Efficient Ensemble Data Assimilation and Forecasting of the Red Sea Circulation

Dissertation by

Habib Toye Mahamadou Kele

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

King Abdullah University of Science and Technology

Thuwal, Kingdom of Saudi Arabia

October, 2020

EXAMINATION COMMITTEE PAGE

The dissertation of Habib Toye Mahamadou Kele is approved by the examination committee

Committee Chairperson: Prof. Ibrahim Hoteit

Committee Members: Prof. Omar Knio, Prof. Tareq Al-Naffouri, Prof. Mohamed Iskandarani

ABSTRACT

Efficient Ensemble Data Assimilation and Forecasting of
the Red Sea Circulation
Habib Toye Mahamadou Kele

This thesis presents our efforts to build an operational ensemble forecasting system for the Red Sea, based on the Data Research Testbed (DART) package for ensemble data assimilation and the Massachusetts Institute of Technology general circulation ocean model (MITgcm) for forecasting. The Red Sea DART-MITgcm system efficiently integrates all the ensemble members in parallel, while accommodating different ensemble assimilation schemes. The promising ensemble adjustment Kalman filter (EAKF), designed to avoid manipulating the gigantic covariance matrices involved in the ensemble assimilation process, possesses relevant features required for an operational setting. The need for more efficient filtering schemes to implement a high resolution assimilation system for the Red Sea and to handle large ensembles for proper description of the assimilation statistics prompted the design and implementation of new filtering approaches. Making the most of our world-class supercomputer, Shaheen, we first pushed the system limits by designing a fault-tolerant scheduler extension that allowed us to test for the first time a fully realistic and high resolution 1000 ensemble members ocean ensemble assimilation system. In an operational setting, however, timely forecasts are of essence, and running large ensembles, albeit preferable and desirable, is not sustainable. New schemes aiming at lowering the computational burden while preserving reliable assimilation results, were developed. The ensemble Optimal Interpolation (EnOI) algorithm requires only a single model integration in the forecast step, using a static ensemble of preselected members for assimilation, and is therefore computationally significantly cheaper than the EAKF. To account for the strong seasonal variability of the Red Sea circulation, an EnOI with seasonally-varying ensembles (SEnOI) was first implemented. To better handle intra-seasonal variabilities and enhance the developed seasonal EnOI system, an automatic procedure to adaptively select the ensemble members through the assimilation cycles was then introduced. Finally, an efficient Hybrid scheme combining the dynamical flow-dependent covariance of the EAKF and a static covariance of the EnOI was proposed and successfully tested in the Red Sea. The developed Hybrid ensemble data assimilation system will form the basis of the first operational Red Sea forecasting system that is currently being implemented to support Saudi Aramco operations in this basin.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

ADT	absolute dynamic topography
AEnOI	Adaptive ensemble Optimal Interpolation
ASCII	American Standard Code for Information Interchange
AVHRR	Advanced Very High Resolution Radiometer
AVISO	Archiving Validation and Interpretation of Satellite Oceanographic Data
BECM	background error covariance matrix
BLUE	best linear unbiased estimator
CMEMS	Copernicus Marine Environment Monitoring Service
CRS	Central Red Sea
CTD	Conductivity Temperature Depth
DART	Data Assimilation Research Testbed
DEU	Dynamic Ensemble Update
EAKF	ensemble adjustment Kalman filter
ECCO	Estimation of the Circulation and Climate of the Ocean
ECMWF	European Centre for Medium-Range Weather Forecasts
EnKF	ensemble Kalman filter
EnOI	ensemble Optimal Interpolation
ETKF	Ensemble Transform Kalman Filter
Fexp	free-run experiment
GCM	general circulation model
GEBCO	General Bathymetric Chart of the Oceans
GHRSST	Group for High Resolution Sea Surface Temperature
GLORYS	Global Ocean Reanalysis and Simulation data
GoA	Gulf of Aden

HPC	High Performance Computing
KAUST	King Abdullah University of Science and Technology
KF	Kalman filter
KPP	K-Profile Parametrization
KSL	KAUST Supercomputing Laboratory
LETKF	Local Ensemble Transform Kalman Filter
MC	Monte Carlo
MDT	mean dynamic topography
MITgcm	Massachusetts Institute of Technology general circulation model
MPD	model physics dictionary
MSE	mean-squared error
MSSH	mean sea surface height
NCAR	National Center for Atmospheric Research
NCEP	National Centers for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
NRS	Northern Red Sea
OBCS	open boundary conditions
OGCM	ocean general circulation model
OI	Optimal Interpolation
OMP	Orthogonal Matching Pursuit
OSTIA	Operational Sea Surface Temperature and Sea Ice Analysis
PDF	probability density function
RADS	Radar Altimeter Database System
RMSE	root-mean-square error
S	salinity
SAMERCK	Saudi Aramco Marine Environmental Research Center at KAUST
SEIK	Singular Evolutive Interpolation Filter
SEnOI	Seasonal ensemble Optimal Interpolation
SLA	sea level anomaly
SRS	Southern Red Sea

SSH	sea surface height
SSS	sea surface salinity
SST	sea surface temperature
STD	standard deviation
SVD	singular value decomposition
T	temperature
THORPEX	The Observing System Research and Predictability Experiment
TIGGE	THORPEX Interactive Grand Global Ensemble
WHOI	Woods Hole Oceanography Institute
WRF	Weather Research and Forecast
XBT	eXpendable BathyThermograph

LIST OF SYMBOLS

$\mathbf{A}_k$	EAKF adjustment operator at time k
$\mathbf{A}$	EAKF adjustment operator
$\mathbf{B}$	Static covariance matrix
$:=$	Symbol for “is defined as equal to”
χ_q^2	Chi-square distribution with q degrees of freedom
$\boldsymbol{\delta}_k$	Innovation vector at time k
$\boldsymbol{\delta}$	Innovation vector
$\mathbb{E}$	Expectation operator
$\boldsymbol{\xi}_{t_k}$	Estimation/representation error at time t_k
$\boldsymbol{\xi}$	Estimation/representation error
$\boldsymbol{\varepsilon}_k^i$	Observation error associated with the ensemble member i, at time k
$\boldsymbol{\varepsilon}_k$	Observation error at time k
$\boldsymbol{\varepsilon}$	Observation error
$\boldsymbol{\eta}$	Model error
$\boldsymbol{\eta}_k$	Model error at time k
$\boldsymbol{\eta}_{t_k}$	Model error at time t_k
$\mathcal{G}_{t_k}$	True dynamical geophysical model to advance the state from time t_k to time t_{k+1}
$\mathcal{G}$	True dynamical geophysical model to advance the state in time
$h_{\mathcal{S}\to\mathcal{O}}$	Map from $\mathcal{S}$ to $\mathcal{O}$
$h_{\mathcal{S}_n\to\mathcal{O}_p}$	Map from $\mathcal{S}_n$ to $\mathcal{O}_p$
$h.o.t.$	higher-order terms
$\mathbf{H}_k$	Matrix associated with the observation operator h_k or the extended observation operator.
h_k	Observation operator at time k
h_{t_k}	Observation operator at time t_k
h	Observation operator

$\mathbf{K}_k$	Kalman gain at time k
l	Joint state-observation space vector size
λ_k	Normalized squared innovation at time k
$\mathcal{M}_{t_k}$	Dynamical model to advance the state from time t_k to time t_{k+1}
$\mathcal{M}$	Dynamical model to advance the state in time
$\mathbf{M}_{t_k}$	Linearization of the model $\mathcal{M}_k$
$\mathbb{N}$	Set of natural numbers
n	Model vector size
N	Ensemble size
N_s	Static ensemble size
$\mathcal{O}_{p_k}$	Observation space of dimension p_k
$\mathcal{O}_{p_{t_k}}$	Observation space of dimension p_{t_k}
$\mathcal{O}_p$	Observation space of dimension p
$\mathcal{O}_{p_{t_k}}(t_k)$	Observation space of dimension p_{t_k} at time t_k
$\mathcal{O}_p(t_k)$	Observation space of dimension p at time t_k
$\mathcal{O}(t_k)$	Observation space at time t_k
$\mathcal{O}$	Observation space
$\mathbf{P}$	Covariance matrix
$\mathbf{P}^{f,e}$	Estimated forecast error covariance matrix
$\mathbf{P}_k^f$	Forecast error covariance matrix at time k
$\mathbf{P}^f$	Forecast error covariance matrix
$\mathbf{P}^{f,H}$	Hybrid forecast error covariance matrix
p_k	Observation vector size at time k
p_{t_k}	Observation vector size at time t_k
$p(\mathbf{x}_0^t)$	Probability distribution of the initial true state vector
$p(\mathbf{x}_k^t \| \mathbf{Y}_{1:k}^o)$	Conditional probability distribution of the true state vector given the observations from time 1 to time k
$\mathbf{R}_k$	Observational error covariance matrix at time k
$\mathbf{R}_{t_k}$	Observational error covariance matrix at time t_k
$\mathbf{R}$	Observational error covariance matrix
r	Single observation error variance

$\mathcal{S}_{n_{t_k}}$	State space of dimension n_{t_k}
$\mathcal{S}_n$	State space of dimension n
$\mathcal{S}_{n_{t_k}}(t_k)$	State space of dimension n_{t_k} at time t_k
$\mathcal{S}_n(t_k)$	State space of dimension n at time t_k
$\mathbf{S}_k$	Innovation covariance matrix at time k
$\mathcal{S}(t_k)$	State space at time t_k
$\mathcal{S}$	State space
$\mathbf{\Sigma}_k^a$	Joint state-observation space analysis error covariance matrix at time k
$\mathbf{\Sigma}^a$	Joint state-observation space analysis error covariance matrix
$\mathbf{\Sigma}_k^f$	Joint state-observation space forecast error covariance matrix at time k
$\mathbf{\Sigma}_{0l}^f$	$\mathbf{\Sigma}_k^f$ entries set to zero, except the last column
$\mathbf{\Sigma}^f$	Joint state-observation space forecast error covariance matrix
$\mathbf{\Sigma}$	Covariance matrix
$\mathbf{\Sigma}^{a,e}$	Estimated joint analysis state-observation space covariance matrix
$\mathbf{\Sigma}^{f,e}$	Estimated joint forecast state-observation space covariance matrix
$\sigma_{l,l}^a$	Covariance of the analysis observation
$\sigma_{l,l}^f$	Covariance of the predicted observation
$\sigma_{j,l}^f$	Covariance between the jth row of the forecast ensemble and the ensemble of predicted observations
$\mathcal{T}_{\mathcal{O}}$	Map from $\mathcal{O}$ to $\mathcal{O}_p$
$\mathcal{T}_{\mathcal{S}}$	Map from $\mathcal{S}$ to $\mathcal{S}_n$
tr	Trace operator
$\|W(z)\|_{max}$	maximum vertical velocity in the ocean column
$\mathbf{X}_k^a$	Ensemble of analysis state vectors at time k
$\mathbf{X}^a$	Ensemble of analysis state vectors
$\boldsymbol{X}'$	Ensemble of anomalies
$\mathbf{X}$	Ensemble of state vectors
$\mathbf{X}_k^f$	Ensemble of forecast state vectors at time k

$\mathbf{X}^f$	Ensemble of forecast state vectors
$\boldsymbol{X}'^{H}$	Hybrid ensemble of anomalies
$\mathbf{X}^H$	Hybrid ensemble
$\boldsymbol{X}'^{s}$	Static ensemble of anomalies
$\overline{\mathbf{x}}_k^a$	Analysis ensemble mean at time k
$\overline{\mathbf{x}}^a$	Analysis ensemble mean
$\mathbf{x}_k^{a,i}$	Analysis member i at time k
$\mathbf{x}^{a,i}$	Analysis member i
$\mathbf{x}^a$	Analysis state vector
$\overline{\mathbf{x}}_k$	Ensemble mean at time k
$\overline{\mathbf{x}}$	Ensemble mean
$\tilde{\mathbf{x}}_{(j)}$	jth row of forecast ensemble $\mathbf{X}^f$
$\overline{\mathbf{x}}_k^f$	Forecast ensemble mean at time k
$\overline{\mathbf{x}}^f$	Forecast ensemble mean
$\mathbf{x}_k^{f,i}$	Forecast member i at time k
$\mathbf{x}^f$	Forecast state vector
$\mathbf{x}$	State vector
$\mathbf{x}_0^t$	Initial true state vector
$\mathbf{x}_k^t$	True state vector at time k
$\mathbf{x}^t$	True state vector
$\mathbf{x}^{a,H,i}$	Hybrid analysis member i
$\overline{\mathbf{x}}^{a,H}$	Hybrid analysis ensemble mean
$\mathbf{Y}_{1:k}^o$	Set of observations from time 1 to time k
$\overline{y}_k^a$	Scalar mean of the analysis ensemble of observations, $\tilde{\mathbf{y}}_k^a$, at time k
$\overline{y}_k^p$	Scalar mean of the predicted ensemble of observations, $\tilde{\mathbf{y}}_k^p$, at time k
$\Delta\overline{y}_k$	Scalar observation mean increment at time k, the difference between $\overline{y}_k^a$ and $\overline{y}_k^f$
Δy_k^i	Scalar observation increment for the member i at time k
$\mathbf{y}_k^o$	Observation vector at time k
$\mathbf{y}_k^{o,i}$	Observation vector member i at time k
$\mathbf{y}^o$	Observation vector
$\mathbf{y}_k^p$	Predicted or expected observation vector at time k

$\mathbf{y}^p$	Predicted or forecasted observation vector
$y_k^{a,i}$	Scalar observation analysis member i at time k
y_k^o	Scalar observation at time k
$y_k^{o,i}$	Scalar observation member i at time k
$y_k^{p,i}$	Scalar predicted observation from the forecast member i at time k
$\mathbf{y}^t$	Representation of $\mathbf{x}^t$ in the observation space
$\tilde{\mathbf{y}}_k^a$	Analysis ensemble of observations at time k
$\tilde{\mathbf{y}}^p$	Ensemble of predicted observations
$\tilde{\mathbf{z}}_{(j)}$	Same as $\tilde{\mathbf{x}}_{(j)}$ for $1 \leq j \leq n$ and $\tilde{\mathbf{y}}^p$ for $n < j \leq n+p$
$\Delta\overline{\mathbf{z}}_{(j)}$	jth component of the joint state-observation space ensemble mean increment
$\Delta\mathbf{z}_{(j)}^i$	Increment added to the jth component of the joint state-observation space ensemble member i
$\overline{z}_{(j)}^f$	jth component of the joint forecast state-observation space ensemble mean, $\overline{\mathbf{z}}^f$
$\Delta\overline{\mathbf{z}}_k$	Joint state-observation space ensemble mean increment at time k
$\overline{\mathbf{z}}_k^a$	Joint analysis state-observation space ensemble mean at time k
$\overline{\mathbf{z}}^a$	Joint analysis state-observation space ensemble mean
$\mathbf{z}_k^{a,i}$	Joint analysis state-observation space member i at time k
$\mathbf{z}^{a,i}$	Joint analysis state-observation space member i
$\overline{\mathbf{z}}_k^f$	Joint forecast state-observation space ensemble mean at time k
$\overline{\mathbf{z}}^f$	Joint forecast state-observation space ensemble mean
$\mathbf{z}_k^{f,i}$	Joint forecast state-observation space member i at time k

$\mathbf{z}^{f,i}$	Joint forecast state-observation space member i
$\mathbf{z}_k$	Joint state-observation space vector at time k
$\mathbf{z}_k^t$	True joint state-observation space vector at time k

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

1.1 The Red Sea

The Red Sea is an extension of the Indian Ocean that resembles a channel running between Africa and Asia. The Bab-al-Mandeb in the South makes the connexion between the Red Sea and the Gulf of Aden which in turn connects it to the Indian Ocean. Two other choke points characterize the Red Sea: the Gulf of Aqaba in the North-East of the Red Sea and the Gulf of Suez in its North-West. The Red Sea therefore joins the Indian Ocean to the Mediterranean Sea and plays a key role in the Old World, serving central shipping trade routes between Africa, Asia and Europe.

The Red Sea water surface is around $17,000$ square miles. The basin length is approximately $2,000$ km and its average width is 280 km. Dozens of islands are scattered along its shores. The Red Sea has an average depth of 490 m and is quite shallow since 40% of the water is under 100 m. The maximum depth is more than $2,200$ m. The water estimated volume is close to $60,000$ cubic miles. Despite the narrowness of the Bab-al-Mandeb, there is a significant water exchange between the Red Sea and the Gulf of Aden [43, 189, 190], while the water flow between the Red Sea and the Mediterranean Sea remains very limited. The Red Sea is among the warmest and saltiest seas in the world [144] with an average salinity of 40 compared to the world seawaters average salinity of 35. Furthermore, its complex terrains and landforms are home to a distinctive ecological system rich of biodiversity [27] and accommodate magnificent coral reefs [43] that thrive despite high temperatures. The

Red Sea is also remarkable for its eddy activities [195]. Many research cruises explored and studied the Red Sea [28, 144], however it remains relatively poorly covered by observations compared to other seas, and many aspects of its physical circulation remain not fully understood, even though the Red Sea considerably contributes to the economic and social growths of the surrounding countries.

A rise in the population and the number of residential areas of the underdeveloped Red Sea region, a surge in economical activities as well as the new initiated mega-projects, NEOM, a futuristic city and the Red Sea Project, a tourist attraction, along the Saudi shores of the Red Sea, underpin the significance and the dynamism of the Red Sea. The environmental impact of these new developments need to be mitigated, through monitoring and intervention systems, to provide a sustainable economic growth. The prediction of the Red Sea circulation is a key component of such systems and would help anticipate extreme events and prevent environmental disasters. It would also assist day to day life, business and maritime operations.

Nowadays oceanographers make use of observations (in situ, satellites, ...) and numerical ocean general circulation models (OGCMs). While observations are the most straightforward way to study the ocean, they remain sparse both in space (because ships follow some given paths in the ocean, satellites coverage is restricted to surface tracks, fixed deployed instruments would be very costly for a dense coverage, ...) and time (because ships and satellites are moving and cannot continuously sample at a given fixed location) [167]. In contrast, numerical models can provide full spatial and temporal coverage through increasing resolution, but may lack for accuracy due to inevitable modeling errors. To benefit from both sources of information, observations and model outputs are combined in various ways, and data assimilation is one of the prominent approaches to do so.

1.2 Ocean data assimilation and forecasting

Observations and model simulations, employed to improve our understanding of the ocean, have their pros and cons (more reliable but sparse for the observations, arbitrary resolution, assuming resources availability, but modeling errors for the model). Data assimilation is a process that extracts the relevant information from both data sources and combine them into a better estimate of the system state [77].

Given measurements and model outputs of the system, there are several approaches to conduct an assimilation, historically categorized into variational (based on calculus of variation [15]) and statistical estimation (filtering) methods. In variational assimilation, grounded in optimal control, a cost function, generally defined as the misfit between the measurements and the model outputs, is built and an optimization procedure leads to the solution [103]. In the filtering methods, the measurements are generally processed sequentially [85] and a statistical estimation is performed to retrieve the solution along with its probability density function (PDF), when expressed in a Bayesian framework. The different assimilation methods might theoretically lead to the same solutions under certain assumptions, but in practice they do not [77].

The Kalman filter (KF) [91] is a special case of Bayesian filtering and estimates the first two moments of the PDF, which is equivalent to finding the PDF for a multivariate Gaussian distribution since it is characterized by its mean and its covariance [151]. Moreover, the KF provides the best linear unbiased estimator (BLUE) (minimum-variance) in the Gaussian (and linear) setup. The dynamical model is integrated and once observations become available, an update is performed, then a new assimilation cycle (forecast - update) can start.

Data assimilation finds application in various fields [15]: meteorology [47, 63, 64], hydrology [3, 65, 66, 111], physical oceanography [170, 171], glaciology [20, 107, 108, 177], marine biology [42, 44], land surface modeling, agroecology [29, 41, 89, 90], natural hazards, medicine, biology, chemistry, physical sciences (fluid dynamics, imaging, acoustics, mechanics), motion tracking (of airplanes, satellites, fluids, ...), human and social sciences (economics, finance, traffic control, urban planning), etc.

In oceanography, where the model outputs and the measurements dimensions are very large, the KF cannot be direcly implemented, because of the prohibitive size of the matrices required for the filter steps. Another issue is the nonlinearity of the involved dynamics and observation systems. Those prompted the introduction of simplifications and generalizations in the KF formulation [80, 142, 178]. The ensemble Kalman filter (EnKF) is the most known example. It is based on Monte Carlo (MC) estimates of the first two moments of the statistics of the Bayesian filtering solution using an ensemble of ocean states [50, 84]. Those forecasted model outputs are updated with incoming observations by applying the standard KF update procedure. The updated (or analysis) ensemble is then advanced, each member separately and in parallel, for computational efficiency. Large ensemble might be needed for better statistical representation [78, 82] required to estimate the covariance matrix and deliver good filter updates, forecasts and predictions.

Non-Gaussian filtering, an active area of research, is theoretically a more sound approach to deal with the nonlinearities from the models, the observation operator, and more generally when the involved PDFs are not Gaussian [77]. Particle filters and Gaussian-mixture filters are examples of non-Gaussian filters. These approximate the prior PDF with weighted Dirac or Gaussian kernels, respectively [23, 77] and are still in development stage.

1.3 Thesis Objectives

The advent of new governmental projects in the region and on the shores of the Red Sea (like the Neom project) and the desire of the authorities to develop new economic hubs is likely to turn things around by boosting the commissioning of a Red Sea operational forecasting system. Such a system is relevant for Saudi Aramco offshore operations in particular, as well as for fisheries and other sea activities (civil and military) in general. Examples of such activities are, but not limited to, under water communications [135, 152] for which the knowledge of water properties (salinity temperature, currents, ...) is needed, and water desalination [37, 38, 39]. The system will also be useful for scientific studies to support the aforenamed activities. These include exploration, examination and analysis of phytoplankton blooms [43, 146], eddies properties and predictability [195, 196], water masses circulation [138, 189, 190], internal waves generation [70], and wind above the Red Sea [99]. Even though the system is first intended for the Red Sea, it could be applied to other water masses making its implementation very useful and of utmost importance.

Despite being among the busiest and most important shipping routes, an operational forecasting system is yet to be developed for the Red Sea. Most of the reported studies of the basin rely on numerical simulations, due to a poor observational coverage of the Red Sea, and used either a simplified model [172, 191, 192], a climatology of general circulation model (GCM) outputs [48, 158, 163, 164], or an OGCM forced with real atmospheric conditions [29, 173, 189, 190, 196]. With more and more data becoming available [191, 195], data assimilation may spring up new discoveries about the Red Sea. Until recently, only one work [31] implemented a primitive assimilation system of the Red Sea, based on an elementary nudging technique assimilating SST and eXpendable BathyThermograph (XBT)/Conductivity Temperature Depth (CTD) data, with a coarse model.

The goal of this thesis is therefore to develop the basis of an operational data assimilation system for forecasting the Red Sea circulation with the established Saudi Aramco Marine Environmental Research Center at KAUST (SAMERCK) (https://iop.kaust.edu.sa). In order to tackle the aforementioned challenges and advance the science, I put my efforts into building, implementing and validating efficient data assimilation schemes with a high-resolution OGCM of the Red Sea, by working on:

- The implementation of a state-of-the-art assimilation system, configured here for the Red Sea. A similar system developed for the Gulf of Mexico was available and our contribution is an adaptation for the Red Sea capable of assimilating various kind of Red Sea collected data, ranging from in situ observations to satellite measurements, where forecasted model elements called members are integrated in parallel, within a High Performance Computing (HPC) environment.

- Enhancing the robustness of the system and make it fault-tolerant, and proposing new data assimilation validation metrics. In a HPC context featuring thousands of cores and aiming at running an ever increasing number of members the risk of failure is even larger. To mitigate these issues interrupting the whole assimilation process, the system should be able to restart by discarding the faulty cores and members. As for the new validation metrics, they should complement the existing ones, and provide simpler implementation and robust verification of the system behavior.

- Pushing the limits of the current ensemble assimilation systems by running the largest ocean ensemble assimilation to date and exploring its impact. Current operational ocean assimilation systems run 50 - 100 members, which may limit those systems performance and often requires the introduction of auxiliary techniques and simplifications. Here we run 1000-member experiments and study their impacts in terms of improvements in the system estimate of the ocean state and reliance on auxiliary techniques and simplifications.

- Reducing the computational load of the system by developing new assimilation schemes tailored for the seasonal variability of the Red Sea. Ensemble assimilation require all the members to be advanced by the model. By advancing only one member, we save the cost incurred by the model integrations and to compensate for the loss of the covariance built from the members, seasonally generated covariances expected to account for the seasonal variability features of the Red Sea are employed.

- Adaptive schemes that adaptively select the desired ensemble members for the seasonal covariances construction. We propose enhanced seasonal schemes that automatically select the ensemble members from a dictionary of ensemble realizations based on some metrics, at each assimilation step.

- A Hybrid scheme that combines the benefits of several schemes while leveraging the resources for cost efficiency. The Hybrid scheme uses a hybrid covariance, a linear combination of a dynamic covariance obtained from the ensemble adjustment Kalman filter (EAKF) and a static covariance. It propagates sufficient ensemble members to incorporate the model dynamic, while thresholding the computational resources and complement the dynamic covariance with a static one.

All the developed schemes will be implemented and tested for data assimilation and forecasting in the Red Sea, and their performance evaluated with the goal of developing the first operational system for the Red Sea.

1.4 Thesis Outline

Chapter 2 presents the DART-MITgcm assimilation system. After introducing the Bayesian filtering and describing the different components of the system, a specific implementation for the Red Sea configuration is discussed. Chapter 3 examines the sensitivity of DART-MITgcm to some assimilation schemes and atmospheric forcing. Then Chapter 4 analyzes the results of the first 1000-member ocean ensemble data assimilation run and demonstrates the robustness of the system. Chapter 5 explores the results of the application of adaptive ensemble Optimal Interpolation schemes, as a consequence of the outcomes of Chapter 3. Finally, Chapter 6 studies the impact of a Hybrid assimilation scheme, which is the combination of the new schemes, on the performance of the assimilation system. Chapter 7 provides concluding remarks and a proposal for a future research work.

Chapter 2

DART-MITgcm Assimilation System

This Chapter presents the Bayesian filtering and the estimation problem (Section 2.1). A description of the ocean general circulation model (MITgcm) is given in Section 2.2 and that of the ensemble assimilation package (DART) in Section 2.3. Section 2.4 focuses on a specific configuration of the DART-MITgcm assimilation system for the Red Sea.

2.1 Estimation problem, Bayesian filtering and ensemble Kalman filtering

Let $\mathbf{x}^t$ be the true state of the system of interest (the ocean here) we would like to estimate and predict. Usually, only limited (sparse in space and time) noisy measurements $\mathbf{y}^o$ related to $\mathbf{x}^t$, directly or indirectly through some conversions or transformations, are available. One way to complement the information from the observations is the use of a numerical model that can provide an estimate $\mathbf{x}$ of the system state at any time and any location, provided that the computational resources are enough to handle high (1-2 km) resolution simulations. Like the observational data, models are prone to uncertainties. So we end up with two (uncertain) sources of information for the unknown true state $\mathbf{x}^t$. We will therefore model $\mathbf{x}^t$ (and $\mathbf{x}$, and $\mathbf{y}^o$ as well) as a stochastic process, i.e. a random variable indexed with a (time) parameter. For more details about the relationship between $\mathbf{x}^t$, $\mathbf{x}$, and $\mathbf{y}^o$, the reader is referred to Appendix A.

Now, let us define the state space model as

$$\mathbf{x}^t_{k+1} = \mathcal{M}_k(\mathbf{x}^t_k) + \boldsymbol{\eta}_k\,, \tag{2.1}$$

$$\mathbf{y}^o_k = h_k(\mathbf{x}^t_k) + \boldsymbol{\varepsilon}_k\,, \tag{2.2}$$

where $\mathbf{x}^t_k$ is the true state of the system at time k, $\mathcal{M}_k$ is a dynamical (forward) model to evolve the system in time, $\mathbf{y}^o_k$ is an observation or measurement of the system at time k, h_k is the observation operator at time k, $\boldsymbol{\eta}_k$ and $\boldsymbol{\varepsilon}_k$ are model and observation errors, generally assumed Gaussian.

In a Bayesian framework we assign a probability distribution $p(\mathbf{x}^t_0)$ to the initial state $\mathbf{x}^t_0$. The goal will be to compute the probability distribution $p(\mathbf{x}^t_k|\mathbf{Y}^o_{1:k})$, that is the conditional probability distribution of the state vector $\mathbf{x}^t_k$ given all the available observations up to time k, $\mathbf{Y}^o_{1:k} := \mathbf{y}^o_1, \mathbf{y}^o_2, \cdots, \mathbf{y}^o_k$, by applying Bayes' rule [170, 82]. To recursively solve the problem, the transition from $p(\mathbf{x}^t_{k-1}|\mathbf{Y}^o_{1:k-1})$ to $p(\mathbf{x}^t_k|\mathbf{Y}^o_{1:k})$ is made by first advancing the system state probability distribution with the dynamical model (2.1), that is applying the Chapman-Kolmogorov equation

$$p(\mathbf{x}^t_k|\mathbf{Y}^o_{1:k-1}) = \int p(\mathbf{x}^t_k|\mathbf{x}^t_{k-1})p(\mathbf{x}^t_{k-1}|\mathbf{Y}^o_{1:k-1})d\mathbf{x}^t_{k-1}. \tag{2.3}$$

Next, equation (2.2) allows us to compute $p(\mathbf{y}^o_k|\mathbf{x}^t_k)$ and Bayes' rule application results in

$$p(\mathbf{x}^t_k|\mathbf{Y}^o_{1:k}) = \frac{p(\mathbf{y}^o_k|\mathbf{x}^t_k)p(\mathbf{x}^t_k|\mathbf{Y}^o_{1:k-1})}{p(\mathbf{y}^o_k|\mathbf{Y}^o_{1:k-1})}. \tag{2.4}$$

Equations (2.3) and (2.4) are the assimilation forecast step and the assimilation update step, respectively. A forecast followed by an update (or an update followed by a forecast) represent one assimilation cycle. After an update, the model runs until the availability of new observations. Then another update is performed and a new assimilation cycle can start.

In the Introduction (Section 1.2), the KF was presented as a specific Bayesian filter and the EnKF as a simplification of the KF for implementation with OGCMs. There are many variants of the EnKF among which the stochastic EnKF where the observations are perturbed [26]. The stochastic EnKF help to reduce the underestimation of the analysis error covariance [184]. The algorithm of the stochastic EnKF is outlined below. First, since we do not have access to $\mathbf{x}^t$, we use $\mathbf{x}$ instead, for practical applications. Also, we define $\mathbf{x}^a$ as the state $\mathbf{x}$ after an assimilation update and we call it the analysis state, and $\mathbf{x}^f$ as the state $\mathbf{x}$ after it has been advanced by the dynamical model $\left(\mathbf{x}_k^f = \mathcal{M}_{k-1}(\mathbf{x}_{k-1}^a)\right)$ and we refer to it as the forecast state. Let us consider a set of model initial conditions:

$$\mathbf{X}_{k-1}^a = [\mathbf{x}_{k-1}^{a,1}, \mathbf{x}_{k-1}^{a,2}, \cdots, \mathbf{x}_{k-1}^{a,N}]. \tag{2.5}$$

The superscript ^{a} stands for analysis, and the numbers are the members ranks. The ensemble $\mathbf{X}_{k-1}^a$ is then advanced with the dynamical model (which is an approximation of equation (2.3) because a finite number of members are being integrated) to provide an ensemble of forecasts:

$$\mathbf{X}_k^f = [\mathbf{x}_k^{f,1}, \mathbf{x}_k^{f,2}, \cdots, \mathbf{x}_k^{f,N}]. \tag{2.6}$$

Next, an ensemble of anomalies $\boldsymbol{X_k'}$ is computed:

$$\boldsymbol{X_k'} = [\mathbf{x}_k^{f,1} - \overline{\mathbf{x}}_k^f,\, \mathbf{x}_k^{f,2} - \overline{\mathbf{x}}_k^f,\, \cdots, \mathbf{x}_k^{f,N} - \overline{\mathbf{x}}_k^f], \tag{2.7}$$

where $\overline{\mathbf{x}}_k^f = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mathbf{x}_k^{f,i}$ is the ensemble mean. The forecast error covariance matrix $\mathbf{P}_k^f$ is approximated by:

$$\mathbf{P}_k^{f,e} = \frac{1}{N-1}\left(\boldsymbol{X_k'}\boldsymbol{X_k'}^T\right), \tag{2.8}$$

and the Kalman Gain is evaluated as:

$$\mathbf{K}_k = \left(\mathbf{H}_k \mathbf{P}_k^{f,e}\right)^T \left[\mathbf{H}_k \left(\mathbf{H}_k \mathbf{P}_k^{f,e}\right)^T + \mathbf{R}_k\right]^{-1}, \tag{2.9}$$

with $\mathbf{R}_k$ the observational covariance matrix associated with the observation $\mathbf{y}_k^o$, and $\mathbf{H}_k$ the matrix associated with the observation operator h_k. The filter update is computed by applying the KF update to each of the forecast ensemble members with a perturbed observation:

$$\mathbf{x}_k^{a,i} = \mathbf{x}_k^{f,i} + \mathbf{K}_k \left(\mathbf{y}_k^o + \boldsymbol{\varepsilon}_k^i - h_k(\mathbf{x}_k^{f,i})\right), \quad i = 1, \cdots, N, \tag{2.10}$$

where $\boldsymbol{\varepsilon}_k^i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_k)$ is the perturbation of size p_k added to the observation $\mathbf{y}_k^o$ to yield the perturbed observation $\mathbf{y}_k^o + \boldsymbol{\varepsilon}_k^i$. Next, from the new ensemble $\mathbf{X}^a$, a new assimilation cycle is conducted by repeating the different steps from equation (2.5).

2.2 The MIT general circulation ocean model (MITgcm) and its Red Sea configuration

The Massachusetts Institute of Technology general circulation model (MITgcm) is a software implementing both atmospheric and ocean general circulation models [1, 121]. Based on the Navier Stokes equations, it employs finite volume for discretization [122] while irregular geometries are dealt with orthogonal curvilinear grids and shaved cells [2]. Its non-hydrostatic capability allows for studying small-scale as well as large-scale processes [123]. Optimization and sensitivity studies are also possible through its tangent linear and adjoint models [120]. The model runs efficiently on several computational platforms and domain decomposition makes it convenient for parallel processing on HPC machines by slicing the ocean domain into vertical columns [122].

In our configuration designed for the Red Sea, the model covers the Red Sea, the Gulf of Suez in the North-West, the Gulf of Aqaba in the North-East and the Gulf of

Aden in the South. The grid of the discretized domain is a 20° × 20° eddy-resolving spherical polar grid, 30°-50°E and 10°-30°N, with a horizontal grid spacing of 0.04° (approximately 4 km), and 50 vertical layers with increasing thickness, starting from 4 m at the surface and reaching 300 m near the bottom. The bathymetry is from the General Bathymetric Chart of the Oceans (GEBCO) (https://www.gebco.net/data_and_products/gridded_bathymetry_data/). The momentum, tracer and free surface equations time steps are set to 200 seconds. The baroclinic multi-stage time stepping is activated. The configuration uses implicit, variable, harmonic and biharmonic horizontal viscosities, a K-Profile Parametrization (KPP) scheme [101] for the vertical mixing, as well as implicit diffusion and implicit, non-linear free surface. The equation of state is a modified UNESCO [55] formula by Jackett and McDougall [88]. The model is run in hydrostatic mode, with exact volume conservation and flux-form Coriolis scheme. A 3^{rd} order advection scheme is applied for temperature and salinity.

2.3 The Data Assimilation Research Testbed (DART)

The Data Assimilation Research Testbed (DART) is a Fortran software implemented at the National Center for Atmospheric Research (NCAR). To cope with large atmospheric and oceanographic models, DART assimilates the observations at a given time one by one and in parallel using a two-steps assimilation strategy [9]. Some auxiliary techniques, such as localization [7, 11] and inflation [10, 12] are also available to enhance assimilation results with small ensembles. Different filters are implemented in DART, the main one being the EAKF [8].

After introducing the joint state space formalism, a clear description on how the EAKF and the EnKF are derived in the two-steps assimilation framework based on [8] is given.

2.3.1 The joint state-observation space nonlinear filter

The true joint state-observation space vector (or true joint state vector) at time k is the vector $\mathbf{z}_k^t = [\mathbf{x}_k^t, h_k(\mathbf{x}_k^t)]$ which length $l_k = n + p_k$ is the sum of the state vector length n and the observation vector size p_k. Similarly we define the joint state-observation space vector $\mathbf{z}_k = [\mathbf{x}_k, h_k(\mathbf{x}_k)]$. It is useful for handling any kind of observation operator, especially nonlinear ones. The Bayesian filter equations (2.3) and (2.4) are then updated by replacing $\mathbf{x}_k^t$ by $\mathbf{z}_k^t$:

$$p(\mathbf{z}_k^t|\mathbf{Y}_{1:k-1}^o) = \int p(\mathbf{z}_k^t|\mathbf{z}_{k-1}^t)p(\mathbf{z}_{k-1}^t|\mathbf{Y}_{1:k-1}^o)d\mathbf{z}_{k-1}^t, \tag{2.11}$$

$$p(\mathbf{z}_k^t|\mathbf{Y}_{1:k}^o) = \frac{p(\mathbf{y}_k^o|\mathbf{z}_k^t)p(\mathbf{z}_k^t|\mathbf{Y}_{1:k-1}^o)}{p(\mathbf{y}_k^o|\mathbf{Y}_{1:k-1}^o)}. \tag{2.12}$$

2.3.2 Computation of the filtering solution using ensemble Kalman filters (EnKF, EAKF)

When solving (2.12) using ensemble methods, we first sample an ensemble from the prior distribution $p(\mathbf{x}_k^t|\mathbf{Y}_{1:k-1}^o)$ then by applying the observation operators h_k we get an ensemble for the joint prior distribution $p(\mathbf{z}_k^t|\mathbf{Y}_{1:k-1}^o)$. So let us sample an ensemble of forecast states $[\mathbf{z}_k^{f,1}, \mathbf{z}_k^{f,2}, \cdots, \mathbf{z}_k^{f,N}]$ with mean $\overline{\mathbf{z}}_k^f$ and sample covariance $\mathbf{\Sigma}_k^f$. When assuming the prior to be Gaussian, then the updated or analysis covariance is

$$\mathbf{\Sigma}_k^a = \left[(\mathbf{\Sigma}_k^f)^{-1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k\right]^{-1}, \tag{2.13}$$

and the analysis mean is

$$\overline{\mathbf{z}}_k^a = \mathbf{\Sigma}_k^a\left[(\mathbf{\Sigma}_k^f)^{-1}\overline{\mathbf{z}}_k^f + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{y}_k^o\right]. \tag{2.14}$$

$\mathbf{H}_k$ is here the matrix associated with the extended observation operator, that is the observation operator that applies to $\mathbf{z}_k$, which is merely a projection from the joint space of dimension l_k onto the observation space $\mathcal{O}_{p_k}$ of dimension p_k. The expected observation value from the joint state vector is therefore $\mathbf{y}_k^p = \mathbf{H}\mathbf{z}_k$.

The filters in DART yield a solution having the mean and covariance expressed by (2.13) and (2.14) (and eventually an additional weight for the kernel filter not discussed here). More details for the EnKF and EAKF filters are given below.

2.3.2.1 Ensemble Kalman filter

DART implements the traditional EnKF with perturbed observations [8, 84]. It generates N perturbed observations $\{\mathbf{y}_k^{o,i}\}_{i=1,\dots,N}$ by sampling the errors (like the $\boldsymbol{\varepsilon}_k^i$ of (2.10)) from the observational distribution $p(\mathbf{y}_k^o \,|\, \mathbf{z}_k^t)$ in (2.12) and adding them to the observation $\mathbf{y}_k^o$. The observations $\{\mathbf{y}_k^{o,i}\}_{i=1,\dots,N}$ mean is adjusted to be $\mathbf{y}_k^o$. $\boldsymbol{\Sigma}_k^a$ is theoretically computed once for all the members, using equation (2.13). To obtain the analysis ensemble $\mathbf{Z}_k^a = [\mathbf{z}_k^{a,1}, \mathbf{z}_k^{a,2}, \cdots, \mathbf{z}_k^{a,N}]$, $\overline{\mathbf{z}}_k^f$ and $\mathbf{y}_k^o$ are replaced by $\mathbf{z}_k^{f,i}$ and $\mathbf{y}_k^{o,i}$, respectively, in equation (2.14) leading to N evaluations of the equation.

2.3.2.2 Ensemble adjustment Kalman filter

The ensemble adjustment Kalman filter is part of the EnKF family. The adjustment refers to the fact that the updated ensemble is adjusted to match the theoretical updated mean (2.13) and covariance (2.14). For that purpose, a linear operator $\mathbf{A}_k$ is applied to the prior ensemble to recover the updated ensemble

$$\mathbf{z}_k^{a,i} = \mathbf{A}_k(\mathbf{z}_k^{f,i} - \overline{\mathbf{z}}_k^f) + \overline{\mathbf{z}}_k^a, \qquad i = 1, \dots, N \tag{2.15}$$

where $\mathbf{z}_k^{f,i}$ and $\mathbf{z}_k^{a,i}$ are respectively the i-th forecast and i-th analysis ensemble member at time k, and the matrix $\mathbf{A}_k$ (of size $l_k \times l_k$) is chosen such that the updated ensemble sample covariance matches the one in (2.13). The detailed computation for a suitable operator $\mathbf{A}_k$ is provided in Appendix B.

2.3.3 Filtering with the two-steps assimilation framework of DART

In practice, equations (2.13) and (2.14) cannot be used to compute the filters update due to the huge sizes of the matrices involved. For the same reason, the adjustment operator $\mathbf{A}_k$ is not appropriate to compute the EAKF filter update, hence the motivation for introducing the two-steps assimilation and make the assimilation implementation possible in DART [9].

2.3.3.1 Two-steps assimilation

Recall that our goal is to compute the distribution of the state conditioned on the observations up to time k (2.4). The joint state-observation space enables us to establish a relation between the distribution of the state vector and the distribution of the predicted observation, through the distribution of the joint state-observation distribution (2.12). The idea of the two-steps assimilation is to first compute the marginal distribution of the predicted observation and then take advantage of the relation provided by the full distribution to derived the distribution of the state [9]. So the predicted observation is first updated then the correction is propagated to the remaining state variables. Here a Gaussian relationship is assumed between the joint state variables to allow for the connection.

We will now present the details of the method. For the sake of simplification, let's drop the time index and assume we have a scalar observation, $\mathbf{y}_k^o = [y^o]$. That is $p_k = p = 1$ and $l_k = l = n + 1$. Indeed, this assumption results from the fact that uncorrelated observations can be processed and assimilated in a sequential fashion [9]. And if there are more than one observation (i.e. $l > 1$), the observations are assimilated either serially or in parallel [14].

Let the distribution of the joint forecast state-observation be Gaussian with mean $\overline{\mathbf{z}}^f$ and covariance $\mathbf{\Sigma}^f$, and the observation distribution be Gaussian with mean $\mathbf{y}^o$ and covariance $\mathbf{R} = [r]$. Because we assumed a single observation, $\mathbf{H} = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}$ is a $1 \times l$ matrix.

$$\mathbf{\Sigma}^f = \left(\begin{array}{cccc|c} & & & & \sigma^f_{1,l} \\ & & & & \sigma^f_{2,l} \\ & \mathbf{P}^{f,e} & & & \vdots \\ & & & & \vdots \\ & & & & \sigma^f_{l-1,l} \\ \hline \sigma^f_{1,l} & \sigma^f_{2,l} & \cdots\ \cdots & \sigma^f_{l-1,l} & \sigma^f_{l,l} \end{array}\right).$$

$\sigma^f_{j,l}, \quad j = 1, \ldots, l$ is defined as $\sigma^f_{j,l} = cov(\tilde{\mathbf{z}}_{(j)}, \tilde{\mathbf{y}}^p), \quad j = 1, \ldots, l$, and

$\tilde{\mathbf{z}}_{(j)}$, for $j = 1, \ldots, l-1$, as the jth row of the ensemble (or matrix) $\mathbf{X}^f$, which contains the jth component of each of the ensemble members, i.e. for $j = 1, \ldots, l-1$, $\tilde{\mathbf{z}}_{(j)} = \tilde{\mathbf{x}}_{(j)} = \begin{bmatrix} x^{f,1}_{(j)}, & x^{f,2}_{(j)}, & \cdots & \cdots, & x^{f,N}_{(j)} \end{bmatrix}$,
and for $j = l$,

$$\tilde{\mathbf{z}}_{(l)} = \tilde{\mathbf{y}}^p = \begin{bmatrix} y^{p,1}, & y^{p,2}, & \cdots, & y^{p,N} \end{bmatrix} = \begin{bmatrix} h\left(\mathbf{x}^{f,1}\right), & h\left(\mathbf{x}^{f,2}\right), & \cdots, & h\left(\mathbf{x}^{f,N}\right) \end{bmatrix} \tag{2.16}$$

is an ensemble of predicted observations.

2.3.3.2 First step: Observation update

At this stage $\mathbf{\Sigma}^a$ ($\mathbf{\Sigma}^f$ respectively) is replaced by the analysis variance $\sigma^a_{l,l}$ (forecast variance $\sigma^f_{l,l}$ respectively) of the ensemble of predicted observations $\tilde{\mathbf{y}}^p$ in equation (2.13). And equation (2.14) is used to update the mean of the predicted ensemble of observations ($\overline{y}^p = \dfrac{1}{N}\sum_{i=1}^{N} y^{p,i}$), before adjusting the predicted ensemble of observations, for the EAKF, and updating each member of the predicted ensemble of observations, for the EnKF. Therefore $\overline{\mathbf{z}}^f$ is replaced by $\overline{y}^p$ (for the EAKF) and $y^{p,i}, \quad i = 1, \ldots, N$ (for the EnKF) to yield the analysis ensemble of observations $\tilde{\mathbf{y}}^a = \begin{bmatrix} y^{a,1}, & y^{a,2}, & \cdots, & y^{a,N} \end{bmatrix}$. Additionally, for the EnKF, y^o is replaced by a perturbed ensemble of observations, $\{y^{o,i}, \quad i = 1, \ldots, N\}$.

That is:

$$\sigma^a_{l,l} = \left[(\sigma^f_{l,l})^{-1} + r^{-1} \right]^{-1} \tag{2.17}$$

and for the EAKF, first update the predicted observation ensemble mean:

$$\begin{aligned} \overline{y}^a &= \sigma^a_{l,l} \left[(\sigma^f_{l,l})^{-1} \overline{y}^p + r^{-1} y^o \right] \\ &= \sigma^a_{l,l} \left[\frac{1}{\sigma^f_{l,l}} \overline{y}^p + \frac{1}{r} y^o \right] \end{aligned} \tag{2.18}$$

then, for $i = 1, \ldots, N$:

$$y^{a,i} = \theta \left(y^{p,i} - \overline{y}^p \right) + \overline{y}^a, \tag{2.19}$$

where $\theta = \left(\dfrac{\sigma^a_{l,l}}{\sigma^f_{l,l}} \right)^{1/2} = \left(\dfrac{r}{r + \sigma^f_{l,l}} \right)^{1/2}$,

or for the EnKF, for $i = 1, \ldots, N$:

$$y^{a,i} = \sigma^a_{l,l} \left[\frac{1}{\sigma^f_{l,l}} y^{p,i} + \frac{1}{r} y^{o,i} \right]. \tag{2.20}$$

Afterwards, the observation increments are computed as:

$$\Delta y^i = y^{a,i} - y^{p,i}, \quad i = 1, \ldots, N. \tag{2.21}$$

2.3.3.3 Second step: Remaining state variables update

$\mathbf{\Sigma}^a$ from (2.13) can be written (by factoring out $(\mathbf{\Sigma}^f)^{-1}$ on the left and using the fact that $(AB)^{-1} = B^{-1}A^{-1}$):

$$\mathbf{\Sigma}^a = \left[\mathbf{I} - \frac{1}{r + \sigma^f_{l,l}} \mathbf{\Sigma}^f_{0l}\right] \mathbf{\Sigma}^f, \tag{2.22}$$

with $\mathbf{\Sigma}^f_{0l}$ being the matrix $\mathbf{\Sigma}^f$ in which all the elements have been set to 0, except the last column, that is

$$\mathbf{\Sigma}^f_{0l} = \left(\begin{array}{ccccc|c} & & & & & \sigma^f_{1,l} \\ & & & & & \sigma^f_{2,l} \\ & & 0 & & & \vdots \\ & & & & & \vdots \\ & & & & & \sigma^f_{l-1,l} \\ \hline 0 & 0 & \cdots & \cdots & 0 & \sigma^f_{l,l} \end{array}\right)$$

Replacing $\mathbf{\Sigma}^a$ by (2.22) in (2.14) yields:

$$\overline{\mathbf{z}}^a = \left[\mathbf{I} - \frac{1}{r + \sigma^f_{l,l}} \mathbf{\Sigma}^f_{0l}\right] \mathbf{\Sigma}^f \left[(\mathbf{\Sigma}^f)^{-1}\overline{\mathbf{z}}^f + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o\right]. \tag{2.23}$$

Because $\mathbf{\Sigma}^f\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o = \mathbf{\Sigma}^f_{0l}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}^o$ and $\mathbf{\Sigma}^f_{0l}\mathbf{\Sigma}^f_{0l} = \sigma^f_{l,l}\mathbf{\Sigma}^f_{0l}$,

$$\overline{\mathbf{z}}^a = \left[\mathbf{I} - \frac{1}{r + \sigma^f_{l,l}} \mathbf{\Sigma}^f_{0l}\right] \overline{\mathbf{z}}^f + \frac{1}{r + \sigma^f_{l,l}} \mathbf{\Sigma}^f_{0l}\mathbf{H}^T\mathbf{y}^o. \tag{2.24}$$

Therefore the increment produced by the assimilation of one observation is:

$$\begin{aligned}
\Delta\overline{\mathbf{z}} &= \overline{\mathbf{z}}^a - \overline{\mathbf{z}}^f \\
&= \frac{1}{r+\sigma^f_{l,l}}\mathbf{\Sigma}^f_{0l}\left(\mathbf{H}^T\mathbf{y}^o - \overline{\mathbf{z}}^f\right) \\
&= \frac{1}{r+\sigma^f_{l,l}}\left(y^o - \overline{z}^f_{(l)}\right)\begin{bmatrix}\sigma^f_{1,l}\\ \sigma^f_{2,l}\\ \vdots\\ \vdots\\ \sigma^f_{l,l}\end{bmatrix}
\end{aligned} \tag{2.25}$$

with $\overline{z}^f_{(l)} = \overline{y}^p$, the lth component of the vector $\overline{\mathbf{z}}^f$.

In order to perform the update of the remaining state variables from the observation increment derived in section 2.3.3.2, let us consider the jth component of the vector $\Delta\overline{\mathbf{z}}$. From (2.25), $\Delta\overline{\mathbf{z}}_{(j)} = \dfrac{1}{r+\sigma^f_{l,l}}\left(y^o - \overline{z}^f_{(l)}\right)\sigma^f_{j,l}, \quad j = 1,\ldots,l-1$ and $\Delta\overline{\mathbf{z}}_{(l)} = \dfrac{1}{r+\sigma^f_{l,l}}\left(y^o - \overline{z}^f_{(l)}\right)\sigma^f_{l,l} = \Delta\overline{y}$.

For $j = 1,\ldots,l-1, \quad \dfrac{\Delta\overline{\mathbf{z}}_{(j)}}{\Delta\overline{\mathbf{z}}_{(l)}} = \dfrac{\sigma^f_{j,l}}{\sigma^f_{l,l}}$ so

$$\begin{aligned}
\Delta\overline{\mathbf{z}}_{(j)} &= \frac{\sigma^f_{j,l}}{\sigma^f_{l,l}}\Delta\overline{\mathbf{z}}_{(l)} \\
&= \frac{\sigma^f_{j,l}}{\sigma^f_{l,l}}\Delta\overline{y}
\end{aligned} \tag{2.26}$$

and to each entry j of the ensemble member $\mathbf{z}^{f,i}, \quad i = 1,\ldots,N,$ the increment

$$\Delta\mathbf{z}^i_{(j)} = \frac{\sigma^f_{j,l}}{\sigma^f_{l,l}}\Delta y^i, \quad j = 1,\ldots,l-1 \tag{2.27}$$

is added to obtain the analysis ensemble $\{\mathbf{z}^{a,i}, \quad i = 1,\ldots,N\}$.

2.4 DART-MITgcm Implementation for the Red Sea

2.4.1 Workflow

DART-MITgcm assimilation system has first been implemented for the Gulf of Mexico [76]. It combines the MITgcm and the DART filter in successive update-forecast cycles. The forecast step of the assimilation is performed by the MITgcm model configured for the Red Sea and the analysis step by the DART filter. Within our DART-MITgcm configuration for the Red Sea, because a parallel MPI job cannot run another parallel MPI job for most MPI libraries, we run the DART filter and MITgcm as separate executables which exchange data through files (see Figure 2.1). Figure 2.1 shows DART flow from the DAReS Perspective (see Section **Schematic of Ensemble Data Assimilation - from the DAReS Perspective** at http://www.image.ucar.edu/DAReS/DART_classic/) and provides an overview of the interaction between the filter and the model.

DART-MITgcm workflow for the Red Sea is as follows (Figure 2.2): first, there is the initialization with the preprocessing of the observations (`obs_seq.out`), the generation of the initial ensemble (`filter_ics`), the specification of the filter namelist (`input.nml`), the MITgcm data namelist (`data`), the MITgcm date and calendar information (`data.cal`), and eventually additional files listed in `input.nml`. Next is the analysis (assimilation) step followed by the forecast step (advancing the model). After that, either a new assimilation cycle starts or the process ends, depending on the number of cycles stated in the filter namelist.

2.4.1.1 System configuration and initialization

Before running the assimilation system, we prepare the input files for each assimilation window. The preparation includes the files for the observations that will be assimilated, the initial conditions for the model states and the different run parame-

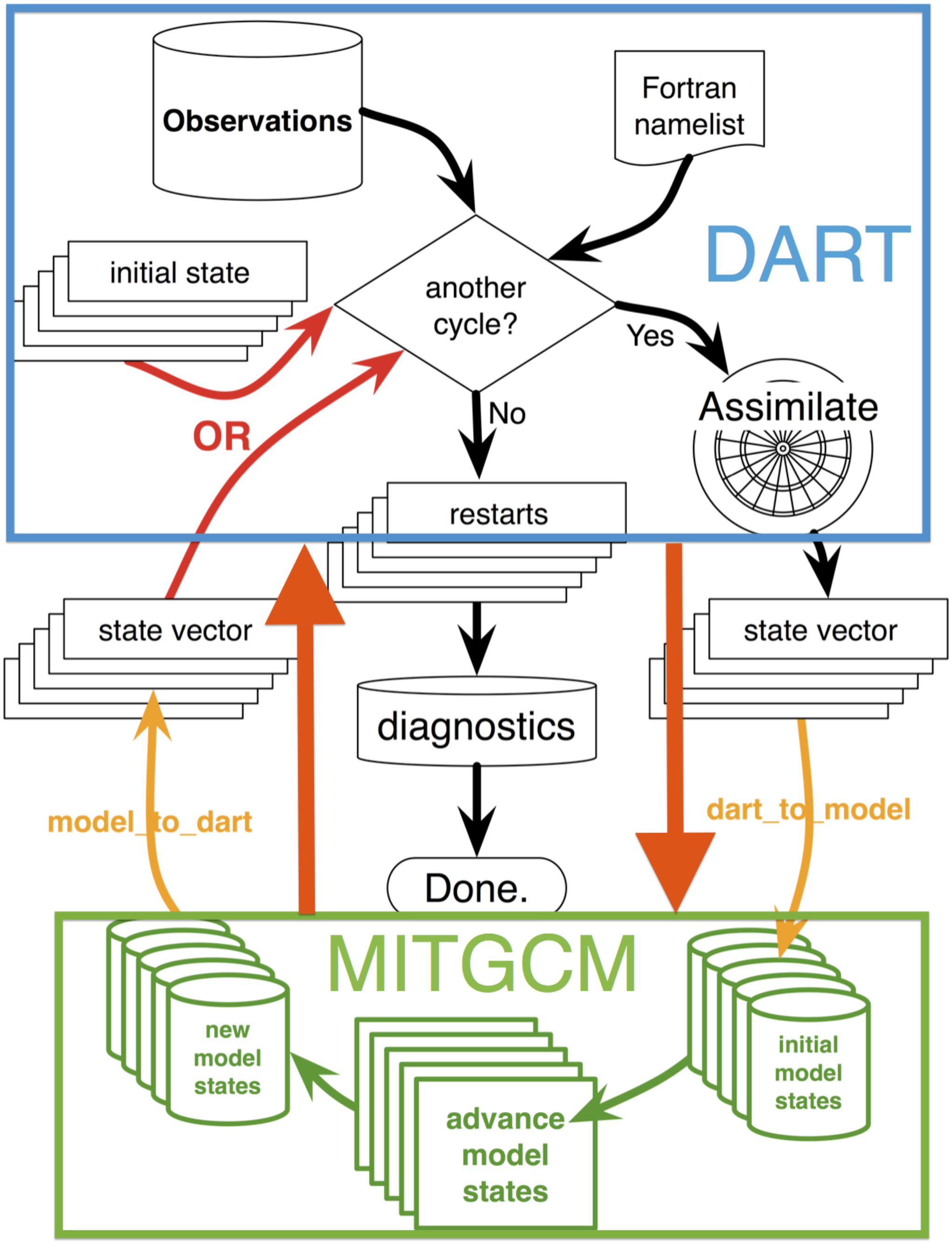


Figure 2.1: Coupling between DART and MITgcm.

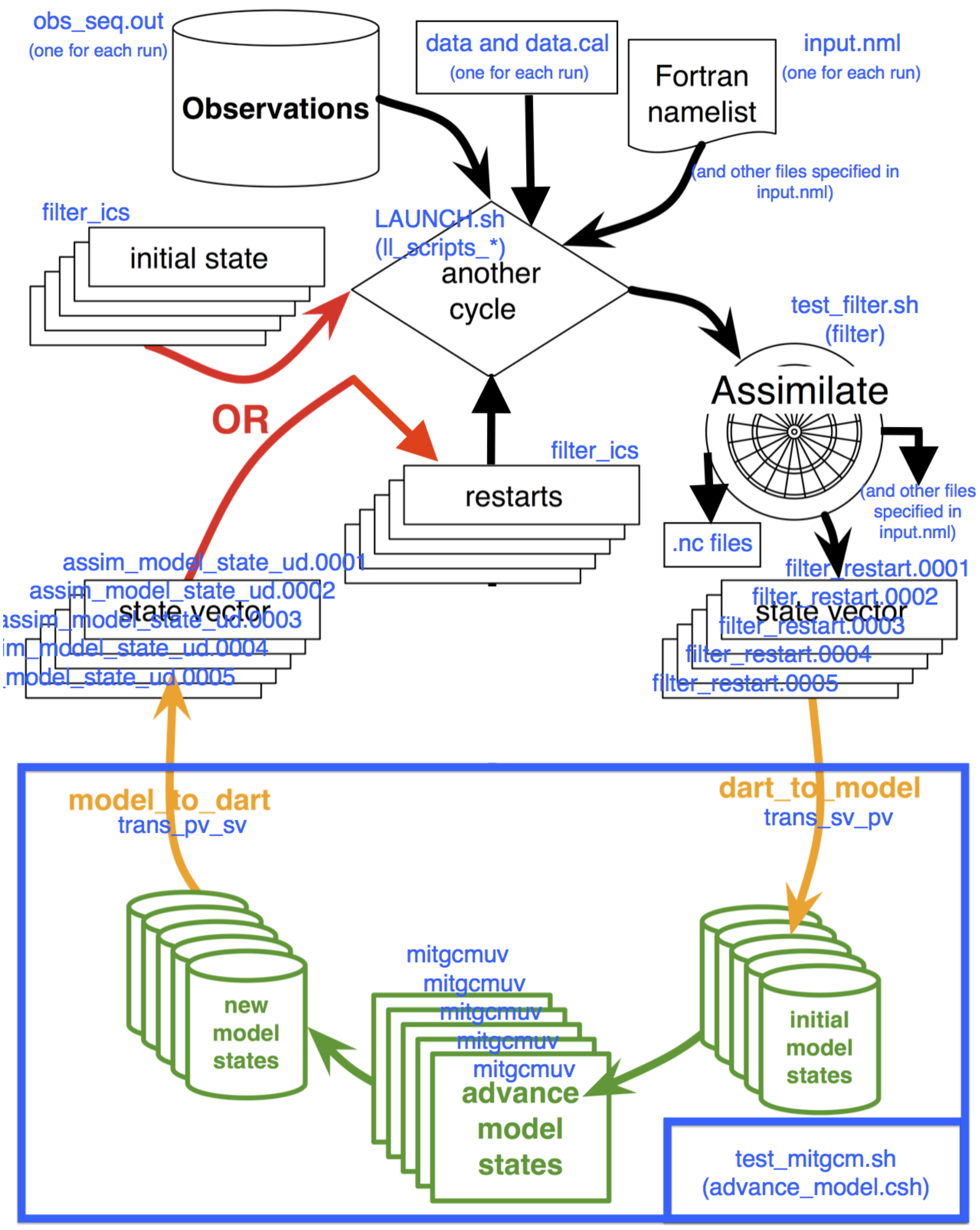


Figure 2.2: Workflow implementation

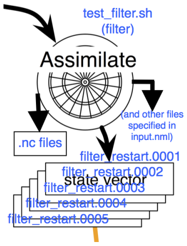
ters stored in the namelist files.

SSH and SST satellite data, temperature and salinity profiles are currently assimilated into the system, but it is designed to accommodate most available in situ and satellite ocean datasets. The data is put in a predefined American Standard Code for Information Interchange (ASCII) format before being converted to a distinct DART format. Once the observation data file for the full assimilation process is ready, a DART program splits the file into multiple files specific to each assimilation window. For an operational usage of the system, the assimilation window observation file should be generated on the fly.

The state vector is composed of the prognostic variables (SSH, salinity, temperature, zonal and meridional velocities) needed to run the model. The initial ensemble of state vectors is generated from a long free run model outputs by keeping the outputs corresponding to the assimilation starting date, then by retaining the elements two weeks (or sometimes one month) before and after, with 3-day spacing (i.e. if the assimilation start on January 1^{st}, we keep the members on December 29, 26, ... and January 4, 7, ...), for all the available years. Once the selection is accomplished, the initial ensemble date is set to the assimilation starting date. This selection method is used to provide more realistic spread and more meaningful physical representation than simply adding random perturbation to a single realization. Other perturbation methods, such as singular/bred vectors or Empirical Orthogonal Functions (EOFs), that may more efficiently account for the current uncertainties (or error-of-the-day), could also be considered to generate the initial ensemble. The ensemble mean is eventually replaced by the free run estimate of the given assimilation starting date. The initial ensemble is generated the same way for all the assimilation schemes and only the subsequent cycles ensembles differ.

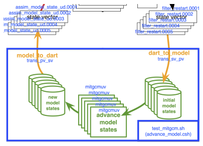
Namelist files are also required to setup and run the assimilation system. The model namelists containing time information (`data.cal` and `data`) are updated at each assimilation step by the program that performs the conversion of the filter outputs from the filter data format to the model data format. The initial assimilation date is read from the filter namelist (`input.nml`) that drives the assimilation process. `input.nml` contains information about the assimilation window period, the ensemble size, the kind of filter to be applied (EAKF, EnKF, Kernel filter, Observation Space Particle filter, Random draw from posterior, Deterministic draw from posterior with fixed kurtosis, Boxcar kernel filter, Rank Histogram filter, or Particle filter) and many other flavors.

2.4.1.2 Analysis (assimilation) step: filter



The DART filter uses the files set at the initialization step, assimilate the observations, and produces `.nc` files containing the analysis ensemble mean and spread (and ensemble members if specified in the filter namelist) prior and posterior to the assimilation. Additionally, the ensemble members are outputted in a specific DART format and will be used at the forecast stage as new initial conditions for the model integrations.

2.4.1.3 Forecast step (advancing the model): MITgcm



During the forecast step, N MITgcm runs are independently integrated by N instances of the forecasting script. The independence naturally allows for parallelization. The number of members that could be run in parallel will however depend on the available resources. For each member, a temporary run directory is created and the required inputs are copied or symbolically linked. Each member is assigned its initial condition (output from the DART filter) from a control file that contains the identification information. Then the initial conditions are converted to the model format, before running MITgcm, whereupon only the model outputs that are part of the state vector are converted back to DART format. The assimilation time is also updated here.

2.4.2 Initialization and submission scripts

The assimilation code runs on HPC platforms (previously Shaheen I (IBM) and currently Shaheen II (Cray)) where the jobs submissions and management is steered by a job scheduler (LoadLeveler for Shaheen I and slurm for Shaheen II).

At the first cycle of the initialization process (initial ensemble, observations and filter namelist generation) in Section 2.4.1.1, the initial ensemble generation is carried out offline with some scripts specific to the selected assimilation method. Since the assimilation needs the files to be in DART format, whereas the model long run outputs are in MITgcm format, there is a first category of scripts that select the model outputs then run some conversion programs that put the model outputs in the format required by the DART filter. To expedite the process, all the model outputs have been converted in DART format and the second category of scripts just need to select the members. There is also an index table that keeps the members identifiers (for each format) and their dates.

For the remaining cycles of the initialization process, the initial ensemble is built up from the model outputs resulting from the forecast step (Section 2.4.1.3), depending on the chosen ensemble selection scheme.

Regarding the observations, as discussed in Section 2.4.1.1, a file containing all the observation is generated. Then offline, before the assimilation starts, after specifying the full assimilation duration and the time lengths of the different assimilations windows, a script launches a program to split the big observations file into small files to fit each assimilation window. The initial version of the script was extracting the observations serially and was taking very long to complete. Consequently, I implemented a parallelized version and obtained a speed up factor of 63. The idea was to make a multiple stage extraction instead of extracting all the needed observations at once.

Indeed, the computation time depends on the length of the observation sequence and not on the length of the extracted sequences, this, because the program traverses the sequence linearly and extracts the observations belonging to the specified time window. To obtain q extracted sequences, the time will be equivalent to q times the time to read the initial sequence. For the multiple stage extraction, let us say s stages, the program extracts q_1 sequences at the first stage, q_2 sequences at the second stage, ..., and q_s sequences at the last stage. At each stage m, the q_m extractions can be performed in parallel. The number of extractions should be small at the beginning for better performance, and once we reach a stage at which the extractions are fast enough, we end by achieving all the q extractions corresponding to the q sequential extractions. For example, a two stage extraction makes two extractions of two intermediate files at the first stage. Next the needed observations files are extracted from the two intermediate files. This two stage extraction was enough to reach the 63 speed up factor, basically turning the hours into minutes.

When it comes to the filter namelists, as for the observations, a file for each assimilation window is generated and parametrized by the starting and ending dates of the assimilation window.

At the analysis (Section 2.4.1.2) and the forecast (Section 2.4.1.3) steps, one filter and N forecast jobs have to be submitted for each cycle. In each job file, the machine resource allocations are specified along with the required input files and the command to run job. Now the questions that arise are: How to write the submission jobs, because it would be painful to write hundreds or thousands of submission jobs by hand? How to manage the dependencies, since each forecast step should follow a filter step? How to launch the workflow, taking into account the machine constraints, and how to monitor it?

To answer the first two questions, a generation script was written in Python for Shaheen I and in Bash shell for Shaheen II to produce a sequence of submission scripts with the desired dependencies. The submission scripts are normally written in one file. But, due to the machine stress caused by the large number of handled jobs, a maximum number of jobs per files has been set. The LoadLeveler version file can be very long since each forecast member is written separately while for the slurm version, the array construct allows for a concise expression by only setting the array size to the number of needed forecast and parametrizing the scripts with the forecast member identifier. In the LoadLeveler version, the dependencies are defined using the job names. Unlike LoadLeveler, slurm cannot use the job names to assign the dependencies, but rather uses the job IDs. Given that the job IDs become available only when the job is submitted, in the slurm version, there is a first job to launch the sequence and collect the IDs in order to set the dependencies. The generation scripts build the launching scripts that will be submitted though they cannot steer and monitor the scripts executions. The machine scheduler manage the dependencies among the jobs but in case of failure, one needs to manually restart the workflow. However, the generation scripts can generate the submission scripts from the desired restarting point. As for the monitoring issue, it is addressed in the next Section.

2.4.3 Consistency checks and monitoring

There are many identical parameters in the DART-MITgcm system that are scattered across the files. To avoid any confusion and surprises, it is very important to set those parameters centrally. Therefore, most of the parameters, as well as the preferred ensemble selection scheme, have been gathered in a template filter namelist. A script first checks for the needed files for the system to run, then it parses the namelist to get the values of the parameters, before setting them in all of the other files, according to constraints related to some parameters combinations. After those checks, the script launches the generation scripts (Section 2.4.2), and eventually submits the jobs. The described consistency check script undertakes the required actions for the jobs submissions but does not monitor the workflow execution. In case of failure or abortion of the workflow, the end user has to identify the issue and relaunch the system. To avoid that demanding process, in partnership with the KAUST Supercomputing Laboratory (KSL) team, I made a further improvement of the assimilation system answering the monitoring question raised in Section 2.4.2. The improvement consists in an extension to the slurm scheduler, *Decimate*, that is able to steer the workflow and automatically relaunch it in case of abortion. When the interruption is due to a numerical failure, *Decitmate* replaces the faulty members by picking up new ones from a precomputed dictionary. *Decimate* is further discussed in Chapter 4.

2.4.4 System and assimilation results validation

To assess a data assimilation system, we first need a rigorous definition (and understanding) of the system and what is meant by a solution of that system. Knowing the properties of the system and the properties of its solution(s) and being able to characterize them is very important. That would allow to derive theoretical solution(s) against which numerical results could be compared in order to validate the assimilation system.

2.4.4.1 Distance between estimates and comparing solutions

First, we need to be able to compare different estimates or solutions to a reference. A commonly used measure is the forecast root-mean-square error (RMSE), at time k, defined as

$$rmse_k^f = \sqrt{\frac{1}{p_k}\sum_{i=1}^{p_k}\left(\left\{h_k(\overline{\mathbf{x}}_k^f)\right\}_{(i)} - \mathbf{y}^o_{k_{(i)}}\right)^2} \quad . \tag{2.28}$$

The subscript $_{(i)}$ refers to the ith component of the corresponding vector. Similarly, the analysis root-mean-square error (RMSE) is defined by replacing the superscript ^{f} by ^{a}. Since the truth is generally unknown, except for validation experiments, the observations are used as reference. But which observations? Usually, the assimilated observations are utilized, and sometimes, the results are benchmarked against independent data or observations that have not been assimilated. Comparing to independent data provides a more reliable assessment.

Figure 2.3 shows the RMSEs for three trials of the same experiment. Keeping in mind that $\overline{\mathbf{x}}_k^f$ is a random variable, an average over a set of trials or outcomes of the same experiment might be required for a better evaluation. This leads to the definition of the forecast mean-squared error (MSE) at time k:

$$\begin{aligned} mse_k^f &= tr\left(\mathbb{E}\left[\left(h_k(\overline{\mathbf{x}}_k^f) - \mathbf{y}_k^o\right)\left(h_k(\overline{\mathbf{x}}_k^f) - \mathbf{y}_k^o\right)^T\right]\right) \\ &= \mathbb{E}\left[\left(h_k(\overline{\mathbf{x}}_k^f) - \mathbf{y}_k^o\right)^T\left(h_k(\overline{\mathbf{x}}_k^f) - \mathbf{y}_k^o\right)\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{p_k}\left(h_k(\overline{\mathbf{x}}_k^f)_{(i)} - \mathbf{y}_{k_{(i)}}^o\right)^2\right]. \end{aligned} \tag{2.29}$$

Notice that the RMSE expression is closely related to the MSE one, but with a normalization factor $\frac{1}{p_k}$, a square root and without the expectation. The MSE can be seen as an "average of (squared-)RMSEs" over multiple experiments. The expectation can be approximated by the sample mean:

$$\begin{aligned} mse_k^f &= \mathbb{E}\left[\sum_{i=1}^{p_k}\left(h_k(\overline{\mathbf{x}}_k^f)_{(i)} - \mathbf{y}_{k_{(i)}}^o\right)^2\right] \\ &= \frac{1}{m}\sum_{j=1}^{m}\left\{\sum_{i=1}^{p_k}\left(h_k(\overline{\mathbf{x}}_k^f)_{(i)} - \mathbf{y}_{k_{(i)}}^o\right)^2\right\}_j, \end{aligned} \tag{2.30}$$

where m is the number of trials. Given the cost for conducting an assimilation experiment with large OGCMs, replications of the experiments are not always viable such that the RMSE is preferred over the MSE.

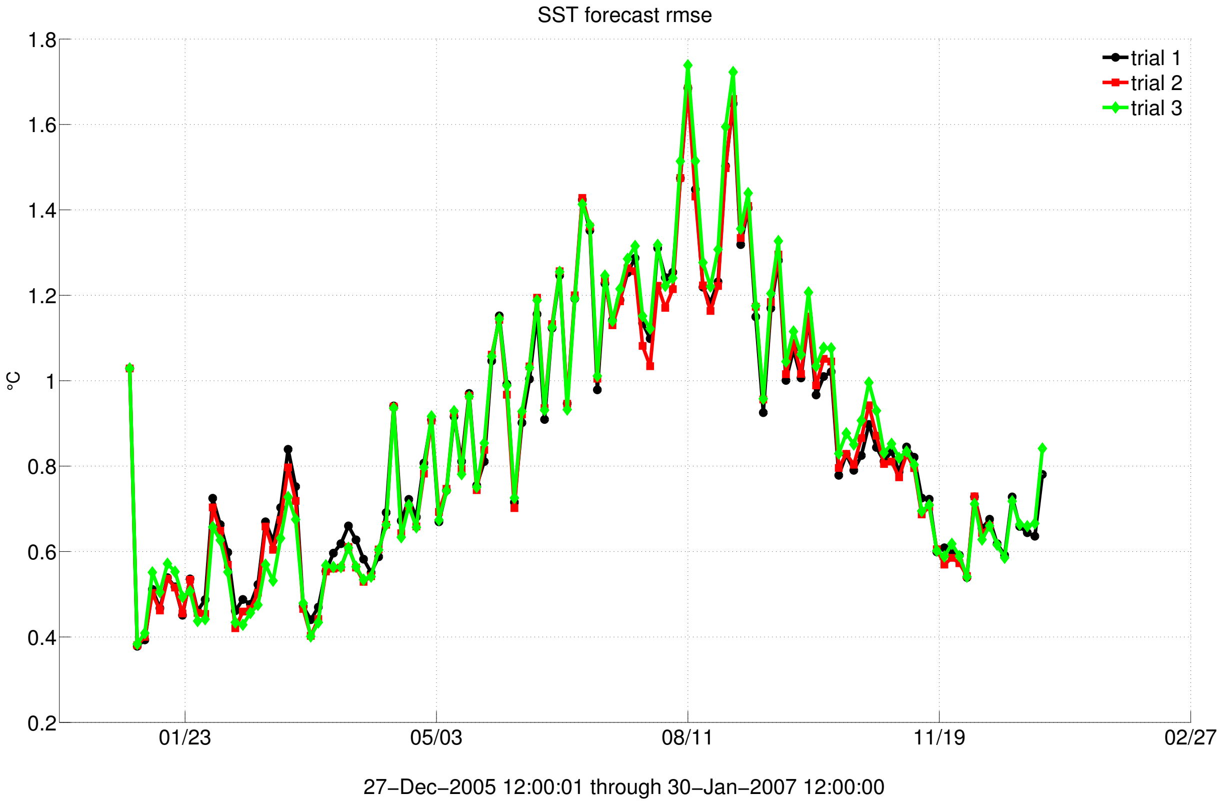


Figure 2.3: RMSEs for three different trials of the same experiment.

2.4.4.2 Twin experiments

The basic idea is to first run an experiment and tag it as the truth. Observations are then sampled from the truth before being assimilated in a second experiment. The assimilation solution is then compared against the true reference solution to confirm that the system is properly working without the effect of model errors.

When looking at the following update equation $\mathbf{x}^a = \mathbf{x}^f + \mathbf{P}^f\mathbf{H}^T(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{y}^o - h(\mathbf{x}^f))$, when $\mathbf{y}^o = h(\mathbf{x}^f)$, it follows that $\mathbf{x}^a = \mathbf{x}^f$. So by setting the observation process identical to the forecast process, we also expect the analysis process to match them. This twin experiment is simple to implement, and does not require a deep understanding of the underlying system. It is very usefull to check and validate the implementation of an assimilation system. Figure 2.4 provides an example in which the method helped spot an implementation issue, due to some non zero values encountered, meaning that the forecast and the analysis are not equal when they should.

2.5 Summary

This Chapter presented the Bayesian filtering and the estimation problem before detailing the components of the DART-MITgcm assimilation system, namely the MIT general circulation ocean model (MITgcm) and the Data Assimilation Research Testbed (DART). The system configuration for the Red Sea was described. In particular, the worklow and its management, as well as the system and the assimilation results validation tools were discussed.

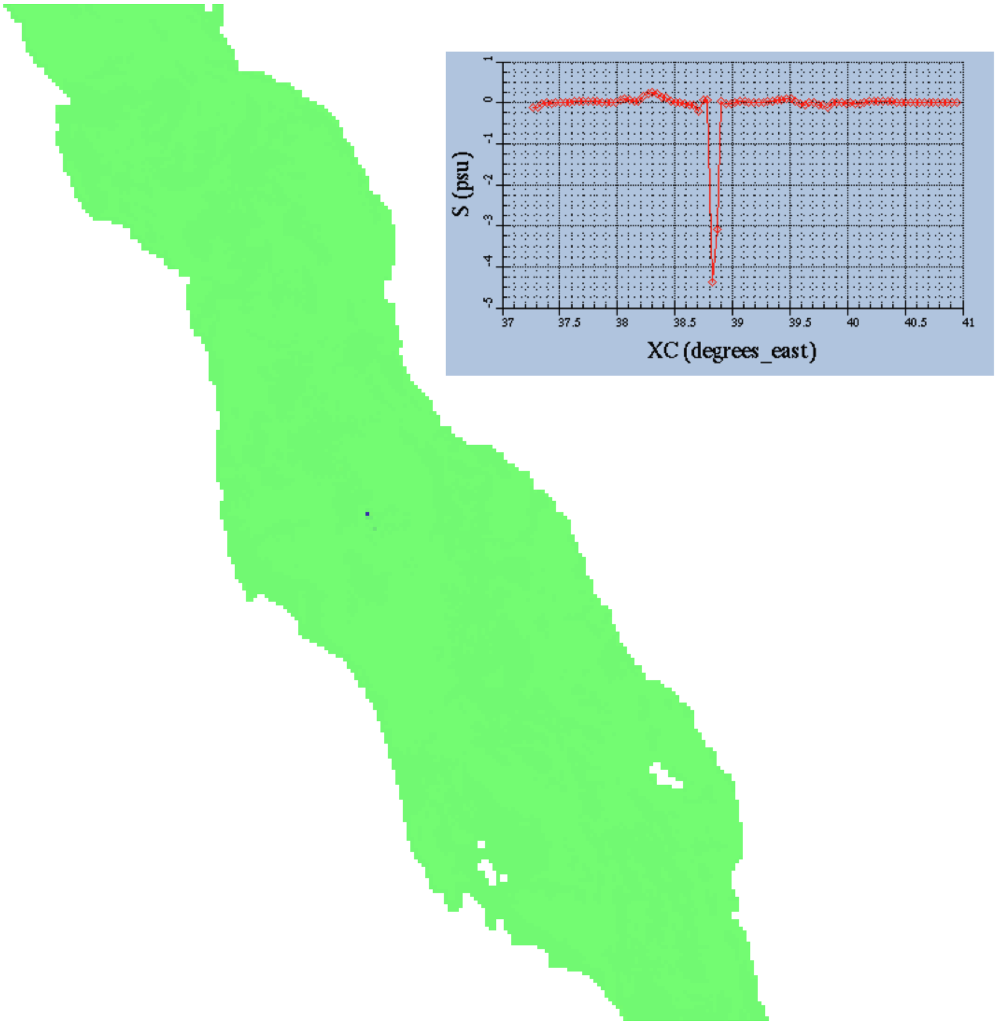


Figure 2.4: Difference between a reference run and a second run in which the observation process is set identical to the reference run. The green color represents (near) zero values. Some non zero salinity values appear in the middle.

Chapter 3

Ensemble data assimilation in the Red Sea: Sensitivity to ensemble selection and atmospheric forcing

This Chapter corresponds to the paper "Ensemble data assimilation in the Red Sea: Sensitivity to ensemble selection and atmospheric forcing" published in *Ocean Dynamics.*

3.1 Introduction

In this Chapter, we examine the overall performance of a deterministic EnKF, the EAKF [8], for assimilating satellite SSH and SST data into a 4 km MITgcm that has been configured and validated to study the circulation of the whole Red Sea [189, 190, 196]. We evaluate the sensitivity of this assimilation system to various parameters and inputs, including filtering scheme and parameters (ensemble size, inflation) and atmospheric fields (NCEP and ECMWF). We are in particular interested in investigating the benefit of using a flow-dependent ensemble against keeping it invariant in time, using an ensemble Optimal Interpolation (EnOI)-like scheme in which the model is used to forecast only the state, and not the ensemble [57, 80, 134, 155]. The latter assumes that the forecast error covariance is well represented by a stationary ensemble, and may lead to drastic reduction (up to 80- 90% less) in the computational burden compared to a flow-dependent ensemble. It may further help maintaining the ensemble spread, which is one of the issues often encountered in EAKF applications, especially when model errors are not directly accounted for in the system [9, 80]. We further assess the possibility of exploiting the dominant seasonal variability of the Red Sea and test the performance of the EnOI scheme with seasonally varying ensemble of model states that are not integrated with OGCM but are readily available from a historical model run [17, 186]. This allows the EnOI scheme to adjust to the seasonal variation of the system without extra computational cost, but may not well-represent the error-of-the day in the most recent estimate compared to an ensemble Kalman filter. We compare the performances of the EAKF, EnOI, and EnOI with seasonal varying ensemble (Seasonal ensemble Optimal Interpolation (SEnOI)) in the Red Sea and study their sensitivities to various settings and atmospheric forcing.

The rest of this Chapter is organized as follows. Section 3.2 gives a brief description of the model and observational data used for model validation and in the assimilation experiments. Details of the assimilation schemes and their implementation are provided in Section 3.3. In Section 3.4, we present the results of several assimilation experiments that have been conducted to evaluate the performances and robustness of the compared ensemble assimilation schemes. A discussion and summary conclude the work in Section 3.5.

3.2 Model and Data

3.2.1 Ocean model and configuration

The model is configured as described in Section 2.2. Moreover, the lateral boundaries are treated with no-slip conditions and a quadratic bottom friction is imposed. On the eastern lateral boundary, SSH, salinity, temperature, zonal and meridional velocities open boundary conditions (OBCS) from the Estimation of the Circulation and Climate of the Ocean (ECCO) [94] are assigned through a 20-km buffer zone.

In the different experiments, the model is forced with 6-hourly atmospheric reanalysis from the National Centers for Environmental Prediction (NCEP) or the European Centre for Medium-Range Weather Forecasts (ECMWF). These include zonal and meridional wind speed, air temperature, specific humidity, precipitation and downward short and long wave heat fluxes. A free model run was integrated over a 32-year period from January 1979 to December 2011 using a time step of 200s (without assimilation), and outputs from 1992 to 2011 were stored for validation and for constructing an (initial or static) ensemble of state realizations.

3.2.2 Observational data

For the assimilation, along-track SSH data is obtained by combining sea level anomaly (SLA) acquired from Radar Altimeter Database System (RADS), available through the web portal http://rads.tudelft.nl, and the mean dynamic topography (MDT) from the Archiving Validation and Interpretation of Satellite Oceanographic Data (AVISO), available at ftp.aviso.altimetry.fr/auxiliary/. RADS is developed by Delft University of Technology and the National Oceanic and Atmospheric Administration (NOAA). It provides merged SLA observations from nine altimeter missions, and is one of the most accurate and complete data bases of satellite radar altimeter data [157]. MDT is a key reference surface for altimeter data, and can be used to calculate the corresponding absolute dynamic topography (ADT) from the altimeter SLA through ADT = MDT + SLA. The ADT is equivalent to the model SSH, which will be assimilated into the ocean model. All the SSH data within the assimilation window, 3 days in this study, were gathered and assimilated once at the middle of the window. Observational errors of these along-track SSH data are specified with different values ranging between 0.05 to 0.1 m for different satellite missions.

It is important to point out that the accuracy of altimetry data in coastal waters could be limited by several factors, including the weaknesses of the altimeters in the range tracking procedure close to the shorelines, intrinsic difficulties in the corrections of the wet tropospheric correction, tides, etc., and issues of land contamination in the altimeter return waveforms [30, 180]. Important efforts are still being carried out to overcome such problems and to extend the capabilities of current and future altimetry data in coastal waters [104, 118, 175, 180, 187]. As a safe and practical approach, the SSH observations over shallow waters (less than 60 m in depth) were excluded. We have also removed the outliers during the assimilation process, which were flagged when the distance between the forecast ensemble mean and the observation value exceeded three times the square root of the sum of the observation variance and the forecast ensemble variance for that observation.

The assimilated SST data is extracted from the Group for High Resolution Sea Surface Temperature (GHRSST) global Level 4 SST analysis produced daily on a 1/4° grid at the NOAA National Climatic Data Center [149] (available at http://podaac.jpl.nasa.gov/dataset/NCDC-L4LRblend-GLOB-AVHRR_OI). These are mapped data from the 4 km Advanced Very High Resolution Radiometer (AVHRR) Pathfinder Version 5 time series (when available, otherwise operational NOAA AVHRR data are used) and in-situ observations. In the assimilation experiments, three-day averaged data is provided at midnight with a 0.25° grid. Observational errors are uniform and set at 1.2°C. These are larger than what is commonly used, but are expected to also account for the spatially correlated nature of this mapped dataset.

3.2.3 Model validation with SSH/SST

A free model run (without assimilation) was integrated over a 32-year period from 1979 to 2011, and its outputs from 1992 to 2011 were stored for validation and for constructing an (initial or static) ensemble of state realizations. As shown in Figure 3.1, the model mean SST (Figure 3.1.e) is in good agreement with the AVHRR data (Figure 3.1.a), exhibiting a clear gradient throughout the basin where highest temperature is found on both coasts of the southern Red Sea. Strong variability of SST near the west coast around 16°N and 23°N, and weaker variability along the east coast between 16°N and 19°N (Figure 3.1.b), are accurately depicted by the model

(Figure 3.1.f). However, the modeled exhibits a weaker SST variability than the AVHRR in the northern basin with a smaller standard deviation. The mean and standard deviation of model SSH are comparable to those of the daily AVISO gridded SSH product (available at http://www.aviso.altimetry.fr/en/data/data-access/aviso-opendap/opendap-adt-products.html). The model well reproduced the south-to-north SSH gradient in the basin (Figure 3.1.g) compared with the AVISO data (Figure 3.1.c). The model SSH variability is slightly weaker than what is observed by AVISO, especially towards the Saudi coast in the central and northern Red Sea. The larger variability of modeled SSH in the central and northern basins is likely the signature of a strong eddy variability (Figure 3.1.h), which is not always represented by AVISO data (Figure 3.1.d). The discrepancy may result from sub-mesoscale features in the model outputs that are not represented by the 0.25° AVISO data. In addition, the merged gridded AVISO product is generally produced from low-coverage of daily along-track data [196], which may underestimate the eddy intensities [191].

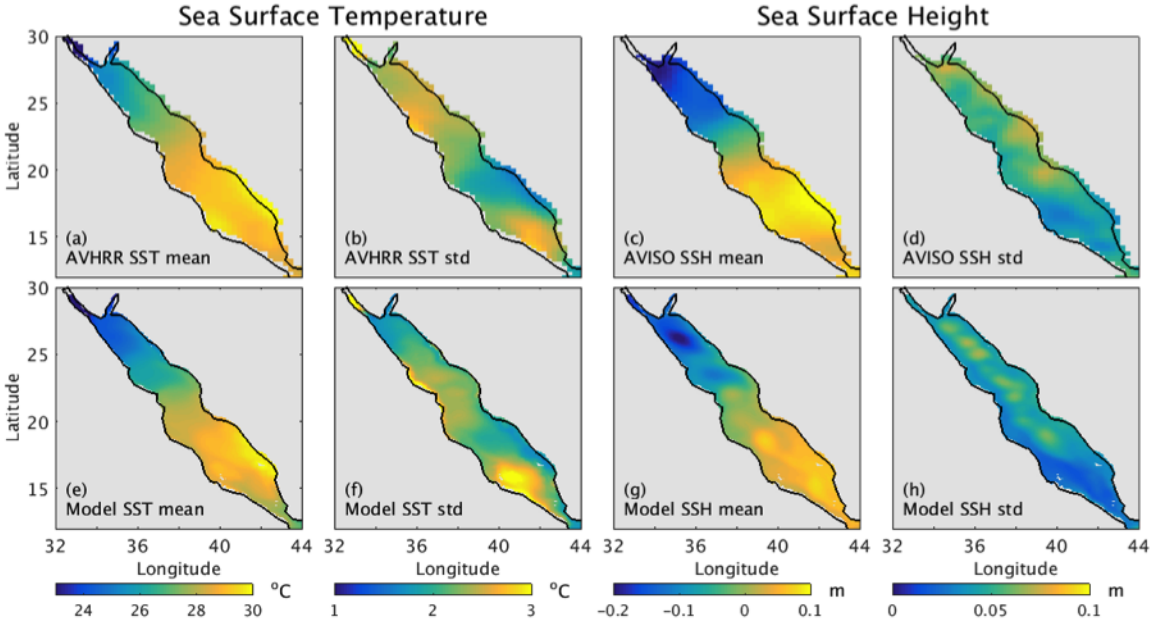


Figure 3.1: Mean and standard deviation of remote sensing observations (upper panel) and model outputs (lower panel) calculated using data from 1996 to 2010.

3.3 Ensemble Assimilation Schemes and Implementation

3.3.1 Ensemble Kalman filtering and Seasonal Optimal Interpolation

Data Assimilation serves to incorporate observational data with numerical models to best estimate the state of the ocean [46]. It is mainly used for forecasting purposes, but also for developing ocean reanalysis products, parameter estimation, uncertainty quantification, etc. State-of-the-art ocean data assimilation schemes are now well established following two directions depending on how the data are assimilated into the model. The variational approach seeks for the deterministic model trajectory that best fits all available observations by tuning some uncertain model parameters. The model-data fit is measured by a well-chosen objective function that is optimized based on its gradients calculated using the adjoint of the ocean model [103]. The filtering approach sequentially updates the model forecasts every time new observations are available based on prior errors estimates on the model forecast and assimilated data [81]. The most widely used sequential assimilation schemes are the ensemble Kalman filter (EnKF) and its variants [46, 53, 80, 169]. These are Monte Carlo-based variants of the famous Kalman filter (KF) designed for nonlinear and computationally demanding models [50]. In contrast with the variational methods, the Kalman methods are non-intrusive (do not require the development of an adjoint model), and are therefore easier to implement.

A crucial aspect of any data assimilation scheme is a good description of the forecast error covariance, often referred to as the background covariance, which describes how the model-observation misfits are projected into the state space to correct the forecast. When the system is linear and errors statistics are Gaussian, the KF provides an optimal way to sequentially estimate the time evolution of the forecast state and its background covariance according to the system dynamics [81]. To enable the implementation of the KF for data assimilation into realistic high-dimensional and nonlinear ocean GCMs, [50] proposed to represent the forecast statistics (first two moments) by an ensemble of state vectors, called ensemble members. Given an ensemble of model forecasts $\mathbf{X}^f = [\mathbf{x}^{f,1}, \mathbf{x}^{f,2}, \cdots, \mathbf{x}^{f,N}]$, estimates of the forecast state and its background covariance are then taken as the sample ensemble mean and covariance,

$$\overline{\mathbf{x}}^f = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}^{f,i} \quad \text{and} \quad \mathbf{P}^f = \frac{1}{N-1}\boldsymbol{X}'(\boldsymbol{X}')^{\mathrm{T}},$$

where $\boldsymbol{X}' = [\mathbf{x}^{f,1} - \overline{\mathbf{x}}^f,\, \mathbf{x}^{f,2} - \overline{\mathbf{x}}^f,\, \cdots, \mathbf{x}^{f,N} - \overline{\mathbf{x}}^f]$ is the ensemble of anomalies. This provided a particularly efficient framework to estimate the forecast error covariance for adequate weighting of the forecast in the assimilation, to account for various sources of model errors, and to quantify the uncertainties in the estimated solution [76]. This study focuses on the ensemble Kalman methods, which we implement here based on the Data Assimilation Research Testbed (DART) package. Our goal is to develop an efficient, in term of computational cost and performances, assimilation system for reconstructing and forecasting the space-time circulation of the Red Sea and to quantify the uncertainties in the estimated fields.

As the KF, EnKFs operate as a succession of forecast and analysis steps. In the forecast step, the ensemble members are integrated with the dynamical model to the time of the next available observations. In the analysis step, the forecasted members are adjusted by the incoming observations using the KF update step:

$$\mathbf{x}^a = \mathbf{x}^f + \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1}\left[\mathbf{y}^o - h(\mathbf{x}^f)\right],$$

where $\mathbf{P}^f$ is the forecast error covariance, $\mathbf{R}$ is the observational error covariance, $\mathbf{y}^o$ is the observational vector, and $\mathbf{H}$ is the linearized form of the observational operator h (in our setting, the assimilated SST and SSH data are model variables, so that h is linear).

EnKF methods were classified into stochastic or deterministic techniques, depending on whether the observations were perturbed, or not, before assimilation [169]. Deterministic filters, such as the EAKF, became more popular for data assimilation in oceanography [46] and meteorology [87], to avoid introducing noise from the undersampling of the observational error covariance with a small ensemble [6, 81, 129]. The performance of an EnKF greatly depends on the representativeness of its ensemble members, which should adequately describes the statistics of the state estimates errors.

The EAKF update step is based on the following equations [8]:

$$\overline{\mathbf{x}}^a = \mathbf{P}^a \left[\left(\mathbf{P}^f \right)^{-1} \overline{\mathbf{x}}^f + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{y}^o \right], \tag{3.1}$$

$$\mathbf{P}^a = \left[\left(\mathbf{P}^f \right)^{-1} + \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \right]^{-1}, \tag{3.2}$$

$$\mathbf{x}^{a,i} = \mathbf{A}(\mathbf{x}^{f,i} - \overline{\mathbf{x}}^f) + \overline{\mathbf{x}}^a, \qquad i = 1, \cdots, N, \tag{3.3}$$

where $\mathbf{x}^{f,i}$ and $\mathbf{x}^{a,i}$ are individual members of the forecast and analysis ensemble and $\overline{\mathbf{x}}^f$ and $\overline{\mathbf{x}}^a$ their respective means, and N is the ensemble size. Equations (3.1) and (3.2) compute the analysis state $\overline{\mathbf{x}}^a$ and its error covariance matrix $\mathbf{P}^a$ from the forecast ensemble mean $\overline{\mathbf{x}}^f$ and covariance $\mathbf{P}^f$, exactly as in the Kalman filter. The analysis members are then generated using Equation (3.3) in such a way to exactly match $\overline{\mathbf{x}}^a$ and $\mathbf{P}^a$, which are the constraints for selecting the matrix $\mathbf{A}$.

The EAKF formulation enables updating the ensemble with the model to track changes in the ocean dynamics, which should be particularly useful in regions subject to important spatial and temporal variability.

EAKFs may suffer from the collapse of their ensembles members (towards the ensemble mean) [58], especially when the forecast model is not integrated with stochastic perturbations to account, for instance, for uncertainties in the forcing and/or parameters. One simple approach that was proven efficient to mitigate this is to amplify the ensemble spread by multiplying by an inflation factor that is larger than 1 [8, 73, 80]. Another challenge in ensemble Kalman methods is the spurious correlations in the forecast error covariance that is inherited from the use of small ensembles in practice [84]. The low-rank nature of the covariances of such ensembles may further extend the impact of an observation to faraway points from its location, which may severely limit the filter ability to fit the data [74]. This problem can be generally efficiently addressed through the localization technique [57, 74, 85, 134]. The basic idea is to restrict the impact of an observation to nearby points only, or equivalently, trim long-range correlations from the ensemble covariance [153]. In practice, large inflation and strong localization may, however, respectively weaken the stability of the assimilation system, and introduce undesirable small-scale features into the analysis when the observations are sparse [72].

Integrating large ensembles with an ocean GCM is computationally demanding. Following the optimal interpolation (OI) approach in data assimilation, which uses a static pre-selected background covariance in the update step, ensemble OI (EnOI) methods were proposed [52, 80, 131]. EnOI is a very cost effective alternative to an EnKF in which the static background covariance is estimated as the sample covariance matrix of an adequately pre-selected ensemble, generally representing the error growing modes, or describing the variability of the studied system. This formulation does not suffer from the ensemble collapse problem, but its performance may be limited during periods of strongly changing dynamics that are generally not well described by a static background [79, 80]. EnOI has a very similar algorithm as an EnKF, except that only the analysis state, and not the entire ensemble, is integrated with the model during the forecast step. The method was found to provide good performances compared to an EAKF at fraction of the computing cost[1] [80, 134, 155].

A completely static background error covariance may not ideally describe the variable patterns of the ocean flow in different seasons, and periods in between. To represent the flow-dependence in term of seasonality, [186] proposed to use EnOI to assimilate Argo profiles in a hybrid coordinate ocean model (HYCOM) with an ensemble selected at every assimilation cycle from monthly climatology fields with a three-month moving window around the assimilation time. The same scheme was latter adopted by [117] and [185] for assimilating SLA data in the South China Sea. In this study, we implemented a similar EnOI scheme but selecting the ensemble on a monthly basis from a climatological dataset of the Red Sea circulation that is assumed to describe the variability of the system. The Red Sea climatology was simulated from a long-term historical model simulation as described in more details in the next section. We refer to the EnOI with seasonally varying ensemble as the seasonal-EnOI (or SEnOI).

3.3.2 Implementation within DART

EAKF, EnOI and SEnOI were implemented in fully parallel mode (at the forecast and the analysis steps) using the Data Assimilation Research Testbed (DART). DART is a portable software for ensemble data assimilation developed at the National Center for Atmospheric Research (NCAR) [12]. It builds on a series of interface routines that incorporate a forecast model and different types of observations, and can be used with a variety of algorithms to update the ensemble, including for instance the (stochastic) EnKF and the EAKF. DART is configured to integrate and update the ensemble members in parallel, exploiting the serial formulation of the Kalman filter update step [10]. It is further equipped with advanced inflation/localization techniques that are important to enhance the performance of an ensemble-based data

[1] in CPU time and not necessarily in real time as the ensemble members can be integrated in parallel.

assimilation system. It has been successfully implemented in various atmospheric and oceanic applications [4, 76, 145].

DART has been already implemented with the MITgcm for data assimilation and forecasting the loop current in the Gulf of Mexico [76]. A similar system is implemented in this study with some specific adjustments to the Red Sea. The model state vector is composed of the prognostic ocean variables that are needed to initialize the MITgcm, i.e. salinity, temperature, horizontal velocity, and sea surface height fields. We used the EAKF as described by [8, 9], and modified some of its routines to enable for EnOI and SEnOI assimilation. The EAKF steps in DART-MITgcm are summarized in the schematic map Figure 3.2. Starting from an analysis step, a given initial ensemble $\mathbf{X}^f$ is delivered to DART, the mean and covariance of $\mathbf{X}^f$ are updated with the filter, based on which the analysis ensemble $\mathbf{X}^a$ is then deterministically generated. This is followed by the forecast step, in which each member of the analysis ensemble $\mathbf{X}^a$ is integrated with the MITgcm to obtain the new forecast ensemble, which enables to start a new assimilation cycle.

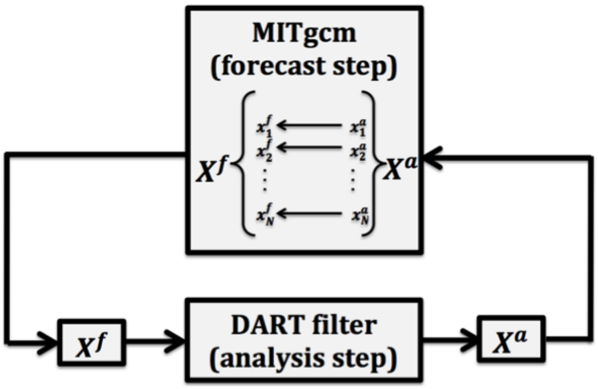


Figure 3.2: DART-MITgcm EAKF scheme flow chart. The forecast ensemble $\mathbf{X}^f$ is first updated with DART using the observation to compute the new analysis ensemble $\mathbf{X}^a$. The latter is then integrated with the MITgcm to obtain the forecast ensemble $\mathbf{X}^f$ from which a new assimilation cycle could be initiated.

In this study, the initial ensemble was selected from the outputs of the long-term model simulation between 1992 and 2004. Three-day outputs were saved and assigned into 12 datasets according to the time-period between their dates and the beginning of each month. For each month, a total of 252 model outputs were retained. As the experiments start from January-1-2006, the 252 sampled outputs were collected from early December and late January in different years, and the first 50, 100 or 250 records were assembled as the initial, or static, ensemble when running an experiment with those number of members.

The EnOI steps in DART-MITgcm are schematized in Figure 3.3. The filter starts from a given state estimate and an ensemble of model outputs from which we remove the mean to obtain an ensemble of anomalies $\boldsymbol{X}'$. When the new observations become available, one would read $\boldsymbol{X}'$ and compute the "new" forecast ensemble using $\mathbf{X}^f = \mathbf{x}^f + \boldsymbol{X}'$, which is then sent to DART to compute the analysis state $\mathbf{x}^a$ (no resampling of a new ensemble is needed here). The MITgcm is then integrated only once to compute the forecast state $\mathbf{x}^f$. A new assimilation cycle could then be initiated. In our implementation, $\boldsymbol{X}'$ is the same as the EAKF initial ensemble and is kept invariant in time. SEnOI is implemented based in EnOI, except that, its ensemble of anomalies is updated monthly by selecting its members from a climatological dataset consisting of long-term model outputs centered at the beginning of each month.

3.4 Experiments Setup and Assimilation Results

The assimilation experiments were performed over a one year period starting from January-1-2006. Along-track SSH and gridded AVHRR SST were assimilated every 3 days at midnight. The model data misfits were calculated as if all the data were observed at the assimilation time. Since we are using sequential data assimilation schemes, the SSH/SST data were binned at the middle of the assimilation window. The experiments were conducted with different ensemble sizes, inflation factors, and atmospheric forcing conditions. The ensemble localization technique is applied to remove eventual spurious long-range correlations that may appear from the use of

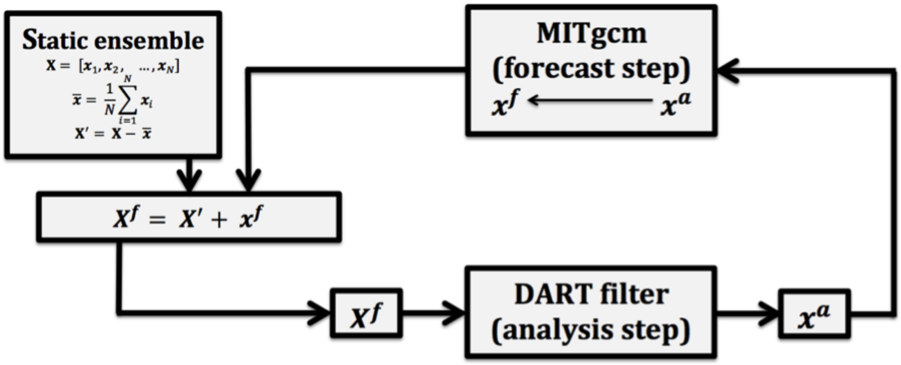


Figure 3.3: DART-MITgcm EnOI scheme flow chart. The forecast ensemble $\mathbf{X}^f$ is first updated with DART based on the observation to compute the new analysis ensemble mean $\mathbf{x}^a$, which is then integrated with the MITgcm to obtain the forecast $\mathbf{x}^f$. The single forecast is added to a pre-selected ensemble of anomalies to build the forecast ensemble $\mathbf{X}^f = \boldsymbol{X'} + \mathbf{x}^f$, from which a new assimilation cycle could be initiated.

small ensembles and to increase the rank of the forecast error covariance. A covariance localization cutoff radius of 0.05 rad (about 300 km depending on the latitude) is chosen from a series of assimilation runs with different localization scales (not shown), providing good and robust assimilation results. To maintain enough ensemble spread and avoid the ensemble collapse, the spread of the anomaly forecast ensemble $\boldsymbol{X'}$, i.e. ensemble covariance, was amplified by an inflation factor, before each analysis step. This is simply implemented by using the following inflated members in the analysis step

$$\mathbf{x}^{f,i,inf} = \alpha(\mathbf{x}^{f,i} - \overline{\mathbf{x}}^f) + \overline{\mathbf{x}}^f, \qquad i = 1, \cdots, N,$$

where α is an inflation factor generally chosen to be slightly greater than 1. The choice of the ensemble from which the forecast error covariance $\mathbf{P}^f$ is estimated is key for designing an efficient sequential ensemble assimilation system. In all the experiments presented in this Chapter, the initial ensemble of the EAKF is selected from a set of January climatological fields, i.e., the members are selected from the January outputs of the long-term model run. In the EnOI, this ensemble is kept invariant in time while in SEnOI the ensemble is reselected from the model outputs for the corresponding month.

3.4.1 Assimilation Results

3.4.1.1 Sensitivity to ensemble size

The objective of an EAKF scheme is to minimize the variance of the analysis error, which is expected to decrease as the ensemble size increases. The choice of the ensemble size is critical to the success of an EAKF assimilation system, and one should balance between ensemble size and computational cost. The ensemble should be large enough to describe well the statistics (mean and spread) of the prior distribution, and to provide a smooth enough covariance between the model state and the observations and avoid severe localization [127]. At the same time, the ensemble size should be reasonable to avoid excessive computing cost. Many studies suggested that, with appropriate localization and inflation, the decrease in the analysis error may stagnate with very large ensembles, suggesting that good performances may be obtained with relatively reasonable size ensembles [148].

To investigate the sensitivity of the EAKF-assimilation system in the Red Sea to the ensemble size, three experiments with 50, 100, 250 ensemble members were performed. An inflation factor of 1.1 and localization radius of about 300 km were considered in all three experiments. The time-evolution of the RMSEs between SSH/SST observations and filter forecast/analysis states are plotted in Figure 3.4. It is clear that for both SSH and SST, the RMSE decreases with larger ensembles. The experiment using 250 members leads to the smallest RMSE for both forecast and analysis, although the RMSE resulting from 50 members is quite reasonable, with an average forecast and analysis RMSEs of 0.71°C/0.08 m and 0.64°C/0.07 m for SST and SSH, respectively. Nevertheless, the improvements resulting from increasing the ensemble size from 100 to 250 are generally not very significant (especially for SST), and this is also reflected in their ensemble spreads. The SST and SSH ensemble spreads stabilize after the first 30 assimilation cycles, (about three months) reaching minimal values of about 0.1°C and 0.01 m in the winter season, before they slightly increase during the summer season. The SST and SSH analysis RMSEs are about 0.1°C and 0.01 m lower than the corresponding forecast RMSEs, respectively, suggesting that the data are properly assimilated into the model. The RMSE of SSH analysis is lower than that of the AVISO gridded data, which has been generated by merging different satellite missions' measurements [45].

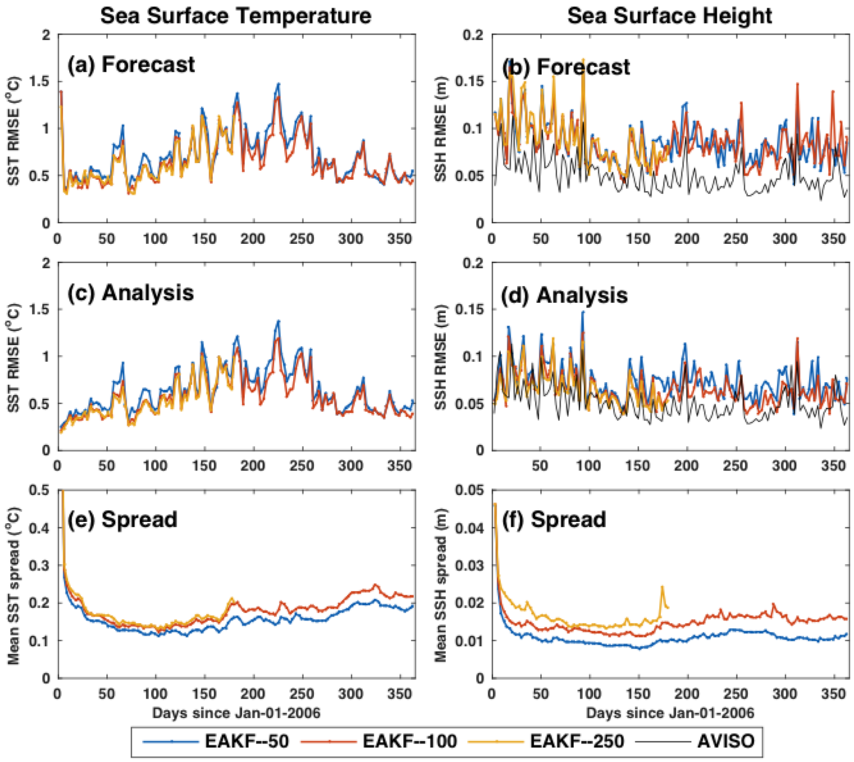


Figure 3.4: Time evolution of the SST/SSH RMSE and ensemble spread for EAKF with different ensemble sizes.

Increasing the ensemble size may increase the risk of collapse of the ensemble assimilation system; this is exactly what happened after 61 assimilation cycles (or six months) for the run with 250 members, when the MITgcm was not able to complete the integration of one of the ensemble members and diverged. In each assimilation cycle, the analysis ensemble increment, which is introduced by the Kalman gain matrix, may introduce some dynamically unstable realizations that are not compatible with the model physics [69]. This imbalance can be further more severe with localization and inflation [134]. Improving the dynamical consistency of the ensemble members and developing efficient online schemes to replace unstable members are two of the directions we are planning to explore in order to enhance the robustness of the system. Hereafter, we will limit the ensemble size to 100 as this seems to provide enough representative ensembles in order to obtain good and robust ensemble assimilation performances.

3.4.1.2 Sensitivity to inflation

The value of the inflation factor may depend on the dynamics of the model and the studied region, and on the configuration of the assimilation system, including the ensemble size and the filter [74]. Here we conduct trial and errors experiments to set the value of the inflation factor. Sophisticated adaptive inflation schemes were suggested for online space-time tuning of the value of the inflation [6, 12, 80], but these also require to be configured to the studied region and were not considered in this study.

To investigate the sensitivity of the MITgcm-EAKF Red Sea assimilation system to the inflation factor, four ensemble assimilation experiments with different values of inflation factors, 1.0 (no inflation), 1.05, 1.1 and 1.2, were conducted using the same ensemble size of 100 members. The time-evolution of SST and SSH RMSEs for the forecast and analysis fields, and the corresponding forecast ensemble spreads are plotted in Figure 3.5. The results suggest that the overall performance of the MITgcm-EAKF assimilation system in the Red Sea is quite dependent on the choice of the inflation factor. A remarkable improvement in the filter performance in term of RMSE is achieved using inflation, as compared to the filter results without inflation (inflation = 1.0). The accuracy of the filter estimates particularly improves with inflation, in summer, for both SST and SSH. However, increasing the inflation factor from 1.1 to 1.2 is not very beneficial, or even occasionally contributing negatively, to the system behavior. The ensemble spread of SSH and SST (Figure 3.5.e and Figure 3.5.f) decreases over time, with the largest decreases during the first few assimilation cycles, before stabilizing in later cycles. The ensemble spreads are then maintained at levels of about 0.15°C and 0.01 m for SST and SSH, respectively. With inflation, the ensemble spread decreases at a slower pace, but tends to diverge after some cycles, depending on the value of the used inflation factor. It is necessary to note that a larger inflation factor noticeably reduced the analysis RMSEs, but not always the forecast RMSEs, particularly for SSH. This probably implies that an EAKF system with larger inflation factor may overfit the observational data, providing analysis fields featuring some dynamically unbalanced features.

Although inflation generally helps maintaining the ensemble spread, large inflation factors may deteriorate the system behavior and even cause divergence. The experiment using an inflation factor of 1.2 stops after 55 assimilation cycles only. The large anomaly imposed by a large inflation factor may cause runaway increase in some states trajectories compared to those purely integrated with the model. This may force the analysis state to overfit the observations in regions where data are available, leading to strong contrast with regions that are not covered by observations [116]. The unphysical and imbalanced inflated variances may lead to large signal-to-noise ratios in the fits, which could impose spurious adjustments to the ensemble and further lead to poor forecasts [76].

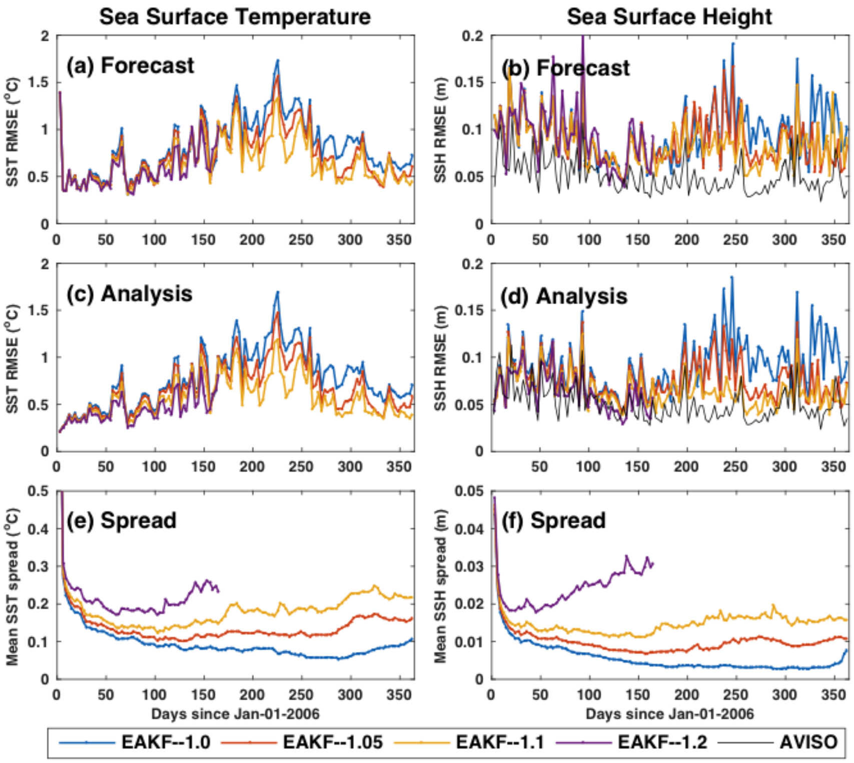


Figure 3.5: Time evolution of the SST/SSH RMSE and ensemble spread for EAKF with different inflation factors.

3.4.2 Sensitivity to the atmospheric forcing: NCEP vs. ECMWF

To test the sensitivity of the assimilation system to the atmospheric forcing product, we compared the results of two assimilation experiments that have been conducted with the EAKF under identical conditions and forced with ECMWF and NCEP fields. The experiments are initialized with an ensemble size of 100 members selected from January climatological fields. The performances of the two experiments are quite comparable in terms of SST and SSH RMSEs (Figure 3.6.a-d). To investigate whether the model without assimilation is sensitive to the atmospheric forcing, we carried out two free model runs respectively forced with ECMWF and NCEP. The free-run results shown in Figure 3.6.e and Figure 3.6.f suggest that the model is sensitive to the atmospheric conditions, and that NCEP and ECMWF do force different circulations in the Red Sea. The disagreement is most pronounced in the SST, probably implying considerable differences in the ECMWF and NCEP heat flux fields in this region. Clearly, data assimilation reduces the SST forecast RMSE (Figure 3.6.a and Figure 3.6.e) in both experiments, with the 1.1°C and 1.2°C RMSE of the free run forced with ECMWF and NCEP respectively reduced to 0.7°C and 0.6°C in the assimilation experiments. Even though the different atmospheric conditions may force different circulation patterns in the Red Sea, the assimilation of remotely sensed SSH and SST data is capable to control the system, at least in the upper layer, and to adjust its circulation according to the available observations.

As atmospheric conditions have a major impact on SST, we further investigate the spatial distribution of the temperature field under different forcing conditions, in the assimilation and free-run experiments. As depicted in Figure 3.7 for two examples in August and in December 2006, the EAKF estimates clearly features stronger vertical variability than the free-run outputs. This is most likely related to the eddies that have been introduced by the filter. In particular, the EAKF forced with both ECMWF

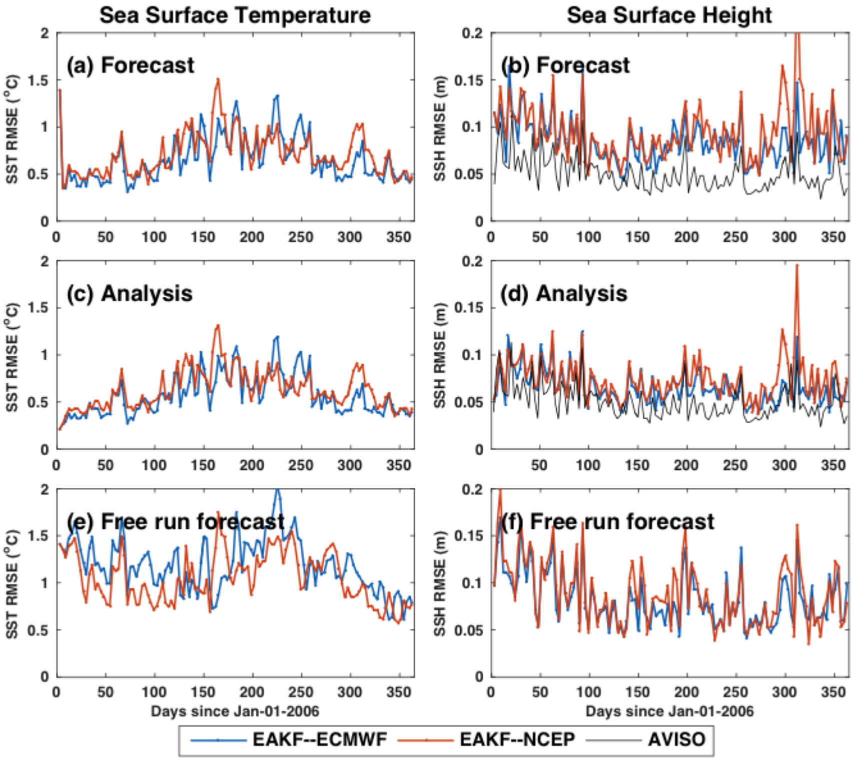


Figure 3.6: Time evolution of the SST/SSH RMSE for EAKF ((a) to (d)) and a free-run simulation ((e) and (f)), both forced with ECMWF and NCEP.

and NCEP exhibit some doming of isothermal in the central Red Sea in August (Figure 3.7.e,f) and in the northern basin in December (Figure12.m,n), which are much weaker in the free run forced with ECMWF (Figure 3.7.g,o) and hardly seen in the free run forced with NCEP (Figure 3.7.h,p). In addition, in the two EAKF assimilation experiments forced with different atmospheric conditions, although little difference is found in their SST RMSEs, and their dominant vertical structures are quite comparable, some of their vertical features are still different. For instance, the slightly depressing isothermal of the southern Red Sea in the ECMWF run (Figure 3.7.e) differs from the doming isothermal in the NCEP run (Figure 3.7.f) in August, suggesting that atmospheric conditions may still influence deep-water structures in an EAKF system as each observation impacts the full water column (no localization was applied in the vertical direction), even when the upper layer is well conditioned by the assimilated data.

3.4.3 Sensitivity to the choice of the background ensemble

3.4.3.1 EAKF vs. EnOI vs. SEnOI

To investigate the behavior of the Red Sea ensemble assimilation system with different choices of ensemble schemes, assimilation experiments were conducted using the EAKF, EnOI and SEnOI. Following the results of Section 3.4.1, the results of the EAKF with 100 ensemble members are used as a reference to evaluate the performances of the two other ensemble schemes.

In contrast with the EAKF, which requires integrating all ensemble members with the MITgcm in the forecast step, EnOI and SEnOI only run the model once, to compute the forecast state starting from the filter analysis, regardless of the ensemble size. One can therefore implement the EnOI schemes with large ensemble members without significant increase in the computational cost. Here we compared the performances of EnOI and SEnOI with 250 ensemble members with those of EAKF with

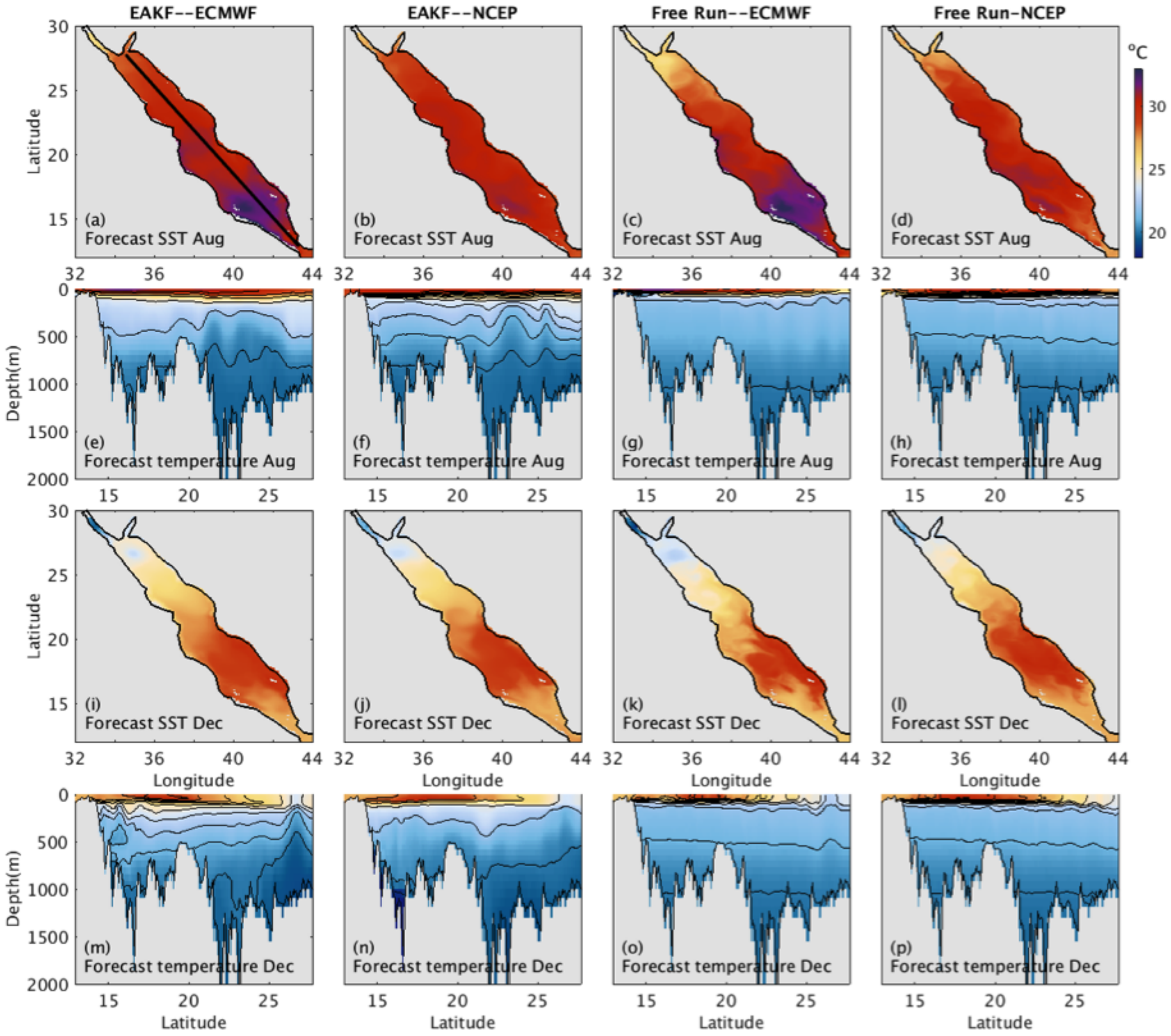


Figure 3.7: Horizontal and vertical distribution of forecast temperature averaged in August (a)-(h) and in December (i)-(p). The left/right two panels plot the forecast fields from an EAKF/free run experiment, forced with ECMWF and NCEP atmospheric conditions, respectively.

100 members.

The results are shown in Figure 3.8. In term of spread, both EnOI schemes exhibit much larger ensemble spreads, whether calculated from a static or a seasonally varying ensemble, as compared to that of EAKF (Figure 3.8.e and Figure 3.8.f). The larger spreads of the forecast ensembles of the EnOI schemes suggest, in some sense, larger forecast errors, which pushes the filter's analysis more towards the observations. This is reflected in the analysis RMSE for both SST (Figure 3.8.c), and SSH (Figure 3.8.d), where the EAKF clearly exhibits a larger RMSE than those of the EnOIs. The differences between the EnOI and SEnOI in terms of their SST and SSH analysis RMSEs are surprisingly not significant, with the latter being comparable to that of the gridded AVISO product.

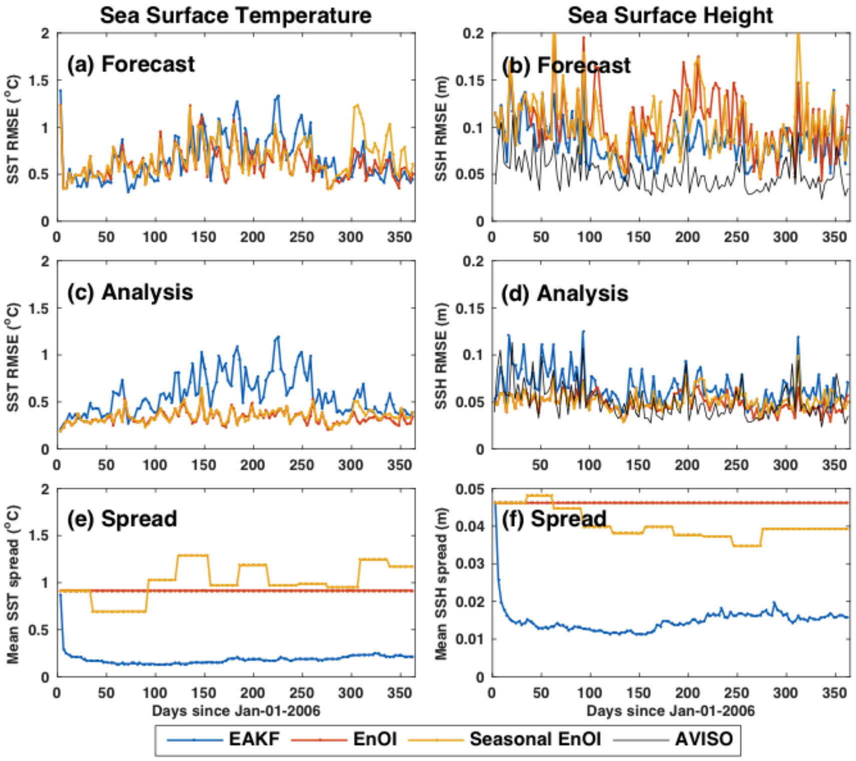


Figure 3.8: Time evolution of the SST/SSH RMSE and ensemble spread for EAKF (ensemble size = 100), EnOI (ensemble size = 250) and SEnOI (ensemble size of 250).

In term of forecast, the performances of the three ensemble schemes are comparable for SST. This is not surprising because the Red Sea SST is dominated by the atmospheric forcing and boundary conditions, which are identical in all three experiments. In forecasting SSH, however, EAKF significantly outperforms both EnOI and SEnOI, which further exhibit an erratic behavior despite using a smaller ensemble. SEnOI generally provides better forecasts than EnOI, except during the winter season, where the static ensemble seems to be well representative of this period.

Remotely-sensed SSH is one of the most used data to describe mesoscale eddies activities in the ocean, and, in practice, provides the most compelling measurements to constrain modeled eddies. The repeat-cycle of satellite altimeters over the Red Sea ranges between 10 to 35 days, which is much longer than the 3-day assimilation window considered in this study. Therefore, unlike the SST observations that are always mapped on the same regular grids, the number and locations of along-track SSH observations vary with the satellites tracks at each analysis step. This means that the forecast SSH RMSE is often evaluated against observations that are not located in the regions where the previous observations were assimilated to produce the most recent analysis (based on which the forecast was computed). One could then consider the SSH altimeter data as independent data to evaluate the filters forecasts. The much better EAKF forecasts suggest better ability to reproduce the hydrodynamics of the Red Sea with a flow-dependent ensemble. In the EnOI and SEnOI experiments, the forecast SSH RMSE (Figure 3.8.b) is quite larger than that of the analysis SSH (Figure 3.8.d), indicating that the ocean model did not adjust fast enough to the (usually large) increments imposed by the no-flow-dependent backgrounds of the EnOI schemes, which probably caused some dynamical imbalances with the ambient water.

We also analyze the forecasts SSH as they result from EnOI and SEnOI. Eddies activity is the most dominant and energetic component of the Red Sea variability, which is usually distinguished from the mean flow as perturbations and can be characterized from the fields of anomalies [196]. This happens to be similar to the generation process of the forecast error covariance $\mathbf{P}^f$ using an ensemble of anomalies, whose members are selected from climatological fields composed of 15-year model outputs. Therefore, the 250 selected ensemble members should be able, to some extent, to represent the eddy variability in the Red Sea. In this case, the constructed $\mathbf{P}^f$ would possibly describe eddy features that happen to be not observed by the sparse altimeter data, or not captured by the forecast state. This could possibly explain why the EnOI with a static ensemble selected in January climatology was able to provide reasonable results. Furthermore, as the Red Sea eddies exhibit a seasonal variability [195, 196], the SEnOI generally out-performs EnOI, particularly in summer, leading to a smaller SSH forecast RMSE. However, the overturning circulation and subsurface intrusion water from the Gulf of Aden also vary seasonally [189, 190], and these features cannot be much improved after assimilating SST and SSH data only. A more robust evaluation would be to also assess the results of the different ensemble assimilation schemes against in-situ profiles in different seasons, but this requires much longer assimilation runs over several years.

As an example, the spatial distributions of the forecast and analysis states on June-6- 2006 as estimated by the three ensemble assimilation schemes are compared with remote sensing observations of SSH (Figure 3.9) and SST/temperature profile (Figure 3.10). The SSH and SST observations are extracted from gridded AVISO and the AVHRR products, respectively. Forecasts from all three schemes agreed well with the remote sensing data, and additionally provided more detailed mesoscale and submesoscale features in the basin than the gridded products. In particular, compared with EAKF, the EnOI schemes introduced stronger eddy activities in the northern Red Sea. This can be clearly seen from the dark blue patches of SSH (Figure 3.9.a-c), the filament features in SST (Figure 3.10.a-c) and the corresponding doming of temperature profile (Figure 3.10.h-j). EnOI and SEnOI also introduced stronger eddies around the altimetry data tracks than EAKF with more pronounced SSH increments (Figure 3.8.h-j). This resulted from the larger ensemble spread, which assigned more weight to the observations in the EnOI schemes (Figure 3.8.e-f). Therefore, EnOIs were more likely to fit the observations and to introduce new features in the analysis fields. These used static ensembles of anomalies that maintain the variability of the Red Sea state (mainly featured with eddy signal) throughout the simulation, while in EAKF, the ensemble tends to converge towards the mean despite the use of inflation, leading to updates that are less impacted by the observations, on basin scales.

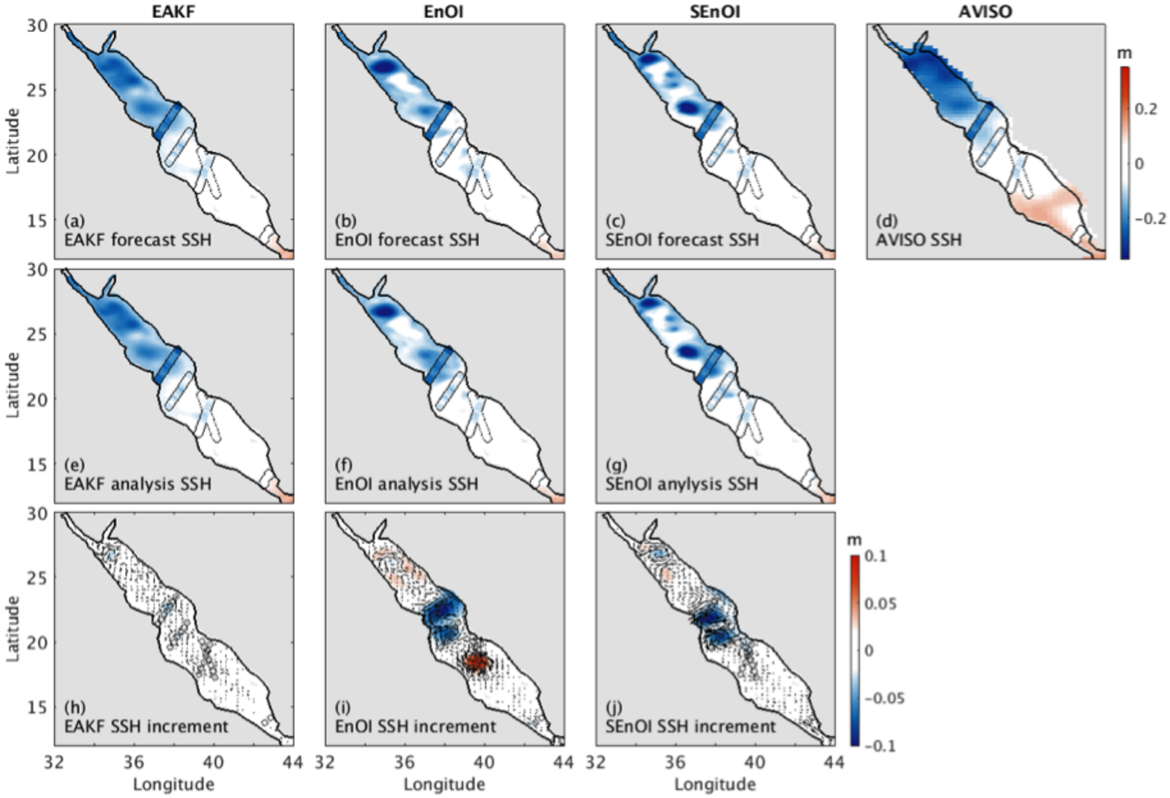


Figure 3.9: (a)-(g): SSH forecast/analysis from EAKF, EnOI and SEnOI compared with gridded AVISO product on June-6-2006, superimposed with the along-track altimeter data. (h)-(j) SSH increment from EAKF, EnOI and SEnOI.

Figure 3.11 shows the spatial distribution of the ensemble spread on June-6-2006

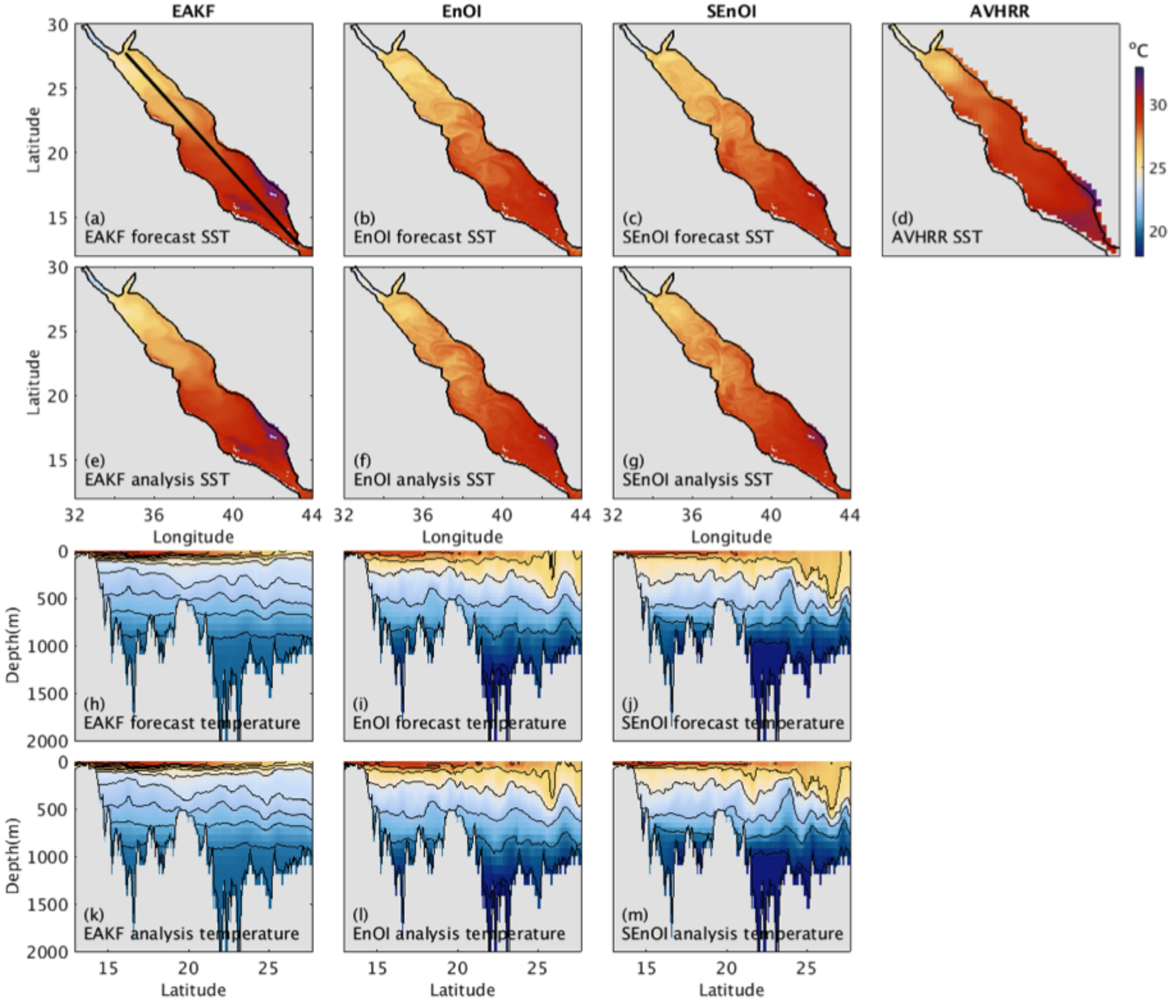


Figure 3.10: (a)-(g): SST forecast/analysis from EAKF, EnOI and SEnOI compared with AVHRR product on June-6-2006. (h)-(m): Vertical structure along a cross section of the Red Sea axis plotted as the black line in (a).

as resulting from the three schemes, showing a significantly weaker spread in EAKF. The EnOI ensemble is selected from January climatology, while the SEnOI ensemble in this example is updated from June climatological fields. We noticed that the ensemble spread of SSH in SEnOI exhibits stronger eddy variability than that in EnOI, but the ensemble spread of temperature in EnOI (Figure 3.11.e and Figure 3.11.h) is larger than that of SEnOI (Figure 3.11.f and Figure 3.11.i), both on the surface and in the upper layers, the latter of which is probably related to a deeper mix layers in winter. The larger SST spread is explained by the stronger interannual atmospheric variability in the winter and the sensitivity of SST to the atmospheric conditions.

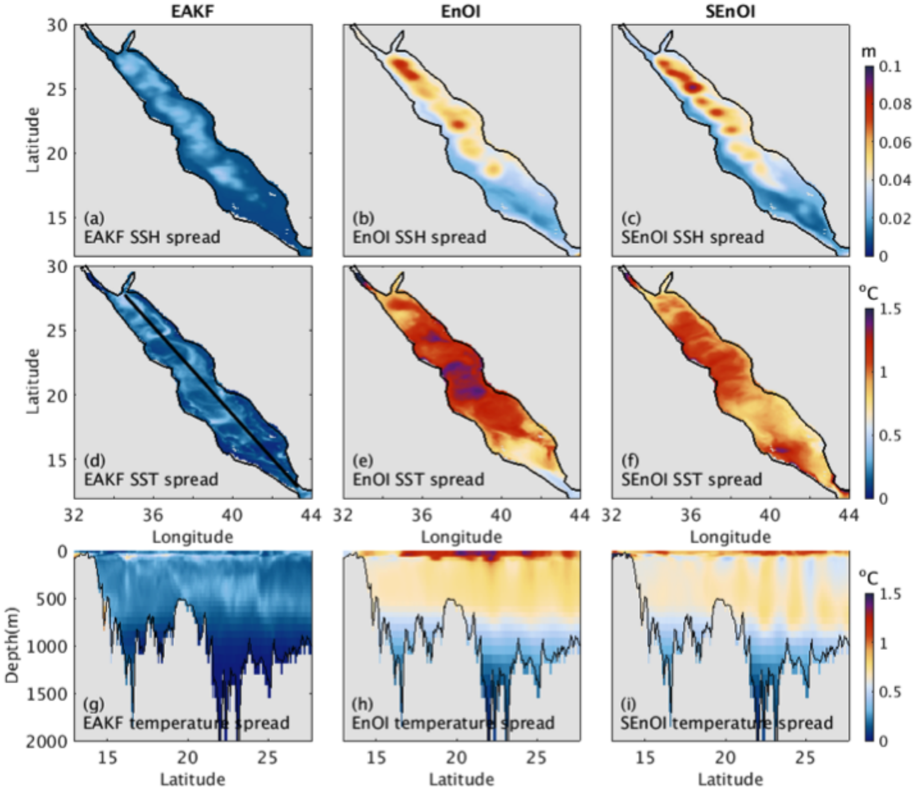


Figure 3.11: Horizontal and vertical distributions of ensemble spread (a)-(c) of SSH, (d)-(f) SST, and (g)-(i) temperature on June-6-2006 as they result from the EAKF, EnOI and SEnOI.

3.4.4 Assessing the dynamical balance of the assimilation solution

It is important to investigate the dynamical balance of the state estimates of the ensemble assimilation schemes. In the analysis fields, the flow is expected to satisfy the geostrophic balance for large-scale and mesoscale phenomenon. In other words, the zonal and meridional residue terms α_x and α_y (Reynolds stress divergences, advection and acceleration terms) defined in the following momentum equations are expected to be small:

$$\alpha_x = \frac{\partial \o}{\partial x} + \mathbb{u} \cdot \nabla u - fv, \qquad \alpha_y = \frac{\partial \o}{\partial y} + \mathbb{u} \cdot \nabla v + fu,$$

where $\o$, $\mathbb{u}$, u, v, f represent the dynamic pressure, the 3-D velocity, the zonal and meridional velocity and the Coriolis parameter, respectively. The residue terms are calculated from the analysis fields, which were updated based on available observations during the assimilation. The zonal and meridional residue terms at a 50 m depth and the relative comparison between the Coriolis term and the horizontal pressure gradient term are plotted in Figure 3.12.

In all these assimilation runs, the geostrophic balance of the analysis fields is well satisfied and the imbalance accounts only for a small portion of the total term. In particular, the new introduced eddies in the EnOI and SEnOI at $20-22$°N (Figure 3.9.h-j) are dynamically balanced. The increments derived in these ensemble-based data assimilation systems can be expressed as $\mathbf{x}^a - \mathbf{x}^f = \mathbf{P}^f\mathbf{H}^T\left(\mathbf{H}\mathbf{P}^f\mathbf{H}^T + \mathbf{R}\right)^{-1}\left[\mathbf{y}^o - h(\mathbf{x}^f)\right] = \boldsymbol{X}'\mathbf{c}$, given an N-dimensional column vector $\mathbf{c}$. The increment at any analysis step is therefore essentially a linear combination of $\boldsymbol{X}'$. The long-term simulation outputs, from which the ensemble members are sampled, are geotropically equilibrated and adjusted with the model dynamics. The same also should hold for the members of the ensemble of anomalies $\boldsymbol{X}'$. Therefore, the increment naturally satisfies the geostrophic balance, which can be seen in Figure 3.9.h-j, where the velocity increments correspond to the SSH increments. Although the large- and meso-scale balances

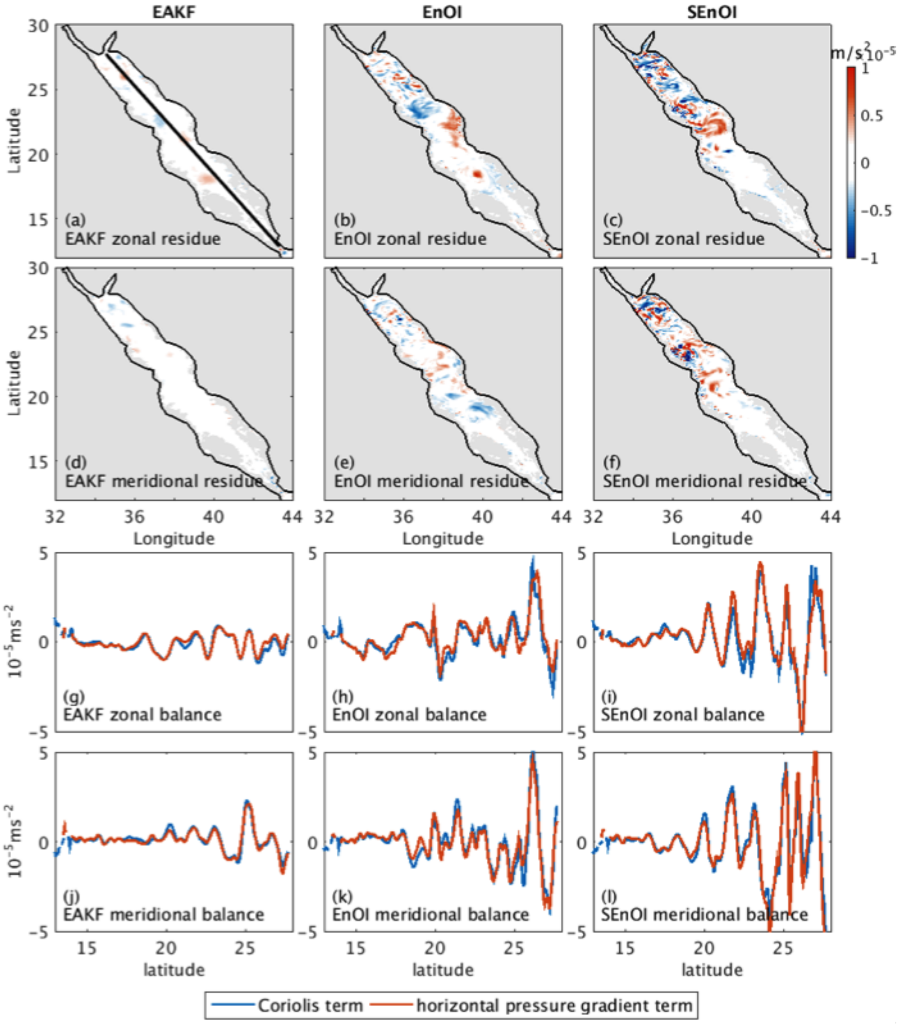


Figure 3.12: (a)-(f) Zonal and meridional residue terms calculated from the analysis fields of EAKF, EnOI and SEnOI on June-6-2006. (g)-(l) Comparison of zonal and meridional Coriolis term and horizontal pressure gradient term for EAKF, EnOI and SEnOI along the Red Sea axis (plotted as the black line in (a)) on June-6-2006.

are warranted, the filter estimates may exhibit some imbalance in small-scale dynamics due to, for instance, stress divergences, acceleration. Localization and inflation may also distort the balance. This inevitable imbalance seems to be generally quite weak and the model is generally able to dynamically adjust it. However, the weak imbalance may be amplified by the magnitude of the increment **c**, which eventually imposes pronounced changes to the analysis fields. If the imbalance is large to top the robustness of the model, the model may sometimes blow up when integrating the members during the forecast step and the assimilation system would break down. In addition, the time-evolution of the basin averaged total residue ($\alpha = \sqrt{\alpha_x^2 + \alpha_y^2}$, not shown) suggests that the EAKF estimates are more dynamically balanced than the OI solutions, often imposing less increments on the forecast because of its smaller ensemble spread.

3.5 Summary and Discussion

This Chapter investigated the sensitivity of the DART-MITgcm system to the choice of the ensemble, and to filtering parameters and atmospheric forcing. Along-track RADS SSH data and gridded AVHRR SST product were assimilated using a three-day window. We investigated the performances of an ensemble Kalman filter, the EAKF as implemented in DART in fully parallel mode, and based on which we have implemented a variant with static covariance, EnOI scheme. The latter does not integrate the ensemble with the MITgcm, offering drastic reduction in computational cost. To deal with the dominant seasonal variability of the Red Sea circulation, the EnOI ensemble was monthly updated by selecting new members from a given climatology of long-term model outputs. This scheme was referred to as seasonal-EnOI, or SEnOI.

An ensemble of 100 members was found enough to obtain good forecasting skills with the EAKF at reasonable computational cost. Increasing the ensemble size to

250 did not improve much the EAKF performances. Inflation is used to artificially increase the ensemble spread and to account for various sources of uncertainties that are either not accounted for or not optimally prescribed in an assimilation system, such as modeling uncertainties, inputs and forcing, filtering approximations, etc. [80]. As such, the value of the inflation factor is system dependent and may vary from one application to another. In our case, we found that an inflation of 1.1 provides the best results, which may give some indications about a suitable value of inflation factor to try in a similar setting. Note that, when the error cross-correlations are well described by the ensemble, the RMSE generally decreases with increased inflation, but of course up to a certain threshold after which one may encounter observations over-fitting issues with more inflation (i.e. larger ensemble spread). A too large inflation might in this case trigger dynamical inconsistency issues, often causing the ensemble to collapse, as we also see in our experiments. The assimilation system was also found not very sensitive, at least for the assimilated surface layer and the dominant vertical structures, to the atmospheric forcing, NCEP or ECMWF fields. Conditioned on the available remote sensing observations, the system is able to adjust the initial state to provide equivalent forecasting products regardless of the forcing product.

Assimilation results from experiments focusing on the relevance of updating the ensemble with the ocean model (EAKF) suggest that, with adequate choice of the static ensemble, (seasonal) EnOI can provide comparable, and even sometimes superior (especially for SST), analysis results. It is important to point out that despite the larger analysis SST RMSE in EAKF (Figure 3.8 (c)), its forecast SST RMSE (Figure 3.8 (a)) is generally better than both EnOI schemes, suggesting proper assimilation of the SST by the EAKF. After all, the best measure of an assimilation system performance is the forecast error, and not the analysis error. The performance of the EAKF with SST could be expected as our analysis suggests that the EAKF ensemble was shown not to carry enough spread for SST, as compared to the other two EnOI schemes (Figure 3.8 (e)). In our experiments, increasing inflation improved the EAKF results, but also often caused the collapse of some of the ensemble members, and thus the filter. A better treatment of the system uncertainties, e.g. through the use of forcing perturbations, is likely to enhance the EAKF performance for SST.

In addition, EAKF clearly provide better forecasts for SSH. A noticeable feature of EnOI and SEnOI, that was seen from the difference between the SSH forecast and analysis RMSEs, is their capability of imposing strong eddies signatures on their analysis fields, but these dynamics seem not quite in line with the ambient water and were often quickly dissipated. Although the geostrophic balance is generally expected in the filters analyses, this may hurt the model forecasting skill when the filter adjustment causes fast-propagating inertial and internal gravity waves. In such an ensemble data assimilation system, the selection of the ensemble subspace is key. It defines the space onto which the model updates are projected. In the Red Sea, the spatial scales of the most energetic and variable components of ocean dynamics (e.g. eddies) are mostly composed of mesoscales or sub-mesoscales (10 km or less) features, which can be described by the anomaly of the mean flow. Such information is reflected from the definition of the forecast error covariance in an ensemble assimilation system. This is a convincing theoretical basis for applying the (seasonal) EnOI method with a static ensemble of state anomalies to enforce eddy variability in the analysis. In contrast, the data-conditioned and flow-dependent ensemble in the EAKF may sometimes lack information about eddies, which may limit its ability to reintroduce them in the analysis.

In the ocean, eddies are often greatly under-sampled due to the sparse available observations, and are therefore likely to be missed given their relatively short spatial scales. In an assimilation system, when eddies happen to be poorly represented in the forecast ensemble, the analysis step is expected to somehow reintroduce these features, provided enough observations coverage. Since the increment between the analysis and the forecast is essentially represented by a linear combination of ensemble members, one should select an ensemble that well represent the eddies, incorporating flow-dependent information to track changes in the system dynamics, and that does not collapse over time. A mixture of dynamically evolving and static ensemble, based on the so-called Hybrid ensemble schemes (Chapter 6), may entertain both features and was one of our targets for the development of the system.

Another limitation of the EAKF is that the assimilation run often terminates when one, or more, members crash during the integration with the dynamical model, probably caused by some imbalances introduced in the analysis step. To enhance the robustness of the EAKF against divergence, we developed the system to enable for automatic replacement of diverged members by new members (Chapter 4) to be selected from a given "dictionary" of system state realizations. This required introducing an optimal selection strategy that is suitable to the unique nature of the circulation and eddy activities in the Red Sea (Chapter 5). Such a strategy would definitely benefit from the availability of independent adequate set of observations (e.g. drifters, HF radar, glider data, etc.), which is presently being deployed in the Red Sea.

Chapter 4

A fault-tolerant HPC scheduler extension for large and operational ensemble data assimilation: Application to the Red Sea

This Chapter corresponds to the paper "A fault-tolerant HPC scheduler extension for large and operational ensemble data assimilation: Application to the Red Sea" published in the *Journal of Computational Science.*

4.1 Introduction

Capabilities in ocean modeling and simulation have witnessed tremendous progress in recent years following the advances in HPC resources [51], the better understanding of the ocean physics, and the availability of ever increasing amount of in situ and remotely sensed data [46, 71].

The celebrated KF computes the best (minimum-variance) estimate of a linear dynamical system given available observations [91], and as such provides a readily efficient algorithm for data assimilation and forecasting [80]. Because of its prohibitive computational requirements when implemented with large scale systems and the nonlinear nature of the ocean dynamics, simplified Kalman filters have been introduced for ocean data assimilation ([80, 142, 178]). One of the most promising Kalman filtering schemes is the EnKF, a MC approach in which the forecast statistics are estimated from an ensemble of model forecasts [84]. An EnKF assimilation system with a high resolution model and large number of observations is expected to require a large ensemble to provide accurate ocean state estimates [78, 82]. Large ensembles should provide more reliable forecast statistics and a smooth forecast covariances for efficient implementation of the filter update steps with the observations.

Increasing the ensemble size would however not only significantly increase the computational load, but would also weaken the robustness of the system and increase the chances of system failure, and thus the workload of the user. Indeed, in case the system crashes, the user will have to manually identify the issue behind its collapse, reconfigure the system and check for consistency before relaunching the jobs. The system failures may be related to a machine problem or may be the result of a dynamical inconsistency between the statistically updated ensemble members and the forecasting model, both of which are unpredictable. The users need therefore to continuously monitor the system execution progress.

In an operational ocean forecasting system, not only huge amount of data need to be processed in a timely manner [124], but the system should also be fault-tolerant in order to recover from failure and deliver real-time responses. In this Chapter, we address these ensemble data assimilation forecasting challenges with our DART-MITgcm assimilation system that we configured for the Red Sea. The system is complex and brings together different components (program executables, data, computational resources). An ensemble of MITgcm runs are integrated in parallel to provide the forecast statistics for the DART filter to perform the assimilation update with the observations. To overcome the aforementioned problems, and build an efficient fault-tolerant ensemble system we coupled the existing DART-MITgcm assimilation system [171] to a scheduler extension named *Decimate* [97]. The system in [171] was neither fault-tolerant nor scalable to ensembles of thousands of members, hence the use of *Decimate* to remediate those limitations. *Decimate* automatically generates the submission scripts along with the dependencies between the jobs and runs them in an environment where checking and restarting functions just need to be defined by the user. It allows an implementation easier to launch and to monitor that can be automatically reconfigured in case of system failure. This work describes the requirements for an operational assimilation system, the coupling of its components and the parametrization of *Decimate*. First results from a high resolution ensemble assimilation system for the Red Sea are presented and discussed.

The Chapter is organized as follows. We first give an overview of the specification and the challenges pertaining to the implementation of an operational DART-MITgcm assimilation system in Section 4.2. Section 4.3, briefly describes *Decimate* on top of which the DART-MITgcm assimilation system was implemented. Section 4.4 presents the results of the assimilation experiments that has been conducted in the Red Sea. Finally, a brief summary and discussion is given in Section 4.5.

4.2 Towards an operational DART-MITgcm system

4.2.1 Specification and challenges for an operational implementation

To run the assimilation workflow in an operational setting, some practical constraints should be taken into account. Figure 4.1 gives a graphical picture of those constraints. Once the initialization is done, the filter starts. In case of failure, the filter is restarted up to *mr* times (*mr* is a shortcut for maximum number of retries). And if after the $mr + 1$ trials the filter still fails, the workflow is aborted and goes in the garbage failure state (not shown in the figure). After a successful filter completion, the N MITgcm instances can begin. The N MITgcm programs run independently and each of them restart up to *mr* times in case of failure. Any successful MITgcm waits in the barrier state for the remaining MITgcms to complete. If any MITgcm still fails to succeed after the *mr* retries, the workflow stops and enters the failure state. Upon successful completion of all the MITgcms, the system is in the barrier state.

Then another assimilation cycle is launched if the required number of cycles is not reached, if not, the system goes to the end state and the workflow finishes successfully. The required time to compute the solution is of course an important factor for an operational system in order to provide needed information for real-time decisions.

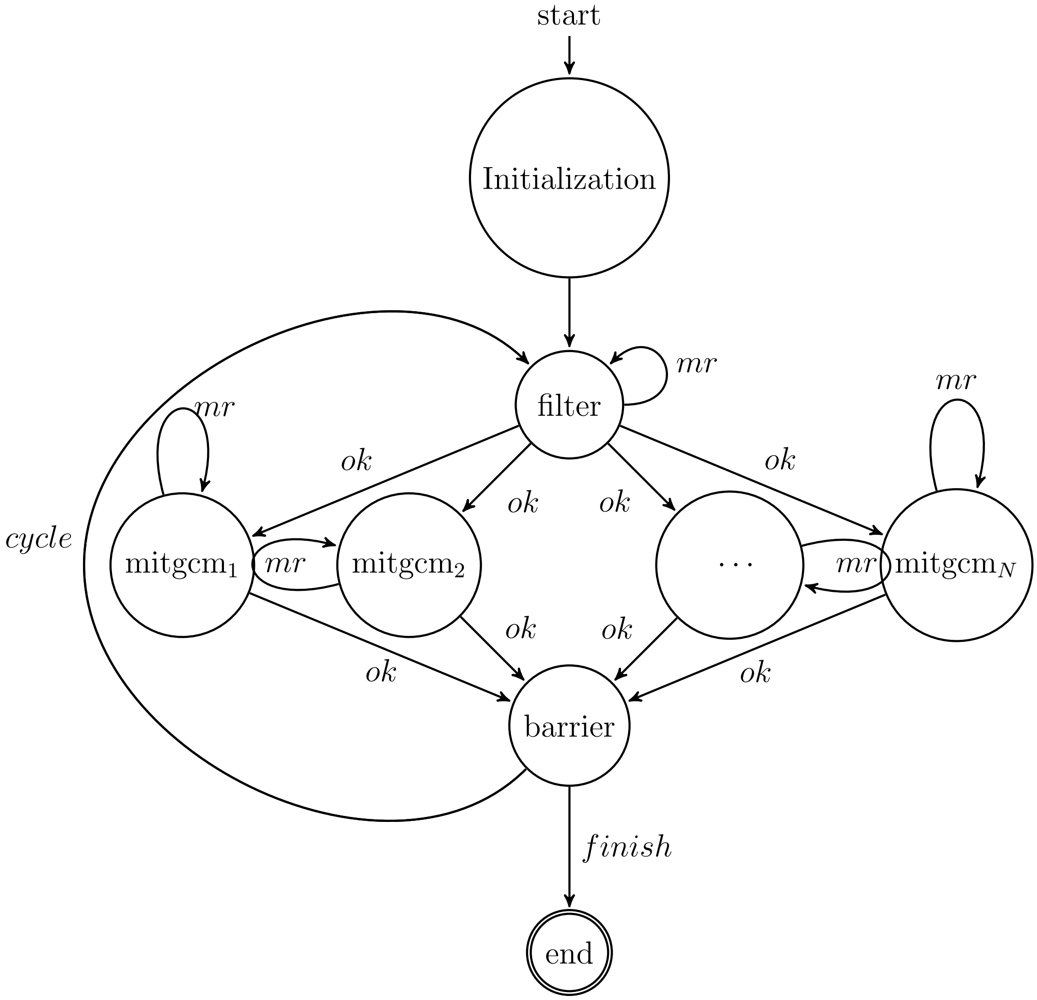


Figure 4.1: Jobs sequence state machine of DART-MITgcm assimilation workflow. *mr* stands for maximum number of retries.

4.2.2 Implementation, issues and limitations

The workflow is designed to run on supercomputers and therefore launching scripts should be written and submitted through a scheduler. Due to the time wall clock policy of many supercomputer centers limiting the execution time of a single job, the full workflow cannot be submitted within one script and needs to be split among many scripts. For that purpose we generate a submission script for each filter state and MITgcm integration. Generating bash scripts helps in writing the submission scripts since a typical run requires ten thousands of jobs and some submission parameters (e.g. the required number of nodes, the time wall clock, the number of cycles, ...) might vary. Indeed, a manual generation of those scripts is not feasible. SLURM (the current scheduler for submitting jobs on KAUST supercomputer Shaheen) allows dependencies handling by means of the command `--dependency`. Moreover the `--array` command is used to submit the MITgcm jobs in parallel with the same scripts lines for all the jobs of a given cycle, making the code more compact. As many other supercomputer centers, our center imposes a limit on the maximum number of jobs per user. Therefore, even though the workflow has been split, it still cannot be submitted if the number of jobs breaks the maximum number of jobs per user limit. Even worse, no assimilation cycle can complete if the ensemble size is greater than the number of jobs limit. Another concern is that the assimilation workflow brings into play huge amount of data so that we need to review and adapt classic data management procedures. The reason is that the stress on the filesystem increases along with the memory usage, which may lead to machine instability and may result in workflow interruption. In other instances the failure could be related to dynamical imbalances in the ocean model as the ensemble assimilation solution is not constrained by the model physics. This is because the linear Kalman update, although statistically optimal (among linear estimators), is not constrained to be dynamically consistent with the physics of the MITgcm (or another ocean circulation model). One may have therefore to often deal with situations during which the MITgcm is not capable of forecasting some of the ensemble members, and the workflow will end. Whatever the failure cause, the workflow needs to restart. The failures are unpredictable and someone has to check from time to time and relaunch the workflow if necessary. The checking process is exhausting, time consuming and inefficient especially for a real-time operational system. To address and solve the mentioned issues, and also to be more compliant with the specification discussed in Section 4.2.1, the existing system is combined with *Decimate*, a scheduler extension described in the next section.

4.3 An efficient implementation of DART-MITgcm workflow based on ***Decimate***, a fault-tolerant scheduler extension

A launching, monitoring and validating tool named *dart_mitgcm* has been specifically designed for developing a fault-tolerant DART-MITgcm ensemble assimilation framework for the Red Sea. Written in Python 2.7, it inspires from the original shell scripts that can be found in DART or MITgcm documentations and relies on our execution framework *Decimate*.

The purpose of this section is not to present *Decimate* itself but to detail how DART-MITgcm was implemented in this environment. More information about the implementation and use of *Decimate* is available at [97]. Distributed as an open source software on Github [95, 96], *Decimate* is freely available and can also be easily installed as a python module distributed from the pypi.python.org repository [98].

4.3.1 ***Decimate***: a robust scheduler extension

In a supercomputing environment, simultaneously accommodating needs of users scalability and capacity is challenging. This often leads to the implementation of a scheduling policy limiting the number of jobs per user in the queue in order to reduce waiting times in queue and optimize turn-around duration. In order to enable efficient use of the computing resources by users producing large number of jobs, *Decimate* was developed by KAUST Supercomputing laboratory to ease the submission, monitoring and dynamic steering of workflow of dependent jobs. Written in Python 2.7, it extends the SLURM scheduler, transparently adding prologue and epilogue to any user script and submit the right job dependency that automatically add new chunks of work or relaunch a job in case of a hardware, software or numerical failure.

Decimate easily allows a user to:

- Submit any number of jobs regardless of any limitation set in the scheduling policy on the maximum number of jobs authorized per user.
- Manage his set of jobs: all the submitted jobs are seen as a single workflow easing their submission, monitoring, deletion or reconfiguration and a centralized log file is created capturing all relevant information about the behavior of the workflow. From Python or shell, at any time and from any jobs, the logging levels info, debug, console and mail are available.
- Via a user-written function, check for correctness of the outputs resulting at the end of a given job and if not make the decision either to stop the whole workflow, to resubmit partially the failing components as is, or to modify it dynamically.

4.3.2 Checking function

Decimate transparently handles the submission, monitoring and resubmission of failed jobs. It is taught what decision has to be made if a part of the workflow failed via a user function written in Python or via a script shell that is passed as a parameter at the initial submission of each MITgcm or DART tasks.

In our case, in a Python function of less than 100 lines, we are checking:

- If output files are produced and contain completeness messages, the task is tagged as COMPLETED and CORRECT. It will not be resubmitted, allowing

Decimate to go on with the next steps of the workflow.

- If no output file is found, the task is tagged as INCOMPLETE and will be resubmitted by *Decimate* if a maximum number of retries has not been reached. This typically happens in the case of hardware failure or if the job required duration has been under-estimated.

- If an error message related to numerical instabilities in the MITgcm forecast step is detected in output or error files, the task is tagged as COMPLETE but INCORRECT. In this case, the "faulty" members are replaced. Many replacement strategies can be implemented. We opted for a dictionary based replacement strategy in which the faulty members are replaced by their closest equivalent among the dictionary members, based on different metrics (l1 norm or l2 norm). These tasks will be resubmitted by *Decimate* if a maximum number of retries has not been reached. In the experiments presented hereafter, the dictionary was constructed from the outputs of a long MITgcm run.

4.4 Experimental setup, application and results

The experimental setup is similar to the one described in [171] and Section 3.2.1, except that only the ECMWF forcing is applied. The assimilation experiments are conducted over a 2 months period starting on January-1-2006 and includes 20 assimilation cycles, one update step every three days. The updates were performed based on a deterministic EnKF, the EAKF [8]. Four experiments are performed (as summarized in Table 4.1): two experiments with 1000 members to assess *Decimate* efficiency, and two experiments with 100 members as references to evaluate the overall behavior of the assimilation system. This is the first reported 1000-members EnKF run with a high resolution general circulation ocean model. Two of these experiments (with 100 members and 1000 members) use a localization cutoff radius of 0.05 rad (about 300 km), while the remaining two do not apply localization. Moreover, an inflation factor of 1.1 is used in all the experiments.

Table 4.1: Experiments.

	100 members	1000 members
localization radius = about 300 km	experiment 1	experiment 2
no localization	experiment 3	experiment 4

4.4.1 *Decimate* assessment

Using *Decimate*, we experienced that handling a workflow involving 1000 members and 20 assimilation cycles was a relatively smooth process. All the experimentation took place on Shaheen from August 21 to September 07 2017 where the average load of the machine was around 90%.

22000 independent successful runs of MITgcm were executed. During the process:

- 617 failed MITgcm runs did not complete due to model failures and were followed by a replacement of members ($\leq 3\%$).
- Roughly 10% of jobs failed because of hardware failures, especially before a maintenance period scheduled on Sept 3, when Shaheen lustre filesystem was highly solicited by other users and instabilities occurred. Our fault-tolerant assimilation system was able to resubmit those jobs.

Table 4.2: Number of members replaced in each Experiment.

Experiment	*100 members*	*1000 members*
With localization	None	9 at cycle 7, 2 at cycle 10, 583 at cycle 11
Without localization	None	7 at cycle 5, 15 at cycle 10, 1 at cycle 20

Before this new implementation, some attempts to handle similar workflow had been made successfully by our team. But while reaching a complete simulation with

100 members, at least 20% of time was spent in the manual steering of the workflow and the multiple manual correction and resubmission of jobs after sporadic hardware failures or numerical issue. Reducing this overhead to a minimum thanks to the automation of restart and decision making in case of glitch, *Decimate* greatly eased the launching and monitoring process and made the system more trackable even for a higher number of members.

4.4.2 Assimilation performance

One key assumption in the EnKF is the distribution of the forecast error to be Gaussian, based on which the members are updated with the observations using the Kalman linear correction step. The distribution of the forecast error, i.e. the prior distribution, is estimated from the statistics of the forecast ensemble anomalies. We first assess the relevance of this assumption in our setup by analyzing the histogram of SST and SSH ensembles at three locations in the northern, central and southern Red Sea at assimilation cycle 4 as shown in Figures 4.2 and 4.3, respectively. The figures suggest that, for both SST and SSH, the prior distribution with 1000 members is clearly more Gaussian than that with 100 members. A smaller ensemble drastically reduces the computational cost associated with the MITgcm ensemble forecast runs, but seems to provide a more scattered ensemble and less Gaussian MC-based approximation of the prior distribution, which may limit the efficiency of the Kalman-based update step of the EnKF.

It is important to monitor both the state estimation forecast and analysis errors to make sure that the filter update is efficient at improving the forecast and that the resulting analysis state is compatible with the MITgcm dynamics. The time-evolution of the RMSEs between SST/SSH observations and filter forecast/analysis states as they result from the different experiments are plotted in Figure 4.5. The analysis RMSEs of both SST and SSH are smaller than their forecast counterparts, suggesting

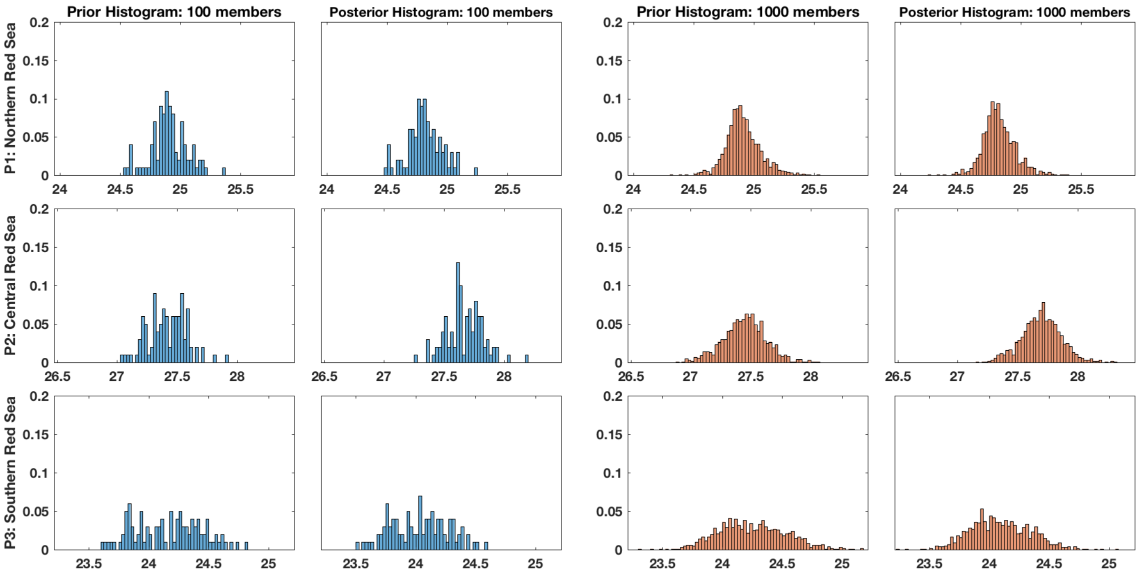


Figure 4.2: The histograms of forecast (Prior) and analysis (Posterior) in experiments 3 ($N = 100$, blue panels) and 4 ($N = 1000$, red panels) based on SST ensembles at three selected locations in the northern, central and southern basins of the Red Sea, as the 1st, 2nd and 3rd rows, respectively.

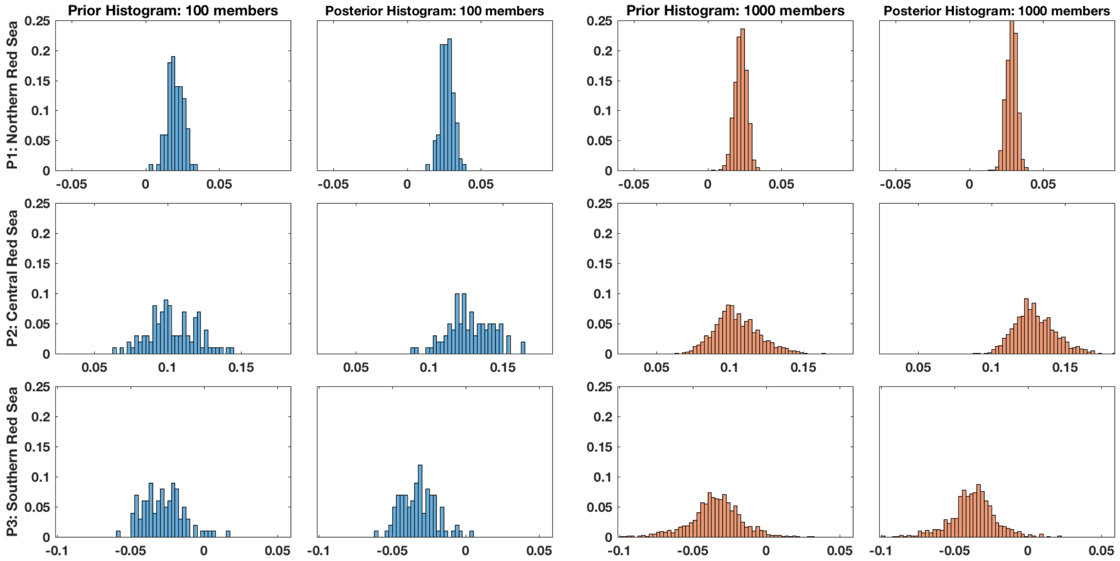


Figure 4.3: The histograms of forecast (Prior) and analysis (Posterior) in experiments 3 ($N = 100$, blue panels) and 4 ($N = 1000$, red panels) based on SSH ensembles at three selected locations in the northern, central and southern basins of the Red Sea, as the 1st, 2nd and 3rd rows, respectively.

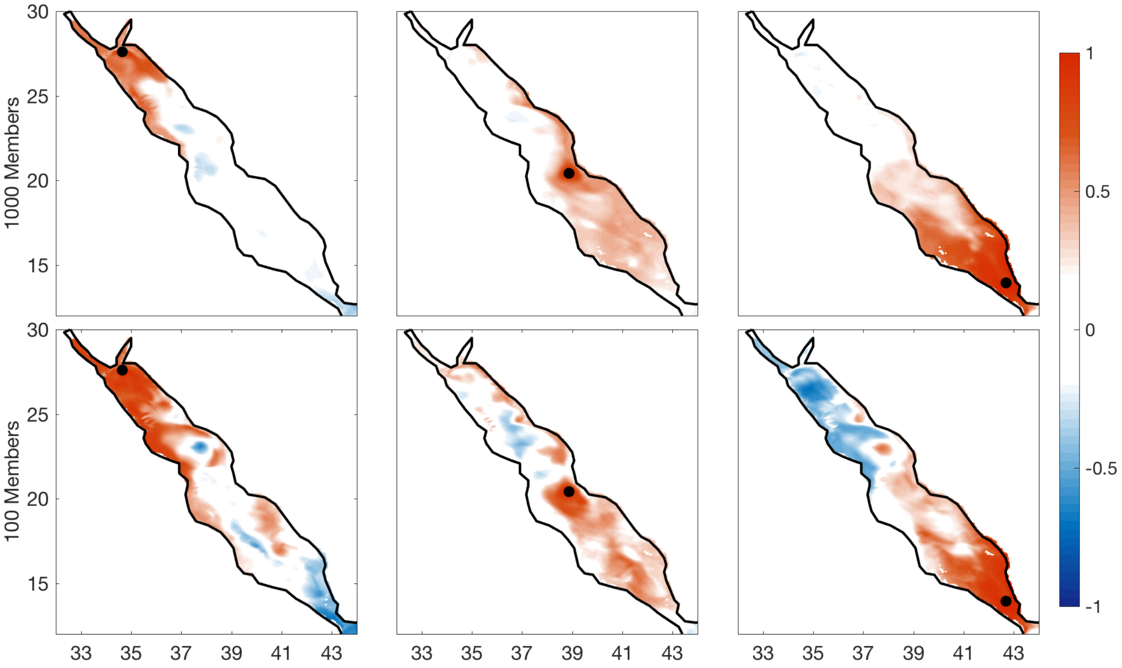


Figure 4.4: Sampled correlations for SSH as computed from the assimilation runs with 1000 (upper panel) and 100 (lower panel) members at three selected locations in the northern (1st column), central (2nd column), and southern (3rd column) basins of the Red Sea, respectively.

the filter's efficiency at providing reliable estimates. Compared with SST RMSE, the SSH RMSE is more fluctuating since the model-data difference was calculated with different along-track SSH observations whose locations vary from one step to another. A comparison between the blue and red curves plotting the filter RMSEs with 100 and 1000 members, respectively, suggests that the RMSEs of both SST and SSH generally decrease as the ensemble size increases from 100 to 1000. Figure 4.5 (e) and (f) suggest that the ensemble spread, an indicator of the filter estimates uncertainties, is quickly reduced after the first few assimilation cycles before leveling off. The ensemble spread is further better maintained with 1000 members, which should impose more pronounced filter's updates.

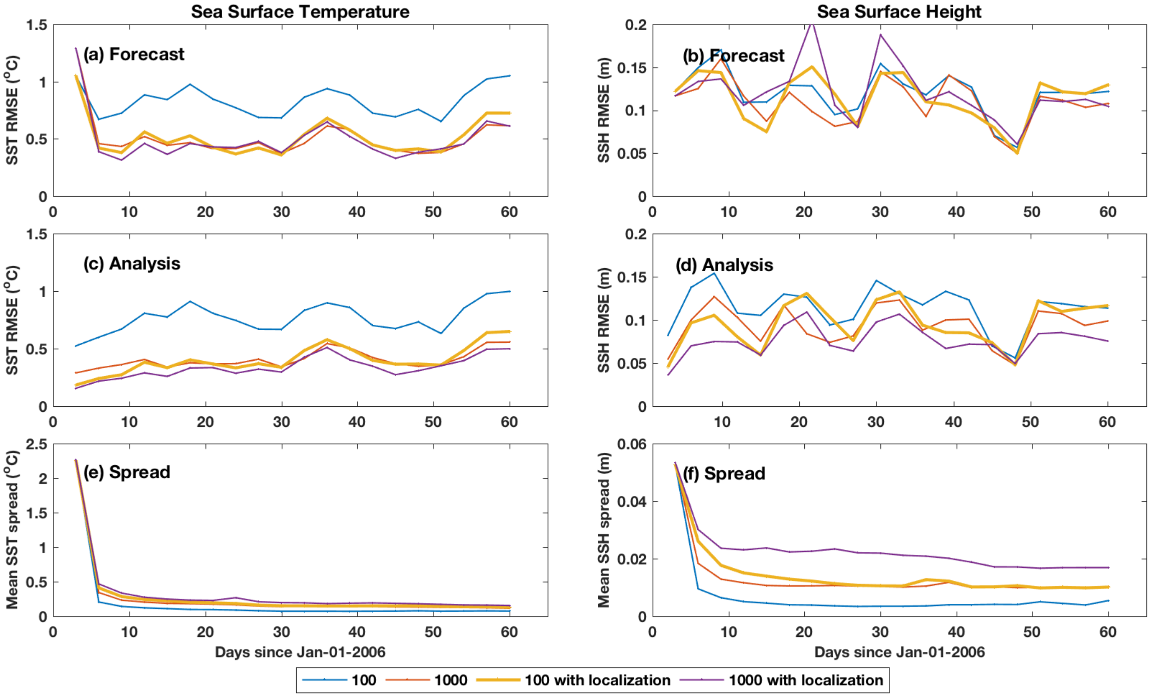


Figure 4.5: Time evolution of the forecast and analysis SST/SSH RMSE and ensemble spread with ensemble size of 100 (blue), 1000 (red), 100 with localization (yellow) and 1000 with localization (purple).

The error covariance in the EnKF, along the modes of which the filter's update is applied, is estimated via the forecast ensemble anomalies. These only provide $N-1$ directions in phase space, which means that the update step will not be able to exploit more that $N-1$ "information" from the observations. To deal with this low rank problem and to remove eventual spurious long-range correlations, a covariance localization [74] cutoff radius of about 300 km is implemented. Localization is a simple technique that enables efficient implementation of an EnKF with small ensembles. As shown in Figure 4.5 (b) and (d), the experiment 1 ($N = 1000$ with localization, purple curve) has the smallest analysis, but not forecast, RMSE of SSH. This means that the data might have been over-fitted in this experiment, and suggests testing with larger covariance localization scale. In addition, a close examination of the results shows that localization also helps to maintain the spread of both SST and SSH with proper tuning. Using more members allows to rely less on localization and improves the filter's performance, but at a significant increase in computational cost. Indeed, Figure 4.4 shows that the correlation range with 100 members is wider than the one with 1000 members, and that the impact at the selected points are more localized with 1000 members.

As an illustration of the system performance, the spatial distribution of the forecast and analysis states and their increment (the difference between the analysis and forecast states) on Jan-12-2006 are compared with remote sensing observations of SST (Figure 4.6) and SSH (Figure 4.7). Overall, the forecast and analysis fields agree well with the remote sensing data. By exploiting the high-resolution model dynamics, the results provide more details of the mesoscale variability in the basin, which is one of the key features of the Red Sea circulation [195, 196]. The distributions of lower SST in the northern Red Sea in experiment 4 ($N = 1000$, Figure 4.6-a) and experiment 1 ($N = 100$ with localization, Figure 4.6-c) are closer to observations (Figure 4.6-d) compared with experiment 3 ($N = 100$, Figure 4.6-b). A better maintained ensemble spread also helps to extract more information from along-track SSH data, as can be seen in the increments fields of SSH in experiment 4 ($N = 1000$, Figure 4.7-h) and experiment 1 ($N = 100$ with localization, Figure 4.7-j) compared with experiment 3 ($N = 100$, Figure 4.7-i). The increment fields of both SST and SSH show that an EnKF with larger ensemble generally leads to smoother analysis states.

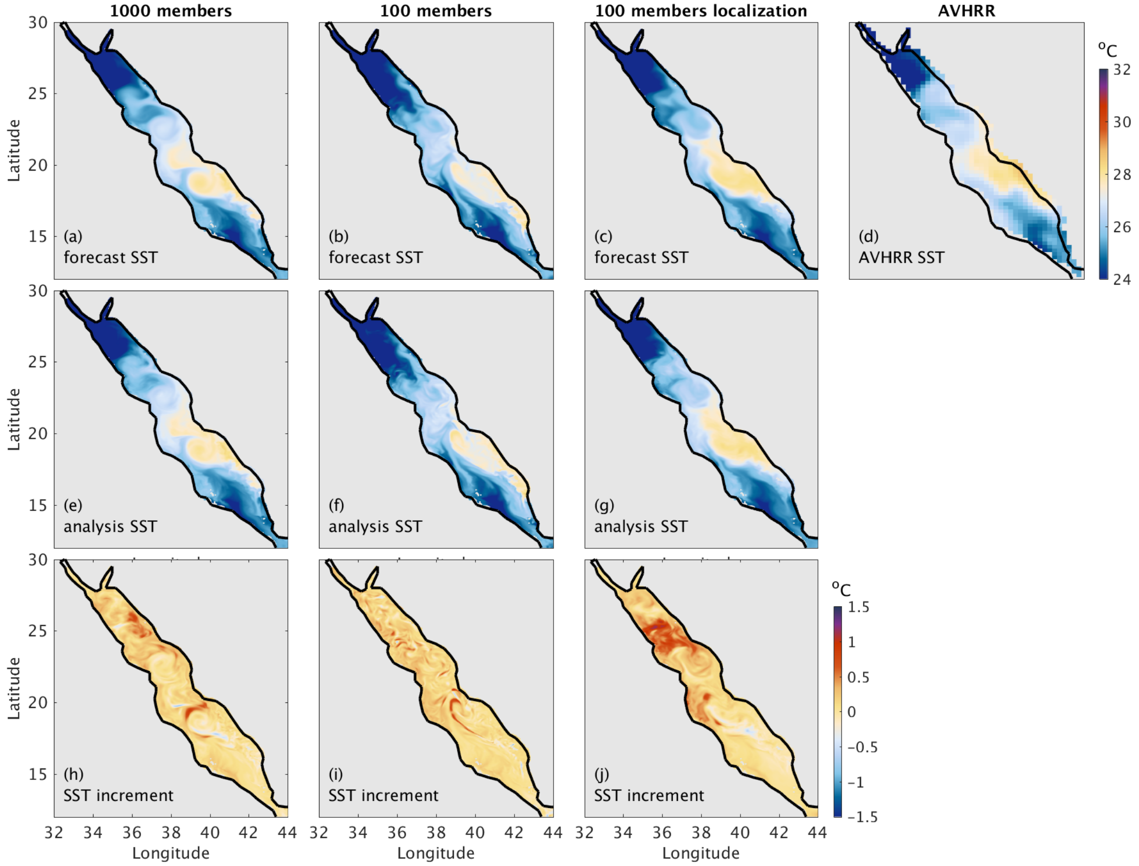


Figure 4.6: SST forecast/analysis/increment from assimilation experiment with ensemble size of 1000, 100 and 100 with localization compared with gridded AVHRR product on Jan 12, 2006.

4.5 Conclusions and discussion

Numerical prediction of oceanic conditions is of foremost importance for navigation, offshore operations, fisheries, and many other marine activities. However, ocean models are never perfect and can be subject to many sources of uncertainties. Data

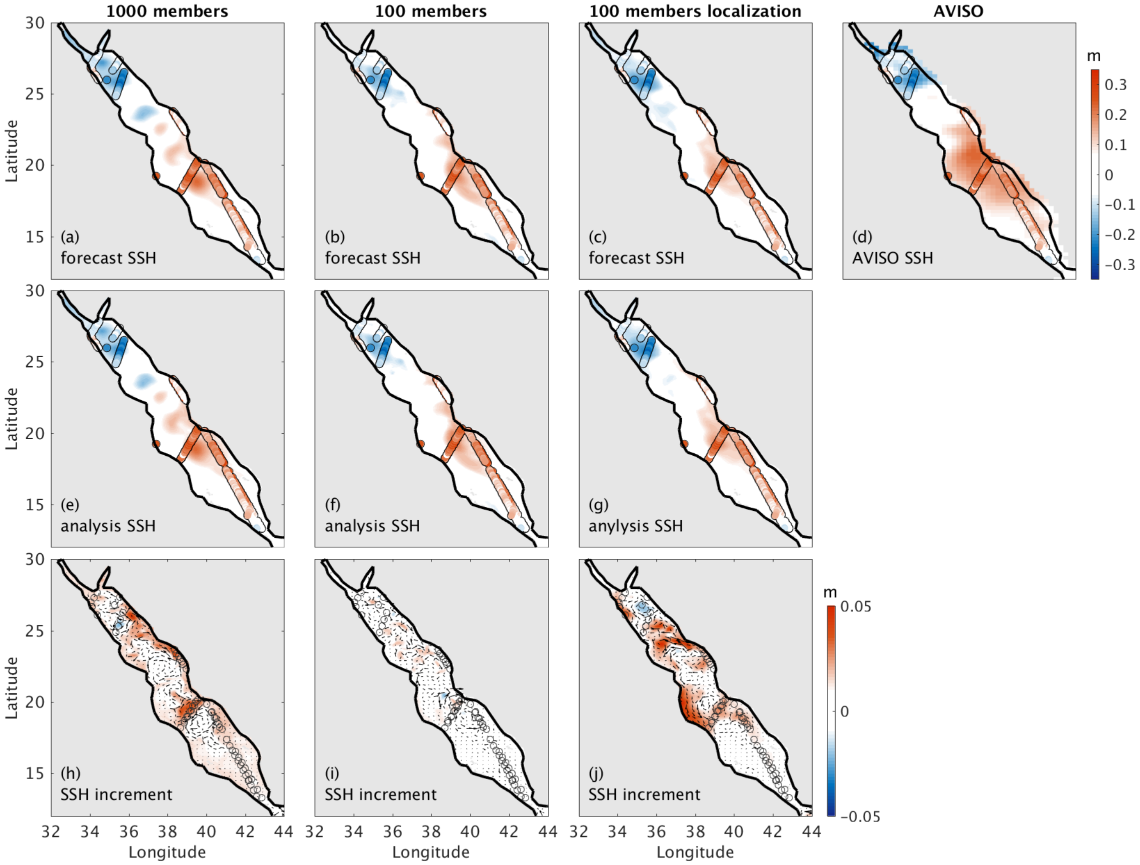


Figure 4.7: SSH forecast/analysis/increment assimilation experiment with ensemble size of 1000, 100 and 100 with localization compared with gridded AVISO product on Jan 12, 2006, superimposed with the alongtrack altimeter data.

assimilation combines a prior knowledge of the ocean state from numerical simulations with the observations to provide best possible ocean state estimates along with their uncertainties. The EnKF, a popular MC data assimilation scheme, is now widely used by the community.

In realistic oceanic applications, the ensemble size is restricted by the computational cost of integrating the numerical ocean model. Using a small ensemble $(10 - 100)$ would limit the filter performance, to fit the data at the update step and to provide reliable spread after the forecast step. Although inflation and localization techniques have been proven efficient at mitigating these problems, physical balances in the model could suffer from arbitrary inflation and localization. With the recent tremendous advances in the developments of HPC resources, an EnKF system with large ensembles $(1000 - 10000)$ becomes feasible and less dependent on these auxiliary techniques.

Nevertheless, increasing the ensemble size introduces new issues and difficulties. Obviously the computational load is the first challenge to deal with. For this study, we were able to mitigate this issue with the KAUST world-class supercomputer, Shaheen. Another challenge when dealing with a large ensemble is the increasing risk of system failure. The assimilation may indeed impose some increments that are not compatible with the model dynamics, and the system will also be exposed to a higher chance of system failure with large number of heavy jobs running. The interruption of any single member will cause a collapse of the whole ensemble assimilation system.

We developed a fault-tolerant ensemble data assimilation system based on the state-of-the-art MITgcm for forecasting and DART for assimilation, and a newly designed scheduler extension, *Decimate*. One key and simple parametrization of *Decimate* consists in describing what to do in case of hardware or numerical failure: here we detailed our choices in this matter. With this parametrization, *Decimate* made it possible for the system to recover from failures, without human intervention. This enabled the implementation of a high resolution ensemble assimilation system for the Red Sea with thousands of ensemble members. This study described the development of the system and its components. From our preliminary experiments with an ensemble with 1000 members, we demonstrated significant improvements in the system performances, and as expected, less dependence on inflation and localization.

Chapter 5

An adaptive Ensemble Optimal Interpolation for cost-effective assimilation in the Red Sea

This Chapter corresponds to the paper “An adaptive Ensemble Optimal Interpolation for cost-effective assimilation in the Red Sea” submitted to the *Journal of Computational Science.*

5.1 Introduction

Ensemble assimilation methods have been proven very efficient in many ocean applications and regions (e.g. [32, 76, 132, 171, 186]). The performance of these methods greatly depends on the representativeness of their ensembles, which should be large enough to describe the directions of errors growth of the system and mitigate the effects of sampling errors [82, 109, 165]. Using large ensembles in an EnKF, such as EAKF, means more numerical model integrations and therefore increased computational cost [171]. EnOI integrates only the filter estimate (i.e. analysis) to compute the forecast and updates the latter with the incoming observations based on the sample covariance of a pre-selected ensemble, as a way to reduce the number of model runs. A stationary ensemble may however not properly capture the striking seasonal variability of the Red Sea dynamics [189, 190]. A Seasonal EnOI, which uses seasonally varying ensembles [186], was successfully implemented in the Red Sea [80, 171]. It was however not very efficient at describing the prevailing eddy and mesoscale activities in the basin [195].

The use of pre-selected time-varying ensembles that represent the different seasons of the studied basin has already been proposed in EnOI [179, 186]. Here we propose to push this idea further by adaptively and automatically selecting, at every filter analysis step, a new ensemble from an available "dictionary" of representative ocean states (e.g. long reanalysis). As in an EnKF, this will enable updating the ensemble of the EnOI scheme, in order to describe the state uncertainties at the time of the analysis step, while avoiding the costly numerical integration of its members. The selection of the ensemble members will be based on the best estimate of the state at the time of the analysis, i.e. the forecast state. This new Adaptive EnOI (AEnOI) scheme will therefore only integrate the model once to forecast the state, and then select an ensemble from the dictionary that represents its current uncertainties according to a certain criteria, based on which the Kalman analysis step will be applied.

Similar ideas have been recently proposed, relying on some kind of dictionary to describe the uncertainties or statistics of the estimate of interest. [168], for instance, suggested a fully data-driven ensemble data assimilation framework that selects the "best" ensemble members from a given "catalog" of possible successive states of the system based on a "analog" or "nearest-neighbor" approach. The nearest-neighbor approach is adopted from the machine learning community and is basically designed to find the closest, according to some metric, possible successor state given the current state of the system. This however amounts to replace the dynamical ocean model by a purely data-driven model. A closely related approach was proposed by [93] based on the so-called Takens approach, replacing the dynamical model with a delay coordinate embedding model. Another technique, known as Dynamic Ensemble Update (DEU) was introduced in the context of an EnKF [156], but uses a particular dictionary of sparse realizations to sparsify the filter estimate. Other approaches also resorted to some dictionaries to account for some missing physics [22], or to simplify the complexity of the dynamical model [44].

The approach we propose here is somehow different; it uses the full dynamical model for forecasting and the dictionary to provide a possible set of realizations (ensemble members) that represents the current uncertainties based on the forecast. We present and discuss different metrics to select the new ensemble members from the dictionary, and test their relevance with a realistic ensemble data assimilation exercise using a high resolution MITgcm model in the Red Sea. The Chapter is organized as follows. Section 5.2 recalls the EnKF and EnOI algorithms. Section 5.3 presents the adaptive EnOI algorithm and discusses approaches to select the ensemble members from the available dictionary. Section 5.4 presents the results of the implementation of the adaptive EnOI algorithm with the Lorenz 63 and 96 models. Section 5.5 describes the general circulation ocean model and its configuration, as well as the assimilated observations. It also outlines the design of the conducted assimilation experiments, and discusses the filters performances and results. Finally, Section 5.6 concludes the work with a summary of the main findings and a discussion on the future directions.

5.2 Ensemble Data Assimilation

The data assimilation problem with the Ensemble Kalman filter is described following the state-space model formulation

$$\mathbf{x}^t_{k+1} = \mathcal{M}_k(\mathbf{x}^t_k) + \boldsymbol{\eta}_k, \tag{5.1}$$

$$\mathbf{y}^o_k = h_k(\mathbf{x}^t_k) + \boldsymbol{\varepsilon}_k, \tag{5.2}$$

where $\mathcal{M}_k$ denotes the model representing the ocean dynamics, $\mathbf{x}^t_k$ is the true state vector at time k, and $\boldsymbol{\eta}_k$ is the model error. $\mathbf{y}^o_k$ is the observation vector, which is related to the state via the measurement operator h_k and $\boldsymbol{\varepsilon}_k$ represents the observational error. Both $\boldsymbol{\eta}_k$ and $\boldsymbol{\varepsilon}_k$ are assumed independent and normally distributed of mean zero and covariance matrices $\mathbf{Q}_t$ and $\mathbf{R}_t$, respectively.

As a variant of the well-known Kalman filter (KF) [91], the EnKF represents the statistics (first two moments) of the system state by a collection of random realizations, or ensemble members [50, 77]. The estimate at any given time is then given by the sample mean and the error covariance is approximated by the sample covariance of the ensemble [50]. Here we adopt a deterministic formulation of the EnKF [77, 81]. Given a forecast ensemble of N members at time step k forming the matrix $\mathbf{X}_k^f = [\mathbf{x}_k^{f,1}, \mathbf{x}_k^{f,2}, \cdots, \mathbf{x}_k^{f,N}]$, with $\mathbf{x}_k^{f,i}$ denoting the i-th ensemble member at time k. The forecast ensemble anomaly is

$$\boldsymbol{X}_k^{f\prime} = \mathbf{X}_k^f - \frac{1}{N}\left(\sum_{i=1}^{N}\mathbf{x}_k^{f,i}\right)\mathbf{e}_{1\times N}, \tag{5.3}$$

with $\mathbf{e}_{1\times N}$ denoting the matrix with ones as elements and size $1 \times N$. At the analysis step, once an observation $\mathbf{y}_k^o$ becomes available, the forecast state $\mathbf{x}_k^f$, which is the mean of $\mathbf{X}_k^f$, is updated using the standard Kalman filter correction step to obtain the analysis state

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k\left(\mathbf{y}_k^o - h_k(\mathbf{x}_k^f)\right), \tag{5.4}$$

where $\mathbf{K}_k = \mathbf{P}_k^f\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{P}_k^f\mathbf{H}_k^T + \mathbf{R}_k\right)^{-1}$ is the Kalman gain. The forecast error covariance $\mathbf{P}_k^f$ is estimated as $\frac{1}{N-1}\boldsymbol{X}_k^{f\prime}\boldsymbol{X}_k^{f\prime T}$, and the associated analysis error covariance as

$$\mathbf{P}_k^a = [(\mathbf{P}_k^f)^{-1} + \mathbf{H}_k^T\mathbf{R}_k^{-1}\mathbf{H}_k^T]^{-1}.$$

In the EAKF formulation, a matrix $\mathbf{A}_k$ is introduced such that $\mathbf{P}_k^a = \mathbf{A}_k\mathbf{P}_k^f\mathbf{A}_k^T$. Based on a judicious choice of $\mathbf{A}_k$, an analysis ensemble is then resampled as $\mathbf{X}_k^a = \mathbf{A}_k\left(\mathbf{X}_k^f - \frac{1}{N}\left(\sum_{i=1}^N \mathbf{x}_k^{f,i}\right)\mathbf{e}_{1\times N}\right) + \frac{1}{N}\left(\sum_{i=1}^N \mathbf{x}_k^{a,i}\right)\mathbf{e}_{1\times N}$ in such a way to match the analysis $\mathbf{x}_k^a$ and covariance $\mathbf{P}_k^a$ before it is integrated by the model (5.1) to compute the next forecast [77], allowing for a dynamical update of the estimation error. A new assimilation cycle starts once the new observation becomes available.

EnOI is the optimal Interpolation (OI) variant of the EnKF [52] in which a pre-selected ensemble remains static during all assimilation cycles, with no feedback from the assimilation system to modify the forecast (background) covariance [155]. In the EnOI forecast step, only the analysis state is integrated by the dynamical model for forecasting, before it gets updated again with the incoming observation based on the pre-selected ensemble [52]. EnOI therefore leads to a drastic computational cost reduction (by almost a factor N) compared to an EnKF, and no resampling step is needed after the analysis. It further does not suffer from the typical ensemble collapse of the EnKFs, which often requires artificial inflation of its ensemble [8]. This makes EnOI a computationally very efficient approach for ensemble assimilation and was shown to be particularly robust in numerous ocean applications [32, 132, 171, 186].

5.3 Adaptive EnOI

The use of representative background covariances is critical for the performance of any data assimilation scheme, as these should describe the spatial and multivariate structure of the subspace in which the update with the observation is performed [72, 112]. In particular, the behavior of ensemble assimilation methods largely depends on the representativeness of their (forecast) ensembles, based on which the background covariance is estimated [165]. The ensemble should (i) describe the directions of estimation errors growth, and therefore be time-variant to follow their dynamical evolution [77, 80], and (ii) be large enough to infer reliable statistics between the observations and the forecast state and to provide enough rank (degrees of freedom) to fit the data [72, 78]. In realistic large scale applications with general circulation ocean models, however, EnKFs can be only implemented with relatively limited ensembles O(10 members) to maintain a manageable computational load [77]. This usually results in rank-deficient background covariances that require various auxiliary techniques, such as covariance localization [85] inflation [72], to infer reasonable forecast increments from the incoming observations. Localization restricts the action of the increments only to grid points close to the observation, which helps increasing the background covariance rank and filters spurious correlations [84]. Another typical concern with ensemble data assimilation systems is the loss of spread in the forecast ensemble, which is associated with the dissipative nature of ocean models [82, 109] and the often misrepresented sources of model errors [83]. This is often mitigated through simple ensemble inflation and/or stochastic perturbations (of the parameters and inputs) techniques [52].

EnOI schemes efficiently resolve the issue of computation load, which enables the use of large ensembles without cost increase. A static ensemble may however not always be representative of the modeled dynamics, especially when dealing with rapidly varying dynamics and those that experience sudden regime changes [80]. To deal with the pronounced seasonality of the South China Sea, [186] suggested preselecting seasonal representative ensembles and then use these in an EnOI according to the season during which the observations are assimilated.

We propose here to push further the idea of using a time-varying ensemble in EnOI, not only by utilizing static ensembles by selecting a new ensemble on a seasonal basis, but at every assimilation cycle to account for the mesoscale and eventually intra-seasonal variability. We propose here to select the new ensemble after every forecast step from an available historical set, or “dictionary”, of ocean states describing the variability of the studied basin. The selection of the ensemble will be based on the best available information at the time of the update step, which in the context of an EnOI is the forecast state. The proposed assimilation workflow is schematized in Figure 5.1 and we will refer to it as the Adaptive EnOI, or AEnOI. The selection of the ensemble members from the dictionary is the corner stone of the proposed approach and is discussed in the next section.

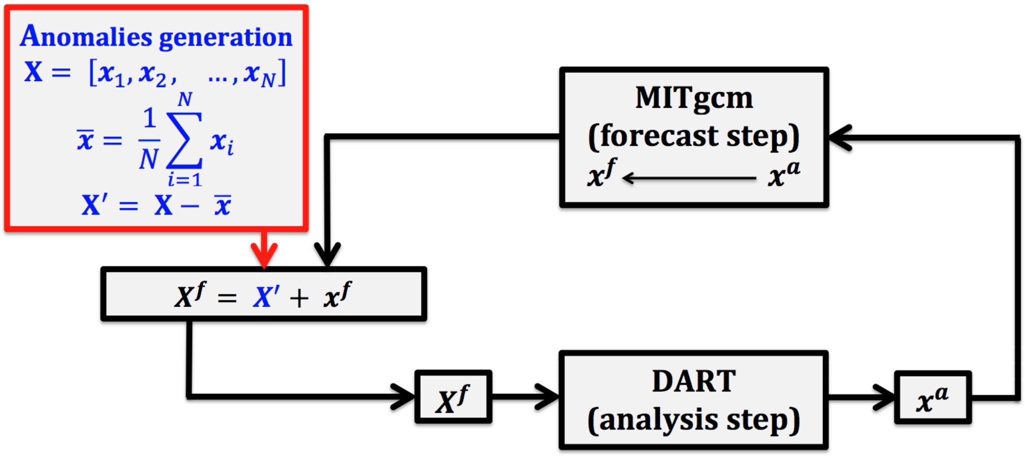

Anomalies generation
$\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ \ldots, \boldsymbol{x}_N]$
$\overline{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$
$\mathbf{X}' = \mathbf{X} - \overline{\boldsymbol{x}}$
MITgcm
(forecast step)
$x^f \leftarrow x^a$
$X^f = X' + x^f$
X^f
DART
(analysis step)
x^a

Figure 5.1: Workflow of the dictionary-based AEnOI schemes as implemented with DART. The forecast state $\mathbf{x}^f$ is used to select an ensemble $\mathbf{X}^f = [\mathbf{x}^{f,1}, \ldots, \mathbf{x}^{f,N}]$ from an available dictionary. This ensemble is then centered around $\mathbf{x}^f$ before it is updated by the upcoming observation to obtain the analysis state $\mathbf{x}^a$, which is then integrated by the model to compute the next forecast.

5.3.1 Ensembles selection

We present two different strategies to select the ensemble members from an available dictionary: (i) select the elements that are the "closest" to the forecast according to a certain distance, (ii) select the elements that describe at best the filtering error subspace [130], based on the so-called Orthogonal Matching Pursuit (OMP) algorithm. After selecting the new members, the mean of the ensemble is replaced by the forecast state in both approaches, so that only the ensemble anomaly is used in the EnOI algorithm. The incoming observations are not used in the selection, so that the data are not involved in the choice of the prior.

5.3.1.1 Distance-based similarity selection

We look for the dictionary elements that bear spatial similarities, or are the closest in some sense, to the forecast state according to a distance measure. The idea is that if an ocean state displays similar spatial features as the forecast state, it is also expected to carry information about the uncertainties around the forecast. One straightforward way to evaluate the distance between the forecast and the dictionary elements is to use the L2-norm, or L1-norm as illustrated in Figure 5.2. In our experiments, the assimilation results were quite close whether using L1-norm or L2-norm, and thus we only report here the results of the latter in the numerical experiments presented in Section 5.5.

Quantifying the similarity between two fields according to some norm may under-represent some localized ocean features in the overall basin-distance. We have also tried to involve correlations in our elements selection, but the strong environmental gradient in the Red Sea [196] dominated the correlations and made it difficult to distinguish the dictionary elements in this basin at the mesoscales.

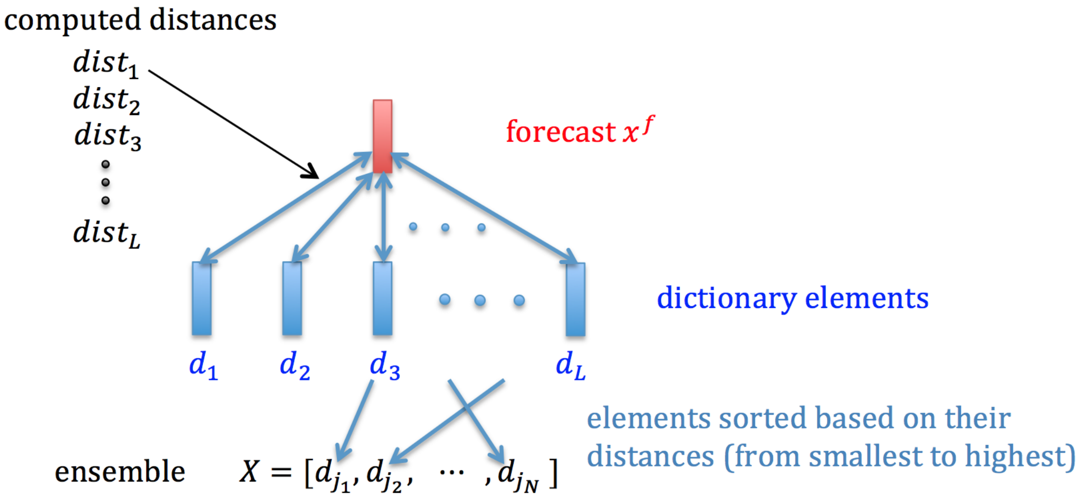


Figure 5.2: Illustration of an ensemble construction based on L2. Compute the L2-distances $(dist_1, dist_2, \cdots, dist_L)$ between the forecast x^f and the dictionary members $(d_1, d_2, \cdots, d_L)$ then select the first N members $(d_{j_1}, d_{j_2}, \cdots, d_{j_N})$ with the smallest distances to the forecast member to generate the ensemble X.

5.3.1.2 Error-subspace selection

The basic idea is to identify a subset of the dictionary elements that represents at best the forecast error subspace in which the Kalman filter update is applied [80, 109, 128]. Here we propose to use a Matching Pursuit (MP) method, an interactive greedy algorithm that finds the *best matching* projections of a high-dimensional signal onto the span of a (complete) dictionary [119]. By selecting the elements that are most correlated with the current residuals (see Figure 5.3 for schematic illustration and algorithm's description), MP attempts to approximately represent a signal using a sparse linear combination of the dictionary elements, called atoms, while minimizing the signal representational error in the dictionary. This is different than selecting the elements that are most correlated with the forecast state, and should lead to an ensemble with more spread describing the forecast state variability, assumingly representative of the filter error-subspace. In the Orthogonal Matching Pursuit (OMP) algorithm, the residual is always orthogonal to the span of the dictionary elements already selected. This can conceptually be implemented using a Gram-Schmidt scheme and results in convergence for a n-dimensional vector after at most p-iterations ($p \leq n$, being the sparsity level) [174]. Enforcing orthogonal elements helps to avoid selecting redundant elements and provides more ensemble spread [174].

5.3.2 Implementation of the ensemble selection strategies

All the selection methods share the same workflow and the difference appears only at the selection stage (3.1. of Algorithm 1). The generic form of the ensemble selection is detailed in Algorithm 2, while Table 5.1 outlines two implementations of Algorithm 2, one with the L2 selection method and the other one with the OMP.

5.4 Preliminary experimentation with Lorenz models

The adaptive EnOI schemes are first tested and compared with the standard EnOI and EnKF algorithms using Lorenz-63 [114] (hereafter L-63), and Lorenz-96 [115] (hereafter L-96) models, two popular prototypes for assessing new assimilation schemes. For each model, all EnOI-based schemes use the same dictionary, constructed from a collection of samples that came with the EnKF-Matlab software [154], and was gen-

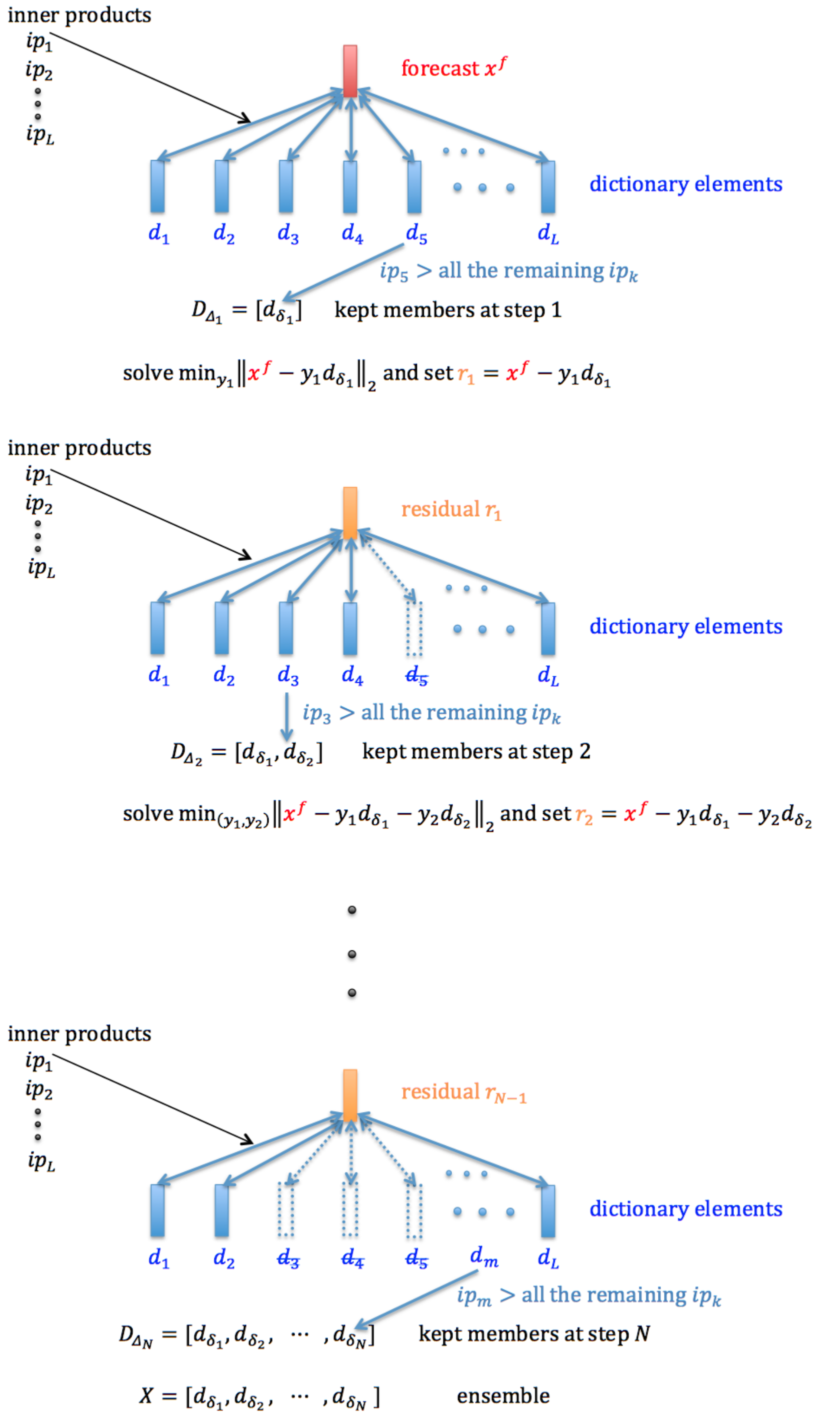

inner products
ip_1
ip_2
ip_L
forecast x^f
dictionary elements
d_1 d_2 d_3 d_4 d_5 d_L
$ip_5 >$ all the remaining ip_k
$D_{\Delta_1} = [d_{\delta_1}]$
kept members at step 1
solve $\min_{y_1} \|x^f - y_1 d_{\delta_1}\|_2$ and set $r_1 = x^f - y_1 d_{\delta_1}$
inner products
ip_1
ip_2
ip_L
residual r_1
dictionary elements
d_1 d_2 d_3 d_4 ~~d_5~~ d_L
$ip_3 >$ all the remaining ip_k
$D_{\Delta_2} = [d_{\delta_1}, d_{\delta_2}]$
kept members at step 2
solve $\min_{(y_1,y_2)} \|x^f - y_1 d_{\delta_1} - y_2 d_{\delta_2}\|_2$ and set $r_2 = x^f - y_1 d_{\delta_1} - y_2 d_{\delta_2}$
inner products
ip_1
ip_2
ip_L
residual r_{N-1}
dictionary elements
d_1 d_2 ~~d_3~~ ~~d_4~~ ~~d_5~~ d_m d_L
$ip_m >$ all the remaining ip_k
$D_{\Delta_N} = [d_{\delta_1}, d_{\delta_2}, \cdots, d_{\delta_N}]$
kept members at step N
$X = [d_{\delta_1}, d_{\delta_2}, \cdots, d_{\delta_N}]$
ensemble

Figure 5.3: Illustration of an ensemble construction based on OMP. Compute the inner products $(ip_1, ip_2, \cdots, ip_L)$ between the forecast x^f and the dictionary members $(d_1, d_2, \cdots, d_L)$ and keep the member having the highest ip value. Solve the least-square problem between the forecast and that member, and then compute the residual r_1. Compute the inner products between the residual r_1 and the remaining dictionary members and keep the member having the highest ip value. Solve the least-square problem between the forecast and the set containing that member and all the previous selected members. Compute the residual r_2. Repeat the process with the successive residuals until N members are selected.

Algorithm 1 Dictionary based schemes Data Assimilation Algorithm

```
0. Initialization: initial ensemble $\mathbf{X}^f$

1. Analysis step:
   Input: $\mathbf{X}^f$
   Output: $\mathbf{x}^a$

2. Forecast step:
   Input: $\mathbf{x}^a$
   Output: $\mathbf{x}^f$

3. Ensemble selection:
   3.1. Anomalies generation
          • Select an ensemble $\mathbf{X}$
          • Compute the anomalies $\boldsymbol{X}' = \mathbf{X} - \overline{\mathbf{x}}$ where $\overline{\mathbf{x}}$ is the mean of $\mathbf{X}$
   3.2. $\mathbf{X}^f$ generation
          Inputs: $\boldsymbol{X}'$ and $\mathbf{x}^f$
          Output: $\mathbf{X}^f = \boldsymbol{X}' + \mathbf{x}^f$

4. Goto 1
```

Algorithm 2 Generic ensemble design Algorithm

```
1. Inputs: a dictionary $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_L]$ of model outputs, the desired ensemble size $N$ (with $L \gg N$ and at least $L \geq N$), and the forecast $\mathbf{x}^f$ is iterated through the dictionary to apply the selection.

2. Sort the elements based on the metric ordering criteria: $\mathbf{d}_{j_1}, \mathbf{d}_{j_2}, \cdots, \mathbf{d}_{j_N}, \cdots, \mathbf{d}_{j_L}$

3. Form the ensemble of the first $N$ members $\mathbf{X} = [\mathbf{d}_{j_1}, \mathbf{d}_{j_2}, \cdots, \mathbf{d}_{j_N}]$ and use it to update the forecast with the incoming observations.
```

Table 5.1: Specific algorithms of the selection methods.

Selection method	**Algorithm**
L2-norm	2. For i from 1 to L compute $\gamma_i = \left\Vert \mathbf{x}^f - \mathbf{d}_i \right\Vert_2$ 3. Sort the γ_i in ascending order: $\gamma_{i_1}, \gamma_{i_2}, \cdots, \gamma_{i_L}$ with $\gamma_{i_1} \leq \gamma_{i_2} \leq \cdots \leq \gamma_{i_L}$ and assign $j_1 = \gamma_{i_1}, j_2 = \gamma_{i_2}, \cdots, j_L = \gamma_{i_L}$
OMP	2.1 Initialization: set $\mathbf{y}_0 = \mathbf{0}$, index set $\mathbf{\Delta}_0 = \emptyset$ and residual $\mathbf{r}_0 = \mathbf{x}^f$ 2.2 For t from 1 to N, • Find the index of the dictionary element having the highest inner product with the residual: set δ_t to one of the indexes j for which the maximum is reached, i.e. $\lvert\langle \mathbf{r}_{t-1}, \mathbf{d}_{\delta_t} \rangle\rvert = \max\limits_{j=1,\cdots,L} \lvert\langle \mathbf{r}_{t-1}, \mathbf{d}_j \rangle\rvert$ • Augment the index set: $\mathbf{\Delta}_t = \mathbf{\Delta}_{t-1} \cup \{\delta_t\}$ • Solve the least-square problem $\min\limits_{y} \left\Vert \mathbf{x}^f - \mathbf{D}_{\mathbf{\Delta}_t} \mathbf{y} \right\Vert_2$ and then choose $\mathbf{y}_t \in \arg\min\limits_{y} \left\Vert \mathbf{x}^f - \mathbf{D}_{\mathbf{\Delta}_t} \mathbf{y} \right\Vert_2$ • Calculate new residual $\mathbf{r}_t = \mathbf{x}^f - \mathbf{D}_{\mathbf{\Delta}_t} \mathbf{y}_t$ End for 3. Assign $j_1 = \delta_1, j_2 = \delta_2, \cdots, j_N = \delta_N$

erated by a long model run. We conduct twin-experiments where the trajectory of a reference run is taken as the "true" trajectory from which synthetic observations are generated by adding zero-mean Gaussian white noise with variance σ_y^2. The filters' performances are evaluated using the RMSE between the reference states and the filters' estimates averaged over all variables and over the whole assimilation period. We implement all filters using the covariance inflation [13]. We further apply a local analysis [84] in the L-96 experiments.

5.4.1 Numerical experiments with L-63

The L-63 model is a nonlinear dissipative dynamical system that mimics an atmospheric chaotic behavior [137]. It is governed by the following systems of differential equations

$$\begin{cases} \dfrac{dx}{dt} = \sigma(y - x), \\[2ex] \dfrac{dy}{dt} = (\rho - z)x - y, \\[2ex] \dfrac{dz}{dt} = xy - \beta z, \end{cases}$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$. The state variables x, y and z measure, respectively, the intensity of convective motion, the temperature difference between the ascending and descending currents, and the distortion of vertical temperature profile from linearity. The model is integrated using a fourth-order Runge-Kutta integration scheme, with a time step of 0.01 time units. After a spin-up period of roughly 20 days to remove any detrimental impact, the simulations are run for a period of five years in model time (i.e., 36500 model steps). We consider the case where all three variables are observed with $\sigma_y^2 = 2$. All schemes are tested with different values of inflation, ranging between 1 and 1.3, and compared, based on their minimum RMSEs. Our numerical experiments suggested that EnOI-based schemes do not require inflation, which was therefore not applied.

Figure 5.4 plots time series of the analysis RMSEs (upper subplot) and the forecast ensemble standard deviations (bottom subplot), as resulting from all schemes between assimilation steps 3000 and 3500, for illustration. Data are assimilated every 4 model steps and the ensemble size is set to 100. The RMSEs time series suggest a clear outperformance of the EnKF, followed by the AEnOI-L2 and the AEnOI-OMP. The reported RMSE values, averaged over the whole simulation period (0.172 for EnKF, 1.032 for AEnOI-L2, 1.119 for AEnOI-OMP, and 1.205 for EnOI), further confirm this and support our expectations about the capabilities of the AEnOI schemes in improving the EnOI performances, although they all fall behind EnKF. This is expected as the EnKF evolves the underlying state distribution by updating the ensemble members with the model dynamics. This however might be computationally demanding, since large ensembles are usually needed to properly describe the state statistics. For example, a one-year filtering run with 200 members, assimilating every 50 model steps (2.5 days), completed in 6.5538 s with the EnKF, 0.5103 s with the AEnOI-L2 and 2.9924 s with the AEnOI-OMP. Regarding the ensemble spread, the results suggest that EnKF, and to a lesser extent AEnOI-L2, exhibit the smallest spreads. EnOI and AEnOI-OMP however have larger spreads, but with different patterns. Indeed, EnOI has of course a constant spread, whereas AEnOI-OMP suggests a strongly variable spread over time. We further study the schemes' sensitivities to different ensemble sizes (N_e) and frequencies of assimilation in Figures 5.5 and 5.6, respectively. Overall, increasing the ensemble size reduces the RMSE values for all schemes, except for AEnOI-L2 which seems to perform better with relatively small ensembles (10, 20 and 40), in contrast with AEnOI-OMP. EnKF performances seem to level-off with increased ensembles size. As the ensemble size increases, the benefit from AEnOI-OMP becomes clearer and its performances approach those of AEnOI-L2 while both schemes remain better than EnOI. Similarly, assimilating the data more frequently further improves the results although the EnOI-based schemes seem less sensitive to the assimilation period than the EnKF.

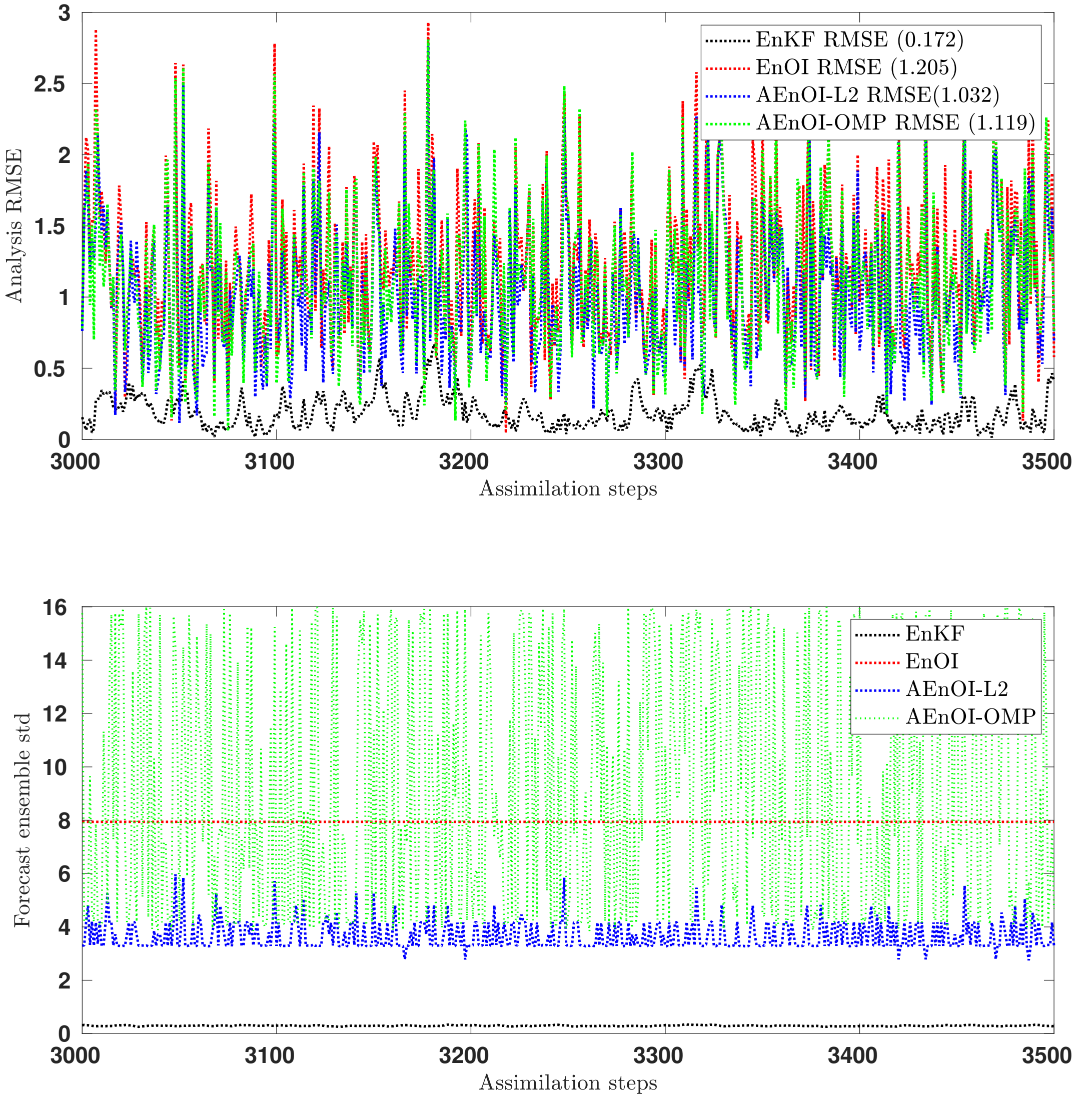


Figure 5.4: Mean analysis RMSE (top) and forecast ensemble standard deviation (bottom) for EnKF, EnOI, AEnOI-L2 and AEnOI-OMP schemes using the L-63 model. The assimilation experiments were performed using 100 members and assimilating all data every 4 model steps.

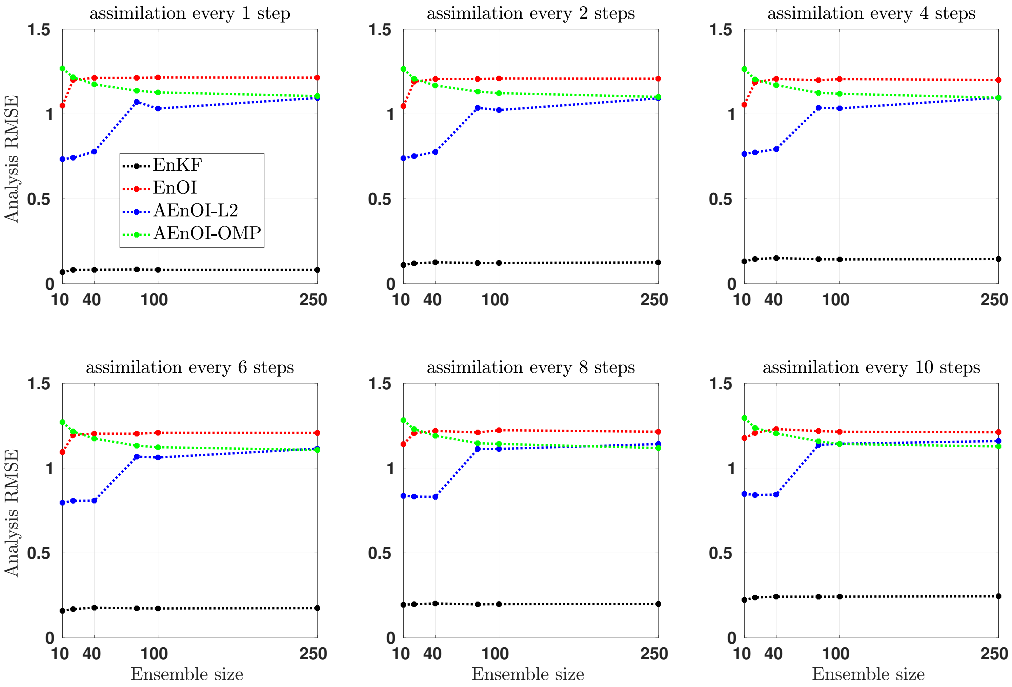


Figure 5.5: Sensitivity of the ensemble schemes to the ensemble size for a given assimilation window using L-63 model.

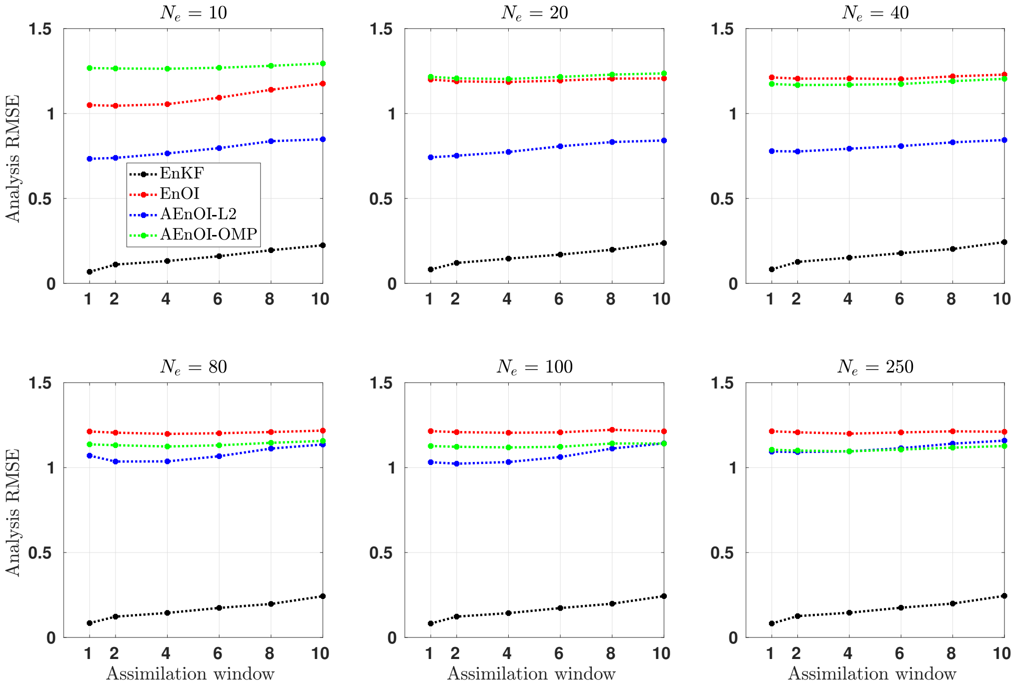


Figure 5.6: Sensitivity of the ensemble schemes to the assimilation period for a given ensemble size using L-63 model.

5.4.2 Numerical experiments with L-96

The L-96 model simulates the time evolution of an atmospheric quantity based on a set of differential equations:

$$\frac{dx_k}{dt} = (x_{k+1} - x_{k-2})x_{k-1} - x_k + F, \qquad k = 1, \cdots, K. \tag{5.5}$$

where x_k denotes the k^{th} element of the state $\mathbf{x}$. The nonlinear (quadratic) terms represent advection and the linear term simulates dissipation. In its most common form, the system dimension is $K = 40$ and the forcing term F is set to 8, a value for which the model exhibits a chaotic behavior. Boundary conditions are periodic (i.e; $x_{-1} = x_{39}$, $x_0 = x_{40}$ and $x_{41} = x_1$). The model is integrated using a fourth-order Runge-Kutta integration scheme, with a time step of 0.05 time units. After a spin-up period of roughly 20 days to remove any transient impact, the simulations are run for a period of five years in model time (i.e., 7300 model steps). We consider the case where all variables are observed with $\sigma_y^2 = 1$. We test the schemes using different values of inflation ranging between 1 and 1.3. We apply the standard local analysis approach by restricting the update of each grid point to only observations falling within some influence radius [153]. The localization support radii vary from 2 (strong localization) to 40 (weak localization) grid points. The schemes are then compared based on their minimum RMSEs over all possible combinations of inflation and localization values. Figure 5.7 gives an idea about the schemes' needs for inflation and localization by plotting the RMSEs as a function of the localization radius and inflation factor.

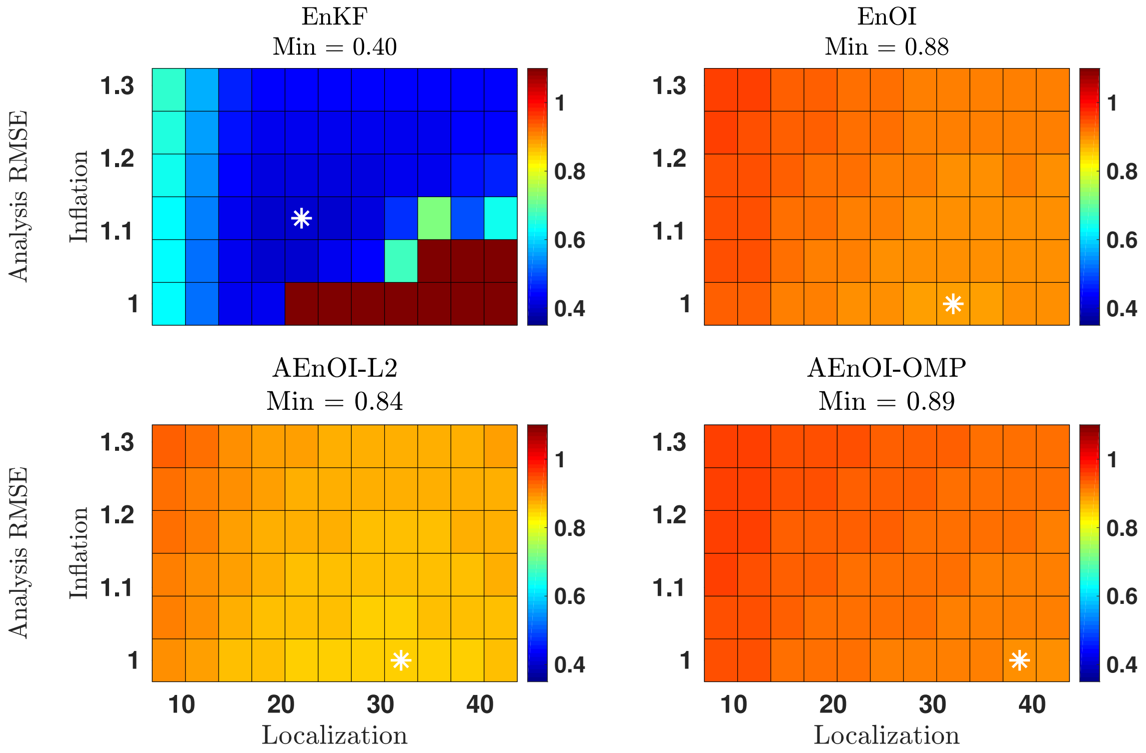


Figure 5.7: Time-averaged RMSE as a function of the localization radius (x axis) and inflation factor (y axis) using L-96 model. All filters were implemented with 40 members, and observations were assimilated every 4 model time steps. The minimum RMSEs are indicated by asterisks, and their associated values are given in the title.

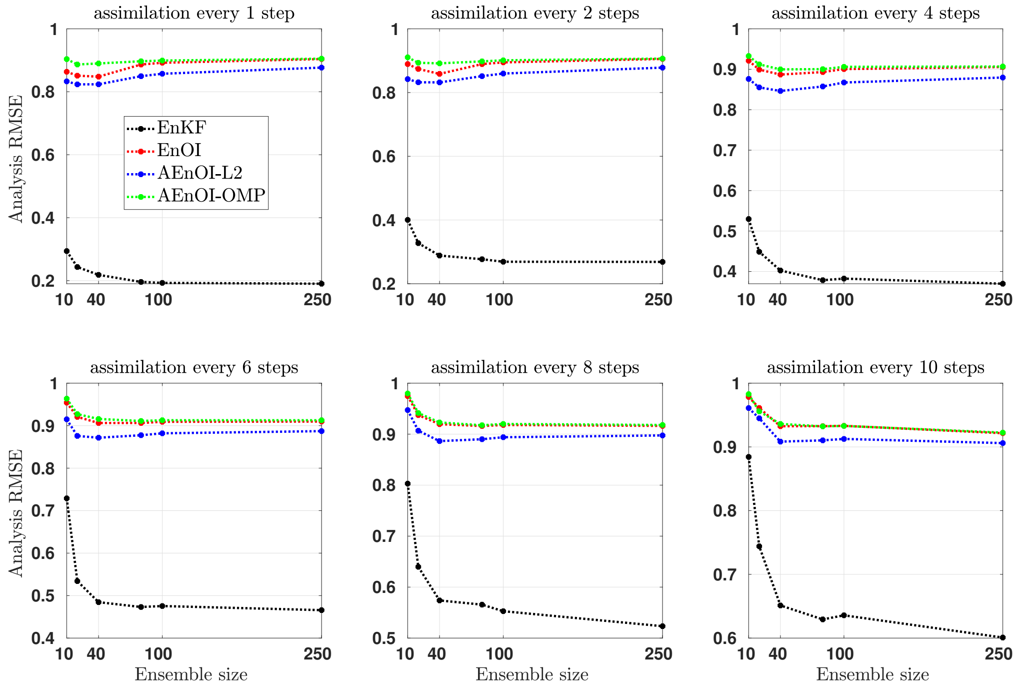


Figure 5.8: Sensitivity of the ensemble schemes to the ensemble size for a given assimilation window using L-96 model.

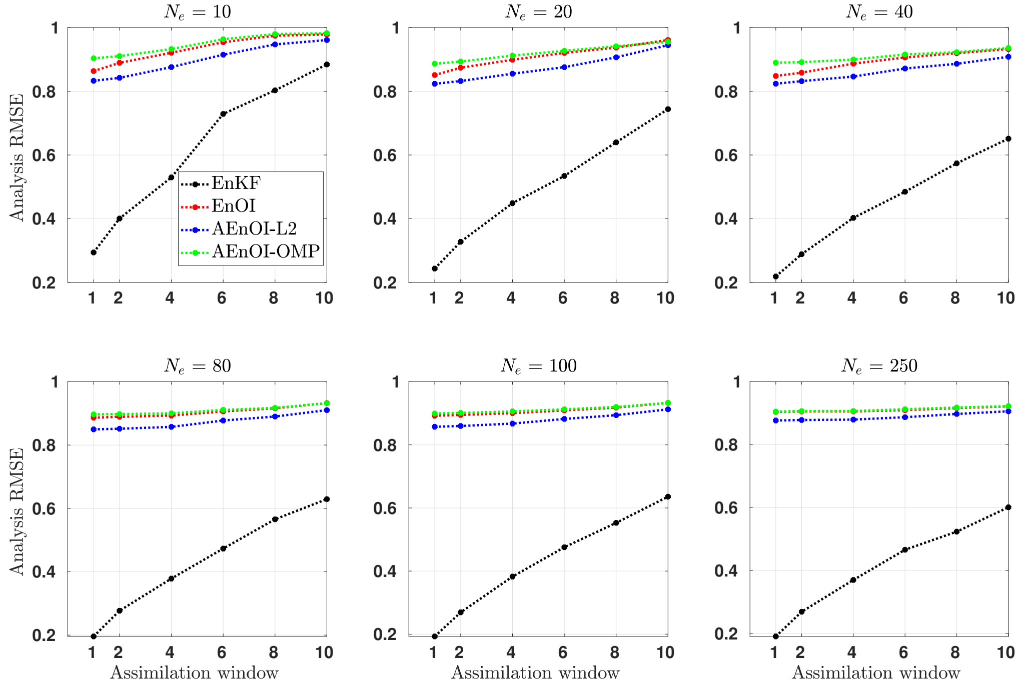


Figure 5.9: Sensitivity of the ensemble schemes to the length of the assimilation period for a given ensemble size using L-96 model.

5.5 Experimentation with an ocean general circulation model in the Red Sea

5.5.1 The ocean model

We employ the MITgcm as in Section 2.2. Additionally, the model uses a direct space time 3^{rd} order scheme for tracer advection, harmonic viscosity with coefficients of 30 m^2/s in the horizontal and 7×10^{-4} m^2/s in the vertical, implicit horizontal diffusion for both temperature and salinity, and the KPP scheme [101] for vertical mixing with a vertical diffusion coefficient of 10^{-5} m^2/s for both temperature and salinity. The open boundary conditions for temperature, salinity, and horizontal velocity are prescribed daily from the Global Ocean Reanalysis and Simulation data (GLORYS; [139]) available on a 1/12° horizontal grid. A sponge layer of 5 grid boxes with a relaxation period of 1-day is implemented for smooth incorporation of open ocean conditions through the eastern boundary. The normal velocities at the boundary are adjusted to match the volume flux of GLORYS, which is estimated from GLORYS SSH variations inside the model domain. The resulting inflow at the eastern boundary ensures consistency between the model and GLORYS basin-scale SSH. The model was spun-up for 31 years starting from 1979 to 2010 using the ECMWF reanalysis of atmospheric surface fluxes of radiation, momentum, freshwater sampled every 6-hour and available on a 75 km × 75 km grid [36]. The model has been extensively validated for the Red Sea by earlier studies (e.g. [67, 171, 189, 190, 193]). For comparison with the assimilation runs (as further discussed in the next section), the same model configuration was integrated forward for the year 2011 using 6-hourly ECMWF atmospheric fields available at 50 km × 50 km resolution. We refer to this model free-run experiment without assimilation as *Fexp*.

5.5.2 Experimental setup

Available observations are assimilated using the EAKF available in the DART-MITgcm package [76, 81] implemented for the Red Sea by [171]. All the experiments, in the present study, assimilate the data every 3 days, using a ∼300 km horizontal localization radius and a multiplicative inflation factor of 1.1, as suggested by [171]. We assimilated observations of SST data generated from a level-4 daily $0.25^\circ \times 0.25^\circ$ resolution product of [149] (which was prepared by blending SST measurements from in situ and advanced very high resolution radiometer infrared satellites), and along-track satellite level-3 merged altimeter filtered SLA corrected for dynamic atmospheric, ocean tide, and long wavelength errors, from Copernicus Marine Environment Monitoring Service (CMEMS) [126]. To compute the innovations between the SLA observations and the model SSH during assimilation, we add the model mean SSH to SLA observations prior to assimilation. Observations errors are assumed uncorrelated, and are prescribed with error variance of $(0.04\ \text{m})^2$ for SLA, and vary between $(0.1\ ^\circ\text{C})^2$ and $(0.6\ ^\circ\text{C})^2$ for SST in accordance with the interpolation errors specified in the level-4 gridded SST product of [149]. Four different assimilation experiments were conducted under the same conditions: EAKF with 50 members, and EnOI, AEnOI-L2, and AEnOI-OMP with 300 ensemble members. They differ only in terms of the underlying method to sample/select the ensemble from a long dictionary of MITgcm outputs simulated during the period 2002-2016. EAKF dynamically evolves the ocean ensemble. Its initial ensemble is generated by first selecting *Fexp* fields corresponding to ±15 days from January 1^{st} and then by adjusting the ensemble mean to the same initial state as *Fexp*. The EnOI uses a static ensemble of 300 members across all assimilation cycles (60 cycles in total) by selecting ocean states of 2002-2016 model hindcasts. AEnOI-L2 and AEnOI-OMP, dynamically select 300 members, based on the SST distance between the current ocean state and the dictionary elements. This choice is motivated by two factors: the SST exhibits a seasonal signal, and the ensemble members selection would have been computationally demanding (especially for OMP) if based on the full ocean state vector (10^7) and a dictionary with large number of elements. Note that we have also tried to select the members based on SSH, but OMP was selecting members from different seasons which affected the dynamical balances of the system. All the EnOI assimilation experiments are conducted over a 6-month period in 2011, starting from January 1^{st}, 2011 using the same initial condition as *Fexp*.

Unless stated, we analyze daily averaged ensemble mean forecasts as they result from the different assimilation experiments. Bias, correlations and RMSEs of the daily averaged forecasts of SST and SSH are computed with respect to the merged satellite level-3 observations of the GHRSST [49] and merged along track level-3 altimeter observations of SSH from CMEMS [126], respectively. The SST bias for a product (OSTIA, *Fexp*, ENOI or another scheme) is given by product-bias(SST) $= \sum_{t=January-June}(<\text{product(SST)}>_t \; - \; <\text{Obs(SST)}>_t)$, where Obs is the level-3 GHRSST data and <> refers to daily average of the ensemble mean over the spatial domain. Standard deviations (STDs) are computed using the following formula: STD $= \sqrt{\frac{1}{\text{time length}} \sum_{t=January-June}(<field>_t \; - \; \overline{<field>})^2}$, where $\overline{<field>}$ refers to time average, and field can be ensemble mean SSH, temperature (T), SST, salinity (S) or sea surface salinity (SSS). In order to demonstrate the relative performance of the assimilation system with respect to interpolated products, we employed level-4 SST and SSH products. The interpolated SST product is a high-resolution daily averaged level-4 SST product from OSTIA (Operational Sea Surface Temperature and Sea Ice Analysis) [40, 166], generated on a 0.054°(∼6 km) grid by combining SST data from various satellites and in situ observations using an Optimal Interpolation (OI) system. The interpolated SSH product is the multi-mission altimeter merged satellite level-4 gridded ADT provided by CMEMS (here after CMEMS-L4; [126]), which is also available daily at a resolution of $0.25° \times 0.25°$. The maximum reported

ADT mapping error (provided along with the CMEMS-L4 ADT product) during the analysis period 1^{st} January-30^{st} June, 2011 is estimated between 1.8 cm - 4 cm in the southern Red Sea and reaches up to 7 cm in the northern Red Sea. In order to use it in the present study, we adjust the CMEMS-L4 ADT by replacing its 15-year average by the model equivalent SSH climatology, similarly to the treatment of the level-3 SLA observations for assimilation.

5.5.3 Experimental results

Figure 5.10 displays spatial maps of SST forecast statistics (computed over the study period, i.e. 1^{st} January to 30^{th} June, 2011), compared to satellite level-3 SST observations, from different assimilation schemes, the free model experiment and the interpolated data product. The results indicate that the STD is large over the Gulf of Aden (reaching 2 °C) and small over the central parts of the Red Sea (below 1.2 °C). Outputs from the free model run captures this contrasting feature. However, it underestimates the SST STD over the whole domain, particularly in the northern and central parts of the Red Sea. Those underestimations of SST STD, as well as SST biases, are improved by assimilation. The improvements are more pronounced in the adaptive and ensemble optimal interpolation experiments relatively to the EAKF. However, the EnOI and AEnOI-OMP suggest increased SST biases in the Gulf of Aden. RMSEs and correlations also deteriorated, particularly in EnOI, with RMSEs increasing from 0.5 °C to 1 °C and correlations dropping from 0.95 to 0.8. Assimilation with the AEnOI-L2 strategy, on the other hand, yields SST improvements, with biases and RMSEs mostly within 0.5 °C and correlations above 0.8, all over the model domain, including Gulf of Aden and the Red Sea. AEnOI-L2 results are even better than the interpolated SST product, particularly in the northern and central Red Sea.

We further analyzed the time evolution of RMSEs for the daily averaged SST forecasts (Figure 5.11a) and for 3-day spaced SST analysis snapshots (Figure 5.11b)

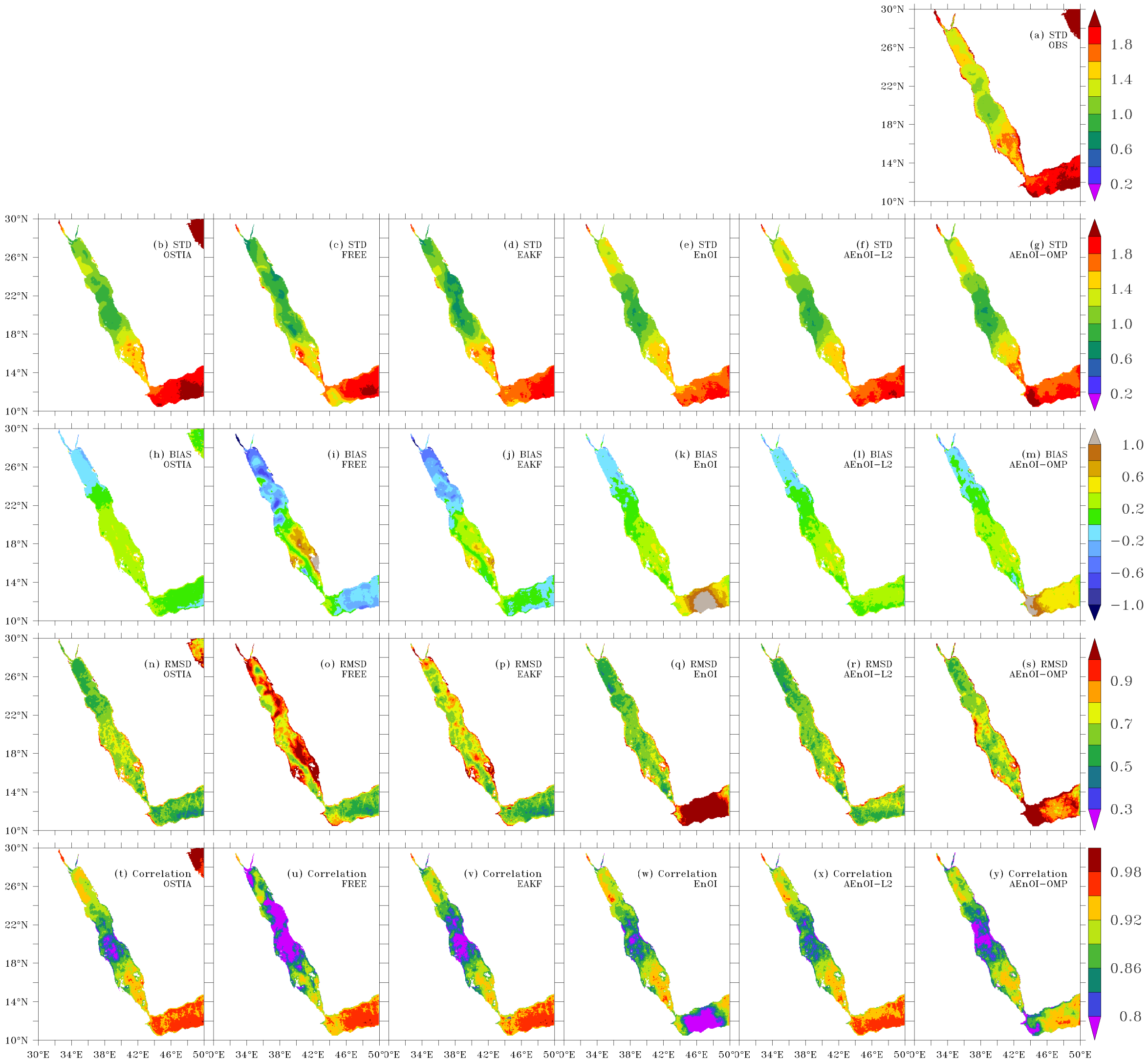


Figure 5.10: Spatial maps of SST STD in °C (b-g), Bias (h-m), RMSE (n-s) and correlations (t-y) for OSTIA (b), *Fexp* (c), EAKF (d), EnOI (e), AEnOI-L2 (f), and AEnOI-OMP (g). All the statistics are with respect to satellite level-3 SST observations. Panel "a" shows STD in the satellite level-3 SST. Negative values of bias indicate model cold biases and vice versa.

corresponding to the studied domain. As shown in Figure 5.11a, RMSEs of SST forecasts from all the model experiments and interpolation products do exhibit time dependence, with SST RMSEs dipping in February and peaking during June, except for EnOI and AEnOI-OMP which showed an additional peak (reaching 2 °C and 1.6 °C in EnOI and AEnOI-OMP, respectively) during the month of March. Interestingly, SST RMSEs resulting from EnOI and AEnOI-OMP are larger than those of *Fexp* until the last week of April. SST RMSEs are almost always less than those of *Fexp* when assimilating observations with EAKF, but they are further improved, even over the interpolated product, with the AEnOI-L2 strategy. Assimilation fits the observations better in AEnOI-L2 than in EAKF (Figure 5.11b), which seem to be due to improved SST spread (as discussed in the subsequent paragraphs using Figure 5.13), explaining the better SST forecasts in AEnOI-L2. The SST analyses of all three EnOIs experiments are indeed almost identical (Figure 5.11b). The failure to yield uniformly low SST forecast RMSEs and the occasional SST degradations in EnOI and AEnOI-OMP compared to the consistent improvements witnessed in AEnOI-L2 and EAKF may be attributed to the repercussion from comparatively larger dynamical imbalances in EnOI and AEnOI-OMP analyses (as discussed in the subsequent paragraphs) [110, 160].

Figures 5.11c and 5.11d, respectively, display the time evolution of SSH RMSEs for daily averaged forecasts and 3-day spaced analysis snapshots from different experiments and interpolated product. SSH RMSEs of *Fexp* exhibit noticeable fluctuations with largest values (reaching 14 cm) during January and smallest values (∼5 cm) during the end of May. Unlike the free model, SSH RMSEs in the interpolated product are stable with values around 5 cm. Assimilating observations with EAKF, EnOI or AEnOI-OMP also yields SSH RMSEs close to 5 cm, but they exhibit fluctuations in SSH RMSEs although not as large as *Fexp*. The fluctuations are reduced in AEnOI-L2, and the SSH RMSEs are generally lower than those of the interpolated product.

In order to spatially investigate the assimilation results for SSH we analyzed the region wise statistics. Since the altimeter coverage is too sparse over the model domain to yield spatial maps of statistics, we tabulated (Table 5.2) statistics for four different regions: Gulf of Aden (GoA; 30°E-50°E and 10°N-14°N), Southern Red Sea (SRS; 30°E-50°E and 14°N-19°N), Central Red Sea (CRS; 30°E-50°E and 19°N-23°N) and Northern Red Sea (NRS; 30°E-50°E and 23°N-28°N). *Fexp* underestimates the STD (up to 3 cm) and the mean (up to 8 cm) of SSH. The underestimations of the mean are largest in the NRS (by 160%) and the largest underestimations of the STD are in the SRS (by 27%). SSH RMSEs (9-11 cm) and correlations (0.4-0.86) are also poor in *Fexp*, particularly in the GoA and the NRS. The interpolated SSH product also underestimates the mean, but provides robust estimates of the STD, with low RMSEs (5-6cm) and high correlations (0.94-0.98) throughout the domain. Assimilation improves the SSH mean and STDs considerably throughout the domain, even better than (or on par with) the interpolated data product. SSH RMSEs (5-7 cm) and correlations (0.54-0.92) are also improved compared to *Fexp*, and still less than the interpolated product. Interestingly, AEnOI-OMP (EAKF) improvements are less pronounced than those resulting from the standard EnOI, which is probably related to the SSH spread of the background ensemble, as further discussed in the subsequent paragraphs. The differences between EnOI and AEnOI-L2 are not so large except for the GoA, in which AEnOI-L2 yields better results than the rest of the assimilation schemes.

We also examine the estimated ocean state in the subsurface to assess the impact of the assimilation strategies in these sparsely observed layers. The ocean state in the subsurface layers is noisy in EnOI (Figures 5.12e and 5.12f) compared to *Fexp* (Figures 5.12a and 5.12b) and to EAKF (Figures 5.12c and 5.12d), consistent with the results of [171], in which the noise in the subsurface was attributed to pronounced dynamical imbalances in the analysis. While the ocean state becomes noisier in AEnOI-OMP

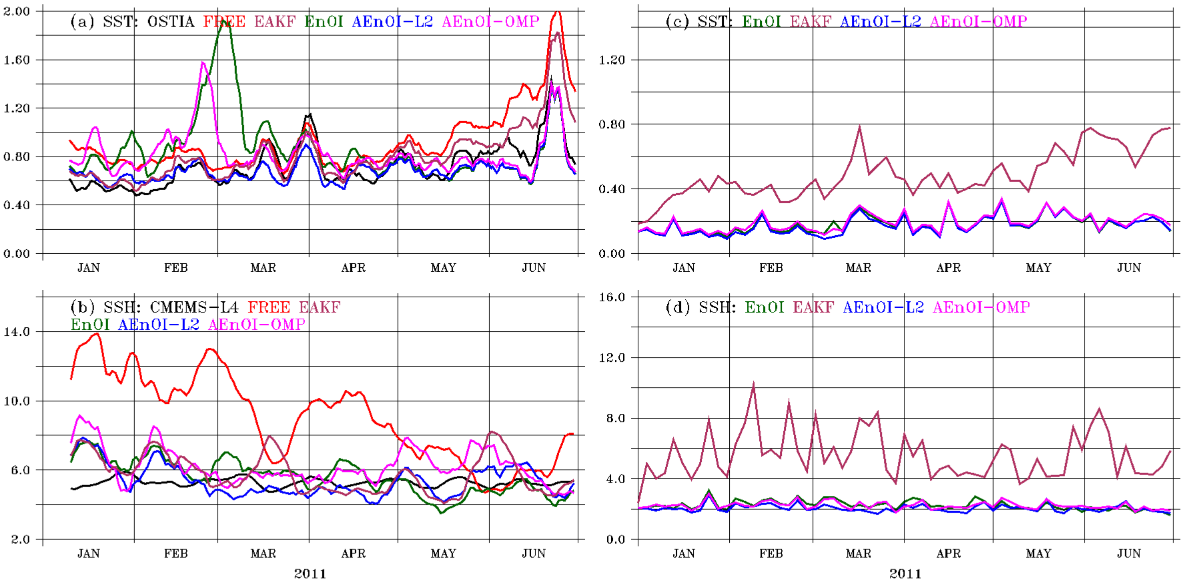


Figure 5.11: Time series of root-mean-square-error (RMSE) for daily averaged forecasts of (a) SST (b) SSH from *Fexp* (red), EAKF (maroon), EnOI (green), AEnOI-L2 (blue), AEnOI-OMP (pink), and level-4 gridded products (OSTIA for SST and CMEMS-L4 for SSH; black). RMSE is computed by collocating the daily averaged model forecasts onto satellite along-track level-3 SST and SSH observations. 10-day smoothing is applied to better emphasize the differences between the assimilation results. Units are in "°C" and "cm" for SST and SSH, respectively. Panels (c) and (d) are similar to (a) and (b) except that the RMSEs are computed for 3-day spaced analyses (snapshots after assimilation) without smoothing.

Table 5.2: Region wise SSH statistics for CMEMS-L4 interpolated product, *Fexp*, EAKF, EnOI, AEnOI-L2, and AEnOI-OMP. Statistics are shown for four different regions, Gulf of Aden (GoA; 30°E-50°E and 10°N-14°N), Southern Red Sea (SRS; 30°E-50°E and 14°N-19°N), Central Red Sea (CRS; 30°E-50°E and 19°N-23°N) and Northern Red Sea (NRS; 30°E-50°E and 23°N-28°N). Units for mean, STD and RMSE are in cm. The assimilation experiment yielding best results for a region is highlighted with bold fonts.

	GoA				**SRS**				**CRS**				**NRS**			
	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr
Observation	12	5			12	11			5	13			-5	13		
CMEMS-L4	7	5	5	0.94	7	11	5	0.98	1	12	5	0.98	-10	13	6	0.98
Fexp	5	5	10	0.40	6	8	9	0.84	-1	10	9	0.86	-13	10	11	0.82
EAKF	13	6	**5**	0.57	14	11	6	0.88	8	12	7	0.86	-4	11	7	0.85
EnOI	13	8	7	0.56	12	11	5	0.91	**6**	13	6	**0.90**	-6	12	6	**0.89**
AEnOI-L2	11	7	6	**0.61**	12	11	5	**0.92**	7	13	6	0.89	-4	12	6	0.88
AEnOI-OMP	10	8	7	0.54	12	11	5	0.88	7	13	7	0.88	-4	11	7	0.85

(Figures 5.12i and 5.12j), AEnOI-L2 (Figures 5.12g and 5.12h) reduces this noise and yields more organized subsurface structures. For instance, EnOI simulates abrupt jumps in the 22 °C isotherm in the months of March, April, and also in May (Figure 5.12k) at (38°E, 22°N), and these are more frequent and larger in AEnOI-OMP. Such abrupt jumps do not appear in the results of AEnOI-L2, indicating a more stable solution. Dynamical imbalances (noise) may result from inappropriate analysis update due to spurious spread, and correlations in the background ensemble. These aspects are further discussed in the next paragraphs.

Figure 5.13 plots the spatial distribution of the ensemble spread on 1-May-2011 from the different filtering schemes. The ensemble spreads of SSH, SST and subsurface temperature are considerably larger in all the EnOIs assimilation experiments compared to those of EAKF. This is because the spread introduced in the ensemble of initial conditions in EAKF fades out after few analysis cycles, and because the

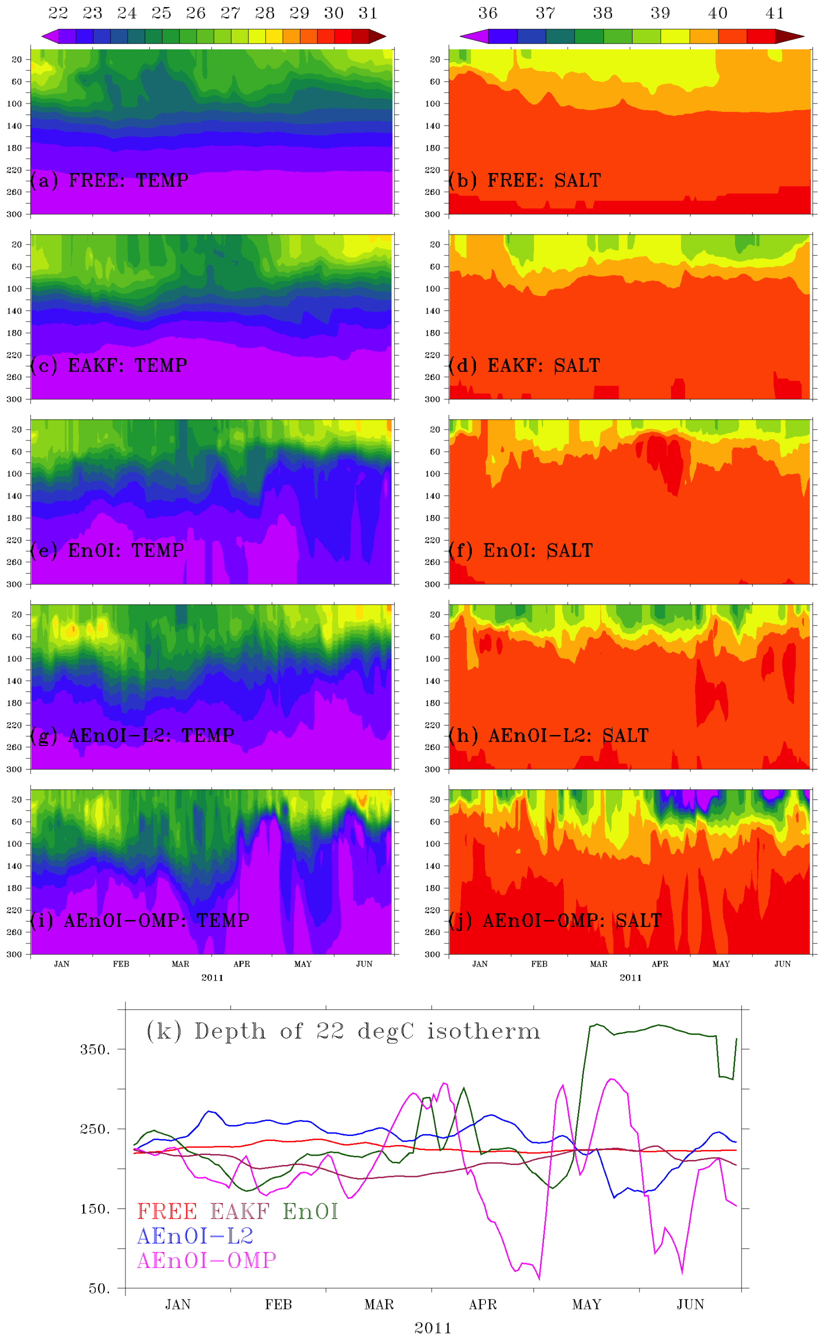


Figure 5.12: Depth-Time evolution of temperature (°C; a, c ,e, g, and i) and salinity (psu; b, d ,f, h, and j) and depth of 22 °C isotherm (meters; k) at (38°E, 22°N) as resulted from *Fexp*, EAKF, EnOI, AEnOI-L2 , and AEnOI-OMP.

EnOIs strategies do not lose spread as they select members, for the Adaptive ensemble Optimal Interpolations (AEnOIs) schemes, from model hindcasts after each analysis cycle. The spreads of SSH, SST and upper layer temperatures resulting from EnOI, AEnOI-L2 and AEnOI-OMP are significantly different. AEnOI-L2 selects the ensemble members from a broader range of months based on their closeness to the forecast SST, which seem to result here in a small ensemble spread (Figure 5.13g and 5.13k). AEnOI-OMP selects the ensemble members based on the correlations of the dictionary elements with residuals of the forecast state in the ensemble subspace (weaker the correlation better are the chances for selection). As a result, the selected members are not necessarily correlated/close to the forecasted SST, and may thus exhibit larger ensemble spread (Figures 5.13d, 5.13h and 5.13l). Large ensemble spreads may cause a data overfit, and amplify the noise in the filter updates, particularly in the data sparse regions [133, 159]. This may explain the more (less) abrupt jumps in the 22 °C isotherm in AEnOI-OMP (AEnOI-L2) compared to EnOI.

One of the key assumptions of an EnKF framework lies on Gaussian forecast errors, based on which the members are updated with the observations using the Kalman linear analysis step [82]. In the ensemble optimal interpolation schemes, the forecast error is estimated based on the anomalies of the selected ensemble. We assess the relevance of the Gaussian assumption in our setup by analyzing the histogram of SST ensembles at three locations in the northern, central and southern Red Sea on 1-May-2011, as shown in Figure 5.14. At all these locations, the prior distributions in EnOI and AEnOI-L2 are clearly more Gaussian than that of the AEnOI-OMP. The OMP scheme provides a more scattered ensemble that is far from a Gaussian distribution, and this may limit the relevance of the Kalman-based update step.

We also analyzed the SST correlation range at three different locations in the northern, central and southern Red Sea for the different EnOI schemes (Figure 5.15). At all locations, the SST correlation range for AEnOI-L2 is narrower and less noisy

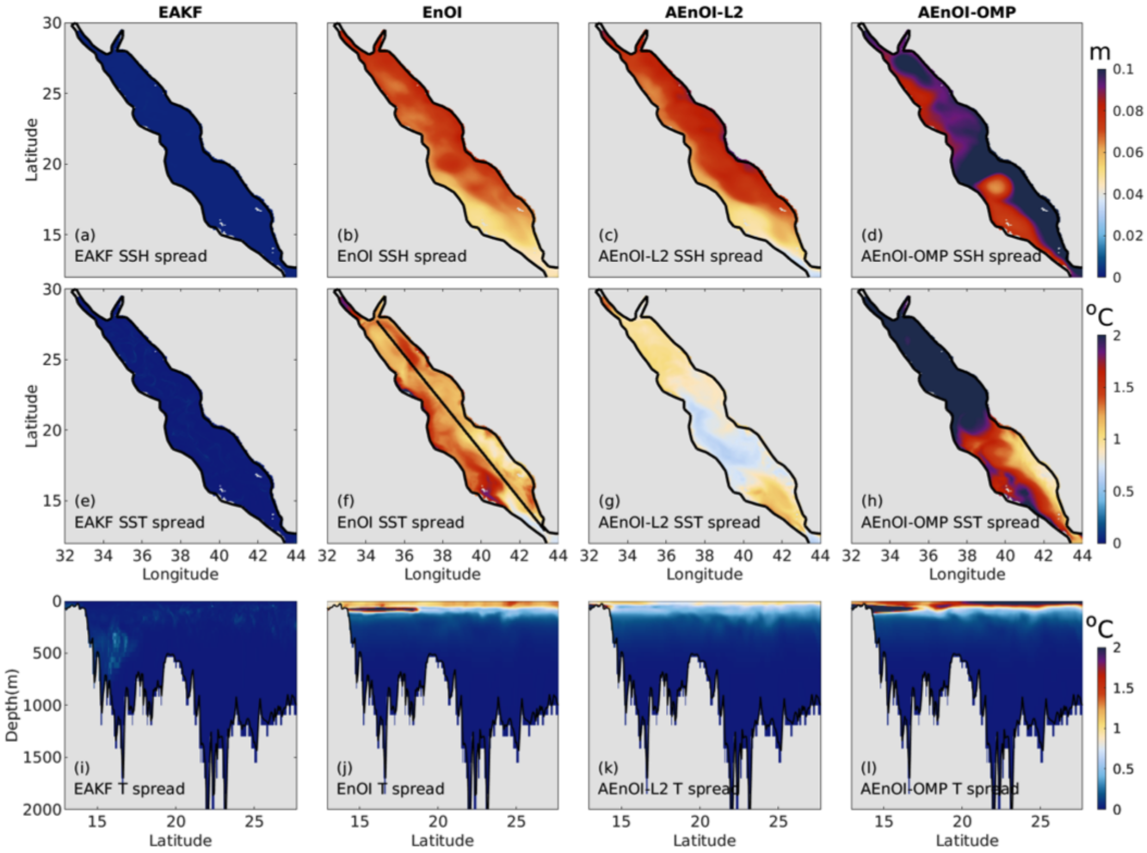


Figure 5.13: Horizontal and vertical distributions of ensemble spread (a - d) of SSH, (e - h) SST, and (i - l) temperature on 1-May-2011 as they result from EAKF, EnOI, AEnOI-L2, and AEnOI-OMP.

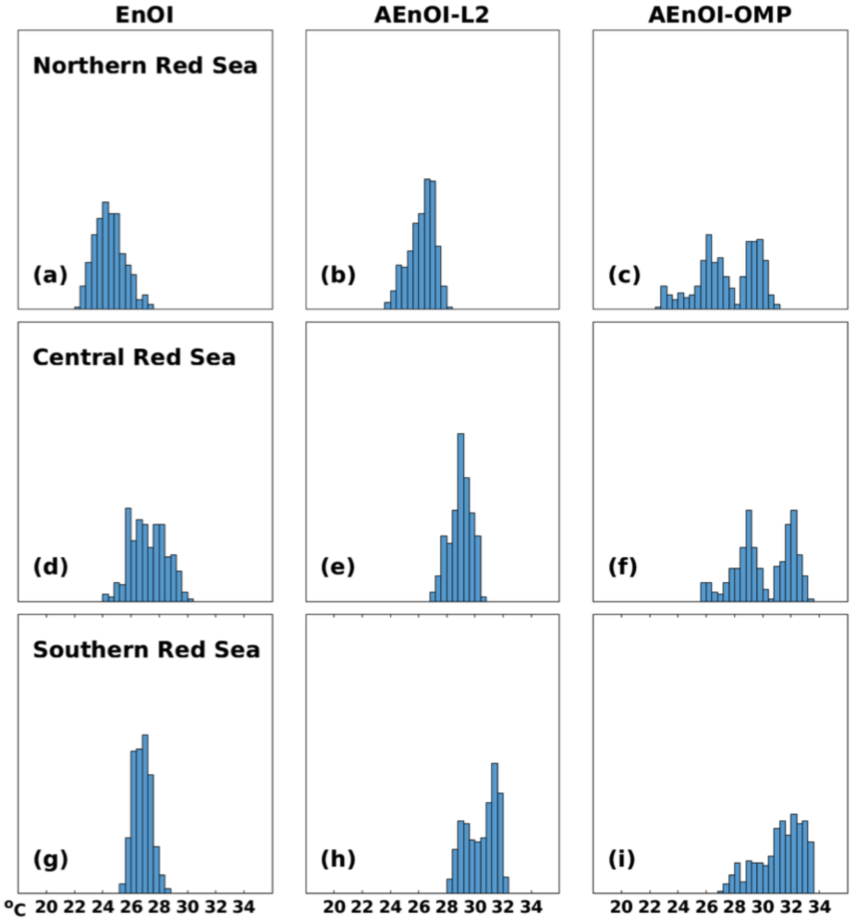


Figure 5.14: The histograms (prior) in experiments using EnOI (1^{st} column), AEnOI-L2 (2^{nd} column) and AEnOI-OMP (3^{rd} column) assimilation experiments at three selected locations (indicated in Figure 5.15) in the northern, central and southern basins of the Red Sea, as the 1^{st}, 2^{nd} and 3^{rd} rows, respectively.

than those of EnOI, and AEnOI-OMP, suggesting less spurious long-range correlations. This means that AEnOI-L2 could be configured with a larger localization radius, which may subsequently result in more dynamically consistent ocean state estimates [54]. Given that AEnOI-L2 only forecasts the analysis state, this would enable using larger ensembles to rely even less on localization [170], without significantly increasing the computational cost. In our specific system, one MITgcm model run requires 4.8 core hours for a 3-day simulation and a DART-filter update requires 111 (25.3) core hours for 300 (50) members. Therefore, one EAKF assimilation step with 300 (50) members consumes 1551 (265.3) core hours. The adaptive schemes involve a single model run for forecasting, and the selection step of its 300-member ensemble requires 21.37 and 20.77 core hours for the AEnOI-L2 and AEnOI-OMP, respectively, followed by a filter update. This amounts to an approximate computational cost of 137 core hours for each of the adaptive schemes and translates to more than a factor 10 (2) cost saving compared to the EAKF.

5.6 Conclusions

The Red Sea is characterized by a marked seasonal variability and strong mesoscales activity. In order to account for these variations at different time scales with reasonable computational burden, we proposed new cost-effective AEnOI schemes for assimilating multivariate data sets of the Red Sea based on DART and the MITgcm.

The AEnOI schemes select the ensemble members from a dictionary that is as complete as possible and describes the underlying system variability. The members selection is based on their similarity to, according to a certain criteria, or to their representativeness of the current forecast state, which represents the best available information at the time of the incoming observations. Two approaches for selecting the ensemble members were proposed: the first is based on the L2-distance between the forecast and the dictionary elements, and the second uses an Orthogonal Matching

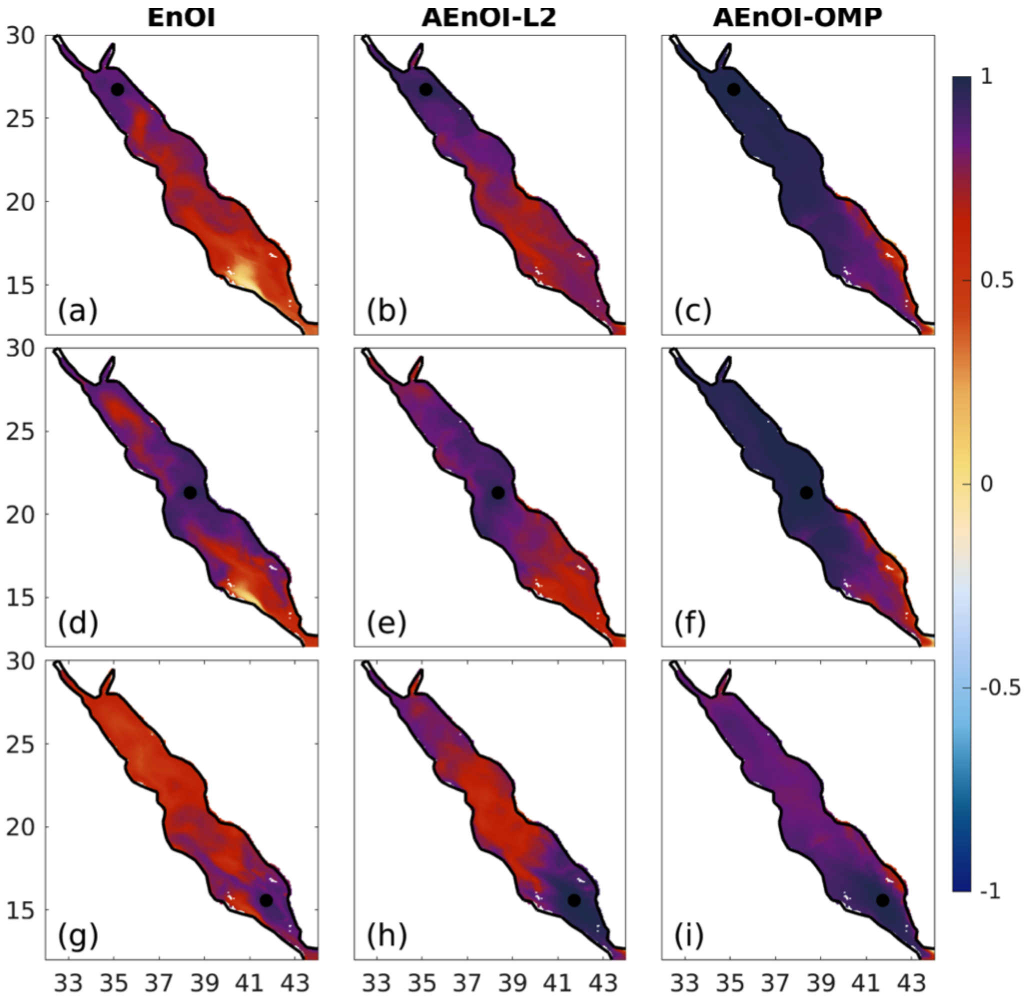


Figure 5.15: Sampled correlations for SST as computed from the assimilation experiments using EnOI (1^{st} column), AEnOI-L2 (2^{nd} column), and AEnOI-OMP (3^{rd} column) schemes at three selected locations in the northern (1^{st} row), central (2^{nd} row), and southern (3^{rd} row) basins of the Red Sea, respectively, before applying localization. The black dot in each panel indicates the selected location.

Pursuit (OMP) algorithm to identify the error-subspace of the forecast state. In term of computational efficiency, EnOI has of course an advantage since the selection process is applied offline and only once, before the start of the assimilation experiments. The AEnOI schemes enable however for adaptive selection of the ensemble members, which could account for instance for inter-seasonal and mesoscale variability.

The AEnOI schemes were first implemented and validated with the Lorenz-63 and the Lorenz-96 models, compared against the EnKF and the standard EnOI. While the EnKF yields the best results, eventually at the expense of applying auxiliary techniques such as inflation and localization, and higher computational cost, AEnOIs generally yield more accurate estimates than the standard EnOI, in terms of RMSE. They are further, particularly AEnOI-L2, computationally very efficient and may provide an alternative to the EnKF in the challenging scenario of small ensembles and large state space.

Within the DART-MITgcm Red Sea assimilation system, the AEnOI schemes operate on a dictionary of ocean realizations describing the multiscale temporal and spatial variability of the basin. Different aspects of the assimilation system have been assessed; including SST and SSH biases, standard deviations, correlations, and root-mean-square errors. AEnOI-L2 yields substantial improvements in certain regions of the Red Sea, whereas the AEnOI-OMP and the EnOI lead, in general, to more or less comparable assimilation results in our particular domain.

The AEnOI schemes, AEnOI-L2 more precisely, provided competitive performances to the computationally much demanding ensemble (adjustment) Kalman filter, especially in situations when the model forward integration is computationally demanding. I will discuss in the next Chapter Hybrid schemes in which a new ensemble member will be selected from a dictionary, eventually regionally, based on the statistics of an (small) evolving ensemble. The resulting ensemble will combine the spread benefit of the EnOI scheme and will constrain it by that of the evolving ensemble that accounts for the error-of-the-day. We are also planning to implement these schemes within a stochastic EnKF framework based on the scheme proposed by [81].

Chapter 6

A Hybrid Ensemble Adjustment Kalman Filter for High-resolution Data Assimilation System in the Red Sea: Implementation and Evaluation

This Chapter corresponds to the paper "A Hybrid Ensemble Adjustment Kalman Filter based High-resolution Data Assimilation System in the Red Sea: Implementation and Evaluation" published in the *Quarterly Journal of the Royal Meteorological Society.*

6.1 Introduction

Ocean forecast models are not perfect owing to uncertainties in their internal physics and inputs, such as initial and boundary conditions, and atmospheric forcing [46, 77]. In data assimilation, these uncertainties are generally accounted for through the so-called background error covariance matrix (BECM), which spreads the observations information to all ocean model variables [77]. While the BECM varies in time (flow dependent) in the ensemble-based Kalman filters (e.g. Singular Evolutive Interpolation Filter-SEIK, Local Ensemble Transform Kalman Filter-LETKF, EAKF), other popular assimilation methods such as the variational and EnOI use static BECMs generated via empirical relations and/or climatological ensembles [46, 77]. In these methods, however, the BECM may not account for the "error-of-the-day", important to obtain reliable ocean state estimates (e.g. [77, 161, 165]).

Because of their high computational requirements in realistic applications, EnKFs are often implemented with limited ensemble size. The EnKFs may also suffer from not accounting for the sources of uncertainties while evolving the ensemble with ocean model due to technical difficulties. The EnKF's BECMs could be therefore severely rank-deficient with noisy ensemble-correlations and a systematic small variance or loss of ensemble spread, see the reviews of [19, 77]. This may greatly limit the EnKF's ability to fit the assimilated observations and produce meaningful ocean state estimates (e.g. [46, 77]). Various auxiliary techniques have been proposed to mitigate the impact of these limitations. Localization, in which long-range correlations are tapered, is a straightforward and efficient way to eliminate spurious corelations and increase the BECMs rank [60, 74, 84]. Covariance inflation, in which covariance of an EnKF forecast or analysis is inflated by some positive factor at each assimilation cycles, is another approach to compensate for the systematic loss of ensemble spread [12, 80, 86]. However, the ad hoc nature of these fixes were shown to degrade the dynamical balance of the filter analysis and to increase the forecast errors in sparsely observed regions (e.g. [160, 161]). The Hybrid ensemble scheme, in which BECMs are estimated as linear combinations of the time-varying ensembles generated by an EnKF and climatological (static) ensembles covariances [72, 113, 165, 183], was proposed as a potential approach to mitigate the aforementioned issues [33, 65, 140, 176, 182]. This increases the EnKF BECMs rank and spread and enforces smoothness, which was found to be particularly beneficial when the filter is implemented with small ensembles [165]. This therefore enables implementing the filter with small ensembles, which would drastically reduce the computational load.

Various Hybrid implementations were successfully implemented for data assimilation into ocean general circulation models [21, 33, 102, 140]. Here, we followed a similar approach as [183] and derived a practical implementation of a Hybrid-EAKF system for the Red Sea with a mesoscale resolving (4km-resolution) MITgcm [122], using DART [8]. The goal is to enhance the EAKF performances, while implementing the system with small flow dependent ensembles to reduce the computational load.

The ocean forecasts resulting from the Red Sea EnOI-based data assimilation system implemented by [170] showed large dynamical imbalances in the subsurface, even after employing monthly-varying climatological ensembles to estimate the BECMs. In a follow up work, [161] demonstrated that by accounting for uncertainties in initial conditions and atmospheric forcing and internal physics the flow dependent EAKF provide dynamically consistent and improved forecasts throughout the ocean column.

The proposed Hybrid-EAKF combines the EAKF system of [161] and EnOI system of [171], and aims at further improving the state estimates of the Red Sea. The remainder of the Chapter is organized as follows. Section 6.2 describes the Hybrid-EAKF. Section 6.3 presents an overview of the numerical experiments setup, including the description of the Ocean model and assimilated observations, and the independent observations used for evaluating the assimilation system solution. Section 6.4 discusses the assimilation results in terms of their forecast statistics, and of reproducing the basin mesoscale eddy features. It further analyzes in details the dynamical balances of the Hybrid-EAKF solutions. A discussion on the computational load of the Hybrid-EAKF system is provided in Section 6.4.3. Summary and conclusions are provided in Section 6.5.

6.2 The Hybrid-EAKF

Data assimilation is the process by which observations are used to update models forecasts to compute the best possible estimate of the state of the ocean [46]. The EAKF is a sequential assimilation scheme that operates as a series of assimilation cycles, each involving a model forecast followed by a filter update.

Let $\mathbf{X}^f$ be an ensemble of N model forecasts and $\mathbf{X}^a$ the analysis ensemble obtained by applying the EAKF. The Hybrid-EAKF has the same algorithm as EAKF, except for the use of a Hybrid forecast error covariance $\mathbf{P}^{f,H}$, expressed as a linear combination of a flow-dependent forecast covariance and a static background covariance [72],

$$\mathbf{P}^{f,H} = (1-\alpha)\mathbf{P}^f + \alpha\mathbf{B}, \text{ with } 0 \leq \alpha \leq 1. \tag{6.1}$$

$\mathbf{P}^f$ is the flow-dependent covariance, computed from the dynamic propagation of the forecast ensemble $\mathbf{X}^f$ with the EAKF, and $\mathbf{B}$, the static background covariance matrix, estimated from a climatological ensemble, for example. The resampling of the analysis members is then performed as in (2.15) (with $\mathbf{x}$ for the state space or $\mathbf{z}$ for the joint state-observation space), but using $\overline{\mathbf{x}}^{a,H}$ instead of $\overline{\mathbf{x}}^a$.

6.2.1 Practical implementation within DART

We implemented the Hybrid-EAKF in DART by calling separately two EAKF update steps in DART, one to update the forecast anomalies based on the flow-dependent forecast covariance and another to update the forecast mean using the Hybrid covariance. Combining the results of the two filters yields a Hybrid analysis ensemble with the desired mean and covariance as follows:

i. The flow-dependent forecast ensemble $\mathbf{X}^f$ is first updated using the EAKF analysis step. This gives an analysis ensemble satisfying equation (2.15), with an analysis mean and covariance respectively given by equations (2.13) and (2.14) (with $\mathbf{x}$ and $\mathbf{P}$ for the state space or $\mathbf{z}$ and $\mathbf{\Sigma}$ for the joint state-observation space). Rewriting (2.15) as

$$\big(\mathbf{x}^{a,i} - \overline{\mathbf{x}}^a\big) = \mathbf{A}(\mathbf{x}^{f,i} - \overline{\mathbf{x}}^f), \qquad i = 1, \ldots, N, \tag{6.2}$$

suggests that the updated anomalies are simply the updated members from which the analysis mean is removed.

ii. In order to update the forecast state, a prior ensemble $\mathbf{X}^H$ is constructed and supplied as input to DART. $\mathbf{X}^H$ is constructed such that by calling a standard EAKF update in DART, the resulting analysis state satisfies equation (2.14). We can show that such an ensemble is expressed as $\mathbf{X}^H = \left[K_d\boldsymbol{X}', K_s\boldsymbol{X}'^s\right] + \overline{\mathbf{x}}^f$, with

$$K_d = \sqrt{\frac{(1-\alpha)(N+N_s-1)}{N-1}} \text{ and } K_s = \sqrt{\frac{\alpha(N+N_s-1)}{N_s-1}}.$$

$\boldsymbol{X}'^s$ is a static ensemble perturbation matrix defined as $\boldsymbol{X}'^s = [\mathbf{x}^{s,1} - \overline{\mathbf{x}}^s, \mathbf{x}^{s,2} - \overline{\mathbf{x}}^s, \cdots, \mathbf{x}^{s,N} - \overline{\mathbf{x}}^s]$ with $\{\mathbf{x}^{s,i}\}_{i=1,\ldots,N_s}$ an ensemble of N_s static members of mean $\overline{\mathbf{x}}^s$ and static covariance $\mathbf{B} = \frac{1}{N_s-1}\left(\boldsymbol{X}'^s\boldsymbol{X}'^{sT}\right)$. One can verify that the $(N+N_s)$ ensemble $\mathbf{X}^H$ has a mean $\overline{\mathbf{x}}^f$ and a covariance $\mathbf{P}^{f,H}$. The sample covariance of $\mathbf{X}^H$ is indeed given by $\frac{1}{N+N_s-1}\left(\boldsymbol{X}'^{\boldsymbol{H}}\boldsymbol{X}'^{\boldsymbol{H}T}\right)$, where $\boldsymbol{X}'^{\boldsymbol{H}}$ is the corresponding perturbation matrix (i.e., $\boldsymbol{X}'^{\boldsymbol{H}} = \left[K_d\boldsymbol{X}', K_s\boldsymbol{X}'^s\right]$). Using the expressions of K_d and K_s, the covariance expression becomes $\frac{(1-\alpha)}{N-1}\boldsymbol{X}'\boldsymbol{X}'^T + \frac{\alpha}{N_s-1}\boldsymbol{X}'^s\boldsymbol{X}'^{sT}$ which is equal to $(1-\alpha)\mathbf{P}^f + \alpha\mathbf{B}$ and thus to $\mathbf{P}^{f,H}$. Finally, since $\mathbf{X}^H$ has a mean $\overline{\mathbf{x}}^f$ and a covariance $\mathbf{P}^{f,H}$, its EAKF update yields an analysis state that matches $\overline{\mathbf{x}}^{a,H}$ given by equation (2.14) with $\overline{\mathbf{z}}_k^a$ and the covariances replaced by $\overline{\mathbf{x}}^{a,H}$ and the Hybrid covariances, respectively.

iii. The Hybrid analysis ensemble is then obtained by adding the anomalies resulting from step (i) to the Hybrid analysis mean state from step (ii) as

$$\mathbf{x}^{a,H,i} = \overline{\mathbf{x}}^{a,H} + \mathbf{A}(\mathbf{x}^{a,i} - \overline{\mathbf{x}}^a), \qquad i = 1, \ldots, N, \tag{6.3}$$

The Hybrid formulation reduces to the EAKF when $\alpha = 0$.

The above algorithm is similar to the Hybrid-ETKF of [183], but we further proposed a practical implementation that makes use of the existing update code within DART.

6.3 Assimilation experiments and results

6.3.1 Model setup

MITgcm is configured as in Section 5.5.1. Figure 6.1 shows the model domain, the bathymetry and the observational coverage. *Fexp* atmospheric ensemble mean is extracted from ECMWF atmospheric ensemble as made available through The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble project (TIGGE, [24]). For more details on the ocean model configuration and its skill, one may refer to [189, 190]. The model has been extensively validated by several earlier studies (e.g. [43, 67, 171, 189, 190, 193, 194]).

6.3.2 Assimilation experiments

The assimilation experiments were conducted based on the routines provided in DART. Except for the differences discussed in the next two paragraphs, the experiments were all performed with the same configuration in terms of assimilated observations, assimilation cycle of 3 days, localization (only in the horizontal direction) using a radius of 300 km, and no-inflation during assimilation following [161]. The system assimilates the [149] level-4 daily SST data available on a $0.25^\circ \times 0.25^\circ$ grid (which was prepared by blending in situ observations with data from the AVHRR infrared satellite; Figure 6.1b), along-track satellite level-3 merged altimeter filtered SLA (corrected for dynamic atmospheric, ocean tide, and long wavelength errors; Figure 6.1c) from the CMEMS [126], and in situ temperature and salinity (Figure 1d) profiles from the EN4.2.1 dataset [68]. For direct comparison with the model SSH during assimilation, SLA is added to a mean sea surface height (MSSH) estimated from a long model run between 2002 and 2016 forced with the best available (5 km) resolution atmospheric forcing that has been dynamically downscaled from the 75 km ECMWF atmospheric reanalysis using an assimilative Weather Research

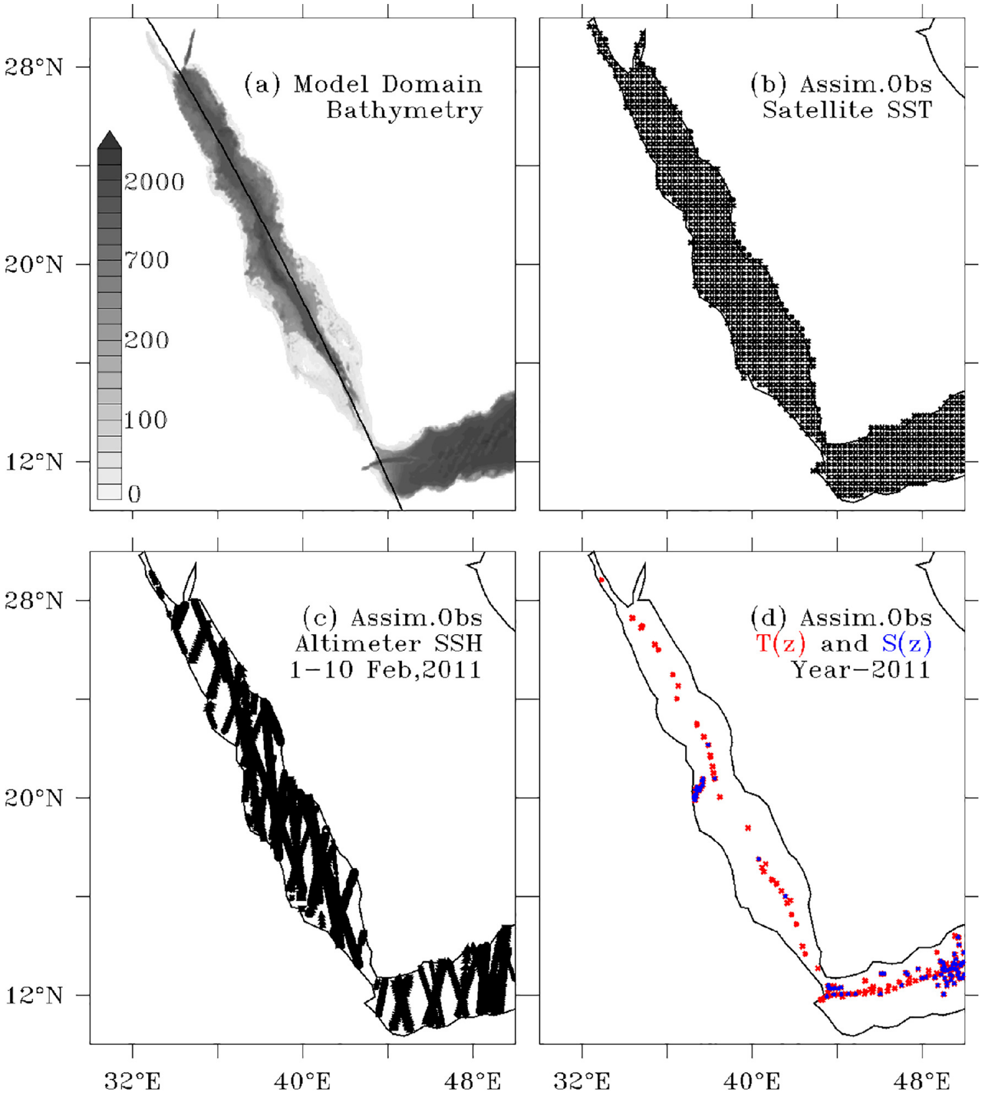


Figure 6.1: (a) Model domain and bathymetry (m). Thick black line represents the Red Sea axis. Panels (b-d) respectively indicate the geographical coverage of assimilated observations of satellite based level-4 SST observations on a typical day, satellite level-3 SSH measurements over a typical altimeter period, and EN4.2.1 in situ temperature (red) and salinity (blue) profiles available over the entire year 2011.

and Forecast (WRF) model [181]. One may also consider estimating the MSSH by assimilating in situ temperature and salinity profiles over a long period [18, 197]. Since our assimilation experiments are only conducted over the year 2011, the MSSH simulated by the high resolution atmospheric forcing is the best available for testing the different assimilation schemes.

The observational error covariance matrix is diagonal with temporally-static and spatially-homogeneous observational error variance values of $(0.04\ \mathrm{m})^2$, $(0.5\ {}^\circ\mathrm{C})^2$ and $(0.3\ \mathrm{psu})^2$ for the satellite along-track SSH, the in situ T and S, respectively, and spatio-temporal error variances for the satellite blended level-4 SST, varying between $(0.1\ {}^\circ\mathrm{C})^2$ and $(0.6\ {}^\circ\mathrm{C})^2$. These relatively large error variances for T and S, which are chosen in accordance with the suggested ranges of in situ observational errors in earlier assimilation studies (e.g. [56, 92, 133, 150], are intended to account for the representational errors due to unresolved scales and processes in the model [159]. The SLA observational error of $(0.04\ \mathrm{m})^2$, which is slightly larger than the suggested altimeter accuracy [16], is selected based on a sensitivity experiment with various values of error variances, $(0.04\ \mathrm{m})^2$, $(0.07\ \mathrm{m})^2$, and $(0.1\ \mathrm{m})^2$ (results not shown here; discussed in [161]). Since our 4km-MITgcm can resolve the scales of the 25 km × 25 km assimilated SST data, only measurements errors of SST data are considered. The specified observational error variances for SST vary in accordance with the analysis errors specified in the level-4 gridded SST product of [149].

Table 6.1 and 6.2 summarize the configurations of the conducted experiments. Three different categories of assimilation experiments are analyzed; *EnOIexp*, *EAKFexp*, and *HyBDexp*. *EnOIexp* employs the same model configuration as *Fexp* and assimilates observations with EnOI, starting from the 1^{st} January, 2011 ocean state obtained from *Fexp*. *EnOIexp* is implemented with a monthly varying 250-member ensembles, generated as in [171] using the last 15-year model hindcasts of the spin-up run. *EAKFexp* assimilates the observations based on EAKF, with a flow dependent background ensemble of 50 members. The initial ensemble in *EAKFexp* is generated by randomly selecting 50 different states corresponding to January's hindcasts of *Fexp* and then re-centering the ensemble mean to the 1^{st} January, 2011 state of *Fexp*. The MITgcm forecasts of the 50 members were forced with the ECMWF ensemble atmospheric forcing. Different model physics were also used for integrating each member, selected from a time-varying ensemble of model physics (hereafter model physics dictionary (MPD)). The MPD encompasses different choices of vertical and horizontal mixing schemes, and viscosity and diffusivity coefficients. These include five types of horizontal diffusion, three schemes of horizontal viscosity, and four schemes of vertical mixing as listed in Table 6.3. More details about the generation of the MPD and of the physical perturbation impact on the assimilation results can be found in [161]. *HyBDexp* is implemented by combining 250-member quasi-static ensemble (used in *EnOIexp*) and 50-member dynamic ensemble with a weighing factor (α) 0.05, selected after examining the sensitivity of the Hybrid system to the value of α. This section examines the sensitivity of the Hybrid system to three different values of the flow-dependent and static covariance weighting factor α (0.15, 0.05, and 0.01). All these sensitivity experiments use the same experimental setting as *HyBDexp*. Examining the results of these experiments, one can notice negligible differences in terms of SST and SSH estimates (e.g. Figure 6.2). Considerable differences are however obtained for subsurface temperature, salinity and SSS. For instance, the large salinity biases in the subsurface salinity layers are significantly improved when the α value is changed from 0.15 (Figure 6.3h) to 0.05 (Figure 6.3f). Similarly, the deepening of 23 °C isotherm is best represented when the Hybrid system is run with $\alpha = 0.05$ (Figure 6.3e). Further decreasing the value of α to 0.01 flattens the 23 °C isotherm (Figure 6.3c) and increases salinity errors (Figure 6.3d) in the surface layers (for instance, Table 6.4 shows that the SSS RMSE is increased from 0.14 to 0.23 when decreasing the value of α from 0.05 to 0.01). Moreover, the SSS features are slightly degraded in the southern Red Sea (Figure 6.4). On an overall note, our Hybrid system with $\alpha = 0.05$ performs the best out of the three sensitivity experiments. *HyBDexp* uses the same initial and atmospheric forcing ensembles, and perturbed internal physics as those of *EAKFexp*. *EnOIexp* was also tested with 300 members and the results were very similar to those of the 250-member case.

Table 6.1: Summary of the experiments conducted to demonstrate the skill of Hybrid system in terms of improved ocean state. In the table "Random" model physics refers to the use of a time-varying ensemble of physics during the model integration of each ensemble member for forecasting.

Experiment	**Initial condition**	**Atm. Forcing**	**Model Physics**	**Assimilated observations**	**Assimilation Category**
Fexp	Single member on 1st January, 2011	Ensemble mean	Standard	None	NA
EnOInoSCLexp	Single member on 1st January, 2011	Ensemble mean	Standard	Reynolds-SST, Altimeter SSH, and in situ temperature and salinity	EnOI before scaling quasi-static-seasonal ensemble of size 300
EnOIexp	Single member on 1st January, 2011	Ensemble mean	Standard	Reynolds-SST, Altimeter SSH, and in situ temperature and salinity	EnOI with scaled quasi-static-seasonal ensemble of size 300
EAKFexp	50-member ensemble on 1st January, 2011	50-member ensemble	Random	Reynolds-SST, Altimeter SSH, and in situ temperature and salinity	50 member EAKF
HyBDexp	50-member ensemble on 1st January, 2011	50-member ensemble	Random	Reynolds-SST, Altimeter SSH, and in situ temperature and salinity	Hybrid with quasi-static-seasonal ensemble of size 250 and dynamic ensemble of size 50
HyBDmPexp	Same as *HyBDexp*	Same as *HyBDexp*	Standard	Same as *HyBDexp*	Same as *HyBDexp*
HyBDmAPexp	Same as *HyBDexp*	Ensemble mean	Standard	Same as *HyBDexp*	Same as *HyBDexp*

Table 6.2: Summary of EAKF and Hybrid assimilation experiments conducted to examine the computational efficiency of the Hybrid system.

Experiment	Initial condition	Atm. Forcing	Assimilation Category
EAKF100exp	100-member ensemble on 1st January, 2011	100-member ensemble	100-member EAKF
EAKF250exp	250-member ensemble on 1st January, 2011	250-member ensemble	250-member EAKF
EAKF500exp	500-member ensemble on 1st January, 2011	500-member ensemble	500-member EAKF
HyBD30exp	30-member ensemble on 1st January, 2011	30-member ensemble	Hybrid with quasi-static-seasonal ensemble of size 270 and dynamic ensemble of size 30
HyBD20exp	20-member ensemble on 1st January, 2011	20-member ensemble	Hybrid with quasi-static-seasonal ensemble of size 280 and dynamic ensemble of size 20
HyBD10exp	10-member ensemble on 1st January, 2011	10-member ensemble	Hybrid with quasi-static-seasonal ensemble of size 290 and dynamic ensemble of size 10

Table 6.3: Dictionary of model physics and associated coefficients considered in the experiments that use perturbed physics. Coefficients of vertical mixing schemes vary according to the standard values in MITgcm, unless otherwise stated. In the table, entries in first row indicate the standard scheme.

Horizontal vicosity	Vertical mixing	Horizontal diffusion
Simple-Harmonic with viscosity coefficient 30 m^2/s	KPP [101]	Implicit diffusion for temperature and salinity
Simple-Bi-harmonic scheme of [75] with viscosity coefficient 10^7 m^4/s	PP81 [136]	Explicit coefficients of 100 m^2/s for temperature and salinity
Harmonic flavor of combined [162] and [106] schemes with viscocity coefficient 30 m^2/s, Smag coefficient 2.5 and Leith coefficient 1.85	MY82 [125]	Gent-McWilliams/Redi [61, 62, 147] using slope clipping of [34], with background diffusion set to 100 m^2/s
	MY82 [59]	Gent-McWilliams/Redi [61, 62, 147] using tapering scheme of [35], with background diffusion set to 100 m^2/s
		Gent-McWilliams/Redi [61, 62, 147] using tapering scheme of [100], with background diffusion set to 100 m^2/s

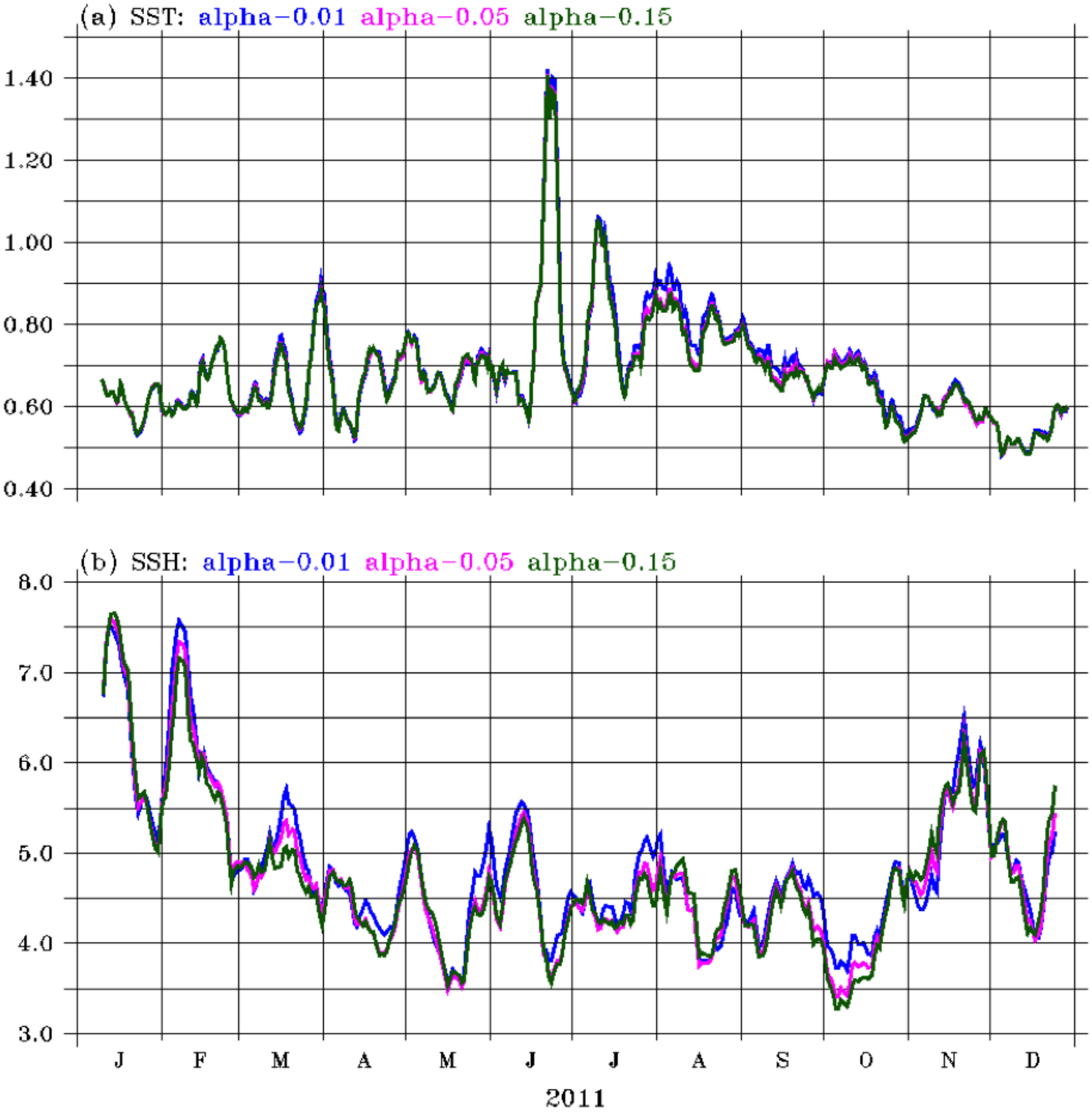


Figure 6.2: Time series of RMSE for daily averaged (a) SST (°C) (b) SSH (cm) from Hybrid experiments with weighting factors α 0.15 (green), 0.05 (pink) and 0.01 (blue). SST RMSE (SSH RMSE) is computed by collocating the daily averaged model forecasts onto level-3 GHRSST (level-3 altimeter observations) product. 10-day smoothing is applied to better highlight the differences among the assimilation results.

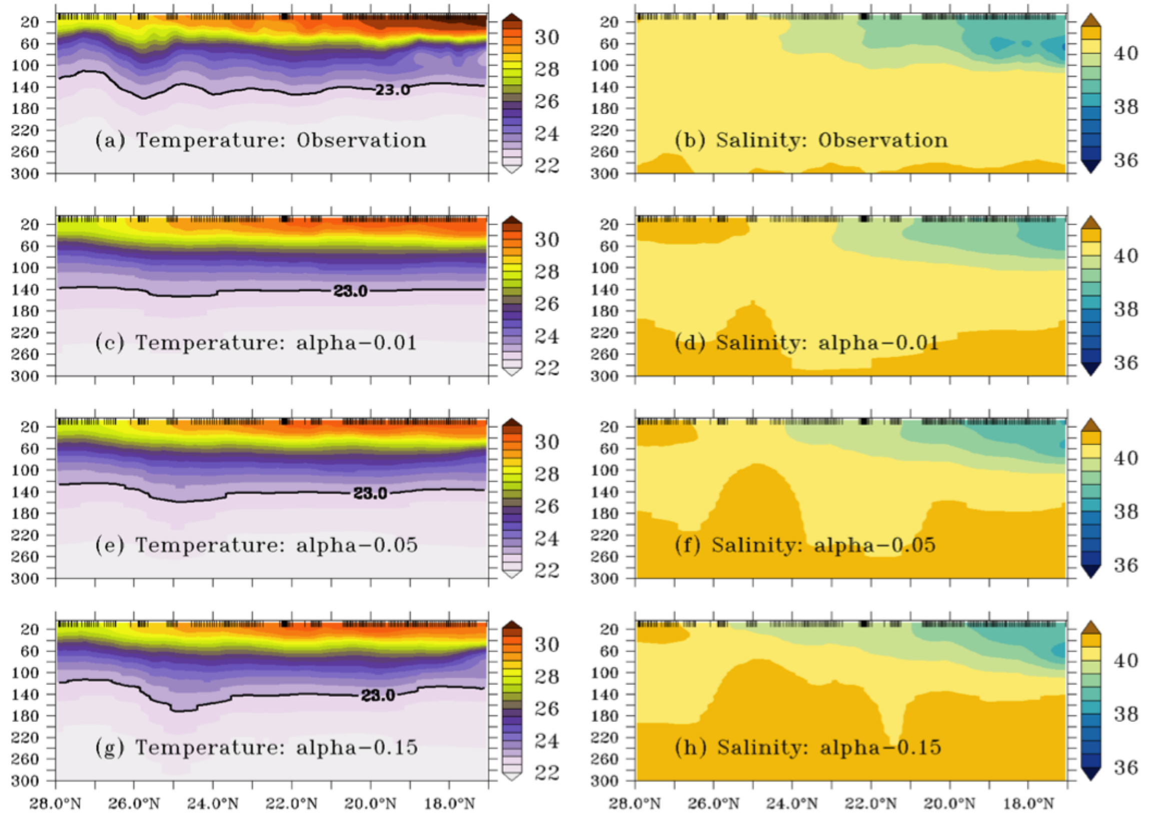


Figure 6.3: Subsurface temperature (°C) and salinity (psu) from in situ CTD observations (a-b) and from the collocated (in space and time, during the WHOI/KAUST summer cruise conducted during 15^{th} September - 8^{th} October, 2011) daily averaged temperature and salinity forecasts as resulted from hybrid experiments with weighting factors α 0.01 (c-d), 0.05 (e-f), and 0.15 (g-h). Temperature and salinity are smoothed by 1 °C and 10 m in latitudinal and vertical direction to better visualize subsurface features. 23 °C isotherm is also indicated in the respective temperature plots by the thick curve. Latitudes corresponding to observation locations are indicated as black vertical dashes at the top of each panel.

Table 6.4: Statistics of the Hybrid (alpha) experiments.

	SST				T (10-700 m)				SSS				S (10-700 m)			
	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr
Observation	29.86	0.95			22.47	2.02			39.64	0.60			40.38	0.36		
***HyBD*-0.01**	29.50	0.83	0.43	0.971	22.58	1.98	0.26	0.993	39.83	0.68	0.23	0.990	40.44	0.29	0.12	0.975
HyBDexp	29.50	0.85	0.44	0.969	22.54	1.93	0.24	0.995	39.66	0.67	0.14	0.982	40.45	0.31	0.13	0.966
***HyBD*-0.15**	29.49	0.86	0.45	0.966	22.51	1.91	0.24	0.994	39.60	0.63	0.20	0.952	40.46	0.31	0.15	0.944
HyBDmAPexp	29.51	0.89	0.41	0.977	22.54	1.98	0.30	0.989	39.54	0.75	0.25	0.962	40.45	0.37	0.13	0.955
HyBDmPexp	29.52	0.86	0.40	0.975	22.54	2.01	0.28	0.991	39.57	0.77	0.24	0.976	40.45	0.36	0.13	0.958

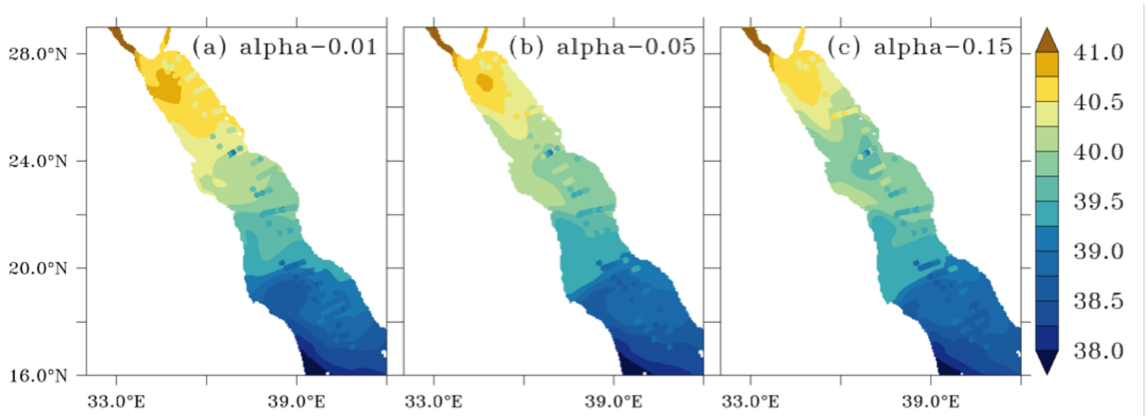


Figure 6.4: Spatial maps of temporally averaged SSS (psu) during the period pertained to the WHOI/KAUST summer cruise (15^{th} September - 8^{th} October, 2011) from Hybrid experiments with weighting factor α 0.15 (a), 0.05 (b), and 0.01 (c). Near surface in situ salinity from the CTD data collected during the summer cruise is also shown with filled circles on each plot.

Figure 6.5 displays the ensemble spread of (a) SST and (b) SSH from *HyBDexp* and *EAKFexp*. It also shows the spread of the quasi-static ensemble before and after scaling its ensemble covariance by a factor of 0.05 (the weighing factor α of *HyBDexp*). The ensemble SSH spread varies between 2-4 cm in *HyBDexp* and *EAKFexp*. The spread of the EnOI-ensemble is significantly larger, but becomes closer to those of *HyBDexp* and *EAKFexp* after the scaling of the quasi-static ensemble covariance. The assimilation results of the experiments using the quasi-static ensemble (here after *EnOInoSCLexp*) and those using the scaling of the quasi-static ensemble covariance (*EnOIexp*) in the EnOI system suggests that this scaling has no significant impact on SST (Figures 6.6a). It however shows noticeable impact on SSH (Figure 6.6b), with the *EnOIexp* exhibiting lower RMSEs compared to *EnOInoSCLexp*. The most pronounced differences between the results of *EnOInoSCLexp* and *EnOIexp* are found in the data-sparse subsurface layers and for under-sampled ocean variables. *EnOInoSCLexp* simulates spurious fresh water anomalies in the surface (Figures 6.7c and 6.7d) and in the subsurface layers (Figure not shown). Such spurious features are likely due to dynamical imbalances, which can be assessed through vertical velocities (e.g. [161]). The maximum vertical speed in the water column, $|W(z)|_{max}$, a proxy for 2D visualization of the abnormal vertical velocities in the ocean column, is suspiciously large in *EnOInoSCLexp* compared to *Fexp* (Figure 6.8a and 6.8b), suggesting important dynamical imbalances in the EnOI before scaling the quasi-static ensemble covariance. *EnOIexp* results in better estimates of the ocean state (Figures 6.6a, 6.6b, and 6.7d) with lesser dynamical imbalances (Figure 6.8c). We therefore evaluate the results of *HyBDexp* against those of *EnOIexp*, the best possible EnOI configuration.

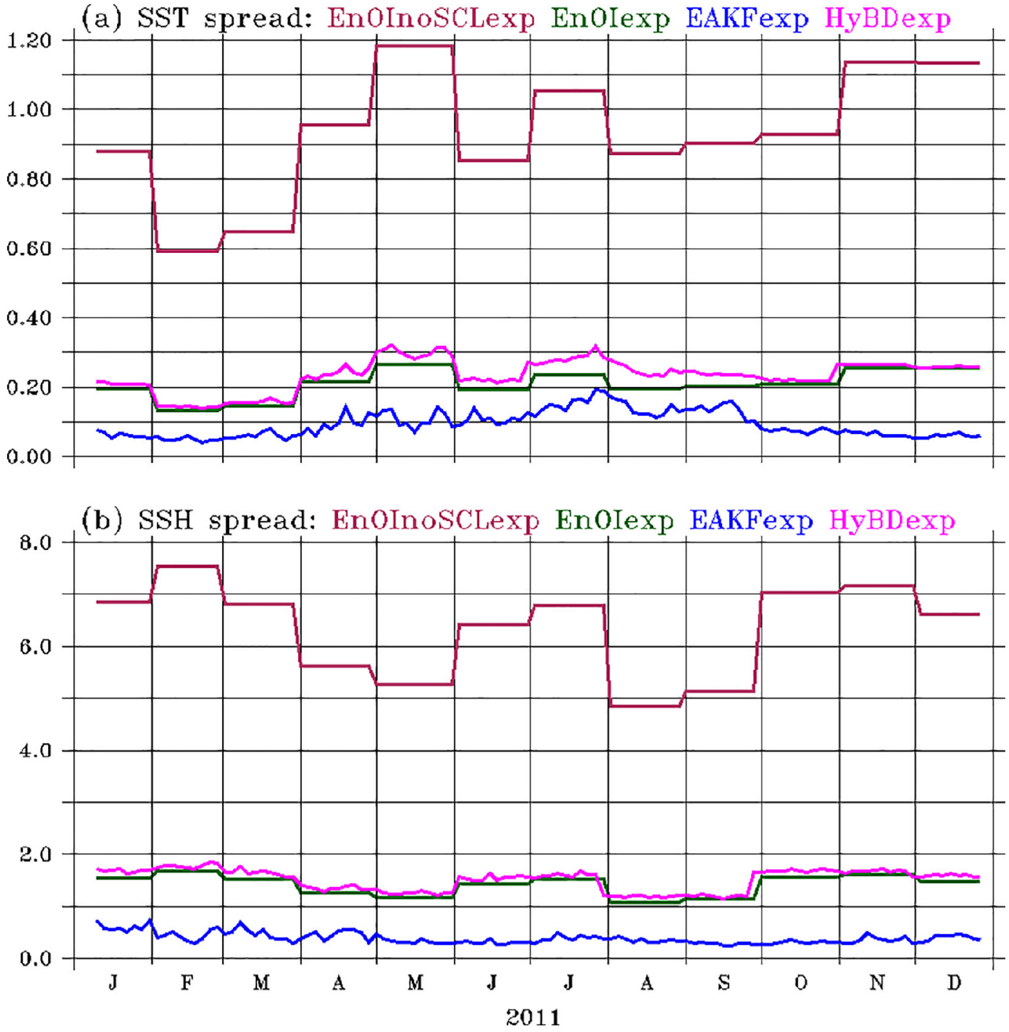


Figure 6.5: Domain averaged ensemble spread of (a) SST and (b) SSH from *EnOInoSCLexp* (green-dash), *EnOIexp* (green-line), *EAKFexp* (blue-line) and *HyBDexp* (pink-line). Units of SST and SSH spread are in °C and cm.

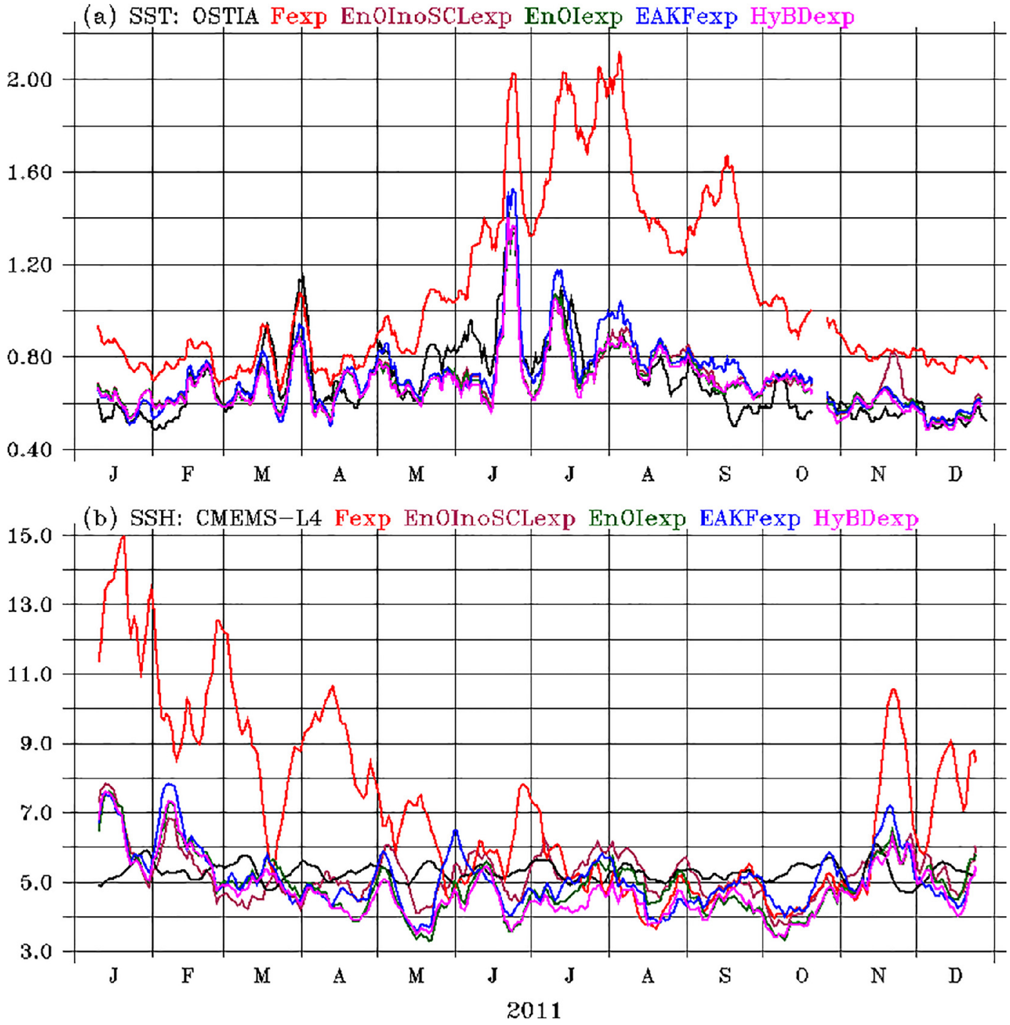


Figure 6.6: Time series of RMSE for daily averaged (a) SST (°C) and (b) SSH (cm) from level-4 gridded products (OSTIA for SST and CMEMS-L4 for SSH; black), *Fexp* (red), *EnOInoSCLexp* (magneta), *EnOIexp* (green), *EAKFexp* (blue), and *HyBDexp* (pink). SST RMSE (SSH RMSE) is computed by collocating the daily averaged model forecasts in the whole model domain onto level-3 GHRSST (level-3 altimeter observations) product. 10-day smoothing is applied to highlight better the differences between the assimilation results.

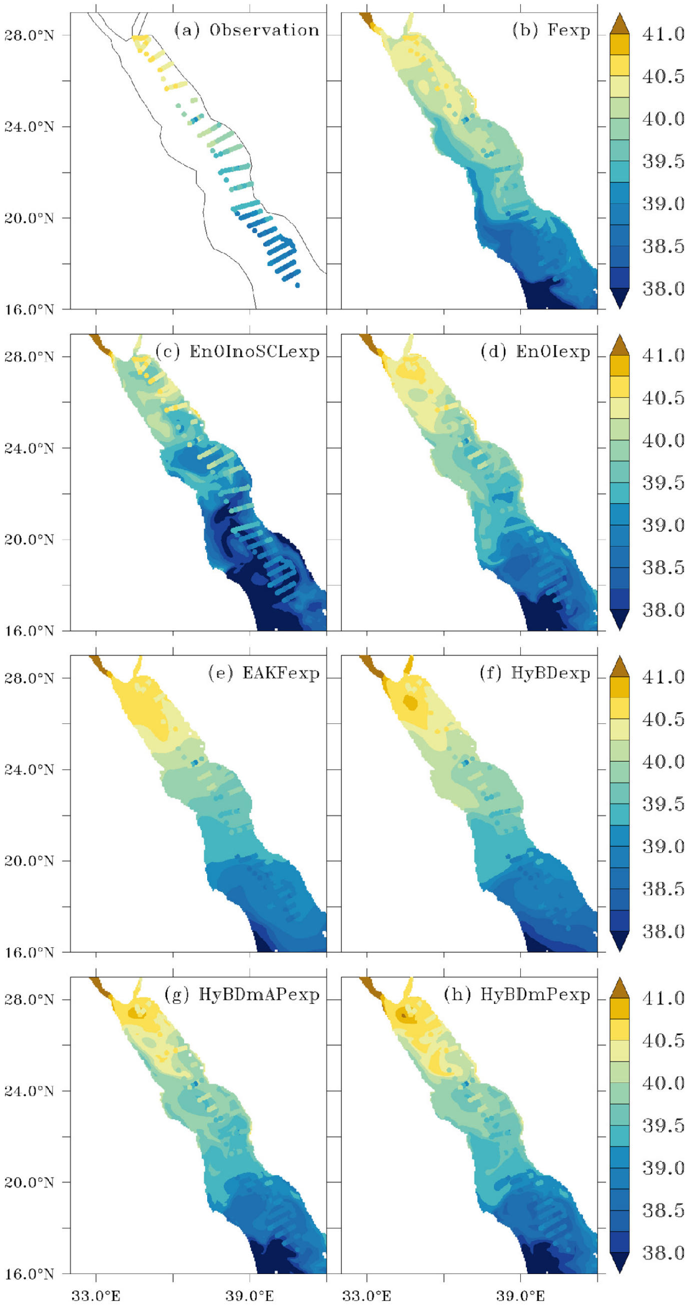


Figure 6.7: SSS (psu) from (a) WHOI/KAUST summer cruise, (b) *Fexp*, (c) *EnOInoSCLexp*, (d) *EnOIexp*, (e) *EAKFexp*, (f) *HyBDexp*, (g) *HyBDmAPexp*, and (h) *HyBDmPexp*. The SSS from the model experiments is the time average between 15th September - 8th October, 2011, pertaining to the WHOI/KAUST summer cruise. SSS of WHOI/KAUST is overlaid on the spatial maps also for model-data comparison.

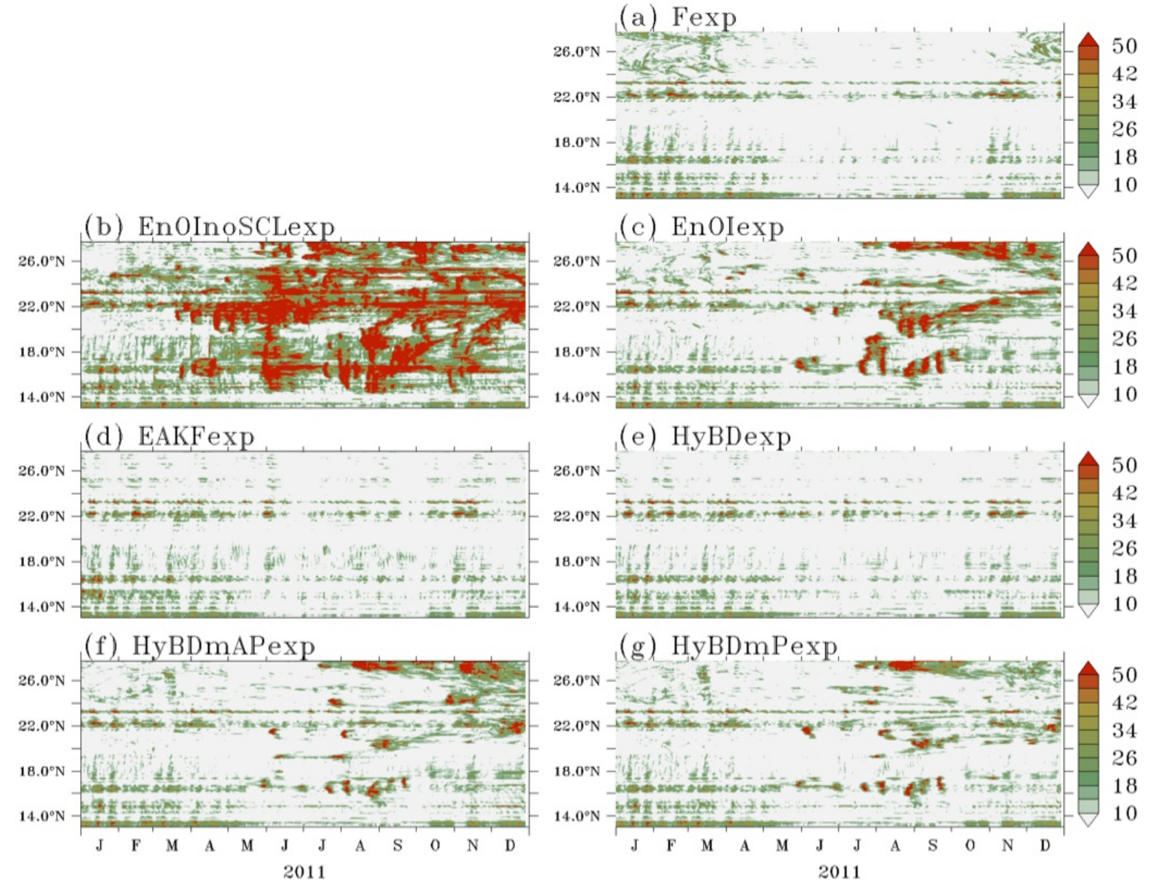


Figure 6.8: Temporal evolution of $|W(z)|_{max}$(m/day) from the daily averaged vertical velocity in the ocean column along the axis of the Red Sea from (a) *Fexp*, (b) *EnOIexp*, (c) *EAKFexp*, (d) *HyBDexp*, (e) *HyBDmAPexp*, and (f) *HyBDmPexp*. Temporal evolution of $|W(z)|_{max}$(m/day) from the daily averaged vertical velocity in the ocean column along the axis of the Red sea (indicated in Figure fig: paper 4 fig 1 accepteda) from (a) *Fexp*, (b) *EnOInoSCLexp*, (c) *EnOIexp*, (d) *EAKFexp*, (e) *HyBDexp*, (f) *HyBDmAPexp*, and (g) *HyBDmPexp*.

6.3.3 Validation data

The daily averaged forecasts from the different experiments are analyzed to assess the performances of the different assimilation configurations. Subsurface estimates are evaluated against CTD observations of T and S profiles collected in the Red Sea between 15th September and 8th October, 2011 (indicated in Figure 6.7a). This dataset includes 206 profiles collected by a joint Woods Hole Oceanography Institute (WHOI) and King Abdullah University of Science and Technology (KAUST) cruise along the eastern Red Sea, with a horizontal spacing of 10 km ([192]; hereafter WHOI/KAUST summer cruise). The data is similar to the one described in the 2nd paragraph of Section 5.5.2, except that the analysis period extends to 31st December, 2011. WHOI/KAUST observations are not assimilated and are therefore used as an independent observations for validation. The assimilated SST and SSH observations were used in the generation of the interpolated level-4 products of OSTIA and CMEMS, and as such these datasets cannot be considered to be fully independent from our assimilated fields.

6.4 Evaluation of the Hybrid system

This section evaluates the outputs of *HyBDexp* compared to *EnOIexp* and *EAKFexp*. We first establish the merits and demerits of *EnOIexp* and *EAKFexp*, and then show how well these are addressed in *HyBDexp*. Figure 6.6 displays the temporal evolution of (a) SST and (b) SSH RMSEs for the entire model domain. RMSEs of SST and SSH are comparatively large and exhibit seasonal dependencies in *Fexp*, with relatively large SST (SSH) RMSEs during summer (winter). The increased SST RMSEs during summer are due to biases in the summer atmospheric fields in the southern Red Sea associated with dust [181]. The increased SSH RMSEs during winter can be related to biases in the surface net heat flux associated with increased atmospheric convective activity [181], which affects strong eddies in the northern Red sea [189, 193, 195].

Assimilation using EnOI or EAKF significantly improves the RMSEs for both SST and SSH, with consistently smaller RMSEs throughout the year. The RMSEs of SST and SSH in these assimilation experiments are even lower than the interpolated products ones. The SST and SSH RMSEs differences between *EnOIexp* and *EAKFexp* are relatively small, with *EnOIexp* yielding slightly better results. For instance, while the SST RMSE (SSH RMSE) corresponding to the whole domain and full year 2011 is 0.68 °C (4.9 cm) in *EnOIexp*, it is 0.71 °C (5.1 cm) in *EAKFexp*. Examining region-wide statistics of SST and SSH suggests that the differences between these two experiments are noticeable only in the Gulf of Aden (Figures/Table not shown), with the EnOI yielding better results compared to EAKF. One may expect the results of *EAKFexp* to improve if uncertainties in the ocean boundary conditions were accounted for (through appropriate perturbations), as this should enhance the ensemble spread in the Gulf of Aden.

To provide more insight into the results for under-sampled regions and ocean variables, we examined the assimilation solution for SSS and subsurface temperature and salinity using independent observations from the WHOI/KAUST. Figure 6.7 displays spatial maps of SSS from the different experiments overlaid with independent observations from WHOI/KAUST summer cruise. Interestingly, the SSS results are very different from the SST and SSH ones, with *EAKFexp* performing significantly better than *EnOIexp*. For instance, the observations indicate a north-south gradient with fresh water in the southern Red Sea and saline-water in the northern Red Sea (Figure 6.7a). Such a prominent north-south salinity gradient is not well reproduced in *EnOIexp* (Figure 6.7d). *EnOIexp* simulates a spurious fresh water pool in the central Red Sea influenced by the advection of anomalous fresh waters from the southern Red Sea, where the SSS differences between observations and *EnOIexp* reach 2 psu. These biases are even larger than that in *Fexp* (Figure 6.7b). The SSS from *EAKFexp*, on the other hand, agrees much better with the observations (compare Figure

6.7e with 6.7b and 6.7c), with the model-data differences being less than 1 psu, and improved north-south spatial gradients of SSS. Figure 6.9 plots the estimated temperature and salinity structures corresponding to the WHOI/KAUST summer cruise observations locations. *EnOIexp* exhibits a salinity bias of 0.5 psu in the subsurface layers throughout the domain (Figure 6.9d), and simulates spurious pockets of high salinity waters in the subsurface layers (180-300m) between 22°-24°N (absolute fields from the assimilation experiments are not shown in the Figure). $|W(z)|_{max}$ is suspiciously large (compared to *Fexp*; Figures 6.8c with 6.8a) in *EnOIexp* almost throughout the Red Sea starting from the middle of the year. As argued in [161], such a large $|W(z)|_{max}$ results from spurious vertical correlations in the quasi-static ensemble of the *EnOIexp*. *EAKFexp* does not show such sporadic behavior. It further improves the subsurface temperature and salinity biases particularly to the north of 20°N. However, as already reported in [161], *EAKFexp* misses high-resolution spatial features such as the deepening of the 23 °C isotherm around 26°N (compare Figure 6.9e with 6.9a), the intrusion of a fresh and cold Gulf-of-Aden water mass around 60 m (which manifest itself as large overestimation of subsurface temperature and salinity south of 20°N; Figure 6.9e-f). *EnOIexp* reproduces these features, but with significant discrepancies in the location of the deeper 23 °C isotherm and in the magnitudes of the temperature/salinity of the Gulf of Aden water mass appearing at the intermediate layers. This is likely related to spurious propagations of surface observations information [159, 160] due to the misrepresentation of the "errors-of-the-day" by the quasi-static ensemble of the EnOI. Such spurious corrections were shown to disrupt the model dynamical balances [12, 25, 80, 105, 141, 161], particularly in the data-sparse subsurface layers (e.g. temperature and salinity) and for under-sampled ocean variables (e.g. SSS).

HyBDexp significantly improves the Red Sea state estimates and also preserves better the dynamical consistency (as can be inferred from reasonable $|W(z)|_{max}$ in

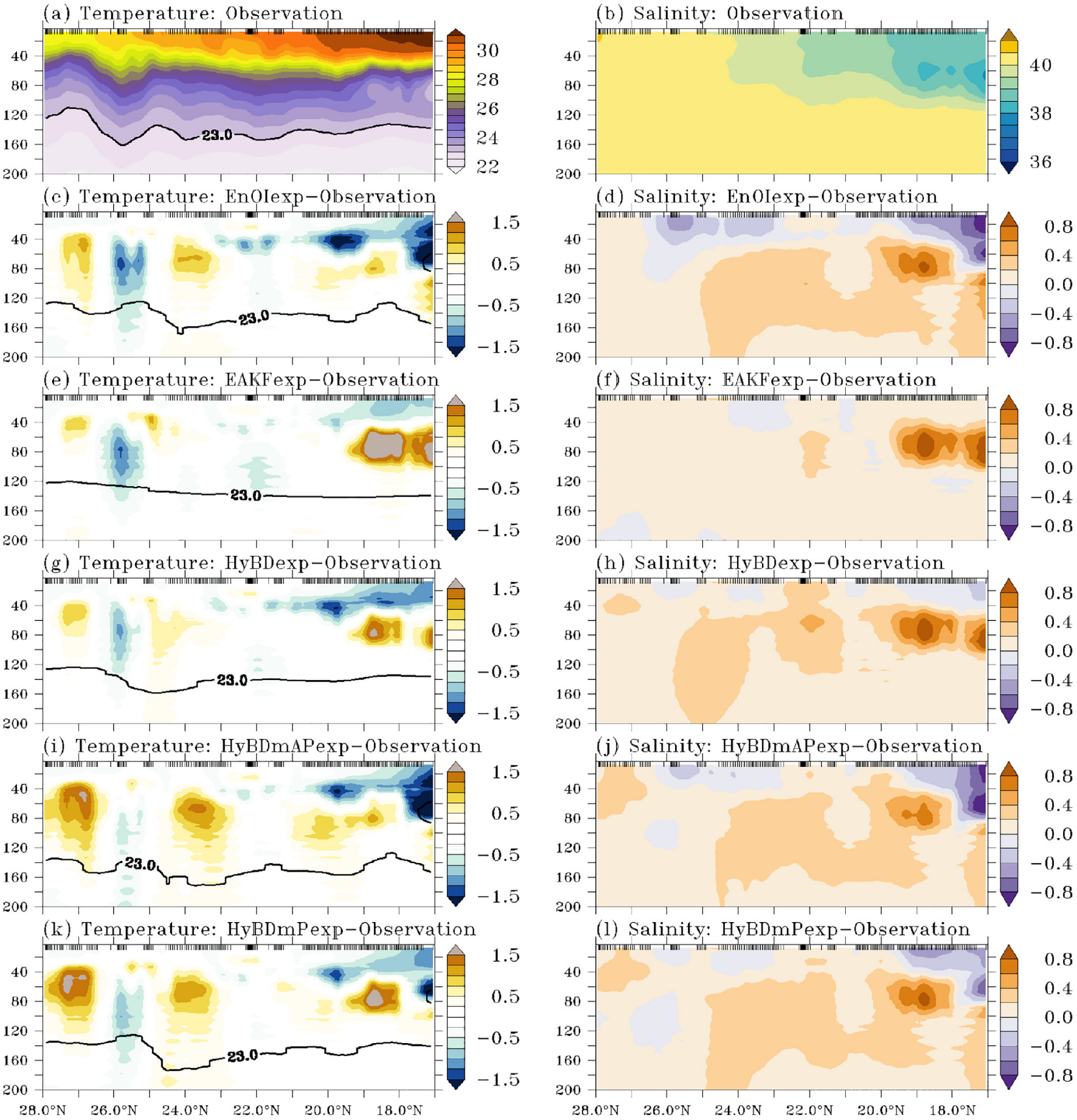


Figure 6.9: Subsurface temperature (°C) and salinity (psu) from in situ CTD observations (a-b) collected during the WHOI/KAUST summer cruise conducted during 15th September - 8th October, 2011. Panels b(c), d(e), f(g), h(i), and j(k) show collocated (in space and time) temperature (salinity) differences between *EnOIexp* and WHOI/KAUST observations, *EAKFexp* and WHOI/KAUST observations, *HyBDexp* and WHOI/KAUST observations, *HyBDmAPexp* and WHOI/KAUST observations, and *HyBmPDexp* and WHOI/KAUST observations respectively. Temperature and salinity observations are smoothed by 1 °C and 10 m in latitudinal and vertical directions to better highlight subsurface features. 23 °C isotherm is also indicated in the respective temperature plots. Latitudes corresponding to observations locations are indicated as black vertical dashes at the top of each panel.

Figure 6.8e). Note that the larger SST and SSH improvements achieved in *EnOIexp* and *EAKFexp* are not compromised in *HyBDexp*. *HyBDexp* indeed does even better than *EnOIexp*, in terms of SSH RMSEs (Figure 6.6b). The SST and SSH RMSEs are improved by 20% in *HyBDexp* compared to *EAKFexp*, reaching 0.2 °C and 1 cm, in terms of differences in SST and SSH RMSEs, respectively. SSS (an undersampled variable) in *HyBDexp*, which was not well simulated by *EnOIexp* and better represented in *EAKFexp*, is closer to the observations in the southern Red Sea with *HyBDexp*. The SSS is even better estimated by *HyBDexp* compared to *EAKFexp* in this region (compare Figure 6.7f with 6.7d and 6.7e). The differences between *HyBDexp* and *EAKFexp* SSS are not very significant over the rest of the domain. Subsurface temperature and salinity are better reproduced by *HyBDexp* (Figure 6.9g-h) compared to *EnOIexp* (Figure 6.9c-d). In addition, *HyBDexp* does better than *EAKFexp* (Figure 6.9e) in capturing the subsurface temperature structure (Table 6.5), particularly the deepening of the 23 °C isotherm in the northern latitudes, which was completely missed in *EAKFexp*. The large subsurface salinity biases introduced by the quasi-static ensemble are however not fully mitigated in *HyBDexp* (Figure 6.9h).

Table 6.5: Statistics of the main experiments.

	SST				**T (10-700 m)**				**SSS**				**S (10-700 m)**			
	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr
Observation	29.86	0.95			22.47	2.02			39.64	0.60			40.38	0.36		
Free	29.15	1.19	0.82	0.953	22.55	1.88	0.271	0.994	39.65	0.57	0.21	0.94	40.40	0.32	0.10	0.967
EnOIexp	29.52	0.90	0.40	0.976	22.49	1.94	0.274	0.991	39.42	0.74	0.33	0.96	40.45	0.38	0.15	0.944
EAKFexp	29.51	0.85	0.41	0.980	22.58	1.95	0.271	0.993	39.71	0.59	0.11	0.99	40.42	0.30	0.10	0.979
HyBDexp	29.50	0.85	0.44	0.969	22.54	1.93	0.237	0.995	39.66	0.67	0.14	0.98	40.45	0.31	0.13	0.966

6.4.1 Meso-scale eddy features

To demonstrate the merits of the *HyBDexp* system in reproducing the Red Sea mesoscale eddy features, spatial maps of SSH snapshots are shown in Figure 6.10, where clear differences can be seen between *HyBDexp* and *EAKFexp* in the northern Red Sea, central Red Sea, and Gulf of Aden. Figure 6.10 displays along-track SSH observations on 6^{th} November, 2011 (top), 15^{th} July, 2011 (middle), and 30^{th} September, 2011 (bottom) overlaid on the corresponding daily averaged spatial maps from CMEMS-L4 (left), *EAKFexp* (middle) and *HyBDexp* (right). The 6^{th} November, 2011 corresponds to a period of an anomalous cyclonic eddy in the northern Red Sea (Figure 6.10a) [138]. It is largely modulated by the local net heat flux and remote sea level perturbations from the southern Red Sea [193]. Around 15^{th} July, 2011, the central Red Sea hosted an anti-cyclonic eddy around 21°N (Figure 6.10b). Such an eddy, whose presence may have been tied to coastline and topographic variations (e.g. [143], appears every June and lasts until July (e.g. [146]). Around 30^{th} September, 2011 the Gulf of Aden experienced a series of eddies (Figure 6.10c), which results from instabilities in the adjacent Somali current and in the nearby large eddies, such as the Great Whirl and Socorta eddy [5, 188].

Comparing the assimilation estimates with SSH observations (for both along-track SSH and interpolated product CMEMS-L4) suggests that the intensity and the size of the eddies are underestimated in *EAKFexp* (Figure 6.10d-f). *EAKFexp*, for instance, completely misses the anti-cyclonic eddy in the central Red Sea. *HyBDexp* significantly improves the eddy features in terms of their intensity and also size, irrespective of the region (Figure 6.10g-i). For instance, the missed anti-cyclonic eddy in *EAKFexp* (Figure 6.10e) is reproduced reasonably well by *HyBDexp* (Figure 6.10h), albeit slightly shifted. The intensities of the series of alternating eddies in the Gulf of Aden, and the anomalous cyclonic eddy in the northern Red Sea, are also better represented in *HyBDexp* (Figure 6.10i) than *EAKFexp* (Figure 6.10f). In fact, *HyBDexp* is a closer match to the observations than CMEMS-L4 (Figure 6.10a-c) in terms of eddies intensities.

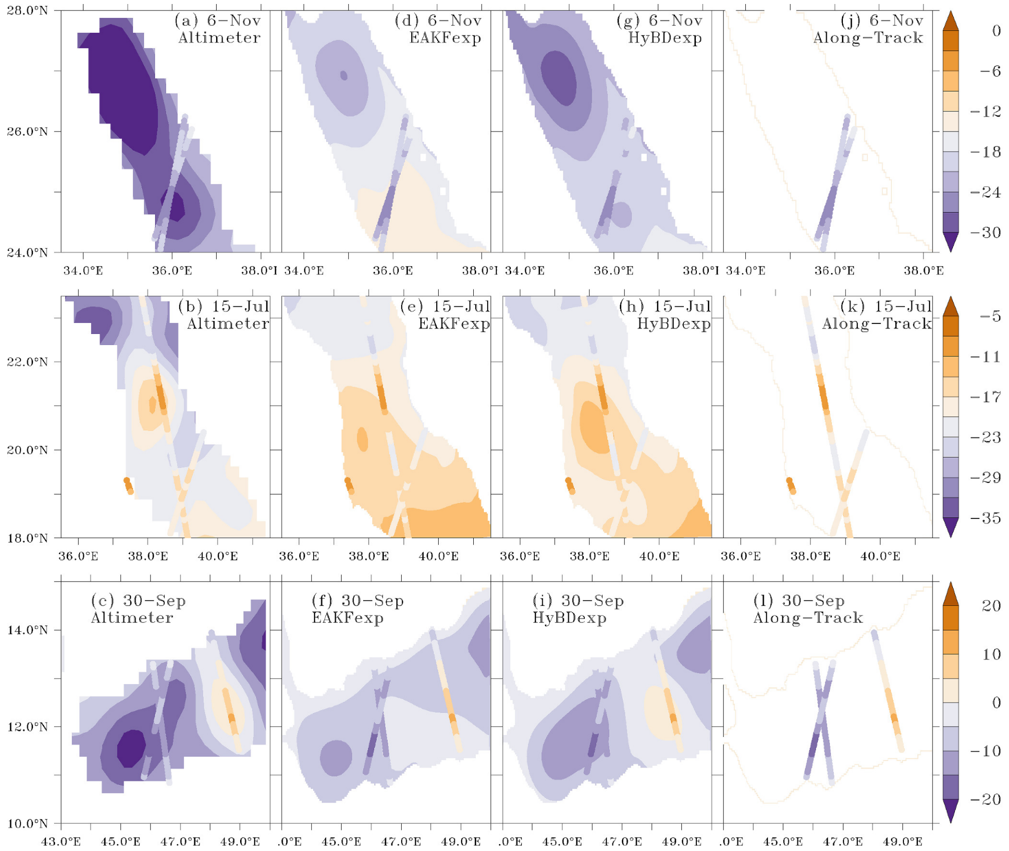


Figure 6.10: Spatial maps of daily averaged SSH (in cm) corresponding to 6th November, 2011 (top), 15th July, 2011 (middle), and 30th September, 2011 (bottom) from (a-c) merged altimeter CMEMS-L4. Panels (d-f), (g-i), and (j-l) show similar plots as resulted from *EAKFexp* and *HyBDexp* forecasts, and along-track observations, respectively. Along track SSH observations of the corresponding day is also overlaid on each map.

6.4.2 Importance of accounting for uncertainties in internal model physics and atmospheric forcing

[161] demonstrated the importance of accounting for background errors due to uncertainties in the internal ocean model physics and atmospheric forcing in the EAKF. To examine the significance of these in the Hybrid system, we have conducted two more *HyBDexp* experiments, *HyBDmAPexp* and *HyBDmPexp*. *HyBDmAPexp* and *HyBDmPexp* are the same as *HyBDexp* except that *HyBDmAPexp* uses the default internal model physics and is forced by the ensemble mean ECMWF atmospheric fields, and *HyBDmPexp* uses the default internal model physics and is forced by the ensemble ECMWF fields. Figures 6.8f and 6.8g display $|W(z)|_{max}$ along the Red Sea axis from *HyBDmAPexp* and *HyBDmPexp*, respectively. Compared to *HyBDexp*, the spread of the large $|W(z)|_{max}$ becomes wider in *HyBDmPexp*, and even wider in *HyBDmAPexp*, suggesting degraded dynamical balances. This is even manifested in SSS, and subsurface temperature and salinity. For instance, as can be seen from Figures 6.7g and 6.7h, the north-south SSS gradient is not well represented in *HyBDmAPexp* and *HyBDmPexp* compared to *HyBDexp*. They also show anomalous fresh waters in the southern parts of the Red Sea. The subsurface temperature (Figure 6.9i and 6.9k) and salinity (Figure 6.9j and 6.9l) also become noisy, and show spurious features and increased biases. These results clearly emphasize the importance of accounting for uncertainties in the atmospheric forcing and internal model physics in the Hybrid system.

6.4.3 Computational gain

This section focuses on assessing *HyBDexp* in terms of computational efficiency, an important aspect of this study. This is achieved by first assessing the sensitivity of EAKF ocean state estimates to gradually increased ensemble size.

We first present results from four different EAKF experiments: the standard 50-

member *EAKFexp*, *EAKF100exp*, *EAKF250exp*, and *EAKF500exp*. The last three experiments are similar to *EAKFexp* but use 100, 250, and 500 ensemble members, respectively. Table 6.2 outlines the configurations of these experiments. The initial ensembles of these experiments are generated as in *EAKFexp*, and the atmospheric forcing is sampled, assuming a normal distribution, using the ensemble mean and spread of the original 50-member ensemble atmospheric forcing of *EAKFexp*. Examining the assimilation results of these experiments suggests little differences in terms of SST and SSH (Figures not shown), which is expected owing to the homogeneous observations coverage of these data sets. No clear differences in the results are found for the sparsely observed temperature variable either. More pronounced differences are obtained however with salinity, the most under-sampled variable, both at surface and subsurface. Increasing the size of the ensemble from 50 to 100 reduces salinity biases in the intermediate layers (Figure 6.11b). Noticeable improvements in the salinity are also achieved in the whole water column when increasing the ensemble size from 100 to 250. Comparing the spatial maps and RMSEs of SSS of these EAKF experiments (Figures 6.12a-d; Table 6.6) with the in situ observations suggest that errors in SSS seem to reach a plateau after using 250 ensemble members. Further increasing the size of the ensemble from 250 to 500 resulted in negligible improvements, suggesting that 250 members are enough to describe the statistics of the filtering errors given the considered uncertainties (from ECMWF ensemble forcing and perturbed physical parameterizations) in the system. The SSS and subsurface salinity in *EAKF250exp* (Figures 6.11c and 6.12c) are clearly comparable or slightly better than those of *HyBDexp* (Figure 6.9h and 6.12e). However, overall, *EAKF250exp* results are still not as good as *HyBDexp*, with substantial differences between the two in terms of SST and SSH, and subsurface temperatures (Figures not shown, as SST and SSH RMSEs and subsurface temperatures in *EAKF250exp* are very similar to those of *EAKFexp*, and Figures 6.6 and 6.9 have already outlined the better performances of *HyBDexp*).

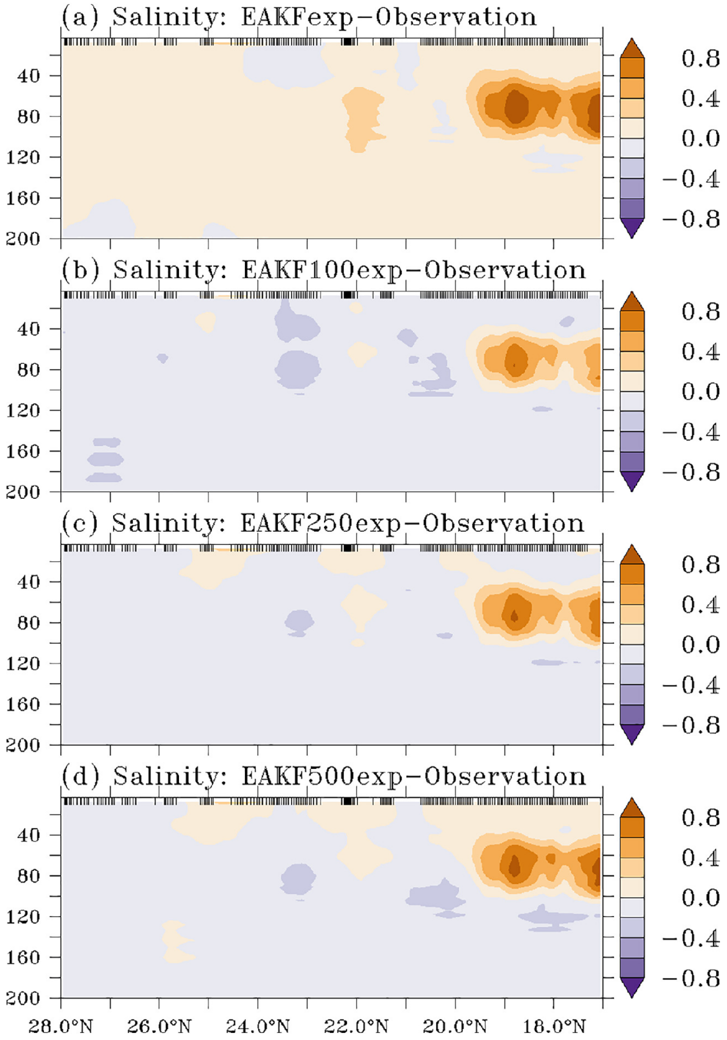


Figure 6.11: Collocated salinity differences (psu) between (a) *EAKFexp* simulations and WHOI/KAUST observations, (b) *EAKF100exp* simulations and WHOI/KAUST observations. (c) *EAKF250exp* simulations and WHOI/KAUST observations, and (d) *EAKF500exp* simulations and WHOI/KAUST observations.

Table 6.6: Statistics of the EAKF experiments.

	SST				**T (10-700 m)**				**SSS**				**S (10-700 m)**			
	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr	Mean	STD	RMSE	Corr
Observation	29.86	0.95			22.47	2.02			39.64	0.60			40.38	0.36		
EAKF50exp	29.48	0.86	0.419	0.987	22.57	1.91	0.26	0.994	39.77	0.71	0.20	0.991	40.43	0.30	0.12	0.9679
EAKF100exp	29.50	0.82	0.430	0.977	22.64	1.97	0.30	0.993	39.54	0.63	0.12	0.995	40.36	0.37	0.09	0.9702
EAKF250exp	29.50	0.83	0.426	0.979	22.65	1.99	0.32	0.992	39.63	0.59	0.08	0.992	40.38	0.35	0.08	0.9733
EAKF500exp	29.50	0.83	0.424	0.979	22.65	2.00	0.32	0.991	39.68	0.57	0.09	0.993	40.38	0.34	0.09	0.9689

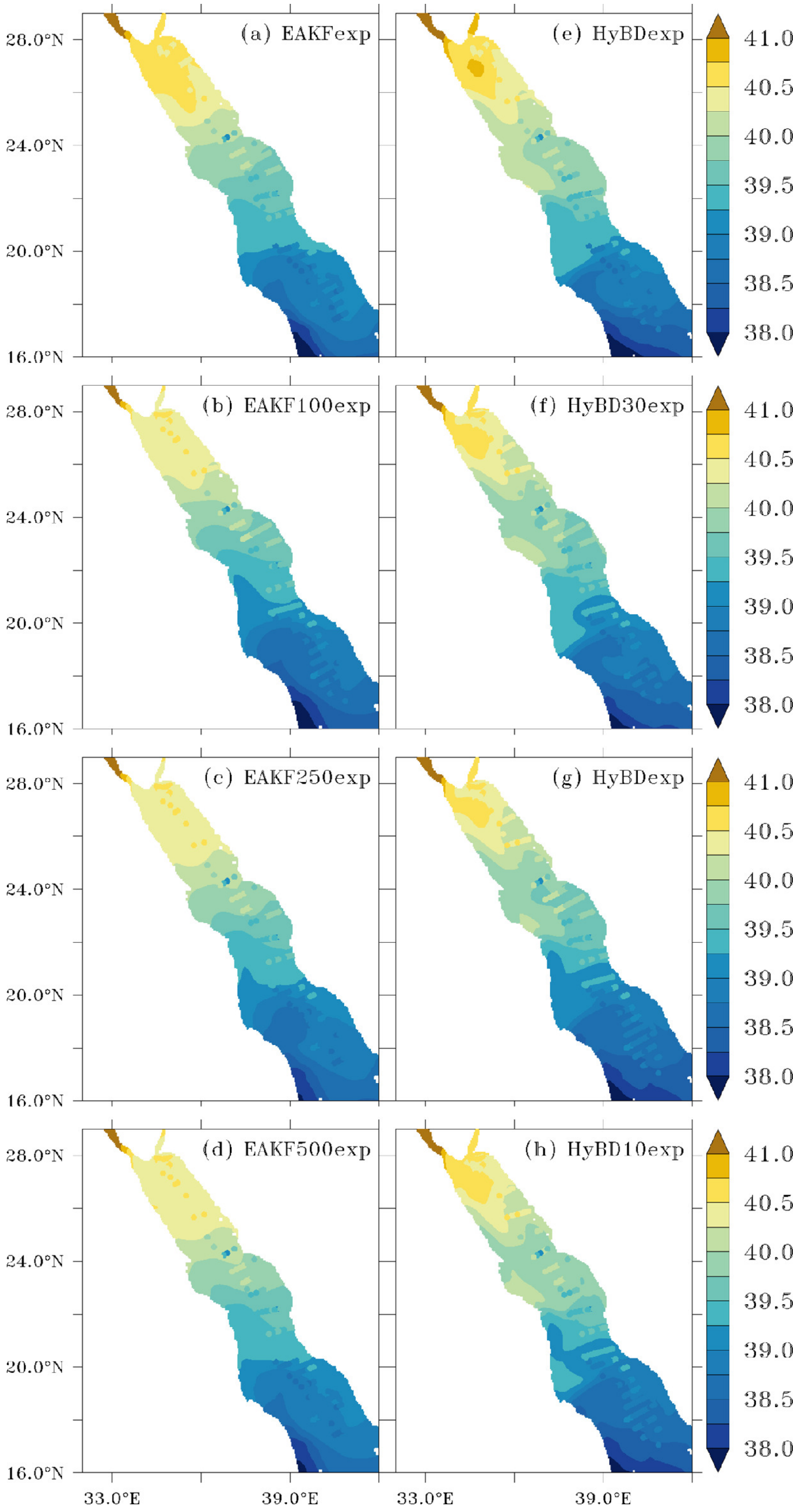


Figure 6.12: Same as Figure 6.7 except that the results are shown for various EAKF (top) and Hybrid (bottom) sensitivity experiments pertained to the size of the ensemble. Panels a-d corresponds to *EAKFexp*, *EAKF100exp*, *EAKF250exp*, and *EAKF500exp*, respectively. Similarly, e-h corresponds to *HyBDexp*, *HyBD30exp*, *HyBD20exp*, and *HyBD10exp*, respectively.

To assess the sensitivity of the Hybrid ensemble system to the flow-dependent and static ensembles sizes, we examined the outputs of *HyBDexp* by gradually decreasing the number of flow-dependent members from 50 to 10 while maintaining the ensemble size at 300 members. As summarized in Table 6.2, these experiments, *HyBD10exp*, *HyBD20exp*, *HyBD30exp*, are the same as the *HyBDexp* (50 dynamical + 250 Static) experiment, but use less flow-dependent members and more static members: 10+290, 20+280, 30+270, respectively. Comparing the results of these experiments (Figures not shown) suggests insignificant changes in the subsurface temperature and salinity and in SST and SSH. Though not substantial, the results differ mainly for SSS. Decreasing the size of the dynamic-ensemble from 50 to 30 slightly degrades SSS (compare Figures 6.12f and 6.12e), particularly in the southern parts of the Red Sea. We also see slight degradations in SSS when the dynamic-ensemble size is decreased from 30 to 20 (Figure 6.12g) and from 20 to 10 (Figure 6.12h). Overall, all *HyBDexp* experiments, including *HyBD10exp*, are at least as good as *EAKF250exp*.

We finally discuss the computational savings achieved by the *HyBDexp* system, comparing the CPU-hours of *EAKF250exp*, *HyBD10exp*, and *HyBDexp* on our high performance supercomputer facility SHAHEEN-II (https://hpc.kaust.edu.sa). Table 6.7 outlines the break-up of computational load of these experiments for an assimilation cycle. The 4km-MITgcm of the Red Sea (array size = 500 x 500 x 50) running with the 200 s integration time step on 3 nodes (each node contains 32 cores with 128 GB flash memory) consumes 4.5 core-hours for a 3-day integration, the length of the assimilation cycle. The update step with DART consumed 40 core-hours for 250-member ensemble when implemented on 20 nodes. *EAKF250exp* consumed 1180 core-hours to complete one assimilation cycle (update+forecast). *HyBDexp* (*HyBD10exp*) calls the DART update step twice, consuming 7+40=47 (3+40=43) core-hours. The total computational cost of *HyBDexp* (*HyBD10exp*), which integrates 50 (10) MIT-gcms, is 275 (89) core-hours. This means that the Hybrid systems (here *HyBDexp* and *HyBD10exp*) led to 76-92% CPU-hours saving with respect to the EAKF-based system (*EAKF250exp*).

Table 6.7: Statistics of computational expenditure, in terms of core-hours, associated with different assimilation experiments. In the table, we show computational expenditure incurred for each component of the assimilation system: the ocean model MITgcm, the first part of assimilation code DART, and the second part of assimilation code, HDART. Note that the *HyBDexp* systems run two assimilation codes, DART and HDART, parallelly.

Experiment	**MITgcm**	**DART**	**HDART**	**Total**	**improvement (%)**
EAKF250exp	1140	40	NA	1180	NA
HyBDexp	228	7	40	275	76
HyBD10exp	46	3	40	89	92

6.5 Summary and conclusions

A new Hybrid data assimilation system was developed for the Red Sea using a 4km-MITgcm and DART. It combines static, but seasonally varying, ensemble members and EAKF-flow-dependent members. The dynamical EAKF members were forecasted with MITgcm forced with atmospheric forcing ensembles and perturbed internal physics. EnOI and EAKF have their own merits and the new Hybrid-EAKF system was able to further improve their performance and helped mitigating their limitations. EnOI was shown to enhance the SST and SSH estimates compared to the EAKF, but degraded the ocean estimates in the under-sampled regions and variables, such as subsurface temperature, salinity and SSS. It further disturbed the dynamical balances of the ocean state. EAKF preserved the dynamical balances better and represented better the under-sampled variables. It was however less efficient at capturing some of the high resolution features, which are important components of the Red Sea circulation. By complementing the flow-dependent ensemble with static members, the Hybrid-EAKF system was able to capture most of the high resolution mesoscale eddy features, and yielded noticeable improvements in SSH, subsurface temperature, and SSS compared to both EnOI and EAKF. In the deeper layers, EAKF salinity estimates remained relatively better than the Hybrid estimates, when evaluated against the few available subsurface observations. Hybrid-EAKF further outperformed EAKF with 250 members. Reducing the number of dynamical members from 50 to 10 did not significantly affect the Hybrid results, but led to drastic (more than 75% in our setup) computational savings compared to the EAKF systems.

The significant improvements, in terms of both quality of ocean state estimates and computational cost, offered by the Hybrid-EAKF system is a motivation for both ocean data assimilation and operational communities developing ensemble data assimilation systems in the Red sea and other regional seas. The fact that the Hybrid-EAKF outperforms the best performing EAKF system (that saturated at 250 members), even when accounting for uncertainties in the atmospheric forcing and internal physics, suggests that the EAKF system is still missing some sources of uncertainties. Uncertainties in the open boundary conditions or bathymetry may be part of these imperfections, and will be considered in our future work.

Chapter 7

Concluding Remarks

7.1 Summary

This dissertation documents our contribution to the implementation of a multipurpose ensemble data assimilation system configured for the Red Sea, a future key asset for the region and the surrounding countries. The system is based on the Massachusetts Institute of Technology general circulation model (MITgcm) to carry out the Red Sea circulation simulations and the Data Research Testbed (DART) package to perform the ensemble data assimilation updates. The DART-MITgcm system was implemented on Shaheen, KAUST world-class supercomputer. Different assimilation schemes have been implemented. The ensemble adjustment Kalman filter (EAKF) integrates the ensemble members in parallel, at the forecast step, and might require large ensembles for its performance. However, using too large ensembles ($\mathcal{O}(100)$) could result in a system crash and failure. While ensembles of 100 members were sufficient for the EAKF to yield satisfactory results, auxiliary techniques such as localization and inflation were employed to mitigate the underestimation of the uncertainties due to the small sizes of the ensembles. Taking advantage of our computational resources, we successfully conducted the first 1000 members ocean ensemble assimilation run and explored the until then uncharted big ensemble territories. This expedition was very fruitful as it revealed the capability of the system to perform outstandingly without the artificial auxiliary fixes. The journey was not without pitfalls as the system instability grew with the ensemble size. This allowed us to develop an enhancement of the system by making it more reliable, fault-tolerant and less prone to disruption.

In instances were gigantic computational resources are out of reach, which is most of the time the case, alternative procedures are ineluctable. In this regard, we designed and consolidated the DART-MITgcm ensemble assimilation system with new assimilation schemes. The ensemble Optimal Interpolation (EnOI) lowers the computational coast by advancing only the mean of the ensemble and makes use of a static covariance. The EnOI sustained the system spread unlike the EAKF, which suffered from the ensemble inbreeding where all the ensemble members converge to the ensemble mean. The static covariance was however not able to account for the prevailing seasonal variability of the Red Sea circulation. The EnOI was therefore enhanced by designing a scheme with seasonally-varying ensembles, the Seasonal EnOI (SEnOI), to remedy the seasonal variability issue. The SEnOI restricts the selected members to set of periods meant to represent the variability of the Red Sea across the year, and one static covariance is built accordingly, for each given season. A further improvement was then suggested to the (S)EnOI schemes by adaptively selecting, with respect to some metrics or criteria, the ensemble members throughout the assimilation process, from a dictionary describing the variability of the Red Sea. The covariance is then constructed from those members related to the current forecast state. Two choices were considered to relate the selected members to the current forecast state: the L2-distance between the forecast and the dictionary elements and the Orthogonal Matching Pursuit (OMP) algorithm. Finally, the key features of the EAKF and the EnOIs schemes have been merged in a Hybrid assimilation scheme that combines a flow-dependent covariance from the EAKF and a static covariance. Another benefit of the Hybrid scheme is to be adjustable to the available resources by increasing or decreasing the size of the flow-dependent component. The Hybrid scheme can lead to striking cost reduction when using small dynamical ensemble with appropriate static ensemble while still providing reliable estimates, without appreciable differences in the results compared to a larger dynamical ensemble.

The system has been validated with numerical experiments forced with real time atmospheric fields from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Center for Environmental Prediction (NCEP), while assimilating real data, namely satellite sea surface height (SSH), sea surface temperature (SST), in situ and cruise data. The results were compared to independent data.

The EAKF provides the best estimates when the underlying assumptions are met. In realistic applications, however, the involved functions are nonlinear, the error are non Gaussian, and the ensemble size is very limited due to computational resources. In these situations in which the EAKF performance is suboptimal, the other schemes might be competitive. Indeed, with suitable static/selected ensembles, the EnOIs schemes can yield results comparable to an optimal EAKF, and so does the Hybrid scheme, except that the Hybrid, combined with perturbed internal physics, further significantly improved the subsurface solution and its dynamical balances.

7.2 Future Research Work

After successfully implementing a fault-tolerant DART-MITgcm ensemble assimilation system tailored for the Red Sea, equipped with cost efficient schemes targeting an operational usage of the system, further steps are to be undertook to reach a full real-time capability. To this end, given that some uncertainties still need to be accounted for, for example the open boundary conditions, the bathymetry, to name a few, we envisage new schemes. Among them is the Stochastic EnKF with second order observation error sampling, expected to improve the estimation of the error covariance since it is designed to match the theoretical error covariance. We also plan to implement Gaussian-mixture filters within DART two-steps assimilation framework to better account for the system nonlinearity and envision using the developed schemes in this thesis (e.g. Hybrid EAKF formulation) for their efficient implementation. We will also work on the synchronization of the system with observation collecting devices and the processing of the observations on the fly to develop the first operational system for the Red Sea.

REFERENCES

[1] Adcroft, A., Campin, J.M., Hill, C., Marshall, J.: Implementation of an atmosphere-ocean general circulation model on the expanded spherical cube. Monthly Weather Review **132**(12), 2845–2863 (2004). DOI 10.1175/MWR2823.1. URL https://doi.org/10.1175/MWR2823.1

[2] Adcroft, A., Hill, C., Marshall, J.: Representation of topography by shaved cells in a height coordinate ocean model. Monthly Weather Review **125**(9), 2293–2315 (1997). DOI 10.1175/1520-0493(1997)125⟨2293:ROTBSC⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1997)125⟨2293:ROTBSC⟩2.0.CO;2

[3] Ait-El-Fquih, B., El Gharamti, M., Hoteit, I.: A bayesian consistent dual ensemble kalman filter for state-parameter estimation in subsurface hydrology. Hydrology and Earth System Sciences **20**(8), 3289–3307 (2016). DOI 10.5194/hess-20-3289-2016. URL https://www.hydrol-earth-syst-sci.net/20/3289/2016/

[4] Aksoy, A., Dowell, D.C., Snyder, C.: A Multicase Comparative Assessment of the Ensemble Kalman Filter for Assimilation of Radar Observations. Part I: Storm-Scale Analyses. Monthly Weather Review **137**(6), 1805–1824 (2009). DOI 10.1175/2008MWR2691.1. URL https://doi.org/10.1175/2008MWR2691.1

[5] Al Saafani, M.A., Shenoi, S.S.C., Shankar, D., Aparna, M., Kurian, J., Durand, F., Vinayachandran, P.N.: Westward movement of eddies into the gulf of aden from the arabian sea. Journal of Geophysical Research: Oceans **112**(C11) (2007). DOI 10.1029/2006JC004020. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006JC004020

[6] Altaf, M.U., Butler, T., Mayo, T., Luo, X., Dawson, C., Heemink, A.W., Hoteit, I.: A comparison of ensemble kalman filters for storm surge assimilation. Monthly Weather Review **142**(8), 2899–2914 (2014). DOI 10.1175/MWR-D-13-00266.1. URL https://doi.org/10.1175/MWR-D-13-00266.1

[7] Anderson, J., Hoar, T., Raeder, K., Liu, H., Collins, N., Torn, R., Avellano, A.: The data assimilation research testbed: A community facility. Bulletin of the American Meteorological Society **90**(9), 1283–1296 (2009). DOI 10.1175/2009BAMS2618.1. URL https://doi.org/10.1175/2009BAMS2618.1

[8] Anderson, J.L.: An ensemble adjustment kalman filter for data assimilation. Monthly Weather Review **129**(12), 2884–2903 (2001). DOI 10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2

[9] Anderson, J.L.: A local least squares framework for ensemble filtering. Monthly Weather Review **131**(4), 634–642 (2003). DOI 10.1175/1520-0493(2003)131⟨0634:ALLSFF⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2003)131⟨0634:ALLSFF⟩2.0.CO;2

[10] Anderson, J.L.: An adaptive covariance inflation error correction algorithm for ensemble filters. Tellus A: Dynamic Meteorology and Oceanography **59**(2), 210–224 (2007). DOI 10.1111/j.1600-0870.2006.00216.x. URL http://dx.doi.org/10.1111/j.1600-0870.2006.00216.x

[11] Anderson, J.L.: Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. Physica D: Nonlinear Phenomena **230**(1), 99 – 111 (2007). DOI http://dx.doi.org/10.1016/j.physd.2006.02.011. URL http://www.sciencedirect.com/science/article/pii/S0167278906002168. Data Assimilation

[12] Anderson, J.L.: Spatially and temporally varying adaptive covariance inflation for ensemble filters. Tellus A **61**(1), 72–83 (2009). DOI 10.1111/j.1600-0870.2008.00361.x. URL http://dx.doi.org/10.1111/j.1600-0870.2008.00361.x

[13] Anderson, J.L., Anderson, S.L.: A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. Monthly Weather Review **127**(12), 2741–2758 (1999). DOI 10.1175/1520-0493(1999)127⟨2741:AMCIOT⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1999)127⟨2741:AMCIOT⟩2.0.CO;2

[14] Anderson, J.L., Collins, N.: Scalable implementations of ensemble filter algorithms for data assimilation. Journal of Atmospheric and Oceanic Technology **24**(8), 1452–1463 (2007). DOI 10.1175/JTECH2049.1. URL https://doi.org/10.1175/JTECH2049.1

[15] Asch, M., Bocquet, M., Nodet, M.: Data Assimilation. Society for Industrial and Applied Mathematics, Philadelphia, PA (2016). DOI 10.1137/1.9781611974546. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611974546

[16] AVISO: SSALTO/DUACS User Handbook: (M)SLA and (M)ADT Near-Real Time and Delayed Time Products (2015). URL https://icdc.cen.uni-hamburg.de/fileadmin/user_upload/icdc_Dokumente/AVISO/hdbk_duacs.pdf

[17] Backeberg, B.C., Counillon, F., Johannessen, J.A., Pujol, M.I.: Assimilating along-track sla data using the enoi in an eddy resolving model of the agulhas system. Ocean Dynamics **64**(8), 1121–1136 (2014). DOI 10.1007/s10236-014-0717-6. URL https://doi.org/10.1007/s10236-014-0717-6

[18] Balmaseda, M.A., Mogensen, K., Weaver, A.T.: Evaluation of the ecmwf ocean reanalysis system oras4. Quarterly Journal of the Royal Meteorological Society **139**(674), 1132–1161 (2013). DOI 10.1002/qj.2063. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2063

[19] Bannister, R.N.: A review of operational methods of variational and ensemble-variational data assimilation. Quarterly Journal of the Royal Meteorological Society **143**(703), 607–633 (2017). DOI 10.1002/qj.2982. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2982

[20] Bazin, L., Landais, A., Lemieux-Dudon, B., Toyé Mahamadou Kele, H., Veres, D., Parrenin, F., Martinerie, P., Ritz, C., Capron, E., Lipenkov, V., Loutre, M.F., Raynaud, D., Vinther, B., Svensson, A., Rasmussen, S.O., Severi, M., Blunier, T., Leuenberger, M., Fischer, H., Masson-Delmotte, V., Chappellaz, J., Wolff, E.: An optimized multi-proxy, multi-site antarctic ice and gas orbital chronology (aicc2012): 120-800 ka. Climate of the Past **9**(4), 1715–1731 (2013). DOI 10.5194/cp-9-1715-2013. URL https://www.clim-past.net/9/1715/2013/

[21] Belyaev, K., Kuleshov, A., Tuchkova, N., Tanajura, C.A.: An optimal data assimilation method and its application to the numerical simulation of the ocean dynamics. Mathematical and Computer Modelling of Dynamical Systems **24**(1), 12–25 (2018). DOI 10.1080/13873954.2017.1338300. URL https://doi.org/10.1080/13873954.2017.1338300

[22] Berry, T., Harlim, J.: Forecasting turbulent modes with nonparametric diffusion models: Learning from noisy data. Physica D: Nonlinear Phenomena **320**,

57 – 76 (2016). DOI https://doi.org/10.1016/j.physd.2016.01.012. URL http://www.sciencedirect.com/science/article/pii/S0167278916000166

[23] Bocquet, M., Pires, C.A., Wu, L.: Beyond Gaussian Statistical Modeling in Geophysical Data Assimilation. Monthly Weather Review **138**(8), 2997–3023 (2010). DOI 10.1175/2010MWR3164.1. URL https://doi.org/10.1175/2010MWR3164.1

[24] Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D.H., Ebert, B., Fuentes, M., Hamill, T.M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.Y., Parsons, D., Raoult, B., Schuster, D., Dias, P.S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., Worley, S.: The thorpex interactive grand global ensemble. Bulletin of the American Meteorological Society **91**(8), 1059–1072 (2010). DOI 10.1175/2010BAMS2853.1. URL https://doi.org/10.1175/2010BAMS2853.1

[25] Bowler, N.E., Clayton, A.M., Jardak, M., Lee, E., Lorenc, A.C., Piccolo, C., Pring, S.R., Wlasak, M.A., Barker, D.M., Inverarity, G.W., Swinbank, R.: Inflation and localization tests in the development of an ensemble of 4d-ensemble variational assimilations. Quarterly Journal of the Royal Meteorological Society **143**(704), 1280–1302 (2017). DOI 10.1002/qj.3004. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3004

[26] Burgers, G., van Leeuwen, P., Evensen, G.: Analysis scheme in the ensemble Kalman filter. Monthly weather review **126**(6), 1719–1724 (1998)

[27] Carvalho, S., Aylagas, E., Villalobos, R., Kattan, Y., Berumen, M., Pearman, J.K.: Beyond the visual: using metabarcoding to characterize the hidden reef cryptobiome. Proceedings of the Royal Society B: Biological Sciences **286**(1896), 20182697 (2019). DOI 10.1098/rspb.2018.2697. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspb.2018.2697

[28] Cember, R.P.: On the sources, formation, and circulation of red sea deep water. Journal of Geophysical Research: Oceans **93**(C7), 8175–8191 (1988). DOI 10.1029/JC093iC07p08175. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC093iC07p08175

[29] Chen, Y., Cournède, P.H.: Data assimilation to reduce uncertainty of crop model prediction with convolution particle filtering. Ecological Modelling **290**,

165 – 177 (2014). DOI https://doi.org/10.1016/j.ecolmodel.2014.01.030. URL http://www.sciencedirect.com/science/article/pii/S0304380014000738. Special Issue of the 4th International Symposium on Plant Growth Modeling, Simulation, Visualization and Applications (PMA12)

[30] Cipollini, P., Beneviste, J., Bouffard, J., Emery, W., Fenoglio-Marc, L., Gommenginger, C., Griffin, D., Hoyer, J., Kurapov, A., Madsen, K., Mercier, F., Miller, L., Pascual, A., Ravichandran, M., Shillington, F., Snaith, H., Strub, T., Vandemark, D., Vignudelli, S., Wilkin, J., Woodworth, P., Zavala-Garay, J.: The role of altimetry in coastal observing systems (2010). URL https://eprints.soton.ac.uk/340378/. Actually deposited by J. Conquer

[31] Clifford, M., Horton, C., Schmitz, J., Kantha, L.H.: An oceanographic nowcast/forecast system for the red sea. Journal of Geophysical Research: Oceans **102**(C11), 25101–25122 (1997). DOI 10.1029/97JC01919. URL http://dx.doi.org/10.1029/97JC01919

[32] Counillon, F., Bertino, L.: Ensemble optimal interpolation: multivariate properties in the gulf of mexico. Tellus A **61**(2), 296–308 (2009). DOI 10.1111/j.1600-0870.2008.00383.x. URL http://dx.doi.org/10.1111/j.1600-0870.2008.00383.x

[33] Counillon, F., Sakov, P., Bertino, L.: Application of a hybrid enkf-oi to ocean forecasting. Ocean Science **5**(4), 389–401 (2009). DOI 10.5194/os-5-389-2009. URL https://www.ocean-sci.net/5/389/2009/

[34] Cox, M.D.: Isopycnal diffusion in a z-coordinate ocean model (1987)

[35] Danabasoglu, G., Mc Williams, J.C.: Sensitivity of the Global Ocean Circulation to Parameterizations of Mesoscale Tracer Transports. Journal of Climate **8**(12), 2967–2987 (1995). DOI 10.1175/1520-0442(1995)008⟨2967:SOTGOC⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0442(1995)008⟨2967:SOTGOC⟩2.0.CO;2

[36] Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato,

C., Thépaut, J.N., Vitart, F.: The era-interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society **137**(656), 553–597 (2011). DOI 10.1002/qj.828. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.828

[37] Dehwah, A.H., Jadoon, K.Z., Al-Mashharawi, S., Missimer, T.M.: Effects of nearshore evaporation rates on the design of seabed gallery intake systems for swro facilities located along the red sea shoreline of saudi arabia. Desalination and Water Treatment **57**(48-49), 22726–22733 (2016). DOI 10.1080/19443994.2015.1098796. URL https://doi.org/10.1080/19443994.2015.1098796

[38] Dehwah, A.H., Li, S., Al-Mashharawi, S., Winters, H., Missimer, T.M.: Changes in feedwater organic matter concentrations based on intake type and pretreatment processes at swro facilities, red sea, saudi arabia. Desalination **360**, 19 – 27 (2015). DOI https://doi.org/10.1016/j.desal.2015.01.008. URL http://www.sciencedirect.com/science/article/pii/S0011916415000120

[39] Dehwah, A.H.A., Missimer, T.M.: Technical feasibility of using gallery intakes for seawater RO facilities, northern Red Sea coast of Saudi Arabia: the King Abdullah Economic City site. Desalination and Water Treatment **51**(34-36), 6472–6481 (2013). DOI 10.1080/19443994.2013.770949. URL https://doi.org/10.1080/19443994.2013.770949

[40] Donlon, C.J., Martin, M., Stark, J., Roberts-Jones, J., Fiedler, E., Wimmer, W.: The operational sea surface temperature and sea ice analysis (ostia) system. Remote Sensing of Environment **116**, 140 – 158 (2012). DOI https://doi.org/10.1016/j.rse.2010.10.017. URL http://www.sciencedirect.com/science/article/pii/S0034425711002197. Advanced Along Track Scanning Radiometer(AATSR) Special Issue

[41] Dorigo, W., Zurita-Milla, R., de Wit, A., Brazile, J., Singh, R., Schaepman, M.: A review on reflective remote sensing and data assimilation techniques for enhanced agroecosystem modeling. International Journal of Applied Earth Observation and Geoinformation **9**(2), 165 – 193 (2007). DOI https://doi.org/10.1016/j.jag.2006.05.003. URL http://www.sciencedirect.com/science/article/pii/S0303243406000201. Advances in airborne electromagnetics and remote sensing of agro-ecosystems

[42] Dreano, D., Mallick, B., Hoteit, I.: Filtering remotely sensed chlorophyll concentrations in the Red Sea using a spacetime covariance model and a Kalman filter. Spatial Statistics **13**, 1 – 20 (2015). DOI https://doi.org/10.1016/j.spasta.2015.04.002. URL http://www.sciencedirect.com/science/article/pii/S2211675315000263

[43] Dreano, D., Raitsos, D.E., Gittings, J., Krokos, G., Hoteit, I.: The gulf of aden intermediate water intrusion regulates the southern red sea summer phytoplankton blooms. PloS one **11**(12), e0168440–e0168440 (2016). DOI 10.1371/journal.pone.0168440. URL https://www.ncbi.nlm.nih.gov/pubmed/28006006

[44] Dreano, D., Tsiaras, K., Triantafyllou, G., Hoteit, I.: Efficient ensemble forecasting of marine ecology with clustered 1d models and statistical lateral exchange: application to the red sea. Ocean Dynamics **67**(7), 935–947 (2017). DOI 10.1007/s10236-017-1065-0. URL https://doi.org/10.1007/s10236-017-1065-0

[45] Ducet, N., Le Traon, P.Y., Reverdin, G.: Global high-resolution mapping of ocean circulation from topex/poseidon and ers-1 and -2. Journal of Geophysical Research: Oceans **105**(C8), 19477–19498 (2000). DOI 10.1029/2000JC900063. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JC900063

[46] Edwards, C.A., Moore, A.M., Hoteit, I., Cornuelle, B.D.: Regional ocean data assimilation. Annual Review of Marine Science **7**(1), 21–42 (2015). DOI 10.1146/annurev-marine-010814-015821. URL https://doi.org/10.1146/annurev-marine-010814-015821. PMID: 25103331

[47] Elsheikh, A.H., Wheeler, M.F., Hoteit, I.: Clustered iterative stochastic ensemble method for multi-modal calibration of subsurface flow models. Journal of Hydrology **491**, 40 – 55 (2013). DOI https://doi.org/10.1016/j.jhydrol.2013.03.037. URL http://www.sciencedirect.com/science/article/pii/S0022169413002461

[48] Eshel, G., Naik, N.H.: Climatological coastal jet collision, intermediate water formation, and the general circulation of the red sea. Journal of Physical Oceanography **27**(7), 1233–1257 (1997). DOI 10.1175/1520-0485(1997)027⟨1233:CCJCIW⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1997)027⟨1233:CCJCIW⟩2.0.CO;2

[49] EUMETSAT/OSI-SAF: GHRSST Level 3P Global Subskin Sea Surface Temperature from the Advanced Very High Resolution Radiometer (AVHRR) on the MetOp-A satellite. https://doi.org/10.5067/GHGMT-3PE01 (2008). DOI https://doi.org/10.5067/GHGMT-3PE01. Ver. 1. PO.DAAC, CA, USA. Dataset accessed 2018.10.01

[50] Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. Journal of Geophysical Research: Oceans **99**(C5), 10143–10162 (1994). DOI 10.1029/94JC00572. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JC00572

[51] Evensen, G.: Ocean and Climate Prediction on Parallel Super Computers, pp. 37–37. Springer Berlin Heidelberg, Berlin, Heidelberg (2001). DOI 10.1007/3-540-70734-4_6. URL https://doi.org/10.1007/3-540-70734-4_6

[52] Evensen, G.: The Ensemble Kalman Filter: Theoretical formulation and practical implementation. Ocean Dynamics **53**(4), 343–367 (2003). DOI 10.1007/s10236-003-0036-9. URL http://link.springer.com/10.1007/s10236-003-0036-9

[53] Evensen, G.: Sampling strategies and square root analysis schemes for the enkf. Ocean Dynamics **54**(6), 539–560 (2004). DOI 10.1007/s10236-004-0099-2. URL https://doi.org/10.1007/s10236-004-0099-2

[54] Flowerdew, J.: Towards a theory of optimal localisation. Tellus A: Dynamic Meteorology and Oceanography **67**(1), 25257 (2015). DOI 10.3402/tellusa.v67.25257. URL https://doi.org/10.3402/tellusa.v67.25257

[55] Fofonoff, P., R. Millard, J.: Algorithms for computation of fundamental properties of seawater. Tech. Rep. Unesco Technical Papers in Marine Science 44, Unesco (1983)

[56] Forget, G., Wunsch, C.: Estimated Global Hydrographic Variability. Journal of Physical Oceanography **37**(8), 1997–2008 (2007). DOI 10.1175/JPO3072.1. URL https://doi.org/10.1175/JPO3072.1

[57] Fu, W., She, J., Zhuang, S.: Application of an ensemble optimal interpolation in a north/baltic sea model: Assimilating temperature and salinity profiles. Ocean Modelling **40**(3), 227 – 245 (2011). DOI https://doi.org/10.1016/

j.ocemod.2011.09.004. URL http://www.sciencedirect.com/science/article/pii/S1463500311001648

[58] Furrer, R., Bengtsson, T.: Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. Journal of Multivariate Analysis **98**(2), 227 – 255 (2007). DOI https://doi.org/10.1016/j.jmva.2006.08.003. URL http://www.sciencedirect.com/science/article/pii/S0047259X06001187

[59] Gaspar, P., Grgoris, Y., Lefevre, J.M.: A simple eddy kinetic energy model for simulations of the oceanic vertical mixing: Tests at station papa and long-term upper ocean study site. Journal of Geophysical Research: Oceans **95**(C9), 16179–16193 (1990). DOI 10.1029/JC095iC09p16179. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/JC095iC09p16179

[60] Gaspari, G., Cohn, S.E.: Construction of correlation functions in two and three dimensions. Quarterly Journal of the Royal Meteorological Society **125**(554), 723–757 (1999). DOI 10.1002/qj.49712555417. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.49712555417

[61] Gent, P.R., Mcwilliams, J.C.: Isopycnal Mixing in Ocean Circulation Models. Journal of Physical Oceanography **20**(1), 150–155 (1990). DOI 10.1175/1520-0485(1990)020⟨0150:IMIOCM⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1990)020⟨0150:IMIOCM⟩2.0.CO;2

[62] Gent, P.R., Willebrand, J., McDougall, T.J., McWilliams, J.C.: Parameterizing Eddy-Induced Tracer Transports in Ocean Circulation Models. Journal of Physical Oceanography **25**(4), 463–474 (1995). DOI 10.1175/1520-0485(1995)025⟨0463:PEITTI⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1995)025⟨0463:PEITTI⟩2.0.CO;2

[63] Gharamti, M., Ait-El-Fquih, B., Hoteit, I.: An iterative ensemble kalman filter with one-step-ahead smoothing for state-parameters estimation of contaminant transport models. Journal of Hydrology **527**, 442 – 457 (2015). DOI https://doi.org/10.1016/j.jhydrol.2015.05.004. URL http://www.sciencedirect.com/science/article/pii/S002216941500339X

[64] Gharamti, M., Hoteit, I.: Complex step-based low-rank extended kalman filtering for state-parameter estimation in subsurface transport models. Journal of Hydrology **509**, 588 – 600 (2014). DOI https://doi.org/10.1016/j.

jhydrol.2013.12.004. URL http://www.sciencedirect.com/science/article/pii/S0022169413008913

[65] Gharamti, M., Valstar, J., Hoteit, I.: An adaptive hybrid enkf-oi scheme for efficient state-parameter estimation of reactive contaminant transport models. Advances in Water Resources **71**, 1 – 15 (2014). DOI https://doi.org/10.1016/j.advwatres.2014.05.001. URL http://www.sciencedirect.com/science/article/pii/S030917081400089X

[66] Gharamti, M.E., Valstar, J., Janssen, G., Marsman, A., Hoteit, I.: On the efficiency of the hybrid and the exact second-order sampling formulations of the enkf: a reality-inspired 3-d test case for estimating biodegradation rates of chlorinated hydrocarbons at the port of rotterdam. Hydrology and Earth System Sciences **20**(11), 4561–4583 (2016). DOI 10.5194/hess-20-4561-2016. URL https://www.hydrol-earth-syst-sci.net/20/4561/2016/

[67] Gittings, J.A., Raitsos, D.E., Krokos, G., Hoteit, I.: Impacts of warming on phytoplankton abundance and phenology in a typical tropical marine ecosystem. Scientific Reports **8**(1), 2240 (2018). DOI 10.1038/s41598-018-20560-5. URL https://doi.org/10.1038/s41598-018-20560-5

[68] Good, S.A., Martin, M.J., Rayner, N.A.: En4: Quality controlled ocean temperature and salinity profiles and monthly objective analyses with uncertainty estimates. Journal of Geophysical Research: Oceans **118**(12), 6704–6716 (2013). DOI 10.1002/2013JC009067. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JC009067

[69] Gottwald, G.A.: Controlling balance in an ensemble kalman filter. Nonlinear Processes in Geophysics **21**(2), 417–426 (2014). DOI 10.5194/npg-21-417-2014. URL https://www.nonlin-processes-geophys.net/21/417/2014/

[70] Guo, D., Akylas, T.R., Zhan, P., Kartadikaria, A., Hoteit, I.: On the generation and evolution of internal solitary waves in the southern red sea. Journal of Geophysical Research: Oceans **121**(12), 8566–8584 (2016). DOI 10.1002/2016JC012221. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2016JC012221

[71] Guo, H.D., Zhang, L., Zhu, L.W.: Earth observation big data for climate change research. Advances in Climate Change Research **6**(2), 108 – 117 (2015). DOI https://doi.org/10.1016/j.accre.2015.09.007. URL http://www.sciencedirect.

com/science/article/pii/S1674927815000519. Special issue on advances in Future Earth research

[72] Hamill, T.M., Snyder, C.: A hybrid ensemble kalman filter3d variational analysis scheme. Monthly Weather Review **128**(8), 2905–2919 (2000). DOI 10.1175/1520-0493(2000)128⟨2905:AHEKFV⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2000)128⟨2905:AHEKFV⟩2.0.CO;2

[73] Hamill, T.M., Whitaker, J.S.: What constrains spread growth in forecasts initialized from ensemble kalman filters? Monthly Weather Review **139**(1), 117–131 (2011). DOI 10.1175/2010MWR3246.1. URL https://doi.org/10.1175/2010MWR3246.1

[74] Hamill, T.M., Whitaker, J.S., Snyder, C.: Distance-dependent filtering of background error covariance estimates in an ensemble kalman filter. Monthly Weather Review **129**(11), 2776–2790 (2001). DOI 10.1175/1520-0493(2001)129⟨2776:DDFOBE⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2001)129⟨2776:DDFOBE⟩2.0.CO;2

[75] Holland, W.R.: The Role of Mesoscale Eddies in the General Circulation of the Ocean?Numerical Experiments Using a Wind-Driven Quasi-Geostrophic Model. Journal of Physical Oceanography **8**(3), 363–392 (1978). DOI 10.1175/1520-0485(1978)008⟨0363:TROMEI⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1978)008⟨0363:TROMEI⟩2.0.CO;2

[76] Hoteit, I., Hoar, T., Gopalakrishnan, G., Collins, N., Anderson, J., Cornuelle, B., Köhl, A., Heimbach, P.: A mitgcm/dart ensemble analysis and prediction system with application to the gulf of mexico. Dynamics of Atmospheres and Oceans **63**, 1 – 23 (2013). DOI http://dx.doi.org/10.1016/j.dynatmoce.2013.03.002. URL http://www.sciencedirect.com/science/article/pii/S0377026513000249

[77] Hoteit, I., Luo, X., Bocquet, M., Köhl, A., Ait-El-Fquih, B.: Data Assimilation in Oceanography: Current Status and New Directions, chap. 17, pp. 465–512. GODAE OceanView (2018). DOI 10.17125/gov2018.ch17. URL https://doi.org/10.17125/gov2018.ch17

[78] Hoteit, I., Luo, X., Pham, D.T.: Particle kalman filtering: A nonlinear bayesian framework for ensemble kalman filters. Monthly Weather Review **140**(2), 528–

542 (2012). DOI 10.1175/2011MWR3640.1. URL https://doi.org/10.1175/2011MWR3640.1

[79] Hoteit, I., Pham, D.T.: An adaptively reduced-order extended kalman filter for data assimilation in the tropical pacific. Journal of Marine Systems **45**(3), 173 – 188 (2004). DOI https://doi.org/10.1016/j.jmarsys.2003.11.004. URL http://www.sciencedirect.com/science/article/pii/S0924796303001581. Marine Environmental Modelling 2001 - Selected Papers from the Fifth International Marine Environmental Modelling Seminar

[80] Hoteit, I., Pham, D.T., Blum, J.: A simplified reduced order kalman filtering and application to altimetric data assimilation in tropical pacific. Journal of Marine Systems **36**(1), 101 – 127 (2002). DOI https://doi.org/10.1016/S0924-7963(02)00129-X. URL http://www.sciencedirect.com/science/article/pii/S092479630200129X

[81] Hoteit, I., Pham, D.T., Gharamti, M.E., Luo, X.: Mitigating observation perturbation sampling errors in the stochastic enkf. Monthly Weather Review **143**(7), 2918–2936 (2015). DOI 10.1175/MWR-D-14-00088.1. URL https://doi.org/10.1175/MWR-D-14-00088.1

[82] Hoteit, I., Pham, D.T., Triantafyllou, G., Korres, G.: A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. Monthly Weather Review **136**(1), 317–334 (2008). DOI 10.1175/2007MWR1927.1. URL https://doi.org/10.1175/2007MWR1927.1

[83] Hoteit, I., Triantafyllou, G., Korres, G.: Using low-rank ensemble kalman filters for data assimilation with high dimensional imperfect models. JNAIAM. Journal of Numerical Analysis, Industrial and Applied Mathematics **2** (2007). URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.418.6672

[84] Houtekamer, P.L., Mitchell, H.L.: Data assimilation using an ensemble kalman filter technique. Monthly Weather Review **126**(3), 796–811 (1998). DOI 10.1175/1520-0493(1998)126⟨0796:DAUAEK⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1998)126⟨0796:DAUAEK⟩2.0.CO;2

[85] Houtekamer, P.L., Mitchell, H.L.: A sequential ensemble kalman filter for atmospheric data assimilation. Monthly Weather Review **129**(1), 123–137 (2001). DOI 10.1175/1520-0493(2001)129⟨0123:ASEKFF⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2001)129⟨0123:ASEKFF⟩2.0.CO;2

[86] Houtekamer, P.L., Mitchell, H.L.: Ensemble kalman filtering. Quarterly Journal of the Royal Meteorological Society **131**(613), 3269–3289 (2005). DOI 10.1256/qj.05.135. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.05.135

[87] Houtekamer, P.L., Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. Monthly Weather Review **144**(12), 4489–4532 (2016). DOI 10.1175/MWR-D-15-0440.1. URL https://doi.org/10.1175/MWR-D-15-0440.1

[88] Jackett, D.R., Mcdougall, T.J.: Minimal adjustment of hydrographic profiles to achieve static stability. Journal of Atmospheric and Oceanic Technology **12**(2), 381–389 (1995). DOI 10.1175/1520-0426(1995)012⟨0381:MAOHPT⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0426(1995)012⟨0381:MAOHPT⟩2.0.CO;2

[89] Jiang, Z., Chen, Z., Chen, J., Liu, J., Ren, J., Li, Z., Sun, L., Li, H.: Application of crop model data assimilation with a particle filter for estimating regional winter wheat yields. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **7**(11), 4422–4431 (2014). DOI 10.1109/JSTARS.2014.2316012

[90] Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., Wang, J.: A review of data assimilation of remote sensing and crop models. European Journal of Agronomy **92**, 141 – 152 (2018). DOI https://doi.org/10.1016/j.eja.2017.11.002. URL http://www.sciencedirect.com/science/article/pii/S1161030117301685

[91] Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering **82(1)**, 35–45 (1960). DOI http://dx.doi.org/10.1115/1.3662552

[92] Karspeck, A.R.: An Ensemble Approach for the Estimation of Observational Error Illustrated for a Nominal 1 Global Ocean Model. Monthly Weather Review **144**(5), 1713–1728 (2016). DOI 10.1175/MWR-D-14-00336.1. URL https://doi.org/10.1175/MWR-D-14-00336.1

[93] Khaki, M., Hamilton, F., Forootan, E., Hoteit, I., Awange, J., Kuhn, M.: Nonparametric data assimilation scheme for land hydrological applications. Water Resources Research **54**(7), 4946–4964 (2018). DOI 10.1029/2018WR022854. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018WR022854

[94] Köhl, A., Stammer, D.: Variability of the meridional overturning in the north atlantic from the 50-year gecco state estimation. Journal of Physical Oceanography **38**(9), 1913–1930 (2008). DOI 10.1175/2008JPO3775.1. URL https://doi.org/10.1175/2008JPO3775.1

[95] Kortas, S.: Decimate beta and development branches. https://github.com/samkos/decimate (2017)

[96] Kortas, S.: Decimate stable branch. https://github.com/KAUST-KSL/decimate (2017)

[97] Kortas, S.: Decimate documentation website. http://decimate.readthedocs.io/ (2018)

[98] Kortas, S.: Decimate python package webpage. https://pypi.python.org/pypi/decimate (2018)

[99] Langodan, S., Cavaleri, L., Viswanadhapalli, Y., Hoteit, I.: The red sea: A natural laboratory for wind and wave modeling. Journal of Physical Oceanography **44**(12), 3139–3159 (2014). DOI 10.1175/JPO-D-13-0242.1. URL https://doi.org/10.1175/JPO-D-13-0242.1

[100] Large, W.G., Danabasoglu, G., Doney, S.C., McWilliams, J.C.: Sensitivity to Surface Forcing and Boundary Layer Mixing in a Global Ocean Model: Annual-Mean Climatology. Journal of Physical Oceanography **27**(11), 2418–2447 (1997). DOI 10.1175/1520-0485(1997)027⟨2418:STSFAB⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1997)027⟨2418:STSFAB⟩2.0.CO;2

[101] Large, W.G., McWilliams, J.C., Doney, S.C.: Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization. Reviews of Geophysics **32**(4), 363–403 (1994). DOI 10.1029/94RG01872. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94RG01872

[102] Larsen, J., Hyer, J., She, J.: Validation of a hybrid optimal interpolation and kalman filter scheme for sea surface temperature assimilation. Journal of Marine Systems **65**(1), 122 – 133 (2007). DOI https://doi.org/10.1016/j.jmarsys.2005.09.013. URL http://www.sciencedirect.com/science/article/pii/S0924796306002880. Marine Environmental Monitoring and Prediction

[103] Le Dimet, F.X., Talagrand, O.: Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. Tellus A

38A(2), 97–110 (1986). DOI 10.1111/j.1600-0870.1986.tb00459.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0870.1986.tb00459.x

[104] Le Hénaff, M., Roblou, L., Bouffard, J.: Characterizing the navidad current interannual variability using coastal altimetry. Ocean Dynamics **61**(4), 425–437 (2011). DOI 10.1007/s10236-010-0360-9. URL https://doi.org/10.1007/s10236-010-0360-9

[105] Lee, Y., Majda, A.J., Qi, D.: Preventing Catastrophic Filter Divergence Using Adaptive Additive Inflation for Baroclinic Turbulence. Monthly Weather Review **145**(2), 669–682 (2017). DOI 10.1175/MWR-D-16-0121.1. URL https://doi.org/10.1175/MWR-D-16-0121.1

[106] Leith, C.: Stochastic models of chaotic systems. Physica D: Nonlinear Phenomena **98**(2), 481 – 491 (1996). DOI https://doi.org/10.1016/0167-2789(96)00107-8. URL http://www.sciencedirect.com/science/article/pii/0167278996001078. Nonlinear Phenomena in Ocean Dynamics

[107] Lemieux-Dudon, B., Bazin, L., Landais, A., Toyé Mahamadou Kele, H., Guillevic, M., Kindler, P., Parrenin, F., Martinerie, P.: Implementation of counted layers for coherent ice core chronology. Climate of the Past **11**(6), 959–978 (2015). DOI 10.5194/cp-11-959-2015. URL https://www.clim-past.net/11/959/2015/

[108] Lemieux-Dudon, B., Blayo, E., Petit, J.R., Waelbroeck, C., Svensson, A., Ritz, C., Barnola, J.M., Narcisi, B.M., Parrenin, F.: Consistent dating for antarctic and greenland ice cores. Quaternary Science Reviews **29**(1), 8 – 20 (2010). DOI https://doi.org/10.1016/j.quascirev.2009.11.010. URL http://www.sciencedirect.com/science/article/pii/S0277379109003734. Climate of the Last Million Years: New Insights from EPICA and Other Records

[109] Lermusiaux, P.F.J., Robinson, A.R.: Data assimilation via error subspace statistical estimation.part i: Theory and schemes. Monthly Weather Review **127**(7), 1385–1407 (1999). DOI 10.1175/1520-0493(1999)127⟨1385:DAVESS⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(1999)127⟨1385:DAVESS⟩2.0.CO;2

[110] Li, Y., Toumi, R.: A balanced kalman filter ocean data assimilation system with application to the south australian sea. Ocean Modelling **116**, 159 –

172 (2017). DOI https://doi.org/10.1016/j.ocemod.2017.06.007. URL http://www.sciencedirect.com/science/article/pii/S1463500317300963

[111] Liu, B., Gharamti, M., Hoteit, I.: Assessing clustering strategies for gaussian mixture filtering a subsurface contaminant model. Journal of Hydrology **535**, 1 – 21 (2016). DOI https://doi.org/10.1016/j.jhydrol.2016.01.048. URL http://www.sciencedirect.com/science/article/pii/S0022169416000664

[112] Lorenc, A.C.: The potential of the ensemble kalman filter for nwp-a comparison with 4d-var. Quarterly Journal of the Royal Meteorological Society **129**(595), 3183–3203 (2003). DOI 10.1256/qj.02.132. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.02.132

[113] Lorenc, A.C., Bowler, N.E., Clayton, A.M., Pring, S.R., Fairbairn, D.: Comparison of hybrid-4denvar and hybrid-4dvar data assimilation methods for global nwp. Monthly Weather Review **143**(1), 212–229 (2015). DOI 10.1175/MWR-D-14-00195.1. URL https://doi.org/10.1175/MWR-D-14-00195.1

[114] Lorenz, E.N.: Deterministic nonperiodic flow. Journal of the Atmospheric Sciences **20**(2), 130–141 (1963). DOI 10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0469(1963)020⟨0130:DNF⟩2.0.CO;2

[115] Lorenz, E.N., Emanuel, K.A.: Optimal sites for supplementary weather observations: Simulation with a small model. Journal of the Atmospheric Sciences **55**(3), 399–414 (1998). DOI 10.1175/1520-0469(1998)055⟨0399:OSFSWO⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0469(1998)055⟨0399:OSFSWO⟩2.0.CO;2

[116] Luo, X., Hoteit, I.: Covariance inflation in the ensemble kalman filter: A residual nudging perspective and some implications. Monthly Weather Review **141**(10), 3360–3368 (2013). DOI 10.1175/MWR-D-13-00067.1. URL https://doi.org/10.1175/MWR-D-13-00067.1

[117] Lyu, G., Wang, H., Zhu, J., Wang, D., Xie, J., Liu, G.: Assimilating the along-track sea level anomaly into the regional ocean modeling system using the ensemble optimal interpolation. Acta Oceanologica Sinica **33**(7), 72–82 (2014). DOI 10.1007/s13131-014-0469-7. URL https://doi.org/10.1007/s13131-014-0469-7

[118] Madsen, K.S., Hyer, J.L., Fu, W., Donlon, C.: Blending of satellite and tide gauge sea level observations and its assimilation in a storm surge model of the north sea and baltic sea. Journal of Geophysical Research: Oceans **120**(9), 6405–6418 (2015). DOI 10.1002/2015JC011070. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC011070

[119] Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Transactions on Signal Processing **41**(12), 3397–3415 (1993). DOI 10.1109/78.258082. URL https://doi.org/10.1109/78.258082

[120] Marotzke, J., Giering, R., Zhang, K.Q., Stammer, D., Hill, C., Lee, T.: Construction of the adjoint mit ocean general circulation model and application to atlantic heat transport sensitivity. Journal of Geophysical Research: Oceans **104**(C12), 29529–29547 (1999). DOI 10.1029/1999JC900236. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999JC900236

[121] Marshall, J., Adcroft, A., Campin, J.M., Hill, C., White, A.: Atmosphere-ocean modeling exploiting fluid isomorphisms. Monthly Weather Review **132**(12), 2882–2894 (2004). DOI 10.1175/MWR2835.1. URL https://doi.org/10.1175/MWR2835.1

[122] Marshall, J., Adcroft, A., Hill, C., Perelman, L., Heisey, C.: A finite-volume, incompressible navier stokes model for studies of the ocean on parallel computers. Journal of Geophysical Research: Oceans **102**(C3), 5753–5766 (1997). DOI 10.1029/96JC02775. URL http://dx.doi.org/10.1029/96JC02775

[123] Marshall, J., Jones, H., Hill, C.: Efficient ocean modeling using non-hydrostatic algorithms. Journal of Marine Systems **18**(1), 115 – 134 (1998). DOI https://doi.org/10.1016/S0924-7963(98)00008-6. URL http://www.sciencedirect.com/science/article/pii/S0924796398000086

[124] Martin, M., Balmaseda, M., Bertino, L., Brasseur, P., Brassington, G., Cummings, J., Fujii, Y., Lea, D., Lellouche, J.M., Mogensen, K., Oke, P., Smith, G., Testut, C.E., Waagb, G., Waters, J., Weaver, A.: Status and future of data assimilation in operational oceanography. Journal of Operational Oceanography **8**(sup1), s28–s48 (2015). DOI 10.1080/1755876X.2015.1022055. URL http://dx.doi.org/10.1080/1755876X.2015.1022055

[125] Mellor, G.L., Yamada, T.: Development of a turbulence closure model for geophysical fluid problems. Reviews of Geophysics **20**(4), 851–875 (1982). DOI

10.1029/RG020i004p00851. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/RG020i004p00851

[126] Mertz, F., Vinca, R., Caroline, M., Yannice, F.: Product user manual, For sea level SLA products. Technical Report CMEMS-SL-PUM-008-032-051, issue 1.1. (2017). URL http://cmems-resources.cls.fr/documents/PUM/CMEMS-SL-PUM-008-032-051.pdf

[127] Mitchell, H.L., Houtekamer, P.L., Pellerin, G.: Ensemble size, balance, and model-error representation in an ensemble kalman filter. Monthly Weather Review **130**(11), 2791–2808 (2002). DOI 10.1175/1520-0493(2002)130⟨2791:ESBAME⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2002)130⟨2791:ESBAME⟩2.0.CO;2

[128] Nerger, L., Danilov, S., Kivman, G., Hiller, W., Schröter, J.: Comparison of the ensemble kalman filter and the seik filter applied to a finite element model of the north atlantic. In: EGU 1st General Assembly, Nice, France, April 25 - 30 (2004). URL http://hdl.handle.net/10013/epic.20984

[129] Nerger, L., Hiller, W., Schröter, J.: PDAF - The parallel data assimilation framework: experiences with Kalman filtering, pp. 63–83. World Scientific (2005). DOI 10.1142/9789812701831_0006. URL https://www.worldscientific.com/doi/abs/10.1142/9789812701831_0006. 0

[130] Nerger, L., Schulte, S., Bunse-Gerstner, A.: On the influence of model nonlinearity and localization on ensemble kalman smoothing. Quarterly Journal of the Royal Meteorological Society **140**(684), 2249–2259 (2014). DOI 10.1002/qj.2293. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.2293

[131] Oke, P.R., Allen, J.S., Miller, R.N., Egbert, G.D., Kosro, P.M.: Assimilation of surface velocity data into a primitive equation coastal ocean model. Journal of Geophysical Research: Oceans **107**(C9), 5–1–5–25 (2002). DOI 10.1029/2000JC000511. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2000JC000511

[132] Oke, P.R., Brassington, G.B., Griffin, D.A., Schiller, A.: Ocean data assimilation: a case for ensemble optimal interpolation. Australian Meteorological and Oceanographic Journal **59**, 67–76 (2010)

[133] Oke, P.R., Sakov, P.: Representation error of oceanic observations for data assimilation. Journal of Atmospheric and Oceanic Technology **25**(6), 1004–1017 (2008). DOI 10.1175/2007JTECHO558.1. URL https://doi.org/10.1175/2007JTECHO558.1

[134] Oke, P.R., Sakov, P., Corney, S.P.: Impacts of localisation in the enkf and enoi: experiments with a small model. Ocean Dynamics **57**(1), 32–45 (2007). DOI 10.1007/s10236-006-0088-8. URL https://doi.org/10.1007/s10236-006-0088-8

[135] Oubei, H.M., Shen, C., Kammoun, A., Zedini, E., Park, K.H., Sun, X., Liu, G., Kang, C.H., Ng, T.K., Alouini, M.S., Ooi, B.S.: Light based underwater wireless communications. Japanese Journal of Applied Physics **57**(8S2), 08PA06 (2018). URL http://stacks.iop.org/1347-4065/57/i=8S2/a=08PA06

[136] Pacanowski, R.C., Philander, S.G.H.: Parameterization of Vertical Mixing in Numerical Models of Tropical Oceans. Journal of Physical Oceanography **11**(11), 1443–1451 (1981). DOI 10.1175/1520-0485(1981)011⟨1443:POVMIN⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1981)011⟨1443:POVMIN⟩2.0.CO;2

[137] Palmer, T.N.: Extended-range atmospheric prediction and the lorenz model. Bulletin of the American Meteorological Society **74**(1), 49–66 (1993). DOI 10.1175/1520-0477(1993)074⟨0049:ERAPAT⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0477(1993)074⟨0049:ERAPAT⟩2.0.CO;2

[138] Papadopoulos, V.P., Zhan, P., Sofianos, S.S., Raitsos, D.E., Qurban, M., Abualnaja, Y., Bower, A., Kontoyiannis, H., Pavlidou, A., Asharaf, T.T.M., Zarokanellos, N., Hoteit, I.: Factors governing the deep ventilation of the red sea. Journal of Geophysical Research: Oceans **120**(11), 7493–7505 (2015). DOI 10.1002/2015JC010996. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC010996

[139] Parent, L., Testut, C.E., Brankart, J.M., Verron, J., Brasseur, P., Gourdeau, L.: Comparative assimilation of topex/poseidon and ers altimeter data and of tao temperature data in the tropical pacific ocean during 19941998, and the mean sea-surface height issue. Journal of Marine Systems **40-41**, 381 – 401 (2003). DOI https://doi.org/10.1016/S0924-7963(03)00026-5. URL http://www.sciencedirect.com/science/article/pii/S0924796303000265. The Use of Data Assimilation in Coupled Hydrodynamic, Ecological and Bio-geo-chemical

Models of the Ocean. Selected papers from the 33rd International Liege Colloquium on Ocean Dynamics, held in Liege, Belgium on May 7-11th, 2001.

[140] Penny, S.G., Behringer, D.W., Carton, J.A., Kalnay, E.: A hybrid global ocean data assimilation system at ncep. Monthly Weather Review **143**(11), 4660–4677 (2015). DOI 10.1175/MWR-D-14-00376.1. URL https://doi.org/10.1175/MWR-D-14-00376.1

[141] Penny, S.G., Kalnay, E., Carton, J.A., Hunt, B.R., Ide, K., Miyoshi, T., Chepurin, G.A.: The local ensemble transform kalman filter and the running-in-place algorithm applied to a global ocean general circulation model. Nonlinear Processes in Geophysics **20**(6), 1031–1046 (2013). DOI 10.5194/npg-20-1031-2013. URL https://npg.copernicus.org/articles/20/1031/2013/

[142] Pham, D.T., Verron, J., Roubaud, M.C.: A singular evolutive extended kalman filter for data assimilation in oceanography. Journal of Marine Systems **16**(3), 323 – 340 (1998). DOI https://doi.org/10.1016/S0924-7963(97)00109-7. URL http://www.sciencedirect.com/science/article/pii/S0924796397001097

[143] Quadfasel D, B.H.: Gyre-scale circulation cells in the red-sea (1993). URL https://archimer.ifremer.fr/doc/00099/21049/

[144] Qurban, M.A., Krishnakumar, P., Joydas, T., Manikandan, K., Ashraf, T., Quadri, S., Wafar, M., Qasem, A., Cairns, S.: In-situ observation of deep water corals in the northern red sea waters of saudi arabia. Deep Sea Research Part I: Oceanographic Research Papers **89**, 35 – 43 (2014). DOI https://doi.org/10.1016/j.dsr.2014.04.002. URL http://www.sciencedirect.com/science/article/pii/S0967063714000521

[145] Raeder, K., Anderson, J.L., Collins, N., Hoar, T.J., Kay, J.E., Lauritzen, P.H., Pincus, R.: DART/CAM: An Ensemble Data Assimilation System for CESM Atmospheric Models. Journal of Climate **25**(18), 6304–6317 (2012). DOI 10.1175/JCLI-D-11-00395.1. URL https://doi.org/10.1175/JCLI-D-11-00395.1

[146] Raitsos, D., Pradhan, Y., Brewin, R., Stenchikov, G., Hoteit, I.: Remote sensing the phytoplankton seasonal succession of the red sea. PLoS ONE 8(6): e64909 (2013). DOI https://doi.org/10.1371/journal.pone.0064909. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0064909

[147] Redi, M.H.: Oceanic Isopycnal Mixing by Coordinate Rotation. Journal of Physical Oceanography **12**(10), 1154–1158 (1982). DOI 10.1175/1520-0485(1982)012⟨1154:OIMBCR⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0485(1982)012⟨1154:OIMBCR⟩2.0.CO;2

[148] Reichle, R.H., McLaughlin, D.B., Entekhabi, D.: Hydrologic data assimilation with the ensemble kalman filter. Monthly Weather Review **130**(1), 103–114 (2002). DOI 10.1175/1520-0493(2002)130⟨0103:HDAWTE⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2002)130⟨0103:HDAWTE⟩2.0.CO;2

[149] Reynolds, R.W., Smith, T.M., Liu, C., Chelton, D.B., Casey, K.S., Schlax, M.G.: Daily high-resolution-blended analyses for sea surface temperature. Journal of Climate **20**(22), 5473–5496 (2007). DOI 10.1175/2007JCLI1824.1. URL https://doi.org/10.1175/2007JCLI1824.1

[150] Richman, J.G., Miller, R.N., Spitz, Y.H.: Error estimates for assimilation of satellite sea surface temperature data in ocean climate models. Geophysical Research Letters **32**(18) (2005). DOI 10.1029/2005GL023591. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2005GL023591

[151] Ross, S.: - random variables. In: S. Ross (ed.) Introduction to Probability Models (Eleventh Edition), 11th edition edn., pp. 21 – 91. Academic Press, Boston (2014). DOI https://doi.org/10.1016/B978-0-12-407948-9.00002-5. URL http://www.sciencedirect.com/science/article/pii/B9780124079489000025

[152] Saeed, N., Celik, A., Al-Naffouri, T.Y., Alouini, M.: Underwater optical wireless communications, networking, and localization: A survey. CoRR **abs/1803.02442** (2018). URL http://arxiv.org/abs/1803.02442

[153] Sakov, P., Bertino, L.: Relation between two common localisation methods for the enkf. Computational Geosciences **15**(2), 225–237 (2011). DOI 10.1007/s10596-010-9202-6. URL https://doi.org/10.1007/s10596-010-9202-6

[154] Sakov, P., Oliver, D.S., Bertino, L.: An iterative enkf for strongly nonlinear systems. Monthly Weather Review **140**(6), 1988–2004 (2012). DOI 10.1175/MWR-D-11-00176.1. URL https://doi.org/10.1175/MWR-D-11-00176.1

[155] Sakov, P., Sandery, P.A.: Comparison of enoi and enkf regional ocean reanalysis systems. Ocean Modelling **89**, 45 – 60 (2015). DOI http://dx.doi.org/10.1016/

j.ocemod.2015.02.003. URL http://www.sciencedirect.com/science/article/pii/S1463500315000219

[156] Sana, F., Katterbauer, K., Al-Naffouri, T., Hoteit, I.: Orthogonal Matching Pursuit for Enhanced Recovery of Sparse Geological Structures with the Ensemble Kalman Filter. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **9**(4), 1710–1724 (2016). DOI 10.13140/RG.2.1.1803.8485. URL http://dx.doi.org/10.13140/RG.2.1.1803.8485

[157] Scharroo, R., Leuliette, E., Lillibridge, J., Byrne, D., Naeije, M., Mitchum, G.: RADS: Consistent Multi-Mission Products. In: 20 Years of Progress in Radar Altimatry, *ESA Special Publication*, vol. 710, p. 69 (2013). URL http://articles.adsabs.harvard.edu/pdf/2013ESASP.710E..69S

[158] Siddall, M., Smeed, D.A., Matthiesen, S., Rohling, E.J.: Modelling the seasonal cycle of the exchange flow in bab el mandab (red sea). Deep Sea Research Part I: Oceanographic Research Papers **49**(9), 1551 – 1569 (2002). DOI https://doi.org/10.1016/S0967-0637(02)00043-2. URL http://www.sciencedirect.com/science/article/pii/S0967063702000432

[159] Sivareddy, S., Banerjee, D.S., Baduru, B., Paul, B., Paul, A., Chakraborty, K., Hoteit, I.: Impact of dynamical representational errors on an indian ocean ensemble data assimilation system. Quarterly Journal of the Royal Meteorological Society **145**(725), 3680–3691 (2019). DOI 10.1002/qj.3649. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3649

[160] Sivareddy, S., Paul, A., Sluka, T., Ravichandran, M., Kalnay, E.: The pre-argo ocean reanalyses may be seriously affected by the spatial coverage of moored buoys. Scientific Reports **7**(1), 46685 (2017). DOI 10.1038/srep46685. URL https://doi.org/10.1038/srep46685

[161] Sivareddy, S., Toye, H., Zhan, P., Langodan, S., Krokos, G., Knio, O., Hoteit, I.: Impact of atmospheric and model physics perturbations on a high-resolution ensemble data assimilation system of the red sea. Journal of Geophysical Research: Oceans **125**(8), e2019JC015611 (2020). DOI 10.1029/2019JC015611. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015611. E2019JC015611 2019JC015611

[162] Smagorinsky, J.: Some historical remarks on the use of nonlinear viscosities. In: In: Galperin, B. and Orszag, S.A. (Eds.) Large Eddy Simulation of Com-

plex Engineering and Geophysical Flows, pp. 3–36. Cambridge University Press (1993). DOI 10.1017/S0022112095231768

[163] Sofianos, S.S., Johns, W.E.: An oceanic general circulation model (ogcm) investigation of the red sea circulation, 1. exchange between the red sea and the indian ocean. Journal of Geophysical Research: Oceans **107**(C11), 17–1–17–11 (2002). DOI 10.1029/2001JC001184. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2001JC001184

[164] Sofianos, S.S., Johns, W.E.: An oceanic general circulation model (ogcm) investigation of the red sea circulation: 2. three-dimensional circulation in the red sea. Journal of Geophysical Research: Oceans **108**(C3), n/a–n/a (2003). DOI 10.1029/2001JC001185. URL http://dx.doi.org/10.1029/2001JC001185. 3066

[165] Song, H., Hoteit, I., Cornuelle, B.D., Subramanian, A.C.: An adaptive approach to mitigate background covariance limitations in the ensemble kalman filter. Monthly Weather Review **138**(7), 2825–2845 (2010). DOI 10.1175/2010MWR2871.1. URL https://doi.org/10.1175/2010MWR2871.1

[166] Stark, J.D., Donlon, C.J., Martin, M.J., McCulloch, M.E.: Ostia : An operational, high resolution, real time, global sea surface temperature analysis system. In: OCEANS 2007 - Europe, pp. 1–4 (2007). DOI 10.1109/OCEANSE.2007.4302251. URL https://doi.org/10.1109/OCEANSE.2007.4302251

[167] Stewart, R.H.: Introduction To Physical Oceanography. Robert H. Stewart Department of Oceanography Texas A & M University (2008)

[168] Tandeo, P., Ailliot, P., Ruiz, J., Hannart, A., Chapron, B., Cuzol, A., Monbet, V., Easton, R., Fablet, R.: Combining Analog Method and Ensemble Data Assimilation: Application to the Lorenz-63 Chaotic System, pp. 3–12. Springer International Publishing, Cham (2015). DOI 10.1007/978-3-319-17220-0_1. URL http://dx.doi.org/10.1007/978-3-319-17220-0_1

[169] Tippett, M.K., Anderson, J.L., Bishop, C.H., Hamill, T.M., Whitaker, J.S.: Ensemble Square Root Filters*. Monthly Weather Review **131**(7), 1485–1490 (2003). DOI 10.1175/1520-0493(2003)131⟨1485:ESRF⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2003)131⟨1485:ESRF⟩2.0.CO;2

[170] Toye, H., Kortas, S., Zhan, P., Hoteit, I.: A fault-tolerant hpc scheduler extension for large and operational ensemble data assimilation: Application to the red sea. Journal of Computational Science **27**, 46 – 56 (2018). DOI https://doi.org/10.1016/j.jocs.2018.04.018. URL http://www.sciencedirect.com/science/article/pii/S1877750317312905

[171] Toye, H., Zhan, P., Gopalakrishnan, G., Kartadikaria, A.R., Huang, H., Knio, O., Hoteit, I.: Ensemble data assimilation in the red sea: Sensitivity to ensemble selection and atmospheric forcing. Ocean Dynamics **67**(7), 915–933 (2017). DOI 10.1007/s10236-017-1064-1. URL https://doi.org/10.1007/s10236-017-1064-1

[172] Tragou, E., Garrett, C.: The shallow thermohaline circulation of the red sea. Deep Sea Research Part I: Oceanographic Research Papers **44**(8), 1355 – 1376 (1997). DOI https://doi.org/10.1016/S0967-0637(97)00026-5. URL http://www.sciencedirect.com/science/article/pii/S0967063797000265

[173] Triantafyllou, G., Yao, F., Petihakis, G., Tsiaras, K.P., Raitsos, D.E., Hoteit, I.: Exploring the red sea seasonal ecosystem functioning using a three-dimensional biophysical model. Journal of Geophysical Research: Oceans **119**(3), 1791–1811 (2014). DOI 10.1002/2013JC009641. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JC009641

[174] Tropp, J.A., Gilbert, A.C.: Signal recovery from random measurements via orthogonal matching pursuit. Information Theory, IEEE Transactions on **53**(12), 4655–4666 (2007)

[175] Tseng, K., Shum, C.K., Yi, Y., Emery, W.J., Kuo, C., Lee, H., Wang, H.: The improved retrieval of coastal sea surface heights by retracking modified radar altimetry waveforms. IEEE Transactions on Geoscience and Remote Sensing **52**(2), 991–1001 (2014). DOI 10.1109/TGRS.2013.2246572

[176] Tsiaras, K.P., Hoteit, I., Kalaroni, S., Petihakis, G., Triantafyllou, G.: A hybrid ensemble-oi kalman filter for efficient data assimilation into a 3-d biogeochemical model of the mediterranean. Ocean Dynamics **67**(6), 673–690 (2017). DOI 10.1007/s10236-017-1050-7. URL https://doi.org/10.1007/s10236-017-1050-7

[177] Veres, D., Bazin, L., Landais, A., Toyé Mahamadou Kele, H., Lemieux-Dudon, B., Parrenin, F., Martinerie, P., Blayo, E., Blunier, T., Capron, E., Chappellaz,

J., Rasmussen, S.O., Severi, M., Svensson, A., Vinther, B., Wolff, E.W.: The antarctic ice core chronology (aicc2012): an optimized multi-parameter and multi-site dating approach for the last 120 thousand years. Climate of the Past **9**(4), 1733–1748 (2013). DOI 10.5194/cp-9-1733-2013. URL https://www.clim-past.net/9/1733/2013/

[178] Verlaan, M., Heemink, A.W.: Tidal flow forecasting using reduced rank square root filters. Stochastic Hydrology and Hydraulics **11**(5), 349–368 (1997). DOI 10.1007/BF02427924. URL https://doi.org/10.1007/BF02427924

[179] Vervatis, V., Testut, C., Mey, P.D., Ayoub, N., Chanut, J., Quattrocchi, G.: Data assimilative twin-experiment in a high-resolution bay of biscay configuration: 4denoi based on stochastic modeling of the wind forcing. Ocean Modelling **100**, 1 – 19 (2016). DOI https://doi.org/10.1016/j.ocemod.2016.01.003. URL http://www.sciencedirect.com/science/article/pii/S1463500316000044

[180] Vignudelli, S., Kostianoy, A., Cipollini, P., Benveniste, J.: Coastal Altimetry, vol. 1 (2011). DOI 10.1007/978-3-642-12796-0. URL https://link.springer.com/book/10.1007%2F978-3-642-12796-0

[181] Viswanadhapalli, Y., Dasari, H.P., Langodan, S., Challa, V.S., Hoteit, I.: Climatic features of the red sea from a regional assimilative model. International Journal of Climatology **37**(5), 2563–2581 (2017). DOI 10.1002/joc.4865. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/joc.4865

[182] Wan, L., Zhu, J., Wang, H., Yan, C., Bertino, L.: A ”dressed” ensemble kalman filter using the hybrid coordinate ocean model in the pacific. Advances in Atmospheric Sciences **26**(5), 1042–1052 (2009). DOI 10.1007/s00376-009-7208-6. URL https://doi.org/10.1007/s00376-009-7208-6

[183] Wang, X., Hamill, T.M., Whitaker, J.S., Bishop, C.H.: A comparison of hybrid ensemble transform kalman filter-optimum interpolation and ensemble square root filter analysis schemes. Monthly Weather Review **135**(3), 1055–1076 (2007). DOI 10.1175/MWR3307.1. URL https://doi.org/10.1175/MWR3307.1

[184] Whitaker, J.S., Hamill, T.M.: Ensemble data assimilation without perturbed observations. Monthly Weather Review **130**(7), 1913–1924 (2002). DOI 10.1175/1520-0493(2002)130⟨1913:EDAWPO⟩2.0.CO;2. URL https://doi.org/10.1175/1520-0493(2002)130⟨1913:EDAWPO⟩2.0.CO;2

[185] Xie, J., Counillon, F., Zhu, J., Bertino, L.: An eddy resolving tidal-driven model of the south china sea assimilating along-track sla data using the enoi. Ocean Science **7**(5), 609–627 (2011). DOI 10.5194/os-7-609-2011. URL https://www.ocean-sci.net/7/609/2011/

[186] Xie, J., Zhu, J.: Ensemble optimal interpolation schemes for assimilating Argo profiles into a hybrid coordinate ocean model. Ocean Modelling **33**, 283–298 (2010). DOI 10.1016/j.ocemod.2010.03.002. URL https://doi.org/10.1016/j.ocemod.2010.03.002

[187] Yang, L., Lin, M., Liu, Q., Pan, D.: A coastal altimetry retracking strategy based on waveform classification and sub-waveform extraction. International Journal of Remote Sensing **33**(24), 7806–7819 (2012). DOI 10.1080/01431161.2012.701350. URL https://doi.org/10.1080/01431161.2012.701350

[188] Yao, F., Hoteit, I.: Rapid red sea deep water renewals caused by volcanic eruptions and the north atlantic oscillation. Science Advances **4**(6) (2018). DOI 10.1126/sciadv.aar5637. URL http://advances.sciencemag.org/content/4/6/eaar5637

[189] Yao, F., Hoteit, I., Pratt, L.J., Bower, A.S., Köhl, A., Gopalakrishnan, G., Rivas, D.: Seasonal overturning circulation in the red sea: 2. winter circulation. Journal of Geophysical Research: Oceans **119**(4), 2263–2289 (2014). DOI 10.1002/2013JC009331. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JC009331

[190] Yao, F., Hoteit, I., Pratt, L.J., Bower, A.S., Zhai, P., Köhl, A., Gopalakrishnan, G.: Seasonal overturning circulation in the red sea: 1. model validation and summer circulation. Journal of Geophysical Research: Oceans **119**(4), 2238–2262 (2014). DOI 10.1002/2013JC009004. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JC009004

[191] Zhai, P., Bower, A.: The response of the red sea to a strong wind jet near the tokar gap in summer. Journal of Geophysical Research: Oceans **118**(1), 421–434 (2013). DOI 10.1029/2012JC008444. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2012JC008444

[192] Zhai, P., Bower, A.S., Smethie Jr., W.M., Pratt, L.J.: Formation and spreading of red sea outflow water in the red sea. Journal of Geophysical Research: Oceans

120(9), 6542–6563 (2015). DOI 10.1002/2015JC010751. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC010751

[193] Zhan, P., Gopalakrishnan, G., Subramanian, A.C., Guo, D., Hoteit, I.: Sensitivity studies of the red sea eddies using adjoint method. Journal of Geophysical Research: Oceans **123**(11), 8329–8345 (2018). DOI 10.1029/2018JC014531. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018JC014531

[194] Zhan, P., Krokos, G., Guo, D., Hoteit, I.: Three-dimensional signature of the red sea eddies and eddy-induced transport. Geophysical Research Letters **0**(0) (2019). DOI 10.1029/2018GL081387. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GL081387

[195] Zhan, P., Subramanian, A.C., Yao, F., Hoteit, I.: Eddies in the red sea: A statistical and dynamical study. Journal of Geophysical Research: Oceans **119**(6), 3909–3925 (2014). DOI 10.1002/2013JC009563. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2013JC009563

[196] Zhan, P., Subramanian, A.C., Yao, F., Kartadikaria, A.R., Guo, D., Hoteit, I.: The eddy kinetic energy budget in the red sea. Journal of Geophysical Research: Oceans **121**(7), 4732–4747 (2016). DOI 10.1002/2015JC011589. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015JC011589

[197] Zuo, H., Balmaseda, M.A., Tietsche, S., Mogensen, K., Mayer, M.: The ecmwf operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. Ocean Science **15**(3), 779–808 (2019). DOI 10.5194/os-15-779-2019. URL https://os.copernicus.org/articles/15/779/2019/

APPENDICES

A Relationship between $\mathbf{x}^t$, $\mathbf{x}$ and $\mathbf{y}^o$

To discuss the relationship between $\mathbf{x}^t$, $\mathbf{x}$, and $\mathbf{y}^o$, let $\mathcal{S}$ be the state space containing $\mathbf{x}^t$. To handle and manipulate $\mathbf{x}^t$, we define a discrete representation of it, i.e. a model, even though we shall also use the term (dynamical) model later for the map that evolves the state in time. Therefore, for $n \in \mathbb{N}$, let $\mathcal{S}_n$ be the state space of dimension n representing $\mathcal{S}$ at n grid points, and let our estimate $\mathbf{x}$ of $\mathbf{x}^t$ be an element of $\mathcal{S}_n$. If we define $\mathcal{T}_{\mathcal{S}}$ to be the map that assigns to $\mathbf{x}^t \in \mathcal{S}$ its restriction to the space $\mathcal{S}_n$,

$$\begin{aligned} \mathcal{T}_{\mathcal{S}} \colon \mathcal{S} &\longrightarrow \mathcal{S}_n \\ \mathbf{x}^t &\longmapsto \mathcal{T}_{\mathcal{S}}(\mathbf{x}^t) := \mathbf{x}^t_{|\mathcal{S}_n}, \end{aligned} \tag{A.1}$$

then $\boldsymbol{\xi} := \mathcal{T}_{\mathcal{S}}(\mathbf{x}^t) - \mathbf{x}$ is the error in estimating $\mathbf{x}^t$ by $\mathbf{x}$ in the space $\mathcal{S}_n$. To make the link between $\mathbf{x}^t$ and $\mathbf{y}^o$, we define $\mathcal{O}_p$ as the observation state of dimension p containing $\mathbf{y}^o$ and $\mathcal{O}$ as the observation space that will contain a representation of $\mathbf{x}^t$, referred as $\mathbf{y}^t$, that can be compared to $\mathbf{y}^o$. We also define the map $h_{\mathcal{S}\to\mathcal{O}}$ to relate $\mathbf{x}^t$ to $\mathbf{y}^t$:

$$\begin{aligned} h_{\mathcal{S}\to\mathcal{O}} \colon \mathcal{S} &\longrightarrow \mathcal{O} \\ \mathbf{x}^t &\longmapsto \mathbf{y}^t = h_{\mathcal{S}\to\mathcal{O}}(\mathbf{x}^t), \end{aligned} \tag{A.2}$$

and, similarly to $\mathcal{T}_{\mathcal{S}}$, we define $\mathcal{T}_{\mathcal{O}}$ to be the map from the infinite dimensional space

$\mathcal{O}$ to the p-dimensional space $\mathcal{O}_p$ that associates to $\mathbf{y}^t \in \mathcal{O}$ its restriction to $\mathcal{O}_p$:

$$\begin{aligned} \mathcal{T}_{\mathcal{O}} \colon \mathcal{O} &\longrightarrow \mathcal{O}_p \\ \mathbf{y}^t &\longmapsto \mathcal{T}_{\mathcal{O}}(\mathbf{y}^t) := \mathbf{y}^t_{|O_p}. \end{aligned} \tag{A.3}$$

Therefore the relation between $\mathbf{x}^t$ and $\mathbf{y}^o$ is summarized by the error $\boldsymbol{\varepsilon} = \mathbf{y}^o - \mathcal{T}_{\mathcal{O}}(\mathbf{y}^t) = \mathbf{y}^o - \mathcal{T}_{\mathcal{O}}(h_{\mathcal{S}\to\mathcal{O}}(\mathbf{x}^t))$. To establish a connection between $\mathbf{x}$ and $\mathbf{y}^o$, we introduce an observation operator $h_{\mathcal{S}_n\to\mathcal{O}_p}$ that sends $\mathbf{x} \in \mathcal{S}_n$ to the space $\mathcal{O}_p$:

$$\begin{aligned} h_{\mathcal{S}_n\to\mathcal{O}_p} \colon \mathcal{S}_n &\longrightarrow \mathcal{O}_p \\ \mathbf{x} &\longmapsto \mathbf{y}^p = h_{\mathcal{S}_n\to\mathcal{O}_p}(\mathbf{x}). \end{aligned} \tag{A.4}$$

$\mathbf{x}$ is then connected to $\mathbf{y}^o$ by the error $\boldsymbol{\delta} = \mathbf{y}^o - h_{\mathcal{S}_n\to\mathcal{O}_p}(\mathbf{x}) = \mathbf{y}^o - \mathbf{y}^p$. $\boldsymbol{\delta}$ is called the innovation and accounts for the mismatch between what we observe and what we predict, i.e. the difference between the observation $\mathbf{y}^o$ and the prediction $\mathbf{y}^p$. The following commutative diagram summarizes the discussed spaces:

$$\begin{tikzcd} \mathcal{S} \arrow[r, "\mathcal{T}_{\mathcal{S}}"] \arrow[d, "h_{\mathcal{S}\to\mathcal{O}}"'] & \mathcal{S}_n \arrow[d, "h_{\mathcal{S}_n\to\mathcal{O}_p}"] \\ \mathcal{O} \arrow[r, "\mathcal{T}_{\mathcal{O}}"] & \mathcal{O}_p \end{tikzcd} \tag{A.5}$$

All the previous interactions have been considered at a given time t_k, for $k \in \mathbb{N}$. To deal with the evolving aspect of the system state, we assign a time component to our system and define a dynamical model that will advance the system forward in time. Let $\mathcal{S}(t_k)$ and $\mathcal{S}(t_{k+1}), \forall k \in \mathbb{N}$, be our state space at time t_k and t_{k+1}, respectively. The same way, we define $\mathcal{O}(t_k)$ to be the observation space at time t_k, $\mathcal{S}_{n_{t_k}}(t_k)$ to be the state space of dimension n_{t_k} at time t_k, $\mathcal{O}_{p_{t_k}}(t_k)$ to be the observation space of dimension p_{t_k} at time t_k. Assuming that $\mathcal{S}$ and $\mathcal{O}$ are not changing over time, we can make some simplifications: $\mathcal{S} := \mathcal{S}(t_k)$ and $\mathcal{O} := \mathcal{O}(t_k), \forall k \in \mathbb{N}$. We also assume that the dimension of the discrete state space $\mathcal{S}_{n_{t_k}}$ is constant such that $\mathcal{S}_n(t_k) := \mathcal{S}_{n_{t_k}}(t_k)$. Unlike $\mathcal{S}_{n_{t_k}}$, the dimension p_{t_k} of $\mathcal{O}_{p_{t_k}}$ changes over time in our application. Let $\mathcal{G}_{t_k}$ be the true dynamical geophysical model that advances the true state in time:

$$\begin{aligned} \mathcal{G}_{t_k} &: \mathcal{S} \longrightarrow \mathcal{S} \\ \mathbf{x}^t(t_k) &\longmapsto \mathbf{x}^t(t_{k+1}) = \mathcal{G}_{t_k}(\mathbf{x}^t(t_k)). \end{aligned} \tag{A.6}$$

$\mathcal{G}_{t_k}$ is unknown and we therefore turn to a representation $\mathcal{M}_{t_k}$ that we expect to capture most of the physics involved in $\mathcal{G}_{t_k}$, in order to minimize the modelling errors:

$$\begin{aligned} \mathcal{M}_{t_k} &: \mathcal{S}_n(t_k) \longrightarrow \mathcal{S}_n(t_{k+1}) \\ \mathbf{x}(t_k) &\longmapsto \mathbf{x}(t_{k+1}) = \mathcal{M}_{t_k}(\mathbf{x}(t_k)). \end{aligned} \tag{A.7}$$

$\mathcal{M}_{t_k}$ operates on the discrete spaces $\mathcal{S}_n(t_k)$ and $\mathcal{S}_n(t_{k+1})$. To apply $\mathcal{M}_{t_k}$ in the place of $\mathcal{G}_{t_k}$ to the true state $\mathbf{x}^t$, we first send it in $\mathcal{S}_n(t_k)$ through $\mathcal{T}_{\mathcal{S}}$. Then we compute $\mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k)))$ which is an approximation of $\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_{k+1})) = \mathcal{T}_{\mathcal{S}}(\mathcal{G}_{t_k}(\mathbf{x}^t(t_k)))$ in the space $\mathcal{S}_n(t_{k+1})$, and we define the corresponding error as the model error

$$\begin{aligned} \boldsymbol{\eta}_{t_k} &:= \mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_{k+1})) - \mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k))) \\ &:= \mathcal{T}_{\mathcal{S}}(\mathcal{G}_{t_k}(\mathbf{x}^t(t_k))) - \mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k))). \end{aligned} \tag{A.8}$$

Notice that $\mathcal{T}_{\mathcal{S}}$ in $\mathcal{T}_{\mathcal{S}}(\mathcal{G}_{t_k}(\mathbf{x}^t(t_k)))$ is $\mathcal{T}_{\mathcal{S}} : \mathcal{S} \longrightarrow \mathcal{S}_n(t_{k+1})$ and the one in $\mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k)))$ is $\mathcal{T}_{\mathcal{S}} : \mathcal{S} \longrightarrow \mathcal{S}_n(t_k)$. Strictly speaking, we should use different notations for each of them, but we make use of $\mathcal{T}_{\mathcal{S}}$ for both since the purpose is to evaluate the variables in the discrete spaces, and also to avoid carrying many additional notations. We already know (cf. page 213) that $\boldsymbol{\xi} = \mathcal{T}_{\mathcal{S}}(\mathbf{x}^t) - \mathbf{x}$ is the error in estimating $\mathbf{x}^t$ by $\mathbf{x}$ in the space $\mathcal{S}_n$. Adding the time index, $\boldsymbol{\xi}_{t_k} := \mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k)) - \mathbf{x}(t_k)$ is the estimation error in $\mathcal{S}_n(t_k)$ and $\boldsymbol{\xi}_{t_{k+1}} := \mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_{k+1})) - \mathbf{x}(t_{k+1})$ the one in $\mathcal{S}_n(t_{k+1})$. From Eq. (A.8),

$\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_{k+1})) = \mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k))) + \boldsymbol{\eta}_{t_k}$ and $\mathbf{x}(t_{k+1}) = \mathcal{M}_{t_k}(\mathbf{x}(t_k))$ by definition of $\mathcal{M}_{t_k}$. So $\boldsymbol{\xi}_{t_{k+1}} = \mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k))) + \boldsymbol{\eta}_{t_k} - \mathcal{M}_{t_k}(\mathbf{x}(t_k))$ and if moreover $\mathcal{M}_{t_k}$ is linear then

$$\begin{aligned}
\boldsymbol{\xi}_{t_{k+1}} &= \mathcal{M}_{t_k}(\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k))) - \mathcal{M}_{t_k}(\mathbf{x}(t_k)) + \boldsymbol{\eta}_{t_k} \\
&= \mathcal{M}_{t_k}(\underbrace{\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t(t_k)) - \mathbf{x}(t_k)}_{=\boldsymbol{\xi}_{t_k}}) + \boldsymbol{\eta}_{t_k} \\
&= \mathcal{M}_{t_k}(\boldsymbol{\xi}_k) + \boldsymbol{\eta}_{t_k}
\end{aligned}$$

In case $\mathcal{M}_{t_k}$ is nonlinear, letting $\mathbf{M}_{t_k}$ be the linearization of $\mathcal{M}_{t_k}$ around $\mathbf{x}(t_k)$, we should add the higher-order terms ($h.o.t.$) of the Taylor expansion of $\mathcal{M}_{t_k}$ in the previous expression of the error:

$$\boldsymbol{\xi}_{t_{k+1}} = \mathbf{M}_{t_k}(\boldsymbol{\xi}_{t_k}) + \boldsymbol{\eta}_{t_k} + h.o.t., \tag{A.9}$$

and obtain the equation governing the errors propagation in time.

We shall also update the observation operators notations. For $h_{\mathcal{S}\to\mathcal{O}}$ there is nothing to do because we assumed $\mathcal{S}$ and $\mathcal{O}$ to be constant over time. As for $h_{\mathcal{S}_n\to\mathcal{O}_p}$, we get $h_{\mathcal{S}_n(t_k)\to\mathcal{O}_{p_{t_k}}(t_k)}$ at time t_k and $h_{\mathcal{S}_n(t_{k+1})\to\mathcal{O}_{p_{t_{k+1}}}(t_{k+1})}$ at time t_{k+1}. We simplify $h_{\mathcal{S}_n(t_k)\to\mathcal{O}_{p_{t_k}}(t_k)}$ as h_{t_k} and $h_{\mathcal{S}_n(t_{k+1})\to\mathcal{O}_{p_{t_{k+1}}}(t_{k+1})}$ as $h_{t_{k+1}}$. A summary of the spaces and variables is given in Table A.1.

Now, let us assume $t_k = k$, $\forall k \in \mathbb{N}$ and $\mathcal{T}_{\mathcal{S}}(\mathbf{x}^t) = \mathbf{x}^t$ by identifying $\mathbf{x}^t$ with its restriction to the space $\mathcal{S}_n$, but keeping in mind that it is an element of $\mathcal{S}_n$. We also place the time variable as subscript in the variables' notations. For example,

Table A.1: Spaces and variables at time t_k and t_{k+1}.

$\mathcal{M}_{t_k}$

$\mathcal{S}$	$\mathcal{S}_n(t_k)$	$\mathcal{S}$	$\mathcal{S}_n(t_{k+1})$
$\mathbf{x}^t(t_k)$	$\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_k))$	$\mathbf{x}^t(t_{k+1})$	$\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_{k+1})) = \mathcal{M}_{t_k}(\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_k))) + \boldsymbol{\eta}_{t_k}$
	$\mathbf{x}(t_k)$		$\mathbf{x}(t_{k+1}) = \mathcal{M}_{t_k}(\mathbf{x}(t_k))$
$\mathcal{O}$	$\mathcal{O}_{p_{t_k}}(t_k)$	$\mathcal{O}$	$\mathcal{O}_{p_{t_{k+1}}}(t_{k+1})$
$\mathbf{y}^t(t_k)$	$h_{t_k}(\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_k)))$	$\mathbf{y}^t(t_{k+1})$	$h_{t_{k+1}}(\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_{k+1})))$
	$h_{t_k}(\mathbf{x}(t_k))$		$h_{t_{k+1}}(\mathbf{x}(t_{k+1}))$
	$\mathbf{y}^o(t_k)$		$\mathbf{y}^o(t_{k+1})$

$\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_{k+1})) = \mathcal{M}_{t_k}(\mathcal{T}_\mathcal{S}(\mathbf{x}^t(t_k))) + \boldsymbol{\eta}_{t_k}$ now becomes $\mathbf{x}^t_{k+1} = \mathcal{M}_k(\mathbf{x}^t_k) + \boldsymbol{\eta}_k$.

B Construction of an adjustment matrix $\mathbf{A}_k$

Let us consider a sample of N members $\{\mathbf{z}_k^{f,i}\}_{i=1,\ldots,N}$ from the prior distribution with sample mean $\overline{\mathbf{z}}_k^f$ and sample covariance $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_k^f$, and assume the quantities $\mathbf{H}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{y}_k^o$ and $\mathbf{H}_k^{\mathrm{T}}\mathbf{R}_k^{-1}\mathbf{H}_k$ are readily available at the assimilation time k, when observations become available. For simplification, the time subscript k will be dropped from here.

Notice that $\boldsymbol{\Sigma}$ and $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ are symmetric positive semi-definite matrices. Therefore their eigenvalue decompositions and their singular value decomposition (SVD) can be computed such that they match. The idea is to rewrite equation (2.13) in the form $\boldsymbol{\Sigma}^a = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}}$ using successive changes of basis by applying SVDs. Indeed, the estimated updated covariance is given by

$$\mathbf{\Sigma}^{a,e} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{z}^{a,i} - \overline{\mathbf{z}}^{a})(\mathbf{z}^{a,i} - \overline{\mathbf{z}}^{a})^{T} \tag{B.1}$$

Then by replacing $\mathbf{z}^{a,i}$ by its expression from (2.15),

$$\begin{aligned}
\mathbf{\Sigma}^{a,e} &= \frac{1}{N-1}\sum_{i=1}^{N}(\underline{\mathbf{A}(\mathbf{z}^{f,i} - \overline{\mathbf{z}}^{f}) + \overline{\mathbf{z}}^{a}} - \overline{\mathbf{z}}^{a})(\underline{\mathbf{A}(\mathbf{z}^{f,i} - \overline{\mathbf{z}}^{f}) + \overline{\mathbf{z}}^{a}} - \overline{\mathbf{z}}^{a})^{T} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{A}(\mathbf{z}^{f,i} - \overline{\mathbf{z}}^{f})(\mathbf{z}^{f,i} - \overline{\mathbf{z}}^{f})^{T}\mathbf{A}^{T}\right) \\
&= \mathbf{A}\left(\underbrace{\frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{z}^{f,i} - \overline{\mathbf{z}}^{f})(\mathbf{z}^{f,i} - \overline{\mathbf{z}}^{f})^{T}}_{=\mathbf{\Sigma}^{f,e}}\right)\mathbf{A}^{T} \\
&= \mathbf{A}\mathbf{\Sigma}^{f,e}\mathbf{A}^{T} \qquad \text{(B.2)}
\end{aligned}$$

Now, let us give the details of the construction.

A SVD of $\mathbf{\Sigma}$ gives:

$$\mathbf{\Sigma} = \mathbf{F}\mathbf{D}^{f}\mathbf{F}^{\mathrm{T}} \tag{B.3}$$

with $\mathbf{F}$ a unitary matrix ($\mathbf{F}^{\mathrm{T}}\mathbf{F} = \mathbf{I}$, $\mathbf{F}^{-1} = \mathbf{F}^{\mathrm{T}}$, $(\mathbf{F}^{\mathrm{T}})^{-1} = \mathbf{F}$), and $\mathbf{D}^{f}$ a diagonal matrix having the singular values of $\mathbf{\Sigma}$ (μ^{f}) on its diagonal because, as previously stated, $\mathbf{\Sigma}$ is symmetric and positive semi-definite. From (B.3),

$$\mathbf{D}^{f} = \mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F} \tag{B.4}$$

By defining $\mathbf{G}$ as the diagonal matrix having the square root of the singular values of $\mathbf{\Sigma}$ ($\sqrt{\mu^f}$) on its diagonal, we get

$$(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1} = \mathbf{I} \tag{B.5}$$

where $\mathbf{I}$ is the identity matrix.

Since we should work on $\mathbf{\Sigma}^{-1}$ and $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ in (2.13) at the same time, we apply a SVD on $\mathbf{G}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{F}\mathbf{G}$ (and not on $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$). So

$$\mathbf{G}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{F}\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}} \tag{B.6}$$

with $\mathbf{U}$ a unitary matrix and $\mathbf{D}$ a diagonal matrix holding the singular values μ. Then,

$$\mathbf{D} = \underline{\mathbf{U}^{\mathrm{T}}\underline{\mathbf{G}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\underline{\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}}\mathbf{F}\mathbf{G}}\mathbf{U}} \tag{B.7}$$

Next, we express $\mathbf{\Sigma}$ in the same basis than $\mathbf{D}$ by applying $\mathbf{U}^{\mathrm{T}}$ and $\mathbf{U}$ to (B.5) (keeping in mind that $\mathbf{U}$ is unitary):

$$\mathbf{I} = \mathbf{U}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1}\mathbf{U} \tag{B.8}$$

In this new basis $\mathbf{\Sigma}^{-1}$ is the identity matrix and $\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}$ is a diagonal matrix, making it easy to compute the updated covariance matrix. Then the transformations must be undone to bring back the updated covariance matrix in the initial basis.

Now, we will apply the different transformations discussed above from the initial basis to get the new expression of $\mathbf{\Sigma}^a$. Let us consider the four following lines as in [8]:

$$\begin{aligned}
\mathbf{\Sigma}^a &= \left(\mathbf{\Sigma}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1} = (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\left[\mathbf{U}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\left(\mathbf{\Sigma}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{F}\mathbf{G}^{-1}\mathbf{U}\right]\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}\\
&= (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\left\{\left[\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}\left(\mathbf{\Sigma}^{-1} + \mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\right)(\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\right]^{-1}\right\}\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}\\
&= (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\left\{\left[\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}\mathbf{\Sigma}^{-1}(\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1} + \mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}(\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\right]^{-1}\right\}\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}\\
&= (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\left\{\left[\left(\mathbf{U}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1}\mathbf{U}\right)^{-1} + \mathbf{U}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{F}\mathbf{G}\mathbf{U}\right]^{-1}\right\}\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}.
\end{aligned}$$

In the first line, we multiply the expression of $\mathbf{\Sigma}^a$ by the transformations matrices $\mathbf{F}$, $\mathbf{G}$ and $\mathbf{U}$. In the second line, we introduce the transformations under the inversion sign of the summation. In the third line, we expand the sum, and then, in the fourth line, we rewrite the expressions as in (B.8) and (B.7). And Finally,

$$\mathbf{\Sigma}^a = (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\left\{\left[\underbrace{\left(\mathbf{U}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1}\mathbf{U}\right)^{-1}}_{=\mathbf{I}} + \underbrace{\mathbf{U}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{F}\mathbf{G}\mathbf{U}}_{=\mathrm{diag}[\mu_1,\mu_2,\ldots]}\right]^{-1}\right\}\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}$$

$$\begin{aligned}\left\{\left[\underbrace{\left(\mathbf{U}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1}\mathbf{U}\right)^{-1}}_{=\mathbf{I}} + \underbrace{\mathbf{U}^{\mathrm{T}}\mathbf{G}^{\mathrm{T}}\mathbf{F}^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{H}\mathbf{F}\mathbf{G}\mathbf{U}}_{=\mathrm{diag}[\mu_1,\mu_2,\ldots]}\right]^{-1}\right\} &= \mathrm{diag}[1/(1+\mu_1), 1/(1+\mu_2), \ldots] \\ &= \mathbf{B}^{\mathrm{T}}\mathbf{B} \text{ with } \mathbf{B} = \mathrm{diag}[(1+\mu_1)^{-1/2}, (1+\mu_2)^{-1/2}, \ldots] \\ &= \mathbf{B}^{\mathrm{T}}\underbrace{(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1}}_{=\mathbf{I}}\mathbf{B}\end{aligned}$$

Then,

$$\mathbf{\Sigma}^a = (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\underline{\mathbf{B}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{F}\mathbf{G}^{-1}\mathbf{B}}\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}$$

$$\mathbf{\Sigma}^a = \underbrace{(\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{B}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}}_{=\mathbf{A}}\mathbf{\Sigma}\underbrace{\mathbf{F}\mathbf{G}^{-1}\mathbf{B}\mathbf{U}^{-1}\mathbf{G}\mathbf{F}^{-1}}_{=\mathbf{A}^{\mathrm{T}}}$$

Finally, $\mathbf{\Sigma}^a = \mathbf{A}\mathbf{\Sigma}\mathbf{A}^{\mathrm{T}}$, where $\mathbf{A} = (\mathbf{F}^{\mathrm{T}})^{-1}\mathbf{G}^{\mathrm{T}}(\mathbf{U}^{\mathrm{T}})^{-1}\mathbf{B}^{\mathrm{T}}(\mathbf{G}^{\mathrm{T}})^{-1}\mathbf{F}^{\mathrm{T}}$.

C Papers

- Toye, H., Zhan, P., Gopalakrishnan, G., Kartadikaria, A.R., Huang, H., Knio, O., Hoteit, I.: Ensemble data assimilation in the Red Sea: Sensitivity to ensemble selection and atmospheric forcing. Ocean Dynamics **67**(7), 915-933 (2017). DOI 10.1007/s10236-017-1064-1. URL https://doi.org/10.1007/s10236-017-1064-1
- Toye, H., Kortas, S., Zhan, P., Hoteit, I.: A fault-tolerant hpc scheduler extension for large and operational ensemble data assimilation: Application to the Red Sea. Journal of Computational Science **27**, 46 - 56 (2018). DOI https://doi.org/10.1016/j.jocs.2018.04.018. URL http://www.sciencedirect.com/science/article/pii/S1877750317312905
- Friederici, A., Toye, H.M.K., Hoteit, I., Weinkauf, T., Theisel, H., Hadwiger, M.: A lagrangian method for extracting eddy boundaries in the Red Sea and the gulf of aden. In: 2018 IEEE Scientific Visualization Conference (SciVis), pp. 52-56 (2018). DOI 10.1109/SciVis.2018.8823600. URL https://ieeexplore.ieee.org/document/8823600
- Sivareddy, S., Toye, H., Zhan, P., Langodan, S., Krokos, G., Knio, O., Hoteit, I.: Impact of atmospheric and model physics perturbations on a high-resolution ensemble data assimilation system of the Red Sea. Journal of Geophysical Research: Oceans 125(8), e2019JC015611 (2020). DOI 10.1029/2019JC015611. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019JC015611
- Toye, H., Sivareddy, S., Raboudi, N., Hoteit, I.: A hybrid ensemble adjustment kalman filter based high-resolution data assimilation system for the Red Sea: Implementation and evaluation. Quarterly Journal of the Royal Meteorological Society n/a(n/a) (2020). DOI 10.1002/qj.3894. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3894
- Toye, H., Zhan, P., Sana, F., Sivareddy, S., Raboudi, N., Hoteit, I.: Adaptive en-

semble optimal interpolation for efficient data assimilation in the Red Sea. Journal of Computational Science, under review **n/a**(n/a), n/a (2020)