Note: The original referee comments are in black, and the authors' responses are in blue.

Response to Reviewer #2：

The manuscript "Attention Enhanced 3D-U-Net++ Ocean Temperature and Salinity Reconstruction in the Northwestern Pacific Based on Transfer Learning" investigates the problem of subsurface ocean temperature and salinity reconstruction in the northwestern Pacific region (0–40°N, 120–160°E). The study proposes an attention-enhanced three-dimensional U-Net++–based framework to reconstruct daily three-dimensional temperature and salinity fields from surface satellite observations. The model produces temperature–salinity estimates on 26 vertical layers at 1/4° horizontal resolution over depths ranging from 5 to 2000 m, using sequences of sea surface temperature (SST) and sea surface height (SSH) as inputs.

The proposed approach combines cross-scale feature aggregation with attention-based gating mechanisms intended to emphasize surface features most relevant for subsurface variability. By integrating 26 consecutive days of SST and SSH, the method aims to alleviate the inherently underdetermined nature of mapping limited surface observations to full-depth ocean structures. In addition, the authors employ a transfer-learning strategy in which the model is pretrained using monthly SST and SSH data and subsequently fine-tuned for daily reconstruction. Model performance is evaluated against in situ temperature and salinity profiles from the World Ocean Database, and the reported results indicate generally good agreement with observations and modest improvements relative to baseline datasets.

While the overall results appear reasonable and the proposed ideas are physically well motivated, the manuscript suffers from a lack of methodological clarity that makes it difficult to fully assess, reproduce, and interpret the approach. In particular, the network architecture is not specified in sufficient detail: the manuscript does not clearly define the input and output tensor dimensions, the dimensionality of the convolutional operations, or what precisely constitutes the "three-dimensional" aspect of the U-Net++ architecture in practice. It remains unclear how spatial, temporal, and vertical

dimensions are represented within the network, and whether time and depth are treated as explicit dimensions or implicitly as stacked channels.

Similarly, the practical implementation of the transfer-learning strategy is valuable but insufficiently described. The pretraining and fine-tuning stages involve substantial changes in spatial resolution (1° to 0.25°), temporal resolution (monthly to daily), and target datasets (IPRC-Argo to GLORYS2V4), yet the manuscript does not clearly explain how these transitions are handled in practice. Key details regarding input normalization, adaptation of temporal windows, and the aspects of the learned representations expected to transfer between stages are either missing or only briefly mentioned in the results section. As a result, the reader is left with an incomplete understanding of how the proposed training strategy operates beyond a high-level conceptual description.

Finally, although the Results section is extensive and includes a wide range of diagnostic figures, it often reiterates similar validation setups and comparisons across multiple subsections. This repetition tends to obscure the main findings rather than sharpen them, and clearer structuring or consolidation of overlapping analyses would improve readability and focus.

Overall, the study addresses an important problem and presents promising results, but substantial improvements in methodological transparency and presentation are required before the contribution can be fully evaluated and appreciated.

Response:

We sincerely thank the reviewer for the careful reading of our manuscript and for the constructive and detailed comments. We appreciate the reviewer's positive assessment that our study addresses an important problem in subsurface ocean temperature and salinity reconstruction, and that the proposed attention-enhanced 3D U-Net++ framework with transfer learning is physically well motivated and yields generally reasonable results with modest improvements over baseline products.

We have carefully studied these comments and have made substantial revisions to the

manuscript to address the concerns raised. The parts that have been revised in the revised manuscript have been marked in yellow. Below, we respond to each comment point by point and indicate the corresponding changes made in the manuscript.

Comment 1:

The manuscript refers throughout to an "attention-enhanced 3D U-Net++" architecture, yet the network design is not described with sufficient precision to assess what makes it three-dimensional in an architectural sense. While Section 2.2.1 provides a general description of U-Net++ and the integration of CBAM attention modules, it remains unclear how this design is extended beyond a conventional two-dimensional framework. In particular, the manuscript does not clearly state whether the third dimension corresponds to depth or time, nor whether either is treated as an explicit spatial dimension within the network or implicitly as stacked input or output channels. It is also unclear whether a single network input consists of full regional SST–SSH maps or of point-wise surface time series. The absence of explicit input and output tensor definitions, together with a clear description of the dimensionality of the convolutional, pooling, and upsampling operations, makes it difficult to evaluate reproducibility, interpretability, and the validity of the architectural claims.

Response:

We sincerely appreciate your comments regarding the accuracy of the network description, the correspondence of data dimensions, and the input/output formats. We realized that the original description was ambiguous regarding the dimensionality. In response to your suggestions, we have redesigned Fig. 1 to more accurately illustrate the neural network structure and the flow of data shapes within the network.

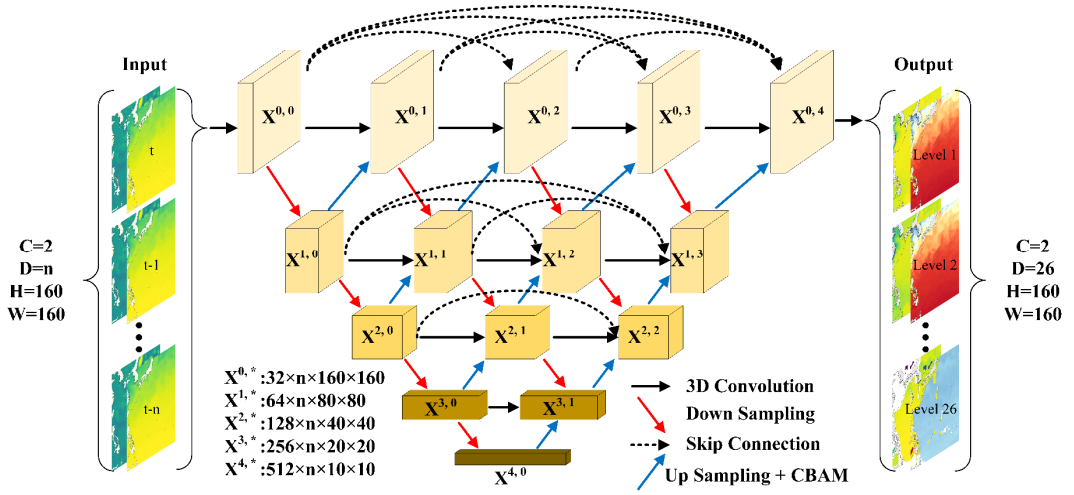

Figure 1. Schematic diagram of the 3D-U-Net++ neural network structure, where $X^{i,j}$ represents the feature map at level i and stage j.

Furthermore, we have included additional details regarding the 3D U-Net++ architecture. Specifically, please refer to **the 3th paragraph of Section 2.2.1** in the revised manuscript:

'**The 3D-U-Net++ network employed in this study is derived from the conventional U-Net++ architecture by replacing all 2D operations—such as convolution, pooling, and up-sampling—with their 3D counterparts. Notably, 3D convolution is capable of simultaneously processing data across four dimensions: depth, channel, height, and width (Tran et al., 2015). This capability renders the network highly suitable for physical oceanographic datasets, which typically encompass spatial extent, depth, and temporal information.**'

and **the 5th paragraph of Section 2.2.1** in the revised manuscript:

"**As illustrated in Fig. 1, the neural network is designed to directly accept a 4D tensor with a shape of (C×D×H×W) as input. In this study, C is set to 2, representing the two input channels: SSH and SST. D is set to 26, denoting the continuous time series of sea surface data from the past 26 days (details regarding D are provided in Section 2.2.3). H and W are both set to 160, corresponding to the spatial dimensions of the study area. The output dimensions of the network are (2, 26, 160, 160). Specifically, the first dimension 2 represents the two target**

variables: seawater temperature and salinity; 26 corresponds to 26 depth levels spanning from 0 to 2000 m; and 160 represents the spatial dimensions of the study area. Within the network architecture, down-sampling is performed using 3D max pooling, while up-sampling is achieved via 3D transposed convolution."

Comment 2:

Related to this, the treatment of temporal information remains insufficiently specified. Section 2.2.3 provides a conceptual motivation for incorporating multi-day SST and SSH inputs to alleviate the underdetermined nature of subsurface reconstruction; however, it remains unclear how temporal information is handled within the network in practice. In particular, the manuscript does not clarify whether time is modeled explicitly (e.g., via temporal convolutions or sequence-aware components) or simply treated as an expansion of the input feature space. The choice to use a 26-day surface input window is reasonable and physically plausible, but the justification for this specific value is limited to its correspondence with the number of reconstructed depth levels. A brief explanation of whether this choice is motivated by physical timescales, empirical tuning, or practical considerations would improve clarity and help readers assess the generality of the approach.

Response:

We agree with the Reviewer that the description of the temporal information processing and the justification for the input window size needed further clarification.

Regarding the handling of time series, relevant explanations have been provided in the response to Comment 1.

Regarding the rationale for the time window size, we have redesigned the experiments and conducted a detailed analysis based on the results. This is presented in **Section 3.2** of the revised manuscript:

'**To determine the optimal length of the input time series for the network, an**

ablation study was conducted. Considering computational costs, time series lengths of 1, 4, 8, 10, 20, 26, 30, and 40 days were selected. The experimental results are illustrated in Fig. 6, which demonstrate a clear negative correlation between the input sequence length and the reconstruction error. When the time step increases from 1 to 20, the RMSE for both variables decreases significantly—Temperature RMSE drops from 0.63563ºC to 0.61222ºC, and Salinity RMSE from 0.09874 PSU to 0.09475 PSU. This indicates that the incorporation of historical surface data helps to mitigate the ambiguity associated with super-resolution tasks. Furthermore, continuous time-series of sea surface data provide robust physical constraints for the reconstruction of underwater 3D T-S fields, thereby enhancing the accuracy of the reconstructed data.

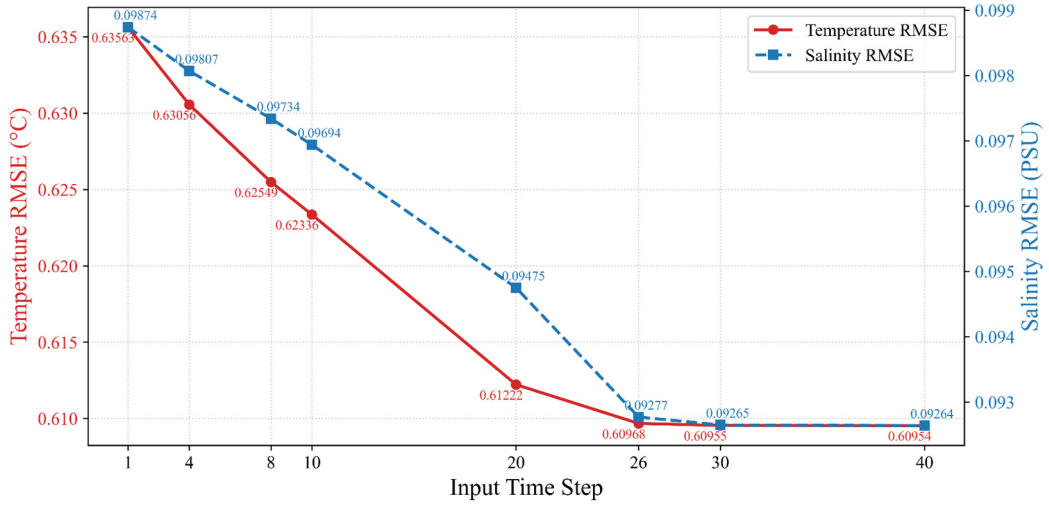

Figure 6. RMSE between the reconstructed data and WOD profiles under different input time series lengths in 2023.

Notably, the performance gain begins to saturate beyond an input length of 26 days. Between time steps 30 and 40, the RMSE curves for both temperature and salinity plateau (stabilizing around 0.6095ºC and 0.0926 PSU, respectively). To balance computational cost and accuracy, the time series length of the input data was set to 26 in this study.'

Comment 3:

The implementation and interpretation of the transfer-learning strategy are beneficial

for the manuscript but also require clarification. Although the manuscript evaluates different layer-freezing strategies and ultimately adopts full fine-tuning, the practical mechanics of transferring between pretraining and fine-tuning remain unclear. In particular, the transition from monthly, coarse-resolution SST/SSH and IPRC-Argo targets to daily, higher-resolution SST/SSH and GLORYS2V4 targets involves substantial changes in temporal resolution, spatial resolution, and target data characteristics. The manuscript would benefit from explicitly describing how input structures, normalization, and temporal windows differ between stages, and what aspects of the learned representation are expected to transfer. Since the final model is fully fine-tuned on GLORYS2V4, a clearer discussion is also needed on whether the resulting product should be interpreted as approximating observational variability or as producing GLORYS-consistent reconstructions with improved alignment to WOD profiles.

Response:

We sincerely thank you for your valuable comments and suggestions regarding the spatiotemporal resolution mismatch and the mechanisms of transfer learning. In response, we have redesigned Table 1 and relevant descriptions to the revised manuscript, which provide a detailed explanation of the data processing and normalization methods applied during the two stages of transfer learning. Specifically, we have added the following description in **Section 2.1.1** of the revised manuscript:

‘**It is worth noting that due to differences in the temporal and spatial resolutions of the label data used in the two stages of transfer learning, the datasets were unified for this study. The specific data processing methods are detailed in Table 1.**

Table 1. Data processing and normalization methods.

Training Stage	Data	Data Processing Method	Normalization Method
Pre-	AVISO SSH	Downsampling to	Standardization

training		**0.25° Monthly averaging**	**followed by Min-Max**
	OISST	**Monthly averaging**	**normalization**
	RPRC Argo	**Linear interpolation to 0.25°**	**Min-Max normalization only**
Fine-tuning	**AVISO SSH**	**Downsampling to 0.25°**	**Standardization followed by Min-Max normalization**
	OISST	**None**	
	GLORYS2V4	**None**	

In the pre-training phase, the input data were averaged monthly to match the temporal resolution of the IPRC Argo temperature and salinity data. Additionally, to be consistent with the OISST data and match the output dimensions of the neural network, both the 1/8º resolution AVISO SSH data and the IPRC Argo data were linearly interpolated to 0.25º. In the fine-tuning phase, since both the training data and label data are daily records, no temporal averaging was performed; however, the 1/8º AVISO SSH data were down-sampled to 0.25º.

Furthermore, the data normalization strategies differed between the two stages. As the pre-training data predominantly contain monthly signals, the label data were simply normalized to the range of [0, 1]. In contrast, for the fine-tuning phase—where the objective was for the network to learn smaller time-scale signals (i.e., abrupt changes) based on the pre-trained model—the label data were first standardized using the mean and standard deviation, and then normalized to [0, 1]. To ensure consistency, the input data for both stages were processed using the same two-step approach: standardization followed by [0, 1] normalization.'

Additionally, we have revised **the 4th paragraph of Section 3.1** to provide a more detailed analysis and a clearer discussion regarding the mechanisms of transfer learning: "**The experimental results demonstrate that the optimal transfer learning strategy**

is global fine-tuning (0% frozen weights), rather than acting as a rigid feature extractor (which would favor partial freezing). This indicates that the fundamental mechanism of transfer learning in this study is providing a physically constrained parameter initialization rather than directly reusable features. Oceanographically, the transfer process mimics a "background-to-perturbation" learning paradigm. During the pre-training phase with monthly, coarse-resolution IPRC Argo data, the network learns the large-scale climatological background, encompassing basic stratification, seasonal cycles, and global vertical covariance structures. By establishing this robust physical framework, the pre-trained weights place the model in a physically plausible region within the high-dimensional optimization landscape. During the fine-tuning phase with daily, high-resolution GLORYS data, global fine-tuning allows the network to bypass the struggle of learning fundamental ocean physics from scratch. Instead, it fully allocates its learning capacity to resolving high-frequency, synoptic-scale dynamics—such as how mesoscale eddies and fronts perturb the pre-established climatological background.

In summary, since the neural network is trained exclusively on GLORYS2V4 data, the error between the reconstructed data and the WOD observational profiles can only asymptotically approach the error between the label data and the WOD profiles. Initializing the network weights using the IPRC Argo dataset allows the model to capture authentic observational information while establishing a background of the ocean environment at a monthly scale. Building upon this foundation, the fine-tuning phase enables the network to learn the complex dynamic mapping rules and smaller time-scale signals inherent in the GLORYS2V4 data. Validation against WOD profile data demonstrates that this transfer learning strategy maintains high physical consistency with GLORYS2V4 while achieving closer agreement with the WOD observations."

Comment 4:

More generally, the training and evaluation strategy relies heavily on model-assisted and empirically reconstructed datasets. This is a reasonable and widely used approach for large-scale subsurface reconstruction; however, its implications for the interpretation of the resulting data product are not discussed in sufficient detail. In the fine-tuning stage, GLORYS2V4 reanalysis fields are used as training labels, and the reconstructed outputs are subsequently shown to closely resemble GLORYS in both spatial structure and error characteristics. While validation against independent WOD profiles is included, additional comparisons are largely performed against other reconstructed or fusion-based products (e.g., HGEM-derived regional datasets and CGOF1.0). Without a more explicit discussion of these dataset dependencies, it remains unclear to what extent the proposed method reconstructs independent oceanic variability versus reproducing the statistical structure of specific reanalysis products. Clarifying this distinction, rather than introducing additional comparative datasets, would help users interpret the dataset appropriately and assess its generalizability to other regions or observational systems.

Response:

We sincerely appreciate this insightful comment. We fully agree that distinguishing between the "learned statistical structure" from the reanalysis data and the "independent oceanic variability" driven by satellite inputs is crucial for the correct interpretation of our data product.

In this study, the proposed neural network employs a transfer learning strategy. During the pre-training phase, SST and SSH serve as inputs, with IPRC Argo temperature and salinity data used as ground truth labels. In the fine-tuning phase, SST and SSH remain as inputs, while GLORYS2V4 temperature and salinity data are utilized as the target labels.

Upon completion of the training, we first analyzed the discrepancy between the reconstructed data (from the validation set) and the GLORYS2V4 dataset. Subsequently,

the World Ocean Database (WOD) was introduced as actual in situ observations to evaluate both the reconstructed data and the GLORYS2V4 data. The results demonstrate that while the reconstructed data maintains high consistency with GLORYS2V4, it exhibits a closer alignment with the actual WOD observations.

Furthermore, to validate the accuracy of the reconstructed data from the perspective of data products, we conducted a horizontal comparison across multiple datasets (GLORYS2V4, HGEM, and CGOF) over a long-term scale (1993–2023), again using WOD observations as the benchmark. The results indicate that the temperature and salinity fields reconstructed by our proposed neural network are closer to the real observations than the other products.

In summary, the primary evaluation metric throughout this study consistently relies on real in situ profile data from WOD. Given that the reconstructed data demonstrates a closer proximity to WOD, we conclude that the reconstructed data can, to a significant extent, accurately reflect real oceanographic phenomena.

We have added a new paragraph in **Section 3.3.2 (the last paragraph)** to explicitly discuss these dependencies. The added text is as follows:

‘**Since the 3D-U-Net++ model learns the mapping from surface (SST, SSH) to subsurface layers based on the statistical relationships embedded in GLORYS2V4, the reconstructed outputs inevitably inherit the structural characteristics of this specific reanalysis product. However, this does not imply that the model merely reproduces the reanalysis climatology. As shown in Figs. 12-13, The high consistency between the reconstruction and the independent WOD profiles demonstrates that the model successfully generalizes the learned relationships to actual oceanic conditions, rather than simply overfitting to the reanalysis statistics. This confirms that the method reconstructs independent oceanic variability driven by surface inputs, while utilizing the reanalysis data to constrain the vertical thermohaline structure. Therefore, the reconstructed product should be**

interpreted as a combination of real-time satellite-observed surface variability and the dynamical vertical structure learned from GLORYS2V4.'

Comment 5:

The Results section is extensive but often repetitive in its presentation of the validation setup. In particular, multiple subsections repeatedly restate the same input data, reference datasets, and evaluation procedures. While this information is important, reiterating it throughout the Results disrupts the narrative flow and obscures the main findings. A clearer separation between a concise description of the evaluation framework (stated once) and the subsequent presentation of results would improve readability and focus without requiring additional analyses.

Response:

We thank the reviewer for pointing out the redundancy in the Results section. We realize that reiterating the validation setup in each subsection disrupted the flow and distracted from the main findings.

In the revised manuscript, we have restructured Section 3 to address this issue:

Removal of Redundancy: We have removed the detailed descriptions of the validation datasets (GLORYS2V4 and WOD), the test period (year 2023), and the evaluation metrics (RMSE, etc.) from Sections 3.1, 3.2, 3.3, and subsequent subsections, , as this information is already specified in the figure captions.

Focus on Results: The subsections now focus directly on the analysis of the results (e.g., performance comparison, error distribution, and vertical profiles) without restating the experimental setup.

These changes have significantly streamlined the text and improved the readability of the Results section.

Comment 6:

The manuscript also makes relatively strong claims regarding the capture of underlying

physical laws and the ability to deliver real-time three-dimensional reconstructions. These claims are not fully supported by the methodological description. The surface inputs are derived from mapped and interpolated satellite products, and the subsurface targets are drawn from reanalysis and empirically reconstructed datasets. In this context, it would be more appropriate to frame the method as producing reanalysis-consistent subsurface fields conditioned on surface information. In addition, the concept of "real-time" reconstruction is not clearly defined in terms of data latency, update cycles, or computational requirements.

Response:

We sincerely thank the reviewer for this insightful comment. We agree that the term 'capturing underlying physics' might not be precisely phrased appropriately, given that our model is trained based on reanalysis products (GLORYS2V4) and gridded satellite observations. Fundamentally, the neural network learns the complex non-linear mappings between surface and subsurface variables within the reanalysis framework, rather than deriving physical laws directly from raw observations.

Following your suggestion, we have revised the manuscript to more accurately describe the model output as 'subsurface temperature and salinity fields constrained by surface information, which are consistent with reanalysis data and demonstrate improved alignment with WOD profiles.'

Regarding the 'real-time' capability emphasized in the manuscript, this is attributed to the fact that the proposed model can achieve high-precision reconstruction of subsurface temperature and salinity using solely SST and SSH data. Since both OISST and AVISO SSH satellite products are available in real-time, the trained model can download the current day's data to perform immediate reconstruction.

We have added a corresponding explanation in the revised manuscript line 90:

'**However, among current satellite datasets, only SST and SSH are available in**

real-time, whereas SSW and SSS data exhibit latencies of approximately 1–3 days and 3–7 days, respectively. Consequently, to realize real-time reconstruction of subsurface temperature and salinity using neural networks, the input variables are restricted to the real-time accessible SST and SSH. To achieve real-time large-depth reconstruction of subsurface temperature and salinity, this study proposes a method that relies solely on real-time available SST and SSH data.'

With respect to the update cycle and computational requirements, we have also provided a clearer elaboration **at the beginning of Section 3** in the revised manuscript: '**Model training and inference were conducted on a supercomputing cluster equipped with an Intel® Xeon® Gold 5218R CPU and an NVIDIA A100-PCIE-40GB GPU. The average inference time required to generate daily 3D T-S fields for the entire region is approximately 5 seconds. In contrast to traditional reanalysis products (e.g., GLORYS and EN4) which typically suffer from latencies ranging from weeks to months due to the assimilation of sparse in-situ data (see Table 2), our method relies exclusively on real-time satellite observations, enabling real-time reconstruction using data from the current day.**

Table 2. Comparison of update cycles and approximate time lag (latency) between the proposed method and mainstream products.

Dataset / Product	Type	Approximate Update Cycle / Latency	Dependence on In-situ Data
Proposed Method	Deep Learning Reconstruction	Daily / No lag	No (Only relies on NRT Satellite SST & SSH)
GLORYS12V1 (CMEMS)	Reanalysis	Monthly / ~2-3 years lag (Delayed Mode)	Yes (High dependence)
PSY4V3R1	Operational	Weekly / ~7 days	Yes

(CMEMS Analysis)	Analysis	lag (Best estimate)	
EN4 (Met Office)	Objective Analysis	Monthly / ~1-2 months lag	Yes
RG-Argo (Scripps)	Gridded Argo Product	Monthly / ~1-2 months lag	Yes
SODA3	Reanalysis	Monthly / Several months lag	Yes

Comment 7:

Finally, the Discussion and Conclusion section largely reiterates the methodological design and reported performance improvements, but offers limited reflection on limitations, uncertainties, and appropriate use cases. Given the dependence on specific training products and the use of spatially sparse in situ profiles for validation, a brief discussion of known constraints (e.g., sampling density, depth-dependent performance, and dataset dependence) would improve transparency and help users interpret and apply the dataset appropriately.

Response:

We sincerely thank the reviewer for this constructive suggestion. We agree that a transparent discussion of the model's limitations, uncertainties, and appropriate use cases is crucial for users to interpret the dataset correctly.

In **the 2nd paragraph of Section 5** of the revised manuscript, we have added the following description of limitations:

'**Despite the promising results, several limitations and uncertainties must be acknowledged. First, the computational cost of training is substantial due to the employment of operations such as 3D convolutions. Second, despite validation against in situ observations, the spatial sparsity of these profiles implies inherent**

uncertainty regarding the model's reliability in unsampled regions.'

Minor Comments:

Comment 1:

The references cited in the Introduction are generally relevant to the topic. However, in several places the way individual studies are positioned in the narrative does not fully reflect the specific emphasis of those works. In some instances, the surrounding text highlights particular methodological aspects, while the cited studies focus on different elements of the approach. This can make it harder for readers to clearly understand how individual contributions relate to the conceptual structure presented. A clearer alignment between the narrative and the cited literature would improve clarity.

Response:

We sincerely thank the reviewer for this insightful comment regarding the literature review in the Introduction. We agree that a precise alignment between the narrative and the specific contributions of cited studies is crucial for outlining the research context clearly.

We have carefully re-examined the Introduction section and verified the specific emphasis of each cited reference. Based on your suggestion, we have revised the text to ensure that the descriptions accurately reflect the methodological focus and unique contributions of the cited works.

Comment 2:

Several in-text citations appear in non-chronological order (e.g., Xie et al., 2025; Wu et al., 2012). Reordering references chronologically within sentences would improve clarity and consistency.

Response:

Thank you for your valuable suggestion regarding the chronological ordering of the references. After making the modifications, the original text reads as follows:

'With the rapid development of deep learning, neural networks have demonstrated great potential in the reconstruction of oceanic temperature and salinity fields due to their powerful nonlinear fitting capabilities (Wu et al., 2012; Xie et al., 2025).'

In addition, we have carefully checked the entire manuscript and corrected similar issues elsewhere (**Line 33, Line 71 in the revised manuscript**).

Comment 3:

In line 93 "A transfer learning strategy" needs a reference.

Response:

We appreciate your suggestion to include the missing reference. We have added the relevant citation at the appropriate location in the revised manuscript.

After making the modifications, the original text reads as follows:

'**A transfer learning strategy (Pan and Yang, 2010) is employed:**'

Comment 4

WOD is referenced before being defined (e.g., "WOD in-situ T–S profiles" in line 104); the acronym should be introduced at first use.

Response:

Thank you for pointing out that the abbreviation was used before being defined. We have modified Line 104 to provide the full name upon its first appearance. Accordingly, we have removed the redundant explanation of the term "WOD" that appeared later in the text.

In addition, we have carefully checked the entire manuscript and corrected similar issues elsewhere (**Line 75, Line 85 in the revised manuscript**).

Comment 5

Figure 3 appears to be cropped on the right-hand side; please verify whether this is intentional.

Response:

We are grateful to you for pointing out that Figure 3 was cropped. This issue arose because the image width exceeded the page margins. We have resized the figure in the revised manuscript (Now it is Figure 4) to ensure it is displayed completely and clearly.

Comment 6

In Section 3.1 (Transfer Learning), it is unclear whether the reported results correspond to the single-day or multi-day input configuration. The specific setup used for the presented results should be stated explicitly.

Response:

Thank you for your comment regarding the lack of clarity in the input configuration for transfer learning. We would like to clarify that we utilized the control variable method across all experiments. Specifically, multi-month inputs were used for the pre-training stage, and multi-day inputs were used for the fine-tuning stage.

We have added the following note at the end of **the 1st paragraph in Section 3.1**:

‘**In the comparative experiments, inputs spanning multiple months and multiple days were employed for the pre-training and fine-tuning stages, respectively. The pre-training phase utilizes data from the preceding *a* consecutive months, whereas the fine-tuning phase employs data from the preceding *a* consecutive days (refer to Section 3.2 for the specific value of *a*).**’

Comment 7

In Figures 5 and 6, monthly correlation profiles are shown using very similar color

maps, making it difficult to distinguish individual months (e.g., December versus February). Using more distinct colors or line styles would improve readability.

Response:

We appreciate your suggestion regarding the color similarity in the figure legends, which made them difficult to distinguish. We have updated the corresponding figures with a distinct color scheme to ensure better readability and distinguishability (Now they are Figure 7 and Figure 8).

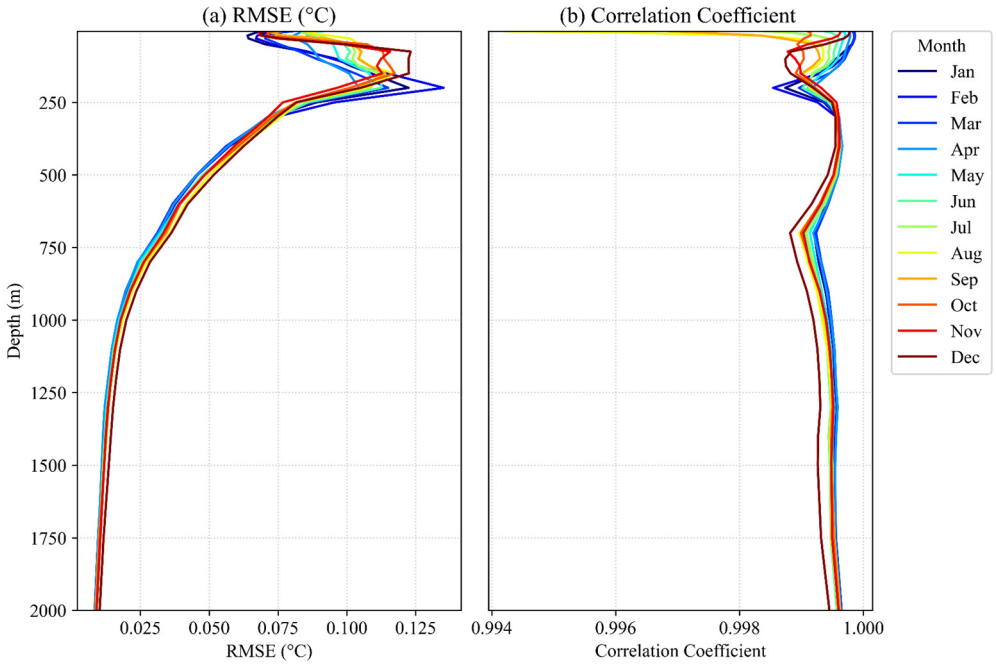

Figure 7. (a) RMSE and (b) correlation coefficient between the model-reconstructed temperature fields and the GLORYS2V4 temperature data in 2023.

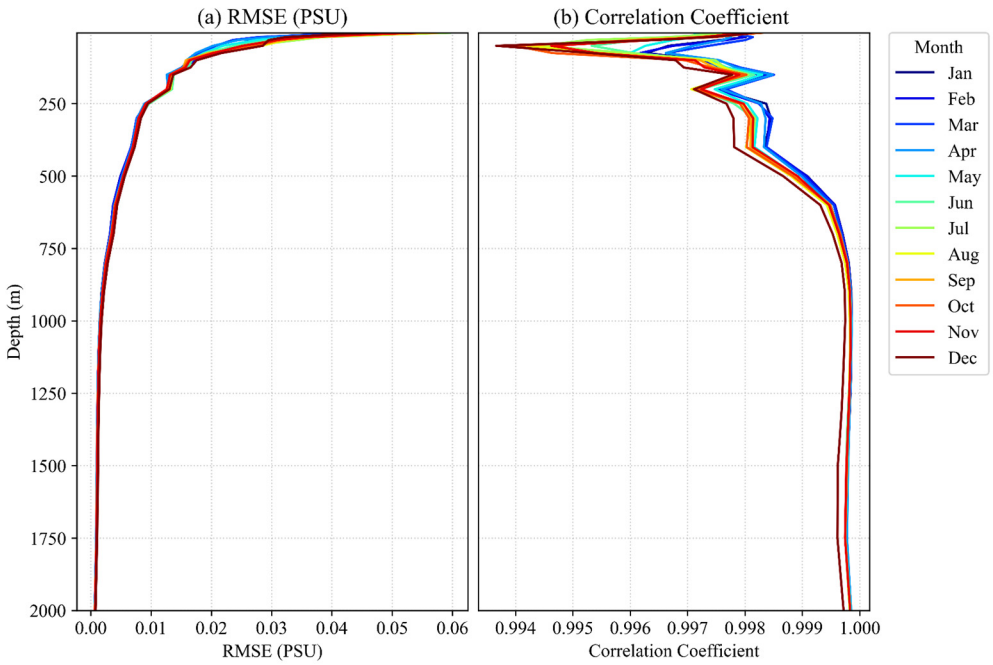

Figure 8. (a) RMSE and (b) correlation coefficient between the model-reconstructed salinity fields and the GLORYS2V4 salinity data in 2023.

Comment 8

Figures 7 and 8 effectively illustrate spatial reconstruction examples at selected depths for a single day. However, the analysis would benefit from accompanying spatial RMSE maps aggregated over the full test period at the same depths, to provide a more representative assessment of performance.

Response:

We sincerely thank the reviewer for this constructive suggestion. We completely agree that while single-day examples demonstrate the model's instantaneous capability, a temporally aggregated assessment is crucial for demonstrating the overall reliability and spatial stability of the reconstructed fields.

Per the reviewer's suggestion, we have calculated the spatial RMSE for both temperature and salinity aggregated over the entire test period at the corresponding depths.

These new spatial RMSE maps have been added to the revised manuscript as Figure 11:

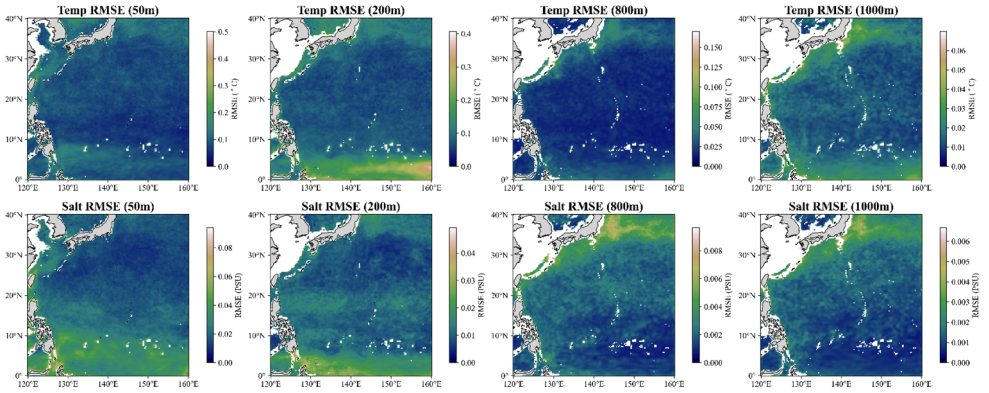

Figure 11. Spatial distribution of the RMSE between the reconstructed data and the GLORYS2V4 data at depths of 50 m, 200 m, 800 m, and 1000 m in 2023.

Figure 11 illustrates the spatial distribution of the RMSE at different depths throughout the validation period. It can be observed that at depths of 50 m and 200 m, the RMSE of the validation set is primarily concentrated in the North Equatorial Countercurrent (NECC) region. In contrast, at 800 m and 1000 m, the RMSE is mainly concentrated in the Kuroshio Extension region. Overall, however, the RMSE across all four depth levels remains within a relatively low range throughout the entire study area.

Comment 9

For the spatial RMSE maps in Figure 10 derived from WOD profiles, it would be helpful to include a map of WOD profile density per grid cell (e.g., in the Appendix or Supplementary Material) to clarify observational support. In addition, visually distinguishing grid cells with missing or insufficient data from genuinely low-RMSE regions (e.g., via masking or a distinct color) would reduce ambiguity.

Response:

We greatly appreciate this constructive suggestion. We agree that verifying the observational support is essential to properly interpret the spatial error distribution and avoid confusing unobserved regions with low-error regions.

To address this, we have implemented the following changes:

Added WOD Profile Density Map: We have calculated the number of WOD profiles falling into each grid cell across the study region. This density map is now included as Figure S1 in the Supplementary Material. It explicitly shows the spatial coverage and observational density of the ground truth data.

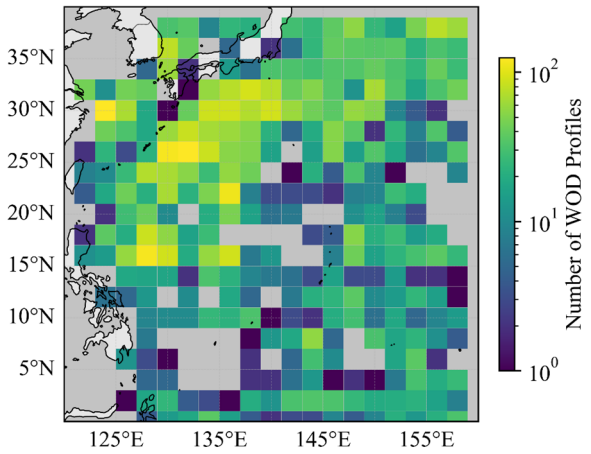

Figure S1. WOD profile density map in 2023, Each grid cell is 2°x2°.

Revised Figure 13: We have updated the spatial RMSE maps. Grid cells with insufficient observational data are now masked with a distinct gray color, while valid data regions use the original color scale to represent RMSE values. This visual distinction ensures that readers can clearly differentiate between areas of high model accuracy (low RMSE) and areas with no observational data.

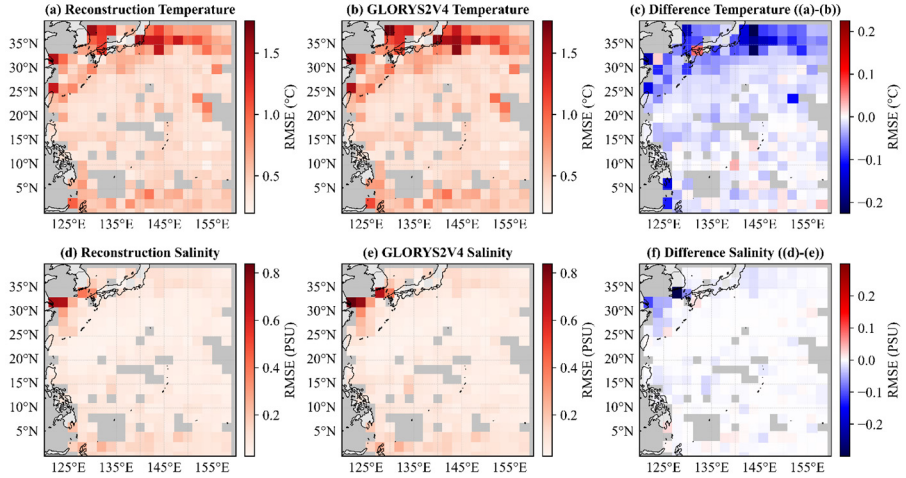

Figure 13. RMSE between reconstruction and WOD profiles (left), RMSE between GLORYS2V4 and WOD profiles (middle), and their difference (right). The first

row shows the results for temperature data, and the second row shows the results for salinity data.

Comment 10

Density scatter plots are presented both for a single validation year (Figures 11–12) and again for a longer period (1993–2023; Figures 16–17) using similar diagnostics. While both analyses are informative, their purposes partially overlap. Clarifying the distinct intent of each or streamlining one of them (e.g., moving it to supplementary material) would improve focus.

Response:

Thank you for pointing out the redundancy regarding the scatter density plots presented in Sections 3.3.2 and 3.4. In the revised manuscript, we have moved the relevant figures from Section 3.3.2 (original Figures 11 and 12) to supplementary material (Figures S2 and S3), and have retained only the scatter density plots in Section 3.4.