General comment

The paper by Matjaž Zupančič Muc et al. (2025) presents a novel ML-based architecture, which includes the combination of a Vision Transformer model with a U-Net, to reconstruct satellite-derived sea surface temperature values not measured by infrared sensors, mostly due to cloud coverage.
While the technical parts concerning the networks involved are mostly well explained and exhaustive, I think there is a lack of attention when dealing with the datasets involved and the presentation of the results.
I suggest publication after addressing some majors concerns explained in details below.

Specific comments

- Already in the **abstract** the authors refer to the difficulties "*to accurately recover high-frequency variability, particularly in SST gradients in ocean fronts, eddies, and filaments, which are crucial for downstream processing and predictive tasks*", but there is no mention in the paper to some evaluation of SST gradients or the scales that the network is able to resolve (e.g., not a single plot of SST gradients or some spectra). The only RMSE is not sufficient, since it is possible to improve the RMSE of a reconstruction only at large scales. I think the authors should present some plots and some metrics that can show if the network effectively resolves the small scales (i.e., submesoscale and mesoscale) of the ocean.

- In the **introduction** I think the authors are missing several papers that have dealt with the reconstruction of fine-scale features when satellite data are missing, mainly citing the papers published by a limited number of researchers. In particular, I think it should be important to consider at least:

  Buongiorno Nardelli, B., Cavaliere, D., Charles, E., & Ciani, D. (2022). Super-resolving ocean dynamics from space with computer vision algorithms. Remote Sensing, 14(5), 1159.

  Fanelli, C., Ciani, D., Pisano, A., & Buongiorno Nardelli, B. (2024). Deep learning for the super resolution of Mediterranean sea surface temperature fields. Ocean Science, 20(4), 1035-1050.

  Lloyd, D. T., Abela, A., Farrugia, R. A., Galea, A., & Valentino, G. (2021). Optically enhanced super-resolution of sea surface temperature using deep learning. IEEE Transactions on Geoscience and Remote Sensing, 60, 1-14.

  Martin, S. A., Manucharyan, G. E., & Klein, P. (2023). Synthesizing sea surface temperature and satellite altimetry observations using deep learning improves the accuracy and resolution of gridded sea surface height anomalies. Journal of Advances in Modeling Earth Systems, 15(5), e2022MS003589.

  Martin, S. A., Manucharyan, G. E., & Klein, P. (2024). Deep learning improves global satellite observations of ocean eddy dynamics. Geophysical Research Letters, 51(17), e2024GL110059.

Young, C. C., Cheng, Y. C., Lee, M. A., & Wu, J. H. (2024). Accurate reconstruction of satellite-derived SST under cloud and cloud-free areas using a physically-informed machine learning approach. Remote Sensing of Environment, 313, 114339.

Moreover, I think authors should spend more words in the introduction explaining which are the limits of infrared measurements for SST data and which are the limits of using standard statistical techniques to reconstruct missing data in satellite images to motivate their paper.

- **Section 2** starts stating that L3 SST observations will be used in the paper but they were never defined. Even if it seems obvious, it is a good practice to explain what L3 images are and which are their characteristics. Moreover, there is no explanation on the motivation of the choices of the datasets, especially about two things:
    - Why do the authors choose one Near Real Time (NRT) dataset and two reprocessed/multi-years (MY) products? The processing chains behind these products can be very different and the datasets can differ among them.
    - Why do the authors choose for the Adriatic a different product with respect to the Mediterranean one (which includes entirely the Adriatic Sea)?

- In Sec. 2.2.1 authors introduce the choice to select sequences of three days to construct the datasets for the training, can they explain the reason for this choice?

- In Sec. 2.2.2, it is not clear to me if the splitting between training/validation/test datasets is in chronological order (i.e., the test is always the last 5% of the temporal series) or they apply a shuffle before splitting the datasets.

- At the beginning of Sec. 3.3, authors state that the CRM part is “self-supervised” but then they define a loss function based on an error between the reconstruction and a “ground truth measurement”. If there is a target, then the network is not “self” supervised, but just supervised.

- Implementation details: How do authors choose N_IRM? And the number of epochs?

- Regarding the performances: why do authors compute the average of the RMSE only for 10 reconstruction?

- Row 204: Authors state that DINCAE2 is the “best SST reconstruction method” but it seems to me that this is more of an opinion and that they have not tested all the methods available in the literature to state something like that.

- Rows 210-212: It is not clear to me how authors can assess that the MAESSTRO network is limited due to the single step approach, can you please elaborate this sentence?

- I do not understand the difference between the “RMSE_all” of Table 1 and row 225 (it seems to me that it is calculated over the entire tested dataset) and the “RMSE_all” above the plots in Figs. 4, 5, 6 and the analogous in Appendix. There has to be different definitions since the values are different but, therefore, the name should

change. It is also strange that all the values “RMSE_all” in the plots are larger than the average in Table 1, are the authors showing the worst outcomes? Moreover, in general Sec. 5.2.1 presents some issues: there are a lot of small panels and only 10 lines of comments of what the images are revealing. I suggest choosing fewer samples and enlarging the size of the images that are significant in order to appreciate the differences between SST fields. Moreover, I suggest changing the colormap for the variance and the RMSE since it is almost impossible to appreciate the variations.

- In general Fig. 4,5, 6 and similar (after a very big zoom) shows a not homogeneous SST field, where the changes in the effective resolution of the SST field due to the network's reconstruction is very clear. Can authors please comment on this issue?

- Throughout the paper, the significance interval for errors is missing. Please, show them to ensure that the differences between methods are relevant.

.

Technical corrections

- Rows 18-19 page 1: Eliminate after “...approaches” the references “(Alvera-Azcarate et al., 2005), (Barth et al., 2020), (Barth et al., 2022), (Fablet et al., 2021), (Beauchamp et al., 2023), (Goh et al., 2024)”. Authors already recall all of those, specifying the techniques used, in next rows.

- Section 2.1:
    a. The way to present the datasets is incorrect. There is a standard way to cite products from the Copernicus Marine Service that can be found here: https://help.marine.copernicus.eu/en/articles/4444611-how-to-cite-copernicus-marine-products-and-services
    b. The sentence “The arc degree resolution of the measurements…” is incorrect for two reasons. First, the L3S products are merged multi-sensors products which are not at the original resolution of the data measured by the sensors, but remapped on a grid at a chosen resolution. Therefore, the products’ resolutions are 0.0625° or 0.05°, not the measurements. Moreover, it is redundant to say “arc degree resolution”, it is “spatial resolution” or “0.05° resolution”.
    c. Authors state that product X “*contains*” from day Y to day Z. Actually, all the products used include temporal series longer (and spatial coverage bigger) than the one stated in this section, so authors should either present the whole temporal series (coverage) or explain why they chose only that temporal (spatial) part.

- The word “occluded” in the title of Section 2.2.1 sounds strange, the common way to define it is “missing” or similar.

- At row 82, authors introduce W and H as dimensions but they never defined them.

- Fig. 2: the caption should explain every variable in the image.

- Row 94: The use of trigonometric functions for the day of the year is a common procedure to take into account the seasonality of SST, it was not proposed by Barth et al. (2020).

- Row 96: Authors never define D_t.

- Row 107: To be consistent throughout the paper, “1 x 8^2” should be “1 x 8 x 8”.

- Row 148: When authors state “...number of ground truth measurements that are not on land”, it confuses me. By definition, if we are talking about SEA surface temperature measurements, they are not on land.

- Row 149: “M_l (·)_(i)” should be “M_l(i) (·)”.

- Fig. 3: Colorbar are missing, even if it is not the intent of the image to show specific values of SST, they should be included, especially for the masks.

- Rows 185-187: This sentence has been already stated before, no need to repeat. Also all the definitions of the matrices.

- Row 207: Please specify what does it mean “under the same conditions”, i.e., datasets, hyperparameters, number of epochs…

- Row 263: I think a “C” is missing when referring to degree Celsius.

- Caption of Table 2: what authors mean with “both dimensionless and bias in °C”. What is dimensionless?