Manuscript os-2022-11
Response to Anonymous Referee #1:

We thank the Reviewer for his/her careful reading of our paper, and for his/her remarks that will help improve the clarity of the manuscript. We did our best to take them into account as explained below. Our replies and comments are given in blue, while the original comments from the Reviewer are in gray.

General comments:

This interesting manuscript focuses on the predictability of small scales in realistic ocean models kept "on track" by data assimilation (although the manuscript does not contain assimilation results). In particular, it proposes a rather novel methodological approach to relate forecast uncertainties to initial uncertainties in the fields, and presents some results quite convincingly in the context of a particular experimental protocol based on a set of 2D "displacements". The topic is scientifically relevant and important and the scientific quality is good, but the focus, clarity and precision could sometimes be greatly improved. I have no reservations about the statistical/probabilistic methodologies implemented, and the results are valid and interesting, but I am not convinced of their generality given the particular experimental protocol (type of uncertainties considered, seemingly "fixed" scale,number of members, etc.): the limits of the ensemble generation approach, and thus thescope and validity domain of the results, should become more apparent. This manuscript should eventually be accepted for publication, but perhaps not quite in its present form.

Specific comments:

The style of the introductory and methodological sections is sometimes rather "literary" and "rhetorical", convoluted to the point of being imprecise (an example: see the comment "lines 56-68" below) -- the approach is often introduced by invoking much more general and theoretical concepts than necessary. On other occasions, the text does not contain enough information or loses the reader. I would recommend (1) adopting a much more "direct", "factual", "scientific" style throughout the text, and (2) improving precision and conciseness. For example, when describing a methodology, the description of what was done in practice could be presented first, accurately and completely (and not in three different places, such as the perturbation scheme in sub-sections 2.2 and 3.1 and Appendix A); then the validity and scope of the approach, including the wider context, can
be discussed, not the other way around.

→ Yes, we agree that, in some places, the text could have been made more concise. We have tried to simplify the text where it was possible without changing the meaning of our arguments. We also believe that positioning the paper in the broader context is important for the reader to understand the method that is presented. In particular, in the case of the description of the perturbation scheme, we have modified the presentation to improve the clarity, but we kept the technical description of the implementation in the appendix. Otherwise, the main text of the paper would be even more lengthy and difficult to read.

However, as the ms. progresses, the style improves, especially in the description of the results, which is often adequate.
The definition of predictability scores (in particular CRPS and predictability diagrams), and the way in which statistical calculations are carried out using all members of the ensemble in turn as a reference (reminiscent of generalised cross-validation) are two aspects of the work that could be generalised to problems beyond the particular experimental protocol. I was particularly interested in the dispersion of the CRPS estimates across the 20 cases (Figures 7,8) --
I would be curious to know what they look like with only the reliability
CRPS component or only the resolution component (the latter possibly giving access to a form of feature-based predictability, i.e. based on whether a particular forecast eddy is present across the members). The decorrelation score is interesting and also seems to be quite general. The location score is of course more related to the particular type of uncertainties in the study.

→ In our application, the verification data that are used in the CRPS come from one additional member of the same ensemble simulation. So, by construction, the ensemble is always perfectly reliable, and the reliability component of the CRPS score should be zero. In practice, numerically, it is non-zero because of the limited size of the ensemble, but it is much smaller than the resolution component, which is what matters in our application to measure predictability.

While section 4 is solid, one should keep in mind that the predictability analyses are conducted within a very specific experimental protocol: that of pseudo-random perturbations based on 2D "displacement" at small scales (10 grid points), having a direct impact on horizontal advection and pressure gradient at these scales (and indirectly on other dynamical processes). (I know that "displacement" is probably not the right term as

you are perturbing the metrics of the model operator, not the grid, but you could give this word that definition in your ms). That's OK, but in retrospect I probably would have liked a more honest introduction and summary framing the study more clearly in the particular experimental protocol (e.g. as described in subsection 2.2 from line 118). Indeed, the results in Section 4 could be very different for other forms of uncertainty. The conclusion is not careful enough in this respect: its first sentence ("The overall aim of this study...") promises too much in relation to the very real and effective work that has been done.

In addition, a limitation of this work that is not mentioned in the conclusion is that the correlation scale of the displacements is set (if I understood correctly) at 10 grid points. So, if I understand correctly, this is a predictability analysis study for a 10 grid point noise. Would a smaller or larger scale noise behave in the same way? What about pseudorandom correlation scales? However, I'm not sure I understood correctly, since the conclusion quotes "10 km wavelength" and not "a scale of 10 grid points" -- which is quite confusing. Similarly, the tenfold use of a Laplacian filter is mentioned -- even more Confusion.

→ Indeed, some confusion might have arised from the text. We have made efforts to improve it at the different places where this aspect was mentioned. The stochastic perturbation is applied on the model grid (~1.4 km), together with a laplacian filter (10 passes) to introduce spatial correlations with neighboring grid points. The ensemble spread progressively develops and cascades upscale as seen for exemple on the spectral metrics in Fig.6.
We tested different numbers of passes for the laplacian filter in the range 3 -10 (not shown), without much difference in the behavior of the stochastic perturbation.

Twenty members is a small size for an ensemble, again a topic not addressed by the conclusion. It is not clear whether we should interpret the discussion in 3.2.3 as evidence that 20 members are "sufficient" for the subsequent predictability study? What about the representation of spatial covariances with 20 members? (These generally converge more slowly than the variances). Also, what is the impact of the ensemble mean, and is it taken into account?

→ Yes, 20 members is usually considered a small size for an ensemble that must be used in data assimilation systems, because they need an accurate and reliable description of the covariance matrix, describing the statistical dependence between model variables (in particular between observed and unobserved variables). In our

application, the objective is to study the evolution of the spread of a given quantity in the ensemble. This quantity may be a model variable, a spectral amplitude or the location of a structure, but this is always just one variable taken from different members. None of the scores described in the paper depend on the ensemble covariance. This is why predictability studies can usually be based on smaller size ensembles as compared to assimilation systems. Nevertheless, it is true that the accuracy of the measure of the spread also depends on the size of the ensemble (but less problematically than correlations), and that this should have been discussed. For instance, with 20 members, the accuracy of the ensemble standard deviation as an approximation to the true standard deviation is about 16%. This is obviously not perfect but sufficient to draw meaningful conclusions. A few words have been added to the conclusion of the paper to discuss this limitation.:

> "Of course, the ensemble size can be a limitation of the accuracy of the conclusions. In our case, with m=20 members, we can expect a 16% accuracy (1/sqrt(2m)) on the ensemble standard deviation as an approximation to the true standard deviation, which is not perfect, but sufficient to draw meaningful conclusions."

This is a scientific paper. Therefore, the emphasis on CMEMS, which is cited several times, and which also comes as the "last word" in the conclusion, seems out of place. Such a study is of interest to all ocean forecasting systems. If appropriate, CMEMS can be mentioned in the acknowledgements.

→ We have now replaced most occurrences of 'CMEMS' by operational systems / operational centers.

Individual comments:

lines 56-58: This appears as a purely rhetorical statement, but perhaps I did not understand what was meant. Models and assimilated observations have errors which do impact the forecasts, we know that. Also, how can model instabilities be used to produce a valuable forecast?

→ Yes, we agree, the sentence was very unclear. It has been replaced by :

> "What matters to the application is then the possibility to produce a valuable forecast with the model that is used (i.e. with its shortcomings and uncertainties), and which may

be quite different from what could be obtained by a perfect deterministic model (as would be done in traditional predictability studies)."

lines 62-63: "initial uncertainties because observation resources are limited": yes, but observations have errors too; and in an assimilating system initial errors are also due to the whole history of all types of errors up to then.

→ Yes, agreed. We clarified with:

"initial uncertainties because observation and assimilation resources are limited, and model uncertainties because model resources are limited."

The introduction has no references on probabilistic skill scores.

→ Our method to quantify predictability could be applied to any kind of score, so we chose to introduce probabilistic scores (i.e. CRPS) in the section where we use them as an example application of our method (i.e. section 4.1)

line 98: "initiated" -> "initialised"

→ Yes, corrected. Thanks.

section 2.1: Which scales can be accurately modeled by MEDWEST60? It is important to have those in mind in relation with the perturbation scales which you will use. Also, in the Mediterranean the internal Rossby radii are quite small.

→ The stochastic scheme used in this work is designed to introduce uncertainty at model-grid scale, with a correlation length scale of 10 grid points, i.e. about 14 km. Uncertainty is thus introduced in the 10-18 km range of the Rossby radius of deformation in the region (e.g. Escudier et al, 2016 , their Fig.5, https://doi.org/10.1002/2015JC011371), which is resolved by ~7 to 13 grid cells in our model. We have now tried to clarify the text on those aspects at the beginning of section 3.2.1.

lines 109-110: "In this context..." -> "In a purely deterministic approach..." to improve clarity.

→ Yes, we clarfied by replacing « in this context » by « *In a purely deterministic approach ».*

But still, you are missing modelling errors here (parameterisation, numerical schemes, missing physics).

→ Yes, we do not claim that we include all possible sources of model uncertainties.

lines 148, 151, in Table 2, etc: "probabilistic model" -> "stochastic model"

→ Fixed.

line 154, legend of Figure 2, etc: "grid size", "size of the model grid" -> "grid spacing" or "mesh spacing". Also what is the distribution law used for the perturbations? (If a noncompact support law is used, did you use an upper bound for the displacement?)

→ Ok. “grid size” has now been replaced in the text as you suggested. Gaussian distributions are used for the perturbations. Since the standard deviation is very small, no bounds were needed.

lines 163-164: "It does rely...": I do not understand the sentence (you wrote the opposite two sentences before). Also: part of this paragraph is descriptive, and part is a discussion in anticipation for another discussion in chapter 4: it is not good to mix everything because you'll get the reader lost.

→ We have now removed the sentence that was unclear.

*Table 2: I do not understand what "identical" initial conditions mean. I would have thought that the spun-up fields would be pseudorandomly displaced using the 20 samples of the displacement fields (for each of 1%, 5% stdev), hence yielding 20 *different* initial conditions across the ensembles.*

→ All 20 member of the ENS-1% and ENS-5% experiments are initialized from “perfect” initial conditions (the same exact ocean state for each member), taken from the spinup simulation (which is a single simulation without any stochastic perturbation). As soon as the ENS-1% or ENS-5% starts, the stochastic perturbation is applied (representing model error) and it makes the members diverge.

lines 182-183: I am a bit confused. The "displacement" is variable, with stdev = 1%-5% of the mesh spacing, but the displacement correlation scale is fixed to exactly 10 gridpoints. Therefore I do not understand the words "on the order of".

→ Yes, the correlation scale is fixed to 10 grid points. This has been rephrased to « with a correlation length scale of 10 grid points, i.e. about 14 km ».

Figure 3: It might be interesting to have a zoomed version on the right (perhaps just for low wavenumbers) to be able to see something.

→ The main point with this figure is to show that the perturbed and unperturbed simulations are nearly undistinguishable from a spectral point of view (meaning that the stochastic perturbation added in the perturbed simulation do not alter the simulation of the physical quantities - here the SSH wavenumber spectrum). We have now modified the text in subsection 3.2.1 to clarify the purpose of this figure.

I did not have time for a full second reading and hence for further individual comments, Sorry.