HORIZON 2020

IMMERSE

(Grant Agreement 821926)

Improving Models for Marine EnviRonment

SErvices Deliverable D7.2 ***Public Report***

H2020-IMMERSE
Deliverable D7.2
Ensemble quantification of short-term predictability of ocean fine-scale dynamics

S. LEROUX[1], J.-M. BRANKART[2],
A. ALBERT[1], J.-M. MOLINES[2], L. BRODEAU[1],
T. PENDUFF[2], J. LE SOMMER[2], P. BRASSEUR[2].

(revised version after internal review, February 12, 2021)

PURPOSE

Deliverable D7.2 (lead Ocean Next): Scientific report on task WP7.1 of H2020-IMMERSE (2019-2020).

CONTENTS

1 *Ocean Next, Grenoble, France*
2 *IGE, MEOM group, Grenoble, France*

D7.2 – EXECUTIVE SUMMARY

The general objective of this task is to quantify how much a high-resolution NEMO modelling system is able to correctly retain and propagate the information in the initial condition, acquired from observations through the assimilation system, during a short and medium range forecast. This question is particularly relevant for ocean dynamics at small scales (< 30 km), where submesoscale dynamics generate a very fast evolution of ocean properties. Relatively little is known indeed about the predictability properties of a high resolution model, and hence about the accuracy and resolution that is needed from the observation system to produce the targeted forecast skill. This task is thus also a contribution to the general ambition of improving the articulation between high-resolution ocean observing systems and CMEMS forecasting models.

For that purpose, a kilometric-scale regional configuration of NEMO for the Western Mediterranean (MEDWEST60, at 1/60° horizontal resolution) has been developed. It is defined as a subregion of a larger North Atlantic configuration at same resolution (eNATL60), which provides the boundary conditions. This deterministic model has then been transformed into a probabilistic model by introducing an innovative stochastic parameterization of model uncertainties resulting from unresolved processes. The purpose is primarily to generate ensembles of initial conditions to be used in the predictability studies, but it has also been applied to assess the possible impact of irreducible model uncertainties on the skill of the forecast.

With this model configuration, 20-member and 2-month ensemble experiments have been performed, first with the stochastic model for two levels of model uncertainty, and then with the deterministic model from perturbed initial conditions. In all experiments, the spread of the ensemble emerges from the small scales (10 km wavelegnth) to progressively upscale to the largest structures. After two months, the ensemble variance has saturated over most of the spectrum (except the largest scales), whereas the small scales (< 30 km) are fully decorrelated between different members. For these scales, these ensemble simulations are thus appropriate to provide a statistical description of the dependence between initial accuracy and forecast accuracy over the full range of potentially useful forecast time lags (typically, between 1 and 20 days).

From these experiments, predictability has then been statistically quantified using a cross-validation algorithm (i.e. using alternatively each ensemble member as a reference truth and the remaining 19 members as ensemble forecast) together with a specific score to characterize the initial and forecast accuracy. From the joint distribution of initial and final scores, it is then possible to diagnose the probability distribution of the forecast score given the initial score, or reciprocally to derive conditions on the initial accuracy to obtain a target forecast skill. Although any specific score of practical significance could have been used, we focused here on simple and generic scores describing the misfit between ensemble members in terms of overall accuracy (CRPS score) or in terms of geographical position of the ocean structures (location score).

For example, our results show that, for our particular region and period of interest, the initial location accuracy required (necessary condition) with a perfect model (deterministic operator) to obtain a forecast location accuracy of 10 km with a 95% confidence is about 8 km for a 1-day forecast, 6 km for a 2-day forecast, 4 km for a 5-day forecast, 1.5 km for a 10-day forecast, and that this target is unreachable for a 15-day or a 20-day forecast (more precisely, in these two cases, the required initial accuracy would be irrealistically small and was not included in our sample). With model uncertainties (stochastic operator), the requirement on the initial condition can be even more stringent, especially for a short-range and high-accuracy forecast. These requirements on the initial condition can then be directly translated into necessary conditions on the design of the ocean observing system, in terms of accuracy and resolution, if a given forecast accuracy is to be expected.

More generally, this study suggests that an ensemble forecasting framework should become an important component of CMEMS systems to provide a systematic statistical quantification of the relation between the system operational target (a useful forecast skill) and the available assets : the observation systems, with their expected resolution and accuracy, and the modelling tools, with their target resolution and associated irreducible uncertainties.

1 INTRODUCTION

The Copernicus Marine Environment Monitoring Service (CMEMS) is dedicated to provide regular analyses and forecasts of the state of the ocean, to serve a wide range of marine scientific and operational applications. Most CMEMS systems rely on the NEMO modelling framework to embed state-of-the-art representations of the various dynamical components of the ocean, with the goal to improve the accuracy and the resolution of the products. However, with the increase of the complexity and resolution of the model, new questions arise regarding the predictability of the system. To what extent is it possible to forecast the very fine scales targeted by the next generation of CMEMS systems using the NEMO dynamical core? How is this forecast sensitive to initial errors or to possible shortcomings or approximations in the model dynamics? These questions are important for CMEMS because they can help rationalizing expectations from the next systems and thus help driving future developments.

Historically, the question of the predictability of dynamical systems has been addressed by considering only the irreducible sources of error, which result from intrinsic model instability combined to inevitable small initial errors. In a deterministic framework, modelling errors can indeed be excluded from the analysis because they can be reduced by additional modelling efforts, so that they do not represent a theoretical limitation to predictability. There is a long history of studies along this line, starting with Lyapunov (1992), who suggested looking for the fastest-growing unstable modes (Lyapunov vectors) and their associated e-fording timescales (Lyapunov exponents). This was extended in meteorology to describe the largest error growth over a finite time (with singular vectors, Lorenz, 1965; Lacarra and Talagrand, 1988; Diaconescu and Laprise, 2012), before it was recognized that linear instability studies were quite often not sufficient to provide a correct picture of the predictability patterns, even for quite short time lags. Nonlinear model integrations are needed to allow the fast instabilities to saturate, and reveal the patterns that really matter over a given forecast time. For this reason, the bred vectors (Toth and Kalnay, 1993; Kalnay, 2003) have been introduced as a practical way to identify the most relevant perturbations to initialize ensemble forecasting systems. In the meantime, ensemble forecast simulations, explicitly performed with the full nonlinear model, have indeed become the standard approach to investigate predictability (e.g. Brasseur et al. , 1996; Palmer and Hagedorn, 2006; Hawkins et al., 2016). Performing an ensemble forecast amounts to propagating a probability distribution in time, which includes the possibility of a non-deterministic model. In this framework, it is thus possible to go beyond the historical assumption that predictability is mainly limited by unstable and chaotic behaviours, and to include the possibility that model uncertainties can be an essential limiting factor to forecast accuracy.

In the last two decades, indeed, more an more studies have suggested that uncertainties are intrinsic to atmosphere and ocean models, as long as they do not resolve the full diversity of processes and scales at work in the system (e.g. Palmer et al. , 2005; Frederiksen et al., 2012; Brankart et al., 2015). Non-deterministic modelling frameworks have been shown for instance very helpful to improve the accuracy of medium-range weather forecast (Buizza et al. , 1999; Leutbecher et al., 2017), to enhance their economical value (Palmer, 2002), to alleviate persistent biases in model simulations (Berner et al., 2012; Juricke et al. , 2013; Brankart, 2013; Williams et al., 2016), and to explain misfit between model and observations in data assimilation systems (e.g. Evensen, 1994; Sakov et al., 2012; Candille et al., 2015). In any case, whether the system can be thought as fundamentally deterministic or not, it is not dubious that, in practice, all CMEMS systems involve substantial modelling uncertainties. What matters to the application is then the possibility to produce a valuable forecast with the model that is presently used, which may be quite different from what is obtained by only considering the unstable or chaotic behaviour of a perfect deterministic model.

For these reasons, our target in this study is to evaluate the predictability of the fine scales in a typical high-resolution NEMO-based CMEMS model, by including the effect of initial uncertainties and model uncertainties, either separately or together. In both cases, it will be assumed that they cannot be made arbitrarily small in CMEMS systems: initial uncertainties because observation resources are limited, and model uncertainties because model resources are limited. Nevertheless, these finite-size uncertainties may have very different origin and may display very different shapes in space and time, so that an assumption is still needed to simplify the problem. In this study, this simplification will be obtained by considering one generic type of model uncertainty that primarily affects the small scales of the system. By tuning the amplitude of the perturbations, we can then simulate different levels of model accuracy, and generate ensemble initial conditions with different levels of spread. With this assumption, we can then compute the forecast accuracy that is obtained for different combinations and levels of initial and model uncertainties.

Reciprocally, we can then expect that this set of experiments can provide insight on the level of initial and model uncertainties that is required to obtain a given forecast accuracy. This might give an idea of the relative importance of the initial and model uncertainties to obtain an accurate forecast of the small scales, and thus the relative weight of the observation and model constraint in the quality of the CMEMS products, and maybe help us understand the level of initial and model accuracy required to produce a useful forecast of the small scales targeted in the future CMEMS systems. However, it will be important to remember that these conclusions will depend on the assumption made to simulate uncertainties in the system. Although generic, and designed to trigger perturbations in the small scales, they are still an idealization and cannot be expected to summarize the full diversity of uncertainties propagating in real operational systems.

THE PLAN OF THIS REPORT IS AS FOLLOWS :

- In section 2, we present the kilometric-scale NEMO regional configuration that will be used throughout this study, emphasizing on the model representation of the fine-scale spectrum.
- In section 3, we introduce our assumption about model uncertainties, which will be used to generate various levels of initial spread and model accuracy in the ensemble experiments.
- In section 4, we describe the ensemble experiments that have performed to evaluate the predictability of the fine scales, with a focus on the effect of the model uncertainties in the behaviour of the simulations.
- In section 5, we apply different sorts of metrics (spectral analysis, probability scores, location errors) to characterize the dependence of the forecast accuracy to initial and model uncertainties.
- We summarize the outcomes of this study in section 6.

2 MEDWEST60: A KILOMETRIC-SCALE REGIONAL MODEL

A strong basis for the present work is the already-existing kilometric-scale simulation eNATL60 performed by Ocean Next and IGE recently over the North Atlantic area (Brodeau et al., 2020). This simulation was designed to model as accurately as possible the surface signature of oceanic motions of scales down to 15km, which is, for example, the expected resolution of the future altimetry mission SWOT (Surface Ocean and Water Topography, Fu and Ferrari (2008) Durand et al. (2010)). It provides a unique scientific material at this resolution to study fine-scale processes (<200 km) and cross-scale interactions in the ocean, from submesoscale processes to basin-scale features. The cost in CPU, memory and storage for such a simulation is however too high to consider performing several sets of ensemble experiments over the entire North Atlantic domain. Instead, we designed in this study a new regional configuration, following as much as possible the eNATL60 setup, but covering a smaller area, and we use the eNATL60 simulation for hourly boundary conditions. The targeted region was selected over the Western Mediterranean Sea, as this area is included in the eNATL60 domain, and minimizes the length of the open lateral boundaries given the geography of the basin (the western lateral boundary is set at the Gibraltar Strait, and the eastern lateral boundary along a line going from north to south through Corsica and Sardinia, see Figure 2.1). The full domain covers 1200 km $\times$ 1100 km, from 35.1°N to 44.4°N in latitude and from 5.7°W to 9.5°E in longitude. Note also that this region includes the crossover point of the future SWOT mission south of the Balearic Islands in the high-sampling phase, and that kilometric-scale ensemble forecast simulations could also provide interesting scientific materials for future analyses in this context.

The MEDWEST60 configuration includes tides and is forced at the western and eastern boundaries with hourly outputs from the reference simulation eNATL60-with-tides (i.e. "eNATL60-TCLB02" in the eNATL60 nomenclature). By design, all technical and parameter choices for the regional configuration MEDWEST60 were made with the idea to remain as close as possible from the reference simulation eNATL60-LBT02. In particular, we use strictly the same horizontal and vertical grids as the reference simulation, meaning that there is no need for spatial interpolation of the lateral boundary conditions from the reference simulation.

As a result, the characteristics of the MEDWEST60 configuration are:

- Numerical code: NEMO 3.6 + XIOS-2.0 (`https://www.nemo-ocean.eu/`)
- Horizontal resolution: 1/60°,
- Grid size: 883 x 803 in the horizontal (1.20 km <Δx<1.55 km),
- Vertical grid: 212 levels along the vertical, those levels are defined(1) exactly as in eNATL60-LBT02 but only 212 levels are actually needed to include the deepest points in the Western Mediterranean region (i.e 3217 m at the deepest), while 300 levels were used in eNATL60 to cover the depth range in the North Atlantic basin.
- Atmospheric forcing: 3-hourly ERA-interim (ECMWF),
- Lateral boundary conditions at the coast: no slip,
- Lateral boundary conditions: hourly outputs from the reference simulation eNATL60-TCLB02 (which explicitly includes tides). The Flow Relaxation Scheme ("frs") is used for baroclinic velocities and active tracers (simple relaxation of the model fields to externally-specified values over a 12 grid point zone next to the edge of the model domain). The "Flather" radiation scheme is used for

(1) The following discretisation is applied to the first 20 meters below the surface: 0.48 m, 1.56 m, 2.79 m, 4.19 m, 5.74 m, 7.45 m, 9.32 m, 11.35 m, 13.54 m, 15.89 m, 18.40 m, 21.07 m.

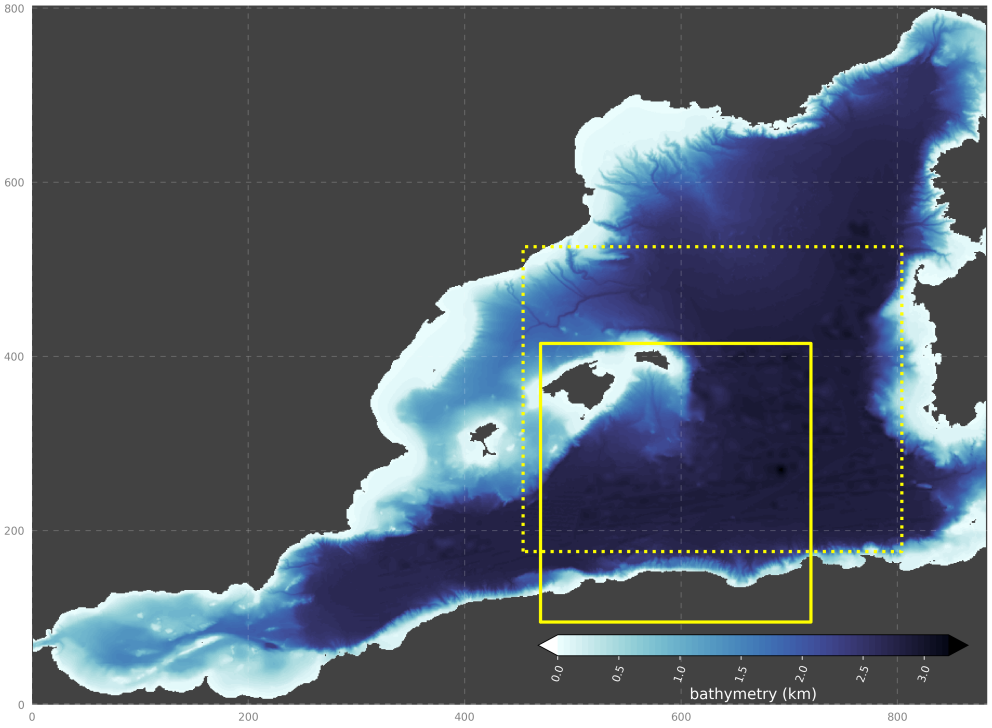


Figure 2.1: Domain and bathymetry (in km) of the MEDWEST60 regional configuration. The full domain covers 883 x 803 grid points in the horizontal, representing 1200 km x 1100 km, from 35.1°N to 44.4°N in latitude and from 5.7°W to 9.5°E in longitude. The two yellow boxes show the subregions over which spectral analysis is performed (dotted line) in the following, and over which zoomed snapshots will be plotted (solid line).

sea-surface height and barotropic velocities (a radiation condition is applied on the normal depth-mean transport across the open boundary).

Doing so, we are able to start the MEDWEST60 regional configuration directly from initial conditions stored from eNATL60-LBT02 (i.e. from NEMO restart files) without the need for a spinup of several months/years as when starting from climatological conditions.

In summary, the only differences between MEDWEST60 and eNATL60-LB02 are:

- the smaller regional domain,
- the lateral boundary conditions,
- there is no additional tidal harmonic forcing at the lateral boundaries in MEDWEST60 since the tidal forcing is already explicitly part of the hourly boundary forcing from eNATL60 outputs,
- the model time-step has been increased by a factor 2 (80 seconds in MEDWEST60 versus 40 seconds in eNATL60) in this regional domain (stability criteria easier to meet in the West Mediterranean region compared to other regions in the North Atlantic).

More details about the starting protocole (spinup) and time-step change are given in section 4.1 along with the ensemble experimental plan.

3 PARAMETERIZATION OF MODEL UNCERTAINTIES

The high-resolution model presented in the previous section is a deterministic model, in the sense that the future evolution of the system is fully determined by the specification of the initial conditions, the boundary conditions and the forcing functions. This type of model is the archetype of the models that are currently used in CMEMS operational forecasting systems (except for the Arctic system, which is not based on NEMO). In this context, forecast uncertainties can only be explained by initial uncertainties, boundary uncertainties or forcing uncertainties, usually amplified by unstable model dynamics. However, as explained in the introduction, the objective of this study is to go beyond this assumption and include the possibility of model errors impairing the predictability of the finest scales.

To transform the deterministic model presented above into a stochastic model, our focus is to simulate uncertainties that primarily affect the smallest scales of the ocean flow, and let them upscale to larger scales according to the model dynamics. These uncertainties are likely to depend mostly on the intimate structure of the model, by embedding misrepresentations of the unresolved scales and approximations in the model numerics. A detailed causal examination of the origin and interactions between these various possible sources of error being quite impossible to achieve, we propose to introduce here a bulk paramterization of these effects, by assuming that one of the most important dynamical consequence of these errors on the finest scales is to generate uncertainty in the location of the oceanic structures (currents, fronts, filaments,...). More details about this assumption is provided in section 3.1 below, and the implementation of this parameterization in NEMO is described in section 3.2.

3.1 Location uncertainties

Location errors in a field $\varphi(\mathbf{x}, t)$, function of the spatial coordinates $\mathbf{x}$ and time t, occur if the field φ displays the correct values but not at the right location. More precisely, this means that the field $\varphi(\mathbf{x}, t)$ can be related to the true field $\varphi^t(\mathbf{x}, t)$ by the transformation:

$$\varphi^t(\mathbf{x}, t) = \varphi\left[\mathbf{x}^t(\mathbf{x}, t), t\right] \tag{1}$$

where $\mathbf{x}^t(\mathbf{x}, t)$ is an anamorphic transformation of the coordinates defining the location where to find the true value of $\varphi(\mathbf{x}, t)$. With respect to the true field φ^t, the values of φ are thus shifted by:

$$\delta\mathbf{x}(\mathbf{x}, t) = \mathbf{x}^t(\mathbf{x}, t) - \mathbf{x} \tag{2}$$

which defines the location error.

If the field $\varphi(\mathbf{x}, t)$ is evolved in time, over one time step Δt, with the model $\mathcal{M}$:

$$\varphi(\mathbf{x}, t + \Delta t) = \mathcal{M}\left[\varphi(\mathbf{x}, t), t\right] \tag{3}$$

we can make the assumption that one of the effect of the model is to generate location uncertainties. In an advection-dominated regime, this means for example that the displacement of the oceanic structures can be different, maybe too large or too small, from what the deterministic model predicts. With this assumption, the model transforms to:

$$\varphi[\mathbf{x} + \delta\mathbf{x}(\mathbf{x}, t + \Delta t), t + \Delta t] = \mathcal{M}\{\varphi[\mathbf{x} + \delta\mathbf{x}(\mathbf{x}, t), t], t\} \tag{4}$$

where the location error $\delta\mathbf{x}(\mathbf{x}, t)$ can be simulated for instance by a stochastic process $\mathcal{P}$:

$$\delta\mathbf{x}(\mathbf{x}, t + \Delta t) = \mathcal{P}\left[\delta\mathbf{x}(\mathbf{x}, t), \varphi(\mathbf{x}, t), t\right] \tag{5}$$

where an explicit dependence to φ and t has here been included to keep the formulation general.

In ocean numerical models, the coordinates $\mathbf{x}$ are usually discretized on a constant grid. To implement the stochastic model in Eq. (4) on this numerical grid, one possibility would be to remap the updated field $\varphi[\mathbf{x}+\delta\mathbf{x}(\mathbf{x},t+\Delta t),t+\Delta t]$ on this constant grid at each model time step. This remapping would amount to a stochastic shift of the model field accounting for the presence of location uncertainties. However, this solution would be computationally very ineffective, and it is much easier to keep track of the modified location of the grid points (described by $\delta\mathbf{x}$), and use this modified grid to implement the model operator $\mathcal{M}$. In practice, to avoid deteriorating the model numerics, this solution may require that location errors remain small with respect to the size of the grid cells, and that their variations over one time step Δt are kept small enough to avoid undesirable numerical effects.

The simple and generic approach that is here proposed to simulate location uncertainties in ocean models has a close similarity to the work of Mémin (2014); Chapron et al. (2018), where it is argued that the effect of unresolved processes in a turbulent flow can be simulated by adding a random component to the Lagrangian displacement $\mathbf{d}X$ of the fluid parcels (as in a Brownian motion):

$$\mathbf{d}X = \mathbf{v}(\mathbf{x},t)\,dt + \boldsymbol{\sigma}(\mathbf{x},t)\,\mathbf{d}B \tag{6}$$

where $\mathbf{v}(\mathbf{x},t)$ is the velocity (as resolved by the model), $\mathbf{d}B$ is a white noise (uncorrelated in space an time) and $\boldsymbol{\sigma}(\mathbf{x},t)$ is a linear operator defining the correlation structure of the random displacement (assumed here correlated in space, but uncorrelated in time). The purpose of these studies is then to examine the effect of this modified material derivative (with the stochastic displacement added) when transformed into an Eulerian framework (i.e. in a constant coordinate system). In a nutshell, from this assumption, the authors manage to derive modified Navier-Stokes equations, with additional deterministic and stochastic terms depending on $\boldsymbol{\sigma}$. These new terms can be summarized to be: (i) an additional deterministic dissipation, and (ii) random fluctuations of the pressure gradient.

3.2 Implementation in NEMO

To implement location uncertainties in NEMO, we explicitly make the assumption that the location errors $\delta\mathbf{x}$ remain small with respect to the size of the grid cells, so that the nodes of the modified grid just follow a small random walk around the nodes of the original grid. Consistently with this assumption, we make the approximation that the model input data (bathymetry, atmospheric forcing, open-sea boundary conditions, river runoffs,...) keep the same location with respect to the model grid, which means that these data are not remapped on the moving grid. Such a tiny shift of the data (much smaller than the grid resolution) would indeed represent a substantial computational burden, with many possible technical complications, and would only produce small additional perturbations to the model solution, which do not correspond to the main effect that we want to simulate. Implictly, this means that the input data are continuously slightly distorted to follow the distortion of the model grid.

Since the model grid is assumed fixed with respect to the outside world, we need only represent the displacement of each model grid point relative to its neighbours. In NEMO, this relative displacement of the model grid points can easily be obtained by transforming the metrics of the grid, which is numerically represented by the distance between the neighbour grid points. A stochastic metrics, describing relative location uncertainties in the model operator $\mathcal{M}$, corresponds to the main effects that we want to simulate, because it can be thought to embed physical and numerical uncertainties that primarily affect the smallest scales. On the one hand, this can be

viewed as an explicit transcription of Eq. (6) in the internal model dynamics, and can thus be argued to describe uncertainties that upscale from unresolved processes. On the other hand, since the metrics is used everywhere in the model to evaluate differential and integral operators, making it stochastic can also be viewed as a simple approach to simulate numerical uncertainties simultaneously in all model components.

In practice, to obtain a stochastic metrics in NEMO, we must transform the arrays describing the size of the grid cells into time-dependent stochastic processes. Thus, if $\Delta\mathbf{x}_i(t) = [\Delta x_i(t), \Delta y_i(t), \Delta z_i(t)]$ is the size of grid cell number i at time t, we must define stochastic processes $\mathcal{P}_i$ such that:

$$\Delta\mathbf{x}_i(t+\Delta t) = \mathcal{P}_i\left[\Delta\mathbf{x}_1(t), \ldots, \Delta\mathbf{x}_j(t), \ldots\right] \tag{7}$$

A very simple approach to define the $\mathcal{P}_i$ is then to use first-order autoregressive processes $\xi_i(t)$ as a multiplicative noise applied to the reference model grid $\Delta\mathbf{x}_i^0$:

$$\Delta\mathbf{x}_i(t) = \Delta\mathbf{x}_i^0 \circ [1 + \xi_i(t)] \tag{8}$$

with

$$\xi_i(t+\Delta t) = \mathbf{a} \circ \xi_i(t) + \mathbf{b} \circ \mathbf{w} \tag{9}$$

where $\circ$ is the Hadamard product, $\mathbf{w}$ is a vector of independent Gaussian white noises, and $\mathbf{a}$ and $\mathbf{b}$ are constant coefficients governing the standard deviation and the correlation length scale of the ξ_i. The three components of ξ_i are thus assumed independent, which means that the grid is deformed independently along the three dimensions.

The use of autoregressive processes $\xi_i(t)$ to simulate the stochastic distortion of the model grid makes the implementation of the scheme straightforward in NEMO, since we can directly apply the tools developed by Brankart et al. (2015) to generate the ξ_i. This tool was indeed meant to be generic enough to trigger various sorts of stochastic parameterizations in NEMO, and has already been used to simulate various sources of uncertainty, including the effect of unresolved scales in the seawater equation of state (Brankart, 2013; Zanna et al., 2019) and in the biogeochemichal equations (Garnier et al., 2016), or the effect of parameter uncertainties in the sea ice model (Brankart et al., 2015) and in the biogeochemichal model (Garnier et al., 2016). This tool only requires specifying a few parameters to characterize the stochastic processes $\xi_i(t)$: the standard deviation (σ), the correlation time scale (τ), the number of passes (P) of a Laplacian filter applied to the ξ_i, and the order (n) of the autoregressive processes. The two last parameters go beyond the formulation of Eq. (9), which describes first order processes (AR1) uncorrelated in space. The application of a Laplacian filter introduces space correlation and makes the distortion of the grid smoother in space, and the use of ARn rather than AR1 processes modifies the time correlation structure and makes the distortion of the grid smoother in time. It must also be noted that the use of ARn processes is also more general than Eq. (7) by making the processes $\mathcal{P}_i$ depend on the n previous time steps, rather than just the previous time step. Fig. 3.1 illustrates the effect of the order (n) of the stochastic process on the time correlation structure.

In the present study, the distortion of the grid has been limited to horizontal displacements of the model grid points, with the same displacements applied to all model fields and along the vertical. This reduces the number of stochastic fields to generate to two two-dimensional fields, one for each of the horizontal coordinates $\Delta x_i(t)$ and $\Delta y_i(t)$. However, since the NEMO fields are shifted according to the rules of the Arakawa C-grid, the stochastic metrics is first computed for the T-grid and then transformed to the other grids to be consistent with the shifted position of the grid points. In the application, the standard deviation is set to a relatively small value σ =1% or 5%, to be consistent with the assumption of small location errors, and the correlation time scale is set to 1440 time steps (1 day) to be consistent with

the assumption of a small variation of the grid over one time step. Some effort is also made to keep the perturbation smooth in space and time by applying $P = 10$ passes of a Laplacian filter and by using second order autoregressive processes ($n = 2$). Fig. 3.2 illustrates the resulting perturbation of the grid relative to the reference model grid.

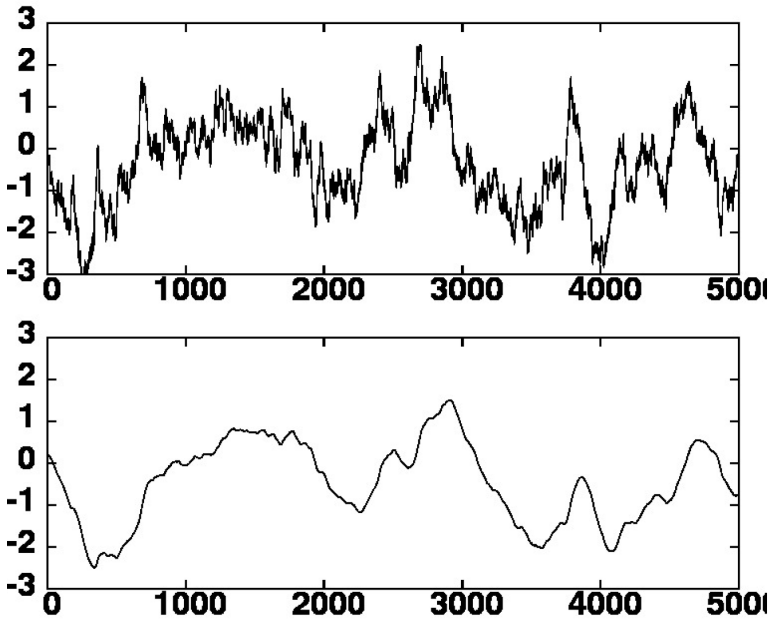


Figure 3.1: First order ($n = 1$, top panel) and second order ($n = 2$, bottom panel) Gaussian autoregressive stochastic process, with zero mean and unit standard deviation. The X-axis is time (labeled in number of time steps), and the correlation time scale is set to $\tau = 180$ time steps.

Finally, before moving to the application, it is interesting to browse the NEMO code and see more concretely where the modification of the metrics can produce a direct effect:

- *Vertical physics.* The horizontal metrics arrays are used everywhere in the code to compute the surface separating two superposed grid cells, and to compute any integrated exchange between the two cells (or between the top cell and the atmosphere, or between the bottom cell and the sea floor). However, since the volume of the cell is modified in the same proportion as the horizontal surface, the perturbation of the metrics does not modify the fluxes at the horizontal interfaces and does not modify the contribution of the vertical physics to the model tendency in each cell.

- *Horizontal diffusion.* The horizontal metrics is also used to compute the horizontal derivatives involved in the computation of diffusion. The effect is to increase diffusion where the grid cells become smaller and to increase diffusion where they become larger. If σ is small, it is equivalent to stochastically increase or decrease diffusivities by a few times σ, and the effect should more or less average to zero after a sufficient time.

- *Horizontal advection.* The effect of the perturbation of the horizontal derivatives in the advection scheme is presumably much less anecdotic. The stochastic part of the material derivative in Eq. (6) is accounted for by the displacement of the grid, but in return, the transformed grid induces modifications in the advection by the resolved scales. This is one of the effect that location uncertainties are meant to simulate, together with possible errors in the numerical scheme. Refering to the work of Mémin (2014), it might be anticipated that

one of the result of the parameterization is to produce an additional cause of dissipation in the model.

- *Horizontal pressure gradient.* In our system, the main effect of location uncertainties is certainly to produce stochastic fluctuations of the horizontal pressure gradient, which is quite consistent with the conclusions obtained by Mémin (2014). These fluctuations should indeed trigger additional ageostrophic motions, with enhanced associated vertical velocities, and may bring a substantial limitation to the predictability of the small scale motions.

In summary, we see that our parameterization of location uncertainties can produce effects in several components of the model. However, what is maybe more important is that they all consistently derive from the same cause, under the common assumption that the updated location of the fluid parcels after a model time step is not exact, but approximate.

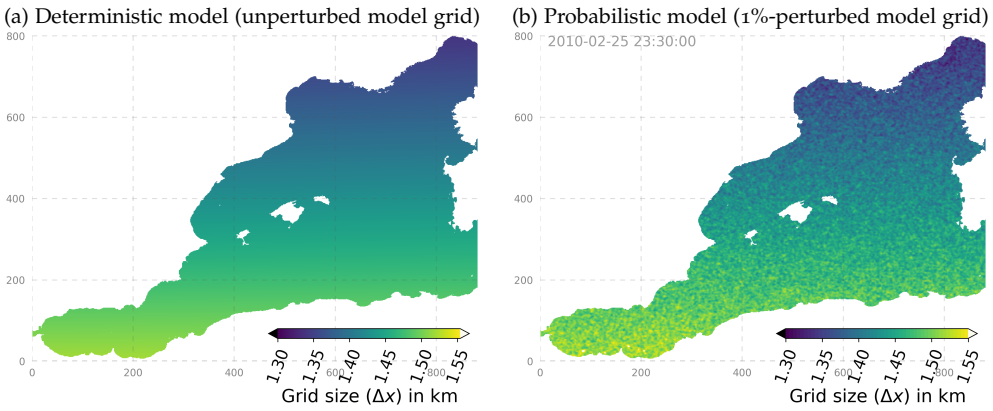


Figure 3.2: Size of the model grid in the horizontal east-west dimension (e1t in NEMO): (a) unperturbed, from the standard NEMO grid at 1/60° resolution, and (b) snapshot of the perturbed metric at a given date (stochastic perturbation set to a level of std=1%).

4 THE MEDWEST60 ENSEMBLE EXPERIMENTS

This study aims to evaluate the predictability of the fine scale dynamics in a typical high-resolution NEMO-based CMEMS model, here the regional MEDWEST60 configuration, by including the effect of (a) initial uncertainties and (b) model uncertainties. In this section, we describe the design and technical details of the four ensemble experiments performed in this context. Three ensemble experiments are performed with the probabilistic model presented in the previous section (i.e., including some model error via a stochastic parameterization of location uncertainties), and starting from identical initial conditions. In practice, two experiments are dedicated to two different amplitudes of model uncertainties, and a third experiment investigates the sensitivity of the probabilistic model to the start date of the experiment (winter versus summer). Finally, a fourth ensemble experiment is also performed with the deterministic model (i.e., no model error) but perturbed initial conditions.

4.1 Experimental protocole and list of the experiments

SPINUP :
As explained in section 2, the set-up of MEDWEST60 is such that the configuration can be started from instantaneous conditions stored from the reference simulation eNATL60-LBT02, i.e. from a balanced 3-D ocean state from a previous NEMO restart file on the same horizontal and vertical grid. The spinup protocole we used is thus as follows:

- From a NEMO restart file archived from eNATL60-LBT02 on a given date we extract the horizontal and vertical domain corresponding to MEDWEST60,
- A single MEDWEST60 simulation is then started from this extracted restart file, using an euler scheme for the first timestep and the exact same timestep as eNATL60-LBT02 (i.e., δt=40 seconds). This single simulation is run for 5 days and a MEDWEST60 restart file is saved after that time.
- This MEDWEST60 restart file is then used to start and run 5 more days of the single simulation, this time increasing the timestep to δt=80 seconds. A final restart file is saved, to be used to start the ensemble experiments (which are run with the same time-step of δt=80 s).

GENERATING THE ENSEMBLE EXPERIMENTS :

In this work, we test the effect of model error of different amplitude. As exposed in section 3, the aggregated effect of model errors for the fine scales is introduced here as a "location uncertainty" using the NEMO tool developed by Brankart et al. (2015). This location uncertainty is implemented as a stochastic perturbation added at each model time step to the horizontal metrics of the model, e1 and e2 (in other words, to the model grid size). In practice, the perturbation expresses a random order-2 auto-regressive process (see Fig. 3.1), of which can be set:

- the amplitude,
- the correlation in time,
- some spatial smoothing (laplacian filter).

The MEDWEST60 ensemble experiments explore two different amplitudes of this stochastic perturbation (standard deviation of 1% and 5%). The latter (5%) can be considered a large perturbation for the horizontal metrics. Larger perturbation (for example 10%) have been eliminated from our experimental plan since unphysical

impacts could be detected as visible noise on the SSH and SST fields. By design, the other parameters of the stochastic module are kept identical in all the experiments: the time correlation is set to 1 day (1080 timesteps), and the laplacian filter is set to 10 grid points.

LIST OF THE MEDWEST60 EXPERIMENTS PERFORMED :

Table 4.1 summarizes the set of experiments performed and their main characteristics. Three MEDWEST60 ensemble experiments are performed with the probabilistic model (i.e. including model error), and starting from identical initial conditions: ENS-1%, ENS-5%, ENS-1%-S, with an amplitude of the stochastic perturbation of: standard deviation 1%, 5%, 1%, respectively. ENS-1% and ENS-5% start from initial conditions in the winter season (05-02-2010) when mesoscale activity in the western Mediterranean sea is expected to be large. ENS-1%-S starts from initial conditions in the summer season (05-02-2010) when mesoscale activity in the western Mediterranean sea is expected to be weaker. For reference, a single simulation (1 member) is also performed with the deterministic model (no model error) on the same winter period: DREF. A forth ensemble experiment, ENS-CI, is finally performed with the deterministic model (i.e. no model error) to study predictability under un-perfect initial conditions. This unperturbed ensemble is initialized from ensemble conditions taken from experiments ENS-1% after 1 day of simulation (i.e. when the state of the 20 members has already slightly diverged on the fine scales, due to the stochastic perturbation introduced in ENS-1%). Note that the choice is made to start experiment ENS-CI with small initial errors, but this experiment also virtually gives access to forecasts initialized with larger errors by considering alternatively day 1, day 2, (...) day 10, etc of ENS-CI as many different start times. This approach will be followed for the predictability diagnostics proposed in section 5. It relies on the strong assumption that varying the initial date of the ensemble forecast does not influence the predictability results more than the initial uncertainty itself. But the alternative approach would have required performing a large number of ensemble forecasts with various levels of uncertainty on the initial conditions (on a same start time), and would have been very substantially more expensive.

	MEDWEST60 experiments				
Name	DREF	ENS-1%	ENS-5%	ENS-1%-S	ENS-CI
File name(1)	GSL03	GSL14	GSL15	GSL16	GSL19
Start date:	05-02-2010	05-02-2010	05-02-2010	01-08-2010	06-02-2010
Length:	60 d	60 d	60 d	30 d	60 d
Type of experiment:	single	ensemble	ensemble	ensemble	ensemble
Ensemble members:	1	20	20	20	20
Ens. initial conditions:	-	identical	identical	identical	perturbed(2)
Restart from:	spinup	spinup	spinup	spinup	ENS-1% restart after 1 day
Model:	deterministic	probabilistic	probabilistic	probabilistic	deterministic
Stochastic perturbation	*none*	e1,e2	e1,e2	e1,e2	*none*
& amplitude	-	std=1%	std=5%	std=1%	-

Table 4.1: Characteristics of the five MEDWEST60 experiments. (1): Original file names (suffix) as stored on the HPC. (2)The "perturbed" initial conditions of experiment ENS-CI are taken from the restart files of experiment ENS-1% (stochastically perturbed) after 1 day of simulation. See text in section 4.1 for more details on the experiments characteristics and the spinup protocole.

4.2 Implementation of the ensemble code in NEMO3.6

IMPLEMENTATION AND PARALLELIZATION :
The ensemble version of NEMO3.6 is run on the `Jean Zay` machine at the IDRIS supercomputing center in Paris, France, using 3 nodes of 40 CPU each (i.e. 120 processors in total) for each member. The MEDWEST60 domain has a horizontal grid of 883 x 803 grid points. The domain is broken down in 12 x 18 subdomains for parallelization purposes (jpni x jpnj), of which 120 subdomains include ocean grid points. A description of the ensemble version of NEMO can be found in Bessières et al. (2017); Leroux et al. (2018). The MEDWEST60 configuration files (namelist, xml, bathymetry, etc...) are shared on the MEDWEST60 github repository: `https://github.com/ocean-next/MEDWEST60`.

CPU COST, STORAGE AND DISK SPACE :
A typical MEDWEST60 ensemble experiment with 20 members is run on 20 x 120 = 2400 NEMO processors. For such an ensemble, we also dedicate 200 processors to the i/o operations on the XIOS-2 servers. Those processors are distributed on 8 dedicated, "depopulated" nodes, where only 25 cores can be activated per node over the 40 available (for memory purposes). The average runtime for 1 simulated day of such an ensemble experiment is about ~0.5 hour. The CPU cost of a MEDWEST60 ensemble experiment of 20 members and 60 days is thus ~ 30h x 2600 CPU = 78 000 hCPU.

The entire 3D outputs are currently stored at the IDRIS supercomputing center in Paris, France. The 3D variables are stored at the hourly frequency in netcdf4 daily files, 1 file per variable : gridT, gridS, gridU, gridV, gridW, gridZ. The 2D variables are stored in gridT-2D, gridU-2D,gridV-2D, flxT files (also at hourly frequency). The disk space occupied by one typical ensemble experiment of 60 days is about 17 To.

All the development and production work with the MEDWEST60 configuration in this project have been performed using HPC resources from GENCI-IDRIS, France (Grant A008-0101279).

4.3 Impact of introducing model uncertainties in MEDWEST60

In this section, we give some illustrations of the behaviour of the probabilistic model and the impact of the stochastic perturbation introduced in ensemble experiments ENS-1% and ENS-5% in comparison to the deterministic model started from perturbed initial conditions (ENS-CI). As discussed in section 3.2, the stochastic perturbation is applied on the model metrics, while the location of the grid points themselves is assumed the same for all members. In other words, the field itself is still considered to be located on the reference grid, for instance with respect to the bathymetry and the external forcing, and the effect of the perturbation is only taken into account in the model operator (e.g. for the differential operations), and it is neglected everywhere else. It implies that ensemble statistics (mean, standard deviation, covariance matrix,...) can be computed as usual on the reference grid. But the perturbed metrics must be used to compute any diagnostics involving a differential operator. In the following, for instance, the perturbed metrics were used to compute relative vorticity from the velocity fields, to be consistent with the perturbed model dynamics, which is specific to each member. For that purpose, the perturbed metrics were archived with time, at the hourly frequency, for each member simulation.

STOCHASTIC VS DETERMINISTIC MODEL :

The stochastic perturbation used in this work was designed to introduce noise at the very small scales, which is then expected to cascade toward larger scales according to the model dynamics (see sections 3 and 4.1).

As a first element of comparison, Fig. 4.1 shows that on average over the 2 months of the ensemble experiments, the simulated mean SSH remains relatively similar in both the experiments based on the deterministic and the probabilistic model. Their mean SSH is also very consistent with the reference simulation eNATL60-BCLB02 (Brodeau et al., 2020) used as boundary conditions.

It is also verified with Fig. 4.2 that even on the hourly frequency, the level of stochastic noise introduced in experiments ENS-1% and ENS-5% remains small enough to avoid any additional visible noise on the hourly fields compared to the deterministic model. The figure only shows a zoom on a subregion of 250 x 250 gridpoints south of the Balearic Islands to better focus on the small-scale features.

To document further this aspect, Fig.4.3 compares the wavenumber power spectrum (Power Spectral Density, PSD) from hourly SSH in the different experiments. The figure shows very consistent SSH spectra in the perturbed and unperturbed models, and both in MEDWEST60 experiments and the eNATL60-BCLB02 simulation. A small bump however appears for scales around λ=10 km in the probabilistic experiment with the largest level of stochastic noise (ENS-5%). This feature is also clearly highlighted in Fig. 4.4 where the ratio is plotted of the mean PSD of each of the two perturbed experiments ENS-1% and ENS-5%, over the mean PSD of the unperturbed experiment ENS-CI. The power spectral density from ENS-5% is around 1.5 times larger than in ENS-CI on the small scales (peaking around λ =8 km), illustrating the direct effect of the stochastic perturbation introduced in the model. This effect is much weaker in the experiment with the smaller level of stochastic noise (ENS-1%).

Note that in Fig.4.3 is also shown the spread of the PSD around the ensemble mean of each experiment (in very thin lines): the members all have a PSD very consistent with their ensemble mean (the spread is smaller than the thickness of the ensemble mean line) on all scales up to ~150 km. For larger scales, some spread is seen between the members and it provides an idea of the sensitivity (significance) of such a spectral analysis on the last few point of the spectrum (aliasing effects). The spectra are computed here over a squared box of L ~450 km (see Fig.2.1), and don't resolve well the spectral scales larger than L/2. The ensemble spread interval appearing in the figure thus provides some guidance as to interpret the significance of the PSD variations in this scale range.

SPREAD GROWTH :

Fig. 4.10 shows the evolution with time of the ensemble standard deviation of the hourly SSH, then spatially averaged over the entire MEDWEST60 domain, for each of the ensemble experiments. As expected, the ensemble spread grows faster in the perturbed experiments (probabilistic model) than in the unperturbed experiment (deterministic model). After about 50 days of simulation, the ensemble spread of all three winter experiments (ENS-CI, ENS-1% and ENS-5%) have converged to a similar value. The spread is still growing at the end of the 60-day experiments but the curves have started to flatten, suggesting that our experimental protocole was successful at initiating divergent-enough ensembles on the targeted time-range (2 months). Similar characteristics of the spread growth are seen in the other surface variables we have examined (SST, SSS, relative vorticity).

After 2 months, the winter experiments (ENS-CI, ENS-1% and ENS-5%) have reached an ensemble spread in SSH of about 2.5 cm in average over the domain, but local maxima of spread values are found around 10 cm (see Fig. 4.11). Those values are close to typical deviation values of hourly SSH over time in the Mediterranean region. Further investigations discussed in the following paragraph (SPATIAL DECORRELATION), also confirm that the spatial decorrelation of the submeso- and meso- scale features has been reached by the end of the 2-month experiments.

After the ~10 first days of simulation, the ensemble spread in the three winter ensembles appears to evolve in a relatively similar manner, in parallel, and almost linearly until day 40-50, where the curves then start to flatten and converge. Only

in the first few days, the presence of model error seems to make a difference in the growth rate, ENS-5% clearly showing a faster growth than ENS-1% (the latter being slightly faster than ENS-CI in the very first few days). This result suggests that in the context of short-range forecasting (1-5 days), model uncertainties might play a role as much as uncertain initial conditions, and should be taken into account in operational systems.

IMPACT OF THE SEASON :
Fig. 4.10 also presents the ensemble spread growth in experiment ENS-1%-S, designed to test the impact of seasonality. This experiment is identical to ENS-1% except for the initial conditions, which are taken in summer (1/08/2010) in ENS-1%-S while in winter (5/02/2010) in ENS-1%. The growth of the ensemble spread is significantly slower with summer initial conditions, than with winter initial conditions, confirming our hypothesis that the seasonal level of mesoscale turbulence activity plays a significant rôle in ensemble spread of the forecast and thus in the quantification of predictability. In the following, we chose not to investigate further the summer experiment (ENS-1%-S) given the slow spread growth it showed, and to rather focus on comparing the winter ensembles.

SPATIAL DECORRELATION :
Figs. 4.5 to 4.9 give an example of how hourly SSH in two different members of experiment ENS-CI diverge with time. Fig. 4.5 first compares hourly snapshots of SSH and SST at the end of the 2-month experiment over the entire MEDWEST60 domain. At that time-lag, the ocean state of the two members appear clearly distinct from each other. Figs. 4.6 to 4.9 provide a sequence of hourly SST and relative vorticity snapshots on a smaller subregion, in order to focus on the smallest simulated features: at a short time-lag of +1-day, the ocean state of the two example members are barely distinguishable from each other. With a +20 day time-lag, differences start to appear on the exact location of the small features and their shape. With +30 and +60 day time-lags, the differences become more and more obvious even on larger features and eddies, and at +60 days, many features don't even have their corresponding feature in the other member.

Fig. 4.12 aims at presenting the wavenumber spectral characteristics of the "forecast error" as a function of forecast time-lag in all three winter experiments. The forecast error is assessed as the difference of the hourly SSH between all pairs of members in the ensemble, and at each time-lag. In other words, each member is alternatively taken as the truth, and compared to the 19 remaining members, taken as the ensemble forecast for that given truth. The power spectral density (PSD) is computed at each time-lag for each pair difference and then averaged over the 20×19 permuted pairs. For reference, on the same figure is also plotted the ensemble-mean PSD of the hourly full-field SSH at time-lag +60 days.

After just one hour of simulation starting from perfect initial conditions with the probabilistic model (ENS-1%, yellow curve), the wavenumber spectrum of the forecast error peaks in the small scales around λ= 15 km and is still two order of magnitude smaller than the reference full-field SSH PSD (in thick black line on the figure). Same behaviour with the larger amplitude of model uncertainty (ENS-5%), except that the forecast-error spectrum is now just one order of magnitude smaller than the reference SSH spectrum after 1h. With increasing time-lag, the shape of the PSD becomes more "red" , with more and more power cascading to the larger scales. By the end of the experiments after 60 days, the PSD of the forecast error has almost converged to the reference full-field SSH PSD, suggesting that the members of the ensembles are more or less decorrelated by that time. Note that we don't necessarily expect a full spatial decorrelation between the members in this type of experiment since all members see the same surface forcing and lateral boundary conditions. From Fig.4.12 it is already interesting to note that on the very small scales (<6km), the spectrum of the forecast error does not seem to have converged

exactly to the SSH spectrum after 60 days. We will examine this aspect further in the next section (section 5.3).

From Fig. 4.12 it is also noteworthy that the evolution in time of the forecast error spectrum in ENS-CI and ENS-1% is very similar in amplitude and shape, except for the first time-lag (+1 hour), where the curve in ENS-CI is already smoother than in ENS-1% and does not show the λ= 15 km peak as in the latter. This is because ENS-CI is by design started from initial conditions from day 1 of ENS-1% (see section 4), and so the impact of the stochastic perturbation on the forecast-error spectrum in the first few hours/days is not present like in ENS-1%. But it is clear on the figure that by a time-lag of 5 days, both ENS-CI and ENS-1% have converged to a very similar forecast error spectrum and evolve in the same manner. Again, we find that model uncertainty might matter to the forecast precision in the first few days of the forecast. For a longer range, the uncertainty on the initial conditions becomes the main factor. This result is consistent with our above discussion about spread growth from Fig. 4.10.

SUMMARY :

- The probabilistic model introducing model uncertainty is validated, showing a very similar behaviour to the deterministic model.
- Only with the largest level of stochastic model error (experiment ENS-5%), the wavenumber spectrum of hourly SSH starts showing a slight, spurious power increase in the smallest scale (around λ=8 km). Improvements to the proposed parametrization of location uncertainty could try to reduce this effect, but this would go beyond the goal of the current study and deliverable.
- In this predictability study, the experiments with the probabilistic model primarily aim to generate ensembles of initial conditions. For that purpose we choose to use the ensemble of ocean states from experiment ENS-1% after 1 day to provide the ensemble initial conditions for experiment ENS-CI. The latter becomes the main experiment of our set and aims to assess predictability under uncertain initial conditions and a perfect (deterministic) model. Experiment ENS-1% and ENS-5% with the probabilistic model will be used to additionally assess the impact of irreducible model errors on the skill of the forecast.
- After 2 months of experiments, we find that ensemble members are spatially decorrelated and the ensemble variance has reached saturation at least for scales in the range ~10-60 km. For these scales, these ensemble experiments are thus appropriate to provide a statistical description of the dependence between initial accuracy and forecast accuracy over the full range of potentially useful forecast time lags (typically, between 1 and 20 days). This is what will be presented in the section 5.
- We confirm that the spread growth in the ensemble experiments is very dependent on the season at which the simulations are started. The test-experiment initialized in summer when mesoscale turbulence in the Mediterranean region is known to be less active has shown a very slowly-growing ensemble spread. Given that result, we choose not to investigate further the summer experiment in this study and focus on the three winter experiments : ENS-CI, ENS-1%, ENS-5%.
- Some first elements of comparison of the experiments with and without model uncertainty (regarding ensemble spread growth and spectral characteristics) suggest that the type of model error introduced here on the small scales matters only in the short-range of the forecast (<5 days). For longer time-lags, the uncertainty on initial conditions becomes the main factor. This aspect will

be illustrated further in the next section (section 5) from actual predictability diagnostics.

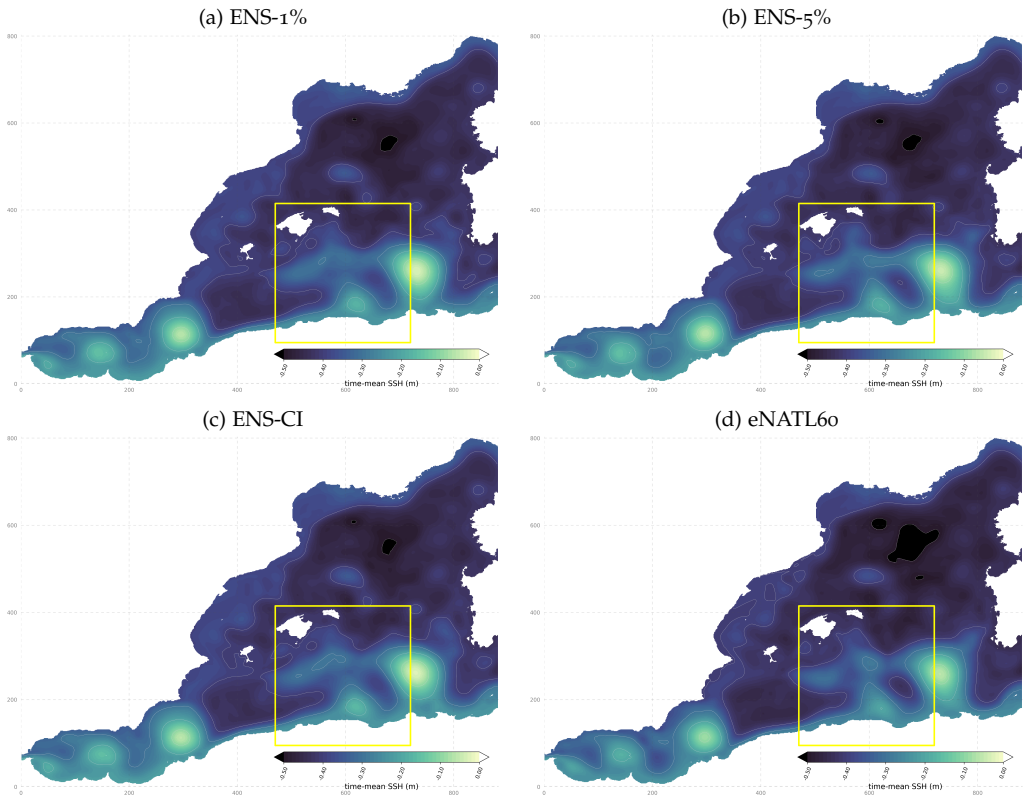


Figure 4.1: Time-mean SSH over 60 days in experiments ENS-CI (deterministic model, no model error), and both ENS-1% and ENS-5% (probabilistic model with stochastic perturbation applied at each time-step) using one example member of each experiment. The comparison is made with the simulation eNATL60-BCLB (Brodeau et al., 2020).

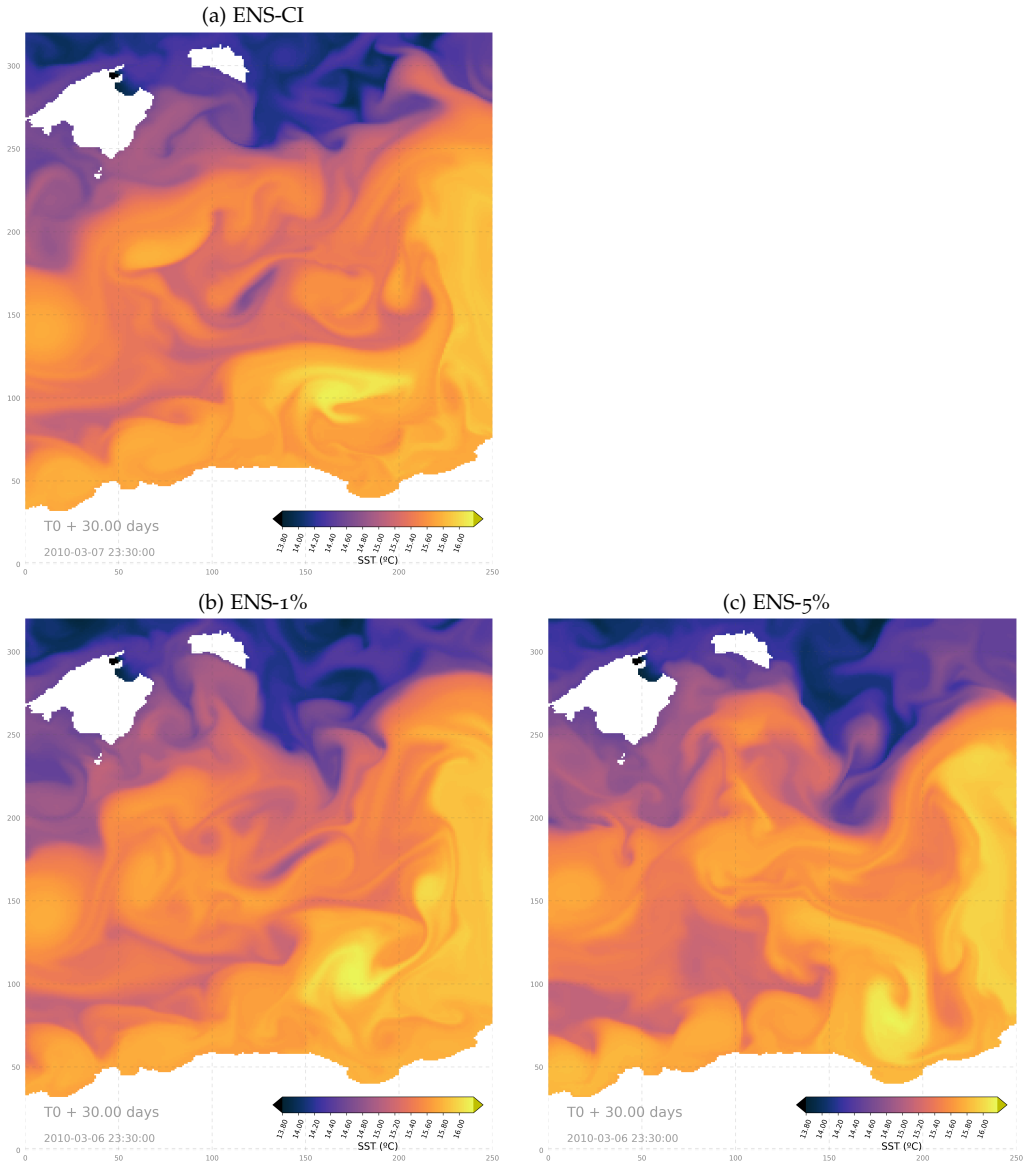


Figure 4.2: Hourly SST after 30 day in experiments ENS-CI (deterministic model, no model error), and both ENS-1% and ENS-5% (probabilistic model with stochastic perturbation applied at each time-step) in the subregion highlighted with the rectangle in Fig. 4.5.

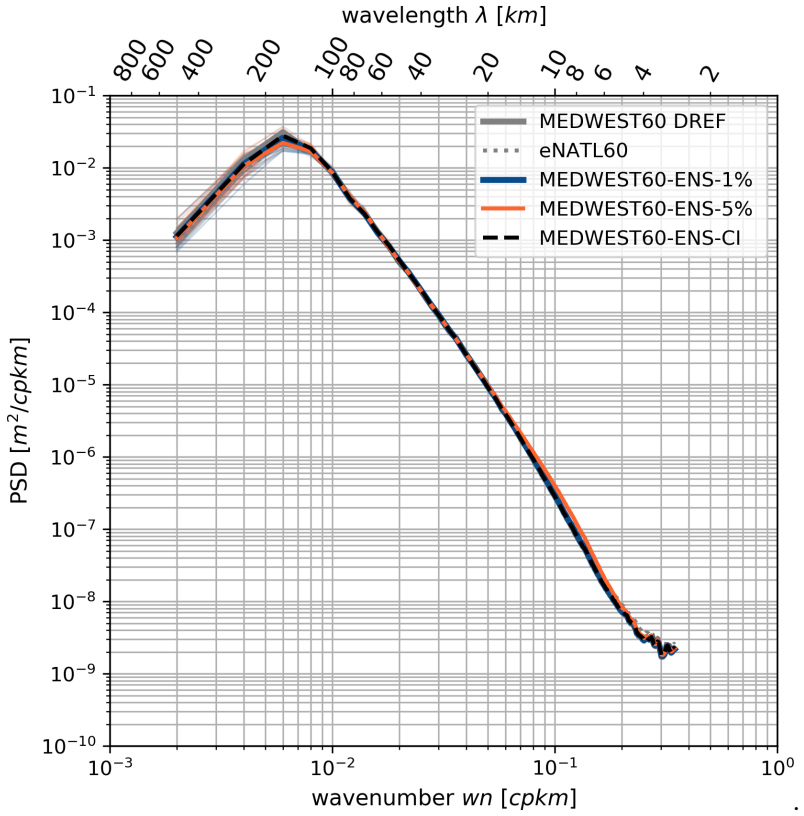


Figure 4.3: Wavenumber spectrum (Power Spectral Density, PSD) from hourly SSH in the MEDWEST60 ensemble experiments and in the reference (unperturbed) MEDWEST60 simulation, in a box of 350 × 350 gridpoints, corresponding to ~ 450 × 450 km (see Fig. 2.1). Comparison is also made with the eNATL60 simulation. The PSD of SSH [m^2/cpkm] is averaged in time over 241 hourly snapshots of SSH, one hourly spectrum every 6h, over the 2 months of simulation and over all members of the given ensemble (thick lines). The PSD of each individual members are also shown in thin lines, in the same color as their ensemble mean. The PSD computation is performed from the gridded model outputs following A. Ajayi's python module PowerSpec (`https://github.com/adeajayi-kunle/powerspec`.)

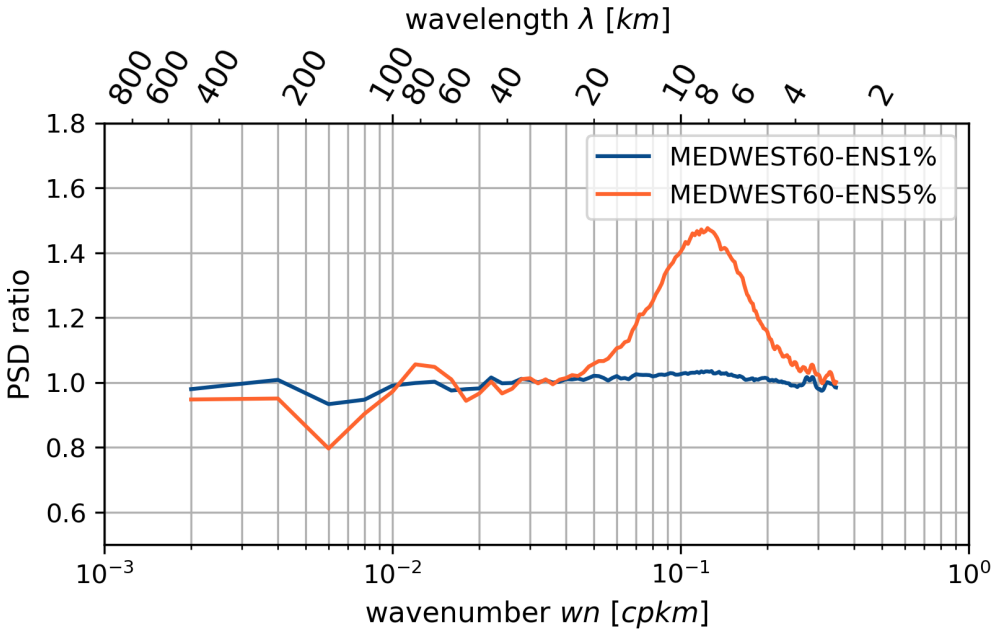


Figure 4.4: Ratio of the mean Power Spectral Density (PSD) of the hourly SSH from the two perturbed ensemble experiments (ENS-1% and ENS-5% with the probabilistic model) over the PSD from the unperturbed experiment (with the determinstic model). Each mean is taken over the 2-month period and over the corresponding ensemble.

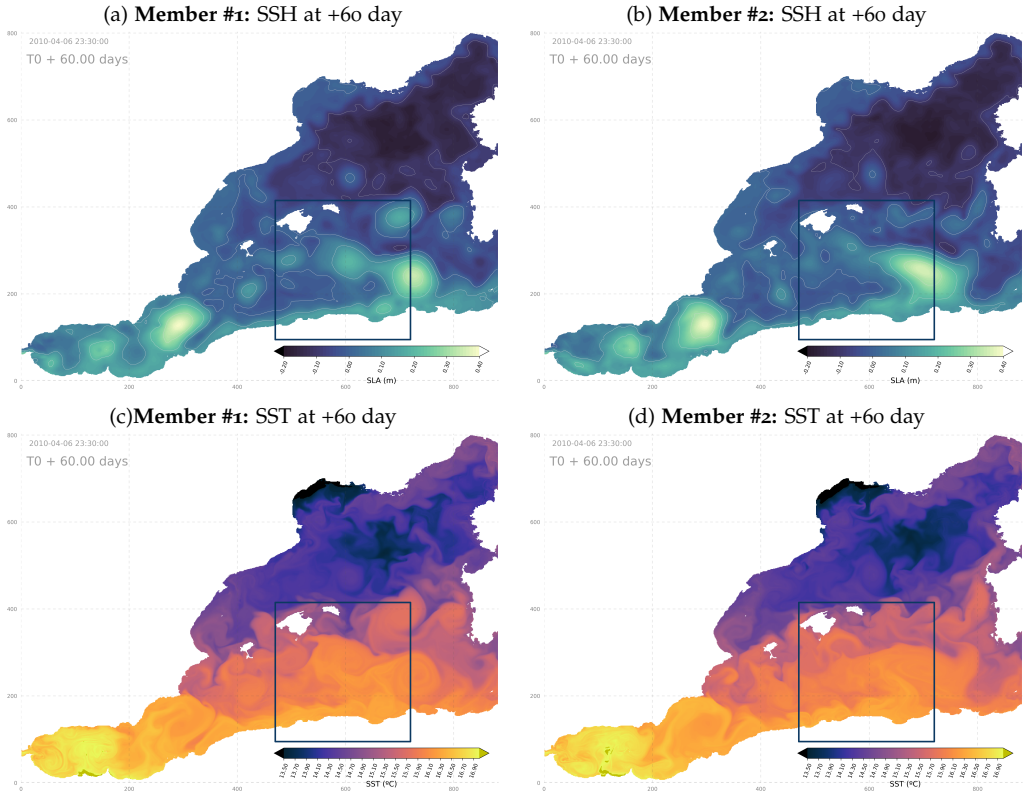


Figure 4.5: Hourly SSH (top row) and SST (bottom row) snapshots from member #1 (left) and member #2 (right) after 60 days in experiment ENS-CI. The blue rectangle indicates the zoom region plooted in Figs. 4.6-4.9.

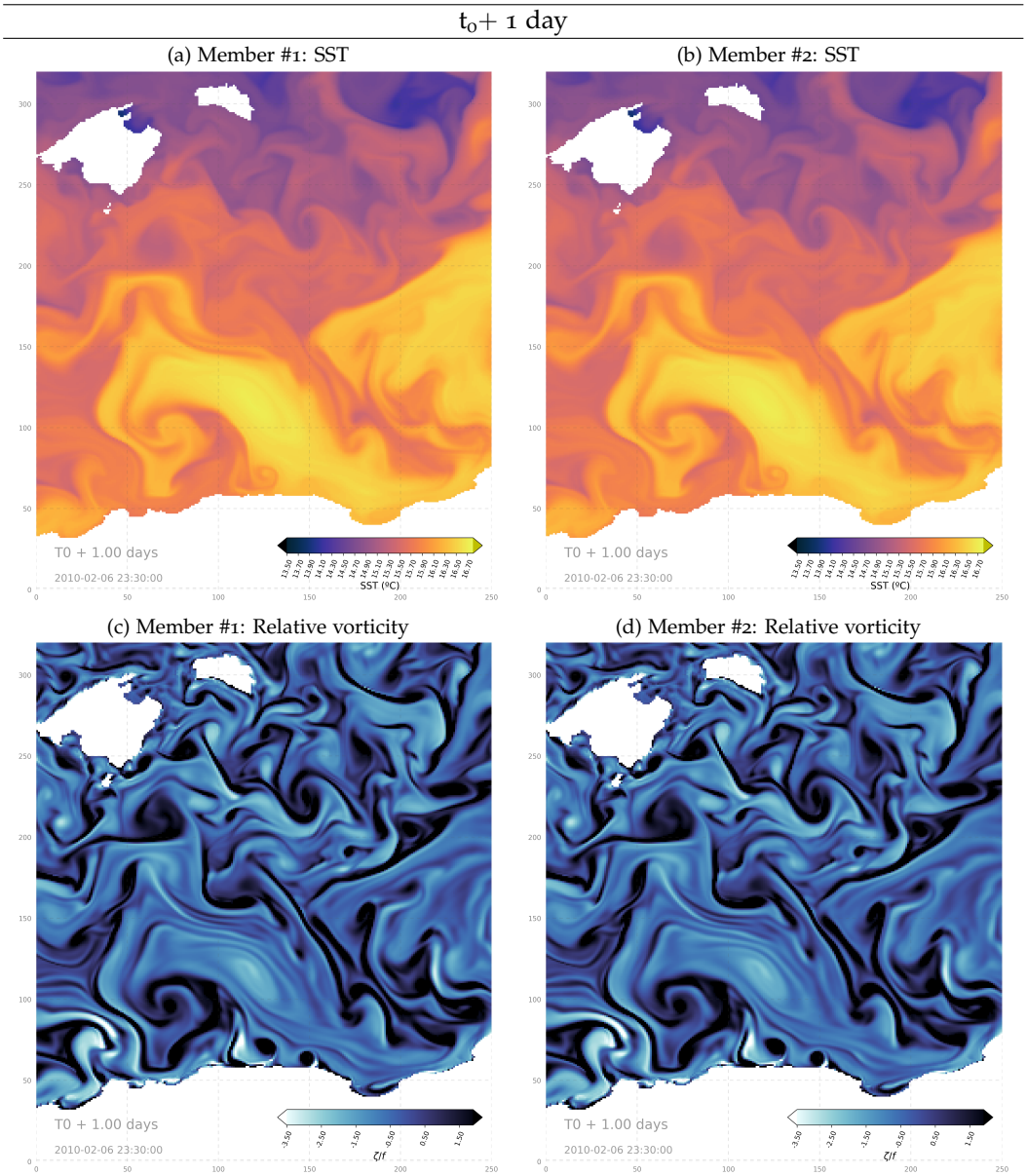
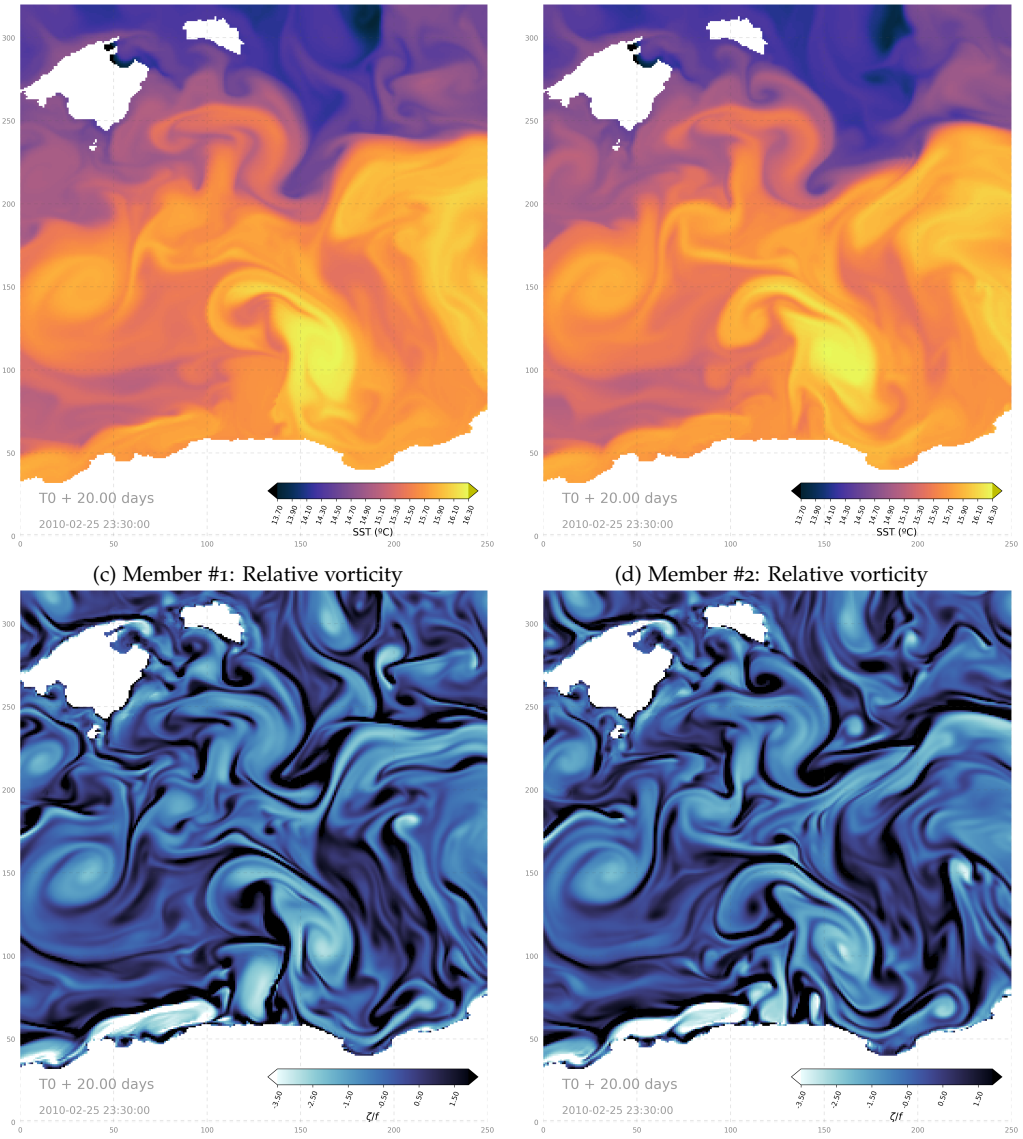


Figure 4.6: Hourly SST (top) and relative vorticity (bottom) snapshots from member #1 (left) and #2 (right) after 1 day in experiment ENS-CI in the subregion highlighted with the blue rectangle in Fig. 4.5.

t_0+ 20 day

(a) Member #1: SST

(b) Member #2: SST



(c) Member #1: Relative vorticity

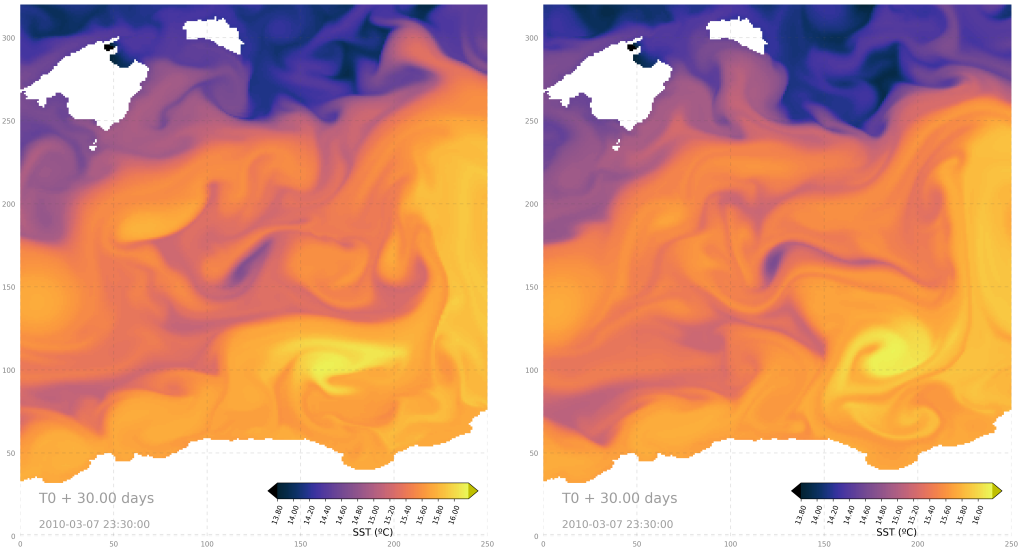
(d) Member #2: Relative vorticity

Figure 4.7: Same as Fig. 4.6 but after 20 days of experiment.

t_0+ 30 day

(a) Member #1: SST

(b) Member #2: SST



(c) Member #1: Relative vorticity

(d) Member #2: Relative vorticity

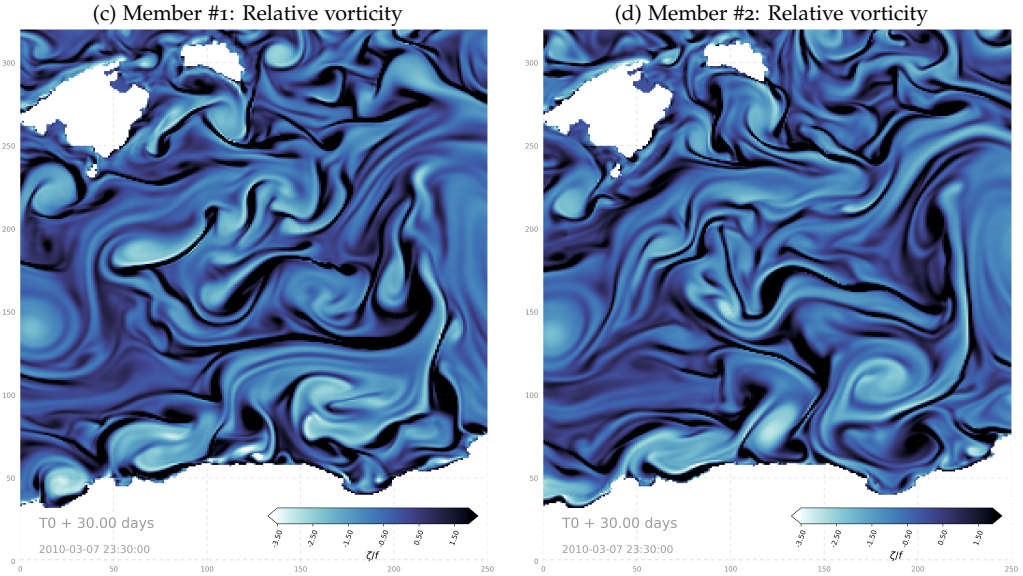


Figure 4.8: Same as Fig. 4.6 but after 30 days of experiment.

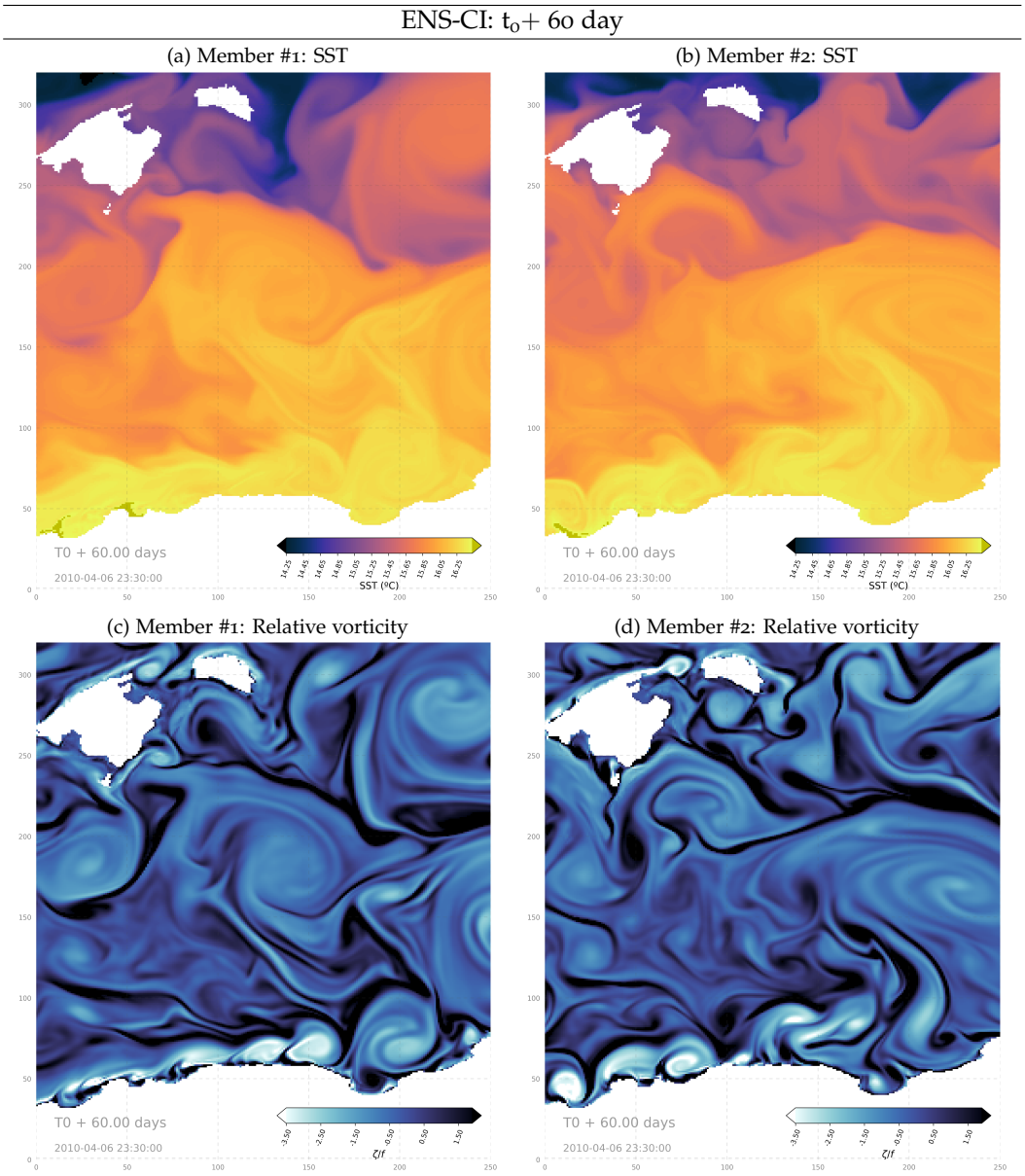
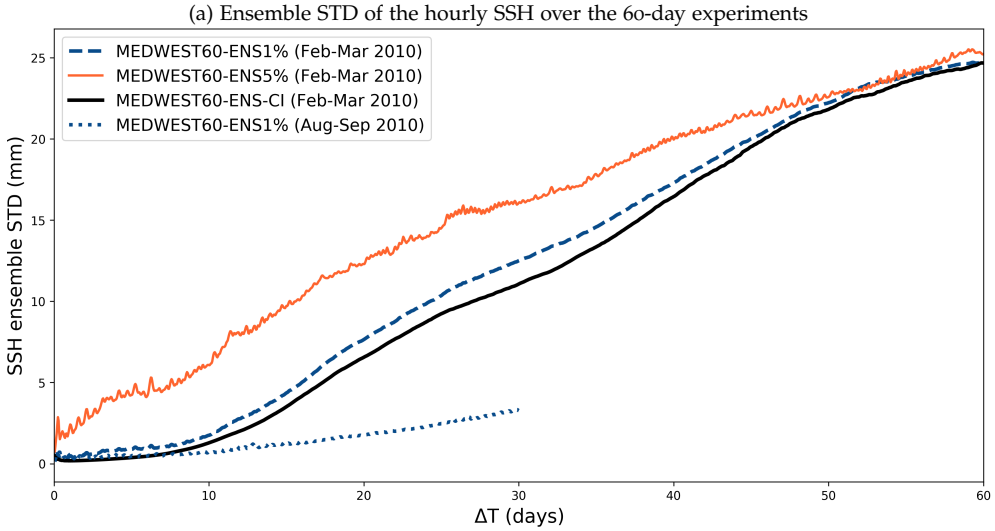


Figure 4.9: Same as Fig. 4.6 but after 60 days of experiment.

(a) Ensemble STD of the hourly SSH over the 60-day experiments



(b) Zoom on the first 10 days

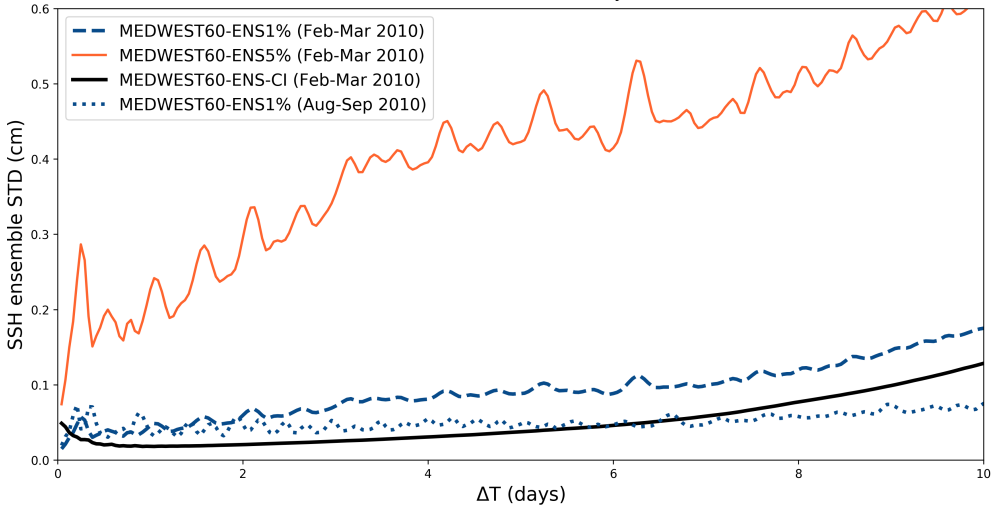


Figure 4.10: Time-evolution of the ensemble standard deviation of the hourly SSH, then spatially-averaged over the entire MEDWEST60 domain for the 4 ensemble experiments (ENS-1%, ENS-5%, ENS-1%-S, ENS-CI): (a) over 60 days, (b) zoom on the first 10 days of simulation where the spread growth is exponnential.

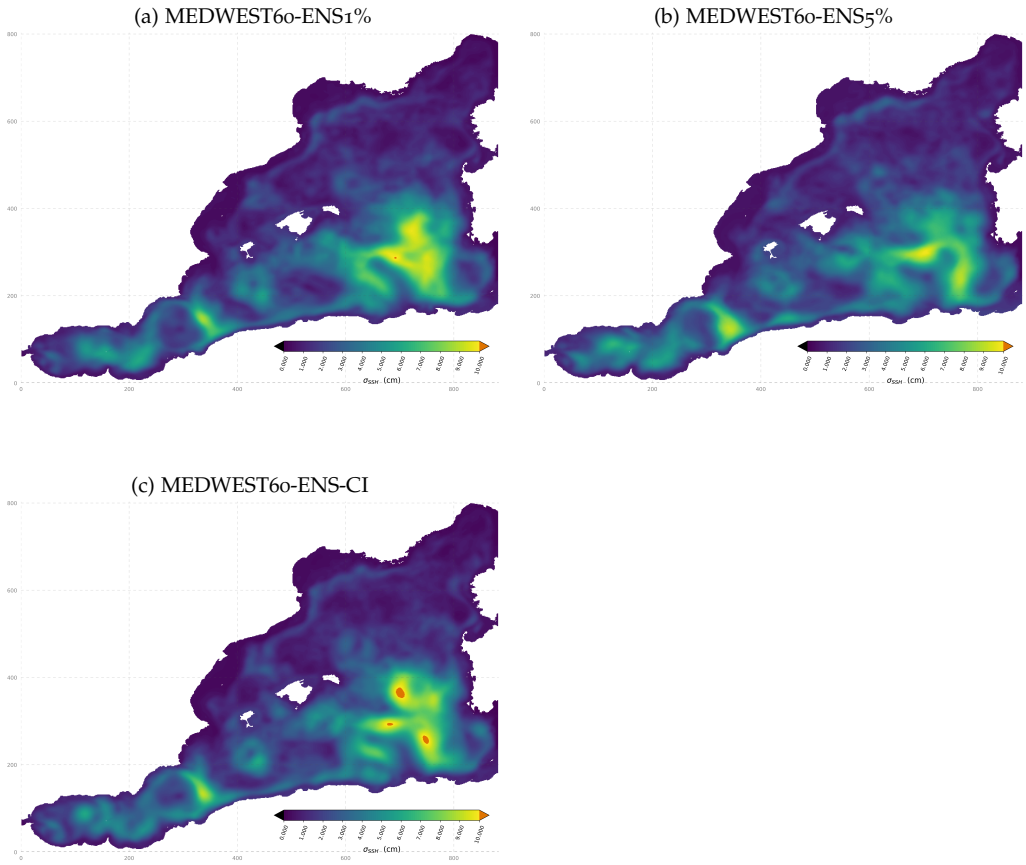


Figure 4.11: Maps of the ensemble standard deviation of the hourly SSH, averaged over the final period (55-60 days) from the ensemble experiments: ENS-1%, ENS-5% and ENS-CI.

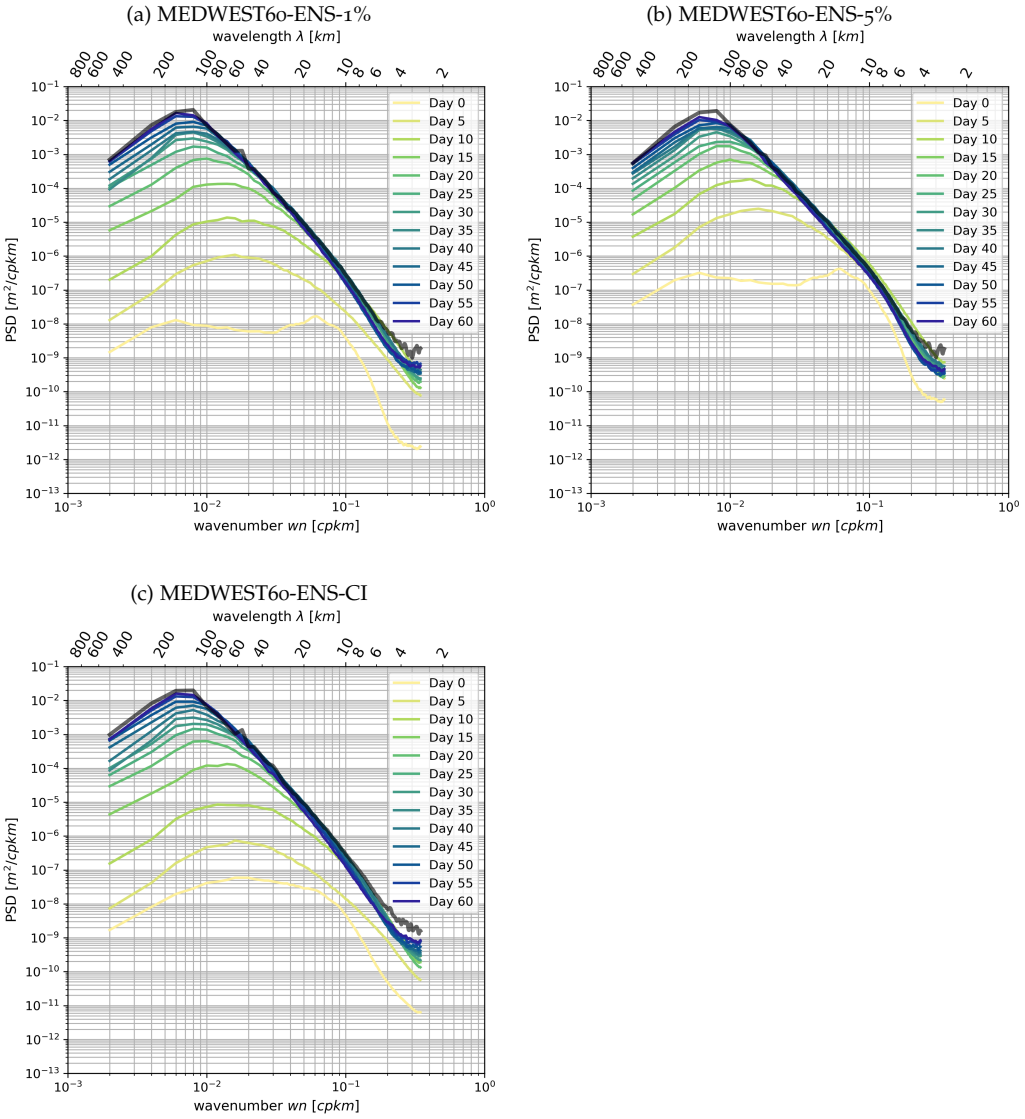


Figure 4.12: Ensemble-mean wavenumber power spectrum density (PSD) of the hourly SSH at day 60 (black thick line), compared to the mean PSD of the forecast error. The forecast error is assessed as the difference of the hourly SSH fields between all pairs of members in the same ensemble, and the mean is taken of the PSDs of all the 20x19 permuted pairs at each time (time increasing from yellow to blue colors). A factor 0.5 is applied to the mean PSD of those differences so that it can be compared in amplitude to the PSD of the full-field SSH (see text in section 5.3 for more details). The time-lag labeled "Day 0" is taken after 1 hour of the experiment. (a,b,c) correspond to the three experiments ENS-1%, ENS-5% and ENS-CI, resp.

5 RESULTS: PREDICTABILITY DIAGNOSTICS

In this section, we present and apply predictability diagnostics where predictability is quantified using a cross-validation algorithm (i.e. using alternatively each ensemble member as a reference truth and the remaining 19 members as ensemble forecast) together with a specific forecast score to quantify both the initial and forecast accuracy. From the joint distribution of initial and final scores, it is then possible to diagnose the probability distribution of the forecast score given the initial score, or reciprocally to derive conditions on the initial accuracy to obtain a target forecast skill. Although any specific score of practical significance could have been used, we focus here on simple and generic scores describing the misfit between ensemble members, in terms of overall accuracy (section 5.1: CRPS score), in terms of geographical position of the ocean structures (5.2: location score) and in terms of spatial decorrelation of the small-scale structures (section 5.3: Spectral score).

5.1 Probabilistic score

A standard approach to evaluate the skill of an ensemble forecast using reference data (Candille and Talagrand, 2005; Candille et al., 2007) is to compute probabilistic scores characterizing the statistical consistency with the reference (reliability of the ensemble) and the amount of reliable information it provides (resolution of the ensemble). For instance, in meteorology, ensemble forecasts can be evaluated a posteriori using the analysis as a reference. In the framework proposed in this study, a consistent approach to assess predictability is thus to compute the probabilistic scores that can be expected for given initial and model errors. In this case, we can use one of the ensemble members as a reference, by assuming that it corresponds to the true evolution of the system, and then compute the score using the remaining ensemble members as the ensemble forecast to be tested. Furthermore, by repeating the same computation with each ensemble member as a reference, as in a cross-validation algorithm, we can obtain a sample of the probability distribution for the score. All members of the ensemble are thus used successively as a possible truth, for which the other members provide an ensemble forecast. This procedure is very similar to the ensemble approach introduced in Germineaud et al. (2019) to evaluate the relative benefit of observation scenarios in a biogeochemical analysis system. In this framework, the probabilistic score can be viewed as a measure of the resulting skill of a given observation scenario.

5.1.1 *CRPS score*

A common measure of the misfit between two probability distributions of a one-dimensional random variable x is the area between their respective cumulative distribution functions (cdf) $F(x)$ and $F_{ref}(x)$:

$$\Delta = \int_{-\infty}^{\infty} \left| F(x) - F_{ref}(x) \right| \, dx \tag{10}$$

In our application, the reference cdf $F_{ref}(x)$ is a Heaviside function increasing by 1 at the true value of the variable, and the ensemble cdf $F(x)$ is a stepwise function increasing by $1/m$ at each of the ensemble values (where m is the size of the ensemble). Thus the further the ensemble values from the reference, the larger Δ, and the unit of Δ is the same as the unit of x.

The continuous rank probability score (CRPS) is then defined (Hersbach, 2000; Candille et al., 2015) as the expected value of Δ over a set of possibilities. In practical applications, the expected value is usually replaced by an average of Δ in space and time. In our application, the cross-validation algorithm would give the opportunity to make an ensemble average and thus be closer to the theoretical definition of CRPS. However, the ensemble size is here too small to provide an accurate local

value of CRPS, so that we prefer computing a spatial average as would be done in a real system, and compute an ensemble of spatially-averaged CRPS scores. In the following, CRPS scores will be computed by averaging over a specific subregion of the Mediterranean basin, south-east of the Balearic islands, as displayed in Fig. 5.5.

5.1.2 Evolution in time

Following the approach we proposed for this study, we start by studying the ensemble experiment that is peformed by applying perturbation on the initial condition only (i.e. experiment ENS-CI). The effect of model uncertainties will be diagnosed in a second step (in section 5.1.4).

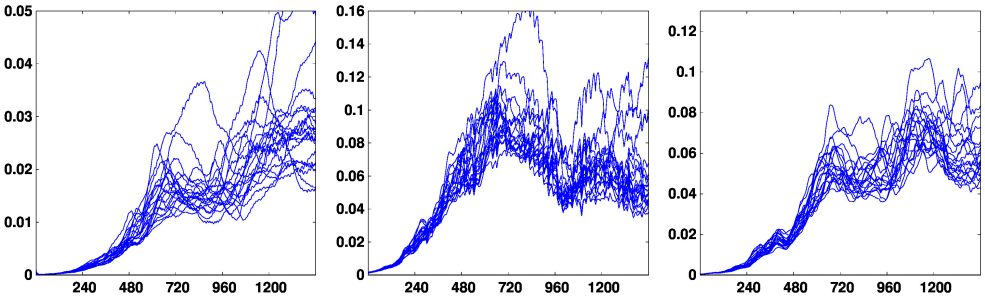


Figure 5.1: Time evolution of the CRPS score (y-axis) for SSH (meters), SST (degree Celsius) and SSS (psu) from experiment ENS-CI as computed using each ensemble member as a possible reference. Time (x-axis) is in hours, with a tickmark every 10 days.

Fig. 5.1 shows the time evolution of the CRPS score for SSH, SST and SSS (from left to right), as obtained in the experiment ENS-CI, i.e. with no model error. The CRPS score thus starts from zero and the initial increase is about exponential, with a doubling time of about 4 days. After typically 20 days, the evolution of the score becomes more irregular, globally increasing, but with possible decreases depending on the particular situation of the system. During the initial exponential increase, the diversity of possible evolutions of the score remains moderate: the score only increases a bit faster or a bit slower according to the member that is used as a reference. Afterwards, however, the evolution becomes very diverse, with the score sometimes increasing with time for a given reference member and decreasing for another reference member. This shows the importance of accounting for the diversity of possible situations in the description of predictability. With time, anomalous situations can emerge, which can produce different predictability patterns. Predictability thus needs to be described as a probability distribution of the score for given conditions of initial and/or model uncertainty.

5.1.3 Predictability diagrams

Using the ensemble time evolution of the CRPS score obtained in the previous section, it is then possible to describe predictability for a given time lag Δt by the joint distribution of the initial and final score $\mathrm{CRPS}(t)$ and $\mathrm{CRPS}(t+\Delta t)$. From this distribution, we can indeed obtain the conditional distribution of the final score given the initial score, and reciprocally the conditional distribution of the initial score required to obtain a given final score.

Fig. 5.2 describes predictability for 3 time lags $\Delta t = 2$, 5, and 10 days (from top to bottom), for SSH, SST and SSS (from left to right), as the CRPS score (y-axis) conditioned on the initial CRPS score (x-axis) for the same variable. The figure has been drawn for ENS-CI, i.e. without model uncertainties as in the previous section. This figure is just a reshuffling of the data from Fig. 5.1, gathering all couples of scores with time lag Δt. In analyzing this figure, it must be kept in mind that it mixes forecasts starting at a different initial time, which can correspond to

various situations of the system, in particular to different atmospheric forcings. The resulting probability distribution thus encompass this set of possibilities, the only conditions being on the time lag Δt and the initial CRPS score. To put a condition on the initial time would have required performing a large number of ensemble forecasts from that initial time with various levels of initial error, and would have been very substantially more expensive.

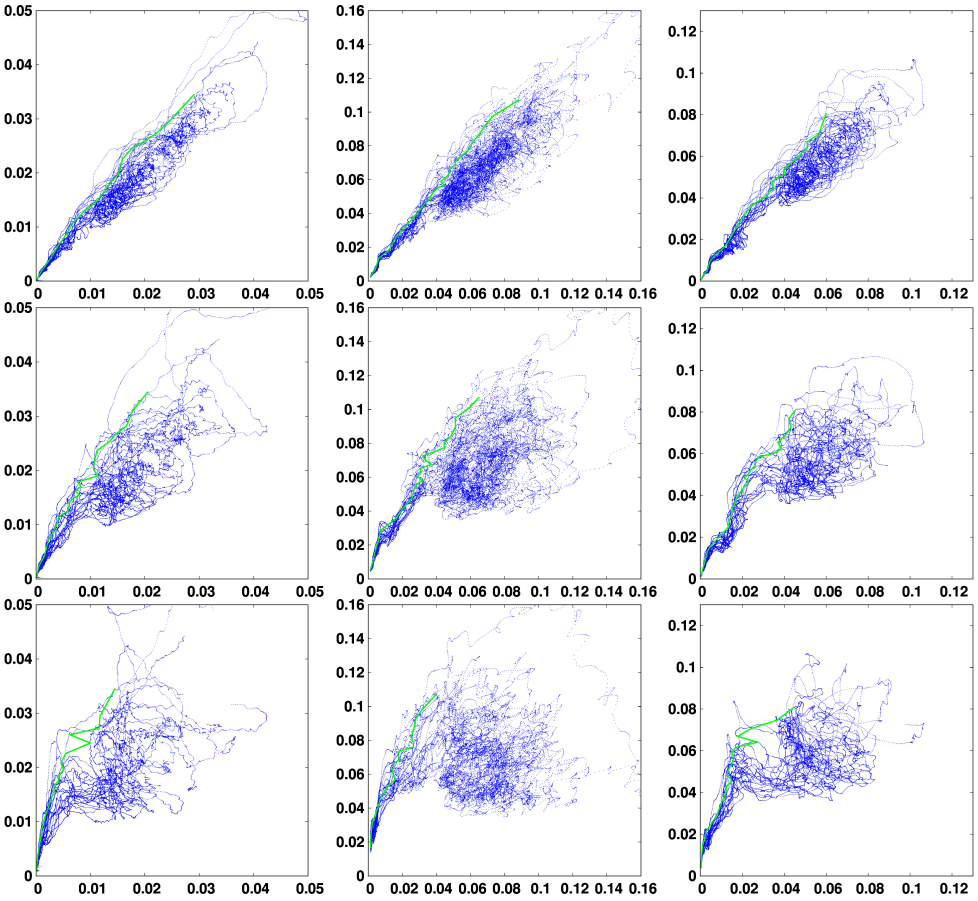


Figure 5.2: Final CRPS score (y-axis) as a function of the initial CRPS score (x-axis), for 3 time lags $\Delta t = 2$, 5, and 10 days (from top to bottom), for SSH (meters), SST (degree Celsius) and SSS (psu). The green line corresponds to the initial score required to have a 95% probability that the final score is below a given value.

The first thing to note from the figure is that for a given initial score, there can be a large variety of final scores after a Δt forecast, which again shows the importance of a probabilistic approach. What we obtain is a description of the probability distribution for the final score given the initial score, or reciprocally, the probability distribution of the initial score to obtain a required final accuracy. These are just two different cuts (along the y-axis or along the x-axis) in the two-dimensional probability distribution displayed in the figure. From this probability distribution, it is then possible to compute the initial score required to have a 95% probability that the final score is below a given value. This result, corresponding to the green curve in the figure, can be viewed as one possible answer to the question raised in the introduction about the initial accuracy required to obtain a given forecast accuracy.

5.1.4 *Effect of model uncertainties*

To explore the possible effect of model uncertainties (as represented by the stochastic scheme described in section 3) on predictability, we can compare the CRPS diagnostics described above for our three ensemble experiments: ENS-CI (no model

uncertainty), ENS-1% (small model uncertainty), and ENS-5% (larger model uncertainty). Fig. 5.3 shows first the time evolution of the CRPS score for these three experiments. In this figure, we observe that forecast uncertainties increase faster with model uncertainties included in the system (especially in ENS-5%), although the asymptotic behaviour of the score is very similar in all three simulations. Model uncertainties mainly matter for a short-range forecast (less than 10 days) when the initial condition is very accurate. Of course, this conclusion only holds for the kind of location uncertainty that we have introduced in NEMO, with short-range time and space correlation. A long standing effect of model uncertainties on predictability would be expected for large-scale perturbations, as in the atmospheric forcing for instance.

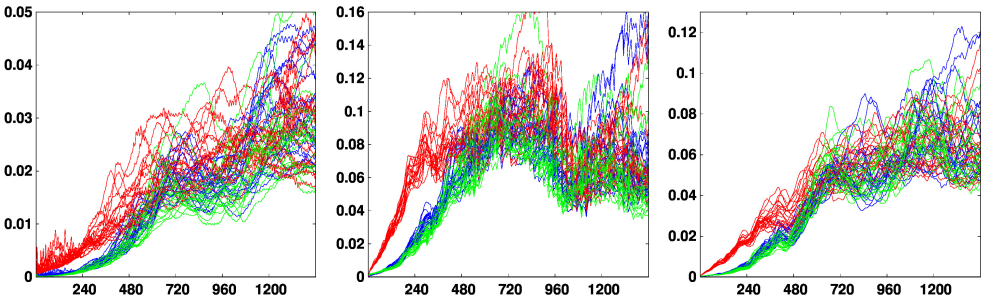


Figure 5.3: Time evolution of the CRPS score (y-axis) for SSH (meters), SST (degree Celsius) and SSS (psu), as computed using each ensemble member as a possible reference. The figure compares the three simulations ENS-CI (no model uncertainty, in green), ENS-1% (small model uncertainty, in blue), and ENS-5% (larger model uncertainty, in red). Time (x-axis) is in hours, with a tickmark every 10 days.

The consequence of this specific impact of model uncertainties is that the predictability diagrams displayed in Fig. 5.2 remain very similar for all three experiments, only becoming a bit more fussy when model uncertainties are included. To see the difference, we need to focus on the short time lag ($\Delta t = 2$ days) and on the small initial and final scores (which correspond to the beginning of the experiments). Fig. 5.4 compares the results obtained for SSH in ENS-CI, ENS-1% and ENS-5%, and we can observe that with larger model uncertainties, a smaller initial score (i.e. a more accurate initialization from observations) is generally needed to obtain a given final score (i.e. a given target of the forecasting system). If these model uncertainties are irreducible (as claimed in section 3 if they represent the effect of unresolved scales), they can thus represent an intrinsic limitation to predictability (at that resolution), at least in the specific case of a short time lag and a small initial error.

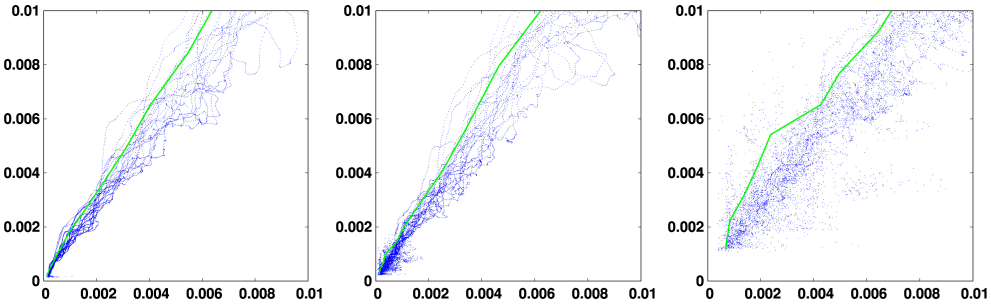


Figure 5.4: Final CRPS score (y-axis) as a function of the initial CRPS score (x-axis) for SSH (meters) and time lag $\Delta t = 2$ days. The green line corresponds to the initial score required to have a 95% probability that the final score is below a given value. The figure compares the three simulations ENS-CI (no model uncertainty, left panel), ENS-1% (small model uncertainty, middle), and ENS-5% (larger model uncertainty, right panel) for the small CRPS scores (smaller than 0.01 m).

5.1.5 Summary

- The CRPS score has been used to measure the discrepancy between an ensemble simulation and a reference truth. It is used to quantify the accuracy of the initial condition on the one hand (which corresponds in CMEMS to the accuracy that can be expected from a given observation and assimilation system), and the accuracy of the forecast on the other hand (which corresponds in CMEMS to the target accuracy of the forecasting system).
- A cross-validation algorithm has been used to obtain an ensemble of possible scores with one single ensemble simulation. This is done by using successively each member of the ensemble as reference truth and the remaining members (without the truth) as a forecast ensemble.
- A predictability diagram has been defined by the joint distribution of all initial and final scores for a given time lag. From this diagram, it is possible to derive (i) the forecast accuracy that can be expected from a given initial accuracy, and (ii) the initial accuracy that is required from the system to obtain a given forecast accuracy.
- As expected, the results show that the initial accuracy plays a major role in driving the forecast accuracy, but irreducible model uncertainties can also play a role for short time lags and accurate initial conditions. See the conclusions in section 6 where a numerical quantification is provided and discussed.

5.2 Location score

In the previous section, a probabilistic score has been used to describe the accuracy of the initial condition that can be associated to any given CMEMS observation/assimilation system. However, in many applications, what matters is not so much the accuracy of the value of the ocean variables, but the location of the ocean structures (fronts, eddies, filaments,...). Moreover, the acuteness of the positioning of ocean structures that can be obtained in the initial condition of the forecast can be thought to be more directly related to the resolution of the observation system that is available in CMEMS (in situ network or satellite imagery).

For these reasons, in this section, we will introduce a simple measure of location uncertainties in an ensemble forecast, which will be used in the same way as the CRPS score in the previous section. The same type of diagnostics will be computed to provide a similar description of predictability, but from a different perspective.

5.2.1 Misfit in field locations

To obtain a simple quantification of the position misfit between two ocean fields (one ensemble member and a reference truth), we are looking for an algorithm to compute at what distance the true value of the field can be found. Ideally, what we would like is to find the minimum displacement that would be needed to transform a given ensemble member into the reference truth. However, it is important to remark that this does not amount to computing the distance between corresponding structures in the two fields. This would indeed require an automatic tool to identify homolog structures in the two fields and would be much more difficult to achieve in practice. In general, if the two fields are not close enough to each other, such identification would even be impossible, since ocean structures can merge, appear, disappear or be transformed to such extent that no one-to-one correspondence can be found.

In addition, to further simplify the problem, we do not consider the original continuous fields, but modified fields that have been quantized on a finite set of values. Fig. 5.5 shows for instance the salinity field from two members of the ENSCI simulation (after 15 days), together with their quantized version. The quantized

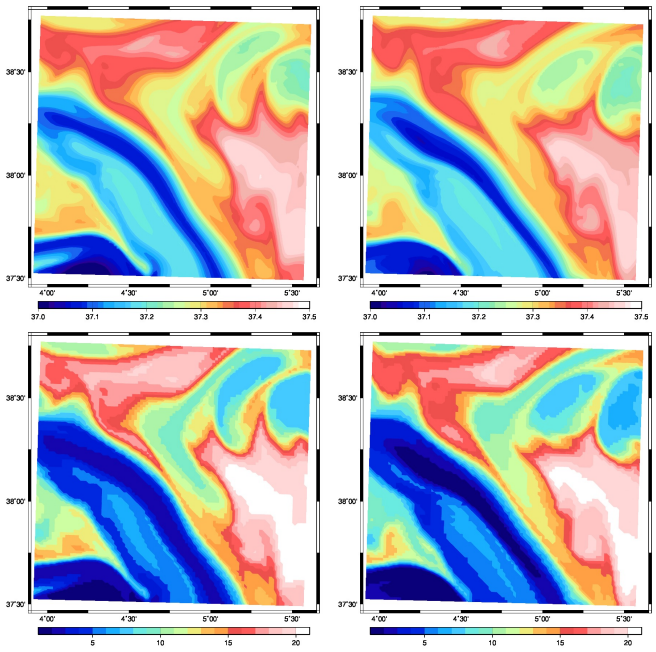


Figure 5.5: Surface salinity fields from two members of ENS-CI after 15 days. The member used as reference truth is displayed in the left panels. The figure displays the original continuous fields (top panels) and their quantized version (bottom panels) in a subregion of the MEDWEST60 domain in the south east of the Balearic Islands.

version is obtained by computing the quantiles of the reference truth (left panel), for instance 19 quantiles here (from the distribution of all values in the map), and then by replacing the value of the continuous field by the index of the quantile interval to which it belongs (between 1 and 20). In this case, a value of 1 means that the field is below the 5% quantile and a value of 20 means that the field is above the 95% quantile. From these quantized fields, it is then easy to find the closest point where the index is equal to that of the reference truth, and thus where the field itself is close to the truth (to a degree that can be tuned by changing the number of quantiles).

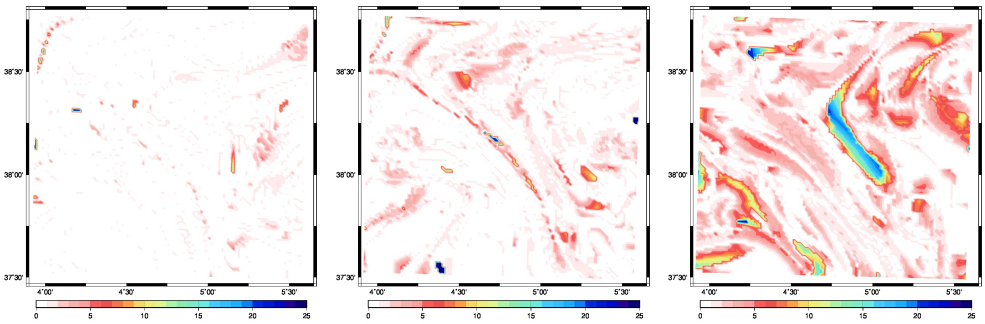


Figure 5.6: Location misfit (in km) between the surface salinity fields from two members of ENS-CI after 5, 10 and 15 days (from left to right).

Fig. 5.6 shows the resulting maps of location misfit for salinity in ENS-CI after 5, 10 and 15 days. We see that the location misfit is increasing with time as the two ensemble members diverge from each other. From such a map, it is then possible to define a single score from the distribution of distances. In this study, the score is defined as the 95% quantile of this distribution, which means that location error has a 95% probability to be below the distance given by the score. In the present study, this score is used as a diagnostic tool, but it may be useful to remark that

it could also be used as an additional constraint in the cost function of the assimilation system, for instance to take a better benefit of the accurate observation of the position of ocean structures in the high-resolution satellite images (similarly to what is attempted in the work of Gaultier et al., 2014; Durán Moro et al., 2017).

5.2.2 Evolution in time

As in section 5.1, we start by analyzing the time evolution of the score in the ensemble experiment ENS-CI, where the only source of uncertainty comes from the initial conditions. In this case, there is thus no model uncertainty; the model is deterministic.

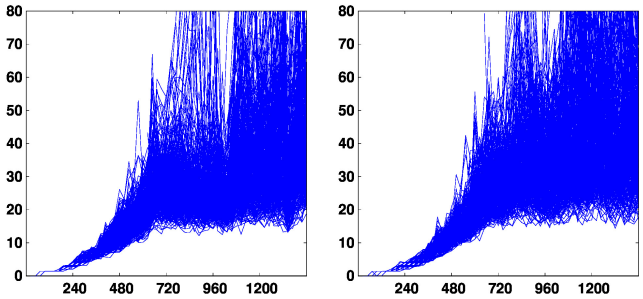


Figure 5.7: Time evolution of the location score (y-axis, in km) for SST and SSS, as computed using each pair of members from the ensemble. Time (x-axis) is in hours, with a tickmark every 10 days.

Fig. 5.9 shows the time evolution of the location score for SST (left panel) and SSS (right panel) for each pair of members in the ensemble, which provides a total of $m(m-1) = 20 \times 19 = 380$ curves displayed in the figure. The first thing that we observe in the figure is that the distribution of time evolutions is about the same for SST and SSS, which indicates that our measure of location uncertainty is consistent for the two tracers. Second, we see that during about the first half of the experiment (the first 30 days), the location score is increasing towards saturation, with a spread that is also increasing with time, whereas in the second half of the experiment, the score has reached the asymptotic distribution, which is characterized by a large location uncertainty and a large spread of the score. This means that there is no more information about the location of the ocean structures in the forecast and that the score can be either moderate (down to 20 km) or very large (up to 80 km and more) depending on chance. In the following, we thus mostly focus to the range of scores, between 0 and 20 km, where a valuable forecast skill can be expected (for the small scale tracer structures that are resolved by our model).

5.2.3 Predicatbility diagrams

From the time evolution of the score described in the previous section, we can then deduce predictability diagrams, using exactly the same approach that was used in section 5.1.3. Fig. 5.8 describes predictability (computed from SST fields) for 6 time lags $\Delta t = 1$, 2, 5, 10, 15 and 20 days), by showing the final location score (y-axis) as a function of the initial location score (x-axis). This figure is just a reshuffling of the data from Fig. 5.9 (left panel), gathering all couples of scores with time lag Δt, using the same assumption already discussed in section 5.1.3. Note that the longest time-lags considered here (>10 days) are relevant only in the present context of forced ocean experiments (as a forecasted atmosphere would also become a major source of uncertainty for ocean predictability in a real operational forecast context at those time lags).

The interpretation of the figure also follows the same logic. However, the structure of the diagrams is here more directly understandable, and the loss of predictability with time can be more easily followed. For instance, if one seeks a fore-

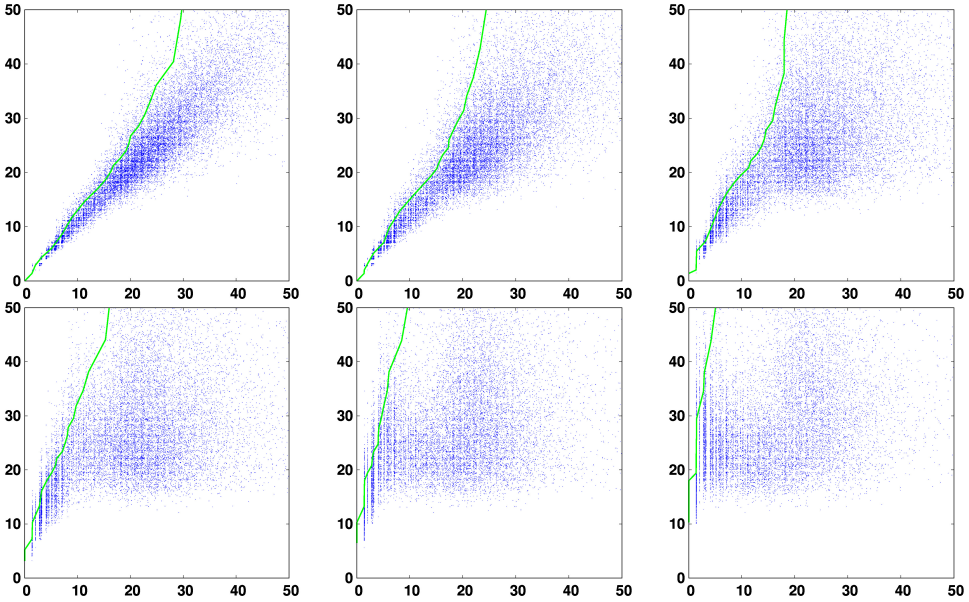


Figure 5.8: Final SST location score (y-axis, in km) as a function of the initial SST location score (x-axis, in km), for 6 time lags $\Delta t = 1, 2, 5, 10, 15$ and 20 days (from top left to bottom right). The green line corresponds to the initial score required to have a 95% probability that the final score is below a given value.

cast accuracy of 10 km with a 95% confidence (i.e. a y-value of the green curve equal to 10 km), then the figure tells that the initial location accuracy required (necessary condition, but not sufficient, see conclusions) is about 8 km for a 1-day forecast, 6 km for a 2-day forecast, 4 km for a 5-day forecast, 2 km for a 10-day forecast, and that this target is impossible to achieve in a 15-day and 20-day forecast. It must be noted however that the last two impossibilities may result from the absence of small enough initial errors in our sample (since ENS-CI was initialized using ENS-1% after 1 day), but this should not make any practical difference since such small initial errors would anyway be impossible to obtain in a real system.

5.2.4 *Effect of model uncertainties*

As for the CRPS score, to explore the possible effect of model uncertainties, we just compare ENS-CI (no model uncertainties) with ENS-1% (small model uncertainties) and ENS-5% (larger model uncertainties). Fig. 5.9 compares first the time evolution of the location score for ENS-CI (in blue) and ENS-5% (in red), and we observe again that model uncertainties mainly matter at the beginning of the simulation by a faster increase of the forecast uncertainties, towards a similar asymptotic behaviour for the two simulations.

As for the CRPS score, the predictability diagrams are thus only substantially different for short time lags and small initial and final scores. This is illustrated in Fig. 5.10 by comparing the diagrams obtained for ENS-CI, ENS-1% and ENS-5% for $\Delta t = 5$ days and scores below 20 km. Again, we can observe hre a moderate effect of model uncertainties (as simulated here) on predictability. For instance, if one seeks a forecast accuracy of 10 km with a 95% confidence, the initial location accuracy required decreases from about 4 km in ENS-CI to about 3 km in ENS-5%.

5.2.5 *Summary*

- A location score has been defined to measure the location misfit between an ensemble member and a reference truth. It is used to quantify the accuracy of the initial condition that can be expected from the resolution of the observation system, and the target location accuracy in the resulting forecast.

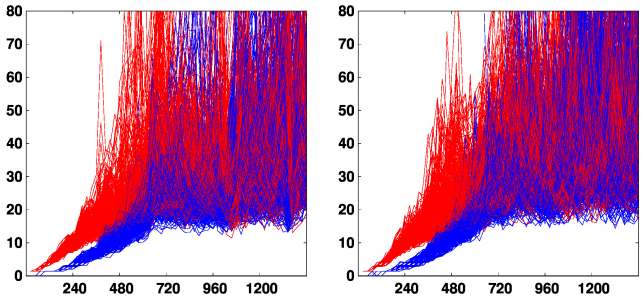


Figure 5.9: Time evolution of the location score (y-axis, in km) for SST and SSS, as computed using each pair of members from the ensemble. The figure compares the two simulations ENS-CI (no model uncertainty, in blue), and ENS-5% (large model uncertainty, in red). Time (x-axis) is in hours, with a tickmark every 10 days.

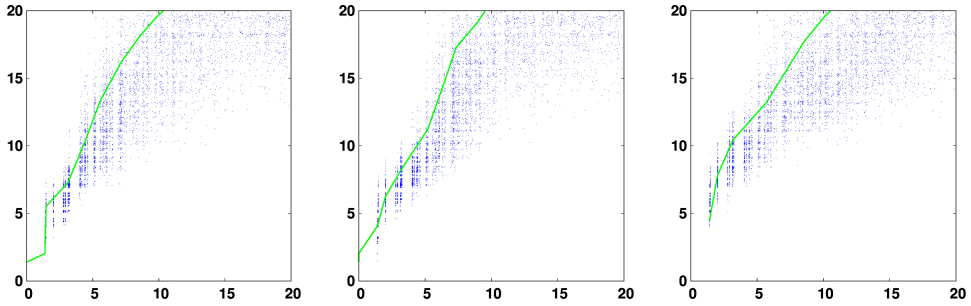


Figure 5.10: Final location score (y-axis, in km) as a function of the initial location score (x-axis, in km) for SST and time lag $\Delta t = 5$ days. The green line corresponds to the initial score required to have a 95% probability that the final score is below a given value. The figure compares the three simulations ENS-CI (no model uncertainty, left panel), ENS-1% (small model uncertainty, middle), and ENS-5% (larger model uncertainty, right panel) for the small location scores (smaller than 20 km).

- Previously defined predictability diagrams (using a cross-validation algorithm) have been used here to describe the joint distribution of initial and final location scores, thus relating the forecast location accuracy to the location accuracy of the initial condition (and thus the resolution of the ocean observing system).
- As expected, the results show that the initial location accuracy plays a major role in driving the forecast location accuracy, but irreducible model uncertainties can also play a role for short time lags and accurate initial conditions. See the conclusions in section 6 where a numerical quantification is provided and discussed.

5.3 Spectral spatial decorrelation

A RATIO R TO QUANTIFY THE SPATIAL SPECTRAL DECORRELATION :
To complement the information provided by the location score above on the "misfit" of the ocean structures, we also investigate here some diagnostics of the spatial spectral decorrelation of the ensemble members. The idea is to compare the spectral content of the the forecast error to the spectral content of the reference field (here considering SSH). The forecast error is assessed as the difference of the hourly SSH fields between a given member taken as the truth and another member considered as the forecast. All the 20x19 combinations of pairs are alternatively considered, following the same cross-validation algorithm as described for the CRPS score above. The "misfit" of the ocean structures is here quantified in spectral space with a ratio R computed for each time-lag as:

$$R = 1 - \frac{< PSD_{diffssh} >}{2\times < PSD_{ssh} >}, \tag{11}$$

where PSD_{ssh} is the Power Spectral Density of the full-field SSH at that given time-lag, and $PSD_{diffssh}$ is the PSD of the forecast error on SSH at that given-time-lag. The brackets $<\ldots>$ denote the ensemble mean operation over the 20 members or over the 20×19 combinations of pairs. By design, R is expected to tend to zero when the ensemble members are fully decorrelated, and to be close to 1 when the members are fully correlated. The factor 2 in the definition of R comes from the fact we compare here the PSD of a difference of two given fields with the PSD of the reference field. For exemple, in the case that the ensemble members are strictly independent and uncorrelated in space on all scales, then for all combinations of a pair of members (t, f) where t would be considered the truth and f the forecast, the space variance (var) of the difference $f-t$ can be expressed as :

$$< var(f-t) > =< var(f) + var(t) - 2covar(f,t) >, \tag{12}$$
$$< var(f-t) > =< var(f) > + < var(t) >, \tag{13}$$
$$< var(f-t) > = 2 < var(f) >, \tag{14}$$

where the factor 2 appears.

EVOLUTION IN TIME AND PREDICTABILITY DIAGRAMS :
In section 4.3, we had already discussed the evolution with time of the spatial spectral content of the forecast error (Figure 4.12). Now Figure 5.11 shows the ratio R, computed at different time-lags from experiment ENS-CI (top panel). By design, values of R are close to 1 when the members are strongly correlated: this is indeed the case on the figure, at very short time lags (<5 days, yellow line). With time increasing, R decreases (the members are less and less spatially correlated), starting from small scales and cascading to larger scales. At the end of the 2-month experiment, R has decreased to zero for scales in the range 10-60 km, consistently with what was seen from Fig. 4.12. Full decorrelation is not yet reached for larger scales, but we don't necessarily expect a full spatial decorrelation between the members in

this type of experiment since all members see the same surface forcing and lateral boundary conditions. Also, note that the size of the box on which the spatial spectral analysis is performed is about 350 km square, so the left hand of the spectrum is not expected to be much significant for scales larger than 150 km (aliasing effect, also see Fig.4.3 and associated text).

On the right-hand side of the spectrum, on very small scales (<10 km), it is noteworthy that R remains larger than 0.5 after 2 months of simulation, suggesting that those scales remain somehow spatially-correlated at long time-lags. This behavior is clearly consistent in the three experiments (see panels a,b,c in Fig.5.11), so it cannot just result from a spurious effect of the stochastic perturbation (which is not present in experiment ENS-CI). The main difference between the three experiments is that R in ENS-5% decreases faster to zero than the two others (consistently with the fact that it undergoes the largest spread). Specific investigations would be needed to understand the reason why these very small scales remain somehow correlated. Note however that the range of wavelengths here (<10 km) concerned scales that are not fully resolved at the resolution of the model ($1/60^o$). This persistent correlation could arise for instance from a systematic numerical noise on the grid scale of the model. Note that it is also seen from relative vorticity spectral ratio (not shown) but limited to even smaller wavelengths (<5 km).

We finally consider the mean ratio $\overline{R}$, averaged over two given ranges of scales (10-30 km and 60-100 km) from experiment ENS-CI to provide an example of predictability diagram (Fig. 5.12), following the same methodology as for the CRPS and location scores. The value of $\overline{R}$ after a given forecast time-lag, $\overline{R}(t+\Delta)$, where Δ is the time-lag, is plotted as a function of the initial value $\overline{R}(t)$. The figure provides, for each given scale range ((a) 10-30 km and (b) 60-100 km), some objective information about the spatial decorrelation between the members, in the case of a perfect model.

In the 10-30 km scale range for example, it appears that even with very small initial errors (initial R close to 1), the members become nearly decorrelated after a time-lag of ~10 days (i.e. $\overline{R}(t+\Delta)<0.5$) on these scales. For larger scales, in the range 60-100 km, the threshold of $\overline{R}(t+\Delta)<0.5$ is reached for time lags above ~15 days. Note however that only the uncertainty on initial conditions is taken into account here. A faster decorrelation would be expected if other types of uncertainties in the forecast system were taken into account, such as uncertainty on the atmospheric forcing.

These type of predictability diagnostics in Fig. 5.12 might also be relevant in the context of preparing for the assimilation of wide-swath high-resolution satellite altimetry such as expected from the future SWOT mission (Fu and Ferrari, 2008). This mission is expected to measure sea surface height (SSH) with high-precision and resolve short mesoscale structures as small as 15 km on a wide swath of 120 km. However the time interval between revisits will be within 11 to 22 days, depending on the location. Our results above tend to show that for time-lags longer than 10 days, the forecasting system considered in the present study will lose most of the information in the initial condition regarding SSH structures in the smallest scale range (10-30 km).

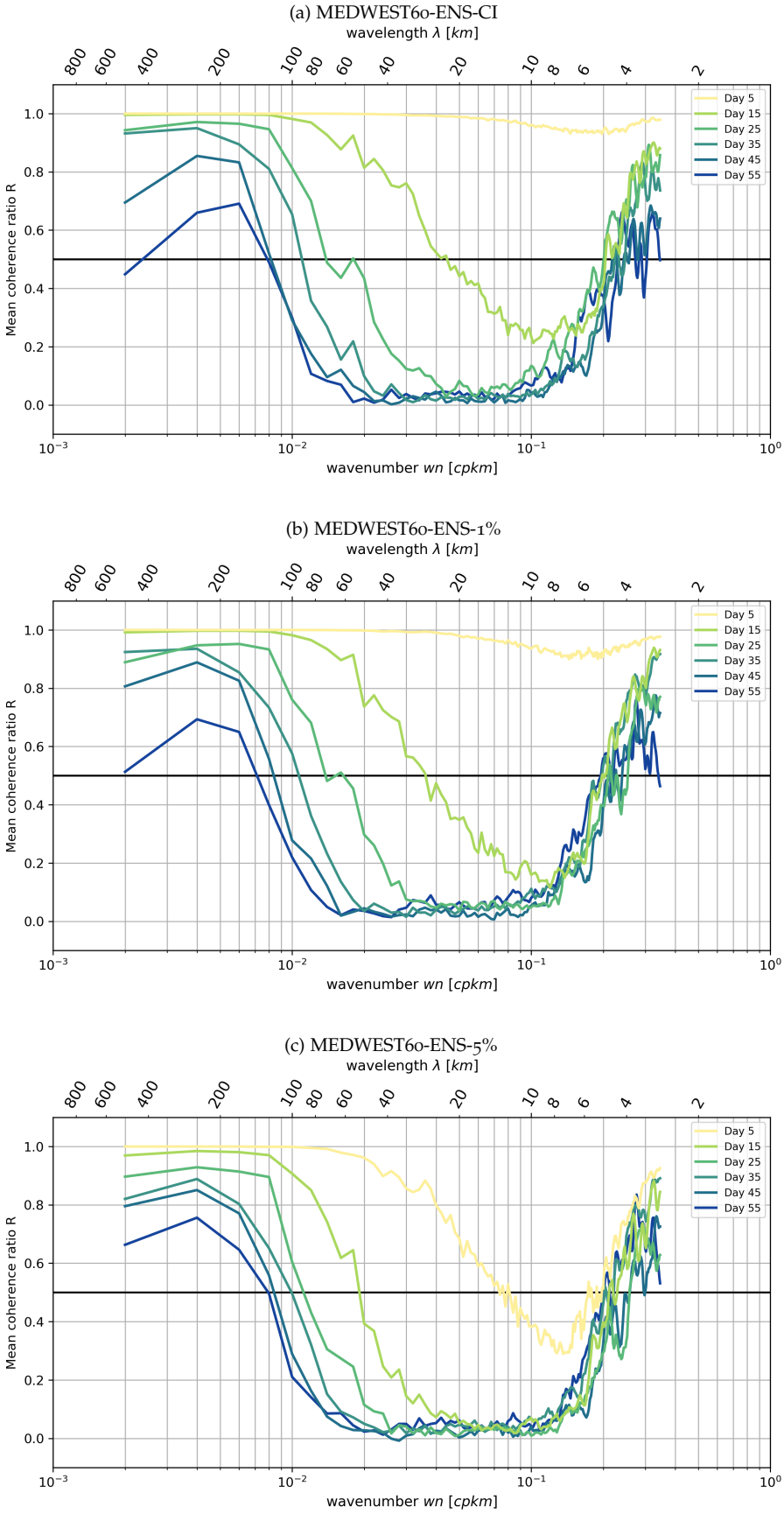
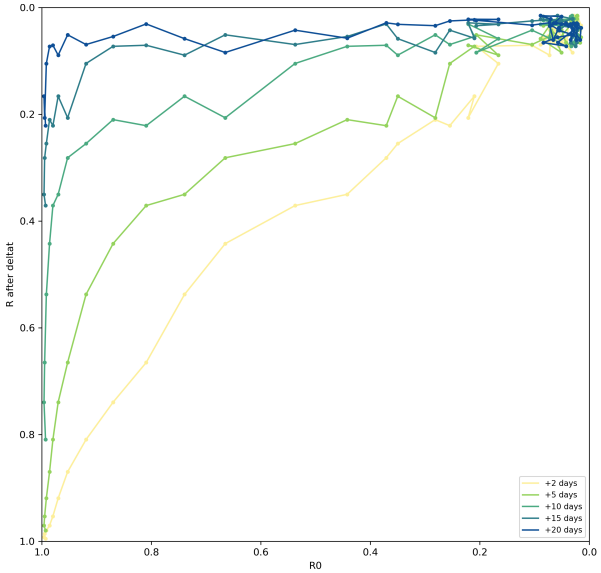


Figure 5.11: Mean coherence ratio R (see text for definition) from experiments ENS-CI, ENS-1% and ENS-5%. The ratio is computed at different time-lags: time increasing from yellow to blue colors.

(a) λ in 10-30 km



(a) λ in 60-100 km

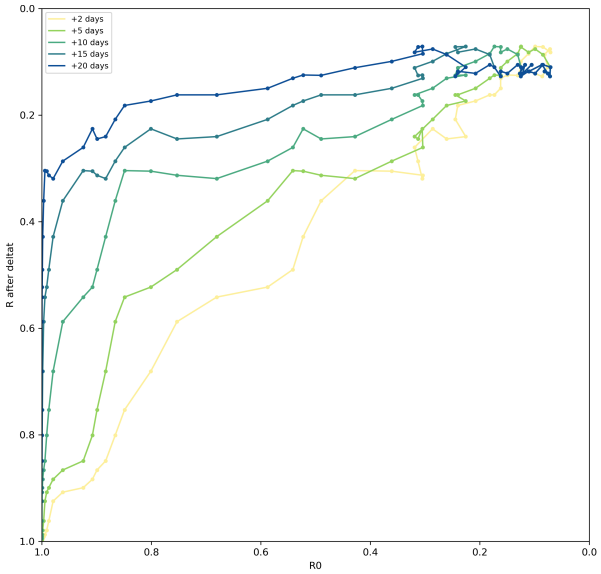


Figure 5.12: Mean wavenumber spectral coherence ratio R of the ensemble forecast as a function of the coherence of the ensemble initial conditions, for different forecast time-lags (+2,5,10,15,20 days), computed from hourly SSH in experiment ENS-CI. The mean ratio R is taken over scales of (a) 10-30 km and (b) 60-100 km in (b).

6 CONCLUSION

The general objective of this study was to quantify how much of the information in the initial condition, acquired from observations, a high-resolution NEMO modelling system is able to correctly retain and propagate during a short and medium range forecast.

For that purpose, a kilometric-scale regional configuration of NEMO for the Western Mediterranean (MEDWEST60, at 1/60° horizontal resolution) has been developed. It has been defined as a subregion of a larger North Atlantic configuration (eNATL60), which provides the boundary conditions. This deterministic model has then been transformed into a probabilistic model by introducing an innovative stochastic parameterization of location uncertainties in the horizontal displacements of the fluid parcels. The purpose was primarily to generate ensemble of initial conditions to be used in the predictability studies, but it has also been applied to assess the possible impact of irreducible model uncertainties on the skill of the forecast.

With this model configuration, 20-member and 2-month ensemble experiments have been performed, first with the stochastic model for two levels of model uncertainty, and then with the deterministic model from perturbed initial conditions. In all experiments, the spread of the ensemble emerges from the small scales (10 km wavelegnth) to progressively upscale to the largest structures. After two months, the ensemble variance has saturated over most of the spectrum (except the largest scales), whereas the small scales (1-10 km) are fully decorrelated between different members. For these scales, these ensemble simulations were thus appropriate to provide a statistical description of the dependence between initial accuracy and forecast accuracy over the full range of potentially useful forecast time lags (typically, between 1 and 20 days).

From these experiments, predictability has then been statistically quantified using a cross-validation algorithm (i.e. using alternatively each ensemble member as a reference truth and the remaining 19 members as forecast ensemble) together with a specific score to characterize the initial and forecast accuracy. From the joint distribution of initial and final scores, it was then possible to diagnose the probability distribution of the forecast score given the initial score, or reciprocally to derive conditions on the initial accuracy to obtain a target forecast skill. Although any specific score of practical significance could have been used, we focused here on simple and generic scores describing the msifit between ensemble members in terms of overal accuracy (CRPS score) or in terms of geographical position of the ocean structures (location score).

	2 days	5 days	10 days
0.025	0.016	0.006	0.001
0.05	0.037	0.027	0.010
0.075	0.056	0.039	0.023
0.1	0.077	0.059	0.033

Table 6.1: Initial SST accuracy required (CRPS score, in °C) to obtain the target final accuracy (CRPS score, in °C, left column) with a 95% confidence for different forecast time lags: 2 days, 5 days and 10 days.

Tables 6.1 and 6.2 illustrate conditions obtained on the initial accuracy to obtain a given forecast accuracy if the model is assumed perfect (as in ENS-CI), using the CRPS score and the location score. For example, Table 6.2 shows that, for our particular region and period of interest, the initial location accuracy required with a perfect model (deterministic operator) to obtain a forecast location accuracy of 10 km with a 95% confidence is about 8 km for a 1-day forecast, 6 km for a 2-day forecast, 4 km for a 5-day forecast, 1.5 km for a 10-day forecast, and that this target is unreachable for a 15-day and a 20-day forecast (more precisely, in these two cases, the required initial accuracy would be irrealistically small and was not included in

our sample). With model uncertainties (stochastic operator, as in ENS-1% or ENS-5%), the requirement on the initial condition can be even more stringent, especially for a short-range and high-accuracy forecast.

	1 day	2 days	5 days	10 days	15 days	20 days
2 km	1.6 km	1.4 km	—	—	—	—
5 km	3.9 km	3.1 km	1.4 km	—	—	—
10 km	7.9 km	6.2 km	4.4 km	1.4 km	—	—
15 km	11.7 km	10.4 km	6.3 km	3.1 km	1.4 km	—
20 km	16.2 km	14.9 km	10.5 km	5.4 km	2.3 km	1.4 km

Table 6.2: Initial location accuracy required (location score) to obtain the target final location accuracy (location score, left column) with a 95% confidence for different forecast time lags between 1 day and 20 days.

However, it is important to remark that this only provides necessary conditions but not a sufficient condition on the initial model state. The reason for that is that the condition is put on one single score for one single variable, whereas the quality of the forecast obviously depends on the accuracy of all variables in the model state vector. In the examples given in the tables, we used the same model variable for both target score and the condition score, but we could have looked as well for a necessary condition on another variable (for instance velocity) to obtain a given forecast accuracy for SST or any other model diagnostic. In this way, for any forecast target, we could have accumulated many necessary conditions on various key properties of the initial conditions, especially observed properties, but this would never become a sufficient condition.

Furthermore, these necessary conditions on observed quantities can then be translated into conditions on the design of the ocean observing system, in terms of accuracy and resolution, if a given forecast accuracy is to be expected. In this case, again, this can still obviously be necessary conditions, because the accuracy of the initial model state also depends on the ability of the assimilation system to interpret properly the observed information and to produce an appropriate initial condition for the forecast. Checking this ability would have required performing observation system simulation experiments (OSSE) using the operational assimilation system, and this was clearly out of the scope of the present work.

More generally, however, what this study suggests is that an ensemble forecasting framework should become an important component of CMEMS systems to provide a systematic statistical quantification of the relation between the system operational target (a useful forecast skill) and the available assets: the observation systems, with their expected resolution and accuracy, and the modelling tools, with their target resolution and associated irreducible uncertainties.

7 ACKNOWLEDMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 821926. The work was performed using HPC resources from GENCI-IDRIS, France (Grant A008-0101279). The MEDWEST60 configuration was set up, initialized and forced at the lateral boundaries using eNATL60 data from (Brodeau et al., 2020)). The CDFTOOLS (https://github.com/meom-group/CDFTOOLS) were used for some of the pre- and post-processing of the model outputs. The Power Spectral Density computation in sections 4.3 and 5.3 were all performed using A. Ajayi's python module PowerSpec (https://github.com/adeajayi- kunle/powerspec) and the predictability diagnostics based on CRPS score and location score were made with the SESAM

(https://github.com/brankart/sesam) and EnsDAM (https://github.com/brankart/ensdam) softwares.

8 MEDWEST60 SOURCE CODES AND DIAGNOSTICS

The source codes of the MEDWEST60 NEMO configuration and some of the diagnostics developed in this study are shared on github in a repository dedicated to MEDWEST60: `https://github.com/ocean-next/MEDWEST60`.

REFERENCES

Berner J., T. Jung and T.N. Palmer, 2012: Systematic model error: the impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *Journal of Climate*, **25**, 4946–4962.

Bessières L., S. Leroux, J.-M. Brankart, J.-M. Molines, M.-P. Moine, P.-A. Bouttier, T. Penduff, L. Terray, B. Barnier, and G. Sérazin, 2017: Development of a probabilistic ocean modelling system based on NEMO 3.5: application at eddying resolution. *Geoscientific Model Development*, **10(3)**, 1091–1106.

Brankart J.-M., 2013: Impact of uncertainties in the horizontal density gradient upon low resolution global ocean modelling. *Ocean Modelling*, **66**, 64–76.

Brankart J.-M., G. Candille, F. Garnier, C. Calone, A. Melet, P.-A. Bouttier, P. Brasseur and J. Verron, 2015: A generic approach to explicit simulation of uncertainty in the NEMO ocean model, *Geoscientific Model Development*, **8**, 1285–1297.

Brasseur, P. and Blayo, E. and Verron, J., 1996: Predictability experiments in the North Atlantic Ocean: Outcome of a quasi-geostrophic model with assimilation of TOPEX/POSEIDON altimeter data. *Journal of Geophysical Research: Oceans*, **101, C6**, 14161-14173, https://doi.org/10.1029/96JC00665.

Brodeau, L., J. Le Sommer and A. Albert, 2020: Ocean-next/eNATL60: Material describing the set-up and the assessment of NEMO-eNATL60 simulations (Version v1). *Zenodo*: http://doi.org/10.5281/zenodo.4032732.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quaterly Journal of the Royal Meteorological Society*, **125**, 2887–2908.

Candille G., and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. Quart. J. Roy. Meteor. Soc., **131**, 2131–2150.

Candille G., C. Côté, P. L. Houtekamer, and G. Pellerin, 2007: Verification of an ensemble prediction system against observations. Mon. Wea. Rev., **135**, 2688–2699.

Candille G., J.-M. Brankart J and P. Brasseur, 2015: Assessment of an ensemble system that assimilates Jason-1/Envisat altimeter data in a probabilistic model of the North Atlantic ocean circulation. *Ocean Science*, **11**, 425–438.

Chapron B., P. Dérian, E. Mémin, V. Resseguieri, 2018: Large scale flows under location uncertainty: a consistent stochastic framework, *Quarterly Journal of the Royal Meteorological Society*, **144(710)**, 251–260.

Diaconescu E. P. and R. Laprise, 2012: Singular vectors in atmospheric sciences: A review. *Earth-Science Reviews*, **113(3–4)**, 161–175.

Durán Moro M., Brankart J.-M., Brasseur P. and Verron J., 2017: Exploring image data assimilation in the prospect of high-resolution satellite oceanic observations. *Ocean Dynamics*, **87(7)**, 875–895.

Durand, M., L. Fu, D. Lettenmaier, D. Alsdorf, E. Rodriguez, and D. Esteban-Fernandez, 2010: The surface water and ocean topography mission: Observing terrestrial surface water and oceanic submesoscale eddies. *Proceedings of the IEEE*, **98 (5)**, 766–779.

Evensen, G. 1994: Sequential data assimilation with a non linear quasigeostrophic model using Monte Carlo methods to forecast error statistics. *J. of Geophys. Res.*, **99(C5)**, 10143–10162.

Frederiksen, J., T. O'Kane, and M. Zidikheri, 2012: Stochastic subgrid parameterizations for atmospheric and oceanic flows. *Physica Scripta*, **85**, 068 202.

Fu, L.-L., and R. Ferrari, 2008: Observing oceanic submesoscale processes from space. *Eos, Transactions American Geophysical Union*, **89 (48)**, 488–488.

Garnier F., J.-M. Brankart, P. Brasseur and E. Cosme, 2016: Stochastic parameterizations of biogeochemical uncertainties in a 1/4° NEMO/PISCES model for probabilistic comparisons with ocean color data. *Journal of Marine Systems*, **155**, 59–72.

Gaultier L., Djath B., Verron J., Brankart J.-M., Brasseur P. and Melet A., 2014: Inversion of submesoscale patterns from a high-resolution Solomon Sea model: feasibility assessment. *Journal of Geophysical Research*, **119(7)**, 4520–4541.

Germineaud, C., J.-M. Brankart, and P. Brasseur, 2019: An Ensemble-Based Probabilistic Score Approach to Compare Observation Scenarios: An Application to Biogeochemical-Argo Deployments. J. Atmos. Oceanic Technol., 36, 2307-2326.

Hawkins E., R.S. Smith, J.M. Gregory and D.A. Stainforth, 2016: Irreducible uncertainty in near-term climate projections. *Clim Dyn* **46**, 3807–3819.

Hersbach H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *W*ea. Forecasting, **1**5, 559–570.

Juricke, S., P. Lemke, R. Timmermann. and T. Rackow, 2013: Effects of stochastic ice strength perturbation on Arctic finite element sea ice modeling, *Journal of Climate*, **26**, 3785–3802.

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

Lacarra J. and O. Talagrand, 1988: Short-range evolution of small perturbations in a barotropic model, *Tellus*, **40A**, 81–95

Leroux S., T. Penduff, L. Bessières, J.-M. Molines, J.-M. Brankart, G. Sérazin, B. Barnier, and L. Terray, 2018: Intrinsic and Atmospherically Forced Variability of the AMOC: Insights from a Large-Ensemble Ocean Hindcast. *J. Climate*, **31**, 1183–1203.

Leutbecher M., Lock S., Ollinaho P., Lang S.T., Balsamo G., Bechtold P., Bonavita M., Christensen H.M., Diamantakis M., Dutra E., English S., Fisher M., Forbes R.M., Goddard J., Haiden T., Hogan R.J., Juricke S., Lawrence H., MacLeod D., Magnusson L., Malardel S., Massart S., Sandu I., Smolarkiewicz P.K., Subramanian A., Vitart F., Wedi N. and Weisheimer A., 2017: Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, **143**, 2315–2339.

Lorenz E.N., 1965: A study of the predictability of a 28-variable atmospheric model, *Tellus*, **17**, 321–333.

Lyapunov A., 1992: The general problem of the stability of motion. *International Journal of Control*, **55:3**, 531–534.

Mémin E., 2014: Fluid flow dynamics under location uncertainty, *Geophysical and Astrophysical Fluid Dynamics*, **108(2)**, 119–146.

Palmer, T.N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Q.J.R. Meteorol. Soc.*, **128**, 747–774.

Palmer T., G. Shutts, R. Hagedorn, F. Doblas-Reyes, T. Jung, and M. Leutbecher, 2005: Representing model uncertainty in weather and climate prediction. *Annu. Rev. Earth Planet. Sci.*, **33**, 163–193.

Palmer T. and R. Hagedorn (Eds.), 2006: Predictability of weather and climate. Cambridge University Press.

Sakov P., Counillon F., Bertino L., Lisæter K.A., Oke P.R. and Korablev, A., 2012: TOPAZ4: an ocean-sea ice data assimilation system for the North Atlantic and Arctic, *Ocean Science*, **8**, 633–656.

Toth Z. and E. Kalnay, 1993: Ensemble Forecasting at NMC: The Generation of Perturbations. *Bulletin of the American Meteorological Society*, **74(12)**, 2317–2330.

Williams P.D., N.J. Howe, J.M. Gregory, R.S. Smith, and M.M. Joshi, 2016: Improved climate simulations through a stochastic parametrization of ocean eddies. *Journal of Climate*, **29(24)**, 8763–8781.

Zanna L., J.-M. Brankart J.-M., M. Huber, S.Leroux, T. Penduff and P. D. Williams (2019): Uncertainty and Scale Interactions in Ocean Ensembles: From Seasonal Forecasts to Multi-Decadal Climate Predictions. *Q J R Meteorol Soc.*, **145(1)**, 160–175.

APPENDIX A

LIST OF FIGURES

LIST OF TABLES