Comments from Reviewer 1:

The authors use a neural network to estimate surface chlorophyll-a, a computationally efficient approach that appears to outperform traditional approaches like mechanistic biogeochemical ocean models. The manuscript presents some compelling results, but the experimental setup is not described well enough, and it is unclear why the comparison of chl-a estimates does not include any coastal regions. We thank the reviewer for their positive comments and appreciate their constructive feedback. We have addressed their comments in detail below.

general comments:

The manuscript is mostly well written and was easy to follow -- with a major exception: the basic setup of the experiments and implementation details are not well described and after reading through the whole manuscript I still do not quite know what, for example, "6-month predictions" are in the manuscript. Does the "6-month" imply a 6-month lead time, a 6-month forecast length, a 6-month time average or something else? Is there a distinction between "prediction" and "forecast" in the manuscript, if so, what is it? Sentences that are meant to explain experiments sometime increase the confusion of the reader, for instance: "These months correspond to lead-times one out of the six months of each forecast." (l. 156). Sentences like this example are confusing to the reader and could be improved considerably by rephrasing and adding some details. Please take the time and space to clarify how the experiments are set up and what is compared at what resolution (this includes space and time). We thank the reviewer for highlighting this, we realize that this was not clearly articulated in our methodology. By “6-month predictions,” we refer to forecasts with a 6-month horizon (i.e., predicting conditions up to 6 months ahead). We also evaluate performance at different lead times by analyzing forecast skill separately for each month within that horizon (i.e. lead times 1-6 *months*). This will be clarified in the revised manuscript.

Even a reader who does not know much about marine chl-a might find it surprising that the regions where performance evaluated, shown in Fig. 3, do not include any "yellow" values and seem to focus only on open-ocean regions (as an aside, a color bar or at least a description of what property is shown in Fig. 3 would be useful). That is, why weren't any coastal regions with high chl-a concentrations included in the comparison? The authors mention "fisheries management" and "harmful algal blooms" but then neglect to evaluate the model in the biologically active regions where most blooms occur and fishery is prevalent. In general, the chl-a estimates were compared mostly as a global average (Fig. 4, 5) or as averages in the large open-ocean regions (Fig. 7, 9); only Fig. 6 shows the performance on a finer spatial scale. We thank the reviewer for this insightful comment. We agree that the current regional evaluation, which focuses on open-ocean areas, does not fully represent the biologically active coastal regions that are relevant to the applications mentioned in the introduction. We are currently considering how best to address this limitation, either through a more detailed analysis of these regions, a revision of the discussion and introduction to better reflect the current scope, or both.

Additionally, we will improve the clarity of Figure 3 by adding a color bar and updating the figure caption.

Even in the computation of the RMSE, a spatial average appears to be used: "The spatially-averaged reconstructed time series has a RMSE of 0.01 ..." (l. 151). Why is the RMSE based on a spatial average? The use of spatial averaging is not explained well or mentioned when the RMSE is introduced. Please ensure that the reader knows at all times how key metrics are being computed. We thank the reviewer for highlighting this. We agree this aspect of the Methodology section was unclear and will revise the manuscript accordingly. We will ensure that the definitions for all key metrics are clearly described to avoid ambiguity.

In addition, I would suggest using nearshore regions in the comparison and evaluating the model performance at a higher resolution, both in space and time. We agree that evaluating performance in nearshore regions could offer valuable insights. While higher spatial or temporal resolution is not possible with the current model configuration (because resolution is fixed), we are exploring the feasibility of assessing model skill in nearshore areas using the existing output. We recognize the importance of this direction and consider it a key area for future development.

Furthermore, the authors later ponder how the decrease in ACC observed in Fig. 6 aligns with little to no increase in RMSE and other metrics in Fig. 9. They explain that "it is likely that the neural network's ability to capture the strong seasonal dynamics in the data (Figs. 7 and 8) is compensating for the decrease in performance with respect to the anomalies" (l. 168). That could well be, but if the RMSE is based on some spatially averaged chl-a, the averaging could have removed most of the effect of the anomalies. Unfortunately, a reader can only guess here, as it is unclear how the RMSE was computed. We thank the reviewer for pointing this out. The RMSE reported in that section is calculated over all spatial points and time steps, but we recognize that the description could be clearer. We will revise the manuscript to explicitly clarify this calculation and its implications.

Due to their distribution, when plotting and comparing chl-a values, they are often log-transformed. The authors mention once that a log-transformation was used, but it is unclear where and to what extent: "The physical ocean data was normalized using min-max normalization and the chl-a data was log-transformed" (l. 82) is the only information the reader gets. Was a log-transformation used when computing the ACC, NRMSE etc., are r_i and p_i in Eq 1-4 log-transformed? How were the climatologies computed? More importantly, perhaps, was a log-transformation used in the loss function for the neural network? The authors mention that they needed to modify the loss function: "so we modified the standard mean squared error (MSE) loss function by adding a small penalty for underestimation." (l. 79). With a log-transformation applied to chl-a, one would expect underestimation to be quite heavily penalized by the MSE. More information is needed to better interpret the results and the setup of the neural network. We thank the reviewer for these important questions and recognize that these details were not clearly described in the original manuscript. To clarify, the evaluation metrics (ACC, NRMSE, etc.) are computed on the original scale, once the neural network output has been transformed back to the natural scale. However, the network predicts log-transformed chlorophyll-a values, and the loss function is calculated on log-transformed. Despite this

approach, we observed underestimation during training, motivating our modification of the standard MSE loss. We will revise the manuscript and update the notation in the equations to clarify these points.

specific comments:

L 1: "Marine chlorophyll-a is an important indicator of ecosystem health, and accurate forecasting, even at the surface level, can have significant implications for climate studies and resource management a lightweight, resource-efficient neural architecture based on the U-Net that reconstructs surface, near-global chlorophyll-a from four physical predictors.": Accurately forecasting/estimating surface chl-a is a good check for "traditional" mechanistic models to verify that they can recreate some key biogeochemical dynamics. How would the output of a neural network model that only estimates surface chl-a be able to inform climate studies and resource management? Maybe this is a point that could be discussed further in Section 4. We thank the reviewer for raising this point. While our model focuses solely on surface chlorophyll-a, this variable can be used as a proxy for phytoplankton biomass and primary productivity and has the advantage of being easily verifiable through observations. Surface chl-a estimates can provide insights into ecosystem health, carbon cycling, and responses to climate variability. However, we acknowledge the importance of subsurface processes and vertical structures, which are not captured by our approach. We will expand the discussion in Section 4 to address these strengths and limitations.

L 59: "The goal of this work is to demonstrate that we can not only estimate chl-a from these four variables, but that by using publicly available forecasts of these as input, we are able to generate an ensemble of skillful chl-a predictions for six months into the future.": Here it would be useful for the reader to be more specific: are the 6-month predictions reliant on a 6-month forecast or are they produced from input 6 months into the past? We thank the reviewer for this question. To clarify, the six-month chlorophyll-a predictions are directly reliant on the six-month forecasts of the physical variables used as inputs. We will revise the manuscript to explicitly state this for clarity.

L 74: "Skip connections link matching layers in the encoder and decoder, facilitating the transfer of information.": Does this mean the first Conv3D layer is linked to the last one, etc.? Not exactly; the fourth convolutional layer in the encoder is concatenated with the output of the first upsampling operation in the decoder, and the second convolutional layer is concatenated with the output of the second upsampling layer. We will revise the Methodology section and Table 1 to clarify this.

Eq 1: It would be good to explain the terms in the equation a bit better (is the data log-transformed?) and move the equation up to where MSE and the terms are introduced. We thank the reviewer for highlighting this point; we recognize that the description could be clearer. We will revise the manuscript to explicitly clarify all metric calculations.

L 85: What motivated the choice of the 12 "monthly" neural networks? How much worse is the use of a single one for all months? This approach was chosen because aligned with the practical operational application that we had in mind, where the user would generate the chl-a

prediction starting at a given month that depended on the physics forecast available. We considered that this would result in smaller, more lightweight models that would be used according to the initialization month of interest. However, we acknowledge that this design choice may limit, by design, the models' ability to generalize across universal temporal relationships. We will revise the manuscript to include this.

L 90: "The optimal architecture found for this task has approximately six million trainable parameters...": Is this for one or all 12 of the networks? We thank the reviewer for this question; we realize that we did not provide sufficient detail on this point in the manuscript. Each one of the 12 networks has approximately six million parameters, and although we employ transfer learning by initializing one network using the weights from others, all parameters remain trainable and independently updated. We are aware that this is not the most efficient approach, as transfer learning could be further leveraged through parameter sharing or freezing to reduce the total count. We will clarify this in the manuscript.

L 97: "...provides daily and monthly data...": Here, or somewhere early on, mention if the networks produce daily or monthly mean estimates. We thank the reviewer for highlighting this point; we recognize that the description could be clearer. We will revise the manuscript accordingly.

L 122: "lead-time two": Does this mean a 2-month lead time? Yes. We will revise the manuscript to clarify this point.

Eq 2-4: How do these metrics compare to the cost function used for training the network, why not report/show that value as well? And mention if any of these chl-a values are log-transformed in these metrics. We thank the reviewer for this comment. The cost function used during training is based on the log-transformed chlorophyll-a values, but the evaluation metrics reported in the manuscript are computed on the original (non-log-transformed) scale. We will revise the manuscript to clarify this distinction. Regarding the training loss, we chose not to report it, as it is not commonly included in the evaluation of machine learning models where the focus is typically on independent performance on data that the model has never seen before.

L 146: "Rather than a direct comparison, we use BIO4 as a benchmark, recognizing that it simulates a wide range of interconnected biogeochemical processes across various depths, whereas our data-driven approach is specifically designed for surface chl-a prediction.": This sentence is a bit confusing. It makes sense to compare the neural network approach to a more classic reference approach for estimating surface chl-a. But why is this dependent on BIO4 also estimating a wide range of other properties? Maybe I just do not understand what "direct comparison" refers to in this context. We thank the reviewer for the comment. Our intention was to note that while we compare surface chlorophyll-a predictions, BIO4 is a more complex model that also simulates a broader range of biogeochemical processes. We will revise the sentence to clarify that BIO4's broader scope is mentioned for context.

L 150: The first sentence of Sec 3 is almost identical to that of Sec 2.2. Unfortunately, it is still not clear to me what a "set of 5-day predictions" means. We thank the reviewer for pointing this out. We will revise both sentences to improve clarity.

L 151: "The spatially-averaged reconstructed time series...": What kind of spatial averaging is performed here, before computing the RMSE etc.? All spatial points are averaged to obtain a single time series, and the metrics are then calculated on this averaged series. However, metric values computed without spatial averaging are reported elsewhere in the manuscript, we will clarify this distinction to avoid confusion.

L 168 and following figures: Are the BIO4 estimates that are shown forecasts as well? For what lead time? No, the BIO4 estimates are not forecasts, they are outputs from the analysis itself. We will clarify this in the manuscript.

***Citation**: https://doi.org/10.5194/egusphere-2025-1246-RC1*