Learning from ocean remote sensing data

Redouane Lguensat

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

UNIVERSITE BRETAGNE LOIRE

THÈSE / IMT Atlantique

sous le sceau de l'Université Bretagne Loire

pour obtenir le grade de

DOCTEUR D'IMT Atlantique

Spécialité : Signal, Image, Vision

École Doctorale Mathématiques et STIC

Présentée par

Redouane Lguensat

Préparée dans le département Signal & communications

Laboratoire Labsticc

Learning from ocean remote sensing data

Thèse soutenue le 22 novembre 2017

devant le jury composé de :

Sylvie Thiria
Professeur, Université Paris VI / présidente

Antonio Turiel
Directeur de recherche, Institut de Ciències des Mar - Espagne / rapporteur

Marc Bocquet
Professeur, Ecole des Ponts- ParisTech / rapporteur

Bertrand Chapron
Chercheur, Ifremer - Plouzané / examinateur

Pierre Ailliot
Maître de conférences, Université de Bretagne Occidentale / examinateur

Clément Ubelmann
Chercheur, CLS – Ramonville Saint-Agne / examinateur

Ronan Fablet
Professeur, IMT Atlantique / directeur de thèse

Acknowledgements

Here I am, during a typical rainy day in my beloved Brittany, writing the last words of my thesis manuscript. This 3-year chapter of my life was full of new experiences, I learned a great amount of information and gained meaningful and valuable work experience. But all of these would not have been possible without the contribution of several persons whom supported me in one or many ways.

My supervisor Prof. Ronan Fablet, whom I knew as a teacher since my engineering studies. I want to sincerely thank him for having me in his group, for all the time and discussions we had, for his patience and his immense knowledge. It was a great pleasure to work under a passionate professor like him. I deeply thank him for always supporting me and for all the chances he offered me.

Dr. Pierre Tandeo, for his invaluable advice and feedback on my research, I appreciate all what he did for me since when he was my supervisor in my pre-thesis project.

Dr. Pierre Ailliot, for his teachings, his encouragements and for providing me with several useful suggestions throughout my thesis.

Dr. Bertrand Chapron, for his relevant comments on my work, his spicy humor and for making me more curious about ocean sciences.

Sincere appreciation to Prof. Marc Bocquet and Dr. Antonio Turiel for accepting to be the examiners of this thesis, and big thanks to Prof. Sylvie Thiria and Dr. Clément Ubelmann for accepting to be part of the jury members.

Thanks to all the members of the Signal and Communications department for maintaining a such great workplace. Particular thanks to my nice office colleagues, Carole, Yupeng, Yann and Aurélien with whom I spent most of the time, he was an excellent colleague and we often had stimulating conversations. Last but not least, I want to thank Phi who did a great job as a

research engineer in our group, collaborating with him was smooth and efficient. I learned from him a lot and I wish him a successful career (hopefully in research).

Warm thanks to Prof. Ge Chen who hosted me in his lab at Ocean University of China, I had the honor to be the first foreign exchange student in his group, and I was well received by the numerous master and phd students in the lab. Special mention to Dr. Miao Sun with whom I interacted the most either in Qingdao and in Brest when she came as an exchange student in our lab. I enjoyed collaborating with her and wish her a great success in her professional and personal life. Thanks to Prof. Junyu Dong, Dr. Tian Fenglin, Ms. Yuting Yang and everyone who made my stay at Qingdao pleasant, 谢谢!

And because it was not easy to undertake a PhD and maintain a top-tier social life, I want to thank every person with whom I spent a good, even so short, time. Many thanks to my friends whom checked in on me during these three years in Brest: Safaa, Rania, Charaf, Ghassane, Abderrahman, Yaser, Reda... the list is long. A heartfelt thank you to Soukaina for her continuous support and encouragement and for sharing with me my ups and downs.

Finally, I must express my high gratitude to my parents for all what they did for me, and I deeply thank them for the amount of sacrifices they did to raise our little family. Thanks to my brother Abdelhadi, my sisters Sofia and Yasmine, and all my extended family. I hope I am making all of you proud. This thesis is dedicated to my great-grandmother who left this world peacefully last year.

Contents

Abstract

Missing data is a widespread characteristic of remote sensing measurements. Various sources are responsible for this problem such as the instrument sampling or the sensitivity to the atmospheric conditions (e.g. cloud cover). The scientific problem of reconstructing geophysical fields from noisy and partial remote sensing observations is a classical problem well studied in the literature. Data assimilation is one class of popular methods to address this issue. It relies on a state-space representation of the physical system by two equations: The *observation equation* which models the measurement process and the *model equation* which explicits the physical model driving the state of the variable in time. In practice, data assimilation is done through the use of classical stochastic filtering techniques, such as ensemble Kalman or particle filters and smoothers. They proceed by an online evaluation of the physical model in order to provide a forecast for the state. The performance of data assimilation heavily relies on the definition of the physical model. The lack of consistency of the model with respect to the observed data and modeling uncertainties are therefore severe limitations of this classical framework. In contrast, the amount of observation and simulation data has grown very quickly in the last decades. Replacing the dynamical model by realistic statistical simulations of the dynamics has become feasible provided that we explore implicit data-driven schemes in such historical datasets using robust and well-suited methods.

My thesis focuses on the potential of exploiting the wealth of archived datasets to perform data assimilation in a data-driven way and this without having access to explicit model equations. Following Tandeo et al. (2014) [139], we particularly investigated a model-free and data-driven methodology. The main contribution of my thesis lies in developing and evaluating the Analog Data Assimilation, which combines analog methods (nearest neighbors search) and stochastic filtering methods (Kalman filters, particle filters, Hidden Markov Models). Through applications to both simplified chaotic models and real ocean remote sensing case-studies (sea surface temperature, along-track sea level anomalies), we demonstrate the relevance of the ana-

log data assimilation for the missing data interpolation of highly nonlinear and high-dimensional dynamical systems from irregularly-sampled and noisy observations.

Driven by the rise of machine learning in the recent years, I dedicated the last part of my thesis to the development of deep learning models for the detection and tracking of ocean eddies from multi-source and/or multi-temporal data (e.g., SST-SSH), the general objective being to outperform expert-based approaches [28, 109].

Keywords: *Analog Data Assimilation, Spatio-temporal interpolation, Ocean remote sensing*

.

Résumé

L'apparition de données manquantes est un phénomène très répandu des mesures de télédétection spatiale. Diverses sources sont responsables de ce problème, comme l'échantillonnage des instruments ou la sensibilité aux conditions atmosphériques (ex. couverture nuageuse). Le problème scientifique de la reconstitution des champs géophysiques à partir d'observations de télédétection bruitées et partielles est un problème classique bien étudié dans la littérature. L'assimilation des données est une des méthodes les plus populaires pour résoudre ce problème. Elle s'appuie sur une représentation espace-état du système physique suivant deux équations : *l'équation d'observation* qui modélise le processus de mesure et *l'équation de modèle* qui explique le modèle physique qui gouverne la dynamique de l'état de la variable dans le temps. En pratique, l'assimilation des données se fait par l'utilisation de techniques classiques de filtrage stochastique, telles que les filtres de Kalman d'Ensemble ou les filtres particulaires. Ils procèdent à une évaluation séquentielle du modèle physique afin de fournir une prédiction de l'état. La performance de l'assimilation des données dépend fortement de la définition du modèle physique. Le manque de cohérence du modèle par rapport aux données observées et les incertitudes de modélisation sont donc des limites sévères de ce cadre classique. D'un autre côté, la quantité de données d'observation et de simulation a augmenté très rapidement au cours des dernières décennies. Remplacer le modèle dynamique par des simulations statistiques réalistes de la dynamique est devenu possible à condition que nous explorions des schémas implicites *basés données* (data-driven) dans ces données historiques en utilisant des méthodes robustes et bien adaptées.

Cette thèse se concentre sur le potentiel d'exploitation de la richesse des données archivées pour effectuer l'assimilation des données de manière pilotée par les données et ce, sans avoir accès à des équations explicites de modèle. Suivant Tandeo et al. (2014) [139], nous avons particulièrement étudié une méthodologie sans modèle et guidée par les données. La principale contribution de ma thèse réside dans le développement et l'évaluation de l'assimilation des données par analogues, qui combine les méthodes analogues (recherche des plus proches voisins) et

les méthodes de filtrage stochastiques (filtres de Kalman, filtres particulaires, modèles de Markov cachés). Des applications aux modèles chaotiques simplifiés et à des études de cas de télédétection océanographique réelle (température de surface de la mer, anomalies du niveau de la mer), démontrent la pertinence de l'assimilation des données par analogues pour l'interpolation des données manquantes de systèmes dynamiques fortement non linéaires et à haute dimension à partir d'observations irrégulières et bruitées.

Poussé par l'essor de l'apprentissage automatique au cours des dernières années, j'ai consacré la dernière partie de ma thèse au développement de modèles d'apprentissage profond (Deep Learning) pour la détection et le suivi des tourbillons océaniques à partir de données multi sources et/ou multitemporelles (ex., SST-SSH), l'objectif général étant de surpasser les approches dites expertes [28, 109].

Mots clés : Assimilation de données par analogues, interpolation spatio-temporelle, télédétection de l'océan.

Acronyms

AMSR-E Advanced Microwave Scanning Radiometer - Earth Observing System

AnDA Analog Data Assimilation

AR Auto-regressive

AVHRR Advanced Very High Resolution Radiometer

AVISO Archiving, Validation and Interpretation of Satellite Oceanographic data

CMEMS Copernicus Marine and Environment Monitoring Service

EnKF Ensemble Kalman Filter

EOF Empirical Orthogonal Function

HMM Hidden Markov Model

KF Kalman Filter

MS-AnDA Multiscale Analog Data Assimilation

OI Optimal Interpolation

PCA Principal Component Analysis

PF Particle Filter

SLA Sea Level Anomaly

SSH Sea Surface Height

SST Sea Surface Temperature

General Introduction

"If you do not know how to ask the right question, you discover nothing"

W. Edward Deming

Introduction and problem statement

The reconstruction of the spatiotemporal dynamics of geophysical systems from noisy and/or partial observations is a major issue in geosciences. Variational and stochastic data assimilation schemes are the two main categories of methods considered to address this issue (see [46] for more details). A key feature of these data assimilation schemes is that they rely on repeated forward integrations of an explicitly-known dynamical model. This may greatly limit their application range as well as their computational efficiency. First, thorough and time-consuming studies may be required to identify explicit representations of the dynamics, especially regarding fine-scale effects and subgrid-scale processes as for instance in regional geophysical models [71]. Such processes typically involve highly nonlinear and local effects [157]. The resulting numerical models may be computationally intensive and even prohibitive for assimilation problems, for instance regarding the generation of members with different initial conditions at each time step. Second, as explained in [153], "with ever-increasing resolution and complexity, the numerical models tend to be highly nonlinear and also observations become more complicated and their relation to the

models more nonlinear". In such situations, standard data assimilation techniques are likely to fail, including nonlinear particle filters which are prone to the "curse of dimensionality". Third, difficulties may occur when geophysical dynamics involve uncertain model parameterizations or space-time switching between different dynamical modes that need to be estimated online [129] or offline [140]. Dealing with such situations may not be straightforward using classical model-driven assimilation schemes.

Meanwhile, recent years have witnessed a proliferation of satellite data, in situ monitoring as well as numerical simulations. Large databases of valuable information has been collected and represent a major opportunity for oceanic, atmospheric and climate sciences. As pioneered by [102], the availability of such datasets advocates for the development of analog forecasting strategies, which make use of "similar" states of the dynamical system of interest to generate realistic forecasts. Analog forecasting strategies have become more and more popular in oceanic and atmospheric sciences [111,115], and have benefited from recent advances in machine learning [163]. They have been applied to a variety of systems and application domains, including among others, rainfall nowcasting [8], air quality analysis [39], wind field downscaling [69], climate reconstruction [134] and stochastic weather generators [160].

In this work, we examine the extension of the analog forecasting paradigm for data assimilation issues. Given a representative dataset of the dynamics of the system, this extension that we call "Analog Data Assimilation" consists of a combination of the implicit analog forecasting of the dynamics with stochastic filtering schemes, namely Ensemble Kalman and particle filtering schemes [47]. This idea was first introduced in [139] where the authors demonstrated the relevance of the proposed analog data assimilation for the reconstruction of complex dynamics from partial and noisy observations. Tandeo et al. derived filtering and smoothing algorithms called the *Analog Ensemble Kalman Filter and Smoother*, which combine analog forecasting and the ensemble Kalman filter and smoother. A similar philosophy was followed independently in [65] where the authors combine ideas from Takens' embedding theorem and ensemble Kalman filtering to infer the hidden dynamics from noisy observations. Hamilton et al. called their algorithm the *Kalman-Takens filter*.

Whereas these two previous works provide proofs of concept, this thesis further investigates and evaluates different analog assimilation strategies and their detailed implementation.

In addition, experiments on Sea Surface Temperature (SST) and Sea Level Anomaly (SLA) missing data interpolation are conducted to investigate the challenges present in realistic applications and to face the curse of dimensionality.

Given that good quality and high resolution SST/SSH maps are crucial to eddy classification and detection, I dedicated the last part of my thesis to the development of deep learning based image segmentation architectures. The aim is to have a pixelwise classification of an SSH map into cyclonic/anticyclonic eddy or absence of eddies. The general objective being to outperform expert-based approaches [28, 109].

This thesis was conducted under the supervision of Prof. Ronan Fablet (LabSTICC, IMT Atlantique), Dr. Pierre Ailliot (LMBA, University of Western Brittany) and Dr. Bertrand Chapron (LOPS, Ifremer). I benefited from a short stay at Ocean University of China, where I started two collaborations with Prof. Ge Chen and Prof. Junyu Dong.

Contributions

The contributions of this thesis are the following:

Presenting a unified framework for the Analog Data Assimilation with new analog forecasting strategies and new analog-based algorithms

The principal objective of chapter 2 is to introduce the Analog Data Assimilation. A brief history of analog methods and their recent implication in data assimilation is presented. The chapter lists the considered analog forecasting strategies, including locally-linear ones that were not considered in previous works, and evaluates their performance for analog data assimilation. Secondly, in addition to the ensemble Kalman algorithms, I propose and examine two novel implementations of the analog forecasting , the first combined with a particle filter and the second with Hidden Markov Models. Finally, in collaboration with Pierre Tandeo and Phi Viet Huynh, we provide a unified computational framework, through both a Matlab Toolbox and a Python Library, to pave the way for practical use and future research (it is available from `https://github.com/ptandeo/AnDA`).

Using the Analog Data Assimilation to solve high-dimensional geophysical problems through the combination of patch-based and EOF-based methods

Chapter 3 and Chapter 4 deal with the challenges of using the AnDA for high-dimensional

problems, more specifically: *i)* Interpolation of Sea Surface Temperature (SST) from cloud contaminated satellite data and *ii)* Sea Level Anomaly (SLA) mapping from along-track data. We circumvent the curse of dimensionality by implementing a patch-based version of the AnDA that breaks the problem into several small subregions, we also used dimensionality reduction throughout the use of EOF decomposition to decrease the dimensionality of the problem. This has a direct effect on the quality of the analogs.

Transfer learning of image segmentation using Deep Learning to the detection and classification of eddies from SSH maps.

Appendix B presents "Eddynet" a deep learning based architectures for automated eddy detection and classification from Sea Surface Height (SSH) maps provided by Archiving, Validation, and Interpretation of Satellite Oceanographic (AVISO). Eddynet's output is a map with the same size of the SSH map input where pixels have the following labels {'0': Non eddy, '1': anticyclonic eddy, '2': cyclonic eddy, '3': land or no data}.

Keras python code, the training datasets and EddyNet weights files are open-source and freely available on `https://github.com/redouanelg/Eddynet`

Publications

Submitted

- **R. Lguensat**, P. Viet, M. Sun, G. Chen, T. Lin, B. Chapron, R. Fablet, "Data-driven interpolation of Sea Level Anomalies using Analog Data Assimilation".
- **R. Lguensat**, M. Sun, R. Fablet, E. Mason, P. Tandeo, G. Chen, "EddyNet: A Deep Neural Network For Pixelwise Classification Of Oceanic Eddies".

Journal papers

- **R. Lguensat**, P. Tandeo, P. Ailliot, M. Pulido and R. Fablet, "The Analog Data Assimilation", *Monthly Weather Review*, 2017.
- R. Fablet, P. Viet, and **R. Lguensat**. "Data-driven Methods for Spatio-Temporal Interpolation of Sea Surface Temperature Images". *IEEE Transactions on Computational Imaging*, 2017.

- Y. Yang, J. Dong, X. Sun, **R. Lguensat**, M. Jian, X. Wang. "Ocean Front Detection from Instant Remote Sensing SST Images". *IEEE Geoscience and Remote Sensing Letters*, 2016

Conference papers

- R. Fablet, P. Viet, and **R. Lguensat**. "Data-driven assimilation of irregularly-sampled image time series" ICIP 2017: IEEE International Conference on Image Processing, Beijing, China.

- **R. Lguensat**, M. Sun, G. Chen, T. Lin, R. Fablet, Spatio-Temporal Interpolation of Altimeter-Derived SSH Fields Using Analog Data Assimilation: A Case-Study In The South China Sea. IGARSS 2017 : IEEE International Geoscience and Remote Sensing Symposium, Fort Worth, Texas, USA.

- **R. Lguensat**, R. Fablet, P. Ailliot and P. Tandeo, An Exemplar-based HMM framework for nonlinear state-space models. EUSIPCO 2016 : IEEE European Signal Processing Conference, Budapest, Hungary.

- **R. Lguensat**, P. Tandeo, P. Ailliot, B. Chaperon and R. Fablet, Using archived datasets for missing data interpolation in ocean remote sensing observation series, MTS/IEEE OCEANS'16, Shanghai, China.

- P. Tandeo, P. Ailliot, B. Chapron, **R. Lguensat** and R. Fablet, The analog data assimilation: application to 20 years of altimetric data, Climate Informatics 2015, Boulder, Colorado.

- **R. Lguensat**, P. Tandeo, R. Fablet and P. Ailliot, Non-parametric Ensemble Kalman methods for the inpainting of noisy dynamic textures. ICIP 2015 : IEEE International Conference on Image Processing, Quebec City, Canada.

- (Pre-thesis project) **R. Lguensat**, P. Tandeo, R. Fablet and R. Garello, Spatio-temporal interpolation of Sea Surface Temperature using high resolution remote sensing data, OCEANS'14, St. John's, Canada.

Invited conferences and workshop talks

- Poster presentation at Data Science and Environment conference (DSE17), July 17, Brest, France.
- Talk on "Toward data-driven methods in geophysics: the Analog Data Assimilation" at European Geophysical Union 2017, Vienna, Austria.
- Poster presentation at Colloque National sur l'Assimilation de données (French national seminar on Data Assimilation) Nov 2016, at Grenoble, France.
- Poster presentation at ACO conference organized by Sea Tech Week, Oct 2016, Brest, France.
- Talk on "Analog Data Assimilation" at the workshop on Stochastic Weather Generators (SWGEN), 17 May 2016, Vannes, France.
- Talk on the use of historical datasets in geophysics at a joint Seminar between the MIT Lab and Vision Lab of Ocean University of China, 19 April 2016, Qingdao, China.
- Poster presentation at the 2nd GlobCurrent User Consultation Meeting on Analog methods applied to 20 years of altimetric data, 4-6 November, Brest, France.
- Talk and Poster presentation on inpainting of noisy and noncomplete image sequences using Analog Ensemble Kalman methods at the first MissData 2015 conference, 18-19 June, Rennes, France.
- Talk at SEACS workshop on Analog Hidden Markov Models: a discrete formulation of the Analog particle filter introduced by Tandeo et al., 26-27 May, in Landeda, France.
- LabSTICC seminar on nonlinear non convex inverse problems, Telecom Bretagne, Brest. 19th March 2015

Part I

Analog methods for state-space problems: Analog Data Assimilation

Chapter 1

Data Assimilation and Analog methods

Knowledge is an unending adventure at the edge of uncertainty.

Jacob Bronowski

In this chapter, we present the main ideas behind the *state-space model* formulation. We will then use the term *data assimilation* which is the term commonly used in the geoscience community for state-space mathematical resolution. Finally, we give a historical overview of the use of the analog methods.

1.1 State-Space models

In many problems encountered in science and engineering, one is interested in estimating an unobserved process $\{\mathbf{x}(t)\}_{t\in[\![1,\cdots,T]\!]}$ given a sequence of observations $\{\mathbf{y}(t)\}_{t\in[\![1,\cdots,T]\!]}$. Examples of such situations include target tracking, signal and image processing, climate modeling, finance, etc... In this section, we review the resolution of such inverse problems using state-space formulations. We may refer the reader to [32] for a comprehensive introduction to state-space models from a theoretical/practical point of view.

State-space methods provide a flexible framework to address this issue. They rely on the definition of two key components. Firstly, the **dynamical model** states the temporal dynamics of process $\{\mathbf{x}(t)\}_{t\in[\![1,\cdots,T]\!]}$, typically Markovian dynamics (as an illustration, we consider here a first-order Markov process). Secondly, the **observation model** relates the unknown state $\mathbf{x}(t)$ at a given time t to the observed variable $\mathbf{y}(t)$ at the same time. Formally, it resorts to:

$$\begin{cases} \mathbf{x}(t) = \mathcal{M}\left(\mathbf{x}(t-1), \boldsymbol{\eta}(t)\right), & (1.1) \\ \mathbf{y}(t) = \mathcal{H}\left(\mathbf{x}(t)\right) + \boldsymbol{\epsilon}(t). & (1.2) \end{cases}$$

Where $\mathcal{M}$ characterizes the dynamical model of the true state $\mathbf{x}(t)$, while $\boldsymbol{\eta}(t)$ is a random perturbation added to represent model uncertainty. Observation error is considered through the random noise $\boldsymbol{\epsilon}(t)$. Here, for the sake of simplicity, we consider an additive Gaussian noise $\boldsymbol{\epsilon}$ with covariance $\mathbf{R}$ in equation 1.2 and the observation operator $\mathcal{H} = \mathbf{H}$ is assumed linear.

To be fully characterized, this state-space setting also involves the definition of the **prior distribution** of $\mathbf{x}(1)$. From a Bayesian perspective, the reconstruction of the unknown state sequence $\{\mathbf{x}(t)\}_{t\in[\![1,\cdots,T]\!]}$ from a partial and/or noisy observation sequence $\{\mathbf{y}(t)\}_{t\in[\![1,\cdots,T]\!]}$ comes to evaluate **filtering** and **smoothing** posteriors, respectively $P(\mathbf{x}(t)|Y_{1:t})$ the probability distribution of state $\mathbf{x}$(t) given all the past and present observations and $P(\mathbf{x}(t)|Y_{1:T})$ the probability distribution of state $\mathbf{x}$(t) given all the past, present and future observations, where the notation $Z_{1:k}$ represents the sequence of states $\mathbf{z}(t)$ from time 1 to time k.Fig.1.1 shows an illustration of the existing conditional dependencies in a state space model

In the following we will present three classical methods for the resolution of state-space models.

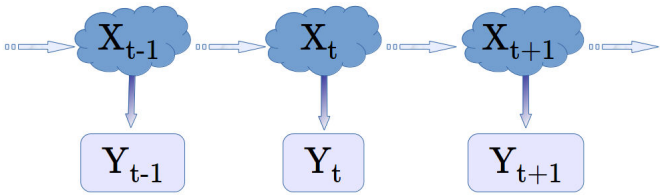


Figure 1.1 – An illustration of a simple SSM: The random variable X_t is the hidden state at time t. The random variable Y_t is the corresponding observation (or measurement) at time t. There is only two kind of conditional dependencies, first between the hidden state X_t at time t and the previous state at time $t-1$ (dynamical model). Second, between the measurement Y_t and the hidden state X_t both at time t (observation model).

1.1.1 The Kalman Filter

Here we consider a Linear Gaussian model *i.e.* we consider an additive Gaussian noise $\boldsymbol{\eta}$ with covariance $\mathbf{Q}$ in equation 1.1 and the dynamical operator $\mathcal{M} = \mathbf{M}$ is assumed linear.

$$\begin{cases} \mathbf{x}(t) = \mathbf{M}\left(\mathbf{x}(t-1)\right) + \boldsymbol{\eta}(t), & (1.3) \\ \mathbf{y}(t) = \mathbf{H}\left(\mathbf{x}(t)\right) + \boldsymbol{\epsilon}(t). & (1.4) \end{cases}$$

In this particular case where conditional are also normal distributions, the Kalman Filter (KF) [82] gives recursive expressions for the mean and variance of the filtering distribution $P(\mathbf{x}(t)|Y_{1:t})$, under the assumption that all parameters in the model are known.

More specifically, the KF recursively estimates:

- $\hat{\mathbf{x}}_{t|t}$ the mean state at time t given the previous observations
- $\mathbf{P}_{t|t}$ the corresponding error covariance matrix.

In the following we derive the equations of the Kalman filter. Consider $Y_{1:t}$ to be the vector of the observations up to and including time t, and to simplify notation we will use $\mathbf{z}_t$ for $\mathbf{z}(t)$. The KF in this case is the MMSE estimator represented by the conditional expectation of $\mathbf{x_t}$ given the known observations $Y_{1:t}$:

$$\hat{\mathbf{x}}_{t|t} = E[\mathbf{x}_t|Y_{1:t}] \tag{1.5}$$

Recall that if $\mathbf{v}_1$ and $\mathbf{v}_2$ are jointly Gaussian with $\begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} \sim \mathcal{N}(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix})$, then $\mathbf{v_1}|\mathbf{v_2} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where:

$$\begin{cases} \tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{v_2} - \boldsymbol{\mu}_2) & (1.6) \\ \tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}. & (1.7) \end{cases}$$

By taking $\mathbf{v_1} = \mathbf{x}_t$ and $\mathbf{v_2} = \mathbf{y}_t$, then conditioning by $Y_{1:t-1}$, Equation 1.5 becomes:

$$\hat{\mathbf{x}}_{t|t} = E[\mathbf{x}_t|Y_{1:t-1}] + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y_t} - E[\mathbf{y}_t|Y_{1:t-1}]) \tag{1.8}$$

where:

$$\boldsymbol{\Sigma}_{xy} = E[(\mathbf{x_t} - E[\mathbf{x}_t|Y_{1:t-1}])(\mathbf{y_t} - E[\mathbf{y}_t|Y_{1:t-1}])^T] \tag{1.9}$$

$$\boldsymbol{\Sigma}_{yy} = E[(\mathbf{y_t} - E[\mathbf{y}_t|Y_{1:t-1}])(\mathbf{y_t} - E[\mathbf{y}_t|Y_{1:t-1}])^T] \tag{1.10}$$

Using the observation equation 1.4 we have:

$$E[\mathbf{y}_t|Y_{1:t-1}] = \mathbf{H}E[\mathbf{x}_t|Y_{1:t-1}] = \mathbf{H}\hat{\mathbf{x}}_{t|t-1} \tag{1.11}$$

which makes equation 1.8 becomes:

$$\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y_t} - \mathbf{H}\hat{\mathbf{x}}_{t|t-1}) \tag{1.12}$$

And using the dynamical equation 1.3 we obtain:

$$\hat{\mathbf{x}}_{t|t-1} = E[\mathbf{x}_t|Y_{1:t-1}] = \mathbf{M}E[\mathbf{x}_{t-1}|Y_{1:t-1}] = \mathbf{M}\hat{\mathbf{x}}_{t-1|t-1} \tag{1.13}$$

Equations 1.12 and 1.13 define the recursive filter. We will now explicit $\mathbf{K_t} = \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}$ called the Kalman gain, then find the updating formulas for the covariance error.

From the observation equation 1.4 we can show that:

$$\begin{cases} \boldsymbol{\Sigma}_{xy} = \mathbf{P_{t|t-1}}\mathbf{H}^T & (1.14) \\ \boldsymbol{\Sigma}_{yy} = \mathbf{H}\mathbf{P_{t|t-1}}\mathbf{H}^T + \mathbf{R}. & (1.15) \end{cases}$$

This results in the following expression for the Kalman gain:

$$\mathbf{K_t} = \mathbf{P_{t|t-1}}\mathbf{H}^T(\mathbf{H}\mathbf{P_{t|t-1}}\mathbf{H}^T + \mathbf{R})^{-1} \tag{1.16}$$

To find the updating formulas for the covariance error i.e. relationship between $\mathbf{P_{t|t}}$ and $\mathbf{P_{t|t-1}}$, we will start by using Equation 1.7, in our case it resorts to:

$$\mathbf{P_{t|t}} = \mathbf{P_{t|t-1}} - \mathbf{\Sigma}_{xy}\mathbf{\Sigma}_{yy}^{-1}\mathbf{\Sigma}_{yx} \tag{1.17}$$

Since $\mathbf{\Sigma}_{yx} = \mathbf{\Sigma}_{xy}^T$, and using Equation 1.14, the previous equation becomes:

$$\mathbf{P_{t|t}} = \mathbf{P_{t|t-1}} - \mathbf{K_t H P_{t|t-1}} = (\mathbf{I} - \mathbf{K_t H})\mathbf{P_{t|t-1}} \tag{1.18}$$

Besides, using the dynamical equation we can show that

$$\mathbf{P_{t|t-1}} = \mathbf{M P_{t-1|t-1} M}^T + \mathbf{Q} \tag{1.19}$$

Finally the KF algorithm can be summarized in Algorithm 1.

Algorithm 1 The Kalman filter algorithm

1: Input: $\mathbf{x}_1 = \mathbf{x}^b$ and $\mathbf{P}_1 = \mathbf{B}$ initial guesses

2: set $t = 2$

3: **Prediction step:**

- predict state estimate $\hat{\mathbf{x}}_{t|t-1} = \mathbf{M}\hat{\mathbf{x}}_{t-1|t-1}$
- predict covariance estimate $\mathbf{P_{t|t-1}} = \mathbf{M P_{t-1|t-1} M}^T + \mathbf{Q}$

4: **Update step:**

- Calculate the Kalman gain $\mathbf{K_t} = \mathbf{P_{t|t-1}H}^T(\mathbf{H P_{t|t-1} H}^T + \mathbf{R})^{-1}$
- update state estimate $\hat{\mathbf{x}}_{t|t} = \hat{\mathbf{x}}_{t|t-1} + \mathbf{K_t}(\mathbf{y_t} - \mathbf{H}\hat{\mathbf{x}}_{t|t-1})$
- update covariance estimate $\mathbf{P_{t|t}} = (\mathbf{I} - \mathbf{K_t H})\mathbf{P_{t|t-1}}$

5: Set $t = t + 1$ then go back to step 3

Despite the attractiveness and the popularity of the classical Kalman Filter (e.g. Apollo navigation computer which took mankind to the moon), it is a basic model with abiding assumptions, and since nature is nonlinear, the need of more general but still computationally plausible methods had arisen. In the literature, two main research directions were followed: the first path considered the use of Monte Carlo methods, especially the particle filter that began to appear and started to be used in the estimation theory field, we present it in section 1.1.2.

While the second group of researchers put their efforts in improving and adapting the KF for nonlinear problems, two popular extensions emerged: the *Extended Kalman filter* (EKF) that simply considers linearization of the nonlinear system around working points, and the *Unscented Kalman Filter* (UKF) [81] which selects some representative points from the state distribution, from which the posterior distribution is then obtained using the propagation of the these representative points through the direct use of the nonlinear system. We refer the reader to the book of Anderson and Moore [1] for a complete description of Kalman filter extensions.

The geoscience community benefited from this variety of research ideas from both sides and widely adopted an interesting method combining the best of both world: The Ensemble Kalman Filter (EnKF) that we discuss in section 1.2.1.

On another side, while the previous models were based on a continuous formulation, Hidden Markov Models (HMM) [126] were introduced as an alternative adapted to discrete systems, we present HMM and the associated filtering and smoothing algorithms in section 1.1.3.

1.1.2 The Particle Filter

Contrary to the Kalman filters, particle filters do not assume a Gaussian distribution for the state. The key principle is to estimate the posteriors of the state from a set of particles (or ensemble members).

Hereinafter, we comply with the notations used in the geoscience community by doing the following replacements:

- $\mathbf{x_{t|t-1}}$ and $\mathbf{P_{t|t-1}}$ are now referred to as the forecast state $\mathbf{x}^f(t)$ and covariance error $\mathbf{P}^f(t)$
- $\mathbf{x_{t|t}}$ and $\mathbf{P_{t|t}}$ are now referred to as the analyzed state $\mathbf{x}^a(t)$ and covariance error $\mathbf{P}^a(t)$

In Algorithm 2, we present one version and probably the most classical of the particle filter, this version is called the Bootstrap [61] (as known as the sampling importance resampling (SIR) particle filter).

The literature comprise several other variants of the particle filter, from which we can cite: firstly, the auxiliary particle filter [124] where the resampling and the prediction step are inverted to give more sampling "chance" to particles close to the observation, secondly, the rejection particle filter [141] which assumes knowing an upper bound of the inferred distribution and then rejects particles that exceed this bound. Although the variety of particle filters, a number of limitations makes the use of particle filter challenging. Firstly, the presence of outliers could neg-

Algorithm 2 The Particle filter algorithm

1: Input: $\mathbf{x}^b$ and $\mathbf{B}$ parameters of the prior Gaussian distribution
2: Generate vectors $\mathbf{x}_i^f(1)\ \forall i \in \{1, ..., N\}$ using a multivariate Gaussian random generator with mean vector $\mathbf{x}^b$ and covariance matrix $\mathbf{B}$. The index i of the state vector corresponds to the i^{th} realization of the Monte Carlo procedure (called member or particle).
3: Set $t = 1$
4: **Prediction step:**

- Apply the model dynamical operator $\mathcal{M}$ to sample new particles $\mathbf{x}_i^f(t)\ \forall i \in \{1, ..., N\}$ from previous filtered particles $\mathbf{x}_i^a(t-1)$
- Compute particle weights $\pi_i(t)$ as

$$\pi_i(t) \propto \phi\left(\mathbf{y}(t) - \mathbf{H}\mathbf{x}_i^f(t); \mathbf{R}\right), \tag{1.20}$$

  where $\phi(\cdot; \mathbf{R})$ is a centered multivariate Gaussian distribution with covariance $\mathbf{R}$.
- Normalize weights $\pi_i(t)$ to total one.

5: **Resampling step:**

- Resample from the multinomial distribution defined by the particles $\{\mathbf{x}_i^f(t)\}$ and their corresponding weights $\{\pi_i(t)\}$.
- Compute the analyzed state $\mathbf{x}^a(t)$ as the sample mean

$$\mathbf{x}^a(t) = \frac{1}{N}\sum_{i=1}^{N} \pi_i(t)\mathbf{x}_i^f(t). \tag{1.21}$$

  but one may also consider as filtered state the posterior mode.

6: Set $t = t + 1$ then go back to step 4

atively affect the importance sampling and mislead the particles, thus the use of many particles is necessary. Secondly, the curse of dimensionality is a serious problem in particle filtering [125]. Actually the need of a large number of particles for a better estimation could be intractable computationally, not to mention that using a very big number of particles means increasing the variance due to the bias-variance trade-off.

For further reading, we point the reader to the well-detailed survey of Chen [30] for a complete review of the different variants of the particle filter, their advantages and limitations, and their convergence guarantees.

1.1.3 Hidden Markov Models

Unlike continuous state space models depicted before, here we're interested in discrete state space models *i.e.* the state has values in a finite set of values. In this case state space models are commonly known as Hidden Markov Models (HMMs).

In the discrete setting considered here, $\mathbf{x}(t)$ is the state at time t of a discrete random variable having N_p number of possible values, and $\mathbf{y}(t)$ the corresponding observation at time t. Two defining properties give to Hidden Markov Models their name: first, $\mathbf{x}(t)$ is supposed to be *hidden* from the observer, $\mathbf{y}(t)$ is all what he observes at time t. Second, the value of the state at time t is independent of all values of the state prior to $t-1$, this is called the *Markov property*,

$$P(\mathbf{x}(t+1) \mid \mathbf{x}(0), \mathbf{x}(1), \ldots, \mathbf{x}(t)) = P(\mathbf{x}(t+1) \mid \mathbf{x}(t)),$$

using these two properties, HMMs describe the joint probability of the hidden and observed discrete random variables. We note by $\Lambda = (A, B, \pi_1)$ the parameters of the HMM where:

- Assuming that $P(\mathbf{x}(t)|\mathbf{x}(t-1))$ is independent of time t, the definition of the time independent **transition matrix** is given by:

$$A = \{a_{ij}\} = P(\mathbf{x}(t) = j|\mathbf{x}(t-1) = i)$$

- The **initial state distribution** (i.e. when $t = 1$) is given by:

$$\pi_1 = \{\pi_i\} \text{ where } \pi_i = P(\mathbf{x}(1) = i)$$

- The **observation matrix** (called also the emission matrix) gives the probability of a certain observation at time t for state j and it is expressed as:

$$B = \{b_j(\mathbf{y}(t))\} \text{ where } b_j(Y_t) = P(\mathbf{y}(t)|\mathbf{x}(t) = j)$$

Given a foreknowledge of an HMM parameters and an observation sequence we can compute the smoothing posterior marginals $P(\mathbf{x}(t)|\, Y_{1:T})$ of all hidden state variables. In the next subsection, we will introduce briefly *The forward-backward algorithm* which is a widely used algorithm to execute this task. Its aim consists of finding the most likely state for any point in time and which results in an estimation of the underlying dynamics of the state.

The forward-backward algorithm

Given $\Lambda = (A, B, \pi_1)$ we are interested in evaluating $\gamma_t(i) = P(\mathbf{x}(t) = i|Y_{1:T})$ which can also be written using Bayes theorem as:

$$\gamma_t(i) = \frac{P(Y_{1:T}, \mathbf{x}(t) = i)}{P(Y_{1:T})} \tag{1.22}$$

Using conditional independence properties between $Y_{1:t}$ and $Y_{t+1:T}$ given $\mathbf{x}(t)$, we can prove easily that:

$$P(Y_{1:T}, \mathbf{x}(t) = i) = P(Y_{1:t}, \mathbf{x}(t) = i).P(Y_{t+1:T}|\mathbf{x}(t) = i) \tag{1.23}$$

Let consider the *forward* variable $\alpha_t(i)$ and the *backward* variable $\beta_t(i)$ defined as:

$$\alpha_t(i) = P(Y_{1:t}, \mathbf{x}(t) = i) \quad \beta_t(i) = P(Y_{t+1:T}|\mathbf{x}(t) = i) \tag{1.24}$$

$\alpha_t(i)$ is the joint probability of observing $Y_{1:t}$ and being in the state $\mathbf{x}(t) = i$ at time t, while $\beta_t(i)$ is the conditional probability of future observation $Y_{t+1:T}$ assuming being in state $\mathbf{x}(t) = i$ at time t. Thus (1.22) and (1.23) lead us to write:

$$\gamma_t(i) = \frac{\alpha_t(i).\beta_t(i)}{P(Y_{1:T})} \tag{1.25}$$

$P(Y_{1:T})$ is a normalization factor that makes $\gamma_t(i)$ a probability measure $i.e$ $\Sigma_{i=1}^{N_p}\gamma_t(i) = 1$, thus we can express (1.25) as:

$$\gamma_t(i) = \frac{\alpha_t(i).\beta_t(i)}{P(Y_{1:T})} = \frac{\alpha_t(i).\beta_t(i)}{\Sigma_{i=1}^{N_p}\alpha_t(i).\beta_t(i)} \tag{1.26}$$

Thanks to the forward-backward algorithm presented in Algorithm 3 we can obtain $\gamma_t(i)$ at each time step. Choosing the most likely state for the system at time t is straightforward by taking the index of the state with the larger probability value:

$$\mathbf{x}(t) = \arg \max_{i=1\cdots N_p} \gamma_t(i) \tag{1.27}$$

A Matlab implementation of the forward-backward algorithm can be found in the Hidden Markov Model (HMM) Toolbox written by Kevin Murphy[1].

[1]See `http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html`

Algorithm 3 Forward-backward algorithm

%% $Input : \Lambda = (A, B, \pi_1)$
%%%%%% $ForwardProbabilities$ %%%%%%
$\alpha_1(i) = \pi_i b_i(Y_1) \quad \forall i \in 1:1:Q$
for $t = 2:1:T$ **do**
    for $j = 1:1:N_p$ **do**
        $\alpha_t(j) = [\Sigma_{i=1}^{N_p} \alpha_{t-1}(i).a_{ij}]b_j(Y_t)$
    end for
end for
%%%%%% $BackwardProbabilities$ %%%%%%
$\beta_T(i) = 1 \quad \forall i \in 1:1:N_p$
for $t = T-1:-1:1$ **do**
    for $i = 1:1:N_p$ **do**
        $\beta_t(i) = \Sigma_{j=1}^{N_p} b_j(Y_{t+1}).a_{ij}.\beta_{t+1}(j)$
    end for
end for
%%%%%% $GammaProbabilities$ %%%%%%
for $i = 1:1:N_p$ **do**
    for $t = 1:1:T$ **do**
        $\gamma_t(i) = \frac{\alpha_t(i).\beta_t(i)}{\Sigma_{i=1}^{N_p} \alpha_t(i).\beta_t(i)}$
    end for
end for

When the finite discrete state-space model is well defined, HMM inference implementation is attractive and modulable with its direct calculations of filtering and smoothing posteriors through dynamic programming. However, the number of finite states and the parametrization could be an issue in high dimensions, if a state has a large number N_p of states, this could affect much the storage and calculation capacities.

It would be useful to mention two other algorithms for HMM inference. Firstly, Viterbi algorithm which computes the most probable path generated by the observation sequence. Secondly, Baum-Welch algorithm that aims to iteratively estimate the parameters of the HMM by performing a series of forward-backward algorithm runs. Details about these algorithms can also be found in [126].

1.2 Data assimilation in geoscience

Data assimilation is generally defined in geoscience as the use of state space models in order to assimilate observations/measurements about a geophysical system of interest. We recommend the book of Asch et al. [6] and the well detailed paper of Carassi et al. [22] for a complete overview of data assimilation techniques in geoscience.

Two types of data assimilation approaches are extensively studied in the literature: variational and stochastic ones. **Variational data assimilation** proceeds by minimizing a cost function based on a continuous formulation of equations (1.1-1.2) [100], while **stochastic data assimilation** schemes rely on the sampling and/or maximization of the posterior likelihood of the state sequence given the observation series [83]. These classical data assimilation schemes are regarded as "model-driven", in the sense that they combine observations with forecasts provided by a numerical model $\mathcal{M}$.

While variational and stochastic schemes are equivalent in the Linear-Gaussian case and resort to the same optimal solution in a MMSE sense, this not the case in general. One advantage of stochastic schemes is that they provide not only an estimation of the state of interest but also its covariance error matrix. Since this thesis fits into the statistical and probabilistic perspective of data assimilation, hereinafter, the focus will be directed to stochastic data assimilation and its methods. More specifically, we are interested in sequential stochastic data assimilation methods. An example of the general procedure of these methods is shown in Figure 1.2, starting from a background state (first-guess) and a background covariance error, the sequential assimilation proceeds in two steps: the **prediction** step uses the transition model (cf. Equation 1.1) to obtain a forecast state, then the upcoming observation is "assimilated" into the model in the **analysis** step. The assimilation is the mathematical resolution of the state-space (1.1)-(1.2). We present in section 1.2.1 the Ensemble Kalman Filter and Smoother [47] one of the popular data assimilation algorithms in geoscience, we also present one of the earliest data assimilation methods that relies on predetermined covariance error matrices instead of dynamical update of the analyzed state, this algorithm is described in section 1.2.2. These two algorithms are the main data assimilation algorithms used in this work.

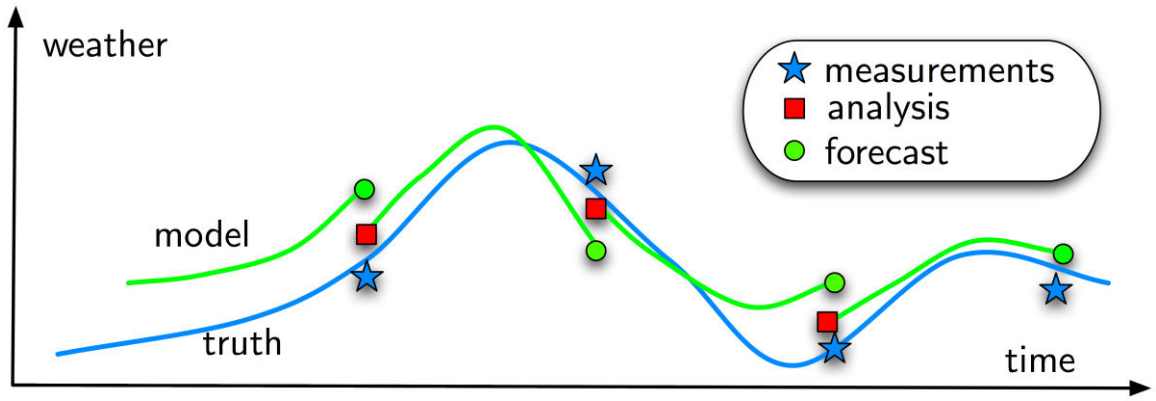


Figure 1.2 – Weather forecast chain, an example of data assimilation procedure. Illustration source [144].

1.2.1 Ensemble Kalman filters (EnKF) and smoothers (EnKS) as an example of stochastic data assimilation

Given the high dimensionality of geophysical problems (Numerical Weather Prediction, Oceanography, Hydrology, etc...), the use of classical Kalman filters is prohibited by computationally expensive matrix inversions (e.g. the error covariance matrix) and storage shortage. Researchers in the field use therefore several techniques to overcome these limitations, in particular, square root implementation of the Kalman filter and the ensemble Kalman filter. The second is appealing from a statistical point of view and was considered in this thesis. In the following, we present step-by-step the Ensemble Kalman Filter and Smoother.

Ensemble Kalman filters (EnKF) and smoothers (EnKS) [21, 46] are particularly popular in geoscience as they provide flexible assimilation strategies for high-dimensional states. They rely on the assumption that the filtering and smoothing posteriors are multivariate Gaussian distributions, such that the following forward and backward recursions are derived.

We describe here the stochastic EnKF algorithm proposed by [21] in which observations are treated as random variables.

Algorithm 4 The Ensemble Kalman Filter algorithm

1: Input: $\mathbf{x}^b$ and $\mathbf{B}$ parameters of the prior Gaussian distribution

2: Generate vectors $\mathbf{x}_i^f(1)\ \forall i \in \{1, ..., N\}$ using a multivariate Gaussian random generator with mean vector $\mathbf{x}^b$ and covariance matrix $\mathbf{B}$. The index i of the state vector corresponds to the i^{th} realization of the Monte Carlo procedure (called member or particle).

3: Set $t = 1$

4: **Prediction step:**

- Apply the dynamical operator to each member of the ensemble following (1.1) to generate $\mathbf{x}_i^f(t)$
- The forecast state is represented by the sample mean $\mathbf{x}^f(t)$ and the sample covariance $\mathbf{P}^f(t)$.

5: **Analysis step:**

- Following (1.2), N samples of $\mathbf{y}_i^f(t)$ are generated from a multivariate Gaussian random generator with mean $\mathbf{H}\mathbf{x}_i^f(t)$ and covariance $\mathbf{R}$.
- The observations are then used to update the N members of the ensemble as $\mathbf{x}_i^a(t) = \mathbf{x}_i^f(t) + \mathbf{K}^a(t)(\mathbf{y}(t) - \mathbf{y}_i^f(t))$ where $\mathbf{K}^a(t) = \mathbf{P}^f(t)\mathbf{H}^T(\mathbf{H}\mathbf{P}^f(t)\mathbf{H}' + \mathbf{R})^{-1}$ is the Kalman filter gain
- The filtering posterior distribution is then represented by the sample mean $\mathbf{x}^a(t)$ and the sample covariance $\mathbf{P}^a(t)$.

6: Set $t = t + 1$ then go back to step 4

A classical Ensemble Kalman smoother, closely related to Rauch-Tung-Striebel smoother (see [35] for more details) is described: Given the forward recursion, the backward recursion starts from time $t = T$ with filtered state, $\forall i \in \{1, ..., N\}$, such as $\mathbf{x}_i^s(T) = \mathbf{x}_i^a(T)$ and $\mathbf{P}^s(T) = \mathbf{P}^a(T)$. Then, we proceed backward from $t = T - 1$ to $t = 1$. At each time t, we compute $\mathbf{x}_i^s(t) = \mathbf{x}_i^a(t) + \mathbf{K}^s(t)(\mathbf{x}_i^s(t+1) - \mathbf{x}_i^f(t+1))$ where $\mathbf{K}^s(t) = \mathbf{P}^a(t)\mathcal{M}^T(\mathbf{P}^f(t+1))^{-1}$ is the Kalman smoother gain. Note that we empirically estimate $\mathbf{P}^a(t)\mathcal{M}^T$ as the sample covariance matrix of the ensemble members as in [122] or [140] in the case of a nonlinear operator $\mathcal{H}$. The smoothing posterior distribution is represented by the sample mean $\mathbf{x}^s(t)$ and the sample covariance $\mathbf{P}^s(t)$.

The asymptotic behavior of the EnKF is studied in [90]. The authors show that the EnKF solution converges to the classical Kalman Filter in the linear Gaussian case, however, in the

non-linear and non-necessarily Gaussian case, the EnKF converges toward a distribution different than the optimal filtering distribution. A hybrid scheme was proposed in [119] that combines ideas from the EnKF and particle filtering schemes, their algorithm called the weighted EnKF outperforms the classical EnKF in various tests, and with a comparable computational complexity. But despite its limitations, the EnKF keeps attracting research interest given its simple implementation and its success in different oceanic and atmospheric operational settings.

1.2.2 Optimal Interpolation

Unlike the previous algorithm where data assimilation is done dynamically, Optimal Interpolation (OI) aims at finding the Best Linear Unbiased Estimator (BLUE) of a field $\mathbf{x}$ given irregularly sampled observations $\mathbf{y}^o$ in space and time and a background prior $\mathbf{x}^b$. The multivariate OI equation was derived in [57] for meteorology and numerous applications in oceanography have been reported since the early work of [17]. Several works used OI to grid sea level anomalies using along-track data (e.g. [37,92]) and it is the method adopted in CMEMS altimetry product.

Considering the following assumptions

- $\mathbf{x}^b = \mathbf{x} + \epsilon^b \quad \epsilon^b$ is the background error
- $\mathbf{y}^o = \mathbf{H}\mathbf{x} + \epsilon^o \quad \epsilon^o$ is the observational error, $\mathbf{H}$ assumed here to be linear is a matrix mapping $\mathbf{x}$ to the observation space.
- Observation and background errors are uncorrelated
- Error covariance matrices $\mathbf{B}$ and $\mathbf{R}$ respectively for background and observations are assumed to be known.

OI aims to solve the following BLUE problem:

$$\mathbf{x} = \mathbf{x}^b + \mathbf{K}(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b) \tag{1.28}$$

The BLUE formula for the optimal weight matrix $\mathbf{K}$ (also called the Kalman gain) is obtained as:

$$\mathbf{K} = E[(\mathbf{x} - \mathbf{x}^b)(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b)^T]E[(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b)(\mathbf{y}^o - \mathbf{H}\mathbf{x}^b)^T]^{-1} \tag{1.29}$$

which can be also written as

$$\mathbf{K} = E[(-\epsilon^b)(\epsilon^o - \mathbf{H}\epsilon^b)^T]E[(\epsilon^o - \mathbf{H}\epsilon^b)(\epsilon^o - \mathbf{H}\epsilon^b)^T]^{-1} \tag{1.30}$$

Since observation and background errors are uncorrelated we can further expand:

$$\mathbf{K} = E[\epsilon^b(\epsilon^b)^T]\mathbf{H}^T(E[\epsilon^o(\epsilon^o)^T] + \mathbf{H}E[\epsilon^b(\epsilon^b)^T]\mathbf{H}^T)^{-1} \tag{1.31}$$

$$\mathbf{K} = \mathbf{B}\mathbf{H}^T(\mathbf{R} + \mathbf{H}\mathbf{B}\mathbf{H}^T)^{-1} \tag{1.32}$$

It is easy to notice that previous equation is the same as the Kalman gain expression of the classical Kalman filter illustrated in Equation 1.16. It might be also relevant to note that this result could be also found using the variational formulation called 3D-VAR [101]. It resorts to minimizing the following cost function:

$$J(x) = (\mathbf{x} - \mathbf{x}^b)^T\mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + (\mathbf{y}^o - \mathbf{H}\mathbf{x})^T\mathbf{R}^{-1}(\mathbf{y}^o - \mathbf{H}\mathbf{x}) \tag{1.33}$$

An advantage of OI over 3D-var is that it gives also $\mathbf{P}^a$ the error covariance of the result (called also the analysis covariance):

$$\mathbf{P}^a = (\mathbf{I} - \mathbf{K}\mathbf{H})\mathbf{B} \tag{1.34}$$

Equations 1.28, 1.29 and 1.34 represent the full set of OI equations.

In geoscience, an important aspect which makes OI popular is the possibility of using localization *i.e.* the value of the interpolated field $x(s,t)$ at location s on time t depends on a small set of observations $\mathbf{y}^o_{i\in\{1,..N\}}$ present in a space-time volume surrounding it. This helps reducing memory and time constraints but needs modeling efforts and parameter tuning.

1.3 Analog forecasting

Analog forecasting is among the very first data-driven techniques used in weather forecasting. Its underlying idea consists in looking for one or many similar situations of the current state that occurred in the past, called *analogs*, then retrieve the *successors* in time of these situations and finally assume that the forecast can be estimated from these successors. Performing analog forecasting needs mainly an archive of historical data and a distance measure.

Even before the start of wide use of computers, some works considered analog forecasting for assessing short-term weather variation in the 50's [44]. Its intuitive and simple formulation

encouraged its adoption by researchers when the early computers were introduced to this field. Probably the most popular and application of analog forecasting was the application for atmospheric predictability by Lorenz in 1969 [103], since then the analog forecasting method was used for several atmospheric, oceanic and climate applications [145], but with the improvements in model integration capabilities, analog-related research dropped significantly overtaken by physically-derived models. However, the idea kept living thanks to some few researchers waiting for the geoscience field to enter the Big data era. In very recent years, the analog forecasting idea started again to attract researchers from not only geoscience but also from data science community, this blend of skills represents an opportunity to advance and reevaluate the method.

A well-known debate has always been surrounding the adoption of analog forecasting methods, the subject of debate relates to the "impossibility" of finding a true analog. Lorenz mentioned that likelihood of finding perfect analogs is small [102], and this was later confirmed by Van den Dool, who also derived an expression to calculate the length of the historical data needed to find a matching analog [151]. In the statistics community, where the analog method is closely related to the K-Nearest Neighbors (KNN) algorithm, it is known that KNN is plagued by the curse of dimensionality *i.e.* fails in high dimensions. Reducing the dimensionality of the problem is a classical strategy used in statistics and pattern recognition to avoid the curse. Literature in dimensionality reduction algorithms is rich, and the most popular algorithm is certainly Principal Component Analysis (PCA). Back to meteorology, this was used in several research papers such as Barnett and Preisendorf [12] where they circumvented the high dimensions through the use of a "climate state vector" which is a projection of a state set of descriptors onto a reduced space using Empirical Orthogonal Functions (EOF), which is the equivalent of PCA in statistics.

Even if we do not consider it in this thesis, it is worthy to mention another classical use of analog forecasting methods. *Analog post-processing* is a way to combine analog methods to numerical weather models. The steps of the analog post-processing consist in first obtaining the forecasts using the numerical model, then retrieving the analog of each forecast, and finally considering the observations corresponding to these analogs to be similar to the observations at the situations forecasted by the numerical model.

1.4 Discussion and conclusion

Over the recent years, the breakthroughs in data storage and computational capacities motivated the increase of research efforts in data-driven methods in general and especially in statistical

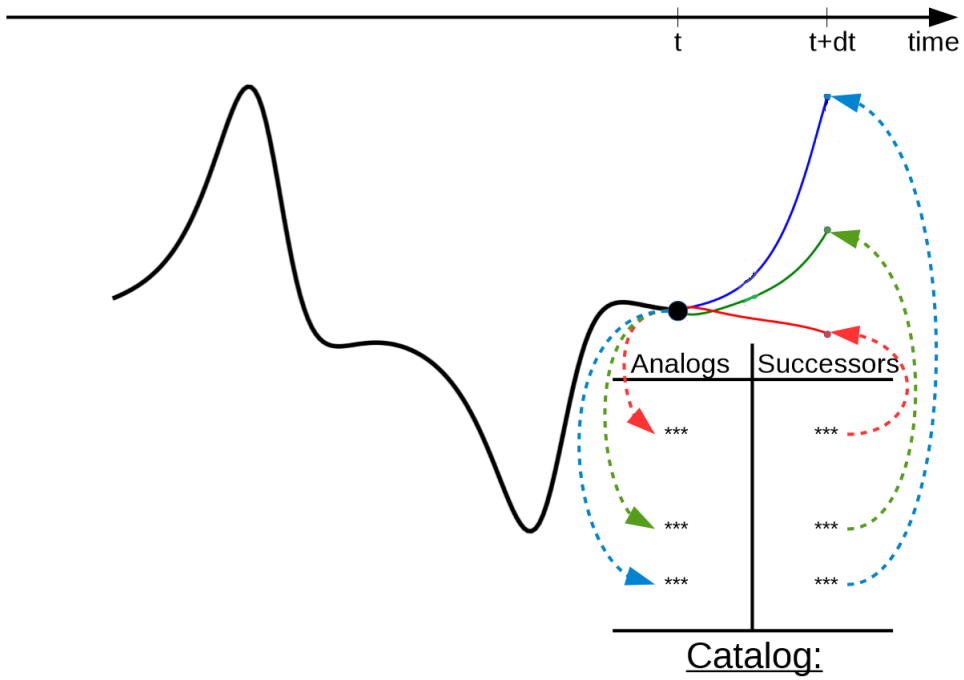


Figure 1.3 – A sketch of the idea behind analog forecasting

learning methods. The story of the revival of neural networks [143] after the Artificial Intelligence (AI) winter[2] and their ongoing impressive results thanks to a technique called deep learning [93], were a catalyzing moment that motivated our interest in data-driven methods but for data assimilation. The next chapter presents the core and most important message this thesis wants to send: An old and not so complicated data-driven method of the numerical weather prediction science community, the analog forecasting, can be plugged in a data assimilation scheme. By learning from historical data, the analog forecasting could mimic the transition equation in a classical data assimilation formulation. Given the results obtained and described all over this thesis, we hope that *the analog data assimilation* would end the analog forecasting winter.

[2] `https://en.wikipedia.org/wiki/AI_winter`

Chapter 2

The Analog Data Assimilation

When the past no longer illuminates the future, the spirit walks in darkness.

Alexis de Tocqueville

Note: Parts of the results described in this chapter have been published as: Lguensat et al., The Analog Data Assimilation, Monthly Weather Review 2017, AMS holds the copyright. [1]

[1] Copyright 2017 American Meteorological Society (AMS). Permission to use figures, tables, and brief excerpts from this work in scientific and educational works is hereby granted provided that the source is acknowledged. Any use of material in this work that is determined

2.1 Data-driven data assimilation

In the previous chapter we presented the classical model-driven data assimilation. Here, we propose an assimilation framework which relies on a similar state-space formulation to model-based data assimilation. Except that, we substitute the explicit dynamical model $\mathcal{M}$ in (1.1) by a "data-driven" dynamical model involving an analog forecasting operator, denoted by $\mathcal{A}$, namely,

$$\mathbf{x}(t) \quad = \quad \mathcal{A}\left(\mathbf{x}(t-1), \boldsymbol{\eta}(t)\right). \tag{2.1}$$

Henceforth, the state-space model (2.1-1.2) will be referred to as Analog Data Assimilation (AnDA). A sequential and stochastic data assimilation scheme is used involving different Monte Carlo realizations of the state at each assimilation time. We sketch the proposed AnDA methodology for one realization in Figure 2.1.

The analog forecasting operator $\mathcal{A}$ requires the existence of a representative dataset of exemplars of the considered dynamics. This dataset is referred to as the *catalog* and denoted by $\mathcal{C}$. The reference catalog is formed by pairs of consecutive state vectors, separated by the same time lag. The second component of each pair is referred to as the successor of the first component hereafter. The catalog may be issued from observational data as well as from numerical simulations. In the last case, one can have a catalog issued from numerical simulations (based on physical equations), and wants to perform data assimilation without running the model again. This is for instance useful for operational prediction centers which do not have the computational resources to integrate a forecast model, but do have access to a large database of numerical simulations or analysis data of a large prediction center. In this respect, we discuss also the situation where the catalog comprises noisy versions of the true states.

Given a catalog $\mathcal{C}$, the analog forecasting operator $\mathcal{A}$ is stated as an exemplar-based statistical emulator of the state $\mathbf{x}$ from time t to time $t + dt$. For any state $\mathbf{x}(t)$, we emulate the following state at time $t + dt$ based on its nearest neighbors in catalog $\mathcal{C}$. Given the analog forecasting operator, we present associated stochastic assimilation schemes, namely the *Analog Ensemble*

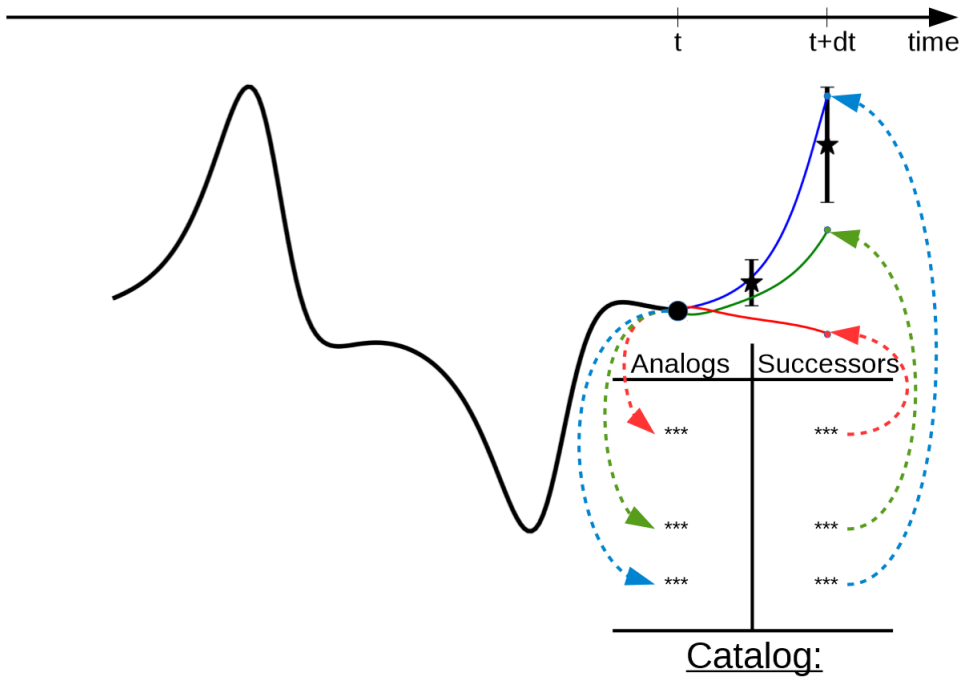


Figure 2.1 – Key principle of the Analog Data Assimilation (AnDA) framework: It consists in implicitly representing the dynamics of the system from exemplars of historical datasets. A catalog with different simulations and/or observations can be considered. Here, we plot the evolution in time of one Monte Carlo realization. The mean of the observations are shown by a black asterisk, and their variance by the corresponding error bar.

Kalman Filter/Smoother [139] and the *Analog Particle Filter*, we also present a discrete HMM-based version called the *Analog Hidden Markov Model*.

2.2 Analog forecasting strategies

2.2.1 Analog forecasting operator

Let us consider a kernel function, denoted by g, in the state-space [135]. Among the classical choices for kernels, we consider here a radial basis function (also referred to as a Gaussian kernel):

$$g(u, v) = \exp\left(-\lambda \|u - v\|^2\right). \tag{2.2}$$

with λ a scale parameter, (u, v) variables in the state-space $\mathcal{X}$, and $\|.\|$ is the euclidean distance or another appropriate distance function. Note that the proposed analog forecasting operator may be applied to other kernels or subspace reduction methods to efficiently retrieve relevant analog situations. This is discussed in Section 2.5.

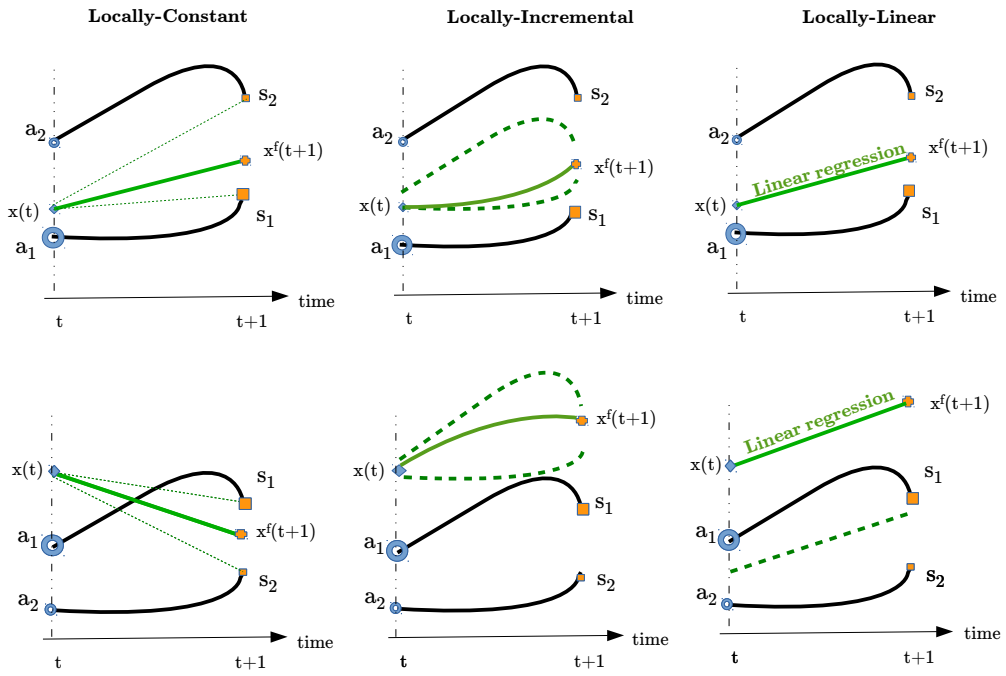


Figure 2.2 – A simplified illustration of the considered analog forecasting strategies in the case of two analogs (nearest neighbors). Two situations for the state $x(t)$ are shown: (top) a situation where $x(t)$ lies in the convex hull spanned by catalog exemplars, (bottom) a situation where $x(t)$ lies farther from its analogs. The second situation is expected to occur more often for high-dimensional space as well as for states, which are less likely. The latter may model extreme events or outliers.

Given the considered kernel, the analog forecasting operator $\mathcal{A}$ is defined as follows: for a given state $\mathbf{x}(t)$, we denote by $a_k(\mathbf{x}(t))$ its kth nearest neighbor (or analog situation) in the reference catalog of exemplars $\mathcal{C}$, and by $s_k(\mathbf{x}(t))$ the known successor of state $a_k(\mathbf{x}(t))$. Hereinafter, we refer by K to the number of nearest neighbors (analogs), and by cov_w the weighted covariance. The normalized kernel weight for every pair $(a_k(\mathbf{x}(t)), s_k(\mathbf{x}(t)))$ is given by:

$$\omega_k\left(\mathbf{x}(t)\right) = \frac{g\left(\mathbf{x}(t), a_k(\mathbf{x}(t))\right)}{\sum_{k=1}^{K} g\left(\mathbf{x}(t), a_k(\mathbf{x}(t))\right)}. \tag{2.3}$$

Several ideas can be explored to define the analog forecasting operator $\mathcal{A}$. The natural first option consists in deriving the forecast using the weighted mean of the K successors. This approach, that we call here the *locally-constant* operator, was considered in many analog forecasting related works [65,111,163], and is also known in statistics as Nadaraya-Watson kernel

regression. One can also think about using the weighted mean of the differences between the K analogs and their successors and adding it to the state to derive the forecast. The operator, referred to as *locally-incremental*, is seen as more physically-sound and relates more closely to a finite-difference approximation of the underlying differential equations. Finally, we introduce in this work a new analog forecasting operator that makes use of local linear regression techniques based on weighted least square estimates. This operator that we call the *locally-linear* operator is known to make an efficient use of small data sets and to reduce biases [33]. Note that the locally-constant and locally-incremental operators are two special cases of the locally-linear operator.

Figure 2.2 shows an illustration of the three analog forecasting operators used in this work. Hereinafter, we denote the forecasted state as $\mathbf{x}^f(t+dt)$. The three analog forecasting operators are defined as follows for two sampling schemes, namely, a Gaussian sampling and a multinomial one. Hereinafter, $\delta_Z(\cdot)$ denotes a delta function centered on Z.

- ***Locally-constant*** *analog operator*: for the Gaussian case, the forecasted state is sampled from a Gaussian distribution whose mean m_{LC} and covariance Σ_{LC} are the weighted mean and the weighted covariance estimated from the K successors and their weights.

$$\mathbf{x}^f(t+dt) \sim \mathcal{N}(m_{LC}, \Sigma_{LC}). \tag{2.4}$$

  where $m_{LC} = \sum_{k=1}^{K} \omega_k(\mathbf{x}(t)) s_k(\mathbf{x}(t))$ and $\Sigma_{LC} = cov_\omega(s_k(\mathbf{x}(t))_{k\in[\![1,K]\!]})$. While in the multinomial case, the forecasted state is drawn from the multinomial discrete distribution that samples the successor $s_k(\mathbf{x}(t))$ with a probability of ω_k

$$\mathbf{x}^f(t+dt) \sim \sum_{k=1}^{K} \omega_k\left(\mathbf{x}(t)\right) \delta_{s_k(\mathbf{x}(t))}(\cdot). \tag{2.5}$$

- ***Locally-incremental*** *analog operator*: instead of considering a weighted mean of the K successors as in the locally-constant operator, we consider the value of the current state plus a weighted mean of the K increments τ_k, *i.e.* differences between analogs and successors $\tau_k(\mathbf{x}(t)) = s_k(\mathbf{x}(t)) - a_k(\mathbf{x}(t))$. The Gaussian sampling is given by:

$$\mathbf{x}^f(t+dt) \sim \mathcal{N}(m_{LI}, \Sigma_{LI}). \tag{2.6}$$

where $m_{LI} = \mathbf{x}(t) + \sum_{k=1}^{K} \omega_k(\mathbf{x}(t))\tau_k(\mathbf{x}(t)) = \sum_{k=1}^{K} \omega_k(\mathbf{x}(t))(\mathbf{x}(t) + \tau_k(\mathbf{x}(t)))$ and $\Sigma_{LI} = cov_\omega((\mathbf{x}(t) + \tau_k(\mathbf{x}(t)))_{k\in[\![1,K]\!]})$ and the multinomial sampling resorts to

$$\mathbf{x}^f(t+dt) \sim \sum_{k=1}^{K} \omega_k\left(\mathbf{x}(t)\right) \delta_{\mathbf{x}(t)+\tau_k(\mathbf{x}(t))}\left(\cdot\right). \tag{2.7}$$

- ***Locally-linear*** *analog operator*: at each current state, we fit a multivariate linear regression between the K analogs and their corresponding successors using weighted least square estimates (see [33]). We obtain regression matrix $\alpha(\mathbf{x}(t))$ and intercept $\beta(\mathbf{x}(t))$ parameters, and residuals $\xi_k(\mathbf{x}(t)) = s_k(\mathbf{x}(t)) - (\alpha(\mathbf{x}(t))a_k(\mathbf{x}(t)) + \beta(\mathbf{x}(t)))$. The Gaussian sampling comes to:

$$\mathbf{x}^f(t+dt) \sim \mathcal{N}(m_{LL}, \Sigma_{LL}). \tag{2.8}$$

  with $m_{LL} = \alpha(\mathbf{x}(t))\mathbf{x}(t) + \beta(\mathbf{x}(t))$ and $\Sigma_{LL} = cov(\xi_k(\mathbf{x}(t))_{k\in[\![1,K]\!]})$, while the multinomial sampling is given by:

$$\mathbf{x}^f(t+dt) \sim \sum_{k=1}^{K} \omega_k\left(\mathbf{x}(t)\right) \delta_{m_{LL}+\xi_k(\mathbf{x}(t))}\left(\cdot\right). \tag{2.9}$$

The choice of one operator over another depends mostly on the computational resource and the complexity of the application. Locally-constant and locally-increment operators are less time and memory consuming than the locally-linear operator, and while they can be of comparable performance in case of a flat regression function, the locally-linear is expected to better deal with curvier regression functions at the expense however of the requirement of a larger number of analogs to fit the regression [66]. Note also that the locally-linear and the locally-incremental are more suitable for samples near or outside the boundary of the select analogs (as depicted in Figure 2.2), this may be particularly relevant in geoscience applications where chaos and extreme events are of high interest.

2.2.2 Global and local analogs

The global analog strategy is the direct application of the introduced analog forecasting strategies to the entire state vector. We also introduce a local analog forecasting operator. For a given state $\mathbf{x}(t)$, the analogs $a_k(\mathbf{x}_l(t))$ in the reference catalog, and their associated successors $s_k(\mathbf{x}_l(t))$ for each component l of the state $\mathbf{x}(t)$ are defined according to a component-wise local neighborhood. The evaluation of the kernel function and the computation of the associ-

ated normalized weights $\omega_k(\mathbf{x}_l(t))$ involve only a portion of the state vector $\mathbf{x}(t)$ defining some component-wise local neighborhood around the l^{th} component of the state vector (typically $\{\mathbf{x}_{l-\nu}(t), \ldots, \mathbf{x}_l(t) \ldots, \mathbf{x}_{l+\nu}(t)\}$ with ν the width of the considered component-wise neighborhood).

The idea of using local analogs is motivated by the fact that points tends to scatter far away from each other in high dimensions, which make the search for skillful analogs nearly impossible. For instance, [151] has shown that finding a relevant analog at synoptic scale over the Northern Hemisphere for atmospheric data would require 10^{30} years of data to match the observational errors at that time. Conversely, he also hinted that lower degrees of freedom of the states lead to better analog forecasting performance. Following this analysis, the analog forecasting of the global state is split as a series of local and low-dimensional analog forecasting operations. Note that such local analogs also help reducing possibly spurious correlations.

2.3 Analog data assimilation

The analog data assimilation is stated as a sequential and stochastic assimilation scheme, using Monte Carlo methods. It amounts to estimating the so-called filtering and smoothing posterior likelihoods, respectively $p(\mathbf{x}(t)|\mathbf{y}(1), \ldots, \mathbf{y}(t))$ the distribution of the current state knowing past and current observations and $p(\mathbf{x}(t)|\mathbf{y}(1), \ldots, \mathbf{y}(T))$ the distribution of the current state knowing past, current and future observations. We investigate both Ensemble Kalman filter/smoother and particle filter.

2.3.1 Analog Ensemble Kalman Filter and Smoother (AnEnKF/AnEnKS)

The AnEnKF and AnEnKS equations are equivalent to those of the EnKF and EnKS described in 1.2.1, except for the update step where we use the analog forecasting operator.

Algorithm 5 The Analog Ensemble Kalman Filter algorithm

1: Input: $\mathbf{x}^b$ and $\mathbf{B}$ parameters of the prior Gaussian distribution

2: Generate vectors $\mathbf{x}_i^f(1)\ \forall i \in \{1, ..., N\}$ using a multivariate Gaussian random generator with mean vector $\mathbf{x}^b$ and covariance matrix $\mathbf{B}$. The index i of the state vector corresponds to the i^{th} realization of the Monte Carlo procedure (called member or particle).

3: Set $t = 1$

4: **Prediction step:**

- Apply the **analog forecasting** $\mathcal{A}$ operator to each member of the ensemble following (2.1) to generate $\mathbf{x}_i^f(t)$
- The forecast state is represented by the sample mean $\mathbf{x}^f(t)$ and the sample covariance $\mathbf{P}^f(t)$.

5: **Analysis step:**

- Following (1.2), N samples of $\mathbf{y}_i^f(t)$ are generated from a multivariate Gaussian random generator with mean $\mathbf{H}\mathbf{x}_i^f(t)$ and covariance $\mathbf{R}$.
- The observations are then used to update the N members of the ensemble as $\mathbf{x}_i^a(t) = \mathbf{x}_i^f(t) + \mathbf{K}^a(t)(\mathbf{y}(t) - \mathbf{y}_i^f(t))$ where $\mathbf{K}^a(t) = \mathbf{P}^f(t)\mathbf{H}^T(\mathbf{H}\mathbf{P}^f(t)\mathbf{H}' + \mathbf{R})^{-1}$ is the Kalman filter gain
- The filtering posterior distribution is then represented by the sample mean $\mathbf{x}^a(t)$ and the sample covariance $\mathbf{P}^a(t)$.

6: Set $t = t + 1$ then go back to step 4

A classical Kalman smoother, here, Rauch-Tung-Striebel smoother (see [35] for more details) is described: Given the forward recursion, the backward recursion starts from time $t = T$ with filtered state, $\forall i \in \{1, ..., N\}$, such as $\mathbf{x}_i^s(T) = \mathbf{x}_i^a(T)$ and $\mathbf{P}^s(T) = \mathbf{P}^a(T)$. Then, we proceed backward from $t = T - 1$ to $t = 1$. At each time t, we compute $\mathbf{x}_i^s(t) = \mathbf{x}_i^a(t) + \mathbf{K}^s(t)(\mathbf{x}_i^s(t + 1) - \mathbf{x}_i^f(t + 1))$ where $\mathbf{K}^s(t) = \mathbf{P}^a(t)\mathcal{M}^T(\mathbf{P}^f(t + 1))^{-1}$ is the Kalman smoother gain. Note that we empirically estimate $\mathbf{P}^a(t)\mathcal{M}^T$ as the sample covariance matrix of the ensemble members as in [122] or [140] in the case of a nonlinear operator $\mathcal{H}$. The smoothing posterior distribution is represented by the sample mean $\mathbf{x}^s(t)$ and the sample covariance $\mathbf{P}^s(t)$. We note that the following way of extending EnKF and EnKS to become analog-based algorithms can be applied in the same way to other flavors of EnKF such as the square-root ensemble Kalman Filter

(EnSRF). We chose stochastic ensemble-based Kalman filters and smoothers as an illustration in this work, even if they are not the first choice in practice for atmospheric and oceanic applications due to issues related to perturbing observations with noise [16]. Besides, the work of [74] where the authors address this issue, suggests that the stochastic EnKF worths a reevaluation for oceanic and atmospheric applications.

2.3.2 Analog Particle Filter (AnPF)

Algorithm 6 The Analog Particle filter algorithm (AnPF)

1: Input: $\mathbf{x}^b$ and $\mathbf{B}$ parameters of the prior Gaussian distribution

2: Generate vectors $\mathbf{x}_i^f(1)\ \forall i \in \{1, ..., N\}$ using a multivariate Gaussian random generator with mean vector $\mathbf{x}^b$ and covariance matrix $\mathbf{B}$. The index i of the state vector corresponds to the i^{th} realization of the Monte Carlo procedure (called member or particle).

3: Set $t = 1$

4: **Prediction step:**

- Apply the **analog forecasting operator** $\mathcal{A}$ to sample new particles $\mathbf{x}_i^f(t)\ \forall i \in \{1, ..., N\}$ from previous filtered particles $\mathbf{x}_i^a(t-1)$
- Compute particle weights $\pi_i(t)$ as

$$\pi_i(t) \propto \phi\left(\mathbf{y}(t) - \mathbf{H}\mathbf{x}_i^f(t); \mathbf{R}\right), \tag{2.10}$$

  where $\phi(\cdot; \mathbf{R})$ is a centered multivariate Gaussian distribution with covariance $\mathbf{R}$.
- Normalize weights $\pi_i(t)$ to total one.

5: **Resampling step:**

- Resample from the multinomial distribution defined by the particles $\{\mathbf{x}_i^f(t)\}$ and their corresponding weights $\{\pi_i(t)\}$.
- Compute the analyzed state $\mathbf{x}^a(t)$ as the sample mean

$$\mathbf{x}^a(t) = \frac{1}{N}\sum_{i=1}^{N} \pi_i(t)\mathbf{x}_i^f(t). \tag{2.11}$$

  but one may also consider as filtered state the posterior mode.

6: Set $t = t + 1$ then go back to step 4

We also implement particle filtering techniques for the proposed analog data assimilation strategy. Given an analog forecasting operator $\mathcal{A}$, we consider an application of the Bootstrap particle filter [152], Algorithm 6 is similar to what we presented in section 1.1.2 apart from the application of the analog forecasting operator in the prediction step.

In theory, particle smoothers may also be considered. Different strategies have been proposed in the past but they showed numerical instabilities in preliminary experiments with the considered analog forecasting operator. We do not further detail the considered implementation but discuss these aspects in Section 2.5.

2.3.3 Analog Hidden Markov Models (AnHMM)

The AnHMM is presented here for a complete vision of the AnDA algorithms but not considered in the experiments shown in this chapter, results relating to AnHMM are to be found in the next chapter.

Unlike the classic state space formulation where $\mathbf{x}(t)$ is a continuous variable, the Analog Hidden Markov Model setting relies on the discrete state space formed by the set of analogs $\mathcal{D}_a$ and successors $\mathcal{D}_s$. Thereby the possible values of $\mathbf{x}(t)$ are restricted to $\mathcal{S} = \mathcal{D}_a \cup \mathcal{D}_s$. The considered exemplar-based state-space model is stated as a discrete HMM with a large number of discrete states. We resort to the Analog HMM characterized by its states $\mathcal{S}$ and by parameters $\Lambda_a = (A, B, \pi_1)$:

$$\begin{cases} \mathbf{x}(t) = s_j | \mathbf{x}(t-1) = s_i \sim A = \{a_{ij}\} \\ \mathbf{y}(t) = y_t | \mathbf{x}(t) = s_j \quad \sim B = \{b_j(y_t)\} \end{cases} \tag{2.12}$$

The parameterization of the transition matrix relies on the determination of transitions between the states. We consider a sparse parameterization of the transition matrix, where each state $s_i \in \mathcal{S}$ involves K possible transitions as follows:

- We search for the K-nearest neighbors of s_i in set $\mathcal{D}_a$ according to a predefined kernel in the state space.
- Let $\{s_n\}_{n \in \mathcal{I}(i)}$ denote the K nearest neighbors (analogs) of s_i, where $\mathcal{I}(i) = \{i_1, i_2, \cdots, i_K\}$ contains the K indices of these analogs. From catalog $\mathcal{C}$, we retrieve their successors $\{s_n\}_{n \in \mathcal{F}(\mathcal{I}(i))}$. $\mathcal{F}$ denotes the operator mapping each analog index to the index of its successor.

- the transition probabilities $a_{ij} = P(X_t = s_j | X_{t-1} = s_i)$ from state $s_i \in \mathcal{S}$ to state $s_j \in \mathcal{S}$ are non-null for successors $\{s_n\}_{n \in \mathcal{F}(\mathcal{I}(i))}$

$$a_{ij} \propto \begin{cases} \exp\left(-\lambda \| s_i - s_{i_k} \|^2\right) & \text{if } j = \mathcal{F}(i_k) \\ 0 & \text{otherwise} \end{cases} \tag{2.13}$$

  where λ can be thought as a scale parameter.

If we denote by W the cardinal of $\mathcal{S}$, the transition matrix is a $W \times W$ matrix with only $W \times K$ non-null values.

The observation matrix of the HMM directly follows from the observation model $P(\mathbf{y}(t)|\mathbf{x}(t))$. The global observation matrix is a $W \times T$ matrix $b_j(y_t) = P(\mathbf{y}(t) = y_t | \mathbf{x}(t) = s_j)$. In the reported numerical experiments, Gaussian observation models are considered:

$$b_j(y_t) \propto \exp(-\frac{(y_t - \mathbf{H}s_j)^T R^{-1} (y_t - \mathbf{H}s_j)}{2}) \tag{2.14}$$

where R is the observation covariance error matrix.

The resolution of the constructed Hidden Markov Model is done through the use of the Forward-Backward algorithm presented in section 1.1.3.

We may consider higher-order Markovian properties with a view to accounting for longer time dependencies. For a given time lag δ, it comes to consider the augmented state:

$$\hat{X} = (\mathbf{x}(t), \mathbf{x}(t-1), \mathbf{x}(t-2), \cdots, \mathbf{x}(t-(\delta-1))) \tag{2.15}$$

Creating the catalog in this case and setting the parameters of the Analog HMM follow the same steps as aforementioned.

2.4 Numerical Experiments

To evaluate the relevance and performance of the proposed analog data assimilation, we consider numerical experiments on dynamical systems extensively used in the literature on data assimilation: Lorenz-63 and Lorenz-96 models. The experiments for evaluating the effect of the size of the catalog, the impact of noisy catalogs and catalogs with parametric model error are conducted

using the Lorenz-63 model. To evaluate the global and local analog forecasting operators we use the Lorenz-96 model, an extended dynamical nonlinear system with 40 variables.

2.4.1 Chaotic models

We first consider the chaotic Lorenz-63 system. From a methodological point of view, it is particularly interesting due to its nonlinear chaotic behavior and low dimension. Several works have used this system, e.g. [5, 31, 75, 113, 122] or [153]. The Lorenz-63 model is defined by

$$\begin{aligned} \frac{\mathrm{d}x_1(t)}{\mathrm{d}t} &= \sigma\left(x_2(t) - x_1(t)\right), \\ \frac{\mathrm{d}x_2(t)}{\mathrm{d}t} &= x_1(t)\left(\gamma - x_3(t)\right) - x_2(t), \\ \frac{\mathrm{d}x_3(t)}{\mathrm{d}t} &= x_1(t)x_2(t) - \beta x_3(t). \end{aligned} \tag{2.16}$$

and behaves chaotically for certain sets of parameters, such as ($\sigma = 10, \gamma = 28, \beta = 8/3$). Here, we use the explicit (4,5) Runge-Kutta integrating method (cf. [42]). As in [153] only the first variable of the Lorenz-63 system (x_1) is observed every 8 integration time steps (i.e., with $\mathrm{d}t = 0.08$). Considering the analogy between the Lorenz-63 and atmospheric time scales, it is equivalent to a 6-hour time step in the atmosphere.

The Lorenz-96 model is another chaotic model largely used for evaluating data assimilation techniques in geophysics [2–4, 73, 118, 156]. It is defined by

$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = \left(-x_{j-2}(t) + x_{j+1}(t)\right) x_{j-1}(t) - x_j(t) + F. \tag{2.17}$$

where, $j = 1, \ldots, n$ and the boundaries are cyclic, i.e. $x_{-1}(t) = x_{n-1}(t)$, $x_0(t) = x_n(t)$ and $x_{n+1}(t) = x_1(t)$. The three right-hand side terms in (2.17) simulate respectively an advection, a diffusion and a forcing term. As in [103], we choose $n = 40$ and external forcing of $F = 8$ for which the model behaves chaotically. Equation (2.17) is solved using Runge-Kutta fourth order scheme. Observations are taken from half of the state vector (20 observed components randomly selected) every 4 time steps (i.e., $\mathrm{d}t = 0.20$).

2.4.2 Experimental details

The considered experimental setting is as follows. To avoid divergence of the filtering methods, we use $N = 100$ members/particles for the Lorenz-63 and $N = 1000$ members/particles for

the Lorenz-96 for both model-driven and data-driven strategies. We use the same covariance matrix $\mathbf{R}$ with a noise observation variance set to 2. To avoid any spin-up effect, the initial state conditions is chosen as the ground truth mean and a covariance matrix $\mathbf{B}$ with noise variance 0.1. To compare the technique performances, we use the Root Mean Square Error (RMSE) on all the components of the state vector and for all assimilation times. As training dataset for the catalog and test dataset for RMSE computation, we respectively use 10^3 and 100 Lorenz times.

The analog forecasting operator involves two free parameters, namely, K the number of nearest neighbors and λ the scale parameter of the Gaussian kernel in (2.2). Two strategies can be considered for K: either a predefined number of nearest neighbors, or a predefined threshold on distance d_{th} to select the analogs which are closer than d_{th}. For the sake of simplicity, we consider in this work the first alternative and set K to 50. Besides, we use for λ the following adaptive rule: $\lambda(x(t)) = \frac{1}{md(x(t))}$, where $md(x(t))$ is the median distance between the current state $x(t)$ and its K analogs. Note that a cross-validation procedure could be used to optimize the choice of K and λ.

2.4.3 Experiments with Lorenz-96 model

Experiment 1: The first numerical experiment consisted only in the application of analog forecasting (without assimilation) from a catalog. We build a database using Lorenz-96 equations, then we split the samples randomly to 2/3 for training the analog forecasting operators and 1/3 for test. Finally, we compare the RMSE w.r.t ground truth data as a function of Lorenz-96 time. For local analogs, we consider $\nu = 2$ the width of the considered component-wise neighborhood. Figure 2.3 shows the results of this experiment using the three choices for the analog forecasting operator $\mathcal{A}$. The locally-linear approach outperforms the two other approaches confirming that its forecasts are with lower bias compared to the other approaches. However, it also involves more parameters which increases the variance of the forecasts. This bias-variance trade-off supports the greater generalization capabilities of the locally-linear operator, when the dynamics can well be approximated locally by a linear operator.

Figure 2.3 also compares local and global analog strategies. When using locally-constant operator, local analogs are always better than global analogs. Searching for nearest neighbors on 40-dimensional vectors results most likely in irrelevant analogs. This affects heavily the locally constant operator more than the two other operators, since it computes a weighted mean of their associated successors. The locally-constant operator also limits novelty creation in the dynamics

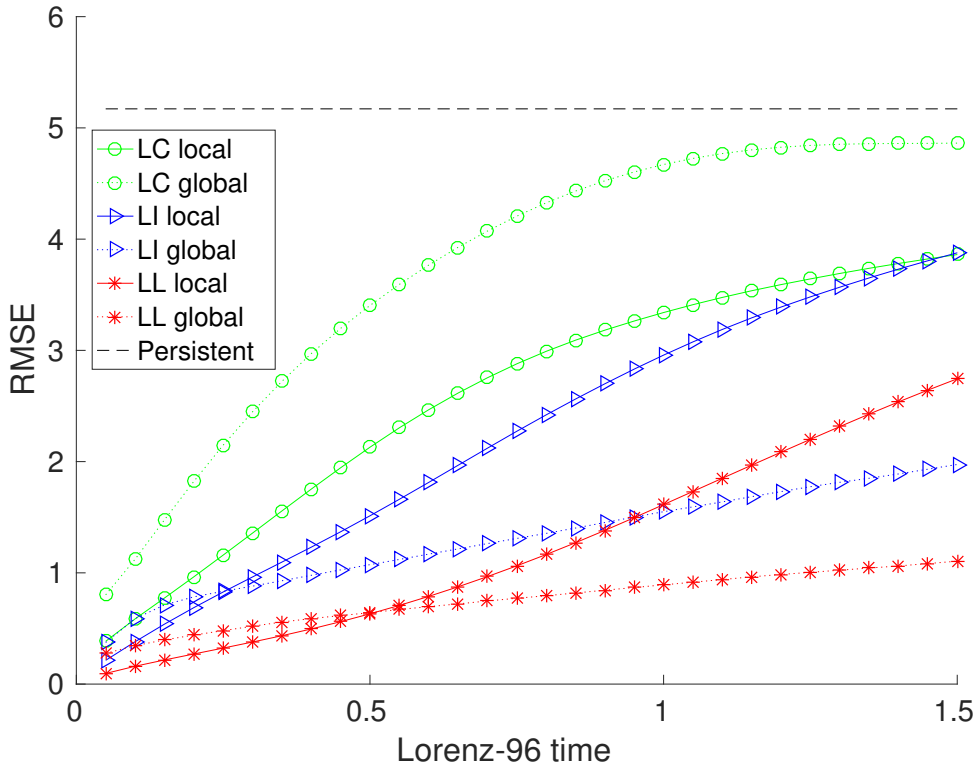


Figure 2.3 – Results of the analog forecasting performance as a function of the horizon. Different analog forecasting methods are plotted: locally-constant (green), locally-incremental (blue) and locally-linear (red) analog operators with local (straight line) and global (dashed line) analog strategies. The black dashed line corresponds to a persistent prediction over time.

by always dragging the forecast near the mean of the K successors, and, according to these experiments, it seems poorly adapted to complex and highly nonlinear systems. Regarding the locally-incremental and locally-linear strategies, local analogs are more relevant than global ones for prediction in a near future (less than 0.5 in Lorenz-96 time for locally-linear operator and less than 0.25 in Lorenz-96 time for locally-incremental).

Experiment 2: We conducted a second experiment for evaluating the impact of analog forecasting in data assimilation using the Lorenz-96 model. We run the AnEnKS with 1000 ensemble members, when only 20 variables are observed every 0.20 time steps. Figure 2.4 shows analog data assimilation experiments with the locally-linear forecasting method using the Lorenz-96 model. Figures 4a and 4b show the true state and the observations, respectively. The reconstructed state with global analogs is shown in Fig 4c and the one with local analogs in Fig 4d. The local analog data assimilation experiment clearly outperforms the global analog data assimilation experiment.

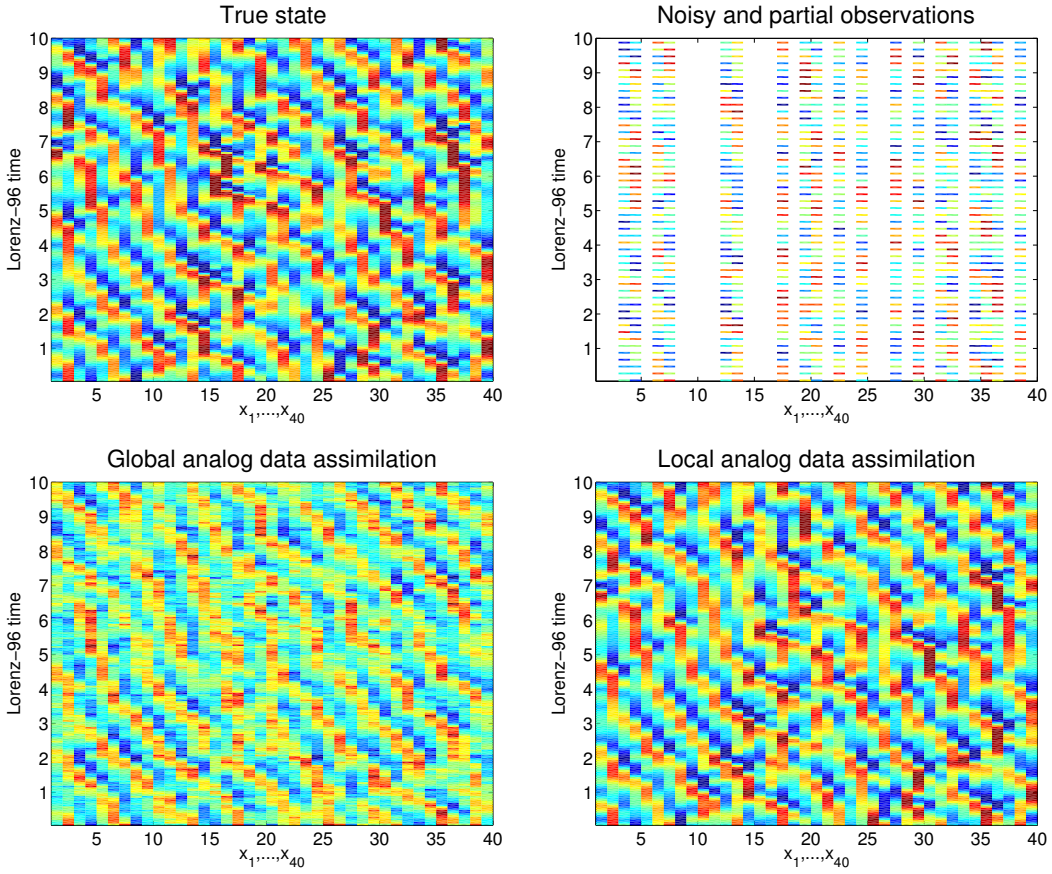


Figure 2.4 – Lorenz-96 trajectories obtained using analog data assimilation procedures with the locally-linear forecasting strategy, when only 20 variables are observed every 0.20 time steps. (top-left) True simulation of the model with 40 variables, (top-right) noisy and partial observations, (bottom-left) reconstructed state trajectories via the AnEnKS with global analogs, (bottom-right) reconstructed state trajectories via the AnEnKS with local analogs (taking into account the 5 ($\nu = 2$) nearest state components). Only the first 10 Lorenz 96 cycles are shown for better visibility.

Table 2.1 – RMSE of the reconstruction of Lorenz-96 trajectories using different forecasting strategies in the analog data assimilation procedures, when only 20 variables are observed every 0.20 time steps. The catalog size corresponds to 10^3 Lorenz-96 times (equivalent to 13 years) and the number of members/particles is N=1000.

Gaussian			
Method	Locally-constant	Locally-incremental	Locally-linear
AnEnKF	1.826	1.785	1.403
AnPF	3.174	4.224	4.4616
AnEnKS	1.320	1.287	**0.970**

Multinomial			
Method	Locally-constant	Locally-incremental	Locally-linear
AnEnKF	1.814	1.774	1.413
AnPF	2.989	4.412	4.729
AnEnKS	1.313	1.288	**1.093**

Experiment 3: A third experiment with the Lorenz-96 system was conducted. For the local analog strategy, we further compare the proposed AnDA algorithms, namely, AnEnKF, AnPF and the AnEnKS using 1000 ensemble members/particles, in Table.2.1. Two main conclusions can be drawn: i) EnKF algorithms outperform the particle filter, ii) the locally-linear analog forecasting operator gives the best reconstruction performance. We noticed that the AnPF suffers in the 40-dimensional Lorenz-96 system from sample impoverishment and degeneracy. Despite the additional experiments with different settings, for instance, w.r.t. the number of ensemble members, the number of analogs as well as using jittering (i.e. perturbing the particles with a small noise), the AnPF still suffered from the aforementioned issues.

2.4.4 Experiments with Lorenz-63 model

Experiment 1: In the proposed AnDA, the size of the catalog is expected to be a critical parameter. For Lorenz-63 dynamics, we conducted different AnDA experiments varying the size of the catalog $S = \{10^1, 10^2, 10^3, 10^4\}$ in Lorenz-63 times. We consider the same setting as in [139] where the locally-constant method with a Gaussian sampling was used for the AnEnKF, then we compare the three AnDA algorithms using 100 ensemble members/particles. As reported in Figure 2.5, the RMSE decreases when the size of the catalog increases for all AnDA algorithms. Regarding filtering-only (i.e. no smoothing) AnDA algorithms, the AnPF (blue) outperforms the AnEnKF (green). This is an expected result since particle filters handle better nonlinear models and non Gaussian probability distributions, although at a high cost in terms of computational complexity and execution time. The AnEnKS (red) clearly gives the lowest RMSE. This supports

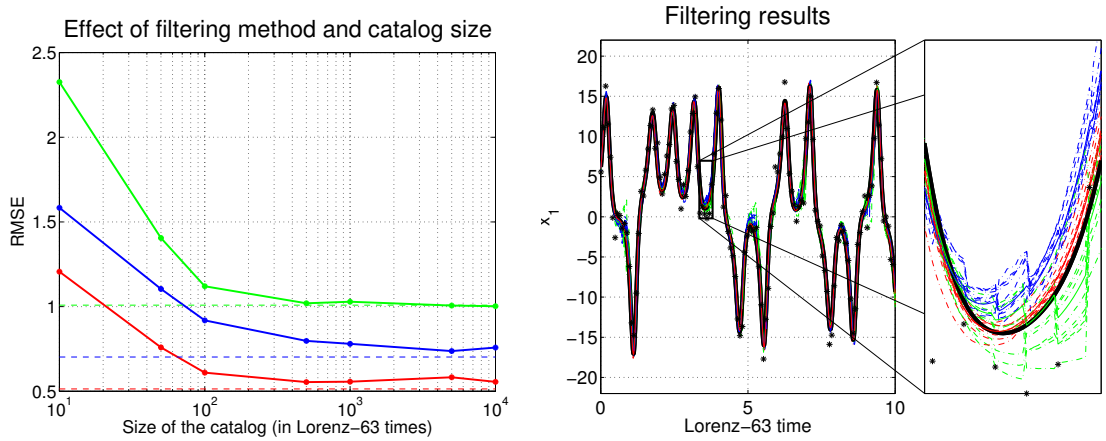


Figure 2.5 – Reconstruction of Lorenz-63 trajectories for different catalog sizes in the analog data assimilation procedures, when only the first component of the state is observed every 0.08 time steps. (Left) RMSE as a function of the size of the catalog for different analog data assimilation strategies: AnEnKF (green), AnPF (blue) and AnEnKS (red). For benchmarking purposes, data assimilation results with true Lorenz-63 equations are given in straight lines. (Right) Time series of the first component of the true state (black solid line), associated noisy observations (black asterisks), mean reconstructed series (solid lines) and 10 analyzed members/particles (dashed lines) with analog data assimilation strategies, namely AnEnKF (green), AnPF (blue) and AnEnKS (red), using a catalog of 10^3 Lorenz-63 times (equivalent to 8 years).

the additional benefit of the smoothing step performed by the AnEnKS. The zoom shown in the right panel of Figure 2.5 highlights how the smoothing step corrects the piece-wise effects resulting from the filtering step.

Experiment 2: Modeling uncertainty is a critical source of error in data assimilation. In this experiment we evaluate whether AnDA can manage a situation in which the catalog is composed by multiple numerical simulations which may have parametric model error. In (2.16), parameters γ and β define the center of the two attractors whereas σ controls the shape of the trajectories. In Figure 2.6, we depict trajectories using three set of parameters with different values for σ: $\boldsymbol{\theta}_1 = (10, 28, 8/3)$ (red), $\boldsymbol{\theta}_2 = (7, 28, 8/3)$ (blue) and $\boldsymbol{\theta}_3 = (13, 28, 8/3)$ (green). We generate three catalogs with Lorenz-63 trajectories for these three set of parameters, with 10^3 Lorenz time steps each. Merging these three catalogs into a global catalog, we apply the proposed AnDA using as observations the "true" integration resulting from Lorenz-63 model with $\boldsymbol{\theta}_1$ parameter values. As a by-product of the analog strategy, we can infer the underlying model parameterization from the observed partial observations. The reported experiments (Figure 2.6) apply the AnPF procedure with the locally-constant analog method and a multinomial sampling scheme using 100 particles. Such a choice was motivated by the desire of keeping track of the particles and

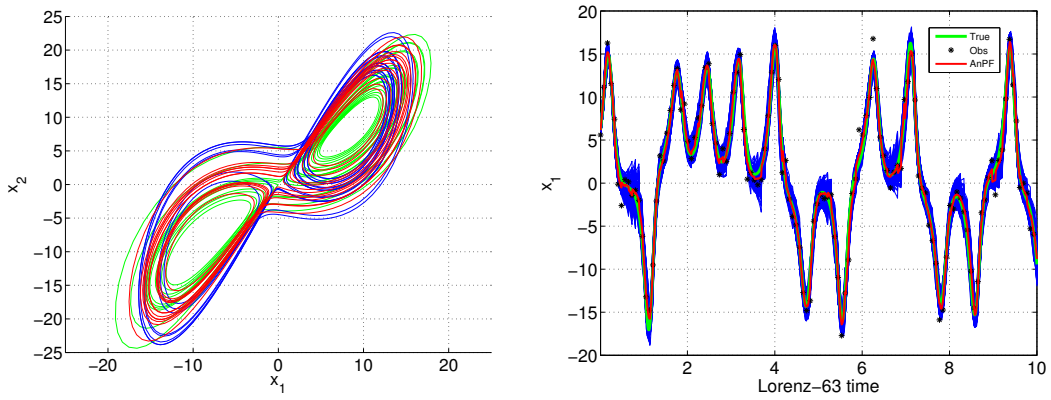


Figure 2.6 – Identification of Lorenz-63 model parameterizations using a multi-parameterization catalog in the analog data assimilation, when only the first component of the state is observed every 0.08 time step. (Left) Examples of Lorenz-63 trajectories generated with three different parameterizations: $\boldsymbol{\theta}_1 = (10, 28, 8/3)$ (red), $\boldsymbol{\theta}_2 = (7, 28, 8/3)$ (blue) and $\boldsymbol{\theta}_3 = (13, 28, 8/3)$ (green). (Right) Result of the AnPF on the first Lorenz-63 variable using the 3 catalogs associated with parameterizations $\{\boldsymbol{\theta}_i\}_{1,2,3}$ for 3×10^3 Lorenz-63 times (equivalent to 3×8 years) when only observations from parameterization $\boldsymbol{\theta}_1 = (10, 28, 8/3)$ are provided. The figure shows the AnPF particles trajectories (blue), the AnPF result (red) and the true trajectory (green).

their source catalog, which is harder to achieve with the other AnDA algorithms, since the particles would be elements from the catalog and the AnPF assigns a weight to each particle. This make it easier to select at each time the particle with the biggest weight and to know from which catalog it came from.

At every assimilation time step, we determine which parameterization most ensemble members come from, and then calculate the proportion of the presence of each parameterization. As expected, the true parameterization (red, parameterization $\boldsymbol{\theta}_1$) is more represented. The proportions for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ are respectively around, 60%, 16% and 24% proving the ability of the methodology to detect the source of the noisy and partial observation (here, only coming from $\boldsymbol{\theta}_1$). In order to analyze more the results, we calculate the RMSE of the reconstruction using: i) the three catalogs as shown before, ii) only the good catalog, iii) only the two "bad" catalogs. The RMSEs are respectively i) 1.287, ii) 1.207, iii) 1.424. These results show that having other catalogs with different parameterization degrade the RMSE but the filter is still performing well. This experiment gives insights on the problem of the assimilation of variables that may switch between different dynamical modes. Analog data assimilation can deal with this problem in a simpler manner than classical data assimilation, through the concatenation of the catalogs issued from different parameterizations into a single catalog.

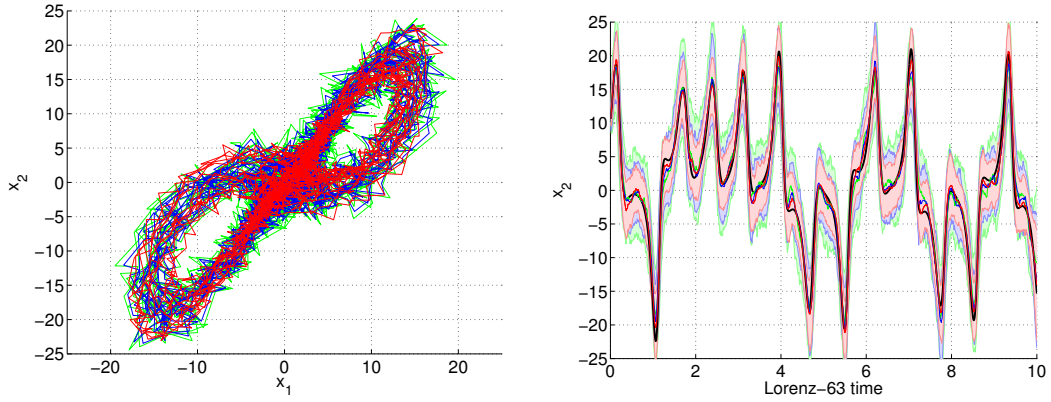


Figure 2.7 – Results of the reconstruction of Lorenz-63 trajectories from noisy catalogs: (Left) Examples of noisy Lorenz-63 trajectories for different noise levels: $\psi_1^2 = 0.5$ (red), $\psi_2^2 = 1$ (blue) and $\psi_3^2 = 2$ (green). (Right) Results of the AnEnKS using noisy catalogs corresponding to 10^3 Lorenz-63 times (equivalent to 8 years) when only observations with variance $R = 2$ are provided. We also plot the 95% confidence interval computed from the smoothing covariances.

Experiment 3: Whereas previous experiments consider catalogs produced from noise-free trajectories, we here evaluate the sensitivity of the AnDA procedures when the catalog may involve noisy trajectories of the considered system. Acquisition systems typically involve such noise patterns, which may relate for instance to both environmental constraints and measurement uncertainties. We simulate noisy catalogs for Lorenz-63 dynamics as follows: we artificially degrade the transition between consecutive states with a Gaussian additive noise. We performed experiments with different noise variances $\psi^2 = \{0.5, 1, 2\}$ to evaluate the sensitivity of AnDA procedures with respect to the signal-to-noise ratio. As illustrated in Figure 2.7, the trajectories of these experiments are extremely noisy. Table 2.2 reports the RMSE of the different AnDA algorithms with the locally-linear analog forecasting operator and 100 ensemble members/particles. As expected, the RMSE increases with the the variance of the additive noise. The AnEnKS clearly outperforms the other AnDA algorithms, which highlights its greater robustness. Figure 2.7 further illustrates that the AnEnKS is able to correctly track the true state of the system, even for highly degraded catalogs ($\psi^2 = 2$, green curve). For high signal-to-noise ratio, *i.e.* low perturbations ($\psi^2 = 0.5$, red curve), reconstructed trajectories are very close to the ones obtained with a noise-free catalog.

Table 2.2 – RMSE of the reconstruction of Lorenz-63 trajectories from noisy catalogs: we vary the variance of an additive Gaussian noise in the creation of the catalogs and apply analog data assimilation procedures with the locally-linear operator with a catalog size of 10^3 Lorenz-63 times, when only the first component of the state is observed every 0.08 time step with observation noise variance $R = 2$.

Method	$\psi_1^2 = 0.5$	$\psi_2^2 = 1$	$\psi_3^2 = 2$
AnEnKF	1.926	2.136	2.681
AnPF	1.652	1.961	2.313
AnEnKS	**1.233**	**1.561**	**2.142**

2.5 Conclusions and perspectives

This chapter demonstrates the potential of data-driven schemes for data assimilation. We propose and evaluate efficient yet simple data-driven forecasting strategies that can be coupled with classical stochastic filters (namely the Ensemble Kalman filter/smoother and the particle filter). We set a unified framework that we call analog data assimilation (AnDA). The key features of the AnDA are twofold: i) it relies on a data-driven representation of the state dynamics, and ii) it does not require online evaluations of dynamical models based on physical equations. The relevance of the AnDA is tangible when the dynamical system of interest demands tremendous and time-consuming physical modeling efforts and/or uncertainties are difficult to assess. In case when large observational or model-simulated datasets of the considered system are available, AnDA can both support or compete with classical data assimilation schemes. As a proof concept, we demonstrate the relevance of the proposed methodology to retrieve the chaotic behavior of the Lorenz-63 and Lorenz-96 models. We performed numerical experiments to evaluate critical aspects of the method, especially the relevant combinations of analog forecasting strategies and of stochastic filters as well as the exploitation of noisy and noise-free catalogs.

All the reported experiments were carried out using the AnDA Python library (available at `https://github.com/ptandeo/AnDA`) and/or the AnDA Matlab Toolbox, which includes the Lorenz-63 and Lorenz-96 systems. In the spirit of reproducible research, the user can conduct the different experiments shown in this chapter. Overall, the reported results demonstrate the relevance of the proposed analog data assimilation methods, even with highly damaged catalogs. They suggest that AnEnKS combined to locally-incremental or locally-linear analog forecasting leads to the best reconstruction performance, the locally-incremental version being the most

robust to noisy settings. Moreover, the flexibility of the analog data assimilation demonstrates the potential for the identification of hidden underlying dynamics from a series of partial observations.

The main pillar of our data-driven approach is the catalog. As such, analog data assimilation deeply relates to the quality and representativity of the catalog. In our experiments, we assumed that we were provided with large-scale catalogs of complete states of the system of interest. While catalogs built from numerical simulations fulfill this assumption, observational datasets (e.g. satellite remote sensing or *in situ* data) typically involve missing data, which may require specific strategies to be dealt with in the building of the catalogs. In this respect, local analogs obviously appear much more flexible than global ones, as partial observations provide relevant exemplars for the creation of catalogs for local analogs.

The application of analog data assimilation to high-dimensional systems is another future challenge. As detailed in [151], the number of elements in a catalog shall grow exponentially with the intrinsic dimension of the state to guarantee the retrieval of analogs at a given precision. This makes unrealistic the direct application of analog strategies to state space with an intrinsic dimensionality above 10. As a consequence, global analog forecasting operators are most likely inappropriate for high-dimensional systems. By contrast, local analogs provide a means to decompose the analog forecasting of the high-dimensional state into a series of local and low-dimensional analog forecasting operations. This is regarded as the key explanation for the much better performance reported for the local analog data assimilation for Lorenz-96 dynamics using catalogs of about a million of exemplars (Fig.2.4). For real world applications to high-dimensional systems, for instance to ocean and atmosphere dynamics, the combination of such local analog strategies to multiscale decompositions [107] arise as a promising research direction as illustrated in [52]. Such multiscale decompositions are expected to enhance the spatial redundancy, with a view to building the requested catalogs of millions to hundreds of millions of exemplars (for an intrinsic dimensionality between 4 and 7, see Appendix A) from observation or simulation datasets over a few decades. Another important aspect that controls the effective size of the catalog is the evolution of the system in time. The more nonlinear the dynamics, the greater the number of requested exemplars in the global catalog to learn the forecast operator and the spread of the prediction.

Part II

Dealing with high-dimensional fields: The Multiscale Analog Data Assimilation

Chapter 3

Interpolation of missing data in Sea Surface Temperature maps

The ocean is a mighty harmonist.

William Wordsworth

Note: Some results described in this chapter have been published as: Fablet, Viet and Lguensat, Data-driven models for the Spatio-Temporal Interpolation of Satellite derived SST fields, IEEE

holds the copyright. Hereinafter, a geophysical field will be noted using capital non-bold letters (X instead of $\mathbf{x}$*)* [1]

3.1 The Multiscale Analog Data Assimilation

3.1.1 Motivation

As depicted in the conclusion of the previous chapter. Two main features make the direct application of AnDA algorithms to spatiotemporal fields poorly efficient: their computational complexity and their ability to jointly capture large-scale and fine-scale structures. Our first application of the AnDA to ocean geophysical fields is detailed in [96] where we considering datasets from AMSRE radiometer Sea Surface Temperature (SST) observations. We applied the AnHMM to infer the interpolated SST maps, and were confronted with the problem of high dimensionality. Since AnDA successful application is affected by the curse of dimensionality, and is limited to relatively low-dimensional spaces (up to a few tens of dimensions). We explored the use of dimensionality reduction techniques. A classical and very popular method in geoscience fields is **Empirical Orthogonal Functions** (EOF), also known (in signal/image processing community) as Principal Component Analysis (PCA). PCA-based decompositions are regarded as relevant representations to encode the spatial patterns exhibited by geophysical fields. Moreover, searching for analogs of a large region decreases the chance of finding good analogs as depicted in the experiment with Lorenz96 (Chapter 2), this advocates for considering an equivalent to the idea of local analogs.

Therefore we directed our efforts to address this issue through the use of **patch-based models** that project images onto large sets of patch exemplars and/or dictionaries. Patch-based techniques are a classical tool used in the image processing/remote sensing community [48,104]. They generally involve small image patches, which help in breaking the spatiotemporal field into a "puzzle" of local regions (typically 10×10 to 20×20 for 2D images). This has the benefit of *i)* making possible the use of parallel computing, and *ii)* supports the idea of localization that was shown to be of importance for the analog data assimilation. Combining the analog data assimilation with patch-based techniques is however not sufficient, and the use of EOF-based techniques is also critical for a successful application. Concretely, the combination of Patch-based and EOF-based methods within the AnDA means that we can circumvent the curse of

[1]

dimensionality by applying AnDA several times on the projection of small images patches into lower dimensions (few tens).

Another important aspect widely known when dealing with the reconstruction of high-dimensional fields is the difficulties faced when trying to infer fine-scales. Algorithms like Optimal interpolation fail to retrieve such scales (generally less than 100km), due to the correlation length of their covariance models. This naturally calls for a multi-scale representation. Formally, we considered a model where field X was decomposed as follows:

$$X = \bar{X} + \sum_{i=1}^{J} dX_i + \xi \tag{3.1}$$

where $\bar{X}$ refers to the large-scale (low-frequency) component of X, dX_i to details at the i^{th} scale and ξ to unresolved scales. The goal of this chapter is then to show that AnDA could be a relevant candidate for the reconstruction of fine scales, and shows that AnDA could support model-based algorithms in order to achieve a better reconstruction of the geophysical field of interest.

3.1.2 Multi-scale data-driven priors

Let take hereinafter $J = 2$ in Equation 3.1. The definition of detail fields dX_1 and dX_2 combines patch-based and PCA-based representations. For scale $i = 1$ or 2, let us consider $P_i \times P_i$ patches, such that $P_1 > P_2$ (typically $P_1 = 40$ and $P_2 = 20$). We proceed as follows for the scale $i = 1$. Given $\bar{X}$ in multi-scale decomposition (3.1), each $P_1 \times P_1$ patch of detail field dX_1 is given by the projection of the associated patch for residual field $X - \bar{X}$ onto a low-dimensional PCA decomposition. This PCA decomposition is learnt from $P_1 \times P_1$ patches of a training dataset of residual fields $X - \bar{X}$. We apply the same procedure for detail field dX_2 from residual field $X - \bar{X} - dX_1$.

Formally, this leads to the following definition of detail fields dX_1 and dX_2:

$$\begin{aligned} dX_1 &= \mathbb{P}_1\left(X - \bar{X}\right) \\ dX_2 &= \mathbb{P}_2\left(X - \bar{X} - dX_1\right) \end{aligned} \tag{3.2}$$

where $\mathbb{P}_{1,2}$ are patch-based PCA image projection operators [38, 142]. They result in the decomposition of any patch $\mathcal{P}_s$ around point s at time t of detail field dX_i as a linear combination of

the principal components of the PCA for scale i:

$$dX_i(\mathcal{P}_s, t) = \sum_{k=1}^{N_E} \alpha_{i,k}(s,t) EOF_{i,k} \tag{3.3}$$

with $EOF_{i,k}$ the k^{th} principal component of the PCA at scale i and $\alpha_{i,k}(s,t)$ the associated coefficient for patch $\mathcal{P}_s$ at time t. $N_{PCA,i}$ refers to the number of vectors of the PCA basis at scale i. The spectral properties of PCA decompositions along with the lower patch size at scale $i = 2$, *i.e.* $P_1 > P_2$, lead to a scale-space decomposition [106]. Contrary to a wavelet decomposition, we only implicitly set the considered scale ranges through the number of principal components kept at each scale. The key interest here is a local adaption with point-specific PCA bases which can also account for any image geometry (e.g., the presence of land points in the considered region).

Given these definitions for detail fields dX_1 and dX_2, we considered an analog (data-driven) formulation of the associated dynamical models (3.10). As stated in the motivation section, analog dynamical models introduced in Chapter 2 do not directly apply to high-dimensional fields and we considered patch-based models. We first assumed that we were provided with representative catalogs $\mathcal{C}_{1,2}$ of patch exemplars of the dynamics of details fields dX_1 and dX_2. Each catalog is composed of a set of patch exemplars $\{dX_i(\mathcal{P}_{s_k}, t_k)\}_k$, referred to hereafter as analogs, and of their temporal successors $\{dX_i(\mathcal{P}_{s_k}, t_k + 1)\}_k$. For a given patch $\mathcal{P}_s$ and scale i, the definition of the analog dynamical model leads to the definition of an exemplar-driven sampling strategy for the distribution of the state at time t, $dX_i(\mathcal{P}_s, t)$, conditionally to the state at time $t-1$, $dX_i(\mathcal{P}_s, t-1)$. Let us denote by $\varphi_i(\mathcal{P}_s, t)$ the vector of the N_E coefficients $\alpha_{i,k}(s,t)$, which represents the projection of $dX_i(\mathcal{P}_s, t)$ in the lower-dimensional EOF space. Formally, we considered Gaussian conditional distributions of the form

$$\varphi_i(\mathcal{P}_s, t) | \varphi_i(\mathcal{P}_s, t-1) \sim \mathcal{G}\left(\mu_i(u, \mathcal{C}_i), \Sigma(u, \mathcal{C}_i)\right) \tag{3.4}$$

where $\mathcal{G}(\cdot)$ is a Gaussian distribution. Mean $\mu_i(u, \mathcal{C}_i)$ is defined as a weighted function of the successors of the K nearest-neighbor of u in catalog $\mathcal{C}_i$. Similarly, covariance $\Sigma(u, \mathcal{C}_i)$ is issued from the weighted covariance of the successors of the K nearest neighbors. Theses weights and the nearest-neighbor search involve a predefined kernel $\mathcal{K}$ as detailed below. Let us denote by $(\mathcal{A}_k(u), \mathcal{S}_k(u))$ the analog-successor pair of the kth nearest-neighbor to u in $\mathcal{C}_i$. Following Chapter 2, we investigate three different analog dynamical models corresponding to different parameterizations of the above mean and covariance:

- **Locally-constant analog model:** mean $\mu_i\left(u, \mathcal{C}_i\right)$ and covariance $\Sigma\left(u, \mathcal{C}_i\right)$ are given by the weighted mean and covariance of the K successors $\{\mathcal{S}_k(u)\}_k$.

- **Locally-incremental analog model:** it proceeds similarly to the locally-constant analog model, but for the differences between the successors and the analogs, such that mean $\mu_i\left(u, \mathcal{C}_i\right)$ is given by the sum of u and of the weighted mean of the K differences $\{\mathcal{S}_k(u) - \mathcal{A}_k(u)\}_k$. $\Sigma\left(u, \mathcal{C}_i\right)$ results in the weighted covariance of these differences.

- **Locally-linear analog model:** given the K analog-successor pairs $\{\mathcal{A}_k(u), \mathcal{S}_k(u)\}_k$, it first comes to the weighted least-square estimation of the linear regression of the state at time t given the state at time $t-1$. Denoting by $\mathbf{A}_i(s,t)$ the estimated local linear operator, mean $\mu_i\left(u, \mathcal{C}_i\right)$ is given by $\mathbf{A}_i(s,t) \cdot dX_i\left(\mathcal{P}_s, t-1\right)$ and covariance $\Sigma\left(u, \mathcal{C}_i\right)$ by the weighted covariance of the residuals of the fitted linear regression.

$$\mathcal{K}_G\left(u(t), v(t)\right) = \exp\left(-\frac{\|u(t) - v(t)\|^2}{\sigma}\right), \tag{3.5}$$

and a cone kernel $\mathcal{K}_C$, recently introduced for dynamical systems in [163]. For any pair of states $u(t), v(t)$, it leads to

$$\mathcal{K}_C\left(u(t), v(t)\right) = \exp\left(-\frac{\mathcal{L}_\zeta\left(u(t), v(t)\right)}{\sigma}\right) \tag{3.6}$$

$$\mathcal{L}_\zeta(u(t), v(t)) = \frac{\|\omega(t)\|^2 \left[\left(1 - \zeta\cos^2\theta\right)\left(1 - \zeta\cos^2\phi\right)\right]^{1/2}}{\|\partial_t u(t)\|\|\partial_t v(t)\|} \tag{3.7}$$

where $\omega(t) = u(t) - v(t)$, $\partial_t u(t) = u(t) - u(t-1)$, $\partial_t v(t) = u(t) - u(t-1)$, $cos\theta = \langle \omega(t), du(t) \rangle$ and $cos\phi = \langle \omega(t), dv(t) \rangle$. Compared to a classical Gaussian kernel, the cone kernel takes into account not only the distance between the two states, but also the alignment of their instantaneous velocities with the difference between the two states. It has been shown in [163] that the cone kernel may be more appropriate for analog forecasting schemes. For the Gaussian (resp. cone) kernels, scale parameter σ is locally-adapted to the median value of the distances $\|u(t) - v(t)\|^2$ (resp. $\mathcal{L}_\zeta\left(u(t), v(t)\right)$) to the K nearest neighbors in the catalogs of exemplars. Parameter ν is set empirically between 0 and 1. In all cases, we take advantage of the considered PCA-based representation of the patches to compute patch similarities within the associated low-dimensional spaces, and not in the original patch space.

3.2 Missing data interpolation in Sea Surface Temperature maps

Satellite-derived products are of key importance for the high-resolution monitoring of the ocean surface on a global scale. A variety of sensors record observations of geophysical parameters, such as Sea Surface Temperature (SST) [26], Sea Surface Height (SSH) [27], Ocean Color [11], Sea surface Salinity (SSS) [86], etc. In all cases, the delivery of L4 gridded products for end-users involves a number of pre-processing steps from the L1 data acquired and transmitted by spaceborne sensors. Due to both the space-time sampling geometry of satellite sensors and their sensitivity to the atmospheric conditions (e.g., rains, aerosols, clouds), ocean remote sensing data may involve very large missing data rates as illustrated in Fig.3.3. Hence, spatio-temporal interpolation is of key importance to deliver gap-free gridded sea surface fields for further analysis.

Optimal interpolation is certainly the state-of-the-art approach for the spatio-temporal interpolation of satellite-derived sea surface geophysical fields [41,45]. Optimal interpolation relies on the modeling of the covariance of the considered spatio-temporal fields. The choice of the covariance model is a critical step [20,41,137,139]. Stationary covariance hypotheses are generally considered, though they might not be verified. For instance, frontal areas as illustrated in Fig.3.5 may involve time-varying and space-varying anisotropical features. In such cases, considering mean covariance model typically results in the smoothing out of the fine-scale SST details. Data assimilation techniques for missing data interpolation may be regarded as another important category of model-driven approaches [10,46,137]. A critical aspect of their implementation lies in the choice of the dynamical model, more precisely the trade-off between its computational complexity and its ability to correctly represent real sea surface dynamics.

The tremendous amount of satellite observation data pouring from space, along with the wider availability of reanalysis and/or numerical simulation datasets supports the development of data-driven approaches as an alternative to model-driven schemes. In this respect, statistical and machine learning models offer new computational means to account for space-time variabilities that cannot be completely captured by simplified physical models. The application of Principal Component Analysis (PCA), also referred to as Empirical Orthogonal Functions (EOF) in the geoscience field, to remote sensing missing data interpolation [13,123] may be regarded as an example of such data-driven schemes, though it proves mainly relevant for large-scale variabilities

[123]. One may also cite the development of exemplar-based models in image processing and their applications to missing data interpolation for single-date remote sensing data [48, 104].

In this study, we investigate such data-driven and exemplar-based models for the spatio-temporal interpolation of missing data in ocean remote sensing time series. We aim to exploit the implicit knowledge conveyed by available multi-annual satellite-derived datasets to improve the interpolation of high-resolution spatio-temporal sea surface geophysical fields. We rely on analog data assimilation [65, 139] and develop, to our knowledge, the first application of analog data assimilation to high-dimensional spatio-temporal fields. Our methodological contributions lie in the introduction of a multiscale analog data assimilation applied to local patch-based and PCA-constrained representations. We demonstrate the relevance of the proposed scheme through an application to SST time series. We report significant gain compared to state-of-the-art approaches, namely optimal interpolation [20, 99] and PCA-based interpolation [13, 123].

3.3 Problem statement and related work

3.3.1 Model-driven approaches

As previously mentioned, model-driven approaches are the state-of-the-art techniques for the spatio-temporal interpolation of missing data in ocean remote sensing observations [41, 46]. In particular, optimal interpolation relates to the following formulation:

$$X \propto \mathcal{G}(X^b, \Gamma) \tag{3.8}$$

$$Y(t,s) = X(t,s) + \epsilon(t,s),\ \forall s \in \Omega_t \tag{3.9}$$

where $\mathcal{G}(X^b, \Gamma)$ is a spatio-temporal Gaussian field with mean background field X^b and covariance function Γ, and ϵ the observation noise assumed to be Gaussian. Ω_t refers to the region domain for which observations are truly available at time t. Given a series of observation fields Y and a known covariance function Γ, optimal interpolation leads to an analytical MAP (Maximum A Posteriori) solution for field X, equivalent to the minimization of a reweighted least-square criterion w.r.t. the covariance of noise ϵ. The choice of the covariance function Γ is a critical step. Exponential and Gaussian covariance models [20, 137] are the most classical choices with both constant parameters as well as space-time-varying parameterization [138].

When dealing with high-dimensional fields, such as ocean remote sensing observations, the numerical computation of the solution of the optimal interpolation may not be feasible, as it involves the inversion of a very large covariance matrix. Sequential approaches, such as ensemble Kalman techniques [46], are then considered. They may be restated as data assimilation formulations. Considering a discrete setting, they amount to the following model for field X:

$$X(t) = \mathcal{M}\left(X(t-1), \eta(t-1)\right) \tag{3.10}$$

where $\mathcal{M}$ is referred to as the dynamical model and η is a random perturbation. Model (3.8) may be restated according to this formulation with a linear model $\mathcal{M}$ and a Gaussian process η derived from the considered Gaussian field with covariance Γ. Other parameterizations of the dynamical model may be derived from fluid dynamics equations, including for instance advection-diffusion models [10]. Ensemble Kalman schemes [46] are the state-of-the-art techniques to numerically solve for the reconstruction of spatio-temporal field X given partial observation field Y under model (3.10). Using a sample-based representation of Gaussian distributions, they provide forward-backward filtering schemes to approximate the optimal interpolation solution. We let the reader refer to [46] and reference therein for additional details on stochastic data assimilation. We may also point out variational data assimilation [10, 89], which exploits a continuous formulations of Model (3.10) and involves a gradient-based minimization of the observation error under model (3.10).

A typical example of the optimal interpolation of an SST field from a series of partial observations is reported in Fig.3.3. An important limitation of model-driven approaches lies in modeling uncertainties. Due to the autocorrelation structure of sea surface geophysical structures and the observation sampling rate, optimal interpolation results to accurate reconstruction of the spatio-temporal fields for spatial scales larger than 100km. However, finer scales are significantly filtered out (see Fig.3.3). This property directly relates to the correlation length of the covariance model (here, 100km). This correlation is a trade-off between the spatial resolution of the observation fields (here, 5km) and the size of the gaps.

As detailed below, we explore data-driven approaches to take advantage of available observation or simulation datasets with a view to improving the reconstruction of the fine-scale structures of sea surface fields.

3.3.2 Data-driven approaches

With the increasing availability of representative observation datasets, data-driven models become more and more appealing to solve inverse image problems, including missing data interpolation. Initially mostly investigated for computer vision and computer graphics applications, such as synthesis, inpainting and super-resolution issues [36, 43], they have also gained interest for applications to remote sensing data [48, 104]. Patch-based and exemplar-based models have emerged as powerful representations to project images onto large sets of patch exemplars and/or dictionaries. Non-local means and non-local priors [19, 121] are state-of-the-art examples of such models for image reconstruction issues. Developments for multivariate time series have also recently been investigated, especially exemplar-driven data assimilation referred to in the geoscience field as analog data assimilation [65, 97, 139]. Two main features make the direct application of these exemplar-based strategies to spatiotemporal fields poorly efficient: their computational complexity and their ability to jointly capture large-scale and fine-scale structures. Patch-based techniques generally involve small image patches (typically, from 3x3 to 11x11 patches for 2D images), which cannot resolve large structures, with a typical scale greater than the width of the patches. In addition, the considered minimization schemes involve repeated iterations over the entire set of exemplars, which may make them extremely computationally-demanding for applications to spatio-temporal data. By contrast, analog data assimilation provides an efficient sequential scheme, but remains limited to relatively low-dimensional space (up to a few tens of dimensions in [65, 139]).

PCA-based models are popular in the geoscience field. They have also gained interest for application to missing data interpolation, especially DINEOF approaches [13, 123]. These involve two key steps: i) the estimation of basis functions, which provide a lower-dimensional representation of the variability spanned by the considered spatial or spatio-temporal data, ii) the interpolation of the missing data from projections onto the basis functions. VE-DINEOF [123] has recently improved compared to the original DINEOF scheme [13]. In both cases, applications to ocean remote sensing data, especially SST, were considered. Applied on a global or regional scale, the lower-dimensional PCA-based representation is mostly relevant to recover large-scale structures and not as appropriate to reconstruct fine-scale details. Overall, PCA-based decompositions are regarded as relevant representations to encode the spatial patterns exhibited by geophysical fields. It may be noted that PCA representations are also often used in patch-based image processing (see for instance [38, 142]).

3.4 Application of the patch-based AnDA

We proceed to the resolution of model (3.1). We might consider a direct discrete gradient-based numerical resolution as the considered parameterization for model (3.1) can be regarded as a spatio-temporal Markov Random field [48, 56]. This would however lead to an extremely-demanding computational scheme. We prefered to exploit the multi-scale nature of our model to develop a coarse-to-fine strategy and cast the global minimization problem as series of smaller problems, which can be solved more efficiently. More precisely, we proceeded as follows. We first solved for the reconstruction of large-scale component $\bar{X}$ using optimal interpolation with covariance model Γ. We then successively solved for the reconstruction of detail fields dX_1 and dX_2. This step runs independent resolution along the temporal dimension for each patch position using sequential data assimilation algorithms, namely an analog Ensemble Kalman Smoother (AnEnKS) and an HMM-based analog smoother (AnHMM). The independent solutions computed for each patch position were recombined using averaging. To reduce the computational complexity, we did not process all possible patch positions, but only overlapping patches (5-pixel overlapping in both directions) with a 35x35 (resp. 15x15) spatial sampling for $P_1 \times P_1$ patches (resp. $P_2 \times P_2$). To remove potential block artifacts, we apply a PCA-based decomposition-reconstruction onto 10×10 patches. As initialization for the analog data assimilation iterations, we use a VE-DINEOF solution [123].

All implementations were run under Matlab. We used [45] for optimal interpolation, and Analog Data Assimilation toolbox [97].

3.5 Results

3.5.1 Experimental setting

Considered case-study: To perform a qualitative and quantitative evaluation of the proposed framework, we used a reference gap-free L4 SST time series from which we create a SST with missing data using real missing data masks. As reference SST, we used OSTIA product delivered daily by the UK Met Office [41] with a 0.05° spatial resolution (approx. 5km) from January 2007 to April 2016. The OSTIA analysis combines satellite data provided by infrared sensors (AVHRR, AATSR, SEVIRI), microwave sensors (AMSRE, TMI) and in situ data from drifting and moored

buoys. For the missing data mask series, we studied an infrared sensor, more specifically METOP, which may involve very high missing data rates as illustrated in Fig.3.3 & 3.5.

As a case-study region, we selected an area off South Africa. This highly dynamic ocean region involves complex fine-scale SST structures (e.g., filaments, fronts) as shown in Fig.3.3. Our evaluation focused on the interpolation of the SST fields for year 2015, other years being used to build a catalog of exemplars for the analog frameworks.

Parameter setting of the proposed approaches: We performed interpolation experiments with both AnHMM and AnEnKF/KS schemes. We exploited a three-scale model: the global scale (entire region), 40x40 patches and 20x20 patches. At each scale, each patch was encoded by its PCA-based decomposition using a 10-component PCA. As initialization for missing data areas, we used an optimal interpolation on the global scale. The parameterizations of the optimal interpolation and of the DINEOF scheme were those used for comparison purposes as detailed below. In the analog setting, the number of neighbors was varied from 10 to 110 and we compared Gaussian and Cone kernels.

Comparison to state-of-the-art approaches: For comparison purposes, we consider an optimal interpolation, which is the interpolation technique used in most operational products (e.g., [41]), VE-DINEOF [123], a PCA-based technique, and a direct region-level application of the analog data assimilation. Their parameter settings were as follows:

- *Optimal interpolation (OI)*: we used a Gaussian kernel with a spatial correlation length of 100km and a temporal correlation length of 3 days. These parameters were empirically tuned for the considered dataset using a cross-validation experiment. We used the optimal interpolation package from [45]. The considered parameter setting was consistent with previous work [41, 137] and stressed the strong temporal correlation of SST field [137]. In our case-study, a direct implementation of the OI would have required a large memory: for a missing data rate of $\sim$70%, the interpolation onto the considered 300×600 grid would have required the inversion of a system of $5\text{T}.10^4$ equations with T the temporal correlation. Given the considered spatial correlation length of 100km, we achieved an optimal interpolation onto a coarser grid with a resolution of 25km and applied a bicubic interpolation onto the targeted high-resolution grid (5km resolution).

- *VE-DINEOF interpolation*: we exploited a direct implementation of VE-DINEOF scheme [123] on the regional scale using 200 PCA components, which amounted to 99.27% of the total variance of the dataset. This VE-DINEOF setting is referred to as G-VE-DINEOF.

We also considered a multi-scale version of the VE-DINEOF procedure using the same three-scale decomposition as the multi-scale analog data assimilation. As for MS-AnEnKS and MS-AnHMM, we used two detail components corresponding to 40x40 patches and 20x20 patches. At each scale, *i.e.* the coarse region scale and the two detail scale, we exploited 10-dimensional PCA decomposition ($N_{PCA,1} = N_{PCA,2} = 10$). The resolution of this multi-scale VE-DINEOF, referred to as MS-VE-DINEOF, applies a coarse-to-fine strategy, such that at each scale, the VE-DINEOF iteratively updated the missing data area from the projection of overlapping patches onto the 10-dimensional PCA basis;

- *Global AnEnKS interpolation*: to evaluate the relevance of the proposed multi-scale decomposition, we tested a direct application of the AnEnKS at the region scale, referred to as G-AnEnKS. Similarly to G-VE-DINEOF, we considered 200 PCA components, which amounted to 99.27% of the total variance of the dataset. From numerical experiments, the best parameter setting combined a locally-incremental analog forecasting with $K = 100$ neighbors and a Gaussian kernel.

It may be noted that variational interpolation techniques, based on the minimization of regularization norms [15], cannot be expected to lead to relevant results given the large missing data rates in the considered dataset (above 70% on average) and were not considered in our experiments.

Qualitative and quantitative evaluation: to assess the quality of the different interpolation schemes, we first achieved a quantitative analysis according to root mean square error (RMSE) statistics for the SST reconstructed SST fields, the associated gradient fields, and the detail fields of a 4-scale dyadic wavelet decomposition of the SST fields. We also computed radially-averaged power spectral densities to analyze the fine-scale patterns of the reconstructed field. In addition, we performed a qualitative analysis of these fields with a focus on the reconstruction of fine-scale structures.

3.5.2 Interpolation performance

We shall begin with the results of our numerical experiments. We first present the quantitative evaluation of interpolation performance, including a comparison to state-of-the-approaches. Second, we further illustrate this performance using interpolation examples. Third, we report a sensitivity analysis of the best analog assimilation setting. We also include an evaluation of

interpolation performance when the creation of the catalogs of exemplars involve observation datasets with missing data.

Quantitative comparison to state-of-the-art approaches: We first report the overall RMSE statistics of the considered interpolation approaches, namely OI, G-VE-DINEOF, MS-VE-DINEOF, MS-AnHMM, G-AnEnKS and MS-AnEnKS, in Tab.3.1. The multi-scale Analog schemes are a clear improvement over the OI and VE-DINEOF reconstruction, with a relative gain in SST RMSE up to 50% for MS-AnEnKS at the finest scale (dX_2). MS-AnHMM also leads to a significant improvement but is clearly outperformed by MS-AnEnKS. It may be noted that the direct application of the analog data assimilation, G-AnEnKS, to field X does not lead to very significant improvement. This is regarded as a direct benefit of the multi-scale decomposition, which greatly increases the representativity of the collected catalogs of exemplars. No such difference is reported for the application of global and multi-scale VE-DINEOF schemes, which further stresses the relevance of the analog dynamical prior exploited by MS-AnEnKS. The analysis of the RMSE statistics at different scales of a dyadic wavelet decomposition indicates that the improvement mainly refers to the third and fourth dyadic scales (*i.e.*, spatial scales greater than 20km). Most of the improvement is brought about by the resolution of component dX_1 (about 40% of relative gain w.r.t. OI), when component dX_2 accounts for about 10% of relative gain w.r.t. OI. The RMSE time series (Fig.3.1) lead to similar observations. Interestingly, AnEnKS depicts a lower time variability of the RMSE compared to OI and VE-DINEOF (standard deviation of 0.06 vs. 0.13), the later being more sensitive to larger missing data rates. This is viewed as a benefit of the exemplar-based time regularization conveyed by the analog framework.

Qualitative analysis of interpolation results from examples: To complement this global analysis, we report interpolation results for two dates, corresponding to relatively low ($\sim 60\%$) and greater ($\sim 90\%$) missing data rates, respectively in Fig. 3.5 and Fig. 3.3. For these two examples, we visually compare OI, MS-VE-DINEOF and MS-AnEnKS interpolations to the groundtruth both for the SST field and the gradient magnitude fields. In Fig.3.3, MS-AnEnKS clearly outperforms OI and MS-VE-DINEOF (SST (resp. SST gradient) RMSE of 0.20 (resp. 0.24) vs. 0.42 (resp. 0.40) and 0.41 (resp. 0.40)). We also highlight areas in which the improvements in the reconstruction of local SST details may be noticed. Visually, the improvement is more noticeable on the gradient amplitude. Whereas OI and MS-VE-DINEOF lead to relatively coarse SST structures, MS-AnEnKS results in finer front details, which are visually more similar

to the groundtruth. This is further emphasized by the analysis of the power spectral densities of the different fields (Fig. 3.6, left). OI clearly underestimates the spectral energy below 100km, as expected from the associated spatio-temporal smoothing with a spatial correlation length of 100km. A similar underestimation is observed for MS-VE-DINEOF for scales ranging between 70km and 150km. By contrast, MS-AnEnKS nicely matches the spectral signature of the groundtruth up to 20km. These results appear consistent with the previous observation that the improvement brought about by the analog assimilation was mainly noticeable in terms of RMSE for scales greater than 20km. The white noise plateau observed from 20km and below for the reference SST field may indicate that the OSTIA field conveys little information for scales lower than 20km for this particular date. This is further illustrated by the analysis of a one-dimensional transect at 36.525^oS accross a strong SST front in Fig.3.4. The MS-AnEnKS interpolation clearly leads to a better estimation of local SST variabilities, where OI and MS-VE-DINEOF tends to oversmooth strong gradients. Overall, the same observation holds for the second example (Fig.3.5), though the lower missing data rate (59%) slightly reduces the differences observed between the different interpolation methods.

We also illustrate the relevance of the post-processing step in the AnEnKS (Fig.3.2). The spatially-independent assimilation of overlapping patches may result in block artifacts at patch boundaries as clearly highlighted by the gradient field. The considered EOF-based filtering for 10×10 patches successfully removes most of these block artifacts and retrieves a visually consistent gradient field as discussed above. It may be noted that a different implementation of the analog assimilation using non-sequential iterative scheme for patch-based image processing [19, 48] would be an alternative, however at the expense of an increased computational complexity. By contrast, the independent assimilation of each spatial patch only involves one forward and one backward iteration, such that each space-time patch is visited only twice. We evaluate more precisely the computational complexity of the different interpolation models in Tab.3.6. MS-VE-DINEOF is clearly involves the lowest computational complexity. In this respect, given relatively similar interpolation performance, VE-DINEOF appears as a relevant alternative to OI for the interpolation of the coarse-scale component. By contrast, even if MS-AnHMM significantly reduces the computational complexity of the analog assimilation, the differences in interpolation performance reported in Tab.3.1 clearly recommend the selection of the MS-AnHMM as the relevant fine-scale analog assimilation scheme for SST fields.

Sensitivity analysis for MS-AnEnKS: Given the overall qualitative and quantitative analysis reported above, we further analyze the MS-AnEnKS setting, especially its sensitivity to the selected parameter setting. In Tab.3.2 we report RMSE statistics while varying the number of neighbors in the analog models. Tab.3.3 reports a similar analysis for different kernel parameterizations. Overall, the best parameterization combines a cone kernel [163] using 100 neighbors and a locally-incremental analog model. It might be noted that the choice of the kernel weakly affects interpolation performance. By contrast, the locally-incremental analog model significantly improves the RMSE of the locally-linear and locally-constant strategies (Tab.3.4) by about 10% and 25%. This is in accordance with the conclusions drawn in [97]. The lower performance of the locally-linear analog model may relate to an unfavourable trade-off between estimation uncertainty and local adaption. We may point out that all these parameterizations of the proposed interpolation framework outperforms both OI and MS-VE-DINEOF.

Creation of catalog $\mathcal{C}$ from observation datasets: In the experiments reported above, the catalog of patch exemplars is built from the gap-free SST time series from 2008 to 2014. This experimental setting is representative of an application context where one aim to exploit previous reanalyses and/or numerical simulations for the interpolation of upcoming observations. The key interest of the analog assimilation is to facilitate the implicit synergy between possibly computationally-expensive high-resolution models and/or reanalyses and satellite-derived observation datasets. A second application context is also investigated. We may also directly build the catalog of exemplars from the satellite-derived observation datasets, which involve missing data. To simulate this experiment, we created a representative catalog from the SST time series with the METOP missing data mask from 2008 to 2014. We proceeded similarly to the scheme described for year 2015 in Section 4.4.3. We only retained SST patches with less than 20% of missing data. We compared the resulting interpolation performance to that of the first experiment in Tab.3.5. Although lower root mean square error (RMSE) values are reported for this second experiment (0.22 vs. 0.20 in terms of root mean square error of the interpolated SST fields), the relative gain compared to OI and VE-DINEOF is still significant (0.22 vs. respectively 0.40 and 0.41). The qualitative analysis of the interpolated fields leads to conclusions similar to those drawn for the first experiment. These results further stress the relevance of the proposed data-driven approach in order to benefit either from high-resolution simulations and/or re-analyses or real satellite-derived observation datasets. It may be noted

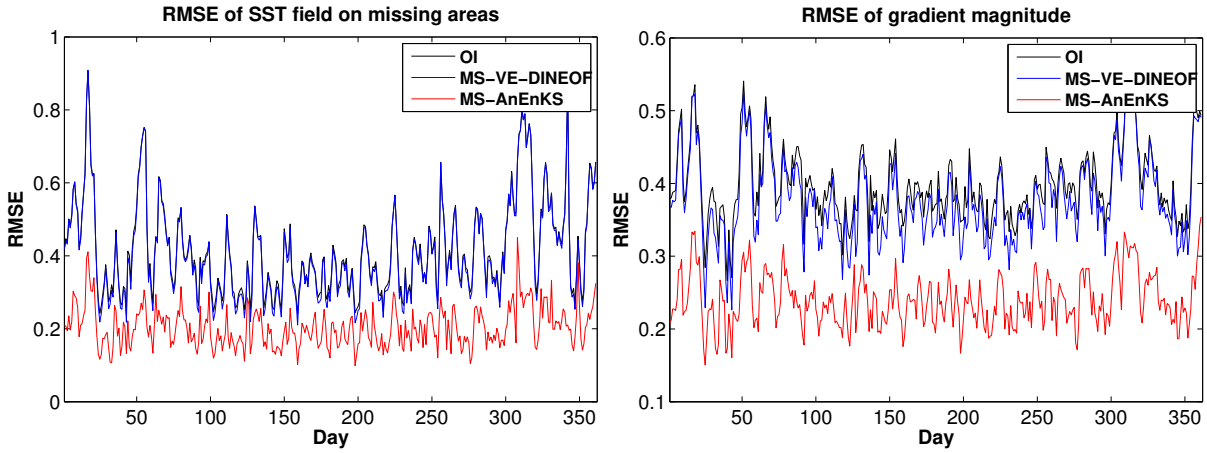


Figure 3.1 – Time series of the RMSE: OI (black,-), VE-DINEOF (blue,-) and AnEnKS (red,-) for the estimated SST fields (left) and gradient magnitude fields (right)

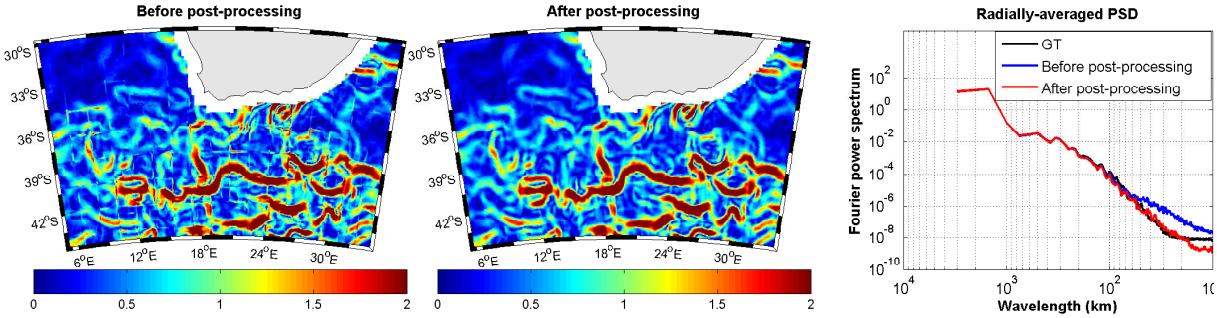


Figure 3.2 – Illustration of the postprocessing step for the removal of blocky artifacts: gradient magnitude field of the an interpolated SST field using MS-AnEnKS before (a) and after (b) the application of the considered PCA-based postprocessing step with 10×10 patches. We also report the radially-averaged power spectral density of the interpolated SST fields w.r.t. the true SST field (GT, black-).

that our multi-scale approach may also allow us to combine observation datasets from different sensors [48].

3.6 Conclusion

In this chapter, we reported the application of the analog data assimilation framework to high-dimensional satellite-derived geophysical fields. We demonstrated its relevance with respect to state-of-the-art techniques, namely optimal interpolation [41] and a PCA-based matrix completion scheme [13, 123]. Our model significantly outperforms these two techniques in terms of reconstruction error, especially for fine-scale structures in the range $[20km, 200km]$. The considered case-study involves real missing data patterns from the METOP-AVHRR sensor. It is

Table 3.1 – Comparison of global interpolation performance: RMSE of OI, G-VE-DINEOF, MS-VE-DINEOF, G-AnEnKS and MS-AnEnKS: we report RMSE statistics in terms of the SST fields, the gradient magnitude of the SST fields and of the detail coefficients for a four-level dyadic wavelet decomposition (noted *wav*). For MS-ANEnKS, we report both the interpolation performance at intermediate scale $i = 1$ (MS-ANEnKS$|dX_1$), *i.e.* with $dX_2 = 0$ in (3.1), and at scale $i = 2$ (MS-ANEnKS$|dX_2$). We let the reader refer the main text for details on the associated parameter setting of the different interpolation models.

Criterion		SST	$\|\nabla\|$	wav=1	wav=2	wav=4	wav=8
OI		0.4157	0.3986	0.0053	0.0212	0.0897	0.1897
G-VE-DINEOF		0.4064	0.3967	0.0124	0.0221	0.0873	0.1969
MS-VE-DINEOF		0.4052	0.3765	0.0052	0.0192	0.0803	0.1697
G-AnEnKS		0.3842	0.3922	0.0120	0.0219	0.0967	0.1902
MS-AnHMM	dX_2	0.3350	0.3529	0.0057	0.0208	0.0838	0.1711
MS-AnEnKS	dX_1	0.2536	0.3349	0.0057	0.0212	0.0848	0.1622
	dX_2	**0.2009**	**0.2357**	**0.0053**	**0.0173**	**0.0579**	**0.1067**

Table 3.2 – Influence of the number of analogs on MS-AnEnKS performance: RMSE of MS-AnEnKS interpolation w.r.t. the number of analogs for the three considered analog strategies.

Number of analogs (K)	10	20	30	40	50	60	70	80	90	100	110
Locally-constant	0.2746	0.2778	0.2822	0.2852	0.2884	0.2904	0.2926	0.2948			
Locally-Linear		0.2449	0.2369	0.2325	0.2301	0.2288	0.2280	0.2278	0.2271	0.2266	0.2266
Locally-incremental	0.2119	0.2113	0.2083	0.2051	0.2030	0.2028	0.2020	0.2012	**0.2009**	**0.2009**	0.2011

Table 3.3 – Influence of the kernel on MS-AnEnKS performance: RMSE of the interpolated SST fields using different kernel parameterizations using a Gaussian kernel and a cone kernel [163].

Gaussian	Cone ζ=0.995	Cone ζ=0.5	Cone ζ=0
0.2030	0.2028	0.2036	**0.2009**

Table 3.4 – MS-AnEnKS performance depending on the selected analog model: we let the reader refer to Tab.3.1 for the description of the considered evaluation criteria

Criterion	SST	$\|\nabla\|$	wav=1	wav=2	wav=3	wav=4
Locally-constant	0.2725	0.3214	0.0063	0.0208	0.0783	0.1529
Locally-Linear	0.2245	0.2730	0.0059	0.0186	0.0637	0.1265
Locally-Incremental	0.2009	0.2357	0.0053	0.0173	0.0579	0.1067

Table 3.5 – Influence of missing data in catalogs $\mathcal{C}_{1,2}$: we let the reader refer to Tab.3.1 for the description of the considered evaluation criteria.

Criterion	SST	$\|\nabla\|$	wav=1	wav=2	wav=3	wav=4
Catalogs $\mathcal{C}_{1,2}$ built from gap-free 2008-2014 data	0.2009	0.2357	0.0053	0.0173	0.0579	0.1067
Catalogs $\mathcal{C}_{1,2}$ built from 208-2015 dataset with missing data	0.2230	0.2643	0.0056	0.0194	0.0653	0.1212

Table 3.6 – Computational complexity of the interpolation models evaluated in Tab.3.1

Method	OI	MS-VE-DINEOF	MS-AnHMM	MS-AnEnKS
Exe. time	≈ 3.5h	≈ 0.5h	≈1.2h	≈ 3h

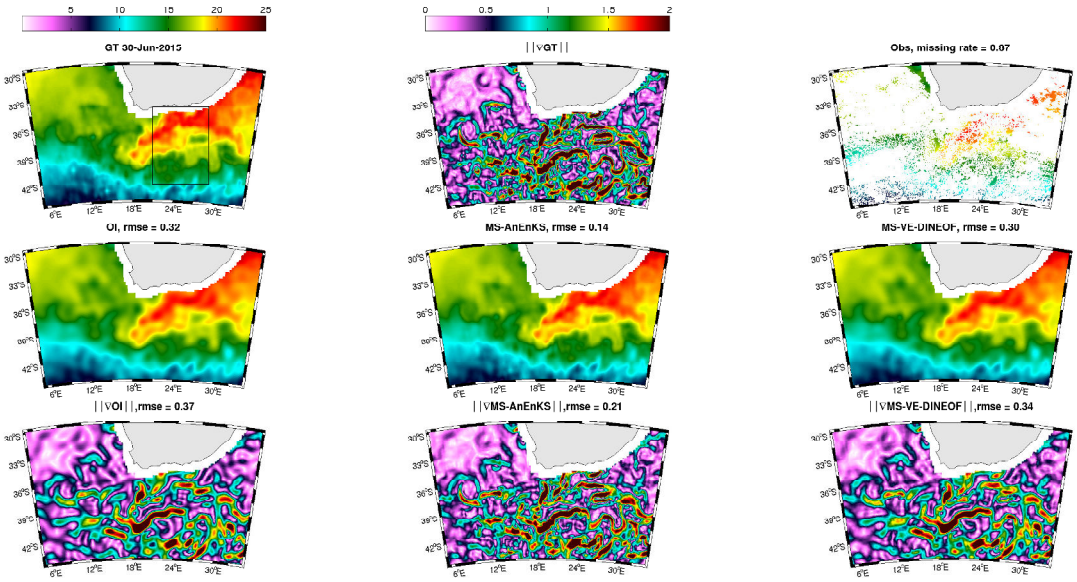


Figure 3.3 – Reconstruction of a SST field on June, 30, 2015 with a large missing data rate (87%): (a) first row, reference SST field (groundtruth (GT)), its associated gradient magnitude, observed field; second row, interpolated fields by OI, MS-AnEnKS, MS-VE-DINEOF; third row, gradient magnitude of the fields depicted in the second row.

.

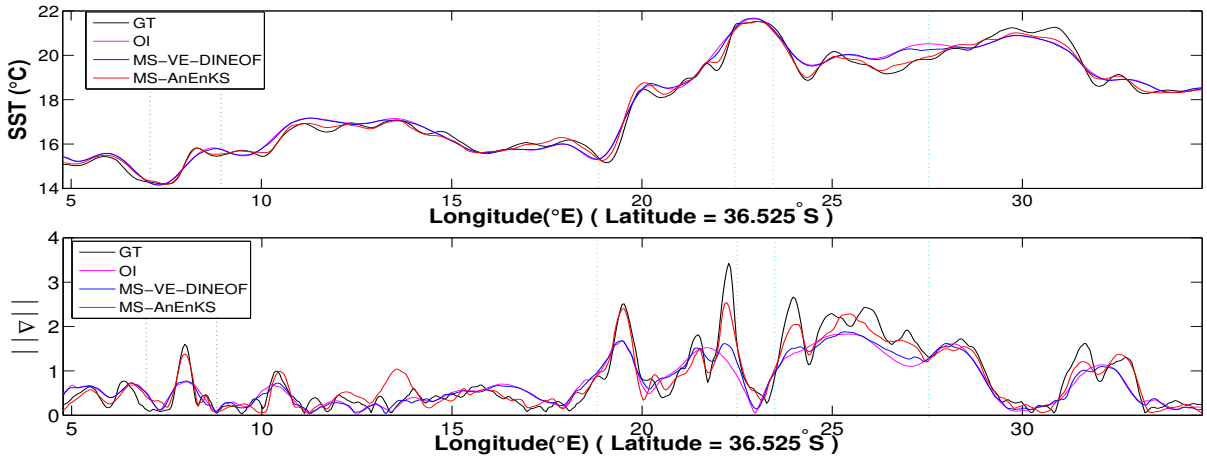


Figure 3.4 – Analysis of a SST transect at 36.525^o S for the interpolation results depicted in Fig. 3.3: we depict a one-dimensional profile at latitude 36.525^o S (c) for both the SST (bottom) and the SST gradient magnitude (top) for the reference SST field (black,-) as well as OI (magenta,-), MS-VE-DINEOF (blue,-) and MS-AnEnKS (red,-) interpolated SST fields.

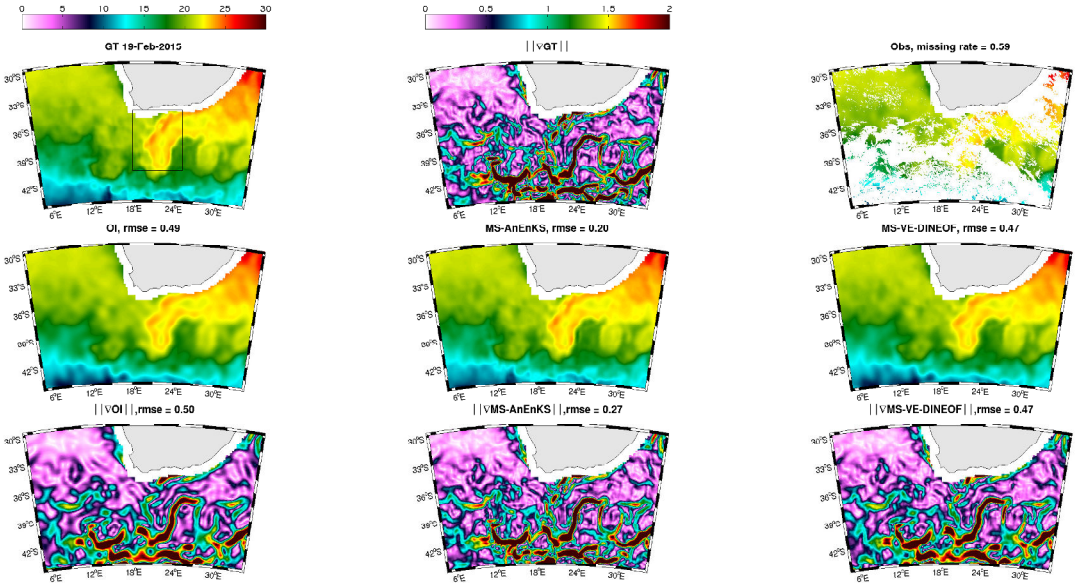


Figure 3.5 – Reconstruction of an SST field on February, 19, 2015 with a relatively low missing data rate (56%): see Fig.3.3 for details.

therefore representative of the irregular space-time sampling of the sea surface associated with infrared satellite sensors. The relative gain in the mean interpolation RMSE of about 50% stresses the potential of data-driven computational models in the exploitation of large-scale observation datasets to improve the reconstruction of geophysical fields from partial satellite-derived observations. We have made our case-study dataset available as a supplementary material to our paper with a view to favoring the benchmarking of interpolation methods for satellite-derived geophysical products[2].

As demonstrated by our experimental evaluation, the first key feature of the proposed model is the use of a multi-scale decomposition. Whereas a classic model-driven interpolation (OI) applies to the coarse-scale component, the reconstruction of the fine-scale components exploit the analog data assimilation [97]. A critical aspect of analog methods is the availability of a representative catalog of exemplars. In this respect, the considered multi-scale decomposition is regarded as a crucial means to stationarize the fine-scale spatial variabilities depicted by sea surface geophysical fields and make more relevant exemplar-based representations of these variabilities. Wavelet analysis is generally the classic scheme to derive a multi-scale decomposition [106]. Here, we exploited PCA-based representations for different patch sizes, so that we naturally combined a multi-scale decomposition to a low-dimensional representation of the spa-

[2]The Python code used for the creation of the considered SST data is available at: https://github.com/rfablet/SSTData_TCI_rfablet

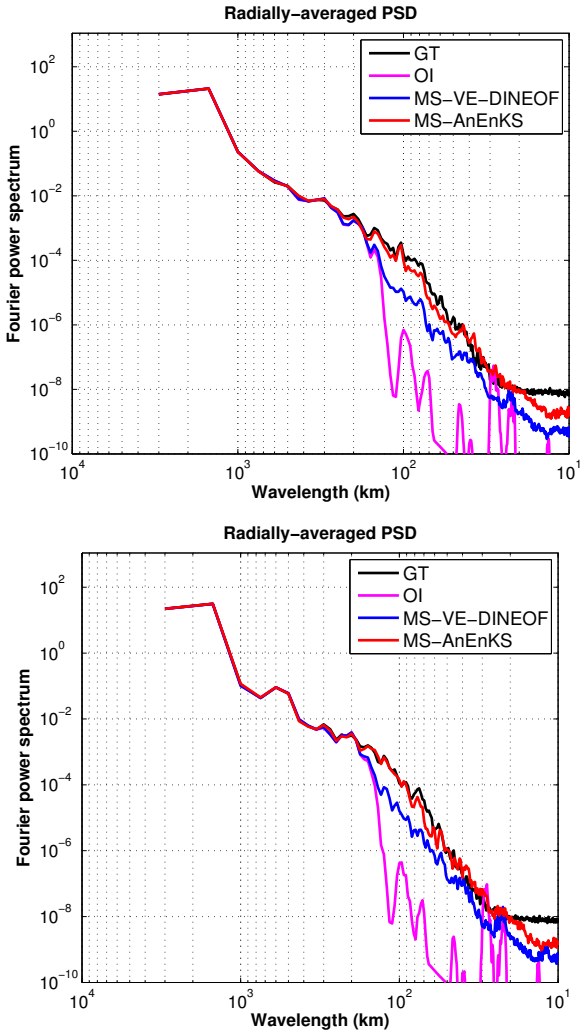


Figure 3.6 – Spectral analysis of interpolation results depicted in Fig.3.3 and 3.5: we report the radially-averaged power spectral densities of the reference SST field (black,-) as well as OI (magenta,-), MS-VE-DINEOF (blue,-) and MS-AnEnKS (red,-) interpolated SST fields for June, 30, 2015 (left) and February, 19, 2015 (right).

tial variabilities on each scale. Such PCA-based representations also efficiently deal with complex image geometries (e.g., the presence of land areas in the considered ocean case-study region). It may be noted that Model (3.1) could be straightforwardly extended to a greater number of scales. For the considered case-study however, numerical experiments did not lead to significant improvements with 3 or 4 detail scales.

We believe that this study opens new research avenues for the development of new data-driven models for the reconstruction of upper ocean dynamics from satellite-derived observations, in the same way that data-driven schemes have led to major advances in other imaging domains such as photography, microscopy, astronomy.... The exploitation of analogs for interpolation may be interpreted in a climatological sense, the key idea being that previously observed fine-scale geophysical variabilities will probably occur again, though not necessarily with the same seasonal timing. The application to other sea surface tracers, such as ocean color, is then natural [133]. The proposed multi-scale analog assimilation also seems particularly appealing for the downscaling of low-resolution satellite-derived products, such as sea surface salinity [150] and sea surface height [51]. From a methodological point of view, multimodal extensions would be of interest to account for multi-sensor observations as well synergies between different tracers [48, 59]. The next chapter of this thesis (Chapter 4) presents an application of the MS-AnDA method to Sea Surface Height.

Analog strategies are particularly appealing when large and representative observation datasets are available, as illustrated in the case-study considered here. By contrast, one may question their relevance in addressing scarce observation datasets as well as extreme events, which are by essence rare events. In this context, the creation of catalogs of analogs from realistic high-resolution numerical simulations [63, 131], which are becoming increasingly available, appears to be a relevant path to be further investigated.

Chapter 4

Analog Spatio-Temporal Interpolation of Sea Level Anomalies from Altimeter-derived Data

But more wonderful than the lore of old men and the lore of books is the secret lore of ocean.

Howard Phillips Lovecraft

Note: This chapter is submitted for publication as: R. Lguensat, P. Viet, M. Sun, G. Chen, T. Fenglin, B. Chapron, R. Fablet. "Data-driven Interpolation of Sea Level Anomalies using Analog Data Assimilation". It is presented as it is with small modifications, except for the motivation section.

4.1 Motivation

In this chapter, we build on the findings and the conclusions drawn from our previous work on SST. Here, we address a more challenging problem: The interpolation of Sea Level Anomaly (SLA) fields from along-track altimeter data. It is challenging because of the high rate of missing data, that are in this case not resulting from cloud coverage or weather conditions, but from the way the altimeter measures the height of the sea surface (SSH). Along-track data are data collected from altimeter passes on its orbit around the globe, moreover, two altimeters (or more) at the same time are needed to perform a relevant reconstruction using Optimal Interpolation. Meanwhile, high resolution SSH fields are available using numerical simulations, we therefore wanted to investigate the use of these numerical simulations as a catalog for the reconstruction of high resolution altimeter-derived fields.

This part of my thesis was done in collaboration with Dr. Miao Sun, Prof. Ge Chen and Dr. Tian Fenglin from the Marine Information Technology lab in Ocean University of China, where I spent one month as a visiting PhD student. Our first attempt in using AnDA for this problem is described in [95], where we used the global multiscale G-MS-AnDA and reached a slight improvement over Optimal Interpolation. In this chapter, we investigate the use of the patch-based version of the multiscale AnDA. Contrarily to the application on SST fields depicted in Chapter 3, we use a two-scale model ($\bar{X}$ and dX_1) since dX_2 did not bring significant improvement and makes time execution and the calculations heavier. We dropped the cone kernel given it's weak influence on the result. While in the previous chapter, we assumed an independence between the scales ($\bar{X}$ and dX_1), this chapter investigates the use of inter-scale dependencies and also the use of additional variables as predictors, here, the SST-SSH relationship.

4.2 Introduction

The past twenty years have witnessed a deluge of ocean satellite data, such as sea surface height, sea surface temperature, ocean color, ocean current, sea ice, etc. This has helped building big databases of valuable information and represents a major opportunity for the interplay of ideas between ocean remote sensing community and the data science community. Exploring machine learning methods in general and non-parametric methods in particular is now feasible and is increasingly drawing the attention of many researchers [25, 62, 80, 88, 162].

More specifically, analog forecasting [102] which is among the earliest statistical methods explored in geoscience benefits from recent advances in data science. In short, analog forecasting is based on the assumption that the future state of a system can be predicted throughout the successors of past (or simulated) similar situations (called analogs). The amount of currently available remote sensing and simulation data offers analog methods a great opportunity to catch up their early promises. Several recent works involving applications of analog forecasting methods in geoscience fields contribute in the revival of these methods, recent applications comprise the prediction of soil moisture anomalies [111], the prediction of sea-ice anomalies [34], rainfall nowcasting [8], stochastic weather generators [160], etc. One may also cite methodological developments such as dynamically-adapted kernels [163] and novel parameter estimation schemes [72]. Importantly, analog strategies have recently been extended to address data assimilation issues within the so-called *analog data assimilation* (AnDA) [97], where the dynamical model is stated as an analog forecasting model and combined to state-of-the-art stochastic assimilation procedures such as Ensemble Kalman filters. The application to high-dimensional fields in Chapter 3 provides the methodological background for this study.

Producing time-continuous and gridded maps of Sea Surface Height (SSH) is a major challenge in ocean remote sensing with important consequences on several scientific fields from weather and climate forecasting to operational needs for fisheries management and marine operations (*e.g.* [67]). The reference gridded SSH product commonly used in the literature is distributed by the Copernicus Marine and Environment Monitoring Service (CMEMS) (formerly distributed by AVISO). This product relies on the interpolation of irregularly-spaced along-track data using an Optimal Interpolation (OI) method [17, 92]. While OI is relevant for the retrieval of horizontal scales of SSH fields greater than $\approx 100km$, its Gaussian assumptions cause the small scales of the SSH fields to be smoothed. This limitation makes it impossible to resolve

finer-scale processes (typically from a few tens of kilometers to $\approx 100km$) which may be revealed by along-track altimetric data. This has led to a variety of research studies to improve the reconstruction of the altimetric fields. One may cite both methodological alternatives to OI, for instance locally-adapted convolutional models [51] and variational assimilation schemes using model-driven dynamical priors [149], as well as studies exploring the synergy between different sea surface tracers, especially the synergy between SSH and SST (Sea Surface Temperature) fields and Surface Quasi-Geostrophic dynamics [51, 77, 78, 85, 146, 148].

In this work, we build upon our recent advances in analog data assimilation and its application to high-dimensional fields [50, 97]. We develop an analog data assimilation model for the reconstruction of SLA fields from along-track altimeter data. It relies on a patch-based and EOF-constrained representation of the SLA fields. Using OFES numerical simulations [110, 132], we design an Observation System Simulation Experiment (OSSE) for a case-study in the South China sea using real along-track sampling patterns of spaceborne altimeters. Using the resulting groundtruthed dataset, we perform a qualitative and quantitative evaluation of the proposed scheme, including comparisons to state-of-the-art schemes.

4.3 Data: OFES (OGCM for the Earth Simulator)

An Observation System Simulation Experiment (OSSE) based on numerical simulations is considered to assess the relevance of the proposed analog assimilation framework. Our OSSE uses these numerical simulations as a groundtruthed dataset from which simulated along-track data are produced. We describe further the data preparation setup in the following sections.

4.3.1 Model simulation data

The Ocean General Circulation Model (OGCM) for the Earth Simulator (OFES) is considered in this study as the true state of the ocean. The simulation data is described in [110, 132]. The coverage of the model is 75°S-75°N with a horizontal resolution of 1/10°. 34 years (1979-2012) of 3-daily simulation of SSH maps are considered, we proceed to a subtraction of a temporal mean to obtain SLA fields. In this study, our region of interest is located in the South China Sea (105°E to 117°E, 5°N to 25°N). This dataset is split into a training dataset corresponding to the first 33 years (4017 SLA maps) and a test dataset corresponding to the last year of the time series (122 SLA maps).

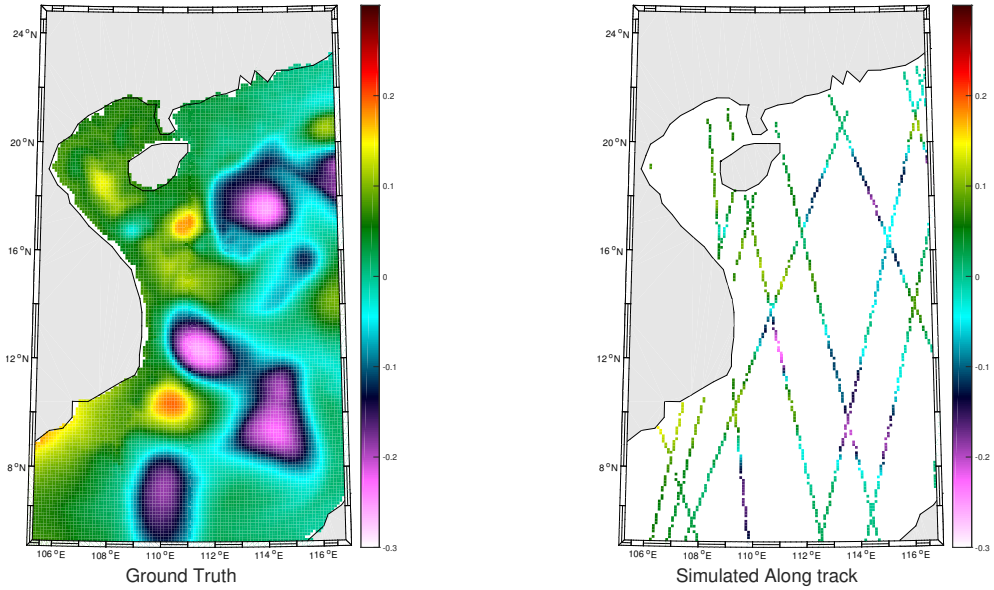


Figure 4.1 – An example of a ground-truth SLA field in the considered region and its associated simulated pseudo-along track.

4.3.2 Along track data

We consider a realistic situation with a high rate of along track data. More precisely we use along-track data positions registered in 2014 where 4 satellites (Jason2, Cryosat2, Saral/AltiKa, HY-2A) were operating. Data is distributed by Copernicus Marine and Environment Monitoring Service (CMEMS).

From the reference 3-daily SLA dataset and real along-track data positions, we generate simulated along-track data from the sampling of a reference SLA field: more precisely, for a given along-track point, we sample the closest position of the 1/10° regular model grid at the closest time step of the 3-daily model time series. As we consider a 3-daily assimilation time step (see Section 4.3.1 for details), we create a 3-daily pseudo-observation field, to be fed directly to the assimilation model. As sketched in Figure 4.2, for a given time t, we combine all along-track positions for times $t-1$,t and $t+1$ to create an along-track pseudo-observation field at time t. We denote by $s3dAT$ the simulated 3-daily time series of along-track pseudo-observation fields. An example of these fields is given in Figure 4.1.

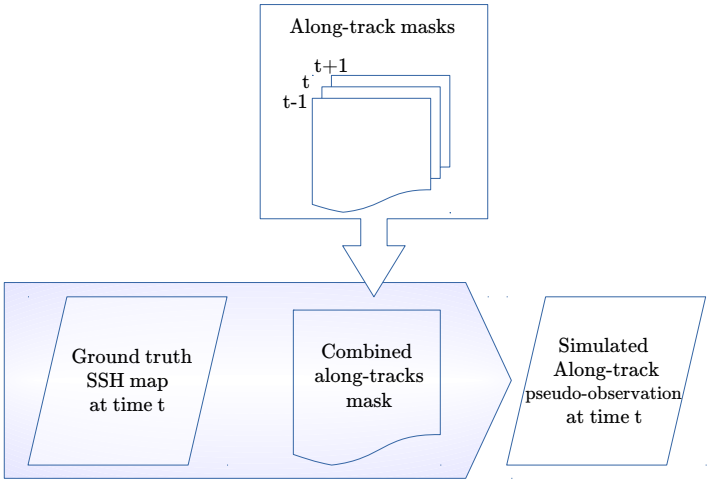


Figure 4.2 – Sketch of the creation of simulated along-track data at a given time t

4.4 Analog reconstruction for altimeter-derived SLA

4.4.1 Patch-based state-space formulation

As stated in the introduction of this chapter, OI may be considered as an efficient model-based method to recover large-scale structures of SLA fields. Following the findings in Chapter 3, this suggests to consider the following two-scale additive decomposition:

$$X = \bar{X} + dX + \xi \tag{4.1}$$

where $\bar{X}$ is the large-scale component of the SLA field, typically issued from an optimal interpolation, dX the fine-scale component of the SLA field we aim to reconstruct and ξ remaining unresolved scales.

The reconstruction of field dX involves a patch-based and EOF-based representation. It consists in regarding field dX as a set of $P \times P$ overlapping patches (*e.g.* $2° \times 2°$). This set of patches is referred to as $\mathcal{P}$, and we denote by $\mathcal{P}_s$ the patch centered at position s. After building a catalog $\mathcal{C}_\mathcal{P}$ of patches from the available dataset of residual fields $X - \bar{X}$, we proceed to an EOF decomposition of each patch in the catalog. The reconstruction of field $dX(\mathcal{P}_s, t)$ at time t is then stated as the AnDA of the coefficients of the EOF decomposition in the EOF space given an observation series in the patch space. Formally, $dX(\mathcal{P}_s, t)$ decomposes as a linear combination

of a number N_E of EOF basis functions with the largest variances:

$$dX(\mathcal{P}_s, t) = \sum_{k=1}^{N_E} \alpha_k(s,t) EOF_k \tag{4.2}$$

with EOF_k referring to the k^{th} EOF basis and $\alpha_k(s,t)$ to the corresponding coefficient for patch $\mathcal{P}_s$ at time t. Let us denote by $\varphi(\mathcal{P}_s, t)$ the vector of the N_E coefficients $\alpha_k(s,t)$, which represents the projection of $dX(\mathcal{P}_s, t)$ in the lower-dimensional EOF space.

4.4.2 Patch-based analog dynamical models

We detail in this section the application of the AnDA framework as presented in Chapter 3 for the sequential reconstruction of fine-scale dX. The proposed patch-based analog assimilation scheme involves a dynamical model stated in the EOF space. As in the previous Chapter we consider the following Gaussian conditional distribution

$$\varphi(\mathcal{P}_s, t) | \varphi(\mathcal{P}_s, t-1) \sim \mathcal{G}(\mu(s,t), \Sigma(s,t)) \tag{4.3}$$

We consider the three analog forecasting operators presented in Chapter 2, namely, the locally-constant, the locally incremental and the locally-linear. The calculation of the weights associated to each analog-successor pair relies on a Gaussian kernel $\mathcal{K}_G$ (Equation 3.5). The search for analogs in the N_E-dimensional patch space (in practice, N_E ranges from 5 to 20) ensures a better accuracy in the retrieval of relevant analogs compared to a direct search in the high-dimensional space of state dX. It also reduces the computational complexity of the proposed scheme.

Another important extension of the current study is the possibility of exploiting auxiliary variables with the state vector Φ in the analog forecasting models. Such variables may be considered in the search for analogs as well as regression variables in locally-linear analog setting. Regarding the targeted application to the reconstruction of SSH fields and the proposed two-scale decomposition (Equation 4.1), two types of auxiliary variables seem to be of interest: the low-resolution component $\bar{X}$ to take into account inter-scale relationship [51], and Sea Surface Temperature (SST) with respect to the widely acknowledged SST-SSH synergies [51,78,85,146]. We also apply patch-level EOF-based decompositions to include both types of variables in the considered analog forecasting models (Equation 4.3).

4.4.3 Numerical resolution

Given the proposed analog assimilation model, the proposed scheme first relies on the creation of patch-level catalogs from the training dataset. This step requires the computation of a training dataset of fine scale data $dX_{training}$, this is done by subtracting a large-scale component $\bar{X}_{training}$ from the original training dataset. Here, we consider the large-scale component of training data to be the result of a global[1] EOF-based reconstruction using a number of EOF components that retains 95% of the dataset variance, which accounts for horizontal scales up to $\sim$ 100km. This global EOF-based decomposition provides a computationally-efficient means for defining large-scale component $\bar{X}$. This EOF-based decomposition step is followed by the extraction of overlapping patches for all variables of interest, namely $\bar{X}_{training}$, $dX_{training}$ and potential auxiliary variables, and the identification of the EOF basis functions from the resulting raw patch datasets. This leads to the creation of a patch-level catalog $\mathcal{C}_{\mathcal{P}}$ from the EOF-based representations of each patch.

Given the patch-level catalog, the algorithm applied for the mapping SLA fields from along-track data, referred to as MS-AnDA, involves the following steps:

- the computation of the large-scale component $\bar{X}$, here, we consider the result of optimal interpolation (OI) projected onto the global EOF basis functions.

- the decomposition of the case study region into overlapping $P \times P$ patches, here, 20×20 patches

- For each patch position s, the application of an analog data assimilation scheme, namely the Analog Ensemble Kalman Smoother (AnEnKS) [97], for patch $\mathcal{P}_s$ of field dX. As stated in (4.3), the assimilation is performed in the EOF space, *i.e.* for EOF decomposition $\Phi(\mathcal{P}_s, t)$, using the operator derived from EOF-based reconstruction (4.2) and decomposition (4.1) as observation model $\mathcal{H}$ and the patch-level training catalog described in the previous section. In the analog forecasting setting, The search for analogs is restricted to patch exemplars in the catalog within a local spatial neighborhood (typically a patch-level 8-neighborhood), except for patches along the seashore for which the search for analogs is restricted to patch exemplars at the same location.

[1] By global, we mean here an EOF decomposition over the entire case study region, by contrast to the patch-level decomposition considered in the analog assimilation setting.

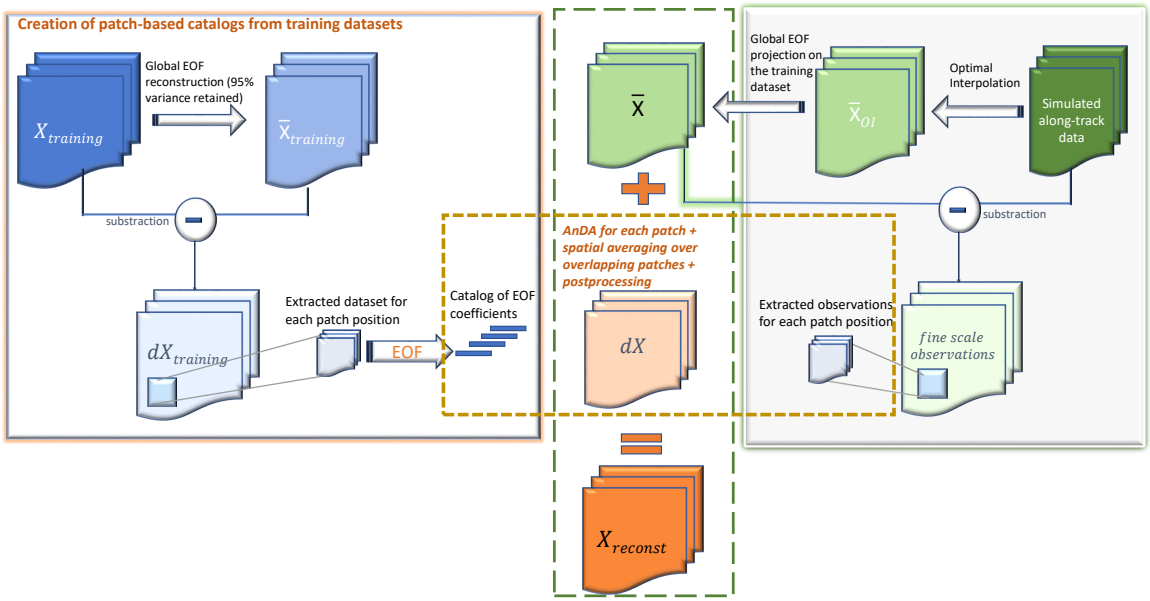


Figure 4.3 – Sketch of the proposed patch-based Multiscale Analog Data Assimilation (MS-AnDA). The left block details the construction of the patch-based catalogs from the training dataset. The right block illustrates the process of obtaining the large-scale component of the SLA reconstructed field. The orange dashed rectangle represents the application of the AnDA using the catalog and the fine-scale observations. Finally, the green dashed rectangle shows the final addition operation that yields the reconstructed SLA field.

- the reconstruction of fields dX from the set of assimilated patches $\{dX(\mathcal{P}_s, \cdot)\}_s$. This relies on a spatial averaging over overlapping patches (here, a 5-pixel overlapping in both directions). In practice, we do not apply the patch-level assimilation to all grid positions. Consequently, the spatial averaging may result in blocky artifacts. We then apply a patch-wise EOF-based decomposition-reconstruction with a smaller patch-size (here, 17×17 patches) to remove these blocky artifacts.

- the reconstruction of fields X as $\bar{X} + dX$.

4.5 Results

We evaluate the proposed MS-AnDA approach using the OSSE presented in Section 4.3. We perform a qualitative and quantitative comparison to state-of-the-art approaches. We first describe the parameter setting used for the MS-AnDA as well as benchmarked models, namely OI, an EOF-based approach [123] and a direct application of AnDA at the region level. We then

report numerical experiments for noise-free and noisy observation data as well the relevance of auxiliary variables in the proposed MS-AnDA scheme.

4.5.1 Experimental setting

We detail below the parameter setting of the models evaluated in the reported experiments, including the proposed MS-AnDA scheme:

- *MS-AnDA*: We consider 20×20 patches with 15-dimensional EOF decompositions ($N_E = 15$), which typically accounts for 99% of the data variance for the considered dataset. The postprocessing step exploits 17×17 patches and a 15-dimensional EOF decomposition. Regarding the parametrization of the AnEnKS procedure, we experimentally cross-validated the number of nearest neighbors K to 50, the number of ensemble members $n_{ensemble}$ to 100 and the observation covariance error (in meters, hereinafter) to $\mathbf{R} = 0.001$.

- *Optimal Interpolation*: We apply an Optimal Interpolation to the processed along-track data. It provides the low-resolution component for the proposed MS-AnDA model and a model-driven reference for evaluation purposes. The background field is a null field. We use a Gaussian covariance model with a spatial correlation length of 100km and a temporal correlation length of 15 days ($\pm$ 5 timesteps since our data is 3-daily). These choices result from a cross-validation experiment.

- *VE-DINEOF*: We apply a second state-of-the-art interpolation scheme using a data-driven strategy solely based on EOF decompositions, namely VE-DINEOF [123]. We implement a patch-based version of VE-DINEOF to make it comparable to the proposed MS-AnDA setting. Given the same EOF decomposition as in MS-AnDA, the patch-level VE-DINEOF iterates patchwise EOF projection-reconstruction of the detail field dX. This scheme is initialized from the along-track pseudo-observation field for along-track data positions and $\bar{X}$ for missing data positions. After each projection-reconstruction, we only update missing data areas. We run this iterative process until convergence.

- *G-AnDA*: With a view to evaluating the relevance of the patch-based decomposition, we also apply AnDA at the region scale, referred to as G-AnDA. It relies on an EOF-based decomposition of the detail field dX. We use 150 EOF components, which accounts for more than 99% of the total variance of the SSH dataset. From cross-validation experiments, the associated AnEnKS procedure relies on a locally-linear analog forecasting model with

$K = 500$ analogs, $n_{ensemble} = 100$ ensemble members and an observation covariance error set to $\mathbf{R} = 0.001$

The patch-based experiments were run on Teralab infrastructure using a multi-core virtual machine (30 CPUs, 64G of RAM). We used the Python toolbox for patch-based analog data assimilation [50] (available at github.com/rfablet/PB_ANDA). Optimal Interpolation was implemented on Matlab using [45]. Throughout the experiments, two metrics are used to assess the performance of the considered interpolation methods: i) daily and mean Root Mean Square Error (RMSE) series between the reconstructed SLA fields X and the groundtruthed ones, ii) daily and mean correlation coefficient between the fine-scale component dX of the reconstructed SLA fields and of the groundtruthed ones.

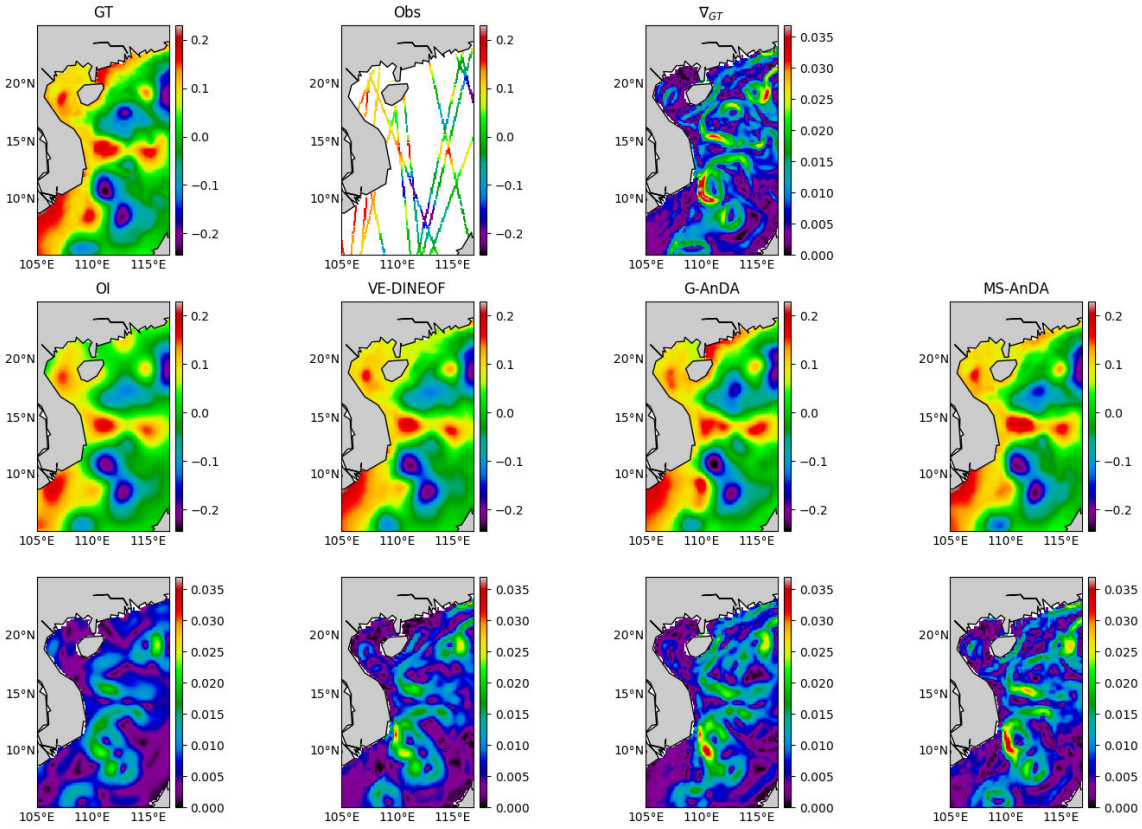
4.5.2 SLA reconstruction from noise-free along-track data

Table 4.1 – SLA Interpolation performance for a noise-free experiment: Root Mean Square Error (RMSE) and correlation statistics for OI, VE-DINEOF, G-AnDA and MS-AnDA w.r.t. the groundtruthed SLA fields. See Section 4.5.1 for the corresponding parameter settings.

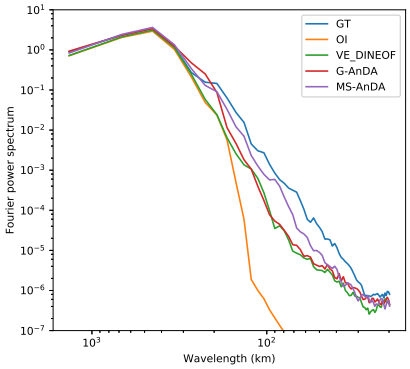
Criterion		RMSE	Correlation
OI		0.026 ± 0.007	0.81 ± 0.08
VE-DINEOF		0.023 ± 0.007	0.85 ± 0.07
G-AnDA		0.020 ± 0.006	0.89 ± 0.04
MS-AnDA	Locally-constant	0.014 ± 0.005	0.95 ± 0.03
	Locally-Increment	0.014 ± 0.005	0.95 ± 0.03
	Locally-Linear	**0.013 ± 0.005**	**0.96 ± 0.02**

We first perform an idealized noise-free experiment, where the along-track observations are noise-free. The observation covariance error takes the value $\mathbf{R} = 0.001$. The interpolation performances for this experiment are illustrated in Table 4.1. Our MS-AnDA algorithm significantly outperforms OI. More specifically, the locally-linear MS-AnDA results in the best reconstruction among the competing methods. We suggest that this improvement comes from the reconstruction of fine-scale features learned from the archived model simulation data. Figure 4.4a reports interpolated SSH fields and their gradient fields which further confirm our intuition. MS-AnDA

interpolation shows an enhancement of the gradients and comes out with some fine-scale eddies that were smoothed out in OI and VE-DINEOF. This is also confirmed by the Fourier power spectrum of the interpolated SLA fields in Figure 4.4b.

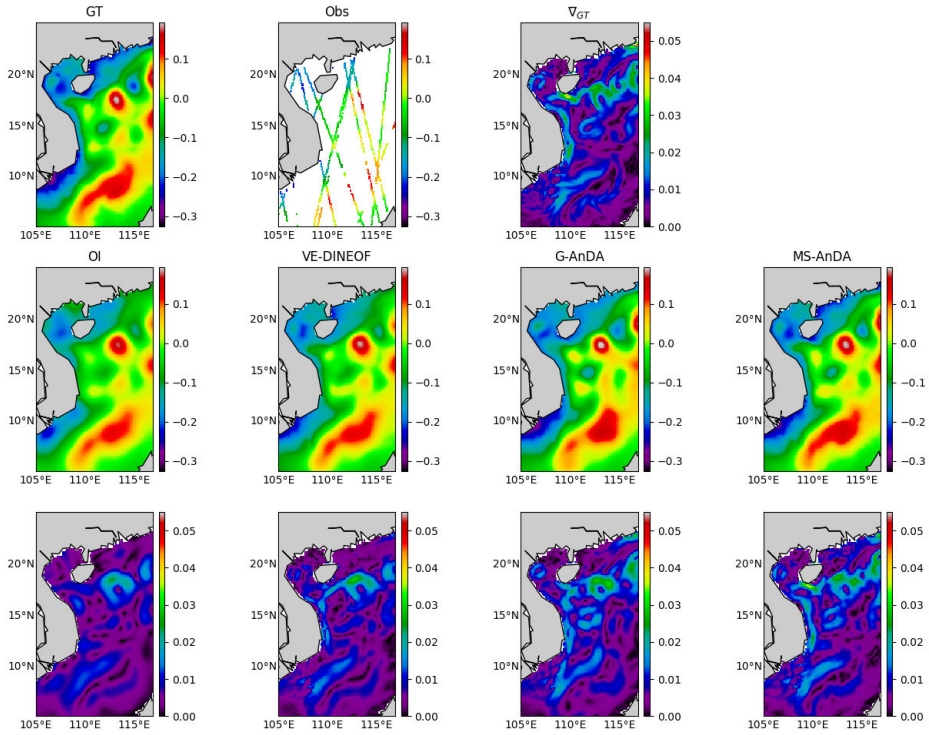


(a)

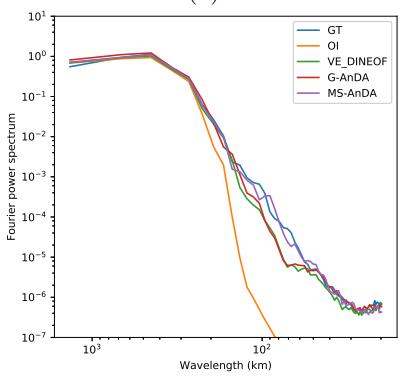


(b)

Figure 4.4 – Reconstructed SLA fields using noise-free along-track observation using OI, DINEOF, G-AnDA, MS-AnDA on February 24^{th} 2012: from left to right, the first row shows the ground truth field, the simulated available along-tracks for that day, the ground thruth gradient field. The second and third rows show each of the reconstruction and their corresponding gradient filed, from left to right, OI, VE-DINEOF, G-ANDA and MS-AnDA. The Fourier power spectrum of the competing methods is also included

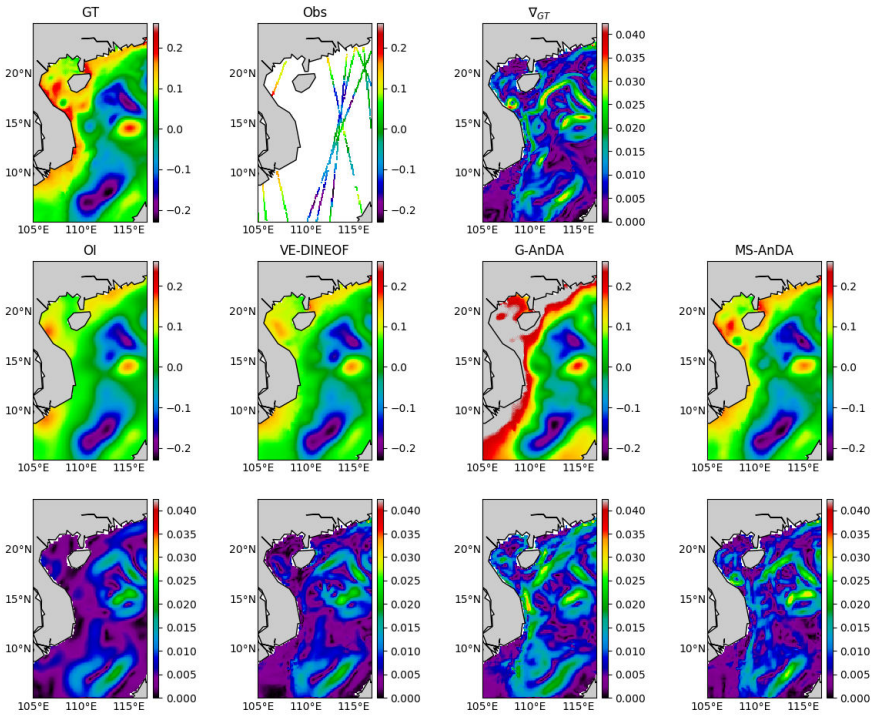


(a)

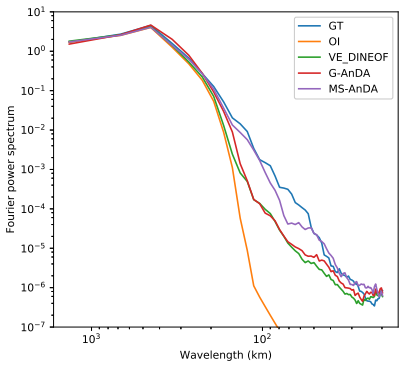


(b)

Figure 4.5 – Reconstructed SLA fields using noise-free along-track observation using OI, DINEOF, G-AnDA, MS-AnDA on August 22^{nd} 2012: from left to right, the first row shows the ground truth field, the simulated available along-tracks for that day, the ground truth gradient field. The second and third rows show each of the reconstruction and their corresponding gradient filed, from left to right, OI, VE-DINEOF, G-ANDA and MS-AnDA. The Fourier power spectrum of the competing methods is also included



(a)



(b)

Figure 4.6 – Reconstructed SLA fields using noise-free along-track observation using OI, DINEOF, G-AnDA, MS-AnDA on December 17^{th} 2012: from left to right, the first row shows the ground truth field, the simulated available along-tracks for that day, the ground truth gradient field. The second and third rows show each of the reconstruction and their corresponding gradient fields, from left to right, OI, VE-DINEOF, G-ANDA and MS-AnDA. The Fourier power spectrum of the competing methods is also included

4.5.3 SLA reconstruction from noisy along-track data

We also evaluated the proposed approach for noisy along-track data. Here, we run two experiments with an additive zero-mean Gaussian noise applied to the simulated along-track data. We consider a noise covariance of $\mathbf{R} = 0.01$ (Experiment A) and of $\mathbf{R} = 0.03$ (Experiment B) which is more close to the instrumental error of conventional altimeters. Given the resulting noisy along-track dataset, we apply the same methods as for the noise-free case study.

We run MS-AnDA using different values for $\mathbf{R}$. For Experiment A, Table 4.2 shows that the minimum is reached using the true value of the error $\mathbf{R} = 0.01$. While for Experiment B, Table 4.3 shows that the minimum is counter-intuitively reached again using value of the error $\mathbf{R} = 0.01$.

Our algorithm is then compared with the results of the application of the competing algorithms considered in this work. Results are shown in Table 4.4. MS-AnDA still outperforms OI in terms of RMSE and correlation statistics in both experiments. The locally-linear version of MS-AnDA depicts the best reconstruction performance. We report an example of the reconstruction in Figure 4.7. Similarly to the noise-free case study, MS-AnDA better recovers finer-scale structures in Fig.4.7.a compared with OI, VE-DINEOF and G-AnDA. In Fig.4.7.b, MS-AnDA also better reconstructs a larger-scale North-East structure, poorly sampled by along-track data and hence poorly interpolated by OI.

Table 4.2 – Impact of variance of observation error R in AnDA interpolation performance using noisy along-track data (R=0.01): RMSE of AnDA interpolation for different values of parameter R. For the same dataset, OI RMSE is **0.039**.

R	0.1	0.05	0.03	**0.01**	0.005	0.001	0.0001
$rmse_{MS-AnDA}$	0.035	0.030	0.028	**0.025**	0.025	0.029	0.044

Table 4.3 – Impact of variance of observation error R in AnDA interpolation performance using noisy along-track data (R=0.03): RMSE of AnDA interpolation for different values of parameter R. For the same dataset, OI RMSE is **0.066**.

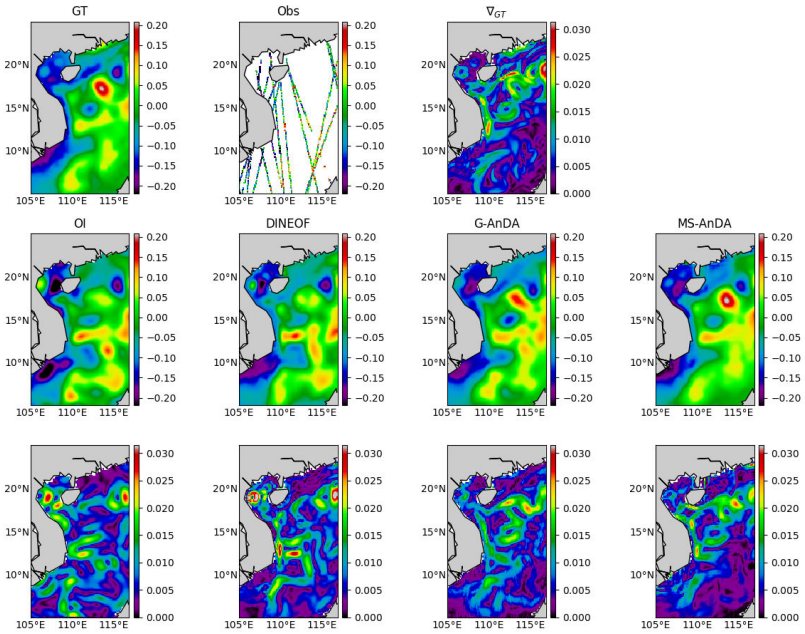
R	0.1	0.05	0.03	**0.01**	0.005	0.001	0.0001
$rmse_{MS-AnDA}$	0.038	0.036	0.035	**0.0349**	0.037	0.046	0.076

Table 4.4 – SLA Interpolation performance for noisy along-track data: Root Mean Square Error (RMSE) and correlation statistics for OI, VE-DINEOF, G-AnDA and MS-AnDA w.r.t. the groundtruthed SLA fields. See Section 4.5.1 for the corresponding parameter settings.

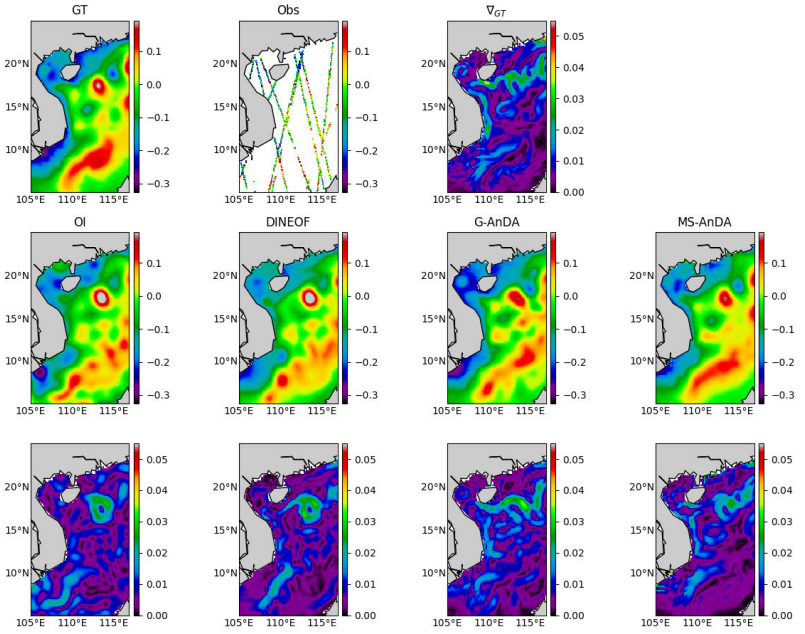
	Criterion		RMSE	Correlation
R=0.01	OI		0.039 ± 0.005	0.64 ± 0.09
	VE-DINEOF		0.035 ± 0.005	0.68 ± 0.09
	G-AnDA		0.030 ± 0.005	0.78 ± 0.06
	MS-AnDA	Locally constant	0.026 ± 0.005	0.82 ± 0.05
		Increment	0.028 ± 0.005	0.81 ± 0.05
		Local Linear	**0.0245 ± 0.005**	**0.83 ± 0.05**
R=0.03	OI		0.066 ± 0.006	0.41 ± 0.09
	VE-DINEOF		0.060 ± 0.006	0.45 ± 0.09
	G-AnDA		0.039 ± 0.006	0.67 ± 0.09
	MS-AnDA	Locally constant	0.035 ± 0.006	0.688 ± 0.064
		Increment	0.036 ± 0.006	0.656 ± 0.07
		Local Linear	**0.032 ± 0.006**	**0.708 ± 0.063**

4.5.4 Conditioning by auxiliary variables

We further explore the flexibility of the analog setting to the use of additional geophysical variable information as explained in Section 4.4.2. Intuitively, we expect SLA fields to involve inter-scale dependencies as well as synergies with other tracers. The use of auxiliary variables provide the means for evaluating such dependencies and their potential impact on reconstruction performance. We consider two auxiliary variables that are used in the locally-linear analog forecasting model: i) to account for the relationship between the large-scale and fine-scale component, we may consider variable $\bar{X}$, ii) considering potential SST-SSH synergies, we consider SST fields. Overall, we consider four parameterization of the regression variables used in MS-AnDA: the sole use of dX (MS-AnDA-dX); the joint use of dX and SST fields (MS-AnDA-dX+SST); the joint use of dX and $\bar{X}$ (MS-AnDA-dX+$\bar{X}$), the joint use of dX and the groudntruthed version of $\bar{X}$ denoted by $\bar{X}^{GT}$, (MS-AnDA-dX+$\bar{X}^{GT}$). The later provides a lower-bound for the reconstruction performance, assuming the low-resolution component is perfectly estimated.



(a)



(b)

Figure 4.7 – (Noisy observation) Reconstruction of SLA fields using OI, DINEOF, G-AnDA & MS-AnDA on day 225^{th} (a) & 228^{th} (b)

We report mean RMSE and correlation statistics for these four MS-AnDA parameterizations in Table 4.5 for the noisy case-study. Considering MS-AnDA-dX as reference, these results show a very slight improvement when complementing dX with SST information. Though limited, we report a greater improvement when adding the low-resolution component $\bar{X}$. Interestingly, a significantly greater improvement is obtained when adding the true low-resolution information. The mean results are in accordance with [51], which reported that large-scale SLA information was more informative than SST to improve the reconstruction of the SLA at finer scales. Though mean statistics over one year leads to rather limited improvement, daily RMSE time series (Figure 4.8) reveal that for some periods, for instance between day 130 and 150, relative improvements in terms of RMSE may reach 10% with the additional information brought by the large-scale component. In this respect, it may noted that MS-AnDA-dX+$\bar{X}$ always perform better than MS-AnDA-dX.

Table 4.5 – MS-AnDA reconstruction performance using noisy along-track data for different choices of the regression variables in the locally-linear analog forecasting model: MS-AnDA-dX using solely dX, MS-AnDA-dX+SST using both dX and SST, MS-AnDA-$dX + \bar{X}$ using both dX and $\bar{X}$, and MS-AnDA-$dX + \bar{X}^{GT}$ using dX and the true large-scale component $\bar{X}^{GT}$.

	MS-AnDA model	RMSE	Correlation
R=0.01	MS-AnDA-dX	0.025 ± 0.005	0.83 ± 0.05
	MS-AnDA-dX+SST	0.024 ± 0.005	0.83 ± 0.05
	MS-AnDA-$dX + \bar{X}$	0.023 ± 0.005	0.84 ± 0.05
	MS-AnDA-$dX + \bar{X}^{GT}$	0.021 ± 0.004	0.87 ± 0.04
R=0.03	MS-AnDA-dX	0.032 ± 0.006	0.708 ± 0.06
	MS-AnDA-dX+SST	0.031 ± 0.006	0.710 ± 0.06
	MS-AnDA-$dX + \bar{X}$	0.029 ± 0.006	0.717 ± 0.06
	MS-AnDA-$dX + \bar{X}^{GT}$	0.026 ± 0.005	0.730 ± 0.05

4.6 Discussion and conclusion

This work sheds light on the opportunities that data science methods are offering to improve altimetry in the era of "Big Data". Assuming the availability of high-resolution numerical simula-

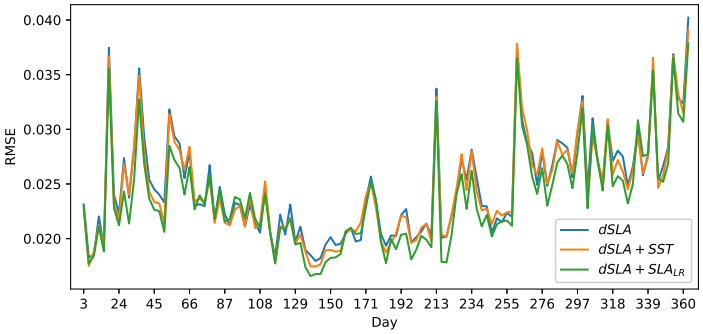
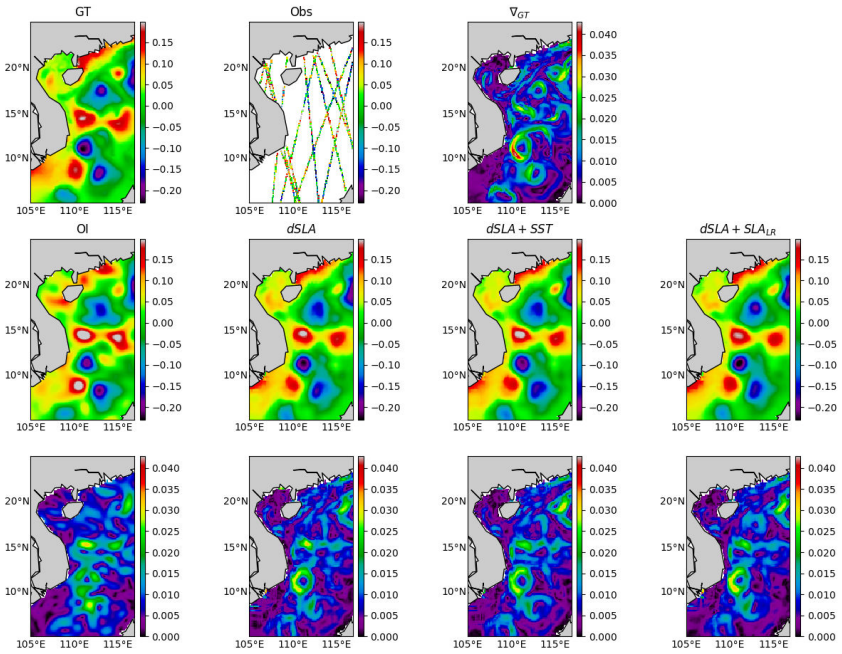


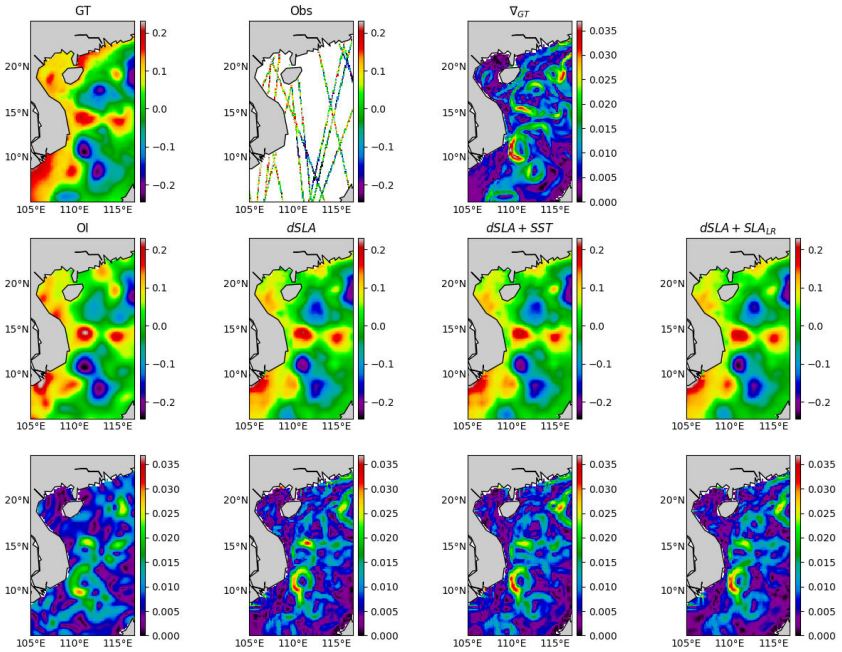
Figure 4.8 – (Noisy observation $\mathbf{R} = 0.01$) Daily RMSE time series of PB-AnDA SLA reconstructions using noisy along-track data for different choices of the regression variables in the locally-linear analog forecasting model: MS-AnDA-dX (light blue), MS-AnDA-dX+SST (orange) and MS-AnDA-$dX + \bar{X}$ (green)

tions, we show that Analog Data Assimilation (AnDA) can outperform the Optimal Interpolation method and retrieve smoothed out structures resulting from the sole use of OI both with idealized noise-free and more realistic noisy observations for the considered case study. Importantly, the reported experiments point out the relevance for combining OI for larger scales (above 100km) whereas the proposed patch-based analog setting successfully applies to the finer-scale range below 100km. This is in agreement with the recent application of the analog data assimilation to the reconstruction of cloud-free SST fields (Chapter 3). We also demonstrate that AnDA can embed complementary variables in a simple manner through the regression variables used in the locally-linear analog forecasting operator. In agreement with our recent analysis [51], we demonstrate that the additional use of local SST and large-scale SLA information may further improve the reconstruction performance for fine-scale structures.

Analog data assimilation can be regarded as a means to fuse ocean models and satellite-derived data. We regard this study as a proof-of-concept, which opens research avenues as well as new directions for operational oceanography. Our results advocate for complementary experiments at the global scale or in different ocean regions for a variety of dynamical situations with a view to further evaluating the relevance of the proposed analog assimilation framework. Such experiments should evaluate the sensitivity of the assimilation with respect to the size of the catalog. The scaling up to the global ocean also suggests investigating computationally-efficient implementation of the analog data assimilation. In this respect, the proposed patch-based framework intrinsically ensures high parallelization performance. From a methodological point of view, a relative weakness of the analog forecasting models may be their low physical

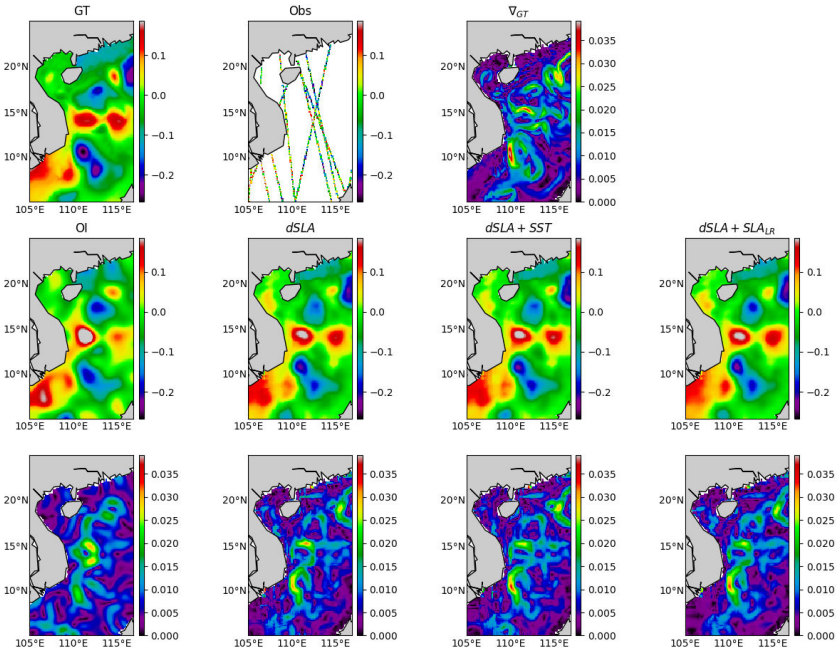


(a)

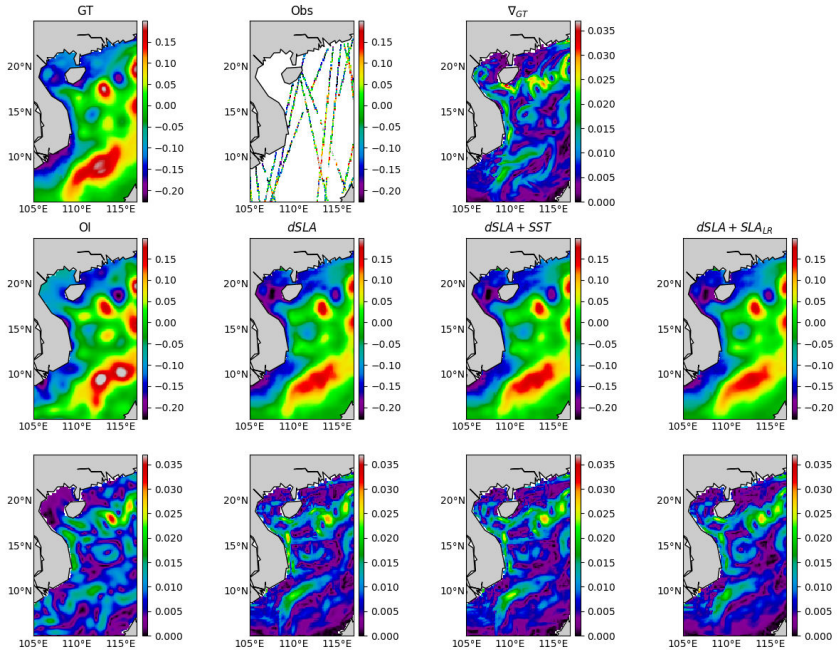


(b)

Figure 4.9 – (Noisy observation) Reconstruction of SLA fields using MS-AnDA with different multivariate regression models on day 51^{th} (a) & 54^{th} (b)



(a)



(b)

Figure 4.10 – (Noisy observation) Reconstruction of SLA fields using MS-AnDA with different multivariate regression models on day 57^{th} & 237^{th} (b)

interpretation compared with physically-derived priors [149]. The combination of such physically-derived parameterizations to data-driven strategies appear as a promising research direction.

Part III

Conclusion

Chapter 5

Conclusions and Perspectives

We can't plan life. All we can do is be available for it.

Lauryn Hill

5.1 Conclusion

In this thesis, we studied the extension of analog forecasting methods to data assimilation issues, and have set the foundations of the Analog Data Assimilation. By seizing the opportunity offered by the increasing amount of geophysical information, our method rely on the exploitation of the available large-scale observation and/or simulation/reanalysis dataset using nearest neighbors schemes to improve the analysis of new observations. The Analog Data Assimilation can be either seen as a data-driven alternative to classical data assimilation in case the latter is difficult to perform, or as a support to classical data assimilation in situations where both can be exploited.

Different tests were performed all along this thesis on different types of datasets, from toy models (Lorenz-63 and Lorenz-96) to realistic datasets (Sea Surface Temperature (SST) and Sea Level Anomaly (SLA)). We have shown the relevance of our method and its potential. In particular, we highlight the benefit of using weighted local linear techniques as an analog

forecasting operator which resorts to the best reconstruction. Experiments conducted on either satellite-derived fields or numerical simulation data illustrate that the resulting reconstructed fields are with higher resolution than the classical Optimal Interpolation algorithm.

Since the analog forecasting depends highly on the K-Nearest Neighbors algorithm, the curse of dimensionality was our biggest challenge. we have shown that breaking the geophysical region field of interest into small subregions using patch-based representation helps in reducing the complexity of our algorithm. Moreover, projecting the patches series using EOF-based representations using few tens of coefficients, yields to a settings where analog forecasting is simple and efficient.

The flexible framework we offer has the advantage of accounting auxiliary variable with less implementation effort. We therefore have shown that considering inter-scale dependencies for the Sea Level Anomaly (SLA) has more benefit than considering synergies of SLA and SST data.

5.2 Perspectives and Future Work

5.2.1 The Analog data assimilation and its applications

We believe that this thesis opens new research avenues for the analysis, reconstruction and understanding of the dynamics of geophysical systems using data-driven techniques. Such techniques will benefit from the increasing availability of large-scale historical observational and/or simulated datasets.

Beyond the wide range of possible applications, future research should further investigate methodological issues. First of all, our study demonstrates the relevance of the analog particle filter, but as mentioned in Chapter 2, the AnPF suffers from degeneracy and sample impoverishment. We may point out that complementary experiments with particle smoother schemes (not shown) resulted in numerical instabilities. The derivation of the Analog Particle Smoother then remains an open question. In addition to advanced particle filters as proposed in [124, 153], one might also benefit from the straightforward applications of the analog procedure in reverse time, which is not generally possible for model-driven schemes. A second direction for future work lies in the design of the kernel used by the analog forecasting operators. Whereas we considered a Gaussian kernel, other kernels have been proposed in the literature, for instance using Procrustes distance instead of the Euclidean distance [111] or different weighing strategies [40]. The explicit

derivation of the mapping associated with a kernel as considered in [163] may also be a promising alternative to state the analog data assimilation in a kernel-derived lower-dimensional space. The theoretical characterization of the asymptotic behavior of analog data assimilation schemes is also an interesting avenue of research. Similarly to the theoretical analysis of ensemble Kalman filters and particle filters [90], the derivation of convergence conditions, possibly associated with reconstruction bounds, would be of key interest to bound the reconstruction performance of the proposed analog schemes with respect to their model-driven counterpart.

For ocean related applications, the results obtained in this thesis call for exploring more research directions that combine the analog strategies with model-derived and/or statistical priors. SST for example is generally assumed to consider an advection-diffusion prior model drifted by the SSH, this information could be used in constraining the local analog regression for the reconstruction of SST. Statistical priors can also be injected into AnDA schemes. In particular, priors on the spatial covariances and the marginal distributions of high resolution details, as done with SST in [49], are expected to result in more geophysically plausible reconstructions. Investigating more synergies between ocean variables can also be of interest [146, 148], an interesting case might be the exploration of relationships between observable and non observable variables. For example we can think of exploiting 4D numerical simulations (3D + depth) to retrieve variables such as vertical velocities or mixed layer depth from satellite-derived observations of ocean surface variables. Preparing the inversion of the future altimetry mission SWOT (CNES/NASA) is a perfect context to carry on such research plans. SWOT mission promises an unprecedented coverage around the globe. More specifically, the large swath is expected to provide a large number of data, urging for the inspection of the potential improvements that this new mission will bring compared to classical along-track data. In the context of analog data assimilation, the interest of SWOT data may be two-fold. First, regarding the observation model, SWOT mission will both significantly increase the number of available observation data and enable the definition of more complex observation models exploiting for instance velocity-based or vorticity-based criterion. Second, SWOT data might also be used to build representative patch-level catalogs of exemplars. Future work should investigate these two directions using simulated SWOT test-beds [58].

Another future research path would be the investigation of the influence of data on the AnDA for remote sensing applications. More specifically, addressing questions we did not answer here, examples comprise the calculation of the number of years of data needed to reach a

consistent reconstruction, the use of nonlinear dimensionality reduction algorithm instead of the EOF/PCA, etc..

While this thesis work was ocean science oriented, it can be clearly seen that the Analog Data Assimilation is domain-free and could be applied to any dynamical system where an archived dataset is available and where the dynamics present a "repeatability" behavior. To support this claim, we applied successfully the AnDA to two non ocean related applications: the interpolation of dynamical textures sequences [98], and the retrieval of missing data in motion capture series [94]. Although we focused on the problem of the interpolation of missing data, applying the AnDA might be relevant to other inverse problems (*e.g.* denoising, deconvolution).

5.2.2 Machine Learning for dynamical systems

Data-driven approaches are starting to reach a good level of maturity with interesting applications in geoscience and satellite remote sensing [24, 25]. Motivated by the increasing and challenging amount of data, researchers from the data science and statistical learning fields are tempted to explore the large avenue of ideas that is finally open to them.

While we placed our faith in analog methods in this thesis, and results were delightfully encouraging. We call for investigating other techniques for emulating the underlying governing equations from data. Examples comprise, but are not limited to, sparsity-promoting techniques [18,128], deep learning techniques [55,87], manifold learning [136], etc. A review work on data-driven methods for dynamical systems would be highly appreciated. Hybrid methods that combine data-driven and model-driven strategies could certainly be of interest.

5.2.3 Deep Learning for detection and classification of eddies from SSH maps

In Appendix B, we describe an example of an ocean remote sensing problem that could be tackled using Deep Learning techniques. Detecting and classifying eddies from SSH maps is a classical example where geometry-based techniques are competing with physical-based techniques. In our work, instead of using geometry-based techniques we treat the problem under a computer vision perspective. We implemented and compared several neural network architectures that are used in image segmentation tasks. Initial results shown in Appendix B are encouraging and calls for considering and putting more efforts into Deep Learning techniques.

It goes without saying that Deep Learning methods are revolutionizing the machine learning and computer vision fields. However, the astonishing promised impact of these methods did not

reach yet the geoscience and remote sensing community. This can be explained by the "black-box" nature of these methods that makes it hard for geoscientists to relate results to theoretical physical and equations-based models. A non negligible effort should then be deployed to improve physical understanding of neural networks based methods.

List of Figures

List of Tables

Bibliography

[1] Brian D. O. Anderson and John B. Moore. *Optimal filtering. Reprint of the 1979 original ed.* Mineola, NY: Dover Publications, reprint of the 1979 original ed. edition, 2005.

[2] Jeffrey L Anderson. An ensemble adjustment Kalman filter for data assimilation. *Monthly Weather Review*, 129(12):2884–2903, 2001.

[3] Jeffrey L Anderson. Exploring the need for localization in ensemble data assimilation using a hierarchical ensemble filter. *Physica D: Nonlinear Phenomena*, 230(1):99–111, 2007.

[4] Jeffrey L Anderson. Localization and sampling error correction in ensemble Kalman filter data assimilation. *Monthly Weather Review*, 140(7):2359–2371, 2012.

[5] Jeffrey L Anderson and Stephen L Anderson. A monte carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts. *Monthly Weather Review*, 127(12):2741–2758, 1999.

[6] Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data assimilation: methods, algorithms, and applications*. Fundamentals of Algorithms. SIAM, 2016.

[7] Mohammad D Ashkezari, Christopher N Hill, Christopher N Follett, Gaël Forget, and Michael J Follows. Oceanic eddy detection and lifetime forecast using machine learning methods. *Geophysical Research Letters*, 43(23), 2016.

[8] Aitor Atencia and Isztar Zawadzki. A comparison of two techniques for generating nowcasting ensembles. part ii: Analogs selection and comparison of techniques. *Monthly Weather Review*, 143(7):2890–2908, 2015.

[9] Nicolas Audebert, Bertrand Le Saux, and Sébastien Lefèvre. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. *arXiv preprint arXiv:1609.06846*, 2016.

[10] S. Ba, T. Corpetti, and R. Fablet. Multi-resolution missing data interpolation in SST Image Series. In *IEEE Int. Conf. on Image processing*, Bruxelles, Belgium, September 2011.

[11] K. Baith, R. Lindsay, G. Fu, and C.R. McClain. Data analysis system developed for ocean color satellite sensors. *Eos, Transactions American Geophysical Union*, 82(18):202–202, May 2001.

[12] TP Barnett and RW Preisendorfer. Multifield analog prediction of short-term climate fluctuations using a climate state vector. *Journal of the Atmospheric Sciences*, 35(10):1771–1787, 1978.

[13] J. M. Beckers and M. Rixen. EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. *Journal of Atmospheric and Oceanic Technology*, 20(12):1839–1856, December 2003.

[14] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

[15] M. Bertalmio, A. L. Bertozzi, and G. Sapiro. Navier-Stokes, fluid dynamics, and image and video inpainting. In *Proc. IEEE Computer Vision and Pattern Recognition, CVPR'01*, pages 355–362, 2001.

[16] Neill E Bowler, Jonathan Flowerdew, and Stephen R Pring. Tests of different flavours of enkf on a simple model. *Quarterly Journal of the Royal Meteorological Society*, 139(675):1505–1519, 2013.

[17] Francis P Bretherton, Russ E Davis, and CB Fandry. A technique for objective analysis and design of oceanographic experiments applied to mode-73. In *Deep Sea Research and Oceanographic Abstracts*, volume 23, pages 559–582. Elsevier, 1976.

[18] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[19] A. Buades, B. Coll, and J. M. Morel. A non-local algorithm for image denoising. In *IEEE Conf. on Computer Vision and Pattern Recognition, CVPR'05*, volume 2, pages 60–65 vol. 2, June 2005.

[20] B. Buongiorno Nardelli, A. Pisano, C. Tronconi, and R. Santoleri. Evaluation of different covariance models for the operational interpolation of high resolution satellite Sea Surface Temperature data over the Mediterranean Sea. *Remote Sensing of Environment*, 164:334–343, July 2015.

[21] Gerrit Burgers, Peter Jan van Leeuwen, and Geir Evensen. Analysis scheme in the ensemble Kalman filter. *Monthly weather review*, 126(6):1719–1724, 1998.

[22] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data Assimilation in the Geosciences - An overview on methods, issues and perspectives. *ArXiv e-prints*, September 2017.

[23] M Castellani. Identification of eddies from sea surface temperature maps with neural networks. *International journal of remote sensing*, 27(8):1601–1618, 2006.

[24] Christopher Chapman and Anastase Alexandre Charantonis. Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps. *IEEE Geosci. Remote Sensing Lett.*, 14(5):617–620, 2017.

[25] Anastase Alexandre Charantonis, Julien Brajard, Cyril Moulin, Bardan Fouad, and Sylvie Thiria. Inverse method for the retrieval of ocean vertical profiles using self organizing maps and hidden markov models - application on ocean colour satellite image inversion. In *NCTA 2011 - Proceedings of the International Conference on Neural Computation Theory and Applications [part of the International Joint Conference on Computational Intelligence IJCCI 2011], Paris, France, 24-26 October, 2011*, pages 316–321, 2011.

[26] D.B. Chelton and F.J. Wentz. Global Microwave Satellite Observations of Sea Surface Temperature for Numerical Weather Prediction and Climate Research - ProQuest. *Bulletin of the American Meteorological Society*, 86(8):1097, 2005.

[27] Dudley B. Chelton, J.C. Ries, B. J. Haines, L.-L. Fu, and P. S. Callahan. Satellite Altimetry. In A. Cazenave and L.-L. Fu, editors, *International Geophysics*, volume 69 of *Satellite Altimetry and Earth SciencesA Handbook of Techniques and Applications*, pages 1–ii. Academic Press, 2001.

[28] Dudley B Chelton, Michael G Schlax, and Roger M Samelson. Global observations of nonlinear mesoscale eddies. *Progress in Oceanography*, 91(2):167–216, 2011.

[29] Dudley B Chelton, Michael G Schlax, Roger M Samelson, and Roland A de Szoeke. Global observations of large oceanic eddies. *Geophysical Research Letters*, 34(15), 2007.

[30] Zhe Chen. Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond. Technical report, McMaster University, 2003.

[31] TM Chin, MJ Turmon, JB Jewell, and M Ghil. An ensemble-based smoother with retrospectively updated weights for highly nonlinear systems. *Monthly weather review*, 135(1):186–202, 2007.

[32] Thierry Chonavel. *Statistical signal processing: modelling and estimation*. Springer Science & Business Media, 2002.

[33] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[34] Darin Comeau, Dimitrios Giannakis, Zhizhen Zhao, and Andrew J Majda. Predicting regional and pan-arctic sea ice anomalies with kernel analog forecasting. *arXiv preprint arXiv:1705.05228*, 2017.

[35] Emmanuel Cosme, Jacques Verron, Pierre Brasseur, Jacques Blum, and Didier Auroux. Smoothing problems in a bayesian framework and their linear gaussian solutions. *Monthly Weather Review*, 140(2):683–695, 2012.

[36] A. Criminisi, P. Perez, and K. Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212, September 2004.

[37] Pierre De Mey and Allan R Robinson. Assimilation of altimeter eddy fields in a limited-area quasi-geostrophic model. *Journal of physical oceanography*, 17(12):2280–2293, 1987.

[38] Charles-Alban Deledalle, Joseph Salmon, and Arnak S. Dalalyan. Image denoising with patch based PCA: local versus global. pages 25.1–25.10. BMVA Press, August 2011.

[39] Luca Delle Monache, Irina Djalalova, and James Wilczak. Analog-based postprocessing methods for air quality forecasting. In *Air Pollution Modeling and its Application XXIII*, pages 237–239. Springer, 2014.

[40] Luca Delle Monache, Thomas Nipen, Yubao Liu, Gregory Roux, and Roland Stull. Kalman filter and analog schemes to postprocess numerical weather predictions. *Monthly Weather Review*, 139(11):3554–3570, 2011.

[41] C. J. Donlon, M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Xindong. The Operational Sea Surface Temperature and Sea Ice Analysis (OSTIA) system. *Remote Sensing of Environment*, 116:140–158, January 2012.

[42] John R Dormand and Peter J Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.

[43] A. A. Efros and W.T. Freeman. Image Quilting for Texture Synthesis and Transfer. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 341–346, New York, NY, USA, 2001. ACM.

[44] Robert D. Elliott. *Extended-Range Forecasting by Weather Types*, pages 834–840. American Meteorological Society, Boston, MA, 1951.

[45] R. Escudier, J. Bouffard, A. Pascual, P.-M. Poulain, and M.-I. Pujol. Improvement of coastal and mesoscale observation from space: Application to the northwestern Mediterranean Sea. *Geophysical Research Letters*, 40(10):2148–2153, 2013.

[46] G. Evensen. *Data Assimilation*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[47] Geir Evensen and Peter Jan Van Leeuwen. An ensemble Kalman smoother for nonlinear dynamics. *Monthly Weather Review*, 128(6):1852–1867, 2000.

[48] R. Fablet and F. Rousseau. Missing data super-resolution using non-local and statistical priors. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 676–680, September 2015.

[49] R. Fablet and F. Rousseau. Joint Interpolation of Multisensor Sea Surface Temperature Fields Using Nonlocal and Statistical Priors. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9(6):2665–2675, June 2016.

[50] R. Fablet, P. H. Viet, and R. Lguensat. Data-driven Models for the Spatio-Temporal Interpolation of satellite-derived SST Fields. *IEEE Transactions on Computational Imaging*, 2017.

[51] Ronan Fablet, Jacques Verron, Baptiste Mourre, Bertrand Chapron, and Ananda Pascual. Improving mesoscale altimetric data from a multi-tracer convolutional processing of standard satellite-derived products. working paper or preprint, October 2016.

[52] Ronan Fablet, Phi Huynh Viet, and Redouane Lguensat. Data-driven assimilation of irregularly-sampled image time series. In *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017.

[53] James H Faghmous, Ivy Frenger, Yuanshun Yao, Robert Warmka, Aron Lindell, and Vipin Kumar. A daily global mesoscale ocean eddy dataset from satellite altimetry. *Scientific data*, 2, 2015.

[54] James H Faghmous, Luke Styles, Varun Mithal, Shyam Boriah, Stefan Liess, Vipin Kumar, Frode Vikebø, and Michel dos Santos Mesquita. Eddyscan: A physically consistent ocean eddy monitoring application. In *Intelligent Data Understanding (CIDU), 2012 Conference on*, pages 96–103. IEEE, 2012.

[55] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. Sequential neural models with stochastic layers. In *Advances in Neural Information Processing Systems*, pages 2199–2207, 2016.

[56] W. T. Freeman and Liu. Markov Random Fields for Super-Resolution. In *Advances in Markov Random Fields for Vision and Image Processing*. MIT Press, a. blake, p. kohli, and c. rother, eds. edition, 2011.

[57] Levi Gandin. Objective analysis of meteorological fields. by L. S. Gandin. translated from the russian. jerusalem (israel program for scientific translations), 1965. pp. vi, 242: 53 figures; 28 tables. £4 1s. 0d. *Quarterly Journal of the Royal Meteorological Society*, 92(393):447–447, 1966.

[58] L. Gaultier, C. Ubelmann, and L.-L. Fu. The Challenge of Using Future SWOT Data for Oceanic Field Reconstruction. *Journal of Atmospheric and Oceanic Technology*, 33(1):119–126, November 2015.

[59] L. Gaultier, Jacques Verron, Jean-Michel Brankart, Olivier Titaud, and Pierre Brasseur. On the inversion of submesoscale tracer fields to estimate the surface ocean circulation. *Journal of Marine Systems*, 126:33–42, October 2013.

[60] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[61] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *IEEE Proceedings F, Radar and Signal Processing*, 140(2):107–113, 1993.

[62] Mbaye Babacar Gueye, Awa Niang, Sabine Arnault, Sylvie Thiria, and Michel Crépon. Neural approach to inverting complex system: Application to ocean salinity profile estimation from surface parameters. *Computers & Geosciences*, 72:201–209, 2014.

[63] J. Gula, M.J. Molemaker, and J.C. McWilliams. Submesoscale Cold Filaments in the Gulf Stream. *Journal of Physical Oceanography*, 44(10):2617–2643, July 2014.

[64] J Hai, Ya Xiaomei, G Jianming, and G Zhenyu. Automatic eddy extraction from sst imagery using artificial neural network. *The international archives of the photogrammetry, remote sensing and spatial information science*, pages 279–282, 2008.

[65] Franz Hamilton, Tyrus Berry, and Timothy Sauer. Ensemble Kalman filtering without a model. *Physical Review X*, 6(1):011021, 2016.

[66] Bruce Hansen. Econometrics textbook, 2000.

[67] N.J. Hardman-Mountford, A.J. Richardson, D.C. Boyer, A. Kreiner, and H.J. Boyer. Relating sardine recruitment in the northern benguela to satellite-derived sea surface height using a neural network pattern recognition approach. *Progress in Oceanography*, 59(2):241 – 255, 2003. ENVIFISH: Investigating environmental causes of pelagic fisheries variability in the SE Atlantic.

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[69] L. He-Guelton, R. Fablet, B. Chapron, and J. Tournadre. Learning-based emulation of sea surface wind fields from numerical model outputs and sar data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 8(10):4742–4750, Oct 2015.

[70] William R Holland. 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.

[71] Song-You Hong and Jimy Dudhia. Next-generation numerical weather prediction: Bridging parameterization, explicit clouds, and large eddies. *Bulletin of the American Meteorological Society*, 93(1):ES6, 2012.

[72] Pascal Horton, Michel Jaboyedoff, and Charles Obled. Global optimization of an analog method by means of genetic algorithms. *Monthly Weather Review*, 145(4):1275–1294, 2017.

[73] Ibrahim Hoteit, Xiaodong Luo, and Dinh-Tuan Pham. Particle Kalman filtering: A nonlinear bayesian framework for ensemble Kalman filters. *Monthly Weather Review*, 140:528–542, 2012.

[74] Ibrahim Hoteit, D-T Pham, ME Gharamti, and X Luo. Mitigating observation perturbation sampling errors in the stochastic enkf. *Monthly Weather Review*, 143(7):2918–2936, 2015.

[75] Ibrahim Hoteit, Dinh-Tuan Pham, George Triantafyllou, and Gerasimos Korres. A new approximate solution of the optimal nonlinear filter for data assimilation in meteorology and oceanography. *Monthly Weather Review*, 136(1):317–334, 2008.

[76] Dongmei Huang, Yanling Du, Qi He, Wei Song, and A. Liotta. Deepeddy: A simple deep architecture for mesoscale oceanic eddy detection in sar images. In *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*, pages 673–678, May 2017.

[77] J. Isern-Fontanet, B. Chapron, G. Lapeyre, and P. Klein. Potential use of microwave sea surface temperatures for the estimation of ocean currents. *GEOPHYSICAL RESEARCH LETTERS*, 33, 2006. L24608.

[78] J. Isern-Fontanet, M. Shinde, and C. Andersson. On the Transfer Function between Surface Fields and the Geostrophic Stream Function in the Mediterranean Sea. *Journal of Physical Oceanography*, 44(5):1406–1423, March 2014.

[79] Jordi Isern-Fontanet, Emilio García-Ladona, and Jordi Font. Identification of marine eddies from altimetric maps. *Journal of Atmospheric and Oceanic Technology*, 20(5):772–778, 2003.

[80] Manel Jouini, Marina Lévy, Michel Crépon, and Sylvie Thiria. Reconstruction of satellite chlorophyll images under heavy cloud coverage using a neural classification method. *Remote sensing of environment*, 131:232–246, 2013.

[81] Simon J. Julier and Jeffrey K. Uhlmann. A new extension of the kalman filter to nonlinear systems. pages 182–193, 1997.

[82] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82(1):35–45, 1960.

[83] Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003.

[84] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *arXiv preprint arXiv:1706.02515*, 2017.

[85] P. Klein, J. Isern-Fontanet, G. Lapeyre, G. Roullet, E. Danioux, B. Chapron, S. Le Gentil, and H. Sasaki. Diagnosis of vertical velocities in the upper ocean from high resolution sea surface height. *Geophysical Research Letters*, 36(12):L12603, June 2009.

[86] V. Klemas. Remote Sensing of Sea Surface Salinity: An Overview with Case Studies. *Journal of Coastal Research*, pages 830–838, July 2011.

[87] Rahul G Krishnan, Uri Shalit, and David Sontag. Deep kalman filters. *arXiv preprint arXiv:1511.05121*, 2015.

[88] David J Lary, Amir H Alavi, Amir H Gandomi, and Annette L Walker. Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10, 2016.

[89] F.X. Le-Dimet and O. Talagrand. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus*, pages 97–110, 1986.

[90] François Le Gland, Valérie Monbet, and Vu-Duc Tran. *Large sample asymptotics for the ensemble Kalman filter*. PhD thesis, INRIA, 2009.

[91] Julien Le Sommer, Francesco d'Ovidio, and Gurvan Madec. Parameterization of subgrid stirring in eddy resolving ocean models. part 1: Theory and diagnostics. *Ocean Modelling*, 39(1):154–169, 2011.

[92] PY Le Traon, F Nadal, and N Ducet. An improved mapping method of multisatellite altimeter data. *Journal of Atmospheric and Oceanic Technology*, 15(2):522–534, 1998.

[93] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[94] Redouane Lguensat, Ronan Fablet, Pierre Ailliot, and Pierre Tandeo. An exemplar-based hidden markov model framework for nonlinear state-space models. In *Signal Processing Conference (EUSIPCO), 2016 24th European*, pages 2340–2344. IEEE, 2016.

[95] Redouane Lguensat, Miao Sun, Ge Chen, Fenglin Tian, and Ronan Fablet. Spatio-Temporal Interpolation Of Altimeter-Derived SSH Fields Using Analog Data Assimilation: A Case-Study In The South China Sea. 2017.

[96] Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Bertrand Chapron, and Ronan Fablet. Using archived datasets for missing data interpolation in ocean remote sensing observation series. In *OCEANS 2016-Shanghai*, pages 1–5. IEEE, 2016.

[97] Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet. The analog data assimilation. *Monthly Weather Review*, 0(0):null, 2017.

[98] Redouane Lguensat, Pierre Tandeo, Ronan Fablet, and Pierre Ailliot. Non-parametric ensemble kalman methods for the inpainting of noisy dynamic textures. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 4288–4292. IEEE, 2015.

[99] J. Li and A. D. Heap. Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software*, 53:173–189, March 2014.

[100] AC Lorenc, SP Ballard, RS Bell, NB Ingleby, PLF Andrews, DM Barker, JR Bray, AM Clayton, T Dalby, D Li, et al. The met. office global three-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society*, 126(570):2991–3012, 2000.

[101] Andrew C Lorenc. Analysis methods for numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 112(474):1177–1194, 1986.

[102] Edward N Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4):636–646, 1969.

[103] Edward N Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on predictability*, volume 1, 1996.

[104] L. Lorenzi, F. Melgani, and G. Mercier. Inpainting strategies for reconstruction of missing data in VHR images. *IEEE Geoscience and Remote Sensing Letters*, 8(5):914–918, 2011.

[105] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):645–657, Feb 2017.

[106] S. Mallat. *A wavelet tour of signal processing, second edition*. Academic Press, 1999.

[107] Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

[108] Evan Mason, Ananda Pascual, Peter Gaube, Simón Ruiz, Josep L. Pelegrí, and Antoine Delepoulle. Subregional characterization of mesoscale eddies across the brazil-malvinas confluence. *Journal of Geophysical Research: Oceans*, 122(4):3329–3357, 2017.

[109] Evan Mason, Ananda Pascual, and James C McWilliams. A new sea surface height–based code for oceanic mesoscale eddy tracking. *Journal of Atmospheric and Oceanic Technology*, 31(5):1181–1188, 2014.

[110] Yukio Masumoto, Hideharu Sasaki, Takashi Kagimoto, Nobumasa Komori, Akio Ishida, Yoshikazu Sasai, Toru Miyama, Tatsuo Motoi, Humio Mitsudera, Keiko Takahashi, et al. A fifty-year eddy-resolving simulation of the world ocean: Preliminary outcomes of ofes (ogcm for the earth simulator). *J. Earth Simulator*, 1:35–56, 2004.

[111] Patrick L. McDermott and Christopher K. Wikle. A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, pages n/a–n/a, 2015.

[112] James C McWilliams. The nature and consequences of oceanic eddies. *Ocean Modeling in an Eddying Regime*, pages 5–15, 2008.

[113] Robert N Miller, Michael Ghil, and Francois Gauthiez. Advanced data assimilation in strongly nonlinear dynamical systems. *Journal of the Atmospheric Sciences*, 51(8):1037–1056, 1994.

[114] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 565–571. IEEE, 2016.

[115] Badrinath Nagarajan, Luca Delle Monache, Joshua P Hacker, Daran L Rife, Keith Searight, Jason C Knievel, and Thomas N Nipen. An evaluation of analog-based post-processing methods across several variables and forecast models. *Weather and Forecasting*, (2015), 2015.

[116] Akira Okubo. Horizontal dispersion of floatable particles in the vicinity of velocity singularities such as convergences. In *Deep sea research and oceanographic abstracts*, volume 17, pages 445–454. Elsevier, 1970.

[117] Stephen M Omohundro. *Five balltree construction algorithms*. International Computer Science Institute Berkeley, 1989.

[118] Edward Ott, Brian R Hunt, Istvan Szunyogh, Aleksey V Zimin, Eric J Kostelich, Matteo Corazza, Eugenia Kalnay, DJ Patil, and James A Yorke. A local ensemble Kalman filter for atmospheric data assimilation. *Tellus A*, 56(5):415–428, 2004.

[119] Nicolas Papadakis, Étienne Mémin, Anne Cuzol, and Nicolas Gengembre. Data assimilation with the weighted ensemble kalman filter. *Tellus A*, 62(5):673–697, 2010.

[120] Ananda Pascual, Yannice Faugère, Gilles Larnicol, and Pierre-Yves Le Traon. Improved description of the ocean mesoscale variability by combining four satellite altimeters. *Geophysical Research Letters*, 33(2), 2006.

[121] G. Peyré, S. Bougleux, and L.D. Cohen. Non-local Regularization of Inverse Problems. *Inverse Problems and Imaging*, 5(2):511–530, 2011.

[122] Dinh Tuan Pham. Stochastic methods for sequential data assimilation in strongly nonlinear systems. *Monthly weather review*, 129(5):1194–1207, 2001.

[123] Bo Ping, Fenzhen Su, and Yunshan Meng. An Improved DINEOF Algorithm for Filling Missing Values in Spatio-Temporal Sea Surface Temperature Data. *PLOS ONE*, 11(5):e0155928, May 2016.

[124] Michael K Pitt and Neil Shephard. Filtering via simulation: Auxiliary particle filters. *Journal of the American statistical association*, 94(446):590–599, 1999.

[125] Paul Bui Quang, Christian Musso, and Francois Le Gland. An insight into the issue of dimensionality in particle filtering. In *Information Fusion (FUSION), 2010 13th Conference on*, pages 1–8. IEEE, 2010.

[126] Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[127] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[128] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

[129] Juan Jose Ruiz, Manuel Pulido, and Takemasa Miyoshi. Estimating model parameters with ensemble-based data assimilation: A review. *Journal of the Meteorological Society of Japan*, 91(2):79–99, 2013.

[130] I Ari Sadarjoen and Frits H Post. Geometric methods for vortex extraction. In *Data Visualization'99*, pages 53–62. Springer, 1999.

[131] H. Sasaki and P. Klein. SSH Wavenumber Spectra in the North Pacific from a High-Resolution Realistic Simulation. *Journal of Physical Oceanography*, 42(7):1233–1241, May 2012.

[132] Hideharu Sasaki, Masami Nonaka, Yukio Masumoto, Yoshikazu Sasai, Hitoshi Uehara, and Hirofumi Sakuma. An eddy-resolving hindcast simulation of the quasi-global ocean from 1950 to 2003 on the earth simulator, 2008.

[133] B. Saulquin, F. Gohin, and R. Garrello. Regional Objective Analysis for Merging High-Resolution MERIS, MODIS/Aqua, and SeaWiFS Chlorophyll- a Data From 1998 to 2008 on the European Atlantic Shelf. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1):143–154, January 2011.

[134] F Schenk and E Zorita. Reconstruction of high resolution atmospheric fields for northern europe using analog-upscaling. *Climate of the Past*, 8(5):1681–1703, 2012.

[135] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[136] Ronen Talmon, Stéphane Mallat, Hitten Zaveri, and Ronald R Coifman. Manifold learning for latent variable inference in dynamical systems. *IEEE Transactions on Signal Processing*, 63(15):3843–3856, 2015.

[137] P. Tandeo, P. Ailliot, and E. Autret. Linear Gaussian state-space model with irregular sampling: application to sea surface temperature. *Stochastic Environmental Research and Risk Assessment*, 25(6):793–804, November 2010.

[138] P. Tandeo, E. Autret, B. Chapron, R. Fablet, and R. Garello. SST spatial anisotropic covariances from METOP-AVHRR data. *Remote Sensing of Environment*, 141:144–148, February 2014.

[139] Pierre Tandeo, Pierre Ailliot, Juan Ruiz, Alexis Hannart, Bertrand Chapron, Anne Cuzol, Valérie Monbet, Robert Easton, and Ronan Fablet. Combining analog method and ensemble data assimilation: application to the lorenz-63 chaotic system. In *Machine Learning and Data Mining Approaches to Climate Science*, pages 3–12. Springer, 2015.

[140] Pierre Tandeo, Manuel Pulido, and François Lott. Offline parameter estimation using enkf and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Quarterly Journal of the Royal Meteorological Society*, 141(687):383–395, 2015.

[141] H. Tanizaki. On the nonlinear and nonnormal filter using rejection sampling. *IEEE Transactions on Automatic Control*, 44(2):314–319, Feb 1999.

[142] T. Tasdizen. Principal Neighborhood Dictionaries for Nonlocal Means Image Denoising. *IEEE Transactions on Image Processing*, 18(12):2649–2660, December 2009.

[143] Sylvie Thiria, Yves Lechevallier, Olivier Gascuel, and Stéphane Canu. *Statistique et méthodes neuronales*, volume 4. Dunod Paris, 1997.

[144] Olivier Thual. Introduction to Data Assimilation for Scientists and Engineers, Open Learn. Res. Ed. INPT 0202 6h, 2013.

[145] Zoltan Toth. Long-range weather forecasting using an analog approach. *Journal of climate*, 2(6):594–607, 1989.

[146] A. Turiel, V. Nieves, E. Garcia-Ladona, J. Font, M.-H. Rio, and G. Larnicol. The multifractal structure of satellite sea surface temperature maps can be used to obtain global maps of streamlines. *Ocean Science*, 5(4):447–460, 2009.

[147] Antonio Turiel, Jordi Isern-Fontanet, and Emilio García-Ladona. Wavelet filtering to extract coherent vortices from altimetric data. *Journal of Atmospheric and Oceanic Technology*, 24(12):2103–2119, 2007.

[148] Antonio Turiel, Jordi Sole, Veronica Nieves, Joaquim Ballabrera-Poy, and Emilio Garcia-Ladona. Tracking oceanic currents by singularity analysis of Microwave Sea Surface Temperature images. *Remote Sensing of Environment*, In Press, 2009.

[149] C. Ubelmann, P. Klein, and L.-L. Fu. Dynamic Interpolation of Sea Surface Height and Potential Applications for Future High-Resolution Altimetry Mapping. *Journal of Atmospheric and Oceanic Technology*, 32(1):177–184, October 2014.

[150] M. Umbert, N. Hoareau, A. Turiel, and J. Ballabrera-Poy. New blending algorithm to synergize ocean variables: The case of SMOS sea surface salinity maps. *Remote Sensing of Environment*, 146:172–187, April 2014.

[151] HM Van den Dool. Searching for analogues, how long must we wait? *Tellus A*, 46(3):314–324, 1994.

[152] Peter Jan Van Leeuwen. Particle filtering in geophysical systems. *Monthly Weather Review*, 137(12):4089–4114, 2009.

[153] Peter Jan Van Leeuwen. Nonlinear data assimilation in geosciences: an extremely efficient particle filter. *Quarterly Journal of the Royal Meteorological Society*, 136(653):1991–1999, 2010.

[154] Michele Volpi and Devis Tuia. Dense semantic labeling of subdecimeter resolution images with convolutional neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):881–893, 2017.

[155] John Weiss. The dynamics of enstrophy transfer in two-dimensional hydrodynamics. *Physica D: Nonlinear Phenomena*, 48(2-3):273–294, 1991.

[156] Jeffrey S Whitaker and Thomas M Hamill. Ensemble data assimilation without perturbed observations. *Monthly Weather Review*, 130(7):1913–1924, 2002.

[157] Robert L Wilby and TML Wigley. Downscaling general circulation model output: a review of methods and limitations. *Progress in Physical Geography*, 21(4):530–548, 1997.

[158] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[159] J Yi, Y Du, Z He, and C Zhou. Enhancing the accuracy of automatic eddy detection and the capability of recognizing the multi-core structures from maps of sea level anomaly. *Ocean Science*, 10(1):39–48, 2014.

[160] P Yiou. Anawege: a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, 7(2):531–543, 2014.

[161] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2528–2535. IEEE, 2010.

[162] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016.

[163] Z. Zhao and D. Giannakis. Analog Forecasting with Dynamics-Adapted Kernels. *arXiv:1412.3831 [physics]*, December 2014. arXiv: 1412.3831.

[164] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer. Deep learning in remote sensing: a review. *ArXiv e-prints*, October 2017.

Appendix A

Operational count of the AnDA applied for high-dimensional applications

This appendix aims at giving an estimate of the operations involved when applying the AnDA for a realistic large-scale application. We discuss the computational cost of the analog forecasting, which is specific to the AnDA. The later directly relates to the cost of the K-Nearest Neighbor (K-NN) step.

In case of large-scale catalogs, an exhaustive search strategy is not suitable and the use of space-partitioning data structures, the most popular ones being *K-d trees* [14] and *Ball trees* [117], appears necessary. These structures speed up the K-NN search, at the expense of an approximate search for nearest neighbors. Let us denote by D the dimension of the system of interest. Making a choice between K-d trees or ball trees depends mostly on the dimensionality of the system. K-d trees are known to perform well in dimensions $D < 20$, while ball trees are more suitable to dimensions higher than 20 but come with a high cost of space-partitioning [158]. In this appendix we focus on the use of K-d trees, which are natural candidates for local analogs with a small component-wise local neighborhood ν or using a preliminary dimensionality reduction algorithm (such as Empirical Orthogonal Functions). A comparison between K-d trees and ball trees is out of the scope of this work.

Let N_{data} be the size of the catalog (the number of samples from where to look for analogs), and K the number of nearest neighbors to be retrieved. Let us recall that ν is the size of the local neighborhood used for the search for local analogs. [151] derived a relationship between

the local neighborhood size and the amount of the data needed to find an analog with a given precision. With the assumption that the components of the states follow a multivariate Gaussian distribution and have the same variance sd^2, finding K samples that have a distance lower than ϵ for all the components of the neighborhood with a probability of 95%, needs the number of data to be on average:

- Global analogs:
$$N_{global} \geq K \frac{ln(0.05)}{ln(1-\alpha^D)} \simeq \frac{3K}{\alpha^D}, \tag{A.1}$$

- Local analogs:
$$N_{local} \geq K \frac{ln(0.05)}{ln(1-\alpha^{2\nu+1})} \simeq \frac{3K}{\alpha^{2\nu+1}}, \tag{A.2}$$

where α is the integral of the standard Gaussian probability density function from $-\epsilon/(\sqrt{2}sd)$ to $\epsilon/(\sqrt{2}sd)$.

We present now the operational count for one ensemble member (or particle) involved in the forecasting, for both global and local analogs. In each case, we distinguish the computational cost of the creation of the K-d trees and the search of K nearest neighbors.

- Global analogs:
  - Creation of the K-d tree: $O(DN_{global}\log(N_{global}))$
  - Search for K global analogs: $O(KD\log(N_{global}))$
- Local analogs
  - Creation of D K-d trees (for every dimension in D): $O(D(2\nu+1)N_{local}\log(N_{local}))$
  - Search for K local analogs of component-wise neighborhood ν: $O(DK(2\nu+1)\log(N_{local}))$

Note that using local analogs requires constructing a Kd-tree for every dimension in D. Construction of the Kd-trees can be done offline (1 "big" Kd-tree for the global strategy and D "small" Kd-trees for the local strategy), then the cost of these construction can be amortized over the high number of queries that needs to be answered during analog data assimilation. However, in terms of memory storage, storing a global Kd-tree could be prohibitive, contrarily to small local Kd-trees that can be created, used, then freed for the creation of the next Kd-tree of the next dimension (if there is no sufficient memory to stock D small local Kd-trees). Keep in mind that we need to have $(2\nu+1) << D$ for local analogs to be of relevance.

Let us take an example using the Lorenz 96 model: $D = 40$, $\nu = 2$. Looking for $K = 50$ analogs, with an $\alpha = 0.15$ we would need $N_{global} \approx 10^{35}$ which is very prohibitive, however we would only need $N_{local} \approx 2.10^6$ samples using local analogs.

Appendix B

EddyNet: A Deep Neural Network For Pixel-Wise Classification of Eddies from SSH maps

Artificial Intelligence is the New Electricity.

Andrew Ng

This is an ongoing work, preliminary results are shown for illustrative purposes.

B.1 Introduction

Going "deeper" with artificial neural networks (ANNs) by using more than the original three layers (input, hidden, output) started the so-called deep learning era. The developments and discoveries which are still ongoing are producing impressive results and reaching state-of-the-art performances in various fields [60]. In particular, Convolutional Neural Networks (CNN) sparked-off the deep learning revolution in the image processing community and are now ubiquitous in computer vision applications. This has led numerous researchers from the remote sensing community to investigate the use of this powerful tool for tasks like object recognition, scene classification, etc... (see [162, 164] and references therein).

By standing on the shoulders of recent achievements in deep learning for image segmentation we present "EddyNet", a deep neural network for automated eddy detection and classification from Sea Surface Height (SSH) maps provided by the Copernicus Marine and Environment Monitoring Service (hereinafter denoted by AVISO-SSH). EddyNet is inspired by ideas from widely used image segmentation architectures, in particular U-shaped architectures such as U-Net [127]. We investigate the use of Scaled Exponential Linear Units (SELU) [84] instead of the classical ReLU + Batch Normalization (R+BN) and show that we greatly speed up the training process while reaching comparable results. We adopt a loss function based on the Dice coefficient (also known as the F1 measure) and illustrate that we reach better scores for the two most relevant classes (cyclonic and anticyclonic) than with using the categorical cross-entropy loss. We also supplement dropout layers to our architecture that prevents EddyNet from overfitting.

Our work joins the emerging cross-fertilization between the remote sensing and machine learning communities that is leading to significant contributions in addressing the segmentation of remote sensing images [9, 105, 154]. To the best of our knowledge, the present work is the first to propose a deep learning based architecture for pixel-wise classification of eddies, dealing with the challenges of this particular type of data.

This paper is organized as follows: Section II presents the eddy detection and classification problem and related work. Section III describes the data preparation process. Section IV presents the architecture of EddyNet and details the training process. Section V reports the different experiments considered in this work and discusses the results. Our conclusion and future work directions are finally stated in Section VI.

B.2 Problem statement and related work

Ocean mesoscale eddies can be defined as rotating water masses, they are omnipresent in the ocean and carry critical information about large-scale ocean circulation [28, 70]. Eddies transport different relevant physical quantities such as carbon, heat, phytoplankton, salt, etc. This movement helps in regulating the weather and mixing the ocean [112]. Detecting and studying eddies helps also considering their effects in ocean climate models [91]. With the development of altimeter missions and since the availability of two or more altimeters at the same time, merged products of Sea Surface Height (SSH) reached a sufficient resolution to allow the detection of mesoscale eddies [53, 120]. SSH maps allow us distinguish two classes of eddies: i) anticyclonic eddies that are recognized by their positive SLA (Sea Level Anomaly which is SSH anomaly with regard to a given mean) and ii) cyclonic eddies that are characterized by their negative SLA.

In recent years, several studies were conducted with the aim of detecting and classifying eddies in an automated fashion [54]. Two major families of methods prevail in the literature, namely, physical parameter-based methods and geometrical contour-based methods. The most popular representative of physical parameter-based methods is the Okubo-Weiss parameter method [116,155]. The Okubo-Weiss parameter method is however criticized for its expert-based and region-specific parameters and also for its sensitivity to noisy SSH maps [29]. Other methods were since then developed using other techniques such as wavelet decomposition [147], winding angle [130], etc. Geometric-based methods rely on considering the eddies as elliptic shapes and use closed contour techniques, the most popular method remains Chelton et al. method [28] (hereinafter called CSS11). Methods that combines ideas from both worlds are called hybrid methods (e.g. [79, 159]). Machine learning methods were also used in the past to propose a solution to the problem [23, 64], recently they are again getting an increasing attention [7, 76].

We propose in this work to benefit from the advances in deep learning to address ocean eddy detection and classification. Our proposed deep learning based method requires a training database consisting of SSH maps and their corresponding eddy detection and classification results. In this work, we train our deep learning methods from the results of the *py-eddy-tracker* SSH-based approach (hereinafter PET14) [109], the algorithm developed by Mason et al. is closely related to CSS11 but has some significant differences such as not allowing multiple local extremum in an eddy. An example of a PET14 result is given in Figure B.1 which shows eddies

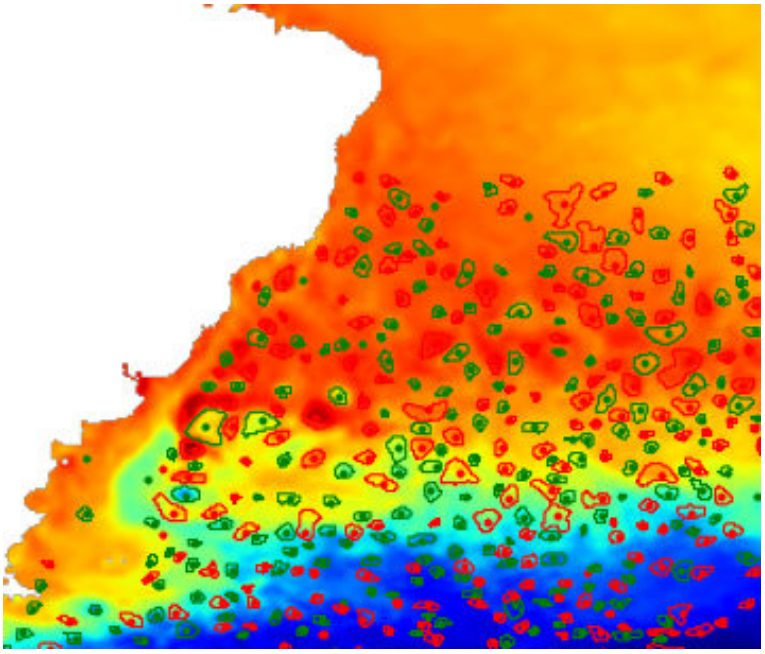


Figure B.1 – A snapshot of a SSH map from the Southern Atlantic Ocean with the detected eddies by PET14 algorithm: anticyclonic eddies (red), cyclonic eddies (green)

identified in the southwest Atlantic (see [108]). The outputs of the eddy tracker algorithm provide the center coordinates of each classified eddy along with its speed and effective contours. Since we aim for a pixelwise classification, i.e., each pixel is classified, we transform the outputs into segmentation maps such as the example shown in Figure B.2. We consider here the speed contour which corresponds to the closed contour that has the highest mean geostrophic rotational current. The speed contour can be seen as the most energetic part of the eddy and is usually smaller than the effective radius. The next section describes further the data preparation process that yields the training database of pixelwise classification maps.

B.3 Data preparation

As stated in the previous section, we consider PET14 outputs as a training database for our deep-neural-network based algorithms. We use 15 years (1998-2012) of daily detected and classified eddies. The corresponding SSH maps (AVISO-SSH) are provided by the Copernicus Marine Environment Monitoring Service (CMEMS). The resolution of the SSH maps is 0.25°.

Due to memory constraints, the input image of our architectures is 128×128 pixels. The first 14 years are used as a training dataset and the last year (2012) is left aside for testing our architecture. We consider the Southern Atlantic Ocean region depicted in Figure B.1 and cut

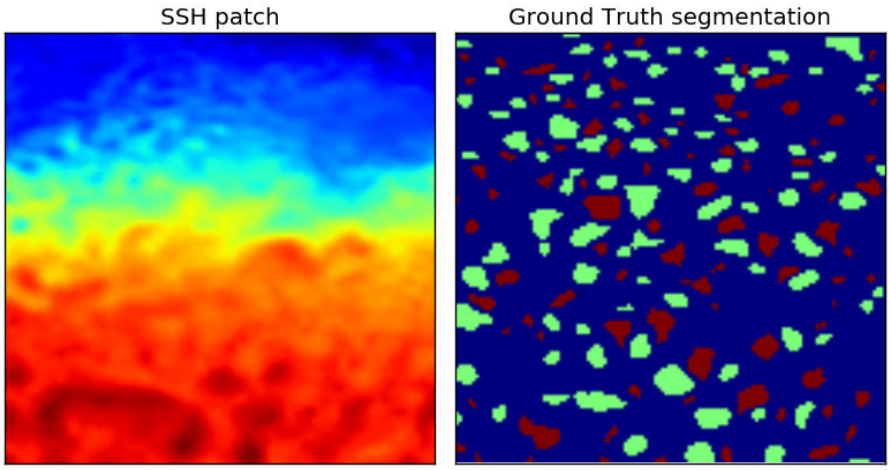


Figure B.2 – Example of a SSH-Segmentation training couple, anticyclonic (green), cyclonic (brown), non eddy (blue)

the top region where no eddies were detected. Then we randomly sample one 128×128 patch from each SSH map, which leaves us with 5100 training samples. A significant property of this type of data is that its dynamics are slow, a single eddy can live for several days or even more than a year. In addition to the fact that a 128×128 patch can comprise several examples of cyclonic and anticyclonic eddies, we believe that data augmentation (adding rotated versions of the patches to the training database for example) is not needed; we observed experiments (not shown here) that even resulted in performance degradation. The next step consists of extracting the SSH 128×128 patches from AVISO-SSH. For land pixels or regions with no data we replaced the standard fill value by a zero; this helps to avoid outliers and does not affect detection since eddies are located in regions with non zero SSH. The final and essential step is the creation of the segmentation masks of the training patches. This is done by creating polygon shapes using the speed contour coordinates mapped onto the nearest lattices in the AVISO-SSH 0.25° grid. Pixels inside each polygon are then labeled with the class of the polygon representing the eddy {'0': Non eddy/land/no data, '1': anticyclonic eddy, '2': cyclonic eddy}. Figure B.2 shows an example of a couple {SSH map, segmentation map} from the training dataset.

B.4 Our proposed method

B.4.1 EddyNet architecture

The EddyNet architecture is based on the U-net architecture [127]. It starts with an encoding (downsampling) path with 3 stages, where each stage consists of two 3×3 convolutional layers followed by either a Scaled Exponential Linear Unit (SELU) activation function [84] (referred to as EddyNet_S) or by the classical ReLU activation + Batch Normalization (referred to as EddyNet), then a 2×2 max pooling layer that halves the resolution of the input. The decoding (upsampling) path uses transposed convolutions (also called deconvolutions) [161] to return to the original resolution. Like U-net, Eddynet benefits from skip connections from the contracting path to the expanding path to account for information originating from early stages. Preliminary experiments with the original architecture of U-Net showed a severe overfitting given the low number of training samples compared to the capacity of the architecture. Numerous attempts and hyperparameter tuning led us to finally settle on a 3-stage all-32-filter architecture as shown in Figure B.3. EddyNet has the benefit of having a small number of parameters compared to widely used architecture, thus resulting in low memory consumption. Our neural network can still overfit the data which shows that it can capture the nonlinear inverse problem of eddy detection and classification. Hence, we add dropout layers before each max pooling layer and before each transposed convolutional layer; we chose these positions since they are the ones involved in the concatenations where the highest number of filters (64) is present. Dropout layers helped to regularize the network and boosted the validation loss performance. Regarding EddyNet_S, we mention three essential considerations: i) The weight initialization is different than with EddyNet, we detail this aspect in the experiment section. ii) The theory behind the SELU activation function stands on the self-normalizing property which aims to keep the inputs close to a zero mean and unit variance through the network layers. Classical dropout that randomly sets units to zero could harm this property; [84] propose therefore a new dropout technique called AlphaDropout that addresses this problem by randomly setting activations on the negative saturation value. iii) SELU theory is originally derived for Feed Forward Networks, applying them to CNNs needs careful setting. In preliminary experiments, using our U-net like architecture with SELU activations resulted in a very noisy loss that even explodes sometimes. We think this could be caused by the skip connections that can violate the self-normalizing

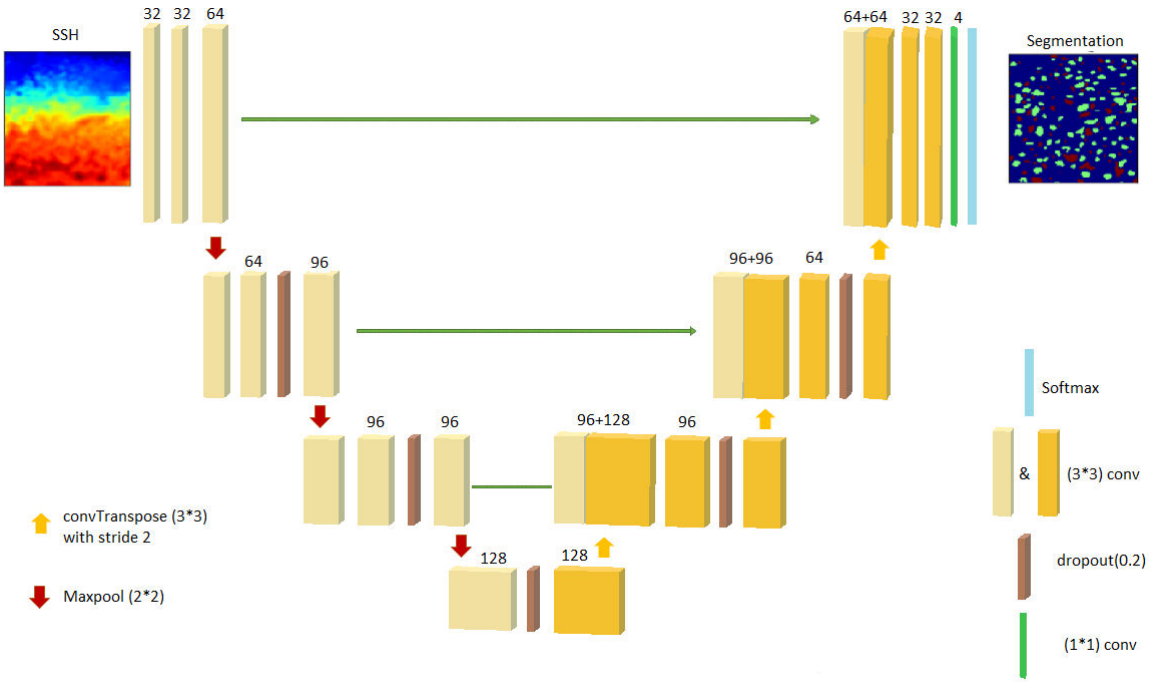


Figure B.3 – EddyNet architecture

property desired by the SELU, and hence decided to keep Batch Normalization in EddyNet_S after each of the maxpooling, transposed convolution and concatenation layers.

B.4.2 Loss metric

While multiclass classification problems in deep learning are generally trained using the categorical cross-entropy cost function, segmentation problems favor the use of overlap based metrics. The dice coefficient is a popular and largely used cost function in segmentation problems. Considering the predicted region P and the groundtruth region G, and by denoting $|P|$ and $|G|$ the sum of elements in each area, the dice coefficient is twice the ratio of the intersection over the sum of areas:

$$\text{DiceCoef}(P, G) = \frac{2|P \cap G|}{|P| + |G|}. \tag{B.1}$$

A perfect segmentation result is given by a dice coefficient of 1, while a dice coefficient of 0 refers to a completely mistaken segmentation. Seeing it from a F1-measure perspective, the dice coefficient is the harmonic mean of the precision and recall metrics.

The implementation uses one-hot encoding vectors, an essential detail is that the loss function of EddyNet uses a soft and differentiable version of the dice coefficient which considers the output

of the softmax layer as it is without binarization:

$$\text{softDiceCoef}(P, G) = \frac{2 \sum_i p_i * g_i}{\sum_i p_i + \sum_i g_i}, \tag{B.2}$$

where the p_i are the probabilities given by the softmax layer $0 \leq p_i \leq 1$, and the g_i are either 1 for the correct class and 0 either. We found later that a recent study used another version of a soft dice loss [114]; a comparison of both versions is out of the scope of this work.

Since we are in the context of a multiclass classification problem, we try to maximize the performance of our network using the mean of three one-vs-all soft dice coefficients of each class. The loss function that our neural network aims to minimize is then simply:

$$\text{Dice Loss} = 1 - \text{softMeanDiceCoef} \tag{B.3}$$

Table B.1 – Metrics calculated from the results of 50 random sets of 360 SSH patches from the test dataset, we report the mean value and put the standard variation between parenthesis.

				Anticyclonic	Cyclonic	Non Eddy		
	#Param	Epoch time	Train loss	Dice Coef			Mean Dice Coef	Global Accuracy
EddyNet	177,571	∼12 min	Dice Loss	**0.708** (0.002)	**0.677** (0.001)	0.929 (0.001)	**0.772** (0.001)	88.60% (0.10%)
			CCE	0.695 (0.003)	0.651 (0.001)	0.940 (0.001)	0.762 (0.001)	89.92% (0.07%)
EddyNet_S		∼7 min	Dice Loss	0.694 (0.003)	0.665 (0.001)	0.933 (0.001)	0.764 (0.001)	88.98% (0.09%)
			CCE	0.682 (0.002)	0.653 (0.002)	0.939 (0.001)	0.758 (0.001)	89.83% (0.08%)

B.5 Experiments

B.5.1 Assessment of the performance

Keras framework with a Tensorflow backend is considered in this work. EddyNet is trained on a Nvidia K80 GPU card using ADAM optimizer and mini-batches of 16 maps. The weights were initialized using truncated Gaussian distributed weights of zero mean and {2/number of input units} variance [68] for EddyNet, while we use weights drawn from a truncated Gaussian distribution of zero mean and {1/number of input units} variance for EddyNet_S. The training dataset is split into 4080 images for training and 1020 for validation. We also use an early-stopping strategy to stop the learning process when the validation dataset loss stops improving in five consecutive epochs. EddyNet weights are then the ones resulting in the lowest validation loss value.

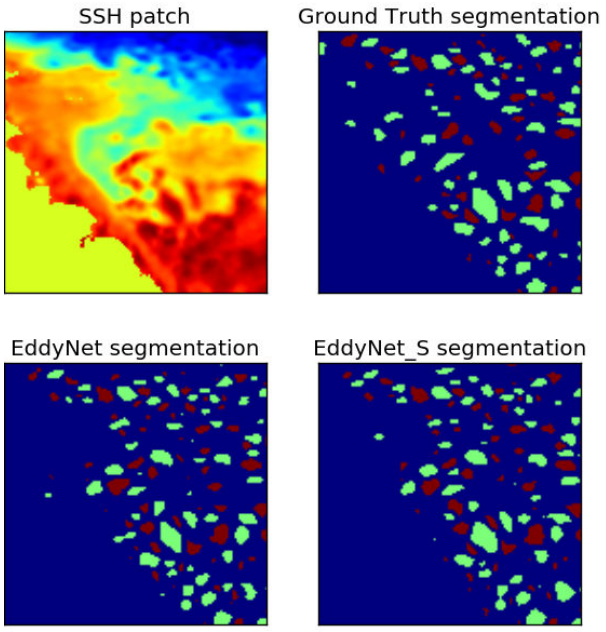
EddyNet and EddyNet_S are then compared regarding the use of the classical ReLU+BN and the use of SELU. We also compare the use of overlap based metric represented by the Dice Loss (Equation B.3), with the classical Categorical Cross-Entropy (CCE). Table B.1 compares the four combination in terms of global accuracy and mean dice coefficient (original not soft) averaged on 50 random sets of 360 SSH 120 × 120 maps from 2012. Training EddyNet_S takes nearly half the time needed for training EddyNet. Comparison regarding the training loss function shows that training with the dice loss results in a higher dice coefficient for our two classes of interest (cyclonic and anticyclonic) in both EddyNet and EddyNet_S; dice loss yields a better overall mean dice coefficient than training with CCE loss. Regarding the effect of the activation function, we obtained better metrics with EddyNet at the cost of a longer training procedure. Visually Eddynet and EddyNet_S give close outputs as can be seen in Figure B.4.

B.5.2 Ghost eddies

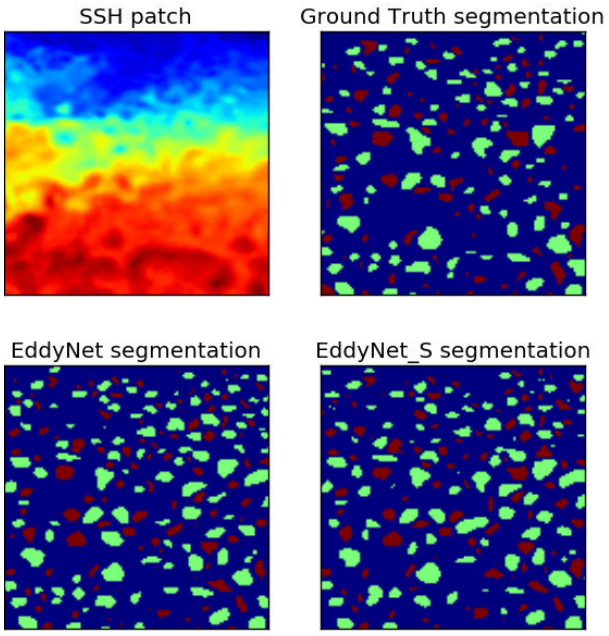
The presence of ghost eddies is a frequent problem encountered in eddy detection and tracking algorithms [53]. Ghost eddies are eddies that are found by the detection algorithm then disappear between consecutive maps before reappearing again. To point out the position of the missed ghost eddies, PET14 uses linear temporal interpolation between centers of detected eddies and stores the positions of the centers of ghost eddies. Using EddyNet we check if the pixels of ghost eddy centers correspond to actual eddy detections. We found that EddyNet assigns the centers of ghost eddies to the correct eddy classes 55% of the time for anticyclonic eddies, and 45% for cyclonic eddies. EddyNet could be a relevant method to detect ghost eddies that are missed out by conventional methods. Figure B.5 illustrates two examples of ghost eddy detection.

B.6 Conclusion

This work investigates the use of recent developments in deep learning based image segmentation for an ocean remote sensing problem, namely, eddy detection and classification from Sea Surface Height (SSH) maps. We propose EddyNet, a deep neural network architecture inspired from architectures and ideas widely adopted in the computer vision community. We transfer successfully the knowledge gained to the problem of eddy classification by dealing with various challenges. Future work involves investigating the use of temporal volumes of SSH and deriving a 3D version inspired by the works of [114]. Adding other surface information such as Sea Surface Temperature might also help improving the detection. Another extension would be the



(a)



(b)

Figure B.4 – Examples of the eddy segmentation results using Eddynet and EddyNet_S: anti-cyclonic eddies (green), cyclonic (brown), non eddy (blue)

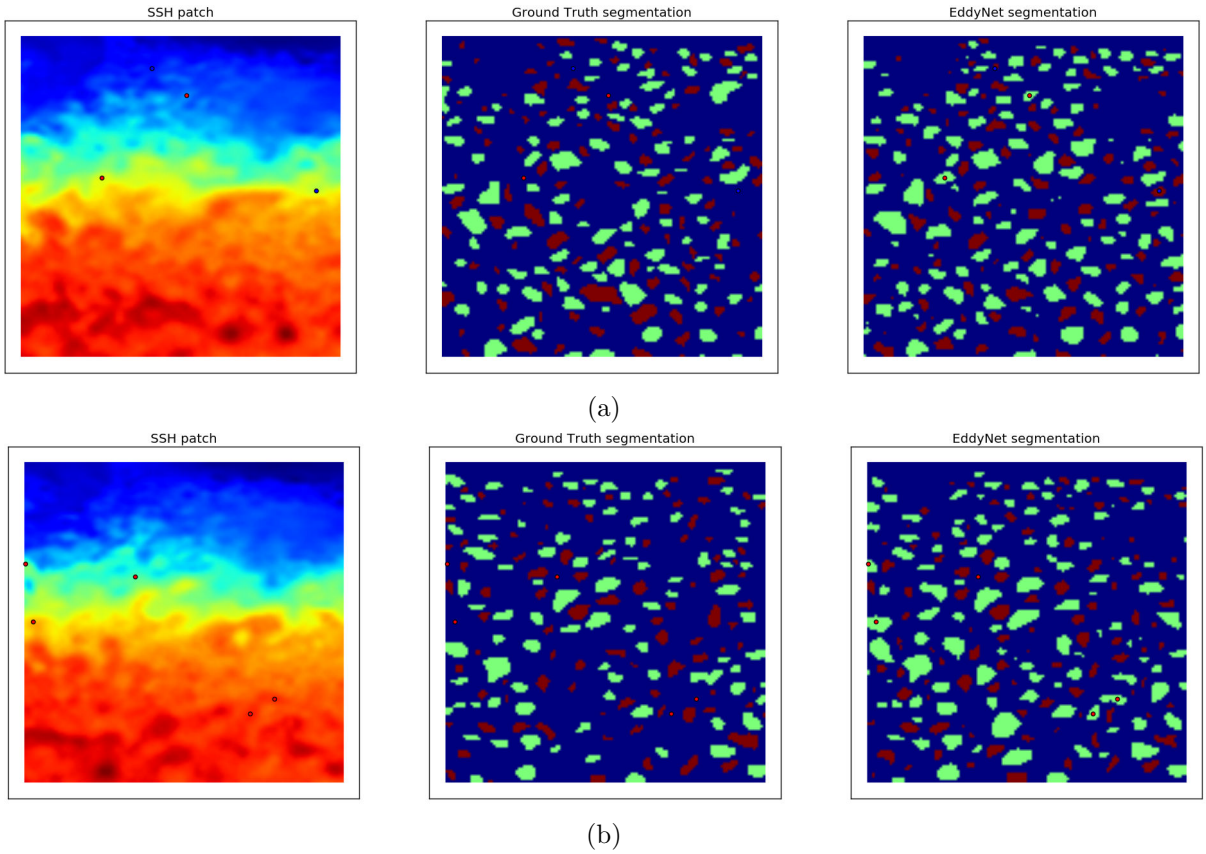


Figure B.5 – Detection of ghost eddies: [left] SSH map with ghost eddies centers: anticyclonic (red dots), cyclonic (blue dots). [center] PET14 segmentation. [right] EddyNet segmentation: anticyclonic (green), cyclonic (brown), non eddy (blue)

application of EddyNet over the globe, and assessing its general capacity over other regions. Post-processing by constraining the eddies to verify additional criteria and tracking the eddies was omitted in this work and could also be developed in future work.

Beyond the illustrative aspect of this contribution, we offer to the oceanic remote sensing community an easy and powerful tool that can save handcrafting model efforts. Any user can employ his own eddy segmentation "ground truth" and train the model from scratch if he/she has the necessary memory and computing resources, or simply use EddyNet provided weights as an initialization then perform fine-tuning using his/her dataset. One can also think of averaging results from classical contour-based methods and EddyNet. In the spirit of reproducibility, Python code is available at `https://github.com/redouanelg/eddynet`, and we also share the training and testing data used for this work to encourage competing methods and, especially, other deep learning architectures.

Résumé

Reconstruire des champs géophysiques à partir d'observations bruitées et partielles est un problème classique bien étudié dans la littérature. L'assimilation de données est une méthode populaire pour aborder ce problème, et se fait par l'utilisation de techniques classiques, comme le filtrage de Kalman d'ensemble ou des filtres particulaires qui procèdent à une évaluation online du modèle physique afin de fournir une prévision de l'état. La performance de l'assimilation de données dépend alors fortement de du modèle physique. En revanche, la quantité de données d'observation et de simulation a augmenté rapidement au cours des dernières années. Cette thèse traite l'assimilation de données d'une manière data-driven et ce, sans avoir accès aux équations explicites du modèle. Nous avons développé et évalué l'assimilation des données par analogues (AnDA), qui combine la méthode des analogues et des méthodes de filtrage stochastiques (filtres Kalman, filtres à particules, chaînes de Markov cachées). Des applications aux modèles chaotiques simplifiés et à des études de cas de télédétection réelle (température de surface de la mer, anomalies du niveau de la mer), nous démontrons la pertinence d'AnDA pour l'interpolation de données manquantes des systèmes dynamiques non linéaires et à haute dimension à partir d'observations irrégulières et bruyantes.

Motivé par l'essor du machine learning récemment, la dernière partie de cette thèse est consacrée à l'élaboration de modèles deep learning pour la détection et de tourbillons océaniques à partir de données de sources multiples et/ou multitemporelles (ex: SST-SSH), l'objectif général étant de surpasser les approches dites expertes.

Mots clés : Assimilation de données, prédiction par analogues, Assimilation de données par analogues, Télédétection de l'océan, Température de la surface de l'océan, Elévation du niveau de la mer, Apprentissage profond

Abstract

Reconstructing geophysical fields from noisy and partial remote sensing observations is a classical problem well studied in the literature. Data assimilation is one class of popular methods to address this issue, and is done through the use of classical stochastic filtering techniques, such as ensemble Kalman or particle filters and smoothers. They proceed by an online evaluation of the physical model in order to provide a forecast for the state. Therefore, the performance of data assimilation heavily relies on the definition of the physical model. In contrast, the amount of observation and simulation data has grown very quickly in the last decades. This thesis focuses on performing data assimilation in a data-driven way and this without having access to explicit model equations. The main contribution of this thesis lies in developing and evaluating the Analog Data Assimilation (AnDA), which combines analog methods (nearest neighbors search) and stochastic filtering methods (Kalman filters, particle filters, Hidden Markov Models). Through applications to both simplified chaotic models and real ocean remote sensing case-studies (sea surface temperature, along-track sea level anomalies), we demonstrate the relevance of AnDA for missing data interpolation of nonlinear and high-dimensional dynamical systems from irregularly-sampled and noisy observations.

Driven by the rise of machine learning in the recent years, the last part of this thesis is dedicated to the development of deep learning models for the detection and tracking of ocean eddies from multi-source and/or multi-temporal data (e.g., SST-SSH), the general objective being to outperform expert-based approaches.

Keywords: Data Assimilation, Analog forecasting, Analog Data Assimilation, Sea Surface Temperature, Ocean Remote Sensing, Sea Surface Temperature, Sea Level Anomaly, Deep Learning

IMT Atlantique
Bretagne-Pays de la Loire
École Mines-Télécom

N° d'ordre : 2017IMTA0050

IMT Atlantique Bretagne-Pays de la Loire - www.imt-atlantique.fr

Campus de Brest
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 03
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

Campus de Nantes
4, rue Alfred Kastler - La Chantrerie
CS 20722
44307 Nantes Cedex 3
T +33 (0)2 51 85 81 00
F +33 (0)2 51 85 81 99

Campus de Rennes
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
T +33 (0)2 99 12 70 00
F +33 (0)2 99 12 70 08