We thank the reviewers for their suggestions and questions. In general, we have expanded the results on real data by including new reconstruction examples, and better emphasized the difference between the simulated and real results. These changes required some reformatting of sections. We have made changes in response to specific points. Our responses are typeset in blue below.

Reviewer 2: General Comments

A. The main claim of the authors seems to be that one can train with hydrodynamical model data alone and apply the trained neural network to real data. But to really make this claim authors should do the training on satellite images and compare the reconstruction of a trained model based on hydrodynamical model to really assess the "sim-to-real" gap in a comparable setting (exactly the same domain, time interval).
We agree that training and testing on real satellite data can supplement the sim-to-real discussion. We have foregone this because an equivalent domain and time interval for satellite data is a very small dataset. After filtering out SST tiles with more than 10% land pixels, there are a total of 1,230,688 tiles, all of which are "clear" when using simulated data. However, using real satellite data, only 7,544 (0.6% of the total) tiles have less than 10% cloudy pixels. Due to the limited size of real satellite training datasets, we instead focused on evaluating different model configurations for synthetic data. We have edited the manuscript to describe the dearth of cloud-free satellite SST data for ML training, and highlighted that simulated SST is beneficial because it results in a significantly larger dataset.

B. The error estimate is also problematic. It seems that the method can only predict the error for a whole tile while other techniques are able to provide an error per pixel. However the error variance is expected to be highly spatial dependent. This limitation is not mentioned in the manuscript. The paper also mentions that one could empirically deduce the RMSE from the cloud coverage and the standard deviation of overlapping tiles. However this approach is not demonstrated and validated quantitatively in a statistical manner. Also it should be noted that empirical correlation coefficients should not be deduced on the test dataset as the test dataset should be independent. It would also be important to check the realism of the error estimation on real data not just simulated data.
We understand the concerns about our error estimates. Our discussion of error estimates is not core to the paper; rather, it is a supplemental result leveraging auxiliary information from our overlapping reconstructions. We have removed this section since it is not a main result, and our analysis on the error estimates is not as thorough as our other analyses. As an aside, we note that not all SST reconstruction methods are able to provide error estimates, and that the spatial dependence of the variance is a consequence of cloud cover patterns and is not unique to our method.

C. The power spectrum analysis should be done on the real observations, instead (or in addition) to the simulated model data. It is indeed an interesting approach to train a neural network on hydrodynamical model data but the validation should focus on the observations (see also point B).
We thank the reviewer for pointing this out - we have added more analysis for the real observations, including PSD and coherence analysis in **Section 5.1**.

D. Comparison with other techniques are not as straightforward as the authors would like it to be as the reported RMS errors are not for the exactly same domains and for the same input data.
Good point - our comparisons were not meant to be direct for the mentioned reasons, so we have removed our comparisons to the RMSEs of previous works. We are still keeping Table 1, which summarizes the domains and inputs of prior works, but without the RMSEs.

Reviewer 2: Specific Comments

L47: Zupančič Muc did compare their work to MAESSTRO, but I would not say that they based their work on MAESSTRO. Zupančič Muc uses a ViT.
We meant that MAESSTRO used ViTs earlier than CRITER, but did not intend to suggest any direct extensions of MAESSTRO. We have clarified this in the text.

Table 1: It is difficult to compare the RMSE from different regions and different datasets. There is too little context to compare the value. A table would be sensible if we would use a standardized dataset but this is not the case.
We have removed the RMSEs from the table, but kept the information about the domain and extent of the prior studies.

L117: “The satellite MW data is bi-linearly upsampled across time to match the 6-hourly temporal resolution of the L3S data.” Do you use a 1d or 2d interpolation?
This is a typo. For the satellite MW data, we perform 1D interpolation across *time*. For both the simulated and satellite MW data, we perform 2D interpolation across *space*.

Equation 7: “We set γ1 = 1 and γ2 = 5 so that the SST and SST gradient losses have similar magnitudes during training.” T and |\nabla_xy T| do not have the same units. So γ1 and γ2 cannot be both adimensional at the same time? Unless, they use as gradient the finite different without dividing by the resolution (but this is not the gradient as it is defined mathematically)
Good point - the unit of the weights are set to make the overall loss adimensional. We have updated the text to reflect this.

L178: replace 1e^{-5} by 10^{-5} (and similar)
Fixed.

Equation 8: I guess that k is the norm of the wave number vector, please clarify. It is also not very clear CSD and PSD is done on time series, 2D field or 3D field. Please also include a reference for the approach as you used it. Reading the source code (assuming it is the file figure_scripts/plot_npo_sst.py) clarifies their approach. But it also shows that there is a lot more to it than what is mentioned in the manuscript: use of Hann window, linear detrending. Also the analysis is done on the x and y axis independently and then averaged is not mentioned (i.e. no real 2D analysis). The MITgcm LLC4320 has a resolution of 1/48° meaning that the resolution in km is different for the x and y axis. Can you clarify how this is taken into account?
We appreciate the feedback on our spectral analysis, and have made the following changes to strengthen our results. The coherence and PSD analysis has been updated to be done on 1D longitudinal cross sections, to avoid the disparity between x and y resolutions. The manuscript has been updated to state the additional steps used in PSD calculation.

L252: “The cross-section also shows a slight oscillating artifact in the IR reconstruction”: What is the origin of those oscillations?
The oscillations are an artifact of the ML model processing. Unfortunately it is hard to pinpoint the exact cause of this, but anecdotally, this artifact only appeared in single-satellite reconstructions with extremely high cloud cover. Our best performing (multi-satellite) model does not exhibit this behavior.

Figure 9: Please add also the spectrum of the original IR data on this figure. This is important information to assess variance is potentially missing at small scales. Also making this comparison to real satellite observations (as opposed to model data) would be much more useful (in addition or instead). Consider using an IR image with no (or very few missing data) where you add artificial clouds for the reconstruction.

We have added the spectrum of the IR data. We have also added spectral analysis on real data.

Table 3 and 4: For correlation of SST, please also provide the correlation with the seasonal cycle removed.
We originally reported two correlation values, one calculated on SST and the other calculated on the high-frequency SST features (SST - microwave SST). The former value tends to be high, so we removed it because it does not meaningfully distinguish model performance. We have clarified that the correlation reported in our revised manuscript has the seasonal component removed.