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).

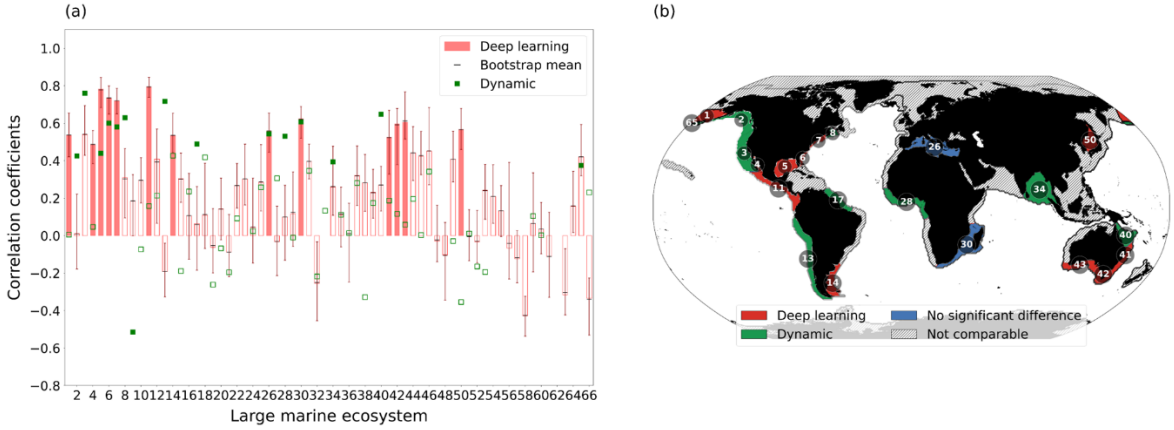

Figure C1: Comparison of chlorophyll prediction skill between deep learning and dynamic models across Large Marine Ecosystems (LMEs). (a) Correlation coefficients between satellite-observed and predicted annual mean chlorophyll anomalies at a 1-year lead time. Red bars show the deep learning model correlation;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

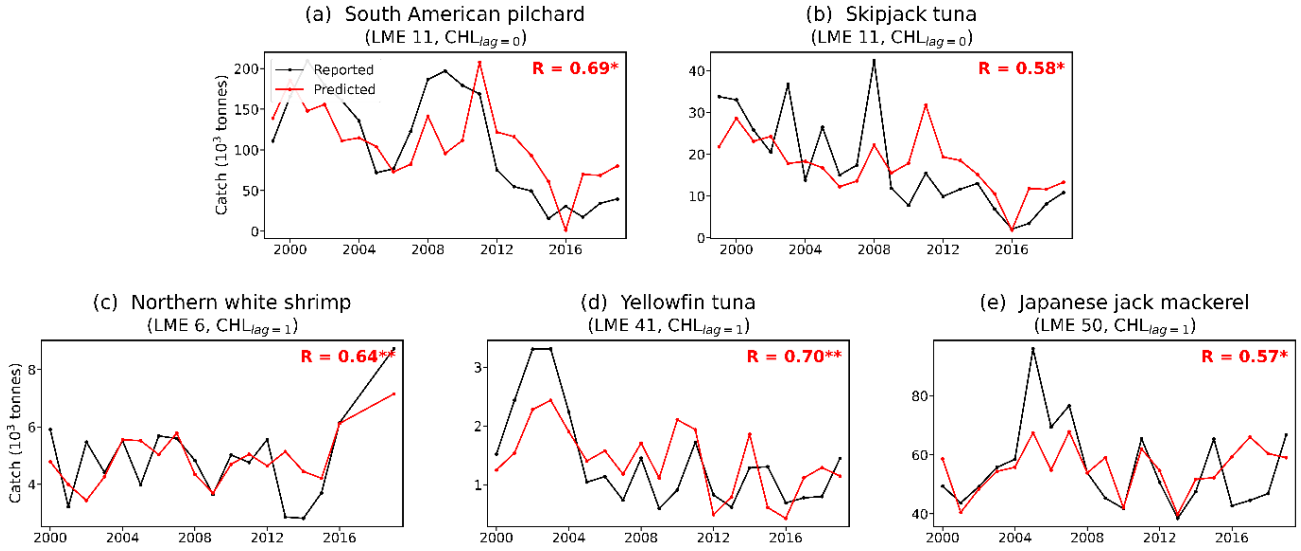

Figure C2: Prediction skill for annual fish catch of individual species in selected Large Marine Ecosystems (LMEs). a-e Time series of reported (black) and estimated (red) annual fish catch (tonnes). (a) South American pilchard, LME 11, lag=0. (b) Skipjack tuna, LME 11, lag=0. (c) Northern white shrimp,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