

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

Response to reviewer #1

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 $\text{BLEND}_{\text{TAS}}$ forecast in addition to HIST_{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**.

General comment

This paper attempts to derive improved prediction information on various time horizons by combining or “blending” subsetting information from historical CMIP simulations and initialized decadal prediction hindcasts with observational constraints. The authors illustrate how difficult it is to improve predictions in general, and how each region and quantity of interest needs its own combination of methods to improve skill. Since this is a methodology paper, the results of course depend on how good the method is that the authors are formulating. As such, the paper stands as a testament to the difficulties and challenges involved with initialized Earth system predictability on regional scales and long leads. The authors do demonstrate improvements in skill with their methodology over some regions and seasons, which is encouraging, though the complexities of applying their method are somewhat daunting.

Detailed comments

1. Since this is a methodology paper, the authors must take care to ensure that the new terminology and acronyms they are introducing are adequately defined and discussed. For example, the key acronyms DEC and HIST should be defined when the concepts are first introduced in lines 85 and 86.

We apologized for these inconsistencies. This is now clarified in the revised manuscript and DEC and HIST are now defined in the section 2.2 Datasets:

“We use 163 non-initialized transient historical simulations (HIST hereafter) from CMIP6 (Eyring et al., 2016) and 92 initialized decadal hindcasts (DEC hereafter) from CMIP6/DCPP Component A (Boer et al., 2016), based on large ensembles from six models (see Table 1).”

2. Lines 162-173: I may be misinterpreting the results in this illustrative example, but it appears that the observed target temperature anomaly they are trying to predict is very small (+0.09C). Therefore, even though all the predicted values from the various methods are only slightly wrong since the values are tiny, they all predict the wrong sign of the observed anomaly (unless I'm missing something). Yes, the spread decreases, but trying to predict an anomaly that is essential zero (or slightly positive) doesn't seem like a fruitful exercise. But I may have missed something here, in which case the authors should clarify that the ultimate goal is to predict the observed value, even when the spread is a key part of the uncertainty.

The objective of the method developed in this study is to provide a reliable forecast of surface temperature over Europe, with reduced uncertainty compared with historical simulations. In this example, although the target temperature anomaly is close to zero, it still provides a relevant test of the method, as it demonstrates the ability to reduce the spread while keeping the observations within the distribution of the HIST_{AMV} and BLEND_{AMV} subsets. Indeed, the HIST distribution ranges from -1 to 1 °C; therefore, these subsets could be negatively or positively biased if the predictors or hindcasts were not relevant for that season and region.

As the previous example did not clearly highlight the usefulness of the method, we have updated the figure to present a 5-year temperature forecast over WCE for JJA during the 1983–1987 period. In this revised example, the observed target temperature anomaly is -0.43 °C. The updated figure is provided below.

Aim: predict the Y-yr average surface temperature over a region of interest
 E.g. region: WCE; season: JJA; Y = 5-yr; start date = 1983

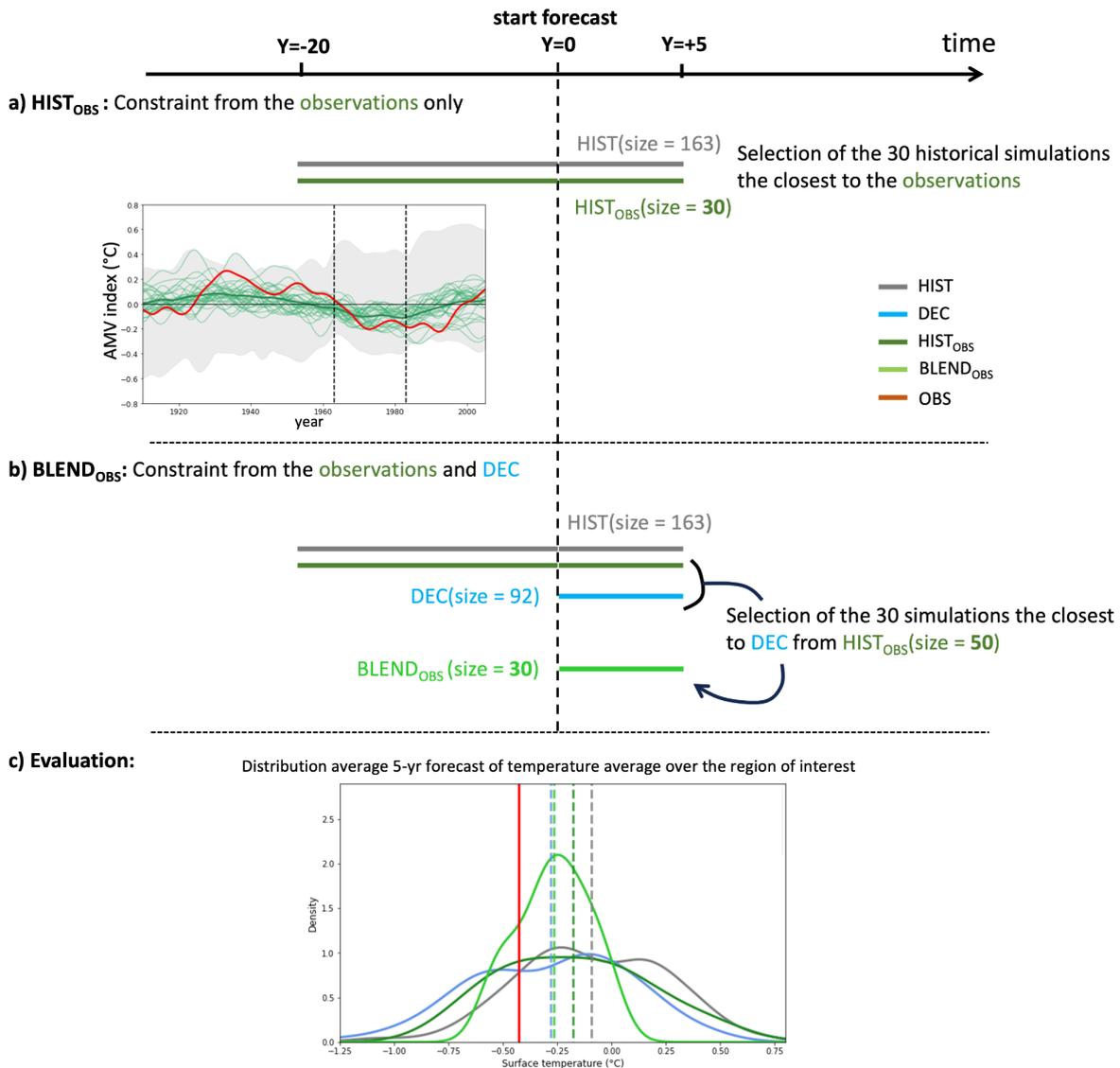


Figure 1. Diagram illustrating the concept and two types of dataset derived from the blending method. For both the historical (HIST) (Table 1) and hindcast (DEC) (Table 2) simulations, the datasets combine all the models (see Section 2.1). In the development of HIST_{OBS} (a) the AMV index (see section 2.2) from the historical simulations (minimum and maximum in gray) is compared to the ERSSTv5 observational dataset (green line), with the selection of the best 20 members in green. In the development of BLEND_{OBS} (b), a first selection of 50 members is then refined to 30 members. The histogram and the ensemble mean (vertical dashed line) of average 5-yr forecast surface temperature from the different datasets are evaluated against the ERA5 reanalysis (red line) (c).

3. Comment: I thank the authors for including Fig. 1 to attempt to diagram their complex methodology. However, in the caption they complicate the issue by using a different acronym “HistOBS” instead of HIST_{obs} as in the figure and text. Also, the observations

are said to be the green line, but then the ERA5 reanalysis observations are the red line. It's very difficult to distinguish the colors in any case, so perhaps the observations could use a dashed black line to better contrast with the solid colored lines from the different methods.

These inconsistencies have been corrected in the revised version of the manuscript. It was a mistake in the legend, the observations are in red, to stay consistent between the Figures. We also changed the solid colored lines from the HIST, DEC, HIST_{AMV} and BLEND_{AMV} into dashed lines for a better contrast. See above for the revised Figure.

4. In Figs. 2 and 3, the most obvious result is reduction in spread and reduced absolute error, especially for winter, at the longer leads (10 and 15 years) for all methods. Maybe the authors commented on this and I missed it, but I presume this is because the greater role of external forcing (increasing GHGs) drives convergence in the predictions at longer leads but internal variability with greater spread and less skill is more at work at the shorter leads (5 yrs)?

Indeed, this overall reduction in the spread and absolute error for all regions and seasons at 10- and 15-yr forecasts in comparison to 5-yr forecast is likely due to the greater role in external forcing, at least for a large part. Part of this reduction may also arise from the averaging out of higher-frequency internal climate variability in the 10- and 15-year forecasts compared with the 5-year forecasts. We agree that this is an interesting point to highlight, so we added a comment in the revised version of the manuscript:

“A reduction in both spread and absolute error (Figures 2 and 3) is observed for the 10- and 15-year forecasts compared with the 5-year forecasts. This is likely due to the stronger influence of external forcing at longer timescales, as well as to the averaging out of higher-frequency internal climate variability in the 10- and 15-year forecasts relative to the 5-year forecasts.”