Milestone 2: Preliminary illustration of the relative merits of the forecast combination and description of the methods identified

D. Verfaillie, Wild, S., Doblas-Reyes, F.J., Donat, M., Mahmood, R., and Bilbao, R.

Summary

Decadal prediction is a rather new area of research that has recently gained momentum in both the climate science and the stakeholder communities. However, some efforts are still needed to assess the forecast quality on 1-10 year timescales, so that users can benefit from the best forecasts available. In this study, we assess the quality of multi-model forecasts compared to single-model forecasts in terms of two characteristics that are crucial for users: reliability and skill. Our results show that, even more important than having very large ensemble sizes, it is crucial to use ensembles composed of different forecasting systems. This is a recommendation that should be passed on to stakeholders through the EUCP project, as we showed that it has important implications not only in terms of forecast skill but also forecast reliability.

Introduction

Lately, initialised decadal climate predictions have been made available for users as a potential source of near-term climate information with the aim of supporting climate-related decisions in key economic and societal sectors such as energy, agriculture and insurance. Pioneering studies of decadal climate prediction (e.g., Smith et al., 2007; Keenlyside et al., 2008; Pohlmann et al., 2009) investigated the capacity of different forecasting systems to accurately “predict” past climate variability in retrospective experiments called hindcasts. At the decadal timescale, the observed climate variability can be understood as the superimposition of an anthropogenically-driven trend on natural fluctuations. While the trend is driven by changes in anthropogenic emissions (mainly GHGs and anthropogenic aerosols), the natural fluctuations

are generated internally by the interactions of the different components of the climate system (atmosphere, ocean and sea ice) or externally by other factors such as volcanic eruptions and solar activity. Provided that these different modes of variability operate on a sufficiently long timescale (multiannual or longer) and can be estimated with a sufficient level of accuracy, they can potentially be a source of skill in a decadal prediction context. In this context, there is a growing interest from many stakeholders for climate services on 1-10 year timescales, but some efforts are still needed from the climate science community to assess the forecast quality on such timescales.

In this report, we assess the relative merits of using a multi-model large ensemble (MM ensemble, from 12 forecasting systems, 103 ensemble members) versus a single-model large ensemble (the NCAR DPLE, 40 ensemble members) in terms of forecast reliability and skill for near-surface air temperature. Forecast reliability is examined using rank histograms for the whole Northern hemisphere, and skill is assessed using deterministic (the Pearson correlation coefficient) and probabilistic (the continuous ranked probability skill score, CRPSS) scores. We further test the impact of removing the NCAR large ensemble from the multi-model ensemble, in two different ways: simply excluding the ensemble members corresponding to NCAR from the MM ensemble (63 ensemble members), and subsetting the MM ensemble to have the same ensemble size as the NCAR DPLE (40 members). This approach allows us to compare the benefits of using a multi-model ensemble in comparison to a single model with an ensemble of the same size.

Data and Methods

Near-surface air temperature data from multiple decadal hindcasts are analysed (1961-2005, yearly start dates initialised in January or November, depending on the model), for annual averages. The table below (Table 1) lists the models used, coming from the CMIP5 (Taylor et al., 2012) and SPECS (http://www.specs-fp7.eu/) projects. The analysis is done for 4 different model combinations:

- the NCAR Decadal Prediction Large Ensemble (DPLE) on its own: ***NCAR***
- the multi-model (MM) ensemble using all 12 models (including NCAR): ***MM***
- the MM ensemble using all models except the NCAR DPLE: ***MM - NCAR***
- the MM ensemble subset to have the same ensemble size as the NCAR DPLE (40 members): ***MM subset***. Subsetting is done by randomly selecting 3 ensemble members from each of the models (except for BCC-CSM1.1, which has only 1 member), plus 1 additional member from the models that have more than 3 members, except one.

Project	Centre	Forecasting system	Ensemble size
CMIP5	BCC	BCC-CSM1.1	1
CMIP5	CCCMA	CanCM4	10
CMIP5	BSC	EC-Earth	5
CMIP5	NOAA-GFDL	GFDL-CM2.1	10
CMIP5	Met Office	HadCM3 (full field)	10
CMIP5	Met Office	HadCM3 (anomaly)	10
CMIP5	MIROC	MIROC5	5
SPECS	IPSL	IPSL-CM5A-LR	3
SPECS	MPI	MPI-ESM-LR (v1)	3
SPECS	MPI	MPI-ESM-LR (v2)	3
SPECS	MPI	MPI-ESM-MR	3
DPLE	**NCAR**	**CESM1-CAM5**	**40**
Multi-model (MM)			**103**
Multi-model subset (MM subset)			**40**
Multi-model without NCAR DPLE (MM - NCAR)			**63**

Table 1. List of models used in this study. The 4 different model ensembles used are indicated in bold text.

For each of the 4 model ensembles described above, the following metrics are analysed, for forecast year 1 and for forecast years 1-5, based on a comparison with the GISS Surface Temperature Analysis (GISTEMP) dataset (GISTEMP Team, 2019; Lenssen et al., 2019):

- **Rank histograms** (Elmore, 2005) computed over the Northern Hemisphere, to assess reliability. They are generated by dividing the hindcasts (pooled for all ensemble members, start dates and grid cells in the analysed region) among a limited number of ranked bins (corresponding to the number of members + 1), thereby defining a set of exhaustive and mutually exclusive events. Then the observed frequencies for these bins are compared with the corresponding forecast probabilities. Rank histograms help to

determine whether the forecast is assumed to be reliable, and in that case it is expected to be flat. Some deviations from uniformity can appear for reliable forecasts because of randomness, however.

- Global maps of the Pearson **correlation** coefficient and the continuous ranked probability skill score - **CRPSS** (Joliffe and Stephenson, 2012), to assess skill. The Pearson correlation coefficient between the ensemble mean and the observational dataset is used as a measure of the linear correspondence between the hindcasts and the reference. This is a deterministic skill score. The CRPS measures the difference between the predicted and observed cumulative distributions, and it can be converted into a (probabilistic) skill score (the CRPSS) that measures the performance of a forecast relative to the climatology. Values below 0 are defined as unskillful, those equal to 0 are equal to the climatology forecast, and anything above 0 is an improvement upon climatology, through to 1, which indicates a “perfect” forecast.

Preliminary Results

Figures 1 and 2 present rank histograms for the Northern Hemisphere for forecast year 1 and forecast years 1-5, respectively, in the 4 model ensembles. All the hindcasts are underdispersive, as suggested by the spikes for extreme ranks (i.e., the observations fall more often, on average, below the smallest forecast value and above the largest forecast value). Apart from this underdispersion, the rank histograms are close to flatness, suggesting rather reliable hindcasts. This is less the case for the *NCAR* ensemble taken on its own, which shows a dome shape for medium ranks (for forecast year 1 and forecast years 1-5, see Figs. 1-2). However, this difference is not reflected much in the *MM - NCAR* ensemble compared to the *MM* ensemble. The *MM*, *MM - NCAR* and *MM subset* ensembles display very similar results in terms of reliability. The same results hold for forecast years 1-5, with more exacerbated characteristics (more underdispersive hindcasts, clearer dome shape in the *NCAR* rank histogram).

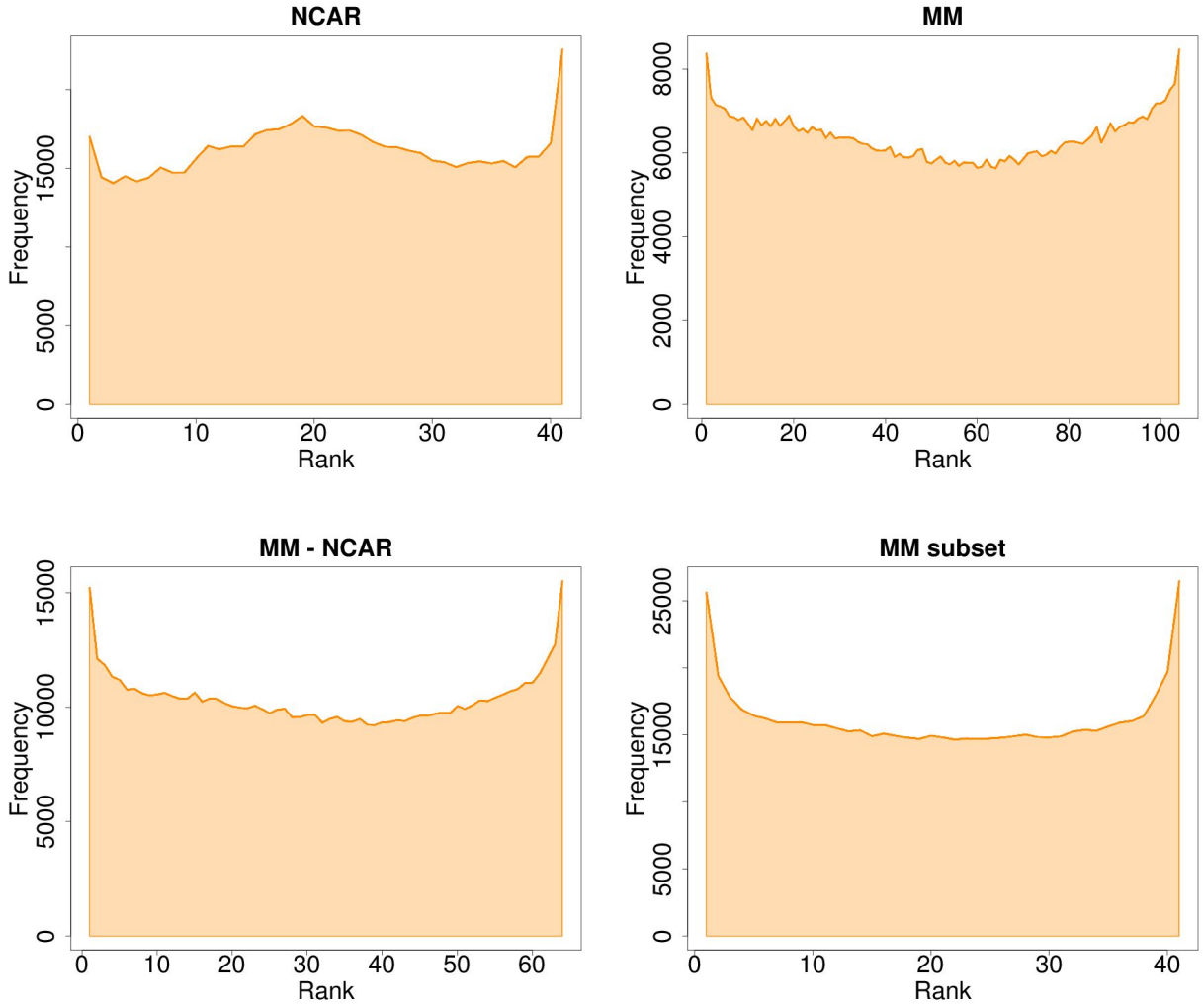

Fig. 1. Annual near-surface air temperature rank histograms for the Northern Hemisphere, for forecast year 1, in the *NCAR* (top left), *MM* (top right), *MM - NCAR* (bottom left), and *MM subset* (bottom right) decadal hindcasts. Hindcasts are verified against GISTEMP. The x axis represents the ranks. The y axis shows the frequency of each rank.

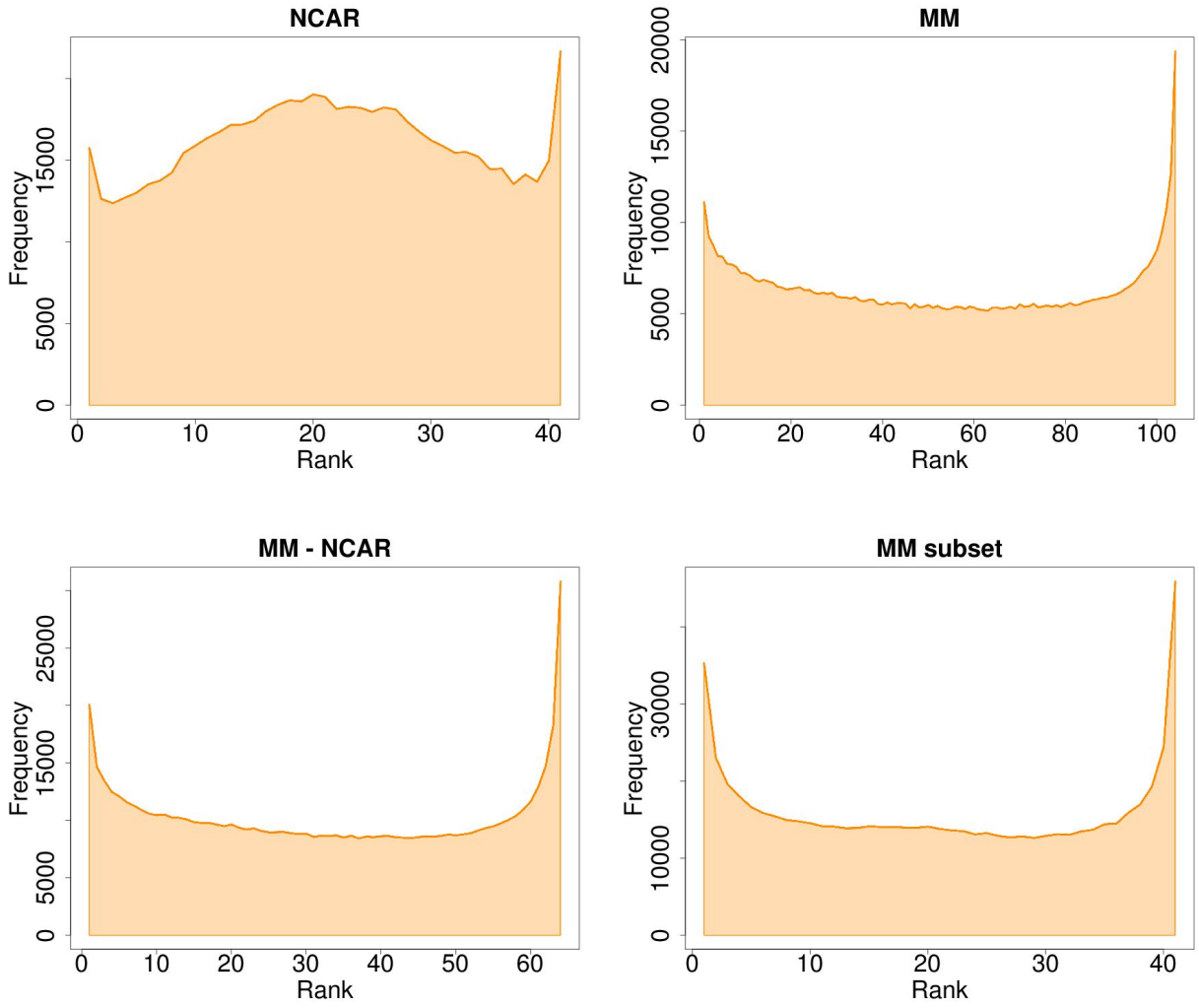

Fig. 2. Same as Fig. 1, but for forecast years 1-5.

The CRPSS of the different model ensembles is displayed as global maps in Figs 3 and 4 for forecast year 1 and forecast years 1-5, respectively. For the *NCAR* ensemble, values below and above 0 are found, depending on the region (Fig. 3). More specifically, the *NCAR* ensemble displays negative values in the ENSO region, which is not found (or not as extensive) in the other ensembles. CRPSS values for the *MM*, *MM - NCAR* and *MM subset* ensembles are very similar, with similar patterns, and generally higher than the ones for *NCAR*. Some very slight improvement of *MM - NCAR* and *MM subset* over *MM* can, however, be noted, e.g., in the tropical regions. They are positive in most regions, showing an improvement of those hindcasts compared to climatology, especially in equatorial regions and most of the North Atlantic. For forecast years 1-5, the difference between *NCAR* and the other ensembles becomes less clear, and CRPSS values are generally more contrasted (Fig. 4). The decadal hindcasts perform better than climatology over this period in most regions, with the notable exception of the North Pacific. The Southern Ocean exhibits negative CRPSS values for both forecast year 1 and

forecast years 1-5, but conclusions in this region can't be drawn with confidence because of the recurrent lack of observations.

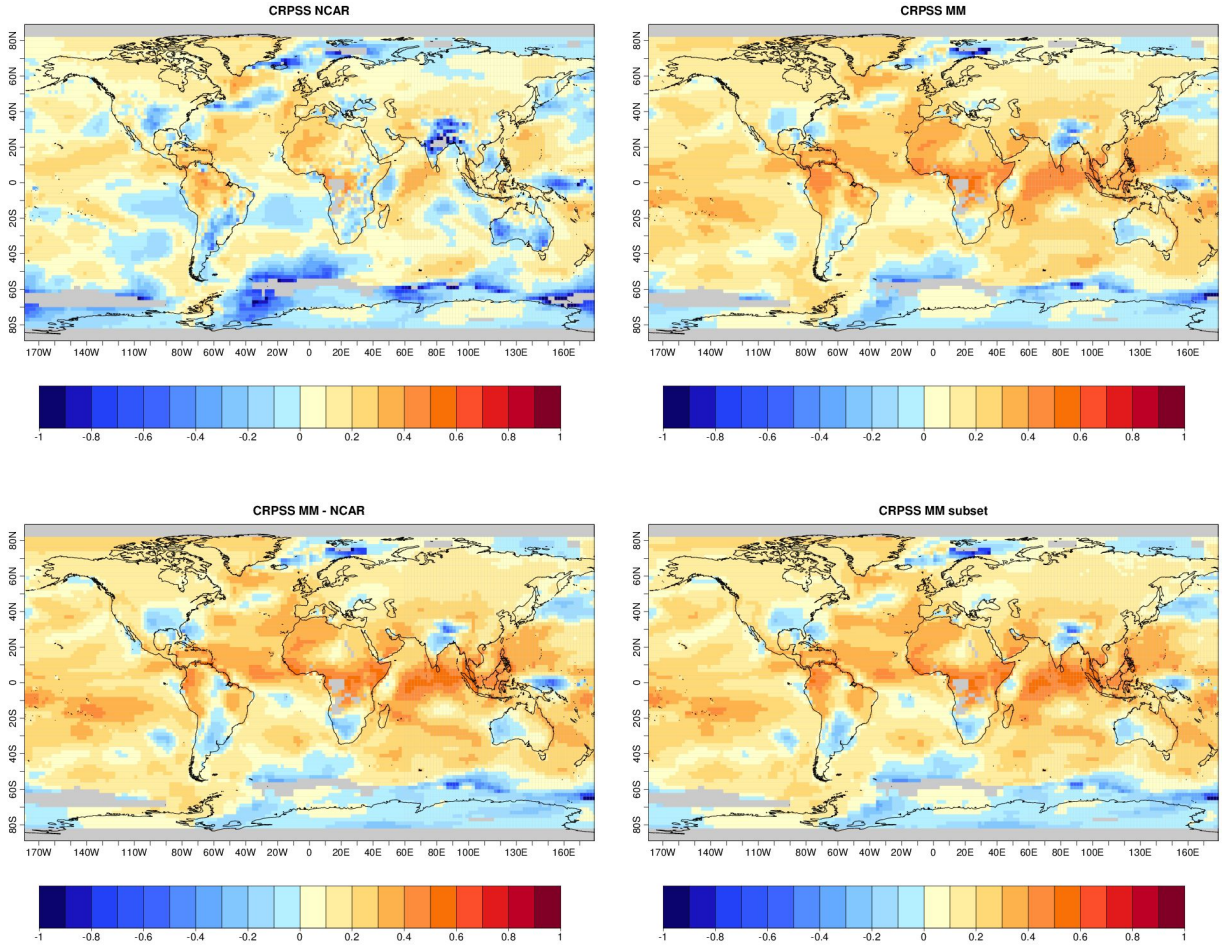

Fig. 3. Global maps of the CRPSS of the *NCAR* (top left), *MM* (top right), *MM - NCAR* (bottom left), and *MM subset* (bottom right) decadal near-surface air temperature hindcasts with the GISTEMP observations for forecast year 1. Missing values are represented in light grey.

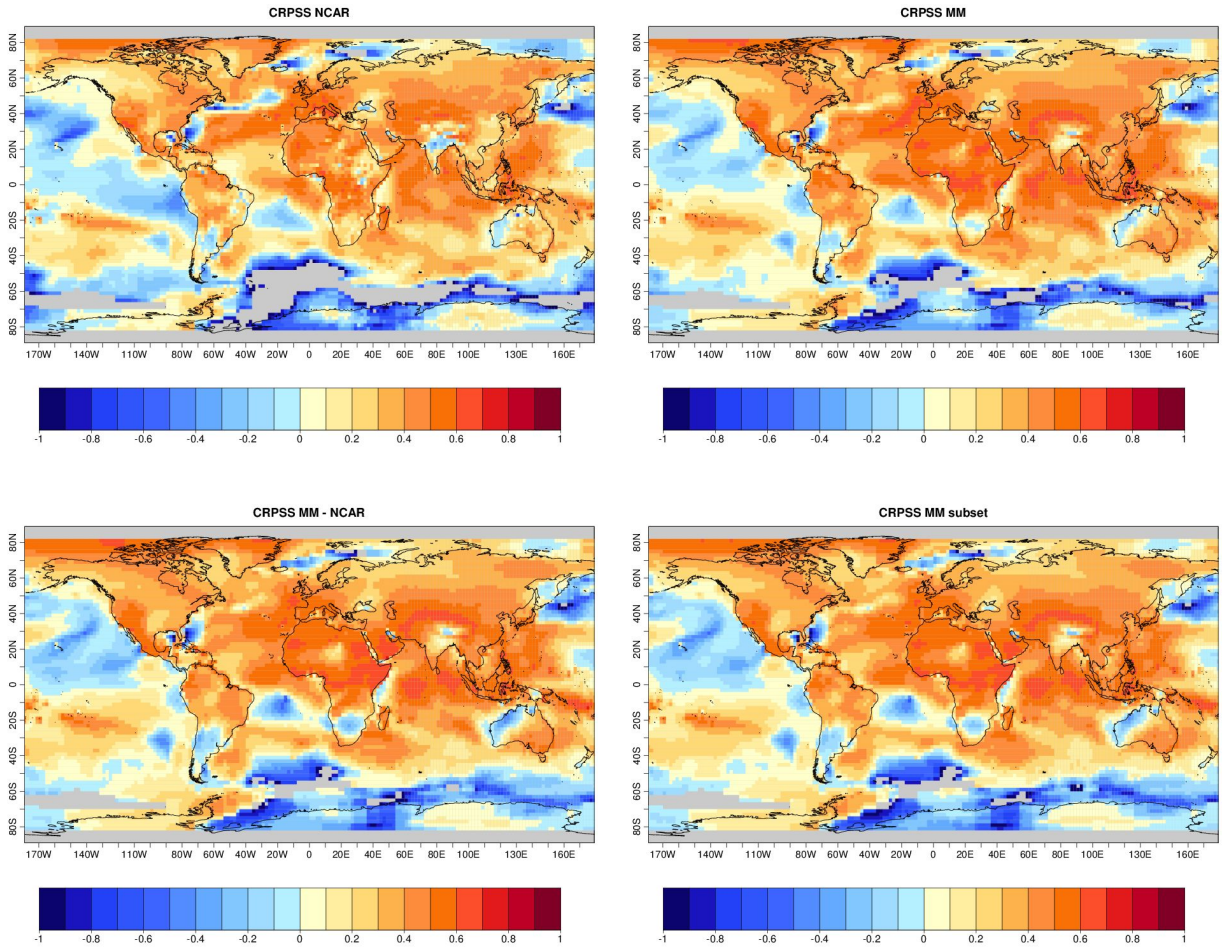

Fig. 4. Same as Fig. 3, but for forecast years 1-5.

Finally, Figs. 5 and 6 show global maps of the correlation between each model ensemble and the GISTEMP observational dataset, for forecast year 1 and forecast years 1-5, respectively. Similar findings as for the CRPSS can be highlighted. Correlation is statistically significant (at the 95% confidence level) for most regions, except for parts of the Southern Ocean (Figs. 5-6), the ENSO region for the *NCAR* ensemble (Figs. 5-6), and parts of the North Pacific for forecast years 1-5 (Fig. 6).

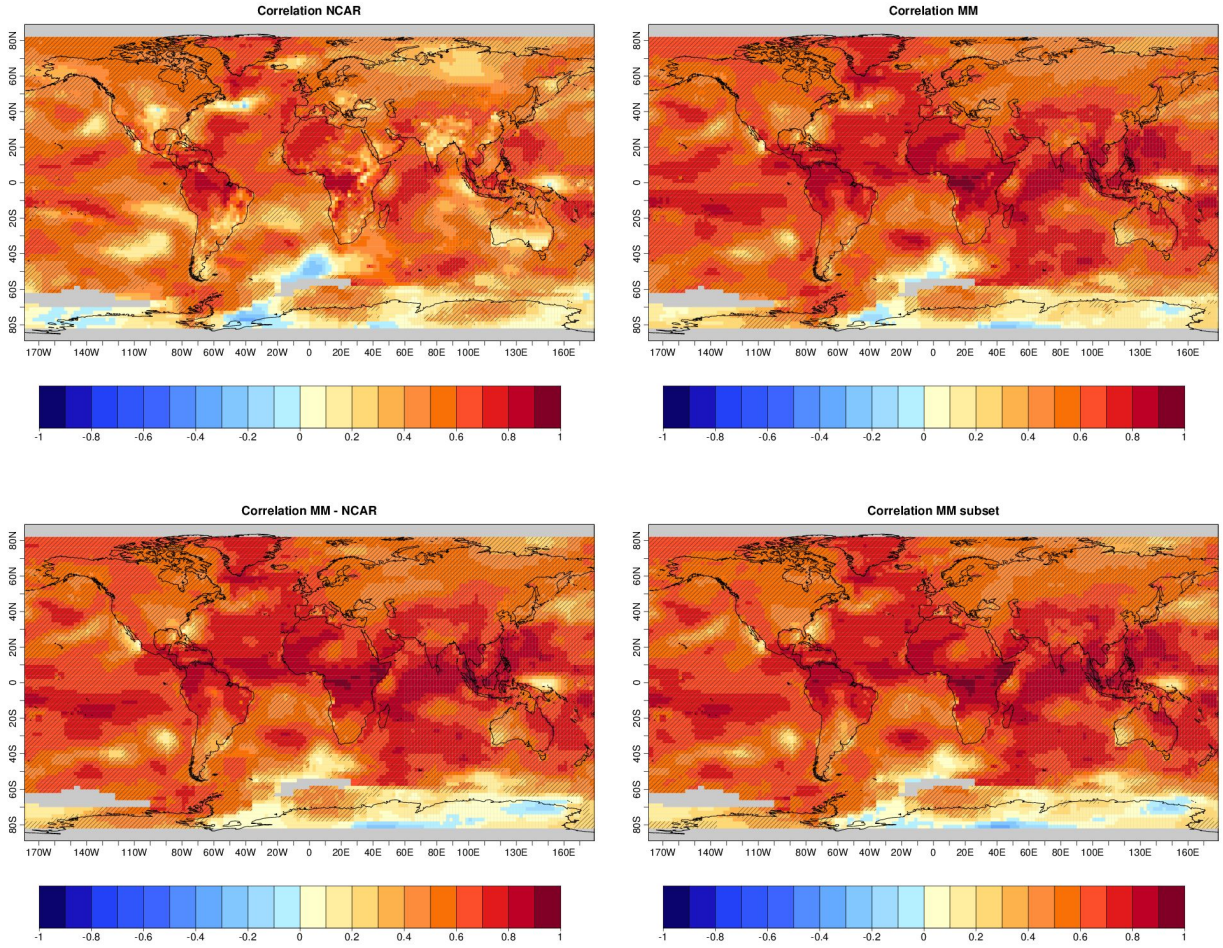

Fig. 5. Global maps of the correlation of the *NCAR* (top left), *MM* (top right), *MM - NCAR* (bottom left), and *MM subset* (bottom right) decadal near-surface air temperature hindcasts with the GISTEMP observations for forecast year 1. Hatching indicates significant correlation (95% level). Missing values are represented in light grey.

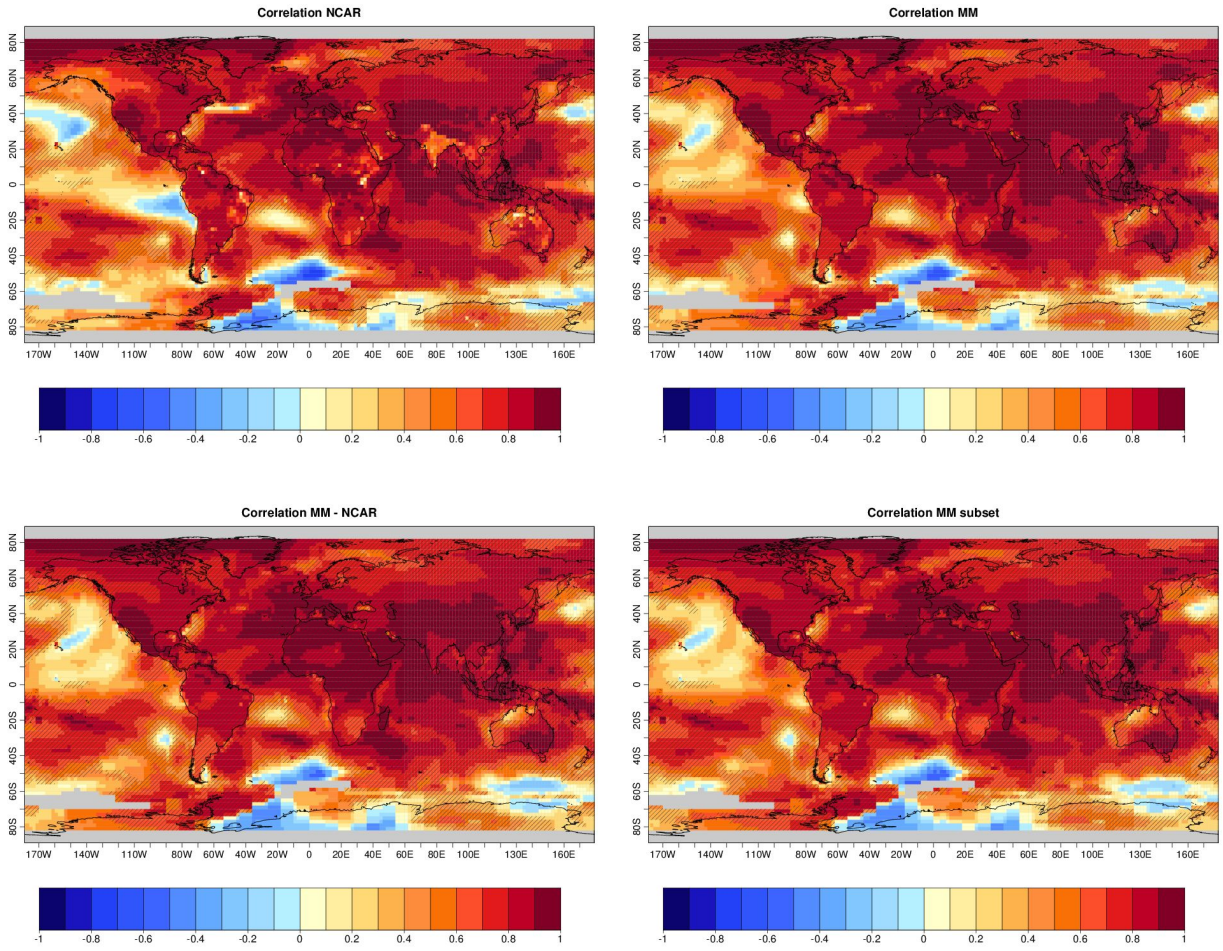

Fig. 6. Same as Fig. 5, but for forecast years 1-5.

Discussion & Conclusions

In this report, we assessed the relative merits of using a multi-model large ensemble (*MM*, 12 forecasting systems, 103 ensemble members) versus a single-model large ensemble (*NCAR*, 40 ensemble members) in terms of forecast reliability and skill. We further tested the impact of removing the *NCAR* large ensemble from the *MM* ensemble, in two different ways: simply removing the ensemble members corresponding to *NCAR* from the *MM* ensemble (*MM - NCAR*, 63 ensemble members), and subsetting the *MM* ensemble to have the same ensemble size as the *NCAR* DPLE (*MM subset*, 40 members).

In general, the multi-model large ensemble shows added-value compared to the single-model ensemble in terms of reliability (flatter rank histograms, see Figs. 1-2), but also in terms of skill (higher CRPSS and correlation values in most regions, see Figs. 3-6). In particular, the *NCAR* ensemble performs worse than climatology and displays low correlation values in the ENSO

region, which is not the case for the *MM* ensemble. Regions where both *NCAR* and *MM* display mediocre performance are the North Pacific and the Southern Ocean, even though the latter is known for its systematic lack of observations, precluding any firm conclusions on the forecast skill.

For longer time horizons (forecast years 1-5), the difference between *NCAR* and the *MM* ensemble becomes less clear. However, each experiment's characteristics are exacerbated compared to results for forecast year 1 (more under/overdispersion in Fig. 2, more spatially contrasting results for CRPSS and correlation in Figs. 4 and 6).

Despite the clear benefit of using a multi-model large ensemble compared to a single-model large ensemble outlined above, removing the *NCAR* ensemble from the *MM* ensemble (*MM - NCAR* and *MM subset*) has a rather limited impact on forecast reliability and skill, mainly visible in the tropical regions (Figs. 3-6). Furthermore, using one or the other method yields quasi identical results, both for reliability and skill.

To conclude, this study showed that, even more important than having very large ensemble sizes (> 40 ensemble members), it is crucial to use ensembles composed of different forecasting systems. Indeed, ensembles made of different models encompass a larger range of model physics and thereby also allow for model error compensation. This is a recommendation that should be passed on to stakeholders through the EUCP project, as we showed that it has important implications in terms of forecast skill, but also forecast reliability.

Acknowledgements

The EUCP project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776613. We acknowledge the use of the s2dverification (http://cran.r-project.org/web/packages/s2dverification), startR (https://cran.r-project.org/web/packages/startR), easyVerification (https://cran.r-project.org/web/packages/easyVerification) and Specs-Verification (http://cran.r-project.org/web/packages/SpecsVerification) R software packages. We also thank Nicolau Manubens, Verónica Torralba and Dragana Bojovic for their technical and scientific support.

References

- Elmore, K.L., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting*, **20**, 789–795, doi:10.1175/WAF884.1
- GISTEMP Team, 2019: GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies. Dataset accessed 2019-04-30 at https://data.giss.nasa.gov/gistemp/
- Joliffe, I.T., and D.B. Stephenson, Eds., 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 292 pp

- Keenlyside, N.S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner, 2008. Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453** (7191), 84. doi: 10.1038/nature06921
- Lenssen, N., G. Schmidt, J. Hansen, M. Menne, A. Persin, R. Ruedy, and D. Zyss, 2019: Improvements in the GISTEMP uncertainty model. *J. Geophys. Res. Atmos.*, **124**, no. 12, 6307-6326, doi:10.1029/2018JD029522
- Pohlmann, H., J.H. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009. Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *Journal of Climate*, **22** (14), 3926–3938. doi: 10.1175/2009JCLI2535.1.
- Smith, D.M., S. Cusack, A.W. Colman, C.K. Folland, G.R. Harris, and J.M. Murphy, 2007. Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317** (5839), 796–799. doi: 10.1126/science.1139540
- Taylor, K.E., R.J. Stouffer, and G.A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. *Bull. Amer. Meteor. Soc.,* **93**, 485–498, doi: 10.1175/BAMS-D-11-00094.1