Authors response to reviews of manuscript “Reduction of uncertainty in near-term climate forecast by combining observations and decadal predictions”

Response to reviewer #2

We would like to thank Reviewer #1 for his interest in our study and the insightful comments that helped us to improve the manuscript.

Before addressing the specific comments, we would like to note that an error was identified in the implementation of the second step of our ensemble selection procedure in the original manuscript. The error affected the indices selected during the second step of the method and consequently led to differences in some of the reported results. This issue has now been corrected, and all analyses, figures, and associated text have been updated accordingly in the revised manuscript. In response to this correction, additional tests were performed to reassess and adjust the number of members selected in the two steps of the method. The revised selection now includes 50 members in the first step and 30 members in the second step. The methodological framework and overall conclusions of the study remain unchanged, although some quantitative results differ from those reported in the original version. We highlight these modifications where relevant in our responses below and in the revised manuscript. We also added the $\mathrm{BLEND}_{\mathrm{TAS}}$ forecast in addition to $\mathrm{HIST}_{\mathrm{TAS}}$ to all the evaluations, which was not available previously. The evaluation has also been extended to 1966-2014 instead of 1967-2014 in the revised version. Additional analysis to test the significance of the different scores used in the evaluation using a resampling approach were also added in the revised manuscript. Finally, we also added a case study to the spatial assessment carried out in section 3.2, to also evaluate 10-year temperature forecasts in summer, especially over WCE, where the method developed in this study provides large forecast improvements in comparison to HIST. This allows for a spatial assessment for each of the seasons evaluated in this study (winter and summer).

Please find below the answers to the comments point-by-point. For clarity, all reviewer comments are in **black** and responses in **blue**.

This study introduces a method that constrains climate projections based on a combination of previous observations and initialised decadal predictions, to reduce the uncertainty in near-term climate projections (for the next 10-20 years defines near-term in the introduction, although later the results only go up to 15 years). The authors demonstrate that this method reduces the spread of the projections, which they claim adds value. The analysis of a selection of skill scores for three European regions reveals a somewhat mixed picture.

The study is an interesting addition to the growing literature that combines observations, initialised predictions and climate projections to provide seamless climate prediction information. I think, however, that additional work is needed to clarify several points in the manuscript, and in particular the interpretation of the results.

Major comments

- Given the primary effect of the blending is a reduction in spread, while the error of the predictions is not significantly affected, it is not obvious to me that reducing the spread/uncertainty of the predictions necessarily ‘adds value’. This is only true if it enhances the reliability of the predictions, but there is a risk that reducing the spread makes the predictions overconfident. This should be tested by including a metric in the analysis that evaluates the reliability of the blended hindcasts.

Thank you for this important comment. We agree that a reduction in spread only adds value if it does not degrade forecast skill and does not lead to overconfident predictions. We therefore evaluate this using the absolute error of the ensemble mean, as well as probabilistic skill scores such as the RPSS and CRPSS, which explicitly account for the forecast spread, so any artificial gain resulting from overconfident spread reduction would be penalized in these metrics.

- Statistical testing of differences or skill/improvements. In Figure 3, the box plots of the absolute error largely overlap, with some minor variations in the different quantiles of the distribution. The authors interpreted a reduced error from the constraint in some cases, however seeing the large spread of the distributions I doubt that this is a robust result. I therefore recommend applying a suitable testing for statistical significance, if using these figures to interpret improvements. Similar for the skill scores presented in Figure 4 – also here the authors should test which of the skill improvements are in fact significant.

We agree that differences in absolute error are difficult to interpret when the distributions largely overlap with those of HIST and DEC.

The other skill metrics used in the manuscript (RPSS, CRPSS, MSSS), are defined relative to a reference forecast and therefore already provide a quantitative measure of the added value and forecast capability compared to that baseline. Therefore, positive skill values directly indicate improved performance over the reference, whereas negative skill values indicate a deterioration in the forecast performance over the reference. This is why we did not initially complement these diagnostics with formal statistical significance tests.

However, we agree that formal significance testing is important to ensure that these improvements are not attributable to sampling variability or noise, as well as to better interpret the results from the box-plots of the absolute error.

To improve the clarity and robustness of the evaluation, we have added a statistical significance assessment using a resampling approach. From the 193 HIST members, we randomly select subsets of 30 members (i.e., the same size as the $\mathrm{HIST_{OBS}}$ and $\mathrm{BLEND_{OBS}}$ subsets). This procedure is repeated 1000 times, yielding 1000 subsets of 30 members. The scores are then computed for each subset to obtain a distribution representing the range of scores expected from a 30-member ensemble drawn from HIST due to sampling variability. A subset derived from BLEND is considered significant when its score exceeds the 95th percentile of the resampled score distribution, corresponding to a significance level of 5%. For the spatial evaluation (Section 3.2), only 500 drawings are held due to computational constraints.

For the ensemble spread and the absolute errors, a subset derived from BLEND is considered significant when its median absolute error or ensemble spread, calculated over the evaluation period, is lower than the 5th percentile of the resampled score distribution, corresponding to a significance level of 5%.

These tests assess whether forecasts derived from the BLEND method significantly outperform forecasts that would be obtained from randomly drawn 30-member subsets of HIST.

The method used to estimate the significance is now described in the section 2.4 “Evaluation Metrics” of the Method section of the revised manuscript. The figures and relative discussions in Section 3.1 and 2.3 have been updated in the revised manuscript accordingly.

Please find below the revised description of the method used for the significance assessment:

“The added value of the forecasts derived from the BLEND method is assessed using a resampling approach. From the 193 HIST members, we randomly select subsets of 30 members (i.e., the same size as the $\mathrm{HIST}_{\mathrm{OBS}}$ and $\mathrm{BLEND}_{\mathrm{OBS}}$ subsets). This procedure is repeated 1000 times, yielding 1000 subsets of 30 members. The scores are then computed for each subset to obtain a distribution representing the range of scores expected from a 30-member ensemble drawn from HIST due to sampling variability. A subset derived from BLEND is considered significant when its score exceeds the 95th percentile of the resampled score distribution, corresponding to a significance level of 5%. For the spatial evaluation (Section 3.2), only 500 drawings are held due to computational constraints.

For the ensemble spread and the absolute errors, a subset derived from BLEND is considered significant when its median absolute error or ensemble spread, calculated over the evaluation period, is lower than the 5th percentile of the resampled score distribution, corresponding to a significance level of 5%.

These tests assess whether forecasts derived from the BLEND method significantly outperform forecasts that would be obtained from randomly drawn 30-member subsets of HIST.”

Please find below the new Figures with updated results and significativity:

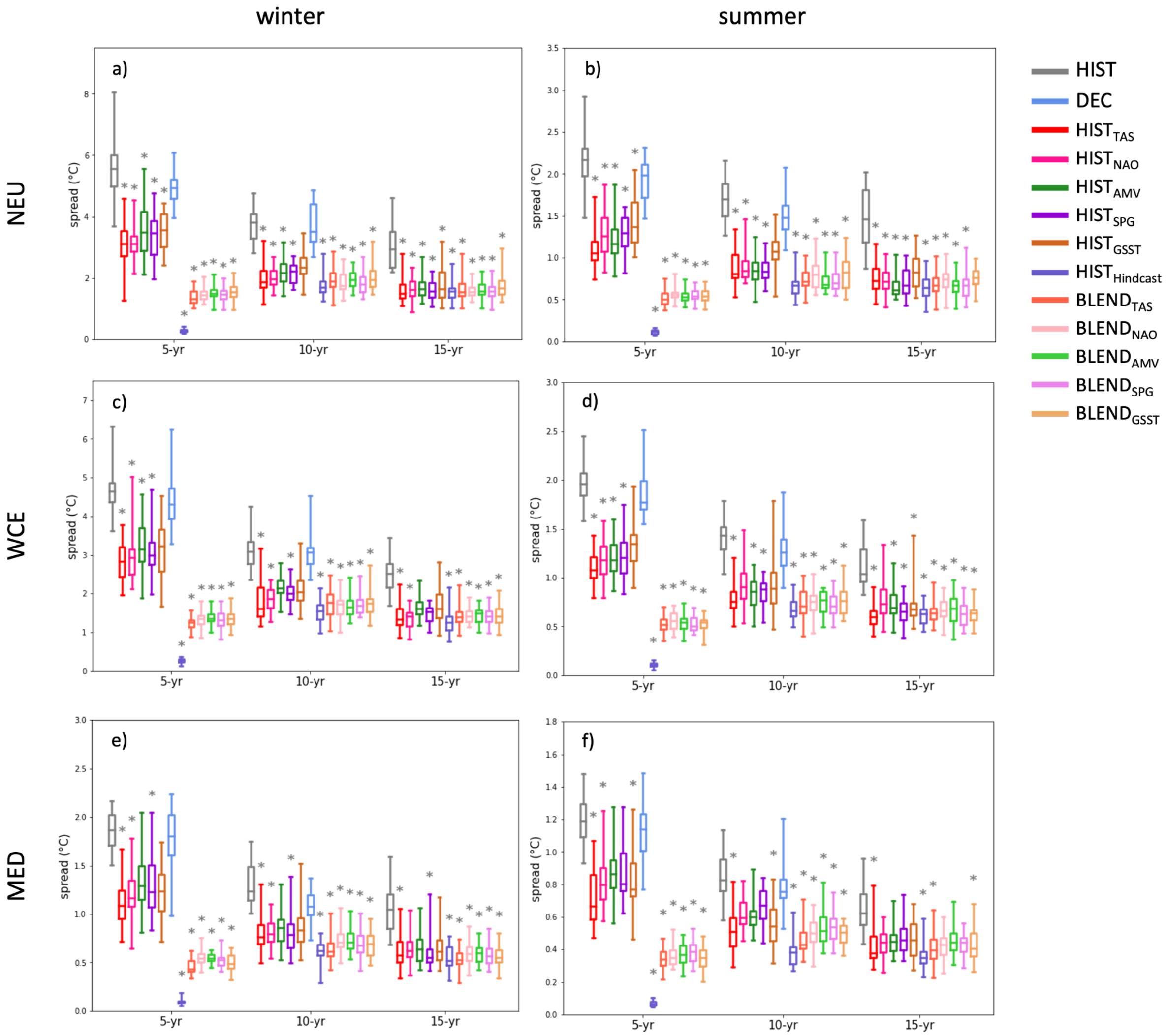


Figure 2: Boxplots of the average surface temperature spread for 5-, 10-, and 15-year forecasts in winter (a, c, e) and summer (b, d, f) over the NEU (a, b), WCE (c, d), and MED (e, f) regions. The spread is defined as the difference between the minimum and maximum and is calculated for each year of the retrospective evaluation period (1966–2000). The boxplots display the minimum, 25th percentile, median, 75th percentile, and maximum. Stars indicate cases where the median of the subsets derived from the method is significantly lower than those obtained from randomly drawn subsets of HIST (p-value < 0.05; see Section 2.4).

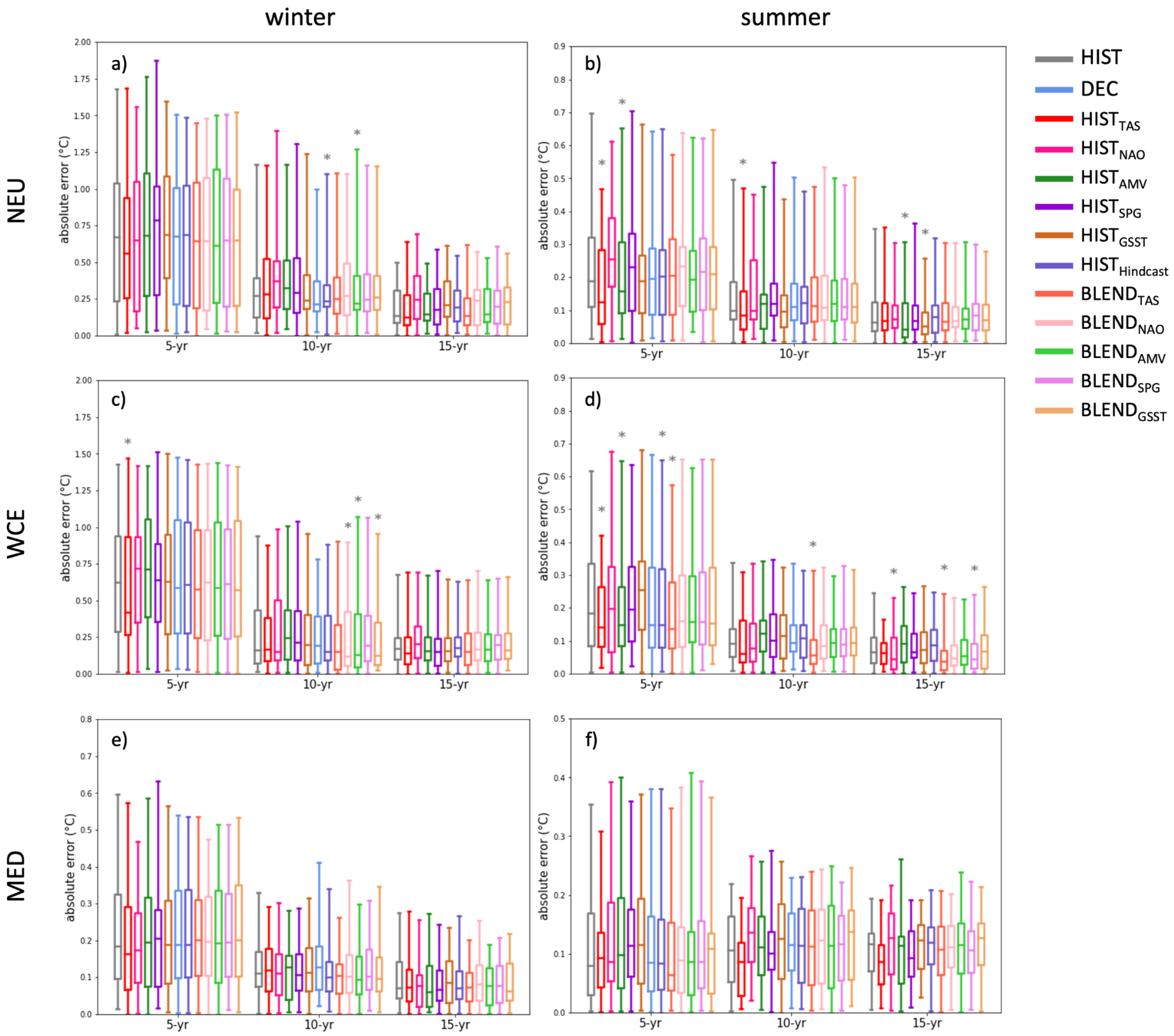


Figure 3: Boxplots of the absolute error of the average surface temperature for 5-, 10-, and 15-year forecasts in winter (a, c, e) and summer (b, d, f) over the NEU (a, b), WCE (c, d), and MED (e, f) regions. The absolute error is calculated each year of the testing period (1966-2000) between the observed surface temperature from ERA5 (Hersbach et al. 2020) and the ensemble mean of the different dataset described section 2.1 and 2.3. The boxplots display the minimum, 25th percentile, median, 75th percentile, and maximum. Stars indicate cases where the median of the subsets derived from the method is significantly lower than those obtained from randomly drawn subsets of HIST (p-value < 0.05; see Section 2.4).

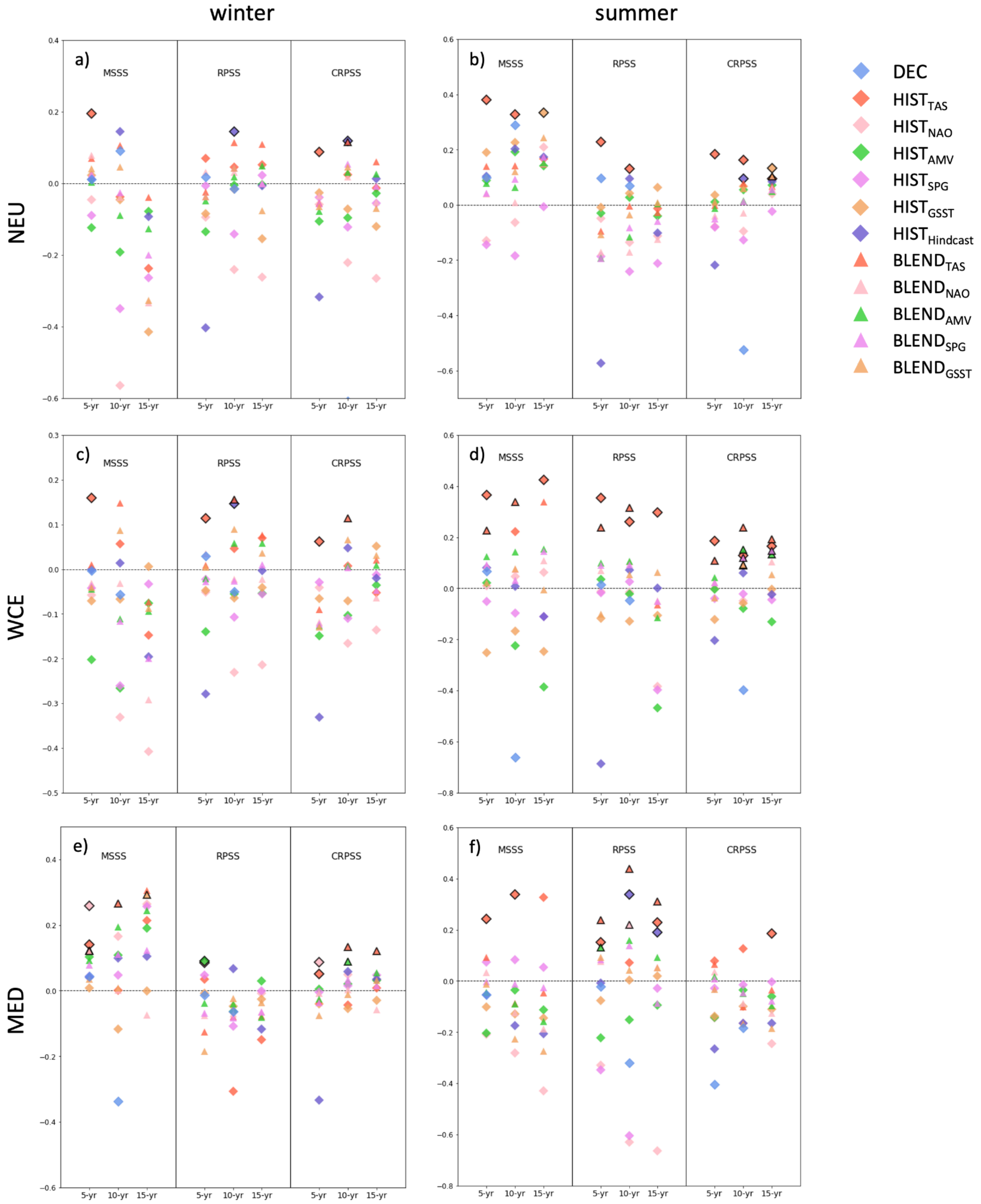


Figure 4: MSSS, RPSS, and CRPSS (see Section 2.4) calculated from the 5-, 10-, and 15-year time series of surface temperature forecasts for winter (a, c, e) and summer (b, d, f) over NEU (a, b), WCE (c, d), and MED (e, f). Scores are shown for the hindcast dataset (see Section 2.1) as well as for the dataset derived from BLEND (see Section 2.3). HIST is used as the reference; therefore, positive values indicate an improvement relative to HIST. Black edges indicate scores significantly better than those obtained from randomly drawn subsets of HIST (see Section 2.4). The scores are calculated over the 1966–2000 period, using the ensemble mean for MSSS and the full ensembles for RPSS and CRPSS.

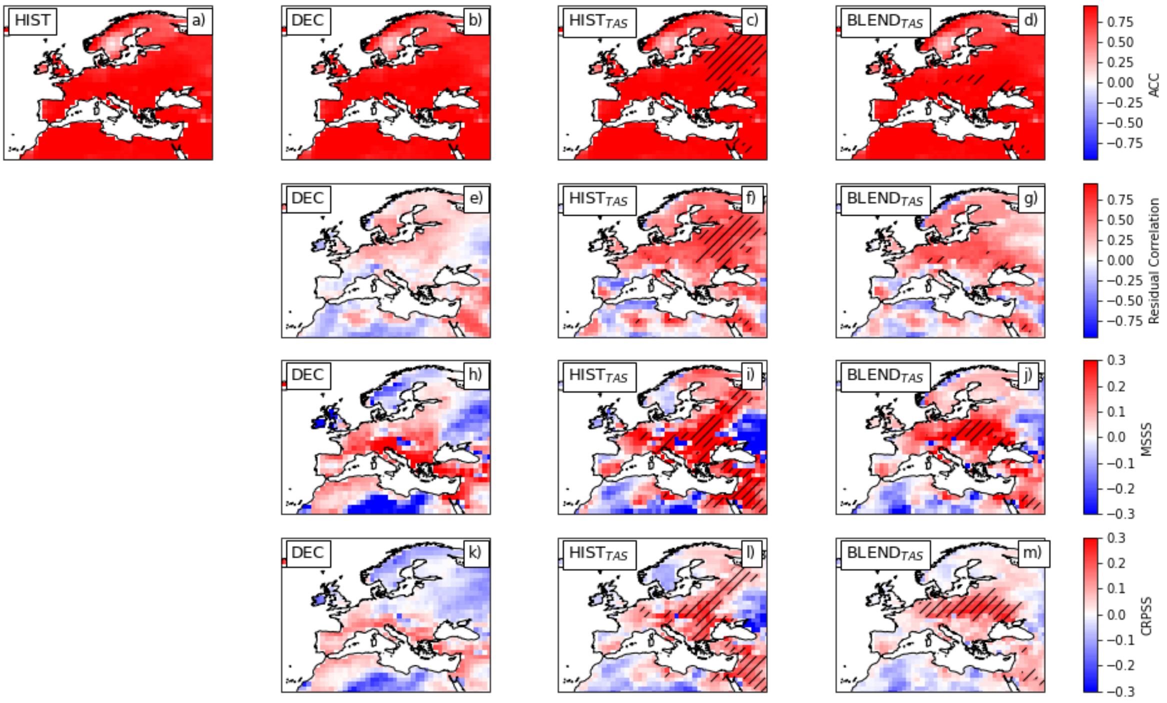


Figure 5: ACC (a–d), Residual correlation(e-g), MSSS (h–j), and CRPSS (k–m) calculated from 10-year forecasts of summer surface temperature over the evaluation period 1966–2000 for HIST (a), DEC (b, e, h, k), HIST$_{AMV}$ (c, f, i, l), and BLEND$_{AMV}$ (d, g, j, m).For BLEND$_{TAS}$, the summer surface temperature averaged over WCE is used for the second-step selection. Hatched regions indicate significant improvement in comparison to randomly drawn subsets of HIST (see Section 2.4).

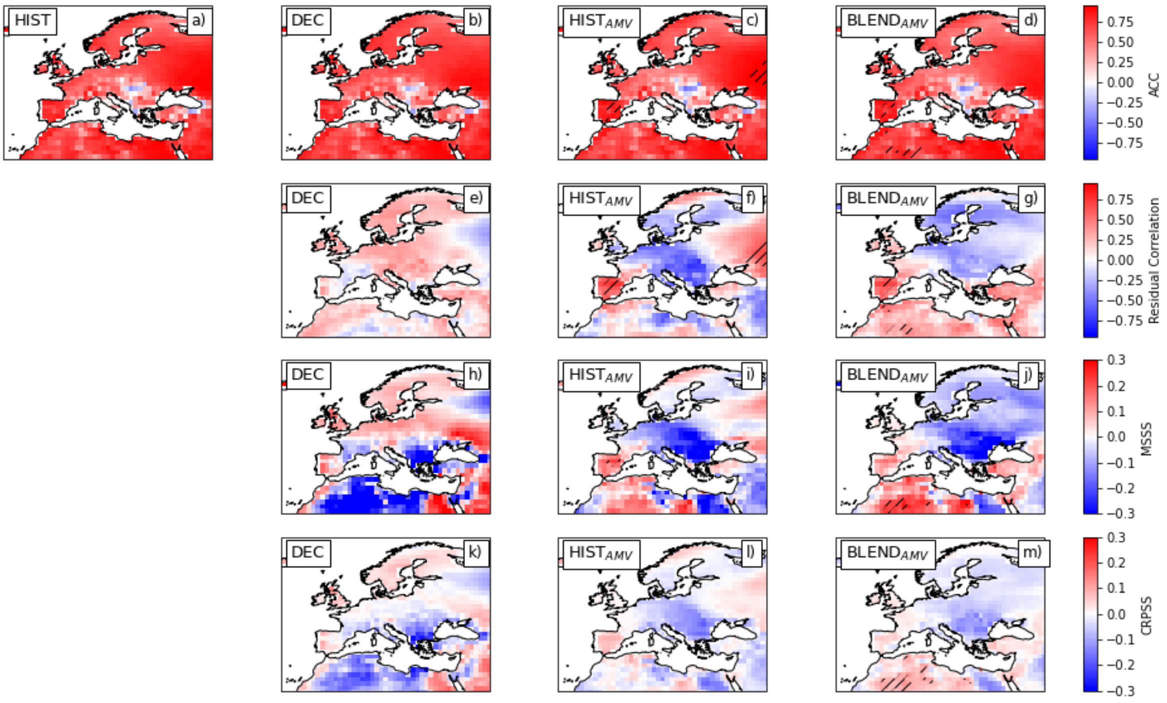


Figure 6: ACC (a–d), Residual correlation(e-g), MSSS (h–j), and CRPSS (k–m) calculated from 10-year forecasts of winter surface temperature over the evaluation period 1966–2000 for HIST (a), DEC (b, e, h, k), $\mathrm{HIST}_{\mathrm{AMV}}$ (c, f, i, l), and $\mathrm{BLEND}_{\mathrm{AMV}}$ (d, g, j, m). For $\mathrm{BLEND}_{\mathrm{AMV}}$, the summer surface temperature averaged over MED is used for the second-step selection. Hatched regions indicate significant improvement in comparison to randomly drawn subsets of HIST (see Section 2.4).

Specific comments

Line 18: Here and elsewhere in the manuscript it is claimed that the method can outperform decadal climate predictions. However it is not clear where this 'outperforming' is shown – presuming that the skill scores in e.g. Fig. 4 are calculated using the historical simulations as reference forecast (this is not explicit, but obviously it can only be either historical runs or decadal predictions as reference forecast here – so I wonder where the results using the other reference forecast are shown?)

Indeed, the MSSS, RPSS, and CRPSS shown in Figure 4 are calculated relative to HIST. We argue that the method outperforms HIST in some regions, as $\mathrm{HIST}_{\mathrm{OBS}}$ and $\mathrm{BLEND}_{\mathrm{OBS}}$ yield MSSS, RPSS, and CRPSS values greater than 1. In several cases, this improvement is comparable to or larger than that provided by DEC, which can exhibit relatively low skill in certain regions (e.g., WCE in JJA). On this basis, we suggest that the method can also outperform DEC in specific contexts.

To better quantify this added value, we have now included statistical significance testing for all the scores considered, as noted in our previous response. The figures, results, and corresponding text have been revised accordingly in the updated manuscript.

We also clarified the legend of Figure 4:

"Figure 4: MSSS, RPSS, and CRPSS (see Section 2.4) calculated from the 5-, 10-, and 15-year time series of surface temperature forecasts for winter (a, c, e) and summer (b, d, f) over NEU (a, b), WCE (c, d), and MED (e, f). Scores are shown for the hindcast dataset (see Section 2.1) as well as for the dataset derived from BLEND (see Section 2.3). HIST is used as the reference; therefore, positive values indicate an improvement relative to HIST. Black edges indicate scores significantly better than those obtained from randomly drawn subsets of HIST (see Section 2.4). The scores are calculated over the 1966–2000 period, using the ensemble mean for MSSS and the full ensembles for RPSS and CRPSS."

Line 36: check wording, seems to imply that the internal variability as such is "resolved" as if it was a process that could be parametrized at low resolution. However, both coarse-scale and high-resolution models do simulate their model-specific internal variability.

We agree that each model does simulate their own internal variability. We were referring to the fact that a large ensemble of non-initialized simulations encompass the full range of uncertainty related to internal climate variability for each model, making internal climate variability explicitly resolved. As this can be confusing, we changed the sentence in the revised manuscript:

“As such, they encompass the full range of uncertainty including that related to internal climate variability.”

Lines 36-37: providing some references to explain initialised predictions could be useful, e.g. Meehl et al.2021 (https://doi.org/10.1038/s43017-021-00155-x) or others

This reference is now added in the revised manuscript.

Line 53: Is it really “crucial” to have seamless climate climate information – cannot decisions also be made by using the best information available for different specific time scales? Can you provide a reference for the statement that it would be “crucial that climate information be seamless”?

Decision-makers often require climate information that spans multiple time scales. In that case, using different types of climate information can result in statistical discontinuity between dataset (as shown in Befort et al. 2022), which can result in inconsistencies when they are combined to inform medium- to long-term planning.

We changed the text and added a reference in the revised manuscript:

“Yet, for effective decision-making and long-term adaptation planning, it is important that climate information be seamless across timescales, ensuring consistency between historical observations, near-term predictions, and long-term projections (e.g. Nissan et al., 2019; Befort et al., 2022).”

Reference:

Befort, D. J., Brunner, L., Borchert, L. F., O’Reilly, C. H., Mignot, J., Ballinger, A. P., ... & Weisheimer, A.: Combination of decadal predictions and climate projections in time: Challenges and potential solutions. Geophysical Research Letters, 49(15), e2022GL098568, https://doi.org/10.1029/2022GL098568, 2022.

Nissan, H., Goddard, L., de Perez, E. C., Furlow, J., Baethgen, W., Thomson, M. C., & Mason, S. J.:. On the use and misuse of climate change projections in international development. Wiley Interdisciplinary Reviews: Climate Change, 10(3), e579, https://doi.org/10.1002/wcc.579, 2019.

Line 69-74: In this context, it may also be worthwhile to mention the recent study by Acosta et al . 2025 (https://doi.org/10.5194/esd-16-1723-2025) that developed seamless seasonal to multi-annual predictions using similar analogue-based methods

This reference is now added in the revised version of the manuscript:

“Similarly, Cos et al. (2024) provide a comparison of these methods to predict near-term mediterranean summer temperature but show instead heterogeneous improvements in comparison to the full non-initialized climate simulations from the Coupled Model Intercomparison Project Phase 6 (CMIP6). Building on the method from Mahmood et al. (2021), Acosta et al. (2025) show that this type of constraining method has the potential to fill the gap between seasonal to multi-annual timescales by delivering seamless climate information while having a very low computational cost.”

Line 92: That is 70% of land in the CNRM-CM6-1 model, or do you use the (regridded) land fraction from each model (as the masks and land fractions may differ between models)?

For consistency, all models were first regridded onto the CNRM-CM6-1 model atmospheric grid. Grid points with at least 70% land fraction were then retained so that the same land points are used across all models. We clarified this in the revised manuscript:

“For simplicity, all data were first regridded to the CNRM-CM6-1 atmospheric grid, and only grid points with at least 70% land fraction were then considered.”

Line 92: “enough grid points” not clear – enough for what?

If only grid points with a land fraction of 100% were retained, the number of grid points along coastlines would be substantially reduced, particularly around the Mediterranean, including regions such as Italy (Figure R1). To avoid losing these coastal areas, we therefore retained grid points with at least 70% of land fraction.

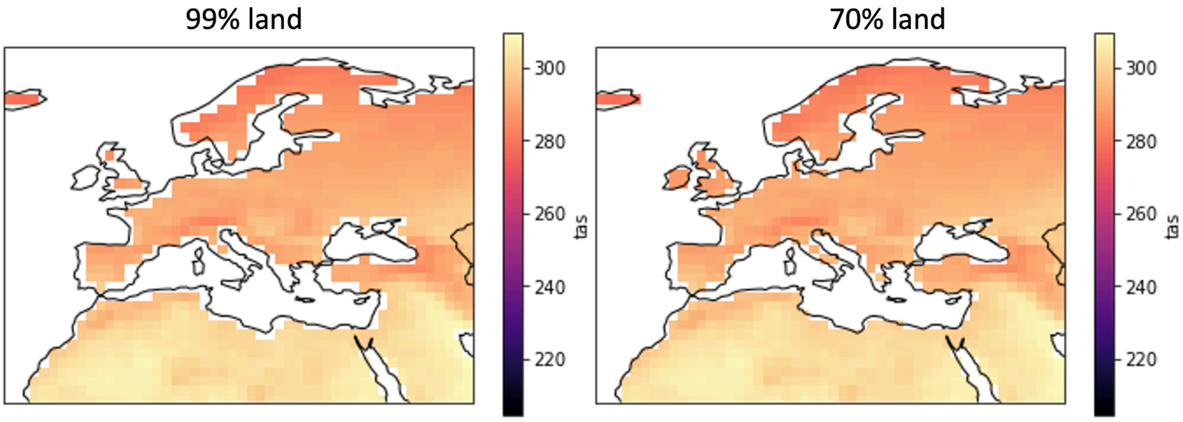


Figure R1: Surface temperature ensemble mean from the CNRM-CM6-1 ensemble averaged over the 1850-2024 period considering grid points with at least (left) 99% of land and (right) 70% of land.

Line 93: lead-time dependent

Done

Line 100 / Table 1: some of the models in the table provide several hindcast experiments, using different initialisation approaches. For reproducibility, please specify which experiments (e.g. which run identifier in ESGF) were used

Table 1 is now updated in the revised manuscript with the members used in the hindcast experiments for each model. Note that some mistakes were made in the previous table, for the MRI-ESM2-0 model with 10 members instead of 12 and the model EC-Earth with 16 members instead of 15. Please find below the updated table:

Model (historical)	Number of members	Model (hindcast)	Simulations

CNRM-CM6-1	30	CNRM-ESM2-1 (5-yr)	25 r(1-15)i1p1f2 r(1-10)i1p1f2

EC-Earth3	15	EC-Earth3
		(10-yr)
			16
			r(1-10i1p1f1)
			r(6-10)i2p1f1)

MIROC6	50	MIROC6 (10-yr)	10 r(1-10i1p1f1)
MRI-ESM2-0	12	MRI-ESM2-0 (5-yr)	10 r(1-10i1p1f1)

NorCPM1	30	NorCPM1
		(10-yr)
			20
			r(1-10i1p1f1)
			r(1-10i2p1f1)

IPSL-CM6A-LR	26	IPSL-CM6A-LR (10-yr)	10 r(1-10i1p1f1)

Table 1: CMIP6 models and associated number of historical simulations used in this study (left) and CMIP6 models and the associated number and list of hindcast simulations, as well as their time length, used in this study (right).

Line 113-115: at which time scale (temporal averaging) did you calculate AMV, did you apply temporal smoothing? (often AMV calculations apply a filter over e.g. 11 years)

We applied a a temporal smoothing to the AMV index as specified line 140-142: “A Lanczos low-pass filter with a cutoff frequency of 1/10 years and 11 weights is applied to all indices to retain only the low-frequency variations, except for the third one.”

To improve clarity, we removed this point from the end of the section and incorporated it into the paragraphs describing each climate index to which this filtering is applied.

Please find below the revised text for the AMV:

“To estimate the evolution of the AMV, we define the AMV index as the average low-pass filtered annual SST over the North Atlantic (0–60° N, 80°W–0° E) after the removal of the externally forced signal following Trenberth and Shea, (2006). A Lanczos low-pass filter with a cutoff frequency of 1/10 years and 11 weights is used for filtering.”

Line 124: I did not think that Befort 2024 constrained based on spatial correlations?

Indeed, Befort et al. (2020) used average SST over the North Atlantic subpolar gyre instead of SST spatial correlation, we removed this reference in the revised manuscript:

“This index has been proposed in previous studies to constrain low-frequency internal climate variability in surface temperature by selecting the simulations that best match the observed SST spatial pattern of sea surface temperature based on spatial correlations at global scale (e.g.Mahmood et al. 2022).”

Line 143: “[1-5]-yr” not clear – does this indicate a 5-year average, or 5 individual years?

It refers to the 5-year average and not individual years. Note that following the recommendations from Reviewer 1, the case study used to describe the method is now based on summer land surface temperature over WCE.

This is clarified in the revised version of the manuscript:

“The method is illustrated here using a case study: the forecast of the summer land surface temperature average over lead years 1 to 5 in the WCE region, as defined by the IPCC (Iturbide et al., 2020), starting in 1983 (Fig. 1).”

Line 155/159: and elsewhere across the text, it seems your dataset acronyms are not consistent, e.g. BLEND_OBS (capital OBS) and BLEND_obs (lower-case obs) – or do these different writings refer to different things? Similar further down for HIST_hindcasts and HIST_Hindcasts. Please check consistency of terminology throughout the text and figures.

We apologize for the inconsistencies. These acronyms refer to the same dataset. The notation has now been corrected and made consistent throughout the revised manuscript

Line 153-160: it seems a bit confusing that in some cases 20 simulations are selected, and in other cases 30 simulations/members. What is the rationale for these different numbers? Overall, this description of the exact (admittedly complex) method seems hard to follow.

All HIST$_{OBS}$ and BLEND$_{OBS}$ subsets used for the evaluation contain 20 members, ensuring a consistent and comparable assessment across methods. The difference arises from how these subsets are constructed: HIST$_{OBS}$ is derived from a single selection step, whereas BLEND$_{OBS}$ results from a two-step selection procedure. As a result, a larger number of simulations is retained in the first step for BLEND$_{OBS}$ before the final subset is defined.

In the first step, members are selected from the historical simulations (HIST) based on their similarity to the observed predictor 20 years prior to the forecast, following an approach similar to that used for HIST$_{OBS}$. A total of 30 members are retained to allow for a second selection step based on the decadal predictions. In this second step, the 20 members whose surface temperature is closest to the decadal prediction ensemble mean over the region of interest are selected to form the final BLEND$_{OBS}$ subset.

Following the correction described above, we have updated the method so that 50 members (instead of 30) are retained in the first selection step for BLEND$_{OBS}$. This provides a more robust intermediate sampling prior to the final selection, which is now of 30 members. So now all HIST$_{OBS}$ and BLEND$_{OBS}$ subsets used for the evaluation contain 30 members. The description of the method in the manuscript has been revised accordingly to clarify this procedure.

Line 164: does 90th percentile range mean the range between the 5th and the 95th percentile?

Yes it means 5th and the 95th percentile, it is clarified in the revised manuscript. Note that as suggested by Reviewer 1, the case study used to describe the Method has been changed. PLease find below the updated text in the revised manuscript:

“In this illustrative example, HIST treated here as the benchmark ensemble predicts a slight cooling with a substantial uncertainty assessed by the spread, namely -0.09 ± 0.81 °C (ensemble mean ± 5-95th percentile range). DEC has an ensemble mean closer to the observations (-0.43°C) than HIST, namely -0.28 ± 1.12 °C. The temperature forecast from HIST_{OBS}, -0.19 ± 0.72°C, shows a decrease in spread and a closer ensemble mean to the observation in comparison to HIST, while $\text{BLEND}_{\text{OBS}}$, -0.23 ± 51°C, reduces the uncertainty and has an ensemble mean even closer to the observation."

Line 164/165: I am not convinced that ±0.6 is robustly smaller (and indicates reduced uncertainty as claimed here) than ±0.62 in the line above, also considering that DEC has a smaller ensemble size.

We did not claim that the reduction in uncertainty was significantly or robustly smaller, and we agree that the difference is negligible. Please refer to the updated text provided in the previous response.

Line 190 / Figure 1 caption: please check the sentence "The data used for the historical simulations and the hindcast simulations mixed all the models together (see section 2.1) are described in Tables 1 and 2." – seems grammatically off

Sorry, it is now corrected in the revised manuscript:

"For both the historical (Table 1) and hindcast (Table 2) simulations, the datasets combine all the models (see Section 2.1)."

line 199-201: unclear: 1967 is 27 years after 1940 when the ERA5 data starts. If using 20-year windows, why not start the hindcasts in 1960? Also the DCPP hindcasts are starting in 1960, so it is unclear why you would decrease your already small sample size further?

The starting date depends on the observations data, as we choose the subset of members from HIST based on their similarities with the observations 20 years before the forecast. We use a lanczos low-pass filter of 11 years prior to the calculation of the metrics to estimate the similarity between the members and the observations, which remove the first and the last 5 years of the observed time series. As the ERA5 reanalysis starts in 1940, it means that the evaluation period could start in 1965. However, we calculate the winter NAO (from December to February), so the time-series starts in 1941 for the first complete winter, which means that the observed low-pass filter winter NAO time-series starts in 1946. Therefore, the evaluation period can start in 1966. In the revised version of the manuscript, we recalculated all the scores and redid the plots based on the new evaluation period 1966-2014.

We clarified this point in the revised manuscript (see below) and modified the text throughout the manuscript accordingly.

"Because the 1966 forecast relies on processed data (e.g., filtering, a 20-year observation-constraining period) starting in 1940—the initial year of the ERA5 product—the evaluation period begins in 1966."

Line 205: can you please specify: these years 1967-2000 are initialisation years – So 1-15 year forecast initialised in 2000 would go until 2014?

Indeed, 1967-2000 are initialisation years, so the 1-15 year forecast initialised in 2000 would go to 2014. We clarified this point in the revised manuscript:

"To this end, we perform a retrospective evaluation of temperature forecasts averaged over the three regions of interest. BLEND is applied each year over the 1966–2000 period. For each initialization year, we evaluate winter and summer temperature forecasts averaged over lead times 1–5, 1–10, and 1–15 year lead times. As some hindcasts simulations start in January, the first winter is computed using only January and February."

Line 206: it would be good to already mention here which predictions you are considering as reference forecast in the skill score calculations.

This is now added in the revised manuscript:

"They indicate the skill of a forecast against a reference forecast, with positive values indicating better skill than the reference. We use HIST as the reference to assess whether the forecast provides added value beyond external forcing."

Line 229: which block size are you using for the boot strapping? Given the auto-correlation in the data, it is important to use block bootstrapping (see e.g. Goddard et al. 2013, https://doi.org/10.1007/s00382-012-1481-2)

Thank you for this comment, we were using a 5-yr block bootstrapping. In the revised manuscript we now apply a resampling approach based on randomly drawing subsets from the HIST ensemble to assess statistical significance. Please see our previous response for a detailed description of the updated significance testing method.

Line 229: Please also specify how you are testing the significance of the other skill metrics (ACC, RPSS, CRPSS, MSSS). Several of the claims about improvements/added values are based on these skill scores, and also the distributions of spread and error, so these statements should be underpinned by suitable tests of the statistical significance of the added values.

The skill metrics used in the manuscript (RPSS, CRPSS, MSSS) are all defined relative to a reference forecast and therefore already provide a quantitative measure of the added value and forecast capability compared to that baseline. Therefore, by definition, positive skill values directly indicate improved performance over the reference, which is why we did not initially complement these diagnostics with formal statistical significance tests.

That said, we agree that explicitly assessing the statistical significance of the reported improvements would strengthen the robustness of the conclusions, particularly for the distributions of spread and error and for comparisons based on the different skill scores. In the revised version of the manuscript, we therefore include a dedicated significance assessment based on a resampling approach (see previous responses). The Figures are now updated in the revised version of the manuscript (see previous responses). This addition allows us to formally quantify the confidence in the reported added values, with corresponding updates to the results where relevant.

Line 239: Please make sure that this methodological detail that analogues are selected solely based on 5-year hindcasts is specified in the Methods section 2.3. I may have missed this information there?

This methodological detail was described in the Method section 2.3 for $HIST_{hindcast}$, but not for the BLEND subset: “Second, $HIST_{hindcast}$, which derives from the selection of the 20 simulations from HIST that are closest to the 5-year [1977-1981] hindcast ensemble-mean surface temperature.”.

We clarify this point in the Method section in the revised version of the manuscript: “Then, 30 simulations out of 50 that show the lowest absolute error with respect to the 5-yr hindcast ensemble mean surface temperature over the region of interest, are retained.”

Line 241: Please briefly explain what you mean by “added value”, and how it is measured

As the aim of decadal prediction is to reduce uncertainty due to internal climate variability, we define “added value” here as a reduction in spread in decadal predictions (DEC) compared to historical simulations (HIST), which encompass the full range of uncertainty associated with internal climate variability. A reduction of the spread due to internal variability is considered an added value, because it enables the provision of near-term climate information with lower uncertainty, which is essential for adaptation planning

We clarified this point in the revised manuscript:

“The spread and absolute errors in DEC are overall relatively close to the one in HIST. This absence of clear spread and error reduction in DEC may be due to several factors.”

Line 244-245: Is this earlier finding f the initialisation shock triggering El Nino events model-specific, or does it systematically affecting the multiple decadal prediction systems that you are using? In other words, how relevant is this for your analysis?

The result reported by Sánchez-Gómez et al. (2016) is based on analyses performed with the CNRM-CM5 model, which belongs to the previous CMIP5 generation and is not used in the present study. To our knowledge, a systematic assessment of initialization shocks triggering El Niño events has not yet been carried out across multiple decadal prediction systems or model generations, and it therefore remains unclear to what extent this behavior is model-specific.

We nevertheless consider this reference relevant for our analysis, as it provides a clear example of how initialization shocks can influence surface temperature predictability, particularly in the early years of decadal forecasts. We nevertheless clarified that this result is only based on the CNRM-CM5 prediction system in the revised manuscript:

“The spread and absolute errors in DEC are overall relatively close to the one in HIST. This absence of clear spread and error reduction in DEC may be due to several factors. One possible explanation is poor hindcast performance, potentially of structural origin—for instance related to initialization shocks that can quasi-systematically trigger El Niño events during the first forecast year, as well as to a negative NAO-type mean bias, as reported for the CNRM-CM5 decadal prediction system by Sanchez-Gomez et al. (2016).”

Line 245: “any add values”?

Done.

Line 246: Not clear what exactly you refer to by “DEC poor scores”?

As mentioned above, DEC simulations aim to reduce uncertainty relative to internal climate variability. Therefore, we refer here to “DEC poor scores” to indicate that DEC simulations provide only limited or no reduction of the spread for certain regions, while not providing a reduction in the absolute errors.

We clarified this point in the revised manuscript:

“The spread and absolute errors in DEC are overall relatively close to the one in HIST. This absence of clear spread and error reduction in DEC may be due to several factors. One possible explanation is poor hindcast performance, potentially of structural origin—for instance related to initialization shocks that can quasi-systematically trigger El Niño events during the first forecast year, as well as to a negative NAO-type mean bias, as reported for the CNRM-CM5 decadal prediction system by Sanchez-Gomez et al. (2016). These effects may overshadow the added value of ocean initialization in DEC, thereby degrading hindcast quality at all lead times. Alternatively, the limited spread reduction in DEC compared to HIST may reflect an intrinsic, or “true,” climate origin. In particular, decadal variability in European surface temperature is relatively weak compared to the strong chaotic atmospheric variability operating at intraseasonal to interannual timescales, meaning that the predictable signal may be masked by noise. Finally, model deficiencies in the forecasting system may also contribute to this behavior.”

Line 242-248: I find this discussion rather confusing, seems to mix up several issues?

The aim of this text is to discuss the potential causes of this lack of spread and absolute error reduction in DEC in comparison to HIST. Although an in-depth analysis is out of the scope of this study, we found it relevant to describe the main potential reason, notably because reducing the spread is one of the goals of the new method that we developed here, as it is relevant for adaptation strategies. We modified this discussion in the revised manuscript to improve the clarity, as well as to the updated results with significance tests. The updated discussion is provided in the previous answer.

Line 265 / Figure 2 caption: should the panel reference for WCE be (c,d)?

Yes, thanks, it is now changed in the revised manuscript

Figures 2/3: it is very hard to see which of the box plots belongs to which dataset. Besides enlarging the legend for better readability, maybe also the choice of colours could be optimised to better distinguish/identify the specific datasets.

We changed the colors in the revised version of the manuscript to improve the readability of the figures. The HIST and BLEND dataset based on the same climate indices (e.g. AMV) are now paired using common color, the one for HIST being darker and the one for BLEND being lighter.

Please see the previous answers with the updated Figures.

Line 274: remove comma after CRPSS; and should add “in” after winter (in NEU)

Thanks, done.

Line 275: “better” in comparison to what?

“Better” refers to performance relative to HIST, which is used as the reference forecast for the calculation of MSSS, RPSS, and CRPSS.

Note that Sections 3.1 and 3.2 have been fully rewritten to reflect the revisions described at the beginning of the review process, including the updated results and the new significance testing procedure.

Line 285: winter “in” WCE?

Done

Line 285: for statements like “perform better than HIST”, etc., please apply a suitable significance testing to avoid interpreting noise

The MSSS, RPSS and CRPSS are defined relative to a reference forecast and therefore already provide a quantitative measure of the added value and forecast capability compared to that baseline. Therefore, positive skill values directly indicate improved performance over the reference, whereas negative skill values indicate a deterioration in the forecast performance over the reference. This is why we did not initially complement these diagnostics with formal statistical significance tests.

However, we agree that formal significance testing is important to more robustly determine whether the reported improvements relative to HIST are not attributable to sampling variability or noise. Please see our previous answers about this point.

Line 297: all the datasets (plural)

Done.

Line 307: all leadtimes (plural)

Done.

Line 321: I am sorry, I do not see a clear error reduction in Figure 3e. The box plots pretty much overlap, any slight shifts may be noise.

We agree that the box plots pretty much overlap and that it is difficult to see directly on the Figure. We were mentioning a lowering of error relative to the median of the distribution. We now added the significance relative to this decrease in the median of the distribution in Figure 2 and 3 and modified the text accordingly.

Line 335: for “tailoring the choice of observational predictors” – so which predictor do you recommend for which region?

We have revised the text to more clearly indicate which observational predictor shows the strongest added value for each region and season. The updated discussion now highlights the most suitable predictor based on the results shown in Supplementary Figures S1–S3, which have also been updated to reflect the new significance testing. Please find below the revised text incorporated into the manuscript.

"These results are summarized in Figures S1, S2, and S3. Overall, the results are quite heterogeneous: in some cases, BLEND provides no clear added value relative to HIST —for example, for the 15-year forecast over NEU in winter. In other cases, however, BLEND shows substantial added value in comparison to HIST. This is particularly evident over WCE in summer, where the HIST_{TAS} and $\text{BLEND}_{\text{TAS}}$ dataset shows large improvements for the 5-, 10-, and 15-year forecasts. Both $\text{BLEND}_{\text{AMV}}$ and $\text{BLEND}_{\text{SPG}}$ also provide some significant improvement in 10- and 15-year forecasts. This is also the case for the HIST_{TAS} dataset, which provides some improvements for summer forecasts over NEU and MED.

Finally, this is also the case for winter forecast temperature over MED, although the BLEND-derived dataset that shows improvement differs depending on the forecast horizon. The HIST_{TAS} and HIST_{NAO} and HIST_{AMV} provide large improvement for 5-year forecasts (Fig. S1), whereas $\text{BLEND}_{\text{TAS}}$ and $\text{BLEND}_{\text{AMV}}$ show significant improvement for 10-year forecasts (Fig. S2). Therefore, BLEND appears to capture part of the low-frequency internal variability of winter temperatures in MED. This is consistent with previous studies (e.g., Mariotti and Dell'Aquila, 2012) that emphasize the role of the AMV in modulating the decadal variability of MED winter surface temperatures.

The fact that HIST_{TAS} provides a significant reduction in absolute errors across several regions may be due either to the use of regional surface temperature inherently capturing the associated low-frequency variability, or to the possibility that the observed surface temperature lies outside, or at the edge of, the distribution of historical simulations, meaning that selecting simulations from this part of the distribution would systematically lead to improved predictions.

The evaluations presented here highlight that tailoring the choice of observational predictors according to region and forecast horizon is key to improving forecast performance."

Fig. 4: Please specify in the caption which reference forecast was used in the skill score calculations? If this figure uses historical simulations as reference(?), did I possibly miss a figure that uses decadal predictions as reference forecast – as several places in the text seem to discuss improvements over the decadal predictions?

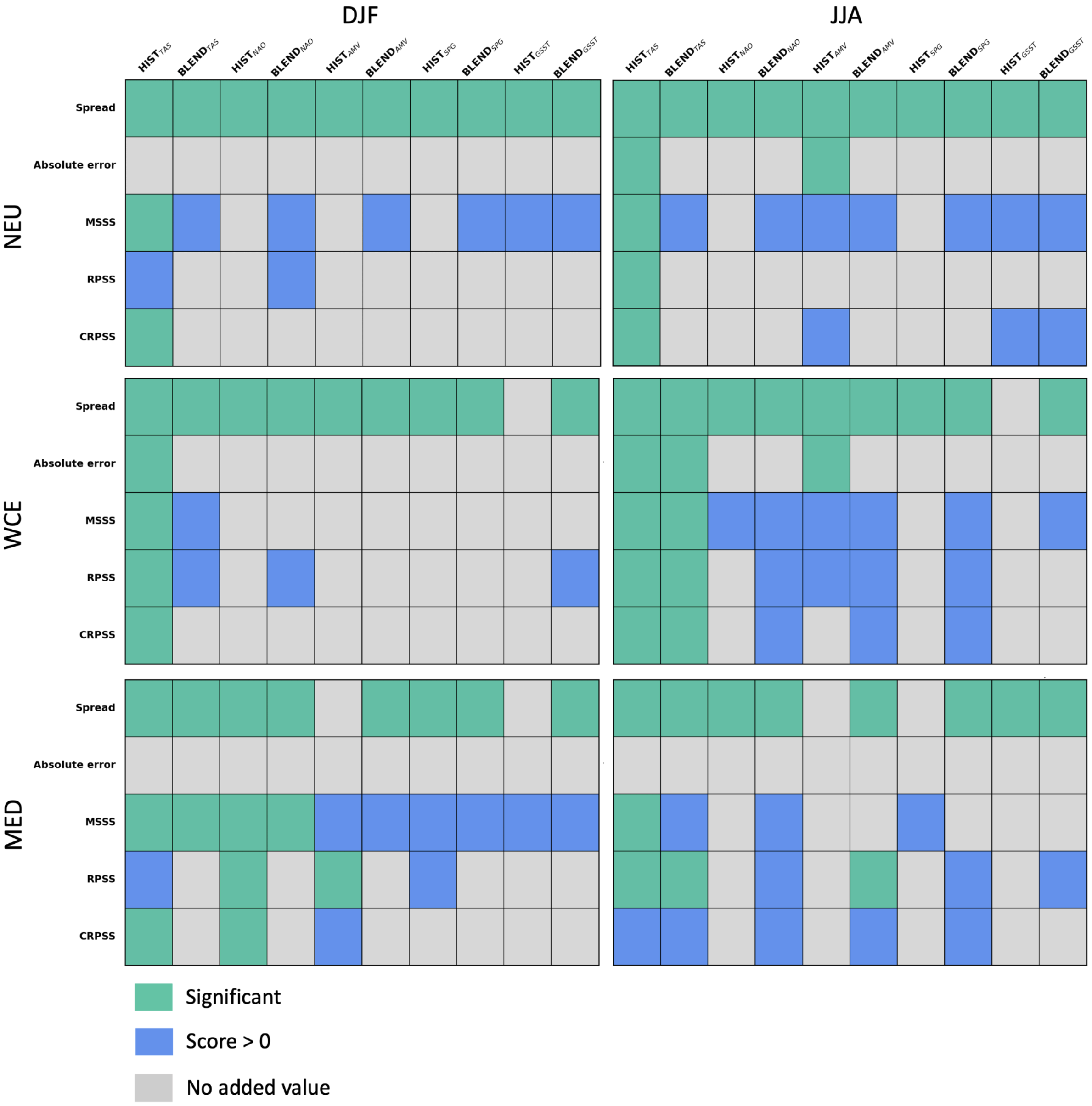
We used the historical simulations as reference. When we talked about improvement in comparison to decadal predictions, we were referring to the fact that some subsets provide higher scores than the decadal predictions, which means that the forecast provides a better forecast improvement relative to the historical simulations than the decadal predictions.

Please find below the new caption of Fig. 4 in the revised manuscript that clarify the reference forecast is used for the score calculation:

"Figure 4: MSSS, RPSS and CRPSS (see section 2.4) calculated from the 5, 10 and 15 years time series of (a, c, e) winter and (b, d, f) summer forecast of surface temperature over (a, b) NEU, (c, d) WCE and (e, f) MED for the hindcasts dataset (see section 2.1), as well as the dataset derived from BLEND (see section 2.3) using HIST as reference. The scores are calculated over the 1966-2000 period, using the ensemble mean for the MSSS and the whole ensembles for the RPSS and CRPSS."

Fig. 4 and related discussion in the text: Can you say something systematic of whether (or how often) the blending improves skill over just the HIST constraint? Also can you highlight for which regions in which seasons the constraint is particularly beneficial, and when not? The current description of these results reads fairly repetitive, and it is hard to grasp the key findings.

We agree that the previous description was repetitive and did not sufficiently summarize the systematic added value of the BLEND approach. In the revised manuscript, we have clarified and reorganized the discussion to better highlight when and where BLEND improves skill relative to HIST alone, distinguishing between regions, seasons, and forecast horizons (please see our previous answer). We also updated Figures S1–S3 (using the new significance testing), which now allow the BLEND-derived datasets with the largest added value relative to HIST to be more clearly identified. Please find below the updated figure from the revised Supplementary Material.



DJF

		HIST$_{TAS}$	BLEND$_{TAS}$	HIST$_{NAO}$	BLEND$_{NAO}$	HIST$_{AMV}$	BLEND$_{AMV}$	HIST$_{SPG}$	BLEND$_{SPG}$	HIST$_{GSST}$	BLEND$_{GSST}$
NEU	Spread	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Absolute error	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	MSSS	🟩	🟦	⬜	🟦	⬜	🟦	⬜	🟦	🟦	🟦
	RPSS	🟦	⬜	⬜	🟦	⬜	⬜	⬜	⬜	⬜	⬜
	CRPSS	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
WCE	Spread	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩
	Absolute error	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	MSSS	🟩	🟦	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	RPSS	🟩	🟦	⬜	🟦	⬜	⬜	⬜	⬜	⬜	🟦
	CRPSS	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
MED	Spread	🟩	🟩	🟩	🟩	⬜	🟩	🟩	🟩	⬜	🟩
	Absolute error	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	MSSS	🟩	🟩	🟩	🟩	🟦	🟦	🟦	🟦	🟦	🟦
	RPSS	🟦	⬜	🟩	⬜	🟩	⬜	🟦	⬜	⬜	⬜
	CRPSS	🟩	⬜	🟩	⬜	🟦	⬜	⬜	⬜	⬜	⬜

JJA

		HIST$_{TAS}$	BLEND$_{TAS}$	HIST$_{NAO}$	BLEND$_{NAO}$	HIST$_{AMV}$	BLEND$_{AMV}$	HIST$_{SPG}$	BLEND$_{SPG}$	HIST$_{GSST}$	BLEND$_{GSST}$
NEU	Spread	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
	Absolute error	🟩	⬜	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜
	MSSS	🟩	🟦	⬜	🟦	🟦	🟦	⬜	🟦	🟦	🟦
	RPSS	🟩	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	CRPSS	🟩	⬜	⬜	⬜	🟦	⬜	⬜	⬜	🟦	🟦
WCE	Spread	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	⬜	🟩
	Absolute error	🟩	🟩	⬜	⬜	🟩	⬜	⬜	⬜	⬜	⬜
	MSSS	🟩	🟩	🟦	🟦	🟦	🟦	⬜	🟦	⬜	🟦
	RPSS	🟩	🟩	⬜	🟦	🟦	🟦	⬜	🟦	⬜	⬜
	CRPSS	🟩	🟩	⬜	🟦	⬜	🟦	⬜	🟦	⬜	⬜
MED	Spread	🟩	🟩	🟩	🟩	⬜	🟩	⬜	🟩	🟩	🟩
	Absolute error	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜	⬜
	MSSS	🟩	🟦	⬜	🟦	⬜	⬜	🟦	⬜	⬜	⬜
	RPSS	🟩	🟩	⬜	🟦	⬜	🟩	⬜	🟦	⬜	🟦
	CRPSS	🟦	🟦	⬜	🟦	⬜	🟦	⬜	🟦	⬜	⬜

🟩 Significant

🟦 Score > 0

⬜ No added value

Figure S1: Summary of the results from the evaluation of the 5-year temperature forecasts of HIST$_{OBS}$ and BLEND$_{OBS}$ for NEU (top row), WCE (middle row) and MED (bottom row) in DJF (left column) and JJA (right column). Each rectangle shows the assessment results for one score of one subset derived from the method against HIST. In green, the forecasts are significantly improved in comparison to randomly drawn 30-member subsets of HIST; in blue, the scores (MSSS, RPSS and CRPSS) are positive, although not significant and in gray, there are no forecast improvements.

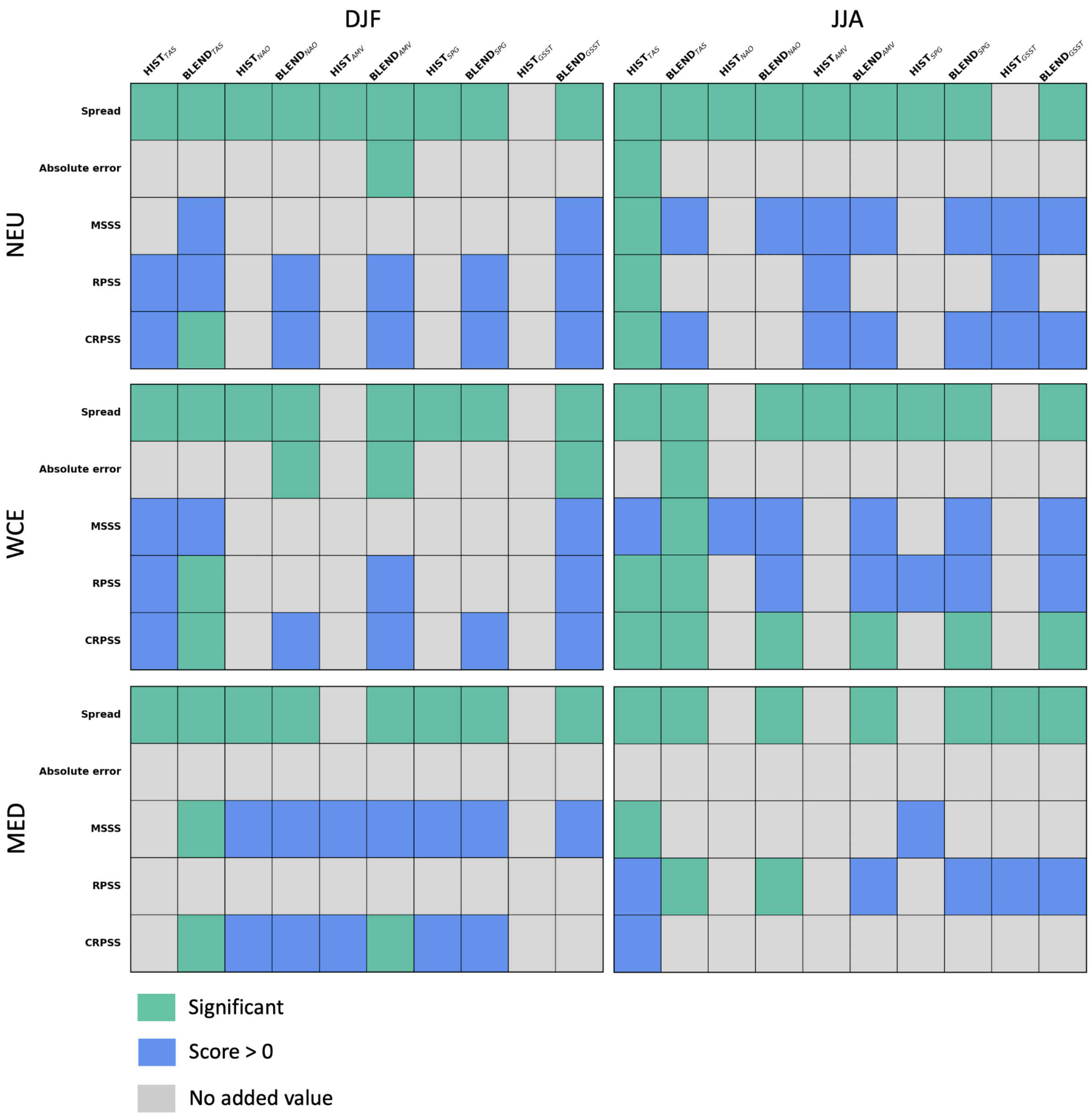


Figure S2: As Figure S1 but for 10yr temperature forecasts.

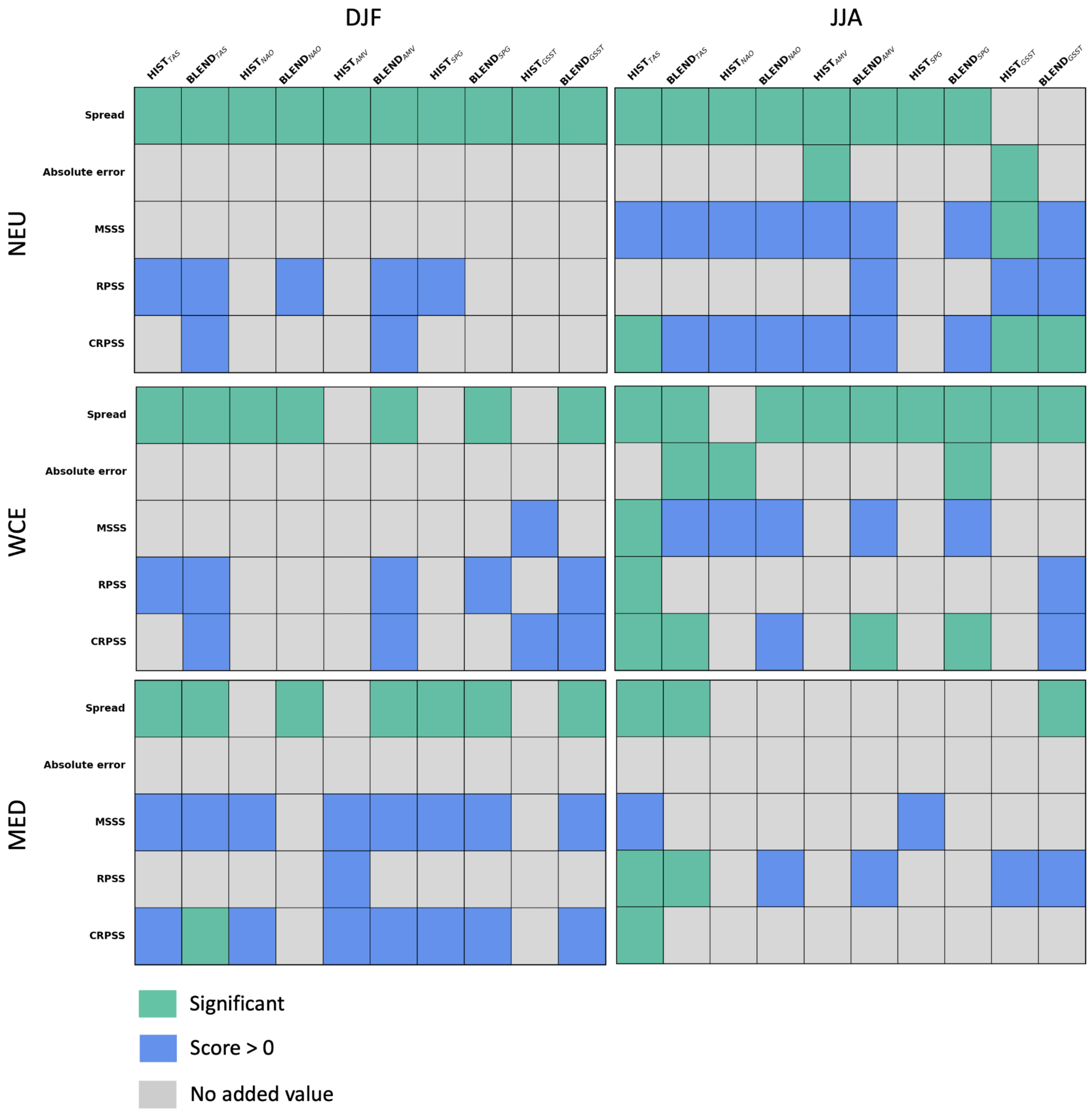


Figure S3: As Figure S1 but for 15yr temperature forecasts.

Line 343 / heading 3.2: the maps discussed here seem to show all of Europe, not just the MED region ?

We chose to present maps for the whole of Europe rather than restricting the analysis to the MED region for two main reasons. First, we aim to assess whether forecasts derived from HIST$_{AMV}$ provide added value beyond the target region, since this subset is based solely on the AMV predictor and not on temperature over the region of interest. Second, we examine whether the region-specific selection used in HIST$_{TAS}$ and BLEND$_{TAS}$ improves forecast skill not only over the target region but also in neighboring areas, or whether it instead degrades performance outside the region of interest.

Line 350: “datasets” here refers to the HIST/DEC and the blended simulations

Yes, we clarified this point in the revised manuscript, where the ACC is now shown to improve the reader's understanding.

"Due to the presence of anthropogenically forced warming trends, which largely dominate the forecast signal, the anomaly correlation coefficient (ACC) between observed temperature and forecast ensembles is very high across much of Europe for HIST, DEC, and the BLEND-derived datasets (Fig.s 5 and 6a–d)."

Line 354: "significant skill improvement" – how is the significance indicated? Similar question applies to line 356.

Thank you for pointing this out. There was an error in the original text, and the results referred to here correspond to Figure S4. This has now been corrected in the revised manuscript. In addition, we have updated the significance testing procedure (see our previous response). We have also added the residual correlation maps within Figures 5 and 6 to improve clarity and facilitate interpretation.

Line 359 – 370: is any of the results discussed here statistically significant?

Please see our previous response for the first part of the discussion. For the second part, no significant test has been made as we used the MSSS and CRPSS scores, which by construction indicates an improvement of the evaluated forecast over the reference for positive values, and therefore provide a direct measure of relative performance.

As indicated previously, a significativity test has been added in the revised manuscript using a resampling method (please see our previous answer).

Line 373-374: not clear, why does the similarity to regional average time series degrde the skill here? I may be missing something to understand the discussed link?

In the second step of the method, members are selected based on the similarity of the temperature average to the hindcast over the target region - here MED. As a result, the selection is optimized specifically for that region. We therefore do not necessarily expect an improvement in forecast skill over other regions (e.g., NEU); skill may even degrade there because the selected members are not specifically constrained to represent temperature variability in those areas.

This behaviour is illustrated in Figures 5 and 6, where the forecasts from $\mathrm{BLEND_{AMV}}$ (which use the two-step selection based on both observations and hindcasts) perform worse in some regions than those obtained using only the first selection step based on observations. This indicates that the second, region-specific selection can improve consistency with the targeted region while reducing performance elsewhere, as it prioritizes similarity to the regional hindcast signal rather than large-scale temperature coherence.

Line 377/378: "compared to both the historical and hindcast ensembles" – not clear, which results show skill over the historical, and which show skill over the hindcast ensemble?

The results in Figure 5 and 6 are calculated using HIST as reference. This was missing in the legend. However, when the score from the method is higher than the score from the hindcast, although it is relative to HIST, it means that the subset provides more added value than the hindcast.

We clarified this point in the legend of Figure 5 and 6 of the revised manuscript:

"Figure 5: MSSS calculated from the time series of 10 years forecast of winter surface temperature over the evaluation period (1967-2000) relative to the reference HIST for (a) the hindcasts dataset (see section 2.1), (b) $\mathrm{HIST}_{\mathrm{Hindcast}}$, (c) $\mathrm{HIST}_{\mathrm{AMV}}$, (d) $\mathrm{BLEND}_{\mathrm{AMV}}$, (e) $\mathrm{HIST}_{\mathrm{NAO}}$ and (f) $\mathrm{BLEND}_{\mathrm{NAO}}$. The winter surface temperature averaged over the Mediterranean region is used for the second step selection in $\mathrm{BLEND}_{\mathrm{AMV}}$ and $\mathrm{BLEND}_{\mathrm{NAO}}$."

Also section 3.2 is very difficult to follow: discussing for each method little improvements (whether significant or not) over sub-regions makes the text appear repetitive, and it is hard to identify the really relevant features and messages. Can this be synthesised at higher level, for potential users to have clearer messages on specific benefits (or lack thereof) of the methods?

We modified the section to better highlight the relevant common results from the methods.

Figures 5 and 6 should highlight where the scores are significant. And ideally focus the discussion in the text on such significant features. Also please indicate in the figure caption which is the reference forecast for these skill calculations.

As indicated previously, the legends of Figure 5 and 6 have been modified in order to clarify the reference forecast (here HIST). A significance testing has also been added to Figures 5 and 6 of the revised manuscript.

Line 403: but this information that you incorporate may be affected by the drift (despite applying basic correction for leadtime-dependent climatologies)?

Indeed, the information derived from the hindcasts may still be affected by model drift, even after applying a lead-time-dependent climatological correction, and this could influence the second selection step. In particular, residual drift may affect the regional mean temperature used as a constraint and potentially lead to a suboptimal selection of members and therefore poor forecast.

However, the final forecasts produced by the method are based on historical simulations rather than on the hindcasts themselves. As a result, they are not subject to initialization drift, which primarily affects decadal predictions. In this framework, the hindcasts are used only to guide the selection toward dynamically consistent states, while the resulting simulations retain the stability and long-term consistency of the historical runs.

We acknowledge that residual drift in the hindcasts may still influence the selection and therefore represents a limitation of the approach. This point has been clarified in the revised manuscript:

"The performance of the BLEND method depends on the quality of decadal prediction systems, which can still be affected by drift even after applying a lead-time-dependent climatological correction, potentially leading to a suboptimal selection of members. It also depends on the ability of models to correctly capture teleconnections between the climate indices used as predictors and the variable of interest, which may be limited in some regions. Nevertheless, our results show that it can provide substantial improvements of 5-, 10- and 15-year winter and summer temperature forecasts over Europe, with reduced uncertainty relative to the historical or hindcast ensembles. This added value is also visible regionally, as illustrated in the case study (see section 3.2), where $\text{BLEND}_{\text{TAS}}$ approaches show large improvements in 10-year forecasts of summer temperature over WCE in comparison to HIST."

Line 408: I am not sure I have seen the added value over hindcasts, presuming the skill scores in Fig. 4, 5, 6 are calculated against the historical simulations as reference forecast?

Please see our previous answers.

Line 408/9: "good added value" compared to what?

Compared to HIST, we clarified this point in the revised manuscript.

Line 412: again, I am not sure where to see the skill calculated against the hindcasts?

As indicated in a previous answer, the skill scores are calculated using HIST as reference. However, we can compare the added value of the hindcasts and of the subsets derived from the BLEND method against this same reference. When the MSSS, RPSS and CRPSS scores are above the hindcast one, it means that the subsets provide a greater added value relative to HIST in comparison to the added value from DEC, which we think is relevant to highlight.

As this can be confusing for the reader, we modified the text in the revised manuscript.
Please find below the new discussion in the revised manuscript:

"The performance of the BLEND method depends on the quality of decadal prediction systems, which can still be affected by drift even after applying a lead-time-dependent climatological correction, potentially leading to a suboptimal selection of members. It also depends on the ability of models to correctly capture teleconnections between the climate indices used as predictors and the variable of interest, which may be limited in some regions. Nevertheless, our results show that it can provide substantial improvements of 5-, 10- and 15-year winter and summer temperature forecasts over Europe relative to HIST, with reduced uncertainty relative to the historical or hindcast ensembles. This added value is also visible regionally, as illustrated in the case study (see section 3.2), where $\text{BLEND}_{\text{TAS}}$ approaches show large improvements in 10-year forecasts of summer temperature over WCE in comparison to HIST."

Line 415: the initialisation can also improve the representation of trends; not all improvements are necessarily related to internal variability.

Indeed, initialization can also improve the representation of trends. We've focused on this point as the near-term future, which is largely influenced by uncertainty associated with internal climate variability, particularly at regional scales. We modified the text in the revised manuscript to nuance this point:

“These large improvements in the 5-, 10-, and 15-year forecasts from BLEND in some regions in both winter and summer, relative to the historical ensemble mean —which reflects only externally forced responses— suggest that the method captures part of the internal climate variability. They may also indicate that the method improves the representation of the forced signal.”

Line 432: wording does not seem to make sense, please check: “the variable exhibits…associated drivers”? Do you mean drivers are identified, or similar?

We modified the text to improve the clarity in the revised manuscript:

“The advantage of the framework proposed here is that it can be readily applied to other regions or variables of interest, provided that the target variable exhibits internal variability on multi-annual to decadal timescales and has identifiable large-scale drivers.”