Response to Reviewer 3

October 20, 2025

Detailed Comments
Responses are marked in blue.

The paper introduces NN4CAST, a Python-based framework designed to identify and investigate drivers of seasonal climate predictability. It shows that NN4CAST provides explainability by attributing predictions to specific regions of the chosen predictor field, thereby quantifying the relative importance of different sources of predictability.

The paper addresses an interesting problem and proposes a framework for understanding sources of predictability. However, the manuscript currently lacks details on the method and justification of key choices, as well as on the interpretation of XAI results to make the framework truly useful for climate services and science. In the perspective of this reviewer, the framework as well as the examples chosen to illustrate its usefulness would benefit from some reconsideration prior to possible resubmission.

General comments

The method chosen to make the predictions is not discussed or justified in the paper. Why is an autoencoder architecture chosen in the first example? It should definitely be discussed whether this makes a difference to the regions identified by the XAI method? Given the short observational record and non-stationarity of the teleconnections, can a deep learning approach always be justified compared to a regularized regression?

We thank the reviewer for this insightful comment. In the revised manuscript, we have clarified the choice of network architecture and its advantages over simpler alternatives such as regularized regression. The structure we adopt (1024–256–64–256–1024) corresponds to a fully connected encoder–decoder, also referred to as an autoencoder-type MLP. This bottleneck design progressively compresses the high-dimensional predictor field into a compact latent representation before reconstructing the target field, which facilitates the extraction of the most relevant nonlinear predictive features while controlling model complexity.

To better situate our approach within the existing literature, we now explicitly reference studies that have applied autoencoder-type networks for seasonal prediction tasks. For example, Ibebychu et al. (2024) employed autoencoders combined with LSTMs to forecast ENSO, demonstrating the suitability of such architectures for capturing physically meaningful patterns in climate data. While our implementation is fully connected rather than recurrent, it follows the same encoder–decoder principle, ensuring interpretability and robustness when dealing with high-dimensional predictor fields. We have clarified this explanation in the text (Lines 238-246).

As for the short observational record and the non-stationarity of teleconnections, we acknowledge this as an important limitation. Our goal here is not to claim that deep learning will always outperform regularized regression, but to demonstrate that the NN4CAST framework is able to identify windows of opportunity and to capture skillful predictions even in challenging cases. For example, in the precipitation application, the framework reveals periods with significant skill despite the known non-stationarity of the ENSO-Europe teleconnection, something that would be difficult to capture with a purely linear model.

Ibebuchi, C. C., & Richman, M. B. (2024). Deep learning with autoencoders and LSTM for ENSO forecasting. Climate Dynamics, 62(6), 5683-5697.

This reviewer agrees with the two other reviewers that tropical Atlantic should not be included in predictor region in the first example.

As also raised by the first reviewer, we addressed this issue by designing an additional experiment in which we explicitly masked the predictor domain to exclude the Caribbean/western tropical Atlantic, while at the same time applying a complementary mask to the predictand field to exclude the Pacific. This setup ensures that there is no overlap between predictor and predictand regions, and thereby allows us to directly test to what extent local SST persistence may be influencing the results.

Importantly, this adjustment does not require any modification of the model code, since the masking can be implemented directly during the preprocessing of the SST fields prior to entering the prediction pipeline.

The results of this sensitivity experiment show that model skill, measured both in terms of ACC and RMSE, remains high even after removing the Caribbean band from the predictor field. We do observe a modest reduction in skill in certain sub-regions (e.g., around the Gulf of Mexico), but the overall performance and attribution patterns remain consistent with those reported in the main text.

In the revised manuscript, we have updated the corresponding figure (new Figure R1) to illustrate these results, and we have also updated the Zenodo repository with the outputs of this new experiment to ensure full transparency and reproducibility.

We sincerely thank the reviewer for this highly relevant and constructive comment. We fully agree that, as you point out, even a narrow band of DJF SST in the western tropical Atlantic can carry substantial persistence into MAM, which in turn may artificially inflate the apparent skill in our first case study (DJF tropical Pacific predictors - MAM TNA SST).

As also raised by the first reviewer, we addressed this issue by designing an additional experiment in which we explicitly masked the predictor domain to exclude the Caribbean/western tropical Atlantic, while at the same time applying a complementary mask to the predictand field to exclude the Pacific. This setup ensures that there is no overlap between predictor and predictand regions, and thereby allows us to directly test to what extent local SST persistence may be influencing the results.

Importantly, this adjustment does not require any modification of the model code, since the masking can be implemented directly during the preprocessing of the SST fields prior to entering the prediction pipeline.

The results of this sensitivity experiment show that model skill, measured both in terms of ACC and RMSE, remains high even after removing the Caribbean band from the predictor field. We do observe a modest reduction in skill in certain sub-regions (e.g., around the Gulf of Mexico), but the overall performance and attribution patterns remain consistent with those reported in the main text.

In the revised manuscript, we have updated the corresponding figure (new Figure R1) to illustrate these results, and we have also updated the Zenodo repository with the outputs of this new experiment to ensure full transparency and reproducibility.

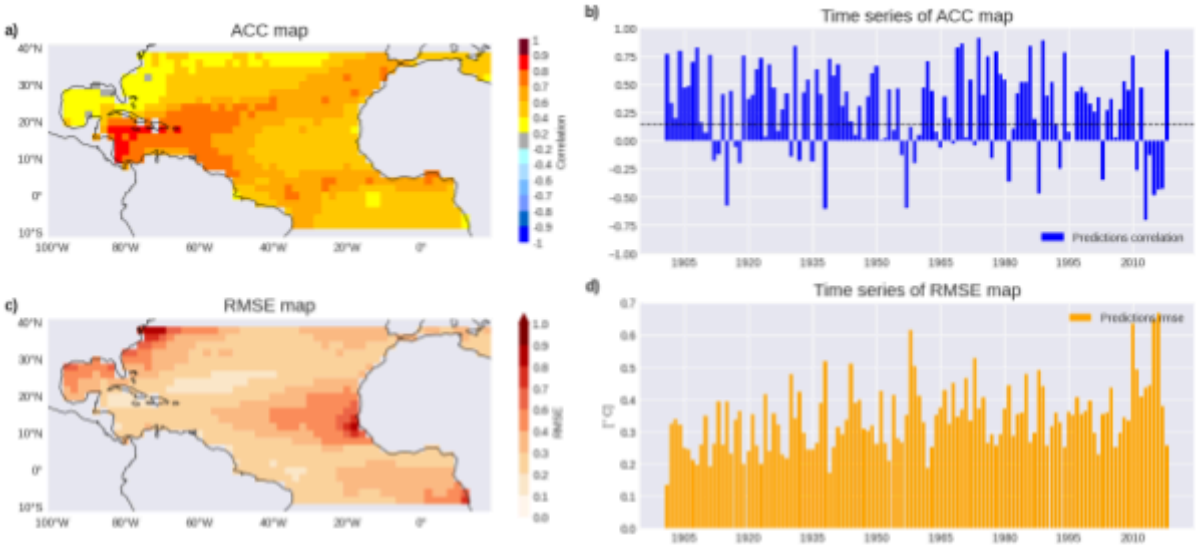

Figure R1. Predictability of tropical North Atlantic SST variability from tropical Pacific anomalies. Panels showing model performance metrics over the full period (1901–2019) using a leave-one-out cross-validation approach for predicting the SST anomaly field of the tropical North Atlantic during MAM, with Tropical Pacific SST from DJF as the predictor. The predictions are compared against observed MAM SST anomalies. Specifically: (a) ACC spatial map, correlating at each grid point the observed and predicted time series (temporal dimension); (b) Time series of ACC maps, correlating for each year the observed and predicted spatial patterns (spatial dimension); (c) RMSE spatial map, computed analogously to (a); and (d) Time series of RMSE maps, computed analogously to (b), all calculated between predicted and observed fields. The ACC (RMSE) time series show the correlation (error) between predicted and observed global mean SST anomalies over time. Statistically significant results, determined using a one-tailed t-test at the 95\% significance level, are indicated by the non-dashed regions in panel (a) and values above the dashed line in panel (b).

Parts of the paper read a lot like a Python package documentation rather than a method or framework description (for example lines 156-164, Table 1, Listing 1-3). Since the paper is presenting a framework and not a package, this reviewer thinks that they might be better suited in the Appendix or Supplementary Material. In particular, the paper contains no details or discussion on the choice of deep learning method, which should be included in the main text - perhaps at the expense of the code description.

We thank the reviewer for this insightful comment. In the revised manuscript, we have moved all Python code listings (previously in the main text, including Listings 1–3) to the Supplementary Material. This change ensures that the main text focuses on the methodological framework and rationale rather than detailed code instructions. The full code and datasets used, are uploaded to the Github and Zenodo repositories, respectively.

Furthermore, we have expanded the main text discussion on the choice of deep learning methodology, including explanations of the network architecture (Lines 238-246). The main text now emphasizes the design decisions and reasoning behind NN4CAST, while detailed code examples for reproducibility are provided in the Supplementary Material.

In further agreement with the other reviewers, the results presented in Figure 3 c and d do not seem particularly convincing to this reviewer, and do not seem to highlight the value of model-based attributions. In the eyes of this reviewer, the composite importances identified by the XAI methods have very low amplitudes and don't show physically interpretable structure or coherence. How would the authors explain this? Furthermore, why is the data first filtered for El-Niño events, and how is the threshold chosen?

We thank the reviewer for this comment and acknowledge the concerns raised. Regarding the relatively low amplitudes of the attribution values, this is a consequence of the high spatial resolution of the predictor field: each grid cell represents one individual input of the model, so its contribution is necessarily small in magnitude. However, when aggregated across regions, these contributions add up to match the predicted signal. A potential extension, which we consider an interesting avenue for future work, would be to spatially aggregate attribution values into larger regions in order to better quantify their relative contribution to the overall forecast.

As for the filtering and thresholding criteria, the filtering was applied because our objective is to analyze interannual variability, which requires isolating this component from lower-frequency variability before the composites are computed. The threshold of ±0.5 standard deviations was adopted consistently across all indices in order to include both moderate and strong events, thereby ensuring a larger and more representative sample size for the composites.

In the revised version of the manuscript, we now compute composites based on the predicted WTNA and SMSCU indices rather than conditioning exclusively on ENSO events. This change provides a more direct link between the model outputs and the attribution maps. In the revised version, we also include large-scale dynamical fields

(SLP, Z200, and surface winds) in the composites, which allows us to analyze how the attribution patterns relate to changes in atmospheric circulation in each case (see Section 4.2 of the new version).

More specific comments

Line 8: What do the authors mean by the 'original files'? Especially since this is in the abstract, a more specific term should be chosen.

We thank the reviewer for this comment. In the revised abstract, we have replaced the ambiguous phrase “original files” with “starting from the raw datasets” to clarify that NN4CAST operates directly on the unprocessed input data.

Line 59: It should be noted that this paragraph talks about AI models at weather timescales.

We thank the reviewer for this comment. We have clarified in the text that this paragraph specifically refers to AI models applied at weather timescales. (Lines 57–59).

Line 67: "The use of DL models to assess seasonal forecast is not so common" - Aside from the spelling error, this statement is very vague. Given the vast emerging literature on deep learning for seasonal forecasting, examples should be cited here, or the sentence should more specifically say what DL models have not been used for.

We thank the reviewer for this comment. In the revised text, we have clarified the statement and provided references to illustrate the limited use of DL in certain aspects of seasonal forecasting. Specifically, we now emphasize that most existing DL studies focus on individual phenomena or regions, rather than providing general-purpose, interpretable models that can handle multiple teleconnections. (Lines 60–67).

Line 132: It would be valuable to state why this method is chosen over others.

Thank you, we have clarified the use of this method due to its properties of sensitivity and implementation invariance, as stated by Sundarajan et al (2017). (Lines 123–125).

Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In International conference on machine learning (pp. 3319-3328). PMLR.

Line 133: "This method addresses the issue of non-linear problems, where the derivative of the output with respect to the inputs is not constant." This sentence is a bit too vague and slightly misleading - other XAI methods address non-linear problems as well, and Integrated Gradients can be applied to linear problems as well.

We thank the reviewer for this comment. In the revised text, we have clarified why Integrated Gradients (IG) is chosen, emphasizing its ability to provide axiomatic, theoretically grounded attributions that satisfy sensitivity and implementation invariance, which makes it particularly suitable for analyzing complex, high-dimensional climate predictors. We also revised the explanation to avoid implying that IG is uniquely applicable to non-linear problems, clarifying that it can be applied to both linear and non-linear models. (Lines 125–128).

Line 167: It is unclear to this reviewer what bullet point one intends to state. Furthermore, points 1-4 would be addressed by a regularized linear regression model as well - it would be valuable to include in this list why a deep learning approach is chosen here.

We thank the reviewer for this comment. We have added a clarification in the text emphasizing that, compared to a simple linear regression model, NN4CAST leverages deep learning to capture complex, nonlinear relationships and spatial interactions, which cannot be fully addressed by linear approaches. (Lines 158–170).

This reviewer is a non-English native speaker and appreciates the difficulties in writing in a second language. However, the paper would benefit from grammatical corrections, including but not limited to the following:

> *Line 1: 'being the changes in tropical sea surface temperature the most influential drivers'*
> *Line 190 "By this way it avoids to introduce"*

We thank the reviewer for this comment. We have carefully reviewed the text and corrected the identified grammatical issues, as well as other minor errors throughout the manuscript, to improve clarity and readability.