman filter. Nonlinear Processes Geophysics 12, 363–371.

Auroux, D., Blum, J., 2005. Back and forth nudging algorithm for data assimilation problems. Comptes Rendus Mathematique 340 (12), 873–878.

Baek, S.-J., Hunt, B. R., Kalnay, E., Ott, E., Szunyogh, I., 2006. Local ensemble Kalman filtering in the presence of model bias. Tellus A 58 (3), 293–306.

Baek, S.-J., Szunyogh, I., Hunt, B. R., Ott, E., 2009. Correcting for surface pressure background bias in ensemble-based analyses. Monthly Weather Review 137 (7), 2349–2364.

Baker, A., Adams, C., Bell, T., Jickells, T., Ganzeveld, L., 2013. Estimation of atmospheric nutrient inputs to the Atlantic Ocean from 50° N to 50° S based on large-scale field sampling: Iron and other dust-associated elements. Global Biogeochemical Cycles 27 (3), 755–767.

Barnes, S., 1964. A technique for maximizing details in numerical map analysis. Journal of Applied Meteorology 3, 395–409.

Barth, A., Alvera-Azcárate, A., Beckers, J.-M., Rixen, M., Vandenbulcke, L., 2007. Multigrid state vector for data assimilation in a two-way nested model of the Ligurian Sea. Journal of Marine Systems 65 (1-4), 41–59.
URL http://hdl.handle.net/2268/4260

Barth, A., Alvera-Azcárate, A., Beckers, J.-M., Weisberg, R. H., Vandenbulcke, L., Lenartz, F., Rixen, M., 2009. Dynamically constrained ensemble perturbations - application to tides on the West Florida Shelf. Ocean Science 5 (3), 259–270.

Barth, A., Alvera-Azcárate, A., Gurgel, K.-W., Staneva, J., Port, A., Beckers, J.-M., Stanev, E. V., 2010. Ensemble perturbation smoother for optimizing tidal boundary conditions by assimilation of High-Frequency radar surface currents - application to the German Bight. Ocean Science 6 (1), 161–178.

Barth, A., Beckers, J.-M., Troupin, C., Alvera-Azcárate, A., Vandenbulcke, L., 2014. divand-1.0: n-dimensional variational data analysis for ocean observations. Geoscientific Model Development 7 (1), 225–241.
URL http://www.geosci-model-dev.net/7/225/2014/

Barth, A., Canter, M., Van Schaeybroeck, B., Vannitsem, S., Massonnet, F., Zunz, V., Mathiot, P., Alvera-Azcárate, A., Beckers, J.-M., 2015. Assimilation of sea surface temperature, sea ice concentration and sea ice drift in a model of the southern ocean. Ocean Modelling 93, 22–39.

Bayes, M., Price, M., 1763. An essay towards solving a problem in the doctrine of chances. by the late rev. Mr. Bayes, frs communicated by Mr. Price, in a letter to John Canton, amfrs. Philosophical Transactions (1683-1775), 370–418.

Bell, M. J., Martin, M., Nichols, N., 2004. Assimilation of data into an ocean model with systematic errors near the equator. Quarterly Journal of the Royal Meteorological Society 130 (598), 873–893.

Bennett, A. F., 1992. Inverse methods in physical oceanography. Cambridge Monographs on Mechanics and Applied Mathematics. Cambridge University Press.

Bergman, J. W., Hendon, H. H., 2015. Calculating monthly radiative fluxes and heating rates frommonthly cloud observations. Journal of the Atmospheric Sciences 72 (9).

Bergthörsson, P., Döös, B. R., 1955. Numerical weather map analysis. Tellus 7 (3), 329–340.

Bertino, L., Evensen, G., Wackernagel, H., 2003. Sequential data assimilation techniques in oceanography. International Statistical Review 71 (2), 223–241.

Bishop, C. H., Etherton, B. J., Majumdar, S. J., 2001. Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects. Monthly Weather Review 129 (3), 420–436.

Blayo, E., Bocquet, M., Cosme, E., Cugliandolo, L. F., 2014. Advanced data assimilation for geosciences. In: International Summer School-Advanced Data Assimilation for Geosciences. p. 608.

Blayo, E., Verron, J., Molines, J. M., Testard, L., 1996. 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.

Bouillon, S., Maqueda, M. A. M., Legat, V., Fichefet, T., 2009. An elastic-viscous-plastic sea ice model formulated on Arakawa B and C grids. Ocean Modelling 27, 174–184.

Brankart, J.-M., Testut, C.-E., Brasseur, P., Verron, J., 2003. Implementation of a multivariate data assimilation scheme for isopycnic coordinate ocean models: Application to a 1993–1996 hindcast of the North Atlantic Ocean circulation. Journal of Geophysical Research: Oceans 108 (C3).

Bratseth, A. M., 1986. Statistical interpolation by means of successive corrections. Tellus A 38 (5), 439–447.

Bryan, F., 1987. Parameter sensitivity of primitive equation ocean general circulation models. Journal of Physical Oceanography 17 (7), 970–985.

Carton, J. A., Chepurin, G., Cao, X., 2000a. A simple ocean data assimilation analysis of the global upper ocean 1950–95. Part II: Results. Journal of Physical Oceanography 30 (2), 311–326.

Carton, J. A., Chepurin, G., Cao, X., Giese, B., 2000b. A simple ocean data assimilation analysis of the global upper ocean 1950-95. Part I: Methodology. Journal of Physical Oceanography 30 (2), 294–309.

Chapron, B., Collard, F., Ardhuin, F., 2005. Direct measurements of ocean surface velocity from space: Interpretation and validation. Journal of Geophysical Research: Oceans 110 (C7).

Chepurin, G. A., Carton, J. A., Dee, D., 2005. Forecast model bias correction in ocean data assimilation. Monthly Weather Review 133 (5), 1328–1342.

Cosme, E., Brankart, J.-M., Verron, J., Brasseur, P., Krysta, M., 2010. Implementation of a reduced-rank, square-root smoother for high resolution ocean data assimilation. Ocean Modelling 33 (1-2), 87–100.

Counillon, F., Bertino, L., 2009. Ensemble Optimal Interpolation: multivariate properties in the Gulf of Mexico. Tellus A 61 (2), 296–308.

Courtier, P., Andersson, E., Heckley, W., Pailleux, J., Vasiljevic, D., Hamrud, M., Hollingsworth, A., Rabier, F., Fisher, M., 1998. The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. Quarterly Journal of the Royal Meteorological Society 124, 1783–1808.

Cressman, G., 1959. An operational objective analysis system. Monthly Weather Review 88, 327–342.

Cucurull, L., Anthes, R., Tsao, L.-L., 2014. Radio occultation observations as anchor observations in numerical weather prediction models and associated reduction of bias corrections in microwave and infrared satellite observations. Journal of Atmospheric and Oceanic Technology 31 (1), 20–32.

da Rocha Fragoso, M., de Carvalho, G. V., Soares, F. L. M., Faller, D. G., de Freitas Assad, L. P., Toste, R., Sancho, L. M. B., Passos, E. N., Böck, C. S., Reis, B., et al., 2016. A 4D-variational ocean data assimilation application for Santos Basin, Brazil. Ocean Dynamics 66 (3), 419–434.

Daley, R., 1991. Atmospheric data analysis, cambridge atmospheric and space science series. Cambridge University Press 6966, 25.

de Boyer Montégut, C., Madec, G., Fischer, A. S., Lazar, A., Iudicone, D., 2004. Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. Journal of Geophysical Research: Oceans 109 (C12).

Dee, D. P., 2005. Bias and data assimilation. Quarterly Journal of the Royal Meteorological Society 131 (613), 3323–3344.

Dee, D. P., Da Silva, A. M., 1998. Data assimilation in the presence of forecast bias. Quarterly Journal of the Royal Meteorological Society 124 (545), 269–295.

Dee, D. P., Todling, R., 2000. Data assimilation in the presence of forecast bias: The geos moisture analysis. Monthly Weather Review 128 (9), 3268–3282.

Dee, D. P., Uppala, S., 2009. Variational bias correction of satellite radiance data in the era-interim reanalysis. Quarterly Journal of the Royal Meteorological Society 135 (644), 1830–1841.

DEL MORAL, P., NOYER, J., RIGAL, G., SALUT, G., 1995. Rec herc hes. Traitement du Signal 12 (4).

Derber, J., Rosati, A., 1989. A global oceanic data assimilation system. Journal of Physical Oceanography 19, 1333–1347.

Derber, J. C., Wu, W.-S., 1998. The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. Monthly Weather Review 126 (8), 2287–2299.

Dimet, F.-X. L., Talagrand, O., 1986. Variational algorithms for analysis and assimilation of meteorological observations: Theoretical aspects. Tellus 38A, 97–110.

Donlon, C. J., 2002. GHRSST-PP Data Poduct Specifivations v2.0. Tech. rep., GODEA Hight Resolution Sea Surface Temperature Pilot Project, GHRSST-PP Reference Document GHRSST/10, http://ghrsst-pp.org.

Donlon, C. J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., Wimmer, W., 2012. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. Remote Sensing of Environment 116 (0), 140–158.

Doron, M., Brasseur, P., Brankart, J.-M., Losa, S. N., Melet, A., 2013. Stochastic estimation of biogeochemical parameters from Globcolour ocean colour satellite data in a North Atlantic 3D ocean coupled physical–biogeochemical model. Journal of Marine Systems 117, 81–95.

Doucet, A., De Freitas, N., Gordon, N., 2001. An introduction to sequential monte carlo methods. In: Sequential Monte Carlo methods in practice. Springer, pp. 3–14.

Dufresne, J.-L., Foujols, M.-A., Denvil, S., Caubel, A., Marti, O., Aumont, O., Balkanski, Y., Bekki, S., Bellenger, H., Benshila, R., et al., 2013. Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Climate Dynamics 40 (9-10), 2123–2165.

Durand, M., Fu, L.-L., Lettenmaier, D. P., Alsdorf, D. E., Rodriguez, E., Esteban-Fernandez, D., 2010. The surface water and ocean topography mission: Observing terrestrial surface water and oceanic submesoscale eddies. Proceedings of the IEEE 98 (5), 766–779.

Edwards, C. A., Moore, A. M., Hoteit, I., Cornuelle, B. D., 2015. Regional ocean data assimilation. Annual review of marine science 7, 21–42.

Errico, R. M., 1997. What is an adjoint model? Bulletin of the American Meteorological Society 78 (11), 2577–2591.

Evensen, G., 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. Journal of Geophysical Research 99 (C5), 10143–10162.

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

Evensen, G., 2007. Data assimilation: the Ensemble Kalman Filter. Springer, 279pp.

Evensen, G., van Leeuwen, P. J., 2000. An Ensemble Kalman Smoother for Nonlinear Dynamics. Monthly Weather Review 128, 1852–1867.

Ferreira, D., Marshall, J., 2006. Formulation and implementation of a "residual-mean" ocean circulation model. Ocean Modelling 13 (1), 86–107.

Fertig, E. J., BAEK, S.-J., Hunt, B. R., Ott, E., Szunyogh, I., Aravéquia, J. A., Kalnay, E., Li, H., Liu, J., 2009. Observation bias correction with an ensemble Kalman filter. Tellus A 61 (2), 210–226.

Fichefet, T., Maqueda, M. A. M., 1997. Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. Journal of Geophysical Research 102, 12609–12646.

Fichefet, T., Maqueda, M. A. M., 1999. Modelling the influence of snow accumulation and snow-ice formation on the seasonal cycle of the Antarctic sea-ice cover. Climate Dynamics 15, 251–268.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W. J., Cox, P., Driouech, F., Emori, S., Eyring, V., et al., 2013. Evaluation of Climate Models. In: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Climate Change 2013 5, 741–866.

Friedland, B., 1969. Treatment of bias in recursive filtering. Automatic Control, IEEE Transactions on 14 (4), 359–367.

Gandin, L. S., Hardin, R., 1965. Objective analysis of meteorological fields. Vol. 242. Israel program for scientific translations Jerusalem.

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

Gent, P. R., McWilliams, J. C., 1990. Isopycnal mixing in ocean circulation models. Journal of Physical Oceanography 20 (1), 150–155.

Ghavidel, A., Schiavulli, D., Camps, A., 2016. Numerical computation of the electromagnetic bias in gnss-r altimetry. IEEE Transactions on Geoscience and Remote Sensing 54 (1), 489–498.

Golub, G. H., Van Loan, C. F., 1996. Matrix computations. 1996. Johns Hopkins University, Press, Baltimore, MD, USA, 374–426.

Gong, J., Wahba, G., Johnson, D. R., Tribbia, J., 1998. Adaptive tuning of numerical weather prediction models: Simultaneous estimation of weighting, smoothing, and physical parameters. Monthly Weather Review 126 (1), 210–231.

Goosse, H., Close, S., Dubinkina, S., Massonnet, F., Zunz, V., Vannitsem, S., Schaeybroeck, B. V., Barth, A., Canter, M., 2015. Understanding and predicting Antarctic sea ice variability at the decadal timescale - "PREDANTAR".
URL `www.elic.ucl.ac.be/users/zunz/site_PREDANTAR/en-project_results.html`

Gottwald, G. A., Majda, A., 2013. A mechanism for catastrophic filter divergence in data assimilation for sparse observation networks. Nonlinear Processes in Geophysics 20 (5), 705–712.

Grewal, M. S., Andrews, A. P., 2010. Applications of Kalman filtering in aerospace 1960 to the present [historical perspectives]. IEEE Control Systems 30 (3), 69–78.

Heimbach, P., Hill, C., Giering, R., 2005. An efficient exact adjoint of the parallel mit general circulation model, generated via automatic differentiation. Future Generation Computer Systems 21 (8), 1356–1371.

Hinzpeter, H., 1967. Der tagesgang der wasseroberflächentemperatur in der nähe des aquators. Meteor Forschungsergeb., Reihe B 1, 41–44.

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

Hunt, B. R., Kalnay, E., Kostelich, E. J., Ott, E., Patil, D. J., Sauer, T., Szunyogh, I., Yorke, J. A., Zimin, A. V., 2004. Four-dimensional ensemble Kalman filtering. Tellus 56A, 273–277.

Hunt, B. R., Kostelich, E. J., Szunyogh, I., 2007. Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. Physica D 230, 112–126.

Ide, K., Courtier, P., Ghil, M., Lorenc, A., 1997. Unifed notation for data assimilation: operational, sequential and variational. Practice 75 (1B), 181–189.

Janjić, T., Nerger, L., Albertella, A., Schröter, J., Skachko, S., 2011. On domain localization in ensemble-based Kalman filter algorithms. Monthly Weather Review 139 (7), 2046–2060.

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

Kaas, E., Guldberg, A., May, W., Déqué, M., 1999. Using tendency errors to tune the parameterisation of unresolved dynamical scale interactions in atmospheric general circulation models. Tellus A 51 (5), 612–629.

Kalman, R. E., 1960. A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering 82 (D), 35–45.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., Zhu, Y., Leetmaa, A., Reynolds, R., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Jenne, R., Joseph, D., 1996. The NCEP/NCAR 40-Year Reanalysis Project. Bulletin of the American Meteorological Society 77, 437–471.

Kalnay, E., Li, H., Miyoshi, T., Yang, S.-C., Ballabrera-Poy, J., 2007. 4-D-Var or ensemble Kalman filter. Tellus A 59, 758–773.

Kalnay, E., Toth, Z., 1994. Removing growing errors in the analysis cycle. In: Tenth conference on numerical weather prediction. pp. 212–215.

Kalnay, E., Yang, S.-C., 2010. Accelerating the spin-up of Ensemble Kalman Filtering. Quarterly Journal of the Royal Meteorological Society 136 (651), 1644–1651.

Kara, A. B., Rochford, P. A., Hurlburt, H. E., 2003. Mixed layer depth variability over the global ocean. Journal of Geophysical Research: Oceans 108 (C3).

Keppenne, C. L., Rienecker, M. M., Kurkowski, N. P., Adamec, D. A., 2005. Ensemble Kalman filter assimilation of temperature and altimeter data with bias correction and application to seasonal prediction. Nonlinear Processes In Geophysics 12, 491–503.

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

Kondrashov, D., Sun, C., Ghil, M., 2008. Data assimilation for a coupled ocean-atmosphere model. Part II: Parameter estimation. Monthly Weather Review 136 (12), 5062–5076.

Kuczera, G., 1983. Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. Water Resources Research 19 (5), 1151–1162.

Landman, W. A., 2015. Statistical Methods in the Atmospheric Sciences , Daniel S. Wilks: book review. Clean Air Journal Tydskrif vir Skoon Lug 25 (1), 19–19.

Large, W., Danabasoglu, G., 2006. Attribution and impacts of upper-ocean biases in CCSM3. Journal of Climate 19 (11), 2325–2346.

Leeuwenburgh, O., 2008. Estimation and correction of surface wind-stress bias in the Tropical Pacific with the Ensemble Kalman Filter. Tellus A 60 (4), 716–727.

Lehmann, E., 1951. A general concept of unbiasedness. The Annals of Mathematical Statistics 22 (4), 587–592.

Lemoine, F. G., Kenyon, S. C., Factor, J. K., Trimmer, R. G., Pavlis, N. K., Chinn, D. S., Cox, C. M., Klosko, S. M., Luthcke, S. B., Torrence, M. H., et al., 1998. The development of the joint NASA GSFC and the National Imagery and Mapping Agency (NIMA) geopotential model EGM96.

Li, H., Kalnay, E., Miyoshi, T., 2009. Simultaneous estimation of covariance inflation and observation errors within an ensemble Kalman filter. Quarterly Journal of the Royal Meteorological Society 135 (639), 523–533.

Locarnini, R., Mishonov, A., Antonov, J., 2013. others (2013) World Ocean Atlas 2013, Volume 1: Temperature. NOAA Atlas NESDIS 73 (9), 40.

Lorenc, A. C., 1986. Analysis methods for numerical weather prediction. Quarterly Journal of the Royal Meteorological Society 112 (474), 1177–1194.

Lorenz, E. N., 1963. Deterministic nonperiodic flow. Journal of the Atmospheric Sciences 20, 130–141.

Lorenz, E. N., 1996. Predictability: A problem partly solved. In: Proc. Seminar on predictability. Vol. 1.

Lorenz, E. N., 2005. Designing chaotic models. Journal of the Atmospheric Sciences 62 (5), 1574–1587.

Lorenz, E. N., Emanuel, K. A., 1998. Optimal sites for supplementary weather observations: Simulation with a small model. Journal of the Atmospheric Sciences 55, 399–414.

Losa, S. N., Kivman, G. A., Ryabchenko, V. A., 2004. Weak constraint parameter estimation for a simple ocean ecosystem model: what can we learn about the model and data? Journal of marine systems 45 (1), 1–20.

Lyard, F., Lefevre, F., Letellier, T., Francis, O., 2006. Modelling the global ocean tides: modern insights from FES2004. Ocean Dynamics 56, 394–415.

Lynch, P., 2006. The emergence of numerical weather prediction: Richardson's dream. Cambridge University Press.

Lynch, P., 2008. The origins of computer weather prediction and climate modeling. Journal of Computational Physics 227 (7), 3431–3444.

Madec, G., 2008. NEMO ocean engine. No. 27 in Note du Pole de modélisation. Institut Pierre-Simon Laplace (IPSL), France.

Madec, G., Delecluse, P., Imbard, M., Cl., L., Décembre 1998. OPA 8.1, Ocean General Circulation Model, Reference Manual.

Madec, G., Imbard, M., 1996. A global ocean mesh to overcome the North Pole singularity. Climate Dynamics 12 (6), 381–388.

Massari, C., Brocca, L., Tarpanelli, A., Moramarco, T., 2015. Data assimilation of satellite soil moisture into rainfall-runoff modelling: A complex recipe? Remote Sensing 7 (9), 11403–11433.

Massonnet, F., Mathiot, P., Fichefet, T., Goosse, H., Beatty, C. K., Vancoppenolle, M., Lavergne, T., 2013. A model reconstruction of the Antarctic sea ice thickness and volume changes over 1980-2008 using data assimilation. Ocean Modelling 64, 67–75.

Mathiot, P., Goosse, H., Fichefet, T., Barnier, B., Gallée, H., 2011. Modelling the seasonal variability of the Antarctic Slope Current. Ocean Science 7 (4), 455–470.

Metropolis, N., Ulam, S., 1949. The monte carlo method. Journal of the American statistical association 44 (247), 335–341.

Miller, R. N., Ghil, M., Gauthiez, F., 1994. Advanced data assimilation in strongly nonlinear dynamical systems. Journal of the atmospheric sciences 51 (8), 1037–1056.

Moradkhani, H., Sorooshian, S., Gupta, H. V., Houser, P. R., 2005. Dual state–parameter estimation of hydrological models using ensemble kalman filter. Advances in Water Resources 28 (2), 135–147.

Nakamura, T., Akiyoshi, H., Deushi, M., Miyazaki, K., Kobayashi, C., Shibata, K., Iwasaki, T., 2013. A multimodel comparison of stratospheric ozone data assimilation based on an ensemble Kalman filter approach. Journal of Geophysical Research: Atmospheres 118 (9), 3848–3868.

Navier, C., 1823. Mémoire sur les lois du mouvement des fluides. Mem. Acad. Sci. Inst. Fr 6 (1823), 389–416.

Navon, I., 1998. Practical and theoretical aspects of adjoint parameter estimation and identifiability in meteorology and oceanography. Dynamics of Atmospheres and Oceans 27 (1), 55–79.

Nerger, L., Gregg, W. W., 2008. Improving assimilation of SeaWiFS data by the application of bias correction with a local SEIK filter. Journal of marine systems 73 (1), 87–102.

Nerger, L., Hiller, W., 2013. Software for ensemble-based data assimilation systems - Implementation strategies and scalability. Computers & Geosciences 55, 110–118.

Nerger, L., Janjić, T., Schröter, J., Hiller, W., 2012a. A regulated localization scheme for ensemble-based Kalman filters. Quarterly Journal of the Royal Meteorological Society 138 (664), 802–812.

Nerger, L., Janjic, T., Schröter, J., Hiller, W., 2012b. A unification of ensemble square root Kalman filters. Monthly Weather Review 140 (7), 2335–2345.

Ngodock, H., Carrier, M., 2014. A 4dvar system for the navy coastal ocean model. part i: System description and assimilation of synthetic observations in monterey bay. Monthly Weather Review 142 (6), 2085–2107.

Ngodock, H. E., Souopgui, I., Wallcraft, A. J., Richman, J. G., Shriver, J. F., Arbic, B. K., 2016. On improving the accuracy of the m 2 barotropic tides embedded in a high-resolution global ocean circulation model. Ocean Modelling 97, 16–26.

Noorbaloochi, S., Meeden, G., 2000. Unbiasedness and Bayes estimators. Tech. rep., Technical Report.

Oke, P. R., Allen, J. S., Miller, R. N., Egbert, G. D., Kosro, P. M., 2002. Assimilation of surface velocity data into a primitive equation coastal ocean model. Journal of Geophysical Research: Oceans 107 (C9).

Panofsky, R., 1949. Objective weather-map analysis. Journal of Meteorology 6 (6), 386–392.

Pham, D. T., 1996. A singular evolutive interpolated Kalman filter for data assimilation in oceanography. Rapport technique RT 163.

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

Radakovich, J. D., Bosilovich, M. G., Chern, J.-d., da Silva, A., Todling, R., Joiner, J., Wu, M.-l., Norris, P., 2004. Implementation of coupled skin temperature analysis and bias correction in the NASA/GMAO finite-volume data assimilation system (FvDAS). In: P1. 3 in Proceedings of the Eighth AMS Symposium on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface. pp. 12–15.

Radakovich, J. D., Houser, P. R., da Silva, A., Bosilovich, M. G., 2001. Results from global land-surface data assimilation methods. In: AGU Spring Meeting Abstracts. Vol. 1.

Richardson, L., 1922. Weather prediction by numerical methods.

Rinne, J., Järvinen, H., 1993. Estimation of the cressman term for a barotropic model through optimization with use of the adjoint model. Monthly Weather Review 121 (3), 825–833.

Rio, M., Guinehut, S., Larnicol, G., 2011. New CNES-CLS09 global mean dynamic topography computed from the combination of GRACE data, altimetry, and in situ measurements. Journal of Geophysical Research: Oceans (1978–2012) 116 (C7).

Rio, M.-H., Hernandez, F., 2004. A Mean Dynamic Topography computed over the world ocean from altimetry, in-situ measurements and a geoid model. Journal of Geophysical Research.

Roberts-Jones, J., Fiedler, E. K., Martin, M. J., 2012. Daily, Global, High-Resolution SST and Sea Ice Reanalysis for 1985-2007 Using the OSTIA System. J. Climate 25, 6215–6232.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., et al., 2003. The atmospheric general circulation model ECHAM 5. PART I: Model description.

Roeckner, E., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kornblueh, L., Manzini, E., Schlese, U., Schulzweida, U., 2006. Sensitivity of simulated climate to horizontal and vertical resolution in the ECHAM5 atmosphere model. Journal of Climate 19 (16), 3771–3791.

Roullet, G., Madec, G., 2000. Salt conservation, free surface, and varying levels: a new formulation for ocean general circulation models. Journal of Geophysical Research: Oceans (1978–2012) 105 (C10), 23927–23942.

Sakov, P., Evensen, G., Bertino, L., 2010. Asynchronous data assimilation with the EnKF. Tellus 62A, 24–29.

Sakov, P., Oke, P. R., 2008. Implications of the form of the ensemble transformation in the ensemble square root filters. Monthly Weather Review 136 (3), 1042–1053.

Santer, B. D., Wigley, T. M., Simmons, A. J., Kållberg, P. W., Kelly, G. A., Uppala, S. M., Ammann, C., Boyle, J. S., Brüggemann, W., Doutriaux, C., et al., 2004. Identification of anthropogenic climate change using a second-generation reanalysis. Journal of Geophysical Research: Atmospheres 109 (D21).

Schluessel, P., Emery, W. J., Grassl, H., Mammen, T., 1990. On the bulk-skin temperature difference and its impact on satellite remote sensing of sea surface temperature. Journal of Geophysical Research: Oceans 95 (C8), 13341–13356.

Schoonover, J., Dewar, W. K., Wienders, N., Deremble, B., 2016. Local Sensitivities of the Gulf Stream Separation. Journal of Physical Oceanography (2016).

Scoccimarro, E., Gualdi, S., Bellucci, A., Sanna, A., Giuseppe Fogli, P., Manzini, E., Vichi, M., Oddo, P., Navarra, A., 2011. Effects of tropical cyclones on ocean

heat transport in a high-resolution coupled general circulation model. Journal of Climate 24 (16), 4368–4384.

Simmons, A., Uppala, S., Dee, D., Kobayashi, S., 2007. ERA-Interim: New ECMWF reanalysis products from 1989 onwards. ECMWF newsletter 110 (110), 25–35.

Simon, E., Bertino, L., 2009. Application of the Gaussian anamorphosis to assimilation in a 3-D coupled physical-ecosystem model of the North Atlantic with the EnKF: a twin experiment. Ocean Science 5 (4), 495–510.

Simon, E., Bertino, L., 2012. Gaussian anamorphosis extension of the DEnKF for combined state parameter estimation: Application to a 1D ocean ecosystem model. Journal of Marine Systems 89 (1), 1–18.

Stark, J., Donlon, C., Martin, M., McCulloch, M., june 2007. Ostia : An operational, high resolution, real time, global sea surface temperature analysis system. In: OCEANS 2007 - Europe. pp. 1–4.

Stewart, R. H., 2008. Introduction to physical oceanography. Texas A & M University Texas.

Stockdale, T. N., 1997. Coupled ocean-atmosphere forecasts in the presence of climate drift. Monthly Weather Review 125 (5), 809–818.

Stroeve, J., Holland, M. M., Meier, W., Scambos, T., Serreze, M., 2007. Arctic sea ice decline: Faster than forecast. Geophysical research letters 34 (9).

Sudre, J., Maes, C., Garçon, V., 2013. On the global estimates of geostrophic and Ekman surface currents. Limnology and Oceanography: Fluids and Environments 3 (1), 1–20.

Talagrand, O., Courtier, P., 1987. Variational assimilation of meteorological observations with the adjoint vorticity equations: Theory. Quarterly Journal of the Royal Meteorological Society 113, 1311–1328.

Tanizaki, H., 2001. Nonlinear and non-gaussian state space modeling using sampling techniques. Annals of the Institute of Statistical Mathematics 53 (1), 63–81.

Ten Brummelhuis, P., Heemink, A., Van Den Boogaard, H., 1993. Identification of shallow sea models. International journal for numerical methods in fluids 17 (8), 637–665.

Terasaki, K., Miyoshi, T., 2014. Data assimilation with error-correlated and non-orthogonal observations: Experiments with the lorenz-96 model. SOLA 10 (0), 210–213.

Timmermann, R., Goosse, H., Madec, G., Fichefet, T., Ethe, C., Dulière, V., 2005. On the representation of high latitude processes in the ORCA-LIM global coupled sea ice-ocean model. Ocean Modelling 8 (1-2), 175–201.

Tippett, M. K., Anderson, J. L., Bishop, C. H., Hamill, T. M., Whitaker, J. S., 2003. Ensemble square root filters. Monthly Weather Review 131 (7), 1485–1490.

Valcke, S., 2006. OASIS3 user guide (prism_2-5). PRISM support initiative report 3, 64.

van Leeuwen, P. J., 2001. An Ensemble Smoother with Error Estimates. Monthly Weather Review 129, 709–728.

Van Leeuwen, P. J., 2009. Particle filtering in geophysical systems. Monthly Weather Review 137 (12), 4089–4114.

Vaughan, D., Comiso, J., Allison, I., Carrasco, J., Kaser, G., Kwok, R., Mote, P., Murray, T., Paul, F., Ren, J., et al., 2013. Observations: Cryosphere climate change 2013: the physical science basis. et al., Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA 1 (1), 1–2.

Vossepoel, F. C., Weaver, A. T., Vialard, J., Delecluse, P., 2004. Adjustment of near-equatorial wind stress with four-dimensional variational data assimilation in a model of the Pacific Ocean. Monthly Weather Review 132 (8), 2070–2083.

Vrugt, J. A., Gupta, H. V., Bouten, W., Sorooshian, S., 2003. A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resources Research 39 (8).

Wieliczka, D. M., Weng, S., Querry, M. R., 1989. Wedge shaped cell for highly absorbent liquids: infrared optical constants of water. Applied optics 28 (9), 1714–1719.

Wiens, J. A., 2016. Climate change and sea-level rise. Ecological Challenges and Conservation Conundrums: Essays and Reflections for a Changing World, 71–75.

Wilks, D. S., 2011. Statistical methods in the atmospheric sciences. Vol. 100. Academic press.

Yang, S.-C., Kalnay, E., Hunt, B., 2012a. Handling nonlinearity in an ensemble Kalman filter: experiments with the three-variable Lorenz model. Monthly Weather Review 140 (8), 2628–2646.

Yang, S.-C., Kalnay, E., Miyoshi, T., 2012b. Accelerating the EnKF spinup for typhoon assimilation and prediction. Weather and Forecasting 27 (4), 878–897.

Zeng, F., Delworth, T., 2015. The Impact of Multidecadal NAO Variations on Atlantic Ocean Heat Transport and Rapid Changes in Arctic Sea Ice. In: AGU Fall Meeting Abstracts.

Zhu, Y., Navon, I., 1999. Impact of parameter estimation on the performance of the FSU global spectral model using its full-physics adjoint. Monthly Weather Review 127 (7), 1497–1517.

Zunz, V., Goosse, H., Massonnet, F., 2013. How does internal variability influence the ability of CMIP5 models to reproduce the recent trend in Southern Ocean sea ice extent? The Cryosphere 7 (2), 451–468.

Zupanski, M., Navon, I. M., Zupanski, D., 2008. The Maximum Likelihood Ensemble Filter as a non-differentiable minimization algorithm. Quarterly Journal of the Royal Meteorological Society 134 (633), 1039–1050.