Review of the paper:

Long-term Prediction of the Gulf Stream Meander Using OceanNet: a Principled Neural Operator-based Digital Twin

by M. Gray et al.

1. General comments

The paper introduces a new deep-learning based model, ***OceanNet***, to predict Sea Surface Height (SSH) fields with physically reasonable predictions over at least 60 days. The authors used a high-resolution ocean reanalysis dataset (1993-2022) to train different neural architectures of their model over a Gulf Stream related domain. They also use Regional Ocean Model Simulations (ROMS) as a baseline for benchmarking in their tests. The paper is overall well written and presents promising results to improve the SSH predictability over a Gulf Stream related area., based on state-of-the-art deep learning strategies. I recommend publication after minor review taking into account some additional questions (2) and correction of typing errors (3). More precisely, I think the authors should consider to improve the presentation of Sections 2.2 to 2.4 by clearly stating from the beginning of these Sections, within a small paragraph for instance, which architectures are tested amongst all the developments. Giving clear names to each architecture would also be helpful I think. An additional table to summarize all their developments (UNet, UNet/integration scheme, UNet/integration scheme with shorter timesteps, FNO) would also help. The same applies for the Section Results where only Figures are displayed but a summary Table would be valuable.

2. Specific comments:

Q.1: l.65. Even if this is more explained in the following sections, can you precise from here what you intend by “predictability” of the SSH, which can be a very general formulation?

Q.2: l.96. In the paper, you state that you used EnKF for producing the reanalysis, which has no knowledge of future observations. I am not sure that it would be a problem to have reanalysis with knowledge from both past and future observations. In the end, what you need for training OceanNet is a sequence of Ground Truth (GT) that you will use to compare OceanNet predictions vs GT. So having the best reanalysis available would be compatible with your framework no?

Q.3: l.96. I would clearly state this as an Equation in the text, something like:
x*_{t+4} = DLOP(x*_t)

Q4. In the legend of Fig.3, the training losses are not yet introduced. You may want to let the readers know that all the related notations are defined in Section 2.5.

Q5. l.172. You precise that the model has 80*10e6 trainable parameters which is rather big. Can you give some information here or in Section 2.5 about the current computational capabilities you needed to train the model and how long it is (not only for inference which is very fast)?

Q6. Fig.4. Can you provide some complementary analyses regarding the plateau reached at day 40 by residual geostrophy? You mentioned it for PV but not for residual geostrophy. I guess there is a clear explanation why the plateau is reached by geostrophy at the same day that the model stops agreeing with reanalysis PV (since they are connected). Maybe for the reader not a specialist in oceanography, it would be great to precise how these variables are connected.

Q7. l199. After reading the entire Section 2, I got the feeling that it would be valuable to add a Table here to summarize all the 4 architectures (*2 with the loss functions after but not yet presented at this point of the paper) tested among DLOP and OceanNet with specific architectures, pros and cons. This would also help the reader having an overview of the methods tested here.

Q8. Eqs 5a and Eqs 5b. You did not precise what are the notations \hat{F}*x and* *hat{F}*y, even if we easily understand this corresponds to the Fourier decomposition. The same for |*hat{F}*x|_k for the kth mode.

Q9. Eq.6. It is not clear to me how *a* and *b* are defined. You state that they are defined to agree with the MSE order of magnitude, which is clear but is it defined once and for all during training while the MSE magnitude may vary from batch to batch no? It is always tricky to define some weighting between multiple losses for sure but some comments added on that point would be great. What would happen if MSE prevails on spectral regularization? And the opposite? Would these parameters could have been trainable?

Q10. l.290. Similarly to Q.7, adding a Table to intercompare the 8 models, together with ROMS and persistence, with the metrics presented would be a good way to summarize everything, not only with Figs.7-9.

Q11. Conclusion. You identified the generalization of your work as an issue. Can you provide some ideas to overcome this problem? Maybe a decomposition of the global domain with specific training for each area? Even better, do you think there is a way to inject more information as inputs. For the specific SSH prediction, what would be great to add for instance to ease transfer learning on other domains (atmospheric forcings?, addition of physical constraints in the losses?)

3. Technical corrections

Please find below a list of grammatical or typing errors to consider before publication:

l.27. reattached the to GSM -> to the
l.175. What is the arrow in front of Eq.3?
l.251. an example of it can… -> One example can

l.254. validations images -> may be replaced by ground truth reanalyses?

l.276. that that

l.278. If the means of two objects of comparisons… -> replace by: if the average metric of two models…?

l.282. can still outperforms -> outperform