Responses to Reviewer #1's Comments

Summary

This paper develops a CNN to predict annual and monthly mean chlorophyll concentrations within Large Marine Ecosytems using surface chlorophyll and SST from the previous three months as predictors, with training primarily based on Earth System Model simulations and reanalysis. Satellite-derived chlorophyll is used for evaluation. For monthly predictions, the authors evaluate lead times from 1 to 24 months.

Overall, I found this paper interesting and potentially useful to the community. The approach of using a data-driven framework for chlorophyll prediction is timely. However, I believe that the manuscript requires substantial revision before publication. In particular, the methods section lacks sufficient detail and clarity to be fully understood and reproduced. The assessment of model skill would be strengthened by comparison to simple baselines such persistence in addition to dynamical forecasts. I would also like to see a more explicit discussion of the limitations of training a CNN on modeled data and how these limitations may affect real-world applicability. Finally, I found that the fish catch prediction section did not convincingly demonstrate utility for marine resource management.

We thank Reviewer #1 for the thorough and constructive comments. Below, we address each comment in the order presented.

Major comments

1) The methods section requires more detail to be understandable and reproducible. While the manuscript describes the data sources and general temporal coverage, key implementation details are ambiguous (see specific comments below). For example, explicitly stating the effective number of training and testing samples would improve transparency.

We have substantially revised the methods section to improve clarity and reproducibility, including explicitly stating the effective number of training and testing samples. 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. Detailed responses to each related comment are provided below in the corresponding sections. As one example of these revisions, the effective sample sizes are now stated in Section 2.1.3:

"The model was trained on CMIP6 historical and preindustrial control simulations (16 models) combined with GFDL-ECDA reanalysis (Park et al., 2018b), totalling 8,013 samples. A subset of CMIP6 simulations and reanalysis data (2,043 samples) is held out for validation during training to monitor convergence and prevent overfitting. For sensitivity experiments (Section 3.1), model performance is evaluated on GFDL-ECDA reanalysis (1998-2017), independent from the

training period. This validation step informed the selection of the reference model configuration. Final model evaluation uses satellite-derived chlorophyll from SeaWiFS and MODIS (1998-2021), fully independent from all model development data."

2) I think that the approach of training a CNN on model data needs stronger justification. I agree with the authors that Earth system models have large uncertainties due to parameterizations, spatial resolution, etc., which make prediction challenging. However, it is not clear how the deep learning approach mitigates these uncertainties when the training data themselves reflect ESM biases. Maybe if training on multiple models, this concern is reduced, but I would appreciate a clear statement on this. The main advantage I see to using the CNN over dynamical forecasts is the greater computational efficiency, which was only briefly mentioned. It would be helpful to include a discussion of how training the CNN on modeled data may limit the applicability to the real world.

We appreciate this important concern. Training on ESM data inevitably introduces model-specific biases into the learning process. Our framework uses 16 CMIP6 models with diverse model physics, biogeochemical parameterizations, and climate sensitivities, which may help the CNN capture physical–biogeochemical relationships that are more broadly consistent across models rather than specific to any single model. Guo et al. (2025) similarly trained on CMIP6 multi-model ensembles and demonstrated that the deep learning model can generalize beyond the limitations of individual models. We acknowledge, however, that we did not systematically test how model subset selection affects prediction skill, and that biases shared across CMIP6 models (e.g., limited representation of coastal processes) may still propagate to CNN predictions.

Regarding computational efficiency, we agree this is a key advantage that was insufficiently emphasized. Once trained, the CNN produces forecasts in seconds, compared to the thousands of simulation years required for dynamical retrospective forecasts. This enables rapid generation of large ensembles and facilitates operational applications. We have expanded this discussion in the revised manuscript as follows:

"A key practical advantage of the deep learning approach is computational efficiency. Once trained, the CNN produces forecasts in seconds, compared to the thousands of simulation years required for dynamical retrospective forecasts (e.g., Park et al., 2019). This enables rapid generation of large ensembles and facilitates operational applications where timely forecast delivery is essential."

We also note that training on ESM data creates an inherent ceiling on CNN performance tied to the fidelity of the training simulations. To partially mitigate this limitation, we incorporated the GFDL-ECDA reanalysis, which assimilates observational constraints into the physical ocean state. As shown in Figure 2, excluding the reanalysis and training on CMIP6 models alone resulted in modestly lower prediction skill, suggesting that observationally-constrained training data helps anchor the CNN to more realistic physical–biogeochemical relationships. In addition, this implies that as ESMs continue to improve across successive generations, the quality of deep learning predictions trained on these outputs can be

expected to improve correspondingly. Séférian et al. (2020) demonstrated that the representation of marine biogeochemistry—including chlorophyll, nutrients, and air–sea CO_2 fluxes—has progressed from CMIP5 to CMIP6, with reduced model–observation biases for several key variables. As such improvements continue in future generations (e.g., CMIP7), the fidelity of training data available to data-driven frameworks like ours is also expected to increase, potentially leading to further gains in prediction skill. We have added this discussion of limitations and future prospects in the revised manuscript as follows:

"Training on CMIP6 simulations creates an inherent ceiling on CNN performance tied to the fidelity of the training data. Training on diverse multi-model ensembles has been shown to improve generalization beyond the limitations of individual models in similar deep learning frameworks (Guo et al., 2025). Building on this principle, our multi-model training strategy (16 CMIP6 models) was designed to leverage the diversity of model physics and biogeochemical parameterizations across the ensemble, with the expectation that this reduces sensitivity to the biases of any individual model. We additionally incorporated the GFDL-ECDA reanalysis, which assimilates observational constraints into the physical ocean state. As demonstrated in the sensitivity experiments (Section 3.1), excluding the reanalysis and training on CMIP6 models alone resulted in modestly lower prediction skill, suggesting that observationally-constrained training data helps anchor the CNN to more realistic physical–biogeochemical relationships. Nevertheless, biases shared across the CMIP6 ensemble, such as limited representation of coastal processes and common biogeochemical parameterization assumptions, may still propagate to CNN predictions, and the forecasts should be interpreted with this limitation in mind. As ESMs continue to improve across successive generations, with documented progress in marine biogeochemistry from CMIP5 to CMIP6 (Séférian et al., 2020), such biases are expected to diminish, offering a pathway toward further gains in prediction skill for data-driven frameworks like ours."

3) The paper would benefit from discussing uncertainties related to studying chlorophyll in LMEs. Low-resolution ESMs do not resolve coastal processes well. There are also large uncertainties in satellite observations of chlorophyll in coastal waters. Additionally, there is huge spatial variability of chlorophyll within LMEs, which limits the applicability to marine resource management. These caveats and room for future work should be clearly articulated.

Following the reviewer's comment, we have expanded the discussion in the revised manuscript to address these caveats more explicitly. Regarding spatial resolution, we recognize that our 1°×1° input data do not resolve fine-scale coastal processes such as submesoscale upwelling, river plume dynamics, and nearshore bathymetric effects. However, as demonstrated by Stock et al. (2015), prediction skill at coastal scales can arise when signals from large-scale processes resolved by the model are strong enough to emerge from noisier local signals. Our results are consistent with this finding, as the CNN achieves significant prediction skill in LMEs where large-scale climate variability dominates chlorophyll variability.

We also acknowledge that satellite-derived chlorophyll observations carry substantial uncertainties in coastal waters due to the optical complexity of these environments. Furthermore, the clear-sky sampling bias of satellite observations introduces an inconsistency with the all-sky ESM training data — a discrepancy that our unified masking strategy mitigates but does not fully eliminate. These limitations may contribute to the reduced prediction skill observed in coastal-dominated LMEs such as eastern boundary upwelling systems. We have acknowledged these caveats in the revised manuscript as follows:

"Additionally, satellite-derived chlorophyll observations carry substantial uncertainties in coastal waters due to optical complexity, and the clear-sky sampling bias of satellite observations introduces an inconsistency with the all-sky ESM training data that our unified masking strategy mitigates but does not fully eliminate."

Finally, the reviewer correctly notes that there is large spatial variability of chlorophyll within LMEs, which limits direct applicability to fine-scale marine resource management. Our LME-mean predictions are most informative for basin-scale environmental conditions rather than localized ecosystem responses. This is also discussed in the revised manuscript as follows:

"These factors, combined with large spatial variability of chlorophyll within LMEs, mean that our LME-mean predictions are most informative for basin-scale environmental conditions rather than localized ecosystem responses."

4) While benchmarking with a dynamic model is a good approach, I believe that this paper would be much stronger if the predictions were also benchmarked against climatological means or persistence. Given the strong autocorrelation of chlorophyll anomalies, it is difficult to assess the added value of the CNN without these comparisons.

We agree that benchmarking against persistence is essential for assessing the added value of the CNN. Following the reviewer's comment, we have added persistence forecasts as a baseline in Figure 4a and 4c (Fig. A1 below). Green open circles indicate combinations where the CNN outperforms persistence. For the Pacific Central-American Coastal LME (LME 11), the CNN consistently outperforms persistence across most forecast start months at lead times up to ~12 months, with boreal winter initializations maintaining skill advantages extending beyond 12-month leads. For the Agulhas Current LME (LME 30), the CNN outperforms persistence primarily at shorter lead times. These results suggest that the CNN provides added value beyond simple autocorrelation, particularly at the seasonal-to-annual timescales targeted in this study.

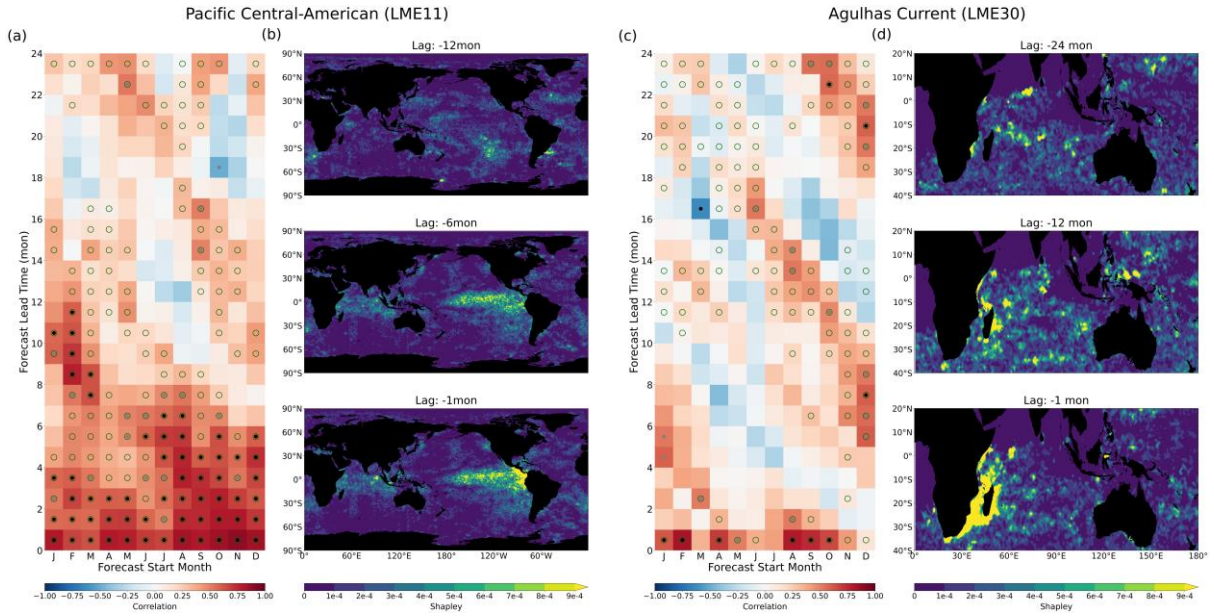


Figure A1: Monthly prediction and mechanism underlying chlorophyll prediction skill. a,c Anomaly correlation coefficient between predicted and satellite-observed 3-month running mean chlorophyll anomalies (LME-averaged) as a function of forecast start month (x-axis) and lead time (y-axis). Black dots 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).

5) I am not convinced that the forecasts presented in Section 3.5 are currently useful for marine resource management. The analysis appears exploratory, with species, LMEs, lag times, and significance thresholds selected in a way that risks cherry-picking statistically significant relationships. Given the large number of combinations explored, it is expected that some relationships will appear significant at the 90% confidence level by chance alone. A more systematic approach is needed. Possible alternatives include focusing on total catch (if available), providing a clear justification for the LMEs and species examined, or targeting regions where fisheries collapses have plausibly been linked to environmental variability. Finally, the authors must acknowledge a major caveat of the fish catch dataset: reported catch depends strongly on fishing effort, management, and reporting practices, not solely on environmental conditions.

We appreciate the reviewer's critical assessment of Section 3.5 and agree that the original framing could be improved. We have revised the manuscript to frame the fish catch analysis as an exploratory demonstration of potential linkages between CNN-predicted chlorophyll anomalies and marine resource variability, rather than a validated fisheries prediction tool. This section is intended to illustrate a possible downstream application of our chlorophyll forecasting framework, consistent with the approach taken in previous seasonal prediction studies (e.g., Park et al., 2019; Tommasi et al., 2017).

Regarding species selection, we recognize that testing multiple combinations of species, LMEs, and lag times increases the risk of identifying spurious relationships. However, a large body of literature has demonstrated that interannual variability in catches of tunas, small pelagic fish, and commercially important invertebrates is strongly modulated by climate modes such as ENSO and IOD through bottom-up forcing pathways (Lehodey et al., 1997; 2006). Our selection process was therefore not purely statistical but rather hypothesis-driven. For each LME where the deep learning model demonstrated significant chlorophyll prediction skill, the ten most frequently caught species were identified and tested via linear regression. The species presented in Figure 6 are those that showed statistically significant correlations among these candidates and for which supporting ecological literature could be identified. We have clarified the selection procedure in the revised manuscript as follows:

"Species–LME combinations were selected based on two conditions: significant CNN chlorophyll prediction skill in the LME, and a statistically significant correlation between predicted chlorophyll and catch anomalies for species with a plausible bottom-up forcing mechanism suggested by ecological literature."

The ecological basis for each species-LME pairing is as follows. Small pelagic fish and tuna in LME 11 respond sensitively to productivity fluctuations driven by ENSO-related convergence zone shifts (Lehodey et al., 1997; Kim et al., 2020), and the lag=0 relationship is consistent with the rapid trophic response of these short-lived or migratory species to concurrent productivity conditions. We note that the correlation for skipjack tuna ($R = 0.58$, $p < 0.1$) is suggestive rather than strongly significant, though the ecological mechanism is well established. Northern white shrimp in LME 6 have annual life cycles, and their abundance is directly linked to prior-year environmental conditions (Diop et al., 2007), making the lag=1 relationship consistent with this recruitment mechanism. Yellowfin tuna in LME 41 preferentially inhabit regions of high primary productivity where epipelagic prey are concentrated near the surface mixed layer and thermocline (Lehodey et al., 1997), and their distribution off eastern Australia is closely linked to productivity and eddy dynamics of the East Australian Current system (Young et al., 2011), with the lag=1 relationship consistent with prior-year productivity conditions influencing available prey fields. Japanese jack mackerel in LME 50 show that larval growth rates are modulated by chlorophyll-mediated prey availability (Takahashi et al., 2016; 2022), and the lag=1 relationship reflects this early life stage sensitivity, though as with skipjack tuna the correlation ($R = 0.57$, $p < 0.1$) is suggestive rather than definitive.

Regarding the operational utility of lag=0 relationships, it is important to clarify that the chlorophyll values used in the lag=0 regression are not observed annual means but CNN-predicted annual mean anomalies. The CNN takes NDJ (November of Year 0 – January of Year 1) satellite observations as input and predicts the annual mean chlorophyll anomaly for Year 1, issued at the beginning of Year 1 well before annual catch data are compiled. The lag=0 relationship therefore still provides operationally useful anticipatory information, and annual-scale environmental predictions are directly relevant to fisheries management given that total allowable catch (TAC) quotas are typically set on an annual basis (Tommasi et al., 2017).

We also acknowledge that our analysis focuses exclusively on bottom-up environmental forcing and does not account for top-down effects such as fishing effort, management interventions, fleet behaviour, and reporting practices, all of which strongly influence reported catch data. Reported catch data were used without reconstruction or correction, consistent with previous studies linking environmental variability to fisheries (e.g., Park et al., 2019), though we acknowledge inherent limitations including underreporting and discards. We further note that the regression relationships shown in Figure 6 represent in-sample associations fitted over the entire analysis period rather than out-of-sample forecasts, consistent with the exploratory nature of this analysis, which aims to demonstrate the potential relevance of CNN-predicted chlorophyll to fisheries variability rather than to provide a validated operational prediction system. Developing robust cross-validated prediction frameworks would be a valuable direction for future work. These caveats have been noted in the revised manuscript as follows:

"While this structured selection reduces the risk of purely spurious associations, the analysis relies exclusively on bottom-up environmental forcing and does not account for top-down effects including fishing effort, management interventions, fleet behavior, and reporting practices, all of which strongly influence reported catch data. 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. Although these relationships were identified in only a subset of LMEs, they demonstrate the feasibility of integrating environmental forecasts into fisheries applications. Such applications will require careful consideration of species-specific ecological mechanisms and regional oceanographic contexts. Developing robust cross-validated prediction frameworks and incorporating additional biogeochemical variables such as NPP or trophodynamic processes would be valuable directions for future work."

References

Lehodey, P., Alheit, J., Barange, M., Baumgartner, T., Beaugrand, G., Drinkwater, K., Fromentin, J.-M., Hare, S., Ottersen, G., Perry, R. I., Roy, C., van der Lingen, C. D., and Werner, F.: Climate variability, fish and fisheries, J. Climate, 19, 5009–5030, https://doi.org/10.1175/JCLI3898.1, 2006.

Minor comments

Abstract

Consider specifying the prediction timescales and lead times in the abstract.

We have revised the abstract to specify that our framework targets monthly to annual chlorophyll prediction with lead times up to two years, as follows:

"Here, we develop a deep learning–based prediction system to forecast surface chlorophyll concentrations across all Large Marine Ecosystems (LMEs) at monthly to annual timescales with lead times up to two years."

Introduction

I think that the need for prediction using deep learning could be more strongly motivated in the introduction.

We have strengthened the motivation for using deep learning in the introduction section. Specifically, we elaborated on how data-driven models can overcome key limitations of ESM-based approaches, as reflected in the revised introduction:

"These constraints have highlighted the need for alternative methodologies that can provide skillful biogeochemical forecasts at the scale of LMEs with greater computational efficiency."

"Deep learning has emerged as a promising alternative for predicting marine biogeochemical variability. 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)."

It would be very helpful to clarify lead times and averaging windows in the introduction. What scales are relevant for ecosystem management?

Our prediction framework targets monthly to annual timescales with lead times of 1–24 months, which align with the temporal scales at which key marine resource management decisions (Stock et al., 2015; Tommasi et al., 2017). We have added this clarification in the introduction as follows:

"our framework takes a purely data-driven approach, ingesting three consecutive months of global sea surface temperature and chlorophyll anomalies to produce monthly or annual chlorophyll forecasts at the LME scale with lead times of 1–24 months, aligning with the temporal scales relevant for marine resource management decisions including seasonal quota setting, harvest control adjustments, and interannual stock assessment planning (Stock et al., 2015; Tommasi et al., 2017)."

Line 37: I'm not sure if it's fair to say that the performance for biogeochemical variables remains limited. There are papers showing skillful predictions of NPP, DIC, and other biogeochemical variables. Please see citations below:

Mogen, S. C., Lovenduski, N. S., Yeager,S., Keppler, L., Sharp, J., Bograd, S.J., et al. (2023). Skillful multi-monthpredictions of ecosystem stressors in thesurface and subsurface ocean. Earth'sFuture, 11, e2023EF003605. https://doi.org/10.1029/2023EF003605

Ilyina, T., Li, H., Spring, A., Müller, W. A., Bopp, L., Chikamoto, M. O., et al. (2021). Predictable variations of the carbon sinks and atmospheric CO2 growth in a multi-model framework. Geophysical Research Letters, 48, e2020GL090695. https://doi.org/10.1029/2020GL090695

Krumhardt, K. M., Lovenduski, N. S., Long, M. C., Luo, J. Y., Lindsay, K., Yeager, S., & Harrison, C. (2020). Potential predictability of net primary production in the ocean. Global Biogeochemical Cycles, 34, e2020GB006531. https://doi.org/10.1029/2020GB006531

Brady, R.X., Lovenduski, N.S., Yeager, S.G. et al. Skillful multiyear predictions of ocean acidification in the California Current System. Nat Commun 11, 2166 (2020). https://doi.org/10.1038/s41467-020-15722-x

We agree that our original phrasing understated recent progress in ESM-based biogeochemical prediction. We have revised the Introduction to explicitly acknowledge these advances, incorporating all references suggested by the reviewer. The revised text now reads:

"recent advances have further shown prediction skill for biogeochemical variables including net primary production (Krumhardt et al., 2020), ocean carbon fluxes (Ilyina et al., 2021), ocean acidification (Brady et al., 2020), ecosystem stressors (Mogen et al., 2023),"

Methods

Line 96: Why not use a gap-filled data product or one that merges more instruments, like OC-CCI or GlobColour? Can you please also clarify how MODIS/SeaWIFS data were accessed and processed?

We used SeaWiFS and MODIS chlorophyll products for two reasons. First, both sensors share a consistent ocean color retrieval framework, ensuring temporal homogeneity across the 1998–2021 record. Second, the ESM-based dynamical forecast system against which we compare our predictions (Park et al., 2019) used the same satellite products, enabling a direct and fair comparison. While merged products such as OC-CCI and GlobColour offer broader coverage, their inter-sensor merging procedures can introduce discontinuities that affect trend estimates and their uncertainties (Hammond et al., 2018). Data were obtained from NASA's Ocean Color Web (oceancolor.gsfc.nasa.gov) as monthly level-3 binned products at 9 km resolution. Following standard practice (Campbell, 1995), the median value within each target grid cell was used during interpolation to account for the lognormal distribution of chlorophyll concentration. The revised Section 2.2 reads as follows.

"Satellite monthly surface chlorophyll-a concentrations were obtained from the SeaWiFS and MODIS ocean color sensors (Esaias et al., 1998; Mcclain, 1998), and sea surface temperature (SST) data were from NOAA's optimally interpolated SST version 2 (OISSTv2) dataset based on the Advanced Very High Resolution Radiometer (AVHRR) (Reynolds et al., 2007). The original chlorophyll and SST data were provided at monthly resolution with fine spatial scales (0.25 degrees for SST and 9 km × 9 km for chlorophyll). For consistency and computational efficiency in deep learning applications, all observational data spanning 1998 to 2021 were interpolated onto a 1° × 1° regular global grid. Following standard practice (Campbell, 1995), the median value within each grid cell was used during spatial interpolation of chlorophyll to account for the lognormal distribution of chlorophyll concentration."

References

Hammond, M. L., Beaulieu, C., Henson, S. A., and Sahu, S. K.: Assessing the presence of discontinuities in the ocean color satellite record and their effects on chlorophyll trends and their uncertainties, Geophys. Res. Lett., 45, 7654–7662, https://doi.org/10.1029/2017GL076928, 2018.

Line 103: I am confused by the zero-filling strategy applied here. While this seams reasonable for polar night regions, where chlorophyll concentration is nearly zero, how can you justify filling in grid cells obscured by clouds with zero? Please clarify this section.

We have substantially revised Section 2.2 to clarify our preprocessing strategy. Rather than filling cloud-obscured grid cells with zero on a per-timestep basis, 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. The same unified mask is applied to simulated CMIP6 chlorophyll fields, ensuring that the spatial domain used for training is identical to that used for evaluation. The revised text reads as follows:

"To ensure spatial consistency across all datasets, we constructed a unified binary mask from the satellite record: any grid cell containing a missing value in any single month during the entire satellite period (1998–2021) was permanently flagged. All flagged grid cells were set to zero across all time steps. The mask itself was not provided as an explicit input channel to the model. The consistently zero-valued regions largely correspond to land-adjacent, polar, or persistently cloud-covered areas where chlorophyll signals are typically absent or negligible, reducing the likelihood that zero-filling introduces spurious learning signals. Land grid cells are also represented as zero in the input fields. 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."

"The same unified mask derived from satellite observations was applied to simulated chlorophyll fields, with masked grid cells set to zero, ensuring that the spatial domain used for training is identical to that used for evaluation."

It may be worth mentioning that the satellite-derived chlorophyll data is biased by selective sampling in clear sky conditions, while the ESM-based training data is not. This may further complicate the applicability of the ESM-trained CNN to real-world predictions.

We have added this point in the revised Discussion, noting the clear-sky sampling bias of satellite observations and discussing how our unified masking strategy helps mitigate—but does not fully eliminate—this inconsistency, as follows:

"Additionally, satellite-derived chlorophyll observations carry substantial uncertainties in coastal waters due to optical complexity, and the clear-sky sampling bias of satellite observations introduces an inconsistency with the all-sky ESM training data that our unified masking strategy mitigates but does not fully eliminate."

Line 112: Please clarify the gap-filling treatment applied to simulated chlorophyll. Are observational data gaps being imposed on the model output? If so, how is physical consistency ensured?

Yes, the same unified binary mask derived from satellite observations is applied to simulated CMIP6 chlorophyll fields, with masked grid cells set to zero. This ensures that the spatial domain used for training is identical to that used for evaluation. Physical consistency of the simulated fields is preserved in non-masked regions, as the CMIP6 output is used without modification beyond regridding to 1°×1° resolution. As described in the revised Section 2.2 in our response to Line 103 above.

Figure 1. Typo in SeaWIFTS

We have updated our figure 1. Thanks for the correction.

Results

Line 154: Did you experiment with different input time windows? Why 3 consecutive months?

The 3-month consecutive input design follows the approach of Ham et al. (2019), who demonstrated its effectiveness for capturing temporal evolution in multi-month climate predictions. While Ham et al. used SST and heat content for ENSO forecasting, we adapt this framework using SST and chlorophyll as key surface ocean states relevant to marine ecosystem prediction. To verify this choice, we additionally tested 1-month and 6-month input windows (Fig. A2). The 1-month window yielded substantially lower skill (2/66 LMEs significant), confirming that temporal context is necessary for capturing evolving climate signals. The 6-month window also showed lower skill (6/66 significant) compared to the 3-month configuration (13/66 significant), suggesting that extending the input window beyond three months does not improve prediction and may introduce noise that dilutes the most recent and predictively relevant signals. The 3-month window was therefore retained as the optimal choice.

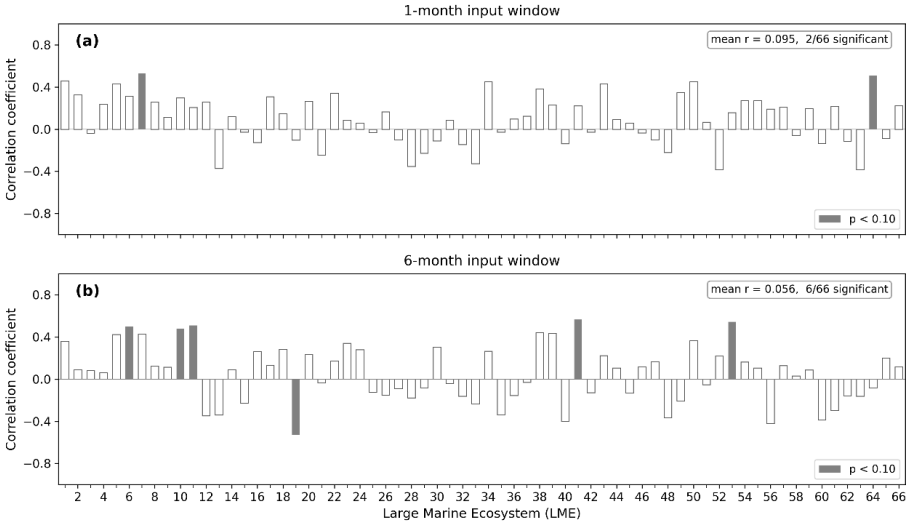
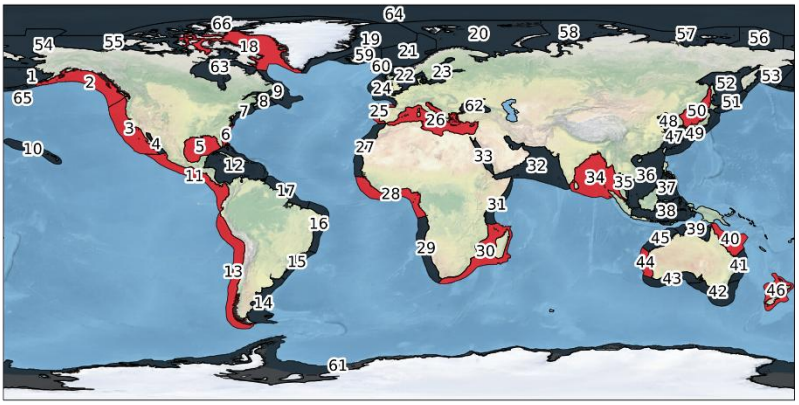


Figure A2. Prediction skill (correlation coefficient between CNN-predicted and satellite-observed chlorophyll anomalies) across all 66 LMEs for (a) 1-month and (b) 6-month input window configurations. Dark grey bars indicate statistically significant skill at p < 0.10. The 1-month window yields significant skill in only 2/66 LMEs (mean r = 0.095), and the 6-month window in 6/66 LMEs (mean r = 0.056), both substantially lower than the 3-month configuration (13/66 significant), supporting the selection of the 3-month input window as the optimal choice.

Line 156: Why were 16 LMEs selected?

The 16 LMEs were selected to provide representative coverage across all major ocean basins while excluding polar regions where persistent data gaps limit reliable evaluation, as illustrated in Figure A3. Additionally, training separate CNN models for each LME is computationally efficient but scales with the number of LMEs; focusing on 16 representative regions allowed thorough sensitivity analysis and model optimization while maintaining computational feasibility.

(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 A3. 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.

It would be extremely helpful to include a labelled map of all LMEs. Perhaps in supplemental information?

Following the reviewer's suggestion, we have added a labelled map of all 66 LMEs in the supplementary information (Figure A3). The map also indicates the 16 LMEs selected for sensitivity analysis with red shading.

Line 173: I would consider revising this sentence. While the inclusion of surface chlorophyll as a predictor improves forecast skill, this may largely reflect the persistence and autocorrelation of chlorophyll anomalies rather than the capture of nonlinear ecological signals.

We acknowledge that chlorophyll autocorrelation likely contributes to prediction skill, particularly at short lead times, and have revised this sentence accordingly. However, the persistence comparison added in Figure 4a,c directly tests this concern: the CNN outperforms persistence across many forecast start months and lead times, suggesting that the model provides skill beyond what autocorrelation alone can sustain, particularly at seasonal-to-annual leads.

Line 194: I find the term “initialized” confusing here, as it implies a dynamical forecasting system. Additionally, were different months tested? I wonder if northern and southern hemispheres would benefit from different input months.

We have revised the phrasing to avoid the implication of a dynamical forecasting system, replacing "initialized" with language that more clearly reflects our data-driven approach. The revised text reads as follows:

“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.”

Regarding different forecast start months, our monthly prediction (Section 3.2, Figs. 4a and 4c) already demonstrates that prediction skill varies substantially depending on the forecast start month, with skill patterns differing across LMEs in ways that reflect regional climate dynamics (e.g., the ENSO spring predictability barrier). This suggests that optimal forecast start timing is indeed region-dependent. For the annual prediction, we acknowledge that the current NDJ start and January–December target window is oriented toward boreal seasons, and southern hemisphere LMEs may benefit from alternative forecast start months and annual mean definitions (e.g., July–June). Systematically optimizing forecast start timing for each hemisphere and LME is a valuable direction for future work.

Line 195: Can you more clearly state here that the forecasts were initialized using real-world observations of SST and chlorophyll for the previous 3 months?

Yes. Following the reviewer’s suggestion, we have revised Section 3.2 to explicitly state that forecasts were generated by providing the model with satellite SST and chlorophyll from three consecutive months in early boreal winter (November to

January), with the model predicting the following calendar year. The revised text is provided in our response to Line 194 above.

Figure 3: I suggest revising this figure. I found it hard to see the stars and a bit blurry when I tried to zoom in.

We have revised Figure 3 with improved resolution and visibility. The revised figure is shown below (Figure A4):

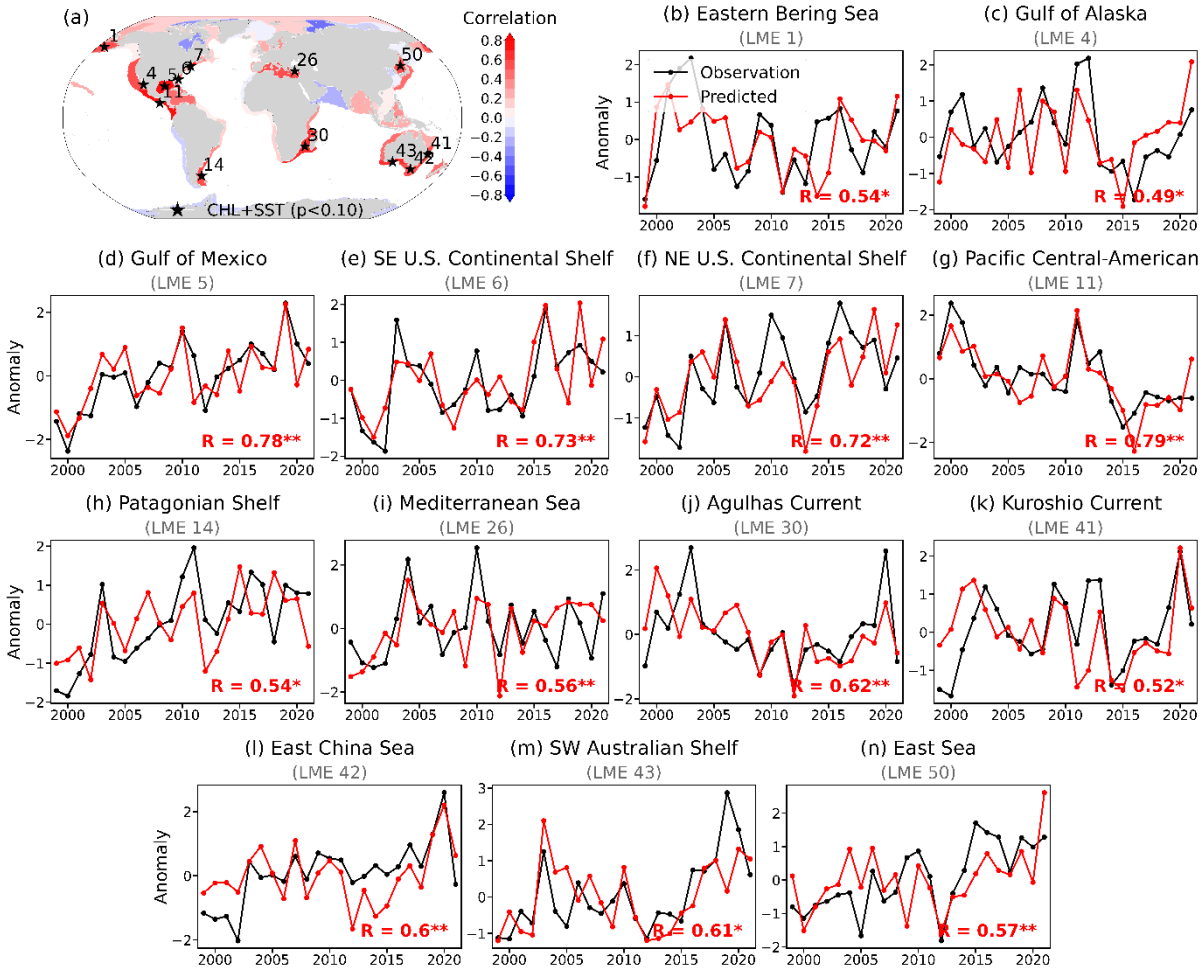


Figure A4: Chlorophyll prediction skill across Large Marine Ecosystems (LMEs). a Correlation coefficients between LME-averaged satellite-derived and predicted annual mean chlorophyll anomalies (1998-2021). The model takes November(Year 0)–December(Year 0)–January(Year 1) satellite observations as input 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).

I would appreciate some clarification on the lead time and time averaging. In the caption, the authors say that the forecast lead time is 1 year. To me, this implies predicting values one year out, not the annual mean values of the upcoming year. It's also not clear which months these annual means include. Do they include the month of January for that year, which is also used as input data?

The model uses November(Year 0)–December(Year 0)–January(Year 1) as input and predicts the annual mean chlorophyll anomaly averaged over January–December of Year 1. January of Year 1 is therefore included in both the input and the prediction target. We have revised the figure caption and Section 3.2 to clarify the temporal definitions of input and prediction windows, as shown in Fig. A4 above. The revised caption reads as follows:

"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."

Line 214: Similar to my previous comment: I'm not sure that it's fair to say this unless you were using a different biogeochemical variable as input to predict chlorophyll. This is why benchmarking with a persistence forecast would be valuable.

As we responded above, this concern is directly addressed by the persistence baseline added in Figure 4a,c (Fig. A1). If the CNN were simply replicating chlorophyll autocorrelation, it would not outperform persistence. The results suggest that the CNN adds value beyond persistence at many lead times, particularly for LME 11 where skill advantages extend beyond 12-month leads where autocorrelation is minimal.

Line 217: A map of LME 11 and LME 30 would be helpful.

As shown in the supplementary information (Fig. A3), we have added maps of LME 11 and LME 30.

Line 217: It would be helpful for the authors to explicitly state how ensemble members are generated in the CNN framework, as the term "ensemble" may otherwise be interpreted in a dynamical modeling sense. "Extending the forecast up to 24 months" also makes it sound like the CNN in stepping forward in time, when I assume that separate models are trained for each lead time. I would appreciate a clearer explanation of what information the CNN is using at these longer horizons.

We have clarified both points in the revised manuscript. Separate CNN models are trained for each combination of forecast start month and lead time, with each model directly predicting the target chlorophyll anomaly from a three-month input window of preceding input variables. This approach differs fundamentally from dynamical forecast systems, which step forward in time and can accumulate error across lead times. Ensemble members are generated by training models with different random weight initializations, reflecting uncertainty in the optimization landscape rather than uncertainty in initial ocean states as in dynamical model ensemble systems. As clarified in Section 2.1.2:

"Separate CNN models sharing the same architecture are trained for each combination of target LME, forecast type (annual or monthly mean), and lead time. For annual forecasts, models predict the annual mean chlorophyll anomaly for the

target LME in the year following the forecast start. Input data consist of three consecutive months of global SST and chlorophyll anomaly fields during boreal winter (November–December–January), gridded at 1° × 1° resolution (360 longitude × 180 latitude) and represented as six input channels. For monthly forecasts, separate CNN models are trained with different forecast start months and lead times (1–24 months ahead). Each model directly predicts a 3-month mean chlorophyll anomaly centered on the targeted month from a three-month window of preceding input variables."

Figure 4: It is interesting that the SHAP values are high in locations that are far away from the LMEs. Has the model learned a teleconnection? I think more discussion on this is needed.

Yes. We agree that the remote SHAP attribution patterns are noteworthy. The high SHAP values in regions distant from the target LME indicate that the CNN has learned to exploit large-scale climate signals for chlorophyll prediction, and we believe this is a key reason why CNN model can provide skillful prediction of LME-scale chlorophyll anomalies. In the revised Section 3.3, we discuss these patterns in detail: SHAP attribution maps reveal that the model utilizes equatorial Pacific SST patterns consistent with ENSO evolution for the Pacific Central-American Coastal LME (Fig. 4b), and remote Indian Ocean signals consistent with westward-propagating Rossby wave dynamics for the Agulhas Current LME (Fig. 4d). These attribution patterns align with established climate dynamics documented in previous studies (Timmermann et al., 2018; Jeon et al., 2022). While SHAP does not infer causality directly, the spatial alignment of attribution patterns with established climate dynamics provides physically interpretable evidence of the mechanisms underlying chlorophyll predictability, supporting the broader conclusion that the model internalizes aspects of coupled physical–biogeochemical dynamics from the training data. This is also reflected in the revised Discussion:

"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."

Line 228: I would consider deferring the discussion of ENSO until the section below.

We have reorganized Sections 3.2 and 3.3 in the revised manuscript following the reviewer's comment. Section 3.2 now focuses on reporting prediction skill patterns, retaining brief references to ENSO only where necessary to describe observed skill structures (e.g., the spring predictability barrier). The detailed mechanistic interpretation, including ENSO-related dynamics and Rossby wave propagation, is presented in Section 3.3 alongside SHAP attribution analysis. The revised Section 3.2 reads as follows:

"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. The model also consistently outperforms persistence forecasts across most initialization months at lead times up to approximately 12 months, with boreal winter initializations maintaining skill

advantages at even longer leads (green circles in Fig. 4a). 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).”

Figure 6: Please address concerns about cherry-picking. Does a chlorophyll lag time of 0 mean that the annual mean of chlorophyll was used to predict fish catch of that same year? Is that useful for real world applications?

We have addressed this concern in detail in our response to Major Comment 5. Briefly, the species presented in Figure 6 were not arbitrarily selected: for each LME with significant chlorophyll prediction skill, the ten most frequently caught species were tested, and only those with both statistically significant correlations and supporting ecological literature were retained. We have clarified the selection procedure in the revised manuscript, noting that this structured selection reduces the risk of purely spurious associations.

Regarding lag=0, this does not mean that the observed annual mean chlorophyll was used to predict fish catch of the same year. Our CNN takes NDJ (November of Year 0 – January of Year 1) observations as input and predicts the annual mean chlorophyll anomaly for Year 1. This prediction is issued at the beginning of Year 1, well before annual catch data are compiled. The predicted chlorophyll anomaly is then used in a linear regression to estimate fish catch for that same year (lag=0) or the following year (lag=1). Therefore, even the lag=0 relationship provides operationally useful anticipatory information, as the environmental prediction is available months ahead of the fisheries data it is related to. We acknowledge that lag=1 relationships (found in LMEs 6, 41, and 50) offer additional lead time for management applications. We have clarified this temporal sequence in the revised manuscript.

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:

590 *"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."

595

(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 600 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, 605 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 610 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 615 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_{0-300m}	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.

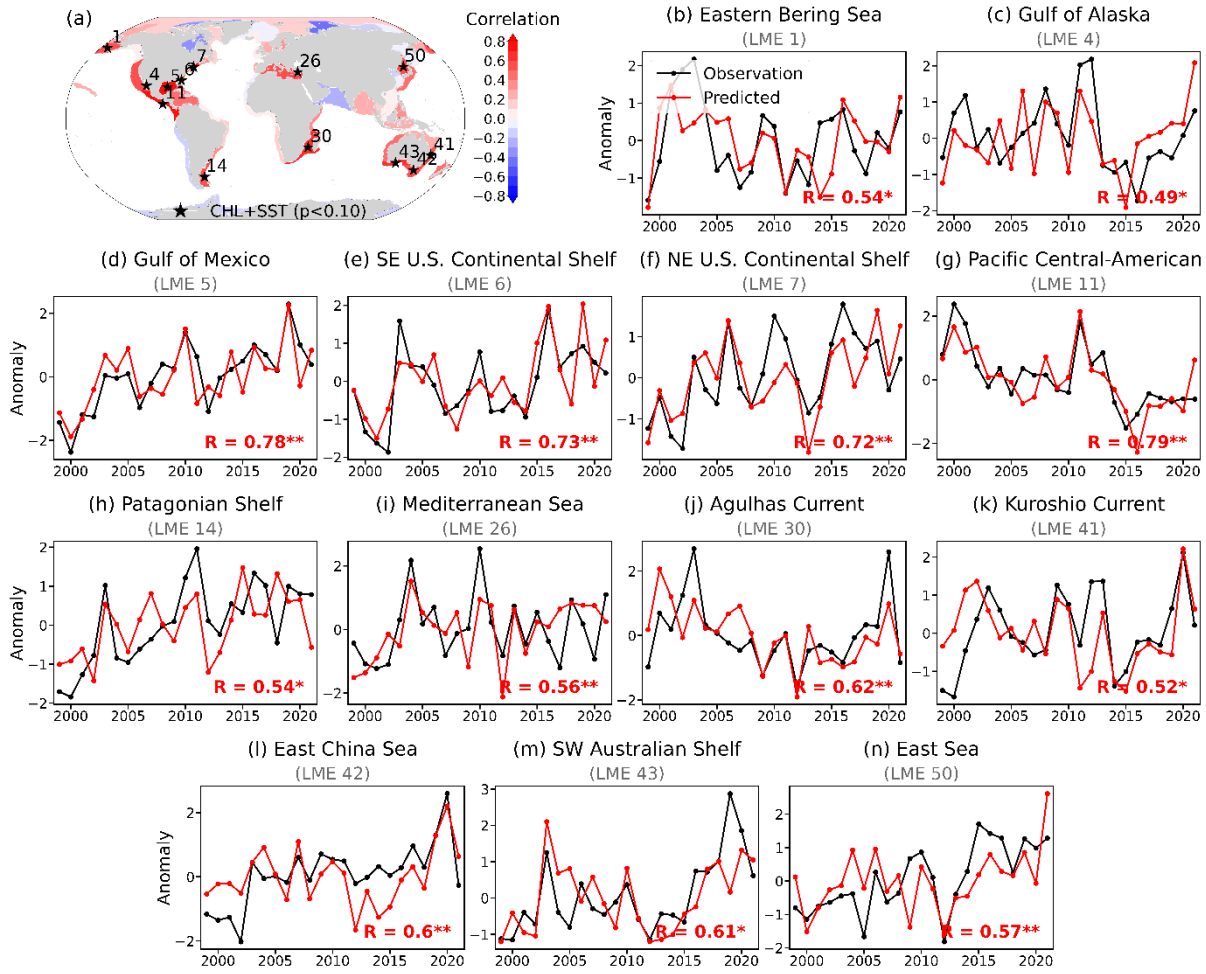


Figure B1: Chlorophyll prediction skill across Large Marine Ecosystems (LMEs). a Correlation coefficients between LME-averaged satellite-derived and predicted annual mean chlorophyll anomalies (1998-2021). The model takes November(Year 0)–December(Year 0)–January(Year 1) satellite observations as input 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).

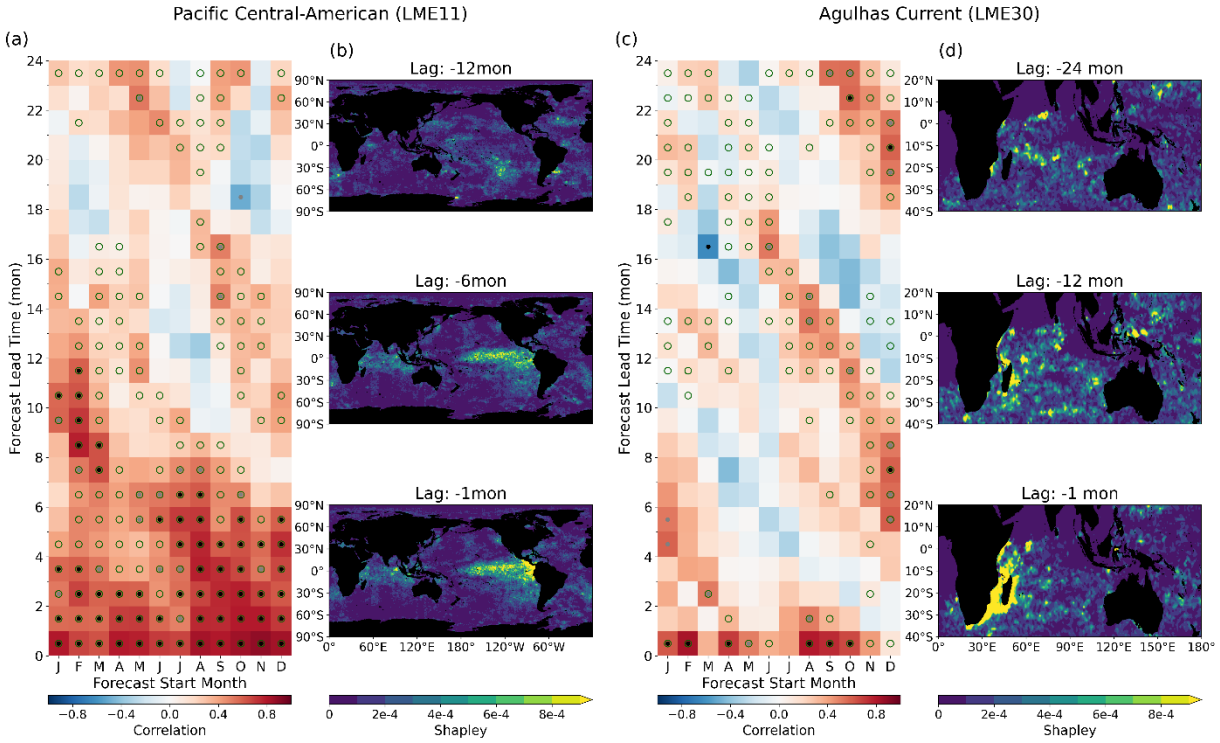


Figure B2: Monthly prediction and mechanism underlying chlorophyll prediction skill. a,c Anomaly correlation coefficient between predicted and satellite-observed 3-month running mean chlorophyll anomalies (LME-averaged) as a function of forecast start month (x-axis) and lead time (y-axis). Black dots 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.

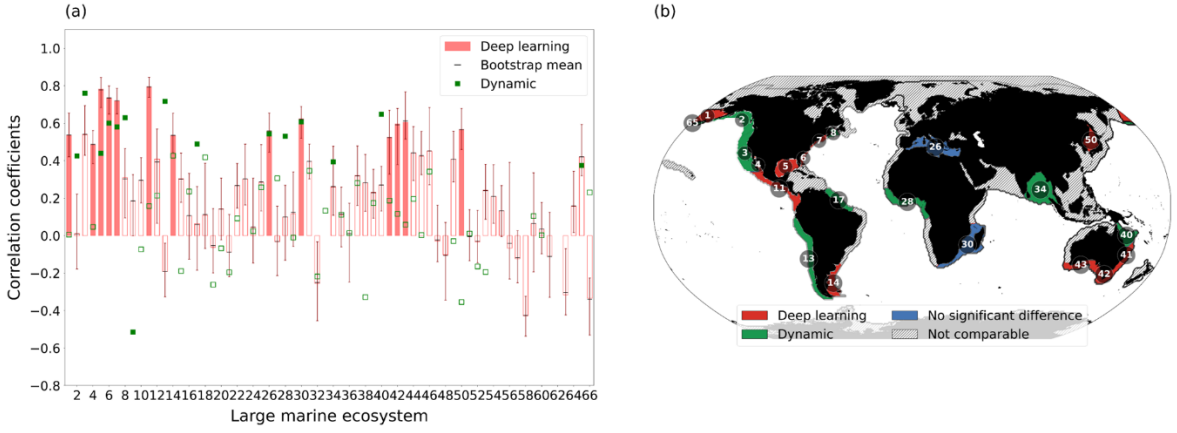


Figure B3: Comparison of chlorophyll prediction skill between deep learning and dynamic models across Large Marine Ecosystems (LMEs). (a) Correlation coefficients between satellite-observed and predicted annual mean chlorophyll anomalies at a 1-year lead time. Red bars show the deep learning model correlation; 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.

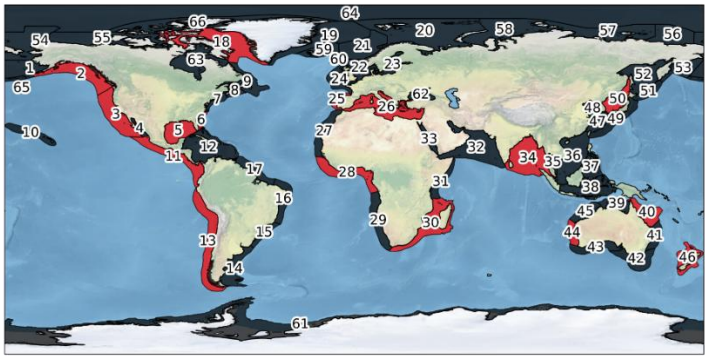
In Fig. 5a, the numbers of significant correlations are 15 for DL and 16 for dynamics. It appears to me to have quite poor performance results. Please reformulate L281-282.

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.

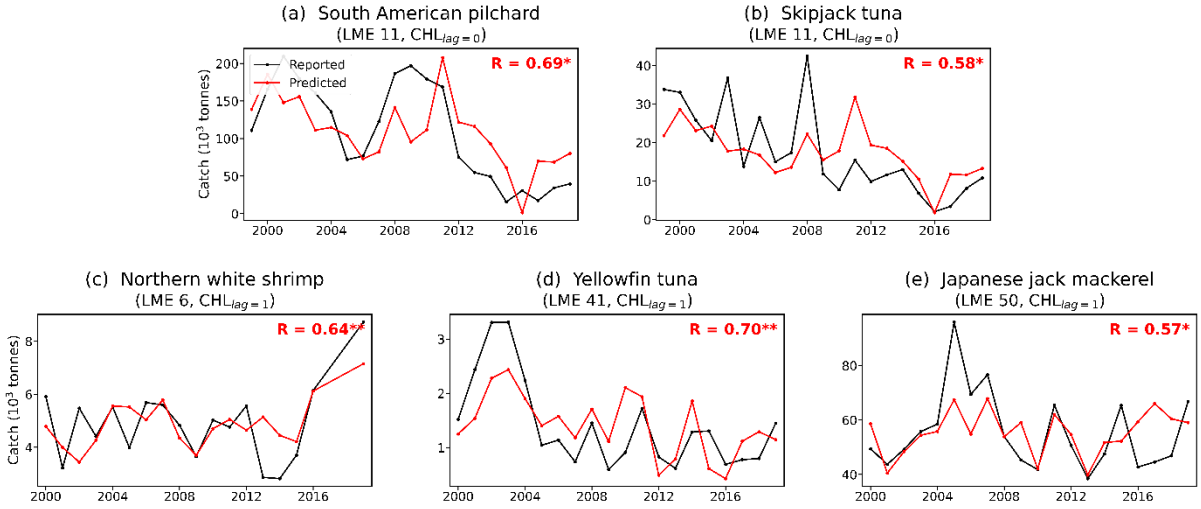


Figure B5: Prediction skill for annual fish catch of individual species in selected Large Marine Ecosystems (LMEs). a-e Time series of reported (black) and estimated (red) annual fish catch (tonnes). (a) South American pilchard, LME 11, lag=0. (b) Skipjack tuna, LME 11, lag=0. (c) Northern white shrimp, 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).”

Responses to Reviewer #3's Comments

I recommend major revision: the paper is promising and well positioned, but the current significance and evaluation framework does not convincingly rule out chance findings across many regions/species/lead times, and several methodological choices need clarification or strengthening.

The manuscript develops a CNN-based system to forecast surface chlorophyll anomalies for Large Marine Ecosystems (LMEs) using three consecutive months of SST and chlorophyll anomaly maps as inputs, trained on CMIP6 simulations plus a coupled reanalysis, and evaluated against SeaWiFS/MODIS satellite products. It additionally benchmarks against an ESM-based dynamical biogeochemical forecast system and explores fisheries relevance via regressions between predicted chlorophyll anomalies and reported fish catches.

Strength points:

- Clear problem framing and a relevant niche: you directly target known limitations in ESM biogeochemical predictability (observation sparsity, structural uncertainty, and computational cost), motivating a data-driven complement.

- Global scope with an application-relevant unit: the LME-scale framing is practical for coastal management and fisheries, and the system is trained/validated/tested on long records spanning CMIP6, reanalysis (1965–1997), and satellites (1998–2021).

- Interpretability attempt linked to dynamics: SHAP attribution is used to relate skill to recognizable mechanisms (ENSO-related patterns for the Pacific Central-American Coastal LME and Rossby-wave–like propagation in the Agulhas region).

We thank Reviewer #3 for the rigorous statistical assessment and constructive suggestions. Below, we address each comment in the order presented.

Major concerns (must address):

- Multiple testing / field significance: annual skill is presented as "significant" using $p < 0.10$ markers across LMEs, and you then show time series for eight LMEs with significant skill when using both SST and chlorophyll inputs. With many LMEs tested, $p < 0.10$ without a multiple-comparison correction (e.g., FDR control) is not sufficient to claim that the set of "significant LMEs" exceeds what would occur by chance; this is especially important because the paper's central headline is "skillful predictions in many LMEs."

We acknowledge that our explanation on the significant test needs to be more elaborated. The $p < 0.10$ significance threshold used in this study is based on effective degrees of freedom corrected for temporal autocorrelation (Bretherton et al., 1999), which substantially reduces the effective sample size from the nominal 23-year test period. This approach is commonly employed in the seasonal-to-decadal climate prediction literature where short verification records and strong temporal autocorrelation limit statistical power (Smith et al., 2019; Joh et al., 2023). We also note that the LMEs identified

as significant tend to coincide with regions having well-established linkages to large-scale climate variability (e.g., ENSO-influenced Pacific LMEs, Indian Ocean LMEs), providing some physical corroboration that these results are not solely attributable to chance. In addition, following the reviewer's comment, we have revised the manuscript to use *"several LMEs"* rather than *"many LMEs"* to more accurately reflect the results.

References

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Hermanson, L., Kharin, V., Kimoto, M., Merryfield, W. J., Mochizuki, T., Müller, W. A., Pohlmann, H., Yeager, S., and Yang, X.: Robust skill of decadal climate predictions, npj Clim. Atmos. Sci., 2, 13, https://doi.org/10.1038/s41612-019-0071-y, 2019.

Joh, Y., Delworth, T. L., Wittenberg, A. T., Yang, X., Rosati, A., Johnson, N. C., and Jia, L.: The role of upper-ocean variations of the Kuroshio-Oyashio Extension in seasonal-to-decadal air-sea heat flux variability, npj Clim. Atmos. Sci., 6, 123, https://doi.org/10.1038/s41612-023-00453-9, 2023.

- Monthly forecast evaluation appears cherry-pickable: monthly forecasts (up to 24-month lead) are shown for only two LMEs (Pacific Central-American Coastal and Agulhas Current), chosen because they "exhibit significant annual mean chlorophyll prediction skill." This selection criterion is not adequate to avoid post-selection bias for the monthly lead-time maps; readers will reasonably ask how typical these two are across all LMEs and whether the same lead-time structure occurs elsewhere. Relatedly, statements like "significant correlations extending up to 12-month lead times for forecasts initialized during boreal winter" (for LME 11) require stronger controls for the large number of initialization-month × lead-time tests shown in Fig. 4.

The two LMEs presented in Figure 4 were selected to examine whether the CNN captures physically meaningful forecast skill consistent with known dynamical mechanism rather than on the basis of post hoc inspection of monthly skill. Previous studies have shown that the Pacific Central-American Coastal LME (LME 11) is strongly influenced by ENSO (Park et al., 2019; Pennington et al., 2006), and the Agulhas Current LME (LME 30) is modulated by Indian Ocean variability and Rossby wave dynamics (Jeon et al., 2022). We have clarified this in the revised Section 3.2:

"We examined monthly forecasts by selecting two representative systems from Pacific and Indian Oceans, both exhibiting significant annual mean chlorophyll prediction skill and well-documented connections to large-scale climate variability in prior literature: the Pacific Central-American Coastal (LME 11) and the Agulhas Current (LME 30)."

Regarding the large number of forecast start month × lead-time tests in Fig. 4: each of the 288 cells (12 months × 24 leads) is assessed using effective degrees of freedom corrected for autocorrelation (Bretherton et al., 1999). The significant cells form coherent spatial structures, including the spring predictability barrier and diagonal reemergence banding, consistent with known climate dynamics. This physical coherence provides supporting evidence that the identified skill patterns are not statistical artifacts.

- Benchmark comparison needs uncertainty quantification: the dynamical benchmark is a strong part of the paper (it is well described as a 12-member, 2-year forecast system initialized monthly over 1991–2017). But the "outperformance" map that uses a correlation-difference threshold (≥ 0.2) at a nominal significance level raises questions: correlation differences should be accompanied by uncertainty estimates and a paired test (e.g., block bootstrap or Fisher-z with effective sample size) to show where differences are robust, not just large.

Following the reviewer's comment, the 0.2 threshold has been replaced with a statistically grounded comparison. In the revised manuscript, we employ a double bootstrap approach that simultaneously accounts for two sources of uncertainty: first, model uncertainty, by resampling with replacement from the five independently trained CNN ensemble members (each with different random weight initializations), and second, temporal sampling uncertainty, by subsampling 20 years without replacement from the 23-year test period to match the dynamical model's verification period (1998–2017). In each of 1,000 iterations, a resampled ensemble mean is correlated with satellite observations over the subsampled years, yielding a bootstrap distribution of CNN correlation skill for each LME. Statistical significance is then assessed by computing a one-sided bootstrap p-value, defined as the fraction of bootstrap samples where the CNN correlation falls at or below the dynamical model's correlation. This directly tests whether the CNN's skill advantage is robust to both ensemble and temporal sampling variability, without relying on an arbitrary threshold. The revised Figure 5 is reproduced below for the reviewer's convenience (Fig. C1).

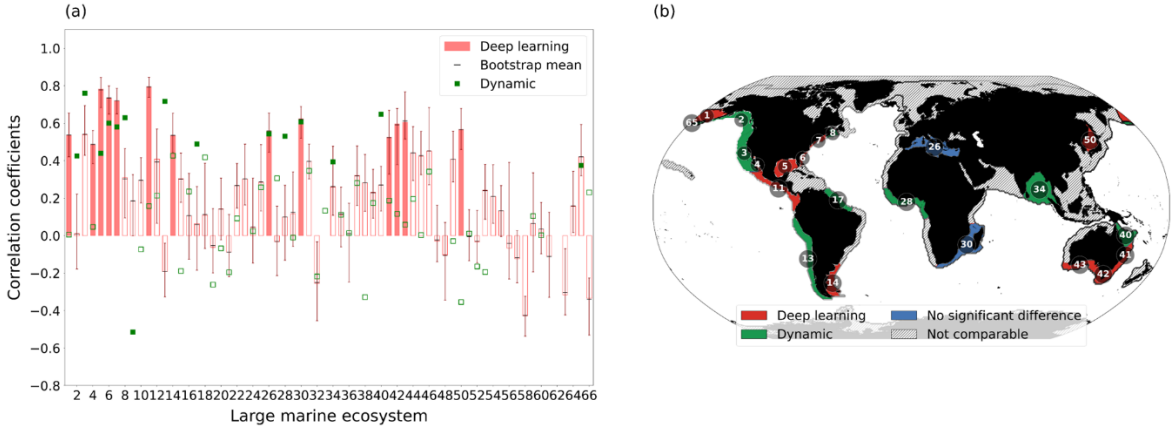


Figure C1: Comparison of chlorophyll prediction skill between deep learning and dynamic models across Large Marine Ecosystems (LMEs). (a) Correlation coefficients between satellite-observed and predicted annual mean chlorophyll anomalies at a 1-year lead time. Red bars show the deep learning model correlation; 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.

- Fish-catch results: selection and multiplicity: fisheries analysis uses Sea Around Us catches, selects top-10 species per LME, and applies linear regression using NDJ-initialized chlorophyll forecasts with different lags, reporting significant correlations for a subset of LME–species pairs. As written, it is unclear whether LMEs, species, and lags were pre-specified or chosen after looking at results, and there is no correction for the very large hypothesis space (LMEs × species × lags). Also, this section currently feels only loosely connected to the core ML/forecasting contribution (and it is linear regression, not a neural network), so it either needs a more rigorous, pre-registered-style evaluation or should be reframed as exploratory/supplementary.

We thank the reviewer for this constructive comment and agree that the fish catch analysis required clearer framing. We have revised Section 3.5 to frame the fish catch analysis as an exploratory demonstration rather than a validated fisheries prediction tool.

Regarding the selection procedure: LMEs were not pre-specified but restricted to those where the CNN demonstrated significant chlorophyll prediction skill, as this is a prerequisite for chlorophyll to serve as a predictable bottom-up driver of fisheries variability. For each such LME, the ten most frequently caught species were identified and all tested via linear regression at lags of 0 and 1 year. Species–LME combinations were retained where statistically significant correlations were found and supporting ecological literature suggested a plausible bottom-up forcing mechanism. We have clarified this structured selection procedure in the revised Discussion as follows:

"Species–LME combinations were selected based on two conditions: significant CNN chlorophyll prediction skill in the LME, and a statistically significant correlation between predicted chlorophyll and catch anomalies for species with a plausible bottom-up forcing mechanism suggested by ecological literature. While this structured selection reduces the risk of purely spurious associations, the analysis relies exclusively on bottom-up environmental forcing and does not account for top-down effects including fishing effort, management interventions, fleet behavior, and reporting practices, all of which strongly influence reported catch data."

Regarding the use of linear regression rather than the CNN itself, the intent of this section is to illustrate that CNN-derived chlorophyll forecasts retain sufficient environmental signal to explain interannual catch variability, consistent with the approach in previous seasonal prediction studies (Park et al., 2019; Tommasi et al., 2017). We agree that more sophisticated approaches could yield improved results, for example by incorporating additional biogeochemical variables such as NPP or by developing trophodynamic-based prediction frameworks that more explicitly represent energy transfer through marine food webs. Such developments would be a valuable direction for future work. The revised Figure 6 is reproduced below as Fig. C2.

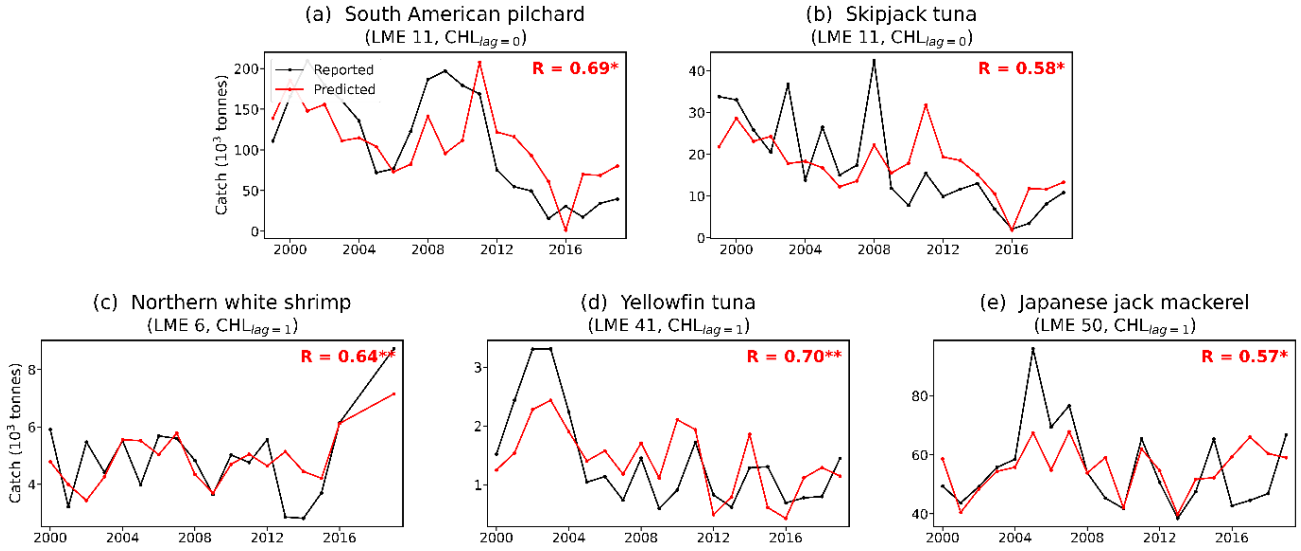


Figure C2: Prediction skill for annual fish catch of individual species in selected Large Marine Ecosystems (LMEs). a-e Time series of reported (black) and estimated (red) annual fish catch (tonnes). (a) South American pilchard, LME 11, lag=0. (b) Skipjack tuna, LME 11, lag=0. (c) Northern white shrimp, 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).

Methodological issues and clarifications:

- Zero-filling of missing ocean color: satellite chlorophyll has missing values (clouds/polar night), and you apply "zero-filling," masking missing pixels and filling them with zeros. This can inject artificial anomalies and create learnable artifacts (especially near persistently cloudy regions and high latitudes); at minimum you should quantify sensitivity (e.g., compare with masked-loss training, add a missingness channel, or use a learned imputation/gap-filled product).

We clarify that our zero-filling strategy is not applied on a per-timestep basis. 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. 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. The same mask is applied to CMIP6 training data to ensure spatial consistency, as described in the revised Section 2.2:

"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."

SHAP attribution maps show near-zero feature contributions from masked regions (Fig. 4b,d), suggesting that the network effectively ignores these areas. We therefore consider the current zero-filling approach adequate for the purposes of this

study, though we acknowledge that alternative approaches could be explored in future extensions, as noted in the revised Discussion.

"Similarly, the current zero-filling approach for missing satellite data, while shown to be effective through SHAP analysis showing near-zero contributions from masked regions (Fig. 4b,d), could be extended in future work through alternative approaches such as missingness indicator channels or masked loss functions."

- "Five ensemble members per initialization" is unclear: monthly forecasts are said to use five ensemble members per start date. For a deterministic CNN, this needs explanation (different random seeds? Monte Carlo dropout? perturbations of inputs?); also report how ensemble mean/spread are used in ACC computation and whether ensemble spread relates to skill.

We have clarified this in the revised manuscript. Ensemble members are generated by training five independent CNN models with different random weight initializations, rather than through perturbed initial conditions as in dynamical modeling systems. The ensemble mean prediction is used for ACC computation, as averaging across members reduces noise from stochastic training variability while retaining the learned climate signal. The revised Section 2.1.3 reads as follows:

"Where ensemble predictions are required, 5-member ensembles are generated by training five models with identical architecture and data but different random weight initializations."

While a formal spread-skill analysis was not conducted, the double bootstrap procedure employed in the dynamical model comparison (Section 2.4) resamples ensemble members to account for uncertainty from stochastic training variability, offering a partial characterization of ensemble-related uncertainty in the context of model comparison.

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.

Requested revisions:

- Add a formal multiple-testing treatment for annual LME skill (e.g., Benjamini–Hochberg FDR on p-values).

As discussed in our response to the major concern on multiple testing above, our significance testing employs effective degrees of freedom corrected for autocorrelation (Bretherton et al., 1999), consistent with standard practice in climate prediction studies Smith et al., 2019; Joh et al., 2023). We note that FDR corrections such as Benjamini–Hochberg assume independence among tests, which does not hold for spatially correlated LMEs sharing common climate drivers. The physical coherence of the significant LMEs, clustering in regions with established climate mode linkages rather than scattered randomly, provides additional support beyond purely statistical criteria.

- For monthly forecasts, provide summary skill maps/statistics across all LMEs (or a clearly pre-specified subset) for forecast start month × lead time, and then discuss the two highlighted LMEs as case studies.

The two highlighted LMEs are physically motivated case studies selected based on established climate–chlorophyll linkages, as described in our response to the reviewer's concern above. Selecting LMEs with well-documented physical mechanisms allows us to assess whether the CNN's monthly forecast skill reflects underlying dynamical processes rather than spurious statistical associations. Computing monthly forecasts for all 66 LMEs at 24 lead times would require training over 19,000 individual models, which is beyond the computational scope of this study. Nevertheless, the coherent skill patterns reflect physically meaningful signals, suggesting that these two LMEs are representative of the broader mechanisms driving chlorophyll predictability across the LME network.

- For DL vs dynamical comparison, replace the ad hoc "0.2 correlation difference" rule with a statistically grounded paired comparison and uncertainty intervals on skill differences.

As described above, we have replaced the ad hoc 0.2 correlation-difference threshold with a double bootstrap test that provides both uncertainty intervals (95% CI on CNN skill) and a formal one-sided p-value for each LME. The revised Figure 5 presents these results, showing where the CNN significantly outperforms or underperforms the dynamical model based on this statistically grounded comparison(Fig. C1 above).

- For fish catch, explicitly predefine the hypothesis set (LMEs, species, lags), and show aggregated results. If this cannot be done within scope, label the section clearly as exploratory and move it to Supplement.

We have revised Section 3.5 to explicitly frame the fish catch analysis as an exploratory demonstration of potential downstream applications, rather than a validated prediction system, as described in our response to the fish catch concern above. Regarding the suggestion to predefine the hypothesis set, the LMEs were restricted to those with demonstrated CNN prediction skill, and all top-10 species in each LME were tested at lags of 0 and 1 year, with results filtered by both statistical significance and ecological literature support. The fish catch analysis is presented in the main text as a direct demonstration of the LME-scale framework's relevance to fisheries and marine resource management, consistent with the motivation outlined in the revised Introduction.

- Clarify what constitutes "ensemble members" for the CNN monthly forecasts and how they affect reported significance

This is addressed in our response to the reviewer's ensemble clarification comment above. In brief, ensemble members consist of five independently trained CNN models with different random weight initializations. For significance evaluation, skill is computed from the ensemble mean time series rather than individual members, which reduces noise from stochastic training variability. Statistical significance is then assessed using effective degrees of freedom corrected for temporal autocorrelation (Bretherton et al., 1999), applied to the ensemble mean correlation coefficients.

Suggested revisions:

- Revisit missing-data handling: include a missingness mask as an input channel and/or use masked losses, and report sensitivity relative to current zero-filled preprocessing.

As described in our response to the zero-filling concern above, our unified masking approach ensures that zero-filled regions maintain constant zero values across all time steps and training samples, carrying no temporal variability, and thus contributing no learnable signal to the CNN. The network therefore effectively learns to rely on grid cells with non-zero, time-varying inputs, and SHAP analysis suggests that these masked regions contribute negligible attribution values to the prediction. We acknowledge that alternative approaches such as missingness channels or masked losses could offer additional flexibility in future extensions of the framework.

With these changes, the manuscript could make a strong, credible contribution: the global LME framing, dynamical benchmarking, and mechanistic interpretability angle are all compelling, but the statistical evidence needs to be made robust before the main claims can be supported.

Responses to Reviewer #4's Comments

Overall assessment (summary for editor and authors):

This manuscript tackles an important direction: using machine-learning emulation to support analysis of chlorophyll variability and uncertainty in Earth System Model (ESM) settings. The topic is timely and relevant to ESD readership, and the paper has potential value if it can clearly justify (i) why chlorophyll is the right motivating problem in an ESM context, (ii) how the proposed training/validation strategy avoids circularity and inherited model biases, and (iii) how interpretability claims (e.g., SHAP) are supported with visible results and biogeochemical discussion.

At present, however, the paper has several conceptual and technical gaps that make the narrative and methodology feel under-justified or internally inconsistent. In particular, the motivation around "chlorophyll problems" and the role of satellite ocean-colour data in ESMs needs to be sharpened; the reliance on CMIP6 simulations for training needs stronger justification given the manuscript's stated concerns about ESM limitations; the evaluation framework and skill metrics need to be stated explicitly; and the interpretability section is currently incomplete (SHAP plots are referenced but not visible).

I therefore recommend major revision. With rigorous clarification, strengthened motivation, clearer methods, and improved validation/interpretability presentation, the study could become a solid ESD contribution.

Major comments

Line 35–48 (motivation / problem statement on chlorophyll):

I am not finding enough rationale behind the "problem" with chlorophyll here. The manuscript should be more precise: what specific deficiency in simulated chlorophyll is being targeted (e.g., seasonal timing, bloom onset/termination, amplitude biases, regional pattern errors, long-term trends, nutrient limitation regimes, or vertical structure)? At the same time, ocean-colour observations are among the richest global observation networks and are widely used for short-term evaluation and reanalysis-type applications. Since the manuscript frames the work in the context of ESMs and climate projections, satellite ocean colour will not "solve" the forward-projection problem (it mainly constrains the historical period and near-real-time monitoring). This makes it difficult to pin down the motivation: why is chlorophyll observation discussed here in the ESM context, and what is the precise gap the proposed ML method fills? The authors should sharpen the problem formulation and explicitly connect it to ESM-relevant uncertainty and projection needs.

We have revised the Introduction (L35–48) to sharpen the problem formulation in the revised manuscript. We now acknowledge that ESMs have demonstrated skillful biogeochemical predictions before noting that prediction skill varies substantially across LMEs and lead times, avoiding framing ESMs as broadly deficient and instead highlights the specific gap our method addresses. We have specified that our framework targets seasonal-to-annual biogeochemical forecasts at the LME scale, directly connecting the methodology to the identified need.

Regarding the role of satellite observations, satellite-derived chlorophyll in our framework serves exclusively as an independent validation dataset rather than a training input. The CNN is trained on multi-decadal CMIP6 simulations and reanalysis fields, which provide centuries of diverse climate states necessary for learning physical–biogeochemical relationships, while satellite validation provides an independent evaluation of real-world applicability of the CMIP6-trained model. The revised Introduction reads as follows:

"Translating this understanding into actionable biogeochemical prediction remains challenging. While Earth System Models (ESMs), which integrate biogeochemical processes within physical climate frameworks (Flato, 2011; Bonan and Doney, 2018), have demonstrated skillful forecasts of oceanic physical variables on seasonal to decadal timescales (Smith et al., 2020; Balmaseda et al., 2024), recent advances have further shown prediction skill for biogeochemical variables including net primary production (Krumhardt et al., 2020), ocean carbon fluxes (Ilyina et al., 2021), ocean acidification (Brady et al., 2020), ecosystem stressors (Mogen et al., 2023), and seasonal to multiannual chlorophyll fluctuations across several regions (Park et al., 2019). Yet prediction skill varies substantially across LMEs and lead times. ESM-based biogeochemical forecasting remains constrained by limited observational records for biogeochemical fields, with satellite-derived chlorophyll-a records extending only since the late 1990s (Henson et al., 2010; Henson et al., 2016), structural uncertainties in biogeochemical models (Séférian et al., 2020; Fennel et al., 2022), large inter-model discrepancies, particularly where observational constraints are insufficient (Mignot et al., 2023; Kwiatkowski et al., 2020), and the substantial computational costs required for ensemble experiments (Balaji et al., 2022). These constraints have highlighted the need for alternative methodologies that can provide skillful biogeochemical forecasts at the scale of LMEs with greater computational efficiency."

Line 42–43 ("Contributions to the substantial uncertainties …"):

This sentence is unclear. If the authors mean parameterisations contribute substantially to uncertainty, please specify which parameterisations (e.g., phytoplankton physiology, photoacclimation, grazing closure, remineralisation, mixing-light coupling) and in what way they contribute. As written, "parameterisations" is too broad and reads vague.

We agree that the original phrasing was imprecise. We have revised this sentence to refer to *"structural uncertainties in biogeochemical models"* and cite Séférian et al. (2020) and Fennel et al. (2022), which document these uncertainties comprehensively, including phytoplankton physiology, grazing formulations, and nutrient cycling representations. Since our framework does not target specific parameterisation deficiencies but rather provides an alternative forecasting approach that bypasses detailed biogeochemical process formulations, we consider a detailed enumeration beyond the scope of this study. The revised text is as follows:

"ESM-based biogeochemical forecasting remains constrained by limited observational records for biogeochemical fields, with satellite-derived chlorophyll-a records extending only since the late 1990s (Henson et al., 2010; Henson et al., 2016), structural uncertainties in biogeochemical models (Séférian et al., 2020; Fennel et al., 2022), large inter-model

discrepancies, particularly where observational constraints are insufficient (Mignot et al., 2023; Kwiatkowski et al., 2020), and the substantial computational costs required for ensemble experiments (Balaji et al., 2022).”

Line 58–59 (context / recent related work):

It may be worth acknowledging that there are very recent works on physics-based AI integration within biogeochemical models (including preprints). For example, a preprint Banerjee et al., 2026 (https://doi.org/10.31223/X5C74R) could be cited as a related direction to position the study in the rapidly evolving landscape.

We thank the reviewer for this suggestion. Banerjee et al. (2026) is already cited in the revised Introduction, where we position our purely data-driven approach relative to hybrid frameworks that embed AI corrections within process-based models. The revised text reads as follows:

“While complementary efforts have explored hybrid approaches that embed AI corrections within process-based models (Banerjee et al., 2026), our framework takes a purely data-driven approach, ingesting three consecutive months of global sea surface temperature and chlorophyll anomalies to produce monthly or annual chlorophyll forecasts at the LME scale with lead times of 1–24 months, aligning with the temporal scales relevant for marine resource management decisions including seasonal quota setting, harvest control adjustments, and interannual stock assessment planning (Stock et al., 2015; Tommasi et al., 2017).”

Line 60 (“three consecutive months …”):

Please clarify what “three consecutive months” refers to (training window? evaluation period? event definition?) and why this temporal criterion is chosen. If it is a key design choice, it needs justification.

This has been clarified in the revised Section 2.1.2. "Three consecutive months" refers to the CNN input window: the model ingests three consecutive monthly fields of global SST and chlorophyll anomalies, yielding six input channels, to produce a chlorophyll forecast for the target LME. This input structure was adapted from prior work on CNN-based climate prediction (Ham et al., 2019), where a three-month window was found effective for capturing evolving climate signals. To justify this choice, we additionally tested 1-month and 6-month input windows (Fig. D1). The 1-month window yielded substantially lower skill (2/66 LMEs with significant predictions), suggesting that temporal context is necessary. The 6-month window also showed lower skill (6/66) compared to the 3-month configuration (13/66), suggesting that extending the window beyond three months does not improve prediction and may introduce noise. These results support the 3-month window as the optimal choice for our framework.

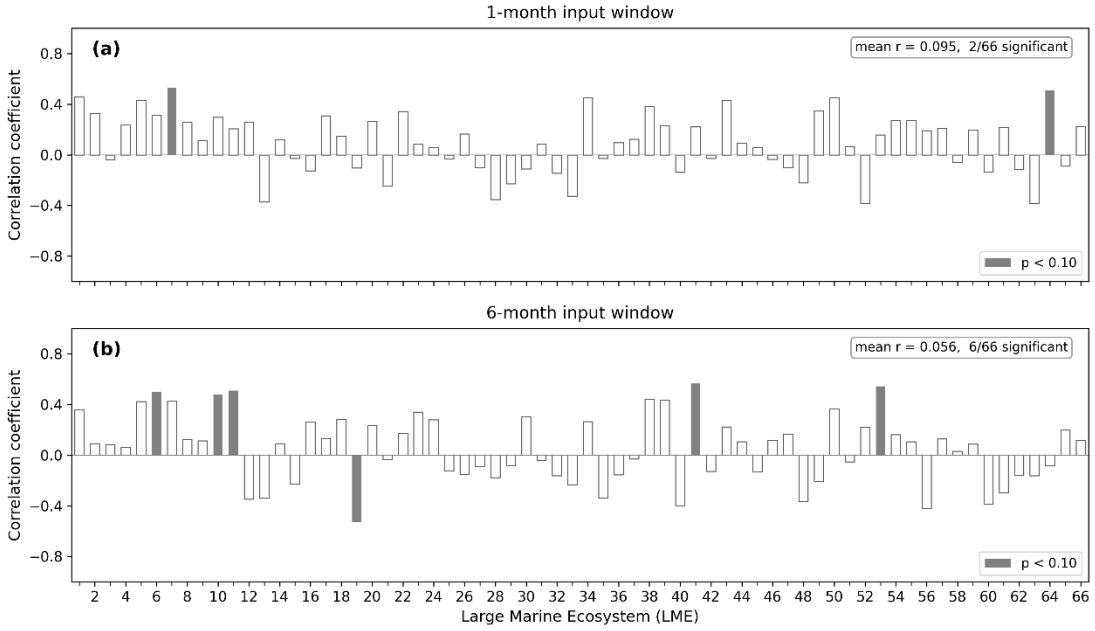


Figure D1. Prediction skill (correlation coefficient between CNN-predicted and satellite-observed chlorophyll anomalies) across all 66 LMEs for (a) 1-month and (b) 6-month input window configurations. Dark grey bars indicate statistically significant skill at p < 0.10. The 1-month window yields significant skill in only 2/66 LMEs (mean r = 0.095), and the 6-month window in 6/66 LMEs (mean r = 0.056), both substantially lower than the 3-month configuration (13/66 significant), supporting the selection of the 3-month input window as the optimal choice.

Line 64 (training with CMIP6 coupled models / conceptual consistency):

Training with CMIP6 model output will inevitably bring CMIP6's own uncertainties and biases into the learned emulator. Earlier, the manuscript discusses limitations of ESMs, yet the ML model is trained entirely on CMIP6 simulations, which feels self-contradictory unless carefully justified. The authors should explicitly explain: (i) what the emulator is learning (model space vs observational truth), (ii) why this is still useful for the stated aims, and (iii) how inherited CMIP6 biases are managed or acknowledged (e.g., domain limitation, bias-aware training, uncertainty propagation, or careful interpretation that results are "CMIP6-consistent" rather than "truth").

We acknowledge that training on CMIP6 output means the CNN learns physical-biogeochemical relationships from model-simulated data rather than from observational truth. However, our framework differs from a single-model emulator in that it learns from 16 structurally diverse models simultaneously, extracting relationships that are consistent across different model formulations rather than reproducing the behaviour of any individual model.

Training on CMIP6 offers two key advantages that satellite observations alone cannot provide. The 16 CMIP6 models plus reanalysis provide over 8,000 samples spanning centuries of diverse climate states, far exceeding what is available from the ~24-year satellite record. Multi-model training exposes the CNN to diverse representations of physics and biogeochemistry, encouraging the learning of relationships that are broadly consistent across models rather than specific to

any single model's biases (Guo et al., 2025). The practical value is demonstrated by the CNN achieving prediction skill comparable to ESM-based dynamical forecasts (Fig. D2) at a fraction of the computational cost.

We recognize that biases shared across CMIP6 models may propagate to CNN predictions. Multi-model diversity reduces the influence of individual model biases, and model predictions are independently validated against satellite-derived chlorophyll that was never used during training, providing a direct assessment of real-world applicability. This distinction is acknowledged in the revised Discussion:

"Nevertheless, biases shared across the CMIP6 ensemble, such as limited representation of coastal processes and common biogeochemical parameterization assumptions, may still propagate to CNN predictions, and the forecasts should be interpreted with this limitation in mind."

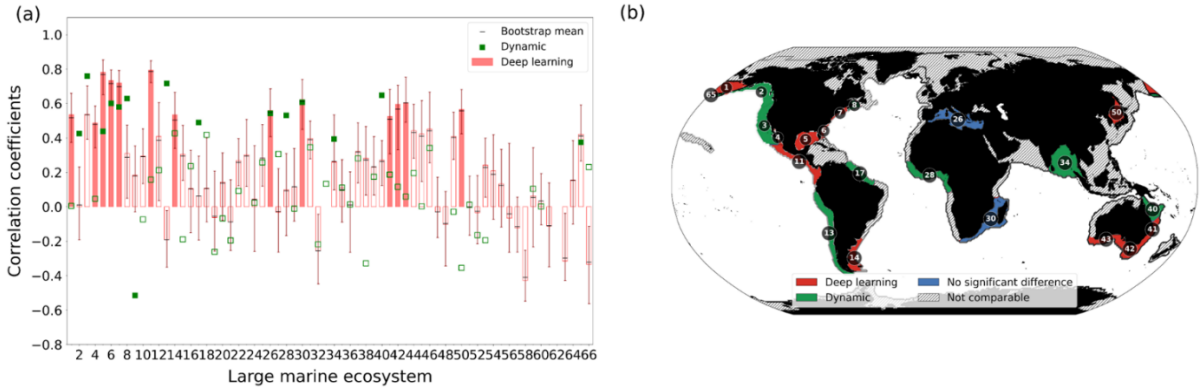


Figure D2: Comparison of chlorophyll prediction skill between deep learning and dynamic models across Large Marine Ecosystems (LMEs). (a) Correlation coefficients between satellite-observed and predicted annual mean chlorophyll anomalies at a 1-year lead time. Red bars show the deep learning model correlation; 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.

Figure 1 (infographic colour palettes):

In the infographic, the colour plots for CHL and SST (even if symbolic) would be clearer if the colour palette is differentiated between SST and chlorophyll. SST appears as a feature (not anomaly), yet it is shown using a coolwarm-type palette, which can be misleading. A more conventional SST palette (or a clearly distinct palette) would improve clarity.

The color scheme reflects the nature of the input fields. That is, both SST and chlorophyll inputs are standardized anomalies rather than raw values, thus the coolwarm palette was chosen to represent anomaly fields.

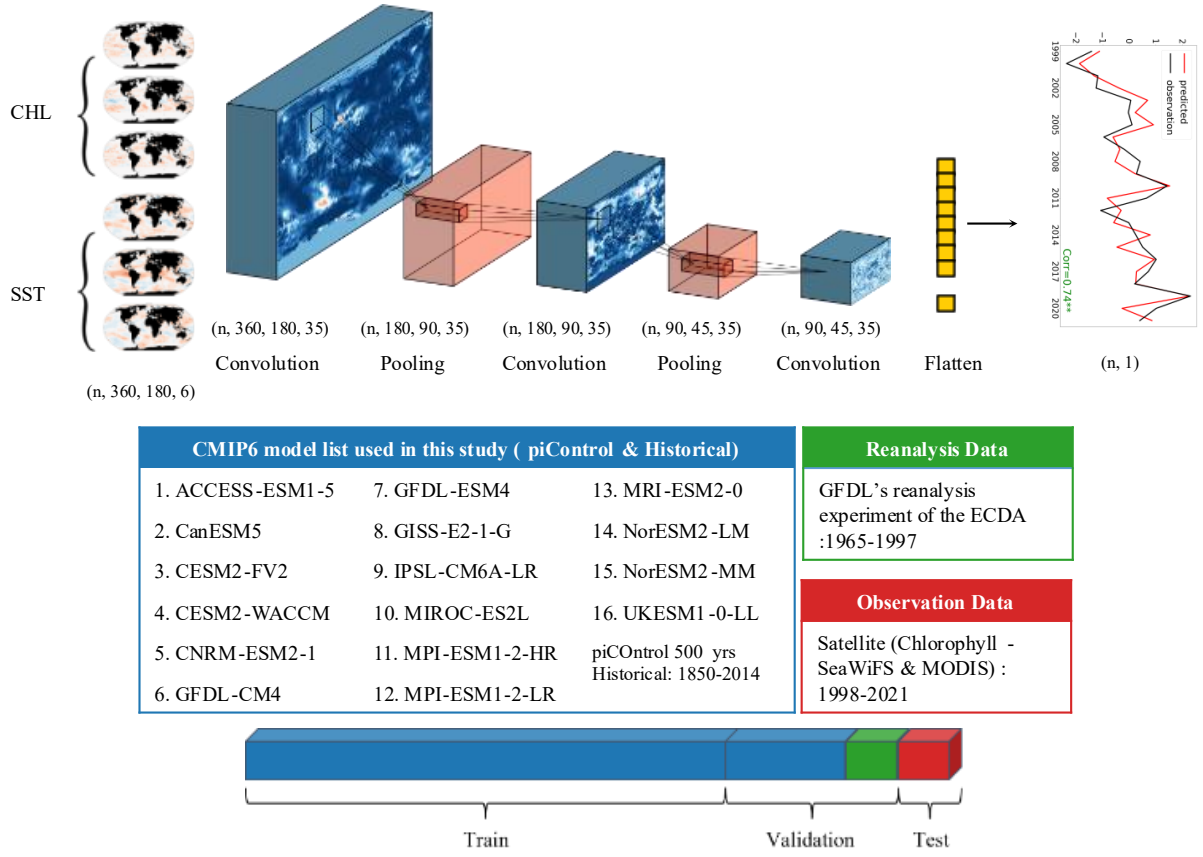


Figure D3: Deep learning model structure. The adapted CNN model comprises three convolutional layers (blue), two max-pooling (MP) layers (red), and one fully connected (FC) layer (yellow). Input data include sea surface temperature (SST) and chlorophyll anomalies for three consecutive months (e.g., November–January), represented as six channels. The model predicts either monthly or annual mean chlorophyll anomalies for each Large Marine Ecosystem (LME). Training data comprise historical and piControl simulations from 16 CMIP6 models, along with physical–biogeochemical reanalysis (1965–1997). Model validation was performed using satellite-based chlorophyll observations from SeaWiFS and MODIS (1998–2021).

Line 86–87 ("Seasonal or annual …"):

Why seasonal or annual mean? At this stage the reader should be specifically informed what temporal aggregation is used and why it is appropriate for the scientific question. Please be explicit.

Our framework targets seasonal-to-annual timescales because forecasts at this scale are directly relevant to fisheries management decisions including seasonal quota setting, harvest control adjustments, and interannual stock assessment planning (Stock et al., 2015; Tommasi et al., 2017). This timescale also aligns with the dominant influence of large-scale climate modes such as ENSO and IOD on marine productivity across LMEs, representing the primary source of predictable chlorophyll variability at the LME scale, and enables direct comparison with ESM-based dynamical biogeochemical forecasts (Park et al., 2019). The temporal aggregation is now explicitly described in the revised Section 2.1.2 as follows:

"Separate CNN models sharing the same architecture are trained for each combination of target LME, forecast type (annual or monthly mean), and lead time. For annual forecasts, models predict the annual mean chlorophyll anomaly for the target LME in the year following the forecast start. For monthly forecasts, separate CNN models are trained with different forecast start months and lead times (1–24 months ahead). 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 87 ("piControl …"):

Not sure a broad readership will understand "piControl" without explanation. Please add a short definition (e.g., long pre-industrial control simulation under fixed forcing, used as a baseline for internal variability).

Following the reviewer's comment, we have added brief definitions in Section 2.2. The reviesed text reads as follows:

"Historical simulations, driven by observed time-varying external forcings over 1850–2014, and preindustrial control (piControl) simulations, run under fixed pre-industrial forcing to provide multi-century records of internal climate variability, were used for training."

Line 88 ("Satellite-derived chlorophyll …"):

Why not train using satellite chlorophyll/ocean-colour data, given the abundance and long record, and instead use it only for testing? If the goal is ESM-relevant emulation, the authors should explain the choice clearly (e.g., differences in variable definition, sampling, error structure, coverage, or mismatch between satellite CHL and model CHL). As written, this choice needs stronger rationale.

The choice to train on CMIP6 rather than satellite observations is driven by several considerations. We note that our framework is not designed as an ESM emulator. It is a data-driven prediction model that learns physical–biogeochemical relationships from CMIP6 simulations and generates its own forecasts evaluated against independent satellite observations.

Satellite-derived chlorophyll records extend from 1998 to the present (~24 years). This is insufficient for the CNN to learn relationships across the full range of interannual-to-decadal climate variability, including diverse ENSO states and decadal modulations. CMIP6 simulations provide over 8,000 training samples spanning centuries of climate variability. By excluding satellite data entirely from training, it serves as a fully independent benchmark for assessing real-world applicability.

Training across 16 CMIP6 models exposes the CNN to a range of physical and biogeochemical representations, promoting the learning of relationships that generalize across model formulations rather than overfitting to the characteristics of a single observational product. These considerations are reflected in the revised Introduction as follows:

"The model is trained on multi-decadal climate model simulations from the Coupled Model Intercomparison Project phase 6 (CMIP6) (Eyring et al., 2016) and physical–biogeochemical reanalysis data (Park et al., 2018b), allowing it to learn from a broader range of climate variability than the satellite record alone provides. Model predictions are evaluated against satellite-derived chlorophyll, and compared with ESM-based dynamical biogeochemical forecasts."

Line 103–104 (handling SST = 0 / masking strategy):

A key methodological concern: if SST values are set to zero (or if missing values are replaced with zero), then the emulator could incorrectly learn that 0°C (or 0 K depending on conventions) is a meaningful physical signal in regions where it is not, and this could contaminate training. The authors should justify the zero-handling strategy and ideally use a robust missing-data approach (masking, NaN-aware methods, explicit land/ice masks, or physically meaningful fill values plus a missingness indicator feature). Please clarify how this is handled and why it is safe.

We clarify that SST fields are not zero-filled. We also note that our framework is not an emulator of any specific ESM. It is a data-driven prediction model that learns from CMIP6 simulations and generates its own forecasts. The optimally interpolated SST product (OISSTv2) provides near-complete global coverage and requires no gap-filling. Zero-filling applies only to chlorophyll fields, where a unified binary mask permanently flags grid cells with any missing value during the satellite period. We consider this approach appropriate for the following reasons. The model uses chlorophyll anomalies, so zero represents "no anomaly" rather than a physical zero concentration. Masked grid cells maintain constant zero values across all time steps and training samples, carrying no temporal variability and thus contributing no learnable signal to the CNN. SHAP attribution maps further suggest near-zero contributions from masked regions (Fig. 4b,d), consistent with the expectation that the network effectively ignores these areas. This is described in the revised Section 2.2:

"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. SST fields were not subject to this masking, as the optimally interpolated SST product provides near-complete global coverage."

Line 106–108 ("Reanalysis data …" validation choice):

Why use reanalysis data as validation, given known potential biases, especially for biogeochemical fields where assimilation and availability of ocean-colour observations vary and can affect the reanalysis itself? If satellite observations exist for chlorophyll, why are they not used more directly for validation? If reanalysis is used, please justify the choice and discuss limitations.

In the revised manuscript, the role of reanalysis in our framework has been clarified. The GFDL-ECDA reanalysis serves two distinct purposes: as part of the training dataset, providing observationally constrained physical–biogeochemical fields alongside CMIP6 simulations, and as a held-out validation dataset for sensitivity experiments (Section 3.1), where a temporally independent subset (1998–2017) is used to compare model configurations and select the reference model.The final model evaluation is performed entirely against satellite-derived chlorophyll (SeaWiFS and MODIS, 1998–2021), which is fully independent from all training and model development data. This two-stage design, with reanalysis for internal model selection and satellite observations for final independent evaluation, supports the interpretation that our reported prediction skill reflects real-world applicability. This is clarified in the revised Section 2.1.3 as follows:

"For sensitivity experiments (Section 3.1), model performance is evaluated on GFDL-ECDA reanalysis (1998-2017), independent from the training period. This validation step informed the selection of the reference model configuration. Final model evaluation uses satellite-derived chlorophyll from SeaWiFS and MODIS (1998-2021), fully independent from all model development data."

Line 118 ("To interpret …" SHAP figures not visible):

The manuscript mentions SHAP-based interpretation, but I am not able to see the SHAP feature plots in the main text. If they are in supplementary material, please ensure they are included and clearly referenced. If not provided, they should be added, as interpretability is part of the paper's core claims.

We note that SHAP attribution maps are presented in the main text as Figure 4b,d. These panels show spatial maps of absolute SHAP values at selected forecast lead times for both the Pacific Central-American Coastal (LME 11) and Agulhas Current (LME 30) regions. We suspect the reviewer may have experienced a PDF rendering issue, as these panels contain colour-mapped spatial fields that may not display correctly in all viewers. We have verified that the figures are correctly embedded in the revised manuscript, and produce Figure 4 as Figure D4.

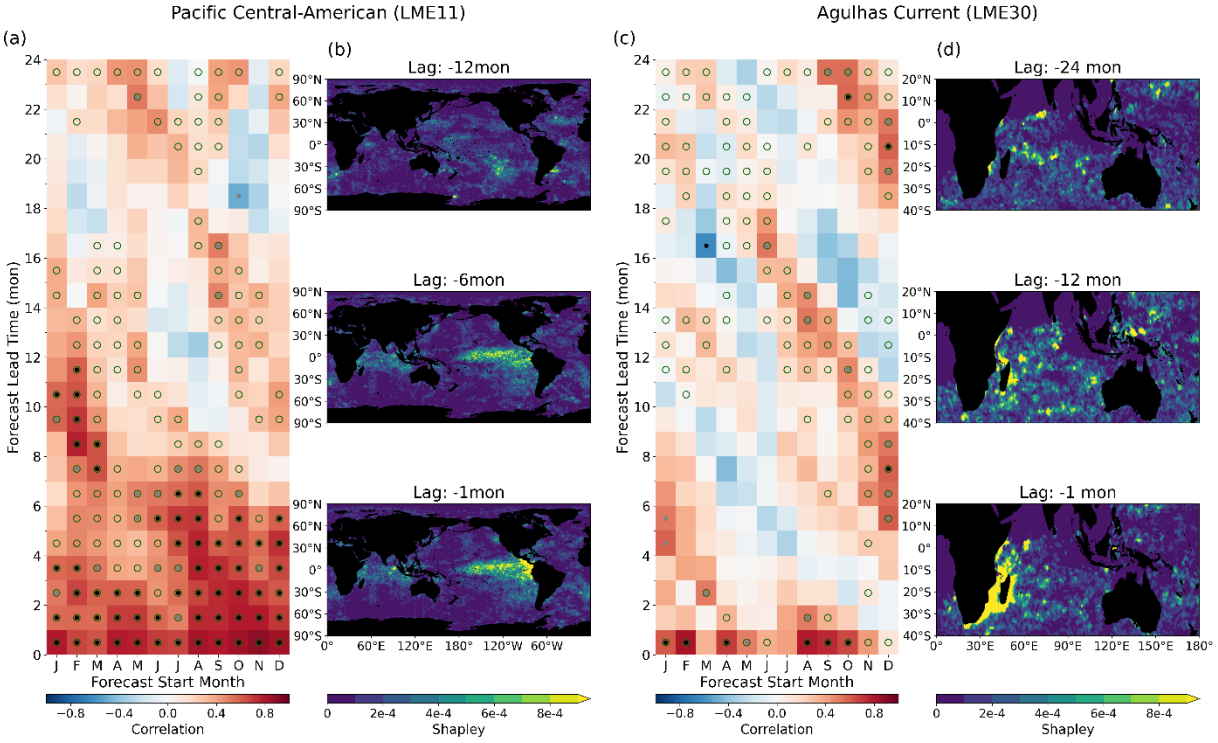


Figure D4: Monthly prediction and mechanism underlying chlorophyll prediction skill. a,c Anomaly correlation coefficient between predicted and satellite-observed 3-month running mean chlorophyll anomalies (LME-averaged) as a function of forecast start month (x-axis) and lead time (y-axis). Black dots 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).

Line 141–142 ("statistical significance …" skill metric definition missing):

I am not able to find the formulae/definitions for the prediction skill metrics used. The authors should define the prediction skill metric(s) clearly in the methods section (and provide the statistical test used for "significance," if applicable). Without explicit definitions, it is difficult to interpret the results.

Following the reviewer's comment, we have added explicit definitions in Section 2.1.2 of the revised manuscript. Prediction skill is evaluated using the anomaly correlation coefficient (ACC), computed as the temporal Pearson correlation between predicted and satellite-observed LME-averaged chlorophyll anomaly time series. Statistical significance is assessed using effective degrees of freedom corrected for temporal autocorrelation (Bretherton et al., 1999). The relevant definitions are provided in the revised Section 2.1.2:

"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."*

Line 181–183 (log-transform claim vs Fig. 2):

Figure 2 appears to show prediction skill is better with log-transformed chlorophyll, which seems contradictory to the claim made in lines 181–183. Please reconcile this: either revise the claim or clarify what specific aspect improves/worsens with the transform (e.g., relative vs absolute error, low-CHL regimes, extremes).

In Figure 2, the reference model (untransformed) shows higher average correlation skill across all 16 LMEs (bars) than the log-transformed variant. When considering only LMEs with statistically significant skill (green dashed line), the difference between the two becomes small. We acknowledge that the original text overstated the advantage of untransformed input, thus have revised L181–185 to reflect this more accurately. The revised Section 3.1 reads as follows:

"While the overall difference is modest, the untransformed input better preserves the dynamic range of chlorophyll variability in productive coastal LMEs, as log-scaling dampens high chlorophyll variability, which can lead to underestimate in regions with high concentrations (Cen et al., 2022)."

Line 210–211 (feature set choice is too narrow):

The feature set appears limited (e.g., mainly SST), but there are well-established publicly available datasets for other drivers that are biogeochemically relevant, such as mixed layer depth (MLD), PAR, winds (proxy for mixing), and potentially nutrients or stratification indices. Why were these not included? Even if the authors aim for minimal features, this choice needs to be justified given known controls on chlorophyll variability.

Our choice of SST and chlorophyll as primary predictors was guided by several considerations. SST serves as an integrated proxy for multiple physical drivers, reflecting upper-ocean stratification, mixed layer dynamics, and large-scale circulation patterns that collectively govern nutrient supply and light availability. The sensitivity experiments in Section 3.1 demonstrate that SST and chlorophyll together provide sufficient information for skillful predictions at the LME scale. We additionally tested subsurface potential temperature (0–300 m mean) as an alternative predictor, but this yielded lower skill than SST alone (Fig. 2), suggesting that surface fields already capture the relevant information. Furthermore, SST and surface chlorophyll are among the most consistently available and observationally constrained fields across the CMIP6 ensemble and the satellite record, making them a natural starting point. Regarding nutrients, global gridded nutrient observations with interannual resolution remain sparse, and CMIP6 nutrient fields exhibit large inter-model spread, making them less suitable as consistent training inputs. We acknowledge this as a limitation and note that incorporating additional physical drivers such as wind stress or mixed-layer depth in future extensions could potentially improve prediction skill in regions where local processes dominate chlorophyll variability. This is noted in the revised Discussion:

"Other key physical drivers, such as wind stress, mixed-layer depth, photosynthetically available radiation, and vertical nutrient gradients, were not systematically evaluated as CNN inputs. SST and chlorophyll were selected as inputs because both variables have consistent availability across the CMIP6 multi-model ensemble and nearly two decades of near-global satellite observations, making them a natural starting point for data-driven biogeochemical forecasting. SST additionally serves as an integrated proxy for multiple physical drivers including upper-ocean stratification and circulation. Regions where local processes dominate chlorophyll variability may nonetheless benefit from incorporating additional physical drivers in future extensions."

Line 242–244 (SHAP plots + biogeochemical interpretation vs black-box):

Again, I cannot see the SHAP plots referenced. In addition, for each SHAP result, the authors should explain whether the learned feature importance makes biogeochemical sense (not just statistical attribution), otherwise it risks being presented as black-box interpretability. The discussion should connect SHAP patterns to plausible mechanisms (e.g., SST as stratification proxy, seasonal light limitation, mixing control, etc.) and acknowledge where interpretation is uncertain.

As noted in our response to Line 118, Figure 4 including SHAP panels is presented in the main text and reproduced here as Fig. D4. We agree that SHAP results should be connected to plausible biogeochemical mechanisms, and Section 3.3 of the revised manuscript does this for both case studies. In the Pacific Central-American region, SHAP maps align with the canonical progression of ENSO-related SST anomalies along the equatorial Pacific, consistent with ENSO's established role

in modulating nutrient supply and primary productivity through thermocline displacement and upwelling variability. In the Agulhas Current, SHAP reveals westward-propagating features consistent with upwelling Rossby wave dynamics previously identified in ESM-based forecasts (Jeon et al., 2022). We note that SHAP does not infer causality directly, and that the interpretation is based on spatial alignment with known physical mechanisms rather than direct causal attribution. This is also reflected in the revised Discussion:

"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."

Major revision.

The study is promising and timely, but it currently does not strike a clear balance between Earth-system modelling theme and the AI-method framing. A major revision that sharpens the ESM-relevant motivation, resolves the CMIP6-training self-consistency issue, clarifies the evaluation framework (metrics + significance), strengthens validation choices, and fully provides/justifies interpretability outputs (SHAP) would substantially improve the manuscript and make it suitable for ESD.