We deeply appreciate the reviewers comments and their effort in reading and reviewing our manuscript. The reviews have helped us improve the study and manuscript. Below is our point-by-point reply to the review comments with the reply marked in blue.

REVIEWER 2

This paper describes an application of a masked autoencoder (MAE) applied to model sea surface temperature to reconstruct missing observations. The performance of the MAE is compared to kriging and radial-basis cubic interpolation. It is shown that the MAE provides a better accuracy than the other methods for the model data, in particular at small scales. Finally the method is tested on a single SST image from VIIRS. The results are quite encouraging.

Major comments:

In general the methodology section of the masked autoencoder does not provide enough details for the typical audience of Ocean Sciences. Please include more information about the network architecture, tensor sizes, and all hyperparameters involved. Did you implement the MAE from scratch or did you adapt an existing implementation? In the later case, please also reference the base implementation.

We thank the reviewer for the important question regarding MAESSTRO's implementation. We have added a table (Table A1) in Appendix A that provides details about the network architecture.

-

The data missing SST are not just random pixels, but they have a spatial extent (the size of clouds). It would be important to test methods in this context. Otherwise the results of the validation would be too optimistic. In the discussion the authors note this themselves, but they have not taken this problem in account (beside noting it as future work).

This is a good point, which we discussed in the original submission but without showing a demonstration. Figure 11 has been updated to reflect this, please also see our response to Reviewer 1 for a detailed description.

The size of the validation/test dataset is not always clear or very small (a single image for the real SST image). Please use a large validation/test dataset to compute the error statistics. Of course it is fine to show a single or only a few representative images in the manuscript.

We indeed have conducted an extensive validation using the simulated data and discussed it in the paper. It was our oversight of not making it clear. The details are now included in table A1. Specifically we used 250,447 images for validation and built the global statistical analysis shown in Figure 7-9.

The MAE decomposes the image in patches of 4x4: how does it work in practice when clouds do not occupy regular patches of size 4x4 ? if within the 16 pixels of a 4x4 patch a single pixel is missing, the entire patch is considered as missing? This can quite significantly increase the amount of missing data.

In our current version, we discard any 4x4 patch that contains even a single missing value due to cloud cover, as shown in Figure 11. This approach is not ideal, but it is deliberately designed to demonstrate the worst-case scenario that MAESSTRO must handle. Our future development will focus on improving the algorithm to manage patches with minimal missing values. This may include applying interpolation to fill in small-sized cloud cover before feeding it to the machine learning model, or working with smaller areas and smaller pixel sizes to retain more valid data points. These enhancements are part of an ongoing effort.

All figures, please make sure to always mention the units of the variable. In particular for the SST gradient.

Thanks. It is done.

I would propose major revision before publication in OS.

Specific comments:

Line 8: “It has exceptional efficiency, requiring three orders of magnitude (a factor of 5000) less time.” Compared to what?

We compared the MAESSTRO with the interpolation methods used in the paper as our benchmarking. We have revised the manuscript accordingly.

Line 84: “To build a model for SST, using real satellite SST imagery as ground truth would be ideal. However, these images often contain noise and are susceptible to bias and errors. As an initial conceptual demonstration, this paper employs synthetic satellite sea surface temperature (SST) data derived from two high-resolution numerical simulations …”

The motivation is not clear. Would you not expect that the errors in models are even larger than in satellite data? Can the masked autoencoder be trained based on gappy data?

To provide further clarification, our assertion is that SST in simulations is unaffected by random noise and measurement errors commonly associated with data collection platforms and corruption. This eliminates the necessity for preprocessing and cleansing of satellite data to eradicate these errors, allowing us to concentrate specifically on assessing the feasibility analysis of MAESSTRO. For actual satellite-derived SST data, our subsequent step involves developing new or using existing pipeline to preprocess satellite SST images before feeding into any ML for training or reconstruction.

Section 2.3, line 115:

“To resize the tiles, a random portion of the full tile is cropped, ranging from 20% to 100% of the original tile, before being resized to the final 128x128 dimensions using bicubic interpolation.” Does this mean that the neural network gets images which are not always at the same spatial resolution? Is the actual resolution provided to the neural network? If not, this can be a problem as the energy/variance is not distributed uniformly across scales.

The reviewer's observation is accurate: altering the size of the original image does indeed modify its spatial resolution. Our machine learning model, constructed with this consideration, is designed to be invariant to scale and resolution. While this method may not adhere to the principles of energy or variance conservation, it effectively captures and replicates the spatial structures of SST across various resolutions and spatial scales. This characteristic underpins the model's robust performance when trained on the higher resolution llc4320 data and subsequently applied to the llc2160 dataset, which possesses half the resolution.

Figure 2: please clarify if this image is from the training, validation or test dataset. (If the image shows the training dataset, please use an image from the validation or test dataset in addition or instead of figure 2)

Figure 2 is now regenerated and the caption was updated to reflect that the images were taken from the validation dataset.

Line 144: “false-color RGB images of SST from the LLC4320 validation”: I don’t understand, is the SST image treated as a 3 channel RGB image rendered using some colorbar? SST is a scalar variable, so a single channel tensor should be sufficient.

We aim to assess the efficacy of the original FB-MAE in reconstructing data without being specifically trained on the SST field. As noted by the reviewer, FB-MAE is designed to process three-channel images, whereas SST is inherently a scalar variable. To address this, we converted single-layer SST measurements into a three-channel image format. This involved normalizing the SST ‘image’, inputting it into FB-MAE, and then applying the inverse of the normalization process to restore the SST original scale. We revised the text to improve the explanations.

Line 152: “While the original MAE implementation He et al. (2022) uses the mean squared error (MSE) between the reconstructed and original pixel values, MAESTRO uses the root-mean-square error (RMSE) in order to recover the same units'': Why should this matter, as the minimum of the loss function is the same? Once we have the MSE, one can compute the RMSE by just applying the square root.

Using either MSE or RMSE does not directly affect the machine learning training and validation. However, we used RMSE for the convenience of comparing error evaluations in units of degrees Celsius, rather than using a variance-type evaluation in units of degrees Celsius squared. But the reviewer is right that we can just take the square root of the MSE to get RMSE. This is not something we need to emphasize. We acknowledge the potential confusion caused by our original statement and have revised the text to mention only the RMSE that was used.

Line 164: “is the cross-spectral density along the x-axis (each row) …” why just considering the x-axis? Can this metric be made rational invariant?

We have tested the evaluation along the y-axis without qualitatively altering our conclusion. This outcome was achieved because we utilized a large number of ensembles in validation, effectively sampling numerous snapshots, making the evaluation direction-agnostic. We have added a sentence to the text to clarify this point.

“Section 3 Evaluation on a sample SST tile from LLC2160”: How the parameters involved in the Kriging operation are chosen and which kriging variant is used (ordinary, simple, universal…) ? In particular, how is the variogram determined and are the observations assumed to be noise free? And If not, what noise level is used?

These are good questions. We worked on the anomaly field from the domain mean in each snapshot and used OrdinaryKriging. We did not conduct a very extensive study on the optimal parameters for Gaussian-Kriging but chose a set of parameters based on a series of evaluations on the snapshot used in Figure 5. The linear-Kriging automatically derived the linear variogram from the provided cloud-masked data. In the Gaussian kernel, we used nugget=0.001degC effectively assuming noise free SST, which is OK for our numerical-simulation-based work here but will be different in reality when dealing with actual satellite measurements. Dealing with noisy observations is one of our ongoing studies.

Line 175: “radial-basis bicubic interpolation” Can you give the equations of this interpolation method. Can it account for noise in the observations?

The mathematical form of the Cubic RBF is given by:

$\phi(r) = r^3$, where $\phi(r)$ is the cubic RBF, r is the radial distance from the center of the function where we would retrieve the interpolation. A linear combination of a set of these cubic RBF will yield a broad smooth function that can be best fit to scattered unstructured data as the cloud-covered SST images. We used the algorithm implemented in scipy.interpolation.RBFInterpolator. The smooth factor was set to 0 to perfectly fit the data at the available data point as we assumed zero noise, but it in principle can accommodate specified noise level by adjusting the smooth factor (whether the fitted surface goes through observations exactly or not). We added some description in the text.

186: “Kriging with linear and Gaussian variogram”: I do not understand what a linear variogram is. A variogram should tend to zero for small distances and to a constant value for large distances. Do you use a piecewise linear function for a variogram? If yes, how you choose the threshold values. It is also not clear how the parameters of the Gaussian variogram were determined. Please provide more information.

The linear variogram was the simplest assumed form. It does not have a sill (constant value for large distances), but the image size, 128x128 in our case, effectively imposes a range. The slope and nugget of the linear variogram is derived from each masked image directly to best fit the available data without a fixed set of prior parameters. The Gaussian variogram on the other hand was chosen to be (0.1ºC,20 pixels, 0.001ºC) for the sill, range and nugget, respectively. This set of values are derived from the snapshot in Figure 4. We have added these discussions in the text

“Table 2: Evaluation metrics for the single-tile example shown in Figures 5 and 4 for different reconstruction methods.” What is actually the training, validation and test split of the dataset here? Please extend the evolution metric to the whole (unseen) test/validation data. Typically the test and validation data are about 10% (or more) of the trending dataset to achieve robust error statistics.

Thanks for the questions. The MAESSTRO was trained on the 1/48-degree resolution simulation (llc4320), while the evaluation images used in figure 4, 5 were taken from the 1/24-degree simulation (llc2160) independent of the 1/24-degree version. The details on the training, validation and test split are not included in table A1. We realized the inadequacy in our first submission and hope that the new text added some clarity.

Line 220: how are the missing pixels chosen for the “Global validation results on LLC2160”. Do the gaps have a realistic spatial extent?

In our validation datasets sourced from LLC2160, we have employed a consistent random masking strategy, where the masks were generated by the machine learning code automatically. Realistic cloud coverage was not incorporated in our first submission. We have now conducted additional tests that include cloud masks derived from VIIRS data. These masks were applied to the simulations to enhance the realism and applicability of our results. (Figure11)

253 “cubic interpolation (the best-performing baseline)...”: It is surprising that cubic interpolation is the best-performing baseline. Must current techniques use optimal interpolation (similar to Kirging). How much effort was placed in optimizing the Kriging interpolation? Please keep the same name of the method as before “radial-basis bicubic interpolation” as an ordinary cubic interpolation does not involve a radial-basis function.

Thanks for the comment. We have not done extensive optimization on the Kriging but did conduct a set of tests to choose the Kriging parameter space based on the case shown in Figure 4 based on a case of 80% missing data. We have revised the text and refer to the radial-basis cubic interpolation consistently as cubic-RBF.

Line 235: “SST gradient typically has a standard deviation of 0.1-0.3 ◦ C” and Figure 9: I do not understand why the gradient of SST has the units °C rather than °C / km (or any other length scale).

Thank you, line 235 has been updated with 0.1-0.3 °C/pixel

Line 269: “To substantiate our methodology, we tested it using real satellite Sea Surface Temperature (SST) data from the Suomi NPP Visible Infrared Imaging Radiometer Suite (VIIRS) in the California Current region, specifically at coordinates (35◦ N, 125◦ W) on January 16, 2021 (Figure 11, left).”. To really validate a technique it is not sufficient to take a single snapshot. One would need to provide error statistics over several images to provide a robust estimate.

We appreciate and agree with the comment. Here we use a single snapshot to demonstrate that the trained ML can apply to satellite data and it is transferable. More extensive statistical evaluation, however, needs substantial effort and worth a standalone study, so we dedicate our next manuscript to a focused analysis with real satellite images.

Line 275: “showing edge artifacts in 4x4 patches” Where do these patches come from? I think it is essential to discuss network architecture.

We removed the need to patchify the one image into 4x4 smaller images. In the revision, the ML was applied to the whole image and the edge artifacts were removed (Figure 11)

Line 309: “while traditional methods require approximately 24 hours” can you be more specific which traditional methods you are comparing to here?

Thank you for your comment, line 309 has been updated. We replaced ‘traditional methods’ with specific reference to the cubic-RBF and Kriging methods we used in the paper.

Figure 11: please show also the clouded image that you used as an input.

Thank you for your suggestion, Figure 11 has been updated to show the cloud masks.

In general, as per OS policy, verify that the image is also accessible to people with vision deficiencies (https://www.ocean-science.net/submission.html). I am not sure if Figure 11 is ok.

Thank you for your comment, Figure 11 has been updated with a new colormap.

(Very) minor comments in the references:

In general please use DOIs when they are available

ALVERA-AZCÁRATE: -> Alvera-Azcárate

Thanks. It is done.

Change https://doi.org/https://doi.org/10.1175/2007JCLI1824.1 -> https://doi.org/10.1175/2007JCLI1824.1 (and other links with the same issue)

Thanks. It is done.

“JPL/OBPG/RSMAS: GHRSST Level 2P Global Sea Surface Skin Temperature from the Visible and Infrared Imager/Radiometer Suite (VIIRS) on the Suomi-NPP satellite (GDS2). Ver. 2016.2., PO.DAAC, CA, USA. Dataset accessed [YYYY-MM-DD] at https://doi.org/

10.5067/GHVRS-2PJ62, 2020."
Please provide year, month and day.

Thanks. It is done.

"Application of dincae to reconstruct the gaps in chlorophyll-a satellite observations in the south china sea and west philippine sea" -> "Application of DINCAE to Reconstruct the Gaps in Chlorophyll-a Satellite Observations in the South China Sea and West Philippine Sea

" Please check the capitalization of your references.
Thanks. It is done.