How Reliable Are Decadal Climate Predictions of Near-Surface Air Temperature?

DEBORAH VERFAILLIE,^{a,c} FRANCISCO J. DOBLAS-REYES,^{a,b} MARKUS G. DONAT,^a NÚRIA PÉREZ-ZANÓN,^a BALAKRISHNAN SOLARAJU-MURALI,^a VERÓNICA TORRALBA,^a AND SIMON WILD^a

^a *Barcelona Supercomputing Center, Barcelona, Spain*

^b *Institució Catalana de Recerca i Estudis Avançats, Barcelona, Spain*

(Manuscript received 2 March 2020, in final form 16 October 2020)

ABSTRACT: Decadal climate predictions are being increasingly used by stakeholders interested in the evolution of climate over the coming decade. However, investigating the added value of those initialized decadal predictions over other sources of information typically used by stakeholders generally relies on forecast accuracy, while probabilistic aspects, although crucial to users, are often overlooked. In this study, the quality of the near-surface air temperature from initialized predictions has been assessed in terms of reliability, an essential characteristic of climate simulation ensembles, and compared to the reliability of noninitialized simulations performed with the same model ensembles. Here, reliability is defined as the capability to obtain a true estimate of the forecast uncertainty from the ensemble spread. We show the limited added value of initialization in terms of reliability, the initialized predictions being significantly more reliable than their noninitialized counterparts only for specific regions and the first forecast year. By analyzing reliability for different forecast system ensembles, we further highlight the fact that the combination of models seems to play a more important role than the ensemble size of each individual forecast system. This is due to sampling different model errors related to model physics, numerics, and initialization approaches involved in the multimodel, allowing for a certain level of error compensation. Finally, this study demonstrates that all forecast system ensembles are affected by systematic biases and dispersion errors that affect the reliability. This set of errors makes bias correction and calibration necessary to obtain reliable estimates of forecast probabilities that can be useful to stakeholders.

KEYWORDS: Climate prediction; Forecast verification/skill; Climate models; Decadal variability

1. Introduction

Trustworthy climate information over the next years to decades has become essential for stakeholders from various economic sectors and societal groups for planning and decision making on investments and climate policies (Buontempo et al. 2014; Vaughan and Dessai 2014). Until recently, the only source of near-term climate change information available to stakeholders was forced climate projections (Meehl et al. 2007; Kirtman et al. 2013). These provide a future outlook on the expected evolution of Earth's climate system, covering a continuous temporal period ranging from the beginning to the end of this century (or beyond). The evolution of a climate system represented in such noninitialized simulations is solely driven by prescribed changes in the atmospheric composition (greenhouse gases mainly) and other external forcings. However, for the

first decade, internal variability is the dominating source of uncertainty (Hawkins and Sutton 2009; Lehner et al. 2020). Lately, initialized decadal climate predictions have been made available for users as a potential source of more accurate climate information for the next decade (e.g., on the Copernicus platform). 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 forecast systems to accurately predict past climate variability in retrospective experiments called hindcasts. At the decadal time scale, the observed climate variability can be understood as the superposition of an anthropogenically driven trend on natural fluctuations. While the trend is driven by changes in anthropogenic emissions, the natural fluctuations are generated internally by the interactions of the different components of the climate system that are explicitly accounted for (atmosphere, ocean, and sea ice) or externally by other factors such as volcanic eruptions and solar activity (Meehl et al. 2007; Kirtman et al. 2013). Provided that these different sources of climate variability operate on a sufficiently long time scale (multiannual or longer) and can be estimated with a sufficient level of accuracy, they can potentially be exploited in the context of decadal predictions.

There is a growing interest from many stakeholders for climate services on 1–10-yr time scales (Soares et al. 2018; Nissan et al. 2019; Solaraju-Murali et al. 2019). However, despite considerable progress made in this area of research recently, some efforts are still needed from the climate science community to demonstrate the added value of initialized decadal prediction (INIT) compared to other sources of future information commonly used by stakeholders (mainly noninitialized projections, termed NoINIT in the following). This requires an

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-20-0138.s1>.

^c Current affiliation: Earth and Life Institute, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.

Corresponding author: Deborah Verfaillie, deborah.verfaillie@gmail.com

DOI: 10.1175/JCLI-D-20-0138.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

in-depth forecast quality assessment of the decadal forecasts. The added value of INIT over NoINIT is generally expressed in terms of forecast accuracy (e.g., [Smith et al. 2019](#)). For example, significant skill has been found for surface air temperature over large areas of the globe, and particularly over the ocean ([Kirtman et al. 2013](#)). Skill was also found for precipitation over land in some regions (e.g., the Sahel region and Canadian Arctic), as well as for sea ice, the stratosphere, aerosols, and some land surface components such as soil moisture, vegetation, snow, and permafrost ([Kirtman et al. 2013](#); [Bellucci et al. 2015](#)). However, many users from different sectors such as energy, agriculture, or insurance use probabilistic information for their decision-making processes, and are thus more interested in the forecasts being reliable than accurate (e.g., [Corti et al. 2012](#); [Torralba et al. 2017](#)). Reliability quantifies the capability to obtain a true estimate of the forecast uncertainty from the ensemble spread. This happens when the observational reference can be considered as statistically indistinguishable from any member of the ensemble. It is essential for the correct representation of the probabilities of a given event (e.g., a heatwave in the Mediterranean region; the occurrence of above-normal or, on the contrary, below-normal precipitation amounts in Europe; or the number of cyclones over the North Atlantic), which is a main concern for many stakeholders because of their need to trust the forecast probabilities as a credible representation of the probability of occurrence of the event.

Previous studies have addressed the reliability of climate forecasts, mainly for seasonal prediction (e.g., [Weisheimer and Palmer 2014](#); [Torralba et al. 2017](#); [Manzanas et al. 2019](#)) and also transient climate simulations (e.g., [Bellprat et al. 2019](#)). Reliability has been considered as well in the assessment of decadal forecast quality, an aspect that is now increasingly studied (e.g., [Corti et al. 2012](#); [van Oldenborgh et al. 2013](#); [Eade et al. 2014](#); [Stolzenberger et al. 2016](#); [Kadow et al. 2017](#); [Pasternack et al. 2018](#); [Smith et al. 2018](#); [Kushnir et al. 2019](#); [Sandgathe et al. 2020](#); [Merryfield et al. 2020](#)). However, those studies generally focused on specific regions or global averages and only few systematic comparisons with noninitialized projections were performed. Previous studies have shown that uncorrected seasonal forecasts ([Weisheimer and Palmer 2014](#); [Torralba et al. 2017](#); [Manzanas et al. 2019](#)) as well as uncorrected decadal forecasts (e.g., [Pasternack et al. 2018](#)) are generally not reliable, which implies that the spread estimated from them cannot be a trustworthy measure of the forecast uncertainty and the error. This is a crucial aspect that needs to be accounted for in decision making, and which will be discussed in this study. Moreover, as explained above, very few studies address reliability in initialized decadal predictions in comparison to uninitialized projections (e.g., [Ho et al. 2013](#); [Caron et al. 2015](#); [Camp and Caron 2017](#)), a gap that this study aims to address.

The main objective of the present study is to assess the added value of decadal predictions for near-surface air temperature in terms of reliability, compared to noninitialized projections. To this end, we performed a comprehensive global reliability assessment of multimodel decadal predictions and noninitialized projections from 12 different Earth system models. To further

investigate the impact on reliability of using different ensemble sizes and forecast system combinations, we used three different forecast system ensembles. We also explored how reliability evolves with forecast time, by looking at results for forecast year 1 and forecast years 1–5, over the period 1961–2010. Finally, we tested the impact of applying several postprocessing techniques to the “raw” temperature anomalies, thereby showing that all forecast system ensembles have issues with reliability, regardless of whether they are predictions or projections, and that bias correction and calibration is fundamental to obtain reliable predictions/projections of the future climate conditions. [Section 2](#) introduces the models and various ensembles used, the 30 different regions, as well as the reliability indicators and postprocessing methods employed. Results are presented in [section 3](#) and discussed in [section 4](#). [Section 5](#) presents the concluding remarks and some perspectives for future studies.

2. Data and methods

The reliability assessment was carried out using rank histograms ([Elmore 2005](#)), test statistics from [Jolliffe and Primo \(2008\)](#) displayed on global maps, and regional time series for 30 different regions around the world. We used three different forecast system ensembles, for forecast year 1 and forecast years 1–5, over the period 1961–2010. We also applied different postprocessing methods to the raw forecasts, described below.

a. Forecast systems and ensembles

In this study, annual average near-surface air temperature anomalies from 12 different forecast systems (or model versions) from the Coupled Model Intercomparison Project Phase 5 (CMIP5; [Taylor et al. 2012](#)) and the European Union's Seventh Framework Programme for Research (FP7) SPECS project (<http://www.specs-fp7.eu/>) were used for both INIT and NoINIT, including those models that provide initialized decadal predictions from yearly start dates during 1961–2005. The full investigation period thus spans the period from 1961 to 2010 (2010 representing forecast year 5 of the 2005 initialization). For the noninitialized runs (NoINIT), we used historical simulations up to 2005 and the RCP4.5 scenario thereafter. The decadal hindcasts (INIT) were initialized every year, either on 1 November or the following 1 January depending on the prediction system. However, to simplify the construction of large multimodel ensembles, we discarded the first two months for forecast systems initialized in November, then calculated annual anomalies (for both INIT and NoINIT) from 1 January to 31 December. To account for model drift resulting from possible initialization shocks, anomalies were calculated with respect to each model's climatology over the period 1971–2000, thus using a different (forecast-time dependent) 30-yr window for each forecast time so that climatology is always calculated for the actual years 1971–2000. [Table 1](#) lists the different forecast systems used and their characteristics. The used ensembles of INIT and NoINIT were adjusted to an equal size for each model (effective ensemble sizes in [Table 1](#), defined by random picking of the ensemble members in INIT or NoINIT)

TABLE 1. List of models used in this study, with their characteristics: project, producing center, model name (and version), INIT ensemble size, and NoINIT ensemble size. The last three rows (in bold text) represent the three different forecast system large ensembles used in this study (effective ensemble sizes are indicated in parentheses).

Project	Centre	Model (version)	INIT ensemble size	NoINIT ensemble size
CMIP5	BCC	BCC-CSM1.1	4	1
CMIP5	CCCMA	CanCM4	10	10
CMIP5	BSC	EC-Earth	5	11
CMIP5	NOAA-GFDL	GFDL-CM2.1	10	10
CMIP5	Met Office	HadCM3 (full field)	10	10
CMIP5	Met Office	HadCM3 (anomaly)	10	10
CMIP5	MIROC	MIROC5	6	5
SPECS	IPSL	IPSL-CM5A-LR	3	4
SPECS	MPI	MPI-ESM-LR (v1)	5	3
SPECS	MPI	MPI-ESM-LR (v2)	3	3
SPECS	MPI	MPI-ESM-MR	5	3
DPLE/LENS	NCAR	CESM1-CAM5	40	40
Multimodel (MM)			111 (103)	110 (103)
Multimodel (MM–NCAR)			71 (63)	70 (63)

if the available ensemble sizes were different. Three different forecast system large ensembles were tested:

- the NCAR large ensembles (DPLE for INIT, LENS for NoINIT) on their own, here called NCAR;
- the multimodel ensembles using the 12 forecast systems available, here called MM; and
- the MM ensembles using all forecast systems except the NCAR DPLE/LENS, here called MM–NCAR.

The NCAR ensemble was generated using round-off perturbation of atmospheric initial conditions (Yeager et al. 2018). The MM ensemble is a collection of different forecast systems, each with their own ensemble generation method. For example, the EC-Earth forecast system uses initial perturbations in both the atmosphere and the ocean (Du et al. 2012; Ménégoz et al. 2018; Bilbao et al. 2020).

b. Regions

Thirty different regions were used to assess reliability in this study. The 21 land regions are similar to those in Weisheimer and Palmer (2014) and Giorgi and Francisco (2000). Additionally, nine ocean regions were defined, corresponding to the main ocean basins (except the Arctic Ocean, where observations are too scarce to be used as reference in the reliability assessment). Both land and ocean grid points were used. Table S1 in the online supplemental material lists the 30 different regions with their acronyms and coordinates, shown on a global map in Fig. 1. To increase the readability of results, Australia (AUS) was chosen as an example, representative of most regions. The results will be presented in section 3 for Australia first, then discussed for all regions.

c. Reliability indicators

Two different types of indicators were calculated to assess the reliability of INIT and NoINIT ensembles against the GISS Surface Temperature Analysis (GISTEMP) dataset (GISTEMP Team 2019; Lenssen et al. 2019). Additionally, the impact of using another reference dataset, HadCRUT4 (Morice et al. 2012), is illustrated in the supplemental material. Both observational

datasets use a combination of near-surface air temperature over land and sea surface temperature over oceans. The impact of assessing forecast reliability for this combination of temperature variables instead of near-surface air temperature only is illustrated in the supplemental material. The indicators were computed and plotted using the SpecsVerification (Siebert 2017), s2dverification (Manubens et al. 2018), ClimProjDiags (BSC/CNS et al. 2020), and boot (Davison and Hinkley 1997; Canty and Ripley 2020) R packages. All forecast anomalies were interpolated to the observational grid (180° longitude $\times$ 90° latitude regular grid for GISTEMP) before calculating the indicators.

1) RANK HISTOGRAMS

In this study, we used the rank histograms, a metric for testing the forecast reliability (Elmore 2005) in a simple way. They are used to assess if the ensemble members and the verifying observation stem from the same probability distribution (i.e., if the observations are predicted as the equiprobable members). To construct the rank histogram an M -member ensemble forecast $[y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,M})]$ and the corresponding observation x_i were considered, for each of the 30 different regions defined in section 2b, and for forecast year 1 and the average of forecast years 1–5. The rank histogram was constructed for N forecast–observation pairs ($i = 1, \dots, N$), depending on the number of grid cells inside each region, the number of start dates (45 years in this study), and the ensemble size. The $M + 1$ possible ranks (bins) were defined by the forecast range. When the observation x_i was smaller than all the ensemble members, it was assigned to the first rank. If it exceeded all the ensemble members then its rank was the $M + 1$. If the ensemble prediction was reliable, then the ensemble members and observations were statistically indistinguishable from each other, and it would be equally probable for the observation to fall in any of the ranks [i.e., the number of counts in each rank would be $N/(M + 1)$]. In this case the rank histogram would be flat (as if observations and the ensemble members were stemming from a uniform distribution). However, because of the limited sample sizes and forecast deficiencies, the

FIG. 2. Near-surface air temperature rank histograms for (top) Greenland (GRL) and (bottom) the North Atlantic Ocean (NAT), for the INIT and NoINIT uncorrected simulations, for forecast year 1, in the MM ensemble using (left) all models available and (right) the NCAR ensemble. Forecasts are verified against GISTEMP. The x axis represents the ranks. The y axis shows the frequency of each rank.

rank histograms for INIT and NoINIT ensembles using the R boot() function (Canty and Ripley 2020), in its nonparametric form, with replacement and 1000 replicates. Distributions of the JP slope and JP convexity coefficients (1000 values for each region) were then obtained, and the distribution of the differences between INIT and NoINIT was derived. Finally, the significance of the differences at the 95% level was obtained by calculating the 2.5th and 97.5th percentiles of the distribution

of differences between INIT and NoINIT. Whenever zero lies outside of this percentile range, the null hypothesis (zero difference) can be rejected and the difference is considered significant at the 95% level.

3) ADDITIONAL ERROR ESTIMATES

Additional measures to address errors in the uncorrected forecasts compared to observations are available in the

Shown in Figs. 4-7 and 9 (+ Suppl. Figs. S1-S6, S8-S27 and S30-S32)

FIG. 3. Schematic of the Jolliffe and Primo (2008) test statistics and the meaning of each component in this study. The χ^2 p value, as well as the contribution of the JP slope and JP convexity components, is shown in Figs. 4-7 and 9 as well as in Figs. S1-S6, S8-S27, and S30-S32.

supplemental material. Those include the difference in trend and the ratio of interannual variability (measured by the variance) between INIT and observations. All error estimates were computed over the full period from 1961 to 2010, for forecast year 1 and the average of forecast years 1–5, and for all grid cells. Trends were calculated for the ensemble means, while variances were computed for each ensemble member separately, then averaged over all members of the ensemble. For errors in interannual variability, time series were detrended first in order to remove the trend component of the error.

d. Postprocessing methods

To check what were the sources for the presence or lack of reliability (incorrect representation of the trend, systematic errors in the variability, and/or lack of ensemble calibration), we tested the impact of applying several postprocessing techniques to the raw uncorrected temperature anomalies from each forecast system (before building the ensembles). Detrending, mean and variance bias correction, and calibration were performed using the *s2dverification* (Manubens et al. 2018) and *CSTools* (Perez-Zanon et al. 2019) R packages.

1) DETRENDING

Detrending of the raw temperature anomalies from INIT and NoINIT was performed by computing the trend along the start dates for the ensemble mean by least squares fitting, and linearly detrending the time series. This detrending has also been applied to the GISTEMP observational reference. If the hindcasts had an incorrect representation of the trend compared to observations, detrending should improve the JP slope coefficient and flatten the corresponding rank histogram because the trend in both the hindcasts and the observations would be removed.

2) MEAN AND VARIANCE BIAS CORRECTION

A bias correction technique was applied to the raw temperature anomalies of each individual ensemble member, using cross-validation. The methodology is described in Leung et al. (1999) and Torralba et al. (2017). Briefly, the method relies on the approximation that the observational (GISTEMP) and predicted distributions are Gaussian. The correction then creates hindcasts with means and standard deviations corresponding to those in the verification dataset. If the hindcasts were affected by a systematic mean (or variance) bias, this bias correction approach should also improve the JP slope coefficient by flattening the corresponding rank histogram.

3) CALIBRATION

The third postprocessing method applied to the raw temperature anomalies is a member-by-member calibration based on variance inflation (von Storch and Zwiers 2001; Doblas-Reyes et al. 2005). This method, used here in cross-validation, produces calibrated hindcasts with interannual variance equivalent to that of the GISTEMP dataset in a similar way to the bias correction method, but at the same time ensuring increased reliability of the probability forecasts. The calibration technique adjusts the mean and the standard deviation, so

there is a common correction in both the bias correction and the calibration. However, it also corrects the underestimation or overestimation of the ensemble spread with no modification of the ensemble-mean correlation. Hence, the calibration adjusts not only the interannual variability, but also the variability of the ensemble spread, which is fundamental to improve the forecast reliability. If the hindcasts had an incorrect spread (e.g., if they were affected by over/underdispersion), the calibration should improve the JP convexity coefficient and the corresponding rank histogram.

3. Results

a. Assessing reliability of uncorrected simulations

1) FORECAST YEAR 1

Figures S1 and S2 and Figs. 4–7 show the JP test statistics (χ^2 , p value, JP slope, and JP convexity) for INIT and NoINIT, and their differences, for forecast year 1, for the MM and NCAR ensembles, respectively. Another figure for the MM–NCAR ensemble can be found in the supplemental material (Fig. S3)

FIG. 4. Maps of the Jolliffe and Primo (2008) (top) slope and (bottom) convexity coefficients, expressed as their contribution to the χ^2 coefficient (%), for near-surface air temperature for the 30 different regions defined in Fig. 1, for forecast year 1 in the INIT MM ensemble. Going from light yellow to dark purple, the colors denote an increasing role of the slope and the convexity terms to decreasing the reliability of the ensemble (diagnosed by the deviations from flatness in the rank histogram). A plus (minus) sign in the convexity coefficient maps represents an underdispersive (overdispersive) forecast. Hatching represents regions where the p value is larger than 0.05, thus where there is no evidence of bias, difference in trend, or error in dispersion (the null hypothesis being that the rank histograms are flat).

FIG. 5. Maps of the difference between INIT and NoINIT Jolliffe and Primo (2008) (top) slope and (bottom) convexity coefficients, expressed as the difference in their contribution to the χ^2 coefficient (%), for near-surface air temperature for the 30 different regions defined in Fig. 1, for forecast year 1 in the MM ensemble. Going from dark blue to dark red, the contribution of JP slope or JP convexity for INIT becomes increasingly larger than for NoINIT (i.e., INIT becomes less reliable than NoINIT). Hatching represents regions where the difference is not significant at the 95% level.

and is very similar to the figures for the MM ensemble. The $\chi^2 p$ values above 0.05 in Figs. S1 and S2 (hatched orange-yellow boxes) indicate regions that are significantly reliable (i.e., without any significant deviation from flatness). The JP slope and JP convexity coefficients are expressed as their contribution to the χ^2 coefficient, a large contribution in Figs. 4 and 6 (purple colors), indicating deviation from flatness mainly due to the slope or the convexity, respectively. For the difference in the contribution to the χ^2 coefficient (Figs. 5 and 7), blue colors indicate a lower contribution of JP slope or JP convexity for INIT than for NoINIT, and thus some added value of INIT over NoINIT. Hatched boxes indicate regions in which the difference is significant at the 95% level.

(i) Results for Australia

For our example region Australia (AUS), neither uncorrected INIT nor uncorrected NoINIT provides significantly reliable estimates, that is, flat rank histograms (the $\chi^2 p$ value is not above 0.05), for near-surface temperature and forecast year 1 (Fig. S1). In the case of Australia, this is because both the slope and the convexity parameters are significantly contributing to the χ^2 coefficient, resulting in unreliable estimates (Fig. 4). INIT shows some significant added value over NoINIT

FIG. 6. As in Fig. 4, but for the NCAR ensemble.

mainly in terms of the convexity coefficient in the MM ensemble (Fig. 5), and mainly in terms of the slope coefficient in the NCAR ensemble (Fig. 7).

(ii) Results for all regions

When analyzing results for all regions, it is clear that no region except one provides significantly reliable estimates (the $\chi^2 p$ value is never above 0.05) for near-surface temperature and forecast year 1 (Fig. S1). For most regions, this is because the slope parameter or the convexity parameter or both are significantly contributing to the χ^2 coefficient, resulting in unreliable estimates (Fig. 4). A notable exception is East Africa (EAF) in the MM (Fig. 4) and MM-NCAR (Fig. S3) INIT datasets. In few cases [e.g., the North Atlantic Ocean (NAT) in INIT in Fig. S1 and Fig. 4], both parameters have a p value exceeding the 0.05 value, but the $\chi^2 p$ value does not. This indicates that the ensemble is not significantly reliable, due to the residual (not shown).

In general, the reliability measured by the JP parameters varies greatly depending on the analyzed regions and forecast system ensembles. For example, the Southern Ocean (SOO) in INIT in Fig. 4 displays a much higher value for the contribution to χ^2 of the JP slope coefficient than the South Pacific Ocean (SPO). However, for the same region (SOO) and the same parameter (JP slope in INIT) but for a different ensemble (Fig. 6), the contribution is much lower. Concerning the added value of INIT over NoINIT, the behavior highlighted for Australia (some significant added value of INIT over NoINIT mainly in terms of the convexity coefficient in the MM ensemble, and mainly in terms of the slope coefficient in the NCAR ensemble) seems to be valid for the broad picture, with some discrepancies depending on the regions.

FIG. 7. As in Fig. 5, but for the NCAR ensemble.

Three regions were selected from Figs. S1 and S2 and Figs. 4 and 5 based on their different reliability characteristics. The corresponding regional rank histograms and regional average time series for the MM and NCAR ensembles are shown in Figs. 2 and 8. Figure 2 (bottom panels) is for the North Atlantic Ocean, a region where INIT seems to be more reliable than NoINIT, which is underlined by the slope coefficient (see Fig. 5). Greenland (GRL) displays rather large contributions to χ^2 of the slope coefficient in both INIT and NoINIT and is shown in Fig. 2 (top panels). Finally, an example of overdispersive ensembles (both INIT and NoINIT) is shown for southern Africa (SAF) in Fig. 8.

As expected, rank histograms reflect the behavior highlighted in the maps of the JP parameters, including that the rank histogram for MM INIT is flatter than the one for MM NoINIT in the NAT (Fig. 2), both MM rank histograms for GRL show a slope (Fig. 2), and both MM rank histograms have a dome shape for temperature over southern Africa (Fig. 8).

The NCAR ensemble, however, does not have the same characteristics as the MM ensemble. For example, both NCAR INIT and NoINIT ensembles are rather overdispersive in the GRL region, whereas the MM ensemble shows a strong slope and is rather underdispersive (Figs. 4–7 and 2). Similarly, MM INIT and NoINIT generally exhibit the same behavior: for example, when INIT is overdispersive (underdispersive), NoINIT is overdispersive (underdispersive) too, but this is not always the case for NCAR INIT and NoINIT (see, e.g., Fig. 8).

The time series in Fig. 8 help to understand the shape of the corresponding rank histograms, even though a larger or smaller spread is not a measure of reliability in itself. For the MM ensemble, INIT and NoINIT have very similar rank

histograms indicating overdispersion, but the spread of the INIT time series seems narrower than the spread of the NoINIT time series (although there are very few cases for which the observations fall outside of the INIT forecast ensemble). The rank histograms for the NCAR ensemble point to underdispersive INIT and overdispersive NoINIT, which is reflected in the spread of the corresponding time series. It should be borne in mind that while the time series correspond to spatially averaged values, the rank histograms were built using values for all the individual grid points in the region.

2) FORECAST YEARS 1–5

The JP test statistics for an average of forecast years 1–5 are displayed in Fig. 9 and in Figs. S4–S6, S8, and S9 for the MM, MM–NCAR, and NCAR ensembles.

(i) Results for Australia

For the Australian region (AUS), the $\chi^2 p$ values for forecast years 1–5 remain below 0.05 for all ensembles, indicating unreliable forecasts (Figs. S4–S6). In fact, the contribution of the JP slope parameter to χ^2 is slightly higher, whereas the contribution of the JP convexity parameter is much lower for forecast years 1–5 than for forecast year 1 in the MM ensemble (Fig. 9 and Fig. S2). The difference between INIT and NoINIT is also slightly smaller (for both parameters) for forecast years 1–5 than for forecast year 1. Another specificity of forecast years 1–5 is the change in the dispersion characteristics of the MM ensemble for Australia, from an overdispersive ensemble for forecast year 1 (Fig. 4) to an underdispersive ensemble for forecast years 1–5 (Fig. S4).

(ii) Results for all regions

Concerning results elsewhere, again no region displays reliable forecasts for any of the ensembles (Figs. S4–S6). The contributions of the JP slope and convexity parameters to χ^2 are generally smaller for forecast years 1–5 than for forecast year 1, in particular in the MM ensemble (Fig. 9 and Fig. S4). The difference between INIT and NoINIT is also often smaller for forecast years 1–5 than for forecast year 1. A puzzling feature of the MM ensemble is the fact that for forecast years 1–5, the majority of regions shows a slightly higher contribution to χ^2 of the JP convexity parameter in INIT compared to NoINIT, whereas for forecast year 1 a clear added value of INIT over NoINIT was found for most regions for this parameter. This feature is less pronounced in the NCAR ensemble. While the number of regions showing overdispersive INIT and NoINIT was about the same as the ones showing underdispersive results for forecast year 1 of the MM ensemble, for forecast years 1–5 almost all the regions (25 out of 30 in the MM ensemble) display underdispersive results (Fig. S4), as seen for Australia.

Figure S7 shows the corresponding rank histograms for the North Atlantic Ocean (NAT) region for forecast years 1–5. As shown in the JP statistics (Fig. 9, Fig. S4–S6, S8, and S9) the added value of INIT over NoINIT is smaller than for forecast year 1. Moreover, both INIT and NoINIT rank histograms for forecast years 1–5 in this region show some deviations from flatness, through the presence of a slope

FIG. 8. (top) Near-surface air temperature rank histograms for southern Africa (SAF), for the INIT and NoINIT uncorrected simulations, for forecast year 1, in the MM ensemble using (left) all models available and (right) the NCAR ensemble. Forecasts are verified against GISTEMP. The x axis represents the ranks. The y axis shows the frequency of each rank. Also shown are corresponding regional average time series of SAF near-surface air temperature anomalies for forecast year 1, in the observation and for every (middle) INIT and (bottom) NoINIT model member separately, for the MM ensemble and NCAR ensemble. The ensemble mean for each forecast system is indicated with a bold line.

and a concave rank histogram (extreme ranks, in this case especially the largest ones, are more populated than the middle ranks).

b. Assessing reliability of the detrended time series

The JP statistics for the detrended MM ensemble are shown in Fig. 9 (in the top triangle of the matrix fields). The

corresponding figures for other ensembles can be found in the supplemental material.

1) RESULTS FOR AUSTRALIA

Detrending the data does not seem to yield reliable forecasts for any of the ensembles in our example region (Australia) for forecast year 1 (Figs. S10–S12). For the MM ensemble in

FIG. 9. Summary of the Jolliffe and Primo (2008) slope and convexity coefficients, expressed as their contribution to the χ^2 coefficient (%), for near-surface air temperature for the 30 different regions defined in Fig. 1, for forecast year 1 and forecast years 1–5 in the MM ensemble. For each forecast time, the top two rows represent the JP coefficients for INIT. Diamonds indicate cases where the p value is larger than 0.05, thus where there is no evidence of bias, difference in trend, or error in dispersion (the null hypothesis being that the rank histograms are flat). The bottom two rows for each forecast time represent the difference between INIT and NoINIT. A diamond indicates a nonsignificant difference at the 95% level. For color codes, please refer to Figs. 4 and 5. Each triangle displays the result for a type of postprocessing (either raw uncorrected values, det = detrended, b-c = bias-corrected, or cal = calibrated).

particular, detrending improves greatly the JP slope coefficient in this region, with a significantly low contribution to χ^2 , and differences between INIT and NoINIT very close to zero (Fig. 9). It however degrades the JP convexity coefficient. For forecast years 1–5 (Fig. 9 and Figs. S8, S9, and S13–S15), detrending also significantly improves the JP slope coefficient in Australia. On the other hand, compared to forecast year 1, the contribution of the convexity coefficient for forecast years 1–5 is slightly higher. After detrending, the difference between INIT and NoINIT for forecast years 1–5 is slightly more positive (i.e., no added value of INIT over NoINIT) than for forecast year 1 in Australia.

2) RESULTS FOR ALL REGIONS

There is no region that becomes significantly reliable (i.e., with a χ^2 p value above 0.05, the null hypothesis being that the rank histograms are flat) after detrending the data (Figs. S10–S15), which suggests that the lack of reliability cannot be only due to a misrepresentation of the observed trend. Similar to the results for Australia, we can see that detrending the data improves greatly the JP slope coefficient for forecast year 1, with low contributions to χ^2 in many regions, and differences between INIT and NoINIT very close to zero. On the other hand, it degrades the JP convexity coefficient in many regions. In a few regions, especially Southeast Asia (SEA), detrending increases the added value of INIT over NoINIT in the MM ensemble. However, despite both the contributions to χ^2 of the JP slope coefficient and the JP convexity coefficient being significantly low for INIT in this region, the χ^2 p value is not above 0.05, indicating that the residual parameter (not shown) is preventing the ensemble from being significantly reliable. The location of these regions with added value of INIT over NoINIT is ensemble dependent (see Fig. 9 and Figs. S8–S12).

For forecast years 1–5 (Fig. 9 and Figs. S8, S9, and S13–S15), detrending also substantially improves the JP slope coefficients, with significantly low contribution to χ^2 in many regions. On the other hand, compared to forecast year 1, the contribution of convexity coefficients for forecast years 1–5 is generally slightly higher. After detrending, differences between INIT and NoINIT for forecast years 1–5 are slightly more positive (i.e., no added value of INIT over NoINIT) than for forecast year 1 for most regions (21 out of 30 in the MM ensemble).

c. Assessing reliability of bias-corrected simulations

Figure 9 also shows the JP statistics for the bias-corrected MM ensemble (bottom triangle of the matrix fields).

1) RESULTS FOR AUSTRALIA

For Australia, bias correction does not make the forecasts significantly reliable (Figs. S16–S21). However, it also greatly improves the JP slope coefficient (Fig. 9). Like detrending, it removes almost all the differences in terms of JP slope coefficient between INIT and NoINIT. However, unlike detrending, for forecast year 1 it does not increase the contribution of the JP convexity coefficient. Bias-corrected results for forecast years 1–5 in Australia are slightly worse than the uncorrected results for the JP convexity parameter (Fig. 9 and Figs. S8, S9, and S19–S21). For the JP slope parameter, the contribution is very low for forecast years 1–5 after bias correction.

2) RESULTS FOR ALL REGIONS

Similarly to detrending, bias correction does not produce any significantly reliable forecast for any region (i.e., with a χ^2 p value above 0.05) when the null hypothesis is that the rank histograms are flat (Figs. S16–S21). This indicates that errors in

the mean variance do not play a significant role in the lack of reliability of the ensembles. As for Australia, the bias correction also greatly improves the JP slope coefficient, increasing reliability for many regions (22 out of 30 for forecast year 1 and 17 out of 30 for forecast years 1–5 in the MM ensemble). Like detrending (Fig. 9), it removes almost all the differences in terms of JP slope coefficient between INIT and NoINIT. However, unlike detrending, for forecast year 1 the contribution of the JP convexity coefficient decreases for some regions and increases for others after bias correction (comparing the bottom triangle and the left triangle of the matrix fields in Fig. 9). Especially in the case of MM NoINIT, for which the contribution of the JP convexity coefficient was generally larger than for MM INIT (17 blue left triangles in Fig. 9), bias correcting improves this coefficient (for 16 out of the 17 previous cases). This implies that the difference between INIT and NoINIT convexity coefficients after bias correction is close to zero. However, this is ensemble dependent, as the results for the NCAR ensemble are quite different (Fig. S9). In the case of NCAR, the JP convexity coefficients for NoINIT were improved more than the ones for INIT, resulting in larger (often positive) differences. Bias-corrected results for forecast years 1–5 are slightly worse (for 20 out of the 30 regions in the MM ensemble) than the uncorrected results for the JP convexity parameter (Fig. 9 and Figs. S8, S9, and S19–S21). For the JP slope parameter, the contributions remain very low for forecast years 1–5 after bias correction. As for detrending, the bias-corrected differences between INIT and NoINIT JP convexity for forecast years 1–5 become slightly more positive (meaning that NoINIT performs better than INIT) and are generally significant for most regions. This is less the case for the NCAR ensemble (Fig. S9) than for the MM or MM–NCAR ensembles (Fig. 9 and Fig. S8).

d. Assessing the reliability of calibrated simulations

The last postprocessing method we tested was calibration. The results of the calibration applied to the MM ensemble are shown in Fig. 9 (right triangle of the matrix fields).

1) RESULTS FOR AUSTRALIA

Calibration does not yield reliable forecasts for any of the ensembles in Australia (Figs. S22–S27). It improves greatly the JP slope coefficient, which becomes significantly low for forecast year 1 in Australia (Fig. 9). For the NCAR ensemble for forecast year 1 (Fig. S9), the improvement in the JP convexity coefficient due to calibration for INIT is larger than for the MM ensemble.

2) RESULTS FOR ALL REGIONS

From a more global perspective, calibrating the various forecast system ensembles yields significantly reliable results in one region, Central America (CAM), for INIT in the MM and MM–NCAR ensembles for forecast year 1 (Figs. S22 and S23). Calibration is the only postprocessing method that leads to significantly reliable ensembles in this region for forecast year 1, indicating that errors in the ensemble spread played a significant role in the lack of reliability of the uncorrected forecasts for this region. As for detrending and bias correction,

calibration greatly improves the JP slope coefficient, which becomes significantly reliable in many regions (17 out of 30 for forecast year 1 and 7 out of 30 for forecast years 1–5 in Fig. 9). Additionally, for forecast year 1 it also generally improves the JP convexity coefficient, even though not in all regions. For example, the North Atlantic Ocean (NAT) and Mediterranean basin (MED) regions, which already had low contributions to χ^2 of the JP convexity coefficients in the uncorrected MM INIT dataset, display higher values for the contribution of convexity after calibration (thus yielding unreliable forecasts). For the NCAR ensemble (Fig. S9), as in the Australian region, the improvement in the JP convexity coefficient for INIT is generally larger than for the MM ensemble. Results for the MM–NCAR ensemble (Fig. S8) are consequently slightly less reliable than for the MM ensemble.

For forecast years 1–5, the contribution of the convexity coefficient to the χ^2 statistic after calibration generally increases compared to the uncorrected ensembles, making the ensemble less reliable, while the slope results remain unchanged (Fig. 9 and Figs. S8, S9, and S25–S27). Calibration, however, does not yield any significantly reliable forecast in any region for forecast years 1–5. For the MM and MM–NCAR ensembles, the differences between INIT and NoINIT JP convexity after calibration are slightly more positive for forecast years 1–5 than for forecast year 1 even though the values remain very close to zero (Fig. 9 and Fig. S8). For the NCAR ensemble, the differences can have opposite signs between forecast year 1 and forecast years 1–5 depending on the region (Fig. S9).

4. Discussion

a. All forecast ensembles are affected by errors

As explained in section 3a, neither uncorrected INIT nor uncorrected NoINIT provides significantly reliable estimates for near-surface temperature and forecast year 1 or forecast years 1–5, except in one region. In the example case of Australia (AUS), the contributions of both the slope and the convexity parameter to χ^2 are high (e.g., in the MM ensemble for forecast year 1 in Fig. 4). The high slope contribution indicates that uncorrected ensembles are either biased (in their mean and variance) or display an inconsistent trend compared to observations. Moreover, the high convexity contribution points to over/underdispersive ensembles. In other words, all forecast system ensembles suffer from errors, whether they are initialized or not.

Relating these errors in the forecasts to the (in)correct representation of specific physical processes would require an in-depth analysis of each of the forecast systems used in the ensembles, which is beyond the scope of this study. However, Figs. S28 and S29 provide some insights into why the forecasts might be unreliable by looking at the errors in trend and interannual variability in the MM and NCAR ensembles for forecast year 1 and forecast years 1–5.

In the case of Australia in the MM ensemble, the high slope contribution can be related to a slight underestimation of the trend in INIT compared to observations, even though the error

is not significant at the 95% level over the whole region (Figs. S28 and S29). On the other hand, the high convexity contribution seems due to an incorrect representation of interannual variability in INIT compared to observations (significant error). INIT displays too large interannual variability compared to observations, which matches its overdispersive character outlined in Fig. 4 (taking into account the fact that systems that overestimate the interannual variability generally also have an excessive spread). One should bear in mind that there could also be some error compensation between different forecast systems included in the MM ensemble.

For the North Pacific Ocean (NPO) region, as for Australia, both the slope and the convexity contributions to χ^2 are high in the MM and NCAR ensembles for forecast year 1 (Figs. 4 and 6), which seems to be due to partly significant (in the case of the MM ensemble) overestimation of the trend and nonsignificant overestimation of the interannual variability in INIT compared to observations (Fig. S28).

The noticeable exception is East Africa (EAF), which displays significantly reliable MM and MM–NCAR INIT ensembles for forecast year 1 (Figs. S1 and S3). In this region, both the slope and the convexity contributions to χ^2 are very low (even though not significantly at the 95% level), indicating very little evidence of bias, difference in trend compared to observations or error in dispersion (Fig. 4 and Fig. S3). This is confirmed to a certain extent by Fig. S28, with small errors in the trend, albeit rather large variance errors highlighted in this region.

Some other regions exhibit a high contribution to χ^2 of one of the parameters, but not the other. This is the case for example in southern Africa (SAF) in the MM INIT ensemble for forecast year 1. This region shows a low contribution of the slope parameter, but a high contribution of the convexity parameter, pointing to errors in dispersion (Figs. 4 and 8). Indeed, the error in trend of MM INIT compared to observations is low and nonsignificant in Fig. S28, but the error in interannual variability is very high. Similar to Australia, southern Africa exhibits forecasts with a too large interannual variability compared to observations, which is coherent with the overdispersive nature of MM INIT in this region.

For the North Atlantic Ocean (NAT) region, both parameters do not contribute to the MM and MM–NCAR ensembles being unreliable because they have a low contribution to χ^2 (Fig. 4 and Fig. S3), reflected in their small errors in trend and variability (Figs. S28 and S29). However, the contribution of the residual parameter (not shown) is high; that is, there are other factors than an incorrect trend or mean biases or over/underdispersion that explain that the rank histograms are not flat (e.g., a combination of different distribution characteristics such as skewness and kurtosis, or a bimodal distribution). Another possibility could be that in those regions, results from points with very different characteristics were gathered. This is confirmed by Fig. S28, which shows grid cells with very different interannual variability errors for the NAT region. However, having smaller regions would reduce the sample and increase the noise. This hypothesis was further confirmed by splitting the NAT region into two subregions, which displayed different rank histograms and thus different contributions to

the forecasts not being reliable (not shown). The low contribution to χ^2 of the slope and convexity parameters in the NAT region for forecast year 1 in the MM ensemble could point to an adequate representation of the Atlantic multidecadal variability (AMV) for most forecast systems, even though the residual parameter has a high contribution to χ^2 .

Using reliability diagrams, rank histograms, and the Brier decomposition, previous studies have shown that uncorrected seasonal forecasts of temperature, precipitation, or wind speed are generally not reliable in most regions of the world (e.g., Europe, Canada, or northern Asia), even for the first forecast season (Weisheimer and Palmer 2014; Torralba et al. 2017; Manzananas et al. 2019). Pasternack et al. (2018) also gave the same conclusions for uncorrected decadal forecasts of surface temperature in the North Atlantic subpolar gyre region and globally, for lead years from 1 to 10, using the ensemble spread score to measure reliability. The lack of reliability of the ensembles implies that the spread estimated from them cannot be a trustworthy measure of the forecast uncertainty and the error. This is particularly important when the forecasts are used in decision making without any further postprocessing or expert adjustment.

b. Added value of INIT over NoINIT in terms of reliability

Some regions exhibit added value associated with the initialization in terms of one or both components of the JP reliability tests. Especially, the North Atlantic and Europe stand out as regions where there is some added value of INIT over NoINIT in terms of reliability, for forecast year 1 (Figs. 5 and 2) but also still present for forecast years 1–5 (Figs. S4 and S7). This is mainly due to the presence of smaller systematic mean biases and/or a better representation of the trend in INIT compared to NoINIT, as reflected in the slope parameter. The location of those regions displaying added value of INIT over NoINIT depends on whether the whole MM ensemble is analyzed, or the MM ensemble without NCAR, or only the NCAR large ensemble. In the NCAR ensemble for forecast year 1, INIT is more reliable than NoINIT in most regions in terms of slope (due to smaller biases and/or a better representation of the trend), but the added value of INIT over NoINIT in terms of convexity is more spatially heterogeneous (Fig. 7). However, except for East Africa (EAF) in the MM (Fig. S1) and MM–NCAR (Fig. S3) datasets, no region is significantly reliable in terms of the χ^2 coefficient. This implies, as explained in section 4a above, that all forecast system ensembles are affected by errors (systematic biases, incorrect trend representation, dispersion errors, etc.), which prevent the uncorrected INIT rank histograms from being significantly flatter than the uncorrected NoINIT ones.

A few previous studies have compared reliability of specific variables in INIT and NoINIT (Ho et al. 2013; Caron et al. 2015; Camp and Caron 2017). Caron et al. (2015) and Camp and Caron (2017) showed that the multimodel INIT ensemble was more reliable than the corresponding NoINIT ensemble in forecasting cyclone activity for forecast years 1–5 in United States landfalling cyclones. However, different variables (cyclone wind damage potential and frequency), model ensembles (only four models from CMIP5 and SPECS), regions (the

United States and North Atlantic), and metrics (reliability diagrams) studied make the comparison with the present study difficult. Ho et al. (2013) showed, by evaluating the spread-to-error ratio, that for short lead times (less than 2 years) INIT ensembles from the Met Office Decadal Prediction System tended to be underdispersive and produce overconfident and hence unreliable forecasts of sea surface temperature. This was also supported by findings from Marini et al. (2016) and Polkova et al. (2019) for single-model initialized ensembles generated with different methods. The NoINIT ensemble in Ho et al. (2013), on the contrary, was overdispersive and similarly unreliable. For longer lead times, both INIT and NoINIT ensembles were predominantly overdispersive (and unreliable), related to excessive interannual variability in the climate model. The single-model ensembles used in Ho et al. (2013) prevent us from making a fair comparison with our multimodel ensembles. However, the comparison with our results for the NCAR single-model ensemble reveal similar differences between INIT and NoINIT for the first forecast year, with INIT generally showing underdispersive behavior and NoINIT generally showing overdispersive behavior in most regions (e.g., Fig. 8).

c. Evolution of reliability over forecast time

The evolution of reliability over time can be appreciated by comparing results for forecast year 1 (Figs. 2–8 and Fig. S3) and for the average of forecast years 1–5 (Figs. S4–S7). In general, the differences between INIT and NoINIT are smaller for forecast years 1–5 than for forecast year 1. This is especially the case for the MM and MM–NCAR ensembles. The smaller differences are linked to the fact that, for forecast years 1–5, the potential beneficial impact of the initialization does not overcome the biases in the trend and variance (Fig. 5; see also Figs. S3–S5, S28, and S29).

The convergence of INIT and NoINIT in terms of reliability over time is somehow similar to what is generally observed in terms of forecast accuracy skill measures for near-surface temperature: INIT usually shows more skill than NoINIT for the first forecast year(s), because of initialization, but then the skill rapidly converges such that their skill after the first few forecast years is similar (e.g., Doblas-Reyes et al. 2013; Marotzke et al. 2016; Smith et al. 2018; Yeager et al. 2018; Liu et al. 2019). As for forecast accuracy, this behavior is largely region dependent.

The convergence of reliability over time especially affects the convexity coefficient (Figs. 5 and 7 and Figs. S3–S6), implying an evolution in the representation of the spread over time (the spread of INIT and NoINIT becomes more similar with time). In the case of the MM and MM–NCAR ensembles, this evolution in spread is linked to an improvement in the representation of the spread for NoINIT when 5-yr averages are considered, reflected in a lower contribution of the JP convexity coefficient (not shown). For the NCAR ensemble, both the spread of INIT and NoINIT converge along the forecast time, with lower JP convexity contribution for some regions and higher JP convexity contribution for others (Figs. 6 and 7 and Fig. S6).

The evolution of the convexity with forecast time should be interpreted as the evolution of the ensemble spread with

respect to the observation, which would usually grow with forecast time until the convexity saturates. The physics behind it are the physics that sustain ensemble growth, which also depends on the way the ensembles have been generated (please see section 4d below).

d. Impact of different forecast system ensembles

In this study, we have used three different (partly overlapping) forecast system ensembles: MM, MM–NCAR, and NCAR. In general, reliability is very similar for the MM and MM–NCAR ensembles, but the NCAR ensemble on its own shows discrepancies. For example, in terms of spread representation, MM INIT and NoINIT generally exhibit the same behavior; for example, when INIT is overdispersive (underdispersive), NoINIT is overdispersive (underdispersive) too, but this is not always the case for NCAR INIT and NoINIT (see, e.g., Fig. 8). Furthermore, the uncorrected NCAR INIT clearly presents deficiencies in terms of spread compared to NoINIT, resulting in limited added value of INIT over NoINIT for this single-model ensemble (Fig. 7).

There are two different contributions to the reliability of a forecast system: the model and the ensemble generation method. On the one hand, the model might have systematic errors in the variability that make the ensemble to be unreliable, and this would have a physical interpretation. On the other hand, the ensemble generation conditions the growth of the ensemble spread, at least in the first few months and years, and can lead to forecasts with different reliability. In a multimodel ensemble, both sources are mixed and determining the dominant one is not straightforward. The only way to address this issue would involve stratifying the forecast systems that enter the multimodel and looking at the difference in behavior of the slope and convexity parameters and the residual, which is beyond the scope of this article.

Despite the clear benefit of using a multimodel large ensemble compared to a single-model large ensemble, removing the NCAR ensemble from the MM ensemble (MM–NCAR) has only a small impact on forecast reliability. This illustrates the point that, even more important than having very large ensemble sizes (typically more than 40 ensemble members), ensembles composed of different forecast systems can avoid some deteriorating effects of deficiencies in individual systems, as already noted for seasonal forecasting (e.g., Hagedorn et al. 2005). Indeed, ensembles made of different forecast systems encompass a larger range of model physics and initialization approaches, and thereby also allow for error compensation.

e. Impact of the reference observational dataset and variables used

Using another reference observational dataset than GISTEMP, as illustrated in Figs. S30 and S31 with the HadCRUT4 dataset for the MM ensemble, has an impact on the JP statistics for some regions. The contribution to χ^2 of the convexity parameter in particular is generally lower when using HadCRUT4 compared to GISTEMP. The slope coefficient seems less impacted, except for the Southern Ocean (SOO) and, to a lesser extent, Australia (AUS). In general, using HadCRUT4 instead of GISTEMP yields more regions with significantly unbiased

results, correct trend representation or correct dispersion results. Consequently, there are more regions that are significantly reliable (in the $\chi^2 p$ value sense) when using HadCRUT4 instead of GISTEMP (Fig. S30). However, using HadCRUT4 instead of GISTEMP produces fewer cases in which INIT has some added value over NoINIT (bottom two rows for each forecast time in Fig. S30), especially in terms of convexity. The differences between results using the two reference datasets outlined above provide a good indication of the observational uncertainty. However, it should be borne in mind that the differences for some of the regions are probably partly due to a lower observational coverage in HadCRUT4 compared to GISTEMP (Morice et al. 2012). In particular, polar regions (SOO, ALA, GRL, NAS, and NEU) and some regions in Africa (SAH, WAF, EAF, and SAF), South America (AMZ), and Asia (CAS and TIB) are likely affected. Moreover, as noted by Marini et al. (2016), part of the observational uncertainty may also arise from using observational datasets that underestimate variability, for example, due to (temporal or spatial) smoothing of observations. The forecast verification would then yield artificially overdispersive forecasts by comparison. HadCRUT4 does not use any form of spatial infilling (Morice et al. 2012), while GISTEMP does (Lenssen et al. 2019), which might explain some differences—at least in terms of convexity—when using both datasets (i.e., more regions with apparently overdispersive INIT in Fig. 4 than in Fig. S28).

Furthermore, the impact on reliability of using a combination of near-surface air temperature (over land) and sea surface temperature (over oceans) instead of near-surface air temperature only is illustrated for the NCAR NoINIT ensemble in Fig. S32. The effect is rather small and limited to some specific regions (e.g., the Southern Ocean or eastern North America), showing that near-surface air temperature and sea surface temperature have similar reliability and that the assessment of forecast systems using near-surface air temperature only is appropriate.

f. Impact of postprocessing

As all forecast system ensembles have issues with reliability (explained in section 4a above), we tested the impact on reliability of applying three different postprocessing methods, namely detrending, bias correction and calibration. We have shown that the three methods have a significant impact on the slope coefficient (Fig. 9). They decrease the contribution of the slope coefficient in most regions in both INIT and NoINIT, for forecast year 1 and forecast years 1–5, resulting in no substantial differences between INIT and NoINIT. Detrending improves the slope coefficient because it removes the linear trend in the verified datasets and the verifying observations alike. Bias correction and calibration have an impact on the slope coefficient through the removal of systematic biases, the former acting upon the variance, and the latter upon the relation between the skill and the spread.

The convexity coefficient, on the other hand, is less easily corrected for. The postprocessing will correct, in a statistical way, some of the systematic errors that are behind the convexity and, hence, have no physical basis beyond the understanding of the sources of those systematic errors (e.g., a high

climate sensitivity that makes the forecast trend to be too strong or an underestimation of the observed variability that would make the spread to be too small). Calibration is the only postprocessing method that yields some improvement in the convexity parameter (Fig. 9). Detrending has some impact, but mostly negative, probably because by removing the trends other problems in the ensemble distributions are uncovered. Bias correction treats the errors in the variance, so it has a small impact on the convexity parameter through its impact on spread. The rather limited improvement (or even worsening in some cases) of spread by the calibration method has already been noted in other studies, in which this improvement was found to be very dependent on the region (Torralba et al. 2017; Manzananas et al. 2019). However, because it corrects the slope coefficient, the convexity coefficient and the residual parameter at the same time, calibration is the only method that significantly improves the general reliability (measurable through the $\chi^2 p$ value in Figs. S22–S24).

Combinations of the different postprocessing methods were also tested (not shown), leading to some improvement in reliability for a few regions compared to the uncorrected results.

By assessing reliability after applying different postprocessing techniques to the raw uncorrected data, we have identified some of the reasons behind the lack of reliability of the different ensembles and shown the necessity of bias correction and calibration in order to stand a chance to obtain more reliable predictions and projections of climate.

5. Conclusions

In this study, we assessed the reliability of multimodel initialized decadal predictions compared to noninitialized projections of near-surface air temperature in 30 different regions of the world, using rank histograms and the Jolliffe and Primo (2008) test statistics. We looked at three forecast system ensembles of different size and model combinations to investigate the impact on reliability. We also explored how reliability evolves with forecast time, by looking at results for forecast year 1 and forecast years 1–5, over the period 1961–2005. Finally, we tested the impact of applying several postprocessing techniques to the raw temperature anomalies, namely detrending, bias correction, and calibration.

Results indicate that both INIT and NoINIT uncorrected output ensembles are largely not reliable, and there is only a rather limited added value of decadal predictions for near-surface air temperature in terms of reliability, compared to noninitialized projections. Indeed, the added value is limited to specific regions and to the first forecast year(s), not unlike skill measures of forecast accuracy for near-surface temperature (e.g., Doblas-Reyes et al. 2013; Yeager et al. 2018; Smith et al. 2019).

Furthermore, using different forecast system ensembles has an impact on reliability, but the model combination inside the ensemble seems to play a larger role than the actual number of ensemble members. As such, we have shown that it is of advantage to use ensembles composed of different forecast systems, as those encompass a larger range of model physics and

initialization approaches, and thereby also allow for error compensation. The ensemble generation method is also a very important aspect that affects reliability, but addressing this goes beyond the scope of this study. We have also shown the impact of using different reference datasets in the analysis, thereby providing a measure of the observational uncertainty. Moreover, this study has demonstrated, as in previous studies, the need for bias correction and calibration of the raw data. This is crucial to obtain reliable predictions and projections of climate that can be useful to stakeholders from various economic sectors and societal groups to obtain more realistic estimates of event probabilities.

This study has focused on near-surface air temperature, but we expect reliability to be variable dependent. Further studies could investigate the reliability of precipitation, for example, or of specific climate indices relevant to users, such as drought indices (Solaraju-Murali et al. 2019), cyclones (Caron et al. 2018) or the Atlantic multidecadal variability (Ruprich-Robert et al. 2018). Additionally, we could explore the impact of the multimodel combinations on reliability, by 1) using the initialized decadal predictions and noninitialized projections that are currently becoming available from phase 6 of the Coupled Model Intercomparison Project (CMIP6; Eyring et al. 2016) and 2) testing other ways of building the multimodel ensembles, for example by ensemble weighting (Mishra et al. 2019; Zhang and Yan 2018). Moreover, the sensitivity of reliability to ensemble generation methods would be something useful to look at in future studies. Addressing this issue would require stratifying the forecast systems that enter the multimodel and looking at the difference in behavior of the slope and convexity parameters and the residual. Finally, the impact on reliability of other bias correction and calibration techniques could be investigated, such as quantile mapping (Déqué 2007; Themeßl et al. 2011), member-by-member bias correction directly from the rank histograms (Stanger et al. 2019), calibration through subensemble selection (Herger et al. 2018), or other member-by-member calibration approaches such as nonhomogeneous Gaussian regression (Van Schaeybroeck and Vannitsem 2015).

Acknowledgments. The EUCP project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement 776613. The research leading to these results has also received funding from the Spanish Ministerio de Ciencia, Innovación y Universidades as part of the CLINSA project with funding reference CGL2017-85791-R. MGD is also supported by the Spanish Ministry of Science, Innovation and Universities Grant RYC-2017-22964. BSM acknowledges financial support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement 713673 and from a fellowship of "la Caixa" Foundation (ID 100010434). The fellowship code is LCF/BQ/IN17/11620038. SW has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433. We acknowledge the use of the s2dverification (Manubens et al. 2018), startR (BSC/CNS and Manubens 2020), SpecsVerification (Siegert 2017), CSTools (Perez-Zanon et al. 2019), ClimProjDiags (BSC/

CNS et al. 2020), and boot (Davison and Hinkley 1997; Cauty and Ripley 2020) R (R Core Team 2013) software packages. We also thank Nicolau Manubens, An-Chi Ho, Pablo Ortega, Rashed Mahmood, Yohan Ruprich-Robert, and Roberto Bilbao from the BSC for their technical and scientific support. We further express our gratitude to Charles Pelletier from ELI for his careful reading of the manuscript. Finally, we would like to thank Antje Weisheimer (University of Oxford) and two anonymous reviewers for constructive comments that helped us greatly improve this article.

Data availability statement. The forecast systems data that support the findings of this study are openly available at <https://esgf-index1.ceda.ac.uk/projects/esgf-ceda/>. The GISTEMP data can be found at <https://data.giss.nasa.gov/gistemp/> and the HadCRUT4 data at <https://www.metoffice.gov.uk/hadobs/> and <https://crudata.uea.ac.uk/cru/data/temperature/>.

REFERENCES

- Bellprat, O., V. Guemas, F. Doblas-Reyes, and M. Donat, 2019: Towards reliable extreme weather and climate event attribution. *Nat. Commun.*, **10**, 1732, <https://doi.org/10.1038/s41467-019-09729-2>.
- Bellucci, A., and Coauthors, 2015: Advancements in decadal climate predictability: The role of nonoceanic drivers. *Rev. Geophys.*, **53**, 165–202, <https://doi.org/10.1002/2014RG000473>.
- Bilbao, R., and Coauthors, 2020: Assessment of a full-field initialised decadal climate prediction system with the CMIP6 version of EC-Earth. *Earth Syst. Dyn. Discuss.*, <https://doi.org/10.5194/esd-2020-66>, in press.
- Boero, G., J. Smith, and K. Wallis, 2005: The sensitivity of chi-squared goodness-of-fit tests to the partitioning of data. *Econom. Rev.*, **23**, 341–370, <https://doi.org/10.1081/ETC-200040782>.
- BSC/CNS, and N. Manubens, 2020: startR: Automatically retrieve multidimensional distributed data sets, version 0.1.4. Barcelona Supercomputing Center, R package, <https://earth.bsc.es/gitlab/es/startR/>.
- , N. Perez-Zanon, and A. Hunter, 2020: ClimProjDiags: Set of tools to compute various climate indices, version 0.1.0. Barcelona Supercomputing Center, R package, <https://CRAN.R-project.org/package=ClimProjDiags>.
- Buontempo, C., C. Hewitt, F. Doblas-Reyes, and S. Dessai, 2014: Climate service development, delivery and use in Europe at monthly to inter-annual timescales. *Climate Risk Manage.*, **6**, 1–5, <https://doi.org/10.1016/j.crm.2014.10.002>.
- Camp, J., and L.-P. Caron, 2017: Analysis of Atlantic tropical cyclone landfall forecasts in coupled GCMs on seasonal and decadal timescales. *Hurricanes and Climate Change*, J. Collins and K. Walsh, Eds., Springer, 213–241, https://doi.org/10.1007/978-3-319-47594-3_9.
- Cauty, A., and B. Ripley, 2020: boot: Bootstrap R (S-Plus) functions, version 1.3-25. R package, <https://CRAN.R-project.org/package=boot>.
- Caron, L.-P., L. Hermanson, and F. Doblas-Reyes, 2015: Multiannual forecasts of Atlantic U.S. tropical cyclone wind damage potential. *Geophys. Res. Lett.*, **42**, 2417–2425, <https://doi.org/10.1002/2015GL063303>.
- , —, A. Dobbin, J. Imbers, L. Lledó, and G. Vecchi, 2018: How skillful are the multiannual forecasts of Atlantic hurricane activity? *Bull. Amer. Meteor. Soc.*, **99**, 403–413, <https://doi.org/10.1175/BAMS-D-17-0025.1>.

- Corti, S., A. Weisheimer, T. Palmer, F. Doblas-Reyes, and L. Magnusson, 2012: Reliability of decadal predictions. *Geophys. Res. Lett.*, **39**, L21712, <https://doi.org/10.1029/2012GL053354>.
- Davidson, A., and D. Hinkley, 1997: *Bootstrap Methods and Their Applications*. Cambridge University Press, 582 pp.
- Déqué, M., 2007: Frequency of precipitation and temperature extremes over France in an anthropogenic scenario: Model results and statistical correction according to observed values. *Global Planet. Change*, **57**, 16–26, <https://doi.org/10.1016/j.gloplacha.2006.11.030>.
- Doblas-Reyes, F., R. Hagedorn, and T. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, <https://doi.org/10.3402/tellusa.v57i3.14658>.
- , and Coauthors, 2013: Initialized near-term regional climate change prediction. *Nat. Commun.*, **4**, 1715, <https://doi.org/10.1038/ncomms2704>.
- Du, H., F. Doblas-Reyes, J. García-Serrano, V. Guemas, Y. Soufflet, and B. Wouters, 2012: Sensitivity of decadal predictions to the initial atmospheric and oceanic perturbations. *Climate Dyn.*, **39**, 2013–2023, <https://doi.org/10.1007/s00382-011-1285-9>.
- Eade, R., D. Smith, A. Scaife, E. Wallace, N. Dunstone, L. Hermanson, and N. Robinson, 2014: Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophys. Res. Lett.*, **41**, 5620–5628, <https://doi.org/10.1002/2014GL061146>.
- Elmore, K., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Wea. Forecasting*, **20**, 789–795, <https://doi.org/10.1175/WAF884.1>.
- Eyring, V., S. Bony, G. Meehl, C. Senior, B. Stevens, R. Stouffer, and K. Taylor, 2016: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.*, **9**, 1937–1958, <https://doi.org/10.5194/gmd-9-1937-2016>.
- Giorgi, F., and R. Francisco, 2000: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HadCM2 coupled AOGCM. *Climate Dyn.*, **16**, 169–182, <https://doi.org/10.1007/PL00013733>.
- GISTEMP Team, 2019: GISS Surface Temperature Analysis (GISTEMP), version 4. NASA Goddard Institute for Space Studies, accessed 30 April 2019, <https://data.giss.nasa.gov/gistemp/>.
- Hagedorn, R., F. Doblas-Reyes, and T. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, <https://doi.org/10.3402/tellusa.v57i3.14657>.
- Hamill, T., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2).
- Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1108, <https://doi.org/10.1175/2009BAMS2607.1>.
- Herger, N., O. Angéllil, G. Abramowitz, M. Donat, D. Stone, and K. Lehmann, 2018: Calibrating climate model ensembles for assessing extremes in a changing climate. *J. Geophys. Res. Atmos.*, **123**, 5988–6004, <https://doi.org/10.1029/2018JD028549>.
- Ho, C., E. Hawkins, L. Shaffrey, J. Bröcker, L. Hermanson, J. Murphy, D. Smith, and R. Eade, 2013: Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. *Geophys. Res. Lett.*, **40**, 5770–5775, <https://doi.org/10.1002/2013GL057630>.
- Jolliffe, I., and C. Primo, 2008: Evaluating rank histograms using decompositions of the chi-square test statistic. *Mon. Wea. Rev.*, **136**, 2133–2139, <https://doi.org/10.1175/2007MWR2219.1>.
- Kadow, C., S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch, 2017: Decadal climate predictions improved by ocean ensemble dispersion filtering. *J. Adv. Model. Earth Syst.*, **9**, 1138–1149, <https://doi.org/10.1002/2016MS000787>.
- Keenlyside, N., M. Latif, J. Jungclauss, L. Kornblueh, and E. Roeckner, 2008: Advancing decadal-scale climate prediction in the North Atlantic sector. *Nature*, **453**, 84–88, <https://doi.org/10.1038/nature06921>.
- Kendall, M., and A. Stuart, 1967: *The Advanced Theory of Statistics: Inference and Relationship*. Vol. 2, Charles Griffin, 690 pp.
- Kirtman, B., and Coauthors, 2013: Near-term climate change: Projections and predictability. *Climate Change 2013: The Physical Science Basis*, T. F. Stocker et al., Eds., Cambridge University Press, 953–1028.
- Kushnir, Y., and Coauthors, 2019: Towards operational predictions of the near-term climate. *Nat. Climate Change*, **9**, 94–101, <https://doi.org/10.1038/s41558-018-0359-7>.
- Lehner, F., C. Deser, N. Maher, J. Marotzke, E. Fischer, L. Brunner, R. Knutti, and E. Hawkins, 2020: Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6. *Earth Syst. Dyn.*, **11**, 491–508, <https://doi.org/10.5194/esd-11-491-2020>.
- Lenzen, 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**, 6307–6326, <https://doi.org/10.1029/2018JD029522>.
- Leung, L., A. Hamlet, D. Lettenmaier, and A. Kumar, 1999: Simulations of the ENSO hydroclimate signals in the Pacific Northwest Columbia River basin. *Bull. Amer. Meteor. Soc.*, **80**, 2313–2330, [https://doi.org/10.1175/1520-0477\(1999\)080<2313:SOTEHS>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<2313:SOTEHS>2.0.CO;2).
- Liu, Y., M. Donat, A. Taschetto, F. Doblas-Reyes, L. Alexander, and M. England, 2019: A framework to determine the limits of achievable skill for interannual to decadal climate predictions. *J. Geophys. Res. Atmos.*, **124**, 2882–2896, <https://doi.org/10.1029/2018JD029541>.
- Manubens, N., and Coauthors, 2018: An R package for climate forecast verification. *Environ. Modell. Software*, **103**, 29–42, <https://doi.org/10.1016/j.envsoft.2018.01.018>.
- Manzanas, R., J. Gutiérrez, J. Bhend, S. Hemri, F. Doblas-Reyes, V. Torralba, E. Penabad, and A. Brookshaw, 2019: Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Climate Dyn.*, **53**, 1287–1305, <https://doi.org/10.1007/s00382-019-04640-4>.
- Marini, C., I. Polkova, A. Köhl, and D. Stammer, 2016: A comparison of two ensemble generation methods using oceanic singular vectors and atmospheric lagged initialization for decadal climate prediction. *Mon. Wea. Rev.*, **144**, 2719–2738, <https://doi.org/10.1175/MWR-D-15-0350.1>.
- Marotzke, J., and Coauthors, 2016: MiKlip: A national research project on decadal climate prediction. *Bull. Amer. Meteor. Soc.*, **97**, 2379–2394, <https://doi.org/10.1175/BAMS-D-15-00184.1>.
- Meehl, G., and Coauthors, 2007: Global climate projections. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 747–846.
- Ménégoz, M., R. Bilbao, O. Bellprat, V. Guemas, and F. Doblas-Reyes, 2018: Forecasting the climate response to volcanic eruptions: Prediction skill related to stratospheric aerosol forcing. *Environ. Res. Lett.*, **13**, 064022, <https://doi.org/10.1088/1748-9326/aac4db>.
- Merryfield, W., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer.*

- Meteor. Soc.*, **101** (6), E869–E896, <https://doi.org/10.1175/BAMS-D-19-0037.1>.
- Mishra, N., C. Prodhomme, and V. Guemas, 2019: Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe. *Climate Dyn.*, **52**, 4207–4225, <https://doi.org/10.1007/s00382-018-4404-z>.
- Morice, C., J. Kennedy, N. Rayner, and P. Jones, 2012: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res.*, **117**, D08101, <https://doi.org/10.1029/2011JD017187>.
- Nissan, H., L. Goddard, E. de Perez, J. Furlow, W. Baethgen, M. Thomson, and S. Mason, 2019: On the use and misuse of climate change projections in international development. *Wiley Interdiscip. Rev.: Climate Change*, **10**, e579, <https://doi.org/10.1002/wcc.579>.
- Pasternack, A., J. Bhend, M. Liniger, H. Rust, W. Müller, and U. Ulbrich, 2018: Parametric decadal climate forecast recalibration (DeFoReSt 1.0). *Geosci. Model Dev.*, **11**, 351–368, <https://doi.org/10.5194/gmd-11-351-2018>.
- Perez-Zanon, N., and Coauthors, 2019: CSTools: Assessing skill of climate forecasts on seasonal-to-decadal timescales, version 2.0.0. R package, <https://CRAN.R-project.org/package=CSTools>.
- Pohlmann, H., J. Jungclaus, A. Köhl, D. Stammer, and J. Marotzke, 2009: Initializing decadal climate predictions with the GECCO oceanic synthesis: Effects on the North Atlantic. *J. Climate*, **22**, 3926–3938, <https://doi.org/10.1175/2009JCLI2535.1>.
- Polkova, I., and Coauthors, 2019: Initialization and ensemble generation for decadal climate predictions: A comparison of different methods. *J. Adv. Model. Earth Syst.*, **11**, 149–172, <https://doi.org/10.1029/2018MS001439>.
- R Core Team, 2013: R: A language and environment for statistical computing. <https://www.R-project.org/>.
- Ruprich-Robert, Y., T. Delworth, R. Msadek, F. Castruccio, S. Yeager, and G. Danabasoglu, 2018: Impacts of the Atlantic multidecadal variability on North American summer climate and heat waves. *J. Climate*, **31**, 3679–3700, <https://doi.org/10.1175/JCLI-D-17-0270.1>.
- Sandgathe, S., B. Brown, J. Carman, J. Infanti, B. Johnson, D. McCarren, and E. McIlvain, 2020: Exploring the need for reliable decadal prediction. *Bull. Amer. Meteor. Soc.*, **101** (2), E141–E145, <https://doi.org/10.1175/BAMS-D-19-0248.1>.
- Siebert, S., 2017: SpecsVerification: Forecast verification routines for ensemble forecasts of weather and climate, version 0.5-2. R package, <https://CRAN.R-project.org/package=SpecsVerification>.
- Smith, D., S. Cusack, A. Colman, C. Folland, G. Harris, and J. Murphy, 2007: Improved surface temperature prediction for the coming decade from a global climate model. *Science*, **317**, 796–799, <https://doi.org/10.1126/science.1139540>.
- , and Coauthors, 2018: Predicted chance that global warming will temporarily exceed 1.5°C. *Geophys. Res. Lett.*, **45**, 11 895–11 903, <https://doi.org/10.1029/2018GL079362>.
- , and Coauthors, 2019: Robust skill of decadal climate predictions. *npj Climate Atmos. Sci.*, **2**, 13, <https://doi.org/10.1038/s41612-019-0071-y>.
- Soares, M., M. Alexander, and S. Dessai, 2018: Sectoral use of climate information in Europe: A synoptic overview. *Climate Serv.*, **9**, 5–20, <https://doi.org/10.1016/j.cliser.2017.06.001>.
- Solaraju-Murali, B., L.-P. Caron, N. Gonzalez-Reviriego, and F. Doblas-Reyes, 2019: Multi-year prediction of European summer drought conditions for the agricultural sector. *Environ. Res. Lett.*, **14**, 124014, <https://doi.org/10.1088/1748-9326/ab5043>.
- Stanger, J., I. Finney, A. Weisheimer, and T. Palmer, 2019: Optimising the use of ensemble information in numerical weather forecasts of wind power generation. *Environ. Res. Lett.*, **14**, 124086, <https://doi.org/10.1088/1748-9326/ab5e54>.
- Stolzenberger, S., R. Glowienka-Hense, T. Spanghel, M. Schröder, A. Mazurkiewicz, and A. Hense, 2016: Revealing skill of the MiKlip decadal prediction system by three-dimensional probabilistic evaluation. *Meteor. Z.*, **25**, 657–671, <https://doi.org/10.1127/metz/2015/0606>.
- Taylor, K., R. Stouffer, and G. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, <https://doi.org/10.1175/BAMS-D-11-00094.1>.
- Thiemeß, J., A. Gobiet, and A. Leuprecht, 2011: Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.*, **31**, 1530–1544, <https://doi.org/10.1002/joc.2168>.
- Torralba, V., F. Doblas-Reyes, D. MacLeod, I. Christel, and M. Davis, 2017: Seasonal climate prediction: A new source of information for the management of wind energy resources. *J. Appl. Meteor. Climatol.*, **56**, 1231–1247, <https://doi.org/10.1175/JAMC-D-16-0204.1>.
- van Oldenborgh, G., F. Doblas-Reyes, S. Drijfhout, and E. Hawkins, 2013: Reliability of regional climate model trends. *Environ. Res. Lett.*, **8**, 014055, <https://doi.org/10.1088/1748-9326/8/1/014055>.
- Van Schaeybroeck, B., and S. Vannitsem, 2015: Ensemble post-processing using member-by-member approaches: Theoretical aspects. *Quart. J. Roy. Meteor. Soc.*, **141**, 807–818, <https://doi.org/10.1002/qj.2397>.
- Vaughan, C., and S. Dessai, 2014: Climate services for society: Origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdiscip. Rev.: Climate Change*, **5**, 587–603, <https://doi.org/10.1002/wcc.290>.
- von Storch, H., and F. Zwiers, 2001: *Statistical Analysis in Climate Research*. Cambridge University Press, 496 pp.
- Weisheimer, A., and T. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, <https://doi.org/10.1098/rsif.2013.1162>.
- Yeager, S., and Coauthors, 2018: Predicting near-term changes in the Earth system: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bull. Amer. Meteor. Soc.*, **99**, 1867–1886, <https://doi.org/10.1175/BAMS-D-17-0098.1>.
- Zhang, X., and X. Yan, 2018: Criteria to evaluate the validity of multi-model ensemble methods. *Int. J. Climatol.*, **38**, 3432–3438, <https://doi.org/10.1002/joc.5486>.