Response to Reviewer 2

October 20, 2025

Detailed Comments
Responses are marked in blue.

The manuscript presents NN4CAST, a Python framework intended to streamline seasonal predictability studies with deep learning. The pipeline covers data preprocessing (region/season selection, anomaly computation, trend removal), model construction with regularization, cross-validation and tuning, and interpretation via an XAI module and EOF analysis. Two case studies are used to illustrate skill: Pacific SST forcing of tropical North Atlantic (TNA) SST in boreal spring, and Pacific SST forcing of European autumn precipitation. The overall aim to facilitate testing sources of predictability and attributing predictions to input regions is very relevant to climate services, but the manuscript in its current form requires revision before it is suitable for publication.

General comments

1. *The first case study (DJF tropical “Pacific” predictors-> MAM TNA SST) formally respects the lag, yet the predictor domain extends into the western tropical Atlantic. Given the well-known persistence of tropical Atlantic SST, even a narrow DJF Atlantic band can carry substantial memory into MAM and thus contribute to the high ACC shown in Fig. 2. In that sense, part of the reported skill may reflect local persistence rather than a Pacific-forced bridge. It would be helpful to clarify whether masking local Atlantic SST alters the ACC/RMSE/importance patterns.*

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 (Figure R1). 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 Figure 2 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.

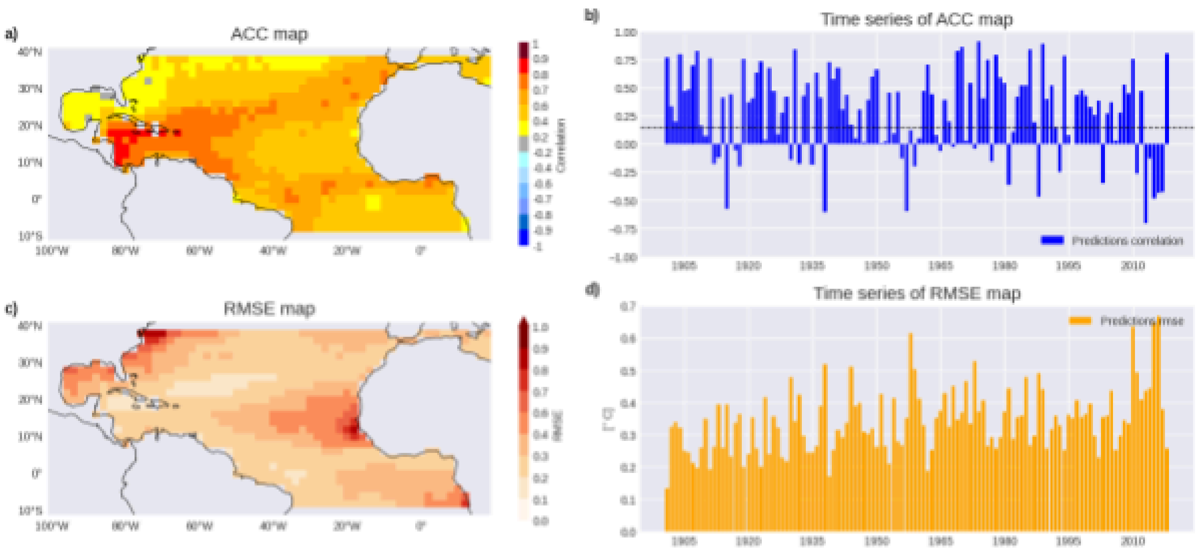

Figure R1. Predictability of tropical North Atlantic SST variability from tropical Pacific anomalies. Panels showing model performance metrics over the full period (1901–2019) using a leave-one-out cross-validation approach for predicting the SST anomaly field of the tropical North Atlantic during MAM,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

2. *In Fig. 3, the comparison between the regression composite for El Niño years (predicted TNA) and the importance composite should be improved. The fact that the attribution map contains cooling features does not, by itself, demonstrate added value, like mentioned in L280. It would help to explain how their sign and placement align with the atmospheric bridge and Wind–Evaporation–SST mechanism (e.g., stronger trades, surface heat-flux anomalies, wind-stress curl…), and whether the lead–lag structure supports that interpretation. As presented, it is difficult to separate a genuine teleconnection signal from collinearity in the SST field or residual Atlantic persistence. Showing that*

attribution hotspots co-locate with observed flux/SLP/wind anomalies, and repeating the analysis with the Atlantic belt removed from the predictors, would clarify whether the cooling patterns reflect a physical mechanism or a model artifact.

We thank the reviewer for this valuable comment. We agree that the attribution patterns in Fig. 3 must be assessed in terms of their robustness and physical consistency. To address this concern, we repeated the experiment with the western tropical Atlantic masked from the predictor field, thereby removing the potential influence of local persistence, as explained in the previous comment. The results confirm that the central Pacific remains the dominant attribution hotspot, with only modest reductions in skill in certain sub-regions of the Atlantic basin, indicating that the cooling features identified in the attribution maps are not artifacts of Atlantic memory.

In addition, and following the reviewer suggestion, we investigated whether the attribution hotspots align with established physical mechanisms. We computed composites of anomalous surface winds, sea level pressure, and geopotential height at 200 hPa, conditioned on Atlantic indices (WTNA and SMSCU) as well as ENSO phases. In the central Pacific the model highlights regions of pronounced wind convergence and SST anomalies that lead to a Gill-type atmospheric response in upper levels (2 anomalous anticyclones at both sides of the equator), which projects eastward into the Atlantic as a Rossby-wave train. This teleconnection produces a negative NAO-like circulation pattern, characterized by a weakening of the North Atlantic subtropical high pressure system and associated trade winds, thereby affecting local surface fluxes and reinforcing SST anomalies in the tropical North Atlantic.

The co-location of attribution hotspots with these observed circulation features supports the interpretation that the attribution maps capture a physically meaningful teleconnection signal rather than statistical artifacts or collinearity in the SST field. To strengthen this point, we also computed analogous composites for negative phases of WTNA and SMSCU, which show consistent patterns but with opposite signs, further confirming the robustness of the mechanism.

In the revised manuscript we now present these new composites in Figure R2, while relocating part of the ENSO-conditioned analysis to the Supplementary Material. This separation highlights how the attribution patterns remain consistent across different conditioning approaches, while allowing readers to directly compare ENSO-based and Atlantic-index–based results.

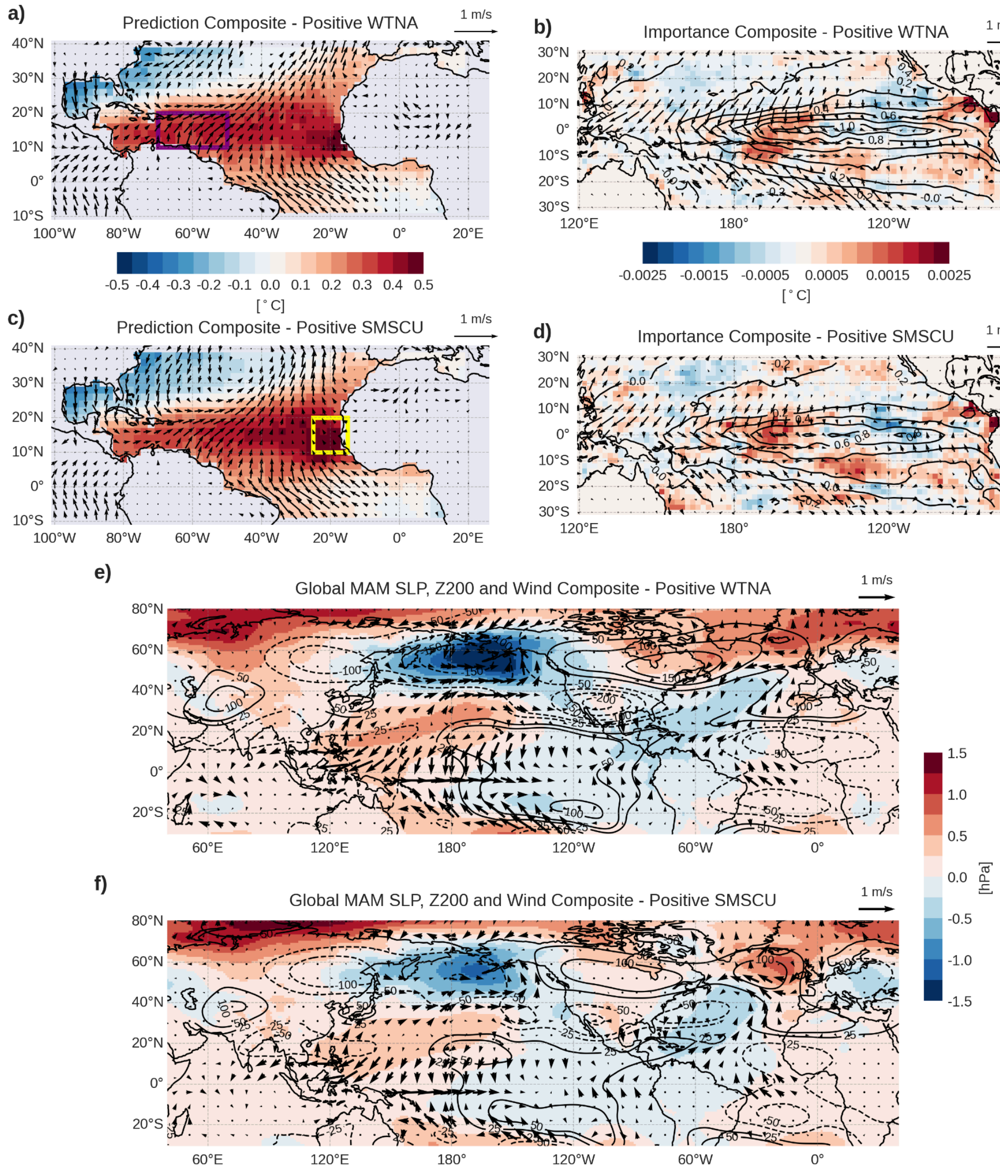

Figure R2. Composites of model anomalous SST predictions, predictor fields, and attribution maps for positive predicted WTNA and SMSCU, based on 28 and 26 events, respectively. Panels a) and c) show the predicted mean SST anomalies in the Atlantic during MAM together with surface wind anomalies indicated 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).

3. *The manuscript often uses the language of "drivers," yet the analysis is primarily associational. This matters for teleconnections, where shared low-frequency covariates can produce strong correlations without isolating a pathway. The discussion around XAI (e.g., Fig. 3) therefore could be improved. For both applications, it would be helpful to control for NAO variability and check how strongly it modulates the two teleconnections, and how this impacts the*

attribution maps. The recent literature arguing for causal-inference tools in teleconnection analysis points in this direction and could be useful (e.g. https://journals.ametsoc.org/view/journals/bams/102/12/BAMS-D-20-0117.1.xml).

We thank the reviewer for raising this important point. We acknowledge that the term “driver” can be interpreted as implying causality, whereas our analyses are primarily associational. We will be more careful and use the term potential drivers when appropriate. In the first application (Pacific SSTs - Atlantic SSTs), our use of “driver” refers to the Pacific SST anomalies during DJF that precede and are statistically linked to Atlantic SST variability in the following MAM, indicating in this way a causality. Because the predictor and predictand are separated by a seasonal lag, the relevant information comes from the Pacific SSTs, while any NAO-related signal captured by the model is likely the component externally forced by SST rather than internally generated variability.

In the second application (Pacific SSTs - European precipitation), there is no lag between the predictor and the predictand field. Thus, the model primarily identifies statistical associations rather than direct physical causality. Here, surface atmospheric circulation is partially embedded in the anomalous behaviour of the SST field, which explains why attribution maps highlight significant SST contributions from extratropical regions in addition to the tropical Pacific. We understand that it is important to discuss the impact of the NAO when explaining precipitation in Europe. To analyze the impact of the NAO variability, we have explored composites conditioned on positive and negative NAO phases (Figure R3 for NAO+). They show how the NAO has a more internal variability in P2 (where no relationship appears with the SST anomalous field), while in P3 there is a clear association with ENSO.

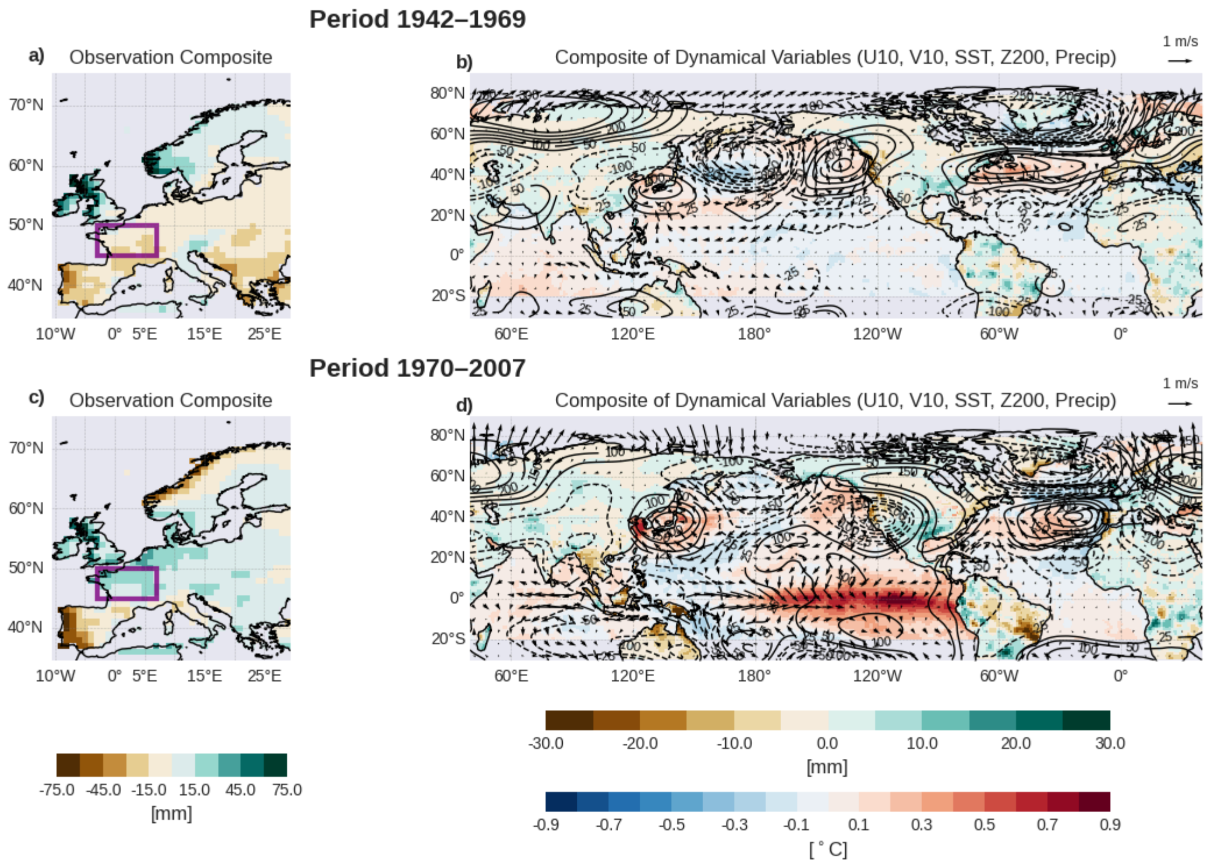

Figure R3. Composites of observed anomalous precipitation and dynamical fields, based on positive observed NAO+ events in western central Europe (index defined as the purple rectangle in a)). Panels show: (a, c) observed precipitation anomalies in Europe during OND for period P2 [1942-1969] and P3 [1970-2007], respectively; (b, d) 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.

To explore the ENSO-NAO relationship further, Table R1 identifies the ENSO, NAO and precipitation cases conditioned, in periods P2 (1940-1969) and P3 (1970-2007) where we find some predictability in our model.

Period	Index	Phase	NAO Negative	NAO Neutral	NAO Positive	Total
1942–1969	ENSO	Negative	3	4	3	10
		Neutral	4	4	4	12
		Positive	3	1	2	6
		Total	10	9	9	28
	Precip	Negative	4	3	3	10
		Neutral	3	4	4	11
		Positive	3	2	2	7
		Total	10	9	9	28
1970–2007	ENSO	Negative	5	5	5	15
		Neutral	5	3	2	10
		Positive	3	4	6	13
		Total	13	12	13	38
	Precip	Negative	5	6	1	12
		Neutral	5	5	4	14
		Positive	3	1	8	12
		Total	13	12	13	38

Table R1. Contingency table of ENSO and precipitation index phases with NAO phases for the periods 1942-1969 and 1970-2007.

For P2, we can see that there is no clear relationship between ENSO and NAO, with half of the NAO events considered ENSO neutral (Table R1). However, for P3, 75% of NAO-like events were related to an ENSO phase of opposite sign (i.e., positive (negative) NAO-like is related to El Niño (La Niña)). This can be seen more clearly in Table R2 on conditional probability. In the P2 period, there is greater internal variability of the NAO, or in other words, more NAOs not associated with ENSO (p=0.42), therefore, in principle, not forced by SSTs in the tropical Pacific and less predictable in our model. However, in P3, the probability of finding NAO-like +/– associated with El Niño/La Niña increases significantly compared to neutral years (i.e. 0.45 in P3 vs 0.22 in P2, Table R2).

Regarding the impact of NAO+ and NAO- on precipitation, First of all, it should be noted that the precipitation pattern associated with the NAO (Figure R3) is very different from the first mode of precipitation variability in Europe identified by our model (Figure 5 in the manuscript). However, central Europe (i.e. the French box) appears to be affected in both maps.

The fact that NAO+/- in P3 are more strongly influenced by ENSO also has an impact on precipitation. The probability of positive precipitation associated with a NAO+ increases in P3 compared to P2 (i.e. 0.67 vs 0.29; Table R3) – Even so, 10% of precipitation variability is not explained by NAO (in P2 and P3, 0.29 and 0.30 respectively, probability of finding anomalous precipitation with neutral NAO). This stronger ENSO-NAO-like and precipitation consistency in P3 is in line with the predicted precipitation anomalous maps (in Fig. 6 of the updated manuscript) and for the observed precipitation (Fig. R4). The composite of positive precipitation patterns in P3 shows a Niño-forced structure with negative SLP anomalies over Iberian Peninsula and British Isles (Fig. 6f and Fig. R4d). This pattern, although not strictly NAO positive, shares pressure anomalies in centres close to the canonical NAO index (Fig. R3d). However, the pattern in P2 shows a circumpolar wave that appears to originate in the Maritime continent region (with positive precipitation anomalies, which can induce an atmospheric wave, as explained in the text) and which projects onto the North Atlantic in a pattern distinct from the NAO (Fig. 6c and Fig. R4b compared with Fig. R3b). We have clarified all this analysis in the updated manuscript (Lines 361-366).

Conditional probability	**1942–1969**	**1970–2007**
p (Niño | NAO +)	0.22	**0.45**
p (Niña | NAO –)	0.30	**0.38**
p (ENSO neutral | (NAO + ∪ NAO –))	0.42	0.29
p (NAO +)	0.32	0.29
p (NAO –)	0.36	0.34

Table R2. Conditional probabilities between NAO and ENSO phases, based on the cases of Table R1.

Conditional probability	**1942–1969**	**1970–2007**
p (NAO+ | Pt+)	0.29	0.67
p (NAO- | Pt-)	0.40	0.42
p (NAO neutral | (Pt+ ∪ Pt-))	0.29	0.30
p (Pt+)	0.25	0.32
p (Pt-)	0.36	0.32

Table R3. Conditional probabilities between Precipitation index and NAO phases, based on the cases of Table R1.

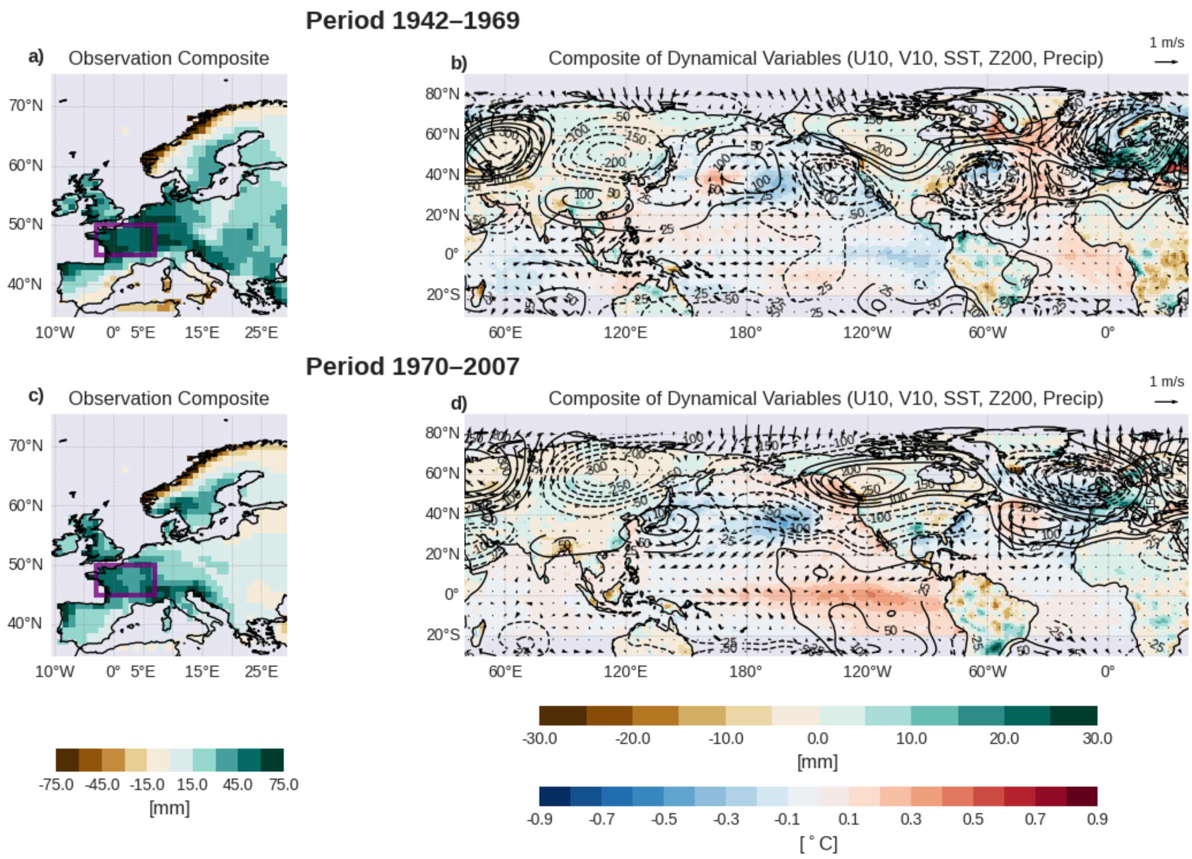

Figure R4. Composites of observed anomalous precipitation and dynamical fields, based on positive observed anomalous precipitation events in western central Europe (index defined as the purple rectangle in a)). Panels show: (a, c) observed precipitation anomalies in Europe during OND for period P2 [1942-1969] and P3 [1970-2007], respectively; (b, d) 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.

4. *The manuscript describes the toolkit as "versatile," yet for identifying dominant spatial modes it offers only EOF analysis of model outputs versus observations. For teleconnection work, this is a narrow diagnostic. At minimum, a versatile layer would include a menu of spatial-mode tools beyond EOF (e.g. maximal covariance or canonical correlation analysis for coupled patterns). Please consider either expanding the diagnostics accordingly or reframing the package as a DL-first pipeline with basic (EOF-based) spatial diagnostics.*

We thank the reviewer for this valuable comment. We agree that for the analysis of teleconnections, EOFs provide only a narrow diagnostic. Our main objective with this toolkit, however, is to enable the modeling of teleconnections through deep learning. The EOF analysis is included primarily as a preliminary diagnostic, while a comprehensive implementation of additional techniques such as Maximum Covariance Analysis (MCA) or Canonical Correlation Analysis (CCA) falls outside the scope of the present work. We note that there are existing packages, such as xMCA (Rieger, 2021) ,Xcast (Hall et al., 2022) or Spy4Cast (Duran et al., 2024), that already offer these functionalities. We therefore see NN4CAST as complementary to such tools: outputs from the simulations produced by our package can be readily used as inputs to these other libraries for further, more specialized, spatial-mode analyses. To avoid overstatement, we will revise the wording in the manuscript, with a clearer description that emphasizes this complementary role. We have clarified this in the updated version of the manuscript (Lines 6-8).

Specific comments:

L1 (...) with the changes in tropical sea surface temperatures (SST) being (...)

We have corrected this.

L8 Please be more specific than writing "(...) performs all the methodological steps".

We have specified these steps. (Lines 8–10).

L27 you already defined SST in the abstract. If you decide to define again, please use lower case as in the abstract.

We have corrected this.

L75-77 Here you introduce the tool for the first time, after a long introduction on seasonal forecasting and ML. I suggest bringing up the goal/what's new about your paper much earlier in the introduction, to help the reader to situate themselves.

We thank the reviewer for this suggestion. In the revised version of the paper, we have substantially reduced the introduction of Artificial Intelligence and Deep Learning, streamlining the background so that the reader can reach the goal and novelty of the paper much earlier.

L86 I don't think bringing up the possibility to combine NN4CAST with ESMValTool in the introduction is relevant. I think this could be mentioned in the conclusions/future work.

Thank you, we have moved this argument to the conclusions.

L89 Please give an example of such tools written in C/C++.

Thank you, we have added as an example tool the Climate Data Operator (CDO)

L92-103 I recommend not giving so many details of the applications to be analysed in this final introduction paragraph. Similarly, mentioning GitHub and code availability here seems misplaced.

We thank the reviewer for this comment. In the revised version, we have streamlined the final paragraph of the introduction, focusing on the main goal and contributions of NN4CAST without including detailed descriptions of the applications or code availability.

L337 I think the sentence should be rewritten, as significant skill is not found in most of the European continent, rather in parts of it.

We have modified the sentence according to the comment.

L375 and L380 are repeated

We have corrected this.

L380, L395 The authors mention that the tool has a primary application to identify windows of opportunity (WoO). However, in the two applications given, there was no framing related to WoO. I recommend improving the discussion towards the context of WoO.

We thank the reviewer for this observation. Following the suggestion, we have clarified the connection between NN4CAST outputs and the concept of windows of opportunity (WoO) in seasonal forecasting. In the European precipitation case study, period P2 (1942–1969) serves as an example of WoO, where the model shows high skill for predicting precipitation. By highlighting this period, NN4CAST can help identify time intervals where predictive skill is higher, providing useful insights for seasonal forecast applications. (Lines 402–406).

L388-L389 I suggest to focus on more specific advantages offered by the NN4CAST in your conclusions. "These complementary approaches offer valuable contributions to the scientific community and support the improvement of current seasonal forecasting systems" seems a bit vague and exaggerated at the same time. In particular for the first application, the authors did not go in depth to highlight any new insights concerning the teleconnection, rather used it as an example to illustrate what the tool does.

We thank the reviewer for this comment. Following the suggestion, we have revised the conclusion to emphasize the specific advantages of NN4CAST. The tool not only provides robust seasonal forecasts through cross-validation, but also systematically identifies the predictor regions contributing to target indices, quantifying their relative importance using explainable AI techniques. Importantly, the analysis of attribution maps allows linking predictive importance to known physical mechanisms, such as ENSO teleconnections. (Lines 402–408).