Review 1

Please find below the answers to the comments point-by-point. For clarity, all reviewer comments are in **black** and responses in **blue**. Modifications in the manuscript are indicated with underlined text.

In this study, a methodology to produce precipitation predictions over France, at high spatial resolution and at the decadal temporal horizon, is presented and evaluated. The method consists of five sequential steps (index selection, index prediction, subsampling, downscaling, and assessment). Key findings include significant skill enhancements over the extended winter season compared to uninitialised simulations, and weaker but significant improvements over the extended summer season.

This work focuses on a specific domain (France), but the authors highlight that their procedure could be adapted to other geographical regions, which I agree with: for this reason, I believe this study fits well within ESD's aims and scope. The methodology is novel, and the results presented are substantial and potentially relevant to stakeholders.

The authors put effort into motivating their methodological choices (e.g. Section 3.4), but the organisation of the material could be streamlined and the level of information provided improved in places. In a couple of cases, further discussion would help the interpretation of the results. Finally, this work requires careful editing as it currently contains several grammatical errors. That being said, I did not find major issues with the manuscript and my overall assessment is positive.

We are grateful to the reviewer for the careful reading of our manuscript and for the constructive comments, which have helped us to substantially improve the clarity and rigor of the work.

Before addressing the specific comments, we would like to note that while preparing the reviews, we have identified a small error in the computation of 8-year mean summer precipitation for both observations and uninitialized projections, resulting in a one-year lag relative to the decadal predictions. This issue has been corrected. All analyses, figures, and associated text have been updated accordingly. This correction slightly affects the quantitative skill scores for summer precipitation predictions (former Figs. 9 and 10), although the overall conclusions of the study remain unchanged. All modifications related to this correction are clearly highlighted in the revised manuscript and discussed where relevant below.

Specific comments

1. Eq.2 There are two free parameters appearing, alpha and beta, which are not discussed in the text and are set to 1. I'd like to understand what their role is, and how the authors decided on their values. Also, since the indices are not standardised (based on definitions

and dimensionality) and have thus different variance, the loss function D(t) may penalise the index with greatest absolute error rather than that with greatest relative error. Is this a desired feature?

We thank the reviewer for pointing this out. The free parameters alpha and beta were not explored and do not add analytical value as presented; they have therefore been removed from the loss function.

Furthermore, we acknowledge an error in the written form of Equation 2 in the original manuscript: the formula did not reflect the standardization that was actually already applied in the computations from the submitted version of the manuscript. Indeed, we agree with the reviewer, that it is crucial that indices with different amplitudes and units contribute equally to the selection criterion. We apologize for this oversight. The correct loss function that we used in the submitted manuscript is now included in the revised manuscript:

"$D_i(t) = \sqrt{(d1_i(t)/std(\{d1_i(t)\}_{i=1..32,\forall t}))^2 + (d2_i(t)/std(\{d2_i(t)\}_{i=1..32,\forall t}))^2}$, $\forall i \in [1,32]$ (2)

where $d_i(t) = u_i(t) - f(t)$ for each index. Standard deviations are computed over the entire historical period rather than at each individual date to provide a stable normalization that is not influenced by the small sample size at any given window."

The correction is therefore just concerning the formula, and does not affect the results, since we were already computing what is proposed by the reviewer, since this is more appropriate to have the equal weight for each index.

2. Fig. 6b The full uninitialised IPSL ensemble shows remarkably high skill for the wNAO (r=0.6). This correlation is comparable with that found in Smith et al. (2020) and Nicolì et al. (2025) for (raw) initialised decadal predictions of the NAO, and much larger than that found in Nicolì et al. (2025) for uninitialised simulations. Clearly there are differences in data and definitions, but still I wonder if the authors could comment on the result further, possibly linking with previous studies if existing.

The relatively high correlation (r = 0.6) found for the IPSL ensemble mean in uninitialized simulations is indeed noteworthy and warrants discussion.

This value is consistent with the results reported by Christiansen et al. (2022), who found a similar ensemble-mean correlation (~0.57) in the CMIP6 historical multi-model ensemble for the NAO (213 members, 1970–2015), comparable to large initialized decadal prediction systems. Klavans et al. (2021) similarly found relatively high skill (r = 0.62, 1962–2015) using the multi-model large ensemble archive (MMLEA, 269 members from six models). Together, these results suggest that a substantial fraction of NAO predictability on decadal timescales arises from the externally forced response, potentially from radiative forcing trends. The r =

0.6 value obtained here for the IPSL large ensemble is therefore consistent with this literature and does not suggest that the model is an outlier within CMIP6. We argue that the high ACC prediction skill obtained from historical simulations may be linked to the influence of aerosols, particularly anthropogenic aerosols, whose concentrations vary on decadal timescales. In the North Atlantic, aerosol forcing has declined since the 1990s as emissions shifted increasingly toward Asia. Furthermore, the increase in greenhouse gas concentrations may also have contributed to a shift of the NAO toward its positive phase, as suggested by the positive trend observed in both the historical simulations and the observations. This externally forced trend likely contributes to the high ACC obtained from the historical simulations.

Nevertheless, initialized decadal predictions allow for further improvement: boosted DCPP forecasts reach ACC = 0.84 in our study, indicating that ocean initialization provides added skill beyond the forced signal — consistent with Klavans et al. (2021), who showed that initialization may be particularly important for specific NAO events. The amplitude of the NAO signal is also enhanced in DCP, highlighting a better signal-to-noise ratio in DCP. Finally we would like to also suggest that internal variability and external forcing might be covarying, either by chance or due to synchronization (e.g. Swingedouw et al. 2013), which might explain the relatively good ACC in the historical.

This discussion is now included in Section 3.2 of the revised manuscript.

“The full uninitialized IPSL ensemble shows moderate skill with wider uncertainty and less agreement with observations (Fig. 6). The relatively high ACC of the uninitialized wNAO ensemble mean (r = 0.6) is consistent with results from comparable large historical ensembles (Christiansen et al., 2022; Klavans et al., 2021), and likely reflects a substantial contribution from the externally forced response–particularly from anthropogenic aerosol forcing and greenhouse gas-driven trends–rather than from ocean initialization alone. Boosted decadal predictions improve upon this baseline with enhanced signal amplitude, confirming that ocean initialization provides added skill beyond the forced component.”

3. L381-382 The claim that “All boosted decadal predictions ACCs outperform corresponding uninitialized IPSL ensembles substantially” is not, in my opinion, supported by the authors’ results for uAMV (see correlation values in Fig. 7 c). Furthermore, it would be meaningful to discuss the high ACC found for the uninitialised ensemble for uAMV, possibly linking with the role of the externally forced signal which the authors opted not to remove (L193-197).

We agree that this sentence lacks clarity. The statement is in fact true only for the atmospheric indices. The uAMV case is different: since we chose not to remove the externally forced signal, the high ACC scores for the uAMV are largely driven by the forced trend. The uAMV is not used as a standalone predictor but only as an additional constraint alongside the sNAO. The inclusion of these two constraints allows for a more effective selection of ensemble members that remain consistent with the predicted AMV, regardless of the underlying source of predictability. In particular, the forced signal is retained, while internal

variability within the historical simulations may otherwise obscure the response to external forcing. Furthermore, covariance between the forced signal and internal variability may also help explain why the ACC values are of similar magnitude in the initialized and uninitialized predictions.

To clarify this, the following text will replace the original sentence in the revised manuscript:

"In the case of atmospheric indices, all boosted decadal predictions ACCs outperform the corresponding uninitialized IPSL ensemble substantially. For the uAMV, the high ACC scores for both the boosted predictions and the uninitialized ensemble seem to be driven by the externally forced trend to a large extent, consistent with a few former studies (e.g. Ting et al., 2009), although covariability between internal variability in observations and forced signal might also obscure the added value of initialized runs. As discussed in Section 2.2a, the externally forced signal was intentionally retained in the uAMV definition, since it contributes to long-term precipitation variability over France. The uAMV is therefore not used as a standalone predictor but exclusively as an additional dynamical constraint alongside the sNAO, in order to better choose members from the large ensemble that include this predictable observed signal, which might be overwhelmed by uninitialized internal variability in some members of the large ensemble."

4. Fig. 8 It is surprising that the subsampling method based on wNAO yields better results than wWEPA, considering that the skill for the two indices is similar and wWEPA is associated with the leading mode of precipitation variability. The authors argue (L483-486) this is because the wNAO signal has larger fluctuations, but how exactly does this explain the difference? Is it because larger pressure variations impose a stricter constraint when sub-selecting ensemble members? Also, I am curious whether the authors have tried combining the wWEPA and wNAO indices for the subsampling procedure. To my eye, it would make sense to try combining the two indices associated with the leading PCs.

We thank the reviewer for pointing this out. The mechanism underlying the better performance of wNAO-based subsampling compared to wWEPA, despite similar index skill and wWEPA's association with the leading precipitation mode, is not fully understood. We offer the following elements of explanation.

First, Fig.R1 shows the skill obtained when the observed (perfect) wWEPA index is used directly as the subsampling criterion. This upper-bound estimate confirms that wWEPA-based subsampling has the potential to yield improved skill–suggesting that the underperformance in the standard configuration is partly attributable to imperfect wWEPA prediction rather than a weak teleconnection per se.

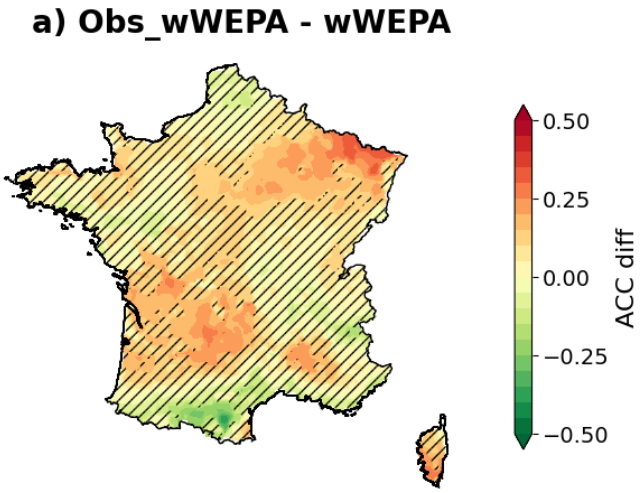

Fig.R1: Forecast skill for 8-year mean winter (ONDJFM) precipitation anomalies over France, based on subsampled forecasts using the observed wWEPA. The map shows the Anomaly Correlation Coefficient (ACC) difference with the reference prediction using wWEPA alone. Red (green) values correspond to improved (deteriorated) predictions. Scores are computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate regions where skill scores are not statistically significant at the 95% confidence level, as assessed with a Steiger's test.

Second, and more fundamentally, the skill of the subsampling method depends not only on the accuracy of the predicted index, but on how well the index and its associated SLP centers of action are represented in the uninitialized model ensemble used for subsampling. To evaluate this second aspect, we compared EOF patterns of extended winter (ONDJFM) SLP over the North Atlantic-European sector (20°-65°N, 35°W-20°E) between ERA5 (1961-2024) and the last 100 years of the IPSL-CM6A-LR pre-industrial control simulation. Both fields are computed from 8-year smoothed seasonal means to match the temporal scale of the subsampling procedure. The results are presented in a new Supplementary S10.

Figure S10 suggests that the spatial pattern associated with the WEPA index is not very well represented in the IPSL-CM6A-LR model as compared to ERA5. This difference in WEPA signature in precipitation over France may weaken the teleconnection between wWEPA and precipitation in the uninitialized ensemble, ultimately limiting the skill gain from wWEPA-based subsampling. The NAO pattern, by contrast, is reproduced more closely, which may partly explain the superior performance of wNAO-based subsampling. This result also suggests that the choice of uninitialized model ensemble is an important methodological consideration: models whose SLP climatology more accurately reproduces the WEPA spatial pattern may yield improved wWEPA-based subsampling skill. However, we prefer to leave the exploration of the impact of using different large ensembles and leave the exploration of alternative large ensembles to future studies, to keep the focus of the present study and avoid adding potentially long further analyses.

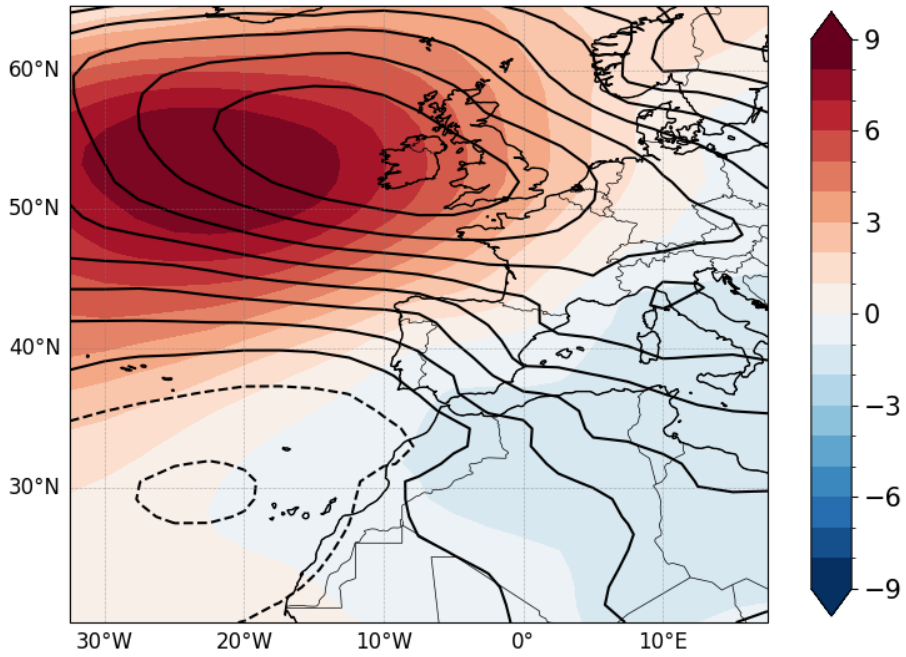

Fig. S10: Second empirical orthogonal function (EOF2) of extended winter (ONDJFM) sea level pressure variability over the North Atlantic-European sector (20°-65°N, 35°W-20°E). Color shading shows the EOF2 spatial pattern derived from the IPSL-CM6A-LR pre-industrial control simulation (last 100 years), explaining 21% of the total variance. Contour lines show the corresponding EOF2 pattern derived from ERA5 (1961-2024), explaining 17% of the variance. Both fields are computed from 8-year smoothed seasonal means to match the temporal scale of the subsampling procedure. EOF2 from ERA5 is closely associated with the wWEPA index (r=0.78). The sign of each EOF pattern is chosen such that positive values correspond to anomalously high pressure over the British Isles.

The following sentence has been modified in the discussion to propose this interpretation (section 4):

“The forecast skill of the indices themselves, as well as their representation and relationship with regional precipitation in climate models, critically influence the quality of precipitation forecasts after subsampling. For example, although WEPA-like indices are closely associated with the principal mode of precipitation variability, wNAO-based subsampling produces higher skills in precipitation predictions. This could be due to the larger amplitude of the NAO predicted signal. However, the difference between predicted NAO and WEPA remains modest, questioning whether this effect might explain all the differences found in terms of prediction of the precipitation over France. Another explanation for this paradox may lie in the model's representation of SLP variability modes (here IPSL-CM6A-LR model). In particular, discrepancies between the observed and simulated centers of action associated with the WEPA pattern could substantially affect its influence over France. To investigate this hypothesis, we performed an EOF analysis of SLP in both the ERA5 reanalysis and a long pre-industrial control (piControl) simulation from the IPSL-CM6A-LR model, which provides a sufficiently long record to robustly isolate internally generated modes of variability.

We find substantial differences in the locations of the centers of action between observations and the IPSL-CM6A-LR model (Supplementary Fig. S10). In the model, the positive pressure center is shifted northward relative to observations, likely causing the associated wind anomalies to affect the UK and Ireland more strongly than France. This contrasts with the observed WEPA pattern, whose influence is more directly felt over France.

These results highlight the potential importance of both large-ensemble selection and model biases in the representation of the underlying circulation patterns. Developing a systematic approach to account for such structural biases when selecting large ensembles could further improve the proposed framework. However, addressing this issue would considerably broaden the scope of the present study and is therefore left for future work."

Regarding the combination of wNAO and wWEPA: we conducted this test and found no significant improvement over wNAO-based subsampling alone, though an improvement over wWEPA-based subsampling was observed. The figure illustrating this test is now included in Supplementary S4. This test has been added to the revised manuscript (section 3.4):

"We also evaluated combinations of wNAO and wWEPA with uAMV (see Supplementary S4), as well as SST averages over the subpolar Gyre, but found no significant improvement. Multi-criteria subsampling based on wNAO and wWEPA was also conducted with no significant improvement over subsampling based on wNAO alone (see Fig.S4.2)."

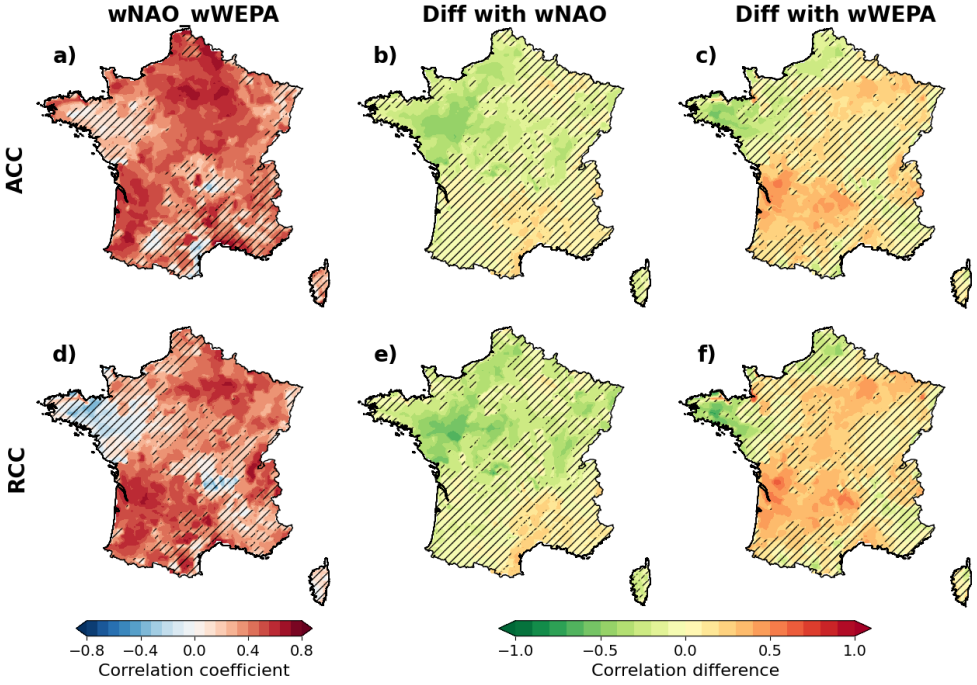

Fig. S4.2: Forecast skill for 8-year mean winter (ONDJFM) precipitation anomalies over France. Skill maps are based on subsampled forecasts using the wNAO+wWEPA (a,d). Panels show (a) Anomaly Correlation Coefficient (ACC), (b,c) the ACC difference with the reference prediction using wNAO (b) and wWEPA (c) alone. Panels (d,e,f) show the same as

(a,b,c) but for Residual Correlation Coefficient (RCC). In panels (b,c,e,f) red (green) values correspond to improved (deteriorated) predictions. Scores are computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate regions where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap (a,d) and with a Steiger's test for ACC and RCC differences (b,c,e,f).

5. Section 3.3 Spatial correlations in the target data can artificially inflate the number of significant grid-points (the multiplicity problem, see e.g. Wilks 2006), and specific statistical tests have been devised to assess the significance of fields. The authors refer to the number of significant grid-points both as a measure of relative goodness between different prediction strategies (e.g. L404-406), which may not warrant further analysis, and as an absolute measure of prediction skill (e.g. L17-19). If the authors wish to use the number of significant grid points in the latter sense, I think field significance should be assessed.

We thank the reviewer for raising this interesting important point for the rigor of our analysis. The multiplicity problem was actually not addressed in the original manuscript (as well as in a number of published articles). To correct for the effects of simultaneous multiple testing, we now apply the False Discovery Rate (FDR) approach throughout the revised manuscript, as recommended by Wilks (2016).

The following analysis has been added to the Supplementary S9 (*"Implementing the False Discovery Rate procedure"*) :

When evaluating the statistical significance of skill scores over metropolitan France, we apply the False Discovery Rate (FDR) approach to address the multiplicity problem arising from simultaneous multiple testing, as recommended by Wilks (2016). The FDR procedure controls the expected fraction of erroneously rejected null hypotheses by evaluating sorted p-values against a linearly increasing threshold rather than a fixed value, as illustrated in Fig.S9.1. As recommended by Wilks (2016), for spatially correlated fields, we adopt $\alpha_{FDR} = 2 \times \alpha_{global}$ (with $\alpha_{global} = 0.05$) rather than $\alpha_{FDR} = \alpha_{global}$ since the latter may be overly conservative when grid points are not spatially independent–as is inherently the case for gridded atmospheric data. Accordingly, $\alpha_{FDR} = 2 \times \alpha_{global}$ yields results closer to the uncorrected pointwise approach ($\alpha_0 = 0.05$), with an average decrease of 3.4 percentage points in the fraction of significant grid points relative to the uncorrected approach (Fig.S9.2).

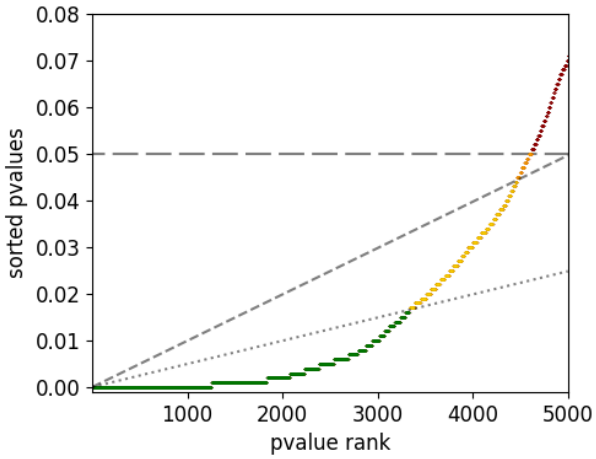

Fig. S9.1: Sorted p-values from RCC scores of summer precipitation predictions from sNAO+uAMV-based subsampling, plotted alongside three significance thresholds: the FDR criterion using $\alpha_{FDR} = \alpha_{global} = 0.05$ (dotted diagonal line), the FDR criterion with $\alpha_{FDR} = 2\alpha_{global} = 0.10$ (dashed diagonal line), and the standard pointwise threshold $\alpha_0 = 0.05$ (dashed horizontal line). For visual clarity, only the 5,000 smallest p-values out of 10,079 local tests are shown. Points falling below a given diagonal line are considered statistically significant under the corresponding FDR threshold.

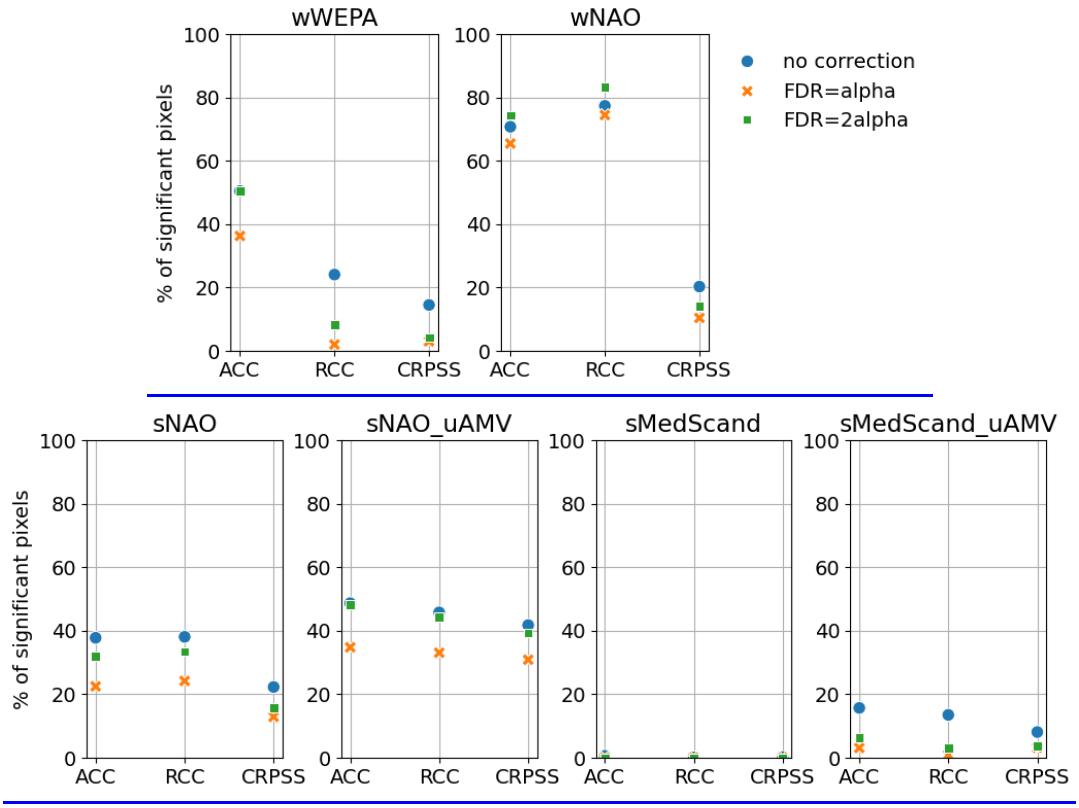

Fig. S9.2: Percentage of significant grid points for ACC, RCC, and CRPSS of winter (top row) and summer (bottom row) precipitation predictions, shown for each subsampling configuration (wWEPA, wNAO, sNAO, sNAO+uAMV, sMedScand and sMedScand+uAMV). Results are shown for the uncorrected pointwise approach ($\alpha_0 = 0.05$, blue dots), the FDR

procedure with $\alpha_{FDR} = \alpha_{global}$ (orange crosses), and the FDR procedure with $\alpha_{FDR} = 2\alpha_{global}$ (green squares)."

Significance masks and percentages of significant grid points have been updated throughout the revised manuscript. A brief description of the method has been added to Section 2.2.e (*"Skill evaluation"*, L270):

"Statistical significance of ACC, RCC and CRPSS is assessed using a block bootstrap with 1,000 samples and block size of eight years to account for temporal autocorrelation (Rousselet et al., 2021; Smith et al., 2020). The p-values are estimated as the proportion of bootstrap scores falling below zero (Goddard et al., 2013). To account for the multiplicity problem arising from simultaneous testing across all grid points (Wilks, 2011), significance is assessed using the False Discovery Rate (FDR) correction with $\alpha_{FDR} = 2 \times \alpha_{global}$ ($\alpha_{global} = 0.05$) following Wilks (2016) (see Supplementary S9)."

Applying the FDR correction results in the following figures, updated in the revised manuscript:

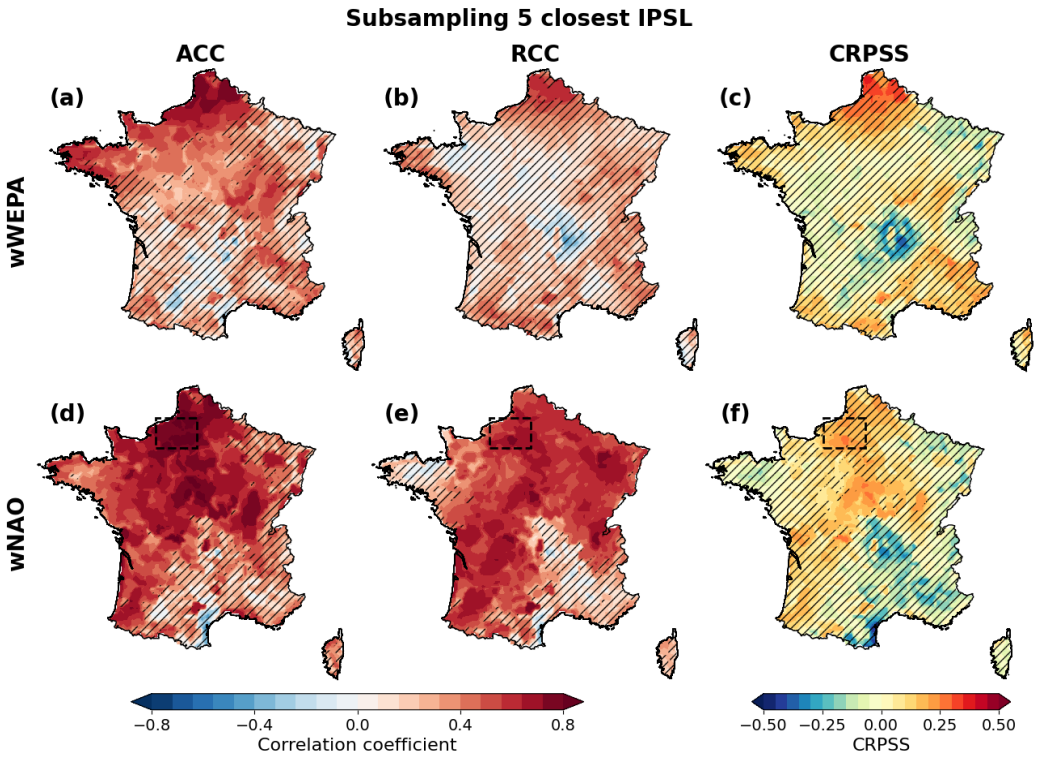

Fig. 8: Forecast skill for 8-year mean winter (ONDJFM) precipitation anomalies over France. Results are derived from subsampled hindcasts based on the wWEPA (a–c) and wNAO (d–f) indices. Panels show (a,d) Anomaly Correlation Coefficient (ACC), (b,e) Residual Correlation Coefficient (RCC)—quantifying skill beyond the forced response—and (c,f) Continuous Ranked Probability Skill Score (CRPSS), all computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate regions where skill

scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

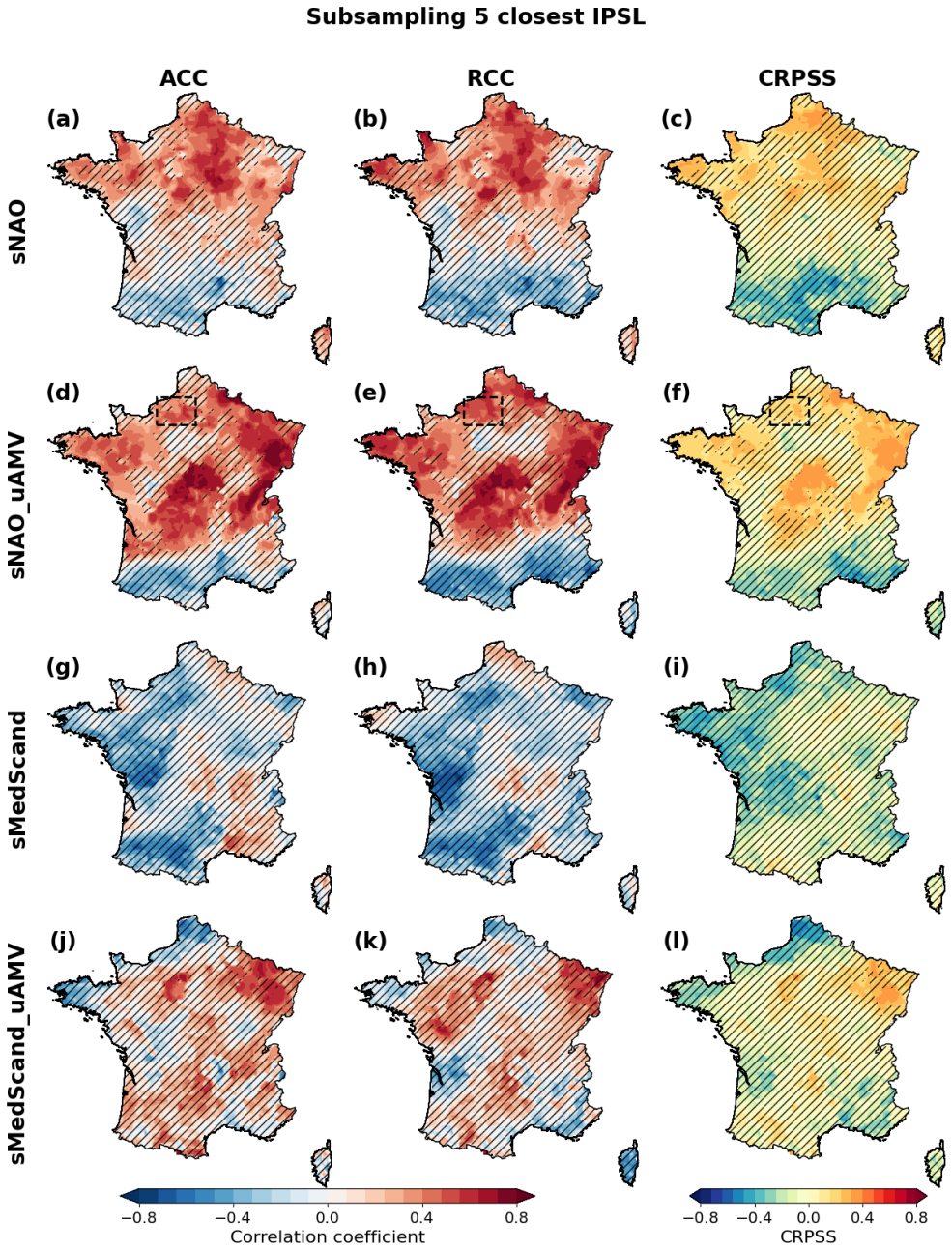

Fig. 9: Forecast skill for 8-year mean summer (AMJJAS) precipitation anomalies over France. Results are derived from subsampled hindcasts based on the sNAO (a–c), sNAO+uAMV (d-f), sMedScand (g-i) and sMedScand+uAMV (j-l) indices. Panels show (a,d) Anomaly Correlation Coefficient (ACC), (b,e) Residual Correlation Coefficient (RCC)—quantifying skill beyond the forced response—and (c,f) Continuous Ranked Probability Skill Score (CRPSS), all computed against SAFRAN precipitation observations over 1966–2019. Hatched regions indicate areas where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

We also take the opportunity to note two observations arising from the corrected summer results. First, the sNAO+uAMV-based subsampling continues to outperform sNAO-only subsampling in the corrected results, confirming the robustness of the joint-index approach and the conclusions of Section 3.3. Second, we note that while the correction improves the standalone sNAO skill — particularly over northern France — the sNAO+uAMV combination appears to degrade skill in this region relative to sNAO alone, while still improving skill elsewhere. The mechanism underlying this spatial heterogeneity remain unclear, and we have noted this explicitly in the revised manuscript as an open question warranting further investigation:

"It is worth noting that while the sNAO+uAMV subsampling improves summer precipitation skill across most of France relative to sNAO-only subsampling, a degradation is observed over northern France. The origins of these regional differences are not fully understood and may reflect competing influences of the sNAO and uAMV teleconnections in this region. This is left as an open question for future investigation."

6. Fig. 9e I do not fully understand how to interpret this result: since the 5 closest IPSL members and the full uninitialised ensemble members have virtually the same skill for uAMV (Fig. 7), and since the RCC measures the added value of initialisation, I would expect no skill improvements according to RCC for sNAO + uAMV compared to sNAO only (Fig. 9b). How can the skill enhancement be explained?

We thank the reviewer for this perceptive observation, which reveals an insufficiently clear explanation in the manuscript.

First, the subsampling procedure is based on minimization of a combined mean squared error (MSE) loss function rather than correlation maximization. Although the ACC of the uAMV for the sub-ensemble and the full uninitialized ensemble are similar (Fig. 7), MSE-based selection additionally constrains the temporal phasing of individual members relative to the predicted index. This distinction matters: two sub-ensembles can have similar ACCs yet differ substantially in how tightly individual members track the predicted trajectory at each start date. We have added a clarification of the selection metric to the caption of Fig. 7.

Second, and more fundamentally, the added value of jointly constraining the subsampling on both sNAO and uAMV does not arise from the uAMV's initialization skill *per se*, but from the increased coherence of the selected sub-ensemble. As illustrated in Fig. 3, members selected on the basis of the sNAO alone span a wide range of uAMV states, some of which are inconsistent with the observed or predicted uAMV phase. Adding the uAMV to the loss function filters out these members that are inconsistent with observed AMV, reducing noise in the associated precipitation response. The improvement in RCC scores for sNAO + uAMV relative to sNAO alone (Figs. 9b and 9e) therefore reflects this tighter dynamical coherence of the sub-ensemble, rather than any direct initialization skill contribution from the uAMV. This mechanism has been clarified in the revised manuscript (Section 3.2):

"It is worth noting that the similar ACC values of the uAMV index for the sub-ensemble and the full uninitialized ensemble (Fig. 7c) do not imply that including the uAMV in the joint subsampling criterion is redundant. Members selected on the basis of the sNAO alone span a wide range of uAMV states (Fig. 7). Incorporating the uAMV into the MSE-based loss function filters out members whose oceanic state is dynamically inconsistent with the predicted uAMV phase, thereby improving the coherence of the sub-ensemble and the resulting precipitation skill (Fig. 9e), independently of the uAMV's direct initialization contribution."

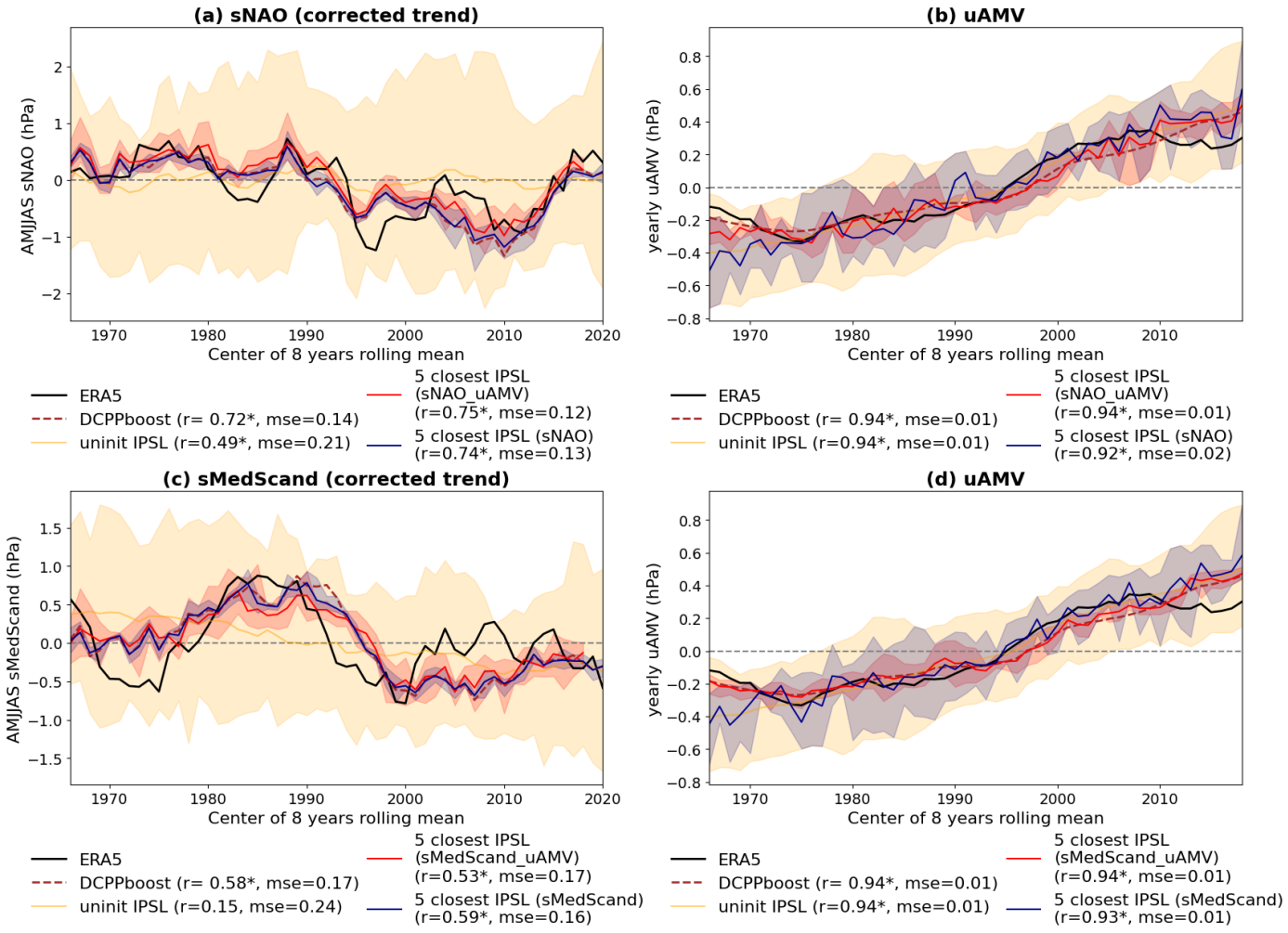

Fig. 7: Observed and predicted decadal variability of summer large-scale indices. Panels (a) and (c) show 8-year running means of the sNAO and sMedScand indices for the extended summer season (AMJJAS); panels (b) and (d) show the uAMV index (note that the uAMV is shown twice, as it serves as the joint constraint in both the sNAO+uAMV and sMedScand+uAMV subsampling configurations). The black line represents ERA5, and the burgundy dashed line shows boosted DCPP predictions. In each 8-year sliding window, five members are selected from the uninitialized IPSL-CM6A-LR ensemble by minimizing the combined MSE loss function (Eq. 2). The red shading and bold red line show the spread and ensemble mean of the five members selected using both indices jointly (sNAO+uAMV in panels a–b; sMedScand+uAMV in panels c–d). The blue shading and bold blue line show the spread and ensemble mean of the five members selected using the atmospheric index alone

(sNAO in panels a–b; sMedScand in panels c–d). The yellow shading and line show the full spread (minimum to maximum) and mean of the entire uninitialized IPSL ensemble. Pearson correlation coefficients (ACC) and mean squared errors (MSE) with ERA5 are reported in each panel legend; stars indicate statistically significant ACC values at the 95% confidence level.

7. L509 I find Fig. S3.1 important as it allows one to uncouple the extent to which final skill is limited by our ability to predict large-scale patterns of variability, or by the impact they have on surface climate. In my opinion this figure would deserve at least a mention in the Results section, but I leave this up to the authors.

We agree that this figure deserves greater prominence. In Section 3.3, the following sentence has been revised accordingly in the manuscript:

"The lower skill of sMedScand-based subsampled hindcasts is consistent with weaker observed correlations and lower predictability for this index (Fig. 7, r=0.57). In contrast, when the observed sMedScand index is used directly as a perfect prediction, subsampled hindcasts show significant ACC and RCC scores across southern France (Fig. S3.1), suggesting that improvements in sMedScand prediction skill would translate directly into improved precipitation forecasts over this region."

Technical corrections

1. The manuscript contains several grammatical and typographical errors. I recommend that a careful editing of the entire text be carried out. Issues include subject-verb agreement (e.g. L17), punctuation (e.g. L57), verb tenses (e.g. "is allowing" L246), typos (e.g. L450), articles (e.g. L130), and syntax (e.g. L117).

We have carefully revised the entire manuscript for grammatical and typographical errors, addressing the specific issues highlighted by the reviewer (subject-verb agreement, punctuation, verb tenses, typos, articles, and syntax). We are grateful for this observation.

2. L65 Perhaps add a very short explanation of the method of Alkama et al. here?

Indeed, this explanation will be added to the new manuscript:

"Building on these advances, a more recent study (Alkama et al., in press.) achieved a further step forward in NAO predictability, by optimizing the averaging step to eliminate temporal misalignment."

3. L82 Here, I would consider mentioning that the 8 km spatial resolution considered in the study is meant to match that of the reference dataset. Also, I think the comments at L238-242 would fit better in the Introduction than in the Methods section.

Indeed this was not clearly stated in the original manuscript. This updated version will be included in the revised manuscript:

"We therefore focus on providing decadal precipitation predictions at approximately 8-km spatial resolution, matching the resolution of the reference dataset (SAFRAN), targeting both winter and summer seasons across regional domains in France."

4. L110-113 I suggest to clarify that the IPSL-CM6A-LR ensemble consists of extended historical (uninitialised) simulations. Currently, this is only explained at L220.

The following revised text will replace the original sentence:

"Once the target indices are predicted, we apply a subsampling method (described below) to the outputs from the large ensemble of extended historical simulations from the IPSL-CM6A-LR model, consisting of 32 members covering the period 1850–2059 (Bonnet et al., 2021)."

5. L113 The citation of Boucher et al. (2020) has no correspondence in the References. I guess the correct citation is Bonnet et al. (2021)?

We thank the reviewer for highlighting this error. The citation will be corrected to Bonnet et al. (2021) in the revised manuscript.

6. Fig. 1I find this figure very informative and easy to follow. The dashed lines with annotations "if poor forecasting skill" and "if limited predictability", though, are never directly referenced in the main text. I can see that some of the choices presented earlier on by the authors (for example the fact that the wNAO index is computed over a different season than the precipitation anomalies, or that wNAO is used instead of EUNS) are explained later in section 3.4. Still, it may be worth mentioning in section 2.2 that try-and-error tests were carried out to reach a particular choice, and refer to 3.4 for full explanation.

Indeed this point was not explicitly commented on. To improve clarity, this comment will be included in the revised manuscript:

"This workflow is applied in this paper to build and evaluate improved precipitation predictions separately for extended-winter (October to March) and extended-summer (April to September). Trial-and-error tests were conducted to identify the most relevant SLP indices, as indicated by the dashed lines (Fig. 1). The five key steps are detailed below:"

7. L166-169 The index of Jianping and Wang (2003) is based on the normalised difference of SLP values, whereas the difference is not normalised here. Please clarify this in the text.

Here is the revised version:

"The North Atlantic Oscillation Index (wNAO), calculated for DJFM (December to March) as the difference in mean SLP between two zonal boxes spanning the North Atlantic-European sector—subtropical (35°-40°N, 80°W-30°E) and subpolar (63°-67°N, 80°W-30°E)—following Jianping and Wang (2003). This formulation reduces sensitivity to east-west shifts in the centers of action that can happen within climate models. It should be noted that unlike Jianping and Wang (2003), here the indices are not normalized."

8. L199-203 To understand the exact procedure followed by the authors, it would be important to make sure readers can access Alkama et al. (currently submitted).

We thank the reviewer for raising this point. Alkama et al. was recently accepted for publication in Science Advances and will be publicly available shortly, ensuring that readers will be able to access the full methodological details.

9. Figs. 2 and 3 Reconciling these figures with the accompanying text requires switching signs a few times (Fig. 2a shows a pattern of negative precipitation anomaly, Fig. 3a shows a negative wWEPA pattern, and Fig. 3b shows an anti-correlation between wWEPA and PC1). I would suggest adding a few words of explanation to help the reader link Fig. 2 with Fig. 3.

To improve clarity, the following explanation has been added.

"Correlation map of SLP with PC1 of precipitation displays a dipole pattern over the North Atlantic-European sector that closely mirrors the WEPA index defined by Castelle et al. (2017). In fact, the first EOF shows negative precipitation anomalies (Fig. 2), and is correlated with a positive wWEPA pattern (Fig. 3), it results in an anti-correlation between PC1 and wWEPA."

10. L321 Do the panels show "spatial correlation" or spatial maps of (time) correlations?

Indeed this formulation was unclear. Here is the modified version:

"Panels (a) and (b) display spatial maps of correlations between the first two precipitation PCs (PC1 and PC2) and SLP anomalies over the North Atlantic."

11. Fig. 7 Shouldn't the unit measure of uAMV be degree Celsius?

We thank the reviewer for highlighting this mistake. The unit is corrected in the revised manuscript.

12. Fig. 8 and 9 I recommend adding an explanation of what the dashed black box is in the caption.

We thank the reviewer for pointing this out. This omission has been corrected by adding the following sentence to the captions of Figures 8 and 9:

"Dashed black boxes indicate a specific region illustrated in Fig. 10."

References

Wilks, D., 2006. Statistical methods in the atmospheric sciences, second ed. Elsevier.

Bonnet, R., Boucher, O., Deshayes, J., Gastineau, G., Hourdin, F., Mignot, J., Servonnat, J., and Swingedouw, D.: Presentation and Evaluation of the IPSL-CM6A-LR Ensemble of Extended Historical Simulations, Journal of Advances in Modeling Earth Systems, 13, e2021MS002565, https://doi.org/10.1029/2021MS002565, 2021.

Christiansen, B., Yang, S., and Matte, D.: The Forced Response and Decadal Predictability of the North Atlantic Oscillation: Nonstationary and Fragile Skills, Journal of Climate, 35, 5869–5882, https://doi.org/10.1175/JCLI-D-21-0807.1, 2022.

Jianping, L. and Wang, J. X. L.: A new North Atlantic Oscillation index and its variability, Adv. Atmos. Sci., 20, 661–676, https://doi.org/10.1007/BF02915394, 2003.

Klavans, J. M., Cane, M. A., Clement, A. C., and Murphy, L. N.: NAO predictability from external forcing in the late 20th century, npj Clim Atmos Sci, 4, 1–8, https://doi.org/10.1038/s41612-021-00177-8, 2021.

Nicolì, D., Gualdi, S., and Athanasiadis, P. J.: Decadal predictions outperform climate projections in forecasting Mediterranean wintertime precipitation, Environ. Res. Lett., 20, 034034, https://doi.org/10.1088/1748-9326/adb59e, 2025.

Smith, D. M., Scaife, A. A., Eade, R., Athanasiadis, P., Bellucci, A., Bethke, I., Bilbao, R., Borchert, L. F., Caron, L.-P., Counillon, F., Danabasoglu, G., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Estella-Perez, V., Flavoni, S., Hermanson, L., Keenlyside, N., Kharin, V., Kimoto, M., Merryfield, W. J., Mignot, J., Mochizuki, T., Modali, K., Monerie, P.-A., Müller, W. A., Nicolí, D., Ortega, P., Pankatz, K., Pohlmann, H., Robson, J., Ruggieri, P., Sospedra-Alfonso, R., Swingedouw, D., Wang, Y., Wild, S., Yeager, S., Yang, X., and Zhang, L.: North Atlantic climate far more predictable than models imply, Nature, 583, 796–800, https://doi.org/10.1038/s41586-020-2525-0, 2020.

Swingedouw, D., Rodehacke, C. B., Behrens, E., Menary, M., Olsen, S. M., Gao, Y., Mikolajewicz, U., Mignot, J., and Biastoch, A.: Decadal fingerprints of freshwater discharge around Greenland in a multi-model ensemble, Clim Dyn, 41, 695–720, https://doi.org/10.1007/s00382-012-1479-9, 2013.

Ting, M., Kushnir, Y., Seager, R., and Li, C.: Forced and Internal Twentieth-Century SST Trends in the North Atlantic*, Journal of Climate, 22, 1469–1481, https://doi.org/10.1175/2008JCLI2561.1, 2009.

Wilks, D. S.: Statistical methods in the atmospheric sciences, 4th edition., Elsevier, 2011.

Wilks, D. S.: “The Stippling Shows Statistically Significant Grid Points”: How Research Results are Routinely Overstated and Overinterpreted, and What to Do about It, Bulletin of the American Meteorological Society, 97, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1, 2016.