HORIZON 2020
THEME SC5-2017

European Climate Prediction system

(GRANT AGREEMENT 776613)

European Climate Prediction system (EUCP)

Deliverable D1.2

Construction of probability forecasts for the near term horizon (up to 10 years) from multiple sources of information for a number of the most commonly used variables and tailored to specific applications

Deliverable Title	*Construction of probability forecasts for the near term horizon (up to 10 years) from multiple sources of information for a number of the most commonly used variables and tailored to specific applications*	
Brief Description	*This deliverable report provides products and methods for the construction of probabilistic forecasts information aimed at this need of climate prediction users. Multiple sources of information of decadal prediction systems from the CMIP5 and CMIP6 are hereby combined to achieve the most credible and reliable and thus most usable prediction product.*	
WP number	WP1	
Lead Beneficiary	*BSC*	
Contributors	*Carlos Delgado-Torres (BSC), Deborah Verfaillie (BSC), Simon Wild (BSC), Markus Donat (BSC), Francisco Doblas-Reyes (BSC), Leon Hermanson (Met Office), Anca Brookshaw (Met Office), Doug Smith (Met Office), Adam Scaife (Met Office), Melissa Seabrook (Met Office), Didier Swingedouw (IPSL/CNRM), Bo Christiansen (DMI), Shuting Yang (DMI), Tim Kruschke (SMHI)*	
Creation Date Version Number Version Date	*22/03/2021* *[v4]* *15/09/2021*	
Deliverable Due Date Actual Delivery Date	*30/11/2021* *30/11/2021*	
Nature of the Deliverable	R	*R – Report*
		P - Prototype
		D - Demonstrator
		O - Other
Dissemination Level/ Audience	x	*PU - Public*

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

Version	**Date**	**Modified by**	**Comments**
V1	22/03/2021	Simon Wild	First outline and overview of already finished and planned work until submission
V2	19/07/2021	Markus Donat et al	First complete version sent to internal review
V3	06/08/2021	Simon Wild & Markus Donat et al.	Addressing the internal reviewer comments
V4	15/09/2021	Simon Wild & Markus Donat et al.	Final version sent to coordinators
Final	21/11/2021	Francisco Doblas-Reyes	Final edited version

Table of Contents

1. Executive Summary

This deliverable report gives an overview of WP1 activities aiming at developing methods and approaches to improve probabilistic climate prediction. Multiple sources of information, primarily stemming from different prediction models, are combined to eventually provide a more digestible and thus more user friendly probabilistic forecast. While this deliverable also includes information from actual predictions for the upcoming years, the assessment of forecast quality and construction of the best possible probabilistic forecast is based on the performance of the multi-model hindcasts of the past climate in comparison to observations. The performance of a forecast system is usually compared to a reference forecast, which can either be the climatological forecast or the free-running, uninitialised simulations for the historical and scenario experiments. With the exception of the real forecasts for the upcoming years, the large amount of raw simulations analysed and combined throughout this deliverable stem from the CMIP5 and CMIP6 archives.

Using the ranked probability skill score as a measure of performance, BSC showed that combining multiple models improves the quality of probabilistic forecasts for forecast years 1-5 for temperature over most regions of the globe with respect to a naïve climatological forecast, while the picture for precipitation is more heterogeneous and improvements are limited to the Asian and African continent. There is usually one model outperforming the overall multi-model. However, the 'best' model depends on region, forecast period and variable and, therefore, the multi-model can be considered as generally superior to the median of the individual models. Without knowing a priori which model will perform best, the combination of more models will be the preferred choice over any single model. Larger ensembles with more members and built from a larger number of different prediction systems can be shown to lead to better results.

This result is confirmed by the BSC study on the reliability of temperature predictions of decadal forecast systems. The benefits of using a multi-model are illustrated in comparison to a large ensemble of a single model. Reliability generally 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 is shown that bias correction and calibration of the raw initialised data is essential to provide reliable predictions.

As the WMO Lead Centre for Annual to Decadal Prediction, the Met Office gathers real forecasts for the upcoming years from contributing partners. Alongside with performance metrics based on simulations of the past, both deterministic and probabilistic forecasts are provided in the address www.wmolc-adcp.org. The most recent set of predictions are for the initial conditions of late 2020. Some of the key forecasts include that annual global mean near-surface air temperature will likely be at least 1ºC above pre-industrial levels in each of the 5 years after 2020 and a 90% chance that one of these 5 years will surpass 2016 as the up until now warmest year on record.

CNRM/IPSL analysed the probability of a rapid change of the North Atlantic sub polar gyre, namely an abrupt cooling event related to a collapse of deep convection in this region. Based on CMIP6 projections the probability of encountering an abrupt cooling in the next decades and before the end of the 21st century is up to 36.4%, slightly lower than the 45.5% estimated in CMIP5. Such changes would have profound impacts on the general ocean circulation.

The area of interest of DMI is also the North Atlantic with a focus on the probabilistic forecast of North Atlantic mean sea surface temperatures (SST). SST forecasts are shown to be reliable based on a single large model ensemble and multi-model decadal predictions. Mean SST forecasts also outperform the climatological reference forecast. This can be primarily attributed to a positive trend in SST over the North Atlantic. The forecast for the upcoming years shows a further warming of the North Atlantic.

SMHI introduced the temporal pooling, a new way to formulate probabilistic forecasts, for the decadal prediction of seasonal means. The multi-model ensemble of this new approach offers skill compared to the climatological reference forecast for extremely high summer temperatures over large parts of the globe. Forecasts for extremely dry boreal summers however lacks skill in many regions with the Sahel being a positive exception.

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

Trustworthy climate information for the near future, i.e. for the next 10 years, has become increasingly important for stakeholders (Buontempo et al. 2014; Hewitt & Lowe 2018) from various economic sectors and societal groups for planning and decision making. While decadal predictions have been shown to be skilful in predicting the sign of anomalies, e.g. for temperature in many regions of the world, but also for atmospheric pressure patterns or precipitation in the North Atlantic region (Athanasiadis et al., 2020; Kushnir et al., 2019; Smith et al., 2019; Yeager et al. 2018), many applications or decision makers require probabilistic information, e.g. about the occurrence probability of specific event classes (e.g. Torralba 2017). As state-of-the-art decadal predictions are probabilistic by their construction, and recent coordinated efforts provide decadal predictions and hindcasts from a growing number of models, the work reported in this deliverable exploits these multi-model decadal prediction ensembles to generate and evaluate probabilistic prediction information. The work covered by this deliverable provides products and methods for the construction of probabilistic forecasts information aimed at this need of climate prediction users. The information of multiple decadal prediction systems from the CMIP5 and CMIP6 archives is hereby combined to achieve more credible and reliable and thus more usable prediction products. Most deliverable contributions exploit the initialised decadal predictions for the probabilistic analyses for predictions of certain event classes. In addition, CMIP6 projections are also used to analyse the probability of very rare events, namely the occurrence of abrupt cooling in the North Atlantic. For this analysis, the prediction period is extended beyond a decade, quantifying the occurrence of such events until the end of this century. It would be interesting to perform a similar analysis based on decadal predictions, but this would need to first resolve issues related to initialisation shock and drift in the predictions (Bilbao et al., 2021), as these may artificially increase the probability for such events to be simulated.

The work in WP1 and for this deliverable has been closely linked to efforts in other work packages. WP1 is providing the information of initialised predictions for the efforts in WP5 including the comparison and blending with non-initialised projections. WP1 products developed for this deliverable also provided a base for the climate information needed for one of the case studies in WP4. There has further been participation in WP6 activities as the question whether initialised climate predictions are action-oriented or not is essential for users.

Some of the work summarised in this deliverable, in particular section 3.1.2, is a continuation of the previous milestone report M2 “Preliminary illustration of the relative merits of the forecast combination and description of the methods identified”.

3.1 Barcelona Supercomputing Center

3.1.1 Probabilistic forecast quality assessment and product generation from decadal predictions

The forecast quality assessment is essential for providing high-quality and reliable forecast products that can be used for decision making in several sectors (Merryfield et al., 2020). This work aims to turn a large number of raw simulations into more digestible information in a probabilistic form and document the effect of using a multi-model ensemble instead of individual forecast systems for the prediction of some of the essential climate variables. In addition, the impact that both the model initialisation and the number of forecast systems utilised for building a multi-model have on the quality of the probabilistic products is assessed.

The 165 available members in the decadal hindcasts (DCPP; the retrospective decadal predictions) and 195 members of the historical simulations (HIST; the retrospective climate projections) from 13 forecast systems contributing to the Component A of the Decadal Climate Prediction Project (DCPP-A; Boer et al., 2016) of the Sixth Coupled Model Intercomparison Project (CMIP6; Eyring et al., 2016) have been used. The variables that have been evaluated are the near-surface air temperature, sea surface temperature, precipitation, and sea level pressure. Besides, two different reference datasets have been used to take into account the observational uncertainty.

The forecast quality assessment has been applied to the anomalies (computed with respect to the 1981-2010 climatology) over the 1966-2014 period. For the decadal predictions, a lead-time dependent climatology has been computed to remove the drift of the simulations towards the climatology of the forecast systems. Different forecast periods have been evaluated for the decadal predictions: forecast years 1, 1-5 (average of the first five forecast years of the hindcasts), 6-10 and 1-10. Four probabilistic multi-model approaches are tested to assess whether the way a multi-model is constructed affects the quality of the probabilistic forecast products. The probabilities for the forecast products of the multi-model-1 are computed as the averaged probabilities (based on the terciles of each system) from all the forecast systems. The probabilities for the multi-model-2 are computed by pooling all the members together, where the terciles are computed from the multi-model pool. The other two multi-model approaches (multi-model-1-calib and multi-model-2-calib) are computed as the multi-models 1 and 2, respectively, but previously calibrating the hindcasts with the method presented in Doblas-Reyes et al. (2005). This is a variance inflation technique that improves the reliability of the forecasts. Such calibration has been performed in cross-validation mode, i.e., excluding the year in which the prediction is made.

The ranked probability skill score (RPSS; Wilks, 2011) has been used to measure the quality of the probabilistic forecast products based on tercile and quintile categories. Its fair version (FairRPSS; Ferro, 2014) has also been calculated to estimate the potential skill that an infinite ensemble size would have. These skill scores have been computed using the climatological forecast, the individual forecast systems, the historical simulations, and a shorter multi-model ensemble as the reference forecasts to address the different objectives of the study. The Random Walk test (DelSole and Tippett, 2016) has been used to assess the significance of the results at the 95% confidence level. In addition, the Relative Operating Characteristic (ROC; Kharin and Zwiers, 2003) score has been used to obtain the skill in predicting each one of the three or five categories. The forecast quality assessment has been performed following the recommendations developed in the C3S_34c contract

(ECMWF/COPERNICUS/2019/C3S_34c_DWD) of the Copernicus Climate Change Service (C3S) operated by the European Centre for Medium-Range Weather Forecasts (ECMWF).

Figure 1 shows the RPSS for tercile categories (below average, average, and above average) obtained with the different multi-model approaches when predicting the near-surface air temperature and the precipitation for the forecast years 1-5 using the climatological forecast as the benchmark. For temperature, a significant positive skill score is found over large regions of the globe, especially over the continents, indicating an added value of the decadal predictions with respect to the climatological forecast. By contrast, there are other regions (e.g., the Pacific Ocean, the subpolar gyre region, the South Atlantic Ocean, India, Northern Asia and Northern Australia) where the decadal predictions do not improve the information of the climatological forecast. For precipitation, the RPSS obtained is much lower, only being significantly positive over limited regions of Africa and Asia. Note that the GPCC dataset used as observational reference for precipitation does not provide data over the ocean regions. Regarding the comparison between the multi-model approaches, there are not substantial differences in the RPSS. Also, the points with statistically significant values are generally in the same regions, excepting the multi-model-2-calib approach for precipitation, which shows fewer points where the climatological forecast is significantly better than the multi-model ensemble (e.g., over Southern Africa, South America, and Asia).

Figure 2 shows the RPSS obtained with the DCPP multi-model-1 using the forecast system that presents the maximum and the median skill as the reference forecast for the near-surface air temperature and precipitation for each grid point. For both variables, the RPSS using the best forecast system as the benchmark is negative over most regions (Figures 2a and 2c), indicating that the multi-model provides worse predictions, particularly for temperature, for which statistical significance is found over several regions. However, to reach the highest possible skill, the best system would have to be chosen individually for each particular region, variable, and forecast period, complicating the creation of the forecast products. Besides, the best forecast system might not be the same if different parts of the period are evaluated, supporting the use of a multi-model approach. The RPSS using the forecast system with the median skill as the benchmark is positive over the whole globe (Figures 2b and 2d), indicating that the multi-model is better than, at least, 50% of the individual systems. For temperature, the areas with the largest benefit of using a multi-model in comparison to the median of the individual systems are the Atlantic subpolar gyre region, South America, Africa and the Indian Ocean, although only a few points present statistical significance. For precipitation, a positive RPSS is found over all the regions, although only a few marginal points are significant. Then, although the best forecast system provides higher skill than the multi-model, the multi-model outperforms the single systems on median, without the need of choosing the best system for each particular region, forecast period, and variable (Doblas-Reyes et al., 2005; Mishra et al., 2018; Hemri et al., 2020). On the other hand, in a climate services context, the best forecast system or multi-model approach could be selected to issue the best possible predictions over a particular region, variable and forecast period.

Figures 3a and 3b display the RPSS obtained with the DCPP multi-model using the HIST multi-model as the reference forecast for the near-surface air temperature and precipitation, i.e., this comparison highlights the effects from initialising the predictions while the forcings are the same between the DCPP and HIST multi-models. For temperature, the regions with a significant improvement due to the model initialisation are the eastern part of the North Atlantic Ocean and

oceanic areas surrounding the southern parts of America and Australia. For precipitation, only a few points show significant values of the RPSS.

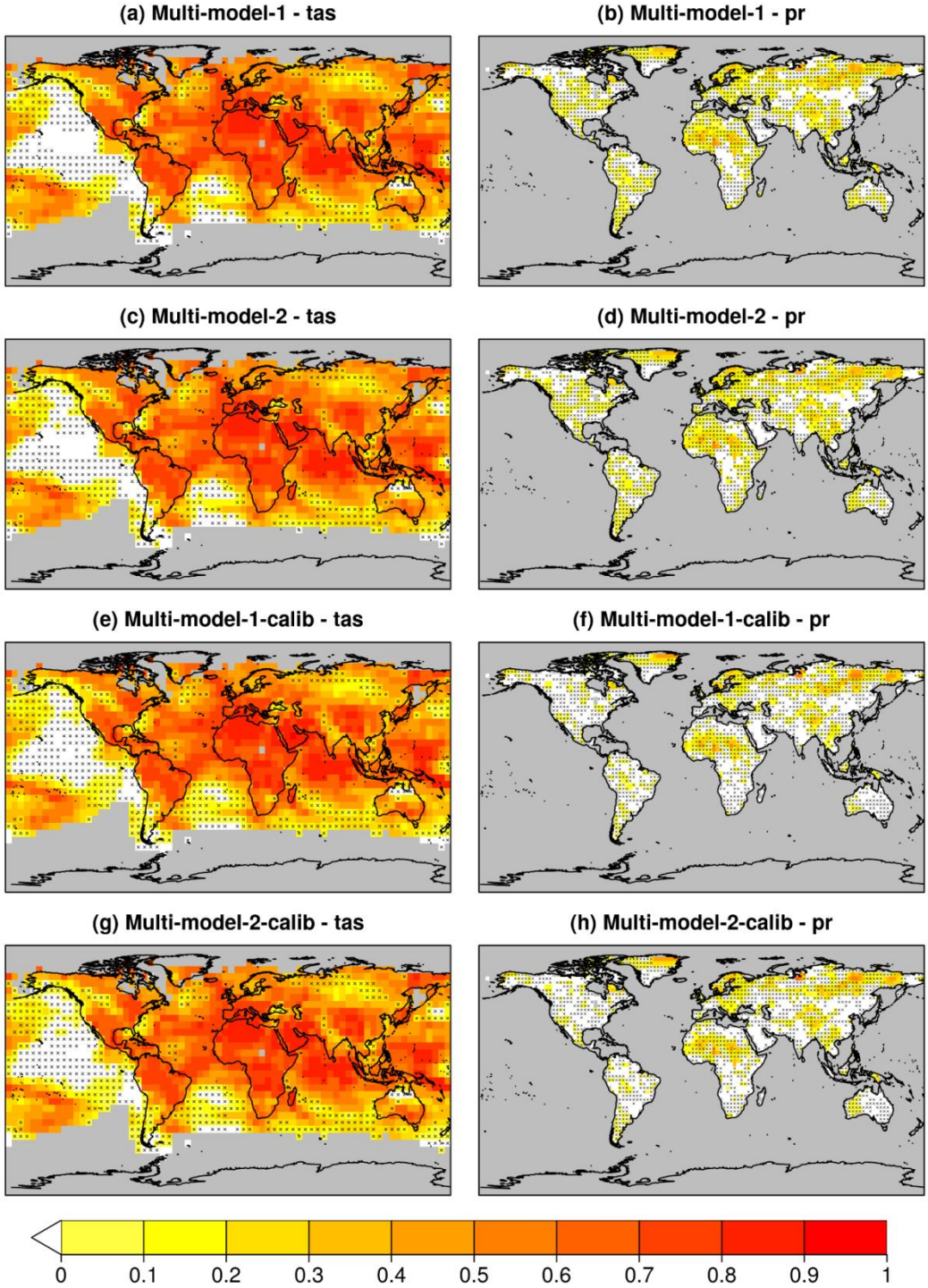


Figure 1. RPSS for tercile categories (below average, average, and above average) obtained with the different DCPP multi-model approaches for the forecast years 1-5 for the near-surface air temperature (first column) and precipitation (second column) using the observed climatology as the reference forecast. The skill has been computed over the 1966-2014 period (start dates 1965-2009). The reference period for the computation of the thresholds between categories is 1981-2010. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets. Crosses indicate that the decadal predictions do not provide significantly better or worse predictions than the observed climatology at the 95% confidence level based on a Random Walk test.

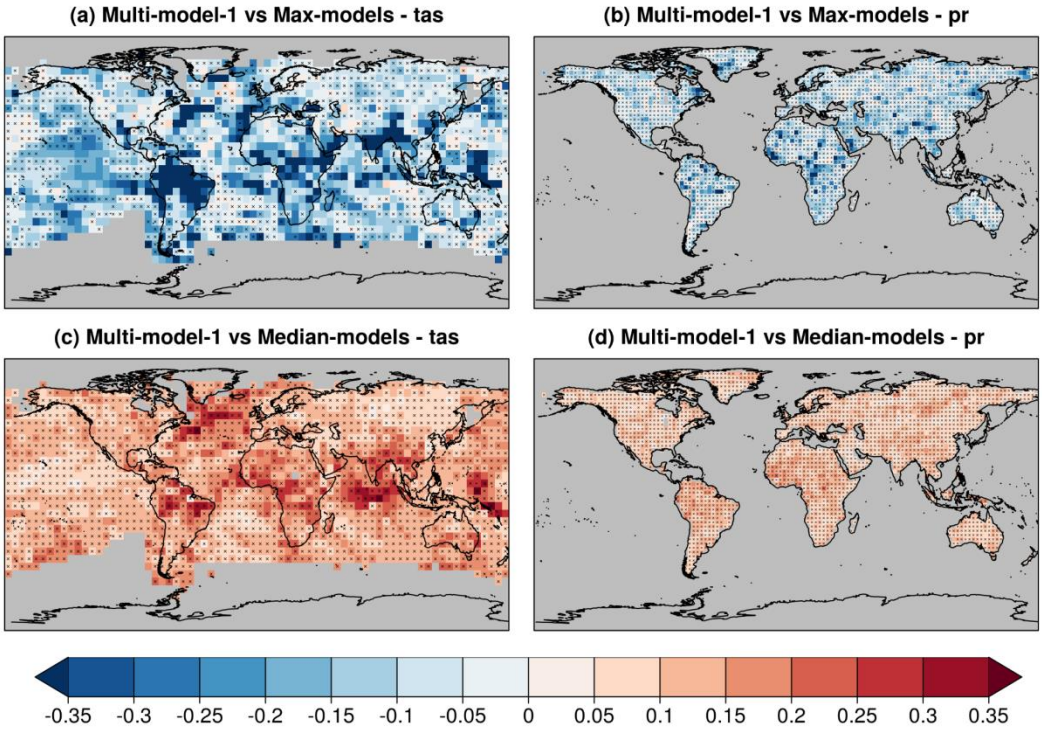


Figure 2. RPSS obtained with the DCPP multi-model-1 using the forecast system that presents the maximum (a, b) and median (c, d) skill as the reference forecast. The skill scores are shown for the near-surface air temperature (first column) and precipitation (second column) for the forecast years 1-5. The skill has been computed over the 1966-2014 period (start dates 1965-2009). The reference period for the computation of the thresholds between categories is 1981-2010. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets. Crosses indicate that the different forecasts do not provide significantly better or worse predictions at the 95% confidence level based on a Random Walk test.

Figures 3c and 3d show the RPSS obtained with the DCPP multi-model-1 (165 members from all the DCPP systems) using the C3S multi-model-1 (36 members from the CMCC-CM2-SR5, EC-Earth3, HadGEM3, MPI-ESM1.2-HR systems) as the reference forecast. The four forecast systems used to create the C3S multi-model (which are also included among the DCPP systems) have been chosen because they provide decadal predictions in near-real time, i.e., quasi-operational predictions that would be available for possible service developments. For temperature, the DCPP multi-model provides better probabilistic products than the C3S multi-model over regions like the Eastern Pacific Ocean, South America, Southern Africa, and oceanic regions of the Southern Hemisphere. For precipitation, there is positive RPSS (i.e., the largest ensemble is more skilful), for example, over parts of Central Africa and the northern area of Southern America. However, it should be noted that most of the areas with an improvement are the ones that present a negative or non-significantly positive RPSS with respect to the climatological forecast (see Figure 1).

The results of this study are being prepared for a scientific publication. As a preliminary assessment of the results, the following outcomes can be highlighted:

- The DCPP multi-model improves the probabilistic information with respect to the climatological forecast, which is the one traditionally available to users, in most of the regions for temperature, while for precipitation the improvements are limited to areas of Africa and Asia.

- The skill of the multi-model is lower than the skill obtained with the best of the individual systems for each grid point, but the multi-model provides typically higher skill than the median of the individual prediction systems.
- There is added value from model initialisation for predicting temperature over some parts of the North Atlantic Ocean and some oceanic areas of the Southern Hemisphere, while no added value is found for precipitation.
- Comparing the multi-models built using a different number of forecast systems the skill of the largest ensemble is higher over several regions. However, most of these regions coincide with those that do not present an improvement with respect to the climatological forecast.

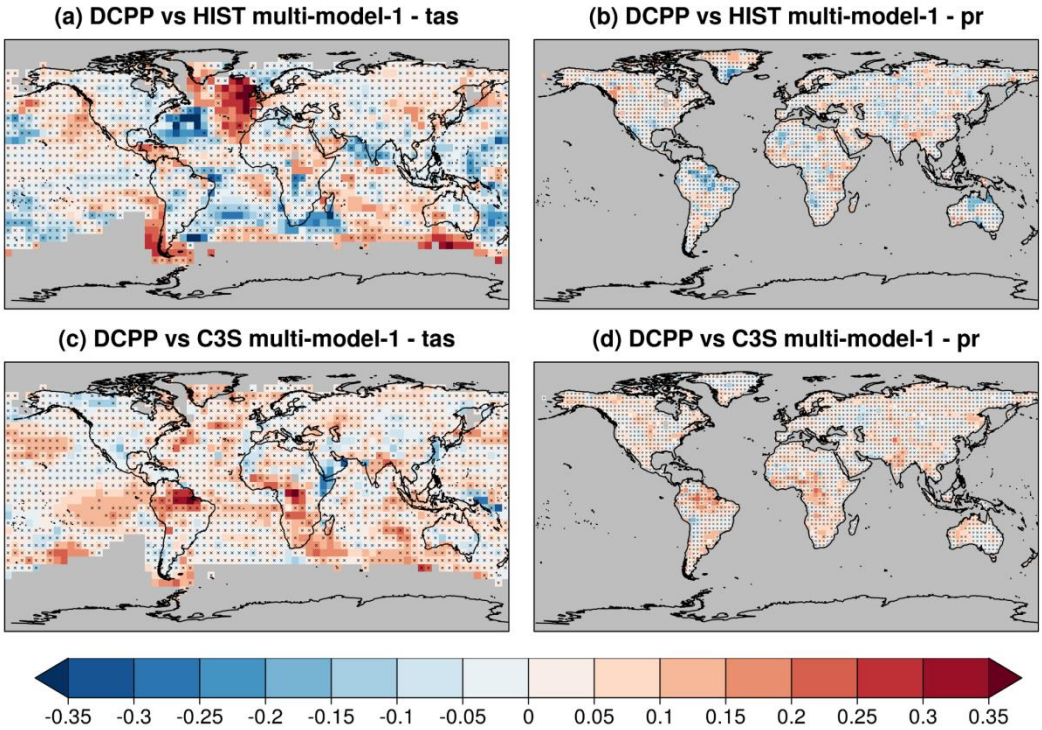


Figure 3. RPSS obtained with the DCPP multi-model-1 using the HIST multi-model-1 (a, b) and the C3S multi-model-1 (c, d) as the reference forecast for the near-surface air temperature (first column) and precipitation (second column). The DCPP multi-model is built with 165 decadal prediction members from 13 forecast systems, the HIST multi-model is built with 195 historical simulation members from 13 forecast systems, and the C3S multi-model is built with 36 decadal prediction members from 4 forecast systems. The skill has been computed for the forecast years 1-5 over the 1966-2014 period (start dates 1965-2009). The reference period for the computation of the climatology and thresholds between categories is 1981-2010. The reference datasets used for the near-surface air temperature and precipitation are, respectively, the GHCNv4 and the GPCC datasets. Crosses indicate that the different forecasts do not provide significantly better or worse predictions at the 95% confidence level based on a Random Walk test.

3.1.2 Evaluating the reliability of decadal climate predictions

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.

The reliability of near-surface air temperature has been analysed for a large ensemble of a single-model (40 members) and in a multi-model of decadal predictions from 12 different climate models

(110 members in total) evaluated against two observational datasets (see Verfaille et al., 2021 for details). The analysed models in this study are slightly different to ones from the previous section 3.1.2. The most recent model simulations, e.g. those used in the previous section 3.1.2 and also in 3.1.4 or 3.1.5, were not yet available at the beginning of this study. The focus of this study lies on the first forecast year after initialisation. These results are however also put in comparison to the multi-year mean of the first five forecast years.

Reliability is assessed using rank histograms (Elmore, 2005) and test statistics proposed by Jolliffe and Primo (2008) on global and regional scale 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 to be 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 reference of an ideal system produces a flat or uniform histogram. However, because of sampling variation the histograms are almost never exactly flat.

The question of whether observed deviations from "flatness" or uniformity be attributed to chance or they indicate deficiencies in the forecasts arises. An overall test of uniformity is provided by the $\chi 2$ goodness-of-fit test. The $\chi 2$ test statistic can be decomposed into several components (Jolliffe and Primo, 2008) that indicate whether the forecasts are biased or show a trend (Jolliffe-Primo test statistic for slope - JP slope), whether they are over- or under-dispersive (Jolliffe-Primo test statistic for convexity - JP convexity), and whether there are any other deviations from flatness, once these 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. The way the reliability evolves with forecast time has also been explored by looking at results for forecast year 1 and forecast years 1-5, over the period 1961-2010. Finally, the impact of applying several post-processing techniques to the "raw" temperature anomalies has been assessed showing that all forecast system ensembles have issues with reliability and that a bias correction and calibration is essential to obtain reliable predictions (see Verfaillie et al., 2021 for more detailed descriptions of the methods).

I. Rank histogram example of (unreliable) predictions

For an ensemble prediction to be reliable, both the ensemble members and the observations should be statistically indistinguishable from each other. It would be equally probable for the observation to fall in any of the ranks and, as a consequence, the rank histogram would be flat (as if both observations and ensemble members stem from a uniform distribution). The particular deviations from flatness of a rank histogram can be used to identify some forecast deficiencies depending on its specific shape: a slope in the rank histogram indicates an incorrect representation of the trend or a mean bias in the forecast as ensemble members mostly occupy the extreme ranks (either the lowest ranks or the highest ranks; see figure 4, left panel). Convex (concave) rank histograms point to an over-dispersive (under-dispersive) ensemble forecast with higher frequencies of the observations corresponding to the middle (extreme) ranks (see the over-dispersive example in the right panel of figure 4).

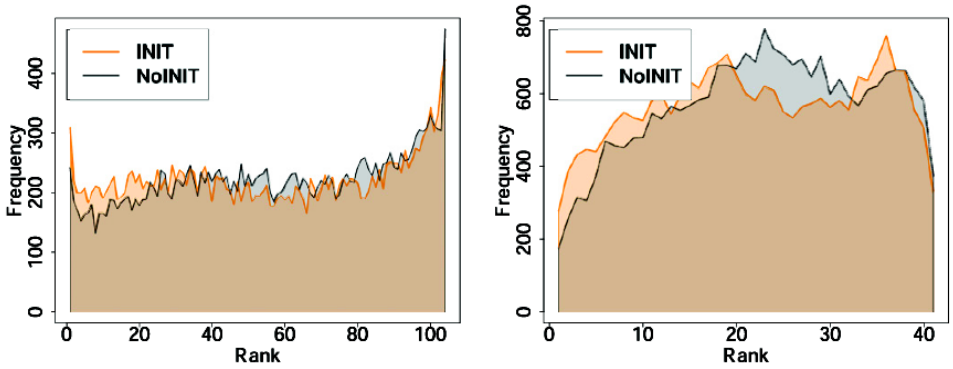


Figure 4: near-surface air temperature rank histograms for uncorrected simulations for the Greenland region for the initialised predictions (INIT) and uninitialised projections (NoINIT), for forecast year 1, in the multi-model ensemble using (left) all models available and (right) the single-model large ensemble. Forecasts are verified against GISTEMP. The x-axis represents the ranks. The y-axis shows the frequency of each rank.

Results in the following refer to the full multi-model of all 110 ensemble members for the first forecast year after initialisation. A brief discussion of how the results compare to the large ensemble single-model and the multi-annual mean of the first five forecast years is also included.

II. Reliability of uncorrected simulations

Based on the rank histograms and the JP test of uniformity, our results show that raw data from current initialised near-term climate prediction ensembles of near-surface air temperature are not significantly reliable for a large majority of the investigated regions (consistent with Doblas-Reyes et al., 2013 or Pasternack et al. 2018). Unreliable predictions due to biases, different trends or over- and underdispersion are expressed as the JP slope and JP convexity coefficients and their contribution to the $\chi 2$ coefficient or the deviations from flatness of the rank histograms.

Uncorrected ensemble predictions do not provide significantly reliable estimates (nor do uninitialised projections - see Verfaillie et al. 2021 for details) , i.e flat rank histograms (the $\chi 2$ p-value is never above 0.05), for near-surface air 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 $\chi 2$ coefficient, resulting in unreliable estimates (purple colours in figure 5). In general, the reliability measured by the JP parameters varies greatly depending on the analysed regions and forecast system ensembles. For example, the Southern Ocean (SOO) in figure 5 (left) displays a much higher value for the contribution to $\chi 2$ of the JP slope coefficient than the South Pacific Ocean (SPO). However, for the same region (SOO) and the same parameter (JP slope) but for the other ensemble (figure 5, right) the contribution is much lower.

The reliability in terms of JP coefficients for the multi-model and single models are quite different for certain regions (compare figure 5 left and right) especially the contributions to $\chi 2$ of the JP convexity parameter for the single model ensemble are often larger (more purple colours in figure 5) compared to the multi-model. In the regions where the single model ensemble has these larger contributions the initialised ensemble is generally less reliable than its non-initialised counterpart ensemble. This points to deficiencies in terms of spread of the forecast system possibly due to the ensemble generation conditioning the growth of the ensemble spread.

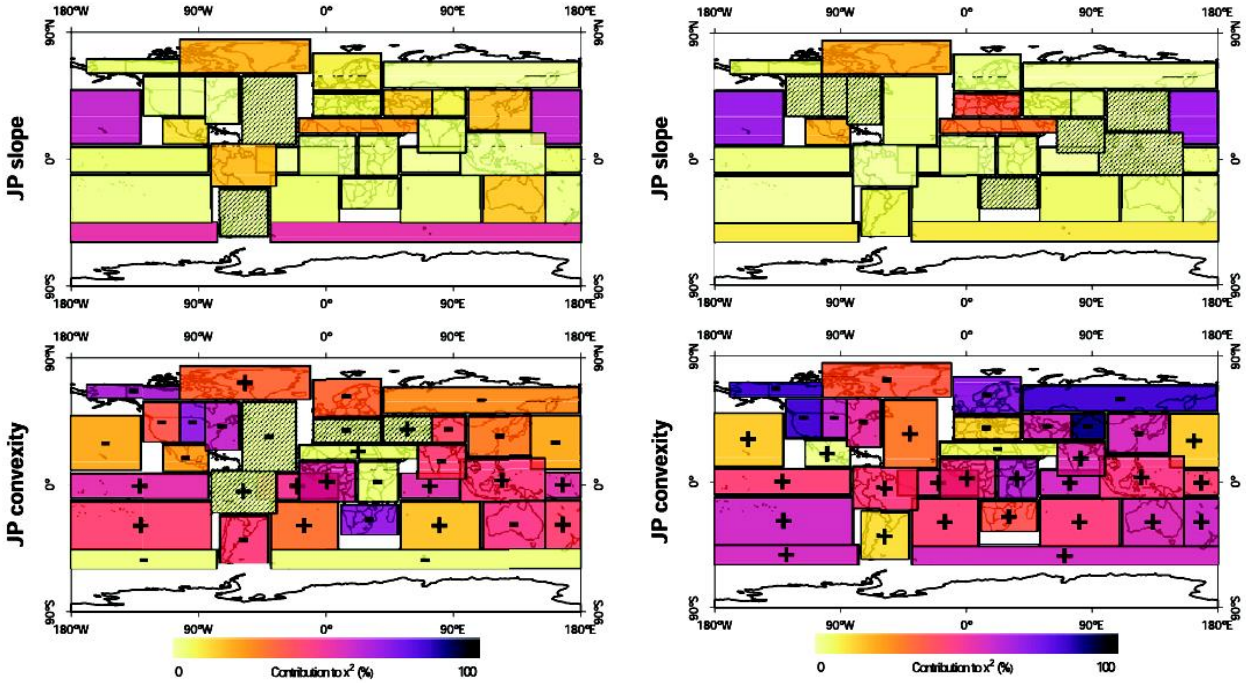


Figure 5: Maps of the Jolliffe and Primo (2008) (top) slope and (bottom) convexity coefficients, expressed as their contribution to the $\chi 2$ coefficient (%), for near-surface air temperature for the 30 different regions, for forecast year 1 in the multi-model ensemble (left) and large single-model ensemble (right). Going from light yellow to dark purple, the colors denote an increasing role of the slope and the convexity terms to decrease the reliability of the ensemble (diagnosed by the deviations from flatness in the rank histogram). A plus (minus) sign in the convexity coefficient maps represents an underdispersive (overdispersive) forecast. Hatching represents regions where the p-value is larger than 0.05, thus where there is no evidence of bias, difference in trend, or error in dispersion (the null hypothesis being that the rank histograms are flat).

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

Given the unreliability of the raw model simulations, we also tested the potentials to correct the reliability of the climate predictions through different post-processing techniques (all results summarised in figure 6). The results for the detrended multi-model ensemble show that detrending the data improves the JP slope coefficient increasing reliability in many regions (Figure 6 top triangle of the matrix fields). However it degrades the JP convexity coefficient in many regions. Reliability is not significant in any region after detrending, suggesting the lack of reliability cannot be only due to a misrepresentation of the observed trend (i.e. with a $\chi 2$ p-value above 0.05, the null hypothesis being that the rank histograms are flat).

The simple bias correction (correcting mean and variance) increases reliability in many regions (lower JP slope coefficient, bottom triangle in figure 6). For this correction method, the JP convexity coefficient does not systematically increase unlike in the detrending correction method above. Similar to detrending, bias correction does not lead to significantly reliable predictions in any region (i.e. $\chi 2$ p-value above 0.05). This indicates that errors in the mean variance are also not primarily responsible for the lack of the reliability of the ensemble predictions.

We further tested the effects of calibration based on variance inflation (Doblas-Reyes et al., 2005). As in the previous two correction methods, calibration greatly improves the JP slope coefficient (Figure 6, right triangle of the matrix fields). Additionally and unlike previously the ensemble

predictions become significantly reliable in one region (CAM, Central America). Calibration also improves the JP convexity coefficient in several regions.

In regions with more reliable raw forecasts, the effect of calibration is not as high or even reverse. For example, the North Atlantic Ocean (NAT) and Mediterranean basin (MED) regions, which already had low contributions to $\chi 2$ of the JP convexity coefficients in the uncorrected thus comparably higher reliability, calibration leads to an increase in the contribution of convexity thus resulting in less reliable forecasts.

Generally 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 at least this region.

Comparing these results to the multi-annual mean forecast for the first years after initialisation the convexity results can be worse than the uncorrected ensembles after post-processing if the contribution of $\chi 2$ of the JP convexity coefficient is already low in the uncorrected ensemble. The slope results remain similar after the post-processing (compare upper and lower panel of figure 6). The evolution of the convexity with forecast time should be interpreted as the evolution of the ensemble spread with respect to the observation, which would usually grow with forecast time until the convexity saturates.

The improvement in the JP convexity coefficient after calibration for the single-model ensemble is generally larger than for the multi-model ensemble (not shown).

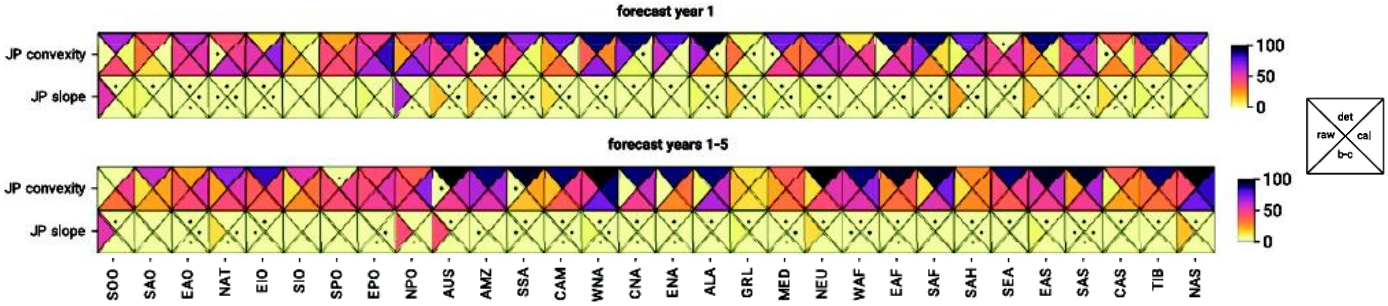


Figure 6:. Summary of the Jolliffe and Primo (2008) slope and convexity coefficients, expressed as their contribution to the $\chi 2$ coefficient (%), for near-surface air temperature for 30 different regions for forecast year 1 and forecast years 1–5 in the multi-model ensemble. For each forecast time, the two rows represent the JP coefficients. Diamonds indicate cases where the p-value is larger than 0.05, thus where there is no evidence of bias, difference in trend, or error in dispersion (the null hypothesis being that the rank histograms are flat). For colour codes, please refer to figure 5. Each triangle displays the result for a type of postprocessing (either raw uncorrected values, det=detrended, b-c=bias-corrected, or cal=calibrated).

IV. Main conclusions

Results indicate that uncorrected output of near-term climate prediction ensembles are largely not reliable for near-surface air temperature, and that initialisation does not significantly improve the reliability in most cases in comparison to uninitialised climate projections.

Using different forecast system ensembles has an impact on reliability. 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 demonstrates the need for bias correction and calibration of the raw data. This is crucial to obtain reliable near-term climate predictions that can be useful to stakeholders to obtain more realistic estimates of event probabilities.

3.2 Met Office

3.2.1 Multi-model multi-year prediction products

The work on combining information from multiple sources to generate forecasts for near-term climate has been centred around the outputs displayed at the WMO Lead Centre for Annual-to-Decadal Climate Prediction (LC-ADCP) which is a portal for international sources of prediction information under the World Meteorological Organisation.

The set of products on display on the Lead Centre's website (www.wmolc-adcp.org) has been updated with data from predictions started from 2020 initial conditions. This time, in addition to the contribution from the designated Global Producing Centres (GPCs) for Annual to Decadal Predictions (BSC, CCCMA, DWD, MOHC), forecasts were obtained from BCCR, CSIRO, SMHI/DMI, GFDL and NRL. The forecasts are provided for temperature, sea-level pressure, precipitation and Atlantic meridional overturning circulation (AMOC) (with the exception of NRL, from which only temperature is available, and DWD, which does not include the AMOC).

Alongside predictions, verification information has also been prepared for the Centre's website and been made publicly available. The assessment metrics currently used are deterministic scores: (Pearson) correlation and probabilistic scores: the Relative Operating Characteristic (ROC) at grid-box resolution. These are calculated for temperature, precipitation and sea-level pressure, for periods of one year (first year in the forecast range) and 5 years (the mean over the first 5 years of the forecast range). The time aggregation is the same as that used in the definition of the prediction products.

Though not yet included in the published products, data from a new contributing centre, CMCC in Italy, has also been collected as input to the LC products and will later be added to the combination. Later in the year, new products will be developed for this multi-model, with global mean temperature time series and multi-year seasonal averages expected soon.

A new issue of the WMO Global Annual to Decadal Climate Update (GADCU) has been prepared and published. This includes new diagnostics evaluating the probability of exceedance of set thresholds. This prediction information is based on a total of ~100 ensemble members from the nine prediction systems that provided predictions initialised in 2020: 10 members from each BCCR, BSC, CCCMA, CSIRO, DWD, GFDL, 15 from DMI/SMHI, 20 from MOHC, and 1 from NRL (NRL only submit temperature data, so results for precipitation and pressure are calculated based on only 95 members from 8 different prediction systems).

The analysis of data from all models participating in the Lead Centre's activity concludes that:

- Annual mean global (land and sea) mean near-surface air temperature is likely (>66% chance) to be at least 1°C warmer than pre-industrial levels (defined as the average over the years 1850-1900) in each of the coming 5 years (2021-2025) and is very likely (>90% chance) to be within the range 0.9 – 1.8°C. It is about as likely as not (40% chance) that at least one of the next 5 years will be 1.5°C warmer than pre-industrial levels, and the chance is increasing with time, but is is very unlikely (<10% chance) that the five-year mean global near-surface air temperature for 2021-2025 will be 1.5°C warmer than pre-industrial levels. The chance of at least one year exceeding the current warmest year, 2016, in the next five years is 90%.

- In 2021, large land areas in the Northern Hemisphere are likely (>66% chance) to be over 0.8°C warmer than the recent past (defined as the 1981-2010 average); the Arctic (north of 60°N) is likely (>66% chance) to have warmed by more than twice as much as the global mean compared to the recent past. Southwestern North America is likely (>66% chance) to be drier, whereas the Sahel region and Australia are likely to be wetter, than the recent past (figure below).

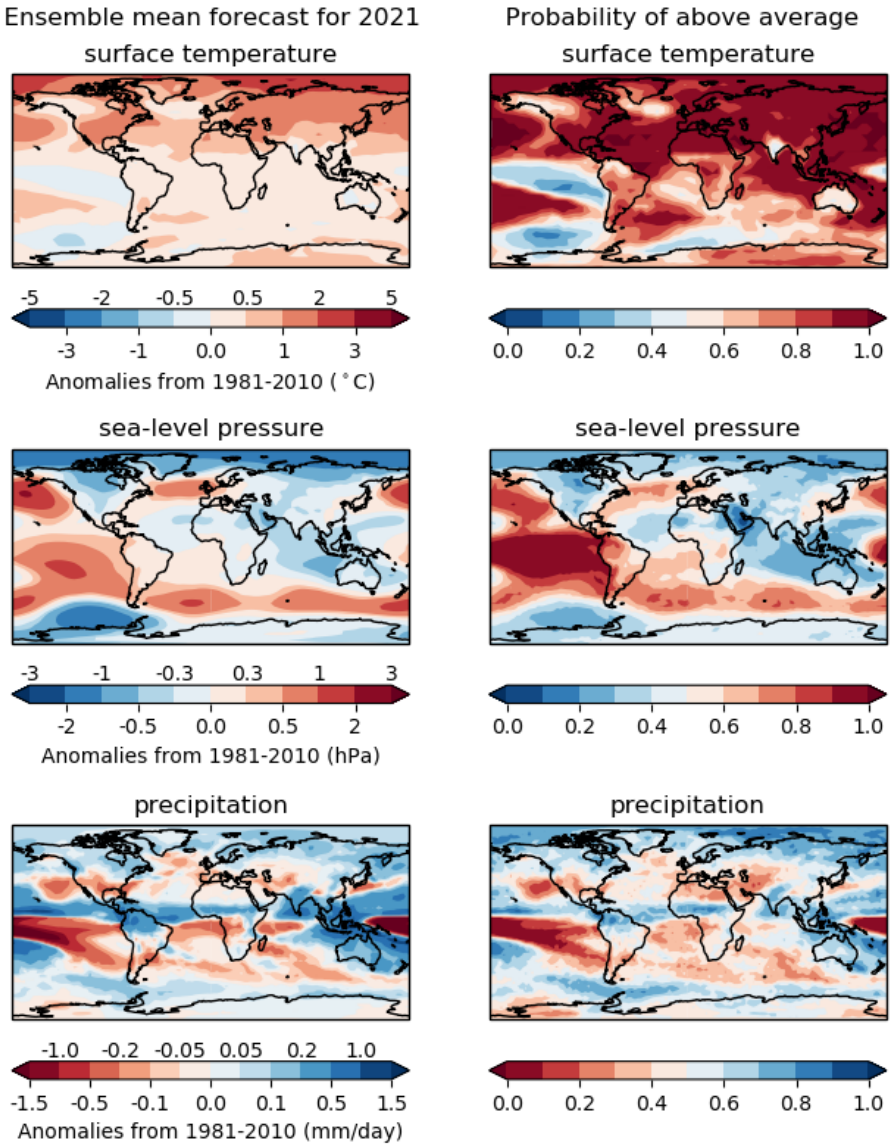


Figure 7: Annual mean anomaly predictions for 2021 relative to 1981-2010. Ensemble mean (left column) for temperature (top, °C), sea level pressure (middle, hPa), precipitation (bottom, mm/day) and probability of above average (right column) based on numbers of ensemble members. As this is a two-category forecast, the probability for below average is one minus the probability shown in the right column.

- Over 2021-2025, almost all regions, except parts of the southern oceans and the North Atlantic are likely (>66% chance) to be warmer than the recent past (defined as the 1981-2010 average); high latitude regions and the Sahel are likely to be wetter than the recent

past; there is an increased chance of more tropical cyclones in the Atlantic compared to the recent past (based on inference from temperature and sea-level pressure predictions).

The GADCU includes newly derived information on the skill of this combination of model output, as well as an evaluation of forecasts covering the most recent past. The skill of predictions of indices is illustrated in the Update using contingency tables, correlation and the Mean Square Skill Score (MSSS). The example in the figure below is for the Pacific Decadal Variability (PDV).

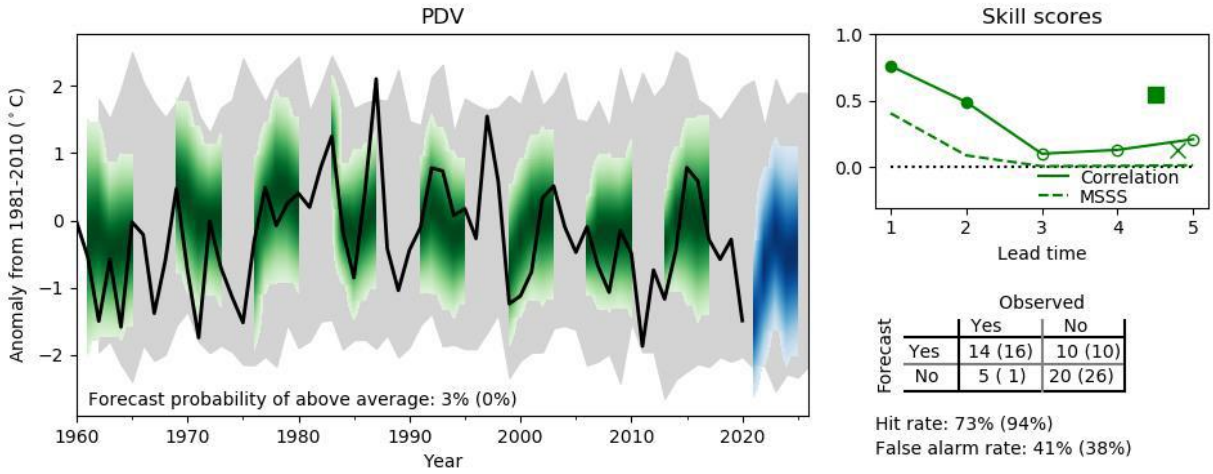


Figure 8: Multi-annual predictions of Pacific Decadal Variability (PDV) - defined as the difference in SST between the eastern tropical Pacific (10°S-6°N, 110°W-160°W) and the North Pacific (30°N-45°N, 145°W-180°W) as in Dong et al (2014) - relative to its 1981-2010 average. Annual mean observations in black, forecast in blue, hindcasts in green and uninitialised simulations in grey. The shading indicates the 90% confidence range. The probability for above average in the five year mean of the forecast is given at the bottom the main panel (in brackets the probability for above average in the next year). Hindcast skill scores are shown in the upper right panel, the square and the cross show the correlation skill and Mean Square Skill Score (MSSS) for five-year means, respectively. Significant correlation skill (at the 5% confidence level) is indicated by solid circles/square. The contingency table for the prediction of above average five year means is shown in the bottom right panel (in brackets values for above average in the next year).

An evaluation of forecasts covering the most recent past (in this instance, the forecast initialised at the end of 2015 covering 2016-2020) is also included. The conclusions of this evaluation can be found in the full text of the Update, on the Lead Centre's website (www.wmolc-adcp.org).

3.3 CNRS/IPSL

3.3.1 Estimating the probabilistic risk of having an abrupt change in the SPG in the on-going century

CMIP5 models have been shown to exhibit rapid cooling (2-3°C in less than 10 years) events in their projections of the North Atlantic subpolar gyre. The exact timing and year of occurrence of those events were model-dependent, and related to non-linear response of the subpolar gyre, which is known to be a tipping element of climate (Swingedouw et al. 2020). As a consequence, the near-term changes in the North Atlantic, and over Europe as well, remains very uncertain and subject to potential decadal events which are difficult to predict since they are highly model dependent and non-linear. Decadal initialised hindcasts for near-term climate have been only performed by a few institutes, notably because these are complex and time-consuming experiments. Furthermore, those hindcasts usually suffer from large drift, due to adjustment of the model to observed initial conditions. In this respect, the DCPP-A database of hindcast is not offering a good way to properly estimate the risk of near-term abrupt events, due to those drifts in model trajectory, as well as the poor sampling of model diversity. This is why, as a first step, it seems more useful to assess in CMIP6 projections, which have been performed with numerous different models, the probabilistic risk of encountering abrupt cooling events in the coming decades.

Here, we have analysed the CMIP6 projections archive, searching for such rapid cooling events in the new generation of models. We were searching in projections, following the approach of Sgubin et al. (2017), events that exceed 3 standard deviations of 10-year differences from pre-industrial control simulations. We were able to analyse surface temperature in the subpolar gyre in 35 models in the first available member of each.

Four models out of 35 exhibit such instabilities (see 3 examples of them in figure 9). The climatic impacts of these events are large on decadal time scales, with a substantial effect on surface temperature over Europe, precipitation pattern in the tropics - most notably the Sahel and Amazon regions - and a possible impact on the mean atmospheric circulation. The mechanisms leading to these events are related to the collapse of deep convection in the subpolar gyre, modifying profoundly the oceanic circulation.

Analysis of stratification in the subpolar gyre as compared to observations (Figure 10) highlights that the biases of the models explain relatively well the spread in their projections of surface temperature trends in the subpolar gyre: models showing the smallest stratification biases over the recent period also show the weakest warming trends. The models exhibiting abrupt cooling rank among the 11 best models for this stratification indicator. Based on this emergent constraint, we evaluate the probability risk of encountering an abrupt cooling in the on-going century of up to 36.4%, slightly lower than the 45.5% estimated in CMIP5 models in Sgubin et al. (2017). The whole study has been published 2021 in the Annals of the New York Academy of Science in a special issue named “a year in climate”.

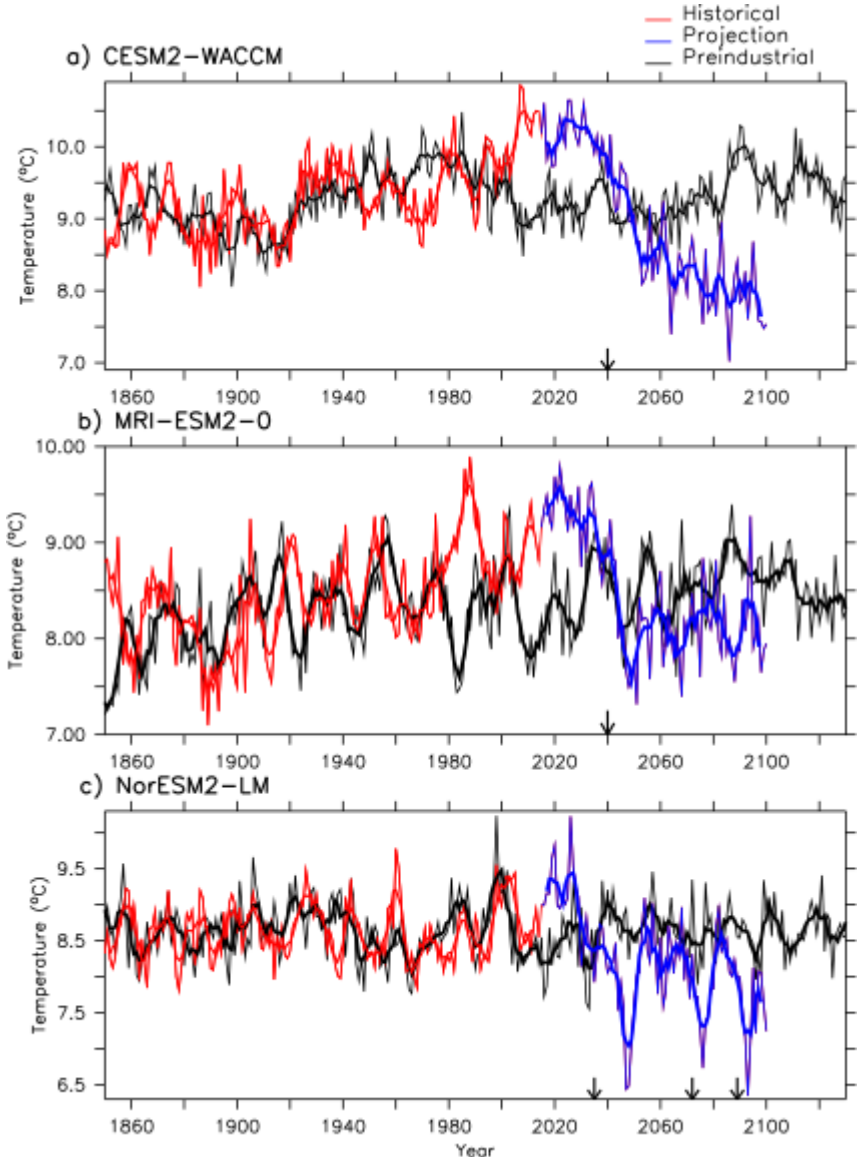


Figure 9: Examples of abrupt changes found in the subpolar gyre. Time series of near-surface air temperature (in °C) averaged over the SPG (70°W-20°W, 45°N-60°N) in annual mean (thin line) and 5-year running mean (thick line). In black is the pre-industrial simulation, in red the historical simulation and in blue the projection considered for a) the CESM2-WACM model with ssp126 scenario, b) the MRI-ESM2-0 model with ssp245 scenario, c) the NorEMS2-LM model with ssp126 scenario. The black arrows represent the approximate starting of the abrupt events.

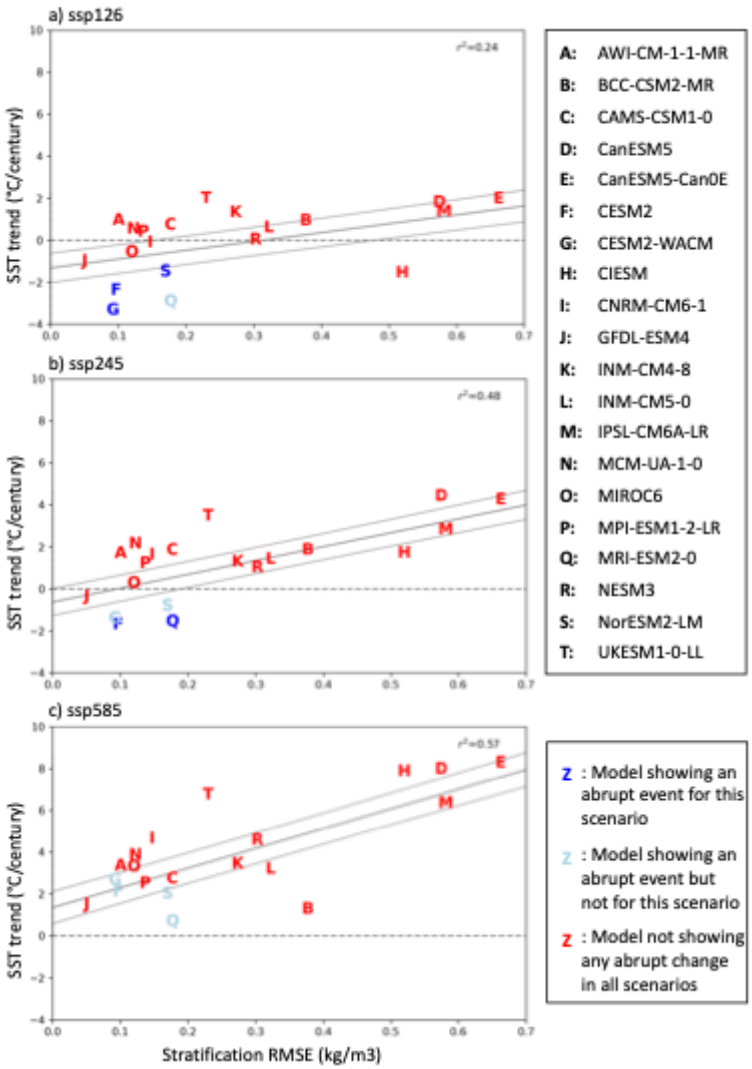


Figure 10: Link between stratification and centennial temperature trend over the SPG. Scatterplot of the root mean square error of the density in the SPG as compared to observation depicted in Figure 7, for the period 1985-2014, averaged over the first 2000 meters of the ocean vs. the linear trend of Surface Atmospheric Air temperature (in °C/century) computed in different emission scenarios, a) ssp126, b) ssp245, c) ssp585. The letters correspond to the models enumerated in the Method section in their alphabetical order. The red letter corresponds to a model showing no abrupt changes, while the blue letters indicate a model showing an abrupt event for the considered scenario, and it is in light blue when it is not occurring in this particular scenario, but still corresponds to a model that does show abrupt changes for other emission scenarios.

Follow up studies might be interested in considering a few more models when available, different members available for some models, as well as the DCPP-A database. From this first study, it appears that among the models showing abrupt changes, only Nor-ESM was participating in DCPP-A (but was not part of the EUCP consortium), and might be able to provide near-term initialised forecasts (which were not required in DCPP-A).

3.4 DMI

3.4.1 Assessment of *probability forecasts* in the North Atlantic

Probability predictions of the North Atlantic Oscillation (NAO) and the Atlantic Multi-decadal Oscillation (AMO) have been investigated. The NAO dominates the North Atlantic winter variability and the AMO influences air temperatures and rainfall over much of the Northern Hemisphere as well as the Atlantic hurricane activity.

We use the CMIP6 multi-model ensemble (Eyring et al. 2016) and the single-model Community Earth System Model (CESM) Large Ensemble Project (LENS, Kay et al. 2015). For the CMIP6 historical experiments we have 215 ensemble members from 51 different models for temperature and 213 ensemble members from 49 models for pressure. For the ssp2-4.5 scenario we have 126/116 (temperature/pressure) members from 35/34 models. For the CMIP6 decadal forecast (DCPP-A, Boer et al. 2016) we have 79 members for temperature and 75 members for pressure from 9 different models. The LENS ensemble has 40 ensemble members for historical, scenario, and forecasts (CESM Decadal Prediction Large Ensemble Project, DPLE, Yeager et al. 2018). As observations, we use Hurrell's station-based index (1865-2020) for the NAO and for the AMO the NOAA PSL index (1856-2020, Enfield et al. 2001). For all ensembles and observations we have used monthly means. Prior to the analysis, models data are interpolated to a common 2.5x2.5 degrees global grid using a simple nearest neighbour procedure.

The AMO is calculated as the area weighted mean of near-surface air temperature over the Atlantic Ocean grid-points between 0-60 N. As the global mean temperature and the Atlantic temperature are much more connected in models than in observations we have not removed the part congruent to the global mean from the AMO. Also, we have not detrended the AMO as the detrending depends on the period involved. Thus, our AMO index includes the full forced signal. Note the similarity between the detrended NOAA PSL index and the non-detrended index calculated from the JRA-55 reanalysis (Fig. 11). The NAO is based on monthly sea-level pressure anomalies. It is calculated as the difference between normalized anomalies between Azores (mean over 20-28 W, 36-40 N) and Iceland (mean over 12-16 W, 63-70 N). The winter means are calculated over December to February. Time-series for the AMO are shown in Figure 11.

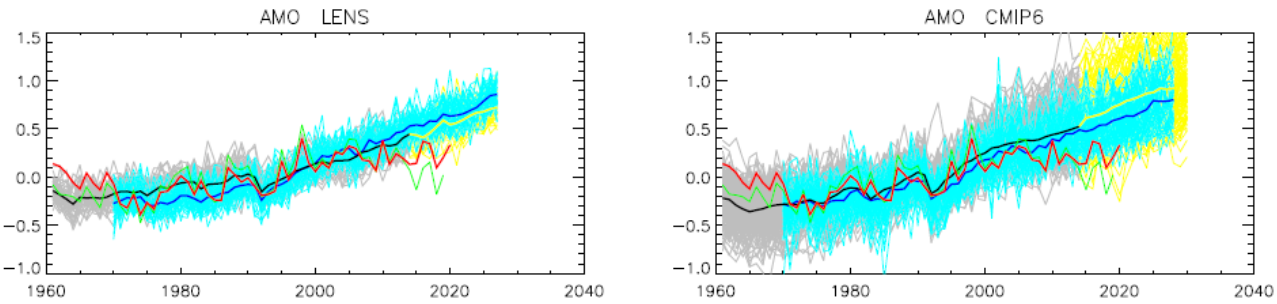


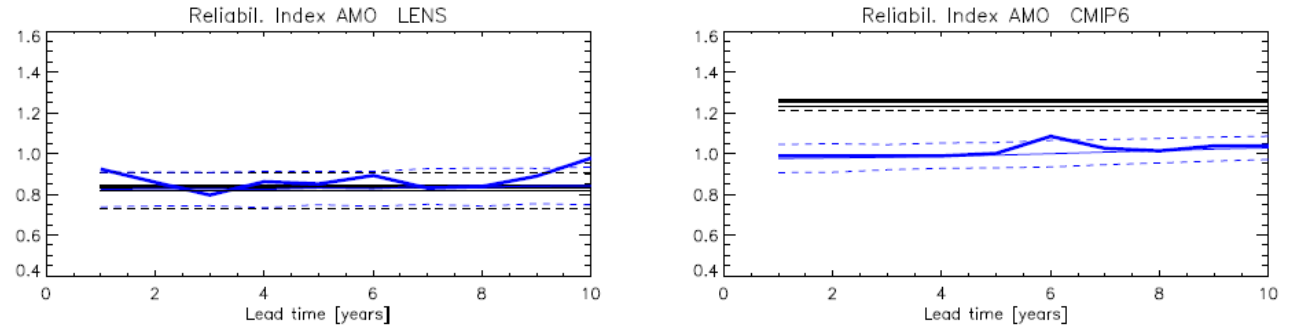
Figure 11. Time-series of AMO [K]. Black: CMIP6 Historical. Yellow: CMIP6 scenario ssp2-4.5. Cyan: DCPP-A Forecast. Thick curves are ensemble means, thin curves the individual ensemble members. Red: Observations (Green: AMO calculated from JRA-55 reanalysis). All series centred to zero in 1961-2015. Lead-time 10 years for forecasts.

The reliability of the ensemble forecasts has been assessed by rank histograms and the related reliability index. This index is defined as the absolute difference between the histogram values and

1/(m+1), where m is the ensemble size, summed over the whole histogram. Low values of the index indicate better reliability, but the index is sensitive to both the sample size and the size of the ensemble (Wilks 2019). We therefore use a Monte-Carlo procedure to estimate the sampling variability for the reliability index of a flat histogram. If the observed index falls within this variability (shown as the mean plus/minus two standard deviations in Fig. 12a), we consider the forecast reliable. As an overall measure of the quality of the probabilistic predictions, we have considered the continuous ranked probability score. This measure has been compared to climatology. The sampling variability of the climatology is estimated by a bootstrap procedure (shown as plus/minus two standard deviations in Fig. 12b).

For the AMO we find that for both LENS and CMIP6 the forecast ensembles are reliable for all lead times (Figure 12a) as they don't deviate significantly (except perhaps for LENS at lead-time 10 years) from the values found for flat histograms. The historical experiment is reliable for LENS, but on the edge of significance for CMIP6 (Figure 12a, black curves). The spread around the ensemble mean is smaller for the forecast ensemble than for the historical ensemble for all lead times in CMIP6 and for the lowest lead times in LENS. The correlations between ensemble means and observations are larger for the forecast ensembles than for the historical ensembles for the shorter lead times in both LENS and CMIP6 (not shown). For the continuous ranked probability score (Figure 12b) we find that both LENS and CMIP6 are superior to climatology (due to the general trend). We also see an improvement in forecast ensemble over the historical ensemble for the first lead times for both LENS and CMIP6.

(a)



(b)

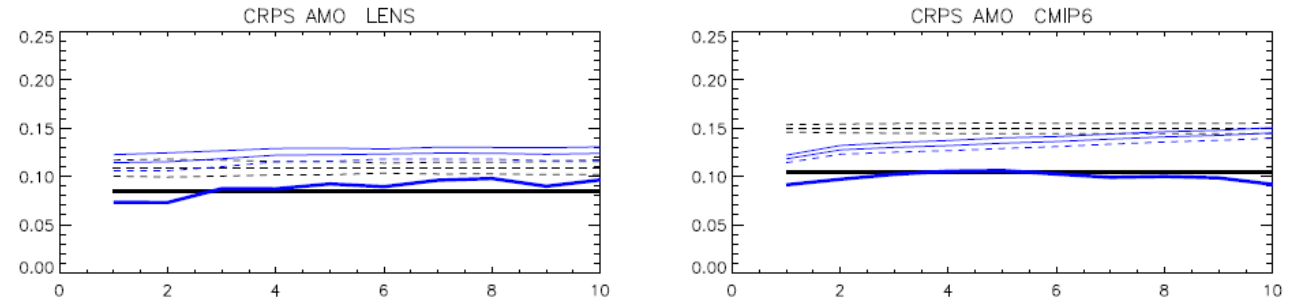


Figure 12. The reliability index (a) and the continuous ranked probability score (b) as function of lead time for AMO. Thick solid black curve: Historical. Thick solid blue curve: Forecast. For the reliability index the thin full and dashed curves give the mean and plus/minus two standard deviations under the null-hypothesis of a flat histogram. For the CRPS the thin full and dashed curves are mean values for the climatology and plus/minus two standard deviations.

AMO predictions have also been obtained for 2022-2027 (Figure 13). We find increasing AMO strength for both LENS and CMIP6. For CMIP6 the spread in the forecast ensemble is smaller than the spread in the scenario ensemble. Note, however, that both ensembles have problems catching the stagnating observed values from 2010-2020 as shown in Figure 11.

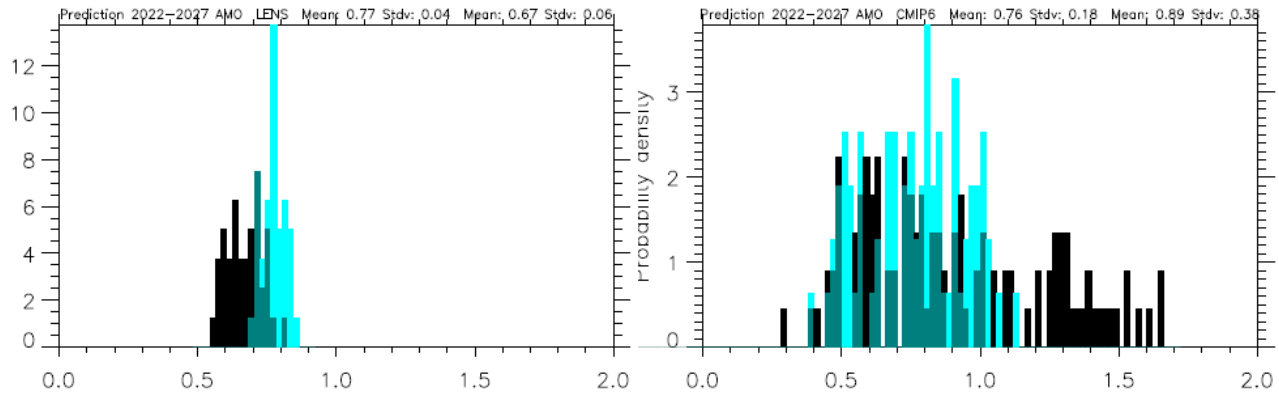


Figure 13. The predictions of the AMO from probability forecasts (cyan) and scenario experiments (black) for the period 2022-2027. For the forecasts we have averaged over all lead times. Left: LENS. Right: CMIP6.

Recently, the NAO has been suggested to be predictable on decadal time-scales. However, the signal in individual model experiments is very weak and it requires averaging over very large ensembles to obtain significant positive correlations with observations (Smith et al. 2019, Smith et al., 2020, Klavans et al. 2021). In particular, we don't find any skill in either the historical ensemble or the forecast ensemble when we apply the reliability analysis. The NAO probability forecast does not distinguish from the scenario ensemble. This is probably what should be expected from the weak signal and the fact that the correlation between observations and model mean depends on the period. Figure 14 shows the correlations between ensemble mean and observations for different 45 years periods as a function of start year for both LENS and CMIP6. The correlations are very variable and for CMIP6 only above 0.5 for start years after 1955 and before 1890. For LENS the correlations are not significant. It is interesting to note that both ensembles show an increase of correlations in the period after 1920. This could indicate that the non-stationarity of the correlations is due to changes in the climate system and not due to chance.

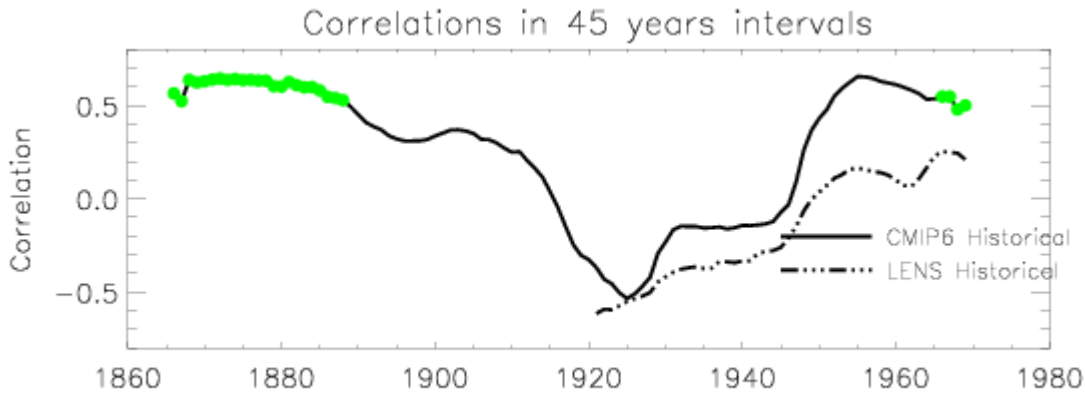


Figure 14. Correlations between ensemble mean NAO and observation in 45 years periods. Plotted as a function of the start year of the 45 years period. Green symbols indicate correlations that are statistically significantly different from zero.

Further analyses will include an attempt to formally decompose the continuous ranked probability score into its components. Other climate indices will also be investigated. We are now finalizing the reported analyses and summarizing the results as a scientific manuscript.

3.5 SMHI

3.5.1 Probabilistic skill assessment based on novel temporal pooling approach and CMIP6-DCPP multi-model ensemble

The peculiarity of our 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 the approach presented here 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 is the first effort to implement this approach for two-dimensional data fields and a large multi-model ensemble (MME) of decadal climate predictions.

The implementation in our study is described in the following. A large multi-model ensemble of CMIP6-DCPP decadal predictions was compiled (eight different models, ten prediction systems, 108 ensemble members in total). Thresholds defining the events of interest were calculated for all models individually to account for existing biases (e.g. extremely hot summer being the case if the JJA-mean 2m air temperature is within the upper sextile of the respective model's climatological distribution).

Different pooling intervals are possible. However, sensitivity tests showed that skill scores typically improve until 4-6 consecutive forecast years are pooled together, afterwards no significant improvement is achieved anymore (not shown). Therefore, it was decided to pool 5 consecutive years, i.e. forecast years 1-5, for this study.

The forecast probability of e.g. summers to be extremely hot within the next five years was empirically derived - that means calculating the ratio of values showing these events based on a total sample size of 540 (108 ensemble members times 5 consecutive years) - for each of the 32 hindcasts started towards the end of the years 1978-2009 (initialisation in 1978 providing data for the summers 1979-1983).

Verification by means of the Brier (Skill) Score was performed against observed probabilities for the given 5-year periods (i.e. probabilities of 0, ⅕, ⅖, ⅗, ⅘, or 1) with ERA5 (for temperature) and GPCPv2.3 (for total precipitation) as observational reference data sets.

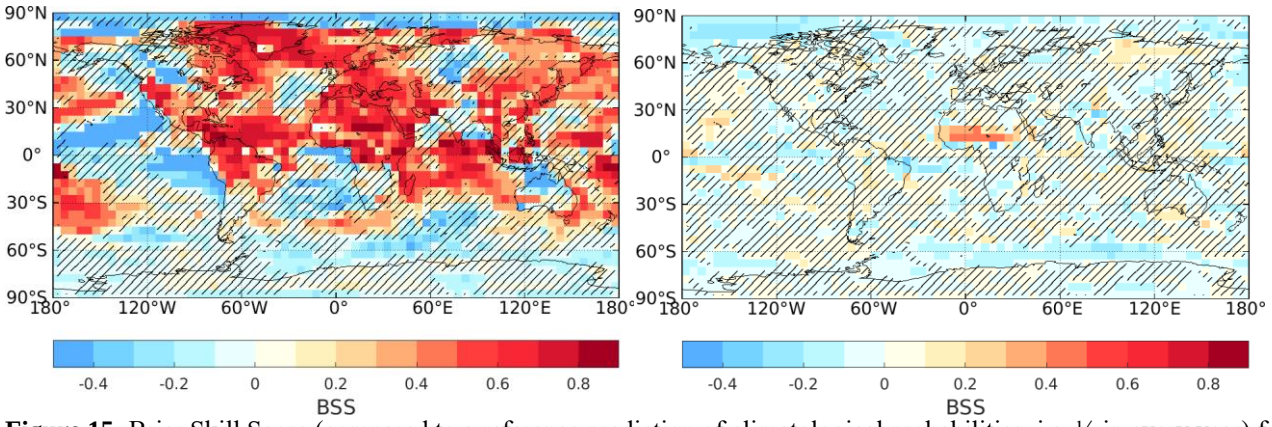


Figure 15: Brier Skill Score (compared to a reference prediction of climatological probabilities, i.e. ⅙ in every year) for the CMIP6-DCPP multi-model ensemble in predicting the probability of a boreal summer (JJA) within the five years after initialization being extremely hot (left: 2m air temperature within local upper sextile) and extremely dry (right: total precipitation within lowest sextile); skill assessment (based on ERA5 and GPCPv2.3 as observational reference datasets) for evaluation period 1979-2014, based on 32 hindcasts s1978-s2009 from eight different models with a total ensemble size of 108 members; hatching marks region where the BSS is not significant (p>0.01).

Fig. 15 shows this first evaluation of skill for the temporal pooling approach and the multi-model ensemble of CMIP6-DCPP-hindcasts forecasting probabilities of boreal summers (JJA) within five years after initialization being extremely hot (Fig. 15 left) and extremely dry (total precipitation within lowest sextile, Fig. 15 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) over 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 level of forecast skill however is comparable to that derived from assessing the standard multi-annual averages (not shown).

At the time of writing this report, an equivalent analysis of uninitialized *historical*-simulations is conducted in order to assess the benefit from initialization for this type of forecast information. This works states a collaboration with WP5 (deliverable D5.2). A peer-reviewed publication of the approach and the results for the CMIP6-DCPP-MME is in preparation.

4. Lessons learnt

Increasingly large ensembles of initialised decadal hindcasts and predictions are becoming available, enabling robust probabilistic analysis and the development of probabilistic forecast products. This Deliverable reports on several complementary activities within EUCP towards understanding probabilistic characteristics of existing decadal predictions, and processing the predictions from multiple systems towards the development of probabilistic predictions. The key conclusions of this work are:

- Decadal hindcast experiments in a multi-model context have been coordinated as part of CMIP5 and CMIP6. The DCPP-A component of CMIP6 now includes decadal hindcasts with annual initialisation during 1961-2015 from ~10 different prediction systems.
- Several of these prediction systems also provide quasi-operational predictions initialised towards the end of each calendar year, and provide annually updated decadal predictions. These activities are coordinated under the WMO Lead Centre for Annual-to-Decadal Climate Prediction, and the latest predictions initialised in 2020 include almost 100 ensemble members from 9 different forecast systems
- Different ways to construct a multi-model ensemble based on the different prediction systems (pooling all ensemble members from all models versus first calculating ensemble averages for each model before combining predictions from the different models) lead to very similar results in terms of probabilistic skill measures, suggesting the way of constructing the multi-model ensemble is not a critical choice
- Further ways to construct a multi-model based on weighting by individual model performance have recently been applied for historical and scenario forced simulations in EUCP WP2. This approach has currently not yet been pursued in the field of initialised climate predictions. Performance weighting methods might however help to increase forecast quality and usability and should thus be considered a possible option for future research. The forecast time performance dependence has to be taken into account and will require additional attention in comparison to the already existing methods.
- Reliability is a crucial probabilistic characteristic of a prediction system, indicating that the predicted probability of an event corresponds to observed frequency of the event type. Reliability is therefore crucial for decision-makers if decisions are based on the prediction of specific events. Different analyses evaluating the reliability of multi-model decadal predictions (in particular Sections 3.1.2 and 3.4.1) come to different conclusions, the first concluding that the multi-model decadal predictions are not reliable, and the second concluding reliability. Both analyses are based on different metrics to evaluate reliability and different prediction systems used. Future research is needed to systematically reconcile the reliability of state-of-the-art prediction systems, using standardised measures, to attribute such differences to choices of metrics and modelling systems used.
- An increase in ensemble members and/or forecast systems for the multi-model is beneficial for the prediction skill and reliability. The combination of different forecast systems is thereby of greater importance than the number of individual ensemble members.
- New methods to derive and evaluate probabilistic information are being developed, and can provide complementary information over previously used methods. A specific example illustrated in Section 3.5.1 pools the information predicted for different individual years rather than averaging over several years. This has the advantage of evaluating certain seasonal or annual events based on a larger sample size. Current results show that the prediction of temperature extremes of seasonal means based on this approach reveals greater

forecast quality than a climatological reference forecast. Future research should further understand the limitations and benefits of this approach.

- The probability risk of encountering an abrupt cooling of the subpolar gyre until the end of the 21st century is in the order of 35% based on most recent climate projections. A remaining challenge will be to estimate this probability with initialised climate predictions. The handling of their forecast drift, in some cases favouring the tendency towards abrupt cooling (Bilbao et al. 2021), might constitute an important obstacle for any future research in this direction.
- Climate predictions initialised in 2020 reveal that annual mean global mean near-surface air temperature is likely to be at least 1°C warmer than pre-industrial levels in each of the coming 5 years. In 2021, large land areas in the Northern Hemisphere are likely to be over 0.8°C warmer than the recent past.Southwestern North America is likely to be drier, whereas the Sahel region and Australia are likely to be wetter, than the recent past.
  The probabilistic information about the near-future temperature will be continuously reported to the public and thus remain a very important aspect in the field of climate prediction.
- The importance of reliable IT infrastructure for data storage, data exchange, efficient computing of climate model output, and subsequent analyses cannot be underestimated. The large amount of data that has to be processed will remain one of the most time-consuming parts of climate prediction research. Efforts for increased efficiency should always be part of any future research agenda in this field.

5. Acronyms

AMO: Atlantic Multi-decadal Oscillation
AMOC: Atlantic Meridional Overturning Circulation
BSC: Barcelona Supercomputing Center - Centro Nacional de Supercomputación
BSS: Brier Skill Score
CESM: Community Earth System Model
CMCC: Centro Euro-Mediterraneo sui Cambiamenti Climatici
CMIP5/6: Coupled Model Intercomparison Project Phase 5 / Phase 6
CNRS: Centre National de la Recherche Scientifique
DCPP: Decadal Climate Prediction Project
DMI: Danish Meteorological Institute
DPLE: Decadal Prediction Large Ensemble Project
DWD: Deutscher Wetterdienst
ECMWF: European Centre for Medium-Range Weather Forecasts
ERA: ECMWF Reanalysis
GADCU: WMO Global Annual to Decadal Climate Update
GHCNv4: Global Historical Climatology Network Version 4
GISTEMP: NASA GISS Surface Temperature Analysis
GPC: Global Producing Centres
GPCC: Global Precipitation Climatology Centre
GPCP: Global Precipitation Climatology Project
IPSL: Institut Pierre-Simon Laplace
JP: Jolliffe-Primo test statistic
JRA: Japanese Reanalysis
MSSS: Mean Square Skill Score
MME: Multi-Model Ensemble
MOHC: Met Office Hadley Centre
NAO: North Atlantic Oscillation
NOAA: National Oceanic and Atmospheric Administration
LENS: Large Ensemble Project
LC-ADCP: Lead Centre for Annual-to-Decadal Climate Prediction
ROC: Relative Operating Characteristic
PDO: Pacific Decadal Variability
PSL: Pressure at Sea Level
RPSS: Ranked Probability Skill Score
SMHI: Sveriges Meteorologiska och Hydrologiska Institut
SPG: Subpolar Gyre
SST: Sea Surface Temperature
WMO: World Meteorological Organisation

6. References

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

Bilbao, R., Wild, S., Ortega, P., *et al.* Assessment of a full-field initialized decadal climate prediction system with the CMIP6 version of EC-Earth, *Earth Syst. Dynam.*, **12**, 173–196 (2021) https://doi.org/10.5194/esd-12-173-2021

Boer, G. J., Smith, D. M., Cassou, C. *et al.* The Decadal Climate Prediction Project (DCPP) contribution to CMIP6, *Geosci. Model Dev*. **9**, 3751–3777 (2016) https://doi.org/10.5194/gmd-9-3751-20

Buontempo, C., C. Hewitt, F. Doblas-Reyes, and S. Dessai, 2014: Climate service development, delivery and use in Europe at monthly to inter-annual timescales. *Climate Risk Manage.*, **6**, 1–5, (2014). https://doi.org/10.1016/j.crm.2014.10.002.

DelSole, T., & Tippett, M. K. Forecast Comparison Based on Random Walks, *Monthly Weather Review* **144**(2), 615-626 (2016) https://doi.org/10.1175/MWR-D-15-0218.1

Doblas-Reyes, F. J., Hagedorn, R., and Palmer, T. N. 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 (2005) https://doi.org/10.3402/tellusa.v57i3.14658

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

Enfield, D. B., Mestas-Nunez, A. M., and Trimble, P. J. The Atlantic Multidecadal Oscillation and its relation to rainfall and river flows in the continental U.S., *Geophys. Res. Lett*. **28**, 2077–2080 (2001) https://doi.org/10.1029/2000GL012745

Eyring, V., Bony, S., Meehl, G. A. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, *Geosci. Model Dev.* **9**, 1937–1958 (2016) https://doi.org/10.5194/gmd-9-1937-2016

Ferro, C. A. T. Fair scores for ensemble forecast, *Quarterly Journal of the Royal Meteorological Society* **140**, 1917-1923 (2014) https://doi.org/10.1002/qj.2270

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

Hemri, S., Bhend, J., Liniger, M. A. *et al.* How to create an operational multi-model of seasonal forecasts? *Climate Dynamics* **55**, 1141–1157 (2020) https://doi.org/10.1007/s00382-020-05314-2

Hewitt, C. D. and Lowe J.A. Toward a European Climate Prediction System. *Bulletin of the American Meteorological Society*, **99(**10), 1997–2002. (2018)

Jolliffe, I., and C. Primo. Evaluating rank histograms using decompositions of the chi-square test statistic. *Mon.Wea. Rev.* **136**, 2133–2139 (2008) https://doi.org/10.1175/2007MWR2219.1.

Kay, J. E., Deser, C., Phillips, A. *et al.* The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability, *Bull. Am. Meteorol. Soc.* **96**, 1333–1349 (2015) https://doi.org/10.1175/BAMS-D-13-00255.1

Kharin, V. V. and Zwiers, F. W. On the ROC Score of Probability Forecast, *Journal of Climate* **16**(24), 4145-4150 (2003) https://doi.org/10.1175/1520-0442(2003)016%3C4145:OTRSOP%3E2.0.CO;2

Klavans, J. M., Cane, M. A., Clement, A. C., and Murphy, L. N.: NAO predictability from external forcing in the late 20th century, *npj Clim. Atmos. Sci.* **4**, 1–8 (2021) https://doi.org/10.1038/s41612-021-00177-8

Merryfield, W. J., Baehr, J., Batté, L., *et al.* Current and Emerging Developments in Subseasonal to Decadal Prediction, *Bulletin of the American Meteorological Society* **101**(6), E869-E896 (2020) https://doi.org/10.1175/BAMS-D-19-0037.1

Mishra, N., Prodhomme, C., and Guemas, V. Multi-model skill assessment of seasonal temperature and precipitation forecasts over Europe, *Climate Dynamics* **52**, 4207–4225 (2018) https://doi.org/10.1007/s00382-018-4404-z

Pasternack, A., Bhend, J., Liniger, M.A. *et al.* Parametric decadal climate forecast recalibration (DeFoReSt 1.0), *Geosci. Model Dev.*, **11**, 351–368 (2018). https://doi.org/10.5194/gmd-11-351-2018

Sgubin, G., Swingedouw, D., Drijfhout, S. *et al.* Abrupt cooling over the North Atlantic in modern climate models. *Nat Commun* **8,** 14375 (2017). https://doi.org/10.1038/ncomms14375

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

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

Swingedouw D., *et al.* On the risk of abrupt changes in the North Atlantic subpolar gyre in CMIP6 models. *Annals of the NY Academy of Sciences* (2021) in press.

Torralba, V., Doblas-Reyes F.J., MacLeod D., *et al.* Seasonal climate prediction: A new source of information for the management of wind energy resources. *J. Appl. Meteor. Climatol.*, **56**, 1231–1247 (2017) https://doi.org/10.1175/JAMC-D-16-0204.1.

Verfaillie, D., Doblas-Reyes, F. J., Donat, M. G. *et al*. How reliable are decadal climate predictions of near-surface air temperature? *J. Climate* (2021) doi: https://doi.org/10.1175/JCLI-D-20-0138.1

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences. *International Geophysics Series, Academic Press* **100**, 301-394 (2011) https://doi.org/10.1016/B978-0-12-385022-5.00008-7

Wilks, D. S.: Indices of Rank Histogram Flatness and Their Sampling Properties, Mon. Weather Rev., 147, 763–769, https://doi.org/10.1175/MWR-D-18-0369.1, 2019.

Yeager, S. G., Danabasoglu, G., Rosenbloom, N. A. *et al*. A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model, *Bull. Am. Meteorol. Soc.* **99**, 1867–1886 (2018), https://doi.org/10.1175/BAMS-D-17-0098.1

List of figures

the bottom right panel (in brackets values for above average in the next year).

- **Figure 9:** Examples of abrupt changes found in the subpolar gyre. Time series of Surface Atmospheric Temperature (in °C) averaged over the SPG (70°W-20°W, 45°N-60°N) in annual mean (thin line) and 5-year running mean (thick line). In black is the pre-industrial simulation, in red the historical simulation and in blue the projection considered for a) the CESM2-WACM model with ssp126 scenario, b) the MRI-ESM2-0 model with ssp245 scenario, c) the NorEMS2-LM model with ssp126 scenario. The black arrows represent the approximate starting of the abrupt events.
- **Figure 10:** Link between stratification and centennial temperature trend over the SPG. Scatterplot of the root mean square error of the density in the SPG as compared to observation depicted in Figure 7, for the period 1985-2014, averaged over the first 2000 meters of the ocean vs. the linear trend of Surface Atmospheric Air temperature (in °C/century) computed in different emission scenarios, a) ssp126, b) ssp245, c) ssp585. The letters correspond to the models enumerated in the Method section in their alphabetical order. The red letter corresponds to a model showing no abrupt changes, while the blue letters indicate a model showing an abrupt event for the considered scenario, and it is in light blue when it is not occurring in this particular scenario, but still corresponds to a model that does show abrupt changes for other emission scenarios.
- **Figure 11**. Time-series of AMO. Black: CMIP6 Historical. Yellow: CMIP6 Scenario ssp2-4.5. Cyan: DCPP-A Forecast. Thick curves are ensemble means, thin curves the individual ensemble members. Red: Observations (Green: AMO calculated from JRA-55 reanalysis). All series centred to zero in 1961-2015. Lead-time 10 years for forecasts.
- **Figure 12**. The reliability index (a) and the continuous ranked probability score (b) as a function of lead time for AMO. Thick solid black curve: Historical. Thick solid blue curve: Forecast. For the reliability index the thin full and dashed curves give the mean and plus/minus two standard deviations under the null-hypothesis of a flat histogram. For the CRPS the thin full and dashed curves are mean values for the climatology and plus/minus two standard deviations.
- **Figure 13**. The predictions of the AMO from probability forecasts (cyan) and scenario experiments (black) for the period 2022-2027. For the forecasts we have averaged over all lead times. Left: LENS. Right: CMIP6.
- **Figure 14**. Correlations between ensemble mean NAO and observation in 45 years periods. Plotted as a function of the start year of the 45 years period. Green symbols indicate correlations that are statistically significantly different from zero.
- **Figure 15**: Brier Skill Score (compared to a reference prediction of climatological probabilities, i.e. ⅙ in every year) for the CMIP6-DCPP multi-model ensemble in predicting the probability of a boreal summer (JJA) within the 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 reference datasets) for evaluation period 1979-2014, based on 32 hindcasts s1978-s2009 from eight different models with a total ensemble size of 108 members; hatching marks region where the BSS is not significant ($p>0.01$).