Review: Deep learning-based chlorophyll prediction: comparison with a dynamic model and applications to fish catch forecasting

The manuscript shows substantial improvement. The introduction is well structured and clearly outlines the research topic. The methods section presents the essential details in a more coherent manner and provides a clearer description of the architectural framework. Similarly, the results section is now more clearly presented, with accurate description of the conducted experiments and more detailed analysis of the presented results. Despite these advances, a few issues remain that still require attention.

We sincerely appreciate the reviewer's encouraging comments on the revised manuscript and the additional points raised. We have addressed each of the remaining points in the revised version, with detailed point-by-point replies provided below.

METHODS:

● Section 2.1 outlines the architecture design and provides several key details necessary for reproducibility. Table B1, referenced in the review, offers a clear and concise summary of these architectural specifications; however, it is not included in the manuscript, and no supplementary material section is currently available. Its inclusion within the manuscript is therefore recommended. Although this information is presented in the results section, the description of the model outputs would benefit from this clarification. In particular, it remains unclear whether the output consists of a single value or a time series. Providing this information more explicitly in the methods section would improve the reader's understanding of the architecture design.

We have now added the architectural hyperparameter table (Table S1) and the dataset summary table (Table S2) to the Supplementary Information of the revised manuscript. We also included cross-references in Sections 2.1 and 2.2 to direct readers to these supplementary tables. Additionally, we updated Section 2.1 to explicitly state that the model output is a single scalar value representing the LME-averaged chlorophyll anomaly, rather than a spatial map or a time series. The modified sentences are as follows:

"…. The complete hyperparameter specifications are provided in the Supplementary Information (Table S1).

The model predicts area-averaged chlorophyll anomalies for individual LMEs from global spatial fields. Input data consist of three consecutive monthly global maps of SST and chlorophyll anomalies, gridded at 1° × 1° resolution (360 longitude × 180 latitude) and represented as six input channels. The model output is a single scalar value representing the LME-averaged chlorophyll anomaly at the target lead time."

"This section describes the data sources used for training, validation, and testing, with sample sizes and temporal coverage detailed in the Supplementary Information (Table S2)."

● In the Introduction, the authors state that "*the network produces monthly or annual chlorophyll forecasts at the LME scale with lead times of 1–24 months*" (line 64). However, in the Methods section it is reported that "*each model directly predicts a 3-month mean chlorophyll anomaly centered on the targeted month from a three-month window of preceding input variables*" (line 93). Clarification of the relationship between these two descriptions would improve the overall clarity and facilitate a better understanding of the project purpose.

We agree that the original phrasing, "3-month mean chlorophyll anomaly centered on the targeted month," was ambiguous. We have corrected this to clarify that the model predicts the chlorophyll anomaly for a single target month using a three-month input window. To ensure full clarity, we have added a formal definition of "lead time" with an illustrative example in the revised Section 2.1. This definition and the corrected target description have been applied consistently throughout the manuscript and in the caption of Figure 4. The relevant passage in Section 2.1 reads:

"For monthly forecasts, CNN models cover all combinations of forecast start months and lead times (1–24 months ahead), each predicting the chlorophyll anomaly for a specific target month. Lead time is defined as the number of months between the final month of the input window and the target month. For example, using inputs from October, November, and December (OND) to predict the January chlorophyll anomaly corresponds to a 1-month lead time."

● Section 2.2 describes the datasets employed, including the training, validation, and test sets, and provides details on input data preprocessing. This information is also partially

introduced in Section 2.1, in paragraph 2.1.2, which presents the training, validation, and test sets in a particularly clear manner. It may therefore be worth considering its relocation to Section 2.2, if feasible, to improve the overall organization of the manuscript. In addition, Table B2 appears to offer valuable support for clarity and readability, but it is not currently included in the manuscript; its inclusion is recommended. Furthermore, the order in which the datasets are presented in Section 2.1.2 differs from that used in Section 2.2. Adopting a consistent order throughout the manuscript could enhance readability and help the reader in better understanding the role and usage of each dataset.

Following the reviewer's comment, we have relocated the descriptions of the training, validation, and test data splits from the previous Section 2.1.2 to Section 2.2. As a result of this relocation, the previous Section 2.1.1 has been promoted to Section 2.1 (Deep learning model and forecast experiment design), and Section 2.2 has been restructured to consolidate all dataset-related information into a single section to enhance readability. Furthermore, we have adopted a consistent ordering of datasets throughout Section 2.2 and Table S2, following the sequence of our modeling workflow: starting with CMIP6 simulations for training, followed by GFDL-ECDA reanalysis for validation and sensitivity testing, and finally satellite observations for evaluation. The dataset summary table (previously Table B2) is now included as Table S2 in the Supplementary Information, as noted in our earlier response.

● SHAP analysis (Section 2.3) quantifies the contribution of each spatial location's input value to the predicted LME-mean chlorophyll anomaly. This is achieved by comparing the model's predictions with and without the grid point of interest, while accounting for all possible subsets of the remaining grid points. However, the procedure used to modify the input data in order to represent the absence of a given grid point is not clearly described. For example, it is unclear whether the value at that location is set to zero or replaced using an alternative strategy. Providing further clarification on this aspect would enhance the transparency and robustness of the analysis.

We have clarified the procedure in the revised Section 2.3, which now explicitly describes how the absence of a grid point is implemented in the SHAP computation. The relevant passage reads as follows:

"This is done by comparing the model's predictions with and without the grid point of interest, considering all possible subsets of the other grid points. The 'absence' of a grid point is simulated not by zero substitution, but by averaging the model's predictions over a range of plausible values for that location, drawn from the input data distribution."

RESULTS:

● Although the implemented revisions have improved the clarity of the explanation, it remains unclear whether the reference model is defined prior to, or derived from, the sensitivity analysis. On one hand, at line 212, the authors state that *"in each sensitivity experiment (blue bars), a single component of the reference model was modified,"* which suggests that the reference model serves as the baseline configuration for the sensitivity experiments. On the other hand, at line 223, it is stated that *"the reference model, optimized through these sensitivity experiments, represents the configuration that achieved the best balance of spatial robustness and computational efficiency,"* implying that it is the outcome of the sensitivity analysis itself. Clarifying the role and development of the reference model—specifically whether it represents the initial baseline or the final optimized configuration—would improve the consistency of the description and strengthen the interpretability of the sensitivity analysis experiments.

We acknowledge the ambiguity in our original description, which conflated two distinct steps: the initial process of identifying the reference configuration and the subsequent visual framework used to present those results in Figure 2. We have updated both Section 3.1 and the Figure 2 caption to distinguish these. In the revised text, we clarify that the reference model was first established by systematically testing modifications against a simple baseline (ReLU activation and MSE loss), and that this optimal configuration then served as the benchmark for the sensitivity analysis presented in Fig. 2. The revised passage in Section 3.1 now reads as follows:

"Starting from a baseline configuration with commonly used settings (ReLU activation functions and mean squared error loss), we systematically evaluated modifications to individual components, including activation functions and loss functions (MAE), kernel sizes, and data composition (Table 1). This allowed us to identify the most robust and efficient combination, which was then adopted as our reference model. To streamline the presentation in Fig. 2, the sensitivity results are organized around this reference. Each blue bar represents

a variant differing by only a single component, providing a direct visualization of how individual choices influence predictive skill."

● At line 223, the authors define the reference model as "*the model which best balances spatial robustness and computational efficiency*", compatible with the results reported in table 1. Although that, figure 2 reports higher predictive skill for the model with 5 layers. This discrepancy suggests that additional criteria may have been considered in selecting the reference model. Providing further clarification on this aspect would improve the transparency, reproducibility, and overall understanding of the experimental design.

We agree that our original phrasing ("best balance of spatial robustness and computational efficiency") was not sufficiently specific, and we have revised the text to make the rationale for selecting the 3-layer reference model more transparent. While the 5-layer configuration shows a marginally higher mean ACC in Fig. 2, the 3-layer model achieves statistically significant skill in a larger number of LMEs (Table 1). Because this study aims to identify regions where chlorophyll forecasts can be reliably generated, we prioritized the consistency of statistically significant skill across LMEs, particularly given the considerable computational cost of training a 5-layer architecture across the many independent models required by our main analysis.

"Although increasing the network depth from 3 to 5 convolutional layers yielded a marginally higher mean ACC, the 3-layer reference achieved statistically significant skill in a larger number of LMEs (5 out of 16) than the 5-layer configuration (4 out of 16). Given that this study aims to identify regions where chlorophyll forecasts can be reliably generated, we prioritized the consistency of statistically significant skill across LMEs over a marginal gain in mean ACC. Combined with the substantial computational cost of training a 5-layer architecture across the large number of independent models required by our main analysis (12 forecast start months × 24 lead times for two representative LMEs in the monthly forecasts, and 66 LMEs in the annual forecasts, all as 5-member ensembles), the 3-layer configuration was adopted as the reference model."

● In the review, the authors provided the rationale for adopting a 2D-to-0D approach. Including the motivation expressed in the answer also in the Discussion section would better

justify the reason behind the choice of this approach, clarifying its advantages and enhancing readability.

We have added the following sentence in the Discussion (Section 4) to clarify the rationale behind the 2D-to-0D design:

"Regional chlorophyll variability in LMEs is often modulated by basin-scale to global-scale spatial patterns associated with large-scale climate variability, yet ESM-based dynamical forecasts at the LME scale remain constrained by limited observational records, structural uncertainty, and high computational cost. To address this, we developed a CNN based deep learning framework that predicts LME-mean chlorophyll anomalies using global SST and chlorophyll fields. By leveraging the large-scale spatial patterns that modulate regional variability, this approach achieves skillful annual chlorophyll predictions across diverse oceanographic regimes in global LMEs."