Responses to Reviewer #2's Comments

This work presents a deep Learning framework to predict surface chlorophyll concentrations and anomalies across large marine ecosystems, with implications for fish catch forecasts.

The abstract and the introduction well describe the core idea and its fundamentals: the interactions between Earth's physical and biogeochemical fields is important for predicting future climate and the marine biogeochemical variability is critical to advance climate predictions based on bio-climate interactions.

Nonetheless, the methods and results section can be deeply improved in order to clarify the purposes of the research, its development and its scientific novelty.

The Method Section offers a detailed overview of the deep learning architecture, together with datasets information, sensitivity analysis and prediction performances information. Despite that, the description of the architecture lacks some important details, and the dataset description, though comprehensive, is presented in a confusionary way, without properly describing the variables collected and their role in training-validation-test phases, a fact that reduces readability and reproducibility.

The architecture developed for this project is a Convolutional Neural Network. Despite the authors having dedicated a paragraph of Methods Section and a paragraph of Results section to the description of the architecture, several fundamental aspects remain unclear—particularly the dimensions of the input and output data—reducing the clarity of the project's objectives and implementation.

The research question behind this project and consequently research purposes (i.e. the relevance of modeling mean chlorophyll within LMEs and the rationale behind using entire 2D maps to derive a single pointwise mean value for each LME) appears confusionary, a lack of clarity that is also reflected in the results. It is not particularly clear the task the manuscript intends to solve, and in particular the objective of some experiments (i.e. mechanisms underlying chlorophyll prediction skills, described in Figure 4, and the capacity to model interannual fish catch variations with chlorophyll anomalies as environmental drivers) and which is the scientific novelty they bring.

Moreover, the description of the experiments and of the results appears not always clear, and more explanations (i.e. a more detailed description of the content of Figure 4, and of the relationship between the anomaly correlation skill behavior described in figs 4a and 4c with the maps of figs 4b and 4d) would improve readability and strengthen the paper as they would support the research question posed by the authors. Finally, the descriptions of certain figures, such as Figures 4 and 5, lack sufficient detail, limiting the comprehension of both the analyses conducted and the importance and relevance of the results obtained.

We thank Reviewer #2 for the detailed and constructive comments. Below, we address each comment in the order presented. We have substantially revised the Methods and Results sections to improve clarity, reproducibility, and scientific motivation. Key changes include: reorganized dataset descriptions with summary tables, explicit definitions of input/output dimensions and anomaly computation, improved figure captions and quality, and strengthened discussion of scientific novelty. Detailed responses to each specific comment are provided below.

In consideration of the previous points, the paper is acceptable for publication after major revisions.

A list of punctual issues is listed below.

ABSTRACT:

(L13-14): Enhance clarity and focus on this sentence to be more consistent with the problem presented.

We agree that this sentence lacked clarity. We have revised it for consistency with the problem statement. The revised abstract now reads:

"Earth System Models (ESMs) capture large-scale physical–biogeochemical coupling, but their biogeochemical prediction skill varies substantially across regions and lead times due to sparse observational records, structural uncertainties in biogeochemical models."

(L20): The sentence emphasizes the relevance of physical–biogeochemical coupling processes; however, it remains unclear whether the network explicitly learns this coupling or merely reproduces its effects, as well as the mechanisms by which such learning or reproduction is achieved.

As the reviewer pointed out that the CNN does not encode physical–biogeochemical coupling equations; rather, it learns statistical relationships from the training data. However, a couple of evidences support this interpretation: first, SHAP attribution maps reveal spatial patterns aligned with established climate dynamics such as ENSO evolution and off-equatorial Rossby wave propagation (Fig. 4b,d); second, the model's seasonal skill variation mirrors the ENSO spring predictability barrier (Fig. 4a); and third, the CNN consistently outperforms persistence forecasts (Fig. 4a,c), suggesting that the model captures predictable signals beyond simple autocorrelation. We have softened the language in the abstract to state that the prediction skill is associated with the previously-identified physical–biogeochemical coupling processes rather than implying direct learning of the coupling itself. The revised abstract reads as follows:

"The prediction skill is associated with physical-biogeochemical coupling processes triggered by large-scale climate variability, consistent with the mechanisms previously identified in dynamical forecasts."

(L22): The term "chlorophyll anomalies" is introduced, but it is not defined, together with the baseline used for its computation. The entire article strengthens on this aspect, but there is no formal definition of anomaly.

Following the reviewer's comment, we have added a formal definition of chlorophyll anomalies in Section 2.1.2 of the revised manuscript, as the abstract is not an appropriate place for such methodological detail. The relevant definition in Section 2.1.2 reads as follows:

"Chlorophyll anomalies are defined as deviations from monthly climatological means, computed separately for each dataset (CMIP6 models, reanalysis, satellite observations) over their respective reference periods."

INTRODUCTION:

(L35): The inclusion of references to the definition of ESMs would facilitate a deeper understanding of the purposes of the project.

We agree with this suggestion and have added references to the definition of Earth system models in the revised introduction.

"Earth System Models (ESMs), which integrate biogeochemical processes within physical climate frameworks (Flato, 2011; Bonan and Doney, 2018),"

(L50): Deep learning models are highly sensible on data coverage. In particular, observational gasps and data-sparse components represent a huge limitation for the majority of deep learning approaches. Even if their usage grows with the increasing availability of data, sparse coverage still represents a limit for these models. A clearer explanation of the statement asserting that deep learning methods are well suited to data-sparse components would strengthen the justification for adopting a deep learning approach for this application.

We acknowledge that the original phrasing was somewhat ambiguous. We do not claim that deep learning methods are inherently robust to data sparsity. Rather, our approach overcomes the limited length of observational records by training the CNN on multi-century simulations from various climate models participated in CMIP6. The revised part of the Introduction is as follows:

"These data-driven models can learn complex, nonlinear relationships and can be trained on data-rich climate model simulations to overcome the limited length of observational records and structural uncertainties in process-based models, making them well-suited for seasonal-to-annual biogeochemical forecasting (Reichstein et al., 2019)."

(L62): The manuscript does not clearly describe the outputs of the deep learning model. Both chlorophyll concentrations and chlorophyll anomalies are presented as model products; however, the definition and interpretation of the anomaly are not provided. Clarification of this aspect would improve the reader's understanding of the overall study. Furthermore, it is unclear whether each LME is modeled independently or whether the model produces a global output from which individual LMEs are subsequently extracted and analyzed.

Following the reviewer's comment, we have clarified the model output definition in the revised Section 2.1.2. The model predicts LME-mean chlorophyll anomalies, with separate models trained for each LME. The anomaly is defined relative to the climatological mean of the training data. The revised Section 2.1.2 reads as follows:

"The model predicts area-averaged chlorophyll anomalies for individual LMEs from global spatial fields. Separate CNN models sharing the same architecture are trained for each combination of target LME, forecast type (annual or monthly mean), and lead time."

"Chlorophyll anomalies are defined as deviations from monthly climatological means, computed separately for each dataset (CMIP6 models, reanalysis, satellite observations) over their respective reference periods."

(L65-68): I think a re-organization of the last sentences of the introduction would enhance the comprehension of the project. The current description of the dataset appears overly detailed for an introductory section, while some key elements, such as a clear definition of the model outputs, are not sufficiently addressed. It is therefore recommended to revise these passages by emphasizing the general characteristics of the proposed algorithms and providing only high-level information about the dataset, while relocating the detailed dataset description to the dedicated method section.

We agree that the introduction would benefit from a more focused presentation. We have streamlined the dataset description in the Introduction and relocated detailed information to Section 2. Specifically, references to the GFDL assimilation system and satellite sensor names have been removed from the Introduction and are now described in full in Section 2.2. The last paragraph of the Introduction now provides only high-level information about the model framework, inputs, and evaluation strategy.

METHODS:

Section 2.1: the architectural description lacks key details required for reproducibility, such as a comprehensive table of all hyperparameters and a clear rationale for the choice of the proposed architecture and its components, such as including the use of GELU activations and the selected loss function. To further improve the clarity of the manuscript, it is recommended to present the network architecture, dataset, and validation strategy in separate subsections.

Following the reviewer's recommendation, we have reorganized Section 2 into separate subsections for network architecture (Section 2.1.1), prediction task and model output (Section 2.1.2), and training and validation strategy (Section 2.1.3). Data sources and preprocessing are addressed in Section 2.2. A comprehensive hyperparameter table has been added to the Supplementary Information (Table S1), and is reproduced below as Table B1. The rationale for key architectural choices, including GELU activations and MAE loss function, is provided in Section 2.1.1, where these are identified as the optimal configuration through systematic sensitivity analysis (Section 3.1).

Table B1. Hyperparameter configuration of the CNN model.

Parameter	Value
Convolutional layers	3
Filters per layer	35
Kernel size	3×3
Max pooling layers	2
Pooling size	2×2
FC layer neurons	50
Activation function	GELU
Loss function	MAE
Optimizer	Adagrad
Learning rate	0.005
Batch size	32
Training epochs	135 (early stopping patience = 30)

(L91): the concept of anomaly correlation coefficient is introduced, but not defined. Including its definition, along with a brief description, would enhance the reader's understanding of the results.

We now have added an explicit definition of the anomaly correlation coefficient (ACC) in the revised Section 2.1.2, along with the method for assessing statistical significance. The relevant passages are as follows:

"Prediction performance was evaluated using the anomaly correlation coefficient (ACC), computed as the temporal correlation between predicted and observed time series of LME-averaged chlorophyll anomalies. Statistical significance was assessed following a method using effective degrees of freedom corrected for temporal autocorrelation (Bretherton et al., 1999):

$$N_{eff} = \frac{N}{\sum_{t=0}^{t=N-1}\left(1-\frac{t}{N}\right) r_t^F r_t^O}, \qquad (1)$$

where N is the number of samples in the forecast (F) and observed (O) time series, and r_t^F *and* r_t^O *are estimates of autocorrelation in each time series at lag t."*

Section 2.2 describes the dataset used, including the input, validation, and test sets, and provides details on input data preprocessing. I recommend reorganizing this section to clarify the distinctions between datasets used for different purposes. Additionally, more detail on the input data preprocessing would improve clarity, as the structure of the input data is not fully specified. Specify source, variables, spatial resolution, temporal frequency of data used; in particular, clarify which dataset

collects the input variables used for training, validation and test. Use a table if it can help. Moreover, it is unclear whether the inputs consist of concatenated global 2D maps of SST and chlorophyll anomalies or of 2D maps defined separately for each LME. Likewise, the description of the network output lacks clarity: it is not evident whether the output represents a mean chlorophyll value across all LMEs or a spatial map over each LME, nor whether the model predicts chlorophyll concentrations, chlorophyll anomalies, or both.

Following the reviewer's recommendation, we have reorganized Section 2.2 and added a summary table clarifying the datasets used for training, validation, and testing, specifying the source, variables, spatial resolution, and temporal coverage of each dataset. A summary table has been added to the Supplementary Information (Table S2) and is reproduced below as Table B2. The inputs consist of concatenated global 2D maps (180×360 at 1° resolution) of SST and chlorophyll anomalies for three consecutive months. The output is a single scalar value representing the LME-mean chlorophyll anomaly at the target lead time. Separate CNN models are trained for each LME.

Table B2. Summary of datasets used for training, validation, and testing.

Dataset	Samples	Description
Training	8013	
CMIP6 piControl	5917	16 models × ~370 yrs each
CMIP6 historical	2096	16 models × ~131 yrs each
Validation	2043	
CMIP6 piControl	1483	16 models × ~93 yrs each
CMIP6 historical	528	16 models × ~33 yrs each
GFDL ECDA reanalysis	32	1965–1997
Test	23	
Satellite Observations	23	1998–2021

(L103): The input mask fills missing values with zeros. It would be helpful if the authors could provide additional insight into the rationale behind this choice. In particular, further clarification on whether missing values and land points are treated differently, and on the network's ability to distinguish between these cases, would enhance the reader's understanding.

Following the reviewer's comment, we have elaborated on the rationale behind the zero filling strategies in the revised manuscript. As described in our preprocessing clarification, we constructed a unified binary mask from the entire satellite record (1998–2021), permanently flagging any grid cell with at least one missing value in any month. The flagged regions

largely correspond to land-adjacent, polar, or persistently cloud-covered areas where chlorophyll signals are typically absent or negligible. Because masked grid cells maintain constant zero values across all time steps and training samples, they carry no temporal variability and thus contribute no learnable signal to the CNN. As a result, the network does not need to distinguish between these cases, as neither provides information relevant to the prediction task. The same mask is applied to CMIP6 training data to ensure spatial consistency between training and evaluation. The revised Section 2.2 reads as follows:

"Because both land and masked ocean grid cells maintain constant zero values across all time steps and all training samples, they carry no temporal variability and thus contribute no learnable signal to the CNN. The network effectively learns to rely on grid cells with non-zero, time-varying inputs."

(L124): Paragraph 2.3 introduces SHAP as a method for interpreting model predictions and identifying dominant spatial drivers (L118). However, the role of SHAP in this context is not entirely clear. Given that the network inputs consist of SST and chlorophyll anomalies, one would expect the analysis to highlight the relative importance of these input variables. Instead, at (L124) it is stated that feature (i) corresponds to a specific grid point in the input map, which introduces some ambiguity regarding what information the SHAP analysis is intended to convey. A clearer explanation of how features are defined and how SHAP results should be interpreted would improve the clarity and understanding of the results.

We acknowledged that the role of SHAP and the definition of features required further clarification. We have revised Section 2.3 to address this. In our framework, each "feature" corresponds to a specific grid point in the input 2D map (i.e., a particular location's SST or chlorophyll anomaly value). SHAP values therefore quantify the contribution of each spatial location's input value to the predicted LME-mean chlorophyll anomaly. This spatial attribution reveals which remote or local regions most influence the prediction for a given LME, allowing us to identify physically meaningful teleconnection patterns (e.g., ENSO-related signals in the tropical Pacific influencing distant LMEs). The revised Section 2.3 reads as follows:

"Each feature corresponds to a specific grid point in one of six input channels: three consecutive months of chlorophyll anomalies and three consecutive months of SST anomalies. We compute SHAP values separately for SST and chlorophyll by aggregating across their respective three monthly channels, then visualize them as spatial attribution maps. These maps reveal which area of the input fields most strongly influence the predicted chlorophyll anomaly in each target LME at different forecast lead times. Additionally, comparing the spatial extent and magnitude of SHAP values between SST and chlorophyll maps allows us to assess the relative importance of physical versus biological drivers for each region's predictability."

Section 2.4: the scope of this section appears scientifically obscure. Chlorophyll (or its anomaly) timeseries from satellites or from ESM models can be directly used to predict catch timeseries. Which are the added values of using NN derived chlorophyll? One would expect, at least, a comparison between catch timeseries predicted using chlorophyll from satellites, or to see the advantage of using the NN derived chlorophyll.

We appreciate this comment and agree that the added value of using CNN-derived chlorophyll needed to be more clearly articulated. Satellite-observed chlorophyll can indeed serve as a bottom-up predictor of fish catch variability, as chlorophyll reflects primary productivity that propagates through marine food webs. However, satellite chlorophyll is available only retrospectively, and at the forecast start time, the only operational alternative would be a persistence forecast using the prior year's annual mean chlorophyll. In contrast, our CNN predicts the target year's chlorophyll anomaly from NDJ (November of Year 0 – January of Year 1) observations, providing forecasts before annual catch data become available The fish catch analysis provides exploratory evidence that these advance predictions retain environmental signal relevant to interannual catch variability. We have clarified this distinction in the revised Section 2.5, which now reads as follows:

"Simple linear regression was applied to predict normalized fish catch anomalies using annual chlorophyll anomaly forecasts generated by providing the CNN with satellite observations from NDJ (November of Year 0 – January of Year 1), at lag=0 (same year) and lag=1 (following year)."

"This analysis is intended as an exploratory demonstration of potential downstream applications of the chlorophyll forecasting framework, rather than a validated fisheries prediction system."

RESULTS:

Section (3.1) presents a very interesting and informative analysis; however, some of the architectural details discussed here would be more appropriately included in the Methods section, within the description of the model architecture. In addition, to improve the comprehensibility of the architecture described in the Methods section and to maintain focus on the model's results, it is suggested to move this sensitivity analysis to a Supplementary Materials section.

We agree that architectural details are more appropriately presented in the Methods section and have relocated architectural details to the revised Section 2.1.1. With these details now in the Methods, Section 3.1 serves a focused role: providing the empirical justification for the reference model configuration adopted throughout the study, demonstrating that surface chlorophyll anomalies provide prediction skill comparable to or exceeding that achieved with subsurface temperature inputs. Retaining this analysis in the main text therefore maintains the transparency of the methodological choices underlying the results presented in subsequent sections.

(L150): In the caption of Figure 2, the baseline model is described as sharing the architecture of the reference model, while differing in certain training settings, such as the loss function. This suggests that the reference model represents an optimized version of the baseline. However, at line 148 it is stated that the sensitivity analysis presented in this paragraph originates from the reference model, with a single component modified in each experiment. Could the authors clarify the contributions of these sensitivity experiments to the reference model, and how the reference model was optimized relative to the baseline? Providing this explanation would help improve the reader's understanding of the experimental design and the relationship between the baseline, reference, and sensitivity models.

We have revised Section 3.1 accordingly for the further clarification of the experimental design. The baseline model uses commonly adopted settings (ReLU activation, MSE loss), and the reference model was identified by systematically modifying individual components from this baseline to find the optimal combination. We have also added Table 1, which summarizes the configuration of each sensitivity experiment (input variables, training data, and architecture), clarifying the relationship between the baseline, reference, and modified models and is reproduced below as Table B3.

Table B3. Sensitivity experiment configurations. Each experiment modifies one component relative to the reference model (bottom row), with all other settings held constant. Sig. LMEs indicates the number of regions with statistically significant prediction skill ($p < 0.10$) out of 16 representative LMEs. Training: CMIP6 historical (1850–2014) + piControl (500 years) + GFDL-ECDA reanalysis (1965–1997). Validation: GFDL-ECDA reanalysis (1998–2017). Abbreviations: CHL – surface chlorophyll anomalies; θ – subsurface potential temperature (0–300 m average); hist – historical; piC – piControl. "—" in the Architecture column indicates the same configuration as the reference model (3×3 kernel, GELU, MAE). Note: Prediction skill measured as ACC averaged across 16 representative LMEs.

Experiment	Predictors	Training Data Source	Architecture	Sig. LMEs
Baseline	SST, CHL	CMIP6 (hist+piC) + Reanalysis	ReLU, MSE	1/16
Kernel size: 5×5	SST, CHL	CMIP6 (hist+piC) + Reanalysis	Larger kernel	3/16
Number of layers: 5	SST, CHL	CMIP6 (hist+piC) + Reanalysis	5 layers	4/16
Resolution (5°)	SST, CHL	CMIP6 (hist+piC) + Reanalysis	Coarser	5/16
SST only	SST	CMIP6 (hist+piC) + Reanalysis	—	4/16
Subsurface temp. only	θ_{o-300m}	CMIP6 (hist+piC) + Reanalysis	—	2/16
SST + Subsurface temp.	SST, θ	CMIP6 (hist+piC) + Reanalysis	—	3/16
Chlorophyll only	CHL	CMIP6 (hist+piC) + Reanalysis	—	5/16
Without piControl	SST, CHL	CMIP6 hist + Reanalysis	—	3/16

Without Reanalysis	SST, CHL	CMIP6 (hist+piC) only	—	5/16
Log-transformed	SST, log(CHL)	CMIP6 (hist+piC) + Reanalysis	—	4/16
Reference	SST, CHL	CMIP6 (hist+piC) + Reanalysis	3×3, GELU, MAE	5/16

(L155): The concept of prediction skill is not defined, and its meaning remains somewhat unclear. In particular, in Figure 2, it is not evident what exactly the prediction skill measures. Including a brief description would enhance both clarity and will facilitate the comprehension of the proposed results.

In the revised manuscript, we have modified the Figure 2 caption to clarify what the prediction skill measures. Each bar represents the ACC averaged across all 16 selected LMEs, and the green dashed line indicates the ACC averaged only across LMEs where prediction was statistically significant ($p < 0.10$) in at least one configuration. The formal definition of ACC and its significance test (based on effective degrees of freedom corrected for autocorrelation; Bretherton et al., 1999) is now provided in Section 2.1.2. The revised caption reads as follows:

"Bars indicate the average correlation skill across 16 selected regions for each model variation. All configurations are derived from the reference model (red bar), which exhibited the highest overall predictive performance. In each sensitivity experiment (blue bars), a single component of the reference model was modified, either a structural aspect (e.g., kernel size, number of layers) or input data configuration (e.g., resolution, predictor variables, log transformation). The baseline model, shown at the top, shares the same architecture as the reference model but uses standard training settings (ReLU activation, MSE loss). The green dashed line shows the average skill across regions where prediction was statistically significant ($p < 0.10$) in at least one configuration. See Table 1 for detailed input variable configurations corresponding to each experiment shown."

In addition, in the process of revising the manuscript, we identified and corrected an error in the computation of annual prediction skill, where correlation coefficients were previously computed per ensemble member and averaged rather than derived from the ensemble mean time series. This correction has been applied throughout the revised manuscript.

(L175): Could the authors clarify the statement, "The inclusion of additional input datasets generally improved the model's prediction skill"? It should be noted that adding input variables does not necessarily guarantee improved model performance; if the additional inputs have weak correlation with the target, their inclusion could potentially lead to overfitting. Providing a reference and a more detailed explanation would help clarify this point and strengthen the interpretation of the results. Moreover, the choice to include chlorophyll as input variable when predicting chlorophyll itself

should be clarified. Moreover, it would be helpful to provide a table which contains for each test the input variables used for it. It is somehow difficult to reconnect the text to names listed in figure 2.

We have revised the text to clarify that additional training data sources (not input variables) improve skill by providing diverse climate states that mitigate overfitting. The use of chlorophyll as both input and predictor target is standard in geophysical forecasting, where the current state of a variable serves as an initial condition for predicting its future evolution. The sensitivity analysis (Fig. 2 and Table 1) suggests that including chlorophyll as a predictor provides substantial additional skill beyond physical variables alone. The input variable configurations for each sensitivity experiment are summarized in Table B3, provided in our response to the comment on L150 above.

From Figure 3a, it appears that the CNN output is represented as a single mean value for the entire LME, resulting in a uniform color. Could the authors clarify whether this interpretation is correct, or if the correlation is instead computed at the level of individual grid points? Providing this clarification would help improve the reader's understanding of the figure and the network's output. Based on Figure 1, the inputs appear to consist of timeseries of two-dimensional spatial fields, whereas the outputs correspond to timeseries of zero-dimensional quantities (i.e., single surface values). If this interpretation is correct, the rationale for adopting a two-dimensional–to–zero-dimensional mapping should be explicitly discussed. In particular, it would be helpful to clarify the intended purpose and advantages of this approach compared to the use of a simple spatial average, as well as to articulate the scientific novelty that this methodology is expected to provide.

As the reviewer mentioned, the model ingests global 2D fields as input and predicts a single LME-averaged chlorophyll anomaly. This design exploits basin-scale to global-scale spatial patterns that modulate regional LME chlorophyll variability. This is supported by SHAP attribution maps (Fig. 4b,d) which suggest that the CNN utilizes spatial information extending well beyond the target LME boundaries, including equatorial Pacific SST patterns associated with ENSO evolution (LME 11) and Indian Ocean signals consistent with Rossby wave propagation (LME 30). In contrast, simply averaging the input fields into scalar indices would discard this spatial structure, preventing the model from distinguishing, for example, between eastern and western Pacific SST anomalies that have opposing effects on regional productivity. The scientific value of the 2D-to-0D approach lies in enabling the CNN to identify which spatial patterns across the global input fields are most relevant for each LME's chlorophyll variability. This is now clarified in the revised Section 2.1.2 as described in our response to the comment on L62 above.

Improve the quality and clarity of the figure 3: y axis is missing the label and unit, and the text should be enlarged.

The axis label with units is now added and the text is also enlarged for improved readability. The revised figure is reproduced below as Figure B1.

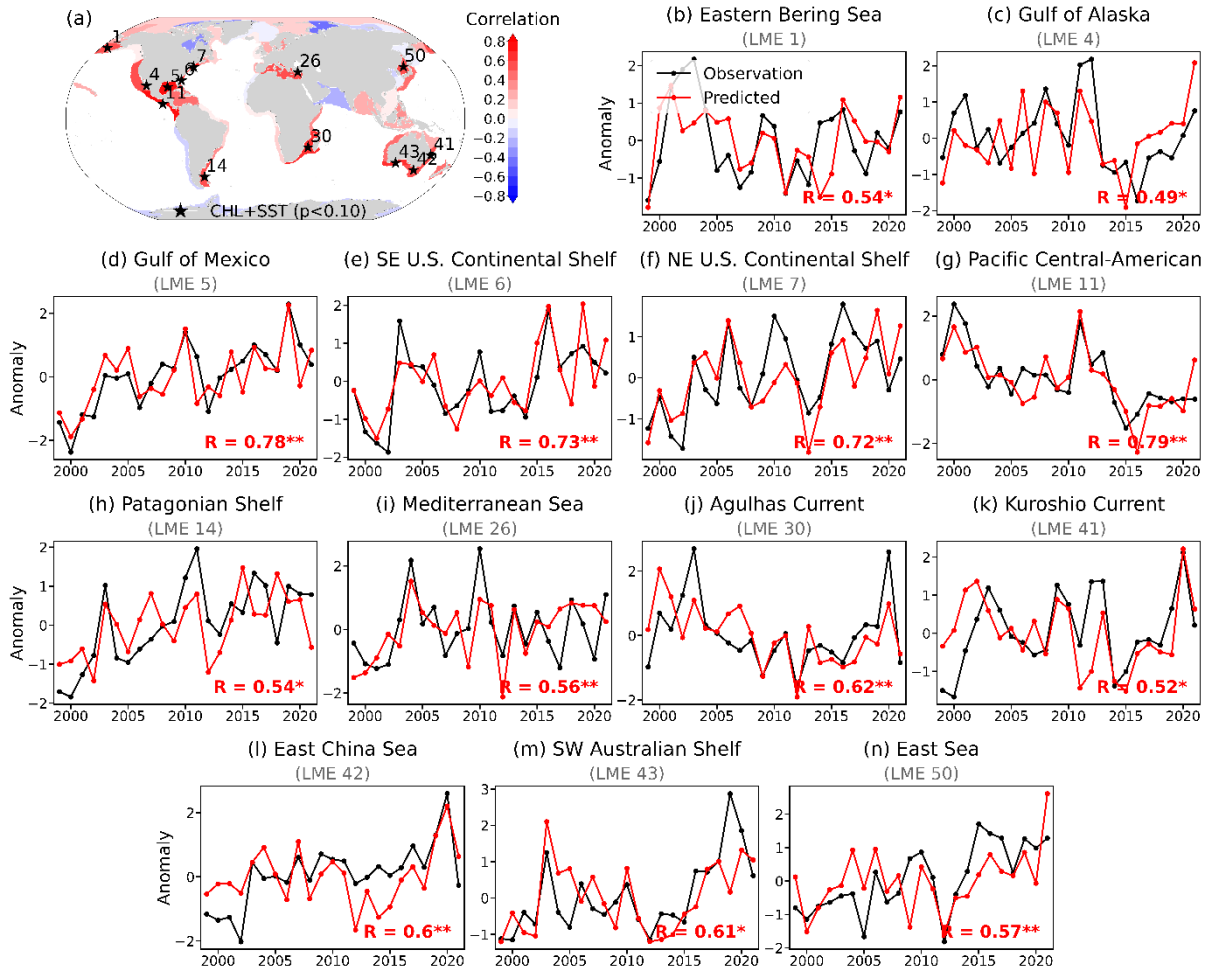


Figure B1: Chlorophyll prediction skill across Large Marine Ecosystems (LMEs). a Correlation coefficients between LME-averaged satellite-derived and predicted annual mean chlorophyll anomalies (1998-2021). The model takes November(Year 0)–December(Year 0)–January(Year 1) satellite observations as input and predicts the annual mean chlorophyll anomaly averaged over January–December of Year 1. Shading shows the prediction skill of the reference model using both chlorophyll (CHL) and sea surface temperature (SST) as input. Black asterisks mark LMEs with statistically significant correlations (P<0.1). b-n Time series of normalized annual mean chlorophyll anomalies from satellite observations (black) and model predictions (red) for the thirteen LMEs with significant prediction skill (corresponding to asterisks in panel a). Correlation values are indicated with significance levels (* : P<0.1, ** : P<0.05).

(L195): The exact number of CNN input variables is not entirely clear. While SST and chlorophyll anomalies are listed as inputs in the introduction (L62), a different description appears later, stating that the model was "tested with a combination of physical and biogeochemical inputs, that is, SST only, chlorophyll only, and both SST and chlorophyll." Could the authors kindly clarify the reason for this apparent discrepancy? If a sensitivity analysis was conducted to determine the

optimal set of input variables, it would be helpful to briefly describe the procedure. Otherwise, specifying the exact input variables used in the current model would improve clarity for the reader.

The sensitivity analysis in Section 3.1 tested various model configurations including input variable combinations (SST only, chlorophyll only, and both SST and chlorophyll), identifying the combined SST+CHL model as the reference configuration. In the original Section 3.2, we additionally presented SST-only and CHL-only results across global LMEs, which created confusion between the sensitivity analysis and the final evaluation. We have revised Section 3.2 and Figure 3 to report only the reference model results, clearly separating sensitivity testing (Section 3.1) from prediction evaluation (Section 3.2). The revised Section 3.2 reads as follows:

"The reference model derived from the sensitivity experiments was applied across all global LMEs to evaluate its skill in forecasting monthly to annual chlorophyll anomalies. Annual forecasts were generated by providing the model with satellite observations of SST and chlorophyll from three consecutive months in early boreal winter (November to January), with the model predicting the following calendar year."

(L213): The manuscript states that "including surface chlorophyll anomalies, either alone or as an additional predictor, substantially increased the number of LMEs where the model achieved high prediction skill." In the introduction, chlorophyll anomalies are already presented as an input to the model, whereas here it appears that they are added subsequently. Could the authors kindly provide a more detailed explanation of how the input data are structured and used? Clarifying this point would improve the reader's understanding of the model setup and the role of different predictors.

The sentence described results from the input variable sensitivity analysis, which now has been moved entirely to Section 3.1. Section 3.2 and Figure 3 have been revised to report only the reference model (SST + chlorophyll) results across all 66 LMEs, eliminating the overlap between sensitivity testing and prediction evaluation. The revised Section 3.2 no longer discusses alternative input configurations, as described in our response to the comment on L195 above.

(L215-220): The caption of Figure 4 lacks clarity, and the prediction task described in lines 215–218 would benefit from a more detailed explanation. In particular, the inputs and outputs of the task should be explicitly specified, and the procedure used to compare predictions with observations should be described more clearly. For example, it is unclear which quantities are being compared at each grid point in Figures 4a and 4c. Additionally, the captions for Figures 4b and 4d are ambiguous; as currently presented, it appears that two sequences of three maps are shown. Consideration could be given to splitting this content into two separate figures in order to improve readability and facilitate the reader's understanding of the task.

We have revised the Figure 4 caption to explicitly state the inputs (3-month SST and chlorophyll anomaly maps), outputs (monthly LME-mean chlorophyll anomaly at each lead time), and the evaluation metric (ACC between predicted and satellite-observed LME-averaged 3-month running mean chlorophyll anomalies). We have also clarified the relationship between Figs. 4a/4c (ACC skill as a function of forecast start month and lead time) and Figs. 4b/4d (SHAP attribution maps showing which input regions drive the prediction skill at selected lead times). Regarding the suggestion to split Figure 4 into

two separate figures, we have considered this carefully but prefer to retain the current layout as it allows direct visual comparison between skill patterns and their physical attribution. The revised Figure 4 is reproduced below for the reviewer's convenience (Fig. B2).

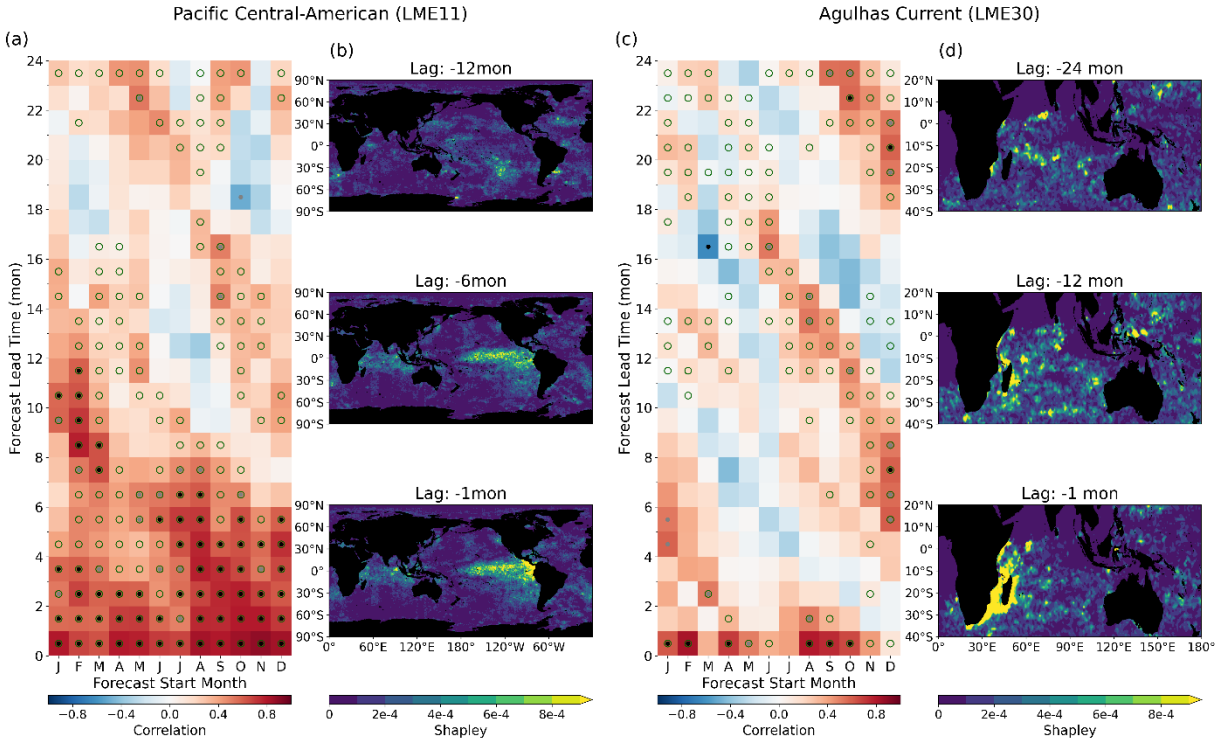


Figure B2: Monthly prediction and mechanism underlying chlorophyll prediction skill. a,c Anomaly correlation coefficient between predicted and satellite-observed 3-month running mean chlorophyll anomalies (LME-averaged) as a function of forecast start month (x-axis) and lead time (y-axis). Black dots indicate significant skill at P < 0.05, while grey dots indicate P < 0.10. Green open circles indicate skill exceeding the persistence model. b,d Spatial maps of absolute Shapley values at selected input lag times (indicated above each panel), illustrating which regions in the input fields contribute most to the predictions. Lag denotes the time offset of input observations relative to the forecast target period. For each LME, the Shapley values are shown for the most dominant predictor variable: SST for LME 11 (b; lags of −1, −6, and −12 months) and chlorophyll for LME 30 (d; lags of −1, −12, and −24 months).

The analysis presented in the latter part of paragraph 3.2 is interesting, and the results shown in Figure 4 are valuable. Nevertheless, the paragraph would benefit from a more detailed explanation of what is the content and the relevance of figures 4a and 4c, together with the implications between Figures 4a and 4b, as well as between Figures 4c and 4d. Clarifying these connections would greatly enhance the reader's understanding of the results and their interpretation.

We appreciate this suggestion and have expanded Section 3.2 and revised the Figure 4 caption to clarify the connection between figures. Fig. 4a/4c show how prediction skill varies with forecast start month and lead time, revealing seasonal dependence (e.g., skill drops for initializations crossing boreal spring, consistent with the ENSO spring predictability

barrier). Figs. 4b/4d show SHAP attribution maps at selected lead times, revealing the spatial regions that drive the predictions. The connection is that high-skill forecast start months in Figs. 4a/4c correspond to SHAP patterns in Figs. 4b/4d that align with known climate variability patterns (e.g., ENSO spatial structure), providing physical interpretability for the model's skill. The revised Section 3.2 reads as follows:

"In the Pacific Central-American region, the model exhibits seasonally varying forecast skill, with statistically significant correlations extending up to 12-month lead times for forecasts initialized during boreal winter (Fig. 4a). Prediction skill for chlorophyll is enhanced during boreal fall and winter, when large-scale climate variability such as ENSO is more predictable, but diminished during boreal spring and early summer, coinciding with the well-documented 'spring predictability barrier' of ENSO."

"*These patterns suggest that the model captures climate-driven signals to enhance chlorophyll prediction in this region, consistent with previous observational and modeling studies of primary productivity in the tropical Pacific (Park et al., 2019; Pennington et al., 2006; Sasai et al., 2012)."*

(L239): The manuscript states that "the recurrence of this pattern in the model's predictions indicates that it captures subsurface ocean memory in addition to surface signals." Could the authors clarify why the recurrence of this pattern is interpreted as evidence of subsurface ocean memory, given that subsurface variables do not appear to have been used or introduced as input to the model? Providing additional explanation would help improve the reader's understanding of this conclusion.

The physical basis for this point is that surface chlorophyll reflects subsurface ocean dynamics, through nutrient supply pathways, effectively carrying an imprint of the subsurface state (Park et al., 2018a; Lim et al., 2022; Lee et al., 2024). The diagonal banding pattern in Fig. 4c (Fig. B2 above), which matches the known winter-to-winter reemergence features in dynamical seasonal prediction, suggests that the CNN leverages this surface-encoded subsurface information from the input satellite observations. We have revised the text to read:

"The recurrence of this pattern in the model's predictions indicates that initial surface conditions reflect underlying subsurface ocean states, consistent with the demonstrated sensitivity of surface chlorophyll to subsurface dynamics (Park et al., 2018a; Lim et al., 2022; Lee et al., 2024)."

(L248): For the sake of comparison, it would be helpful to include the ENSO dynamics in a Supplementary Material section, providing a baseline for reference alongside Figures 4b and 4d.

For the Pacific Central-American region (Fig. 4b), the SHAP attribution patterns can be compared with the canonical spatial evolution of ENSO SST anomalies documented in Timmermann et al. (2018, Nature, Fig. 3f–m). For the Agulhas Current region (Fig. 4d), the westward-propagating patterns are consistent with the large-scale dynamics documented in Jeon et al. (2022, Fig. 3b–e). We have added these references in the revised text to provide the reader with a direct basis for comparison.

Reference

Timmermann, A., An, S.-I., Kug, J.-S., Jin, F.-F., Cai, W., Capotondi, A., Cobb, K. M., Lengaigne, M., McPhaden, M. J., Stuecker, M. F., Stein, K., Wittenberg, A. T., Yun, K.-S., Bayr, T., Chen, H.-C., Chikamoto, Y., Dewitte, B., Dommenget, D., Grothe, P., Guilyardi, E., Ham, Y.-G., Hayashi, M., Ineson, S., Kang, D., Kim, S., Kim, W., Lee, J.-Y., Li, T., Luo, J.-J., McGregor, S., Planton, Y., Power, S., Rashid, H., Ren, H.-L., Santoso, A., Takahashi, K., Todd, A., Wang, G., Wang, G., Xie, R., Yang, W.-H., Yeh, S.-W., Yoon, J., Zeller, E., and Zhang, X.: El Niño–Southern Oscillation complexity, Nature, 559, 535–545, https://doi.org/10.1038/s41586-018-0252-6, 2018.

L265-270: move in material and method the description of models.

Following the reviewer's comment, we have moved the description of the GFDL dynamical forecast system to the Methods section (Section 2.4), where it now appears at the beginning of the comparison methodology. A cross-reference to Section 2.4 has been added in Section 3.4 to guide readers to the model description.

Explain better how correlation between satellite chlorophyll and predictions by DL and dynamics are computed.

In the revised manuscript, we have elaborated on how the correlation analysis was conducted. Both the deep learning and dynamical models are evaluated against satellite-observed annual mean chlorophyll anomalies using the ACC metric. The deep learning model is evaluated over 1998–2021, while the dynamical model is evaluated over 1998–2017, reflecting the availability of retrospective predictions. To account for the difference in evaluation periods and ensemble variability in the deep learning model, we employed a double bootstrap procedure that resamples both ensemble members and temporal periods, yielding 95% confidence intervals for the deep learning model's correlation skill. This procedure is described in detail in the revised Section 2.4. The revised Figure 5, reproduced below as Fig. B3, illustrates the updated comparison with bootstrap confidence intervals and revised classification criteria.

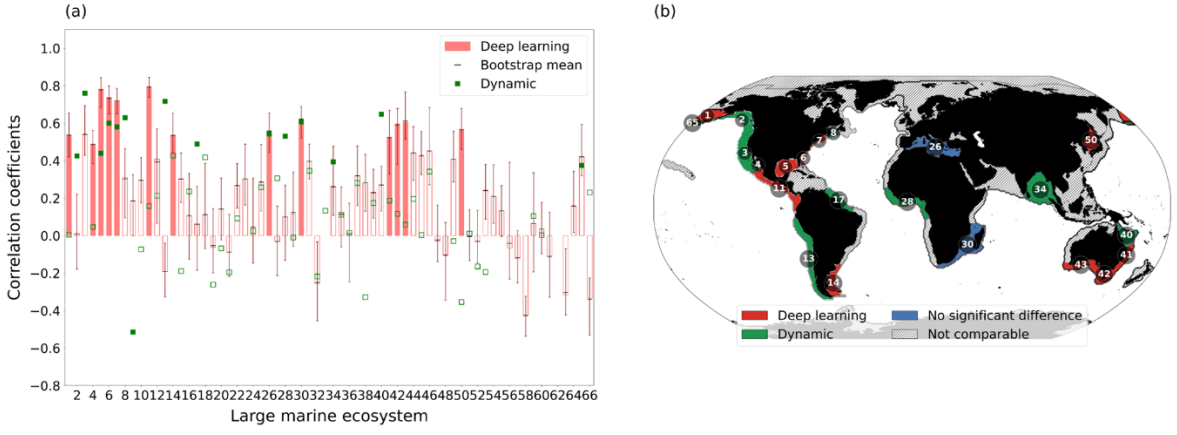


Figure B3: Comparison of chlorophyll prediction skill between deep learning and dynamic models across Large Marine Ecosystems (LMEs). (a) Correlation coefficients between satellite-observed and predicted annual mean chlorophyll anomalies at a 1-year lead time. Red bars show the deep learning model correlation; filled bars indicate significance at p < 0.10. Error bars show the 95% bootstrap confidence interval from a double bootstrap procedure accounting for both ensemble and temporal sampling

uncertainty, with black dashes indicating the bootstrap mean. Green markers show the dynamic model correlation (1998–2017); filled markers indicate significance at $p < 0.10$. (b) Map comparing prediction skill. Red shading indicates LMEs where the deep learning model significantly outperforms the dynamic model (bootstrap $p < 0.05$) or is the only model with significant skill. Green indicates the same for the dynamic model (bootstrap $p > 0.95$). Blue indicates LMEs where neither model significantly outperforms the other. Hatched regions indicate LMEs where both models lack significant skill or data are unavailable.

In figure 5, use labels (DL and dynamics) that are consistent throughout the paper and are clear; if figure 5a and 5b provide the same information, consider simplification and use only one, otherwise, clarify the distinction.

Following the reviewer's suggestion, we have revised the figure labels for consistency throughout the paper. Fig. 5a and 5b present complementary information that cannot be reduced to a single panel. Fig. 5a displays the correlation coefficients for all 66 LMEs with bootstrap confidence intervals, enabling a quantitative comparison of prediction skill between the deep learning and dynamical models across the full set of LMEs. Fig. 5b synthesizes this information geographically, mapping which model significantly outperforms the other based on a bootstrap statistical test, replacing the previous ad hoc correlation difference threshold (≥ 0.2). The revised figure is shown in Fig. B3 above.

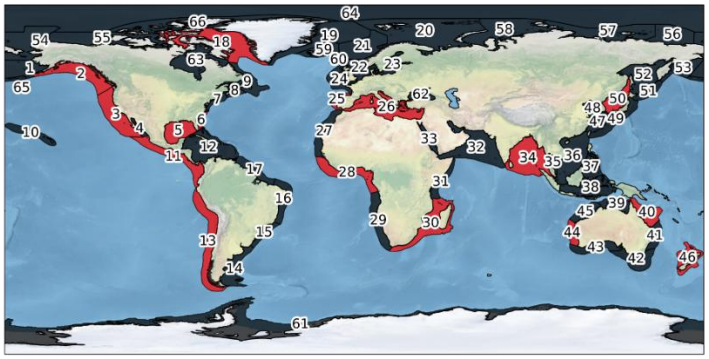
In Fig. 5a, the numbers of significant correlations are 15 for DL and 16 for dynamics. It appears to me to have quite poor performance results. Please reformulate L281-282.

While we acknowledge that the number of LMEs with significant skill appears limited, predicting LME-averaged chlorophyll anomalies from low-resolution ESM inputs represents a challenging problem. One of the key factors contributing to the successful prediction of LME-averaged chlorophyll anomalies is that both dynamic and deep learning models can capture the large-scale signals that drive chlorophyll variability in each LME region. The level of skill obtained here is consistent with that reported for a dynamical model (Park et al., 2019), which achieved significant skill in a comparable number of LMEs and has been recognized as a meaningful result in the field of marine biogeochemical prediction.

(L284): Some regions are listed as examples of comparable performance habits, but the Fig 5a does not show explicitly these regions. Indicating at which bars of the plot they correspond would increase the clearness of the results. Moreover, a map with the 66 LME is missing in the paper.

Following the reviewer's recommendation, we have added LME numbers in parentheses throughout Section 3.4 when referring to specific regions, enabling direct identification in Fig. 5a. LME numbers are also labeled in the categorical map (Fig. 5b) for all LMEs where at least one model shows significant skill. Additionally, we have added a labeled map of all 66 LMEs in the Supplementary Information for general reference. The labeled LME map is reproduced below for as Fig. B4.

(a)



(b)

1	East Bering Sea	18	Canadian Eastern Arctic	35	Gulf of Thailand	52	Sea of Okhotsk
2	Gulf of Alaska	19	East Greenland Shelf	36	South China Sea	53	West Bering Sea
3	California Current	20	Barents Sea	37	Sulu-Celebes Sea	54	Chukchi Sea
4	Gulf of California	21	Norwegian Shelf	38	Indonesian Sea	55	Beaufort Sea
5	Gulf of Mexico	22	North Sea	39	North Australian Shelf	56	East Siberian Sea
6	Southeast U.S. Continental Shelf	23	Baltic Sea	40	Northeast Australian Shelf	57	Laptev Sea
7	Northeast U.S. Continental Shelf	24	Celtic-Biscay Shelf	41	East-Central Australian Shelf	58	Kara Sea
8	Scotian Shelf	25	Iberian Coastal	42	Southeast Australian Shelf	59	Iceland Shelf
9	Newfoundland-Labrador Shelf	26	Mediterranean Sea	43	Southwest Australian Shelf	60	Faroe Plateau
10	Insular Pacific-Hawaiian	27	Canary Current	44	West-Central Australian Shelf	61	Antarctica
11	Pacific Central-American	28	Guinea Current	45	Northwest Australian Shelf	62	Black Sea
12	Caribbean Sea	29	Benguela Current	46	New Zealand Shelf	63	Hudson Bay
13	Humboldt Current	30	Agulhas Current	47	East China Sea	64	Arctic Ocean
14	Patagonian Shelf	31	Somali Coastal Current	48	Yellow Sea	65	Aleutian Islands
15	South Brazil Shelf	32	Arabian Sea	49	Kuroshio Current	66	Canadian High Arctic
16	East Brazil Shelf	33	Red Sea	50	East Sea		
17	North Brazil Shelf	34	Bay of Bengal	51	Oyashio Current		

Figure B4. a Global map of the 66 Large Marine Ecosystems (LMEs) examined in this study. Red shading indicates the 16 LMEs selected for sensitivity analysis, chosen to provide representative coverage across all major ocean basins while excluding polar regions where persistent data gaps limit reliable evaluation. Numbers correspond to LME identifiers referenced throughout the manuscript. b List of all 66 LMEs with corresponding identifiers.

To evaluate the validity of using NN chlorophyll predictions instead of observed chlorophyll data for fish catch prediction, it would be informative to include a comparison, for example with results obtained from a linear regression model using satellite chlorophyll observations. Alternatively, please clarify the reason for this methodological choice.

The rationale for this methodological choice is detailed in our response to the related comment on Section 2.4 above, where we explain that satellite chlorophyll is available only retrospectively and cannot serve as an operational forecast tool. The fish catch analysis is intended as an exploratory demonstration that CNN-derived chlorophyll forecasts retain sufficient environmental signal to explain interannual catch variability, rather than as a validated fisheries prediction system.

Figure 6 shows only 2 LME. Providing additional information about the correlation between chlorophyll and fish catch in the other LMEs could strengthen the results.

We have expanded Figure 6 in the revised manuscript. It now presents results for four LMEs with five species: South American pilchard (LME 11, lag=0), skipjack tuna (LME 11, lag=0), northern white shrimp (LME 6, lag=1), yellowfin tuna (LME 41, lag=1), and Japanese jack mackerel (LME 50, lag=1). These represent all species–LME combinations where statistically significant correlations were identified among the ten most frequently caught species in each LME, and for which supporting ecological literature could be identified. The revised Figure 6 is reproduced below as Fig. B5.

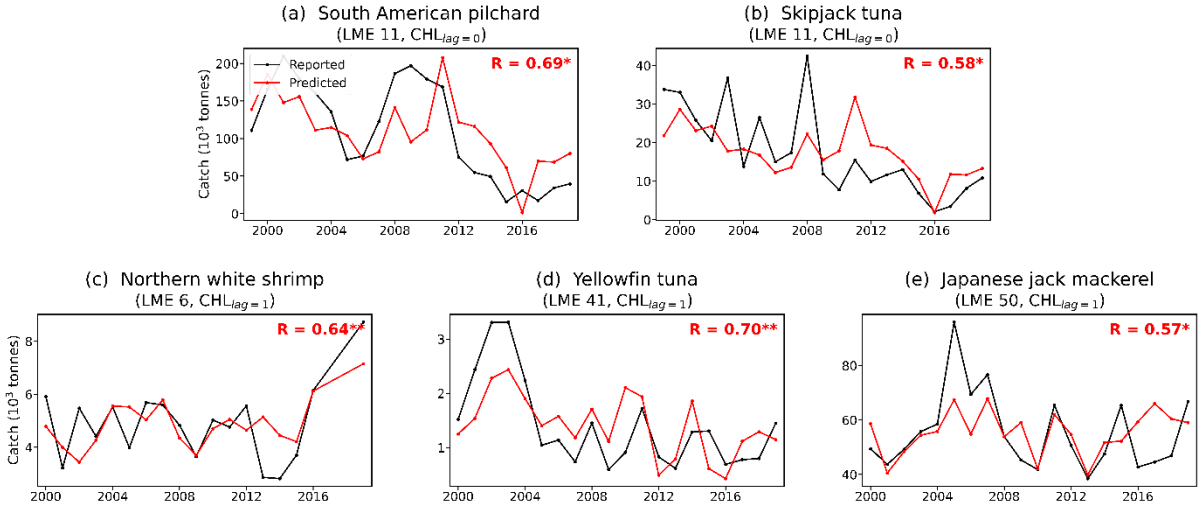


Figure B5: Prediction skill for annual fish catch of individual species in selected Large Marine Ecosystems (LMEs). a-e Time series of reported (black) and estimated (red) annual fish catch (tonnes). (a) South American pilchard, LME 11, lag=0. (b) Skipjack tuna, LME 11, lag=0. (c) Northern white shrimp, LME 6, lag=1. (d) Yellowfin tuna, LME 41, lag=1. (e) Japanese jack mackerel, LME 50, lag=1. Lag=0 and lag=1 indicate regression against CNN-predicted chlorophyll of the same year and the preceding year, respectively. Asterisks denote statistical significance (*p < 0.1, **p < 0.05).

Improve the quality and clarity of the figure 6: y axis is missing the label and unit and the text needs to be enlarged. Does y-axis represent the correlation coefficient between fish catch and chlorophyll anomalies, or the comparison between predicted and observed fish catch?

Following the reviewer's comment, we have revised Figure 6 by adding y-axis labels with units of annual fish catch in tonnes. Each panel shows time series of reported catch (black) and catch estimated from CNN-predicted chlorophyll anomalies via linear regression (red), with correlation coefficients and significance levels indicated. We note that the regression relationships are fitted over the entire analysis period and thus represent in-sample associations, consistent with the exploratory nature of this analysis. The revised figure is shown in Fig. B5 above.

DISCUSSION & CONCLUSION:

The conclusion and discussion section clearly summarizes strengths and limitations of the approach and the value of the sensitivity analysis. However, a few aspects could be better presented.

(L340): The phrase "while capturing physically interpretable signals underlying chlorophyll variability" could benefit from clarification. Since the CNN inputs are SST and chlorophyll anomalies, it would be helpful to specify whether this comment refers specifically to SST or to other physical signals. Providing this clarification would improve the reader's understanding of the model's interpretation.

We have revised this statement to specify that the physically interpretable signals refer primarily to ENSO-related SST patterns and their spatial teleconnections, as identified through SHAP attribution analysis (Section 3.2, Fig. 4b,d; see also Fig. B1, provided in our response to the comment on L215-220 above). The revised Discussion reads as follows:

"Further analysis of monthly prediction skill in two representative LMEs, selected a priori based on well-documented connections to large-scale climate variability, revealed that this skill arises from physically interpretable signals, including ENSO-driven SST variability and wintertime reemergence mechanisms, suggesting that statistical learning can internalize aspects of coupled physical-biogeochemical dynamics from training data."

(L340): The statement that "the model successfully reproduces the known ocean–climate process" could benefit from further elaboration. Providing a brief explanation of which specific ocean–climate processes this sentence refers to would help strengthen the interpretation of the results and improve clarity for the reader.

As the reviewer mentioned, we have elaborated on which specific ocean–climate processes are referred to, as reflected in the revised text quoted above. These include the ENSO spring predictability barrier, ENSO teleconnection patterns, and the winter-to-winter reemergence signal (see also Fig. B1, provided in our response to the comment on L215-220 above).

(L362): The statement that "sensitivity tests show that surface chlorophyll anomalies captured subsurface variability" would benefit from further clarification. From the manuscript, it appears that the sensitivity analysis was primarily performed to optimize the network architecture and input data. It is therefore not immediately clear how this analysis supports the conclusion regarding subsurface variability. Providing a more detailed explanation of the connection, or the underlying correlations, would help the reader better understand the interpretation of the proposed results.

We have revised this statement to clarify the reasoning. The sensitivity analysis showed that models using surface chlorophyll as input achieved comparable or higher prediction skill than models using subsurface temperature (0–300 m average), suggesting that surface chlorophyll anomalies encode information about subsurface ocean states through the physical linkage between nutrient supply, vertical mixing, and phytoplankton growth. This finding is consistent with prior studies showing that surface chlorophyll carries an imprint of subsurface ocean dynamics through nutrient supply pathways, contributing to improved predictability of ocean biogeochemistry (Park et al., 2018a; Lim et al., 2022; Lee et al., 2024). The revised Discussion reads as follows:

"In particular, models using surface chlorophyll as input achieved comparable or higher prediction skill than models using subsurface temperature (0–300 m average), suggesting that surface chlorophyll anomalies encode information about

subsurface ocean states through the physical linkage between nutrient supply, vertical mixing, and phytoplankton growth (Park et al., 2018a; Lim et al., 2022; Lee et al., 2024).”