Response to Dr. Rachel Furner:

We thank Dr. Rachel Furner for your thoughtful and constructive review. We greatly appreciate your recognition of the relevance of our work and your positive feedback on our methods and presentation.

Your suggestions are very valuable, and we have carefully addressed them in our manuscript revision and our responses below. We are grateful for your time and effort, which have helped us improve this manuscript.

Specific Comments:

- Currently the abstract lacks information on performance and impact. For example, the computational gains of OceanNet are significant but not mentioned here, neither are the results.

Following reviewer's suggestion, we have added the following sentence to the abstract:

"OceanNet can generate a 120-day forecast of the Gulf Stream Meander within seconds, offering significant computational efficiency."

- Line 8, As far as I can tell the model doesn't track, or identify, the GSM, but predicts it. This is key and should be the focus, i.e. 'trained to predict…' rather than 'trained to identify and track…'.

Following the reviewer's suggestion, we have revised the sentence as follows:

"…OceanNet (a neural operator-based digital twin for regional oceans) was trained to predict the GS's frontal position over subseasonal-to-seasonal timescales."

- Figure 1: It's not clear why all the subpanels cover different dates and/or averaging periods. It would be more cohesive if they were all the same (possibly separating into 2 figures, one with long term mean and one with an example daily mean), unless there's a sensible reason for the different views, in which case briefly explaining this would be useful.

Different dates were selected to highlight the dominant ocean feature in each sub-region. For the Gulf of Mexico, the focus of OceanNet is to predict the Loop Current and the Loop Current Eddies, which are not visible in the mean fields of this region. The date for the Gulf Stream Meander image was chosen to display a continuous feature with eddies on both sides of the northern boundary.

- Line 68-70 claims there are no global data-driven ocean models. This instead should cite Wang 24 and the XiHe model, https://arxiv.org/abs/2402.02995 which is a global data-driven ocean model. More generally there are a growing number of related ocean approaches, it would be beneficial to include some of these, i.e. Xiong23, Bire23, Subel24.

Following the reviewer's comment, we have revised this sentence to include the reference of Xihe.

"These studies are certainly a step in the right direction, but there is ample need for more data-driven ocean models (e.g., Wang et al., 2023) similar to the data-driven global weather models seen in Pathak et al. (2022), Bi et al. (2023b), and Lam et al. (2023)"

- Line 99: What is meant by 'as would be the case for observations'? Please can the authors expand a bit on this. Do they mean as would be the case in a realistic forecast scenario, rather than when training on past data?

For clarification, we have revised this sentence as follows in the revision:

"Unlike the 4D-var method, the EnKFDA method does not rely on future timestep observations or require forward and adjoint model iterations during data assimilation. This approach enables the efficient creation of a data-assimilative ocean reanalysis, allowing OceanNet to be trained on a time-space continuous reanalysis dataset."

- Line 109-115: Please add a sentence making it clear what inputs and outputs are to the model --- it seems this is just a 2d SSH field for input (no other variables, and no temporal info, or explicit spatial info etc), and SSH field output. Clarifying this would be helpful.

Following the reviewer's suggestion, we have revised this sentence as follows in the revision:

"DLOP is a relatively simple U-Net and overall simpler than DLWP, but the core idea is the same: pixel-wise connections of two-dimensional fields of physical variables (i.e., SSH fields in this study) between timesteps are sufficient to predict the evolution of such fields (SSH in this study) through time."

- Figure 2: It seems from the caption that 'Run 57' refers to the detail of the DLOP run, and has no meaningful relevance to the reanalysis. Ideally this would be replaced with the date which is being assessed, but otherwise omitted or used in the DLOP title, but not the reanalysis title.

Following the reviewer's suggestion, we have changed the title of the subplots to reflect the date they are valid

- Line 155-156: I think this misses the very important step of the convolution being applied? i.e. a "Fourier transform is performed… *convolutions are applied in Fourier space*… an inverse Fourier transform…."

In our design of FNO, convolutions are not applied in the Fourier space. Please see Figure 3.

- Equation 2: I want to clarify if N and H as referred to in figure 3 and elsewhere (i.e. line 171) are *exactly* the parts of the equation as described in equation 2, or if they (at least N, possibly H as well) are the learnt approximations to this. It seems that we learn N (and to some extent maybe H), and so some differentiation between their true meaning/value, and the learnt approximation of them is needed if so. (Otherwise, if the learnt value as described in figure 3, and the use as described in equation 2 are *exactly* the same thing, where in the set up is the approximation, and where do errors come in when calculating X(t+delta_t)?) Similarly, equation 3 should clarify that this is the approximation to X(t+delta_t), for example with use of a ^.

We acknowledge that some of our choices of notation for equations and figures were inconsistent. Per the reviewer's suggestion, we have adopted notation that is more common in machine learning disciplines. We have reviewed all equations and figures for consistency.

- Figure 3 is quite busy. I think it would benefit from the loss function being included as an equation rather than within the figure.

We acknowledge that Figure 3 is quite detailed, but we think it effectively consolidates all the relevant information in one place for the convenience of readers. After careful consideration, we have decided to retain the current format.

- Line 176-180: Again please clarify what the model inputs are, and what the model outputs are, and if PV and GEO are calculated from output variables, or explicitly output by the model (I assume the former, but it would be good to be explicit about this).

We have removed this section in the revision. Please see our response to the next comment of the reviewer.

- Line 176-187 and figure 4: this comparison could do with a baseline or comparison to clarify what 'good' means here. Can an example from an alternative method be added? (perhaps not, given persistence is not of any interest here, and numerical models are built to conserve so an unfair comparator. But some way of clarifying whether these differences are large or small in comparison to some scale of impact, or of current predictability, would be useful if one can be found). Also, the framing of the discussion on conservation (Lines 176-187 and fig. 4) is confusing to me. Is this related to the DLOP, or the FNO model. Or is this something between the two? If the FNO model, then there is no need to refer to the DLOP model. If its related to the DLOP model and isn't the FNO, then it would seem more sensible to discuss this earlier in the paper. Clarification as to whether this model is an expansion of the FNO, or an expansion of the DLOP, and the exact nature of the difference would be helpful please.

We agree with the reviewer that the framing of the discussion on conservation (Lines 176-187 and fig. 4) is not as clear, and more importantly, it is not the focus of this study. We have removed Figure 4 and its related discussion in the revision.

- Equation 4a and 4b: Use of N vs N1 is unclear to me. What's the difference here? Particularly re the left hand side of equation 4b. It also seems that H is a function of X(t) as well as of N(X(t)). I think this needs a bit more thought, and mathematical rigour, both here and throughout the paper (i.e. in the description of the loss function, in figure 3, etc), ensuring it is consistent, clear, and correct throughout. Is it best to define that H is applied to N? Or better to say that H is a function applied to X, and that N is a part of that function, i.e. $Xt+1 = H(Xt) = Xt + \mathcal{N}\ (Xt + ½\ \mathcal{N}(Xt, \varphi), \varphi)$ With $\mathcal{N}(Xt, \varphi)$ is the output from…

We have clarified this by adding to the revision, ““…where N1 represents the output of the model from the previous timestep”

- Line 201 describes that the training data undergoes a 5 day mean processing step. Is this meaned data what is used as 'truth' for comparison, and in the loss function etc, or is the time-meaned data only used as training inputs? (I would hope the former). Can this be clarified here please, i.e. something like 'all data used for training, validation and testing undergoes a 5 day running mean…'

The reviewer is correct about the 5-day running means. Following your suggestion, we have revised this sentence to:

"OceanNet for the GSM was trained on five-day running mean SSH reanalysis fields from 1993 to 2018, which helped remove high-frequency features like tides. The years 2019 and 2020 were reserved for validation and testing. All data used for training, validation, and testing underwent the same five-day running mean procedure."

- Line 260-262: It's stated that ROMs would be run with persistence if no atmospheric forecast was available, I've never heard of this being the case, especially for 120 day forecast periods. Ocean models are commonly forced using predictions from atmospheric models, and regional models are forced at the boundary by global models. feel it should be noted in the text that this is not a realistic comparator to common use cases for this kind of prediction problem. This set up of ROMS may serve as an additional baseline comparator (alongside climatology and persistence), but the limitations of this comparison should be much better noted (or ideally, forecasts from the coupled version of ROMS, run with boundary forcing from a wider domain model, should be used for the full forecast period). There's a big difference between no forcing, as applied to the ML methods, and incorrect forcing (which is the case if using persistence for 120 days to force ROMS). I don't think ROMS would be used in this way for predicting over 120 days (or even over 10 days), but if there are examples of this it would be good to reference them in the text here.

In regional ocean forecasting, defining surface and boundary forcing is a significant challenge, particularly when accurate and continuous global ocean and atmosphere forecasting data for extended periods is unavailable. Currently, almost all global ocean and atmosphere forecasts extend only 7-10 days, whereas our research requires forecasts with a

duration of 120 days. None of the operational global models routinely provide such long-term forecasts.

Persistence, in this context, refers to the assumption that future conditions will resemble past conditions. This approach can be used to define boundary and surface forcing for regional-scale models. However, as the reviewer correctly pointed out, it is important to understand the limitations of persistence and how it is applied in practice.

Persistence is commonly used in short- to medium-term ocean forecasting due to its simplicity (e.g., Jacox et al., 2020). However, it does not account for changes in climatic conditions, such as those driven by El Niño or other large-scale climate phenomena. While persistence can provide a baseline, it is not expected to capture full variability or trends in long-term forecasts.

We acknowledge the limitation of using persistent forcing in driving ROMS forecasts. This limitation lies not with ROMS, as a dynamical model, but with the specific ROMS forecast configuration that we adopted in this study. We have incorporated this discussion and the reference below in the revision, and we thank the reviewer for this insightful comment.

Reference:

Michael G. Jacox, Michael A. Alexander, Samantha Siedlecki, Ke Chen, Young-Oh Kwon, Stephanie Brodie, Ivonne Ortiz, Desiree Tommasi, Matthew J. Widlansky, Daniel Barrie, Antonietta Capotondi, Wei Cheng, Emanuele Di Lorenzo, Christopher Edwards, Jerome Fiechter, Paula Fratantoni, Elliott L. Hazen, Albert J. Hermann, Arun Kumar, Arthur J. Miller, Douglas Pirhalla, Mercedes Pozo Buil, Sulagna Ray, Scott C. Sheridan, Aneesh Subramanian, Philip Thompson, Lesley Thorne, Hariharasubramanian Annamalai, Kerim Aydin, Steven J. Bograd, Roger B. Griffis, Kelly Kearney, Hyemi Kim, Annarita Mariotti, Mark Merrifield, Ryan Rykaczewski (2020), Seasonal-to-interannual prediction of North American coastal marine ecosystems: Forecast methods, mechanisms of predictability, and priority developments, Progress in Oceanography, Volume 183,2020, 102307, ISSN 0079-6611, https://doi.org/10.1016/j.pocean.2020.102307.

- Figure 8: I think the pdf on the left is unnecessary and therefore a bit confusing. If kept it needs more clarification as to what it is (presumably, it’s what was used to calculate the saturation metric?), but I don’t think it’s needed.

Yes, the PDFs on the left side of Figure 8 are used to calculate the saturation metrics, which determined 95% of the means derived from the random pair. We have followed the reviewer’s suggestion and removed the PDFs on the figure. The figure (now figure 7) retains the saturation value on the plot, but is explained more thoroughly in the caption:

“OceanNet’s performance metrics in the northwest Atlantic: RMSE (top), CC (middle), and MHD (bottom), compared to the persistence forecast and ROMS dynamical model forecast. The performance statistics, calculated based on forecasts of 0-120 days, are displayed as mean values (lines) with standard deviations (shading). The black horizontal dashed lines denote saturation values, which are determined as 95% of the means derived from 1,000 pairs of

random images in the reanalysis dataset. These representations illustrate how each method's statistics compare with the target SSH from the reanalysis dataset".

- Figure 9, 10, 11 need a bit more clarity – having the legend on only the FNO plot is misleading, better to add this as a label down the side, similar to the DLOP/FNO labels across the top. Or at least as a legend on both DLOP and FNO plots.

Following the reviewer's suggestion, we have revised Figure 9-11 by adding legends to DLOP plots as well.

- In figure 9, why don't panes c and d have the value at day 60? Presumably because of the level of instability, but worth noting this briefly in the caption, assuming it's related to the instability of this model.

Due to instabilities in the model runs shown in (c) and (d), their RMSE values are so large that they cannot be meaningfully represented in the plot. We have noted this issue in the revised caption of Figure 9.

- Figure 9 caption (and various places through the paper, i.e. figure 12) refer to the various runs as ensemble members, instead I would recommend using 'model runs' or something else. Ensemble members predict for the same period and are in some way perturbed to give a variety of predictions for one specific time, but this isn't what's being done here as far as I can tell. These instead are multiple instances of model runs, from different start times. Using 'ensemble members' makes it confusing as to what they are.

Following the reviewer's suggestion, we have changed "ensemble members" to "model runs" in the revision.

- Line 288: there isn't a run with no integration scheme ('a lack thereof'), presumably a simple euler first order is used, or some simple addition? The most basic form of integration is still an integration scheme. It needs to be clarified here what that was, even if its trivial.

We have refined this sentence in the revision to improve the readability:

"The RMSE, anomaly correlation coefficient (ACC), and MHD are compared across different iterations of the DLOP and FNO models, focusing on integration schemes and loss function terms. The two integration schemes compared were the absence of integration and PEC. The loss function terms compared were MSE and MSE with spectral regularization. This combination of model types, integration schemes, and loss function terms results in eight models to compare, following the same approach as before (ensembled metrics, Fig. 7), against each other and with ROMS and persistence predictions."

- Line 292 -294 Needs a bit of clarification.

We have revised this sentence as follows, to improve its clarity:

"RMSE not only indicates the magnitude of values present but also serves as a measure of accuracy and stability. A high RMSE suggests that the magnitudes in the analyzed field are, on average, less realistic. If RMSE continues to increase over time, it implies that the model is becoming unstable."

- Line 331 needs to be clear that this is ROMS with incorrect persistence forcing, rather than the more meaningful application of ROMS.

Please see our responses above to reviewer's comments on line 260-262.

- Line 342-345: This sentence doesn't make sense to me. More importantly, again, regional models are most often forced with predictions from global models. I don't know of any cases where predictions (especially long sub-seasonal to seasonal), are made using persistence for boundary conditions, as seems to be implied here. My bigger concern though is the persistence for atmospheric forcing --- I suspect this has even greater impact, especially on SSH fields, than the boundary forcing, and isn't mentioned at all here. I think the caveat needs to be clear, and include the atmospheric aspect.

Please see our responses above to reviewer's comments on line 260-262.