Combination of Decadal Predictions and Climate Projections in Time: Challenges and Potential Solutions

D. J. Befort, L. Brunner, Leonard F. Borchert, C. H. O'Reilly, Juliette Mignot, Andrew P. Ballinger, G. C. Hegerl, J. M. Murphy, Antje Weisheimer

Geophysical Research Letters

RESEARCH LETTER
10.1029/2022GL098568

Key Points:

- Concatenating climate predictions and projections can introduce inconsistencies
- Concatenation is less problematic for quantiles close to the median compared to more extreme quantiles of the ensemble distribution
- A simple calibration method reduces but does not eliminate inconsistencies

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

D. J. Befort,
daniel.befort@physics.ox.ac.uk

Citation:

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

Received 8 MAR 2022
Accepted 20 JUN 2022

Combination of Decadal Predictions and Climate Projections in Time: Challenges and Potential Solutions

D. J. Befort[1], L. Brunner[2,3], L. F. Borchert[4,5], C. H. O'Reilly[6], J. Mignot[4], A. P. Ballinger[7], G. C. Hegerl[7], J. M. Murphy[8], and A. Weisheimer[9,10]

[1]Department of Physics, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK, [2]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland, [3]Department of Meteorology and Geophysics, University of Vienna, Vienna, Austria, [4]LOCEAN Laboratory, Sorbonne Universités (SU/CNRS/IRD/MNHN), Institut Pierre Simon Laplace (IPSL), Paris, France, [5]École Normale Supérieure, Laboratoire de Météorologie Dynamique (LMD), IPSL, Paris, France, [6]Department of Meteorology, University of Reading, Reading, UK, [7]School of Geosciences, University of Edinburgh, Edinburgh, UK, [8]Met Office Hadley Centre, Exeter, UK, [9]National Centre for Atmospheric Science, University of Oxford, Oxford, UK, [10]European Centre for Medium-Range Weather Forecasts, Reading, UK

Abstract This study presents an approach to provide seamless climate information by concatenating decadal climate predictions and climate projections in time. Results for near-surface air temperature over 29 regions indicate that such an approach has potential to provide meaningful information but can also introduce significant inconsistencies. Inconsistencies are often most pronounced for relatively extreme quantiles of the CMIP6 multi-model ensemble distribution, whereas they are generally smaller and mostly insignificant for quantiles close to the median. The regions most affected are the North Atlantic, Greenland and Northern Europe. Two potential ways to reduce inconsistencies are discussed, including a simple calibration method and a weighting approach based on model performance. Calibration generally reduces inconsistencies but does not eliminate all of them. The impact of model weighting is minor, which is found to be linked to the small size of the decadal climate prediction ensemble, which in turn limits the applicability of that method.

Plain Language Summary Continuous and consistent information about the evolution of climate in next 1–40 years is crucial for the development of mitigation and adaptation strategies. Historically, scientific products providing weather and climate information have been designed for specific time scales, for example, seasonal or decadal climate predictions (1–10 years), or long-term climate projections (1–100 + years). As a consequence, currently only climate projections can be used to provide continuous information beyond the 10 years time scale. However, as climate projections are not initialized using observations, they are likely – depending on the variable and region – less skillful than initialized predictions for shorter lead times. In this study we assess if meaningful continuous climate information can be obtained by concatenating decadal climate predictions and climate projections in time. Results suggest that significant inconsistencies (materialized in *jumps* in the resulting time series) can be introduced by this approach, however, for some regions temporal concatenation might provide meaningful climate information beyond decadal time scales. Furthermore, potential ways to reduce inconsistencies when concatenating are discussed.

1. Introduction

Seamless climate information for the near-term future (up to ~40 years) is crucial for the development of mitigation and adaptation strategies (Hewitt & Lowe, 2018; Jones & Mearns, 2005; Nissan et al., 2019). Historically, scientific products providing weather and climate information have been designed for specific time scales, for example, daily weather forecasts, seasonal or decadal climate predictions (from hereon called decadal predictions), or long-term climate projections. This is motivated by the fact that short-term predictions (up to ~5 years) require accurate assimilation of the observed state at the start of the model integration, whereas for predictions/projections beyond 5–10 years the correct representation of external forcings is thought to be more important (Branstator & Teng, 2010). As initialized decadal predictions are only providing data for the following 10 years after initialization, only climate projections can be used to provide seamless information beyond the 10 years time scale. However, as climate projections are not initialized using observations, they are likely – depending on variable and region – less skillful than initialized predictions for shorter lead times (Boer et al., 2016; Meehl et al., 2009, 2014). Users of climate information require the most accurate, reliable and consistent climate

information within and beyond the next 10 years, which demands the temporal combination of initialized predictions and uninitialized projections.

The initialized products with the longest forecast range currently available are decadal prediction ensembles providing climate information up to 10 years into the future (Boer et al., 2016; Meehl et al., 2009, 2014). Several studies analyzed the skill of decadal predictions to forecast past observed variability and their added value over climate projections (e.g., Borchert, Koul, et al., 2021; Boer et al., 2016; Doblas-Reyes et al., 2013; Meehl et al., 2014; Smith et al., 2019; Yeager & Robson, 2017). One approach to make use of the added value of initialization carried by decadal predictions at interannual lead times, and to provide seamless climate information beyond decadal time scale, has been recently developed by Befort et al. (2020). In this framework, uninitialized climate projections are constrained using decadal predictions by selecting only those projections that are closest to the decadal prediction ensemble mean. Befort et al. (2020) show that the constrained projection ensemble is significantly more skillful at predicting near-surface air temperatures over the North Atlantic (NAT) Gyre region compared to the unconstrained climate projection ensemble, even beyond decadal time scales when no information from decadal prediction is available anymore. A similar approach based on the idea of sub-sampling climate projections was applied to a large single model ensemble (Mahmood et al., 2021). The advantage of using sub-sampling approaches is that the resulting constrained ensembles are physically consistent as they are based solely on climate projections which are available for the entire period from the present up to the end of the century.

In this study we examine a new approach to provide seamless climate information beyond decadal time scales for near-surface air temperatures, based on the temporal concatenation of decadal predictions with climate projections. A particular focus is assessing to what extent temporal concatenation of prediction and projection time series introduces significant inconsistencies, for example, jumps. We also discuss ways to reduce potential inconsistencies by applying a calibration as well as a model weighting method.

2. Datasets

Near-surface air temperature (SAT) data are used from eight decadal ensemble prediction systems from the CMIP6 archive (Boer et al., 2016) and their corresponding uninitialized historical simulations (Eyring et al., 2016) (see Table S1 in Supporting Information S1). Each decadal prediction system consists of 10 members except CanESM5, which has 20 members. For NorCPM1, data for two different initialization techniques (i1 and i2) with 10 ensemble members each are available, which are treated as two different ensembles. To make prediction and projection ensembles comparable, the same number of ensemble members for each model is used, except for HadGEM3-GC31-MM as only four projection ensemble members are available but still 10 prediction members are used. Decadal prediction systems are corrected for lead-time dependent biases following Boer et al. (2016), using the years 1970–2014. For each historical climate projection ensemble, anomalies are calculated with respect to the single model ensemble average from 1970 until 2014.

All analyses are carried out for annual mean SAT (January to December averages) over 29 geographic regions: the 26 Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) regions (Field et al., 2012), the tropical central Pacific (NINO3.4) region, the NAT Ocean basin (identical to the N. Atlantic-Ocean region in Iturbide et al., 2020), and the global average (see Figure 3a for region boundaries). Note that no land-sea masking has been applied for any of the 29 regions.

3. Methodology

3.1. Framework for Temporal Concatenation

The main objective is to assess whether the temporal concatenation of decadal predictions and climate projections is sensible and to what extent it introduces inconsistencies, such as unrealistic year-to-year variations at the time of the concatenation (the precise definition is presented below). Figure 1 shows an example for the simple concatenation of predictions and projections after forecast year 10 for Northern Europe (NEU) annual mean SAT using the prediction initialized in 1975. Figure 1a presents time series of the climate projection multi-model ensemble (MME) from 1976 until 2001, whereas Figure 1b shows the decadal prediction MME from 1976 until 1985 and the concatenated projections thereafter. For this example, the decadal prediction MME is largely different to the

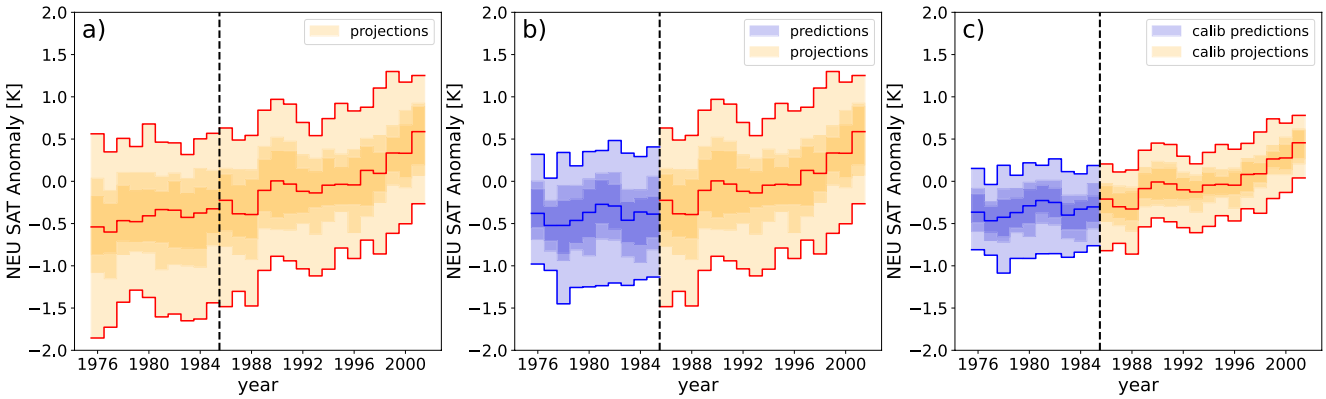

Figure 1. Example of inconsistencies introduced by concatenation of decadal predictions and climate projections. Time series of near-surface air temperatures over Northern Europe region (NEU) for (a) uninitialized climate projections from 1976 to 2001 (b) decadal predictions from 1976 to 1985 (initialized in 1975) and climate projections from 1986 to 2001 and (c) same as (b) but for calibrated decadal prediction and calibrated climate projection, using the variance inflation method. Solid lines in figures show 10th/90th percentile and median for the respective ensemble. Shading illustrates 10th/90th percentiles as well as tercile and quartile boundaries. Percentiles are calculated by ranking the different realizations of the ensemble.

climate projection MME, which leads to inconsistencies (materialized in jumps in the time series) for different quantiles of the ensemble distributions around the year 1986 when both datasets are concatenated in time. These discontinuities in the time series appear to be larger than those present in the climate projection MME between 1985 and 1986 (Figure 1a). However, the climate projection MME itself also exhibits some degree of interannual variability, which raises the question of how to objectively quantify significant inconsistencies introduced by the temporal concatenation of these different datasets.

3.2. Metrics to Assess Inconsistencies

The approach used here is based on the SAT interannual increments between 1970 and 2014, which are calculated as the difference in SAT between year n and year $n + 1$ ($1970 \leq n \leq 2013$). First the interannual increments are calculated for climate projection data only (an example is shown in Figures 2a and 2b). The resulting increment distribution consists of 44 values and is from now on referred to as *baseline* (Figure 2c). Second, increments between year n of the decadal predictions and year $n + 1$ from the projection (pairs of predictions and projections) as they would occur when concatenated are calculated (see Figures 2d and 2e for an example). The resulting distribution (of 44 values) is from now on called *merged* (Figure 2f). To investigate inconsistencies for different lead times, the *merged* increment distributions are calculated for different forecast years (ranging from 1 to 10). The *baseline* and *merged* increment distributions are further calculated for different quantiles of the prediction and projection MMEs (see Figure 1 for examples for different quantiles in the case of NEU). In this study 11 different quantiles ranging from 0 (minimum) to 1 (maximum) in 0.1 steps are analyzed, which are calculated by ranking the different realizations of the ensemble. From here on we define quantiles between 0.4 and 0.6 as *central* quantiles, 0.2–0.3, 0.7–0.8 as *moderate* quantiles, and 0.0–0.1, 0.9–1.0 as *extreme* quantiles.

To quantify whether concatenating produces larger interannual increments than the *baseline* values, four different metrics are defined and calculated for each quantile separately (Figure 2).

Metric 1 (M1) aims to quantify the difference between the *merged* and *baseline* increment means. M1 is calculated using the absolute differences of the mean values of *merged* and *baseline* divided by the standard deviation of the *baseline* distribution (this is the interannual standard deviation of the increment time series). The value of M1 is also known as the standard score (z-score). In the example of Figure 2g, the absolute z-score is 0.8, meaning that the differences in the means of *merged* and *baseline* distributions equals 0.8 times the interannual standard deviation of the *baseline* distribution. A small z-score close to 0 indicates that *merged* and *baseline* distributions are consistent with each other whereas larger z-scores hint toward inconsistencies.

Metric M2 aims to assess differences of the whole *merged* and *baseline* distributions. This is done by calculating the overlap between *merged* and *baseline* increment distributions after estimating their probability density functions. Here, a value of one indicates a complete overlap (perfect consistency between *merged* and *baseline*

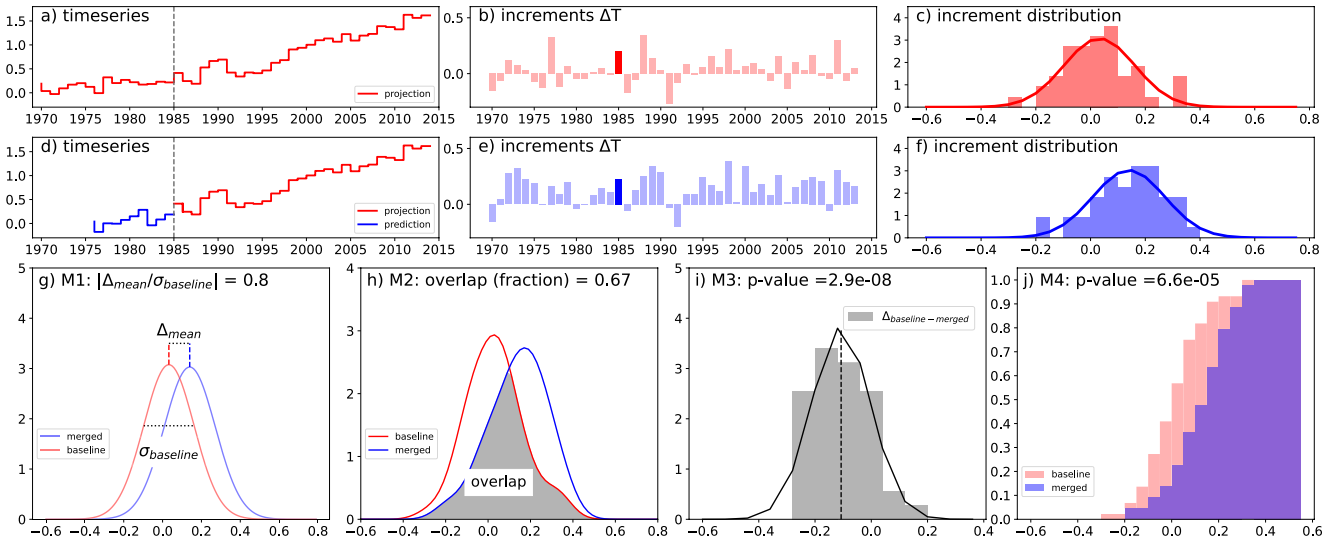

Figure 2. Approach used to assess inconsistencies when concatenating decadal predictions and climate projections illustrated using data for the 80th percentile of Northern Europe surface air temperature. (a) climate projection timeseries. (b) interannual increments of the projection timeseries. The dark red bar indicates the increment shown as concatenating example in (a). (c) resulting *baseline* increment distribution (red curve shows Gaussian fit). (d) example for concatenating decadal prediction data from 1976 to 1985 and climate projections afterwards. (e) interannual increments when concatenating prediction using forecast year 10 and projections for all years (1970–2014). The dark blue bar indicates the increment shown as concatenating example in (d) and (f) resulting *merged* increment distribution. Overview of metrics: (g) M1: *baseline* and *merged* increment distributions (Gaussian fits are plotted for visualization). Location of the respective means is shown as dashed lines, whereas the dotted line illustrates the standard deviation of the *baseline* increment distribution and the difference between the distribution means (Δ_{mean}) respectively. The value of M1 (absolute z-score) is given on top. (h) M2: Probability density functions (pdf) for *baseline* and *merged* distributions as well as their overlap (gray shading). The value of M2 is displayed on top (i) M3: Distribution of the difference between *baseline* and *merged* increment timeseries (b and e). The p-value is derived from the one-sided *t*-test. (j) M4: Cumulative distributions for *baseline* and *merged* and the associated Kolmogorov-Smirnov test *p*-value. Note that panels (g) and (h) represent the same distributions using a parametric (g) and non-parametric (h) approach to estimate the pdfs.

distributions) whereas a value of 0 indicates no overlap. To assess statistical significance, we calculated the overlap between the *baseline* distribution and 1,000 samples of the distribution generated by randomly sampling 44 values (as there are 44 increments) from the *baseline* distribution with replacement. For the example in Figure 2h, the overlap is about 0.67, which is smaller than the 10th percentile of the 1,000 randomly bootstrapped samples.

Metric 3 (M3) is based on the *p*-values of a *t*-test to estimate the significance of the mean difference between the *merged* and *baseline* distributions. Here we test the difference distribution (*baseline–merged*) against an expected value of 0. For the example of the 80th percentile of NEU averaged SAT (Figure 2i), M3 is below 0.01 indicating that the mean of the difference distribution is significantly different from 0.

Metric 4 (M4) is based on the *p*-value of the Kolmogorov-Smirnov test statistics comparing *baseline* and *merged* increment cumulative distributions (Figure 2j). Similarly to M3, a small *p*-value indicates significant differences between the two increment distributions.

Note that the bias-correction applied to projections and predictions directly leads to small (close to 0) differences between the mean increment for the MME median for the *merged* and *baseline* distributions. In other words metrics M1 and M3 for the median (quantile: 0.5) will likely indicate non-significant differences between *merged* and *baseline* distributions. This is not necessarily the case for M2 & M4, which compare the full distributions.

4. Results

4.1. Examining the Temporal Concatenation of Predictions and Projections

Results of the metrics M1 & M2 when concatenating decadal predictions after forecast year 10 with climate projections are shown in Figure 3b. These suggest there are relatively small inconsistencies for *central* quantiles (0.4–0.6) of the MME for almost all regions. For M1 this is to a large extent related to the bias-correction applied to predictions and projections (see Section 3). However, the high M2 values indicate that the whole increment distributions are also in agreement with each other. The only exceptions are NAT, NEU and Greenland (CGI)

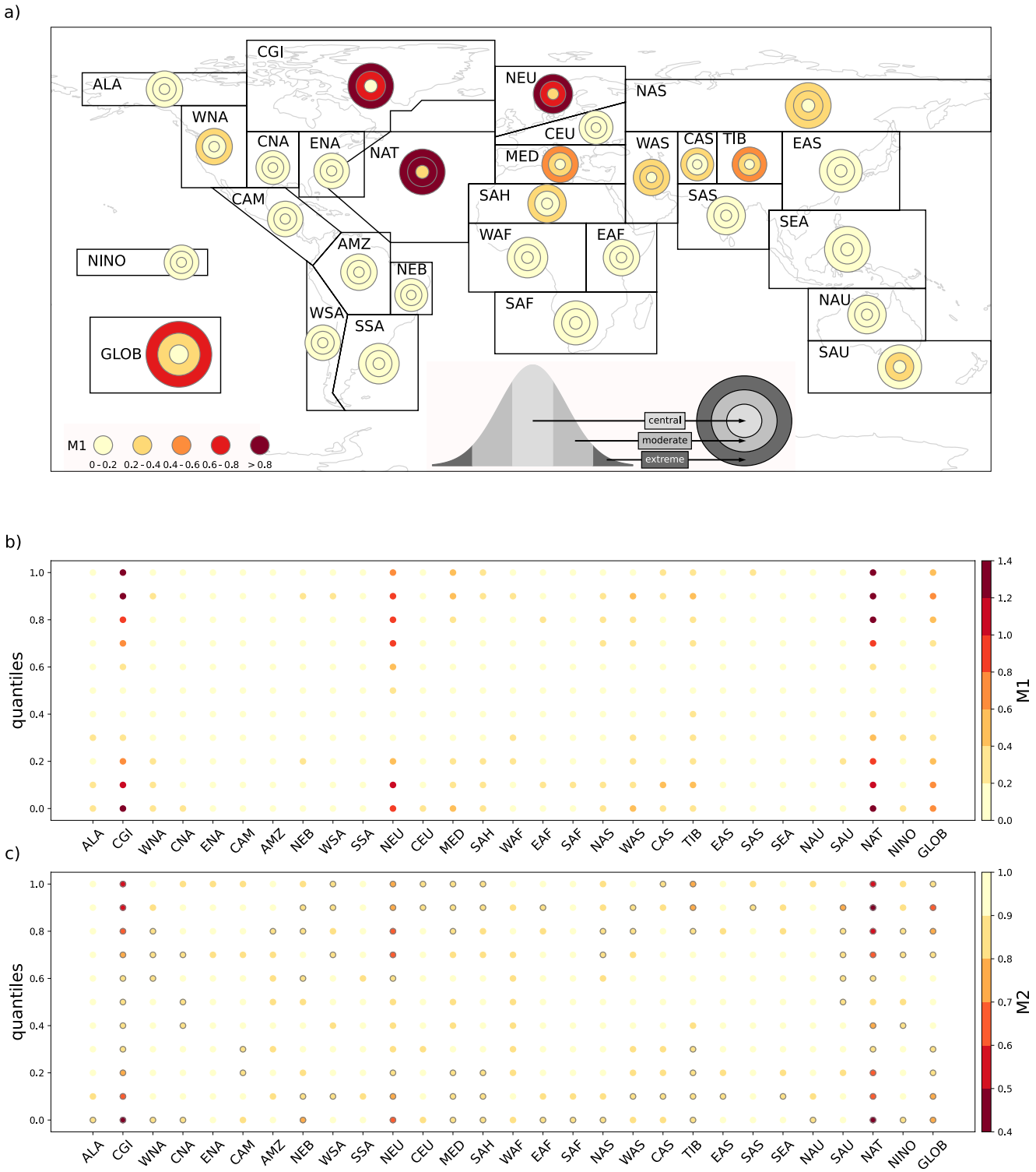

Figure 3. (a) Map illustrating significant inconsistencies for the 29 regions used in this study. Each region is represented by a "dartboard"-plot, with the rings representing results for different quantiles. The inner ring represents *central* quantiles (0.4–0.6), the middle ring represents *moderate* quantiles (0.2–0.3, 0.7–0.8) and the outer ring represents the *extreme* quantiles (0–0.1, 0.9–1). The color in the ring indicates the level of inconsistencies introduced when concatenating decadal predictions and climate projections after forecast year 10. Here, the color indicates the average M1 metric value over the respective quantiles. (b) Results for metrics M1when concatenating decadal predictions after forecast year 10 and climate projections. Regions are shown on the x-axis, while the y-axis displays the quantiles that are examined. (c) same as (b) but for metric M2. The circles in M2 indicate those values which are significantly different to the 10th percentile of randomly sampled *baseline* distribution overlaps (see method section).

regions, for which the overlap of the *baseline* and *merged* distributions is significantly smaller for several *central* quantiles than what is expected from sampling the sensitivity of the underlying *baseline* distribution. This result for *central* quantiles is generally supported by M3 and M4 (Figure S1 in Supporting Information S1), which also indicate no significant inconsistencies with the exception of the NAT, NEU and CGI regions for M3.

For *moderate* and *extreme* quantiles (0–0.3 & 0.7–1.0) larger inconsistencies are indicated by M1 for some regions. These include CGI, NEU, NAT and globally averaged SAT (GLOB). This is supported by M2, which also indicates significant differences between the full distributions for these regions. Additionally, significant differences for M2 are also found for example, the Mediterranean (MED), Western Asia (WAS) and Tibet (TIB) regions. For other regions M2 also indicates significant differences for some quantiles, however, the relatively small sample size of only 44 increment values needs to be taken into account, which might affect the robustness of statistical tests as well as the estimation of the probability density functions. Thus, we put focus on those regions for which several different metrics suggest the appearance of significant inconsistencies as we regard the trustworthiness of these results as increased. Results for M3 and M4 are in line with those from M1 and M2 (Figure S1 in Supporting Information S1), significant inconsistencies found for CGI, NEU, NAT, GLOB in both metrics whereas inconsistencies are found for several *moderate* and *extreme* quantiles for MED, WAS, TIB for M3.

Overall, inconsistencies appear to be largely symmetric for all regions meaning they tend to be larger for *extreme* quantiles rather than for *central* quantiles. This potentially indicates that these are related to differences of the spreads of the decadal prediction and the climate projection ensembles.

Figure 3a synthesizes the main results found for concatenating SAT predictions after forecast year 10 with climate projections. Based on metric M1 it is found that for all regions concatenating both datasets is generally unproblematic for *central* quantiles of the MME, which is partly related to the bias correction performed. Furthermore, inconsistencies for *moderate* and *extreme* quantiles are small for most of the American continent, eastern Asia and Australia. Elsewhere, inconsistencies tend to be larger for these quantiles, being most pronounced for CGI, NAT, NEU, and GLOB.

In principle, the climate projection ensemble could be concatenated to each forecast year from 1 to 10 from the predictions. The effect using shorter lead times is shown in Figure S2 of Supporting Information S1. As expected, inconsistencies for all quantiles tend to decrease with longer lead times. For example, over the Amazon and South East Asia regions inconsistencies are largest when using forecast years 1 to about 3, whereas for forecast years 4–10 metric M1 indicates similarly low levels of inconsistencies in the increment distribution means. For NEU and GLOB, inconsistencies decrease with lead time as well, but they are still apparent at forecast year 10 in line with the previous results. Results for all regions for forecast year 2 and 7 are shown in Figure S3/S4 of Supporting Information S1. When concatenating after forecast year 2, inconsistencies are more pronounced over the tropics compared to the extratropics (with the exception of CGI, NAT, NEU). This might be partly related to the fact that the tropical regions are linked to sea surface temperatures (SSTs) over the tropical Pacific Ocean, which are to some extent predictable beyond annual time scales by current decadal prediction systems (e.g., Befort et al., 2021). Inconsistencies introduced when concatenating after forecast year 7 are similar to those when using forecast year 10. Figure S5 in Supporting Information S1 illustrates for each region the first forecast year at which predictions and projections can be concatenated without introducing significant inconsistencies based on metric M4.

4.2. Testing the Effect of Calibration & Model Weighting on the Concatenation of Predictions and Projections

Our results suggest that the concatenation inconsistencies are primarily due to large differences in the predictions and projections ensemble spreads. To examine this further, we calibrated the MMEs using the variance inflation (VINF) method which scales the ensemble spread and signal using observed variability (see Doblas-Reyes et al., 2005; O'Reilly et al., 2020, and Supporting Information S1 for further information). VINF is applied to the climate projection MME from 1970 to 2014 and separately for each forecast year to the decadal prediction MME (using leave-one-out cross-validation, Doblas-Reyes et al. (2005); Gangstø et al. (2013)). An example of the VINF calibration is shown in Figure 1c. Compared to the uncalibrated time series, the calibrated time series have a smaller ensemble spread, most notably for the projections. The inconsistencies between the prediction in 1985

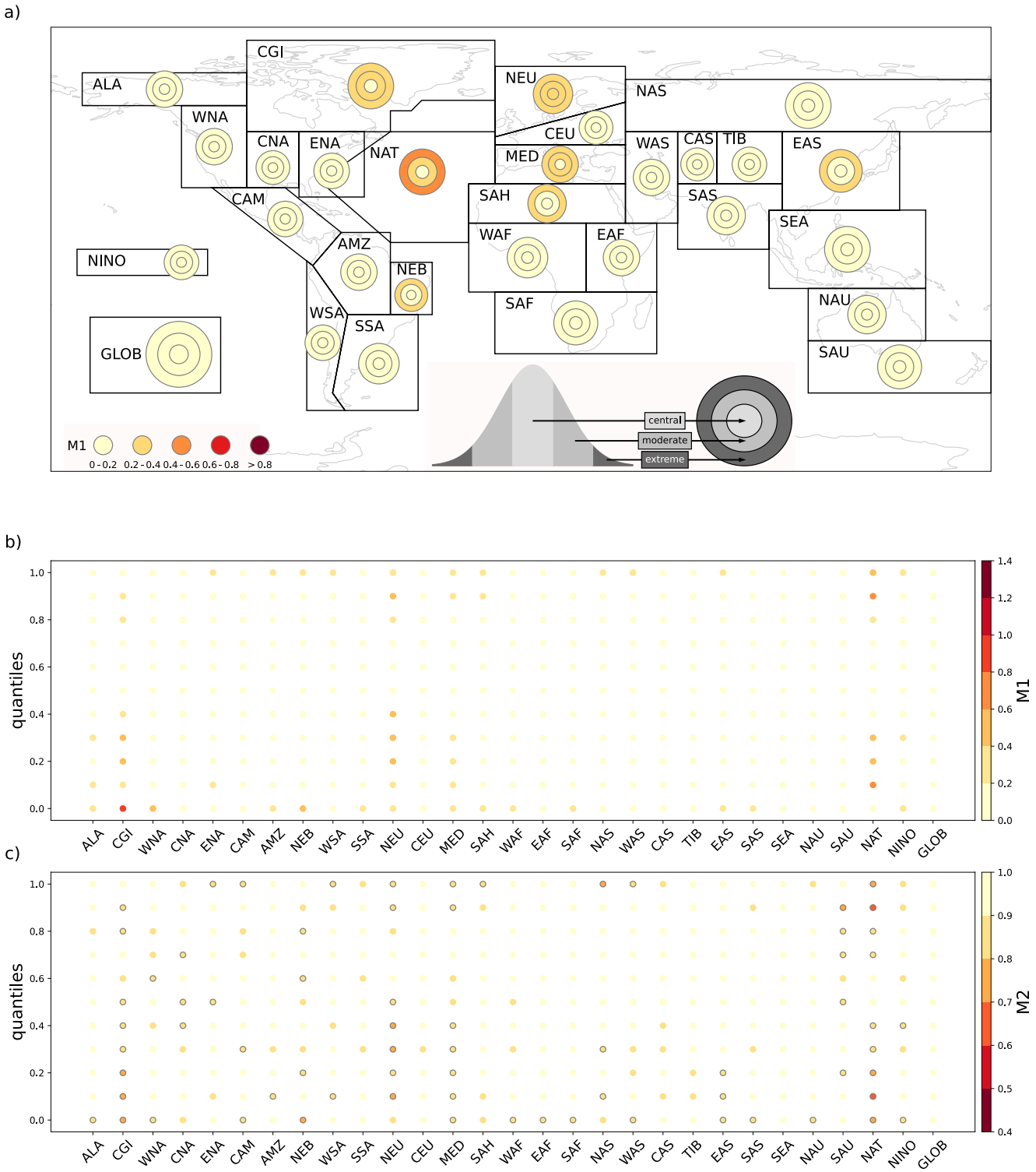

Figure 4. As Figure 3 but for calibrated decadal predictions and calibrated climate projections.

(forecast year 10) and the climate projection in 1986 are smaller than the concatenated uncalibrated data sets (i.e., Figure 1b) and this seems particularly pronounced for more extreme quantiles of the distribution.

Results from the concatenation of calibrated decadal predictions and calibrated climate projections for all regions are shown in Figure 4b. Compared to uncalibrated data (i.e., Figure 3b) the magnitude of inconsistencies in M1

and M2 are reduced, especially for GLOB, WAS and TIB. Smaller inconsistencies are also found for both metrics for NAT, CGI and NEU. However, even though M2 indicates that the overlap is large for most regions/quantiles (and increased compared to uncalibrated data), significant inconsistencies are still present, especially for NAT, CGI and NEU. The inability of the VINF method to entirely eliminate inconsistencies could be linked to increased predictability from initialization but might also be related to non-stationary prediction and projection ensemble spreads. In line with results for M1/M2, M3 and M4 indicate fewer significant inconsistencies between both increment distributions, particularly for *moderate* and *extreme* quantiles (See Figure S6 in Supporting Information S1). Overall, the positive effect of calibration on the consistency of temporal concatenation for most regions indicates that most of the inconsistencies encountered when concatenating predictions and projections arise from differences in the MME spreads. However, the application of the VINF method is not a panacea, as significant inconsistencies remain for some regions and quantiles, especially for CGI, NAT, NEU and MED (Figure 4a).

In addition to the VINF method we also test the potential of a weighting approach that assigns weights to each model based on its ability to capture past observed variability (e.g., Knutti et al., 2017). This model-weighting approach has been shown to better quantify and ultimately reduce model uncertainty in projections of mid-to end-of-century future changes on regional and global scales (Brunner et al., 2019, 2020; Lorenz et al., 2018). Here, we apply this model weighting scheme, for the first time, to initialized decadal predictions. Due to the small number of models available the method can not be applied with the typically used model-as-truth test which requires at least several tens of models (see Supporting Information S1 for further information). This test, in the standard approach, regulates the strength of the weighting (Brunner et al., 2020). Here, we use a simplified version of the weighting method that just calculates the weights as the normalized inverse distance of the models to the observations (see Supporting Information S1 for more details on the calculation of the distances).

For decadal predictions the weights are generally higher for shorter lead-times as can be expected due to the information from the initialization (see Figure S7 in Supporting Information S1 for a distribution of weights for GLOB). Considerable differences in the weights between forecast years 8–10 and the projections for several models are found, also indicating potential problems for concatenating. Results for average SAT over all regions when concatenating weighted predictions and weighted projections are shown in Figure S8 of Supporting Information S1. These results suggest that the weighting has only minor influence on inconsistencies introduced compared to using bias-corrected data only (Figures 3b & S1 in Supporting Information S1). This limited effect of the weighting is most probably related to the small number of models, which all receive relatively similar weights.

5. Summary & Discussion

In this study we have assessed the possibility to provide meaningful seamless climate information beyond decadal time scales by concatenating (bias-corrected) initialized decadal predictions and uninitialized climate projections from CMIP6. To analyze to what extent this simple concatenation introduces significant inconsistencies, the decadal predictions are concatenated to the climate projection in the following year, for all years between 1970 and 2013. This is done for near-surface air temperatures over the 26 SREX regions, in addition to the NAT, the NINO3.4 region and globally averaged, as well as for 11 different quantiles (ranging from 0 to 1 in 0.1 steps) in order to evaluate the full distribution. Four different metrics designed to assess the level of inconsistencies introduced when concatenating predictions and projections are used.

For the bias-corrected decadal prediction and climate projection MMEs it is found that for most regions the simple concatenation after forecast year 10 is not problematic for *central* quantiles close to the MME median, which is to a large extent attributed to the bias correction applied. However, significant inconsistencies are found for *moderate* and *extreme* quantiles and over some regions, most prominently over the NAT, CGI and NEU. This suggests that the simple bias correction commonly applied to predictions and projections does not entirely eliminate inconsistencies between prediction at forecast year 10 and projections for subsequent years, implying that merging information from both types of experiments is not trivial.

Results for most regions suggest that inconsistencies are caused by differences in the prediction and projection ensemble spreads. This motivated the application of the variance-inflation (VINF) calibration technique to both systems as this technique scales the signal and ensemble spread to obtain a reliable (calibrated) ensemble. Overall, the application of the VINF method mostly reduces inconsistencies, especially for *moderate* and *extreme*

quantiles, compared to using bias-corrected data only. The fact the there are fewer inconsistencies when concatenating the calibrated prediction and projection MMEs indicates similar levels of skill of prediction in forecast year 10 and projections as the ensemble spread of the calibrated ensemble is determined by the correlation between ensemble mean and observations (see Supporting Information S1). However, the VINF method alone is not sufficient to eliminate all significant inconsistencies, for example, over CGI, NAT and NEU. Thus, while results based on the VINF method confirm that differences in spread drive some of the inconsistencies, other factors, for example, differences in skill or non-stationary ensemble spreads might also contribute to inconsistencies. Further research on how to apply calibration methods to facilitate temporal concatenation of predictions and projections is necessary.

The impact of a second approach, using a weighting method based on model performance, on inconsistencies when concatenating predictions and projections has been tested. It is found that results for weighted predictions and projections are very similar to those for the equally weighted MMEs. The main reason for this is the small number of models that provide both decadal predictions and projections which meant that the weighting method could not be applied in an optimal setup. Further research is necessary to investigate how to potentially overcome this issue to make this weighting approach applicable to prediction ensembles consisting of few individual models.

Inconsistencies are most pronounced over the NAT, CGI, NEU and globally averaged (GLOB). Over the NAT region (as well as neighboring regions: CGI and NEU as no land-sea masking has been applied), might be partly linked to erroneous SST hindcasts over parts of this region due to initialization as found for CanESM5 (Sospedra-Alfonso et al., 2021). Initialization shock and model drift might also be a reason for differences between decadal predictions and climate projections as found for EC-Earth (Bilbao et al., 2021). Another explanation for the inconsistencies after forecast year 10 could be differences in skill between decadal predictions and climate projections. Based on CMIP5 data it has been found that decadal predictions are more skillful compared to climate projections in predicting SAT over the North Atlantic Subpolar Gyre region (SPG) (see Yeager & Robson, 2017). However, the skill difference for SAT between initialized decadal predictions and climate projections is smaller in CMIP6, mainly due to an improvement of the climate projections in predicting SPG SATs (Borchert, Menary, et al., 2021).

However, if inconsistencies are linked to added value from initialization, other methods, for example, constraining projections using decadal predictions might be more applicable (e.g., Befort et al., 2020; Mahmood et al., 2021). In the case that inconsistencies are linked to higher skill of climate projections over decadal predictions, neither simple concatenation nor constraining approaches are likely to provide robust seamless climate information beyond the 10 year time scale.

The explicit use of decadal prediction data for the first 10 years (and climate projections thereafter) is the main advantage of the proposed concatenation framework. While further work is needed to reconcile remaining inconsistencies for the more extreme percentiles, results indicate that the proposed approach can already be applied to provide seamless climate information for expected median changes for some regions. In such cases the approach allows a potential user to benefit from initial condition information on short time scales while also being able to draw on climate projection for longer time scales.

Data Availability Statement

In this study, CMIP6 dcppA-hindcast as well as historical climate projection data is used from the following models: CanESM5 (Sospedra-Alfonso et al., 2021), EC-Earth3 (Bilbao et al., 2021; Haarsma et al., 2020), IPSL-CM6A-LR (Boucher et al., 2020), MIROC6 (Kataoka et al., 2020; Tatebe et al., 2019), HadGEM3-GC31-MM (Andrews et al., 2020; Williams et al., 2018), MPI-ESM1-2-HR (Mauritsen et al., 2019; Müller et al., 2018) and NorCPM1 (Bethke et al., 2021). All datasets can be accessed through the Earth System Grid federation https://esgf-node.llnl.gov/search/cmip6/.

Acknowledgments
All authors were supported by the European Climate Prediction project funded by the European Commission's Horizon 2020 programme, Grant Agreement number 776613. L.F. Borchert also acknowledges support from the ANR-TREMPLIN ERC Project HARMONY, Grant Agreement Number ANR-20-ERC9-0001. C.H. O'Reilly was supported by a Royal Society University Research Fellowship. The authors thank Klaus Pankatz and Wolfgang Muller for providing additional members (6–10) of MPI-ESM1-2-HR decadal hindcasts, which at the start of this study hadn't been available through the esgf. The authors thank Francisco Doblas-Reyes and one anonymous reviewer for their helpful comments.

References

Andrews, M. B., Ridley, J. K., Wood, R. A., Andrews, T., Blockley, E. W., Booth, B., et al. (2020). Historical simulations with HadGEM3-GC3.1 for CMIP6. *Journal of Advances in Modeling Earth Systems*, *12*(6), e2019MS001995. https://doi.org/10.1029/2019MS001995

Befort, D. J., O'Reilly, C. H., & Weisheimer, A. (2020). Constraining projections using decadal predictions. *Geophysical Research Letters*, *47*(18), e2020GL087900. https://doi.org/10.1029/2020gl087900

Befort, D. J., O'Reilly, C. H., & Weisheimer, A. (2021). Representing model uncertainty in multiannual predictions. *Geophysical Research Letters*, *48*(5), e2020GL090059. https://doi.org/10.1029/2020gl090059

Bethke, I., Wang, Y., Counillon, F., Keenlyside, N., Kimmritz, M., Fransner, F., et al. (2021). NorCPM1 and its contribution to CMIP6 DCPP. *Geoscientific Model Development*, *14*(11), 7073–7116. https://doi.org/10.5194/gmd-14-7073-2021

Bilbao, R., Wild, S., Ortega, P., Acosta-Navarro, J., Arsouze, T., Bretonnière, P.-A., et al. (2021). Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth. *Earth System Dynamics*, *12*(1), 173–196. https://doi.org/10.5194/esd-12-173-2021

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., et al. (2016). The decadal climate prediction project (DCPP) contribution to CMIP6. *Geoscientific Model Development*, *9*(10), 3751–3777. https://doi.org/10.5194/gmd-9-3751-2016

Borchert, L. F., Koul, V., Menary, M. B., Befort, D. J., Swingedouw, D., Sgubin, G., & Mignot, J. (2021). Skillful decadal prediction of unforced southern European summer temperature variations. *Environmental Research Letters*, *16*(10), 104017. https://doi.org/10.1088/1748-9326/ac20f5

Borchert, L. F., Menary, M. B., Swingedouw, D., Sgubin, G., Hermanson, L., & Mignot, J. (2021). Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. *Geophysical Research Letters*, *48*(3), e2020GL091307. https://doi.org/10.1029/2020gl091307

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2019MS002010. https://doi.org/10.1029/2019ms002010

Branstator, G., & Teng, H. (2010). Two limits of initial-value decadal predictability in a CGCM. *Journal of Climate*, *23*(23), 6292–6311. https://doi.org/10.1175/2010jcli3678.1

Brunner, L., Lorenz, R., Zumwald, M., & Knutti, R. (2019). Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environmental Research Letters*, *14*(12), 124010. https://doi.org/10.1088/1748-9326/ab492f

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. *Earth System Dynamics*, *11*(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

Doblas-Reyes, F. J., Andreu-Burillo, I., Chikamoto, Y., García-Serrano, J., Guemas, V., Kimoto, M., et al. (2013). Initialized near-term regional climate change prediction. *Nature Communications*, *4*(1), 1715. https://doi.org/10.1038/ncomms2704

Doblas-Reyes, F. J., Hagedorn, R., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting — II. Calibration and combination. *Tellus A: Dynamic Meteorology and Oceanography*, *57*(3), 234–252. https://doi.org/10.3402/tellusa.v57i3.14658

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Field, C., Barros, V., Stocker, T., & Dahe, Q. (Eds.) (2012). *Managing the Risks of extreme Events and Disasters to advance climate change adaptation: Special Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781139177245

Gangstø, R., Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2013). Methodological aspects of the validation of decadal predictions. *Climate Research*, *55*(3), 181–200. https://doi.org/10.3354/cr01135

Haarsma, R., Acosta, M., Bakhshi, R., Bretonnière, P.-A., Caron, L.-P., Castrillo, M., et al. (2020). HighResMIP versions of EC-Earth: EC-Earth3P and EC-Earth3P-HR – Description, model computational performance and basic validation. *Geoscientific Model Development*, *13*(8), 3507–3527. https://doi.org/10.5194/gmd-13-3507-2020

Hewitt, C. D., & Lowe, J. A. (2018). Toward a European climate prediction system. *Bulletin of the American Meteorological Society*, *99*(10), 1997–2001. https://doi.org/10.1175/bams-d-18-0022.1

Iturbide, M., Gutiérrez, J. M., Alves, L. M., Bedia, J., Cerezo-Mota, R., Cimadevilla, E., et al. (2020). An update of IPCC climate reference regions for subcontinental analysis of climate model data: Definition and aggregated datasets. *Earth System Science Data*, *12*(4), 2959–2970. https://doi.org/10.5194/essd-12-2959-2020

Jones, R., & Mearns, L. O. (2005). Assessing future climate risks. In B. Lim, E. Spanger-Siegfried, I. Burton, E. Malone & S. Huq (Eds.), *Adaptation policy frameworks for climate change: Developing strategies, policies and measures* (pp. 119–143). Cambridge University Press.

Kataoka, T., Tatebe, H., Koyama, H., Mochizuki, T., Ogochi, K., Naoe, H., et al. (2020). Seasonal to decadal predictions with MIROC6: Description and basic evaluation. *Journal of Advances in Modeling Earth Systems*, *12*(12), e2019MS002035. https://doi.org/10.1029/2019ms002035

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. *Geophysical Research Letters*, *44*(4), 1909–1918. https://doi.org/10.1002/2016gl072012

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and caveats of weighting climate models for summer maximum temperature projections over North America. *Journal of Geophysical Research: Atmospheres*, *123*(9), 4509–4526. https://doi.org/10.1029/2017jd027992

Mahmood, R., Donat, M. G., Ortega, P., Doblas-Reyes, F. J., & Ruprich-Robert, Y. (2021). Constraining decadal variability yields skillful projections of near-term climate change. *Geophysical Research Letters*, *48*(24), e2021GL094915. https://doi.org/10.1029/2021gl094915

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO2. *Journal of Advances in Modeling Earth Systems*, *11*(4), 998–1038. https://doi.org/10.1029/2018ms001400

Meehl, G. A., Goddard, L., Boer, G., Burgman, R., Branstator, G., Cassou, C., et al. (2014). Decadal climate prediction: An update from the trenches. *Bulletin of the American Meteorological Society*, *95*(2), 243–267. https://doi.org/10.1175/bams-d-12-00241.1

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., et al. (2009). Decadal prediction: Can it be skillful? *Bulletin of the American Meteorological Society*, *90*(10), 1467–1486. https://doi.org/10.1175/2009bams2778.1

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., et al. (2018). A higher-resolution version of the Max Planck Institute Earth System model (MPI-ESM1.2-HR). *Journal of Advances in Modeling Earth Systems*, *10*(7), 1383–1413. https://doi.org/10.1029/2017ms001217

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

O'Reilly, C. H., Befort, D. J., & Weisheimer, A. (2020). Calibrating large-ensemble European climate projections using observational data. *Earth System Dynamics*, *11*(4), 1033–1049. https://doi.org/10.5194/esd-11-1033-2020

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., et al. (2019). Robust skill of decadal climate predictions. *npj Climate and Atmospheric Science*, *2*, 13. https://doi.org/10.1038/s41612-019-0071-y

Sospedra-Alfonso, R., Merryfield, W. J., Boer, G. J., Kharin, V. V., Lee, W.-S., Seiler, C., & Christian, J. R. (2021). Decadal climate predictions with the Canadian Earth system model version 5 (CanESM5). *Geoscientific Model Development*, *14*(11), 6863–6891. https://doi.org/10.5194/gmd-14-6863-2021

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., et al. (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6. *Geoscientific Model Development*, *12*(7), 2727–2765. https://doi.org/10.5194/gmd-12-2727-2019

Williams, K. D., Copsey, D., Blockley, E. W., Bodas-Salcedo, A., Calvert, D., Comer, R., et al. (2018). The Met Office Global Coupled Model 3.0 and 3.1 (GC3.0 and GC3.1) Configurations. *Journal of Advances in Modeling Earth Systems*, *10*(2), 357–380. https://doi.org/10.1002/2017ms001115

Yeager, S. G., & Robson, J. I. (2017). Recent progress in understanding and predicting Atlantic decadal climate variability. *Current Climate Change Reports*, *3*, 112–127. https://doi.org/10.1007/s40641-017-0064-z

References From the Supporting Information

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., et al. (2019). ESD Reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth System Dynamics*, *10*(1), 91–105. https://doi.org/10.5194/esd-10-91-2019

Amos, M., Young, P. J., Hosking, J. S., Lamarque, J.-F., Abraham, N. L., Akiyoshi, H., et al. (2020). Projecting ozone hole recovery using an ensemble of chemistry–climate models weighted by model performance and independence. *Atmospheric Chemistry and Physics*, *20*(16), 9961–9977. https://doi.org/10.5194/acp-20-9961-2020

Fortin, V., Abaza, M., Anctil, F., & Turcotte, R. (2014). Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology*, *15*(4), 1708–1713. https://doi.org/10.1175/jhm-d-14-0008.1

Massonnet, F., Fichefet, T., Goosse, H., Bitz, C. M., Philippon-Berthier, G., Holland, M. M., & Barriat, P.-Y. (2012). Constraining projections of summer Arctic sea ice. *The Cryosphere*, *6*(6), 1383–1394. https://doi.org/10.5194/tc-6-1383-2012

Merrifield, A. L., Brunner, L., Lorenz, R., Medhaug, I., & Knutti, R. (2020). An investigation of weighting schemes suitable for incorporating large ensembles into multi-model ensembles. *Earth System Dynamics*, *11*(3), 807–834. https://doi.org/10.5194/esd-11-807-2020

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al. (2021). An updated assessment of near-surface temperature change from 1850: The HadCRUT5 data zet. *Journal of Geophysical Research: Atmospheres*, *126*(3), e2019JD032361. https://doi.org/10.1029/2019jd032361