Supplement of Earth Syst. Dynam., 17, 1177–1199, 2026
https://doi.org/10.5194/esd-17-1177-2026-supplement

Earth System Dynamics
Open Access

Supplement of

Toward robust fine-scale decadal precipitation forecasts through dynamically consistent subsampling

Joanne Couallier et al.

Correspondence to: Joanne Couallier (joanne.couallier@suez.com)

S1. Correcting trends in indices decadal predictions

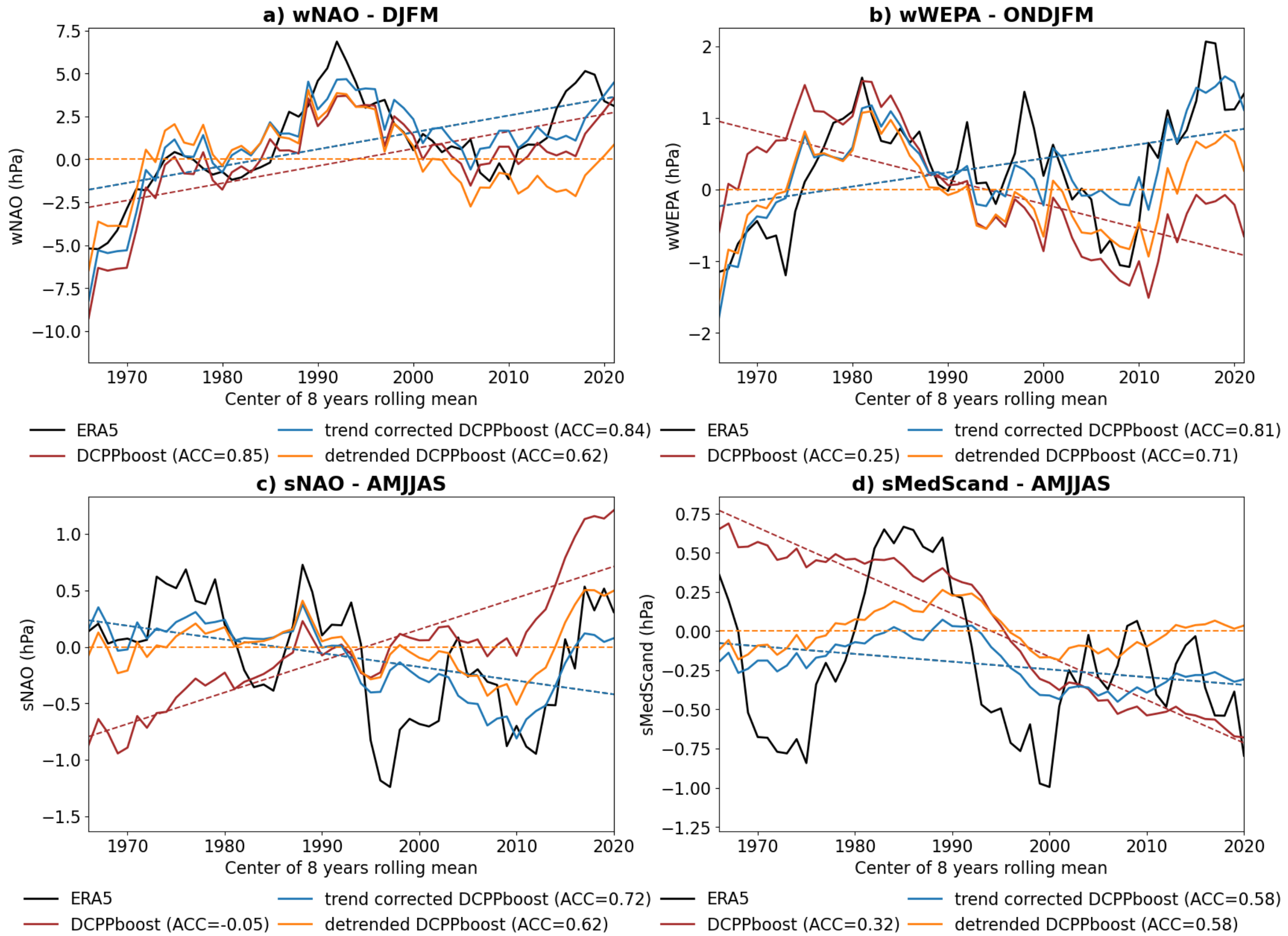

Fig. S1: Observed and predicted decadal variability of the wNAO (a), wWEPA (b), sNAO (c), sMedScand (d). The black line represents ERA5 reanalysis, while the burgundy dashed line shows predictions from the boosted DCPP method. The orange and blue lines represent corrected predictions after removing or correcting the trend, respectively. Correlation coefficients between predictions and observations are reported in parenthesis in each panel's legend.

S2. Uninitialized IPSL-CM6A-LR Skill scores maps

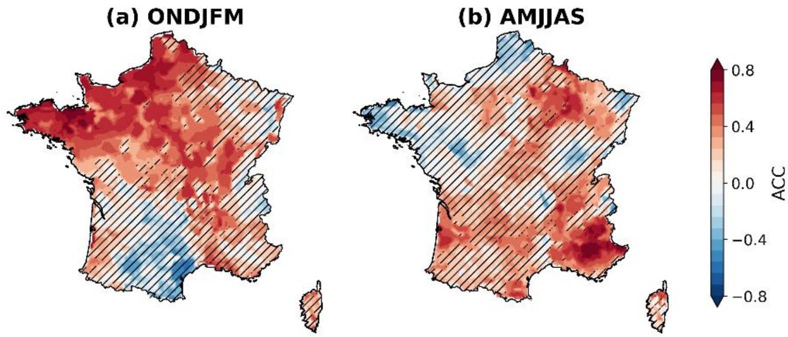

Fig. S2: Forecast skill for 8-year mean winter (a) and summer (b) precipitation anomalies over France. Skill maps are based on the uninitialized large ensemble mean IPSL-CM6A-LR. Panels (a-b) show the Anomaly Correlation Coefficient (ACC) computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

S3. Subsampling from observed sMedScand

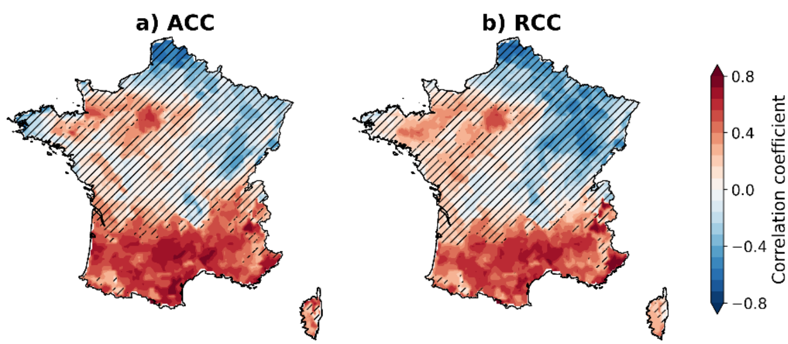

Fig. S3: Forecast skill for 8-year mean summer (AMJJAS) precipitation anomalies over France. Skill maps are based on subsampled forecasts using the observed sMedScand index. Panels show (a) Anomaly Correlation Coefficient (ACC) and (b) Residual Correlation Coefficient (RCC)—quantifying skill beyond the forced response—, both computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

S4. Multi-criteria subsampling for winter precipitation prediction

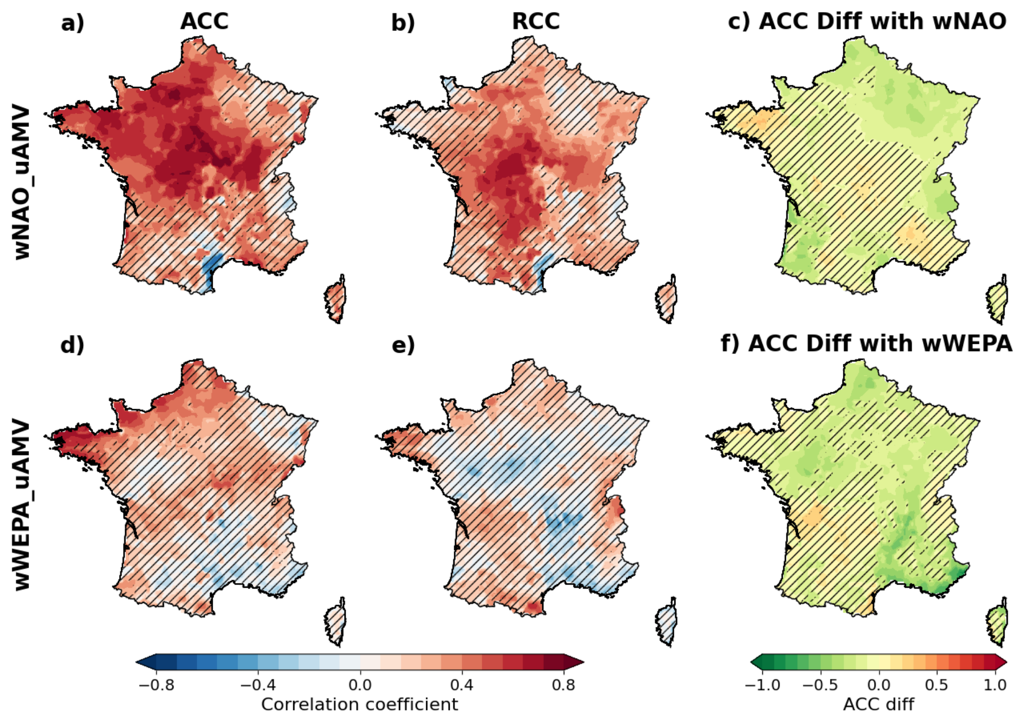

Fig. S4: Forecast skill for 8-year mean winter (ONDJFM) precipitation anomalies over France. Skill maps are based on subsampled forecasts using the wNAO+uAMV (a–c) and wWEPA+uAMV (d–f) indices. Panels show (a,d) Anomaly Correlation Coefficient (ACC), (b,e) Residual Correlation Coefficient (RCC)—quantifying skill beyond the forced response—and (c,f) the ACC difference with the reference prediction using wNAO (c) and wWEPA (f) alone. Scores are computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap (a-b, d-e) and with a Steiger's test for ACC differences (c,f).

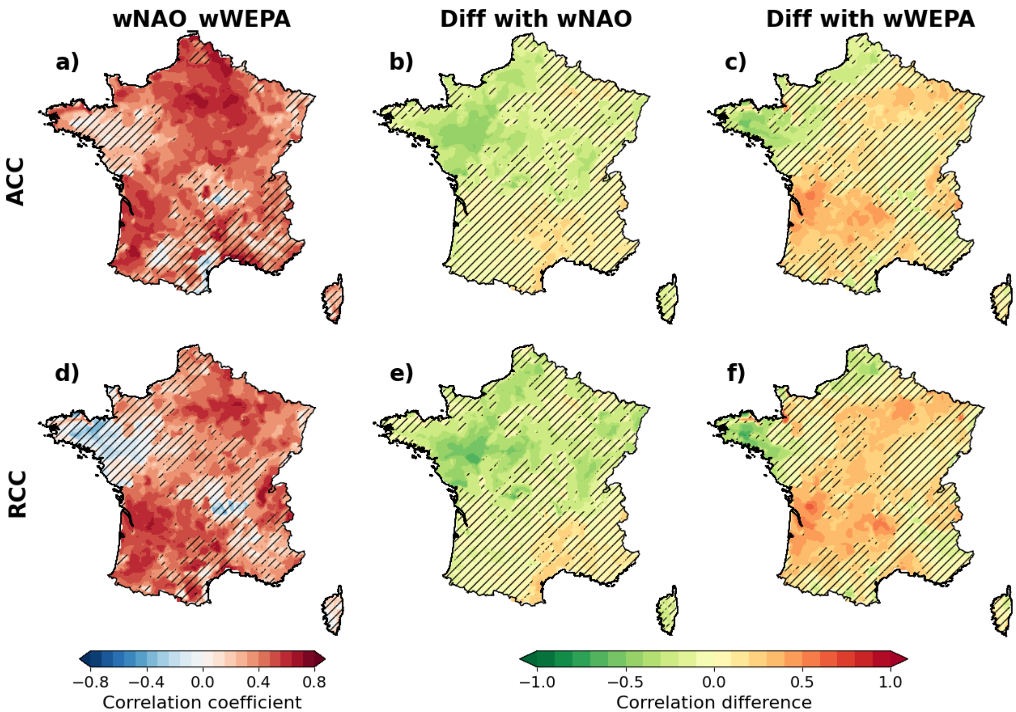

Fig. S5: Forecast skill for 8-year mean winter (ONDJFM) precipitation anomalies over France. Skill maps are based on subsampled forecasts using the wNAO+wWEPA (a,d). Panels show (a) Anomaly Correlation Coefficient (ACC), (b,c) the ACC difference with the reference prediction using wNAO (b) and wWEPA (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).

S5. Exploring other indices linked to winter precipitation second principal component

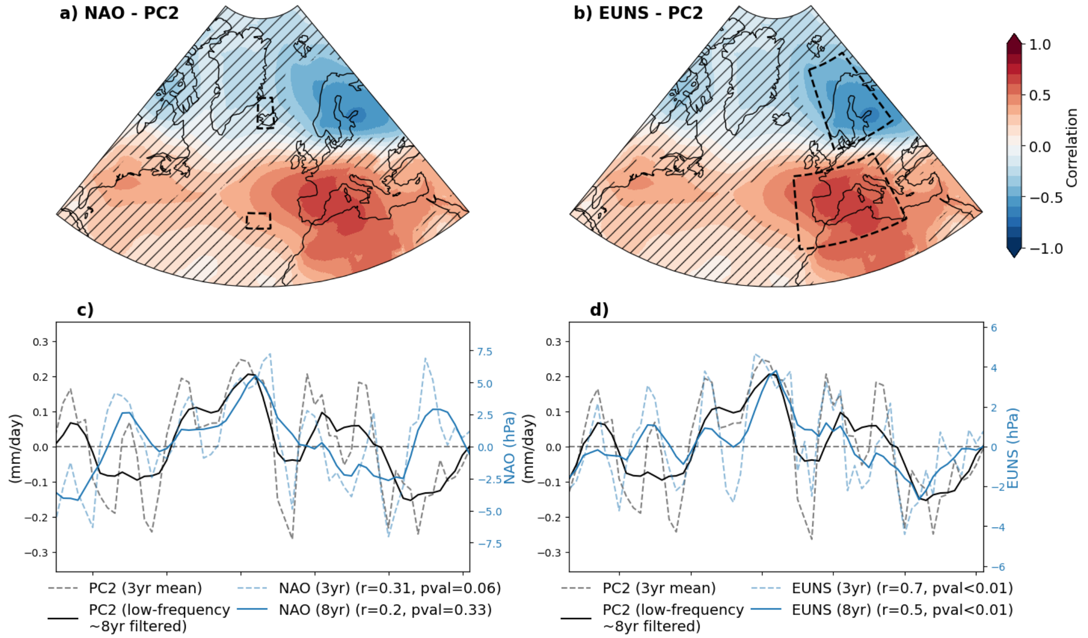

Fig. S6: Relationship between winter (ONDJFM) precipitation variability and large scale North Atlantic circulation. Panels (a) and (b) display spatial correlation coefficients between the second precipitation PC (PC2) and SLP anomalies over the North Atlantic. Hatched areas denote statistically non-significant correlations at the 95% confidence level based on a block-bootstrap test. Black dashed boxes outline the spatial domains used to compute the NAO (panel a) and EUNS indices (panel b). Panels (c) and (d) compare the time series of precipitation PC2 (black lines) and SLP indices (blue lines), including 3-year running mean (dashed) and low-frequency (~8-year) filtered (solid). Reported correlation coefficients (r) correspond to both smoothing scales. Units for the precipitation PCs are mm/day, and for the SLP indices are hPa.

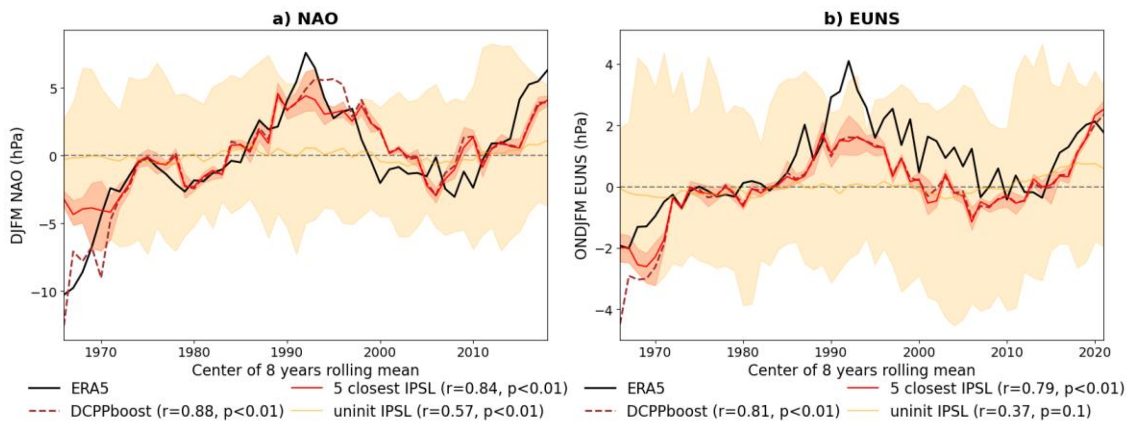

Fig. S7: Observed and predicted decadal variability of winter atmospheric indices. Panels (a) and (b) show 8-year running mean of the classic NAO and the Europe North-South (EUNS) indice for extended winter seaons (DJFM and ONDJFM, respectively). The black line represents ERA5 reanalysis, while the burgundy dashed line shows predictions from the boosted DCPP method. The red shading indicates the spread of the five subsampled members, best matching DCPP indices and selected from the uninitialized IPSL-CM6A-LR ensemble in each 8-year time window, with the bold red line marking their ensemble mean. The yellow line and shading

represent the mean and full spread (minimum to maximum) of the entire uninitialized IPSL ensemble, respectively. Correlation coefficients and p-values between predictions and observations are reported in parenthesis in each panel's legend.

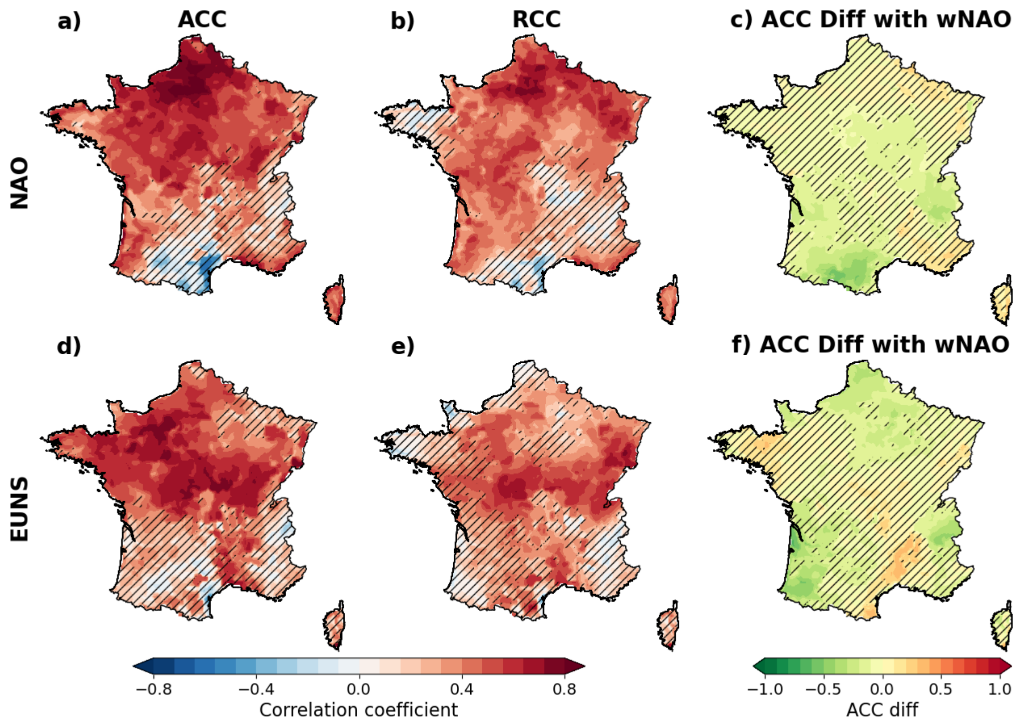

Fig. S8: Forecast skill for 8-year mean winter (ONDJFM) precipitation anomalies over France. Skill maps are based on subsampled hindcasts using the NAO (a–c) and Europe North-South (EUNS) (d–f) indices. Panels show (a,d) Anomaly Correlation Coefficient (ACC), (b,e) Residual Correlation Coefficient (RCC)—quantifying skill beyond the forced response—and (c,f) ACC difference with the reference prediction using wNAO. Scores are computed against SAFRAN precipitation observations over 1966–2019. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap (a-b, d-e) and with a Steiger's test for ACC differences (c,f).

S6. Results without scaling to the observed variance

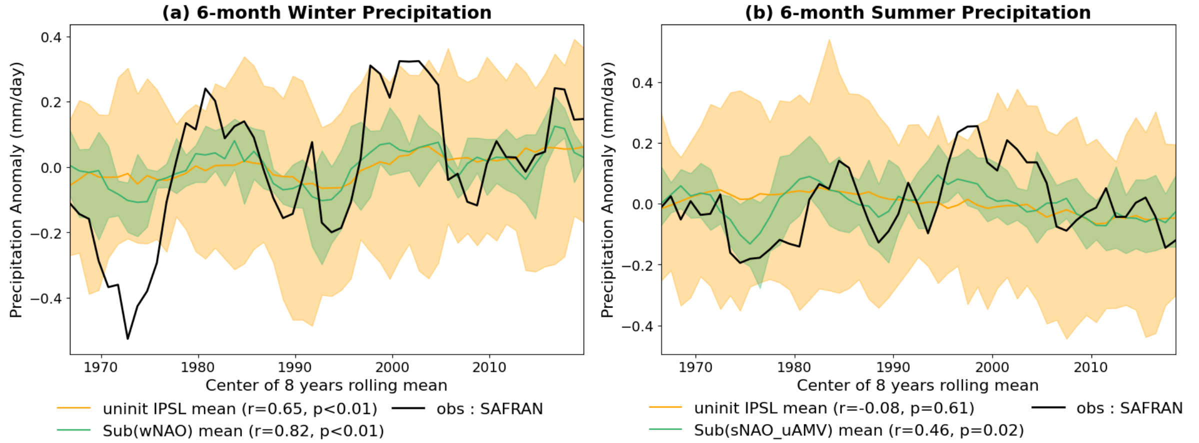

Fig. S9: Panel (a) and (b) show 8-year running mean precipitation anomalies for winter (ONDJFM) and summer (AMJJAS), respectively, averaged over the region outlined by dashed boxes in Figs. 8 and 9. The black line shows SAFRAN observations. The green line represents the mean of the five-member sub-ensemble selected based on the wNAO index (panel a) and combined sNAO+AMV indices (panel b), without scaling predictions to the observed variance. The yellow line shows the mean of the full uninitialized IPSL ensemble. Shaded areas indicate the ensemble spread, defined as the 5th to 95th percentile range.

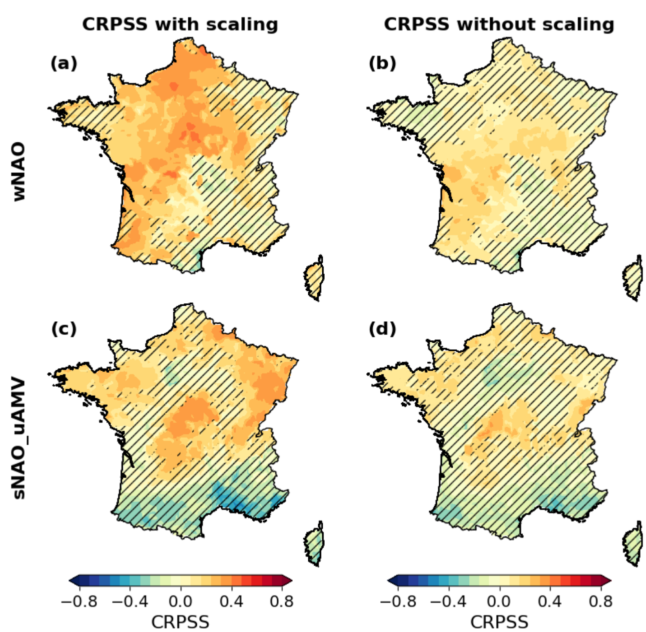

Fig. S10: Forecast skill for 8-year mean winter (ONDJFM) and summer (AMJJAS) precipitation anomalies over France. Results are derived from subsampled hindcasts based on the wNAO (a–b) and sNAO+AMV (c-d) indices. Panels show Continuous Ranked Probability Skill Score (CRPSS) computed over 1966-2019, with (a,c) or without (b,d) scaling the predictions to the observed variance. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

S7. Predicting 5-year mean indices and precipitation

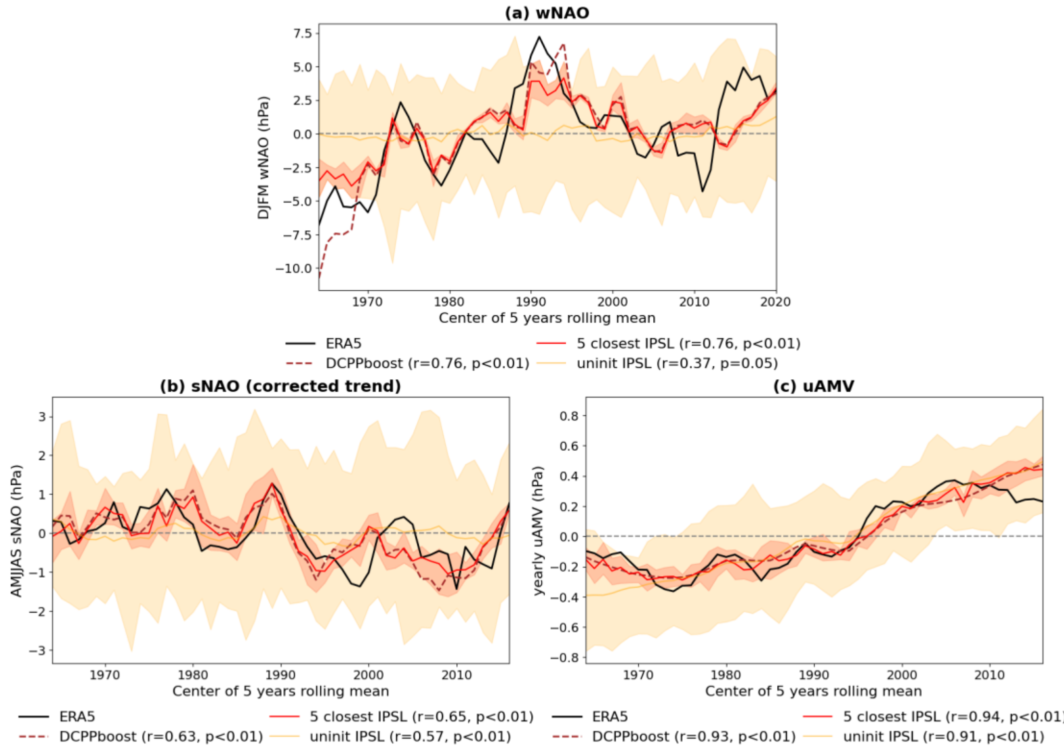

Fig. S11: Observed and predicted decadal variability of atmospheric and oceanic indices. Panels show 5-year running mean of the wNAO (a), sNAO (b) and uAMV (c). The black line represents ERA5 reanalysis observations, while the burgundy dashed line shows predictions from the boosted DCPP method. The red shading indicates the spread of the five index-matched members selected from the uninitialized IPSL-CM6A-LR ensemble in each 5-year time window, with the bold red line marking their ensemble mean. The yellow line and shading represent the mean and full spread (minimum to maximum) of the entire uninitialized IPSL ensemble. Correlation coefficients and p-values between predictions and observations are reported in parenthesis in each panel's legend.

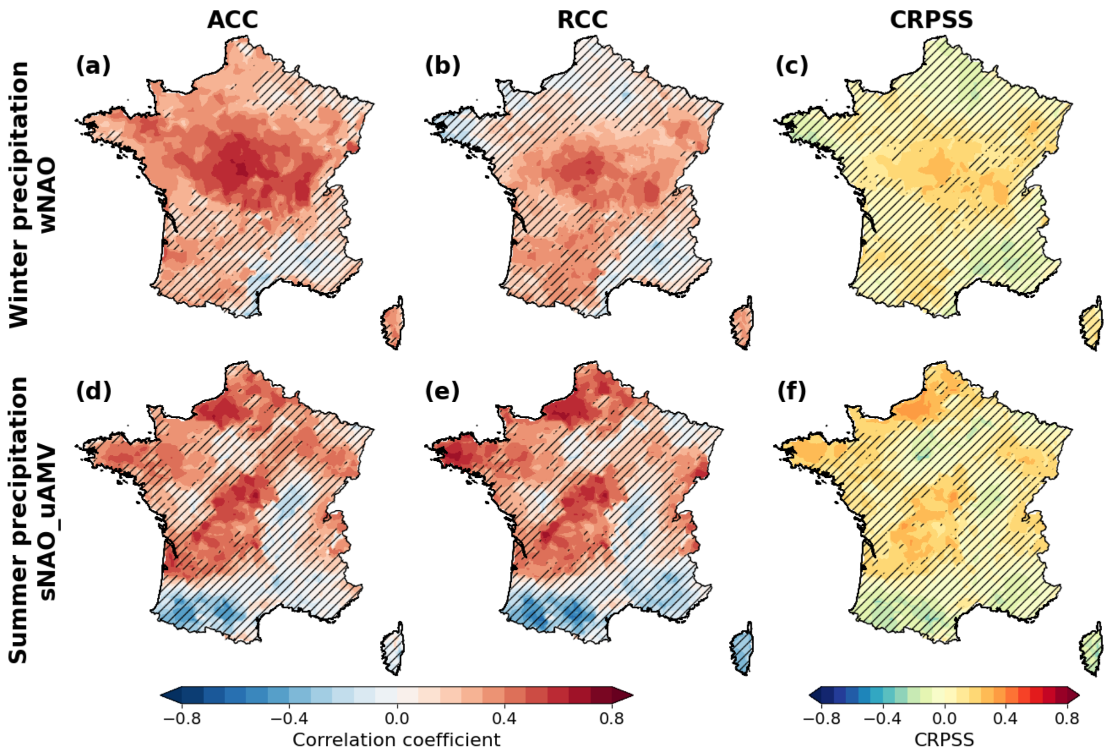

Fig. S12: Forecast skill for 5-year mean winter (ONDJFM) and summer (AMJJAS) precipitation anomalies over France. Results are derived from subsampled hindcasts based on the wNAO (a–c) and sNAO+uAMV (d–f) indices. Panels show (a,d) Anomaly Correlation Coefficient (ACC), (b,e) Residual Correlation Coefficient (RCC)—quantifying skill beyond the forced response—and (c,f) Continuous Ranked Probability Skill Score (CRPSS), all computed against SAFRAN precipitation observations over 1964–2020. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

S8. Sensitivity to sub-ensemble size in the case of sNAO+uAMV subsampling

The forecast skill from the subsampling procedure may depend on the number of members in the selected sub-ensemble. Figures S8.1 and S8.2 compare ACC and CRPSS skill scores for sNAO+uAMV-based subsampling of 8-year mean summer precipitation anomalies, for sub-ensemble sizes ranging from two to eight members.

It should be noted that CRPSS estimates are less reliable for small sub-ensemble sizes, as the probabilistic skill score is sensitive to the number of ensemble members used to construct the forecast distribution: with very few members, the spread is systematically underestimated, which can artificially inflate or deflate the CRPSS independently of the true forecast quality. Results for CRPSS with sub-ensemble sizes of two or three members should therefore be interpreted with caution. ACC, which is based solely on the ensemble mean, is not affected by this limitation and provides a more robust measure of skill across all ensemble sizes considered here.

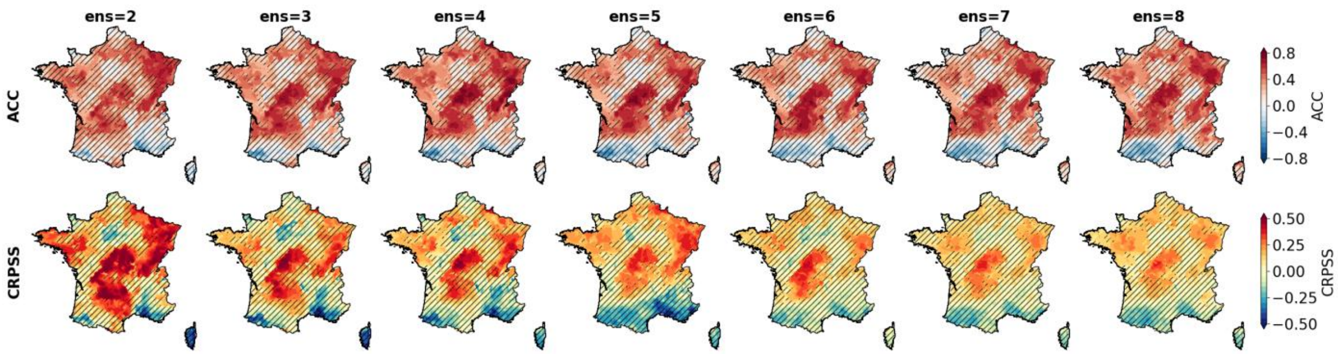

Fig. S13: Forecast skill for 8-year mean summer (AMJJAS) precipitation anomalies over France. Results are derived from subsampled hindcasts based on the sNAO+uAMV indices, with sub-ensemble size from 2 to 8. Upper panels show Anomaly Correlation Coefficient (ACC) and lower panels show Continuous Ranked Probability Skill Score (CRPSS), over 1966–2019. Hatched areas indicate region where skill scores are not statistically significant at the 95% confidence level, as assessed using a 1,000 sample block bootstrap.

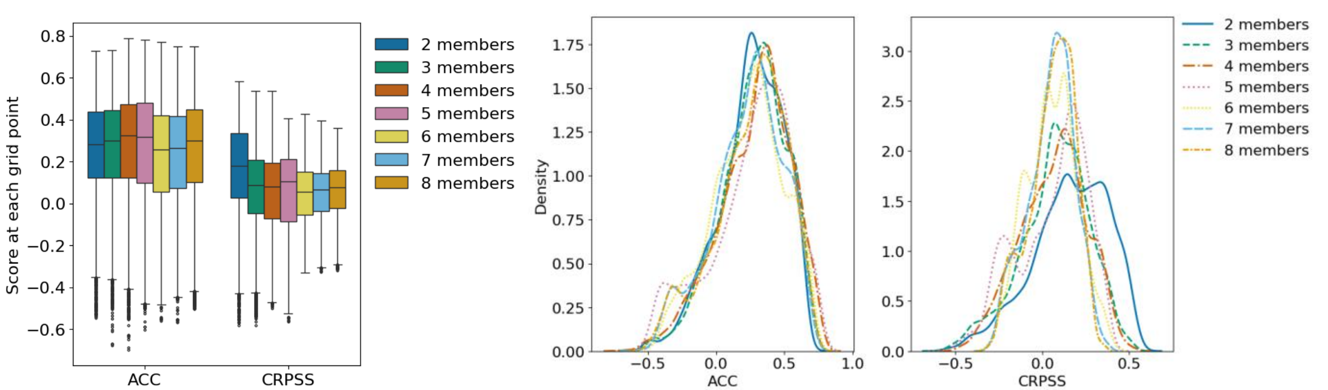

Fig. S14: Distribution of forecast skill for 8-year mean summer precipitation anomalies over France derived from subsampled forecasts using sNAO+uAMV indices with sub ensemble size from 2 to 8. Left panel shows the Anomaly Correlation Coefficient (ACC) and Continuous Ranked Probability Skill Score (CRPSS) for each grid point as boxplots. Right panel indicate ACC and CRPSS as continuous probability density curves. Skill scores are computed against SAFRAN precipitation observations over 1966–2019.

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).

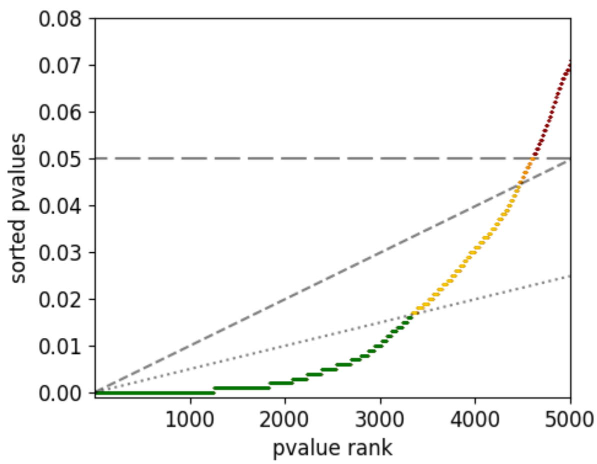

Fig. S15: Sorted p-values from RCC scores of summer precipitation predictions from sNAO+uAMV-based subsampling, plotted alongside thress significance thresholds: the FDR criterion using $\alpha_{FDR} = \alpha_{global} = 0.05$ (dotted diagonal line), the FDR criterion with $\alpha_{FDR} = 2 \times \alpha_{global} = 0.10$ (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.

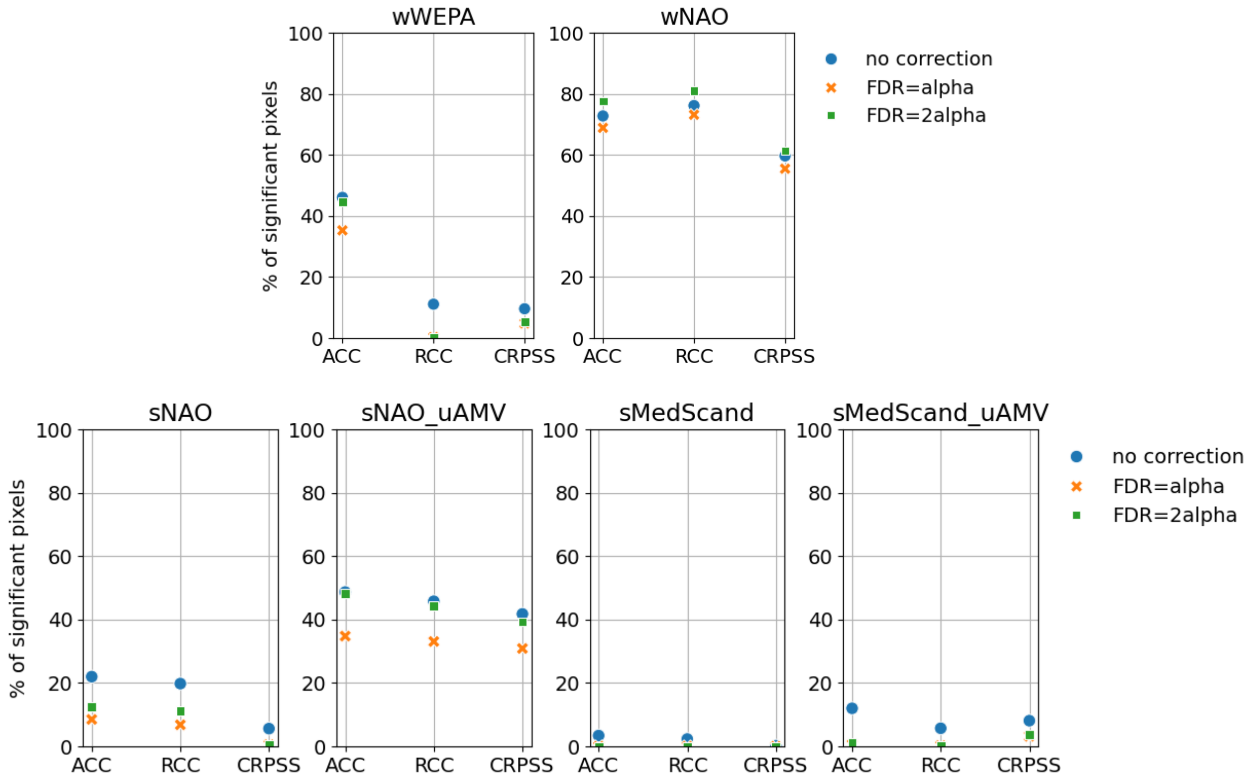

Fig. S16: Percentage of significant grid points for ACC, RCC, and CRPSS of winter (top row) and summer (bottom row) precipitation predictions, shown for each subsampling configuration (wWEPA, wNAO, sNAO, sNAO+uAMV, sMedScand and sMedScand+uAMV). Results are shown for the uncorrected pointwise approach ($\alpha_0 = 0.05$, blue dots), the FDR procedure with $\alpha_{FDR} = \alpha_{global}$ (orange crosses), and the FDR procedure with $\alpha_{FDR} = 2 \times \alpha_{global}$ (green squares).

S10. Representation of the WEPA spatial pattern in IPSL-CM6A-LR

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 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.

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.

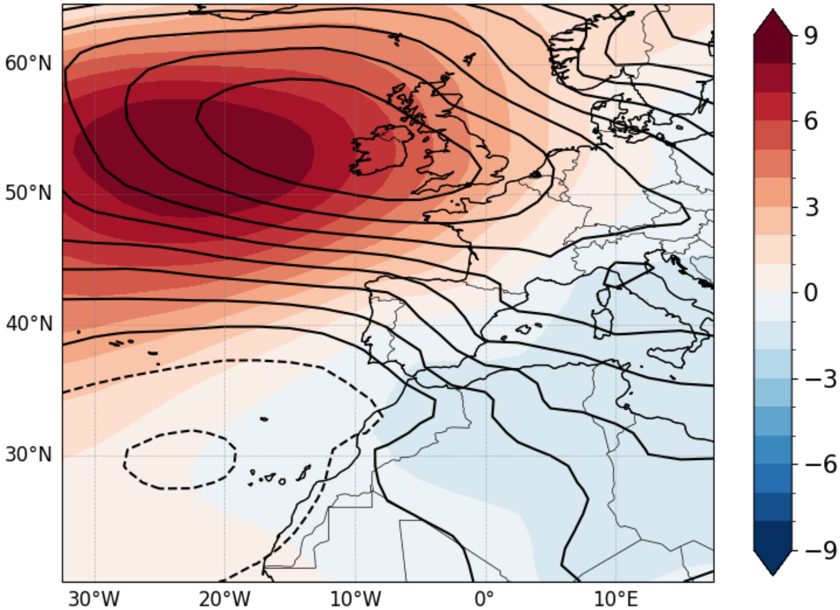

Fig. S17: Second empirical orthogonal function (EOF2) of extended winter (ONDJFM) sea level pressure variability over the North Atlantic-European sector (20°-65°N, 35°W-20°E). Colour shading shows the EOF2 spatial pattern derived from the IPSL-CM6A-LR pre-industrial control simulation (last 100 years), 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.