General Comments:

This paper addresses a very relevant topic, within the scope of Ocean Science, and does a very interesting job of assessing this. I do question if the work might be better placed in Geoscientific Model Development, but it certainly isn't out of place here.
The novel concept of data-driven ocean models is discussed, with a focus on the Gulf Stream through use of (predominantly) a Fourier neural operator based model, and with some comparison to a UNet based approach.
Methods are solid, a thorough assessment of results is given, and suitable conclusions derived. However, I feel results need to be caveated more given the limitations of the ROMS comparison used (see specific comments below).
Overall the work is really interesting. The paper is nicely presented, on the whole the narrative is clear, and figures compliment the text.
I feel the paper would benefit from a little more rigour in places. Specifically, labels and text on figures need a bit of careful consideration (see some specific notes below). Most importantly, the mathematics needs to be reassessed to ensure it is precise. It's important that throughout the paper variables are clearly explained, and used consistently, especially with differentiation between predicted values and true values (it seems that in places the same descriptor is used interchangeably for the true value and predicted value --- use of the hat symbol would greatly help readability in that respect).

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.

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…'.

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.

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.

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?

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.

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.

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…."

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 ^.

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

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).

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.

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.

$$X_{t+1} = H(X_t) = X_t + \mathcal{N}\left(X_t + \frac{1}{2}\mathcal{N}(X_t, \varphi),\ \varphi\right)$$

With $\mathcal{N}(X_t, \varphi)$ is the output from…

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…'

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. I 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.

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.

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.

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.

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.

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.

Line 292 -294 Needs a bit of clarification.

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

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.

Technical comments:
In many places (i.e. Lines 3, 55, 115, 117, 132, 139, 252, and very possibly elsewhere) a hyphen is used when it should be an em-dash. This needs to be checked throughout the paper, not just the lines mentioned here as my search was not extensive.

Spelling of Chattopadhyay for 2023 paper missing the first a throughout; in both the bibliography and when cited in text (i.e. it seems an issue in the way its stored in the referencing software)

Line 86/87 citation for He, and Wu and He, are both missing years.

Line 86, I don't think the reference to figure 1 is relevant here, without a mention to the domain of OceanNet.

Line 127, implies lower CC, I assume this isn't the case(!)

Line 129-131, '….an inaccuracy of interest…'. This sentence is very hard to process, can it be amended please.

Equation 5a and 5b need $\hat{\mathcal{F}}$ to be defined.

Line 242, I would say these metrics are '…widely employed in assessment of data driven atmospheric forecasts' (rather than just 'forecasting'), given the references being pointed to.

Line 250, simply say 'an example of this contour can be seen in figure 1b'. The comment about it not being shown in other figures in the paper is unnecessary and a bit confusing.

Line 270, clarify that the data is (presumably) spatially averaged before averaging across multiple runs.

Line 282 outperforms -> outperform

Line 259, comma needed: '…compared were MSE, and MSE with…'