Response to Reviewer 1

October 20, 2025

Detailed Comments
Responses are marked in blue.

General Comments:

The authors provide a tool that may be utilized to research seasonal predictability using basic deep learning methods. The code library provides a pipeline to preprocess data, train the model, evaluate, and calculate some metrics/attributions, based on a user-defined namelist and input files. Although the model will not achieve state-of-the-art skill, it does have potential for mechanistic studies through explainable AI. However, I do not believe the manuscript in its current state effectively communicates this message.

*1. The analysis of the teleconnection between DJF Pacific tropical SST and MAM tropical Atlantic SST and related evaluation of the model is not valid, due to the region of the input predictor field, which includes parts of the western tropical Atlantic. Looking through the individual Integrated*Gradient attribution samples on Zenodo, it is clear that the largest attributions are most often in this area, rather than in the tropical Pacific. This is also confirmed by calculating correlations between areal-averaged SST in the target WTNA or SMSCU region with the input SST field. This leads to unrealistic, inflated skill in Figure 2, which is a result of the inclusion of the west Atlantic in the input fields, rather than the Pacific-Atantic teleconnection, as stated in the text (line 266-267).*

We thank the reviewer for this insightful comment. We agree that including the small region of the Caribbean in the predictor field could introduce information that is more directly related to the persistence of the predictand. To assess the impact of this effect, we performed an additional experiment in which the Caribbean region has been masked in the predictor field. In addition, a complementary mask has been also applied to the predictand field to exclude areas in the Pacific. This allows us to evaluate the model skill over the Caribbean without any overlap between predictor and predictand domains.

Importantly, this adjustment did not require modifications to the core model code, as the masking can be applied directly at the preprocessing stage of the SST datasets before they are introduced into the model pipeline.

The results of this experiment show that the model skill is slightly reduced in certain areas around 20ºN, such as the Gulf of Mexico, but maintains the spatial structure with correlation scores over 0.6-0.7 and similar RMSE (Figure R1) than in the original experiment.

In the revised manuscript we have updated figure 2 to include the results of this new experiment instead of the original experiment. In addition, the outputs of the model on

Zenodo have also been updated to incorporate these new simulations, ensuring full reproducibility of the results.

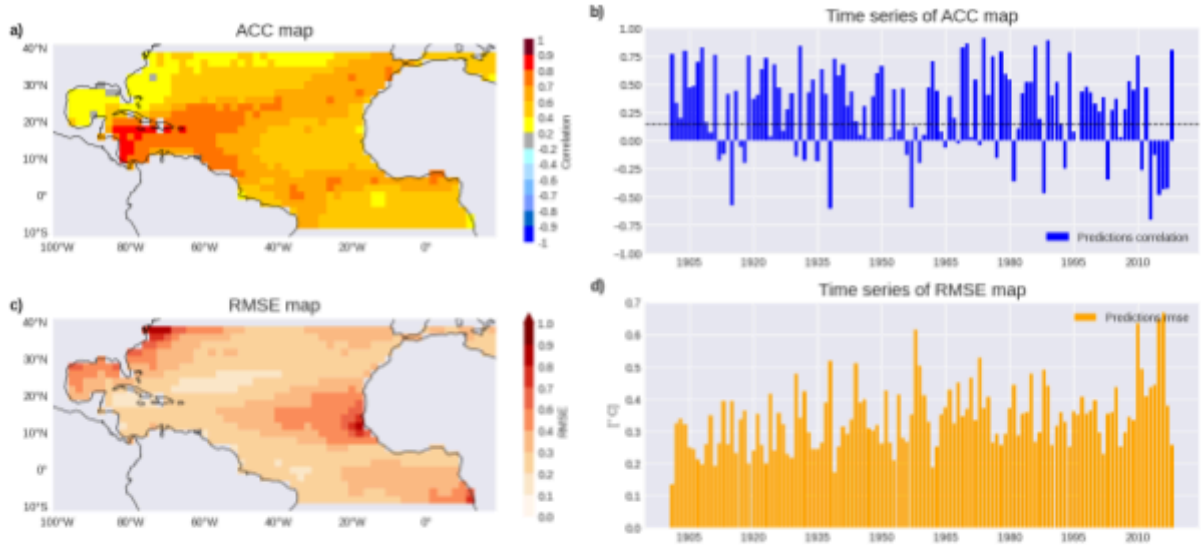

Figure R1. Predictability of tropical North Atlantic SST variability from tropical Pacific anomalies. Panels showing model performance metrics over the full period (1901–2019) using a leave-one-out cross-validation approach for predicting the SST anomaly field of the tropical North Atlantic during MAM, 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).

2. The discussion surrounding XAI in Figure 3 is unconvincing. Although the model attribution plot (Fig 3c) shows more spatial variability than the simple regression (Fig 3e), this does not necessarily mean there is added value. The work would benefit from further exploring the physical mechanisms associated with the Integrated Gradients attribution. There is not a clear connection between the spatial variance in Fig 3c and the citation of Wade et al. 2023 in the text.

We are very grateful for these detailed and constructive comments. They have prompted us to deepen our analysis and to clarify important points regarding mechanisms, robustness, and sample dependence.

First, concerning Wade et al. (2023): their study shows that SST variability in the Senegalese upwelling is connected to Pacific ENSO anomalies, particularly during 1960–1990 (their Figure 6). This coincides with our results (Figure R1b), where the model skill is higher in those decades, and with anomalies in the Pacific having maximum values on the central Pacific when the model predicts a warming in the region of the upwelling Figure R4d).

In addition, attribution results reveal that areas located in the central (180ºW–150ºW) and easternmost equatorial Pacific (110ºW–90ºW) significantly contribute to SST anomalies over the SMSCU (Fig. 3c). The reference to Wade et al. (2023) is made precisely because they identify this equatorial Pacific signal as a key remote driver of Senegalese SST variability (see their Fig. 6). In our case, NN4CAST shows significant skill throughout the century, with higher scores during 1960-1990 (Fig. 2b), coinciding with the period in which Wade et al. found a strong relation between coastal upwelling and central Pacific SSTs (their Fig. 3). The Integrated Gradients attribution, confined mainly to this region, therefore confirms the central equatorial Pacific as a key remote driver of SST variability in both WTNA and SMSCU. This has been clarified in the new version of the manuscript (Lines 272-280).

Wade, M., Rodríguez-Fonseca, B., Martín-Rey, M., Lazar, A., López-Parages, J., & Gaye, A. T. (2023). Interdecadal changes in SST variability drivers in the Senegalese-upwelling: the impact of ENSO. *Climate Dynamics*, *60*(3), 667-685.

What is the sample size? There is only a ~100 year record that is being used, with even fewer El Niño's, so I am skeptical of the robustness of model attribution. How much does the attribution pattern change with different initial seeds?

Regarding the sample size: our training period covers approximately 120 years (1901–2019). Within this span, there are on the order of 35-40 El Niño and 45-50 La Niña events, meaning that the strongest events remain relatively few. However, 100–120 years of observational record is generally considered sufficient for studies of climate variability (e.g., Trenberth (1997); Ray & Giese (2012)). Furthermore, the NN4CAST model has been tested against observations to reproduce known teleconnection patterns, providing confidence in the robustness of the attribution results. In the revised manuscript, the composites are constructed based on the WTNA and SMSCU indices rather than directly on ENSO events, which reduces dependence on the relatively few strongest ENSO events.

Concerning variation with initial seeds: we have now conducted ten simulations in which the only difference is the random seed (which we have made explicit as a hyperparameter in the library; previously it was implicit). Thus, seed variability does not appear to compromise the robustness of the main attribution findings. To illustrate the spread in the results according to initial seeds, we have computed the longitudinal and latitudinal averages of importances across the predictor field for the positive events of WNTA for the different initializations (see Figure R2), as well as the spatial average of the importance for those positive events across the model initializations (see Figure R3). The results show that while small-scale details vary somewhat across seeds, the large-scale attribution patterns remain stable. In particular, although there is some variability in the sign of attributions in certain regions, in the areas that the models assign the highest importance for their predictions, the attributions are consistent across all

initializations. Notably, the central Pacific consistently emerges as an important region (Figure R3), in agreement with the previous results (Figure 3).

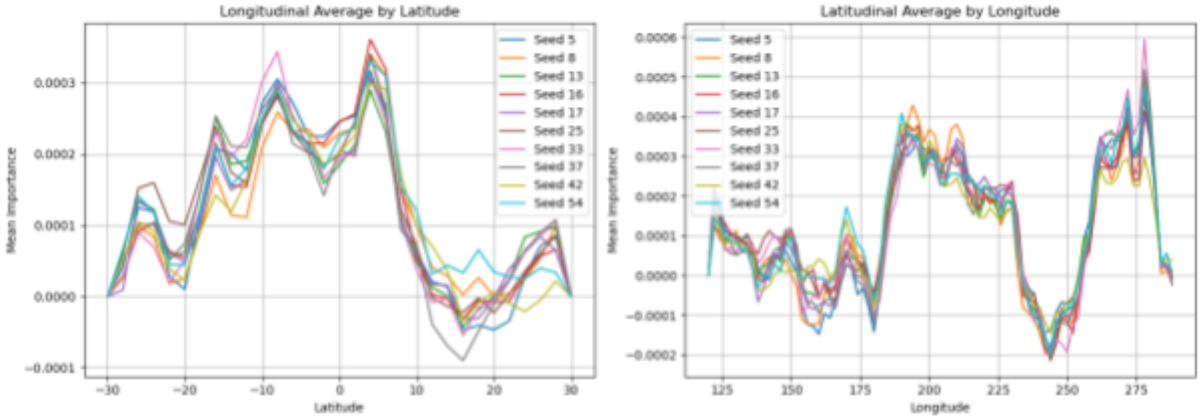

Figure R2. Mean longitudinal and latitudinal distributions of variable importance for predicting positive WTNA events. The values represent averages across 10 independent model initializations obtained by varying the random seeds.

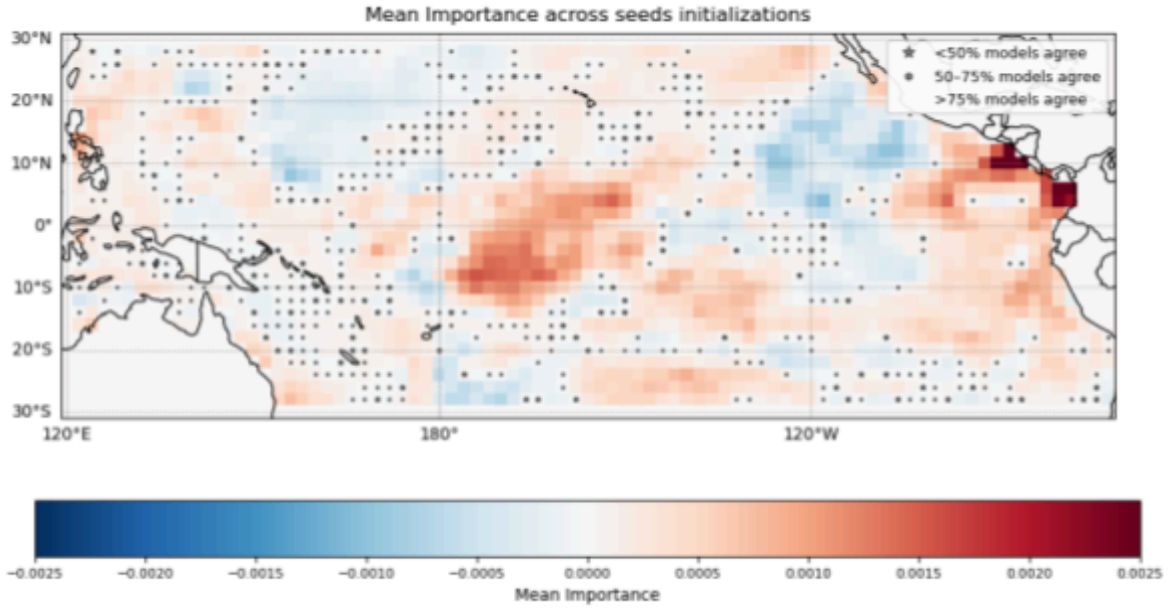

Figure R3. Spatial distribution of the mean importance of each predictor for forecasting positive WTNA events. The values are obtained by averaging across multiple model initializations with different random seeds. Markers indicate the level of agreement in sign among models: '*' for less than 50% agreement, '.' for 50–75% agreement, and no marker for more than 75% agreement.

Trenberth, K. E. (1997). The definition of el niño. *Bulletin of the American Meteorological Society*, *78*(12), 2771-2778.

Ray, S., & Giese, B. S. (2012). Historical changes in El Niño and La Niña characteristics in an ocean reanalysis. *Journal of Geophysical Research: Oceans*, *117*(C11).

Have you tried calculating attribution plots, compositing on a warm WTNA or SMSCU, rather than ENSO?

Following your suggestion, we have computed composite maps based on the predicted Atlantic indices (WTNA and SMSCU) in addition to the ENSO-based composites originally reported. These new composites are presented in Figure R4. In the first column of Figure R4 we show composites of the model SST predictions for both indices (panels a and c) together with surface wind anomalies for those events in MAM, to examine local dynamical changes that may underpin the mechanism. For example, SMSCU-positive composites reveal a local strengthening of southwesterly winds, blowing along the Senegalese coast, which can contribute to a reduction of the coastal upwelling and strong coastal SST warming (Figure R4d). In contrast, the WTNA box shows weaker wind anomalies close to the African coast, consistent with the weaker predicted SST signal (Figure R4a). To better understand which regions contribute to the SST signals in each index, we show both the predictor-field composites and the attribution maps. For both WTNA and SMSCU, the central Pacific emerges as an important region in the attribution maps, specifically around 170ºE-150°W, indicating that the model often leverages Pacific-centered anomalies when predicting these Atlantic indices (Figure R4b,d).

To understand the physical mechanism and atmospheric pathways and to corroborate the robustness of the relation found with ENSO, we have generated two additional composites of anomalous surface wind, mean sea level pressure (SLP), and geopotential height at 200 hPa (Z200) (Figure R4, panels e and f). In these fields, the region highlighted by the model in the central Pacific corresponds to an area of pronounced wind convergence and an Gill atmospheric response to an equatorial warming (Gill, 1980), which is characterized by 2 symmetric anticyclones at both sides of the equator in upper levels This tropical atmospheric response is part of a broader wave response, which propagates to the extratropics towards the Atlantic as an extratropical Rossby-wave, producing a negative NAO like pattern over the North Atlantic (more clear for WNTA events). This associated weakening of the subtropical high pressure system during a negative NAO weakens the trade winds over the TNA region (Figure R4g,h). This physical mechanism linking central Pacific SST anomalies to the tropical North Atlantic indices is in accordance with the literature (Horel & Wallace (1981); Czaja et al. (2002)). We have also computed analogous composites for negative phases of WTNA and SMSCU with consistent results. These are included in the revised manuscript for completeness. In addition to the extratropical Rossby wave, a secondary Gill response also appears over the equatorial Atlantic, as a result of the anomalous upper level convergence from the anomalous Walker circulation. This signal, which is baroclinic, also contributes to the weakening of the trades and upwelling, in agreement with García-Serrano et al (2017). The difference between WTAN and MSCU is the extension of this Gill response, which is more regional for the MSCU

These WTNA/SMSCU-based composites are complementary to the ENSO-based analysis shown in the previous version of the manuscript. ENSO-conditioned composites provide information about the model performance of the Pacific-Atlantic teleconnection

specifically under ENSO events, whereas the Atlantic-index based composites show which remote features the model exploits to predict Atlantic indices independently of ENSO. To maintain clarity of presentation, we have focused on the WTNA/SMSCU composites in the main text (Figure R4). We agree with the referee that the index-based composites provide valuable information about the key predictor regions of SST for WTNA and SMSCU. In the revised manuscript, the ENSO-based composites are presented in the Supplementary Material, while selected WTNA/SMSCU composites are shown in Figure 3, allowing readers to compare ENSO-conditioned and Atlantic-index–conditioned attribution results.

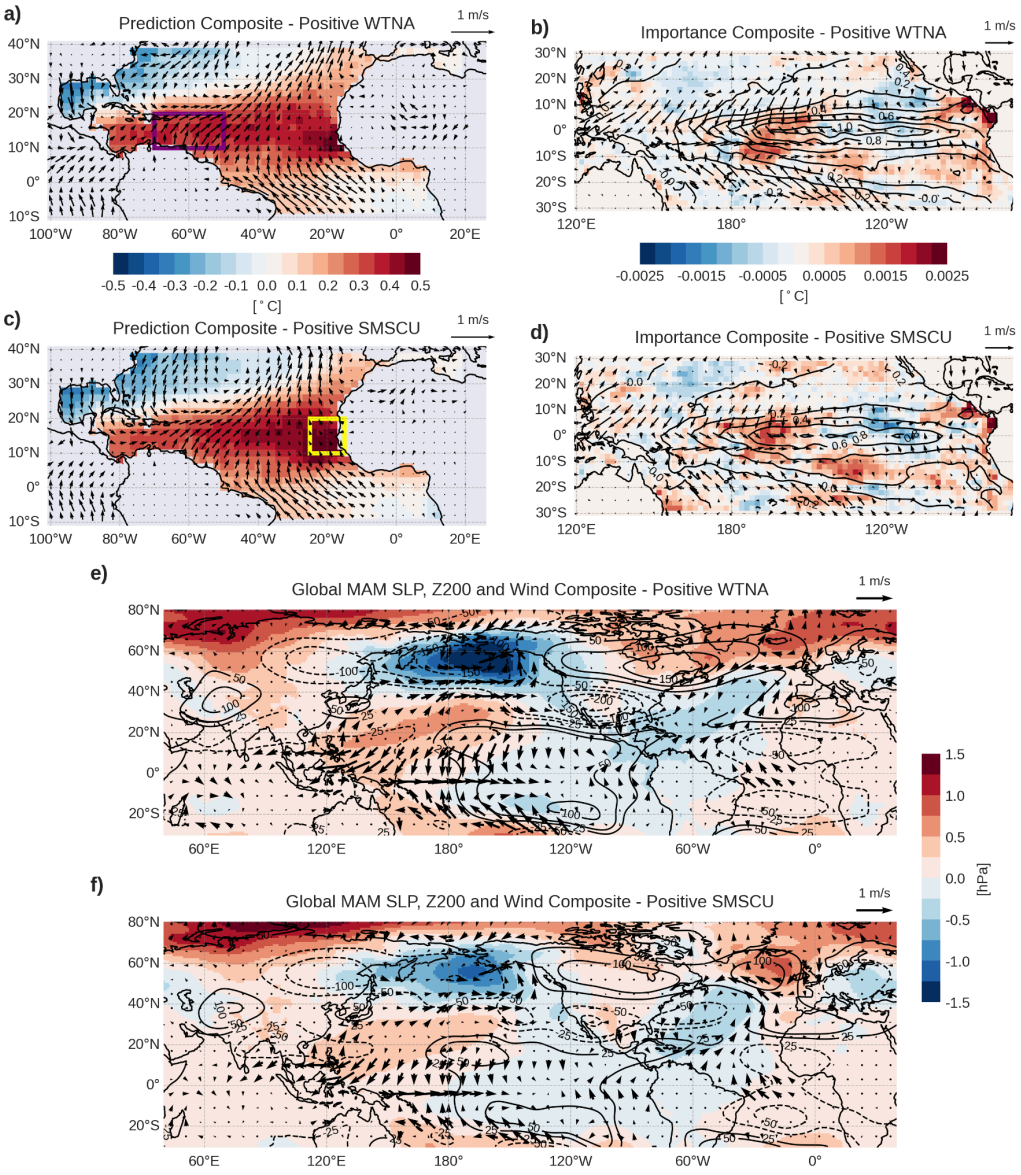

Figure R4. Composites of model anomalous SST predictions, predictor fields, and attribution maps for positive predicted WTNA and SMSCU, based on 28 and 26 events, respectively, during the period 1901-2019. Panels a) and c) show the predicted mean SST anomalies in the Atlantic during MAM together with surface wind anomalies indicated by arrows. Panels b) and d) show the attribution maps over the predictor fields with SST in contours and surface winds in arrows. Panels e) and f) display global composites of MAM anomalies in sea level pressure (shading), 200 hPa geopotential height (contours), and surface winds for positive WTNA and SMSCU events. Attribution maps indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region, with the sum of the values within each map matching the predicted anomaly in the corresponding index region (i.e., the sum of values in panel c matches the WTNA anomaly within the purple box in panel a).

Gill, A. E. (1980). Some simple solutions for heat-induced tropical circulation. *Quarterly Journal of the Royal Meteorological Society*, *106*(449), 447-462.

Horel, J. D., & Wallace, J. M. (1981). Planetary-scale atmospheric phenomena associated with the Southern Oscillation. *Monthly Weather Review*, *109*(4), 813-829.

Czaja, A., Van der Vaart, P., & Marshall, J. (2002). A diagnostic study of the role of remote forcing in tropical Atlantic variability. *Journal of Climate*, *15*(22), 3280-3290.

García-Serrano, J., Cassou, C., Douville, H., Giannini, A., & Doblas-Reyes, F. J. (2017). Revisiting the ENSO teleconnection to the tropical North Atlantic. Journal of Climate, 30(17), 6945-6957.

3. The analysis of European precipitation is useful for showing how the predictability varies between different periods. However, the regression analysis in Figure 6 is a little confusing, as you could perform the exact same regression with only observational data, yielding more faithful results and yielding the same conclusion regarding ENSO and European precipitation. Figure 5 shows the model can reproduce some of the same trends as observations, but doesn't reveal any new insights not available from solely observations.

We thank the reviewer for this comment. The primary goal here is to assess whether the model can reproduce the variability of European precipitation and its decadal changes, including its modulation by ENSO impact. To this end, comparing the regression using model predictions with the regression using observations serves as a consistency check: it validates the model ability to capture this teleconnection and its temporal evolution (non-stationarity behavior). Importantly, the purpose of this figure is not to provide a new observational analysis, but to demonstrate that the model itself reliably reproduces these patterns under the leave-one-out cross-validation framework.

Similarly to the previous analysis, it does not seem like the model is directly capturing a connection between ENSO and European precipitation, based on the individual attribution plots on Zenodo, which mostly show the model thinks SST anomalies in the extratropical Pacific and Atlantic Ocean are important. What could maybe be useful is to look at the attribution plots for precipitation in skillful regions during 1942-1969? Maybe there is a change in the background state (e.g. the extratropical jet), which changes the

propagation of the extratropical Rossby wavetrains that affect European precipitation and thus predictability?

We thank the reviewer for this helpful comment. Following the suggestion, we analyzed composites of positive and negative events based on a predicted precipitation index over western–central Europe (purple box in Fig. R5a), where the model shows significant skill (Figure 6 in the previous version of the manuscript) for the periods P2 (1942–1969) and P3 (1970–2007) (figure R5). On this basis, we first show composites of European predicted precipitation anomalies (Figs. R5a,d and R6a,d), along with global anomalies of wind (U10, V10), geopotential height (Z200), SST, and precipitation (Fig. R5c,f and R6c,f) to analyze the teleconnection mechanisms associated with the anomalous rainfall predicted by the model. We also present the attribution maps (Fig. R5b,e and Fig. R6b,e) to assess the regions contributing to the European precipitation signal.

The results reveal distinct mechanisms between the two periods. During P2 (1942–1967), a weak La Niña appears to induce enhanced convection and positive precipitation anomalies over the Maritime Continent, which act as a source of Rossby waves and generate a Gill-type response in the upper troposphere. The circulation anomalies suggest the presence of two Rossby wave trains, one propagating westward from the eastern Pacific and another emanating from the Maritime Continent region in association with the tropical precipitation anomalies (Fig. R5c). In contrast, during P3, a strong El Niño event dominates, with a clearer Gill-type response that generates an atmospheric extratropical Rossby wave train propagating into the extratropics (Figure R5f). An analogous analysis for negative precipitation anomalies yields an approximately opposite mechanism (see Fig. R6).

The differences in the mechanisms driving anomalous European precipitation between these two periods could be related to changes in the background state (Figure R7). For example, Fig. R7b shows that P2 is characterized by a weaker meridional SST gradient in both the Pacific and the Atlantic compared to P3, resulting in a weakened and southward-displaced extratropical jet.These climatological changes could explain the differences in the teleconnection patterns observed in Fig. R5: in P2, the Pacific-Europe link is relatively weak, with the apparent Rossby wave source located over the Indochina region (due to anomalous convection), whereas in P3, a stronger meridional gradient allows a clear tropical Pacific wave source, consistent with a Gill-type response, to influence European precipitation

The attribution maps allow us to clarify the regions contributing to the European precipitation signal. In P2, the maps indicate that most of the predictive signal comes from an extratropical region around 40°N and 160°E, reflecting the weakening of the Aleutian Low. In P3, in addition to this extratropical region, a tropical Pacific contribution emerges. As noted previously, to specifically assess the predictability arising from SST anomalies alone (rather than from dynamical factors that are indirectly reflected in the SST field), the simulations could be repeated with an increased lag between the predictor and predictand fields. This approach would allow a clearer separation of the SST-driven signal from atmospheric circulation effects.

In the revised manuscript, this analysis has been clarified by not only analyzing the dynamics of the mechanisms and its relation with the importances of the model, but also highlighting the impact of the changes of the mean state in the teleconnection mechanism (Lines 335-366).

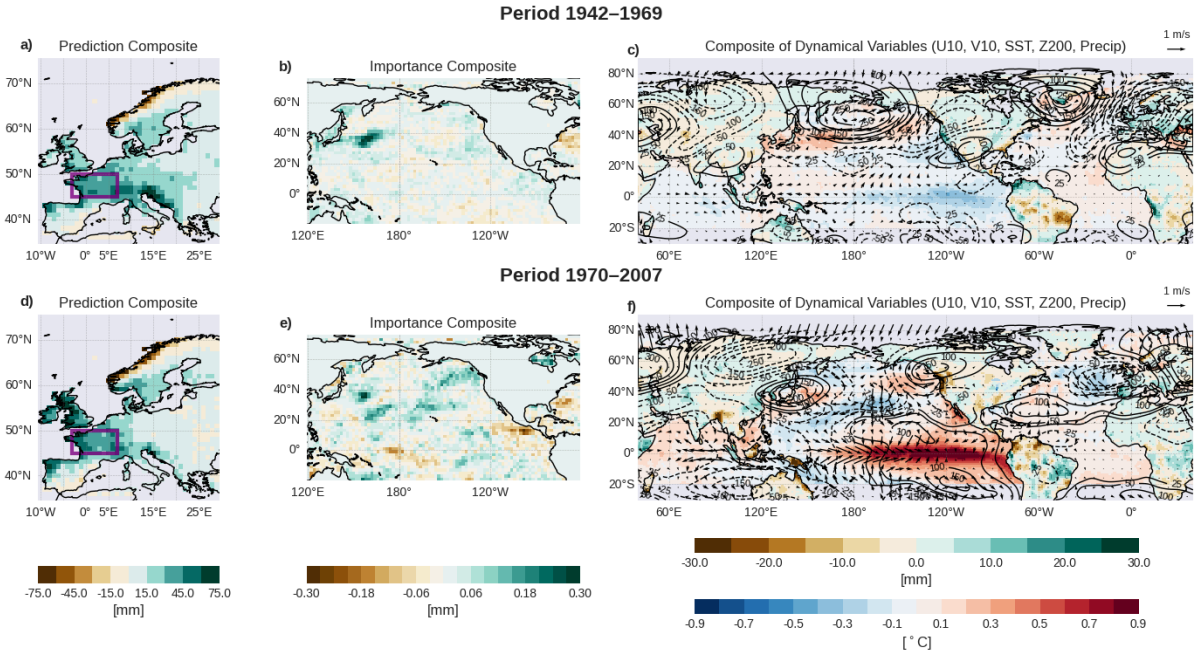

Figure R5. Composites of model anomalous precipitation predictions, predictor fields, and attribution maps based on positive predicted precipitation events in western central Europe (index defined as the purple rectangle in a)). Panels show: (a, d) precipitation anomalies in Europe during OND for period P2 [1942-1969] and P3 [1970-2007], respectively; (b, e) attribution maps over the predictor field corresponding to positive events for periods P2 and P3, respectively. Panels (c, f) display global composites of OND anomalies in sea surface temperature and precipitation (shading), 200 hPa geopotential height (contours), and surface winds for positive events for periods P2 and P3, respectively. Attribution maps (b, e) indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region. The sum of the attribution values within each map equals the predicted anomaly in the corresponding index region (i.e., the sum of values in panel b) matches the precipitation anomaly within the purple box in panel a).

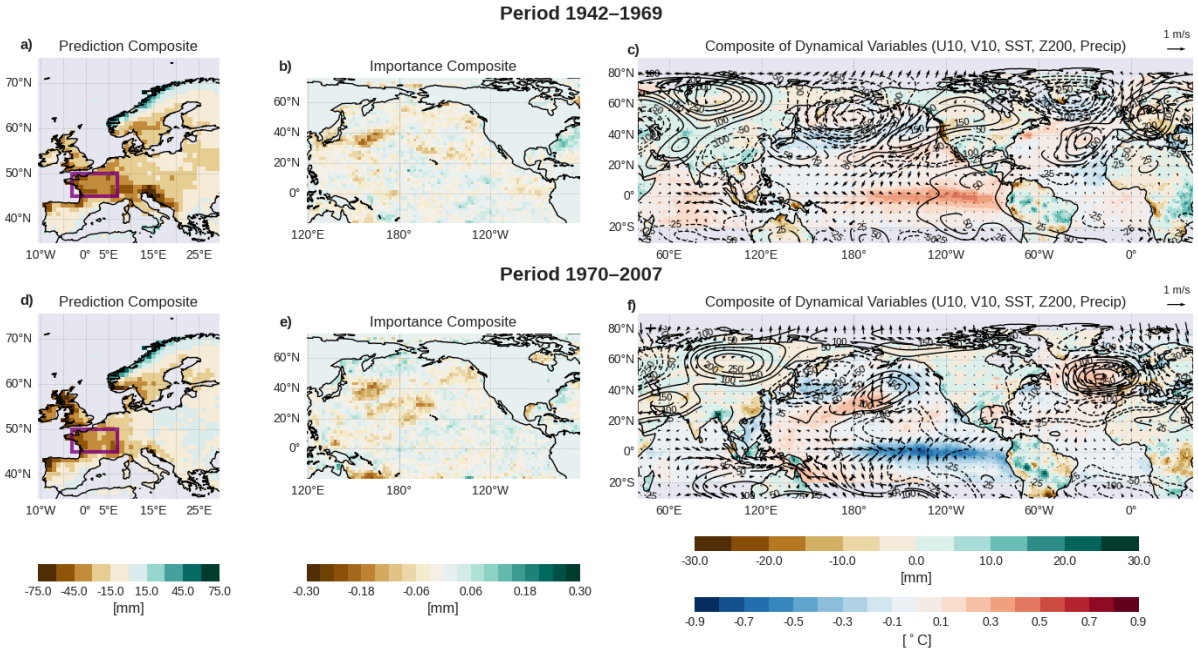

Figure R6. Composites of model anomalous precipitation predictions, predictor fields, and attribution maps based on negative predicted precipitation events in western central Europe (index defined as the purple rectangle in a)). Panels show: (a, d) precipitation anomalies in Europe during OND for period P2 [1942-1969] and P3 [1970-2007], respectively; (b, e) attribution maps over the predictor field corresponding to positive events for periods P2 and P3, respectively. Panels (c, f) display global composites of OND anomalies in sea surface temperature and precipitation (shading), 200 hPa geopotential height (contours), and surface winds for positive events for periods P2 and P3, respectively. Attribution maps (b, e) indicate the relative contribution of each grid point in the predictor field to the forecasted value in the target region. The sum of the attribution values within each map equals the predicted anomaly in the corresponding index region (i.e., the sum of values in panel b) matches the precipitation anomaly within the purple box in panel a).

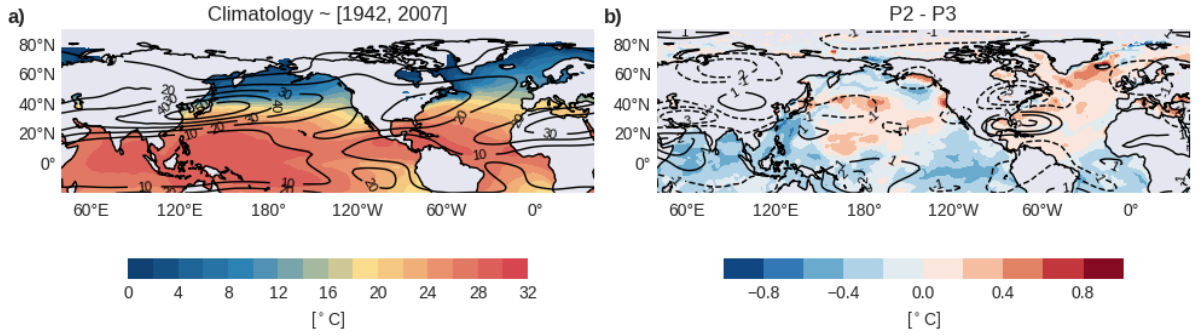

Figure R7. (a) Climatology of SST (shading) and U200 (contours) for the common period of P2 (1942–1969) and P3 (1970–2007). (b) Differences between the climatologies of P2 and P3 for SST and U200, using the same representation as in panel a).

4. In the introduction it is stated that “The idea behind NN4CAST is to mitigate the risk of treating deep learning methods as “black boxes”, thereby enabling users to identify sources of predictability and assess the sensitivity of predictions to variations in the training period and/or to the predictor region.” (line 80). However, the current manuscript does not really analyze the sensitivity to the training period or predictor region.

We thank the reviewer for pointing out this issue. We agree that the current wording in the introduction may give the impression that the present manuscript directly analyzes sensitivities to the training period and predictor regions. To avoid this confusion, we will reformulate the text to clarify that these are functionalities that the NN4CAST framework enables in general, but that they are not explored in detail in the two case studies presented here. The focus of the current manuscript is instead on evaluating model skill and attribution patterns, while the broader flexibility of the framework will be emphasized more clearly as a potential for future applications. This point is now clarified in the revised manuscript (Lines 73-77).

Specific comments:

1. The description of how the ACC is calculated could be a little more clear on what dimension is being averaged over, spatially or temporally. For when it is spatial, it is also typical that an areal weighting is applied to account for latitudinal variations in grid area.

   We thank the reviewer for this comment. In the revised manuscript, the figure captions now clarify how the ACC and RMSE are calculated, explicitly indicating whether correlations are computed across time (for spatial maps) or across space (for temporal series) (Figures 2 and 4). Regarding areal weighting, we have not applied it in the current plots. However, the model outputs are provided in a format that allows users to apply such weighting a posteriori if desired.

2. The different Listing’s showing the python code are probably somewhat redundant. It would be more useful to show what architecture is implemented, which is not easily derived from the text. For example, there is an option for convolutional layers, but how is this implemented alongside the option for dense layers?

   We thank the reviewer for the suggestion regarding the description of the network architecture, specifically the integration of convolutional layers with dense layers. In the revised manuscript, we have clarified this in the hyperparameter table (Tab. 1). Specifically, we now indicate that if at least one convolutional layer is applied, it is added at the input of the network, and its output features are flattened and concatenated with the dense layers, ensuring the combination of both convolutional and fully connected representations.

Additionally, all the Python listings previously included in the manuscript have been moved to the Supplementary Material. The main text now focuses on the description of the methodology and hyperparameters, while the Supplementary Material provides the detailed code examples for reproducibility.

3. Figure 4 and 6 colorbar is not uniformly spaced. Most of the values are near 0 or in the 0.2-0.4 range. Would be better to have separate colors for 0.2-0.3 and 0.3-0.4, to evaluate the skill.

   We thank the reviewer for this suggestion. We have accordingly updated Figures 4 and 6, adjusting the color scales to better resolve the 0.2–0.4 range and improve clarity.

4. It is stated that linear statistical models are weakened by limited observational record and nonstationarity (line 369). However, it should be clear these are also limitations for the deep learning model.

   We thank the reviewer for this comment. We have revised the sentence in the manuscript to clarify the advantages of NN4CAST. (Lines 373-378).

5. The subpanel titles in figures 3 and 6 are not clear. For example, “Regression predicted TNA on Niño years” could be something like “Input SST regressed against TNA index during El Niño”

   We thank the reviewer for this suggestion. We have revised the subpanel titles in Figures 3 and 6 to make them clearer and more informative.