HORIZON 2020
THEME SC5-2017

European Climate Prediction system

(GRANT AGREEMENT 776613)

European Climate Prediction system (EUCP)

Deliverable D5.2

Evaluation of forecast quality over the 1-40 year time span for both global initialised forecasts and the non-initialised projections

Deliverable Title	*Evaluation of forecast quality over the 1-40 year time span for both global initialised forecasts and the non-initialised projections*	
Brief Description	*This deliverable report compares the performance of initialised and non-initialised global predictions for overlapping prediction time scales, with a view to develop merging methodologies in task 5.2. A particular focus is the assessment of model reliability and forecast spread, potential economic value, and the role of forced and natural variability.*	
WP number	*WP5*	
Lead Beneficiary	*UOXF*	
Contributors	*Daniel J. Befort (UOXF), Christopher H. O'Reilly (UOXF), Antje Weisheimer (UOXF), Deborah Verfaillie (BSC), Markus Donat (BSC), Francisco Doblas-Reyes (BSC), Simon Wild (BSC), Gabi Hegerl (UEDIN), Andrew Ballinger (UEDIN), Juliette Mignot (CNRS/IPSL), Leonard Borchert (CNRS/IPSL), James Murphy (UKMO), Tim Kruschke (SMHI)*	
Creation Date Version Number Version Date	*15/01/2021* *[v1]* *30/03/2021*	
Deliverable Due Date Actual Delivery Date	*31/03/2021* ***TO BE DONE***	
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	15/01/2021	Daniel J Befort	First version of D5.2 using structure from outline submitted in Nov. 2020
V2	18/02/2021	Daniel J Befort et al.	Include contribution from all partners
V3	23/02/2021	A.Weisheimer/D. Befort	Finalized draft for internal peer review
V4	26/03/2021	A Weisheimer, D Befort	Comments int. review
V5	29/03/2021	D. Befort	Comments int review II
V6	30/03/2021	A Weisheimer, D Befort	Final report for submission

Table of Contents

1. Executive Summary

This deliverable compares the performance of initialised and non-initialised global predictions for overlapping prediction time scales, with a view to develop merging methodologies in task 5.2. As will be demonstrated, the potential benefit of initializing models with the observed climate state does not, however, always translate into measurable performance improvements on multi-annual time scales. One of the reasons is likely related to the so-called initialisation shock, that is the inconsistency between the observed state of the climate system and the model climate attractor which tends to pull the forecast model towards its intrinsic state. Another reason is the externally forced long-term trend found over large areas of the globe for surface temperatures, which explains large parts of the total variance and which is captured to a large extent by climate projections and initialized predictions.

While details of the fundamental reasons behind these findings will need to be fully understood in the future beyond the EUCP project, the first and last part of this report with contributions from UOXF, BSC and UKMO focuses on the assessment of essential characteristics of probabilistic forecasting systems. Reliability quantifies the agreement between the predicted probabilities and observed relative frequencies of a given event. Reliability is therefore a key requirement for the predictions to be useful to decision-makers, who base their decisions on the prediction of certain event types. It has been demonstrated that bias correction and calibration of the raw initialised and non-initialised data is crucial to obtain reliable predictions and projections, in line with a recent study showing the benefit of applying calibration methods to large-ensembles on European climate projections in WP2. In addition to reliability, the sharpness of the forecast probability distributions is shown to be a key characteristic for a forecast to have value to users. Here, a simplified cost-loss decision model has been explored to demonstrate the potential economic value of initialised and non-initialised forecasts.

A perfect model framework has been used by BSC to better understand where decadal real-world predictions can be potentially improved further. For example, from comparing initialised and non-initialized predictions, model long-term changes in the Pacific that are inconsistent with the observed climate have been identified to likely cause the current lack of skill over this region.

The North Atlantic region and its sea surface temperatures (SST) are important drivers of European climate and have shown relatively large added value from the initialised forecasts compared to non-initialised projections in the CMIP5 simulations. CNRM-IPSL's contribution highlights the modelling advances over the North Atlantic in the latest generation of climate predictions and projections in CMIP6. While the new models, both initialised and non-initialised, are generally more skilful in reproducing observed SST variations in the North Atlantic subpolar gyre region (specifically from the 1980s onwards), the added value of initialised runs is, however, strongly reduced. It was found that natural forcings, in particular volcanic forcings, play a stronger role than previously thought.

Assessing the role of natural forcings has also been a focus of UEDIN's contribution to this deliverable. The impact of the North Atlantic Oscillation (NAO) on precipitation and surface air temperature across the different Special Report on Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation (SREX) regions over Europe has been studied to understand the role of internal and forced variability. The analysis suggests that the models substantially underestimate the change in precipitation. This is important because the multi-decadal NAO variability in the past is generally not reproduced by CMIP class models and can hence confound estimated trends due to forcing, affecting forward projections. The NAO is also relevant for seamless predictions, as it is predictable over months to possibly years, although with insufficient amplitude

UK Climate Projections (UKCP18) is the latest generation of national UK climate scenarios. UKMO have explored the extent to which initialised predictions from CMIP6 can improve upon, or augment, the information on near-term risks available from the probabilistic projections of UKCP18, with a specific focus on the performance over England and Wales. The findings suggest that differences in spread between the initialised CMIP6 forecasts and UKCP18 are likely to depend mainly on differences in the amplitude of internal variability simulated in their constituent modelling systems.

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	Develop an ensembles climate prediction system based on high-resolution climate models for the European region for the near-term (~1-40 years)	x	
2	Use the climate prediction system to produce consistent, authoritative and actionable climate information	x	
3	Demonstrate the value of this climate prediction system through high impact extreme weather events in the near past and near future	X	
4	Develop, and publish, methodologies, good practice and guidance for producing and using EUCP's authoritative climate predictions for 1-40 year timescales		x

3. Detailed Report

Introduction

The aim of this deliverable is to assess key characteristics of initialised and non-initialised global simulations. These characteristics include the skill of predictions and projections, measured using standard deterministic as well as probabilistic approaches. Large emphasis is put on the potential added value of initialisation which is analysed for those CMIP5 and CMIP6 models that provide both initialised and non-initialised simulations. An advantage of CMIP6 over CMIP5 is that all prediction systems are initialised annually (whereas in CMIP5 some prediction systems have only been initialised every 5 years) and mostly provide 10 or more ensemble members, which potentially allows a more robust assessment of their skill and their ability to represent observed features as e.g. teleconnections.

Due to the combined use of decadal predictions and climate projections, the work carried out for this deliverable has been closely linked to efforts in WP1 and WP2. Whereas WP1 mainly focuses on climate information on decadal time scales provided by initialised prediction systems, WP2 deals with climate information provided by non-initialised climate projections. WP5 aims to join information from both strands and provide an assessment of key characteristics in predictions and projections. Such knowledge is crucial for the development of successful methodologies for merging climate information from projections and predictions in WP5.

Institutions contributing to this deliverable also have strong linkages to either WP1 or WP2 or both, so that several different approaches to evaluate the quality of predictions and projections have been explored. The analyses presented in this report have also triggered a number of cross-WP collaborations, aiming to apply methods originally developed solely for climate projections or initialised predictions, to initialised and non-initialised simulations in a systematic and comparative way.

The milestone MS19 “Internal workshop on length of initialised skill to review research quantifying the time at which added skill from initialised predictions fades out, for multiple climate variables, regions and seasons” which was planned to form a crucial element of the efforts in task T5.1, could unfortunately not be held in the anticipated format of a face-to-face workshop in Oxford in June 2020 due to the global pandemic. Instead, a stripped-down online workshop was held. While initially we were still optimistic to be able to postpone the workshop until spring 2021, it has now become clear that the earliest realistic time to hold the workshop is in the autumn 2021.

University of Oxford

1. Potential economic value of initialised predictions and non-initialised projections

Added skill of initialised predictions over non-initialised climate projections is a prerequisite for the temporal merging of predictions and projections (*Befort et al., 2020*). Here, key characteristics of CMIP5 and CMIP6 initialized predictions and non-initialised projections are assessed. These include common deterministic and probabilistic metrics including anomaly correlation coefficients (ACC) as a measure of skill, reliability, sharpness (range of forecasted event probabilities). Furthermore, we aim to assess the value of initialised predictions over non-initialized projections from a user perspective using the potential economic value (PEV) framework. The added value of initialisation is quantified using the approach presented by *Smith et al. (2019)*. In this method, the variability explained by the climate projection ensemble mean is removed from the decadal prediction ensemble as well as from the observational data. The anomaly correlation coefficient (ACC) between these two residual time series quantifies the added value of initialisation. Analyses presented here are for surface temperature averaged over 2-9 years lead time.

For the analyses, CMIP5 and CMIP6 initialised predictions and non-initialised projections are used. To assess the added value of initialisation, only models for which both predictions and projections are available have been selected: CMIP5 – 7 models with 64-member prediction and 46-member projection ensemble; CMIP6 – 7 models with 90-member prediction and 84-member projection ensembles. Predictions are corrected for lead-time dependent biases following *Boer et al. (2016)* using the baseline period from 1970 until 2006. The mean over the same baseline period is removed from climate projections and observations (HadCRUT5; *Morice et al., 2021*). For projections the climatological mean is calculated for each ensemble member separately. To compare the performance of CMIP5 and CMIP6 models to each other, only data for the common time period from 1961 to 2014 is used (initialised predictions from 1960 to 2004).

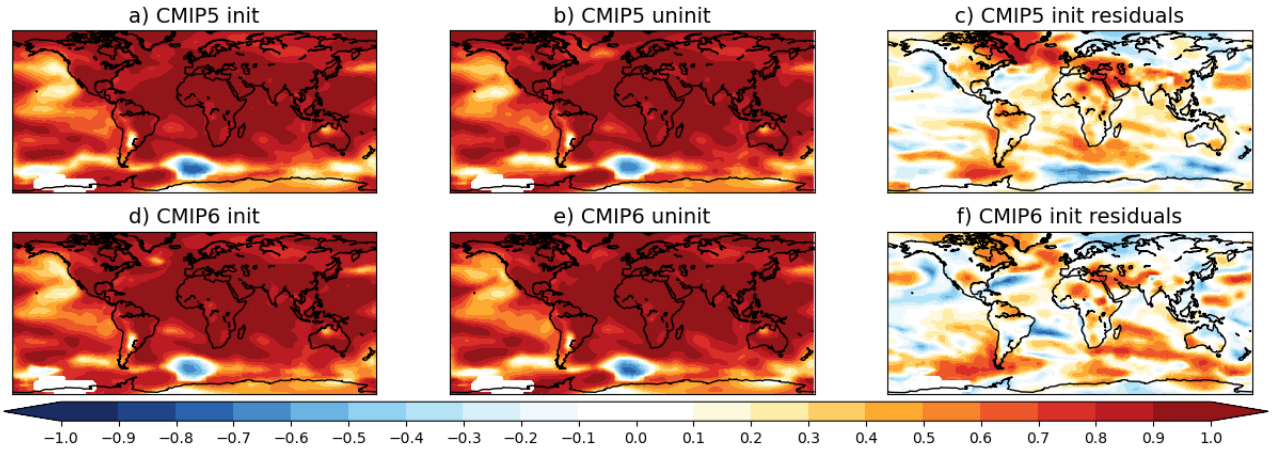

Figure 1: ACC for surface temperatures averaged for forecast times 2-9 yrs. a) CMIP5 initialised predictions (init) b) CMIP5 non-initialised projections (uninit), c) CMIP5 residuals, d) CMIP6 initialised predictions (init) e) CMIP6 non-initialised projections (uninit), f) CMIP6 residuals. HadCRUT5 is used as reference.

CMIP5 and CMIP6 decadal predictions (Figure 1a & d) as well as their projection ensembles (Figure 1b & e) show high ACC for surface temperatures averaged over forecast years 2-9 over most of the globe, except parts of the Pacific Ocean. The added value of initialisation can be quantified using the method presented in *Smith et al. (2019)*. In CMIP5 added value is most dominant over the North Atlantic region and parts of central and southern Europe (Figure 1c & f). A similar pattern is found for CMIP6; however, ACC of residuals is smaller over the North Atlantic, in agreement with *Borchert et al. (2021)* [see IPSL contribution]. Added value also is reduced over Europe. However, it needs to be noted that ensemble sizes are relatively small for all datasets (only realizations from 7 models for CMIP5 and CMIP6).

Eventually, we calculate the potential economic value (PEV) using the framework presented in *Richardson (2003)* for upper and lower tercile events (warm and cold events). This framework is based on the cost-loss decision model, in which a user experiences a certain loss (L) in case of the occurrence of an event (e.g., costs associated with Tropical Cyclone) and experiences a cost (C) if the user prepares for that event (Table 1). Please note that the assumption here is that all the potential loss (L) is prevented if taking action. The cost-loss model resembles the contingency table for a deterministic forecast (Table 1), meaning that it is possible to directly assess the costs and losses a user would experience if he would use that forecast to either take action or not take action. The monetary benefit of using the forecast over either taking action or not taking action, depends on the skill of the forecast itself (parameters a, b, c, d in Table 1) and is referred to as the potential economic value (PEV) of the forecast. Analogue to a deterministic forecast, the PEV can also be calculated for any probabilistic forecast (*Richardson, 2003*). PEV values of a forecast are scaled with the PEV of a climatological forecast/perfect forecast meaning that a value of 0 means no PEV over using a climatological forecast, whereas a PEV value of 1 indicates a perfect forecast. Each user is entirely characterized by their cost-

loss ratio, meaning that PEV is different for different users as PEV varies for different cost-loss ratios.

		Event occurs	
		Yes	No
Action taken	Yes	Cost (C)	Cost (C)
	No	Loss (L)	0
Event forecasted	Yes	a	c
	No	b	d

Table 1: Cost-loss decision framework

It is found that CMIP5 and CMIP6 predictions and projections have high PEV for upper tercile events of surface temperature (warm events) averaged over forecast years 2-9 (Figure 2). However, for residuals the PEVs are significantly lower. The PEV is linked to the reliability of the ensemble as well as to its sharpness (the range of forecasted event probabilities).

Figure 3 shows the level of skill based on 4 reliability categories ranging from dangerously useless to useful (categories adapted from *Weisheimer and Palmer, 2014*). "*Useful"* forecasts of the highest category, are those for which the uncertainty range of the reliability slope in a reliability diagram (Figure 5a) falls entirely into the area where forecast probabilities contribute to a positive Brier Skill Score (BSS; *Mason, 2004*). The raw CMIP5 and CMIP6 projections and predictions provide "*useful*" information based on these categories over large areas, except the Pacific Ocean, with the pattern resembling the one found for ACC (Figure 1). A mixed picture is found for residuals, with a "*useful*" skill score over the North Atlantic and parts of Europe. Sharpness is measured by the standard deviation of the forecast probabilities normalised with the standard deviation of a perfectly sharp (deterministic) forecast, meaning that a sharpness value of 0 indicates a climatological forecast and a value of 100 a perfectly sharp forecast. High levels of sharpness are found for CMIP5/6 predictions and projections, which together with the high reliability result in large PEV values (Figure 4). For residuals, the sharpness is reduced substantially, also over the North Atlantic and the European region.

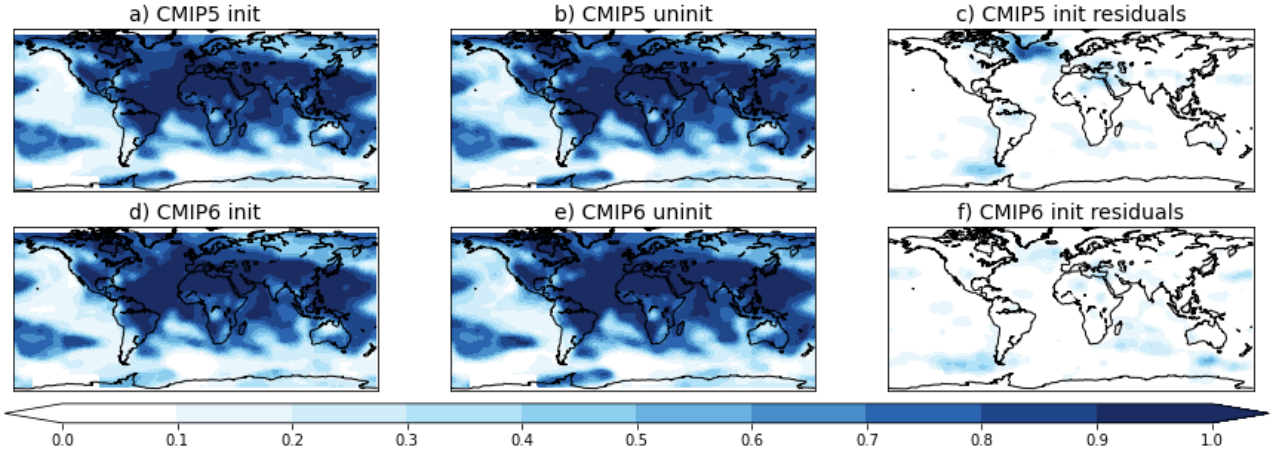

Figure 2: PEV for surface temperatures averaged over forecast time 2-9 years for upper tercile events.

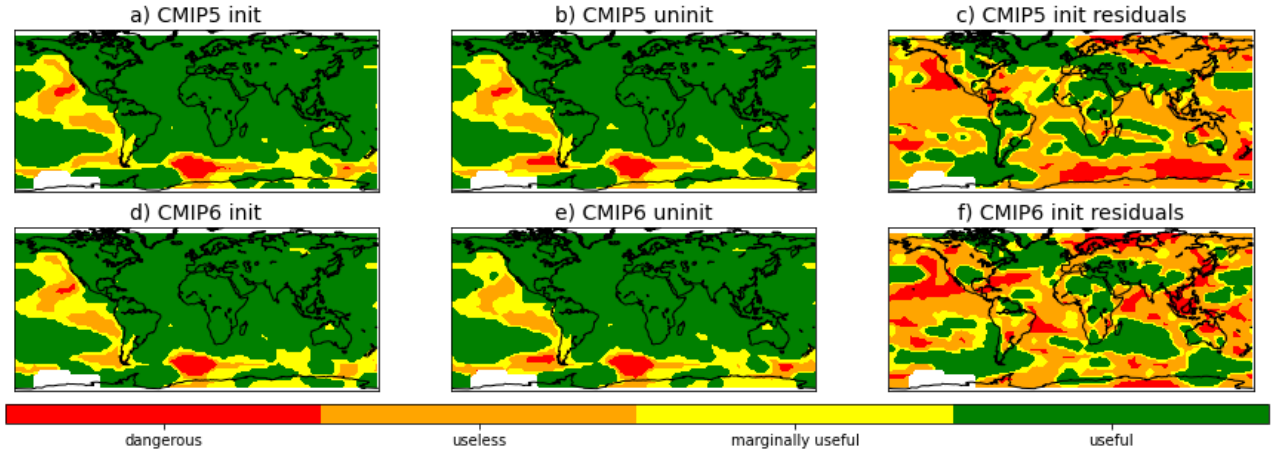

Figure 3: Reliability categories for upper tercile events of surface temperatures averaged over forecast time 2-9 years (categories adapted from Weisheimer and Palmer, 2014).

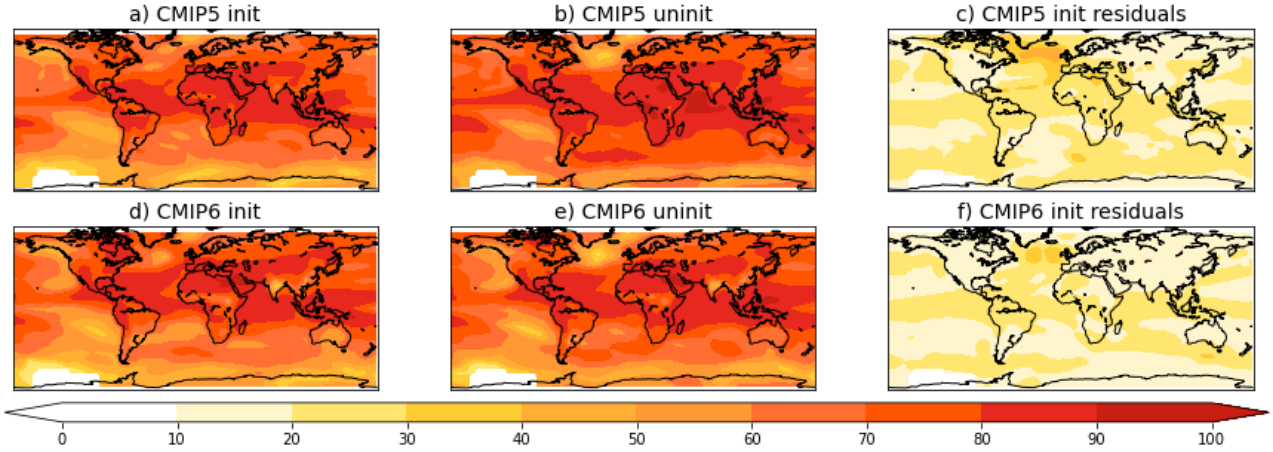

Figure 4: Sharpness for upper tercile events of surface temperatures averaged over forecast time 2-9 yrs.

These findings indicate that the reduction in PEV over the North Atlantic and the European region mainly results from a lack of sharpness of the residuals rather than a lack of skill in the reliability. Figure 5 illustrates reliability, sharpness and PEV for the North Atlantic Subpolar Gyre region (50-10W; 45-60N) in CMIP6 initialised predictions and their residuals. Reliability is very similar for both ensembles, with both being underconfident/overdispersive but adding positive to the BSS. For CMIP6 initialized residuals the forecasted event probabilities are centred around the climatological tercile event frequency (0.33), whereas for the initialized predictions forecasted probabilities are more evenly distributed between 0 and 100%. The higher sharpness of the latter results in a much larger PEV (Figure 5c)

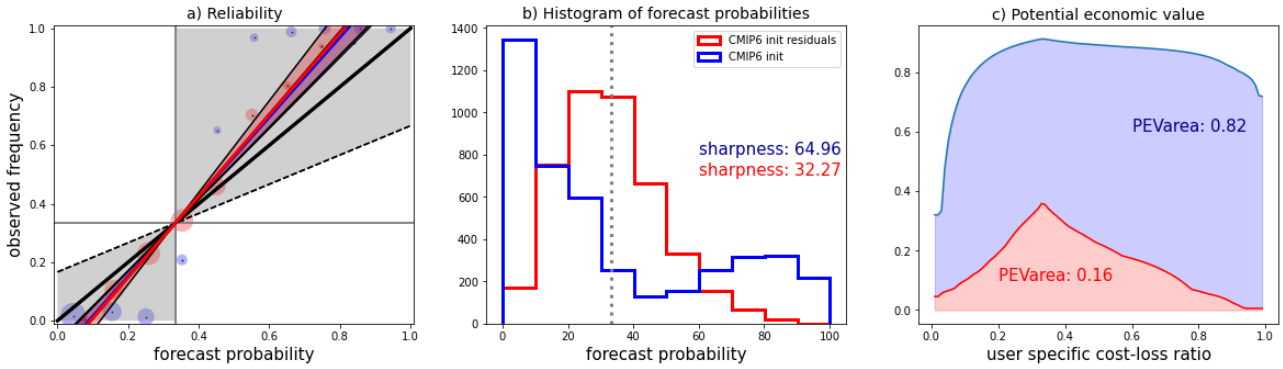

Figure 5:a) Reliability diagram, b) histogram of forecast probabilities and c) PEV for initialized CMIP6 predictions and their residuals for upper tercile events of surface temperatures averaged over forecast time 2-9 yrs. For a perfectly reliable ensemble forecast probabilities match observed frequency (regression line has a slope of 1), whereas regression lines with slopes < 1 indicate overconfident and slopes >1 underconfident forecasts. If the uncertainty of reliability slope falls completely into the grey shaded area in a), the reliability is categorized as "useful" in Figure 3. The histogram of forecast probabilities b) shows the distribution of the number of occurrences the ensemble issues a forecast of the tercile event with a certain probability. For a specific initialization, a probability of 0 indicates that each member of the entire ensemble simulates no event for, whereas a value of 1 indicates that the entire ensemble forecasts the event. For a deterministic forecast only forecast probabilities of 0% and 100% are possible, whereas a climatological forecast only issues probabilities matching the climatological event rate of occurrence (1/3 for tercile events; see grey dotted line in b). Here sharpness is defined such that a deterministic forecast has a sharpness of 100, whereas a climatological forecast has a sharpness of 0. In c) the PEV is shown for different cost-loss ratios (characterizing different users). Small cost-loss ratios indicate users for which preparation costs for an event are lower than losses associated with the event (if no action has been taken prior). Vice versa for high cost-loss ratios, which indicate users for which costs associated with preparing for an event are of similar magnitude as the losses associated with the event (with no prior action).

Analyses have been carried out for different seasons and also for precipitation. For precipitation, PEV is much smaller than that found for surface temperatures, in line with smaller deterministic skill. It is planned to summarise these findings in a scientific publication.

2. Collaboration SMHI/UOXF: probabilistic skill assessment based on novel temporal pooling approach with respect to decadal climate predictions of probabilities for seasonal temperature and precipitation extremes [this work is also part of WP1, Deliverable D1.2]

   UOXF further contributed to the probabilistic skill assessment based on a novel temporal pooling approach that has been developed at SMHI. The peculiarity of this approach is that climate prediction information of consecutive years is not averaged over time as usually done. Instead,

the three-month averages over specific seasons of consecutive years are pooled together and treated as exchangeable within the respective pooling window.

The primary motivation behind this approach is to derive a completely new kind of forecast information, complementary to standard multi-annual averages and hence, potentially useful for different types of stakeholders. Thus, this approach addresses users interested in the probability of the occurrence of extreme seasons within the next few years rather than information on a multi-annual average.

This alternative approach of analysing climate predictions has already been suggested by *Fricker et al. (2013)* but - to our knowledge - has never since been further applied in the context of decadal climate prediction. Hence, our study - which is illustrated with an example in the following - is the first effort to implement this approach in a comprehensive manner for a large multi-model ensemble (MME) of decadal climate predictions.

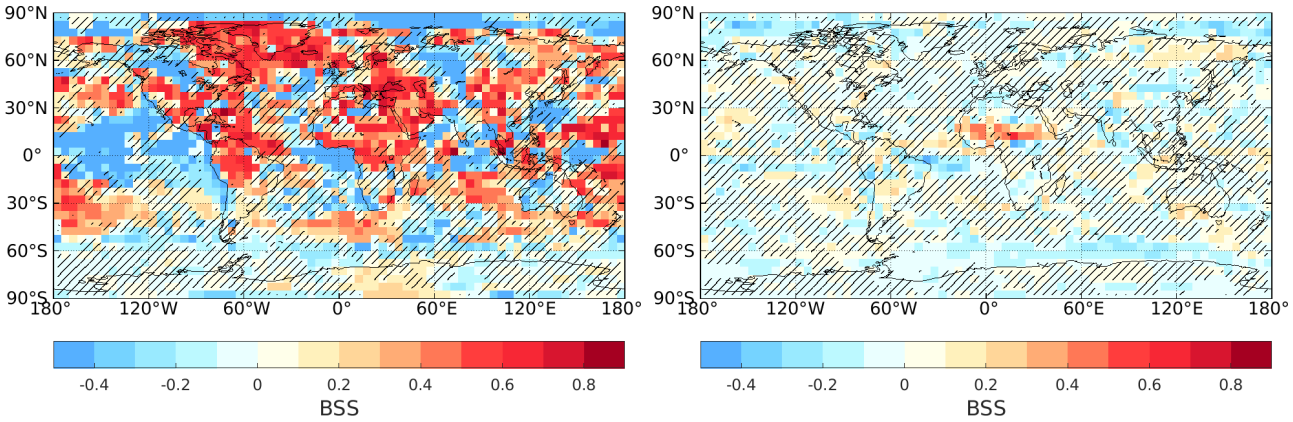

Figure 6: Brier Skill Score (compared to a reference prediction using climatological probabilities, i.e. 1/6 in every year) for CMIP6-DCPP multi-model ensemble in predicting the probability a boreal summer (JJA) within the next five years after initialization being extremely hot (left: 2m air temperature within local upper sextile) and extremely dry (right: total precipitation within lowest sextile); skill assessment (based on ERA5 and GPCPv2.3 as observational references) for evaluation period 1979-2014 based on 32 hindcasts s1978-s2009 from 8 different models with a total of a 108 ensemble members; hatching masks regions where BSS is not significant (p>0.01)

Figure 6 shows a first evaluation of skill when making use of the temporal pooling approach for the multi-model ensemble of currently available CMIP6-DCPP-hindcasts to forecast probabilities of boreal summers (JJA) within five years after initialisation being extremely hot (2m air temperature within upper sextile; Figure 6 left) and extremely dry (total precipitation within lowest sextile, Figure 6 right). Positive values of the Brier Skill Score (BSS) indicate that the multi-model probability forecast is more skilful than a climatological forecast.

The MME offers skill compared to the climatological reference forecast for extremely high summer temperatures over large parts of the Americas, Greenland and the North Atlantic, Europe, and Africa. The respective forecast for extremely dry boreal summers however lacks skill for most parts of the globe. The only positive exception is the Sahel region. The general skill pattern is qualitatively in very good agreement with BSS-results for less extreme thresholds, although lower and with more insignificant areas. Still this result confirms that it is possible to derive robust probabilistic predictions for such seasonal extremes (at least temperature-related)

and many parts of the globe when making use of our novel temporal pooling approach and therefore provide new climate prediction information useful for potential user requirements beyond the standard multi-year averages. The analyses are currently extended by including also CMIP6 projections and a scientific publication of this approach and its application is currently in preparation.

3. Representation of model uncertainty in multiannual predictions

This is cross-cutting work on the representation of model uncertainty in multi-annual predictions extending results presented in MS4 in WP1. To obtain reliable predictions on any time scale it is inevitable to account for model uncertainties caused by unresolved processes. One prominent way to do this is by combining simulations from different models into a MME. However, in numerical weather prediction, it has been shown that using stochastic physics, which aims to represent the effect of the unresolved processes, is another possibility to account for model uncertainties within a single model. Here we compare the two different approaches developed to account for model uncertainties, namely stochastic physics and multi-model ensembles. The work has recently been published in GRL (*Befort et al., 2021*).

Two hindcasts have been conducted using ECMWFs coupled model system. The setup of the two hindcasts is identical: i) 28-month forecast, ii) initialized in November 1981 to 2014, iii) 10 members. The only difference between both hindcasts is that one includes the stochastic physics perturbed tendency scheme (SPPT) whereas this scheme is switched off in the other hindcast simulation (ECMWF-noSPPT). Data from 5 different decadal prediction systems are used: NCAR-DPLE, EC-Earth, MPI-ESM-1-2-HR, MIROC6, HadGEM3-GC31-MM. These are all initialized in November and consist of 10 ensemble members (for NCAR-DPLE only the first 10 members are used). All ensembles are corrected for lead-time dependent biases. From these a 10-member MME is built using 2 members from each single model ensemble. The sensitivity to the choice of ensemble members is assessed by using a 10000-sample bootstrap.

Skill of all ensembles is assessed using anomaly correlation coefficient (ACC), whereas reliability is measured using the spread over error statistic (SoE), defined by the ratio of the average ensemble spread and the root mean square error. For a reliable ensemble, SoE equals 1, whereas values below 1 indicate overconfident and values above 1 indicate underconfident ensembles.

It is found that reliability of surface temperatures is largely improved for the MME and ECMWF-SPPT compared to ECMWF-noSPPT up to 28 months, showing that both approaches are suitable to account for model uncertainties. Largest improvements are found over the tropics (see Figure 7 for ENSO region), which is in line with results for seasonal forecasts (*Weisheimer et. al, 2011*). Besides higher reliability in ECMWF-SPPT, there is also more skill until about the 2nd winter for surface temperatures over the NINO3 region. Due to the teleconnections between tropical Pacific Ocean SSTs over this region and sea-level pressure (SLP) over the North Pacific (*O'Reilly, 2018*), we analysed the skill and reliability of the North Pacific Index (SLP over 180-120W; 30-65N). Significant skill is found for all ensembles until the first summer. In contrast to the MME and the ECMWF-noSPPT, skill returns in the 2nd winter in ECMWF-SPPT, which is probably related to the higher skill and improved reliability for SSTs over the NINO3 region. This study

motivates the more widespread use of stochastic physics in climate predictions on multiannual time scales, especially as stochastic physics is computationally cheap.

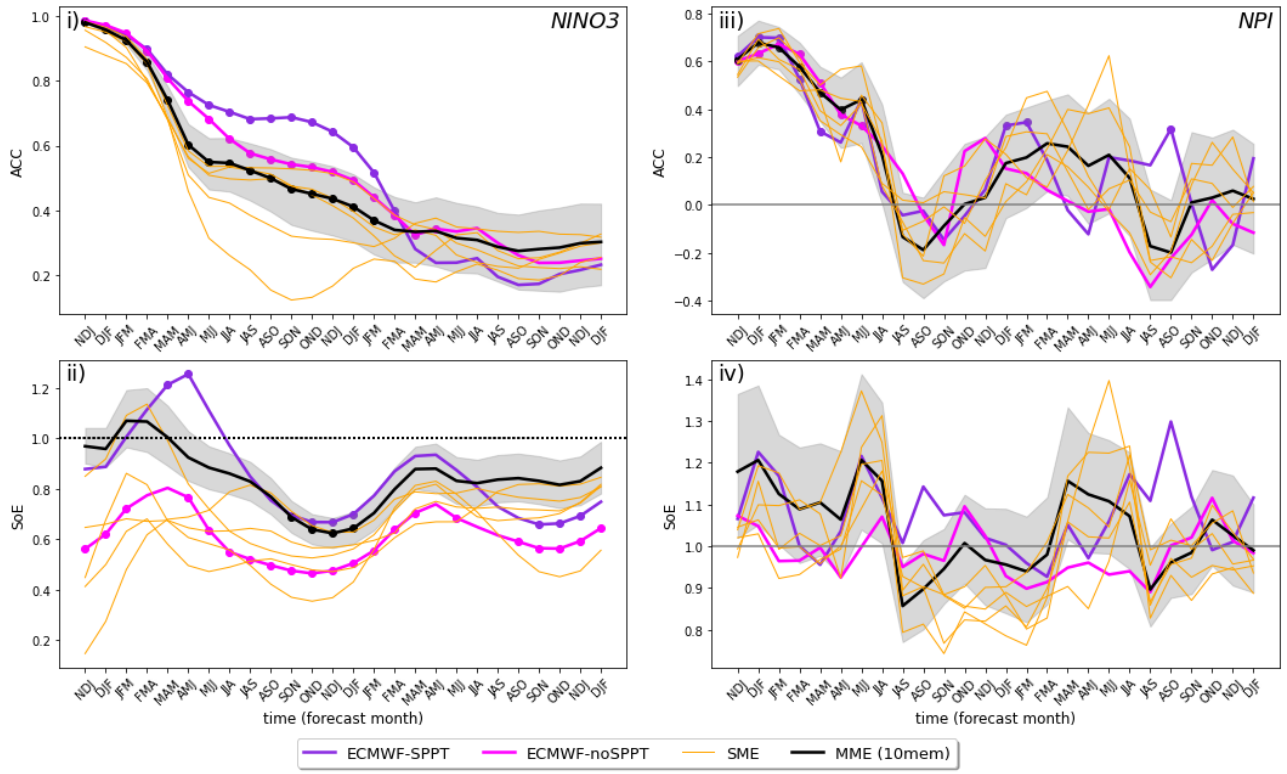

Figure 7: i) Anomaly correlation coefficients for SSTs over NINO3 region using ERA5 as reference, ii) same as i) but for SoE, iii) same as a) but for the North-Pacific Index (NPI), iv) same as iii) but for SoE. Grey shading for the MME indicates 2.5 and 97.5 percentile derived from randomly sampling (10000 samples) 2 members from each single model ensemble. Dots in i) & iii) indicate forecast times for which the respective ensemble is significantly larger than 0, whereas dots in ii) & iv) indicate forecast times for which the respective ensemble is significantly different from 1 (95% confidence, 10000 samples). Samples have been generated by bootstrapping over years for ECMWF-SPPT and ECMWF-noSPPT ensembles and over years and members for the MME. Figure adapted from Befort et al. (2021).

Barcelona Supercomputing Center

1. Evaluating the reliability of decadal predictions and projections

Reliability is an essential characteristic of climate simulation ensembles, quantifying the agreement between the predicted probabilities and observed relative frequencies of a given event. Reliability is therefore a key requirement for the predictions to be useful to decision-makers, who base their decisions on the prediction of certain event types.

We analyse aspects of reliability in initialised decadal predictions (INIT) in comparison to non-initialised projections (NoINIT), for near-surface air temperature. We use multi-model decadal predictions and projections from 12 different climate models (110 members in total, see *Verfaille et al, 2021*) evaluated against observational data. The reliability assessment was carried out using rank histograms (*Elmore, 2005*), test statistics from *Jolliffe and Primo (2008)* displayed on global maps and regional time series for 30 different regions around the world. Rank

histograms are used to assess if the ensemble members and the verifying observation stem from the same probability distribution (i.e. if the observations are predicted as the equiprobable members), in which case the forecast ensemble is considered reliable and the rank histogram is flat. In addition to the qualitative information provided by a visual inspection of the shape of rank histograms, information on the forecast deficiencies can be further quantified using goodness-of-fit test statistics. The histogram of an ensemble forecast system and the corresponding observational verification of an ideal system produces a flat or uniform histogram. However, because of sampling variation the histograms are almost never exactly flat. The question then arises: can observed deviations from "flatness" or uniformity be attributed to chance, or do they indicate deficiencies in the forecasts? An overall test of uniformity is provided by the χ2 goodness-of-fit test. The χ2 test statistic can be decomposed into components (*Jolliffe and Primo, 2008*) that indicate whether the forecasts are biased (Jolliffe-Primo test statistic for slope - JP slope), whether they are over- or underdispersed (Jolliffe-Primo test statistic for convexity - JP convexity), and whether there are any other deviations from flatness, once these two possibilities are accounted for. Other decompositions are also possible. Note that the statistic and its decomposition does not target the forecast pdf nor assesses the adequacy of its sharpness in the sense the resolution component of the Brier score does. We also explored how reliability evolves with forecast time, by looking at results for forecast year 1 and forecast years 1 to 5, over the period 1961-2010. Finally, we tested the impact of applying several post-processing techniques to the "raw" temperature anomalies, thereby showing that all forecast system ensembles have issues with reliability, regardless of whether they are predictions or projections, and that bias correcting and calibrating them is fundamental to obtain reliable predictions/projections of the future climate conditions (see *Verfaille et al., 2021)* for more detailed descriptions of the methods).

I. Reliability of uncorrected simulations

The results indicate that in both initialised (INIT) and non-initialised (NoINIT) ensembles in most regions are not significantly reliable. The JP slope and JP convexity coefficients are expressed as their contribution to the X^2 coefficient, a large contribution (purple colours in, left panels) indicating some deviation from flatness mainly due to the slope or the convexity, respectively. For the difference in the contribution to the X^2 coefficient, blue colours in Figure 8 (right panels) indicate a lower contribution of JP slope or JP convexity for INIT than for NoINIT, thus some added-value of INIT over NoINIT.
Neither uncorrected INIT nor uncorrected NoINIT provide significantly reliable estimates, i.e, flat rank histograms (the X^2 p-value is never above 0.05), for near-surface temperature and forecast year 1. For most regions, this is because either the slope parameter or the convexity parameter or both parameters are significantly contributing to the X^2 coefficient, resulting in unreliable estimates. In general, INIT shows some added-value over NoINIT mainly in terms of the convexity coefficient in the large multi-

model ensemble (Figure 8 right panel), with some discrepancies depending on the regions. Also, for the forecast years 1-5 the ensembles provide generally no reliable simulations of regional temperatures. The difference between INIT and NoINIT is generally smaller for forecast years 1-5 than forecast year 1, indicating no measurable added value from initialisation on this prediction time scale.

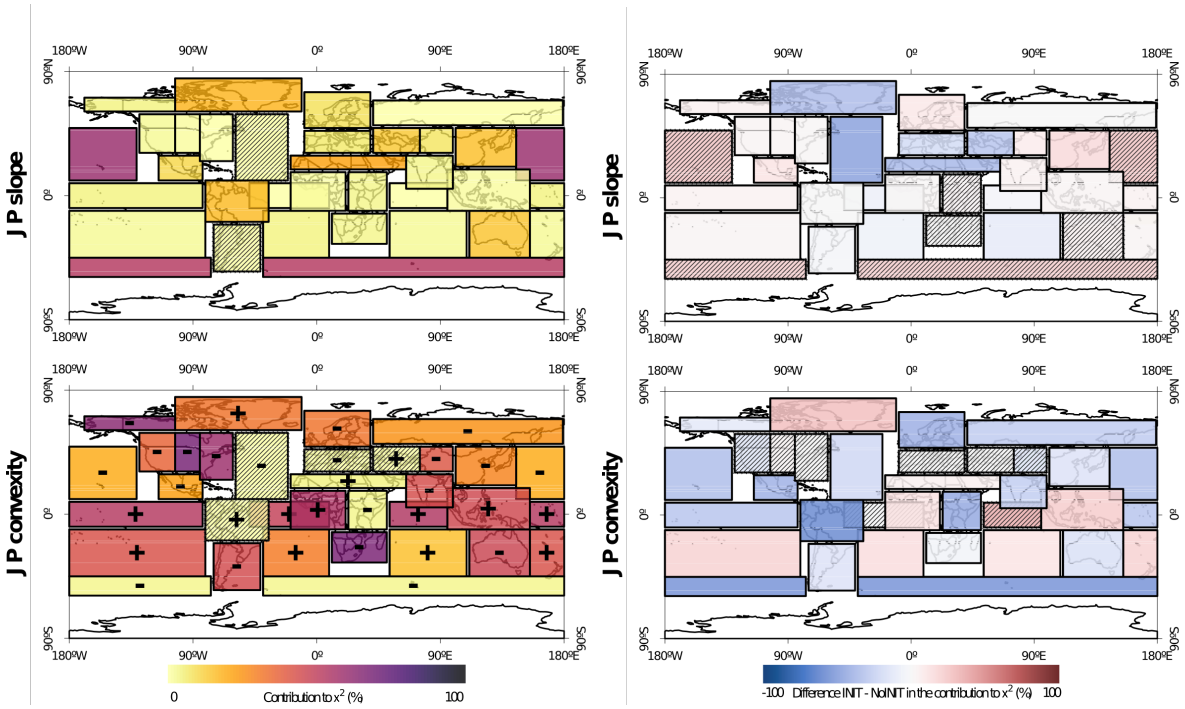

Figure 8: [left] Maps of the Jolliffe and Primo (2008) slope (top) and convexity (bottom) coefficients, expressed as their contribution to the X² coefficient (in %), for near-surface air temperature for the 30 different regions considered in this study, for forecast year 1 in the INIT MM ensemble. Going from light yellow to dark purple, the colours denote an increasing role of the slope and the convexity terms to decreasing the reliability of the ensemble (diagnosed by the deviations from flatness in the rank histogram). A plus (minus) sign in the convexity coefficient maps represents an underdispersive (overdispersive) forecast. Hatching represents regions where the p-value is larger than 0.05, thus where there is no evidence of bias, difference in trend, nor error in dispersion (the null hypothesis being that the rank histograms are flat).
[right] Difference between INIT and NoINIT slope (top) and convexity (bottom) coefficients for near-surface air temperature, for forecast year 1 in the MM ensemble. Going from dark blue to dark red, the contribution of JP slope or JP convexity for INIT becomes increasingly larger than for NoINIT (i.e., INIT becomes less reliable than NoINIT). Hatching represents regions where the difference is not significant at the 95% level.
Figure taken from Verfalle et al. (2021), copyrights: American Meteorological Society.

II. Reliability after de-trending, bias correction and calibration

Given the unreliability of the raw model simulations, we also tested the possible corrections from different post-processing techniques (all results summarised in Figure 9). The JP statistics for the detrended multi-model ensemble are shown in Figure 9 (top triangle of the matrix fields). The results show that detrending the data improves the JP slope coefficient, which becomes reliable in many regions, with differences between INIT and NoINIT very close to zero. On the other hand, it degrades the JP convexity coefficient in many regions. In a few regions, especially Southeast Asia (SEA), detrending increases

the added value of INIT over NoINIT, but the location of regions with added value of INIT over NoINIT differs between different sub-ensembles (see *Verfaille et al., 2021*). There is no region that becomes significantly reliable (i.e., with a X^2 p-value above 0.05, the null hypothesis being that the rank histograms are flat) after detrending the data, which suggests that the lack of reliability cannot be only due to a misrepresentation of the observed trend.

The simple bias correction (correcting mean and variance, bottom triangle in Figure 9) also improves the JP slope coefficient, increasing reliability for many regions. Like detrending, it removes almost all the differences in terms of JP slope coefficient between INIT and NoINIT. However, unlike detrending, it does not systematically increase the contribution of the JP convexity coefficient. Especially for NoINIT, for which the JP convexity coefficient was generally worse than for INIT (Figure 9), bias correcting improves this coefficient. This implies that the difference between INIT and NoINIT convexity coefficients after bias correction is close to zero. Similar to detrending, bias correction does not produce any significantly reliable region, i.e., with a X^2 p-value above 0.05 when the null hypothesis is that the rank histograms are flat, indicating that errors in the mean variance do not play a significant role in the lack of reliability of the ensembles.

We also tested the effects of calibration (Figure 9, right triangle of the matrix fields). As for detrending and bias correction, calibration improves greatly the JP slope coefficient, which becomes significantly reliable in many regions. Additionally, it also generally improves the JP convexity coefficient, although not in all regions. For example, the North Atlantic Ocean (NAT) and Mediterranean Basin (MED) regions, which already had low contributions to unreliability of the JP convexity coefficients in the uncorrected multi-model INIT dataset, display higher values for the contribution of convexity after calibration (thus contributing more to unreliability). For the NCAR single-model large ensemble (not shown), the improvement in the JP convexity coefficient for INIT is generally larger than for the MM ensemble. Calibrating the various forecast system ensembles yields significantly reliable results in one region (Central America, CAM) for INIT. For forecast years 1 to 5, the convexity results after calibration are generally worse than for the uncorrected ensembles, while the slope results remain good (Figure 9). The difference between INIT and NoINIT JP convexity after calibration is slightly more positive for forecast years 1-5 than for forecast year 1 even though the values remain very close to zero. Calibration is the only post-processing method that yields significantly reliable ensembles for one region, indicating that errors in the ensemble spread play a significant role in the lack of reliability of the forecasts for this region.

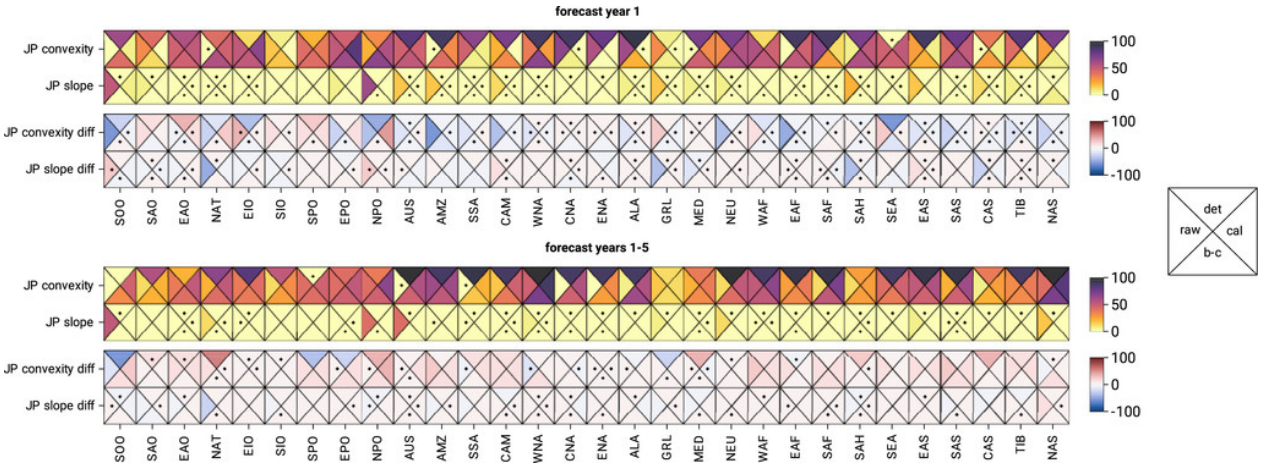

Figure 9: Summary of the Jolliffe and Primo (2008) slope and convexity coefficients, expressed as their contribution to the X^2 coefficient (in %), for near-surface air temperature for the 30 different regions used in this study, for forecast year 1 and forecast years 1-5 in the MM ensemble. For each forecast time, the top two rows represent the JP coefficients for INIT. Diamonds indicate cases where the p-value is larger than 0.05, thus where there is no evidence of bias, difference in trend, nor error in dispersion (the null hypothesis being that the rank histograms are flat). The bottom two rows for each forecast time represent the difference between INIT and NoINIT. A diamond indicates a non-significant difference at the 95% level. For colour codes, please refer to Fig. 9. Each triangle displays the result for a type of post-processing (either raw uncorrected values, det = detrended, b-c = bias-corrected, or cal = calibrated). Figure taken from Verfalle et al. (2021), copyrights: American Meteorological Society.

III. Main conclusions

Results indicate that both INIT and NoINIT uncorrected output ensembles are largely not reliable, and there is only a rather limited added value of decadal predictions for near-surface air temperature in terms of reliability, compared to non-initialised projections. Indeed, the added value is limited to specific regions and to the first forecast year(s), similar to skill measures of forecast accuracy (e.g., *Doblas-Reyes et al., 2013; Yeager et al., 2018; Smith et al., 2019*). Furthermore, using different forecast system ensembles has an impact on reliability, but the model combination inside the ensemble seems to play a larger role than the actual number of ensemble members. As such, we have shown that it is of advantage to use ensembles composed of different forecast systems, as those encompass a larger range of model physics and initialisation approaches, and thereby also allow for error compensation. Most importantly, this study has demonstrated the need for bias correction and calibration of the raw data. This is crucial to obtain reliable predictions and projections of climate that can be useful to stakeholders to obtain more realistic estimates of event probabilities.

2. Evaluating added skill from initialisation in perfect-model predictions

We have also explored the added value of initialisation in perfect-model decadal predictions (*Liu et al., 2019*). The idea for this work is to determine the model-specific predictability in the absence of effects that deteriorate the skill in predictions of the real-world climate (these are primarily the limited knowledge of the initial states due to incomplete observational coverage, and initialisation shocks as a consequence of inconsistencies between model climate and real-world climate).
To this end, we initialised a 5-member ensemble of decadal runs from each year of a historical climate simulation, and evaluated the skill of the decadal runs in predicting the historical run from which they were initialised. These simulations were performed with the CESM1 coupled climate model.

I. Perfect-model skill and added skill from initialisation

The non-initialised perfect model shows significant skill over parts of Middle Asia, the Indian Ocean, the West Pacific, North America, and most of the Atlantic in lead years 1 and 2 (Figure 10). The skill improves considerably in lead years 2–5 and 2–9 in large areas of the globe. After initialisation, most areas of the globe show significant skill in lead year 1, but the skill in lead year 2 relative to year 1 reduces particularly over the Pacific (Figure 10). The skill for lead years 2–5 and 2–9 increases almost everywhere, and the general skill patterns of the initialised runs are very similar to the case with no initialisation. The regions in Figure 10 where the skill of the initialised runs is improved over the uninitialised runs are mostly in the tropics (in the first 2 years) and the North Atlantic.
The difference in skill between the initialised perfect model and the uninitialised perfect model decreases with lead time: In lead year 1, 25% of areas have significant skill contributed by initialisation, followed by 11% in lead year 2. The number drops below 3% in lead years 2–5 and then slightly increases to 7% in lead years 2–9.

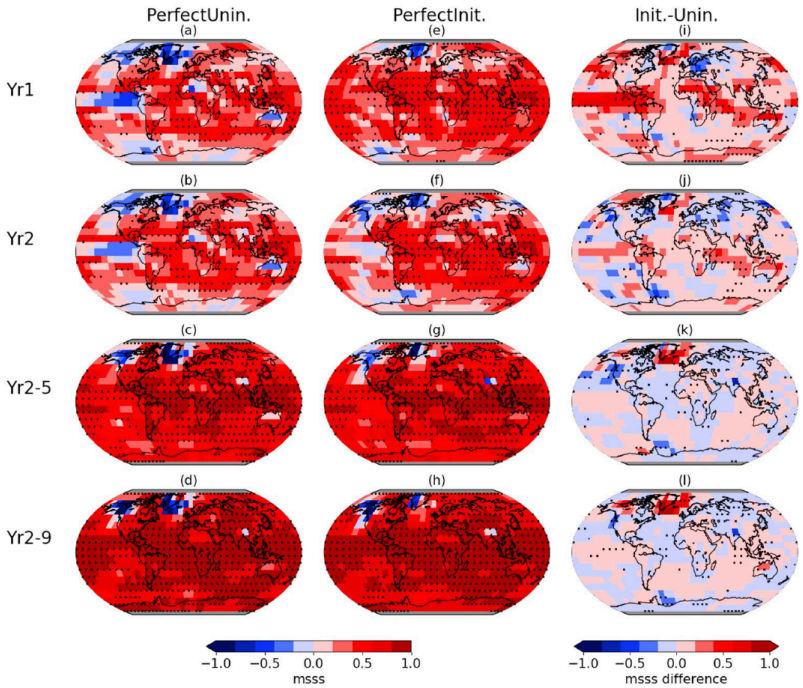

Figure 10: The mean squared skill score maps of the mean near-surface temperature for the Community Earth System Model (CESM) (a–d) uninitialised perfect model, (e–h) the initialized perfect model; and the (i–l) difference map between left and middle columns. The first row shows the skill for lead year 1, the second-row lead year 2, the third row lead years 2–5 average temperatures, and the fourth row lead years 2–9 average temperatures. Stippling indicates grid cells where the skill score or skill differences are significant at $p \leq 0.05$ (two-tailed), based on 200 bootstrap realizations. Figure taken from Liu et al. (2019), copyrights: American Geophysical Union/Wiley.

II. Comparing perfect-model predictability and skill in decadal predictions of the real-world climate

We next compared the potential skill of the model in predicting its own climate evolution with the skill of decadal predictions of the observed climate. To this end we used the decadal predictions and historical simulations provided by the same model within CMIP5.

The non-initialised simulations show significantly increased skill in predicting an independent realisation of the same model ('uninitialised perfect model') than in predicting the observed climate in particular over large parts of the Pacific for the forecast times 2-5 years and 2-9 years (Figure 11). This points to different long-term changes between model and observations, either due to inconsistent responses to forcing or different representations of multi-decadal variability in the model. In general, the model simulations tend to show too strong warming in this region compared to observations.

Comparing the initialised decadal perfect-model predictions and the hindcasts of the observed climate, we find very similar patterns of skill differences as in the non-initialised perfect model, where the decadal temperature predictions in the perfect model have higher skill than the real-world predictions in particular over the Pacific. This similarity in

skill differences for the initialised and non-initialized predictions led us to conclude that the limited skill in the Pacific region is primarily a consequence of inconsistencies between the model and the real-world. This suggests that improving the representation of Pacific response to forcing or multi-decadal variability will have the potential to improve decadal predictions of the real-world climate.

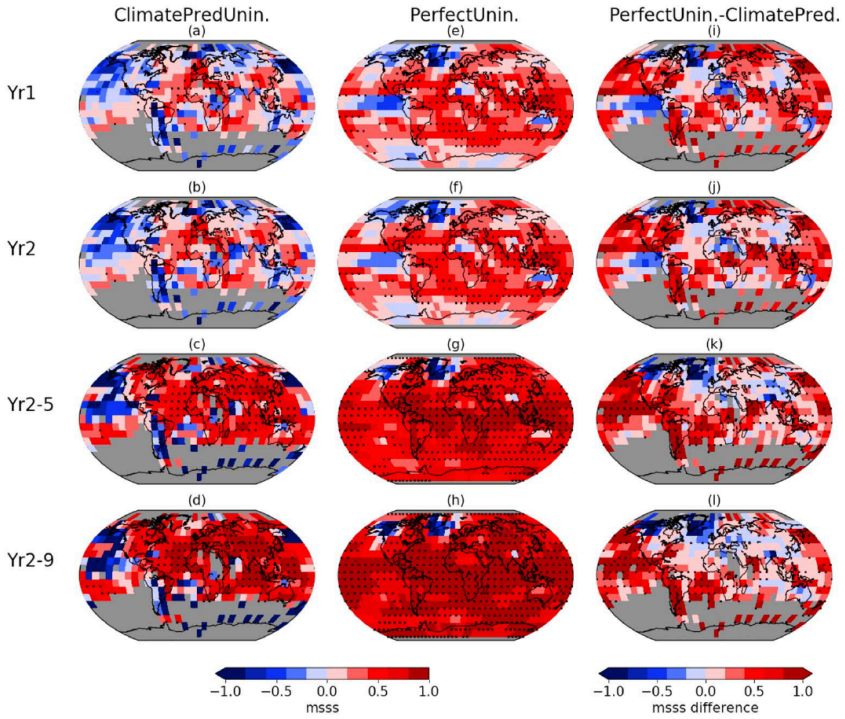

Figure 11: Same as Figure 10 but for the (a–d) uninitialized real-world climate predictions (left column) and the (e–h) uninitialized perfect model (same as Figure 10 left column). (i–l) Difference between perfect model and real-world predictions. Figure taken from Liu et al. (2019), copyrights: American Geophysical Union/Wiley.

III. Conclusions

Applying ideal initialisation (in terms of perfectly known initial state that is fully consistent with the model's climate attractor) in perfect-model predictions with the CESM1 model shows that initialisation improves skill in large areas of the globe in particular in the first 2 forecast years (*Liu et al., 2019*). Comparing the skill of perfect-model predictions with the skill of real-world hindcasts with the same model, we conclude that the decadal real-world predictions can be potentially improved in particular in the Pacific regions. From comparing initialised and non-initialised predictions we identify that the current lack of skill in this region is likely a consequence of the model showing long-term changes in the Pacific that are inconsistent with the observed climate.

We also plan to run a similar set of perfect-model predictions with the EC-Earth3 model, to understand the model-dependence of predictability. An experiment that we already performed for this purpose was suffering from errors in the initial condition files and needs to be rerun.

3. Other studies in progress to evaluate initialised predictions compared to and uninitialized projections

I. Decadal predictions of Atlantic-European Weather Regimes

We have performed a detailed evaluation of the representation and predictability of Euro-Atlantic Weather Regimes in the CMIP6/DCPP-A decadal hindcasts and projections performed with the EC-Earth3 model (10 members in both the initialised and non-initialised ensembles). For this work, we classified the large-scale atmospheric pressure patterns (using mean sea level pressure, MSLP) in the Euro-Atlantic domain into clusters representing weather regimes (WR). We used the k-means clustering algorithm, and following previous work that identified 4 clusters being a good representation of the flow regimes in this region, used this algorithm to group each day into one of the 4 patterns: Positive NAO (NAO+), negative NAO (NAO-), Atlantic Ridge (AR), and Blocking (BL). We evaluated the WR for the summer (JJA) and winter (DJF) seasons, as well as extended summer (MJJASO) and extended winter (NDJFM) seasons.

We find that both the initialised predictions and non-initialised simulations with EC-Earth3 show a good representation of climatological features (i.e., patterns and mean frequencies) in particular for the NAO+, NAO- and Blocking regimes. The annual to decadal variations of WR are found to be mostly not predictable, but there is weak indication for moderately skilful predictions of blocking during summer for forecast years 4 to 5 in agreement with *Athanasiadis et al. (2020)*. The late emergence of weak skill may be due to strong SST drift deteriorating the representation of teleconnections at shorter forecast times.

This work is currently being revised for resubmission to JGR-Atmospheres. In particular we are working to address sensitivities related to the clustering method and the relatively small ensemble size for predicting noisy fields such as extra-tropical atmospheric circulation patterns.

II. Evaluation of forecast quality of European winter windstorms in a multi-model decadal prediction framework

We assess the forecast quality of windstorm events for different forecast times of 5 decadal prediction systems contributing to CMIP6/DCPP-A which provide sub-daily wind speed data required for the windstorm tracking (*Leckebusch et al. 2008*). This particular tracking algorithm has the benefit of low computational costs. Despite the high impact of windstorms over Europe and previous encouraging results (*Kruschke et al. 2015*) only few efforts have been made recently in analysing decadal prediction skill of windstorm frequency (*Schuster et al., 2019*) or intensity. Studies regarding decadal prediction of windstorms from a multi-model perspective are completely missing.
We find that the number of winter (DJFM) windstorms as captured by the decadal prediction hindcasts is two-fold. The skill analyses reveal regions with anomaly correlations of up to 0.5 over some parts over the North Atlantic, especially north of 55°N in the Nordic Sea for forecast years 2-5. However, the multi-model hindcast shows no or very little skill over the central North Atlantic, the region with climatologically the highest number of windstorm events. Over the European continent and primarily Central Europe, the region with highest interest from an impact perspective, the multi-model also shows positive skill for the forecast range 2-5 years.
Over almost the entire North Atlantic and most parts of continental Europe the multi-model ensemble of the non-initialised projections also shows positive correlation values due to external forcings such as transient greenhouse gas values. The greatest correlation values over land can be seen over the Iberian Peninsula for both initialised and non-initialised. Windstorm events in this region are usually associated with a negative phase of the NAO.

Current focus of ongoing work is on the analysis of the development of the skill over different forecast years and the assessment of deviations of individual models from the multi-model.

CNRS-IPSL

The added value of initialisation for decadal (1-10 yrs) North Atlantic SST predictions is analysed for CMIP5 and CMIP6 with particular attention to the subpolar gyre (SPG) region, highlighting advances in most recent model development. Using a total of 58 global climate models (30 from CMIP5 and 28 from CMIP6; 6 and 7 initialised model systems, respectively), the representation of decadal variations of SPG SST is compared between initialised predictions and non-initialised historical simulations across CMIP5 and CMIP6, and for different time periods between 1961-2014. This research was recently published in Geophysical Research Letters (*Borchert et al., 2021*).

CMIP6 models are generally more skilful in reproducing observed SPG SST variations (Figure 12). This is true for initialised and non-initialised simulations. Specifically, the characteristic "skill-hole" in the SPG region in historical simulations that was consistently found in non-initialised historical simulations with CMIP5 models (Figure 12b) and fixed through hindcast initialisation (Figure 12a) (e.g. *Marotzke et al., 2016*) is not found in CMIP6 models, with or without initialisation (Figure 12 c,d). Among other things, our work indicates that the added value of initialisation for predictions of SPG SST, previously a poster child for the need for initialisation to achieve skilful decadal predictions, is strongly reduced in CMIP6. These findings indicate a stronger role of forcing for observed SPG SST variations than previously thought, and an increased capability of CMIP6 models to reproduce these variations compared to CMIP5, both of which we analyse more closely in the following.

Drawing on the time-dependence of hindcast skill, i.e. the change of skill over time which is often called windows of opportunity (*Mariotti et al., 2020; Borchert et al., 2019a*), we illustrate phases in the period 1965-2014 for which CMIP5 and CMIP6 initialised and non-initialised models are particularly capable of reproducing observed SPG SST changes (cf. Figure 2a in *Borchert et al., 2021*). We use correlation analysis to find that both model generations capture SPG SST variations after ~1980 particularly well, but CMIP6 models show higher skill than CMIP5 models. The post-1980 period thus explains much of the improvement from CMIP5 to CMIP6. This finding is only partly explained by increased ensemble size in CMIP6, indicating a profound underlying physical reason for this improved skill in CMIP6 (cf. Figure 2b in *Borchert et al., 2021*).

Using a 9-member ensemble from the Detection and Attribution MIP (DAMIP; *Gillett et al., 2016*) contribution to CMIP6, we assess the contribution of different forcings to the full signal in the historical CMIP6 simulations, to find the forcing that is responsible for the high post-1980 skill in CMIP6 historical simulations. Our findings indicate that anthropogenic aerosol and greenhouse gas forcings have a limited impact on observed SPG SST changes during that time, explaining 0% and 16% of the observed variability based on correlation analysis, respectively. On the other hand, natural forcing explains 55% of the observed SPG SST variations, which is thus identified as the main forcing responsible for the high skill (cf. Figure 3 in *Borchert et al., 2021*). Out of the two forcing factors included in natural forcing in CMIP6, volcanic and solar forcing, we find volcanic forcing to be consistent with the high historical SPG SST skill we find. Indeed, a simple model of harmonic Atlantic meridional overturning circulation (AMOC) response to major volcanic eruptions (*Swingedouw et al., 2015*) shows a robust lagged relationship between AMOC and SST 10 years later after 1980 (cf. Figure S1 in *Borchert et al., 2021*). This finding is in line with published literature (e.g., *Yeager & Robson, 2017*) and highlights the importance of AMOC for our findings. We therefore conclude that the improved correlation skill of historical CMIP6 simulations compared to their CMIP5 equivalent for SPG SST arises from an improved AMOC-related lagged response of CMIP6 models to natural, particularly volcanic, forcing since the 1980s.

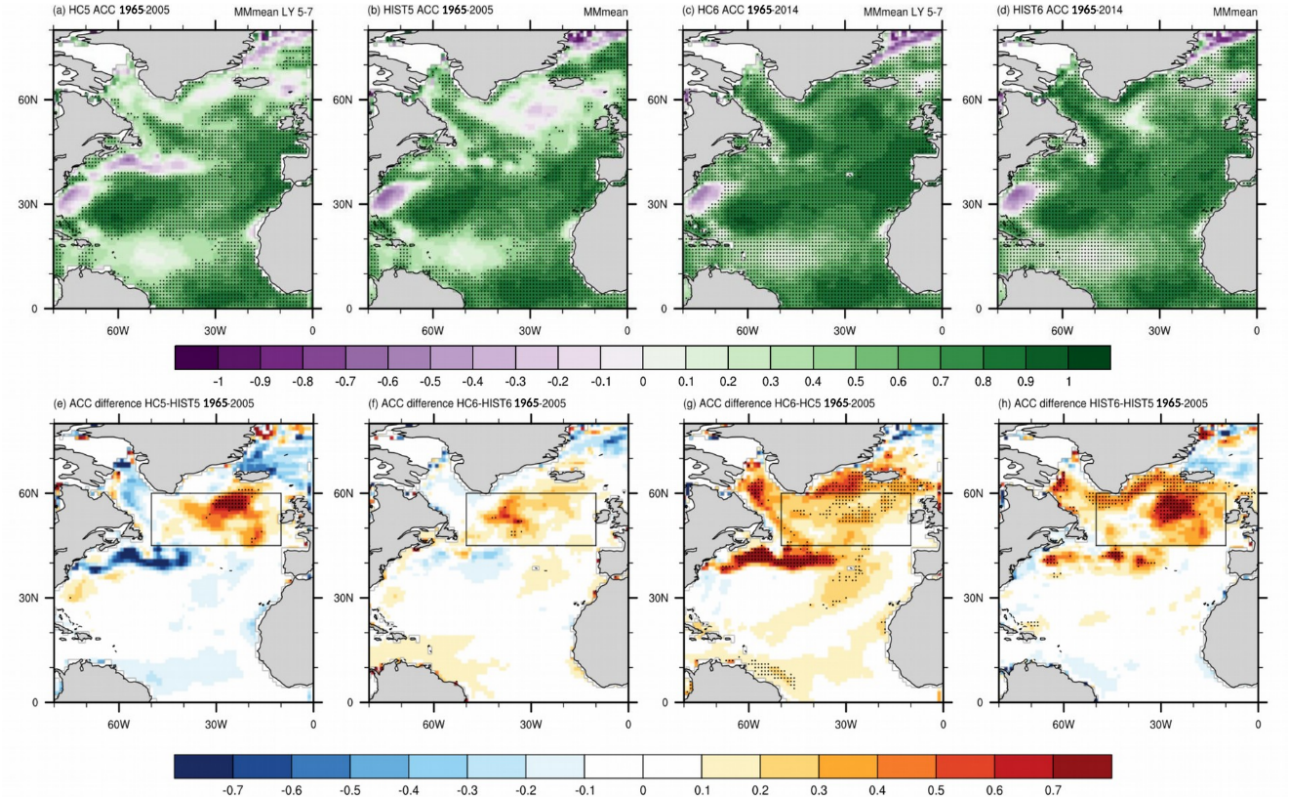

Figure 12: Multi-model ensemble mean decadal prediction skill (anomaly correlation coefficient; ACC) for annual mean non-detrended SST. (a) Skill of CMIP5 initialised decadal hindcasts at lead year 5-7 for the period 1965-2005, compared to skill in the multi-model ensemble mean of (b) CMIP5 historical simulations (1965-2005), (c) CMIP6 initialised hindcasts (1965-2014) and (d) CMIP6 historical simulations (1965-2014). The historical ensemble means are based on the same model subset as the HC5 and HC6 means, which were selected based on availability of simulations in HC5 (6 models) and HC6 (7 models). Skill differences are shown for the common period 1965-2005 between (e) HC5 and HIST5, (f) HC6 and HIST6, (g) HC6 and HC5, and (h) HIST6 and HIST5. Stippling shows where correlation or correlation differences are significantly different from zero (95% confidence). The box outlined in black in (e-h) shows the area used to calculated the SPG index. Adapted from Borchert et al. (2021), their Figure 1.

The previously discussed findings are achieved using correlation analysis to assess the skill of decadal hindcasts. Going beyond this commonly-used metric, we estimate the robustness of these skill estimates using residual ACC (*Smith et al., 2019*) and Mean Square Skill Score (MSSS) analysis. While residual ACC explicitly measures the correlation skill that is achieved beyond the forced component (by subtracting the time series from historical simulations from both initialised hindcast time series and observations), MSSS measures the extent to which the absolute values of observed observations can be reproduced by model simulations. Both residual ACC and MSSS analysis show pronounced skill increase through initialisation, and skill improvement from CMIP5 to CMIP6. We thus highlight in our work that hindcast initialisation improved its value in CMIP6 over CMIP5 for capturing the full amplitude of SPG SST variations and the signal beyond the forced component, indicating continued need for initialisation to really predict North Atlantic SST on the decadal time scale.

These findings hold potential improvement for similar predictions over Europe, as predictions of SPG SST have in the past been shown to influence predictions over Europe (e.g. *Borchert et al., 2019b*). Analysis of this possibility is currently ongoing. A first look at the results suggests, however, that skill for seasonal surface temperature over Europe is comparable between CMIP5 and CMIP6 models. A possible reason for this is the too large atmospheric noise simulated by climate prediction systems (*Smith et al., 2020*), which inhibits an accurate transfer of the predictable signal in the ocean to the atmosphere and over Europe.

University of Edinburgh

Introduction

UEDIN has been assessing the role of forced and internal variability in precipitation projections over Europe. Here we focus on the NAO, the leading mode of climate variability over the North Atlantic region, affecting temperature and rainfall over timescales from days through seasons and decades. Various studies have shown that multi-decadal variations in the NAO yield significant trends in European temperature and rainfall, especially in winter (e.g. *Deser et al., 2016, 2017; Iles & Hegerl, 2017*). This has important implications for deriving observational constraints and the application of these scaling factors to projections of European rainfall. This is important, because the multi-decadal NAO variability in the past is generally not reproduced by CMIP class models (*Iles & Hegerl, 2017; Schurer, in prep.*) and can hence confound estimated trends due to forcing and affect forward projections using the so-called Allen Stott Kettleborough "ASK method" (*Allen et al., 2000; Stott & Kettleborough, 2002*) method. The NAO is also relevant for seamless predictions, as it is predictable over months to possibly years, although with insufficient amplitude (*Scaife & Smith, 2018*).

Methods and Datasets

For this study, we characterise the NAO by the first Empirical Orthogonal Function (EOF1) of sea level pressure (SLP) over the North Atlantic sector (20N-90N; 90W-40W). For the observations, SLP is taken from HADSLP2 (*Allan & Ansell, 2006*), and European rainfall is from the gridded E-OBS v19.0e dataset *(Haylock et al., 2008)*, with monthly values computed from the daily data (Figure 13, black lines). The first principle component time series (associated with EOF1) is computed for each month's (1950-2014) anomalous SLP history (separately) in order to construct a monthly time series of the NAO. The Northern European rainfall is then regressed against the NAO time series (1950-2014, separately for each month) in order to compute the component of rainfall associated with the NAO. The residual component (Figure 13, red lines) is thus an estimate of the rainfall time series with the NAO (as defined) removed. Note that removing the influence of the NAO on Northern European rainfall reduces the variance of the time series, especially in winter.

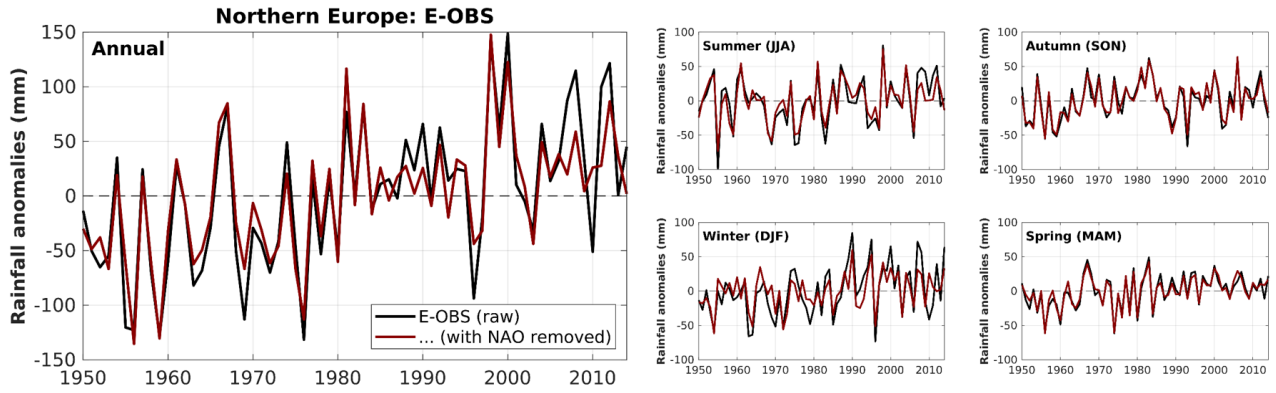

Figure 13: Time series of Northern European annual (left panels) and seasonal (right panels) rainfall anomalies (1950-2014 reference period). Black lines show the observed (E-OBS v19e) time series; red lines show the time series with the NAO-component removed (see text).

In order to compare with model results we utilise CMIP6 historical simulations (Eyring et al., 2016), computing the individual NAO time series in the same way as was done for the observations, separately for each of the ensemble members (41 models, 163 ensemble members), after spatially regridding to a regular 2.5-deg lat-lon grid, retaining only the gridboxes over land. Each model's rainfall was subsequently regressed against its own NAO time series (1950-2014, month by month). The resulting CMIP6 multi-model mean regression patterns exhibit similar features to the corresponding observed patterns and shows a modest reduction in variance of the interannual variability that arises mainly from the cold season after removal of the NAO.

Results

We have analysed the NAO's contribution to both rainfall and surface air temperature across the different SREX regions (NEU/CEU/MED/EUR; regional masks as in *Brunner et al., 2020*), and these various results will be discussed in the forthcoming paper (in prep). Here we demonstrate our results for Northern European rainfall. Figure 14 illustrates the variability and trends associated with the NAO compared to that in all data for observations (a) and models (b). A positive long-term trend in the rainfall anomalies associated with the NAO (blue distribution, most pronounced in winter) is not seen in models. Removing these observed trends (red distribution) brings models and data closer together.

To explore the impact on attribution results, and with it, observationally constrained predictions, we have constructed two sets of multi-model-mean spatiotemporal fingerprints of European rainfall change: one set that retains NAO variability, and another set that excludes the variability associated with the NAO (removing it using the simple regressions from observations, and individual climate model simulations). The 5-yr smoothed time series of Northern European annual rainfall in shown in Figure 15, along with the four seasons (in the small subpanels). Following the ASK method (*Allen et al., 2000; Stott & Kettleborough, 2002; Kettleborough et al., 2007, Shiogama et al., 2016*), we conduct total-least-squares regressions using the two different sets of single all-forced fingerprints

against the observations in order to analyse the impact of removing the NAO in potentially enhancing the signal-to-noise ratio. A confidence interval for each of the scaling factors describes the range of magnitudes of the model response that are consistent with the observed signal, and hence are relevant for constrained predictions. A forced model response is *detected* if the range of scaling factors are significantly greater than zero, and can be described as being *consistent with observations* if the range of values contains the magnitude of one (=1), where uncertainty is estimated using model estimates of internal variability (following *Schurer et al., 2018*).

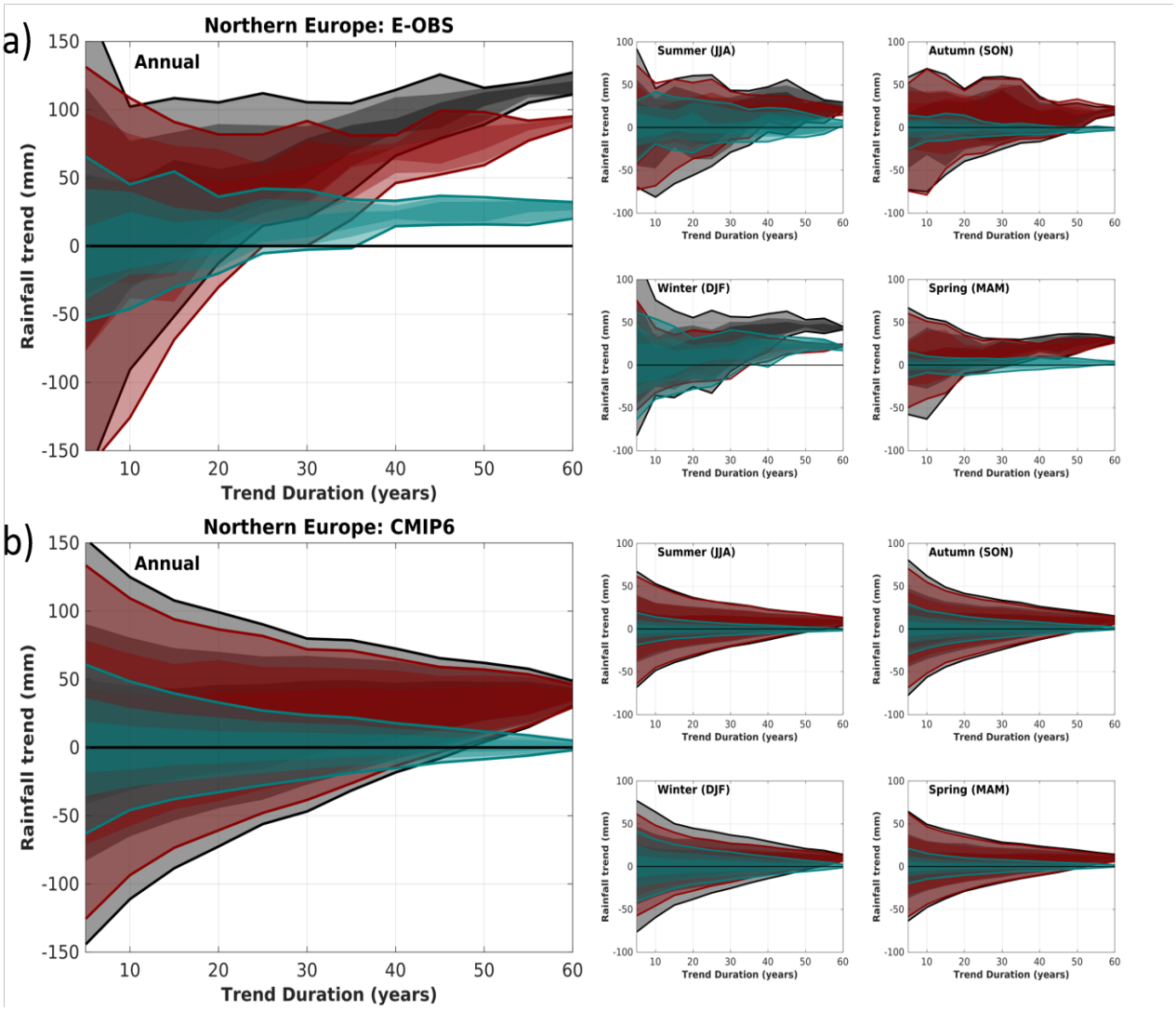

Figure 14: Distributions of the linear trends (of 5-,10-,…,60-yr duration, shown along the x-axis) in Northern European annual (left panels) and seasonal (right panels) rainfall, sampled from: a) E-OBS v19e observations, and b) CMIP6 historical simulations (41 models, 163 ensemble members). Grey shading indicates the trends in the raw time-series; blue shading indicates the trends in the component of rainfall associated with the NAO; red shading indicates the trends of the time-series with the NAO removed. The lightest shading spans the minimum to maximum trends, and the darker levels of shading indicates the 10th-90th and 25th-75th percentile ranges of sampled trends. Trend distributions are randomly sampled over the period 1950-2014, and the CMIP6 panels display the multi-model mean of ensemble means. Units are given as total accumulated (annual or seasonal) rainfall (in mm) per trend duration (in years, along the x-axis).

The results show a detectable rainfall signal in the annual time series (along with the summer and winter seasons). The raw analysis (Figure 15a) suggests that the models underestimate the change in NEU precipitation by a factor of 1.4-3.8 in annual, and even more in the winter and summer. However, this model data discrepancy reduces after removing the NAO, consistent with Figure 14, with results now indicating that the multi-model forced response (brown bar) together with model-simulated climate variability reflected in uncertainty ranges (light brown) explain the observed change. It also indicates that the future simulated multi-model mean forced change is expected to be realistic and within uncertainty ranges if correcting for the NAO. Along with a shift in the magnitude that comes from the modified observations, the constraint also tightens. This should bolster confidence in the constrained projections. A paper incorporating this work is in preparation (*Ballinger, Hegerl, et al.*). Our result of underestimated multidecadal NAO variability is consistent with the finding, from seasonal predictions, that NAO signals in models are underestimated.

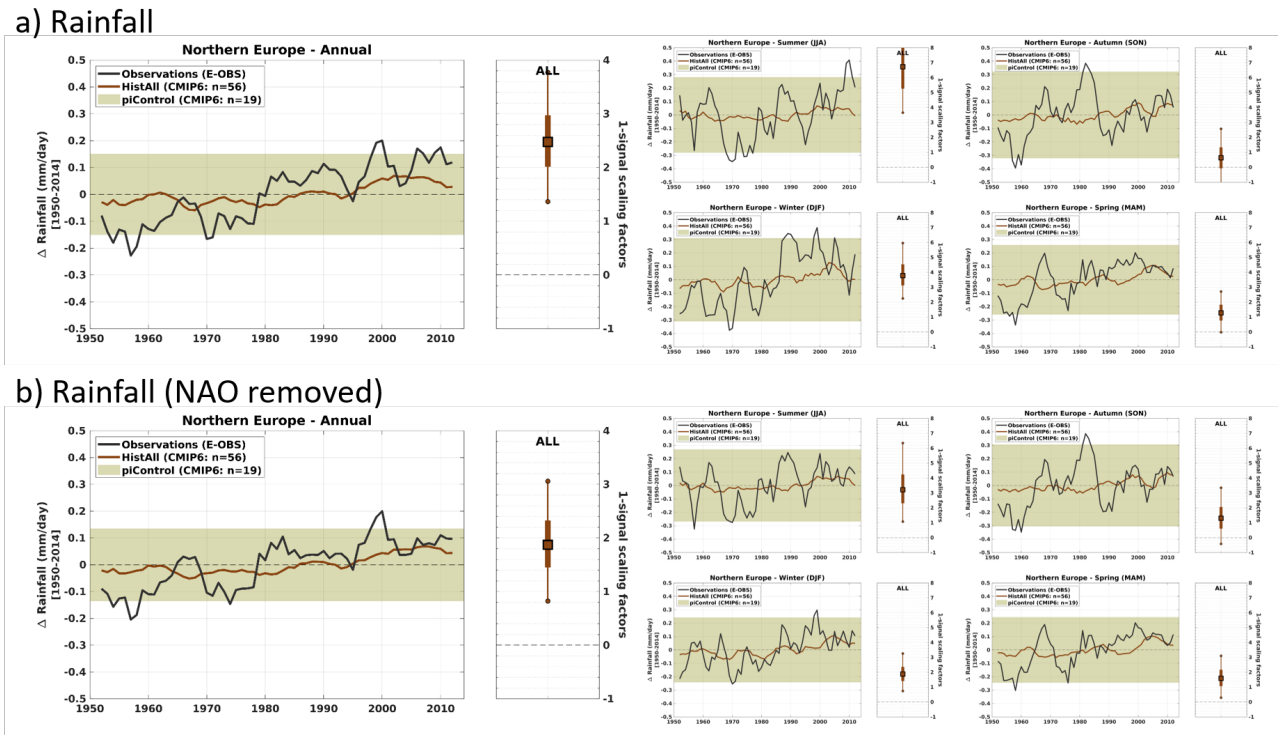

Figure 15: Annual and seasonal time series of Northern European rainfall anomalies (relative to 1950-2014) from observations (E-OBS v19, black line) and CMIP6 historical simulations (all forcings, brown line, displaying the multi-model mean of ensemble means (19 models, 56 total ensemble members); a) original time series, and b) 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).

Met Office

The latest generation of national UK climate scenarios (*Lowe et al., 2018*) includes a probabilistic projections product, alongside sets of global, regional and local projections that consist of raw climate model data (*Murphy et al, 2018; Kendon et al, 2019*). Here, we compare the probabilistic

projections against decadal hindcasts covering 1960 to present day. The probabilistic projections are derived from centennial climate change simulations in which internal climate variability is simulated but not predicted, as no attempt is made to initialise their future components using recent observations. In contrast, the hindcasts are initialised by assimilating observations of the ocean and (in some cases) atmosphere, thus creating potential to capture additional predictability from successfully forecasting aspects of low-frequency variability, or correcting model responses to previous changes in external forcing. This report assesses the extent to which initialised predictions can improve upon, or augment, the information on near-term risks available from the probabilistic projections.

The initialised hindcasts consist of ensembles from eleven systems contributing to the CMIP6 Decadal Climate Prediction Project (*Boer et al., 2016*). Ten of these consist of 10 members, with five members from CMCC. We combine these to form a 105-member multi-model ensemble in which all simulations are assumed to represent equally likely outcomes. Hindcasts were started every winter from 1960 to 2017, including changes in external forcing agents derived from observations to 2015, switching subsequently to the SSP2-4.5 scenario. Prior knowledge of volcanic eruptions is assumed in the hindcasts (CMIP6-init), as in the probabilistic projections (UKCP-pdf).

The probabilistic projections are derived by combining 348 perturbed parameter ensemble simulations (using several configurations of the HadCM3 model) with 12 CMIP5 earth system model simulations. This is done using a Bayesian statistical framework (*Murphy et al., 2018, Harris et al., 2021*), updated from the earlier implementation of *Sexton et al (2012)* and *Harris et al. (2013)*. It includes emulation techniques to estimate results from points in parameter space for which no climate model simulation is available, and alternative projections are weighted using a set of historical performance metrics that include spatial fields of climatological averages and changes observed during the 20th century in carbon dioxide concentration, upper ocean heat content and global patterns of surface temperature. UKCP-pdf is the primary source of uncertainty information in UKCP18, and is being used extensively to inform future impacts and hazards related to climate change, alongside other UKCP datasets (e.g. *Arnell et al., 2021*).

Evidence from previous hindcast experiments shows that initialisation can improve the skill of predictions of global mean surface temperature (GMST) for a few years ahead, beyond that attributable to the simulation of externally forced warming trends (e.g. *Kirtman et al., 2013*). There is also evidence of enhanced skill in the North Atlantic sector. *Borchert et al (2020)* find that initialisation of CMIP6 models increases the skill of sea surface temperature hindcasts in the sub-polar gyre, and also that CMIP6 hindcasts perform better than their CMIP5 counterparts.

Below, we consider GMST and winter and summer hindcasts of surface air temperature and precipitation anomalies for England and Wales. Anomalies are calculated relative to a baseline

period of 1971-2000. In order to facilitate comparison with UKCP-pdf, the initialised hindcasts are expressed as a multi-model frequency distribution, percentiles of which are estimated by a simple ranking of the 105 members. Table 2 shows anomaly correlation (ACC) and mean-square-skill (MSSS) scores for ensemble median hindcasts from the two datasets.

Variable	Score	Year 1		Years 2-9	
		CMIP6-init	UKCP-pdf	CMIP6-init	UKCP-pdf
GMST (°C)	ACC	0.97	0.95	0.98	0.98
	MSSS	0.94	0.85	0.91	0.94
England/Wales winter temperature (°C)	ACC	0.34	0.28	0.63	0.59
	MSSS	0.10	0.07	0.29	0.34
England/Wales summer temperature (°C)	ACC	0.44	0.55	0.88	0.86
	MSSS	0.18	0.29	0.65	0.69
England/Wales winter precipitation (%)	ACC	0.28	0.04	0.36	0.34
	MSSS	0.06	0.00	0.11	0.11
England/Wales summer precipitation (%)	ACC	0.13	-0.30	0.41	-0.63
	MSSS	-0.04	-0.12	-0.29	-0.76

Table 2: Skill of predicted anomalies relative to 1971-2000, for medians of the frequency distribution of CMIP6-init and the probability distribution of UKCP-pdf, for a lead time of one year, and the multiyear average of years 2-9. MSSS measures the mean square forecast error relative to that of climatology, zero indicating no relative skill and unity perfect skill. The England & Wales region is defined here by combining the "Wales" and "southern England" grid boxes of the HadCM3 model (see Fig. 3 of Harris et al., 2010).

For one year ahead, enhanced skill in GMST is found in CMIP6-init, as measured by MSSS. The ACC scores are high for both systems, being dominated by the multidecadal climate change trend that both predict well. Skill on the decadal time scale is assessed by considering the average of years 2-9. For GMST we find similar scores for both systems. For years 2-9 observations during the warming hiatus period of ~2000-2015 lie within the range of the UKCP-pdf distributions, but below the median (not shown). At this range, the CMIP6-init median is very similar to that of UKCP-pdf during 2000-2015, and is not therefore closer to observations. However, at year 1, the CMIP6-init median tracks the observations quite well during the hiatus period. Both the UKCP-pdf and CMIP6-init distributions typically capture the observations at year 1, but the CMIP6-init information has higher skill and a narrower spread (not shown). This indicates a higher level of confidence, implying added value compared to the non-initialised information.

Figure 16 (left) shows time series of winter surface temperature anomalies for England & Wales, for years 2-9. Observations show a warming of ~0.8°C during the period, again captured quite well by the prediction systems. The observations also include pronounced variability on the decadal time scale. This is (at least in part) associated with variability in the phase of the winter NAO, which was predominantly negative during 1960-1990 (*Kendon et al., 2020*). The CMIP6-init median is slightly cooler than that of UKCP-pdf during this period. However, the lower (10th percentile) end of the CMIP6-init range fails to encompass the most negative observed events in the mid-1960s and early 1980s. This is also the case for UKCP-pdf. The skill scores (Table 2) are similar for both systems (also true for summer England & Wales temperature). The median NAO for years 2-9 from CMIP6-init hindcasts (ACC=0.21; MSSS = -0.16, based on data currently available for 90 hindcasts) gives small

anomalies (not shown). These do not capture the amplitude of the largest anomalous events found in observations. This may partly explain the lack of a clear benefit of initialisation in Figure 16 (left).

For NAO in years 2-9, *Smith et al. (2020)* obtained a higher ACC of 0.48 (using a combination of CMIP5 and CMIP6 decadal prediction systems). They increased this to 0.79, by scaling the ensemble-mean anomalies to match the observed variance (on the basis that the relatively high ACC indicated a systematic underestimate of the predictable signal), and combining the latest hindcast with previous ones verifying 1,2 and 3 years previously. The lower ACC obtained here does not necessarily justify applying this type of scaling to our dataset, but the question of how to interpret and exploit dynamical signals in seasonal to decadal hindcasts remains an important question.

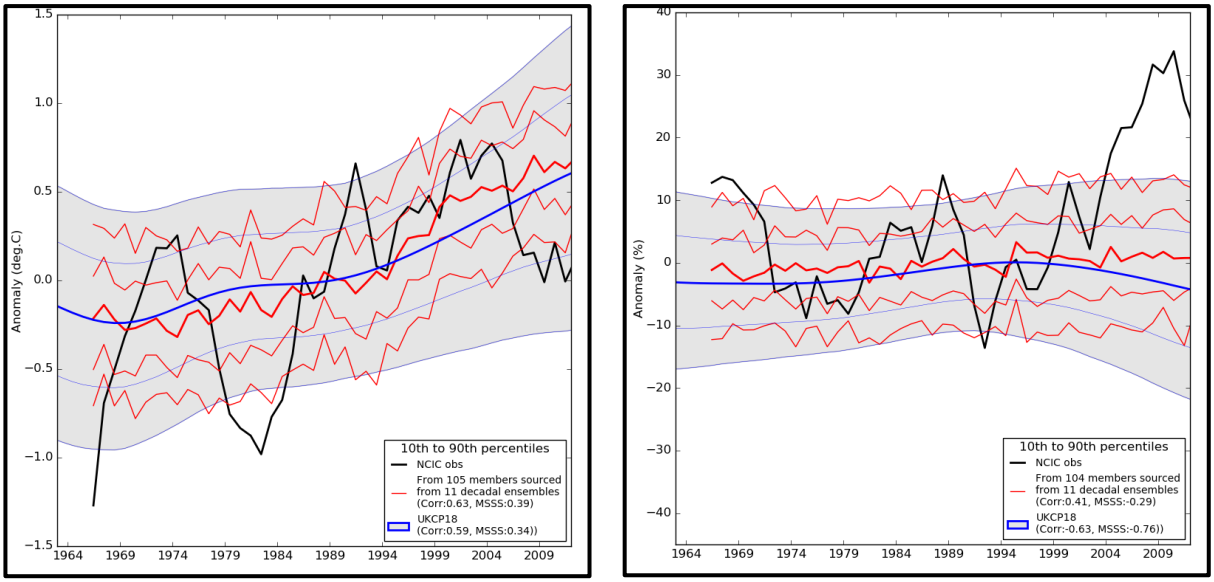

Figure 16: left: Distributions of anomalies for England & Wales surface air temperature (°C) for December to February, from UKCP-pdf (grey shading and blue lines) and CMIP6-init (red lines). The CMIP6-init results are eight-year means over years 2-9 of each hindcast, with corresponding UKCP-pdf results formed by averaging 3000 samples of evolving annual anomalies that comprise its climate projections (Murphy et al., 2018). The black line shows verifying observations from the UK National Climate Information Centre (NCIC, https://www.metoffice.gov.uk/weather/climate/uk-climate). Thick blue and red lines show the 50th percentile (median) of UKCP-pdf and CMIP6-init respectively, and thin lines show the 10th, 25th and 75th percentiles of the corresponding distributions. Right: As left but for eight year-means of precipitation anomalies (%) for June to August, comparing CMIP6-init hindcasts for years 2-9 against UKCP-pdf and observations.

Summer precipitation time series for England & Wales anomalies are shown in Figure 16 (right), also for years 2-9. The UKCP-pdf results show an incipient drying from ~2000 onwards, which is the early stage of a projected climate change response that strengthens during the 21st century (*Murphy et al., 2018*). The observations do not show an obvious long-term trend, but do show considerable decadal variability, notably in the substantial wet anomalies that developed after 2003. These lie well outside the 10-90% range of UKCP-pdf. Wet anomalies over northern Europe are associated with the positive phase of the Atlantic Multidecadal Oscillation (AMO) (*Sutton and Dong, 2012*), which has been prevalent since the mid-1990s and is skilfully predicted by initialised hindcasts

(*Smith et al., 2020*). However, there is no clear evidence of an AMO-driven wet signal in the CMIP6-init distributions of Figure 16 (right). The median is generally positive during the post-2000 period, but the median anomalies are small. The 90th percentile values are similar to those of UKCP-pdf, and to CMIP6-init results in earlier decades. More work is needed to understand these results, including checking the circulation and precipitation teleconnection patterns associated with the AMO in the CMIP6 systems. For example, *Ruggieri et al. (2021)* studied the extratropical response of the Atlantic jet stream and storm track in winter, using seven models contributed to the CMIP6 decadal prediction experiment. In the positive phase of the AMO, relative to the negative phase, they found an equatorward shift of the low-level jet in the eastern Atlantic, accompanied by reduced storm activity over the UK and north-west Europe. However, the dynamical response was found to be uncertain, and dependent on biases in the simulation of the climatological distribution of the jet stream. *Simpson et al. (2019)* considered precipitation hindcasts for March, from the NCAR system contributed to CMIP6-init. Whilst the hindcasts failed to reproduce the observed response of the jet stream to the AMO, Simpson et al. found that skilful hindcasts for UK precipitation could be obtained by combining hindcasts of SST anomalies with the observed SST-precipitation relationship. A mixed statistical-dynamical approach of this type may be an option for improving the precipitation distributions derived from CMIP6-init, in future work.

Another interesting question is whether the lack of a clear signal in CMIP6-init results from compensation between a drying trend driven by climate change (as seen in UKCP-pdf), and a wet signal driven by the Atlantic decadal variability. The CMIP6-init scores show some skill in ACC, but not in MSSS, while UKCP-pdf shows negative skill in both scores (Table 2). Note also that the scores are potentially sensitive to the choice of climatological baseline. Our 1971-2000 was chosen to support inclusion of recent climate trends in our assessment. However, it was an unusually dry period in the historical record of England & Wales summer rainfall (*Kendon et al., 2020*), potentially because the phase of AMO was predominantly negative during this period. If a more recent baseline was chosen, the observed wet anomalies diagnosed in Figure 16 (right) would be significantly smaller.

The spread of the two predictive distributions is also an important consideration, as this informs the range of outcomes that should be considered in assessments of near-term risks. For winter England & Wales temperature CMIP6-init shows a narrower spread than UKCP-pdf, particularly beyond 2000 (Figure 16, left). For summer precipitation the spread is broadly comparable between the two systems prior to 2000, and narrower in CMIP6-init subsequently (Figure 16, right). Differences in spread can potentially be driven by predictability arising from initialisation in CMIP6-init, or by different sampling of uncertainties due to internal climate variability and/or climate change signals. For GMST, the spread within the 10-member ensembles of the individual CMIP6-init systems (not shown) typically grows as a function of hindcast lead time out to 4-5 years ahead, and then saturates. This is consistent with a steady decay of predictability, ending in an asymptotic level

of spread consistent with unconstrained internal variability. For the England & Wales variables, there is no clear evidence that the spread in individual CMIP6 systems grows systematically with lead time (e.g., Figure 17, which shows summer temperature as an example). For the DePreSys4_GC3.1 system, an independent estimate of unconstrained internal variability is also shown (grey vertical bar), that was derived from four non-initialised historical climate change simulations. The initialised system shows a similar spread to the non-initialised results (within sampling uncertainties) at all lead times. These results suggest that initialised predictability has little impact in constraining uncertainties due to internal variability at the national scale, in this example.

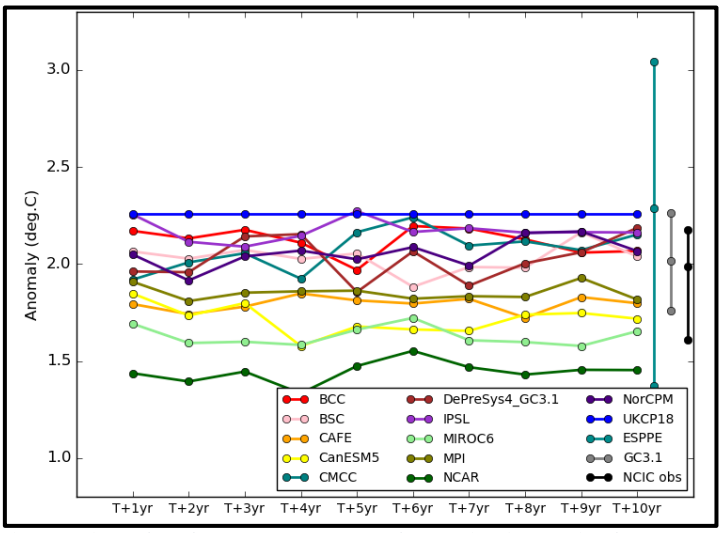

Figure 17: Spread in seasonal anomalies of surface air temperature for England & Wales for June to August (°C), as a function of hindcast lead time. Coloured lines show spread for the individual systems contributing to CMIP6-init, compared against the corresponding spread from UKCP-pdf (blue line). The UKCP-pdf values are the average distance between the 10th and 90th percentiles of its probability distributions for the set of verification dates relating to each lead time. Corresponding 10-90% ranges for each initialised system are obtained by averaging the standard deviations of their ensemble hindcasts for each verification date, and assuming a normal distribution. Vertical bars show estimates of spread due to unconstrained climate variability on 1-30 year time scales, obtained by applying a Butterworth filter to multidecadal time series of historical climate. In each vertical bar, the central dot shows the median estimate of spread, for NCIC observations (black); non-initialised simulations (grey) from HadGEM3-GC3.1 (the climate model used in the DePreSys4_GC3.1 initialised system); non-initialised simulations from the ESPPE (green), a 57-member perturbed parameter ensemble of earth system model simulations that provides the climate variability information incorporated in UKCP-pdf. In each case, the lower and upper dots show 10% and 90% limits of the confidence interval for the spread, estimated from block-bootstrap resampling of filtered anomalies. The confidence interval for the ESPPE is much wider than for observations and GC3.1, because different PPE members simulate internal variability with different characteristics.

Also noteworthy is that the saturation level of spread varies between the different systems, though values do lie mostly within the uncertainty range associated with the observed level (black vertical bar in Figure 17). In UKCP-pdf the sampling of internal variability is derived from a 57-member ensemble of perturbed variants of an earth system model (vertical green bar in Figure 17), the median level of which lies close to the spread found in the UKCP-pdf results (blue line in Figure 17). This shows that uncertainty arising from internal variability dominates uncertainty in climate change signals as a source of spread, in these near-term results. This does not apply in the multidecadal projections of UKCP-pdf, because the spread in the distributions grows during the 21st century as the influence of long-term climate change develops (*Murphy et al., 2018*). Overall, the results of

this analysis suggest that differences in spread between CMIP6-init and UKCP-pdf are likely to depend mainly on differences in the amplitude of internal variability simulated in their constituent modelling systems.

4. Lessons learnt

- The CMIP6 data base provides a useful framework to analyse the factors driving climate variability across internal variability (DCPP), forced response (historical), and isolated individual forcing (DAMIP) in a coherent set of models.
- Forcing seems to have played a larger role than previously thought in modulating observed decadal-scale North Atlantic temperature variations since the 1980s. This could have implications for the estimation of predictive capability of climate models over Europe.
- Novel temporal pooling approach allows assessing exceedances of more extreme quantiles compared to multi-annual averaging and robust estimates of skill regarding probabilities of seasonal extremes in upcoming years (based on a large multi-model ensemble)
- For UK climate anomalies, differences in spread between initialised and non-initialised predictions appear to arise mainly from differences in internal variability between the underlying climate models, rather than from potential constraints offered by initialisation. The credibility of projected risks therefore depends strongly on how well regional variability is simulated in the relevant ensemble systems.
- Multi-model ensembles of decadal predictions and climate projections are largely not reliable, as measured by the flatness of their rank histograms. Initialisation only has small effects to improve reliability in the first forecast years, and does rarely lead to significantly reliable prediction ensembles. This implies that statistical post-processing (of both decadal predictions and projections) is needed to obtain probabilistically useful information from the climate simulations, in particular calibration leads to reliable information.
- Perfect-model prediction experiments can provide a useful framework to estimate the potentially achievable skill of predictions assuming ideal initial conditions which avoid initialisation-related problems that affect current decadal prediction systems. Comparing the potential skill in perfect-model predictions with real-world prediction skill can help identify regions where models and climate observations behave inconsistently.
- CMIP5 and CMIP6 decadal predictions and climate projections hold large potential economic value for upper tercile (warm) events of surface temperature averaged over lead years 2-9. This large potential economic value is explained by the strong trend in surface temperatures over large areas of the world, which is to a large extent captured by initialized and uninitialized simulations.
- The added value of initialization measured using the potential economic value and the framework presented in Smith et al. (2019) is small over most parts of the world. Results indicate

that added value measured using correlation scores does not entirely translate into potential economic value.

- Stochastic physics is a computationally cheap way to improve tropical SST reliability in ECMWFs coupled model system on time scales up to 28 months. This motivates the use of stochastic physics in current decadal prediction models, which mostly do not make use of such schemes to represent model uncertainty

5. Links built

- Work on the contents of this deliverable sparked collaborative efforts between UOXF, ETHZ, CNRS/IPSL and UEDIN in the development of prediction-projection merging techniques. This collaboration applies a method usually applied to uninitialized projections (Brunner et al., 2020) to decadal initialized predictions, and thus spanning activities in WP1, WP2 and WP5.
- A collaboration between Met Office and CNRS/IPSL on constraining European summer climate using large-scale North Atlantic climate is underway, spanning EUCP WPs 2 and 5.
- Collaboration between UOXF and SMHI regarding verification of predictions of seasonal extremes' probabilities, spanning EUCP WPs 1 and 5

6. Acronyms

BSC - BARCELONA SUPERCOMPUTING CENTER - CENTRO NACIONAL DE SUPERCOMPUTACION
CNRS – CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE CNRS
SMHI – SVERIGES METEOROLOGISKA OCH HYDROLOGISKA INSTITUT
UEDIN - THE UNIVERSITY OF EDINBURGH
UKMO - MET OFFICE
UOXF – THE CHANCELLOR, MASTERS AND SCHOLARS OF THE UNIVERSITY OF OXFORD

7. References

Allan, R., & Ansell, T. (2006). A New Globally Complete Monthly Historical Gridded Mean Sea Level Pressure Dataset (HadSLP2): 1850–2004, *Journal of Climate*, *19*(22), 5816-5842.

Allen, M. R., P. A. Stott, J. F. Mitchell, R. Schnur, and T. L. Delworth (2000). Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, 407, 617–620, https://doi.org/10.1038/35036559

Athanasiadis, P.J., Yeager, S., Kwon, YO. et al. (2020). Decadal predictability of North Atlantic blocking and the NAO. *npj Clim Atmos Sci* 3, 20. https://doi.org/10.1038/s41612-020-0120-6

Arnell, N.W., Kay, A.L., Freeman, A., Rudd, A.C., Lowe, J.A. (2021). Changing climate risk in the UK: A multisectoral analysis using policy-relevant indicators. *Climate Risk Management* 31, 100265.

Befort, D.J., C.H. O'Reilly, and A. Weisheimer (2020). Constraining projections using decadal predictions *Geophys. Res. Lett.*, doi:10.1029/2020GL087900

Befort, D. J., O'Reilly, C. H., & Weisheimer, A. (2021). Representing model uncertainty in multiannual predictions. *Geophysical Research Letters*, 48, e2020GL090059. https://doi.org/10.1029/2020GL090059

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. (2016). The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, *Geosci. Model Dev.*, 9, 3751–3777, https://doi.org/10.5194/gmd-9-3751-2016.

Borchert, L. F., Düsterhus, A., Brune, S., Müller, W. A., & Baehr, J. (2019a). Forecast-oriented assessment of decadal hindcast skill for North Atlantic SST. *Geophysical Research Letters*, 46, 11444–11454. https://doi.org/10.1029/2019GL084758

Borchert, L. F., Pohlmann, H., Baehr, J., Neddermann, N.-C., Suarez-Gutierrez, L., & Müller, W. A. (2019b). Decadal predictions of the probability of occurrence for warm summer temperature extremes. *Geophysical Research Letters*, 46, 14042– 14051. https://doi.org/10.1029/2019GL085385

Borchert, L. F., Menary, M. B., Swingedouw, D., Sgubin, G., Hermanson, L., & Mignot, J. (2021). Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. *Geophysical Research Letters*, 48, e2020GL091307. https://doi.org/10.1029/2020GL091307

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., Coppola, E., de Vries, H., Harris, G., Hegerl, G. C., Knutti, R., Lenderink, G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi, P., & Undorf, S. (2020). Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework. *Journal of Climate*, 33(20), 8671–8692. https://doi.org/10.1175/jcli-d-19-0953.1

Deser, C., Hurrell, J. W., & Phillips, A. S. (2017). The role of the North Atlantic Oscillation in European climate projections. *Climate dynamics*, 49(9-10), 3141-3157.

Deser, C., Terray, L., & Phillips, A. S. (2016). Forced and Internal Components of Winter Air Temperature Trends over North America during the past 50 Years: Mechanisms and Implications, *Journal of Climate*, *29*(6), 2237-2258.

Doblas-Reyes, F., Andreu-Burillo, I., Chikamoto, Y. *et al. (2013).* Initialized near-term regional climate change prediction. *Nat Commun* **4,** 1715. https://doi.org/10.1038/ncomms2704

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

Elmore, K. (2005). Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and forecasting*, 20 (5), 789–795, doi:10.1175/WAF884.1.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev*., 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016

Fricker, T. E.; Ferro, C. A. T. & Stephenson, D. B. (2013). Three recommendations for evaluating climate predictions. *Meteorol. Appl*., 20, 246–255

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., et al. (2016). The Detection and Attribution Model Intercomparison Project (DAMIP v1.0) contribution to CMIP6, *Geosci. Model Dev.*, 9, 3685–3697, https://doi.org/10.5194/gmd-9-3685-2016.

Harris, G.R., Sexton, D.M.H., Booth, B.B.B., Collins, M., Murphy, J.M. (2013). Probabilistic projections of transient climate change. *Clim. Dyn*. 40:2937-2972.

Harris, G.R., Murphy, J.M., Sexton, D.M.H., Booth, B.B.B. (2021). Probabilistic projections for regional climate change accounting for Earth System modelling uncertainty. In preparation.

Haylock, M., N. Hofstra, A. M. G. Klein Tank, E. J. Klok, P. D. Jones, and M. New (2008). A European daily high-resolution gridded dataset of surface temperature, precipitation and sea-level pressure, *J. Geophys. Res*., 113, D20119, doi:10.1029/2008JD010201.

Iles, C. E., & Hegerl, G. C. (2017). Role of the North Atlantic Oscillation in decadal temperature trends. *Environmental Research Letters*, 12(11), 114010.

Jolliffe, I., and C. Primo (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136 (6), 2133–2139, doi:10.1175/2007MWR2219.1.

Kendon, E.J., Fosser, G., Murphy, J., et al. (2019). UKCP Convection-permitting model projections: Science report. Available from https://www.metoffice.gov.uk/research/collaboration/ukcp/guidance-science-reports

Kendon, M, McCarthy, M.P., Jevrejeva, S., Matthews, A., Legg, T. (2019). State of the UK climate 2018. *Int. J. Climatol*. 39 (Suppl 1): 1-55.

Kettleborough, J. A., B. B. Booth, P. A. Stott, and M. R. Allen, (2007). Estimates of uncertainty in predictions of global mean surface temperature. *J. Climate*, 20, 843–855, https://doi.org/10.1175/JCLI4012.1.

Kirtman, B., et al. (2013). Near-Term Climate Change: Projections and Predictability, in Climate Change 2013: The Physical Science Basis. Contribution of Working Group 1 to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge Univ. Press, Cambridge.

Kruschke, T., Rust, H.W., Kadow, C., Müller, W.A., Pohlmann, H., Leckebusch, G.C. and Ulbrich, U., (2016). Probabilistic evaluation of decadal prediction skill regarding Northern Hemisphere winter storms, *Meteorol. Z*, 25, pp.721-38.

Leckebusch, G. C., Renggli, D. Ulbrich, U. (2008). Development and application of an objective storm severity measure for the Northeast Atlantic region. *Meteorologische Zeitschrift* Vol. 17 No. 5, 575 - 587.

Liu, Y., Donat, M. G., Taschetto, A. S., Doblas-Reyes, F. J., Alexander, L. V., & England, M. H. (2019). A framework to determine the limits of achievable skill for interannual to decadal climate predictions. *Journal of Geophysical Research: Atmospheres*, 124, 2882– 2896. https://doi.org/10.1029/2018JD029541

Lowe, J.A., et al. (2018). UKCP18 Science Overview Report. Available from https://www.metoffice.gov.uk/research/collaboration/ukcp/guidance-science-reports

Mason, S. J. (2004). On Using "Climatology" as a Reference Strategy in the Brier and Ranked Probability Skill Scores, *Monthly Weather Review*, *132*(7), 1891-1895.

Mariotti, A., Baggett, C., Barnes, E. A., Becker, E., Butler, A., et al. (2020). Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond, *Bulletin of the American Meteorological Society*, *101*(5), E608-E625. https://doi.org/10.1175/BAMS-D-18-0326.1

Marotzke, J., Müller, W. A., Vamborg, F. S. E., Becker, P., Cubasch, U., et al. (2016). MiKlip: A National Research Project on Decadal Climate Prediction, *Bulletin of the American Meteorological Society*, *97*(12), 2379-2394. https://doi.org/10.1175/BAMS-D-15-00184.1

Morice, C. P., Kennedy, J. J., Rayner, N. A., Winn, J. P., Hogan, E., Killick, R. E., et al. (2021). An updated assessment of near-surface temperature change from 1850: the HadCRUT5 data set. *Journal of Geophysical Research: Atmospheres*, 126, e2019JD032361. https://doi.org/10.1029/2019JD032361

Murphy, J.M., Harris, G.R., Sexton, D.M.H., Kendon, E.J et al. (2018). UKCP18 Land Projections: Science Report. Available from https://www.metoffice.gov.uk/research/collaboration/ukcp/guidance-science-reports.

O'Reilly, C. H. (2018). Interdecadal variability of the ENSO teleconnection to the wintertime North Pacific. *Climate Dynamics*, 51 (9), 3333{3350. doi: https://doi.org/10.1007/434 s00382-018-4081-y

Richardson, David S (2003). Predictability and economic value. *Seminar on Predictability of weather and climate, 9-13 September 2002, https://www.ecmwf.int/node/11922*

Ruggieri P et al. (2020). Atlantic Multidecadal Variability and North Atlantic Jet: a multi-model view from the Decadal Climate Prediction Project. *J of Climate*, doi: 10.1175/JCLI-D-19-0981.1

Scaife, A. A., Arribas, A., Blockley, et al. (2014). Skillful long-range prediction of European and North American winters. *Geophysical Research Letters*, 41(7), 2514-2519.

Scaife, A. A., & Smith, D. (2018). A signal-to-noise paradox in climate science. *npj Climate and Atmospheric Science*, 1(1), 1-8.

Schurer, A., Hegerl, G., Ribes, A., Polson, D., Morice, C., & Tett, S. (2018). Estimating the transient climate response from observed warming. *Journal of Climate*, 31(20), 8645-8663.

Schuster, M., Grieger, J., Richling, A., Schartner, T., Illing, S., Kadow, C., Müller, W. A., Pohlmann, H., Pfahl, S., Ulbrich, U. (2019). Improvement in the decadal prediction skill of the North Atlantic extratropical winter circulation through increased model resolution, *Earth Syst. Dynam*., 10, 901–917, https://doi.org/10.5194/esd-10-901-2019

Sexton, D.M.H., Murphy J.M., Collins, M., Webb, M.J. (2012). Multivariate prediction using imperfect climate models part I: outline of methodology. Clim Dyn 38:2513-2542.

Shiogama, H., Stone, D., Esmori, S. et al. (2016). Predicting future uncertainty constraints on global warming projections. *Sci Rep* 6, 18903. https://doi.org/10.1038/srep18903

Simpson, I.R., Yeager, S.G., McKinnon, K.A. et al. (2019). Decadal predictability of late winter precipitation in western Europe through an ocean–jet stream connection. *Nat. Geosci*. 12, 613–619. https://doi.org/10.1038/s41561-019-0391-x

Smith, D.M., Eade, R., Scaife, A.A. et al. (2019). Robust skill of decadal climate predictions. *npj Clim Atmos Sci* **2, 13**. https://doi.org/10.1038/s41612-019-0071-y.

Smith, D.M., Scaife, A.A, Eade, R. et al. (2020). North Atlantic climate far more predictable than models imply. *Nature* 583, 796-800.

Stott, P. A., and J. A. Kettleborough (2002). Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, 416, 723–726, https://doi.org/10.1038/416723a.

Sutton, R.T., Dong, B. (2012). Atlantic Ocean influence on a shift in European climate in the 1990s. *Nat. Geosci*. 5:788-792.

Swingedouw, D., Ortega, P., Mignot, J. *et al.* (2015). Bidecadal North Atlantic ocean circulation variability controlled by timing of volcanic eruptions. *Nat Commun* **6,** 6545. https://doi.org/10.1038/ncomms7545

Verfaillie, D., Doblas-Reyes, F. J., Donat, M. G., Pérez-Zanón, N., Solaraju-Murali, B., Torralba, V., & Wild, S. (2021). How Reliable Are Decadal Climate Predictions of Near-Surface Air Temperature?, *Journal of Climate*, *34*(2), 697-713.

Weisheimer, Antje, and T. N. Palmer (2014). On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface* 11.96: 20131162.

Weisheimer, A., Palmer, T. N., and Doblas-Reyes, F. J. (2011). Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles, *Geophys. Res. Lett.*, 38, L16703, doi:10.1029/2011GL048123.

Yeager, S., and Coauthors (2018). Predicting near-term changes in the Earth System: A large ensemble of initialized decadal prediction simulations using the Community Earth System Model. *Bulletin of the American Meteorological Society*, 99 (9), 1867–1886, doi:10.1175/BAMS-D-17-0098.1.

Yeager, S.G., Robson, J.I. (2017). Recent Progress in Understanding and Predicting Atlantic Decadal Climate Variability. *Curr Clim Change Rep* **3,** 112–127. https://doi.org/10.1007/s40641-017-0064-z

List of tables

List of figures