Manuscript os-2022-11
Response to Anonymous Referee #2:

We thank the reviewer for his/her careful reading of our paper, and for his/her appreciation of the work done in this paper. We did our best to take the reviewer's remarks into account as explained below. Our replies and comments are given in blue, while the original comments from the Reviewer are in gray.

Review of Leroux et al. "Ensemble quantification of short-term predictability of the ocean dynamics at kilometric-scale resolution: A Western Mediterranean test-case."
The manuscript presents an analysis of an ensemble of ocean model simulations at very high resolution using a novel idea for intrinsic model errors based on concepts of location errors. The article uses very solid and interesting concepts and methodology and makes both a refreshing and useful contribution to the operational ocean forecasting community (where I belong).
The exploitation part of the research is very well developed and thoroughly explained, which will certainly help popularise probabilistic diagnostics into the oceanographic community, but is so extensive as to almost entirely eclipse the core stochastic model developments, which constitute the novel aspect of the paper. It is indeed seldom that one sees a theoretical advance (that from the papers from Mémin and Chapron) brought into a realistic ocean model, so it is of general interest to see for the first time the effects of the stochastic perturbations on the model solution.

However there is no discussion of the numerical effects of these perturbations and no visual from the perturbed model (illustrations are disappointingly always extracted from the CI control simulation without stochastic noise), leaving an uncomfortable impression that something is hidden from the Readers.
→ The stochastic scheme used in this work is designed to introduce uncertainty at the model grid scale (plus smoothing by a 10-passes laplacian filter to introduce spatial correlations with a few neighboring grid points). This uncertainty is then expected to develop and cascade spontaneously toward larger scales, through the model dynamics. Thus the stochastic perturbation introduced for that purpose should alter as less as possible the behavior of the physical quantities simulated by the model. This is why, by design, there is almost no visual difference between field

snapshots from the unperturbed model and from the perturbed models. For the sake of brevity we had mainly provided illustrations from the unperturbed experiment, which we consider as our main experiment in this study.
But we acknowledge the interrogations that it might have induced and we have now added a new figure (figure 3 in the revised manuscript, showing snapshots of the SST and relative vorticity simulated by the unperturbed model and the 2 perturbed models). We have also modified the text to clarify these issues at the beginning of section 3.2 and 3.2.1:

> "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. This uncertainty is then expected to develop and cascade spontaneously toward larger scales through the model dynamics. The design should be such that the introduced perturbation alters as less as possible the behaviour of the physical quantities simulated by the model. Figure \ref{fig.mapSST} illustrates that indeed the simulated fields in the perturbed model remain nearly unaltered and undistinguishable from the same fields in the unperturbed model. Only in the zoomed snapshot of relative vorticity (i.e. taking the Laplacian of Sea Surface Height, thus emphasising gradients) from experiment ENS-5\% (Fig. \ref{fig.mapSST}f), some visual alterations starts to appear on the smallest scales (note that this is why we did not propose any additional experiment with a stronger perturbation than 5\% in our study)."

We also propose for the reviewer some additional movies of the evolution in time of the SST field and relative vorticity field from experiments ENS-CI and ENS-GSL15 (unperturbed, and perturbed with a 5% std of the stochastic scheme) over the 2 months of simulation. These movies are in open access on vimeo: https://vimeo.com/showcase/9695743.

Another aspect that is not discussed is the somewhat binary response of the model to the amplitude of the stochastic noise. The 1% case corresponds to 15m/d displacements (according to my own back-of-the-envelope calculation) and is most often indistinguishable from the CI (0%) case. On the contrary, the 5% perturbations corresponds to 75 m/d, which also seems tiny, turns out completely different from the other two cases and generates kilometers of feature location uncertainties within one single day. What happens between 1 and 5% that causes such a binary response? I believe that tidal amplification is the culprit and suggest an additional experiment in the detailed comments below, where the stochastic noise is turned off in the model nesting zone. The doubts on the stochastic perturbation method do not impair the main findings of the paper, because the latter probably stand with the

CI control ensemble alone, and the diagnostic methods can be applied to any stochastic model, but there is a risk that the manuscript is used to advocate for a stochastic model perturbation method that it does not truly validate.

→ We understand that the presentation of the results may give the impression that the response of the model to the perturbation is binary, with almost no effect with a 1% perturbation and a large effect with a 5% perturbation. But the response is actually gradual, and in both cases, the dynamical behavior of the model is about the same as in the unperturbed model. By looking at one single member (whatever model variable), it is hardly possible to say if it is a perturbed or an unperturbed member (except by looking specifically at very fine details or at the spectrum in the fine scales in the case of a 5% perturbation). It is only by looking at the spread of the ensemble (misfit between members) that the effect of the perturbations can be very clearly detected, which is precisely what we do in the paper. And this effect on the spread is again visually magnified by looking at location misfits, which is a very sensitive diagnostic. So the reason why the effect on location misfits is much larger with a 5% perturbation is just that the perturbation itself has a 5 times larger standard deviation.

Another general remark about the use of the probabilistic diagnostics is that some of them can be generalised to deterministic forecasts under ergodicity assumption: spatially averaged statistics (CPRS, PSD) can be interpreted as expectations and could be applied to forecast systems that have invested in high model resolution rather than in ensembles.

→ It is true that some ensemble statistics can be reached also from a deterministic simulation under ergodicity assumption. But it is not clear to us how you could generalize our approach and our diagnostics to a deterministic *forecast:* the time-lag after which, starting from a given initial uncertainty, the final forecast score is quantified, has to be the same in all the realizations.
In addition, this would also require assuming the stationarity of the predictability statistics, which is far from obvious in complex non-autonomous systems. The system can be more or less predictable depending on its current state or on the atmospheric forcing conditions. This can only be assessed using an ensemble approach.

Overall the paper is very good and makes a very enjoyable read. I am impressed by the

enormous amount of thoughts and work that went into it. The structure, the style and the illustrations are all excellent, and will certainly make a splash in the operational community. So I recommend its publication after revisions that I would call "major" because of a possible problem in the implementation of the stochastic method.
The paper is maybe a little on the long side but I will suggest some reduction of the illustrations and point out a few repetitions in the text. Ideally the manuscript should be split into two separate papers, one demonstrating a new stochastic perturbation method and the other on the ensemble forecast diagnostics, but I will not insist on this if the authors can shed more lights on the stochastic perturbation method without adding pages of text.

→ Thank you again for the appreciation of the work done in this paper. We hope that we have provided enough new material to convince the reader that we only apply very small perturbations to the model operator, which produce only little effect on the model behavior (even if these small perturbations induce a substantial effect on the ensemble spread). That is why we did not expand much the description of the stochastic effect in the model (which is barely visible by itself), but only in the description of the effect produced on predictability (which is non-negligible as compared to initial uncertainties).

Detailed comments:
Title, abstract and introduction: no remark. All are representing well the actual contents of the paper.

Section 2
- Figure 1: Why do you need to define as many as 3 subregions?
→ We agree that it might not be optimal for the sake of clear presentation, but we ended up with 3 defined subregions in this work for technical reasons. (a) is the largest squared region to apply the spectral analysis, (b) was meant to be a zoom to illustrate fine-scale features on the snapshots, and (c) is a small (100x100 points) region without land to apply our example score on the location of the features. We have now added *"and used for various diagnostics or visualizations"* in the caption of Figure 1 to be more explicit.

- Line 90: I understand that the eNAT60 configuration is not only a boundary condition but a baseline to which the different experiments should revert if there were no stochastic

perturbations at all. Please make it explicit and come back to it whenever the different experiments are compared to eNAT60.

→ We have now added an explicit mention to the eNATL60 experiment in paragraph 3.2.1 where a comparison of the wave-number spectra are made.

- indicate which method is used to impose lateral boundary conditions (the Flather conditions?).

→ 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 sea-surface height and barotropic velocities (a radiation condition is applied
on the normal depth-mean transport across the open boundary).
We have now added these technical details as a note in Table 1.

- Line 95-98: a) and c) are not strictly a "difference" and b) should not lead to any difference as long as the model is numerically stable. Please rephrase.

→ We have now rephrased this sentence to avoid using "difference" although we do think it is important to mention those technical aspects for the sake of reproducibility:

> "Compared to eNATL60 which was forced at the lateral boundaries by the daily GLORYS reanalyse \ref{LELL21} and an additional tidal harmonic forcing from the FES2014 dataset \ref{LYAR20}, in MEDWEST60 we add no additional tidal forcing since it is already explicitly part of the hourly boundary forcing taken from the eNATL60 outputs. The model time-step in MEDWEST60 is also increased by a factor 2 compared to eNATL60 (80 seconds in MEDWEST60 versus 40 seconds in eNATL60."

- Line 114-119: This argument is contorted. Any intrinsic or extrinsic errors (in the vertical mixing or winds for example) may as well affect the smallest scales of the ocean, if they are set up to do so. It would clarify the argument if you state upfront that you consider location errors exclusively and that other types of errors can be added at will.

→ Yes, indeed, other sources of errors can directly affect the small scales. We have modified the text of the paper to correct this point:

"These uncertainties are likely to depend on many possible sources, by embedding for instance misrepresentations of the unresolved scales and approximations in the model numerics, but also many others. "

- Line 134: Indicate the physical scales of 1% and 5% with respect to the temporal autocorrelation: displacements of 15 m/d and 75 m/d respectively.

→ Yes, this information is indeed very helpful. In view of the typical grid size (1.4 km in average) and the correlation timescale of the perturbations (1 day), the typical velocity of the grid points is indeed about 14 meters per day (for the 1% perturbation) and 70 meters per day (for the 5% perturbation) in the two horizontal directions. This has been added in the text of the paper.

- Line 139: "quite consistent" does not sound too good. Can you recall which conclusion of Mémin (2014) is comforted by the present study?

→ We agree that the reference to Mémin in this sentence leads to confusion. It has been removed. What we do in this paper is not equivalent to what is done in the work of Mémin (2014). We just say that we use a « similar approach ». So, none of the conclusions obtained by Mémin (2014) can be comforted by this study. As explained in the paper, in the work of Mémin (2014) the noise is introduced in the continuous equations (as a random Lagrangian displacement of the fluid parcels) to obtain modified Eulerian equations (with additonal terms accounting for the noise), while in our study, the noise is directly introduced in the discrete model by a perturbation of the grid. The underlying idea is the same but we do not claim that it is equivalent. In addition, in the work of Mémin (2014), the noise is assumed uncorrelated in time (Brownian motion) as a basic assumption, while we assumed a 1-day decorrelation time scale.

Section 3

- L. 158: what does CI stand for in ENS-CI? Control Integration?

→ It stands for 'Conditions Initiales' (i.e. the source of uncertainty in the experiment) as opposed to ENS-1% and ENS-5% where the source of uncertainty comes from the stochastic perturbation. We have now made it more explicit in subsection 3.1 of the manuscript.

- Figure 2b indicates that even after Laplacian smoothing, the square model grid is

distorted and deviates from orthogonality, which may lead to numerical noise and eventually instabilities. The ROMS user community is advised to keep the grid cells orthogonality above 95% in practice, and especially at the lateral boundaries of the model, to avoid errors propagating inside the model grid. My recommendations would therefore be to dampen the model grid perturbations in the nesting zone of the model (in the first 5 or 10 grid cells) to avoid inconsistencies between the outer an inner model solutions, in particular the barotropic mode. I will come back to this at Figure 4.

→ Yes, it is true that too much distortion of the model grid cell can deteriorate the accuracy of the numerical schemes. On the other hand, our scheme is also intended to describe uncertainties in the numerics and thus to produce some spread at the numerical level. One perspective of development to alleviate possible difficulties might be to re-interpolate the model solution on the reference grid every while, or even at every timestep.

- Table 2: Define e1 and e2 in relation to the appendix.

→ Ok we have now replaced e1,e2 by Delta x Delta y in the Table.

- Figure 3 shows indistinguishable lines, and no indication of what is good or bad. You could either plot the difference of PSD from the eNATL60 reference or solely indicate the maximum difference in the text and skip the figure altogether. If you keep the figure, I recommend to remove the part for wavelength > 250km because of the small domain.

→ It seems important to us to keep this Figure, as it shows from a spectral point of view that the perturbed and unperturbed simulations are undistinguishable (meaning that the stochastic perturbation added in the perturbed simulation do not alter the simulation of the physical quantities (here the SSH wavenumber spectrum). In fact it comes back to your previous comment saying that there was not enough comparison of the perturbed and unperturbed simulations (see our answer to this comment). We have now modified the text in subsection 3.2.1 to clarify the purpose of Fig.3 (spectra) and new Fig.3 (snapshot).

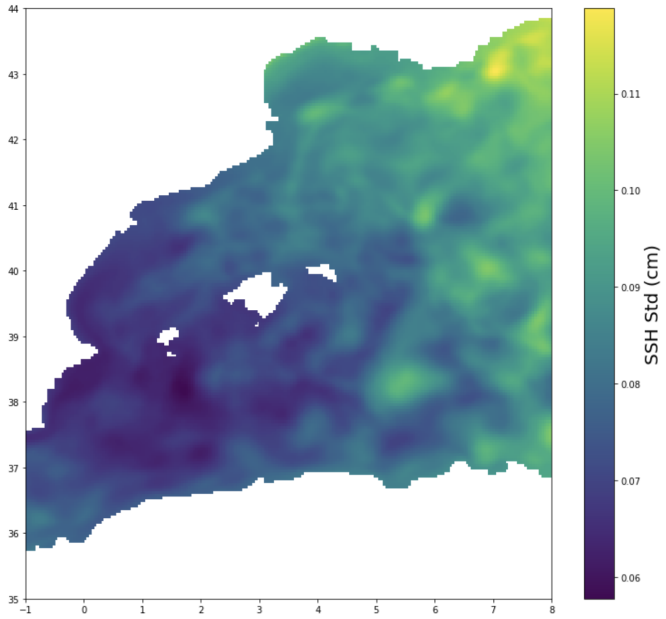
- Figure 4 exhibits an oscillatory signal in the ensemble spread, whereas intuitively I expect the spread to grow monotonously. The oscillations are most visible in the 5% case but also in the 1% case. I also noted that the oscillations peak at the same time in the 1% and the 5% cases, about 4 times a day. Unless you have used the same random seed in

the 1% and the 5% case - which would be odd - the coherent oscillations indicate an amplified resonance of tidal signals, which brings me back to my previous remark about barotropic lateral boundary conditions: the nesting routines (radiation condition or Flather conditions, whichever you use) should allow tidal and other barotropic signals to be evacuated out of the domain, but if the perturbations make this boundary condition imprecise, the tides may be reflected at the lateral model boundary and resonate inside the nested model domain. I have a suspicion that this could be avoided if the perturbations were attenuated near the model boundaries (and maybe in shallow waters as well).

→ Yes, it is difficult to exclude the possibility that spurious numerical effects due to grid distortions have some impact on the solution, but it is difficult to speculate on this without running specific test cases (that would require significant additional computing resources).

- Line 209: This claim could be confirmed by a look at the accuracy numbers from the MED MFC QuID document on the Copernicus Marine website.

→ We have not found any reference to which we could compare based on *hourly* SSH in the region. But instead we have directly computed the time Std from the hourly SSH outputs from the up-to-date CMEMS Mediterranean Forecasting System at 1/24º and including tides (https://doi.org/10.25423/CMCC/MEDSEA_ANALYSISFORECAST_PHY_006_013_EAS6) and we found values consistent with our study. An example of plot is provided below (time Std of the hourly SSH over feb-may 2022 from the above dataset). Maximum signals are locally ~ 10cm which is consistent with our experiments.



→ We have now modified the text as followed in section 3.2.2:

"Those values are close to typical deviation values of hourly SSH over time in the Mediterranean region found in the CMEMS Mediterranean Forecasting System \citep{CLEM21} at same period of year (not shown)."

- Figure 5 makes a stunning impression, but is uninformative. I would have preferred to see the 5% case to have a visual impression of the effect of random perturbations (there are otherwise none in the whole paper).

→ As discussed above already, the stochastic perturbation was designed so that there is no visual effect of the perturbation on the physical fields. The point of this figure was rather to show the divergence between 2 members of the same ensemble (here we chose ENS-CI for the sake of brevity). See attached two supplementary figures (FIG06new_ENS-1.pdf and FIG06new_ENS-5.pdf) illustrating how 2 members diverge in ensembles ENS-1% and ENS-5%.

Section 4

- L. 268-280 is a nice introduction of the ensemble diagnostics, but seem like a methodological overkill: the diagnostics are initially intended for location-dependent comparison to observations, but in the absence of observations like in the present study, some more basic diagnostics may be simpler to use than a cross-validation with each ensemble member. This is the case for the CRPS which is aggregated spatially for all

members to a single number and does not seem to add more information than a standard deviation. Please replace by the ensemble spread if this is a simpler diagnostics that provides the same insights.

→ Our argument in the paper is that cross-validation is useful if the objective is to measure predictability by comparison of different indicators (which can be more or less complex) to a reference truth, and not only by the standard deviation of the ensemble spread. To obtain general conclusions, it is then necessary to use each ensemble member as the reference truth, hence the cross-validation algorithm.

As a first simple indicator, we could indeed have used the rms misfit with respect to the reference truth rather than the CRPS score. And, in this case, the result would indeed probably not have been very different from simply looking at the ensemble spread. But for more complex indicators, the cross-validation algorithm is usually needed.

In practice, computing the CRPS score is not more complicated or more expensive to compute than the rms misfit. It was used on purpose to illustrate the fact that, with the cross-validation algorithm, predictability can be evaluated using any type of score of practical interest to the user. The only thing that is needed is an operator to measure some kind of misfit between a forecast and a reference truth.

- L298-299 are repeated in the figure caption.

→ It is on purpose that the text is repeated in the caption, as we wish the captions to be as informative as possible, even for a Reader that would only browse quickly the text and focus mainly on the figures.

- Figure 8. It would seem fair to mention that beyond 5 days of lead time, the 95% percentile is dependent on the model trajectory and does not make a robust statistic, a larger ensemble or a different perturbation method may improve that.

→ With the cross-validation algorithm, the result does not depend on the model trajectory since every member is used successively as the reference truth. The accuracy is thus only limited by the size of the ensemble. When the spread becomes large, the error is also larger in amplitude, which explains the irregular behaviour that can be seen in the figures.

- The small lines in Figure 10 are not very informative. The three figures could be compressed into one by showing the three 95% quantile only and plotting the differences from the initial CRPS.

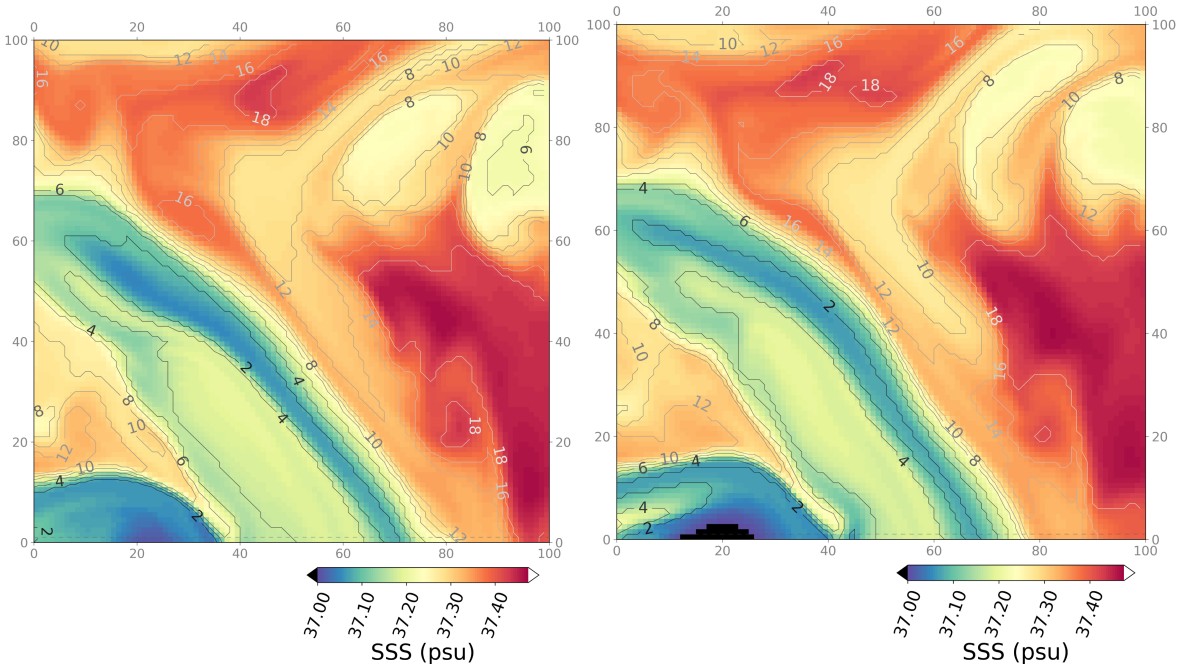
→ The green line (i.e. showing the initial score required to have a 95% probability that the final score is below a given value) only gives an illustration of how our probabilistic definition for predictability can be read and used for quantitative results. We think that it is worth comparing the full probability distribution from the 3 experiments in the Figure, and not just the example of application (the green line).

- Section 4.2.1: I guess there are technical difficulties with the location score in the presence of islands or complex coastlines. This could be mentioned.

→ Yes, we fully agree that the location score used in the paper is just a first simple approach to further illustrate the point that any score can be used to evaluate predictability. (For instance, here, it would not be possible to measure the ensemble spread, cross-validation is really needed.) As it is, it has many shortcomings and should clearly be generalized if it must be used in practical applications. This is now acknowledged in the manuscript.

- Figure 11 (top against bottom) is nearly showing the same thing. You could remove the two lowermost panels by adding the 20 isolines in the top panels.

→ We tried to follow the Reviewer's suggestion (see Figure below) by adding the quantiles as contours on top of the SSS field in shading. But we think the resulting figure is less easy to understand than the initial figure, so in the end we prefer to keep the initial one in the revised manuscript.



- L. 433: Why choose SSH this time?

→ We choose SSH for this last example score because SSH is an observed quantity, and SSH spectral analysis in space domain is often applied in studies focusing on submesoscale-permitting realistic ocean models (e.g. Ushida 2022, Adjayi 2021). We think the kind of probabilistic approach and score we illustrate here might be of interest for a larger audience than just the operational modeling community. This is why we also discuss the potential relevance of this kind of predictability diagrams in the context of the future SWOT altimetry mission, at the end of section 4.3.3.

- L. 460: scales above 150 km should be removed from the figure.

→ We have now added some grey shading in all the spectral figures for scales that are not fully resolved within the considered region (lambda>L/2 where L is the size of the region and lambda the spatial scale) and we also added some comments in the captions and text.

- L. 461: I would suspect that checkerboarding (numerical noise) would easily cause the correlation of small scales. Numerical noise is ubiquitous in all ocean models although viscosity makes it almost invisible. If the authors use a high-contrast colour scale (like "details" in Ncview), they would probably see some checkerboarding in the model output, which would inevitably appear coherent at the smallest wavelengths of the model output.

→ Yes. We had mentioned the possibility for numerical truncation errors in the text. We have now generalized to “numerical noise”.

- Figure 18: Add the diagonal line for T=0.

→ We have now added a diagonal line for R_0 = R_forecast in the Figure.

- L. 485: The authors could indicate which SWOT revisit time would be necessary to maintain the small-scale structures (if the data assimilation were ideally good).

→ We have now added a few lines in section 4.3.3:

> “With a perfect model and a very good assimilation system that would ensure an initial ratio R_0 close to 1 (say 0.9 for the sake of the numerical application here) the spectral coherence ratio R of the forecast after 5 days drops down to 0.5 for scales in the range 10-30 km, while it remains above 0.8 for scales in the range 60-100 km at same time-lag. Or to put it differently, if the target for the spectral decorrelation was to remain above R=0.5 for all scales in the range 10-100 km, then a revisit time of the satellite between 5 and 10 days would be necessary.”

Appendix A1:

- L. 553: “Anamorphic transformation” is a pleonasm.

→ Yes, but it is commonly stated like this. We modified the text to avoid the pleonasm: *“is a transformation of the coordinates (anamorphosis)”*

- L. 582: the link between the theoretical papers from Mémin and Chapron and this one is not obvious. How does the sigma value translate into the stochastic process P?
Appendix A2

→ Yes, we agree that the connection is not direct. The point is that there is no equivalence and thus no direct correspondence to find with the work of Mémin/Chapron, only close similarities. In the theoretical papers of Mémin/Chapron (leading to a continuous Eulerian model formulation), the noise is assumed uncorrelated in time, but they have a general formulation for the spatial correlation structure. On the other hand, in our simple pragmatic implementation (directly introduced as a Lagrangian displacement of the grid in the discrete model), the noise is assumed correlated in space and time, but with a very simple assumption for the space/time correlation structure.
The text of the appendix has now been modified as:

" σ(x,t) dB is a stochastic process uncorrelated in time, but correlated in space, with a general formulation of the spatial correlation structure. "

- L. 595 to 599: "can be thought", "can be be viewed" and "can be argued" make a very embarrassed logical chain to line 600, which I would promote upfront to motivate the Approach.

→ Here, the point is that we do not want to reduce the interpretation of the scheme to numerical uncertainties (line 600), but also to physical uncertainties (unresolved scales). We have tried to simplify the text to improve the clarity of the argument. The text has now been modified as followed in the appendix

" A stochastic metrics, describing relative location uncertainties in the model operator M, corresponds to the main effects that we want to simulate, because it can represent both physical and numerical uncertainties. On the one hand, the stochastic metrics is an explicitly Lagrangian transcription of Eq. (A6) in the model dynamics, which describes physical uncertainties that upscale from unresolved processes. "

- L. 610: Mention that a^2 + b^2 =1 to maintain the variance constant.

→ Depending on the situation, the variance is not always expected to be constant.

- L. 611: The "assumed independence" of the perturbation is later contradicted by the Laplacian filter in Line 620.

→ No, the application of the Laplacian filter does not modify the independence between the x and y components of the noise.

- L. 618: the citation to Garnier et al. (2016) is repeated.

→ Yes, sorry, the repetition has been removed.

- L. 620: does the Laplacian filter maintain the standard deviation?

→ The Laplacian filter does not maintain the standard deviation but a correction factor is applied afterwards to restore the specified standard deviation. This is now explicitly stated in the text of the appendix.

- L. 620: is the value of sigma linked to the sigma in Mémin/Chapron?

→ No, there is no link with the notation used in Mémin/Chapron. Here, it is just the standard deviation of the noise. In Mémin/Chapron, it is something like a square root of the spatial covariance of the noise.

- L. 629: Transformed to the other grids: do you mean a linear interpolation?
→ Yes, this is done by linear interpolation. This is now explicitly stated. The T-points are moved by the noise, and the U-points, V-points etc, are moved accordingly.

- L. 632: Only here is it possible for the reader to calculate the typical scale of the perturbations (about 15 m/day for 1%). This information is important to realise how much the model amplifies the location noise into location errors (roughly by a factor of 100 to 1000 in a single day, which is mind-boggling) and should be discussed in the main text.
→ Yes we agree that the scale of the perturbation is very important, but it was already provided in section 2.2 (where we gave the standard deviations 1% and 5%). The text has now also been improved by giving explicitly the typical grid velocity (as suggested above by the reviewer).

Typos:

———

- l. 133: remove the second "that". —>FIXED.
- L. 239: "characterizing" —>FIXED.
- L. 343 Fussy -> Fuzzy —>FIXED.
- Section 4.3.1: "pf" -> "of" —>FIXED.
- L. 531: Beying -> Beyond—>FIXED.