Master's Internship Project

Report

Academic year 2019/2020

Implementation of the Ensemble Kalman Filter on an Earth System Climate Model

Submitted by

Romain BOUTELET

Student ID : 11710545

Département de Mathématiques

Université Claude Bernard Lyon 1

Under the guidance of

Dr. Juan RESTREPO

Professor

Department of Mathematics

Oregon State University

Dr. Andreas SCHMITTNER

Professor

College of Earth, Ocean, and Atmospheric Sciences

Oregon State University

Supervised by

Dr. Clément MARTEAU

Professor

Département de Mathématiques

Université Claude Bernard Lyon 1

Acknowledgements

I would like to express my deep and sincere gratitude to my research supervisors, Dr. Juan RESTREPO and Dr. Andreas Scmittner for giving me the opportunity to work on this project and for their guidance.

Besides my research supervisors, I would like to deeply thank Dr. Clément MARTEAU, my academic supervisor who has made this internship possible and for giving me support while the situation was very complicated and unpredictable.
I would also thank the administration of Université Claude Bernard Lyon 1 for accommodating to the rapidly changing situation just before the beginning of my internship.

Finally, I would like to give my deep gratitude to my family and my loved ones for their support and love during the duration of my internship.

Contents

Chapter 1

Introduction

The ocean's carbon cycle is an important part of Earth's climate system because it impacts the concentration of CO_2 in the atmosphere. This mechanism is not well understood because it is driven by complex physical, chemical and biological processes. Moreover, it is important to look at past climate periods to better understand how these processes work and how this carbon cycle mechanism can change. The aim of this project is therefore to use a data assimilation method, the Ensemble Kalman Filter, and to implement it on an Earth System Climate Model, the UVic ESCM, to constrain the ocean circulation during the last deglaciation and give a better understanding of the past changes and improve possible future projections.

The first part of this report will provide some background on the Earth's climate system and Paleoceanography. That will provide a foundation to understand the later results and to efficiently work with the UVic Earth System Climate Model (ESCM).
We will then present the theory about Bayesian filtering and more particularly the Kalman Filters. This will enable us to then introduce the Ensemble Kalman Filter (EnKF) and apply it on a chaotic system. Finally, we will detail how the implementation of the EnKF on the UVic ESCM was made, some of its results. We will also discuss possible improvements of the implementation as well as future work that can be done.

Chapter 2

The Atlantic Meridional Overturning Circulation and Paleoceanography

In this chapter, we will first introduce the paleoceanographic and paleoclimatic context around our the dast deglaciation, which occurred between the end of the Last Glacial Maximum approximately 19,000 years ago (19 ka) and the early Holocene (11 ka).
We will next develop more in detail the Atlantic Meridional Overturning Circulation (AMOC), its effects on the climate, and its sensitivity to external forcings and to periodic changes.
Finally, we will explain how the past oceanographic and climate conditions are reconstructed with the use of Earth System Climate Model (ESCM) such as the UVic 2.9 model and observations from proxy records.

2.1 The Atlantic Meridional Overturning Circulation

The ocean plays a crucial role in the global climate : absorbing, storing and giving back energy from the Sun thanks to water's high heat capacity and fluidity. This ability to absorb solar radiation is also due to the darkness of the ocean, leading to a low albedo (the reflectivity of a body) on contrary to ice sheets, which have a really high albedo and therefore reflecting most of the solar radiation [17]. This mechanism is a very important aspect of global climate but the ocean also plays a role in the global climate in different ways.
In addition to having a big impact on global climate thanks to its absorption of solar radiation and storage of heat, the ocean's circulation also plays a crucial role in the internal transport of heat (for example the Gulf Stream, a wind-driven ocean current thats carries warm water from the tropics to the North Atlantic), shaping regional climates.

2.1.1 General description

The thermohaline circulation, also known as the Meridional Overturning Circulation is a large-scale ocean circulation driven by temperature and salinity gradients.

This circulation involves a northwards surface ocean current in the Atlantic ocean, which sinks in the deep ocean at high latitudes, forming the North Atlantic Deep Water (NADW). This deep water then travels southwards and emerges in the Southern Ocean, the Indian Ocean or in the North Pacific.

The Atlantic Meridional Overturning Circulation (AMOC) is a part of the thermohaline circulation. The AMOC is localized in the Atlantic ocean, and involves a large system of ocean currents that carry warm surface water from the tropics northwards into the North Atlantic, evaporating en route and increasing the salinity of this water. When cooling down as it goes towards the pole, its density surpasses the surrounding surface ocean and ocean's upper layers. That gradient of density leads this water to sink in the Labrador Sea and in the Nordic Seas (during our current period), forming the NADW.

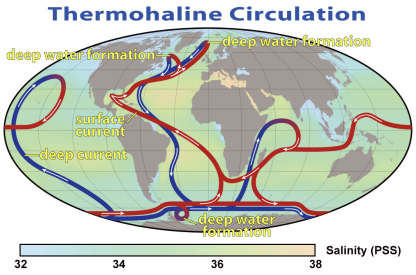

Figure 2.1: Pattern of the thermohaline circulation. Source : NASA

The AMOC is very important for the global climate, particularly for the North Atlantic zone. It can directly impact the climate, and it has been suggested that changes in the AMOC behavior could have led to abrupt climate changes (Greenland recorded warming of 5-10° C within a few decades [17]).

2.1.2 The AMOC and its impact on greenhouse-gas variations

Other consequences of variations of the AMOC should also be acknowledged. Indeed, the oceans store a very large part of the Earth's total carbon through different mechanisms (solubility pump, carbonate pump and biological pump). The ocean has absorbed over 90% of the CO_2-induced global warming over the past several decades, and the AMOC contributes to this heat uptake [20]. It also needs to be noticed that this storage involves ocean circulations that can keep the water in the deep ocean during several hundred of years to a few thousand of years before re-emerging to the surface, leading to a sort of smoothing of the discharge of greenhouse-gas from anthropogenic sources. Although that means that the carbon in the atmosphere can be stored for long times, it also means that the carbon from the deep ocean re-emerges after being kept in the deep ocean.

It has been shown that changes in the AMOC circulation strength can lead to important variations of CO_2 and N_2O, two greenhouses-gases, in the atmosphere on time scales of several hundreds of years. Studying the response of different AMOC strengths on the atmospheric concentration of greenhouse-gas during the last glacial period, evidence has been brought supporting that the collapse of the AMOC resulted in a rapid decrease in the production of N_2O, and an increase in the CO_2 concentration during this same period [21]. Another project supports the idea that a weakening or complete shutdown of the AMOC during the last deglaciation could have decreased the biological pump [22].

2.1.3 Sensitivity of the AMOC

As explained before, the AMOC is driven by water density gradients. Highest density in the surface ocean are reached where water is the coldest, that is at extreme latitudes, even though the water is less salty than in warmer regions such as the tropics and subtropics (due to evaporation). Salinity is however a very important factor in the thermohaline circulation and is more specifically involved in the nonlinearity of the AMOC system [17].

The AMOC is indeed fueled by a positive feedback process : higher salinity in deep-water formation zones boosts the circulation, which in return carries more high-salinity water towards the region of deep-water formation. Moreover, this region being a zone of positive net precipitation (precipitation - evaporation), the water in this region would not be dense enough to sink in the ocean without the inflow of high-salinity water.

These properties make the AMOC a system very sensitive to the amount of freshwater flowing into the North Atlantic, leading to two possible equilibrium cases for the AMOC : with and without the NADW formation [16].
On the opposite diagram (Fig 2.2), the bistable regime of the AMOC implies a breaking point (the Stommel bifurcation) at which the circulation stops completely (horizontal line in the diagram). On the other hand, when the AMOC is shut down, a negative freshwater forcing in the North Atlantic would lead to a restart of the circulation (with respect to the equilibrium curve).

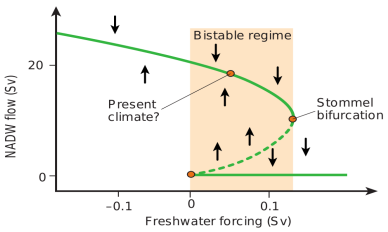

Figure 2.2: Hysteresis diagram of the AMOC in response to freshwater forcing in the North Atlantic. Source : Rahmstorf 2002 [17]

2.1.4 The three possible AMOC patterns

Studies of the properties of the ocean from ancient periods with the use of proxies (extracted from ice cores, sediment cores, ...) has opened the path to a better understanding of the ocean circulation under different conditions than the recent conditions. It has then been inferred that the AMOC exhibits three differents conceptual modes from observations of the last glacial cycle [11].

- Today's AMOC, characterized by a strong and deep circulation, with the convection zones in the Greenland-Iceland-Norwegian Seas (plot **a.** of Fig 2.3) and can also be designated as the warm mode of the AMOC;
- A shallower AMOC, or cold mode, (less than 2km deep) with convection sites displaced to the subpolar North Atlantic is supported with strong evidence [11] to have occurred during the Last Glacial Maximum (plot **b.**);
- A shutdown, or of mode, of the AMOC where there is no deep-water formation in the North Atlantic (plot **c.**) is linked to events

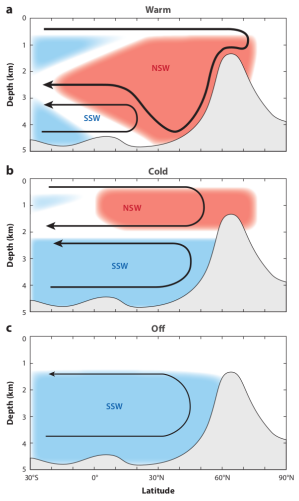

Figure 2.3: Representations of the AMOC on different states (warm, cold and off modes) in the North Atlantic. Source : Lynch 2017 [11]

of abrupt climate changes although the proxies are too noisy to
be certain that such events would be due or related to a shutdown of the AMOC.

These changes in the AMOC pattern and strength have been the subject of many studies in Paleoceacnography to understand the links between changes in the AMOC states and climate changes that happened (more particularly during the Last Glacial Period, since the proxies from this period is less noisy than for more ancient periods).

2.2 An introduction to Paleoclimatology and Paleoceanography

Paleoclimatology and Paleoceanography are the study of, respectively, the climate and the ocean during past periods, more precisely during periods where no direct measurements are available. Studying the climatic and oceanographic conditions over longer periods than the period we have been able to measure in the recent history (first reliable records of climate were made around the 1880s) allows a new understanding of the mechanisms of the ocean and of the climate over very long periods.

2.2.1 From the Last Glacial Maximum to the early Holocene

The period that we want to study more specifically lies between the end of the Last Glacial Period (115,000 - 20,000 years ago) and in the following deglaciation until the early Holocene (11,700 years ago to present).
The transition from the Eemian **interglacial** (an interglacial period is a geological period lasting thousands of years and defined by a warmer climate and that separates successive glacial periods within an ice age) to the beginning of the Last Glacial Period occurred very rapidly [17]. This shift opened the way to a glacial period which was itself delimited in **stadials** (cold phases) and **interstadials** (warm phases). Stadials and interstadials differ from glacials and interglacials : stadials are cold periods in Greenland and the North Atlantic region associated with reduced AMOC, whereas glacials are global cold periods. During the Last Glacial Period (LGP), these stadials and interstadials were characterized by pronounced climate change during certain periods (the **Dansgaard-Oeschger events**). Other noticeable events of the LGP include the **Heinrich events**, and the **Younger Dryas**.

Dansgaard-Oeschger events

The Dansgaard-Oeschger events are abrupt climate changes (in a matter of decades) seen in the Greenland ice records as 5-10°C warmings, and which happened almost periodically during the LGP. During these interstadials, the North Atlantic is warming while the Southern Hemisphere is cooling in a "see-saw" effect.
These events were closely related to AMOC changes, but the causality between these events and AMOC changes is complicated to determine. Two hypotheses for the D-O events exist to explain the variations in the AMOC state (cold mode to warm mode). Whereas one of the reasons involve the bistability of the AMOC, the other one relies on an internal oscillation of the AMOC which would weaken the AMOC until an external trigger would start the D-O events. The last reason, which would imply the existence of a threshold mechanism is more in compliance with the quasi-periodicity of the D-O events, leading to the explanation of a stochastic resonance, where the climate variability is combined to the

Milankovitch cycles, producing a weak external cycle of freshwater forcing One hypothesis suggest that the switches could be forced by freshwater input whereas the other hypothesis relies on an internal cycle of the AMOC in addition to some other external forcing such as the Milankovitch cycles [17],[15]. The Milankovitch cycles describe variations in the position of the Earth that would result in variations in the solar radiation, implying an external forcing influencing the heat budget of the Earth.

Heinrich events

Heinrich events are another type of major climatic event that happened during the LGP. These events can be clearly recognized in sediment cores, as it is related to massive discharges of icebergs into the North Atlantic, leading to drastic changes in the sediments properties in the ocean basin.
Proxies evidence show that NADW formation ceased, or was weakened, during these events. That would be in agreement with the idea of linking Heinrich events to a switch from cold mode to an off mode [11]. D-O and Heinrich events are also not unrelated [17] :

- Each Heinrich event is followed by a particularly warm D-O event, even though not all of the D-O events are preceded by a Heinrich event. Moreover, the D-O events get cooler and cooler until a Heinrich event happens;
- Heinrich events always occur during stadials, suggesting that these events are triggered by temperature or sea-level changes.

Younger Dryas

The end of the last ice age led to the climate that we know now. This change included a warming of the global climate starting from 20ka, according to Antarctic ice cores. However, the proxies from Greenland tell a different story, plausibly caused by AMOC changes as freshwater forcing were happening in the North Atlantic with the melt of the ice sheets in North America [17].
After the succession of the second to last Heinrich event (HS1) and the following D-O event (also called Bølling warming), it gave way to the Younger Dryas (Fig 2.4), being the last large cold event until today. The Younger Dryas disrupted the overall warming trend, and the cause of this global climate change has been initially attributed to a shutdown of the AMOC due to freshwater forcing. However, new explanations involving a combination of an AMOC change, a decrease of solar radiation, and a shift in the atmospheric circulation was most likely to be the causes of such a global cooling [19].

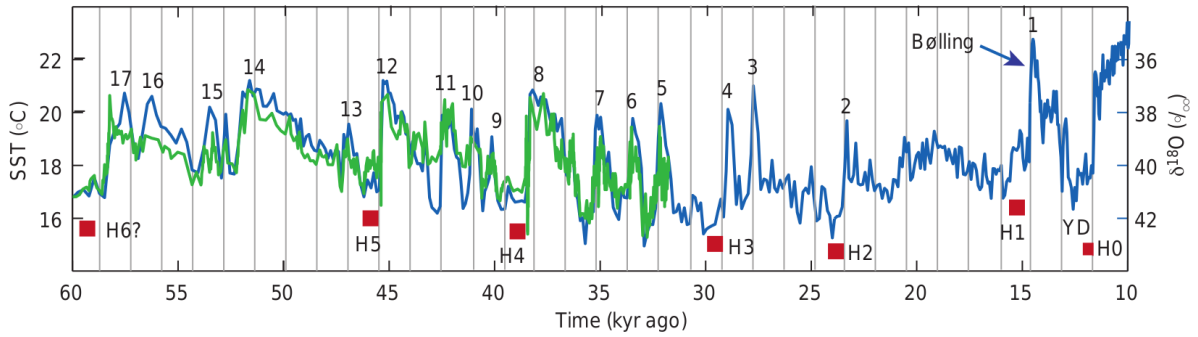

Figure 2.4: Temperature reconstructions from ocean sediments in subtropical Atlantic (green) and from Greenland ice core GISP 2 (blue). D-O events are numbered above the curve, Heinrich events are timed with red boxes. YD : Younger Dryas. Source Rahmstorf, 2002 [17]

2.2.2 Using proxies for reconstructing the past ocean and climate

Paleoceanography has been able to improve the understanding of the ocean during past periods thanks to a new source of information : proxies from sediment cores, etc...
These proxies give information about the geochemistry conditions that happened at a certain depth of these cores. Using radioactivity observations (carbon-14) from these same proxies it is possible to age the geochemistry observations. These geochemistry observations (δ13C, and other isotopes ratios and radioactive components), make it possible to trace certain geochemistry processes. More precisely, certain properties of the deep ocean such as its age, concentration in some particular elements, the biological productivity, or even large-scale flows can be traced thanks to these data.
As a result from this information, paleoceanography can reconstruct credible states of the ocean during past periods.

2.2.3 Using Earth System Climate Models for reconstructing the past ocean and climate

With the development of highly performant computers and the combination of various geophysics fields to a more global and universal science, the climatology, has come the emergence of numerical models for simulating the Earth System. The numerical models have become essential in the study of the dynamics of climate change, even though these simulations do not perfectly represent the real states of the Earth System. Indeed, some of these models include isotopes, which allows comparisons.
Reconstructing radically different past climates such as the LGM have been successful using coupled ocean-atmosphere models from prescribed orbital, CO_2 and continental ice-sheet forcing [17]. Coherent reconstruction of such past climates is now a requirement for new numerical climate models to be used for estimating the effects of anthropogenic forcing on the present and future climate.

2.3 The UVic 2.9 climate model : an Earth System Model of intermediate complexity

2.3.1 Different level of climate models

Climate models are not representing perfectly the reality, inevitable errors arise from the approximation made in the numerical methods as well as the uncertainties on the initial and boundary conditions on the model grid. However error can be limited by the use of a better resolution for model grids and in the increase of the system integration, requiring more computing power. The integration defines the level of interaction that geophysics components have between each other and the frequency of these interactions. Consequently, climate models have different levels of complexity, going from zero-dimensional models (only modeling the radiative equilibrium of the Earth with parameters, such as the external forcing, global albedo,...) to the Global Climate models (discretizing the Earth in grids and integrating the fluid motions and other internal processes as well as external forcing).
The Climate model that we are using in this project is an Earth System model of intermediate complexity (EMIC). EMICs are a compromise between the conceptual models and the general circulation models which work with a high spatial and temporal resolution.

2.3.2 The UVic ESCM model more in the details

The UVic ESCM model is a three-dimensional dynamical ocean governed by the primitive equations and coupled with a two-dimensional atmosphere, which solves heat and moisture balance equations. Interaction between the ocean and atmosphere are modelled with interactive moisture and heat fluxes and momentum fluxes from a prescribed wind field. The model grid is set with 19 vertical levels a $3.6° \times 1.8°$ horizontal resolution.
This ocean-atmosphere model is combined with a dynamical sea-ice model, a dynamical land vegetation model and also includes a carbon cycle. An ocean biogeochemistry model (the Model of Ocean Biogeochemistry and Isotopes, MOBI) is also driven by the ocean circulation model to simulate the ocean carbon cycle [27, 14].
The UVic ESCM has been used in many studies [22, 14, 21] since its description and performance assessments were first released in 2001 [27].

Chapter 3

About the Kalman Filters

3.1 Bayesian filtering

Bayesian filtering is a probabilistic approach for estimating the joint probability distribution of a collection of variables describing a dynamics system using measurements and a process model. For this approach the measurements and process model are assumed to be subject to random noise and therefore to have probability distributions that we assumed to be known.

The noise in the measurements means that they are uncertain; even if we knew the true system state the measurements would not be deterministic functions of the state, but would have a distribution of possible values. The time evolution of the state is modeled as a dynamic system which is perturbed by a certain process noise. This noise is used for modeling the uncertainties in the system dynamics. In most cases the system is not truly stochastic, but stochasticity is used for representing the model uncertainties.

3.1.1 Applications of Bayesian filtering

The Bayesian filtering can thus be used for phenomena that are time-varying and modelled as above. A (non exhaustive) list of various applications [26] :

- **GPS** : The widely used navigation system typically uses an Extended Kalman Filter or some other Bayesian filtering algorithm for computing the position and velocity of the object from satellite information as measurements and the laws of physics for modelling the dynamics of the system. In practice, the measurements are time delays of satellite signals (at least 3 satellites for a consistent model) and the optimal filter computes the position of the object.
- **GPS/INS navigation** : This navigation system has a comparable functioning than the GPS system above but adds internal sensors for refining the measurements information. This allows the system to be effective even when the satellite signals aren't available (tunnels, indoors,...)
- **Spread of infectious disease** : These problems can often be modelled with differential equations representing the number of infections/deaths/recovery. When uncertainties are added in the dynamics model and the measurements, the problem can be turned into an optimal filtering problem.

- **Stochastic optimal control** : The objective here is to control time-varying stochastic systems. Such stochastic controllers can be found in airplanes, cars or rockets. Here the control signal is constructed to minimize a performance cost and a Bayesian filter computes this optimal state acknowledging that random noise with known probability distribution affects the evolution and observation of the state variables.
- **Physical systems** : Bayesian filters can also be used for time-varying physics systems which would be thought as stochastic state space models with non-ideal sensors.

Bayesian inference for filtering

Bayesian filtering can be seen as a statistical inversion problem where then unknown variable is the real state of the system, described as a time series $\{x_0, x_1, ...\}$, and which is solved using a measurement model and a time series of observations $\{y_1, y_2, ...\}$.

The purpose of this inversion is therefore to estimate $x_{0:T} = \{x_0, x_1, ...\}$ from $y_{1:T} = \{y_1, y_2, ...\}$, i.e. computing the joint posterior distribution of all the states given all the measurements. Using the Bayes' rule, we obtain that:

$$p(x_{0:T} \mid y_{1:T}) = \frac{p(y_{1:T} \mid x_{0:T})p(x_{0:T})}{p(y_{1:T})}$$

where :

- $p(x_{0:T})$ is the **prior distribution** defined by the dynamics model;
- $p(y_{1:T} \mid x_{0:T})$ is the **likelihood model** for the measurements.
- $p(y_{1:T}) = \int p(y_{1:T} \mid x_{0:T})p(x_{0:T})dx_{0:T}$ is the **normalization constant**.

Using this approach directly is not possible as the size of the posterior distribution increases with time, and this would imply that the computational complexity also increases at each time step.

A way to get around this problem is to consider estimating only the current state x_k posterior distribution at each time step instead of the full posterior distribution. The complexity of the computational problem would become lighter, allowing this method to be feasible in practice. For adapting the method, we also need to restrict the dynamics model to stochastic Markov sequences and the probabilistic state space model would be :

$$\begin{aligned}
x_0 \sim p(x_0), &\quad \text{the initial state } x_0 \text{ following the prior distribution } p(x_0) \\
x_k \sim p(x_k \mid x_{k-1}), &\quad \text{the current state } x_k \text{ following the transition} \\
&\quad \text{probability distribution } p(x_k \mid x_{k-1}) \\
y_k \sim p(y_k \mid x_k), &\quad \text{the observation } y_k \text{ of the current state, following the likelihood} \\
&\quad \text{of the observations given the current state } p(y_k \mid x_k)
\end{aligned}$$

From this we can derive three different types of recursive Bayesian estimation :

- The **Bayesian filtering** which estimates the current state given *past* and *present* observations, i.e. the marginal distribution is $p(x_k \mid y_{1:k})$, $k = 1, ..., T$;

- The **Bayesian smoothing** which estimates a past state given *past* and *present* observations, i.e. the marginal distribution is $p(x_k \mid y_{1:T})$, $k = 1, ..., T$;
- The **Bayesian estimation** which estimates the probable future state given *past* and *present* observations, i.e. the marginal distribution $p(x_{k+n} \mid y_{1:k})$, $k = 1, ..., T, \quad n = 1, 2, ...$

We will in the following only consider the Bayesian filtering even though the Bayesian prediction will be indirectly induced in some following results (indeed the cases when the data is not sufficiently informed in time need some Bayesian prediction).

3.1.2 The Bayesian filtering equations

We will here consider the general equations for computing the Bayesian filtering solutions. These equations will be derived from Bayesian inference and Bayesian filtering methods seen previously.

Probabilistic state space models. We will place ourselves in a probabilistic state space model with the aim of estimating the state of the system.

Definition 3.1.1 (Probabilistic state space model)**.** *A **probabilistic state space model** consists of a sequence of conditional probability distributions :*

$$\begin{aligned} x_k &\sim p(x_k \mid x_{k-1}) \\ y_k &\sim p(y_k \mid x_k) \end{aligned} \tag{3.1}$$

The underlying state model is assumed to be Markovian, which means that the states of the system $(x_k)_k$ of the model form a sequence that must follow the following properties.

Property 3.1.1 (Markov property of states)**.** The sequence $(x_k)_k$ of the states of the system is Markovian.
That is, conditional on the present state of the system, its future and past states are independent :

$$\begin{aligned} p(x_k \mid x_{1:k-1}, y_{1:k-1}) &= p(x_k \mid x_{k-1}) \\ p(x_k \mid x_{k+1:T}, y_{k+1:T}) &= p(x_k \mid x_{k+1}) \end{aligned}$$

Property 3.1.2 (Conditional independence of the measurements)**.** The current measurements y_k given the current state x_k is independent of the states and measurements histories:

$$p(y_k \mid y_{1:k-1}, x_{1:k}) = p(y_k \mid x_k)$$

The posterior distribution of the states of the system can then be computed with the Bayes' rule :

$$p(x_{0:T} \mid y_{1:T}) = \frac{p(y_{1:T} \mid x_{0:T}) p(x_{0:T})}{p(y_{1:T})}$$

where $p(y_{1:T})$ is a normalization constant, which is not taken into consideration in practice; and $p(x_{0:T})$ is the prior distribution of the states of the system as assumed through assumptions and expectations.

Moreover, with the Markovian assumptions and the model given by the equations (1.1), we obtain that the full joint prior distribution of the states and the likelihood of the measurements are :

$$p(x_{0:T}) = p(x_0) \prod_{k=1}^{T} p(x_k \mid x_{k-1})$$
$$p(y_{1:T} \mid x_{1:T}) = \prod_{k=1}^{T} p(y_k \mid x_k)$$

However, as we have seen in paragraph **1.2**, the computation of the full posterior joint distribution is not feasible in practice and that problem is bypassed by solely estimating the current posterior distribution of the current state of the system.

Bayesian filter equations The purpose of Bayesian filtering is to compute the marginal posterior distribution of the state x_k at each time step k given the observations up to time step k.

Theorem 1. *The recursive algorithm for computing the predicted distribution $p(x_k \mid y_{1:k-1})$ and the filtering distribution $p(x_k \mid y_{1:k})$ are given by the following recursive **Bayesian filtering equations**:*

- Initialization . *The algorithm starts from the prior distribution of the state $p(x_0)$;*
- Prediction . T*he predicted state x_k at time step k is computed given the dynamic model :*
$$p(x_k \mid y_{1:k-1}) = \int p(x_k \mid x_{k-1}) p(x_{k-1} \mid y_{1:k-1}) dx_{k-1}$$
- Update . *The posterior distribution of the current state given the measurements at time step k can be computed with the Bayes' rule :*
$$p(x_k \mid y_{1:k}) = \frac{1}{Z} p(y_k \mid x_k) p(x_k \mid y_{1:k-1})$$
where $Z = \int p(y_k \mid x_k) p(x_k \mid y_{1:k-1}) dx_k$ is the normalization constant.

Proof.
- *Prediction* . We can find the joint distribution of x_k, x_{k-1} given $y_{1:k-1}$ by :
$$p(x_k, x_{k-1} \mid y_{1:k-1}) = p(x_k \mid x_{k-1}, y_{1:k-1}) p(x_{k-1} \mid y_{1:k-1}) = p(x_k \mid x_{k-1}) p(x_{k-1} \mid y_{1:k-1})$$
Using the Chapman-Kolgmogorov equation on the previous equation and integrating over x_{k-1} gives us the result.
- *Update* . The equation derives from the conditional independence of the measurements $(p(y_k \mid x_k, y_{1:k-1}) = p(y_k \mid x_k)$.

□

3.2 The concept of the Kalman Filters

3.2.1 General idea

The Kalman Filter is an algorithm of data assimilation, using the Bayesian filtering equations, and which therefore takes a series of measurements and a process model - both containing statistical noise - and produces a more accurate estimate from this information. In a nutshell, the idea is to estimate the joint probability distribution over the variables at each timestep [26].
This method that the noise is assumed gaussian and that process and measurements covariances are assumed to be known (or at least having good approximation of it).

3.2.2 Overview of the method

The algorithm is a recursive filter - meaning that at each time step it reuses the output of the previous timestep - estimating the current state of the system in two steps at every timestep : **Predict** and **Update**.
I will now introduce some notations for the sake of simplicity :

The model is assumed to be evolving from time $k-1$ to time k as :

$$x_k = F_k x_{k-1} + v_k$$

where :

- x_k is the **true state** of the system at time k;
- v_k is the **process noise**, assumed to be following a gaussian distribution of zero mean and covariance Q_k : $\mathcal{N}(0, Q_k)$.
- F_k is the **state transition model** of the system at time k (a matrix in the linear case);

The observation z_k of the true state x_k at time k is modelled as :

$$z_k = H_k x_k + w_k$$

where :

- z_k is the **observation** of the true state x_k of the system at time k;
- w_k is the **observation noise**, assumed to be following a gaussian distribution of zero mean and covariance R_k : $\mathcal{N}(0, R_k)$.
- H_k is the **observation model** of the system at time k (an observation matrix);

The initial state and the noise vectors are assumed to be independent.

3.2.3 The algorithm in a nutshell

The evolution of the estimate of the step at time k can be conceptualized in two successive steps : the predict step and the update step.

Predict The idea of the predict step is to run one timestep of the model taking the output of the filter at the time $k-1$ to predict the joint probability distribution of the system at time k.
The algorithm is then :

$$\begin{aligned} \hat{x}_{k|k-1} &= F_k \hat{x}_{k-1|k-1} && \text{Predicted (a priori) state estimate} \\ P_{k|k-1} &= F_k P_{k-1|k-1} F_k^T + Q_k && \text{Predicted (a priori) estimate covariance} \end{aligned}$$

The a priori estimate of the covariance can be seen as the addition of the expected covariance of the system after going through the model and the covariance of the uncertainties of the model.

Update The idea of the update step is to readjust the joint probability distribution of the system using the a priori estimate and the observations.

The algorithm is then :

$\tilde{y}_k = z_k - H_k \hat{x}_{k\|k-1}$	Pre-fit residual (difference between the observation and the expected measures of the predicted state)
$S_k = H_k P_{k\|k-1} H_k^T + R_k$	Pre-fit residual covariance
$K_k = P_{k\|k-1} H_k^T S_k^{-1}$	Optimal Kalman gain (optimize $\mathbb{E}\big[\|\|x_k - \hat{x}_{k\|k}\|\|^2\big]$)
$\hat{x}_{k\|k} = \hat{x}_{k\|k-1} + K_k \tilde{y}_k$	Updated (a posteriori) state estimate
$P_{k\|k} = (I - K_k H_k) P_{k\|k-1}$	Updated (a posteriori) estimate covariance

The covariances $P_{k|k-1}$ and S_k are ensured by the Lemma **A.0.1** (see **Appendix A**).

From this we can note that the importance given to the observation or the system model can be tuned by varying the trace of the covariances Q_k/R_k. Indeed, this ratio influences the Optimal Kalman gain and therefore the computing of the a posteriori state. The higher the ratio Q_k/R_k is the closer will be the estimate to the observations (this will be highlighted in the following wit the experiments).

The Extended Kalman Filter (EKF)

For nonlinear dynamic and measurements models.

This algorithm works for dynamic and measurements models that are linear gaussian. If either or both the dynamic and/or measurement models are nonlinear, this algorithm needs to be adapted.

Then we would have f - a state transition function - and h the measurement function. In the algorithm, the state transition matrix F_k and the observation matrix H_k are replaced with the Jacobian matrices $F_x(m_{k-1|k-1})$ and $H_x(m_{k|k-1})$, where F_x and H_x are the Jacobian matrices of f and h, respectively, and $m_{k-1|k-1}$ and $m_{k|k-1}$ are the means of $x_{k-1,k-1}$ and $x_{k|k-1}$, respectively.

3.3 Kalman Filters and the Lorenz 96 model

3.3.1 Lorenz 96

The Lorenz96 model is a dynamical system formulated by Edward Lorenz in 1996. This model is very interesting for practicing on a chaotic system as it is rather simple in its formulation and the chaoticity of the system can be tuned. It is defined as :

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F$$

where F is the forcing constant, and $x_0 = x_N$, $x_{-1} = x_{N-1}$, $x_{N+1} = x_1$, etc...

Here x_i corresponds to the i^{th} variable of the state x of the model.

For the implementation of the Lorenz 96 model, I had first implemented a rather simple but not very robust code to integrate the solution along time. This method was not viable for a forcing constant $F > 10$.

I have then used the *Python* function **odeint** which is using the collection **ODEPACK** of *Fortran* solvers. The solving method is way more robust but needs a lot more of computational power.

The setup that I have implemented for the Lorenz 96 model gives 36 dimensions, the initial condition is $x_0 = F + 0.01x_{19}$ (which is the forcing that I have set up at 8 at first), and observations of the model adding white noise following a gaussian joint law $\mathcal{N}(0, 0.5 * I_N)$.

We can see the graph representing the perfect model and its measurement below (Fig.1).

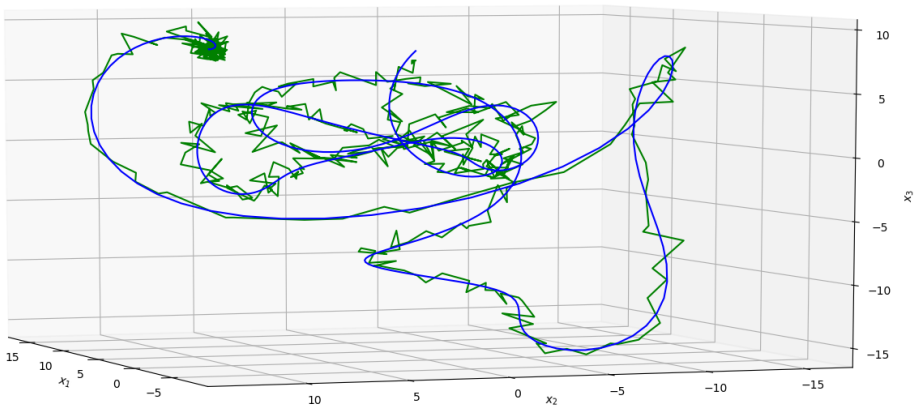

Figure 3.1: The perfect model is in blue and the measurements are in green.

Kalman Filter : Varying the process covariance assumption It is to be noticed that if not noticed otherwise, the Lorenz 96 model is solved with the first non-robust implementation.

We first want to highlight the importance and the consequences of varying the ratio process covariance/observations covariance.

For that we will represent in the following figures the results for varying process covariances. We look at the trace of the covariance matrices and the Root Mean Squared Error for comparing the accuracy of each corrected model. (RMSE).

The trace of the data covariance matrix is set up to $Tr(C_{data}) = 9$, (indeed we have set $C_{data} = I_N * 0.25$).

Our first run of the filter has been made with the assumption of process covariance being $C_{process} = I_N * 0.01$ (Fig.2).

We obtain that the filtered model has these features : $Tr(C) = 3.73$, $RMSE = 0.16$.

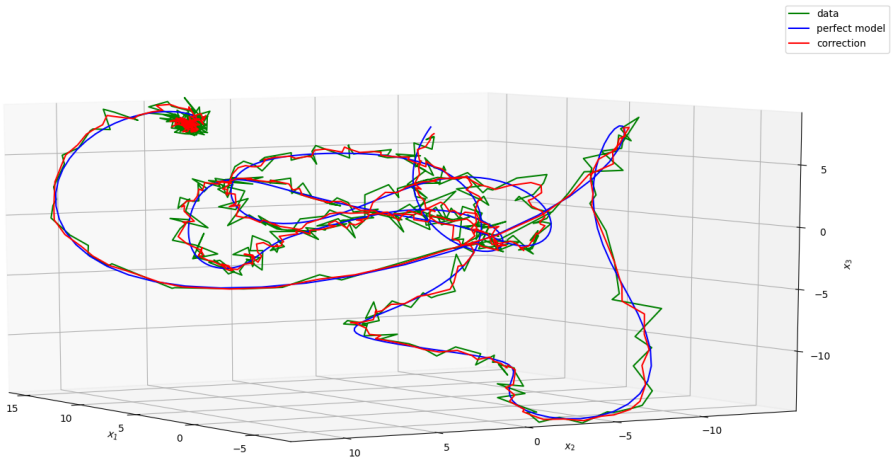

Figure 3.2: Graphic of the perfect model (in blue) the observations signal (in green) and the filtered signal (in red). The process covariance is here $C_{process} = I_N * 0.01$

Our second run of the filter has been made with the assumption of process covariance being $C_{process} = I_N * 0.0001$ covariance (Fig.3). We therefore would expect to have a filtered signal to be very close to the perfect model (since there is no bias in the numerical model nor in the observations).

We obtain that the filtered model has these features : $Tr(C) = 0.24$, $RMSE = 0.1$.

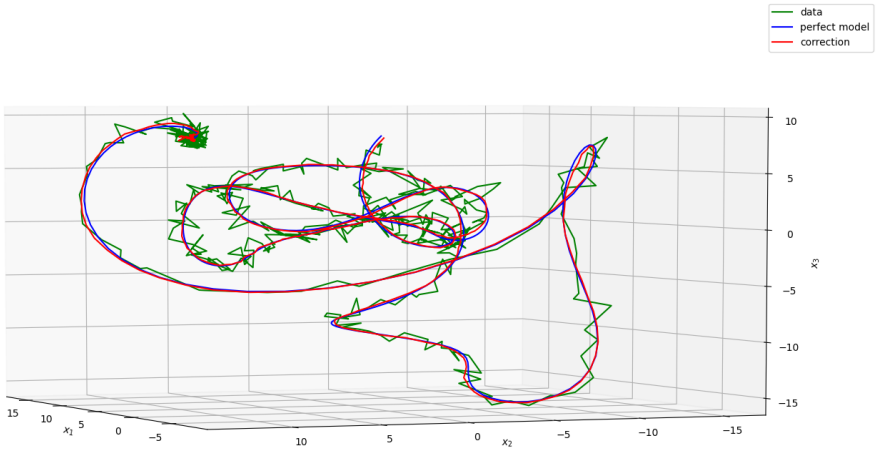

Figure 3.3: Graphic of the perfect model (in blue) the observations signal (in green) and the filtered signal (in red). The process covariance is here $C_{process} = I_N * 0.0001$

Our third run of the filter has been made with the assumption of process covariance being $C_{process} = I_N$ covariance (Fig.4). We therefore would expect to have a filtered signal to be very close to the observations.

We obtain that the filtered model has these features : $Tr(C) = 7.45$, $MSE = 0.41$.

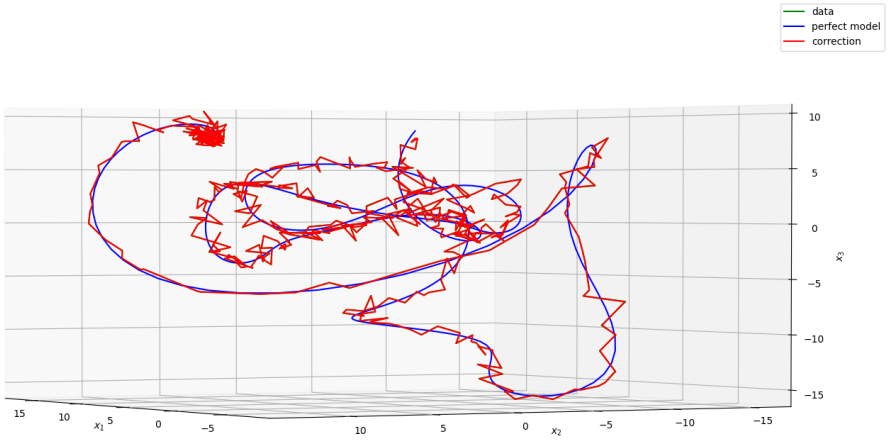

Figure 3.4: Graphic of the perfect model (in blue) the observations signal (in green) and the filtered signal (in red). The process covariance is here $C_{process} = I_N * 1$

We can see how the ratio between assumed process covariance and data covariance changes the way the filter behaves. In the experience that I have made, the measurement and process biases are almost nonexistent and therefore the real model can be retrieved easily by adding weight on the measurement covariance for "smoothing" the posteriori model. However, such a situation is not possible in what we will study and in the following I try to add more uncertaintainties by adding biases to the numerical model.

We will now focus on the Ensemble Kalman Filter as it is what we will use for the ocean model with a brief comment on the Unscented Kalman Filter.

3.4 Ensemble Kalman Filter and Unscented Kalman Filter

3.4.1 General idea of EnKF and UKF

The problem of the standard and Extended Kalman Filters appears when the underlying model is highly nonlinear. Indeed, the EKF linearizes the propagation of the state and covariance results in poor performance from the EKF. Moreover, maintaining the covariance matrix for the state of the system turns out to be computationally too heavy for high dimensional systems.

For these reasons, the EnKF and UKF use an ensemble $(x_k^i)_{1\leq j\leq n}$ of estimated state vectors of the system, where n is the number of state vectors. These state vectors give information concerning the mean and covariance of the filtered signal [1, 24]. Here is a simplified scheme of the functioning of the EnKF in practice (Fig.5) [7].

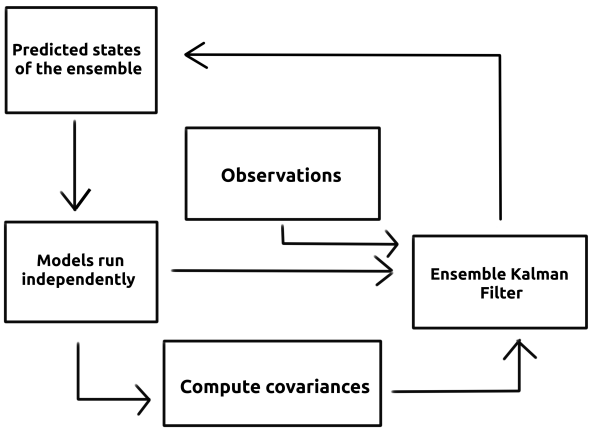

Figure 3.5: Scheme of the functioning of the EnKF

In practice the algorithm for the EnKF becomes :

Predict

$$\hat{x}^j_{k|k-1} = f(x^j_{k-1|k-1}) + v^j_k \qquad \text{Predicted (a priori) state estimate}$$
$$P_{k|k-1} = \hat{P}_{k|k-1} \qquad \text{Predicted (a priori) estimate covariance}$$

$v_k \sim \mathcal{N}(0, Q_k)$ represents a stochastic process deviation .
$\hat{P}_{k|k-1}$ is the sample covariance of the ensemble state vectors.

Update

$$\hat{z}^j_k = h(\hat{x}^j_{k|k-1}) \qquad \text{Measurements of the predicted states}$$
$$S_k = P^z_{k|k-1} + R_k \qquad \text{Observations sample covariance}$$
$$C_k = \frac{1}{n-1}\Big(\sum_{j=1}^{n} \hat{x}^j_{k|k-1} - m_{k|k-1}\Big)\Big(\sum_{j=1}^{n} \hat{z}^j_k - \mu_k\Big)^T \qquad \text{Cross covariance}$$
$$K_k = C_k S_k^{-1} \qquad \text{Optimal Kalman gain (optimize } \mathbb{E}\big[||x_k - m_{k|k}||^2\big]\text{)}$$
$$\hat{x}^j_{k|k} = \hat{x}^j_{k|k-1} + K_k(\tilde{y}_k + w^j_k - \hat{z}^j_k) \qquad \text{Updated (a posteriori) states estimate}$$
$$P_{k|k} = P_{k|k-1} - K_k S_k K_k^T \qquad \text{Updated (a posteriori) estimate covariance}$$

$w_k \sim \mathcal{N}(0, R_k)$ represents a stochastic measurement deviation .
$P^z_{k|k-1}$ is the sample covariance of the measurements of the predicted state vectors.

The algorithm of the UKF is mostly the same except for the deviations of the state vectors (δ^j, σ^j for the EnKF) that are deterministic in the UKF algorithm and drawn according to the covariance columns $(P^i_{k|k-1})_i$ but I will not get into the details.

3.4.2 Implementation of the algorithm

For the UKF I have implemented an algorithm that can run it. However, the choice of the coefficients for the sigma points is something that I haven't really been able to handle. I have therefore not been

able to have any satisfying result with the UKF.

On the other hand the implementation of the EnKF has been successful and I have been able to draw good results from that.

3.4.3 EnKF with sparse in time data

In this paragraph we will consider the results of the EnKF on the Lorenz 96 model using only sparse in time data.

I have run the EnKF with different process covariances and keeping the data covariance at 0.25 and we use 50 ensembles. The process covariance is the identity matrix multiplied by a constant going from 0.01 to 1. We also consider a sparse in time data, having gaps in time of 5,10 or 20 time steps. Here are the results - the RMSE between the output of the EnKF and the perfect model and with the covariance of the output - plotted (Fig.6).

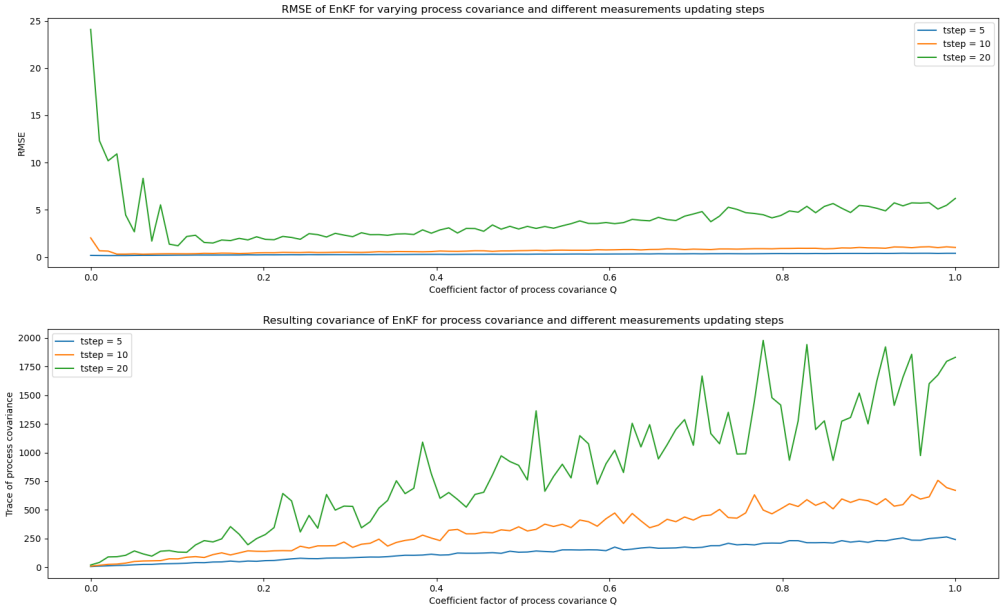

Figure 3.6: RMSE and covariance of the corrected model depending on the process covariance. We also vary the sparsity in time of the data.

We can observe that in this model there is a threshold for the sparsity in time of the data. The covariance and RMSE become way too large when we go above gaps of 10 time steps. Moreover, and more importantly from these results we can deduce that taking a process covariance of $Q = 0.1 * I_n$ seems to be the best option.

We can also notice that when the data is sparse in time, having a small process covariance results in great errors compared to the perfect model. Indeed, trusting the model here would lead to underestimating the chaos system that is underlying and the randomization of the ensembles (or the range of possibilities it can cover) shrinks too much to lead to an accurate estimation.

I have then run the EnKF varying the number of ensembles (Fig.7), still in the idea of studying under

the scope of time sparsity.The same experiment is made where the number of ensembles goes from 10 to 110 and we choose the process covariance to be $Q = 0.1 * I_n$ following the previous results.

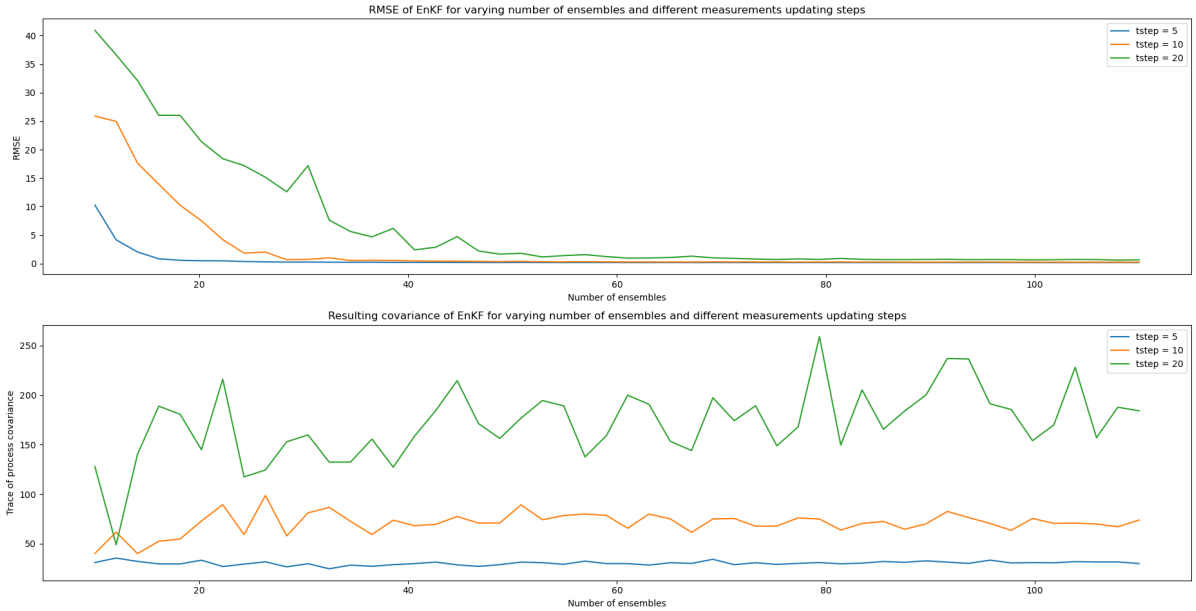

Figure 3.7: RMSE and covariance of the corrected model depending on the number of ensembles. We also vary the sparsity in time of the data.

From this we can deduce a reasonable number of ensembles that we need depending on the sparsity of the data.

3.4.4 EnKF with sparse in dimensions data

In this paragraph we will consider the results of the EnKF on the Lorenz 96 model using only some dimensions of data.

To study the influence of the sparsity in variables of the data, we configure an observation matrix that removes some variables from the filtering process.

The first way to do it is to remove 1 variable every k ones (example : 111110 111110 111110 ; we remove 1 variable each 6 variables).

We first run the EnKF with 50 ensembles and we vary the process covariances the same way as above. We then look at 3 different degrees of sparsity : we decide to remove 1 variable each 8,4 or 2 variables. Here are the results for the RMSE and the resulting covariance (Fig.8) :

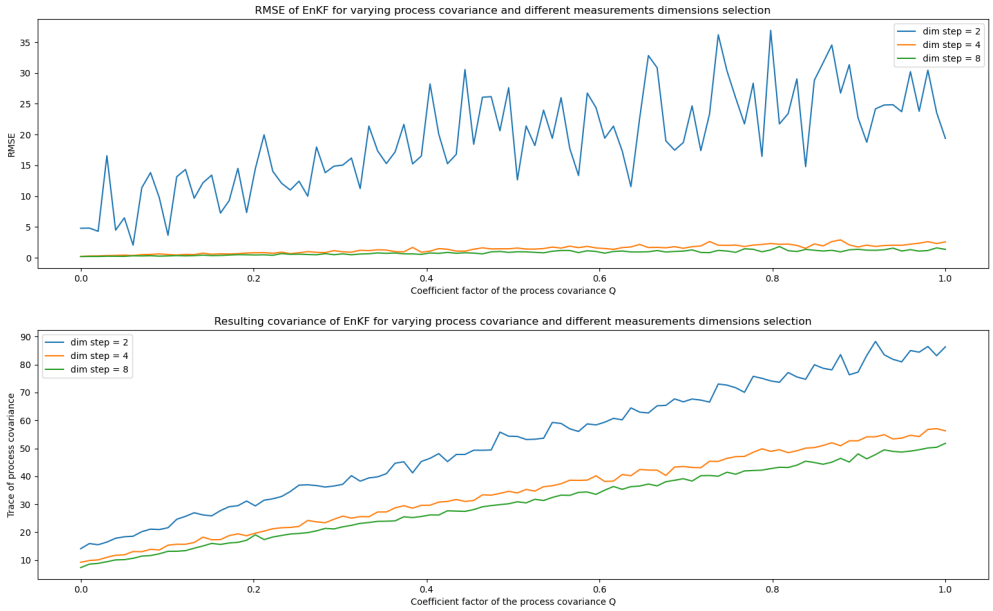

Figure 3.8: RMSE and covariance of the corrected model depending on the process covariance. We also vary the sparsity of the data.

The same experiment but varying the number of ensembles and setting the process covariance at $Q = 0.1 * I_n$ gives us these results (Fig.9) :

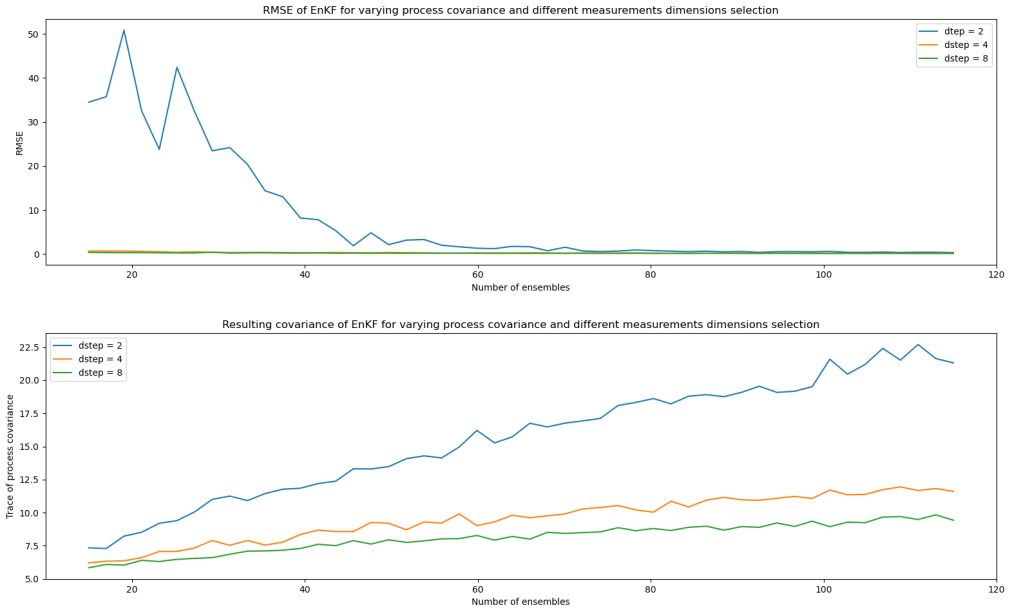

Figure 3.9: RMSE and covariance of the corrected model depending on the number of ensembles. We also vary the sparsity of the data.

When having a sparse in variables data, it would seem that for this model a low (around 0.1)

coefficient before the process covariance and at least 50 ensembles are necessary to obtain sufficiently accurate results.

One important thing to remark at first is that the way that we choose the dimensions to be part of the data can make big differences. Indeed, in the Lorenz 96 model, the dimensions are interacting "four in rows" so when choosing the data, it will have big impacts on the accuracy of the EnKF model whether we choose to take the data by packs of 2,3,4 or more. (a pack of data would look like '111 000 111 000 111 000 ...' where only the ones are chosen and where the same numbers of dimensions would be used for data). That is actually what we can observe from the experiment, where there is a big increase of the RMSE when the data are taken into "packs of 4 or more dimensions".

That is the way we have chosen to remove dimensions in the data in the following experiment. We have removed a row of k variables each $4k$ variables so that we obtain a quarter of the data being removed. We choose k being 1,2,3 or 4 in the following experiment and vary the process covariance and the number of ensembles (Fig.10 & 11), the same way as above :

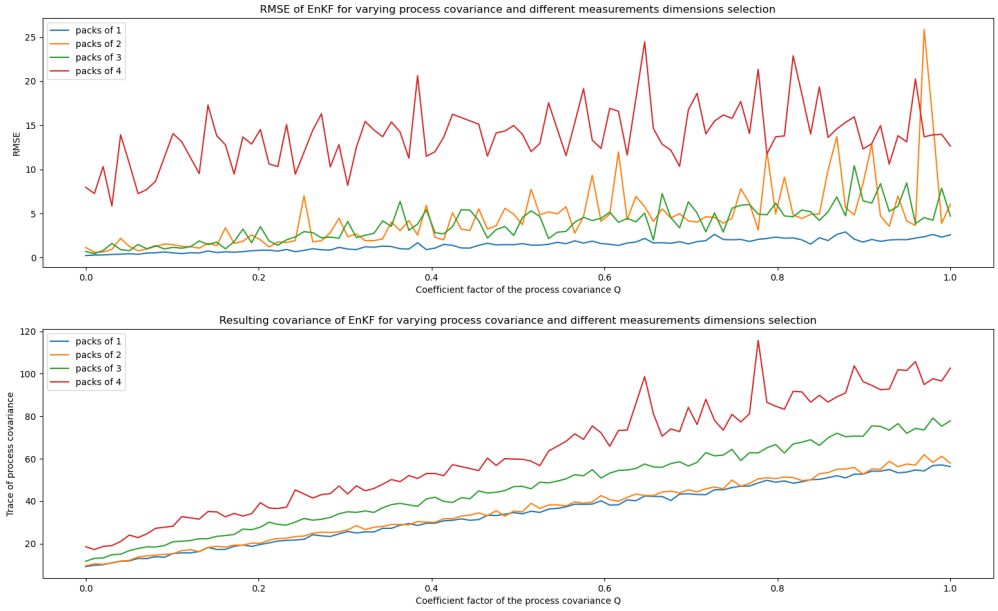

Figure 3.10: RMSE and covariance of the corrected model depending on the process covariance. We also vary the sparsity of the data.

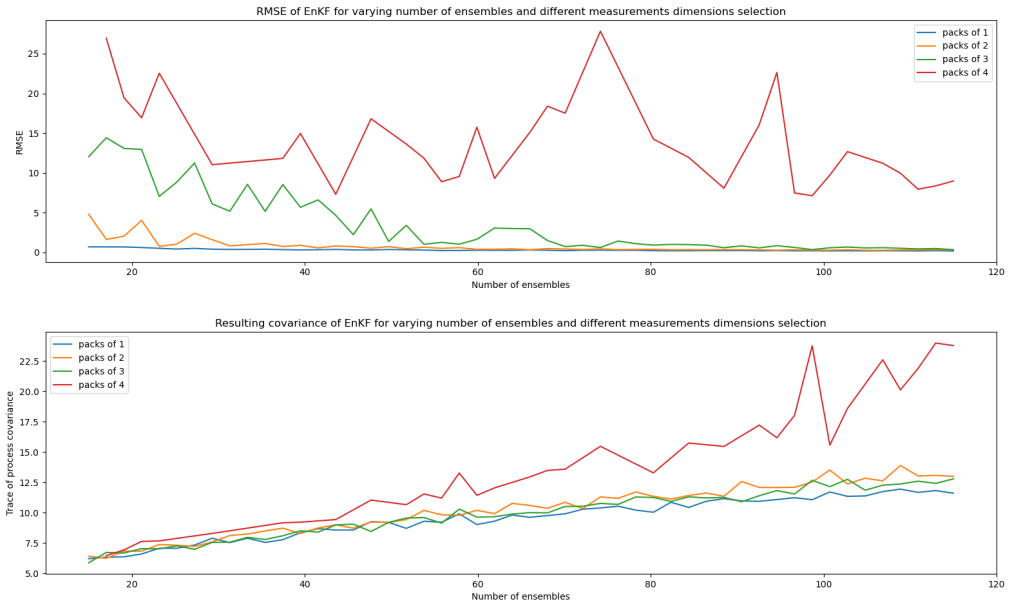

Figure 3.11: RMSE and covariance of the corrected model depending on the number of ensembles. We also vary the sparsity of the data.

As mentioned before, we find that the packs of 4 become way less accurate as expected. We can also deduce that a bigger number of ensembles will be needed if the data is too sparse in term of variables.

However, we can discuss some limitations of the Lorenz 96 model and the way we have studied it.

Here the model is very well solved and the physics of the model as well. Moreover, all the variables have the same impact on the model, and that is underpinned by the fact that changing which variables are removed doesn't change the results. However the way these are removed affect the results.

Moreover, I have always worked here with a diagonal process covariance. On a quick experiment and without going any further, I have used a matrix that takes into account the interaction between variables without leading to any improvement in the filtered model.

3.5 Aspects of the EnKF that will need additional work

After this preliminary work and with the addition of state of the art work in this field, we can identify the possible weaknesses of the EnKF that we could encounter when implementing it on the UVic ESCM model.

3.5.1 Filter divergence

In some cases, the ensemble model can blow up in finite-time, with an exponential inflation, even when the system has bounded solutions. This is known as **catastrophic filter divergence**. This kind of deficiency appears as the sampling of the covariance contains error due to the limited number of ensemble members, restricting the rank of the covariance. The 'model error' subspace spanned by this sample

covariance therefore does not accurately represent the system covariance.

This problem is very well described in the Fig 3.12. The model used here involves a rotational map on $\mathbb{R}^2$ and a readjustment map which links the ensemble members on the nearest grid point. (A) The posterior ensemble lies in a subspace spanned by the forecast ensemble (the vertical subspace). (B) The observation model is $y_n = Hx_n + \xi_n$ matrix $H = \begin{pmatrix} 1 & 0 \\ \epsilon^{-2} & 1 \end{pmatrix}$ where ϵ will be chosen very small for a sake of divergence and ξ_n is a normal random variable. This observation model shifts the ensemble members on the vertical axis. When applying the rotational map (C) and the readjustment map (D), the ensemble members turn out to spread further on the horizontal axis [8].

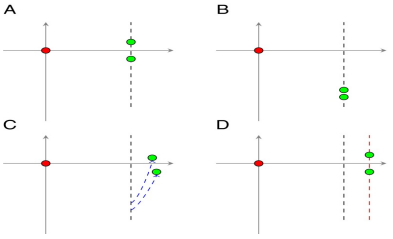

Figure 3.12: Representation of catastrophic filter divergence. The red dot represents the 'true state' and the ensemble members are represented by the green dots. Source : Kelly et al. 2015 [8]

3.5.2 Covariance inflation

The problem mentioned just before occurs when the number of ensemble members is too limited to correctly represent the model covariance.
A first solution would be to use more ensemble members, adjusting this amount to the number of variables involved in the EnKF. However, when working on complex systems (and particularly in Earth System modeling), the number of variables is too high and one can not reasonably use enough ensemble members to avoid filter divergence.
A second solution is to artificially increase the subspace spanned by the ensemble members. This is called **covariance inflation** [28, 10]. This method leads to three different approaches : multiplicative inflation, additive inflation, or both.
For the multiplicative inflation, the idea is to inflate the standard deviation of the ensemble by a fixed factor r :

$$x = r(x - \overline{x}) + \overline{x}$$

On the other hand, the idea of additive inflation is to inflate the covariance by a random vector of fixed covariance Q :

$$x = x + \nu \quad , \quad \nu \sim \mathcal{N}(0, Q)$$

Both of these approaches need to be chosen and applied depending on the system involved. For example, models which involve very underestimated covariances would need an additive inflation to be efficient.

These adjustments will be taken into account when constructing the EnKF for the UVic ESCM model.

Chapter 4

Implementing the Ensemble Kalman Filter on the UVic 2.9 model

The next step of this project was to use the previous work done on the Lorenz 96 model and its results to build a code that would be run with the UVic 2.9 climate model. Previous applications of the EnKF (and others close versions of it like the LETKF or EAKF) to weather forecasting, but also to climate predicions on longer periods [5, 25, 6, 29]

4.1 The Ensemble Kalman Filter for constraining the input of freshwater flux in the Northern Atlantic Ocean

Our objective is to use the Ensemble Kalman Filter to constrain the freshwater flux in the Northern Atlantic Ocean. Indeed, the amount of freshwater added in the North Atlantic plays a key role in the behavior of the AMOC [12, 13, 18]. Together with the UVic 2.9 oceanic and atmospheric model, and with observations from sediment cores, ice cores and pollen cores we aim to refine the state of the art understanding of the AMOC behavior during the Last Deglaciation (19k-11k years ago).

4.2 Applying the Ensemble Kalman Filter with intrinsic variables offered by the UVic 2.9 model

The UVic Earth System Climate Model (ESCM) consists of a three-dimensional ocean general circulation model coupled to a thermodynamic/dynamic sea-ice model, an energy-moisture balance atmospheric model with dynamical feedback, and a thermo-mechanical land-ice model. It works with a collection of submodules written in Fortran and compiled altogether forming the compiled file which will be run, modeling the physics model of oceanic and atmospheric circulation on the Earth.
The objective was to write a code that would run the model coupled with the EnKF to correct the model simulation with observations throughout the computer simulations. Once the EnKF is done, the variables corrected by the EnKF are then set as initial conditions and the model is run again and this process is reiterated again until the given end of the simulation. The overall functioning of this code would use the UVic 2.9 model to run the physics simulation for a certain number of years, apply the

EnKF on specific variables, and then running the model again for a same number of years with the corrected variables as initial conditions. That process would be done over and over until we reach our given date of end of the simulation.

4.3 Building a wrapper for running the UVic 2.9 model at each time step

Writing the EnKF code inside the already working UVic 2.9 model would have been very complicated as the structure and functioning of the code is very complex. More precisely, it would have required a lot more time to analyse and understand all of the subtleties of all the different submodules that are called by the main code.

From this point, the method that I chose was to write a code (a wrapper) which would run the model, update the information, apply the EnKF, implement this correction in the initial conditions and run the model again. The main challenge for implementing the wrapper for The EnKF on the UVic 2.9 model was to find a way to gather all the resulting variables from the model runs, and to run the model again with the updated state variables. Indeed, the specific format used in the UVic 2.9 model had to be used to get the state variables from the model runs and being able to input the corrected state variables as initial conditions in the correct format so that the UVic 2.9 model can be ran. Moreover, the data files were netCDF files, which functioning is very similar to dictionaries, except that you cannot modify or delete information from the file, you can only add variables or append data in the already existing variables. With all of these requirements taken into account, I chose to write the code in Python, using the **netCDF4** library for working on the netCDF files, and using **subprocess, os** for interacting with the operating system interface.

4.4 A first attempt to build an EnKF on all grid points of certain variables

This first try of implementing the EnKF was meant to gather information about how to implement the EnKF practically, understand the functioning of the UVic 2.9 model and outline the possible problems that would be encountered later.

4.4.1 Setting up the EnKF

In this paragraph, I will describe in a more practical way how I adapted the EnKF to the working environment of the UVic 2.9 model.

Choice of the state variables.

I will here present what state variables I chose and how I constructed the data to be used as observation in the EnKF.

First, I needed to choose state variables that influence by the AMOC or are influenced the AMOC since the aim of this project is to use the EnKF for constraining the possible behavior of the AMOC during the Last Deglaciation.
The AMOC intensity is primarily influenced by the temperature and salinity distributions in the ocean as well as atmospheric forcings. Using evidence of such oceanic temperature and salinity is therefore important for constraining the AMOC behavior.
A wealth of information on past ocean circulation is provided by the analysis of ice and sediment cores. Indeed, the observations that we will use for constraining the AMOC will come from these kinds of cores in the Atlantic ocean.
The traces of organic geochemistry can help for reconstructing the past ocean temperatures, using also species abundances of fossil planktons, trace-metal ratios or the distribution of oxygen isotopes [17, 11]. This information is taken from the sediment and ice cores aforementioned.
The ocean salinity can also be reconstructed with organic geochemistry and oxygen isotopes, mostly [11]. Additionally, indirect evidence for ventilation rates of the AMOC comes from the distribution of isotopes such as ^{13}C, and from the radiocarbon content of the atmosphere.

In the end, we pick 5 state variables for the EnKF, namely : **temperature, salinity, DIC, ^{13}C DIC, ^{14}C.**

For a sake of clarity, we will use the term of **state variable** for referring to temperature, salinity, ... , ie. physical variables.

$$x_i = (x_i^j)_{j=1}^{m_i}$$

where m_i represents the number of grid points where the state variable i is being used.

The term of **variable** will refer to a state variable at a specific point of the grid, unless otherwise stated

$$x = (x^j)_{j=1}^{M}$$

where M represents the total number of variables used in the EnKF.

We then needed to link these state variables with a typical variation that would be used in the EnKF when adding the stochastic component to the predicted state system.
For each variable, we used a formula that would capture the variability and order of magnitude of the state variables. Using the state variables values on the whole grid of the UVi2.9 model, we used this formula :

$$\eta_i = \frac{\max_{1\leq j\leq m_i}(x_i^j) - \min_{1\leq j\leq m_i}(x_i^j)}{100} \tag{4.1}$$

where η_i is the typical variation of the state variable i.

Preparing the observations.

Now that the state variables were chosen and associated with their typical variation, we needed to construct observations to be used in the EnKF.

For that I ran a model using initial conditions previously stored, and produced at each model year a file containing the values of the state variables. This could be considered as the 'perfect model run'. which would be considered as a target. I then introduced observations noise to this data in the form of a stochastic component, using the typical variations for each state variable.
The stochastic component added to each position was independent to the other positions and I used different amplitude for the stochastic component standard deviation so that the standard deviations for the stochastic components were set as : $\sigma_i^j \in \{0.25\eta_i,\ 0.5\eta_i\}$, where σ_i^j is the stochastic component for the variable i at position j.

I also constructed a biased model run, adding $0.5°C$ on the whole ocean temperatures in addition to the observation noises. Although, this model run would not give meaningful physical results, it would reveal if the EnKF constructed here could adapt to this bias in an ideal situation like this. I call this an ideal situation as we use all of the possible grid points of the UVic 2.9 model as measurements position, which will not be the case in reality. Since we also change the temperature and salinity, key factors in the AMOC behavior, in the ensemble runs we have a direct and insantenate impact on the state system whereas we aim to only directly change the amount of freshwater input in the North Atlantic in our project objective.

Creating directories.

After having created the data to be used as observations we needed to construct the structure that would be used for the ensemble members.

First of all, a directory needs to be specifically set up to contain the n directories that will be used as ensemble members for running the successive model runs. Each directory will contain the code needed for the simulation and must be named as : '**ensemble_name**i', where $i \in \{0, ..., n-1\}$.

The data from the observations also need to be stored in a directory within this same directory.

Extracting the state variables for the EnKF.

At each iteration, the EnKF algorithm requires to gather all the state variables that we have previously chosen and store it in a 'state system' array.

The extraction of these state variables from the model is done by looking into the **restart.nc** files for the last conditions of the system, or into the **tavg.nc** files for a historic of the evolution of these state variables, averaged with an average step of 30 days, for taking into account the seasonality of the climate system.
Each of these state variables is filtered with the Ocean mask for only using meaningful information and then flattened to be stored in an array. Then all of the flattened arrays containing the state variable information are concatenated in the 'state system' array which will be used in the EnKF analysis process.

.

4.4.2 The step-by-step development of the code

The first idea for implementing the EnKF was to use certain variables from the **restart.nc** file, and to use all of the information that would be extracted from these state variables, meaning using the data from all of the grid points of the model. As mentioned previously but was not developed, some problems arose from the adaptation of the UVic 2.9 model to our objective.

Ocean mask.

As previously mentioned, the first problem was that not all of the grid points could offer useful data. Indeed, a lot of the grid points were placed on geographical positions where no ocean could be found, but only rock. An easy solution was to construct a mask for the ocean with which we could filter the locations on which we would extract the data.

Managing a large amount of variables.

Now that we have only useful information, this information could be passed through the EnKF so that we could maybe have some results, or more likely a new problem standing out.
Of course the second happened as the dimension of the state variables (all of the grid points of the ocean mask) used in the EnKF was too big and therefore the linear algebra, matrix products and matrix inversions, could not be performed with a limited RAM memory. These calculations would indeed have required 200Mo of RAM memory, which is an indecent amount.
Following the advice of Pr. Restrepo I have therefore decided to reduce the time of calculation by considering that the variables would be completely independent to each other and therefore, we would only consider the standard deviations of each variable. This different approach induced a changein the way of computing the sample covariances (which were previously very time consuming) by only computing the squared sample standard deviations and storing it in a 'covariance' array. The system of equations for deriving the covariances becomes the following.

Model sample covariance $\hat{p}_{k|k-1}$ and a priori model covariance q_k :

$$\hat{p}_{k|k-1} = \Big(\frac{1}{n-1}\sum_{j=1}^{M}(\hat{x}^j_{k|k-1} - \overline{x}_{k|k-1})^2\Big)_{j=1}^{M} \tag{4.2}$$

$$q_k = (\sigma_k^j)_{j=1}^{M} \tag{4.3}$$

where :

- $\hat{x}^j_{k|k-1}$ is the j-th variable of the a priori estimate,
- $\overline{x}_{k|k-1}$ is the mean of $(\hat{x}^j_{k|k-1})_{j=1}^{M}$,
- σ_k^j is the prior estimate of the model standard error for the $j-th$ variable.
- m is the number of variables.

Predicted measurements sample covariance $p^z_{k|k-1}$ and a priori observations covariance r_k:

$$s_k = \Big(\frac{1}{n-1}\sum_{j=1}^{M}(\hat{z}_k^i - \overline{z}_k)^2\Big)_{j=1}^{M} \tag{4.4}$$

$$r_k = (\rho_k^j)_{j=1}^{M} \tag{4.5}$$

where :

- $\hat{z}_k^j = h(x_{k|k-1}^j)$ is the measurement of i-th variable of the a priori estimate,
- $\overline{z}_k$ is the mean of $(\hat{z}_k^j)_{j=1}^{M}$,
- ρ_k^j is the prior estimate of the observations standard error for the $j-th$ variable.

Cross covariance of the predicted states and their measurements :

$$c_k = \Big(\frac{1}{n-1}\sum_{j=1}^{M}(\hat{x}_{k|k-1}^i - \overline{x}_{k|k-1})(\hat{z}_k^i - \overline{z}_k)\Big)_{j=1}^{M} \tag{4.6}$$

$v_k^j \sim \mathcal{N}(0, Q_k)$ represents a stochastic process deviation .

With these covariances we can derive a similar algorithm to the EnKF method that we have seen above. Instead of using matrix operations, we will only use element-wise operations.

Algorithm for overnumerous state variables

Predict

$$\hat{x}_{k|k-1}^j = f(x_{k-1|k-1}^j) + v_k^j \qquad \text{Predicted (a priori) state estimate}$$

where $v_k^j \sim \mathcal{N}(0, q_k^j)$ represents a stochastic process deviation.

Update

$$\begin{aligned}
\hat{z}_k^j &= h(\hat{x}_{k|k-1}^j) && \text{Measurements of the predicted states}\\
s_k &= p_{k|k-1}^z + r_k && \text{Observations sample covariance}\\
k_k &= c_k s_k^{-1} && \text{Optimal Kalman gain}\\
\hat{x}_{k|k}^j &= \hat{x}_{k|k-1}^j + k_k(\tilde{y}_k + w_k^j - \hat{z}_k^j) && \text{Updated (a posteriori) states estimate}\\
p_{k|k} &= p_{k|k-1} - k_k s_k k_k^T && \text{Updated (a posteriori) estimate covariance}
\end{aligned}$$

where $w_k^j \sim \mathcal{N}(0, r_k^j)$ represents a stochastic measurement deviation .

Since this first application of the EnKF is being performed with ideal conditions and do not reflect what will be done in the final work of this project, we for now define the prior estimate of model and observations covariances as :

$$(q_k^j)_{j=1}^{M} = Q \times (\eta_{i_j})_{j=1}^{M} \quad , \quad (r_k^j)_{j=1}^{M} = R \times (\eta_{i_j})_{j=1}^{M}$$

where

- η_{i_j} is the typical variation related to the variable j.
- Q, R are the coefficients used for determining, respectively, the prior model covariance and the observations covariance.

The calculations with the new method were done in 5mn, a decent time.

Formatting the initial conditions.

Now that we can actually launch the EnKF with the state variables from the UVic 2.9 model, we can aim to run it and hope that the code will end with no failure, whatever the results would be. The objective at this point is to have a code that runs our EnKF and try to make it run for several years with the ensemble runs and the observations. We used the results of a previous UVic 2.9 model run as observations.
However, this new implementation stopped during the first iteration during the preliminary checks made by the **UVic_ESCM** executable. The problem lied in the initial conditions, as the UVic 2.9 model could not support the type of file containing the initial conditions. The **restart.nc** files which contain the initial conditions are stored as netCDF files, I therefore used the *netCDF4* package of Python to create and modify the initial conditions. However, I figured out that important to extract from this is that we need the output of my Python code was in the format netCDF4 and that the UVic 2.9 model needed this file in the format *Classic netCDF*. The solution that I found for this problem is to use the command *nccopy* directly in the Terminal which allows to copy a *netCDF* file from one specific format to another.

Model not converging : the leap-frog numerical scheme.

Now that the initial conditions could be input in the UVic 2.9 executable, the code was going through the analysis step of the EnKF and then running the model with the corrected state variables. However, it resulted in an almost immediate crash of the model, returning that the model could not converge. When referring about this problem to the Pr. Schmittner, he explained to me that the UVic 2.9 model uses a **leapfrog** numerical scheme, that requires the updating of two separate code variables for each physical variable.

Leap-frog integration :

Let the position of an object at time k be x_k. We also want to consider the velocity v_k of the object at time k.
The outside forcing is represented by the acceleration $A(x_k)$ at the position x_k and at time k.

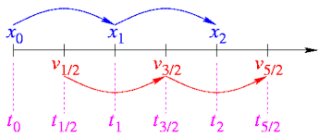

Figure 4.1: Functioning of the leapfrog scheme

$$v_{k+1/2} = v_{k-1/2} + A(x_k)\Delta t$$
$$x_{k+1} = x_k + v_{k+1/2}\ \Delta t$$

where Δt is the size of each time step.

In this setting, we would have *'temp1'* and *'temp2'* in the UVic 2.9 model, representing the temperature at two distinct moments of the simulation. From that I could understand that it is important to add the same stochastic component, and apply the same Kalman correction to both of these variables - *'temp1'* and *'temp2'* - during the process of the update step. In the end, this conservation of the difference between *'temp1'* and *'temp2'* can be seen as a conservation of the inertia of the climatic and atmospheric system. It is therefore clear that changing the difference between *'temp1'* and *'temp2'* would lead to an extreme change in the trajectory of the temperature, leading in the end to the crashing of the UVic 2.9 model.

Model not converging : temperature not converging.

Until now, the way that the random noise was added in the EnKF algorithm was as independent random normal variables on each point of the grid considered, such as :

$$v_k^j \sim \mathcal{N}(0, q_k^j) \quad , \quad w_k^j \sim \mathcal{N}(0, r_k^j)$$

However, when applying the EnKf and running the UVic 2.9 model with the corrected state variables, the solver of the UVic 2.9 raised a problem, namely : " *Warning: land surface temperature not converging.*
The source of this problem was unknown but I could only assume that the problem came from the distribution of the temperature on the model grid. If the UVic 2.9 model had a function checking the temperature differences on adjacent points of the model grid, then this problem could be solved by smoothing out the stochastic components added to the temperature on the whole grid. In more mathematical terms, this would mean to use a gaussian process over the model grid with a given Matérn covariance function for the model stochastic component v_k.

.

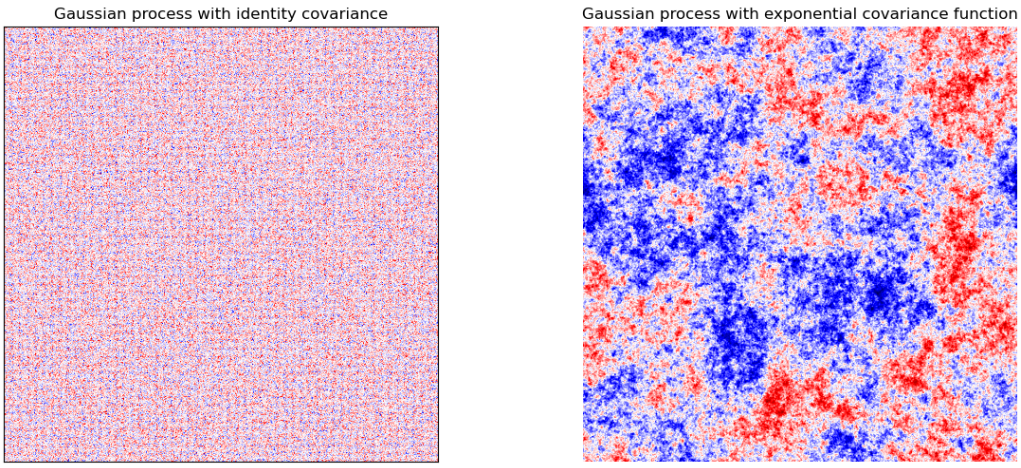

Figure 4.2: Heatmaps for gaussian processes with two different covariance functions : identity covariance function on the left and exponential covariance function on the right

Further details about the development of this gaussian process are to be found in the Appendix ?

Since the independence of the variable is necessary to apply the EnKF as this ensures the unbiasedness of the sample covariance, applying this gaussian process in the EnKF would be problematic.
With the EnKF as we built it here, this independence requirement would be avoided since the "sample covariance" only takes into account the standard deviation. The dependence of the stochastic component of a same state variable would therefore not be acknowledged by the algorithm as we built it.

This worked and the EnKF was performed successfully afterwards.

Later in the internship, I came back to using the EnKF with the stochastic component as a 'cluster random field' since I refined the standard deviations used for the random field. This adjustment resulted in a success as the EnKF could be performed to the end.

4.4.3 Performing the EnKF.

Now that the different problems encountered during the adaptation of the EnKF to the UVic 2.9 environment were solved, and that the wrapper was running to the end without crashing, we look at the results of the EnKF. We first performed it with the 'perfect model run' observations. The aim of this was to check if the EnKF diverges when coupled with the UVic 2.9 model. We then performed it with the 'biased model run' observations to evaluate if the EnKF could adapt to this biased information. Finally, we performed it with 'biased model run' and sparse spatial variables, only using a small part of the model grid points.
Only temperature and salinity were used here. Indeed, in the specific way that we have built the EnKF the variables are totally independent to each other. The state variables **DIC,** ^{13}**C DIC,** ^{14}**C** having no or very few short-term direct impact on the AMOC or the ocean and atmospheric environment, would have no benefit to include it in the state variables used in the EnKF. Moreover, removing these state variables can shorten the time needed for the analysis step of the EnKF.

Some precisions on the setting up of the EnKF

First of all, we used an additive covariance inflation component as the climate numerical model is very stable and therefore needss this additional component to have a reasonably accurate ensemble subspace. The additive factor is a covariance matrix representing the supposed standard deviations of the state variables of the system.In our case, we used two different values : $\sigma_1 = 0.1°\text{C}$ for the temperature and $\sigma_2 = 5 \times 10^{-6}$ for the salinity. These values will also be multiplied for some factors to study the impact of such changes.

The observations were constructed with the data from a previous run of the UVic 2.9 model. We performed 3 different observation sets :

- The first includes a gaussian random noise on both the temperature and salinity which standard deviations are $0.5\sigma_1$ for the temperature and $0.5\sigma_2$. This random noise is performed on all grid points independently.

- The second and third include a random noise with the same parameters as the first one, with a bias introduced in the temperature. This bias is an increase in the temperature of 0.5°C for the second observation set and 1°C for the third observation set. This bias was performed on all the model grid points.

Although the biased observation sets don't have a physical sense, it gives us an indication of the ability of the EnKF to adapt to inaccurate initial conditions.

The random field added in the EnKF for representing the observations uncertainties is chosen such as the standard deviation is $0.5\sigma_1$ for the temperature and $0.5\sigma_2$ for the salinity. This values will be changed to study the impact of such changes.

The analysis step was performed every year.

Cluster vs Gaussian

We first performed a 'perfect model run' using the non biased observation set. We used different amplitudes of random fields applied on the model state variables and on the observations to represent different level of uncertainties. The ratio of 'trust' between model and observations was always given equal ($R = Q$).
We also needed to determine which of **cluster random field** or **gaussian random field** gave better results when used in the EnKF.
For this purpose we performed the EnKF with 6 different settings. We performed the EnKF with 3 different amplitudes of uncertainties for both random field type (gaussian and cluster). The amplitude used for the standard deviations were of 0.25, 0.5 and 1, of the typical standard deviations σ_1 and σ_2. The results on the temperature, salinity and maximum flow of the Meridional Overturning Circulation (MOC).

We can observe in Fig 4.3 that the variance of the ensemble members for all of these variables is way larger when using the gaussian random field. That can be expected as the cluster random field tend to represent a random noise over the model grid. This anomaly would be in a sort smoothed out rapidly by the physics process of the UVic 2.9 model. On the other hand, the random component from the gaussian field express a deviation at a regional level, which would be less rapidly smoothed out and would give a larger ensemble spread than the cluster field as the UVic 2.9 model runs.
Moreover, we can see that the amplitude of the random fields used in the EnKF impact the convergence of the model for salinity and maximum MOC. Indeed, the ensemble mean does not converge to the target run values for salinity and maximum MOC as the highest the random field amplitude, the higher the error in the salinity is. We can also remark that the ensemble mean salinity is consistently being underestimated, suggesting that the amplitude of the uncertainties are probably underestimated, and that other phisical mechanisms are leading this deviation in global salinity values. We have therefore looked more closely at the changes of global ice and snow volume, and at the changes in net precipitation for our ensemble runs and compared it to the observation runs values. (see **Appendix C**). The results indicate that higher random noise amplitues lead to a larger decrease in terms of ice and snow volumes, explaining the differences in tha salinity that we can observe in Fig 4.3.
The error in the maximum MOC also increases with the amplitude of the random fields, but here the

error tends to overestimate the maximum MOC. This could be explain by the gradients of temperature and salinity induced by the introduction of a random field, and such gradients being the main mechanism driving the MOC.
The maximum MOC consequences show that the processes working in the UVic 2.9 model are non-linear, and we need to take the physical process of the model into account when interpreting the results from the EnKF.

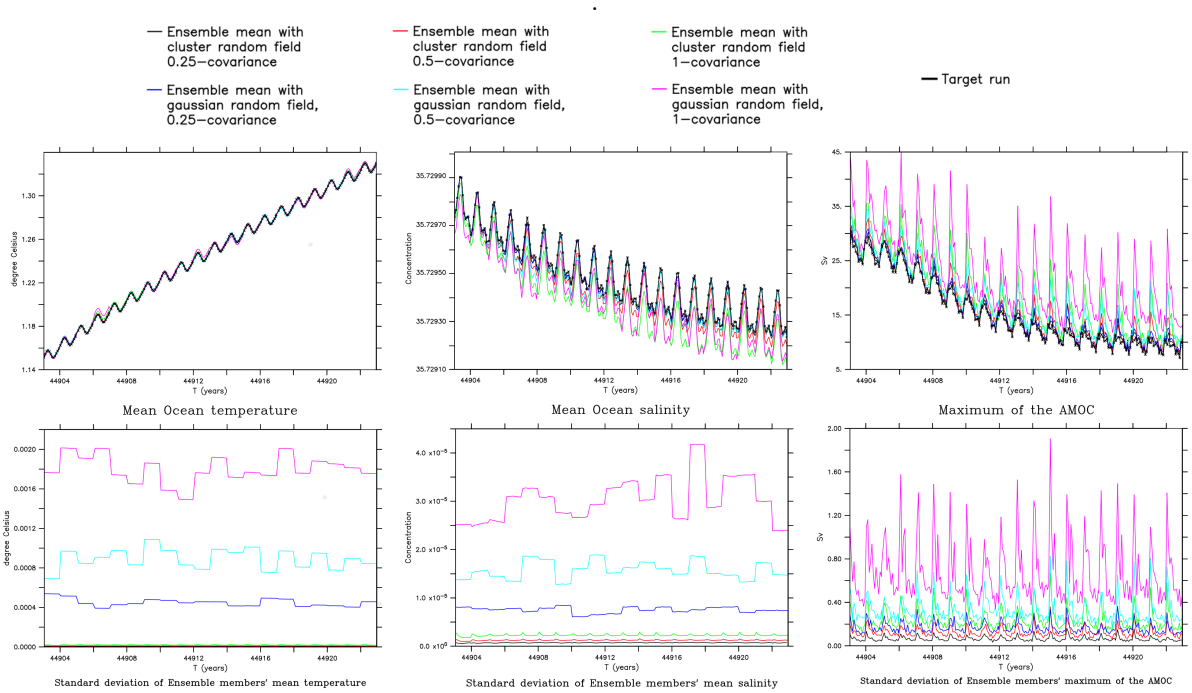

Figure 4.3: Plots of (respectively from left to right and top to bottom) the global mean temperature, global mean salinity, maximum MOC, and ensemble standard deviations of the same quantities. The lines represent different observation runs and ensemble means.

Biased observations and random field type

Our second objective was to study the ability of the EnKF to adjust its system state when the initial conditions appear to be inaccurate. For that we used in this experiment the two biased model runs constructed as observations. We again used the two different types of random fields - gaussian and cluster - to better understand how this difference could influence the recovery from initial conditions. The uncertainties on model and observations are given an even ratio and the amplitude of the random fields representing the uncertainties are here set up at $0.5\sigma_1$ and $0.5\sigma_2$.

In Fig 4.4 we can see that the EnKF readjusts rapidly to the 'error' in the initial temperatures. The difference in standard errors of the ensemble members between the cluster and gaussian random field give us the same information as previously, thus that the gaussian random field gives a larger ensemble 'spread' than the cluster random field.
We can however see that the salinity is not well constrained by the EnKF. We can first see that the

increase in the temperature causes a decrease in the salinity. However this change in the salinity of the ocean is not seen in the ensemble members, or at least is not capturing the salinity of the observations as well as the temperature. Once again this result can be due to uncertainties in the salinity being underestimated. The change in the salinity that we can see would therefore be mainly due to the change in the temperature in the form of changes in ice and snow melting for the most part.
On the other hand, the maximum MOC does not show important differences between the different runs (except for the target run which is represented here for a matter of comparison.

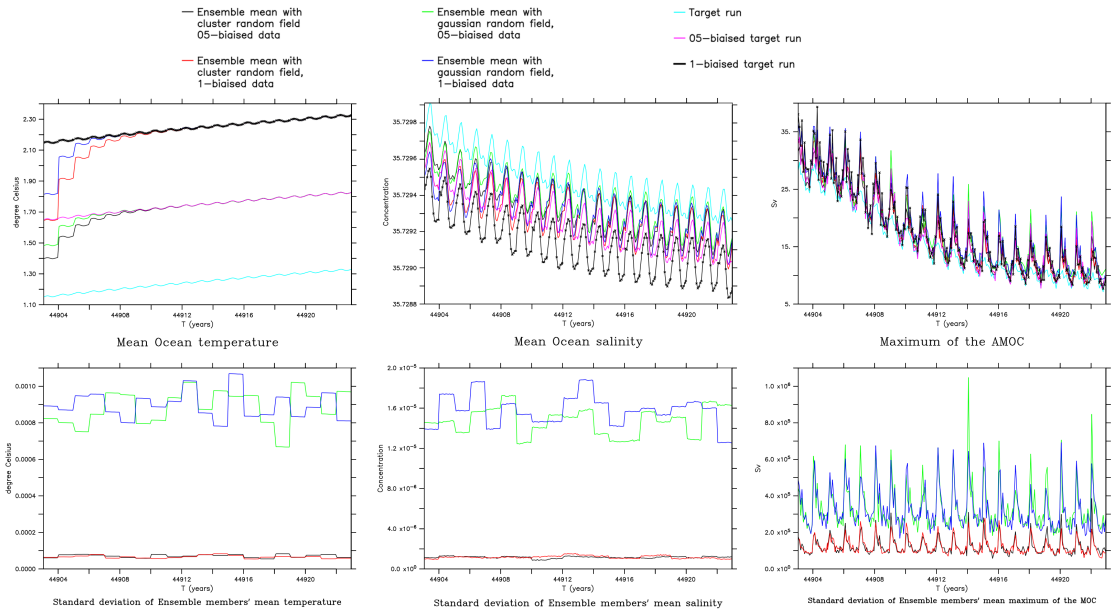

Figure 4.4: Plots of (respectively from left to right and top to bottom) the global mean temperature, global mean salinity, maximum MOC, and ensemble standard deviations of the same quantities. The lines represent different observation runs and ensemble means.

Biased observations and Q/R ratio

In the following experiment, we wanted to experiment the impact of different Q/R ratios on the EnKF while using the 0.5-biased observations as the target run. The observations being consistently biased, we expect that the ensemble readjusts to the biased state after a few iterations, or even several ones, depending on the suggested observations/model uncertainties.
We set up three different Q/R ratios : 5/1, 1/1, 1/5.

As we could have expected, increasing the ratio Q/R leads to an ensemble state that shifts faster to the observations state.
The results obtained for the ensemble means confirm the comments in the previous experiments.
Additionally we can remark that the standard deviation is higher when the trust is oriented to the model state. A reasonable explanation would be that in the 5/1 ratio run the transition period during which the model state is slowly converging to the observation state lasts longer than for the other runs. During this period, the stochastic component representing the observations uncertainties is multiplied by the

Kalman gain K which represents the 'direction' that the ensemble should follow, and this K is higher during this transition period.

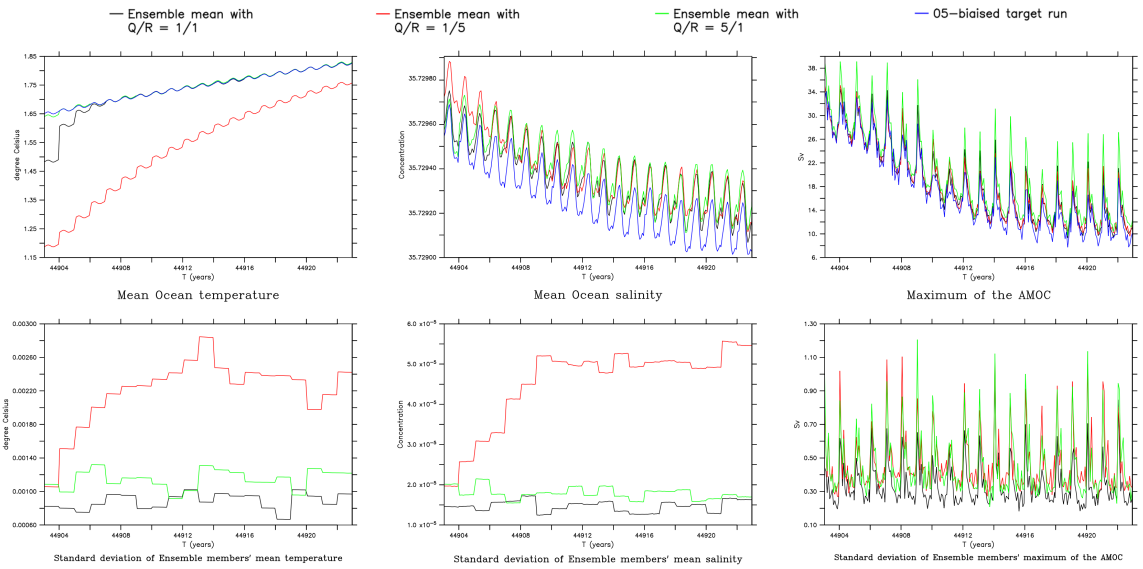

Figure 4.5: Plots of (respectively from left to right and top to bottom) the global mean temperature, global mean salinity, maximum MOC, and ensemble standard deviations of the same quantities. The lines represent different observation runs and ensemble means.

Biased observations and sparse in space variables

Our last objective our preliminary implementation of the EnKF of the UVic 2.9 model was to perform the UVic 2.9 model with sparse in space variables. Indeed, the real-world observations can not be obtained on all the points of the grid. Therefore, we performed the EnKF using only a small part of the grid points and using the biased data to test the robustness of the EnKF. The sparsity was performed such that we chose only 1 model grid point every 1,2,5 or 10 model grid points. (browsing the grid points in depth, then longitude and then latitude). Only these grid points were changed by the EnKF since our EnKF implementation with 'covariance arrays' do not enable intervariability between different variables. We also used the gaussian random fields with an even Q/R ratio and amplitudes the stochastic components of 0.5.

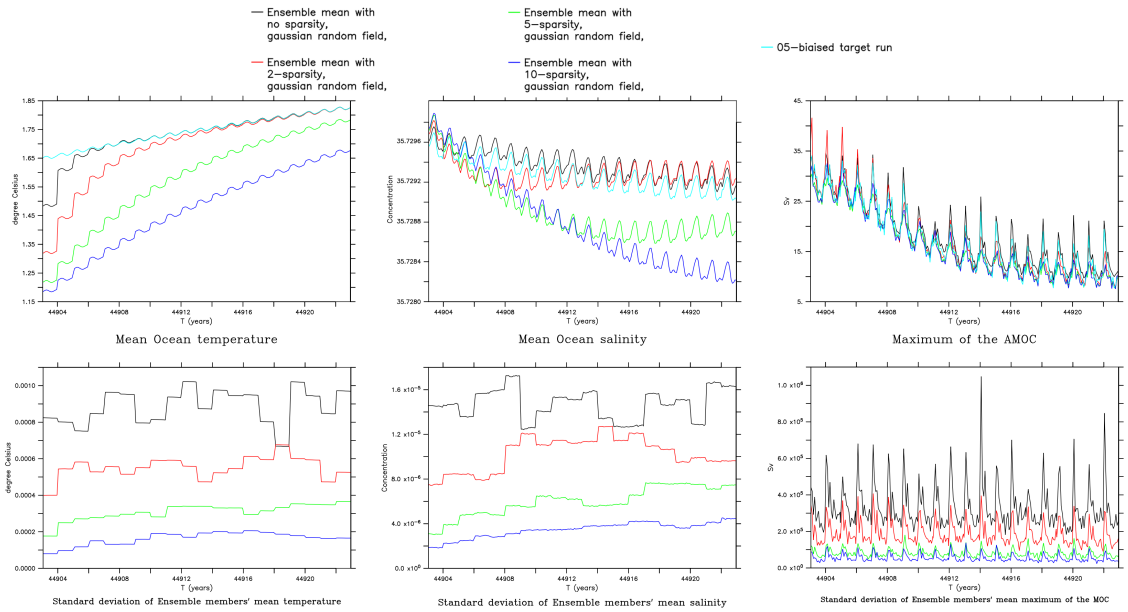

Figure 4.6: Plots of (respectively from left to right and top to bottom) the global mean temperature, global mean salinity, maximum MOC, and ensemble standard deviations of the same quantities. The lines represent different observation runs and ensemble means.

The results of this experiment are really in agreement with what we would have expected : the transition period lasts longer as the sparsity is higher. Moreover, the standard deviations decrease when the sparsity is higher, which is reasonable to support since the changes in temperature and salinity are made at less many grid points, inducing that the variation is less important.
However, the results of the salinity are less straightforward. Indeed, it seems that the more sparse the variables are, the less the global salinity would be.

4.4.4 Conclusions of this first application of EnKF to the UVic 2.9 model.

In the end, we have shown that the EnKF implemented this way is very robust, can adapt to inaccurate initial conditions and to sparse data.
However, we have seen that the salinity uncertainties was probably underestimated, and that other processes were therefore mainly impacting the global salinity of the ensemble members. The salinity is not an observation that we will use in the final implementation of the EnKF on the UVic 2.9 model, we have therefore decided to not investigate this point furthermore.

4.5 Adapting the EnKF for real-world observations

The next step in our project was to implement the EnKF on the UVic 2.9 model with real-world observations.

4.5.1 Obtaining observations and its uncertainties

The data from the observations is taken from different proxy records such as sediments cores, ice cores, or pollen cores. These proxy records use geochemistry observations to reconstruct the estimated temperatures, and radioactivity from the Last Deglaciation period and form a time series for these variables. However this data, being calculated from cores observations, has a lot of uncertainties concerning the age model. That means that the values are more expressing a smoothed result of the actual state system during a certain period and so defining a state variable at a certain time should carry uncertainties from the age model error. The uncertainties given are therefore given in terms of age model error in most of the data.

4.5.2 Linking the locations of the cores to the model grid

The next step to obtain useful information from the observations is to link the cores location to the adequate model grid points. It was also necessary to avoid the grid points that were not in the ocean mask of the model.
I used the standard distance in spherical coordinates, except that I chose the closest depth coordinate first and then to search for the closest grid point with this fixed depth. The reason for that is that the closest grid point was not always having the closest depth with the core location depth.

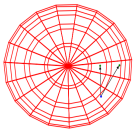

Figure 4.7: The points are at the same longitude (view from a pole)

4.5.3 Linking the estimated age of the observations with the temporality of the UVic 2.9 model

After joining the observations with the simulation in a geographical way, we need to use the observations at the appropriate time in such a way that it would fit with the runs of the UVic 2.9 model. We use two different methods, depending of how the observations are given in terms of uncertainties :

- For the variables with uncertainties on its values, we look into the **tavg.nc** file at the specific date of the observations. This is used only for the sea level observations.
- For the variables with uncertainties on the age associated with the observations, we look into the **tavg.nc** file but averaging over the year and a certain number of years before and after, depending on the age model error.

4.5.4 Setting up the EnKF

Now that the data and state variables have been gathered, we are using the EnKF to estimate the true state of the system given the results from the simulations and the observations. We use the state variables of the temperature, radiocarbon activity and sea level to constrain the input of freshwater in the Northern Atlantic Ocean. The location of this discharge is derived from the known location of the Laurentide Ice Sheet (above the 37°N latitude) [3]. A stochastic component on the freshwater input in this region is directly implemented in the UVic 2.9 model, in a matter that this stochastic component

could be set up in different ways. The first possibility was to use either a random walk, causing the freshwater input to be actually varying while the UVic 2.9 model was running. The other one was to use a normal variable that would remain the same during the whole run of the UVic 2.9 model. The second option was chosen for a sake of simplicity. Indeed, since the observations from freshwater input involved the sea level (integration of the freshwater input over time), having a fixed freshwater input during the run made the sea-level easier to be traced and to be compared with the observations.

4.5.5 Performing the EnKF for constraining the freshwater input

With the implementation as described before we ran the EnKF with observations from a previously simulated UVic 2.9 model run with a constant 0.1 Sv/year freshwater input.
The state variables were used such that observations were compared with the state variables extracted from the UVic 2.9 model at each EnKF time step, and assimilated with the observations to obtain the updated freshwater input that would be used during the next time step.

A first try

At this point, the EnKF was using x such that x consisted of all of the state variables used for observations and of the freshwater input. From this ensemble state we used all these same variables to be compared with the observations, except the freshwater input. The observations consisted of the same variables at the exception of the freshwater input.
More formally, we had :

- The state variables were : x =(**fwf**, **o_temp**, **a_temp dic**, **dic13**), where **fwf** is the freshwater input, **o_temp** is the ocean temperature, **a_temp** is the atmospheric temperature, **dic** is the dissolved inorganic carbon, and **dic13** is the dic which isotope is ^{13}C;
- The observations were : Hx =(**sl**, **o_temp**, **a_temp**, δ**13c**), where **sl** is the sea level and δ**13c** is calculated with **dic** and **dic13**.

Additional observations that we will use in the final project but that is not used here is the ^{14}C. The reason we did not use it is because we used a version of the UVic 2.9 that do not take into account all the carbon cycle processes for a matter of time. Indeed, the time needed to compute a year was divided by three this way.

The standard deviations were set up with the same method than in paragraph **4.4.2**. Moreover, as the number of variables in the EnKF was a lot more large than in our first implementation of the EnKF, we were able to use the "normal" EnKF, representing the full covariance matrix of the ensemble state.
With these settings, we first tried to perform the EnKF for a duration of 100 years with the time steps lasting 10 years.

However, the ensemble resulting from this implementation was very unstable and was diverging. A lot of these runs could not go to the end of the 100 years as the UVic 2.9 model crashed before the end because it could not allow negative freshwater input of more then 2 Sv.

Changing the method

From these failures, it was easy to understand that we needed to find a solution to this divergence.
First of all I recognized that the way I implemented the ensemble state variables in the EnKF was erroneous. Indeed, the only state variable that we wanted to influence is the freshwater input. We therefore changed the ensemble state to : $x = \mathbf{(fwf)}$, and kept the same observations.
Additionally, Dr. Schmittner pointed out that the impact of the input of freshwater was not instantaneous, and that it would be better to have a longer EnKF time step to have an ensemble state that gives a more meaningful response to the freshwater input.
We also constructed a new observation set by running a UVic 2.9 simulation with a varying freshwater input : the first 50 years saw the freshwater input linearly increase from 0 to 0.1 Sv, then stayed constant at 0.1 Sv for 250 years, and finally instantly dropped at 0 Sv and stayed null for the remaining 200 years. Finally , we used a Q/R ratio of 1/5 to represent as accurately as possible the uncertainties that we would have with the real-world measurements. The random noise for the freshwater input was set at 0.01 Sv per year.

We then performed the EnKF with these new settings and obtained the following results.

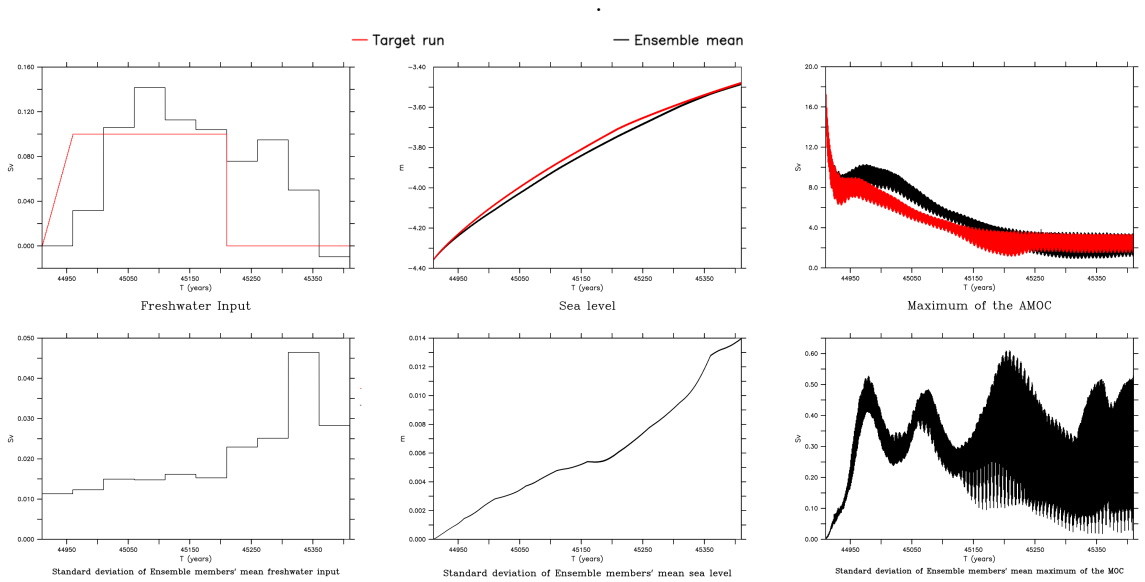

Figure 4.8: Plots of (respectively from left to right and top to bottom) the freshwater input per year, sea level, maximum MOC, and standard deviations of the same quantities. The red line represents the target run and the black line represents the ensemble mean.

In Fig 4.8, we can observe that the freshwater input is being accurately captured by the EnKF but with a delay. That delay can be explained by the delay induced in the slow process that leads the freshwater input to have an actual impact on the ensemble state. We can however observe that the sea level is constrained by the end of the period of the 500 years.
Moreover, we can observe a steady increase in the standard deviation, showing a possible filter divergence on the freshwater input. This issue will be ackowledged in the future experiments by implementing a

covariance inflation in the EnKF.

Additional information can be found in **Appendix C**, describing the evolution of the standard deviations of the observation variables from the ensemble members and standard errors of the same variables when compared to the observation model. The standard deviations and standard errors are found to be increasing along time. This behavior would not be very surprising since the freshwater input in the ensemble runs does not capture accurately the freshwater input used in the observation run. The impact of that on state variables should therefore logically increase as time passes. However, we can expect that on a longer time scale this standard error would stabilize around the standard deviations given for the observations of the real-world.

Using covariance inflation

Another lead to improve the EnKF was given by the covariance inflation process. Using an additive covariance component, the ensemble state subspace could be better assimilated in the EnKF.
We therefore performed a new run, adding this covariance inflation. The standard deviations were taken as identical to the standard deviations used for the observations and has been scaled such as we have an overall Q/R ratio of 1/5.
The results that we particularly wanted to look at was the behavior of the standard deviation of the freshwater input as we want that it stays stationnary. We also introduced a more noisy random noise for the freshwater input, the new random noise amplitude was set at 0.05 Sv per year. That change was implemented to test hwo the system would react to this high apmlitudes of random noise.
The results of this experiment are displayed below in Fig 4.9.

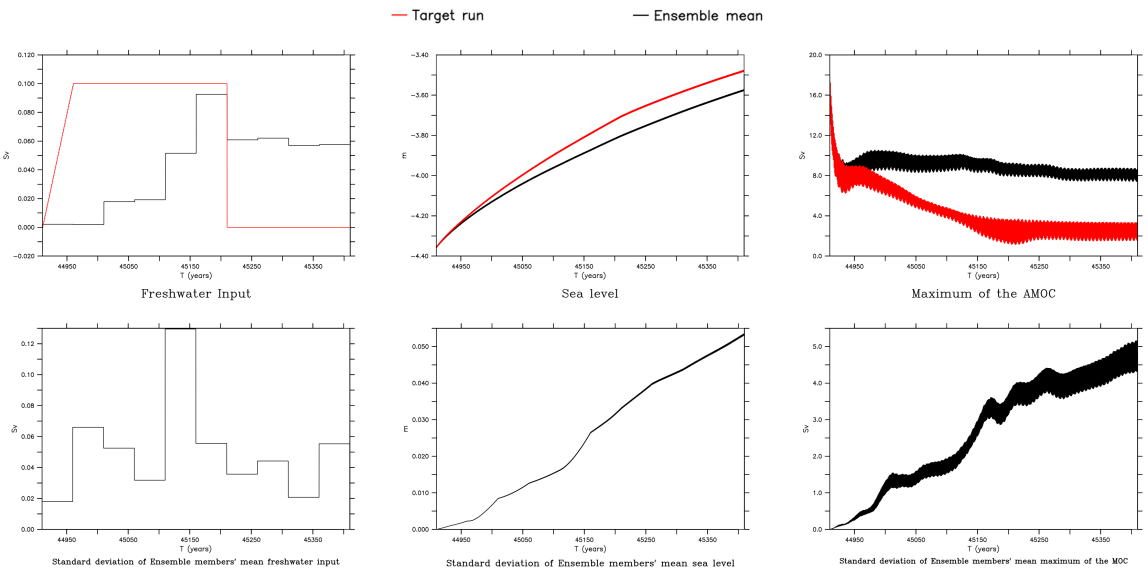

Figure 4.9: Plots of (respectively from left to right and top to bottom) the freshwater input per year, sea level, maximum MOC, and standard deviations of the same quantities. The red line represents the target run and the black line represents the ensemble mean.

In contrary to the previous experiment, this one does not show a strong reactivity to quicks changes in the freshwater forcing. This lack of reactivity is impacting the sea level and maximum MOC, which end up not being accurate with the observations.
However, a positive result from this experiment is given in ths stationnary behavior of the freshwater forcing standard deviation. Additionally, the standard deviation of the sea level, even though seeming to be steadily increasing, is not really surprising as the sea level change depends on the integration of the freshwater forcing. This behavior is therefore not surprising, so thatis the behavior of the standard deviations of the maximum MOC since this process is complex and nonlinear.

At the regards of these results, we can suggest that further experiments need to be done for estimating standard deviations that need to be used for the stochastic components in the EnKF to obtain optimal results.

4.6 Limitations and future research directions

These results are promising for a future experiment on observations from the UVic 2.9 model but on a longer period and adding the complete UVic carbon cycle.
However, additional work needs to be done to determine how to introduce the uncertainties from the real-world observations in the EnKF as well as properly setting up the initial conditions in the UVic 2.9 model.

Chapter 5

Conclusion

In this project, we have improved our understanding of the oceanographic context surrounding our study. This understanding has turned out to be crucial to understand the consequences of different physical processes on our results.
We have then implemented the EnKF on the Lorenz 96 model, and used the results from the implementation to have a first insight of the possible problems and their solutions in the future implementation of the ENKF on the climate model.
Finally, we have implemented the EnKF on the UVic 2.9 model. A first implementation was built to understand how adapt the EnKF to the structure of the UVic 2.9 model. After that, we constructed what is the final implementation of the EnKF for constraining the past ocean circulation with the UVic 2.9 model. This final implementation still is to be developed to be performed with real-world observations and on a long time scale.

Bibliography

[1] Gerrit Burgers, Peter Jan Van Leeuwen, and Geir Evensen, *On the analysis scheme in the ensemble kalman filter*, Monthly Weather Review **126** (1998).

[2] Jean-Paul Chilès and Pierre Delfiner, *Geostatistics: Modeling spatial uncertainty*, vol. 95, p. 695, 03 2000.

[3] Arthur Dyke, A. Moore, and Lennart Robertson, *Deglaciation of north america*, 01 2003.

[4] Jocelyne Erhel, Mestapha Oumouni, Géraldine Pichot, and Franck Schoefs, *Analysis of continuous spectral method for samplingstationary gaussian random fields*, https://hal.archives-ouvertes.fr/hal-02109037/document, 01 2019.

[5] Pieter Houtekamer and Herschel Mitchell, *Data assimilation using an ensemble kalman filter technique*, Monthly Weather Review - MON WEATHER REV **126** (1998).

[6] Alicia R. Karspeck, Steve Yeager, Gokhan Danabasoglu, Tim Hoar, Nancy Collins, Kevin Raeder, Jeffrey Anderson, and Joseph Tribbia, *An Ensemble Adjustment Kalman Filter for the CCSM4 Ocean Component*, Journal of Climate **26** (2013), no. 19, 7392–7413.

[7] Matthias Katzfuss, Jonathan Stroud, and Christopher Wikle, *Understanding the ensemble kalman filter*, The American Statistician **70** (2016).

[8] David Kelly, Andrew Majda, and Xin Tong, *Concrete ensemble kalman filters with rigorous catastrophic filter divergence*, Proceedings of the National Academy of Sciences **112** (2015), 201511063.

[9] Christian Lantuéjoul, *Geostatistical simulation. models and algorithms*, 01 2002.

[10] Hong Li, Eugenia Kalnay, and Takemasa Miyoshi, *Simultaneous estimation of covariance inflation and observation errors within an ensemble kalman filter*, Quarterly Journal of the Royal Meteorological Society **135** (2009), 523 – 533.

[11] Jean Lynch-Stieglitz, *The atlantic meridional overturning circulation and abrupt climate change*, Annual Review of Marine Science **9** (2017).

[12] Syukuro Manabe and Ronald Stouffer, *Simulation of abrupt climate change induced by freshwater input to the north atlantic ocean*, Nature **378** (1995), 165–167.

[13] Syukuro Manabe and Ronald Stouffer, *Coupled ocean-atmosphere model response to freshwater input: Comparison to younger dryas event*, Paleoceanography **12** (1997), 321–336.

[14] Juan Muglia, Luke Skinner, and Andreas Schmittner, *Weak overturning circulation and high southern ocean nutrient utilization maximized glacial ocean carbon*, Earth and Planetary Science Letters **496** (2018).

[15] W. Peltier and Guido Vettoretti, *Dansgaard-oeschger oscillations predicted in a comprehensive model of glacial climate: A "kicked" salt oscillator in the atlantic*, Geophysical Research Letters **41** (2014).

[16] Stefan Rahmstorf, *The thermohaline ocean circulation: A system with dangerous thresholds?*, Climatic Change **46** (2000), 247–256.

[17] Stefan Rahmstorf, *Ocean circulation and climate during the past 120,000 years*, Nature **419** (2002), 207–14.

[18] Asa Rennermalm, Eric Wood, Andrew Weaver, Michael Eby, and Stephen Déry, *Relative sensitivity of the atlantic meridional overturning circulation to river discharge into hudson bay and the arctic ocean*, Journal of Geophysical Research **112** (2007).

[19] Hans Renssen, Aurelien Mairesse, Hugues Goosse, Pierre Mathiot, Oliver Heiri, Didier Roche, Kerim Nisancioglu, and Paul Valdes, *Multiple causes of the younger dryas cold period*, Nature Geoscience **8** (2015).

[20] Raymond Schmitt, *The ocean's role in climate*, Oceanography **31** (2018).

[21] Andreas Schmittner and Eric Galbraith, *Glacial greenhouse-gas fluctuations controlled by ocean circulation changes*, Nature **456** (2008), 373–6.

[22] Andreas Schmittner and David Lund, *Early deglacial atlantic overturning decline and its role in atmospheric co2 rise inferred from carbon isotopes (δ13c)*, Climate of the Past **11** (2015), 135–152.

[23] M. Shinozuka and C.-M. Jan, *Digital simulation of random processes and its applications*, Journal of Sound and Vibration **25** (1972), no. 1, 111 – 128.

[24] Vincent Sircoulomb, Ghaleb Hoblos, Houcine CHAFOUK, and José Ragot, *Analysis and comparison of nonlinear filtering methods*, (2006).

[25] Istvan Szunyogh, Elizabeth Satterfield, José Aravéquia, Elana Fertig, Gyorgyi Gyarmati, Eugenia Kalnay, Brian Hunt, Eric Kostelich, David Kuhl, Edward Ott, and James Yorke, *The local ensemble transform kalman filter and its implementation on the ncep global model at the university of maryland.*, (2007).

[26] Simo Särkkä, *Bayesian filtering and smoothing*, 10 2013.

[27] Andrew J. Weaver et al., *The uvic earth system climate model: Model description, climatology, and applications to past, present and future climates*, Atmosphere-Ocean **39** (2001), no. 4, 361–428.

[28] Guocan Wu and Xiaogu Zheng, *The error covariance matrix inflation in ensemble kalman filter*, 02 2018.

[29] Yuhua Zhou, Dennis McLaughlin, and Dara Entekhabi, *Assessing the Performance of the Ensemble Kalman Filter for Land Surface Data Assimilation*, Monthly Weather Review **134** (2006), no. 8, 2128–2142.

Appendix A

Additional materiel

Lemma A.0.1 (Joint distribution of Gaussian random vectors)**.** *Let $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ be two randow gaussian vectors such that :*

$$x \sim \mathcal{N}(\mu, P),$$
$$y \mid x \sim \mathcal{N}(Hx + u, R)$$

Then we have the following joint distribution for x, y :

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ H\mu + u \end{pmatrix}, \begin{pmatrix} P & PH^T \\ HP & HPH^T + R \end{pmatrix} \right)$$

Proof. • $\mathbb{E}[y] = \mathbb{E}[\mathbb{E}[y \mid x]] = \mathbb{E}[Hx + u] = H\mu + u.$

- The covariance matrix is obtained with the following:

$$\begin{aligned} \mathbb{E}[(x - \mathbb{E}[x])(y - \mathbb{E}[y])^T &= \mathbb{E}[x\mathbb{E}[y|x]^T] - \mathbb{E}[x\mathbb{E}[y]^T] - \mathbb{E}[\mathbb{E}[x]y^T] + \mathbb{E}[x]\mathbb{E}[y] \\ &= \mathbb{E}[x(Hx + u)^T] - \mathbb{E}[x(H\mu + u)^T] - \mathbb{E}[\mu y^T] + \mu(H\mu + u)^T \\ &= \mathbb{E}[xx^T]H^T + \mu u^T - \mu\mu^T H^T - \mu u^T - \mu\mu^T H^T - \mu u^T + \mu\mu^T H^T + \mu u^T \\ &= (\mathbb{E}[xx^T] - \mathbb{E}[x]\mathbb{E}[x]^T) \\ &= PH^T \end{aligned}$$

We obtain that $\mathbb{E}[(y - \mathbb{E}[y])(x - \mathbb{E}[x])^T = HP$ in a very similar way.
Finally, since $\mathbb{E}[yy^T|x] = R + (Hx + u)(Hx + u)^T$:

$$\begin{aligned} \mathbb{E}[yy^T] = \mathbb{E}[\mathbb{E}[yy^T|x]] &= R + \mathbb{E}[(Hx + u)(Hx + u)^T] \\ &= R + H\mathbb{E}[xx^T]H^T + uu^T + H\mu u^T + u\mu^T H^T \\ &= R + HPH^T + (H\mu + u)(H\mu + u)^T \end{aligned}$$

Therefore we obtain that $Var(y) = R + HPH^T$.

□

Appendix B

Gaussian processes for simulating smoothed out random noise in the EnKF

The simulation of random functions over spatial field has been studied in the framework of the modeling space uncertainties. In our particular problem, the simulation of a gaussian random process aims to adapt our stochastic component and obtain one having more physical sense. Indeed, such mathematical and statistical tools are being widely used in geophysical problems. Even though geophysical problems often require more powerful tools - such as conditional simulation of Gaussian processes or Gaussian process regression (kriging) - the simulation of Gaussian process with an adapted covariance function will be a very good solution for our problem. Such work have been developed by Chilès & Delfiner [2], Lantuéjoul [9].

We aim in this section to introduce the reader to gaussian processes and get further into the details of how gaussian processes with a certain covariance function C can be simulated.

B.1 Simulation of Gaussian processes

Before introducing the simulation of Gaussian processes, we will recall (or introduce) the reader to some background from multivariate probability and stochastic processes in space.

B.1.1 Gaussian processes

Multivariate random variable.

A multivariate random variable (or **random vector**) is a generalization to n dimensions of a random variable. Whereas a random variable links an event ω to a real number, a multivariate random variable X links a event ω to a real vector.

Definition B.1.1. *Let $(\Omega, \mathcal{F}, P)$ be a probability space, Ω being the ensemble of the events, $\mathcal{F}$ a collection*

of the events of Ω*, and* P *being the probability measure.*
Then a ***random vector*** *from* Ω *to* $\mathbb{R}^n$ *is a function such that for each* $\omega \in \Omega$*,*

$$X(\omega) = (X_1(\omega), ..., X_n(\omega)) \in \mathbb{R}^n$$

Multivariate Gaussian random variable.

A **Gaussian random vector** is a random vector such that any linear combination of its random components is a gaussian variable.

Definition B.1.2. *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. Then a random vector from* Ω *to* $\mathbb{R}^n$ *is a* ***Gaussian random vector*** *if and only if for all* $(a_1, ..., a_n) \in \mathbb{R}^n$*,*

$$a_1 X_1 + ... + a_n X_n$$

is a gaussian variable

We then note $X \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, where $\mu = (\mu_1, ..., \mu_n)$ is the mean of the vector X and $\mathbf{\Sigma}$ is the covariance matrix.

Stochastic process.

A **stochastic process** is a collection of random variables defined on a same **probability space** $(\Omega, \mathcal{N}, F)$, taking values in the same **state space** E and indexed on an **index set** T, that is a stochastic process can be written as :

$$\{X_t\}_{t \in T} \in E^T$$

where each random variables X_t is defined on $(\Omega, \mathcal{N}, F)$ and takes values in E :

$$X_t(w) \in E$$

In our case, the **index set** is $\mathbb{R}^3$, reprensenting the Earth ocean and atmosphere, and the **state space** is $\mathbb{R}$, since the state variables are measured as real numbers. In the case of the representation of a spatial domain, the term **random field** is more often used.

Gaussian random field.

A **Gaussian random field** (or Gaussian field) is random field such that any finite sub-collection of the random field is a Gaussian vector.

Definition B.1.3. *Let* $(\Omega, \mathcal{F}, P)$ *be a probability space. Then a random field defined on* $(\Omega, \mathcal{F}, P)$ *and taking values in* E *is a* ***Gaussian field*** *if and only if for all subset* $\{t_1, ..., t_k\} \in T$*,*

$$\{X_{t_1}, ..., X_{t_n}\}$$

is a Gaussian vector

A key fact is that Gaussian fields are completely defined by their first two moments - the mean function and the covariance function - meaning that the spatial distribution of a Gaussian field can be fully determined by its mean and covariance function.

B.1.2 Choosing a mean and a covariance function

Now that we know that Gaussian fields distributions are only dependent on their mean and covariance functions, we want to study more specifically what type of mean function and covariance function we would use in the scope of adding a smoothed out random noise over the UVic 2.9 model grid.

Strict stationarity.

A first import fact is that we want to simulate a white noise over the model grid, meaning that the mean of the Gaussian field needs to be null over the whole grid. Moreover, we also want the covariance function to be invariant on translation and isotrope, meaning that the covariance only depend on the distance between two points.
That leads us impose a constraint of strict stationnarity on the Gaussian field.

Definition B.1.4. *A stochastic process $\{X_t\}_{t\in T}$ is said to be **strictly stationary** if for all $n \leq 1$ and for all $(t_1, ..., t_n) \in T^n$:*

$$\mathbb{P}(X_{t_1} \in A_1, ..., X_{t_n} \in A_n) = \mathbb{P}(X_{t_1+\tau} \in A_1, ..., X_{t_n+\tau} \in A_n) \quad ; \quad \forall \tau \in T \ , \ \forall A_1, ..., A_n \in \mathcal{F}$$

This is the strongest definition of stationarity for stochastic processes although there exist weaker definitions (2^{nd} order stationarity, weak stationarity, ...).

With a strict stationary random field, we can define the covariance function between two points of T by the distance between these two points. Defining the covariance function between two points $t_1, t_2 \in T$ as :

$$C(t_1, t_2) = Cov(X_{t_1}, X_{t_2})$$

we obtain in the case of strict stationarity that the covariance is entirely defined by the distance between t_1 and t_2 :

$$C(t_1 - t_2) := C(t_1 - t_2, 0) = C(t_1, t_2)$$

Matérn covariance functions.

Now that we have developed a frame for these gaussian fields, we want to define the covariance functions that we will use in our wrapper code.

The Matérn covariance function is a type of covariance function that is widely used in the domain of geostatistics and spatial statistics and when using Gaussian fields.
It is defined as the following :

$$C(h) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{h}{\rho} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{h}{\rho} \right) \tag{B.1}$$

where h is the distance, Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and ρ and ν are non-negative parameters of the covariance.

Since we will use the spectral density of this covariance function for the simulation of Gaussian fields, we will define this spectral density on $\mathbb{R}^n$ as the following :

$$S(f) = \frac{2^n \pi^{\frac{n}{2}} \Gamma(\nu + \frac{n}{2})(2\nu)^\nu}{\Gamma(\nu)\rho^{2\nu}} \left(\frac{2\nu}{\rho^2} + 4\pi^2 f^2 \right)^{-\left(\nu + \frac{n}{2}\right)} \tag{B.2}$$

where f is the frequency.

In the Matérn covariance function, the parameters ν the regularity of the covariance function in the neighboorhood of 0, whereas ρ controls the range of the covariance function. We will present some specific examples of covariance functions and their representation.

Spherical	$C(h) = \left(1 - \frac{3}{2}\frac{\lvert h\rvert}{a} + \frac{1}{2}\frac{\lvert h\rvert^3}{a^3}\right) 1_{\lvert h\rvert \leq a}$
Exponential	$C(h) = \exp\left\{-\frac{\lvert h\rvert}{a}\right\}$
Stable	$C(h) = \exp\left\{-\sqrt{\frac{\lvert h\rvert}{a}}\right\}$
Hyperbolic	$C(h) = \frac{1}{1 + \frac{\lvert h\rvert}{a}}$
Gaussian	$C(h) = \exp\left\{-\frac{\lvert h\rvert^2}{a^2}\right\}$
Cardinal sine	$C(h) = \frac{\sin\left\{\frac{\lvert h\rvert}{a}\right\}}{\frac{\lvert h\rvert}{a}}$

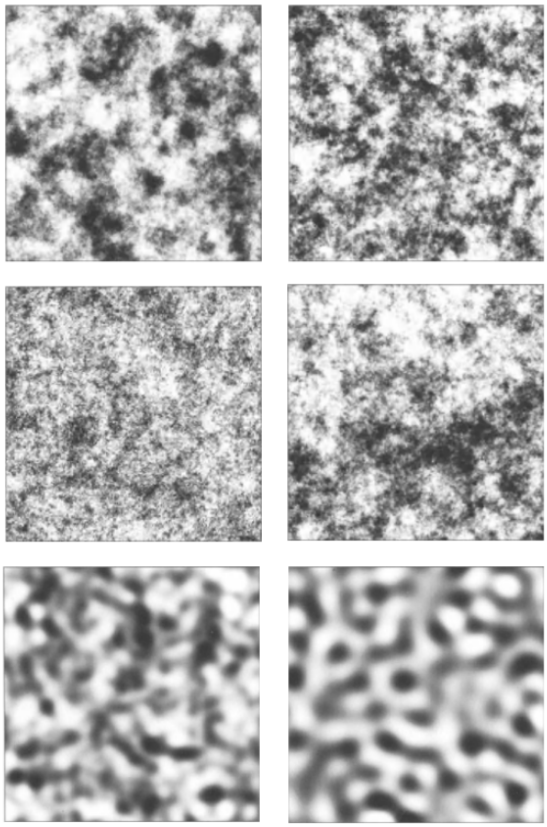

Figure B.1: Realizations of Gaussian fields with different covariance functions. From top to bottom and left to right : spherical , exponential, stable, hyperbolic, gaussian et sinc covariances. Source : Lantuéjoul, 2002

From these examples, the exponential covariance and the gaussian covariance are derived from the Matérn covariance function (with respectively $\nu = 1/2$ and $\nu = +\infty$).
We decide to use the Matérn covariance function since its spectral representation is given explixitly. Moreover, we choose to use the exponential covariance since we want to obtain a shaped enough random field that could capture the essential idea of the white noise, being the spatial heterogeneity of its realization.

B.1.3 Simulating a Gaussian process with covariance function C

That being said we now want to simulate a Gaussian field of zero mean with exponential covariance function.

Fourier transform and probability measure.

Our first step in the simulation of these Gaussian fields is to derive a probability measure from the covariance function.

Theorem 2 (Bochner Theorem)**.** *For any continuous positive-definite complex function f, such that f(0)=1, there exists a unique probability measure $\mathbb{P}_f$ such that*

$$f(h) = e^{i<h,u>} d\mathbb{P}_f(u)$$

ie, f is the Fourier transform of a unique probability measure $\mathbb{P}_f$.

We obtain straightforwardly that C, our exponential covariance, is a positive-definite complex function and we normalize C such that $C(0) = 1$.

According to the Bochner Therorem, we therefore have that there is a unique probability measure $\mathbb{P}_f$ such that C is the Fourier transform of $\mathbb{P}_f$.

Spectral method.

The spectral method for simulating a Gaussian random field from its covariance function relies on the previously mentionned probability measure $\mathbb{P}_f$.

The spectral method, developed by Shinozuka and Jan [23], tells us that for V a random variable with distribution $\mathbb{P}_f$ and U a random variable of uniform law on $[0, 1]$, the random function :

$$Y(t) = \sqrt{2}\cos(< V, x > +2\pi U) \tag{B.3}$$

is a normalized Gaussian field with covariance C.

B.2 Algorithm for simulating Gaussian processes

The development of the code for simulating a Gaussian field is done in 4 phases (Ehrel et al. [4]):

1. Constructing a grid of complex frequencies $\{f_{i,j,k}\}_{i,j,k=0}^{N_1,N_2,N_3}$. Sample $\mathbb{P}_f$ on this grid.
2. Generate a collection of $2N_1 \times 2N_2 \times 2N_3$ Gaussian complex variables : $\{\xi_{i,j,k}\}_{i,j,k=0}^{N_1,N_2,N_3}$
3. Compute X the collection of complex random variables such that :

$$X_{i,j,k} = \xi_{i,j,k}\sqrt{\Delta p_1 \Delta p_2 \Delta p_3 \mathbb{P}_f(p_{i,j,k})} \ , \quad i = 0, ..., N_1 \ , \ j = 0, ..., N_2 \ , \ k = 0, ..., N_3$$

$$X_{i,j,k} = \xi_{i,j,k}\sqrt{\Delta p_1 \Delta p_2 \Delta p_3 \mathbb{P}_f(p_{2N_1-i,2N_2-j,2N_3-k})} \ , \quad i = N_1+1, ..., 2N_1+1 \ , \ j = N_2+1, ..., 2N_2-1$$
$$k = N_3+1, ..., 2N_3-1$$

4. Finally perform the 3D FFT of X and obtain a complex random field over our the grid. The real, or imaginary part, of this random field yields a Gaussian field with covariance function C

Appendix C

Additional figures

C.1 Global snow and ice volume changes and net precipitation changes

Following the observations from pararagraph **4.4.3**, specifically the runs presented in the "**Cluster vs Gaussian**" part, we present here the evolution of global ice volume, global snow volume, and average net precipitation (precipitation - evaporation) over the ocean on the following figure.

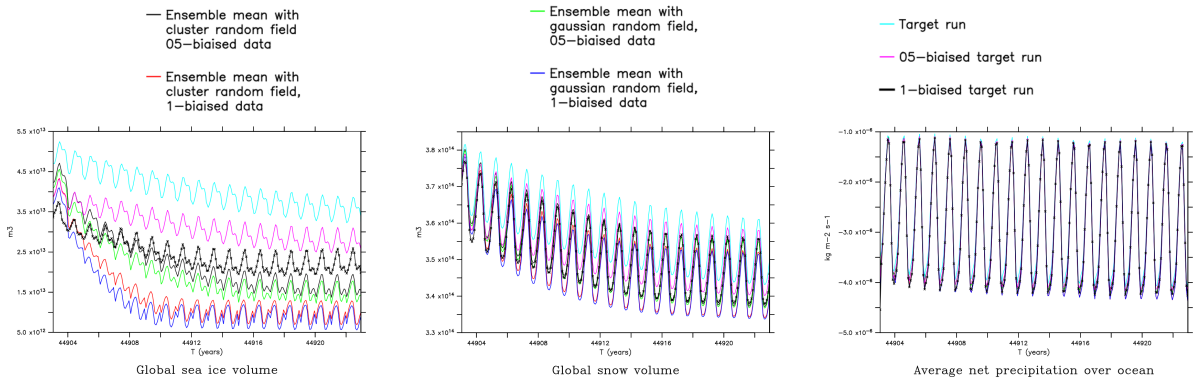

Figure C.1: Plots of (respectively from left to right and top to bottom) the global ice volume, global snow volume, and average net precipitation over the ocean. The lines represent different observation runs and ensemble means.

C.2 Standard deviation and standard errors on the observations' locations

The following figures C.2 and C.3 show the ensemble standard deviations for the temperature, and DIC values, averaged over the locations of the same observations. These values are estimated for the two EnKF runs performed, and described in the paragraph **4.5**.

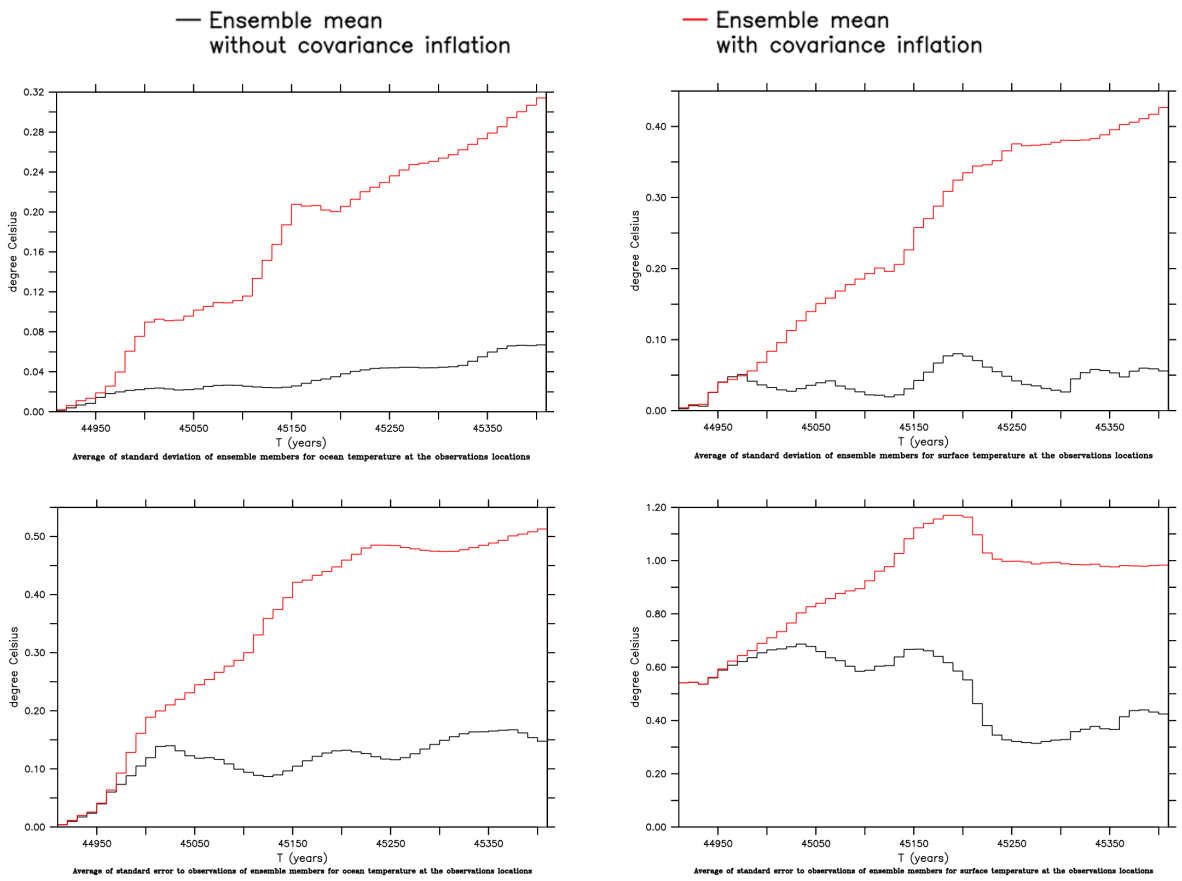

Figure C.2: Plots of (respectively from left to right and top to bottom) the ensemble standard deviations of ocean and surface temperature, and ensemble standard error of the same quantities with respect to the observation model. The lines represent different ensemble means.

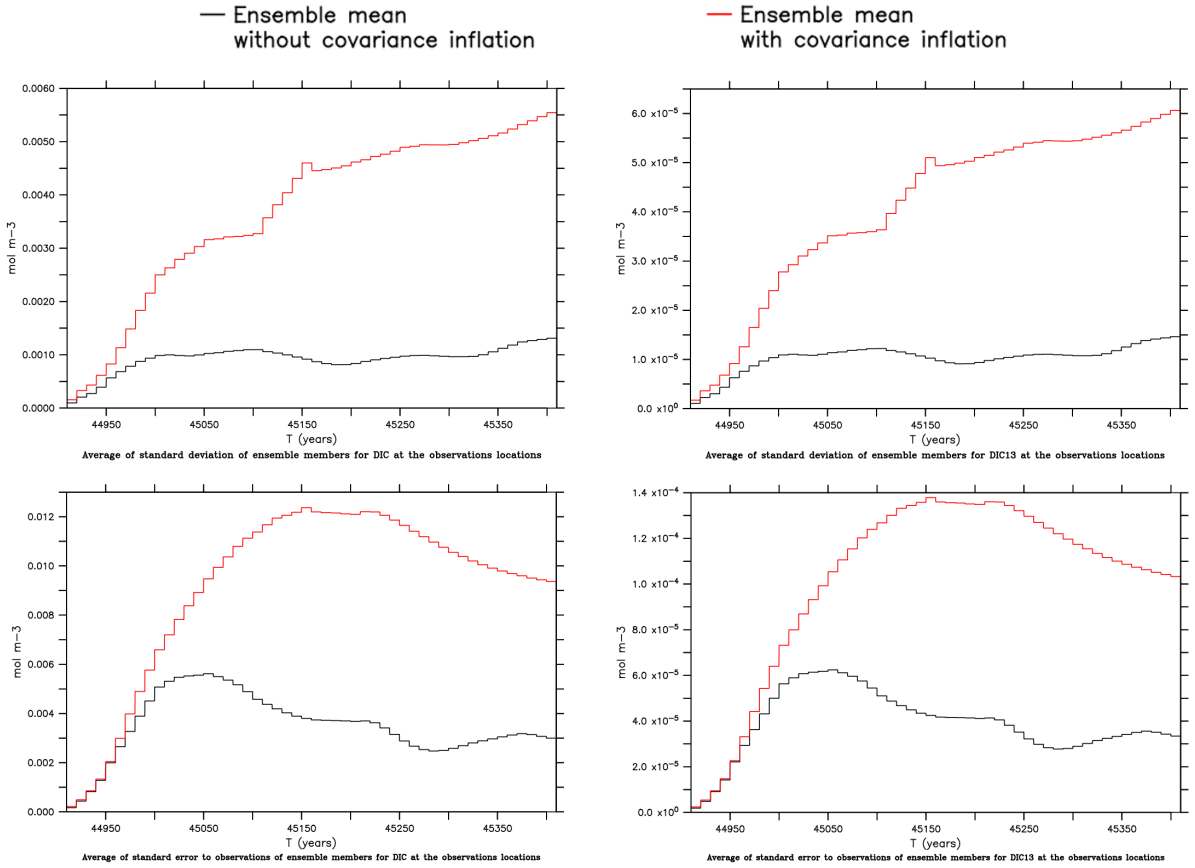

Figure C.3: Plots of (respectively from left to right and top to bottom) the ensemble standard deviations of DIC and DIC13 concentrations, and ensemble standard error of DIC and DIC13 with respect to the observation model. The lines represent different ensemble means.

The similar behavior of DIC and DIC13 can supposedly be caused by the fact that the UVic 2.9 model run that we have used here does not modelize te whole carbon cycle. Using a more complete carbon cycle in the UVic 2.9 model may give results more contrasted for DIC and DIC13.