HORIZON 2020
THEME SC5-2017

European Climate Prediction system

(**Grant Agreement 776613**)

European Climate Prediction system (EUCP)

Deliverable D5.3

Development of methods to merge probabilistic forecasts based on global initialised and non-initialised predictions to provide a seamless prediction system

Deliverable Title	*Development of methods to merge probabilistic forecasts based on global initialised and non-initialised predictions to provide a seamless prediction system*	
Brief Description	*This report summarises the pioneering work to develop methods to combine information from decadal predictions and climate projections, towards providing seamless climate information of highest possible skill for the next multiple decades.*	
WP number	WP5	
Lead Beneficiary	*BSC*	
Contributors	*Rashed Mahmood (BSC), Markus Donat (BSC), Francisco J. Doblas-Reyes (BSC), Daniel Befort (UOXF), Antje Weisheimer (UOXF), Christopher O'Reilly (UOXF), Lukas Brunner (ETHZ), Leonard Borchert (IPSL), Dominic Matte (UCPH), Jens H. Christensen (UCPH), Andrew Ballinger (UEd), Gabriele Hegerl (UEd)*	
Creation Date Version Number Version Date	*13/04/2021* *[v3]* *19/10/2021*	
Deliverable Due Date Actual Delivery Date	*31/10/2021* *20/10/2021*	
Nature of the Deliverable	R	*R – Report*
		P - Prototype
		D - Demonstrator
		O - Other
Dissemination Level/ Audience	x	*PU - Public*
		PP - Restricted to other programme participants, including the Commission services

		RE - Restricted to a group specified by the consortium, including the Commission services *CO - Confidential, only for members of the consortium, including the Commission services*	

Version	**Date**	**Modified by**	**Comments**
V1	13/04/2021	Rashed Mahmood	Initial outline and overview of already finished and planned work until submission
V2	10/08/2021	Rashed Mahmood, Markus Donat, with input from all partners	Finalising a first complete draft of the report for internal review
V3	11/10/2021	Rashed Mahmood, Markus Donat	Addressing the internal reviewer comments

Table of Contents

1. Executive Summary

This deliverable presents research towards providing seamless climate information for the next multiple decades, by combining information from initialised decadal predictions and longer-term climate projections. This is a novel field of research, for the first time attempted within the EUCP project. The work presented in this deliverable explores to what extent projections can be improved beyond the first 10 years by constraining large ensembles of climate projections according to their agreement with decadal predictions, and evaluates differences in the distribution functions between decadal predictions and projections to understand key inconsistencies that need to be resolved in approaches that would concatenate decadal predictions and projections.

To improve climate projections for the near-term future but beyond the first 10 years covered by decadal predictions, different approaches have been developed that constrain the projections based on their agreement with decadal predictions. A first pioneering study (Section 3.1.1 and published in Befort et al., 2020) illustrated the concept and feasibility of such variability constraints for the North Atlantic Subpolar Gyre region, where decadal predictions typically show the largest added value over projections. This study demonstrated added value in the constrained projections of temperatures in the North Atlantic region beyond the time period covered by decadal predictions, up to 15 years after initialisation of the decadal predictions.

Building on this initial demonstration of potential benefit, other constraining methods have been developed that take into account global patterns of climate variability (documented in Sections 3.1.2 and 3.1.3). Applied to a large single-model ensemble (Section 3.1.2), these methods are able to improve climate projections in larger regions of the globe including the North Atlantic, the tropical Pacific and land regions in Eurasia and Africa. In particular, added value is demonstrated for temperature projections of the following 20 years in several of these regions. Applying the constraint to predict the near-term future until 2035, a tendency towards warmer conditions in the North Atlantic is found, and increased warming of summer temperature in related regions such as the Sahel and Southern Asia. In particular, the application to very large multi-model ensembles (with more than 200 members, Section 3.1.3) demonstrates the strong potential of this global constraining approach. It is shown to improve 20-year temperature projections in large areas of the world.

Other ongoing work focuses on the differences between decadal predictions and projections, which may complicate the merging of actual output data from both data sources. In particular, the work demonstrates significant inconsistencies in the distributions from both data sources, which would introduce inhomogeneities when ‘stitching’ decadal predictions and projections together after year 10 of the decadal predictions. Statistical corrections such as calibration or weighting are being explored in ongoing work and evaluated for their efficacy in minimising such inconsistencies. The deliverable also briefly discusses how inconsistencies in the representation of low-frequency NAO variability in models in projection mode may bias observationally constrained projections, something that could be addressed by selection from very large ensembles.

Finally, we have explored a way to merge climate projections and predictions using a simple scaling approach (i.e. scaling the grid-point change by the global average 2-meter temperature change). As the time lead increases, we find that the overall spatial pattern shows some correlation with the scaled pattern from climate projections. This work also highlighted some key regions such as the

North Atlantic subpolar gyre and Antarctica that remain quite different when comparing climate predictions to climate projections. This suggests that the pattern scaling may not be an optimal approach for merging climate projections with initialized predictions.

This deliverable documents the first efforts to merge information from decadal predictions and longer-term climate projections. It indicates strong opportunities for improving climate information for the next multiple decades, and also highlights some challenges that remain to be addressed in future work. Based on these first studies, it appears worthwhile to continue and intensify the research to combine the information from decadal predictions and projections to further improve estimates of near-term climate in the next multiple decades. Particular focus could be on improving projections of extreme climate events, which would be of large relevance to support the development and implementation of targeted adaptation strategies, and improvement in climate model simulated low-frequency variability in atmospheric circulation modes

2. Project Objectives

WITH THIS DELIVERABLE, EUCP HAS CONTRIBUTED TO THE ACHIEVEMENT OF THE FOLLOWING OBJECTIVES (DESCRIPTION OF ACTION, SECTION 1.1):

No.	Objective	Yes	No
1	Estimate relative merits of initialised and non-initialised methods in global simulations during the first 10 years when data from both approaches exist.	x	
2	Estimation of added value for combined predictions in terms of merge point and the merged forecast for different variables and regions compared to non-initialised forced-only simulations.	x	
3	Test methods traditionally used to quantify uncertainty and combine different members in projections and assess the forecast quality of the resulting predictions.	x	

3. Detailed Report

Introduction

Accurate, reliable and actionable information about near-term future climate (up to 30–40 years ahead) is important for policy making and planning ahead to minimize the potential impacts of the ongoing climate variability and change on various sectors including human lives, livestocks, agriculture, ecosystems and other large-scale infrastructures. Such future climate information can be obtained either from transient climate model projections (e.g. Eyring et al., 2016) or initialized decadal predictions (Boer et al., 2016). The projection simulations are integrated for a century or more and simulate the evolution of future climate based on different scenarios of future greenhouse

gas concentrations, assuming different pathways of socio-economic and political development. These projection simulations are, however, strongly affected by uncertainties arising from internal climate variability, which can reduce their usefulness for making decisions about adapting to near-term climate change in the next few decades. Decadal predictions, on the other hand, are initialized towards the observed climate state, aligning the simulated and observed phases of climate variability, and thereby reducing the uncertainty related to internal variability. However, decadal predictions are computationally very expensive (they require about 10 times as many computing resources as climate projections at same ensemble size as every year would be simulated by 10 different simulations started 1–10 years before the year of interest), and the current decadal prediction systems are typically integrated for 10 years into the future and thus leave a gap towards providing climate information on timescales beyond a decade.

Within the framework of the EU Horizon 2020 EUCP project we aim to address this gap by developing new methodologies to combine information from climate projections and the initialized decadal predictions to provide seamless climate information covering the next few decades. This deliverable report summarises the progress made at different participating institutions in achieving this goal.

Combining future climate information from different sources, initialised decadal predictions and projections, is a new field of research, and therefore the approaches presented here are pioneering work towards developing temporally seamless climate information that optimises the skill on multiple (decadal to multi-decadal) timescales. The works reported in this deliverable exploit single-model large ensembles from the Community Earth System Model (CESM) based projection simulations (Kay et al., 2015) and decadal predictions (Yeager et al., 2018) and also the large multi-model ensembles from CMIP5 and CMIP6 simulations. UOXF has recently published (Befort et al., 2020) a methodology for constraining projections using decadal predictions in the North Atlantic. BSC has submitted a manuscript for peer-reviewed publication (Mahmood et al., 2021, in revision) that takes large-scale to global variability patterns into account for constraining projection simulations. Further publications are in preparation.

The work covered by this deliverable is a part of EUCP WP5. Earlier results of the work presented in this Deliverable report were also summarized in the milestone MS20 *"Developing tools to reduce uncertainty of climate change estimates for the coming decades"*.

3.1 Constraining projections using decadal predictions *(University of Oxford, Barcelona Supercomputing Center)*

3.1.1 Constraining projections using decadal predictions in the Subpolar North Atlantic *(University of Oxford)*

At University of Oxford (UOXF), a framework to constrain uninitialized climate projections using initialized decadal predictions has been developed and applied to model simulations from the CMIP5 archive. These results were published in Befort et al., 2020.

The main difference between decadal predictions and climate projections is that the latter includes only external forcings as e.g., varying greenhouse gas concentrations whereas the former is

additionally initialized at the start of its integration using observational data. The correct representation of the initial state is crucial for successful shorter-term predictions (e.g. seasonal forecasts) but for long-term projections the correct representation of external forcings is of larger importance.

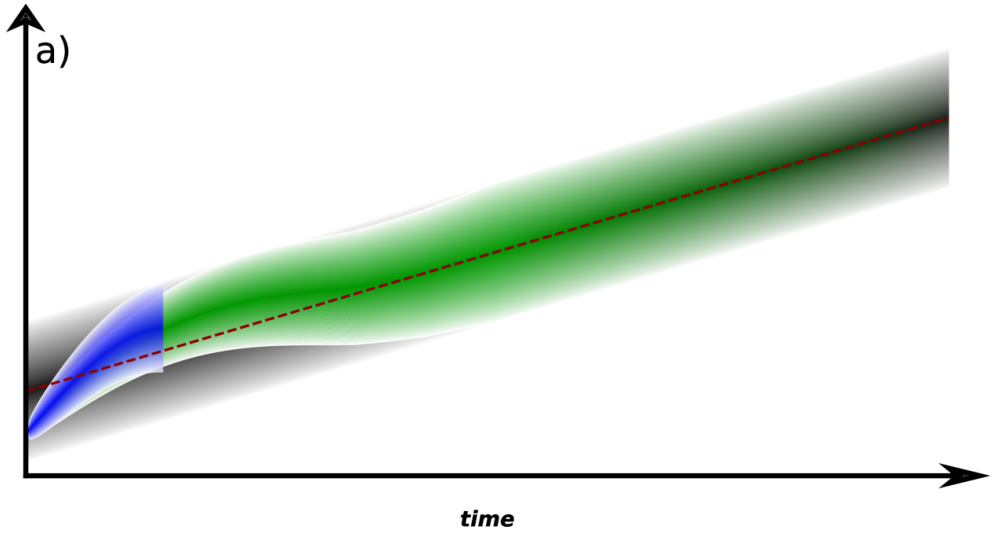

Figure 1: *Schematic illustration of framework used to constrain uninitialized projections (gray) using decadal predictions (blue). The selected ensemble based on proximity to the decadal pre*diction *ensemble mean are indicated in green. Shaded areas indicate the range of the respective ensemble. Figure adapted from Befort et al. (2020) their Figure 1a.*

A schematic demonstrating the framework in Befort et al. (2020) is shown in Figure 1. Here the initialized decadal prediction ensemble (blue) differs from the climate projections (grey) at the start of the integration. Over time the spread of the decadal prediction increases but even after 10 years the distributions of the projections and predictions differ significantly. Assuming that the initialized decadal prediction is more skillful than the climate projection, it is sensible to assess to what extent the prediction can be used to constrain the projection. Here, the constraining is based on a sub-selection of those climate projections which are closest to the decadal prediction ensemble average over the first 10 forecast years. The closest climate projections are chosen separately for each start date (for which decadal predictions are available). As the selected climate projections are available also after 10 years when the decadal predictions are unavailable, the constrained ensemble (green in Figure 1) provides consistent seamless information beyond decadal timescales.

The main research question is whether such a constraining method can provide more skillful climate information compared to the unconstrained (overall) climate projection ensemble. For this, a basic necessity is the existence of added value of initialization for preferably long lead times. Variables and regions for which decadal predictions are more skillful than projections up to ten years are, however, rare. One exception in CMIP5 are surface temperatures over the North Atlantic Gyre region (GYRE). Figure 2 shows anomaly correlation coefficients (ACC) and root-mean-square-error (RMSE) for climate projections and decadal predictions for different forecast year ranges. It is found that while ACC is only slightly larger for decadal predictions for forecast years 6-10, RMSE is lower for forecast times up to ten years. This makes the GYRE region the perfect testbed for the proposed framework.

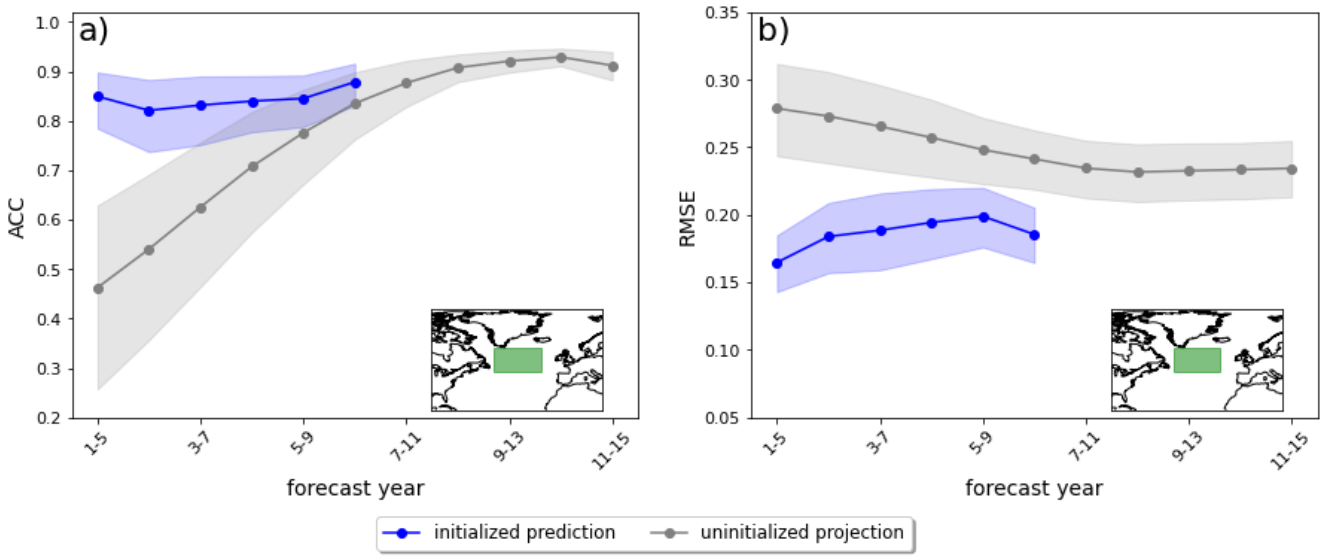

Figure 2: *(a) Anomaly correlation coefficient (ACC) between observations and uninitialized projection (gray)/decadal predictions (blue) for surface temperatures over the North Atlantic Gyre region for a given forecast year range (5-year averages). Shaded areas show 10–90% confidence intervals (based on a 10,000 sample bootstrap). (b) Same for root-mean-square error (RMSE). Figure adapted from Befort et al. (2020) their Figure 2c and d.*

The method has been applied to near-surface temperatures (SAT) over the GYRE region. Prior, lead-time dependent biases calculated over the baseline period (1970-2006) are removed from the decadal prediction ensembles, whereas for observations and climate projections the mean over this baseline period is removed. Next, for each decadal prediction start year between 1960 and 2000 those 35 climate projection members with the smallest mean absolute error (MAE) to the decadal prediction ensemble mean over the upcoming ten years are selected. The resulting skill (ACC, RMSE) for the decadal prediction ensemble as well as for the unconstrained and constrained climate projections is shown in Figure 3. It is found that the ACC for the first 10 years is similar for the decadal predictions and the constrained projections, which both show higher values than the unconstrained projections. After 10 years, ACC values are similar between constrained and unconstrained projections. This is less surprising given that skill at forecast years 6–10 is very similar for decadal predictions and climate projections. As discussed before, larger differences between decadal predictions and unconstrained projections are found for RMSE. Similar to the results for ACC, RMSE for the constrained ensemble is only slightly larger than for the decadal predictions. However, in contrast to ACC results RMSE is also significantly smaller than for the unconstrained projections up to 15 years ahead. This demonstrates that it is possible to obtain more skillful seamless climate information beyond decadal timescales using the proposed framework.

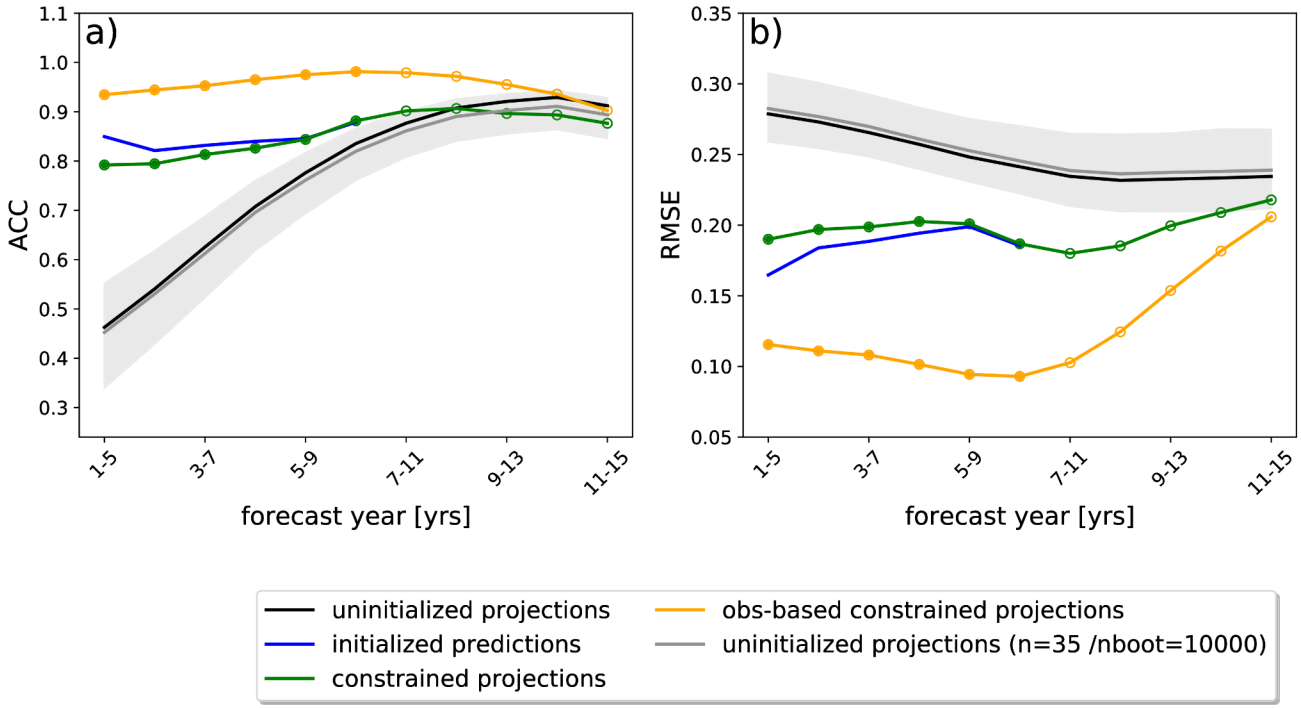

Figure 3: *(a) ACC and (b) RMSE for unconstrained projections (black), constrained projections using decadal ensemble mean (green), constrained projections using observations (orange), unconstrained projections with fixed ensemble size n = 35 (gray) and initialized predictions (blue) for annual mean surface air temperatures over the North Atlantic Gyre. Statistics are given for 5-year averages. The gray area indicates the 10–90% confidence intervals (based on a 10,000 sample bootstrap). Closed circles indicate those periods for which all forecast years have been used to constrain the respective ensemble, whereas open circles indicate those periods for which at least 1 year has not been used for constraining. Figure taken from Befort et al. (2020) their Figure 3.*

Besides constraining climate projections using the decadal prediction ensemble mean, it has been investigated to what extent skill can be increased when using observations instead (obs-based constrained projections). Please note that this approach is used to determine the upper limit of the method, presuming the feasibility of perfect decadal temperature predictions, and cannot be pursued in real-time as it would require knowledge of observational variability in the future. It is found that constraining using observations improves ACC skill up to about 13 years but afterwards ACC values are similar to the unconstrained projections. Similarly, RMSE for the obs-based constrained ensemble are smaller for the first 13 years but are similar at year 11–15. This suggests that even with a perfect decadal prediction system more skillful information through a constrained projection ensemble using the proposed framework can only be provided up to about 15 years for this region and variable. However, it should be noted that these results are based on a limited sample size (as decadal predictions in CMIP5 have been initialized between 1960 and 2000 only).

The framework has also been applied to the European region and the NINO3.4 region but the benefits are smaller compared to the GYRE region, which is most likely linked to the smaller added value found in decadal predictions over climate projections for both regions.

3.1.2 Constraining climate projections using initialised decadal predictions based on global anomaly patterns, using a large single-model ensemble *(Barcelona Supercomputing Center)*

At BSC we developed a novel method to constrain a large ensemble of projection simulations based on their agreement with anomaly patterns in selected fields of the initialized decadal predictions. The objective of the approach is to select those projection ensemble members more in phase with the climate variability of the observed climate. Since the projection simulations cover timescales up to centuries long, our approach can thus provide improved information, with reduced uncertainty from internal variability, for extended periods beyond the typical 10 year predictions of the current initialized prediction systems.

As a first test of this constraining approach we used climate model simulations from the National Center for Atmospheric Research (NCAR)'s Large Ensemble (LENS) simulations of the historical climate extended with the RCP8.5 scenario after 2005 (Kay et al., 2015; hereafter referred to as UNINIT40) and the initialised decadal predictions (Yeager et al., 2018: hereafter referred to as DPLE40). Both UNINIT40 and DPLE40 ensemble simulations are performed using the community earth system model (CESM). The DPLE40 hindcasts are run for 122 months starting in November of every year over the period 1954–2014.

The constraining procedure we developed is based on spatial sea surface temperature (SST) anomaly pattern correlations between DPLE40 ensemble mean and the individual members of the historical simulation (Figure 4). For each start date, we select the 10 historical members with the highest pattern correlations (hereafter referred to as "Best10"). We also select the 10 members with the lowest pattern correlations (referred to as "Worst10") to evaluate further the effectiveness of the constraint. These selected ensembles are used to construct 20–year mean retrospective forecasts in order to evaluate the skill of the constrained sub-ensemble in comparison to climate observations. The use of a 20–year window is somewhat arbitrary and is a compromise that matches the time spans often used by the Intergovernmental Panel on Climate Change (IPCC) to analyse projected climate changes (Collins et al., 2013), and is a first step to evaluate the added value for multi-decadal climate change estimates. However we emphasize that it is conceptually possible to make forecasts on longer timescales as long as the respective historical/projection simulations are available (keeping in mind that the added value from phasing in climate variability decreases with increasing lead time).

The selection of projection members can be achieved using a variety of different choices. The most important being (1) the region over which the SST anomalies are compared and (2) the forecast range considered from the initialised predictions for the selection. We used five selection domains (see Table 1) and eleven forecast periods (from first five months to ten years) to investigate some of the associated sensitivities. We also compare the Best10 ensemble against a distribution of an equivalent 10–member ensemble from the UNINIT40 for which the 10 members are always randomly sampled for each start date (hereafter referred to as "Random10"). Note that this

Random10 distribution is valid to assess the significance of all the Best10 ensembles, independent of the chosen spatial and temporal selection criteria. A detailed analysis of these sensitivities is given in Mahmood et al. (2021, in review) while a brief overview is provided in the following.

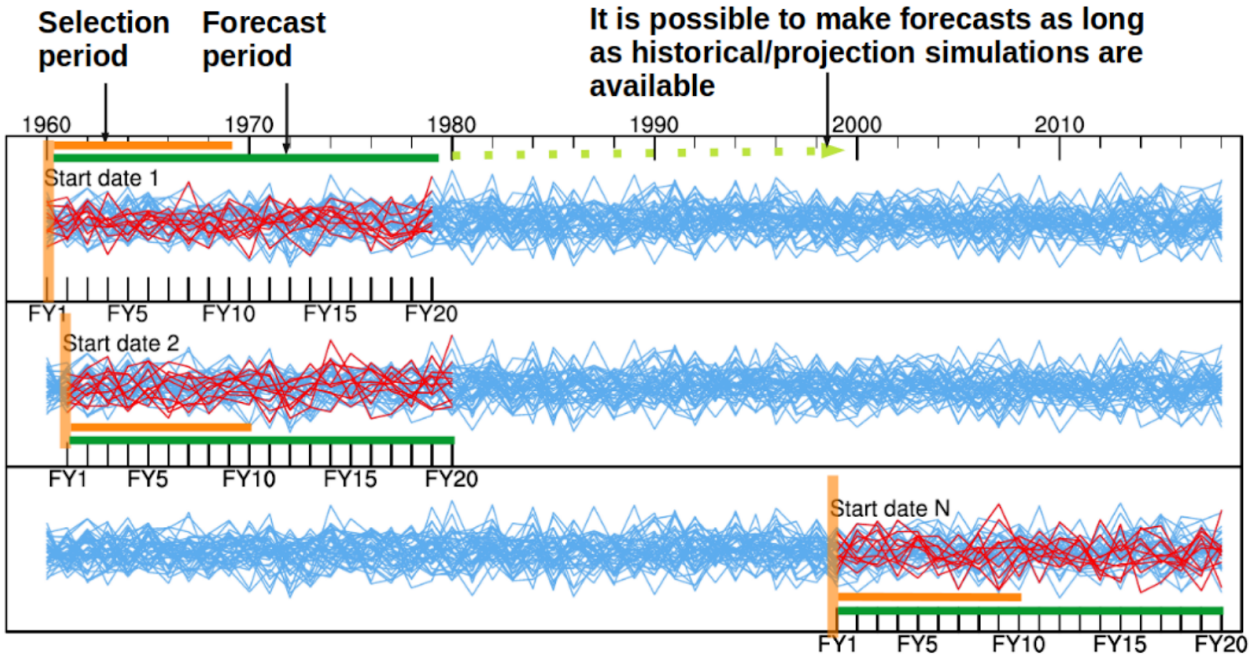

Figure 4: *Schematic diagram explaining the large ensemble subselection methodology (Idealized data). For each start date, SST anomaly of individual projection members is compared with the ensemble mean of SST anomalies of initialized prediction using area-weighted spatial pattern correlation. The selection period (shown as orange line) can be any time interval within the forecast range of the initialized prediction. The projection members are ranked based on the pattern correlation coefficients and the top N members (N can be any subset of the projection ensemble) are chosen as "BestN" for each start date. The temporal trajectories of these BestN members are then used to predict the climate over the 20 years after initialisation, as depicted by the green line. Note that by design these "BestN" members can be a different subset of the projection ensemble at each start date.*

Table 1 | SST regions used for selecting members.

Selection domain name	Latitude range	Longitude range
Global	All	All
NoPolar	60ºS–60ºN	All
Atl+Pac	25ºS–60ºN	120ºE–360ºE
Pac	60ºS–60ºN	140ºE–275ºE
NAtl	0–60ºN	280ºE–360ºE

We first evaluate the skill of the constrained ensemble in predicting three large-scale SST indices including global mean SST (GMSST), Atlantic Multidecadal Variability (AMV) and Interdecadal Pacific Oscillation (IPO). These constrained ensembles are evaluated against observations from the Met Office Hadley Center's sea ice and sea surface temperature (HadISSTv1.1; Rayner et al., 2003). Figure 5 shows the sensitivity of the anomaly correlation coefficient (ACC) to different selection time-intervals and SST regions. For GMSST and AMV the highest ACC skill of Best10 is achieved when the latter are defined based on 3–year or longer time averages. Similarly, the results are also sensitive to the selecting region as the highest ACC for GMSST is achieved when applying the constraints based on Global, NoPolar and Pac regions, whilst the AMV skill is maximum when the Best10 are defined using the NAtl region. The effectiveness of the constraint can also be seen from the low ACC values of the Worst10 sub-ensemble. We did not find any appreciable skill for IPO since it is also not well captured by initialised predictions.

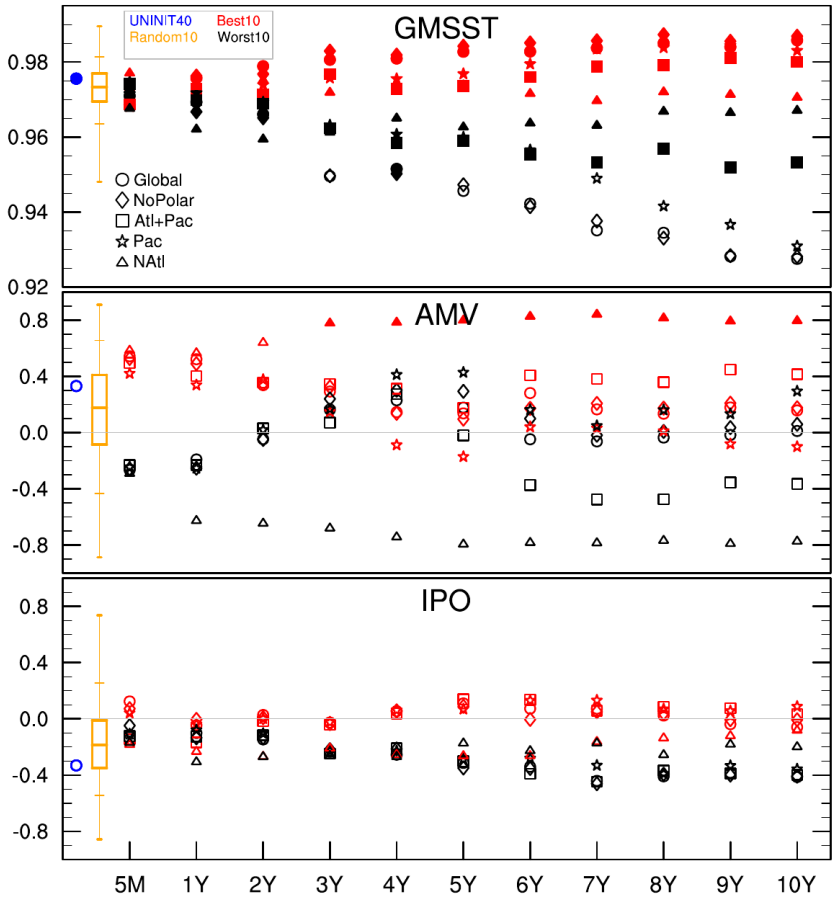

Figure 5: *ACC of large-scale SST indices for all time periods and domains used for sub-selection. Each marker represents the Best10 and Worst10 members chosen using different selection regions (cf. legend with markers in the top panel). The box-and-whisker plots represent the range of skill scores for 40,000 randomly selected 10-member ensemble means, while whiskers represent the minimum and maximum correlation and the horizontal line inside the box represents the median value. The lower and upper boundaries of the box represent the 25th and 75th percentiles, respectively. Small horizontal dashes on upper and lower whiskers represent 95th and 5th percentiles. Filled markers, for positive ACC, represent correlation significant at 95% confidence level based on two-sided Student's t-test against the null hypothesis of no correlation between the two variables.*

Since the constrained ensembles are most skillful when selected based on longer time-intervals (see Figure 5) and to maintain brevity of the deliverable, all the subsequent analysis in this section (as well as in section 3.1.3) is based on 9–year mean constraints. More detailed analysis are being prepared for peer-reviewed publications. As GMSST is subject to significant warming trends, leading to the high correlations seen in Figure 5a, we also analyse the residual correlation after removing the effect of the forced warming trend to identify the added value of the constraint (see Smith et al. 2019 for details). Figure 6 shows that the residual correlations of Best10 are statistically significant and are also higher than the 95th percentile of Random10 (i.e. when selected based on Global, NoPolar and Pac SST regions) while the Worst10 residual correlations are negative also indicating the effectiveness of the constraining approach.

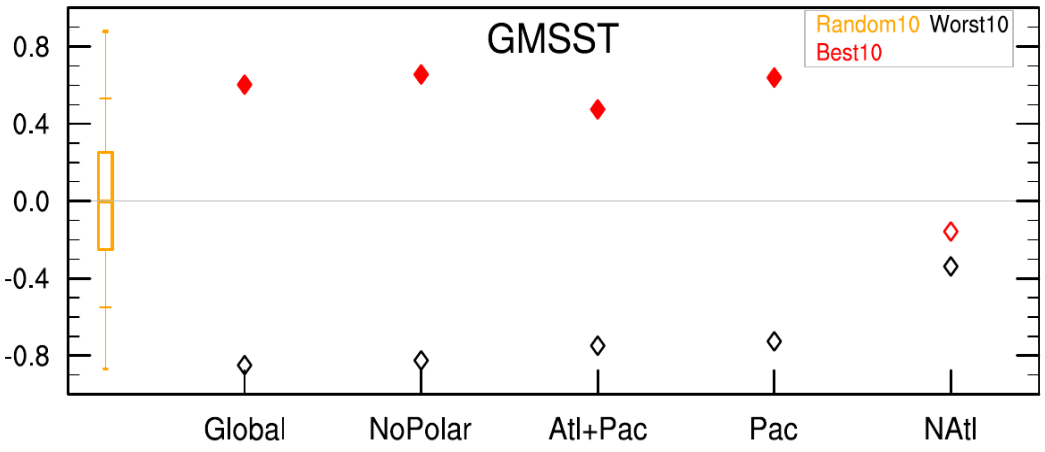

Figure 6*: Residual Correlations for global mean SST following Smith et al. (2019) for 9–year based constraints. The box-and-whisker plot represents residual correlation distribution of ensemble mean of 10 members selected randomly 40,000 times, where whiskers represent the minimum and maximum correlation and the horizontal line inside the box represents the median value. The lower and upper boundaries of the box represent the 25th and 75th percentiles, respectively. Small horizontal dashes on upper and lower whiskers represent the 95th and 5th percentiles. Filled markers, for positive residual correlations, represent correlation significant at 95% confidence level based on two-sided Student's t-test against the null hypothesis of no correlation between the two variables.*

We also evaluated the spatial characteristics of the constrained ensembles skill using ACC, root mean square skill score (RMSSS) and spread-over-error ratio (SOE; Ho et al., 2013). Figure 7 evaluates the regional characteristics of the different skill scores for predicting the 20–year mean surface air temperature anomalies. The ACC skill is high for most global regions due to the presence of strong warming trends in observations and model simulations (Figure 7a). We also obtained a similarly high ACC skill for the UNINIT40 ensemble mean (not shown). Due to these very high ACC values, we again use residual correlation to identify skill improvements in Best10 over UNINIT40 (Figure 7d). Apart from the Atlantic subpolar gyre region, where decadal predictions typically show added value from initialization, we also find significant positive residual correlations in areas of the tropical Pacific and Indian ocean, and tropical and North Atlantic ocean,

including some adjacent land regions, e.g., in South America, Africa, and Southern Asia (Figure 7d), indicating added value of the global pattern constraint.

The Best10 ensemble also shows skill in terms of RMSSS (Figure 7b), and added value compared to UNINIT40 over most of the areas highlighted by the residual correlations (Figure 7d). RMSSS improvements by the constraint are indicated by values above the 95th percentile of the Random10 distribution in several locations including the eastern Atlantic, western Europe, and parts of Africa and Asia (Figure 7e). Measured by the SOE, the Best10 ensemble is underconfident over North America and large parts of Europe and overconfident over east Asia, Africa, parts of south America, the North Atlantic, tropical central Pacific and the Southern Ocean (Figure 7c). In general both the Best10 and UNINIT40 ensembles are overconfident or underconfident in the same regions (not shown). In order to identify regions where the Best10 ensemble is more reliable than UNINIT40, regardless of being overconfident or underconfident, Figure 7f indicates if the SOE distance to the ideal value of 1 is reduced (shown by negative values) or increased (shown by positive values) in Best10 with respect to UNININT40, and by how much. The Best10 indicates improved reliability compared to UNINIT40 in large areas of the eastern North Atlantic, tropical eastern Indian ocean and some land regions in Asia, northern Africa and eastern parts of South America (indicated by negative values in Figure 7f).

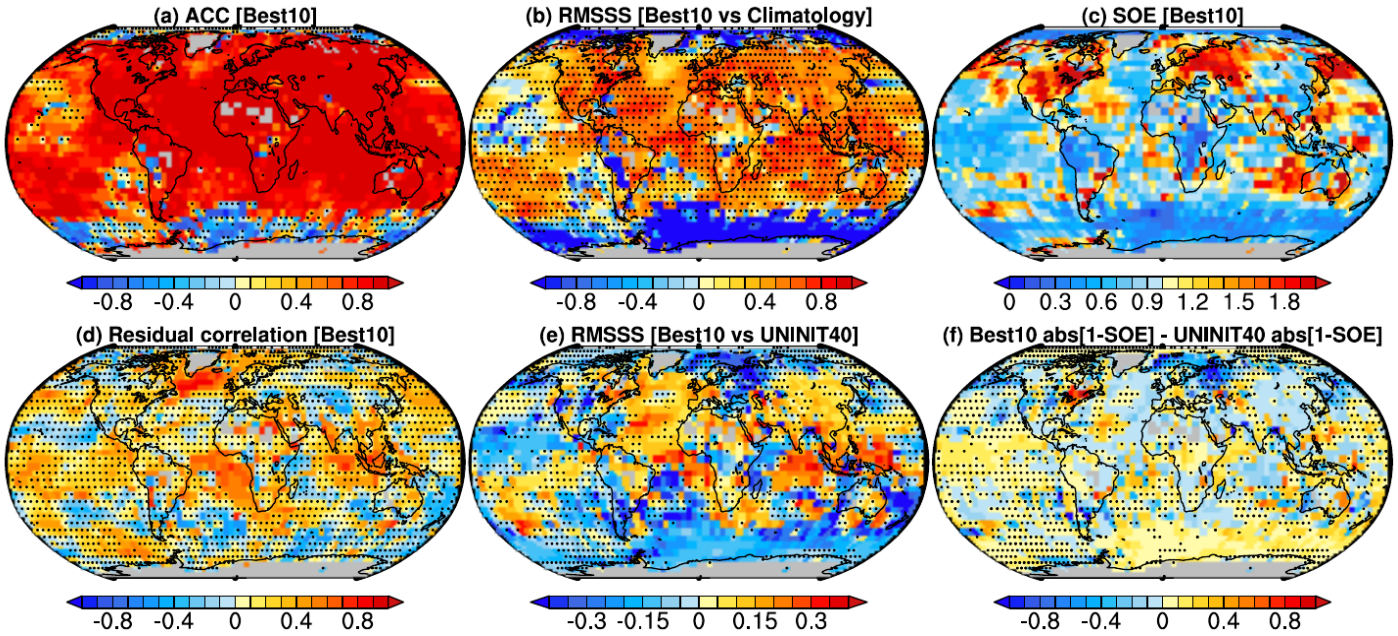

Figure 7: *(a) ACC for Best10 20-year near-surface temperature projections, (b) RMSSS and (c) spread over error ratio. Best10 selections are based on anomaly pattern correlations in UNINIT40 and DPLE over the first 9 forecast years. (d) Residual correlation for Best10 (e) RMSSS skill of Best10 relative to UNINIT40 and (f) difference of abs [1-SOE] (indicating the distance of SOE to the ideal value of 1) between Best10 and UNINIT40 (negative values indicate Best10 is more reliable than UNINIT40). The stippling in (a and d) indicates where correlation is not significant at the 95% confidence level. Similarly on panel (b) stippling represents RMSSS values not significant at the 95% confidence level using Fisher's f-test. For (e) and (f) stippling represents regions where the skill of Best10 lies in between 5th and the 95th percentile of the corresponding skill of Random10 distribution.*

Finally, we illustrate the application of the constrained Best10 ensemble for projecting near-term (2016–2035) summer (June-July-August) temperatures over the subpolar North Atlantic region (SPNA; 45ºN–60ºN; 310ºE–340ºE) and Western Asia (15ºN–50ºN; 40ºE–60ºE). For both regions we find improved skill for the Best10 ensemble compared to UNINIT40 over the hindcast period as shown by the skill metrics in Figure 8. The distributions of future projections indicate that the variability constraint excludes members of UNINIT40 with weaker warming leading to higher estimates of warming including a larger ensemble mean in the Best10 ensemble compared to UNINIT40. For instance, the minimum change over SPNA is +0.37K in Best10 compared to -0.05K in UNINIT40, while the maximum of the range remains unchanged at +0.81K resulting in larger ensemble mean warming signal in Best10 (+0.56K) compared to UNINIT40 (+0.47K). Similarly for Western Asia region we found higher minimum (and mean) value in the Best10 ensemble while the maximum value remains the same as for UNINIT40.

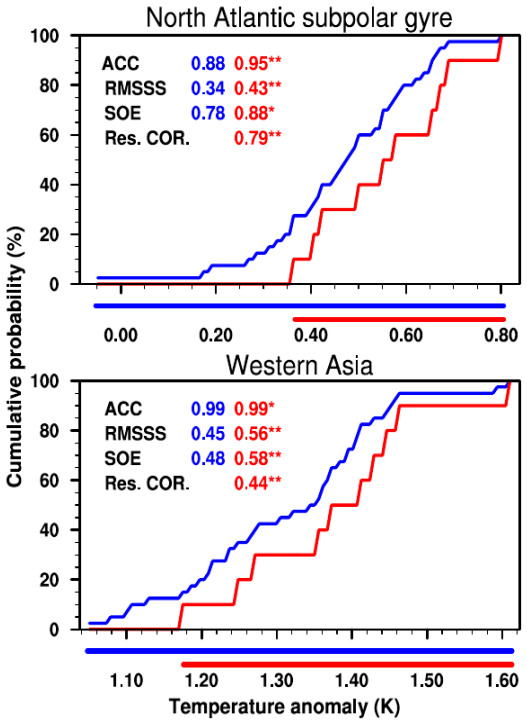

Figure 8: *Near-term summer temperature projections. Cumulative distribution functions of 20-year average (i.e. 2016–2035) projections of summer (June-July-August) near-surface air temperature anomalies (relative to 1961 to 1999) over the Subpolar North Atlantic (45ºN–60ºN; 310ºE–340ºE) and the IPCC SREX region of West Asia (over land areas). Best10 results (in red) are based on the values of the ensemble selected with the decadal prediction initialised in 2015. Selections are based on 9 year mean global SST anomaly patterns. The distribution of the unconstrained full UNINIT40 ensemble is shown in blue. The horizontal bars at the bottom of each panel show the range (minimum to maximum) of the 20-year average projections. The inset table summarises the different skill measures of hindcasts of 20-year average values from 1955–1974 to 1999–2018. For the Best10 skill measures (except for SOE), a single (double) star indicates that the skill is better than the 90th (95th) percentile of the corresponding skill of the Random10 distribution. For the SOE, a single (double) star indicates that the abs(1-SOE) of Best10 ensemble is lower than the 10th (5th) percentile of abs(1-SOE) distribution of Random10.*

These results suggest that, when aligning the climate variability phases in the projections with the decadal predictions, the climate change projections up to 2035 indicate enhanced warming compared to projections that are not constrained by the predictions. A more detailed discussion of these results (including results for other regions) are presented in a peer-reviewed paper (Mahmood et al., 2021, in review).

3.1.3 Merging climate projections and predictions based on global pattern agreement using large multi-model ensembles from CMIP6 *(Barcelona Supercomputing Center)*

To further explore the efficacy of the constraining methodology and to assess the added value in the constrained projections we apply the above-described technique to a much larger ensemble of initialized and projection simulations obtained from the Coupled Model Intercomparison Project phase 6 (CMIP6; Eyring et al., 2016). We use SST and surface air temperature data from 223 historical members simulated by 32 different models and 94 members of initialized predictions performed by 9 models. These ensembles are chosen based on the availability of the data on the Earth System Grid Federation (ESGF) at the start of our analysis. The historical CMIP6 simulations include observed time-varying natural and anthropogenic forcings until 2014 and projected forcings afterwards based on Shared Socioeconomic Pathways (SSP; O'Neil et al., 2014). For simulations after year 2014 we use SSP2-4.5 based future projections and predictions. The CMIP6 initialised predictions follow a common protocol under the so-called Decadal Climate Prediction Project component A (DCPP-A) experiments (Boer et al., 2016). These initialized decadal predictions (for brevity referred to as DCPP instead of DCPP-A) also include time-varying forcings but are started every year from observational climatic states and are integrated for at least 10 years.

The constraining procedure is exactly the same as discussed in section 3.1.2, however the hindcast period for this analysis is 1961–2000 as some of the DCPP simulations start from January 1961. In addition, here we select "Best30" for the highest ranking 30 members and similarly "Worst30" for the lowest ranking 30 members. We also compare the constrained ensemble in comparison to the skill distributions obtained by randomly selecting 30 members (referred to as "Random30"). The choice of 30 members for the constrained ensemble is arbitrary. However, we also tested selections of 10 or 50 members and did not find any major differences in the results compared to 30 member selections (not shown). In this section, the constrained ensembles are based on 9–year mean SST anomaly pattern correlations since the optimum skill is obtained when using longer constraining periods (see section 3.1.2). The sensitivity of the constrained ensembles to different selection regions is evaluated using three large scale SST domains (i.e. Global, Pac, and NAtl; see Table 1).

We evaluate the skill of the constrained ensemble compared to the unconstrained (i.e. all members, henceforth referred to as "UNINIT") and DCPP ensemble mean for 10 year mean hindcasts (i.e. forecast years 1 to 10). Figure 9(a-d) shows that the DCPP is highly skillful in predicting GMSST, AMV, and SPNA, providing a skillful basis for constraining the projections. In particular the constraining approach based on global SSTs leads to improved skill of Best30 for GMSST and AMV with ACC values that are statistically significant and also higher than the 95th percentile of

the Random30. For SPNA, the ACC of Best30 is close to the corresponding skill of the initialised predictions when selected based on NAtl SST. The lowest skill is found for the Worst30 adding further confidence to the efficacy of the constraining methodology. The IPO is only marginally better predicted in DCPP than in the UNINIT and therefore the skill of the Best30 is also not statistically significant but it is still higher than the 75th percentile of the Random30. These results show that the constraining depends on the skill of the initialized predictions which is consistent with previous studies (Befort et al., 2020; Mahmood et al., 2021, in review). For IPO, we also note that the skill of Best30 is higher than for the initialized predictions suggesting that the DCPP skill is not a limiting factor for the skill of the constrained projections (Figure 9d). Similarly for the 20-year mean hindcasts the constrained ensembles show significantly improved skill compared to the uninitialized ensembles (Figure 9e-h). On these timescales the ACC skill for IPO is also statistically significant for Best30 when constrained based only on Pacific SST (Figure 9h).

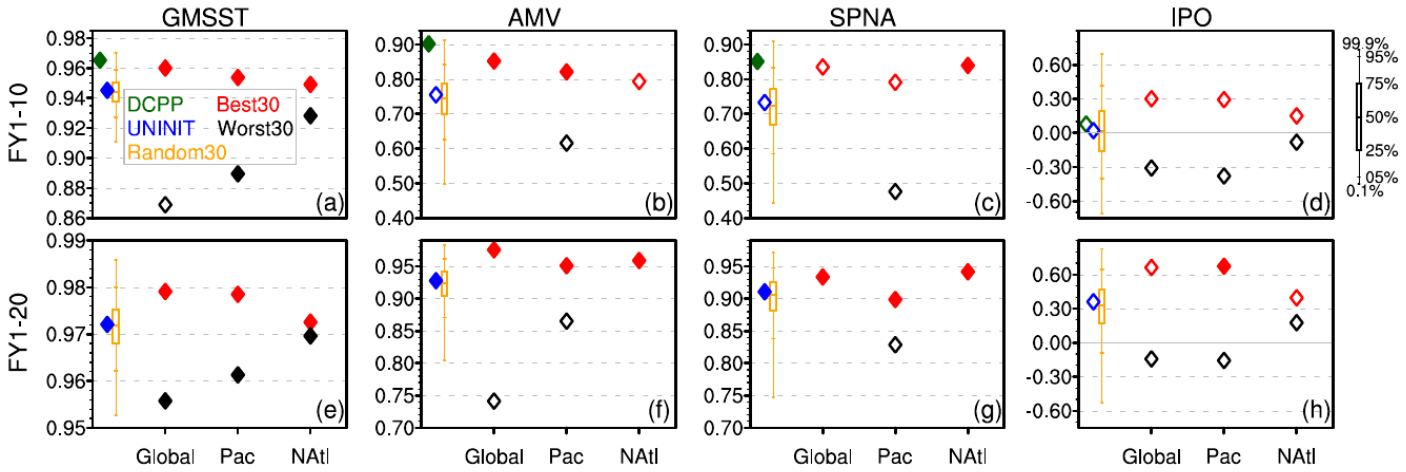

Figure 9: *ACC of large scale SST indices for first 10 year mean hindcasts. Best30 and Worst10 are defined based on 9–year mean anomaly pattern correlations with initialized predictions. Each marker represents the Best30 members chosen using different selection regions as shown on x-axis. The box-and-whisker plots represent the range of skill scores for 100,000 randomly selected 30-member ensemble means. The lower and upper boundaries of the box represent 25th and 75th percentiles respectively and the horizontal line inside the box represents median value. Small horizontal dashes on upper and lower whiskers represent 95th and 5th percentiles respectively. ACC for initialized prediction (in green) and uninitialized projection (in blue) ensembles is also shown. For AMV and SPNA, some of the ACC values for Worst30 lie outside the plot limits.*

The regional characteristics of the constrained ensemble skill are evaluated using residual correlations (Figure 10). For the forecast years 1-10, we find that the skill in Best30 is in most regions at least as high as the corresponding skill of the DCPP (cf. 10a, and 10c). This suggests that the skill in the constrained ensemble may not just be due to reducing noise by selecting those members that are closest to the DCPP ensemble mean. To further investigate this we also evaluate the skill of a constrained 30 member DCPP ensemble (henceforth referred to as "DCPP30"). Similar to the constrained projections the DCPP30 sub-ensemble is defined, for each start date, by ranking the 9–year mean global SST based anomaly pattern correlation of individual members of DCPP with it's ensemble mean. Figures 10a and 10b show that the skill of DCPP30 is marginally

improved over the corresponding skill of DCPP especially in tropical pacific, where the DCPP30 shows relatively higher correlation values and also extended areas of significant skill. But overall the DCPP30 skill map is very similar to the full DCPP ensemble, and the constrained Best30 projections have added value over the decadal predictions in several regions for predicting the first 10 years after initialisation. This may be indicative of other problems, e.g. related to model drift, affecting the initialised decadal predictions but not the projections.

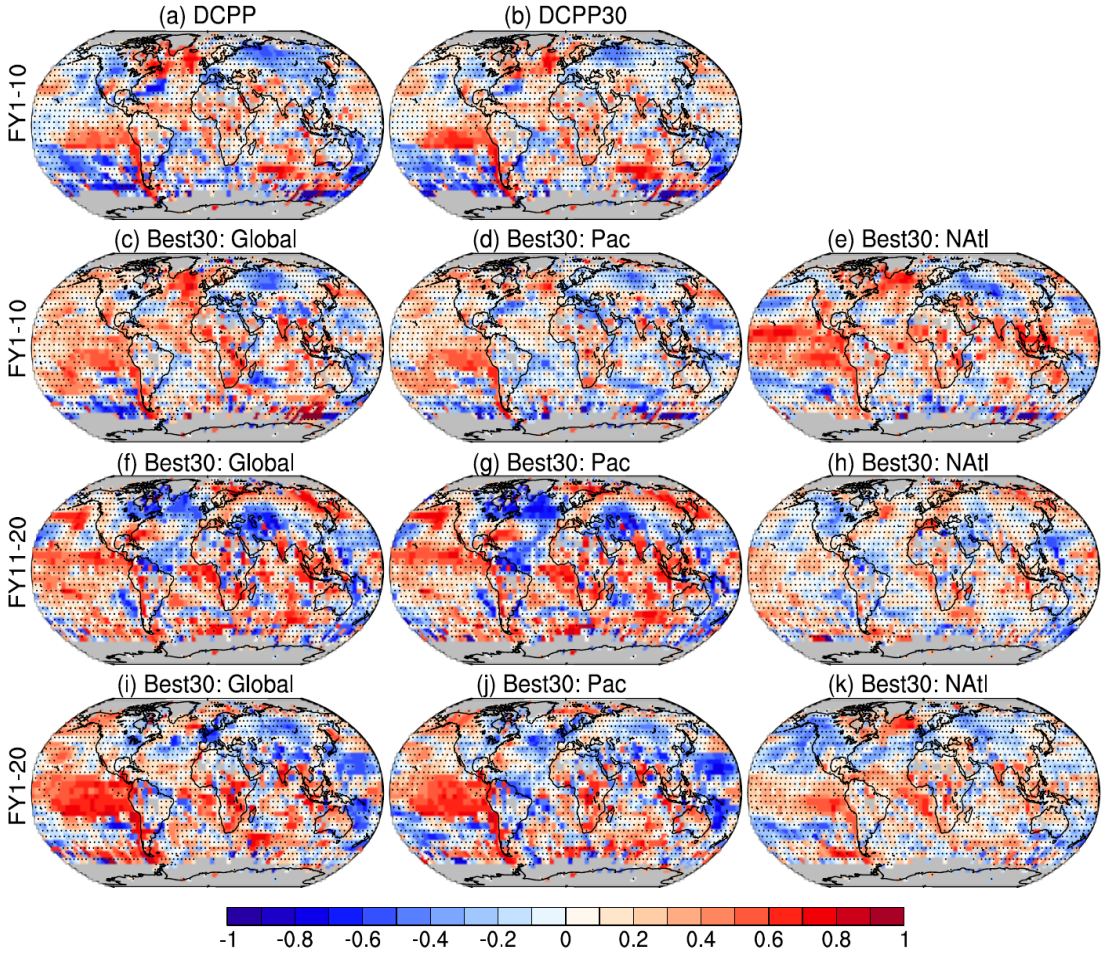

Figure 10: *Residual correlations of initialized predictions and the constrained projections for the hindcast periods of years 1-10 (first two rows), 11-20 (third row) and 1-20 (bottom row). The top row shows residual correlations for DCPP ensemble mean and a sub-selected 30 members initialised ensemble, as DCPP30. The second, third and fourth rows show residual correlations for the Best30 constrained by 9 year mean SST anomalies over three regions. Stippling shows regions where the residual correlations are not statistically significant for 95% confidence level based on Student's t-test.*

The Best30 is also skilful in several global regions for forecast years 11 to 20 (Figure 10f–h) and 1 to 20 (Figure 10i–k). These results show that the regional skill depends, apart from using longer time periods, on the choice of the constraining SST regions. For example, the global SST based constraining can provide skillful 20 year mean constrained projections over the Pacific, Atlantic and Indian oceans as well as over several land regions including Africa, South and Southeast Asia,

Australia and western North America. Similarly the Best30 is relatively more skillful in the North Atlantic region when constrained based on NAtl.

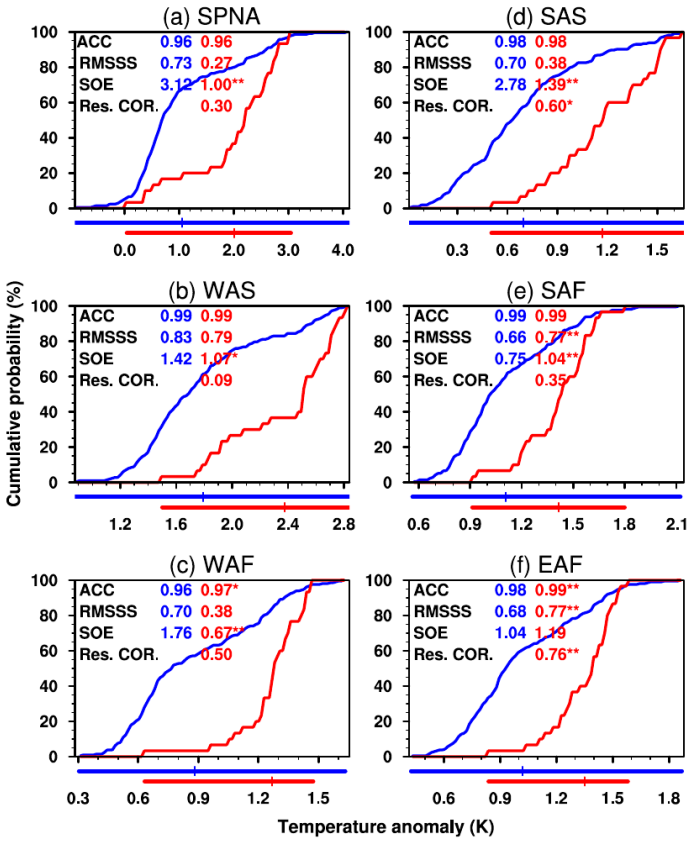

Figure 11: *Near-term summer temperature projections. Cumulative distribution functions of 20-year average (i.e. 2015-2034) projections of summer (June-July-August) near-surface air temperature anomalies (relative to 1961 to 2000) over the Subpolar North Atlantic (45°N–60°N; 20°W–50°W) and the five IPCC SREX regions. Best30 results (in red) are based on the selections using 9 year mean (i.e. 2015 to 2023) global SST anomaly patterns. The distribution of the unconstrained full UNINIT ensemble is shown in blue. The horizontal bars at the bottom of each panel show the range (minimum to maximum) of the 20-year average projections with small vertical dash line representing ensemble mean. The inset table summarises the different skill measures of hindcasts of 20-year average values from 1961–1980 to 2000–2019. For the Best30 skill measures (except for SOE), a single (double) star indicates that the skill is better than the 90th (95th) percentile of the corresponding skill of the Random30 distribution. For the SOE, a single (double) star indicates that the abs(1-SOE) of Best30 ensemble is lower than the 10th (5th) percentile of abs(1-SOE) distribution of Random30.*

Similar to section 3.1.2, we also show here the applicability of the constraining methodology to provide future projections of 20 year mean (2015–2034) summer (JJA) temperature anomalies over SPNA and five IPCC's SREX regions (Figure 11). The CMIP6 ensembles show warmer projections of summer temperature anomalies compared to the CESM single model large ensemble (cf. Figures 8 and 11a-b). Furthermore, the Best30 ensemble means tend to project higher temperature anomalies for all regions compared to the UNINIT ensemble mean which is consistent with single model large ensemble based constraining.

Overall, these results indicate great potential for improving near-term climate change estimates of the next few decades by constraining large multi-model ensembles of climate projections based on their agreement with initialised decadal predictions. In ongoing work we explore the added skill in more detail, for other variables, and beyond the current 20-year prediction horizon. This work is being prepared for a journal publication to be submitted in autumn 2021.

3.2 Temporal merging of decadal predictions and climate projections *(University of Oxford, in collaboration with ETHZ and IPSL)*

Another possible method to obtain seamless climate information beyond decadal timescales is by explicitly merging, or "stitching together", decadal predictions and climate projections after forecast year 10 (or before). In a collaboration between UOXF, ETHZ and IPSL it is assessed to what extent such a temporal merging of both data sources introduces inconsistencies in the resulting time series. The analysis has been carried out for annual mean near-surface temperatures (SAT) in each of the SREX (IPCC 2012) regions (including land and ocean grid cells). Data from eight different decadal prediction systems (CanESM5, MPI-ESM1.2-HR, EC-Earth3 (i1), HadGEM3-GC31-MM, IPSL-CM6A-LR, MIROC6, NorCPM1 (i1), NorCPM1 (i2)) and their corresponding historical projections are used. Each prediction/projection consists of 10 members, except CanESM5 for which 20 members are available and HadGEM3-GC31-MM for which only 4 projections are available. Thus, the decadal prediction ensemble consists of 90 members, whereas the uninitialized climate projection ensemble consists of 84 members. Lead-time dependent bias corrections following Boer et al. (2016) using the years 1970 until 2014 are applied to all decadal prediction single-model ensembles. For historical projections, anomalies were calculated against the climatological average from 1970 until 2014.

The problems potentially arising when stitching together decadal predictions and climate projections after forecast year 10 are illustrated in Figure 12 for the example of SATs over the Northern Europe region (NEU) using the multi-model ensemble (MME) of decadal predictions initialized in 1975 (first complete forecast year is 1976). Figure 12a shows the time series for projection data only, whereas Figure 12b shows the decadal predictions for the first 10 years 1976 until 1985 and the projections afterwards. In contrast to Figure 12a, a clear inconsistency is found in Figure 12b when stitching together both data sources in 1985. Large differences are found for different percentiles of the multi-model distribution, especially the more extreme ones, e.g. the 10th percentile, which has much lower values in the projections for 1986 than in the predictions for 1985. The apparent inconsistencies of the distributions before and after stitching may lead to difficulties for potential end-users when interpreting those datasets. However, some level of interannual variability is expected, and indeed is present in the projection time series, creating some variability around the stitching years. In order to assess the extent to which inconsistencies arise when stitching the predictions and projections, the following analysis has been performed.

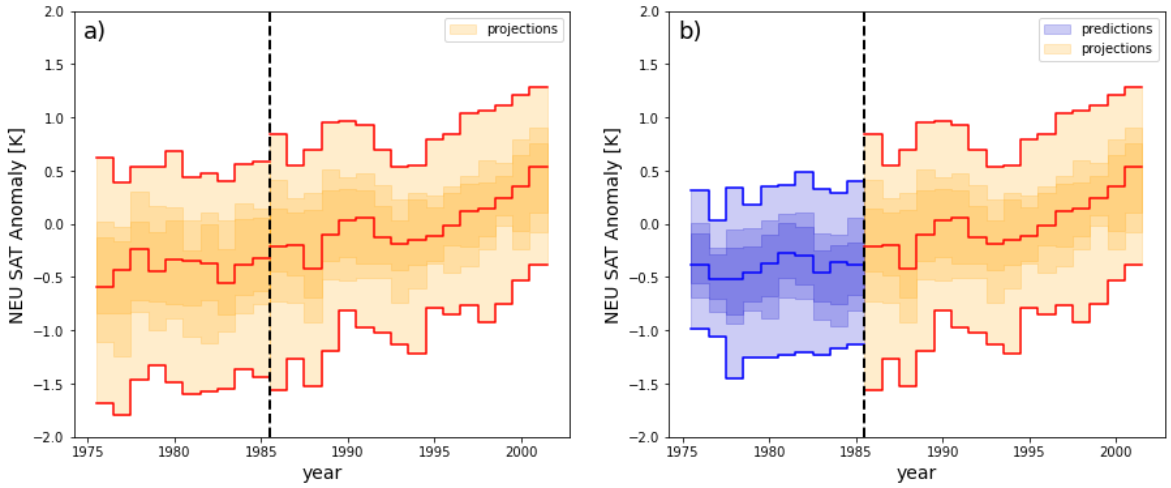

Figure 12: ***(a)*** *Northern Europe near-surface temperature (SAT) anomalies in the projection multi-model ensemble. The shading indicates the 10th, 25th, 33th, 66th, 75th and 90th percentiles of the distribution, whereas the solid lines indicate the 10th, 50th (median) and 90th percentiles.* ***(b)*** *same as (a) but using the decadal prediction multi-model ensemble (blue) initialized in 1976 up to 1985 and projections thereafter. The dashed vertical line indicates the time at which predictions and projections would be combined.*

Firstly, the expected interannual variability of SAT over a specific region is defined based on the 1-year increments between each year in the period 1970 to 2013 and the corresponding following year using data from the projections (subsequently called *baseline*). These 44 increments are calculated for different quantiles of the MME distribution, e.g. median (see Figure 12), as different users might be interested in different aspects of the distribution.

In addition to the *baseline* increments, the increments introduced by the stitching are calculated using decadal prediction data from forecast year 10 over the period 1970 until 2013 and projection data for the respective following year (1971–2014, n=44). We therefore assume that a potential end user would use the decadal prediction information until lead year 10 and information from the projection simulations afterwards (forecast year 11 onwards). This approach is called *stitching* subsequently. We use 4 metrics to assess differences between *baseline* and *stitching*, which are illustrated in Figure 13 for the 10th percentile of global SAT.

Metric 1 (M1) is defined by the p-values of the t-test of the differences between the 44 *baseline* and *stitching* increments (Wilks, 2006). Thus, M1 assesses whether mean increment differences between *stitching* and *baseline* are significantly different from 0. The example in Figure 13a shows very small p-values for the t-test, indicating that the differences between *baseline* and *stitching increments* are significant. Metric 2 (M2) is based on the p-value of the Kolmogorov-Smirnov test statistic comparing *baseline* and *stitching* increment cumulative distributions (Wilks, 2006). Thus, M2 in contrast to M1 compares the whole distributions of the *stitching* and *baseline* increments rather than comparing differences of the distribution means. Similarly to M1, the small p-value of M2 indicates significant differences between the two distributions. Metric 3 (M3) and Metric 4 (M4) aim to quantitatively measure differences when stitching predictions and projections. M3 is based on the quantile value of the *stitching* median in the *baseline* distribution. If differences

between *stitching* and *baseline* are small, the *stitching* median is expected to match the median of the *baseline* distribution. M3 (minus 50) thus provides a measure of the distance of the median of the *stitching* distribution to the median of the *baseline* distribution. For the 10th percentile of global SAT M3 equals about 30 (Figure 13c), meaning the median of the *stitching* distribution equals approximately the 20th percentile of baseline distribution (please note that median and percentiles in Figure 13c are based on the real underlying data, whereas only the gaussian fit is plotted for simplicity). M4 is calculated using the absolute differences of the mean values of *stitching* and *baseline* divided by the standard deviation of the *baseline* distribution (this is the interannual standard deviation of the increment time series). In the example of Figure 13d, the normalised difference is 0.7, meaning that the differences in the means of *stitching* and *baseline* distributions equals 0.7 times the interannual standard deviation of the *baseline* distribution.

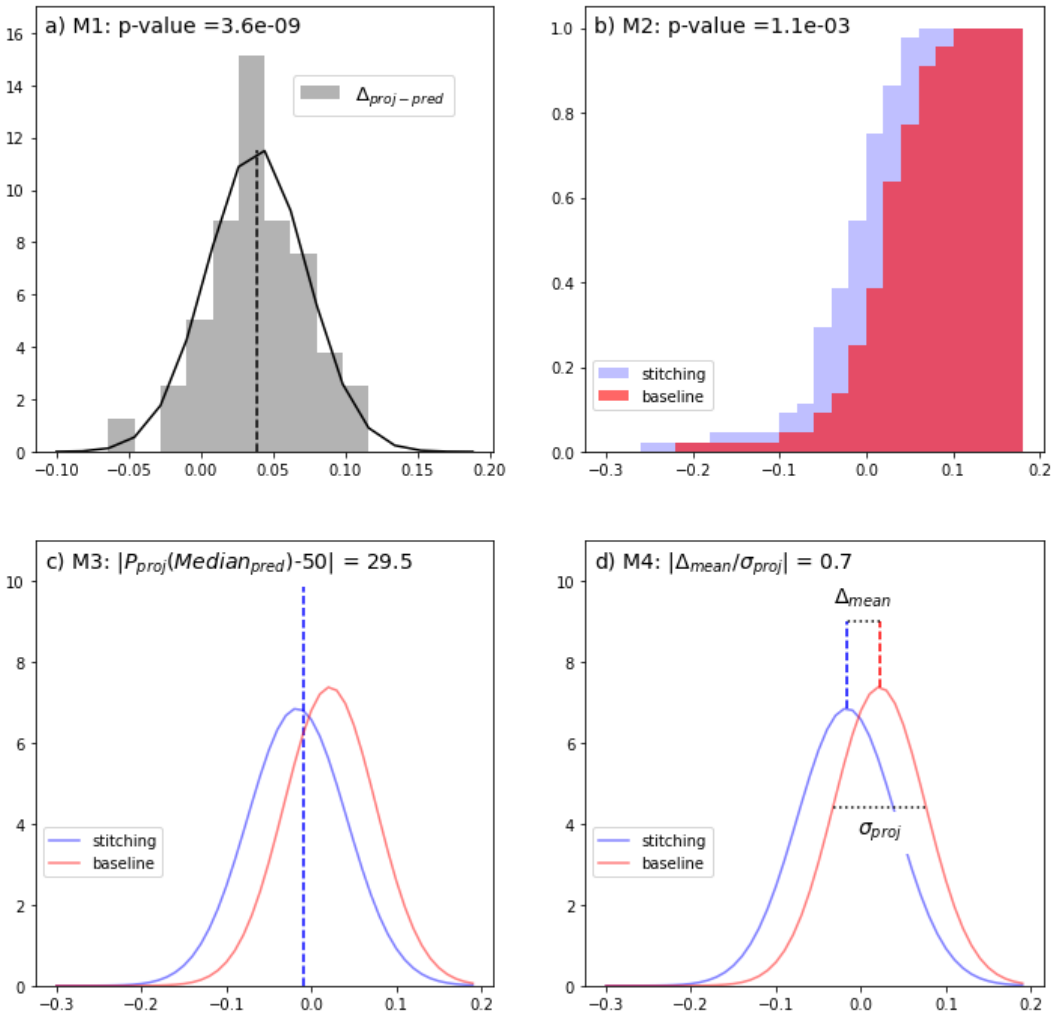

Figure 13: *Metrics used to assess inconsistencies in increments when combining decadal predictions and projections after forecast year 10. The example is for the 10th percentile of global SAT.* ***(a)*** *M1: Distribution of differences between baseline and stitching values* ***(b)*** *M2: cumulative distributions of stitching and baseline,* ***(c)*** *M3: fitted normal distributions to pdf's of stitching and baseline. (d) same as (c). Values given on top of each plot are results for each metric for the 10th percentile of global SAT. Dashed lines in (c) indicate the median of the stitching distribution, whereas dashed lines in (d) represent the means of the stitching and baseline distributions.*

Results for all SREX regions and global near-surface temperatures are shown in Figure 14. It is found that combining predictions and projections after forecast year 10 rarely introduces inconsistencies for the MME median of the timeseries. This is different for higher/lower percentiles and especially pronounced for the global (GLOB), NEU (Northern Europe) and CGI (Eastern Canada/Greenland/Iceland) regions. For these 3 regions M1/M2 indicate significant differences in the means and the cumulative increment distributions of *stitching* and *baseline*, whereas M3 and M4 indicate that for these regions quantitative differences are also large. Besides these 3 regions, inconsistencies are also found for other regions, e.g. Mediterranean (MED), Western Asia (WAS) or Tibetan Plateau (TIB), but these are limited to more extreme percentiles, less pronounced, and less robust across the different metrics. Overall, these results indicate that stitching predictions and projections together after forecast year 10 might introduce large inconsistencies over some regions, particularly for more extreme percentiles of the multi-model distributions. However, it also shows that for other regions the simple stitching approach could potentially work in order to obtain seamless climate information.

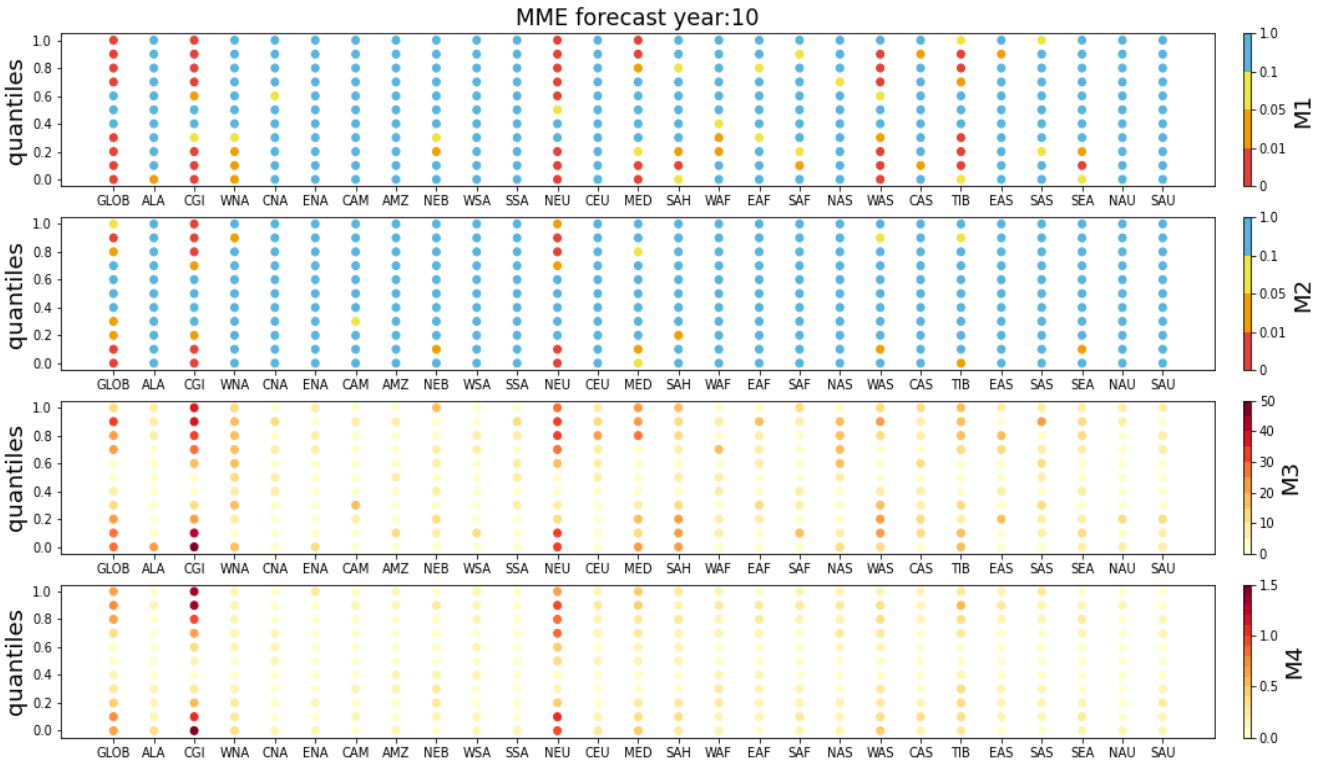

Figure 14: *Results for metrics M1 to M4 for different SAT quantiles over all SREX regions and globally averaged SAT. Color-coding for M1 and M2 indicates p-value derived from the t-test (M1) and ks-test (M2) respectively. Darker colors in M3 and M4 indicate that inconsistencies are larger for those regions and quantiles. The y-axis shows the different percentiles.*

We currently assess to what extent a simple calibration method or a weighting scheme based on model performance (Knutti et al., 2019) can be used to minimise the inconsistencies when stitching predictions and projections. As for the calibration method, we use the variance inflation method (VINF), which is described in Doblas-Reyes et al. (2005) and has been recently applied to

large-ensemble projections (O'Reilly et al., 2020). The application of the VINF method scales the signal and ensemble spread which yields a (statistically reliable) calibrated ensemble on interannual timescales. A scientific publication on these results is in preparation.

3.3 Discussion of the relevance of observational constraints for merging approaches *(University of Edinburgh)*

The team at the University of Edinburgh focused on the use of observational constraints on predictions and projections, which are largely discussed elsewhere. Deliverable 5.1 (a paper based on it published as Hegerl et al., 2021) discusses both the need to use observational constraints consistently across predictions and projections, and the benefit of using such constraints. These can be used both to select climate models whose simulated changes are consistent with observations (and disregard those which are not; e.g. Tokarska et al., 2020), or multi model averages can be weighted for improved performance (see Hegerl et al., 2021). This work is well developed in projections, and still being explored for predictions, yet first results are promising. Some observational constraints, even when applied to initialized predictions in the near term, are influenced by the emerging forced signal, while model-to-data agreement in the historical period can also be strongly affected by short-term forcings such as volcanism. The latter shows that since constraints on predictions and projections are influenced by different factors, they may pull in different directions if not used consistently. Results of these analyses are now published (Hegerl et al., 2021 and references therein).

Some factors that are important for initialized predictions, such as the state of decadal observed internal variability can confound observational constraints on projections, in the sense that it can be a confounding factor. For example, the observed long-term variability in the North Atlantic Oscillation has influenced both temperature and precipitation trends, particularly in the cold season (Iles and Hegerl,. 2017), and climate models do not reproduce such low-frequency variability (O'Reilly et al., in press; Schurer et al., in draft). Thus results using the ASK method (Stott and Kettleborough, see Brunner et al., 2020; Ballinger et al., in prep.) suggests that observed trends in precipitation are stronger in some regions than in climate models, (e.g. stronger contrast between increase in Northern Europe; weakening precipitation in Southern Europe). If the influence of the North Atlantic Oscillation is removed prior to analysis from both climate model data and observations (based on observed and simulated sea level pressure), then the observational constraint is in agreement with the multi-model mean (as indicated by scaling factors closer to '1' in Figure 15, which indicates that the multi-model mean projection is more consistent with observations). Since long-term variability in the NAO can also influence the ocean state (e.g., Iles and Hegerl, 2017), the failure of climate models to reproduce the observed NAO variability may challenge merging attempts (see also O'Reilly et al., 2021). The effect of such variability on constraints is illustrated in Figure 16, where constrained projections considering this confounding factor (purple) are more similar to the multi model mean projections than where it is not (blue). In summary, observational constraints should be considered when merging. However, where variability in observations is not consistent with that simulated, observational constraints may produce misleading results, and initialized simulations may show different variability from free running projections.

A publication that includes the NAO results is in preparation, both using a data assimilation technique (Schurer et al., in prep), and analyzing observational constraints on European regions, considering the influence of the NAO (Ballinger et al., in prep). Both should be available later in 2021.

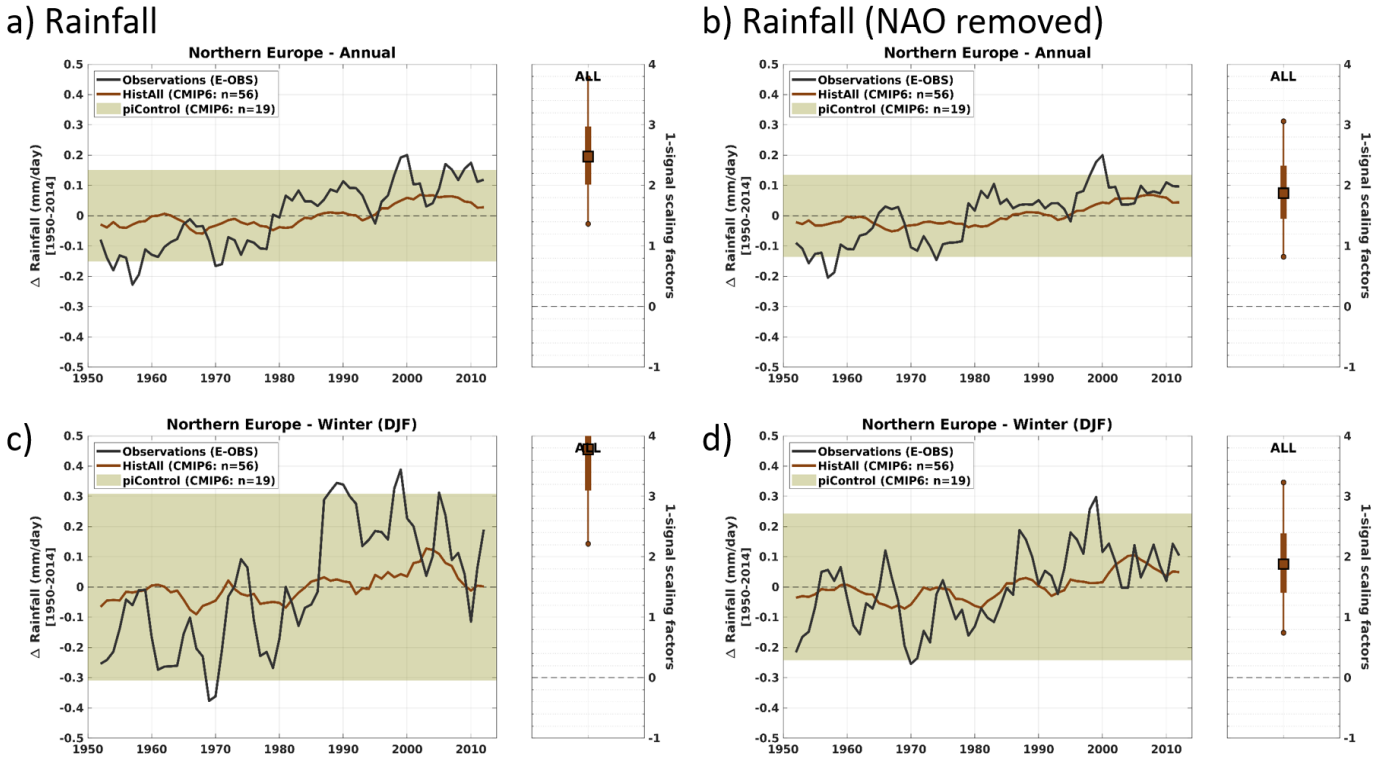

Figure 15: *Annual (a, b) and Seasonal (DJF, c,d)) time series of Northern European rainfall anomalies (relative to 1950-2014) from observations (E-OBS v19, black line) and CMIP6 historical simulations (all forcings, brown line, displaying the multi-model mean of ensemble means (19 models, 56 total ensemble members); a), c) original time series, and b),d) time series with the NAO removed. Time series are smoothed with a 5-yr running mean, and the shaded region denotes the mean variability (±1 standard deviation) of the associated unsmoothed piControl simulations. The 1-signal scaling factor is derived from a TLS regression of the CMIP6 model fingerprint and the observations, indicating to what extent the multi-model mean fingerprint needs to be scaled to best match observations (central square marker) and can be scaled to still be consistent with observations (5-95% range).*

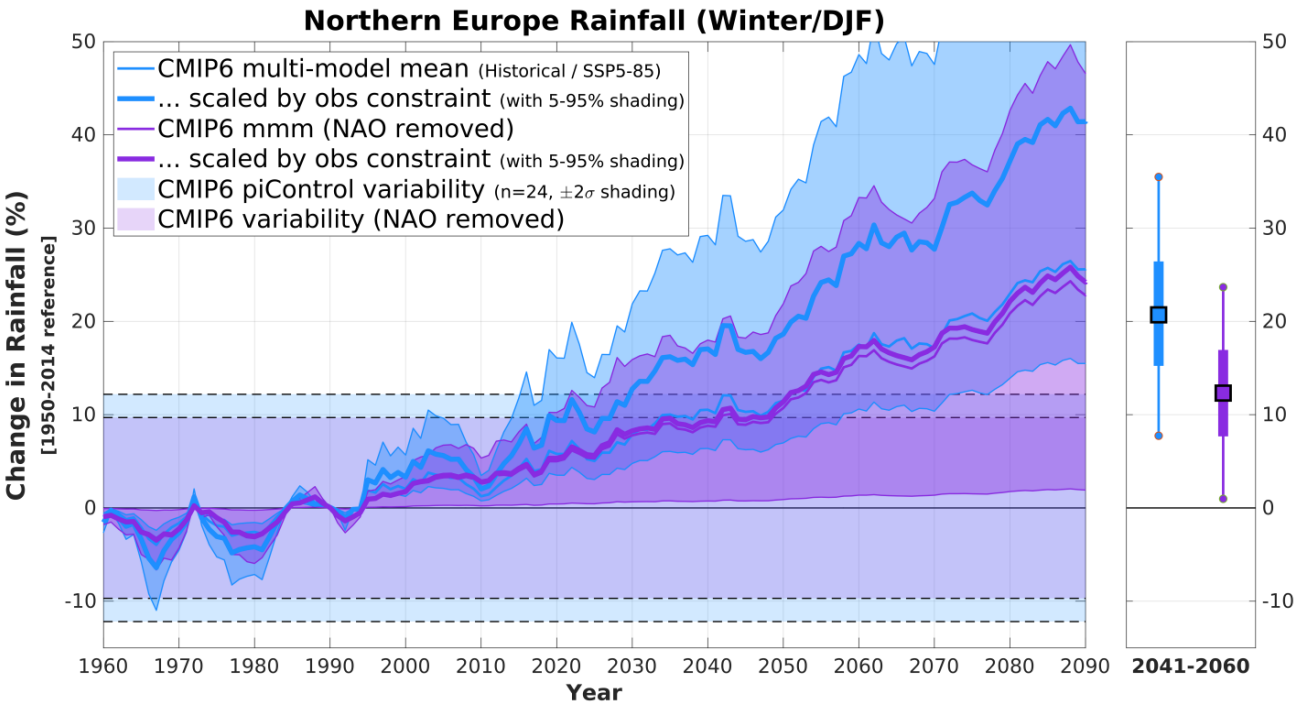

Figure 16: *The Impact of accounting for NAO variability in the observational constraint on projections. The thin lines show the CMIP6 multi-model mean of ensemble means (66 total simulations from 24 models, forced with historical emissions and the future SSP5-85 scenario from 2015) of northern European winter (DJF) rainfall, shown as a percentage change relative to a 1950–2014 baseline, with a 5-yr running mean. The thick line and shaded region shows this multi-model mean scaled by the best estimate (and 5th–95th percentile range) of the scaling factor required for the historical simulations to be consistent with past observed winter rainfall. The blue lines/shading show the multi-model mean and constrained projection using the raw (total) winter rainfall, whereas the purple lines/shading show the results after first regressing out the component of rainfall that is associated with the NAO (1st EOF of SLP) from both models and observations. The dashed coloured regions indicate the mean variability of the 24 individual piControl model simulations, with and without the NAO. Right panel: The associated constrained projections of rainfall change for the period 2041–2060; square marker indicates the 50th percentile, and the thick (thin) bar the 25th–75th (5th–95th) percentile.*

3.4 Exploring pattern scaling as a way to merge climate predictions with climate projections *(University of Copenhagen)*

At the University of Copenhagen, we have explored the concept of scaled climate change pattern. This method is scaling at a grid-point level the climate change using the spatial average of the 2-meter temperature global change. This latter approach has been explored in order to merge climate predictions with climate projections using the CanESM2 (projection) and CanCM4 (prediction) model. Those results were presented in a few workshops but not published.

Christensen et al. (2019) showed that pattern scaling has similar outcomes from several coordinated experiments (PRUDENCE, ENSEMBLES and CORDEX). Those outcomes suggested that pattern scaling is robust across modeling initiatives. Furthermore, they show a high correspondence to a similar scaled pattern deduced from an observational dataset. Recently, Matte et al. (2019) have shown that such patterns emerge earlier within the simulation than was previously thought. A

plausible explanation behind the temporal gap between the observed scaled pattern and the emerging simulated one is that the simulation needed time to stabilize upon the external forcing.

Since climate predictions are initialized, the hypothesis was that the observed warming trend pattern mentioned in Christensen et al. (2019) would be present and should evolve in a similar way than the projection as the lead-time increases. In other words, Christensen et al. (2019) have shown that the observed warming trend of the 2-meter temperature over a historical period was similar to the scaled pattern from climate projection, the initial idea was then to show that the same pattern is emerging in climate prediction as lead time increases. The expected impact was to explore a possible way to temporally merge the projection with the prediction.

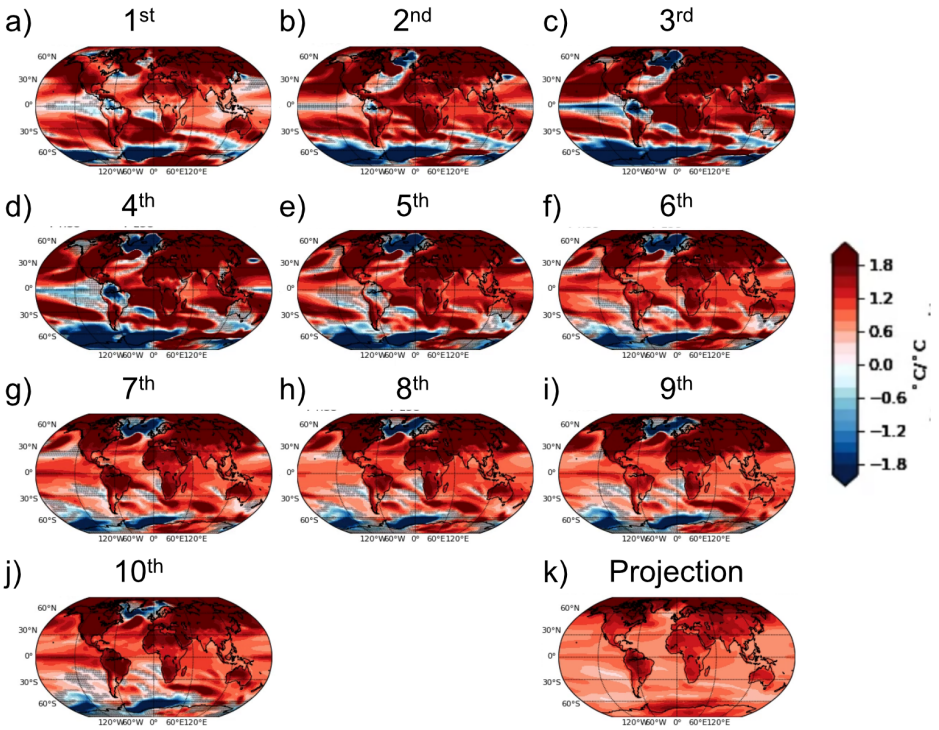

Figure 17: *1990–2009 near surface-air temperature scaled patterns deduced from CanCM4 (predictions) for lead years 1 to 10 (a to j, respectively) and the end-century (2071–2100) scaled patterns from CanESM2 (k). All patterns have been computed using the 30-year climatology reference of 1961–1990.*

The scaled pattern from climate prediction was calculated as follows. A time series was built using only the first lead year of all start-dates (1960–2014) where the grid-point warming/cooling over a moving window of 20 years was divided by the global change over the same window. The protocol was repeated for all lead years and compared to the end century (2071–2100) scaled patterns from climate projections.

A subset of the results is shown in Figure 17 where scaled patterns extracted from different lead times of the climate prediction (Figure 17a–j) and climate projection (Figure 17k) are mapped. The pattern correlation between the lead years and the end-of-century scaled patterns increased from 0.35 to 0.53 from lead year 1 (Figure 17a) to the lead year 10, respectively. There are major differences in key regions such as the North Atlantic subpolar gyre and around Antarctica. Although those differences in the early lead time could be due to a shock caused by initialization,

they remain for longer lead time which led to a hitch that put an end to the exploration using this technique.

4. Lessons learnt

Combining the information from initialised decadal predictions and climate projections offer promising pathways towards providing improved and seamless climate information for the next multiple decades. In particular, initialisation of decadal predictions aligns the phases of simulated and observed climate variability, which reduces the uncertainty of near-term climate change estimates. Some added value from the initialisation can persist beyond the 10 years covered by decadal predictions. This deliverable report presents pioneering research towards developing seamless climate information from combining decadal predictions and projections. Different conceptual approaches are followed, which (i) constrain large ensembles of climate projections based on their agreement with decadal predictions, and (ii) explore inconsistencies between decadal predictions and projections that can prevent a direct merging of both data sources.
Key conclusions are:

- First studies have demonstrated the concept to constrain climate projections based on their agreement of decadal predictions. These have shown the potential for such constraints to improve the skill of the climate projections beyond the first decade covered by decadal predictions. In particular, constraining based on temperatures over the North Atlantic Subpolar Gyre has shown to improve projections of temperatures over the North Atlantic Subpolar Gyre for up to 15 years after initialisation. Constraining based on global patterns of temperature anomalies has been shown to improve regional projections of 20-year average temperatures in several regions including the North Atlantic ocean and teleconnected land regions, such as Western Asia.
- The actual data provided by decadal predictions and projections can be inconsistent (e.g. as a consequence of initialisation shocks), which could cause inhomogeneities when merging data from the different types of simulations. Statistical correction methods, such as calibration or model weighting, can help to reduce such inconsistencies, and their optimal implementation is subject to ongoing research.

Given the promising results from these first implementations to combine information from predictions and projections, there is scope to continue developing these methods. Of particular interest for future work are:

- Combining different types of constraints, that also consider performance metrics, or the representation of key processes, in addition to the phases of climate variability. This has the potential to further reduce the uncertainty of climate projections, addressing e.g. model uncertainty in addition to uncertainty from climate variability.
- Explore the effect of the constraining or merging methods on other climate variables beyond average temperatures. Developing methods to effectively reduce the uncertainty of new-term predictions for impact-relevant variables, including climate extremes, will be most relevant for informing the development of adaptation strategies to the expected climate in the next few decades.

Given the pioneering character of this work, it is still too early to provide guidance on possibly 'preferable' approaches towards combining decadal predictions and projections for specific applications. An initial discussion of potential advantages and disadvantages is provided below:

- Univariate statistical corrections of distributional inconsistencies between predictions and projections are specific to the variables for which they are applied (e.g. temperature), and would likely not be applicable as such to other variables (e.g. precipitation). In contrast, the constraining methods that sub-sample large ensembles of projections are physically consistent across different variables, and may therefore be preferable when providing seamless climate information for multiple climate variables.
- Application of constraints to 154 projection ensemble members from CMIP5 (Section 3.1.1, and Befort et al., 2020) suggested that the skill of the decadal predictions may act as a limit for the skill of the constrained projections also during the first decade - and provide a motivation to statistically correct the inconsistencies between both types of data sources, to make sure to use the most skillful information for all prediction times. However, recent applications to even larger ensembles (more than 200 members from CMIP6, see Section 3.1.3) indicate that the skill of the constrained projections can in fact provide similar skill to the decadal predictions used to constrain the projections. In fact, the constrained ensemble exhibited in some cases significant added value over the projections where decadal predictions did not exhibit added value. This possibly surprising result needs to be better understood. Possible explanations may be related to improving the signal-to-noise ratio when sub-selecting ensemble members close to the mean signal of the decadal predictions (similar to Smith et al., 2020). Other possible reasons may be related to problems potentially deteriorating the skill in decadal predictions, such as initialisation shock and related climate drift, but not affecting the projections as such.

5. Links built

WP5 established a sequence of very fruitful meetings, in which the different approaches towards providing seamless climate information were constructively discussed between partners from a number of institutions contributing to WP5. With an interest in the development of temporally seamless climate information for the next multiple decades, active exchanges have been established in particular between UOXF, IPSL, ETHZ, UEd, UCPH, and the BSC. We hope that these exchanges and collaborations also persist beyond the work presented in this deliverable, as we expect there is still large potential to further improve the approaches pioneered as part of the research presented in this deliverable. This work was also supporting wider discussions between work on decadal predictions (WP1) and constraining projections (WP2).

6. Acronyms

ACC - Anomaly Correlation Coefficient
AMV - Atlantic Multidecadal Variability
BSC - Barcelona Supercomputing Center - Centro Nacional De Supercomputacion
CMIP - Climate Model Intercomparison Project
CNRS - Centre National De La Recherche Scientifique

CESM - Community Earth System Model
DCPP - Decadal Climate Prediction Project
DPLE - Decadal Prediction Large Ensemble
ESGF - Earth System Grid Federation
ETHZ - ETH Zurich
EUCP - European Climate Prediction system
GMSST - Global Mean Sea Surface Temperature
IPCC - Intergovernmental Panel on Climate Change
IPO - Interdecadal Pacific Oscillation
IPSL - Institut Pierre Simon Laplace
MAE - Mean Absolute Error
NAO - North Atlantic Oscillation
NCAR - National Center for Atmospheric Research
RMSE - Root Mean Square Error
RMSSS - Root Mean Square Skill Score
RCP - Representative Concentration Pathways
SAT - Surface Air Temperature
SOE - Spread-Over-Error ratio
SPNA - Sub-polar North Atlantic
SSP - Shared Socioeconomic Pathways
SST - Sea Surface Temperature
UCPH - University of Copenhagen
UEd - University of Edinburgh
UOXF - University of Oxford
VINF - Variance Inflation method

7. References

Befort, D. J., O'Reilly, C. H., & Weisheimer, A. (2020). Constraining projections using decadal predictions. Geophysical Research Letters, 47, e2020GL087900. https://doi.org/10.1029/2020GL087900

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Mueller, W. A., Taylor, K. E., Zwiers, F., Rixen, M., Ruprich-Robert, Y., and Eade, R.: The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, Geosci. Model Dev., 9, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016 , 2016.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., Wehner, M. F., Allen, M. R., Andrews, T., Beyerle, U., Bitz, C. M., Bony, S., & Booth, B. B. B. (2013). Long-term Climate Change: Projections, Commitments and Irreversibility. Climate Change 2013 - The Physical Science Basis: Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, 1029–1136.

Christensen, J. H., Larsen, M. A., Christensen, O. B., Drews, M., & Stendel, M. (2019). Robustness of European climate projections from dynamical downscaling. *Climate Dynamics*, *53*(7), 4857-4869.

Doblas-Reyes, F.J., Hagedorn, R. and Palmer, T.N. (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. Tellus A, 57: 234-252. https://doi.org/10.1111/j.1600-0870.2005.00104.x

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

Hegerl GC, Ballinger AP, Booth BBB, Borchert LF, Brunner L, Donat MG, Doblas-Reyes FJ, Harris GR, Lowe J, Mahmood R, Mignot J, Murphy JM, Swingedouw D and Weisheimer A (2021) Toward Consistent Observational Constraints in Climate Predictions and Projections. *Front. Clim.* 3:678109. doi: 10.3389/fclim.2021.678109

Ho, C. K., Hawkins, E., Shaffrey, L., Bröcker, J., Hermanson, L., Murphy, J. M., Smith, D. M., & Eade, R. (2013). Examining reliability of seasonal to decadal sea surface temperature forecasts: The role of ensemble dispersion. Geophysical Research Letters, 40(21), 5770–5775. https://doi.org/10.1002/2013GL057630

Iles C., Hegerl G.C. (2017): Role of the North Atlantic Oscillation in Decadal Temperature Trends; Env. Res. Lett. 12, 114010.

IPCC 2012 *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change* (Cambridge: Cambridge University Press)

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., … Vertenstein, M. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability. Bulletin of the American Meteorological Society, 96(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., and Eyring, V. (2017), A climate model projection weighting scheme accounting for performance and interdependence, Geophys. Res. Lett., 44, 1909– 1918, doi:10.1002/2016GL072012.

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*, in review.

Matte, D., Larsen, M. A. D., Christensen, O. B., and Christensen, J. H. (2019). Robustness and scalability of regional climate projections over Europe. *Frontiers in Environmental Science*, *6*, 163.

O'Reilly, C. H., Befort, D. J., and Weisheimer, A.: Calibrating large-ensemble European climate projections using observational data, Earth Syst. Dynam., 11, 1033–1049, https://doi.org/10.5194/esd-11-1033-2020, 2020.

O'Reilly, C.H., D.J., Befort, A. Weisheimer, T. Woollings, A. Ballinger and G. Hegerl (2021). Projections of northern hemisphere extratropical climate underestimate internal variability and associated uncertainty. Commun Earth Environ, doi:10.1038/s43247-021-00268-7

O'Neill, B. C., Kriegler, E., Riahi, K., Ebi, K. L., Hallegatte, S., Carter, T. R., Mathur, R., & van Vuuren, D. P. (2014). A new scenario framework for climate change research: The concept of shared socioeconomic pathways. Climatic Change, 122(3), 387–400. https://doi.org/10.1007/s10584-013-0905-2

Smith, D. M., Eade, R., Scaife, A. A., Caron, L.-P., Danabasoglu, G., DelSole, T. M., Delworth, T., Doblas-Reyes, F. J., Dunstone, N. J., Hermanson, L., Kharin, V., Kimoto, M., Merryfield, W. J., Mochizuki, T., Müller, W. A., Pohlmann, H., Yeager, S., & Yang, X. (2019). Robust skill of decadal climate predictions. Npj Climate and Atmospheric Science, 2(1), 1–10. https://doi.org/10.1038/s41612-019-0071-y

Smith, D.M., Scaife, A.A., Eade, R. *et al.* North Atlantic climate far more predictable than models imply. *Nature* 583**,** 796–800 (2020). https://doi.org/10.1038/s41586-020-2525-0

Tokarska K., Hegerl G.C., Schurer A.P., Forster P. and Marvel K. (2020): Observational Constraints on the effective climate sensitivity from the historical record. Environ. Res. Lett. 15 (2020) 034043 https://iopscience.iop.org/article/10.1088/1748-9326/ab738f/pdf

Wilks DS. Statistical methods in the atmospheric sciences. Academic press; 2006.

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A., Strand, W., Bates, S. C., Meehl, G. A., Karspeck, A. R., Lindsay, K., Long, M. C., Teng, H., & Lovenduski, N. S. (2018). Predicting Near-Term Changes in the Earth System: A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model. Bulletin of the American Meteorological Society, 99(9), 1867–1886. https://doi.org/10.1175/BAMS-D-17-0098.1

List of tables

List of figures

Figure 3: *(a) ACC and (b) RMSE for unconstrained projections (black), constrained projections using decadal ensemble mean (green), constrained projections using observations (orange), unconstrained projections with fixed ensemble size n = 35 (gray) and initialized predictions (blue) for annual mean surface air temperatures over the North Atlantic Gyre. Statistics are given for 5-year averages. The gray area indicates the 10–90% confidence intervals (based on a 10,000 sample bootstrap). Closed circles indicate those periods for which all forecast years have been used to constrain the respective ensemble, whereas open circles indicate those periods for which at least 1 year has not been used for constraining. Figure taken from Befort et al. (2020) their Figure 3.*

Figure 4: *Schematic diagram explaining the large ensemble subselection methodology (Idealized data). For each start date, SST anomaly of individual projection members is compared with the ensemble mean of SST anomalies of initialized prediction using area-weighted spatial pattern correlation. The selection period (shown as orange line) can be any time interval within the forecast range of the initialized prediction. The projection members are ranked based on the pattern correlation coefficients and the top N members (N can be any subset of the projection ensemble) are chosen as "BestN" for each start date. The temporal trajectories of these BestN members are then used to predict the climate over the 20 years after initialisation, as depicted by the green line. Note that by design these "BestN" members can be a different subset of the projection ensemble at each start date.*

Figure 5: *ACC of large-scale SST indices for all time periods and domains used for sub-selection. Each marker represents the Best10 and Worst10 members chosen using different selection regions (cf. legend with markers in the top panel). The box-and-whisker plots represent the range of skill scores for 40,000 randomly selected 10-member ensemble means, while whiskers represent the minimum and maximum correlation and the horizontal line inside the box represents the median value. The lower and upper boundaries of the box represent the 25th and 75th percentiles, respectively. Small horizontal dashes on upper and lower whiskers represent 95th and 5th percentiles. Filled markers, for positive ACC, represent correlation significant at 95% confidence level based on two-sided Student's t-test against the null hypothesis of no correlation between the two variables.*

Figure 6*: Residual Correlations for global mean SST following Smith et al. (2019) for 9–year based constraints. The box-and-whisker plot represents residual correlation distribution of ensemble mean of 10 members selected randomly 40,000 times, where whiskers represent the minimum and maximum correlation and the horizontal line inside the box represents the median value. The lower and upper boundaries of the box represent the 25th and 75th percentiles, respectively. Small horizontal dashes on upper and lower whiskers represent the 95th and 5th percentiles. Filled markers, for positive residual correlations, represent correlation significant at 95% confidence level based on two-sided Student's t-test against the null hypothesis of no correlation between the two variables.*

Figure 7: *(a) ACC for Best10 20-year near-surface temperature projections, (b) RMSSS and (c) spread over error ratio. Best10 selections are based on anomaly pattern correlations in UNINIT40 and DPLE over the first 9 forecast years. (d) Residual correlation for Best10 (e) RMSSS skill of Best10 relative to UNINIT40 and (f) difference of abs [1-SOE] (indicating the distance of SOE to the ideal value of 1) between Best10 and UNINIT40 (negative values indicate Best10 is more reliable than UNINIT40). The stippling in (a and d) indicates where correlation is not significant at the 95% confidence level. Similarly on panel (b) stippling represents RMSSS values not significant at the 95% confidence level using Fisher's f-test. For (e) and (f) stippling represents regions where the skill of Best10 lies in between 5th and the 95th percentile of the corresponding skill of Random10 distribution.*

Figure 8: *Near-term summer temperature projections. Cumulative distribution functions of 20-year average (i.e. 2016–2035) projections of summer (June-July-August) near-surface air temperature anomalies (relative to 1961 to 1999) over the Subpolar North Atlantic (45°N–60°N; 310°E–340°E) and the IPCC SREX region of West Asia (over land areas). Best10 results (in red) are based on the values of the ensemble selected with the decadal prediction initialised in 2015. Selections are based on 9 year mean global SST anomaly patterns. The distribution of the unconstrained full UNINIT40 ensemble is shown in blue. The horizontal bars at the bottom of each panel show the range (minimum to maximum) of the 20-year average projections. The inset table summarises the different skill measures of hindcasts of*

20-year average values from 1955–1974 to 1999–2018. For the Best10 skill measures (except for SOE), a single (double) star indicates that the skill is better than the 90th (95th) percentile of the corresponding skill of the Random10 distribution. For the SOE, a single (double) star indicates that the abs(1-SOE) of Best10 ensemble is lower than the 10th (5th) percentile of abs(1-SOE) distribution of Random10.

Figure 9: *ACC of large scale SST indices for first 10 year mean hindcasts. Best30 and Worst10 are defined based on 9–year mean anomaly pattern correlations with initialized predictions. Each marker represents the Best30 members chosen using different selection regions as shown on x-axis. The box-and-whisker plots represent the range of skill scores for 100,000 randomly selected 30-member ensemble means. The lower and upper boundaries of the box represent 25th and 75th percentiles respectively and the horizontal line inside the box represents median value. Small horizontal dashes on upper and lower whiskers represent 95th and 5th percentiles respectively. ACC for initialized prediction (in green) and uninitialized projection (in blue) ensembles is also shown. For AMV and SPNA, some of the ACC values for Worst30 lie outside the plot limits.*

Figure 10: *Residual correlations of initialized predictions and the constrained projections for the hindcast periods of years 1-10 (first two rows), 11-20 (third row) and 1-20 (bottom row). The top row shows residual correlations for DCPP ensemble mean and a sub-selected 30 members initialised ensemble, as DCPP30. The second, third and fourth rows show residual correlations for the Best30 constrained by 9 year mean SST anomalies over three regions. Stippling shows regions where the residual correlations are not statistically significant for 95% confidence level based on Student's t-test.*

Figure 11: *Near-term summer temperature projections. Cumulative distribution functions of 20-year average (i.e. 2015-2034) projections of summer (June-July-August) near-surface air temperature anomalies (relative to 1961 to 2000) over the Subpolar North Atlantic (45°N–60°N; 20°W–50°W) and the five IPCC SREX regions. Best30 results (in red) are based on the selections using 9 year mean (i.e. 2015 to 2023) global SST anomaly patterns. The distribution of the unconstrained full UNINIT ensemble is shown in blue. The horizontal bars at the bottom of each panel show the range (minimum to maximum) of the 20-year average projections with small vertical dash line representing ensemble mean. The inset table summarises the different skill measures of hindcasts of 20-year average values from 1961–1980 to 2000–2019. For the Best30 skill measures (except for SOE), a single (double) star indicates that the skill is better than the 90th (95th) percentile of the corresponding skill of the Random30 distribution. For the SOE, a single (double) star indicates that the abs(1-SOE) of Best30 ensemble is lower than the 10th (5th) percentile of abs(1-SOE) distribution of Random30.*

Figure 12: ***(a)*** *Northern Europe near-surface temperature (SAT) anomalies in the projection multi-model ensemble. The shading indicates the 10th, 25th, 33th, 66th, 75th and 90th percentiles of the distribution, whereas the solid lines indicate the 10th, 50th (median) and 90th percentiles.* ***(b)*** *same as (a) but using the decadal prediction multi-model ensemble (blue) initialized in 1976 up to 1985 and projections thereafter. The dashed vertical line indicates the time at which predictions and projections would be combined.*

Figure 13: *Metrics used to assess inconsistencies in increments when combining decadal predictions and projections after forecast year 10. The example is for the 10th percentile of global SAT.* ***(a)*** *M1: Distribution of differences between baseline and stitching values* ***(b)*** *M2: cumulative distributions of stitching and baseline,* ***(c)*** *M3: fitted normal distributions to pdf's of stitching and baseline. (d) same as (c). Values given on top of each plot are results for each metric for the 10th percentile of global SAT. Dashed lines in (c) indicate the median of the stitching distribution, whereas dashed lines in (d) represent the means of the stitching and baseline distributions.*

Figure 14: *Results for metrics M1 to M4 for different SAT quantiles over all SREX regions and globally averaged SAT. Color-coding for M1 and M2 indicates p-value derived from the t-test (M1) and ks-test (M2) respectively. Darker colors in M3 and M4 indicate that inconsistencies are larger for those regions and quantiles. The y-axis shows the different percentiles.*

Figure 15*: Annual (a, b) and Seasonal (DJF, c,d)) time series of Northern European rainfall anomalies (relative to 1950-2014) from observations (E-OBS v19, black line) and CMIP6 historical simulations (all forcings, brown line, displaying the multi-model mean of ensemble means (19 models, 56 total ensemble members); a), c) original time series, and b),d) time series with the NAO removed. Time series are smoothed with a 5-yr running mean, and the shaded region denotes the mean variability (±1 standard deviation) of the associated unsmoothed piControl simulations. The 1-signal scaling factor is derived from a TLS regression of the CMIP6 model fingerprint and the observations, indicating to what extent the multi-model mean fingerprint needs to be scaled to best match observations (central square marker) and can be scaled to still be consistent with observations (5-95% range).*

Figure 16*: The Impact of accounting for NAO variability in the observational constraint on projections. The thin lines show the CMIP6 multi-model mean of ensemble means (66 total simulations from 24 models, forced with historical emissions and the future SSP5-85 scenario from 2015) of northern European winter (DJF) rainfall, shown as a percentage change relative to a 1950–2014 baseline, with a 5-yr running mean. The thick line and shaded region shows this multi-model mean scaled by the best estimate (and 5th–95th percentile range) of the scaling factor required for the historical simulations to be consistent with past observed winter rainfall. The blue lines/shading show the multi-model mean and constrained projection using the raw (total) winter rainfall, whereas the purple lines/shading show the results after first regressing out the component of rainfall that is associated with the NAO (1st EOF of SLP) from both models and observations. The dashed coloured regions indicate the mean variability of the 24 individual piControl model simulations, with and without the NAO. Right panel: The associated constrained projections of rainfall change for the period 2041–2060; square marker indicates the 50th percentile, and the thick (thin) bar the 25th–75th (5th–95th) percentile.*

Figure 17: *1990–2009 near surface-air temperature scaled patterns deduced from CanCM4 (predictions) for lead years 1 to 10 (a to j, respectively) and the end-century (2071–2100) scaled patterns from CanESM2 (k). All patterns have been computed using the 30-year climatology reference of 1961–1990.*