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.

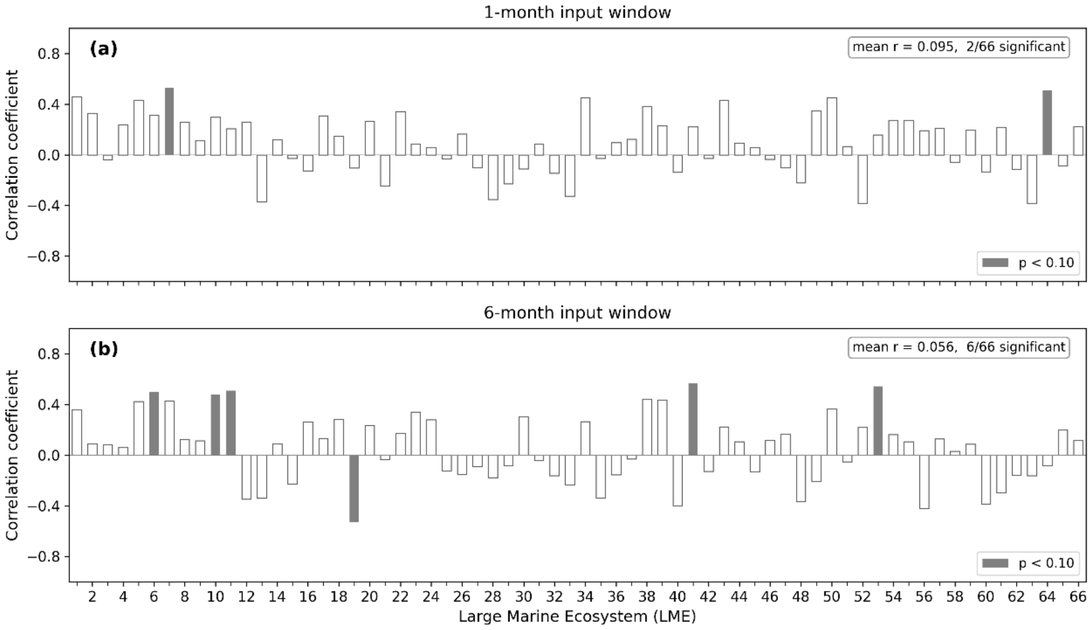


Figure D1. Prediction skill (correlation coefficient between CNN-predicted and satellite-observed chlorophyll anomalies) across all 66 LMEs for (a) 1-month and (b) 6-month input window configurations. Dark grey bars indicate statistically significant skill at p < 0.10. The 1-month window yields significant 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."

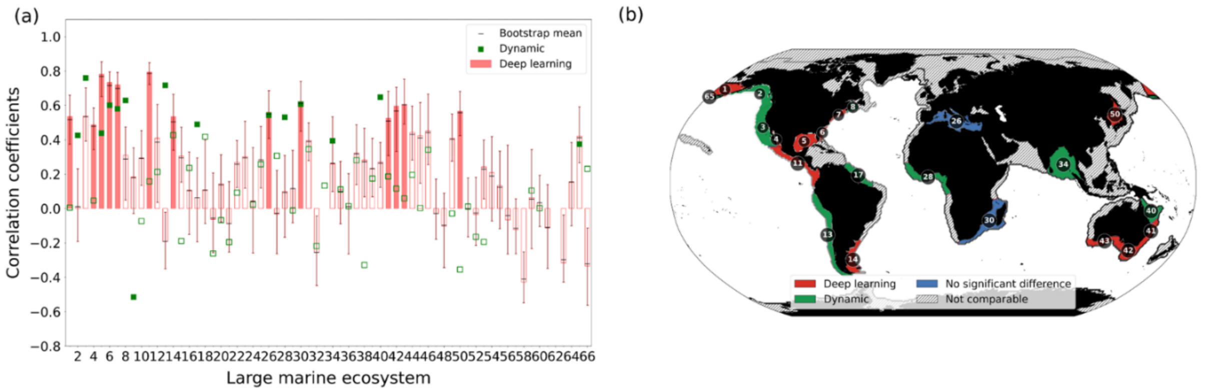
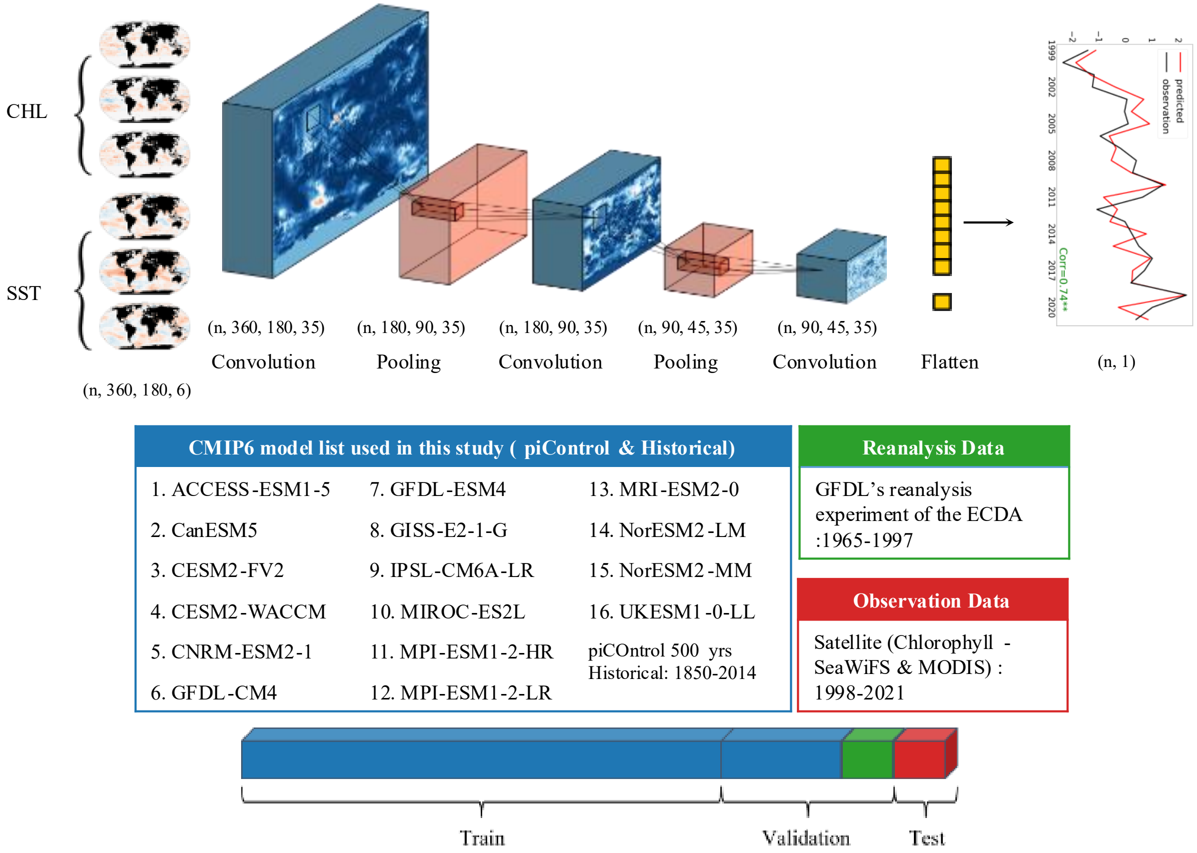


Figure D2: Comparison of chlorophyll prediction skill between deep learning and dynamic models across Large Marine Ecosystems (LMEs). (a) Correlation coefficients between satellite-observed and predicted annual mean chlorophyll anomalies at a 1-year lead time. Red bars show the deep learning model correlation;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



CMIP6 model list used in this study (piControl & Historical)

1. ACCESS-ESM1-5
2. CanESM5
3. CESM2-FV2
4. CESM2-WACCM
5. CNRM-ESM2-1
6. GFDL-CM4
7. GFDL-ESM4
8. GISS-E2-1-G
9. IPSL-CM6A-LR
10. MIROC-ES2L
11. MPI-ESM1-2-HR
12. MPI-ESM1-2-LR
13. MRI-ESM2-0
14. NorESM2-LM
15. NorESM2-MM
16. UKESM1-0-LL

piCOntrol 500 yrs
Historical: 1850-2014

Reanalysis Data

GFDL’s reanalysis experiment of the ECDA :1965-1997

Observation Data

Satellite (Chlorophyll - SeaWiFS & MODIS) : 1998-2021

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.

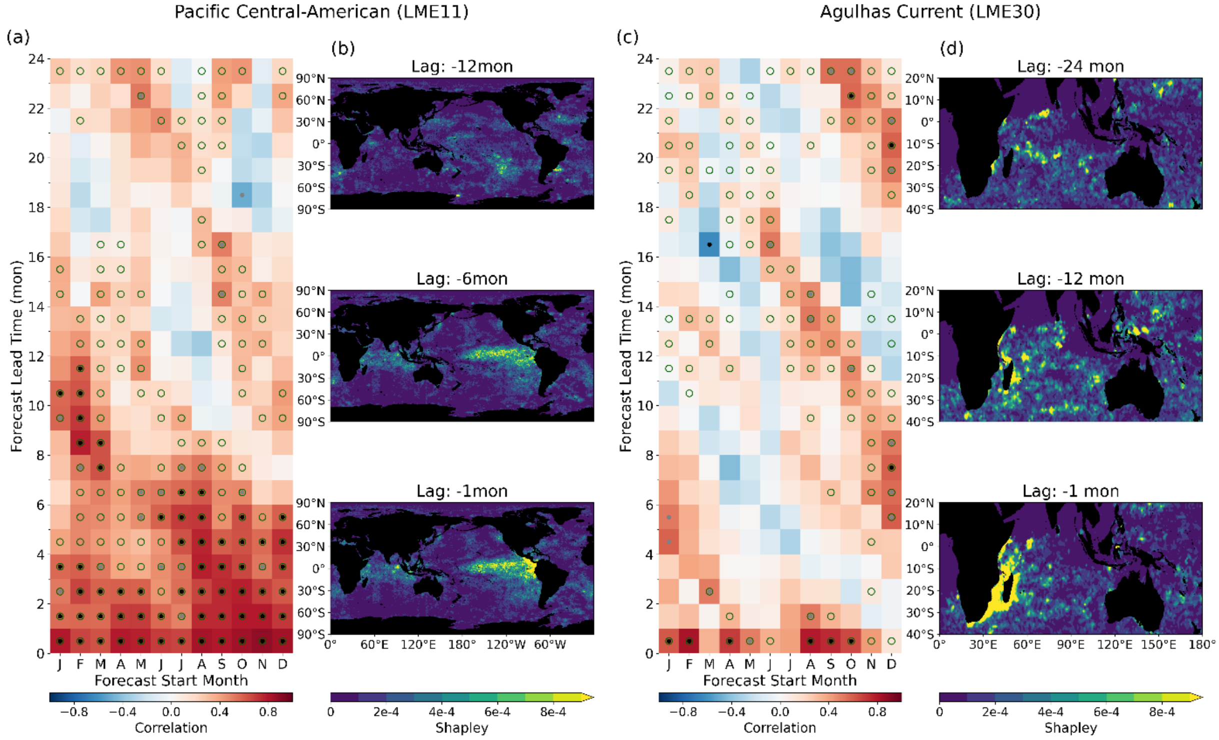


Figure D4: Monthly prediction and mechanism underlying chlorophyll prediction skill. a,c Anomaly correlation coefficient between predicted and satellite-observed 3-month running mean chlorophyll anomalies (LME-averaged) as a function of forecast start month (x-axis) and lead time (y-axis). Black dots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