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 1: General Comments

A. Authors mention foundation SST only once in the introduction, but never in the main text. It is an important concept for SST especially if linked to the numerical model SST. I suggest they discuss what they mean by foundation SST, how it is linked to the LLC4320 first model layer. How it is defined in the L3S and L4 products used. How the diurnal cycle represented model's hourly outputs projects onto the foundation SST?
We appreciate the reviewer's points. Our reference to "foundation SST" in the introduction was intended to mean "large-scale SST", and we have updated the text to say this instead. Indeed, the foundation SST from L3S products was used but the model (LLC4320) does have diurnal cycle from the hourly snapshots. This difference theoretically imposes inconsistency but the effect on model inference is small as shown in our results. It is most likely due to the finite depth of the model top layer with ~1 m thickness. The diurnal cycle in LLC4320 is much smaller than the diurnal cycle shown in the skin temperature (top ~10 micrometers) in satellites. We added the following text in the **Training Dataset** section:
Note that the model (LLC4320) SST represents a thin top layer with thickness about 1 meter. Its hourly temperature snapshots still have a diurnal cycle, but less significant than the diurnal cycle in (sub)skin temperature measured by satellites that represents the upper 10s micrometers to millimeters. The dominant variability in the SST images are in their spatial structures than the diurnal bias and the results are not very sensitive.

B. The study is clearly conducted for climate applications. However, using future data is a limitation for short term forecasting applications such as MHW. It would be useful to discuss if the distribution of data and the skill of MUSE would change if only the past data would be used as input. In other words, would the skill be similar if the last day of the input is reconstructed instead of the day in the middle?
As noted by the reviewer, reconstruction of the last day is useful for forecasting applications, but may result in worse performance than reconstructing the middle day for our model. Reconstructing the last day is an autoregressive problem, which can be approached using different ML architectures for better results. We note that this work is useful beyond forecasting because it focuses on global high-resolution SST reconstruction, which is helpful for studying submesoscale ocean dynamics. We added text about this to our **Discussion** section.

Reviewer 1: Specific Comments

Figure 1: No test or validation in boreal summer? Do you expect an impact of land distribution on the separation of dataset?
Our full dataset consists of 14 months of data, and was split to prevent overlapping months in the train and test dataset, which resulted in no boreal summer data in the test set. Future work can use multi-year data so that the test set can include boreal summer data. Despite the lack of boreal summer data, MUSE is able to generalize to unseen test months in this study (December, January). We are unsure of what "separation of dataset" means in this context. If "separation" refers to test/train split, the split is done across dates and not space, so the land distribution is balanced across the dataset splits.

L112: Is there a reason why clouds from forcing aren't used instead of the L3 product? Wouldn't it be more consistent with the model SST during training?

The LLC4320 dataset includes atmosphere heat flux product from ECMWF, but cloud cover was not explicitly simulated or represented. For this reason, we apply the L3S cloud masks.

L209: The middle timestep is maximally correlated to all input timesteps, so reconstructing the middle timestep is easier than reconstructing all input timesteps.
We have updated our explanation of why middle step reconstruction is easier, and moved discussion of the ablation study results to the **Discussion** section.

L218: Why the channel method does not learn enough deserves a more solid justification. Is it the "multivariate" nature of the channel approach that degrades the correlations? Is MW used only on the gaps or everywhere when used as a channel?
For the channel method, the entire MW image is used. There are a few reasons why the MW channel method is worse. First, as noted by the reviewer, the multivariate nature of the channel method degrades performance because both time and data modality vary across the channel dimension. Second, the channel method doubles the input channel dimension, and ML models can struggle to learn long-range dependencies across the input. We have added these reasons to the **Discussion** section.

L292: The SST RMSE is 0.037°C, and the gradient SST RMSE is 0.012 °C km-1. Please discuss if these errors are realistic.
The errors reported here are for the simulated SST, which don't translate to real data, but establish a potential lower-bound on real SST reconstruction error. We've clarified this in the text. Errors on real data, shown in Table 4, are indeed higher (next response explains why this happens).

L337: The RMSE using real observations instead of simulated ones is 3-5 times more. Can the degradation of the skill using real observations be due to the mismatch between the foundation SST and model SST? Would it be better in case subskin SST is used? What is the first model depth?
The LLC4320 model depth is 1 m, while the real data we use is foundation (L4 MUR 0.25°) and subskin (L3S LEO STAR 0.02°) SST. As the reviewer notes, there is a mismatch in SST depth for our simulated and real dataset, which degrades performance. We combatted this issue by calculating an additive bias between the L4 and L3S SST, which is described in **Section 2.2.2**. Despite the SST depth differences, the real data RMSE are still low (0.156°C), indicating that the model is robust to some degree of dataset differences.

L324: Please explain precisely how daily L3 SST becomes an input to an 8-time model.
We use the L3S SST, which is approximately 6-hour snapshots because we combine observations from AM and PM datasets, and we use 2 days of data to get an input with 8 timesteps. We also use the L4 MUR 0.25° *daily* SST, which is upsampled linearly to match the 6-hourly snapshots of the L3S dataset. Dataset preparation is described in **Section 2**, and we streamlined the text to improve clarity.

L341: What are other ways of mitigating OOD beyond preprocessing? If there are ways, authors should justify why they haven't used them.
Another method of mitigating the OOD effect is to *fine-tune* the model by training on a small amount of satellite data after training on the full simulated dataset. This study focuses on the methodology for preparing synthetic data for ML training, and we present a preliminary analysis on real data that has some major limitations (see the response about SST depth mismatch). This paper is a proof-of-concept for training on global synthetic SST, and our future work will use model fine-tuning to improve performance.