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”