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.

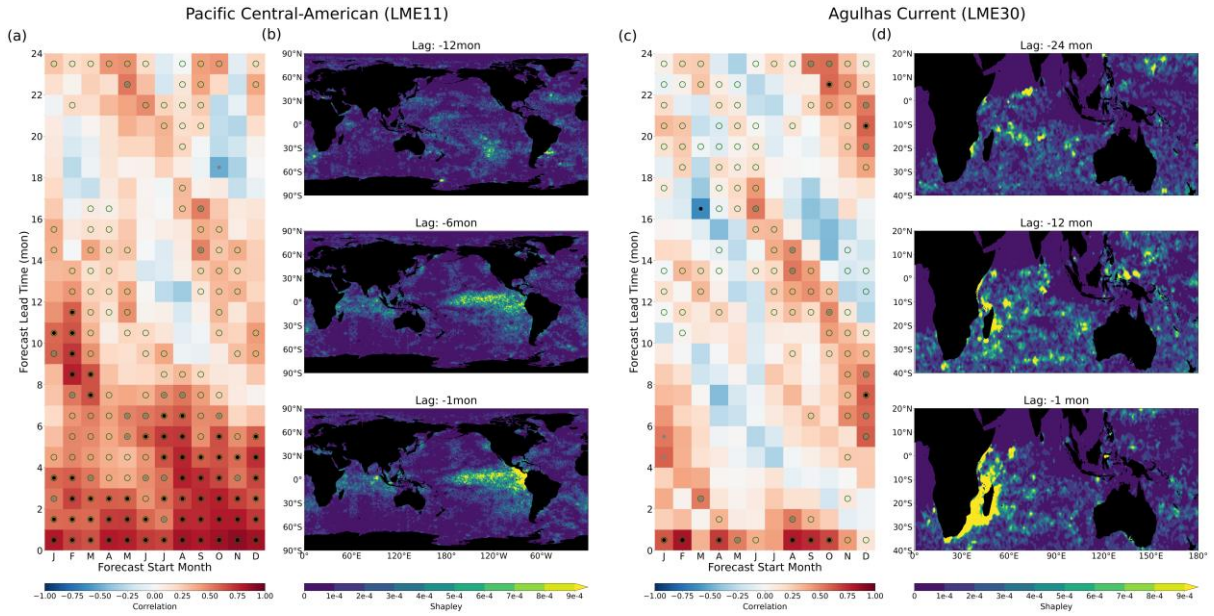


Figure A1: Monthly prediction and mechanism underlying chlorophyll prediction skill. a,c Anomaly correlation coefficient between predicted and satellite-observed 3-month running mean chlorophyll anomalies (LME-averaged) as a function of forecast start month (x-axis) and lead time (y-axis). Black dots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

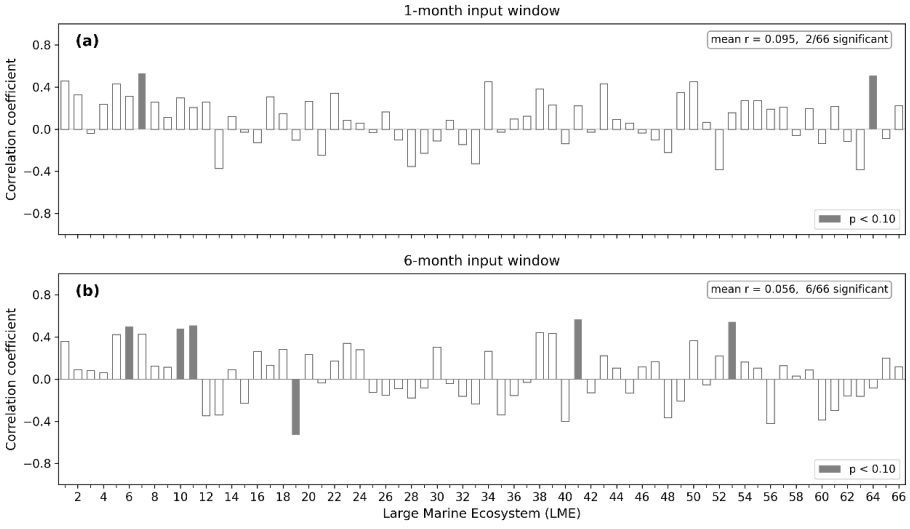
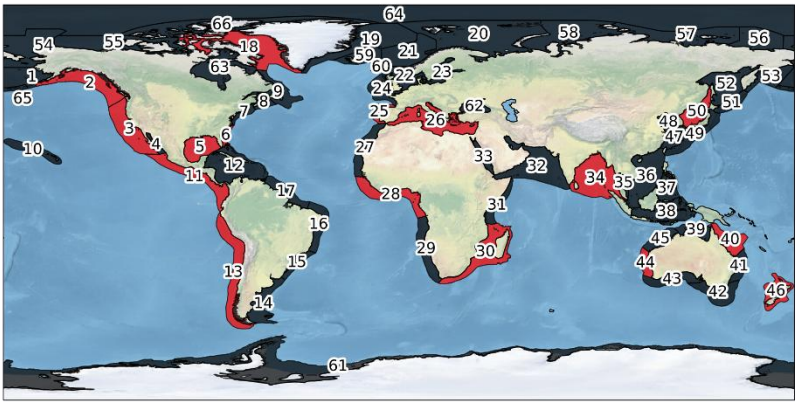


Figure A2. Prediction skill (correlation coefficient between CNN-predicted and satellite-observed chlorophyll anomalies) across all 66 LMEs for (a) 1-month and (b) 6-month input window configurations. Dark grey bars indicate statistically significant skill at p < 0.10. The 1-month window yields significant 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):

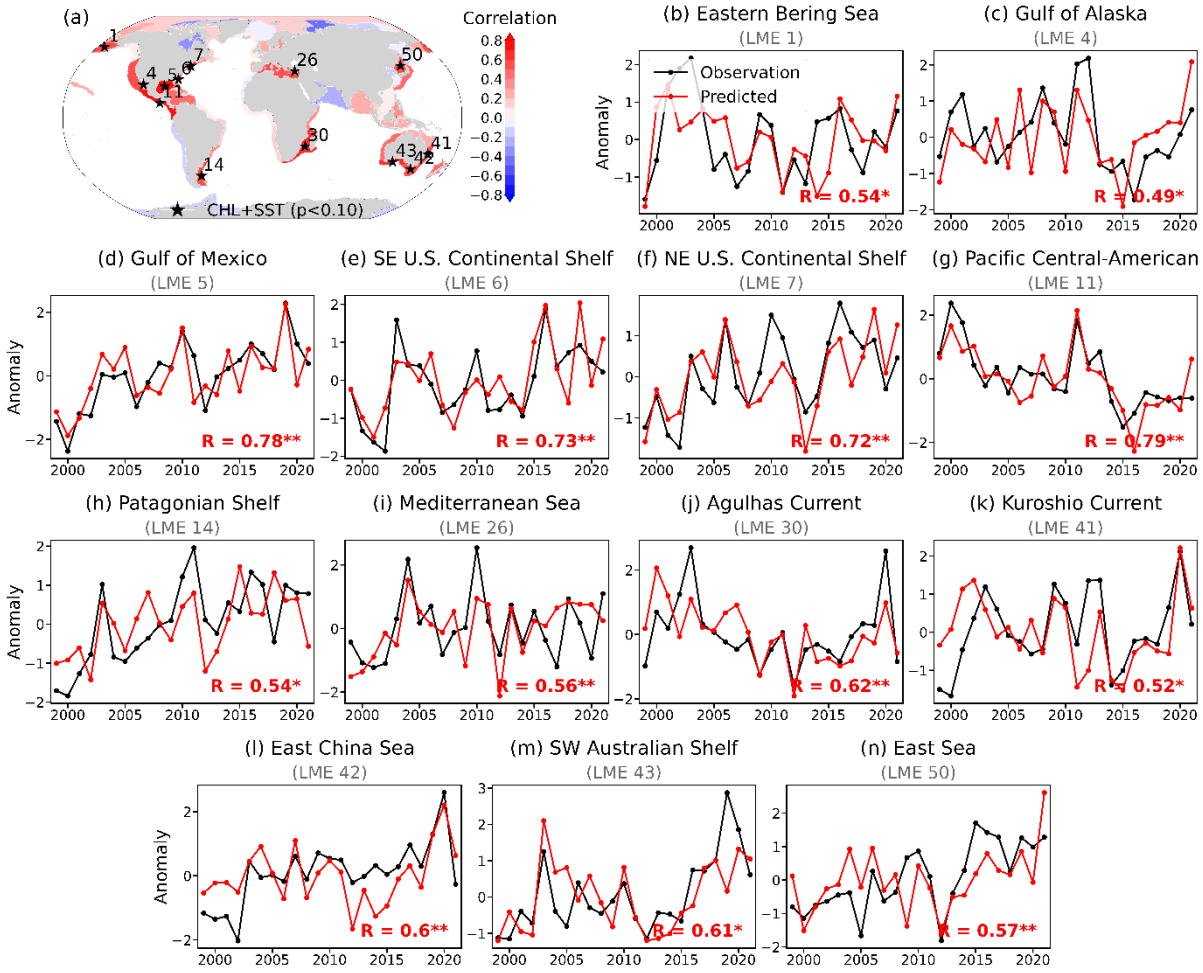


Figure A4: Chlorophyll prediction skill across Large Marine Ecosystems (LMEs). a Correlation coefficients between LME-averaged satellite-derived and predicted annual mean chlorophyll anomalies (1998-2021). The model takes November(Year 0)–December(Year 0)–January(Year 1) satellite observations as input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