Author's Response of second review of Assessing seasonal climate predictability using a deep learning application: NN4CAST

January 23, 2026.

Detailed Comments
Responses are marked in blue.

Response to Reviewer 2

I appreciate the authors' careful revisions, which satisfactorily address my 1st, 2nd, and 4th general concerns. However, aspects of my 3rd general concer–relating to the distinction between associational and causal language in the teleconnection analysis–remain only partially resolved.

The revisions addressing the causal-language issue are appreciated, and the shift toward more cautious terminology (e.g., "potential drivers") improves the framing. However, parts of the response still lean on mechanistic interpretations that exceed what can be inferred from the associational analyses. For example, describing the DJF-MAM lag as "indicating causality," or asserting that the NAO component captured by the model is "likely the SST-forced part," goes beyond what is formally demonstrated, as NAO is not explicitly controlled for within the prediction framework.

In addition, the manuscript relies on conditional relationships (e.g., ENSO-NAO frequencies, period-dependent correlations, NAO-stratified composites) to infer mechanism–for instance, suggesting that NAO is "more internally driven" in P2 or "more strongly forced by SST anomalies" in P3, or interpreting attribution maps as identifying regions physically "contributing" to precipitation variability. These diagnostics are informative but remain descriptive; they do not isolate NAO as a confounder, nor do they establish distinct physical pathways beyond the SST-precipitation composites and the observed covariation with ENSO.

No additional analyses are needed. A short clarification in the manuscript would fully resolve the issue: namely, to state explicitly that the lagged correlations, conditional diagnostics, and attribution maps reveal associational patterns rather than isolated causal mechanisms, and that NAO-related variability is not independently controlled for within the ML framework. Adding this caveat where mechanism is discussed (e.g., in the P2–P3 comparison and the interpretation of attribution maps) would prevent any overinterpretation while keeping the scientific message intact.

We thank the reviewer for the clarification around this topic. To clarify this topic, we have added this clarification around the discussion of model attributions in the interbasin and

precipitation studies. (Lines 279-281 and 370-372 the reviewed manuscript, in bold, highlighted the new changes).

"Attribution maps highlight that the central Pacific is a major contributor for both WTNA and SMSCU, specifically around 170ºE-150°W, with additional positive contributions from the eastern part of the Pacific basin as well as negative contributions from a region located around 130°W-110°W (Fig. 3b-d). **Notice that these spatial patterns should be interpreted as reflecting statistical associations within the learning framework, capturing co-varying large-scale climate signals rather than direct causal influences from individual regions.**"

"These composites highlight how changes in both tropical Pacific forcing and the extratropical background state contribute to variations in model skill and in the mechanisms linking ENSO to European precipitation, all of which are captured by the simulated rainfall. Composites for negative precipitation events reveal a similar but reversed mechanism (see Fig. S6 in Supplementary Material). **We note that the diagnosed relationships represent statistical associations learned by the NN model. While the patterns resemble known modes of variability (e.g., ENSO, NAO), they do not imply causal forcing, which would require dedicated sensitivity experiments to isolate individual contributions.**"

Response to Reviewer 4

Remarks: Authors proposed a framework for assessing seasonal climate predictability using deep learning. They have discussed two case studies with their proposed framework. Overall, the work is good and I think it has potential to be published. However, there are some major issues in the present form, which requires clarifications. Please see below some major comments:

Major Comments:

1- Fig. 2 and 4: Authors have discussed Anomaly correlation and RMSE maps and they have discussed their formulation. However, I do not understand how they have estimated a time series of ACC and RMSE (shown in Fig. 2). According to eq. (2), m sample and I guess sample here is meant to be total time period. So, using eq. (2), spatial anomaly correlation map is right field and map, but how they have estimated time series, that not clear. If authors meant pattern correlation, which is different from anomaly correlation? Please clarify, what is ACC and RMSE time series meaning?

We thank the reviewer for pointing out this ambiguity. In our formulation, for computing the time series of ACC and RMSE, each sample in Eq. (2) does not refer to the entire time

period, but rather to a single year, represented by a full spatial field over the analysis domain. That is, for each year, the predicted and reference anomaly fields are compared across all spatial grid points, yielding one ACC and one RMSE value per year. Repeating this procedure for all years in the dataset produces a time series of ACC and RMSE, as shown in Fig. 2.

Therefore, the ACC time series represents the temporal evolution of the spatial anomaly correlation between predicted and observed fields on a year-by-year basis, and it should not be interpreted as a pattern correlation computed over the full time period. We have clarified this definition explicitly in the new version of the manuscript to avoid confusion. (Lines 259-260 and 318-320, in bold, highlighted the new changes).

“Notably, the skill in the TNA region remains positive and ranging from 0.4 to 0.8 in most of the years, with improved performance during certain decades, as during the 1980s**, where each yearly value represents the spatial anomaly correlation computed over the full domain** (Fig. 2b)”

“The temporal evolution of the ACC and RMSE (Fig. 4b-d), **where each yearly value represents the ACC and RMSE computed over the full domain,** supports the latter hypothesis; there are periods such as the 1950s-1970s when the model displays considerable skill (ACC > 0.4), contrasted with periods like the 2000s where skill is minimal (ACC ≈ 0).”

2- Please see following literature, which could be relevant (https://www.nature.com/articles/s41612-025-01198-3)

We thank the reviewer for the suggested reference. We agree it is highly relevant: Kent et al. (2025) demonstrate that a machine-learning weather model trained on reanalysis (ACE2) can produce skilful seasonal predictions and capture large-scale modes such as the NAO, while also highlighting interpretability challenges. We have added a short citation and sentence in the manuscript to place our ML-based attribution results in this broader context. (Lines 74-78).

3- Fig 5: The predicted precipitation pattern is based on the data driven forecast model. Why authors are applying EOF on their predicted rainfall? If they do not apply EOF, then what is the prediction skill of the predicted precipitation pattern compared to the observations (without EOF)? i strongly suggest to please show the skill of the predicted precipitation pattern comparing with observation.

We thank the reviewer for this comment. We agree that EOF analysis is not a direct measure of deterministic prediction skill. The purpose of applying EOFs to the predicted precipitation field is not to evaluate pointwise forecast accuracy, but rather to assess whether the data-driven model is able to reproduce the dominant modes of large-scale precipitation variability observed in the reference data. Precipitation is a highly intermittent and spatially noisy variable, for which grid-point skill scores are often low even in physically based models. In this context, examining the leading EOF patterns and their associated temporal

coefficients provides insight into whether the model captures the main coherent spatial structures and their temporal evolution.

Therefore, the EOF analysis is intended as a diagnostic of dynamical consistency between predicted and observed precipitation variability, rather than only as a measure of forecast skill. The prediction skill of the model in reproducing the observed precipitation patterns without applying EOF is already quantified and shown in Fig. 4, where spatial maps of ACC and RMSE are presented, and further summarized by periods in Figs. 5e-g. These figures provide a direct comparison between predicted and observed precipitation fields and therefore address the model skill independently of the EOF decomposition.

4- Fig. 2: I see authors have mentioned about statistical significance of the maps, but that is not clear.

We thank the reviewer for raising this issue. The assessment of statistical significance is explicitly described in the caption of Figure 2. As indicated there, statistical significance is assessed using a one-tailed t-test with a 95% confidence level, applied to the ACC values. We have clarified the caption of Figures 2 and 4 to be more explicit about the significance representation:

“Statistically significant results, determined using a one-tailed t-test at the 95% significance level, **are indicated by the shading while dashed black lines indicate non-significant areas in panel (a), and values above the dashed line in panel (b**).”

In the case of Figure 2, the model ability to predict SST is consistently high across the entire tropical North Atlantic domain. As a result, all ACC values for the grid points exceed the significance threshold and, therefore, no shaded (i.e., non-significant) regions appear in panel (a). This indicates that the entire forecast field is statistically significant.

In contrast, in the case of the precipitation forecasts shown in Figure 4a, the model's ability is more spatially heterogeneous and, in general, weaker. Consequently, regions where the ACC does not reach statistical significance are explicitly indicated by shading in that figure. This distinction reflects differences in predictability between SST and precipitation fields, rather than differences in the methodology used to assess significance.

5- Which precipitation dataset is adopted as observations, used in Fig. 5?

We thank the reviewer for this question. For all precipitation-related analyses presented in Figs. 4-6, the observational reference dataset is the Climatic Research Unit gridded Time Series (CRU TS) precipitation dataset. The same CRU TS dataset is consistently used as the observational benchmark throughout the precipitation analysis, including Fig. 5. This dataset is described at the beginning of the corresponding section (Lines 305-306):, where we state:

“For this analysis, SST data from the HadISST dataset were used as the predictor field, while precipitation data from the Climatic Research Unit gridded Time Series (CRU TS) dataset served as the predictand.”

6- Authors are claiming this is an explainable AI approach, which is not very much. How they are demonstrating that the approach they have adopted is more like explainable AI approach? Please explain clearly. Can authors explain the results in eq. (3) framework?

We thank the reviewer for this comment. As described in the Introduction, our approach employs eXplainable AI (XAI) techniques to quantify the contribution of each predictor feature to the model's output for a given target region. In particular, we use the Integrated Gradients (IG) method, which provides theoretically grounded feature attributions satisfying axioms such as sensitivity and implementation invariance (Sundarajan et al.2017).
Equation (3) formalizes this attribution framework: $R_{i,n}$ represents the relevance of the predictor feature at grid point i for sample n, computed as the product of the distance between the feature and a chosen baseline $\hat{x}$ and the average of the gradients along the straight-line path connecting the baseline to the feature. This allows us to interpret the predicted output in terms of the relative contributions of individual input features.

Importantly, these attributions reflect associational relationships captured by the model rather than direct causal mechanisms. By applying this methodology, we can identify which regions of the predictor field have the strongest statistical influence on the predicted target, and assess their relative importance and spatial or temporal variations. Therefore, although our model is a deep learning-based predictor, the use of IG provides feature-level attributions that are interpretable while remaining associational rather than causal.

We have added this clarification in the new version of the manuscript, just after the definition of Equation 3 (Lines 134-141). Concretely:

"The attributions computed using this equation provide a quantification of the contribution of each input feature to the predicted output for a given target region. By integrating gradients along the path from a baseline input to the actual predictor field, we can identify the most influential regions and quantify their relative importance. These attributions reflect statistical associations captured by the model and should not be interpreted as evidence of direct causal relationships. This approach allows the model to remain entirely data-driven, while providing interpretable information at the feature level on the factors that determine seasonal predictability, in line with XAI principles. The next section describes how this theoretical framework is implemented in the NN4CAST tool."