Reviewer 1

This paper describes a deep neural network to reconstruct subsurface thermohaline profiles from a combination of in-situ measurements and satellite data, focused on the Northwestern Pacific.

The technique by itself presents very relevant and innovative solutions to extract information from time series of satellite-based sub-images (centered on the target profile location) and combine them to correct available climatological data. However, the manuscript presents some misleading statements and some questionable claims that need to be corrected before publication. Additionally, several technical clarifications are also necessary, as detailed in the following.

Major remarks:

In different sections, the authors cite "geostrophy" as a limiting factor in designing 3D reconstructions based on satellite altimetry. This is a conceptual imprecision, as geostrophy is an approximation used to derive horizontal currents from horizontal pressure gradients and has no influence on the sea level observations (which reflect all processes driven by full dynamics). Moreover, while it is commonly stated by the community that the sea surface elevation represents an integral measure of subsurface processes, this is actually true only in purely baroclinic flows and it is not true in general. Surface elevation is useful to infer the vertical distribution of hydrographic properties because the ocean is predominantly baroclinic and surface intensified (so that pressure differences are detectable at the surface). While this is discussed in section 4, it is not correctly introduced in the Abstract. Consequently, the second and third points listed in section 4 are redundant and should be merged removing any reference to non-geostrophy and eventually referring to non-baroclinic (the two terms should not be used interchangeably).

Response: We agree that our use of the term "geostrophy" as a limiting factor was imprecise, as satellite altimetry measures the total sea surface height driven by full dynamics, not just the geostrophic component. We also agree that the connection between sea surface elevation and subsurface properties relies on the dominance of baroclinic (steric) processes, whereas barotropic (mass-loading) signals act as noise for

our specific reconstruction task.
Accordingly, we have revised the manuscript to address these points:

- Abstract (Lines 4–5): We corrected the sentence to specify that surface elevation is useful for inferring subsurface properties in regions where baroclinic processes dominate.
- Discussion (Section 4): As suggested, we have merged the second and third points into a single section titled “Barotropic and non-steric components”. We removed the term non-geostrophy and instead focused on the distinction between baroclinic signals and barotropic/non-steric signals.

The FiLM layer is not detailed to the level that would allow replicating the study. In fact, the neural network that modulates the scaling and shifting factors (conditioning network) is described nowhere (number of layers/neurons, etc.). It might also be worth adding a dedicated panel in figure 3.

Response: We agree with your comment. To address detail and ensure reproducibility, we have revised the manuscript as follows:

- Figure 3: We have added a new panel (b) to Figure 3, which provides a schematic of the conditioning layer and illustrates how the scaling (γ) and shifting (β) factors are applied.
- (Lines 134–139) We have expanded the description in Section 2.3 to specify the architecture of the conditioning network, including the number of layers and neurons.

I understood the target of the TS-cast regression model is represented by the observed T/S profiles (2 channels, 128 layers). What is the “Error Variance” 3x128 tensor in figure 3? How is it computed and what are the three channels? Are the predicted/target variables normalized before entering the model? If yes, how (full profile standardization, depth-by-depth, min/max., etc.)? How does this eventually impact the loss functions defined in section 2.4?

Response: Regarding the "Error Variance" tensor (3×128) in Figure 3, we revised the manuscript (Section 2.4) and would like to clarify the following (Lines 145–154):

- Components of the Tensor: The three channels correspond to the predicted uncertainties for Temperature (T), Salinity (S), and Density (ρ), respectively. Although the model primarily reconstructs T and S profiles, it also predicts the uncertainty for density (σ_ρ^2) to weight the physical constraint term defined in the loss function ($L_T, L_S,$ and L_ρ). The predicted density profiles are computed from the predicted TS profiles based on the equation of state for seawater as seen in the manuscript.
- Computation: These values are not post-computed statistics but are learnable outputs directly generated by the network's variance head. The model predicts the logarithmic error variance ($\log \delta^2$) for numerical stability at each depth level.
- Normalization: We did not apply explicit normalization to the target variables before feeding them into the model. Instead, the scale differences between variables and variations across depths are handled by the uncertainty-aware loss function (Section 2.4). As seen in $L_T, L_S,$ and L_ρ, the squared error term is divided by the predicted error variance ($2\delta^2$). This mechanism acts as an adaptive, depth-dependent normalization, where the model learns to assign higher error variances to regions or variables with larger inherent scales or variability, effectively balancing gradients during training.

Following from previous point: in Section 2.4 the authors write: "TS-Cast network directly estimates temperature ($\hat{T}$) and salinity ($\hat{S}$) profiles, but also their associated depth-dependent uncertainty, represented by the logarithmic variance ($\log \delta^2$)". Are the authors speaking of the true variance or error variance? How is this practically computed? I suggest the terms "uncertainty" and "variance" are used much more carefully throughout the text as they are not synonyms. Again, please provide formulae to explain your operational definitions. In the present form, this section is very confusing.

Response: We agree with your comment. To resolve the confusion, we have revised the manuscript to consistently use the term "predicted error variance" and clarified its definition (Section 2.4).
The model outputs the predicted error variance (δ^2), which represents the estimated variance of the residual distribution at each depth, rather than the true natural variance of

the ocean state. It is a learnable parameter directly output by the network's variance head. It is optimized during the training via the negative log-likelihood loss.

By the way, if the authors are relying on variance and not error variance, the Loss function used forces the model to learn more from the less dynamically active depth levels, which seems quite a sub-optimal choice, requiring further justification. In fact, this seems in contrast to the objective of the model, which would likely be to "resolve" the most relevant dynamics.

Response: We agree that weighting by the natural variance of the ocean state would be suboptimal. However, we clarify that our loss function relies on the predicted error variance rather than the natural variance.
The model predicts this variance based on local climatological context, thereby learning the region's and season's inherent variability. Even in dynamically active regions where natural variance is high, this mechanism ensures that the loss is appropriately scaled by the typical variability of that water mass. This prevents the gradients from being dominated by regions with inherently large scales, allowing the model to effectively learn dynamics across all regimes, both active and stable, without biasing the learning process towards one or the other.

The physics-informed loss function accounting for the equation of state has been first introduced in Sammartino et al., https://doi.org/10.1016/j.envsoft.2025.106660, and that paper should thus be cited in this section.

Response: (Line 153) We have added the citation to Sammartino et al. (2025) in Section 2.4.

How did the authors compute coherence?

Response: We used the magnitude-squared coherence estimated via Welch's method.

- We used a Hanning window with a length of 128 days and a 50% overlap.

- The 95% confidence level was calculated based on the effective degrees of freedom determined by the windowing and overlap settings (Thompson and Emery, 2014).

(Lines 100–103) We have added a brief description of this calculation method to Section 2.2 (Validation data) or the relevant part of the Methods section.

The authors' critique of what they name 'closed-loop' validation may be overly broad (both in the introduction and in the concluding remarks). Validation against independent data from the same platform (e.g., held-out Argo floats) is a standard and robust practice. While time-series validation at mooring sites provides valuable complementary information on temporal dynamics, it has its own spatial limitations. I suggest reframing this as a complementary strength of the present work rather than a limitation of prior approaches.

Response: We agree with the reviewer's assessment. Accordingly, we have revised the relevant sections as suggested:

- Introduction (Lines 49–53): We removed the sentences regarding "closed-loop". Instead, we now present our timeseries validation as a necessary step to assess temporal coherence, which complements the spatial assessment provided by Argo profiles.
- Conclusion (Lines 294–297): We emphasized that combining standard spatial validation (Argo) with high-frequency temporal validation (Moorings) provides a more robust evaluation of ocean reconstruction models.

Minor points:

Response: We thank you for the comments, and all have been revised.

Line 1: "estimates of ocean surface states" -->"estimates of ocean surface state"

Line 3: "at sufficient space and time scale" --> "at sufficient space and time resolution"

Line 4: "While ADT represents" --> "While ADT contains"

Line 13 and following: "data-assimilated" --> "data-assimilating"

Line 18: remove "fundamental variables,"

Line 28: remove "primarily"

Line 39: change to "like multilinear regressions combined with optimal interpolation"

Line 44: change to “designed to capture sequential dependencies” (at least the first paper is not using LSTM over time, but over depth)

Line 46: add reference to Sammartino et al. (2025) - see above

Line 61 and following: considering the model domain, what is the advantage of transforming Lat-lon coordinates to 3 different variables?
Response: (Line 62–64) The primary advantage is to address the periodicity of the spherical domain and avoid numerical discontinuity.

L262: “explains”-->”might explain”

Line 284: “non-geostrophic”-->”non-baroclinic”

Reviewer 2

This study proposes a CNN-based U-Net architecture named TS-cast for reconstructing subsurface temperature and salinity data. By integrating remote sensing products with in-situ observations to fill data gaps, the research topic holds certain value. However, the current manuscript requires substantial improvements in details such as data preprocessing and vertical interpolation. Specific comments are as follows:

Major Comments:

1) Many studies have established relationships between surface remote sensing data products and subsurface temperature-salinity profiles, enabling subsurface temperature-salinity reconstruction. While the abstract states that ADT represents integrated information for subsurface water properties, it remains challenging to correlate SST, SSS, and ADT with subsurface water profiles due to their complex spatial and temporal variations. The research challenges are not sufficiently clarified. Authors are advised to further clarify the research questions, innovative aspects of the study, and the problems addressed.

Response: We have revised the abstract to explicitly distinguish our approach from existing methods. We clarified that, unlike conventional models, TS-Cast introduces a physics-informed architecture and uncertainty quantification.

Revised text (Lines 4–11): While ADT reflects total ocean dynamics, its steric component represents the integrated information for subsurface water properties. However, relating surface variables to subsurface profiles remains challenging because surface signatures are often non-linearly related to interior structures, and satellite data contain inherent non-steric signals. To address these limitations, we introduce TS-Cast, a novel uncertainty-aware deep neural network. Unlike direct regression models, TS-Cast is designed to adjust monthly climatological profiles as a physical prior and learns to dynamically adjust them. By using a 31-day sequence of satellite inputs (SST, SSS, and ADT) and quantifying prediction uncertainty, the model effectively captures the temporal variation of mesoscale dynamics. It was trained on approximately 150,000 Argo and ship-based thermohaline profiles in the northwestern Pacific.

2) The author's X, Y, Z coordinate definitions are incorrect: Since cosine and sine transformations of longitude and latitude require consideration of periodicity, their current

transformations are entirely erroneous and lack any geographic meaning. Consequently, the model inputs encoded geographic information incorrectly.

The correct transformation should be: $X = \sin\left(lat \times \frac{\pi}{180}\right)$, $Y = \sin\left(lon \times \frac{\pi}{180}\right) \times \cos\left(lat \times \frac{\pi}{180}\right)$, and $Z = -\cos\,(lon \times \frac{\pi}{180}) \times \cos\,(lat \times \frac{\pi}{180})$

Response: We have revised the equation in Section 2.1 to explicitly include the conversion factor (180/π) as suggested by the reviewer.

3) High-resolution data can be interpolated to lower resolutions, but is it correct to interpolate 1/4° low-resolution climatological data to 1/8° high resolution? The text states: “All in-situ and climatological profiles were linearly interpolated onto 128 evenly spaced vertical layers between 10 and 700 dbar.” This is a significant issue. Due to vertically discrete sampling, direct simple linear interpolation introduces substantial errors. On one hand, the interpolated results fail to match actual vertical variations—an inherent error of the method itself. On the other hand, if some profiles have only sparse observations (perhaps just one or two sampling points) between 10 and 700 dbar, interpolating these directly into 128 layers yields completely erroneous results.

Response: Regarding the horizontal regridding of the WOA23 climatology (1/4° to 1/8°), this was a necessary step to align the dimensions of the background climatological prior with the high-resolution satellite input data (1/8°) within the neural network architecture. We do not claim this process increases the information content of the climatology itself; rather, it allows the TS-Cast model to use the coarse climatology as a baseline and refine the spatial details using the high-resolution information provided by the satellite observations (SST, ADT).

We agree that applying linear interpolation to sparse data is problematic, and apologize for the textual error. The sentence “Profiles containing data gaps were retained by applying masks to use only valid-level observations for model training and error estimation” (Lines 72–73) should be corrected as “… for model test and error estimation”. However, we would like to clarify our data selection process, which was designed to prevent this issue.

We used the CORA 5.2-Easy CORA dataset provided by CMEMS (Lines 65–67). This product is a processed version of the global in situ dataset, with rigorous quality controls already applied. We also applied a quality-control criterion to the training and test datasets. To avoid the linear interpolation artifacts, we selected only profiles containing more than 50 valid observation points within the sampled depth range (typically 10-1500 dbar). Given this sampling density, linear interpolation was used as a vertical resampling step. Profiles with gaps were excluded from training. We have revised Section 2.1 to explicitly state this selection criterion.

Revised Text (Section 2.1, Lines 65–67): For the training dataset, thermohaline profiles from CTD and Argo floats were sourced from the Coriolis Ocean dataset for Reanalysis version 5.2 (CORA5) Easy CORA product provided by CMEMS (ID: INSITU_GLO_PHY_TS_DISCRETE_MY_013_001). This is a delayed-mode dataset with preprocessed and quality-controlled profiles from CTD, Argo floats, and other platforms.

Revised Text (Section 2.1, Lines 74–77): "Only profiles containing more than 50 valid observation points within the sampled depth range (typically 10-1500 dbar) were used. For these profiles, linear interpolation was applied as a vertical resampling step to align the data onto the 128 standard vertical levels. Profiles with gaps were excluded from training. For the test dataset, profiles with gaps were retained by applying masks to use only valid-level observations for error estimation."

4) Due to the significant vertical interpolation bias in this study, their training and test datasets underwent identical interpolation operations. Since both datasets share the same bias issue, their results are relatively close. However, by using this biased test dataset as a benchmark to compare against GLORYS, HYCOM, and ARMOR3D datasets, they inevitably present the erroneous conclusion that TS-Cast is more accurate. If this comparison method is applied, then all results presented in Fig. 5 and subsequent figures are invalid.

Response: We respectfully disagree with the claim that our comparative analysis is invalid. This concern appears to stem from a misunderstanding of how the datasets were used in different figures.

1. Figure 4 (Internal Evaluation): The "identically processed" test dataset (interpolated profiles) was used solely for Figure 4. Crucially, Figure 4 does not present any comparison with other models (GLORYS, HYCOM, or ARMOR3D). It only illustrates the internal error distribution of our model. Therefore, the claim that we used a biased dataset to benchmark against other models is factually incorrect.

2. Figure 5 and Subsequent Figures (Model Comparison): The comparisons with other models (Figure 5 onwards) use independent in-situ mooring observations (KEO, EC1, PIES), which are not vertically interpolated. These are time-series measurements at specific

depths or vertically integrated values. If our model learned "linear interpolation artifacts" as suggested, it would fail to capture the complex variations observed in these in-situ datasets. The fact that TS-Cast outperforms or matches other models on these independent observations confirms that it has learned realistic states rather than artifacts.

Minor comments:

Page 2, Lines 25–26: This creates a critical “observational void” between the broad coverage of satellites and in-situ measurements, hindering a complete understanding of the ocean's 3D dynamics. This sentence is incorrect and should be rewritten. The disparity between abundant surface data and sparse internal data reflects differences in observational techniques and data acquisition. Describing this as an “observational void” is inappropriate. Furthermore, the obstacle to understanding the ocean's three-dimensional dynamic processes stems from the sparsity of three-dimensional environmental data, not from the aforementioned differences.

Response (Line 28): We agree with your comment and have rewritten the sentence, “This sparsity of subsurface observations limits our ability to fully resolve the ocean's 3-D dynamics.”

Pages 2–3, Lines 54–60: The four summaries of the objectives of this paper contain some redundancies and should be further refined and enhanced.

Response (Lines 58–59): We have revised the objectives to eliminate redundancies, particularly by integrating the validation steps. The second and third objectives are integrated and rewritten as “Assess the model's capability to reproduce continuous ocean dynamics by validating against multi-year, high-frequency timeseries mooring data, using spectral coherence analysis to quantify performance across different frequency bands.”

Page 7, Lines 112–114: Here, linear interpolation is used to interpolate the profiles into 128 uniformly spaced vertical layers. This processing yields erroneous results, with significant interpolation errors causing the outcomes to deviate substantially from the true temperature-salinity distribution.

Response: By interpolating the 10–700 dbar range into 128 layers, we achieve a vertical spacing of approximately 5.4 dbar. This resolution is sufficiently fine to capture sharp vertical gradients, including the main thermocline and halocline, with negligible interpolation error. This preprocessing step was necessary to generate fixed-dimension input tensors for the U-Net architecture.