Manuscript os-2022-11
Response to the Reviewer's comments:

We thank the Reviewer for reading the revised manuscript, for his/her interest for the work that has been done, and for accepting our answers to his questions about the perturbation scheme, which, as we understood, was covering the major concern of his/her first review. Please find below our answers to his/her second review below:.

—

1/ We are sorry to see that some misunderstanding seems to remain about the objective of our paper. The Reviewer misinterprets them as: "(1) the introduction of a new perturbation method for ocean models and (2) the popularisation of ensemble forecast metrics based on cross-validation in operational oceanography."

In fact, the objective of our paper, as stated in the introduction, is *"to evaluate, the predictability of the ocean fine scales in a high-resolution (kilometric scale) NEMO-based model"*. And the methodology proposed in our paper to reach this objective is to compute necessary conditions on initial accuracy and model accuracy to achieve a given forecast accuracy (as explained in the introduction and summarized in the conclusion).

So, of course, to reach this objective, we need to make specific assumptions (as would be necessary in any practical system) about the metrics that is used to evaluate the forecast accuracy (interpreted by the Reviewer as objective (2) of the paper) and about possible sources of irreducible model uncertainty (interpreted by the reviewer as objective (1) of the paper).
The limitation of the study resulting from these choices are already acknowledged in the paper:

a. The fact that the choice of the metrics/score is specific is acknowledged at the beginning of section 4. Any practical application needs to define its own goal. Cross-validation is a powerful tool to produce a predictability score not depending on a particular reference trajectory, but it is presumably not the only possibility.

b. In section 2.2, the model perturbation scheme is presented as a possible generic solution to simulate uncertainties upscaling from the smallest scales of the model. It is used in the paper to illustrate the possible impact of irreducible model uncertainties on predictability. The fact that the conclusions depend on this assumption is already acknowledged in the introduction with: *"However, it is important to keep in mind 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 approximation and cannot be expected to account for the full diversity of uncertainties of different kinds propagating in real operational systems."* .The suggestion by the reviewer to use atmospheric data as a possible source of irreducible model error (thus affecting predictability, and not just operational forecast accuracy) is debatable.

—

2/ The Reviewer also points out the "shortcoming in the experimental setup, vastly underestimating uncertainties from operational forecast systems".

We would like to emphasize that, following the logic presented in the introduction, the purpose of the paper is to evaluate the predictability of the system and usually this question is addressed by computing a lower bound for the forecast uncertainty resulting from irreducible uncertainty in the system (i.e. a vanishingly small initial error and/or irreducible model uncertainties). In this paper, we suggest that an alternative approach to investigate predictability in realistic systems is to look for an *upper bound* for initial and/or irreducible model uncertainties, *to obtain* a targeted forecast accuracy. In this respect, the results

obtained in the paper are realistic. As acknowledged in the text, they are only specific to the particular choice of the metrics, and the particular choice of irreducible model uncertainties (which only decreases the upper bound obtained with initial uncertainty only). Our goal is clearly not to simulate the same kind of initial error as in operational forecasting systems, which depends on the observation and assimilation system. This is clearly out of the scope of this study, as we have explained in the last two paragraphs of the conclusion.
We have now tried to clarify the objectives by adding the following text in the introduction :
"In other words, the objective of this paper is thus to compute an upper bound (or more generally, necessary conditions) for the initial uncertainties, in order to obtain a targeted forecast accuracy. We do so by using different types of metrics to quantify the forecast accuracy, in order to emphasize that the definition of this metrics is still a subjective choice, which depends on the goal of every particular application. The influence of one possible source of irreducible model uncertainty on this upper bound will also be illustrated."

And : *"It should be emphasized that the goal of the present study remains to quantify the intrinsic predictability of the system (as defined by Lorenz, 1995) and should not be confused with that of quantifying the prediction skill of any given current operational forecasting system, that would then incorporate all sources of error, such as extrinsic errors that would result from coupling with the atmosphere, sea ice etc (e.g. Robinson et al., 2002). However, deriving predictability as an upper bound or 'necessary conditions', as it is proposed in the present case study, can provide useful guidance for the design of the future generations of operational systems that (...)"*

—

Detailed comments :

- l. 225: The comparison of ensemble STD at one given time (your figure 5) to time-averaged std from the CMEMS system is not relevant for the discussion at hand.
First, 2.5 cm is the ensemble spread after two months of simulations while Clementi et al. run 10 days forecasts only. The accuracy indicated by Clementi et al is 4.2 cm as analysis error (day one) for daily data in the region of interest (Table 4, Region 2 in Clementi et al. 2021). Since daily averaging reduces the spread, the ensemble setup seems really far below target, say, by a factor of 20 to 100 rather than "comparable". Adjust the text here and dependencies elsewhere to indicate that the experiments are underestimating the operational uncertainties but the methods remain applicable.

We agree that some confusion might arise from this line of text and from this comparison and we have now tried to clarify..
In short, Figure 5 is used to illustrate the spread growth in our 3 experiments, and one of our comment about the figure is to say that the spread (ensemble STD) has reached saturation after 2 months, with an amplitude at saturation of about 2.5 cm in average over the domain, with local maxima of spread values are found around 10 cm.
This amplitude seems to be consistent with time standard deviation of hourly SSH timeseries from CMEMS analyses available from this dataset :
https://doi.org/10.25423/CMCC/MEDSEA_ANALYSISFORECAST_PHY_006_013_EAS6), which is what we can expect if the ensemble members are fully decorrelated after 2 months (saturation of the spread).
In any case, we feel that it is not the main point of our study, and we agree that our initial comment in the text and the reference to (Clementi et al. , 2021) for the above dataset might have brought some confusion. So we propose to simplify by removing the confusing sentence "*Those values are close to typical deviation values of 235 hourly SSH over time in the Mediterranean region found in the CMEMS Mediterranean Forecasting System (Clementi et al. , 2021) at same period of year (not shown).*"

What is perhaps more important to clarify is that with FIG.5, we were not commenting on the amplitude of the initial error (as the Reviewer might have misunderstood?). It should be noted that given the objective of the paper (i.e study predictability), it would be inappropriate to generate levels of initial errors that are comparable to what exists in currently-used operationnal systems. In fact, for our predictability analysis, we consider each timestep in the ensemble experiment as a virtual start date (thus bearing some initial spread of various amplitude), and we analyze the forecast accuracy at a given time-lag, *relative* to each of these virtual start dates. So in other words, it is not just the small initial condition of the ensemble

simulations shown in Fig. 5 that is used as possible initial error in this study, but the all range of ensemble spread available over the two months of these experiments (see our explanation in section 3.1: *"Note that the choice is made to (…)"*).
By looking at the CRPS score for SSH in Fig. 9 (a,d,g) for instance, it appears that there are situations where the initial error is as large as 4cm. The fact that current operational systems are unable to produce a sufficiently low level of initial accuracy would just mean that they would be unable to reach the level of forecast accuracy that we tested. But again, this depends on the observation and assimilation systems, and it is out of the scope of this predictability study.

- Detailed comment on Figure 5 (previously Figure 4) even though there is no time for new simulations, the authors should at least confirm the tidal (or inertial or other) nature of the oscillations in the ensemble spread and speculate their origin. This can be done by eyeballing the timing of the maxima in the curve, no need for additional experiments.

We have now added this comment in the text: *"From the figure, it appears that the presence of model uncertainty is associated with some oscillatory behavior in the ensemble spread evolution, at a period close to half a day, and with amplitude growing with the amplitude of the model error (barely visible in ENS-1\% but appearing more clearly in ENS-5\%). These oscillations might reflect some slight spurious numerical effects due to the horizontal grid distortions imposed in these experiments with the parametrization. The period close to tidal or inertial period, might suggest some effect related to partial wave reflexion in the buffer zone at the lateral boundaries of the domain, but further investigation would be needed to be able to conclude."*

Figure 8 (previously 7): The text only pleads very generally for using ensemble techniques over deterministic simulations in order to use the metrics, which has limited interest. Looking at the curves, my impression is that there is no use in integrating the ensemble past 20 days because the CPRS stops increasing. A reflexion on this point would be welcome.

We think that it is important that the experiment is long enough so that the ensemble spread has time to saturate. This depends on the metrics that is used. On the CRPS scores for SSH, we agree that the spread saturates quite quickly, but on the larger scales, the saturation needs more time to be reached. This is what is shown on Figure 18, where scales larger than L=40 km are not yet fully decorrelated between the ensemble members after 20 days (mean coherence ratio above 0.5 for scales L>40 km).

On Figure 9 (previously 8): The authors have missed my point: the 95% probability isoline is very erratic when the CPRS scores are clustered in bulgy-looking scatterplots. This illustrate my general comment that the exercise is academic and that the readers should be cautious when the results are tightly linked to a given trajectory. These scatterplots would likely look more homogeneous if several sources of errors were considered, and the text should note the limitation by the experimental setup.

Yes, the accuracy of the results is limited by the size of the ensemble simulation that has been produced, especially where the spread of the scatterplot is large. We have now acknowledged this in the text of the paper (end of subsection 4.1.3): « *(with imperfect accuracy where the spread is large, as a result of the limited size of the ensemble)* ».

- L. 506: The revisit time is given for a perfect model and an almost perfect initial state and is on the optimistic side. Indicate that the necessary revisit time should be shorter to suit today's forecasting systems.

Yes. We have now added "*(and even shorter with current imperfect models).*" at the end of the sentence.

- L. 521-525: The authors should summarise their recommendations for further - more realistic - experiments.

As mentioned above, the objective of this study is not to reproduce realistic initial and modelling error as they are in realistic operational systems. In the context of a predictability study, our experiments are actually quite realistic. The only two specific choices that can be adjusted are (i) the metrics that is used to quantify the forecast accuracy, and (ii) the type of irreducible model uncertainties that is accounted for (as a possible further limitation to standard predictability). Also see our previous responses above.