HORIZON 2020

THEME SC5-2017

EUCP
European Climate Prediction system

(Grant Agreement 776613)

European Climate Prediction system (EUCP)

Deliverable D5.1

A report on the relevance of observational and emerging constraints and their reflection in the PDFs

Deliverable Title	*Deliverable 5.1: Relevance of observational and emerging constraints and their reflection in the prediction distribution functions*	
Brief Description	*This deliverable investigates different observational constraints available from WP2 and WP1 work - by this we refer to any use of observations to evaluate, weight, or select models or ensemble members for prediction and projections. It evaluates constraints in out-of-sample tests, evaluates combining constraints, and challenges to the consistency of constraints across the prediction/projection merging boundary*	
WP number	5	
Lead Beneficiary	*Gabi Hegerl (University of Edinburgh, UK)*	
Contributors	*Andrew Ballinger (University of Edinburgh, UK)* *Leonard Borchert, Juliette Mignot, Didier Swingedouw, Giovanni Sgubin (IPSL/CNRS, France)* *Lukas Brunner, Reto Knutti (ETH Zurich, Switzerland)* *Markus Donat, Rashed Mahmood (BSC, Spain)* *James Murphy (Met Office, UK)*	
Creation Date Version Number Version Date	*01/09/2020* *2.0* *12/11/2020*	
Deliverable Due Date Actual Delivery Date	*30/11/2020* *28/11/2020*	
Nature of the Deliverable	x	*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**
1.0	30/10/2020	co-authors	for EUCP internal review
2.0	27/11/2020	co-authors	incorporated review comments

Table of contents

1. Executive summary

The goal of this deliverable is to assess the usefulness of observational constraints on predictions and projections in EUCP. Accounting for both is key to the longer-term objective of creating a meaningful merged prediction/projection system that has skill on both short and longer timescales. The team has addressed this by analyzing constraints from observations on climate projection ensembles on the one hand and forecast quality measures across prediction and projection methods on the other hand. We present new work that explores combinations of constraints between different methods with the aim to arrive at a more robust use of observational constraints. The team has further explored objectively assessing the value and reliability of observational constraints by using imperfect model studies where the projected future response of a withheld model is predicted based on its historical performance. Observational constraints on decadal prediction skill scores are used to infer information on physical mechanisms that drive decadal climate variability in space, time, and the ensemble dimension. This activity sets the scene for future merging attempts by breaking ground on unifying methods and vocabulary between the projection and prediction communities.

For projection systems, this deliverable explores the information added from different observational constraints, building on a number of existing methodologies used in D2.1/D2.2. We focus on observational constraints using performance weights obtained from the Climate model Weighting by Independence and Performance (ClimWIP) method, and on an estimate of the magnitude of the forced signal in observations to use in forward projections (ASK). Previous work has been extended to more seasons and variables, and to CMIP6 simulations. We explore using successive constraints by combining constraint information from two different WP2 approaches (ClimWIP and ASK). This provides new insights into the value of combining information from different observations. The team has evaluated if sequential application may overcome limitations in signal-to-noise ratio of the ASK method or in reach of projections timescales for ClimWIP. These first results are encouraging for summer, yet there is some seasonal variability in the results that requires further exploration.

Using UKCP18 methodology, we further demonstrate that sequential application of a number of different observations combined to narrow the projection uncertainty, but the addition of successive observations (after the first couple) has a diminishing impact on the constrained range.

To assess the value in combining observations in projections, the team conducted an out-of-sample test using CMIP5 models as pseudo observations and evaluated RMSE, spread and reliability (percent time prediction was within error bars) using a constrained projection system based on CMIP6 simulations. Results of this perfect/imperfect model test suggest that observational constraints improve the root mean square error (RMSE) of projections and predictions compared to unconstrained CMIP6 simulations most clearly in summer, and over central and Northern Europe regionally. Results in other seasons are more mixed, with the RMSE generally better for ClimWIP in spring and winter. This might be in part due to confounding strong winter variability and resulting difficulty to separate the response to different external forcings in ASK. Some of the observed change, for example, is explained by North Atlantic Oscillation (NAO) tendencies, which we plan to further evaluate. In some seasons and cases, there are strong improvements in combining both strands of observational constraints, and first pilot studies for precipitation are promising.

In summary, observational constraints on projections show promise, both individually and in combination, but need to be evaluated prior to use for the target of projection, for example, using (im)perfect model studies evaluating the reliability, error reduction and spread of the constraint for the target of projection.

For initialized predictions, skill scores and model performance indicators have been used to evaluate which models perform well and what might contribute to this performance. Skill for both subpolar gyre sea surface temperatures (SSTs) and surface temperature over Europe varies over forecast time between models, and some time periods in the hindcast period show consistently higher skill than others. This suggests that skill estimates vary with natural and internal variability and climatic boundary conditions. Understanding how these modes of variability and climatic boundary conditions affect the skill of decadal prediction systems can enable a cleaner estimation of the point in time at which prediction and projection simulations will have to be merged to produce seamless prediction. Additionally, we here tested for the first time the use of SREX regions as focus regions for decadal temperature predictions over Europe, as are commonly used in analysis of climate projections, setting the scene for possible future merging activities.

Furthermore, constraints based on reproduction of the observed evolution of natural variability, using global SST patterns were explored to understand the effects of aligning climate variability when constraining climate projections. This work proved to be a good testbed for subsequent work towards merging initialised predictions and projections in a large ensemble context, and we recommend exploring this framework further.

Overall, a common performance-based model metric or constraint from the observed signal that relates clearly to either and even more both projection and prediction skill across different climate models is yet to be found. This means that when predictions and uninitialized projections are merged in task 5.2, multiple observational constraints will need to be carefully managed across the merged product and applied on a case-by-case basis dependent on the target of prediction, projection, or the merged product. Objective comparisons of different methodologies, and hybrid approaches between them that have been pursued in this deliverable, go beyond what has previously been done, and suggest pathways towards more robust climate projection methodologies.

The results shown here will be valuable towards the Task 5.3 goals, namely to evaluate to what extent the constraints on decadal subpolar gyre predictability and subsurface stratification are reflected in the merged seamless information; to assess the risk of abrupt (<10 years) subpolar gyre change in the seamless information; and to evaluate to what extent the constraints on seasonal-mean and extreme temperature and seasonal precipitation over large European regions are reflected in the merged seamless information. These full objectives will be targeted once merged pilot predictions are available. Towards these further goals, we will extend our analysis to temperature and precipitation extremes on monthly timescales, and if time allows, shorter timescales.

2. Project objectives

These deliverables have contributed to the following EUCP 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

3.1 Introduction

This deliverable aims to exemplify WP5's brief of pulling together threads from different EUCP work packages (typically WP1 and WP2) that use observations to weight (or select) simulations from various climate models. It assesses the value of observational constraints and skill scores in both prediction and projection systems, as accounting for both is key to creating a skilful merged prediction/ projection system in the future that has skill on both short and longer timescales. Specifically, observational constraints used here are based on metrics involving observed mean climate and climate evolution, including skill scores based on hindcasts of observations and metrics focusing on mechanisms of predictability. This deliverable reports on the crosscutting relevance of observational constraints and skill derived by comparing observations to modelling systems. It reflects on the consistency of observational constraints across prediction/projection timescales and approaches, and pilots opportunities for building upon multiple constraints. Its ultimate goal is to inform the use of observational constraints in the merged prediction and projections products on timescales from one to about 50 years originating from WP5 and EUCP.

Due to its relatively early timing in WP5, this deliverable is a discussion document introducing different constraints used across the project, and their level of consistency with each other, rather than seeking to produce a final conclusion. This deliverable also reflects on how observational constraints may be used in uncertainty characterization across EUCP, and the challenges in keeping them consistent across prediction timelines if using different observational constraints.

Focus regions considered in this deliverable are similar to those used in Brunner et al. (2020a), namely the SREX regions Northern Europe (NEU), Central Europe (CEU) and Mediterranean (MED), and the main focus is on predicting change in the spatial averages of temperature and precipitation in those regions.

The focus of this deliverable is on constraints and merging in time. However, observational constraints may also inform on the realism of differences in projections and predictions between high resolutions, particularly convection resolution, climate models and standard projection tools. An example might be changes in summer rainfall where results may vary between modelling tools. This latter point is added to recommendations but not further discussed here, as this work has not yet been done, but could be considered in the future.

This deliverable aims to:

- Compare and cross analyze the use of observational constraints and model selections across predictions and projections and determine if multiple observational constraints may further improve skill in predictions and projections.
- Determine consistency of observational constraints across lead times and from predictions to projections.
- Develop suggestions how to reflect observational constraints in merged prediction/projection products.

In order to compare observational constraints across a consistent framework, the study sought to target a coherent selection of climate model projections from the Coupled Model Intercomparison Project Phase 6 (CMIP6; Eyring et al., 2016), initialized predictions from the Decadal Climate Prediction Project (DCPP; Boer et al., 2016), and individually forced historical simulations from the Detection and Attribution Model Intercomparison Project (DAMIP; Gillet et al., 2016). The overlap between all three types of simulations, however, was not large enough to evaluate the different constraints, particularly when accounting for the use of different model versions between DCPP and DAMIP in several model systems (table 3). As a result, slightly inconsistent sets of simulations have been used between predictions, projections and DAMIP analyses, drawing from the largest possible subset of CMIP6 model output for each application. If a larger family of models were available than used here (see Table 1), the closeness of different models to each other would also need to be considered (Knutti et al., 2013).

3.2 Building upon WP2 constraints presented in deliverable D2.1/2.2 (ETH)

The work in WP2 is closely related to the WP5 efforts and therefore some relevant results are described here. One of the main objectives of WP2 is the production of future climate projections and related uncertainties on time-scales beyond 10 years and thus it provides the long-term part for the merging of climate projections and predictions discussed in this deliverable. The most relevant WP2 tasks in this context include (i) assessing the relevance of constraints for future climate projections, (ii) comparing and discussing the results of different constraining methods, and (iii) investigating the implications of these constraints for European climate.

These tasks are still ongoing but first results were presented in a joined deliverable (D2.1 and D2.2) submitted in December 2019 and summarized here. The core part of this deliverable was a scientific paper in *Journal of Climate*, which was published recently (Brunner et al. 2020a). This paper constitutes a comparison of eight different methods used to constrain future European climate projections using a consistent framework. In addition, reasons for agreements and disagreements across these methods are discussed, with direct implications for the work in WP5.

The results in Brunner et al. (2020a) focus on temperature and precipitation changes between 1995-2014 and 2041-2060 under RCP8.5 (i.e., using CMIP5). A common framework was developed to allow a consistent comparison between the different methods, including a set of European sub-regions. This proved to be an essential effort as the lack of coordination across methods in the presentation of results was identified as a main obstacle for comparison: “a consistent comparison may be hindered by subtle differences in [the methods] setup such as domain and grid resolution, season and time period, models and ensemble members included, or reported results (such as mean versus median or standard deviation versus percentile range). In such cases the results may diverge not only due to assumptions and characteristics inherent to the methods but also due to these differing setups” (Brunner et al. 2020a).

Figure 1 shows some results of the comparison, for a detailed discussion of the results and the underlying methods we refer to Brunner et al. (2020a). While all methods clearly show the anthropogenic warming signal, the comparison reveals different levels of agreement based on the region considered and the metric of interest (e.g., median versus 80% range). In general, methods

more often agree in the central estimate while uncertainty ranges can be fairly different, particularly for the more extreme percentiles (see, e.g., figure 1d). However, for some regions, the median values can also differ across methods, and in isolated cases, methods also disagree on the direction of the shift from the unconstrained distributions. Methods also constrain to different extents, with some methods leading to a stronger reduction in spread while others hardly reduce it. This can be due to using observations more or less completely and efficiently, but can also reflect neglected structural uncertainties (such as model error).

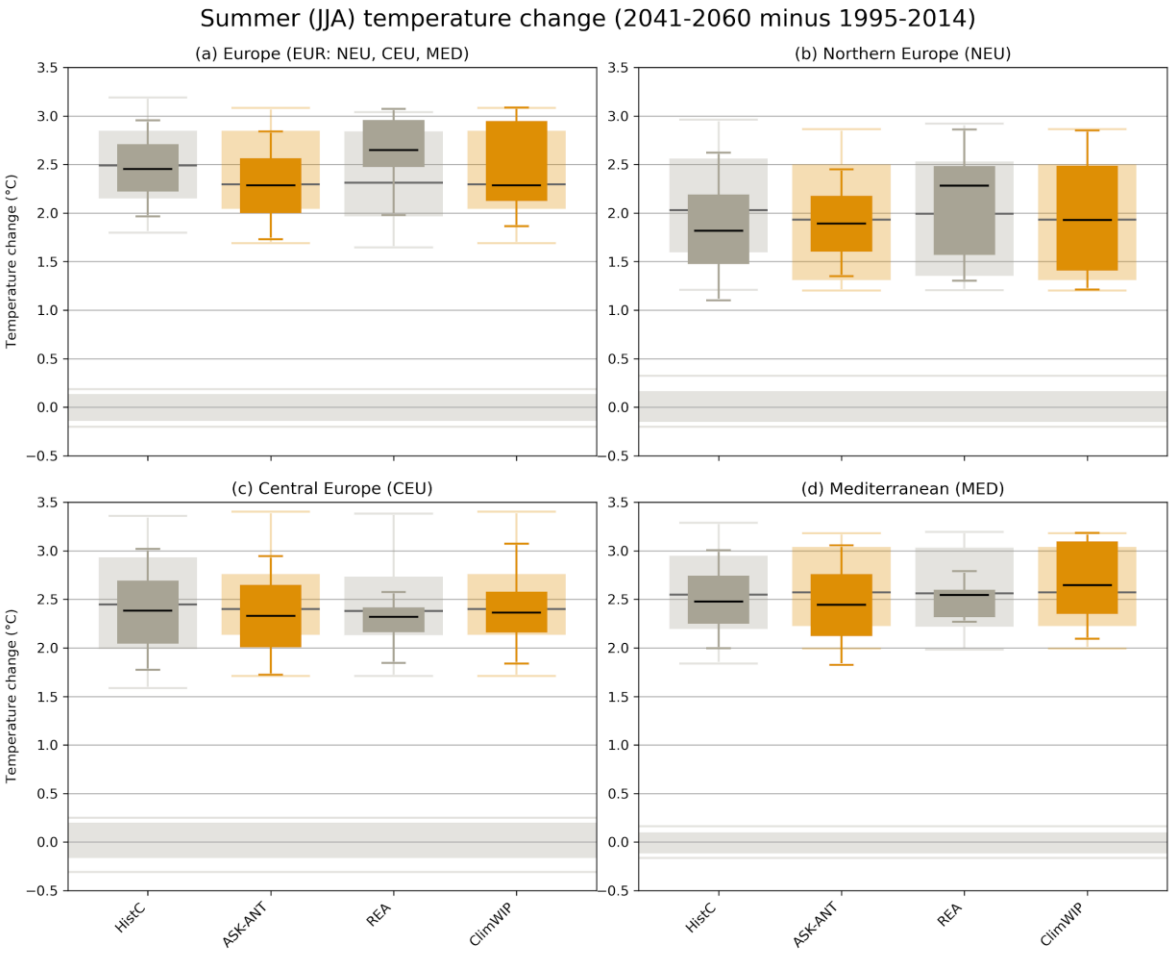

Figure 1: Summer (July–August) temperature change 2041–60 relative to 1995–2014 for (a) the combined European region as well as (b)–(d) the three European SREX regions (Northern Europe, Central Europe, and the Mediterranean) using RCP8.5. The lighter boxes give the unconstrained distributions originating from model simulations; the darker boxes give the observationally constrained distributions. Shown are median, 50% and 80% range. The horizontal light grey box and lines centered on zero in the diagram show the same percentiles of 20-yr internal variability based on observations. Methods used in this deliverable are coloured, additional methods are in grey. Adapted from figure 6 in Brunner et al. (2020a).

Given the diversity of methods investigated in WP2, some differences are not unexpected and Brunner et al. (2020a) dive into several avenues exploring why the methods might produce different projection ranges:

1. Differences in underpinning assumptions
2. Differences in uncertainties accounted for
3. Differences in the application of constraints
4. Physical and spatial consistency of characteristics of the outputs

This discussion of the source of differences will, at least in part, be relevant also for discussing predictions and their merge with projections in WP5, for example by helping to avoid (possibly hidden) discontinuities between the prediction and projections time horizons through the use of appropriate constraining methods. Finally, Brunner et al. (2020a) identify ways forward, which, again, have direct relevance for the existing and future work of WP5 as well as for interpreting the results. In summary, several avenues can be explored to make best use of the created uncertainty ranges (from projections as well as from predictions):

1. Considering the decision context in which the PDFs are used
2. Using cases where all (most) methods agree
3. Combining methods either before or after the PDF calculation
4. Selecting reliable and skillful methods based on a consistent test score, for example, in 'perfect' or 'imperfect' model tests.

An important lesson from Brunner et al. (2020a) when using constraints is that the extent to which internal climate variability (shown in figure 1 for 20-year mean values of temperature) is considered varies between approaches. Therefore, when merging initialized and non-initialized predictions over time, it needs to be clear what the role of internal variability is in the merged product, and ensure that there is a smooth transition for both the underlying forced signal at the merge point (which is constrained by the approaches shown above), and natural variability. Internal variability as well as the forced component are the target of prediction, yet at the merge point, the point where the importance of internal variability diminishes to insignificance, it needs to smoothly transition to the full variability shown in the non-initialized simulations. Methods that do not include an envelope of natural variability but constrain the signal only (such as ASK) need to add this variability later.

3.3. Lessons from the UKCP18 observational constraint (UK Met Office)

The UK's national scenarios, UKCP18 include sets of 28 global and 12 regional projections from alternative climate models, including perturbed parameter ensembles (PPEs) derived from the latest UK climate model, and (in the case of the global projections) CMIP5. Also included is a set of probabilistic climate projections, derived from a larger set of 360 model simulations, again based on a combination of PPE and CMIP5 runs. These have been combined to make the pdfs, using a Bayesian framework that includes the formal application of several observational constraints. These consist of climatological spatial fields for a set of standard climate model evaluation variables, plus changes observed during the 20th century in carbon dioxide concentration, upper ocean heat content and global patterns of surface temperature. The UKCP18 method was one of the approaches covered in Brunner et al. (2020a). Here, we provide examples of the impact of these constraints, on the uncertainty ranges that emerge from the pdfs. The method is detailed in Murphy et al., 2018; based on Murphy et al., 2009, Sexton et al., 2012; Booth et al., 2017 and Harris et al., 2013.

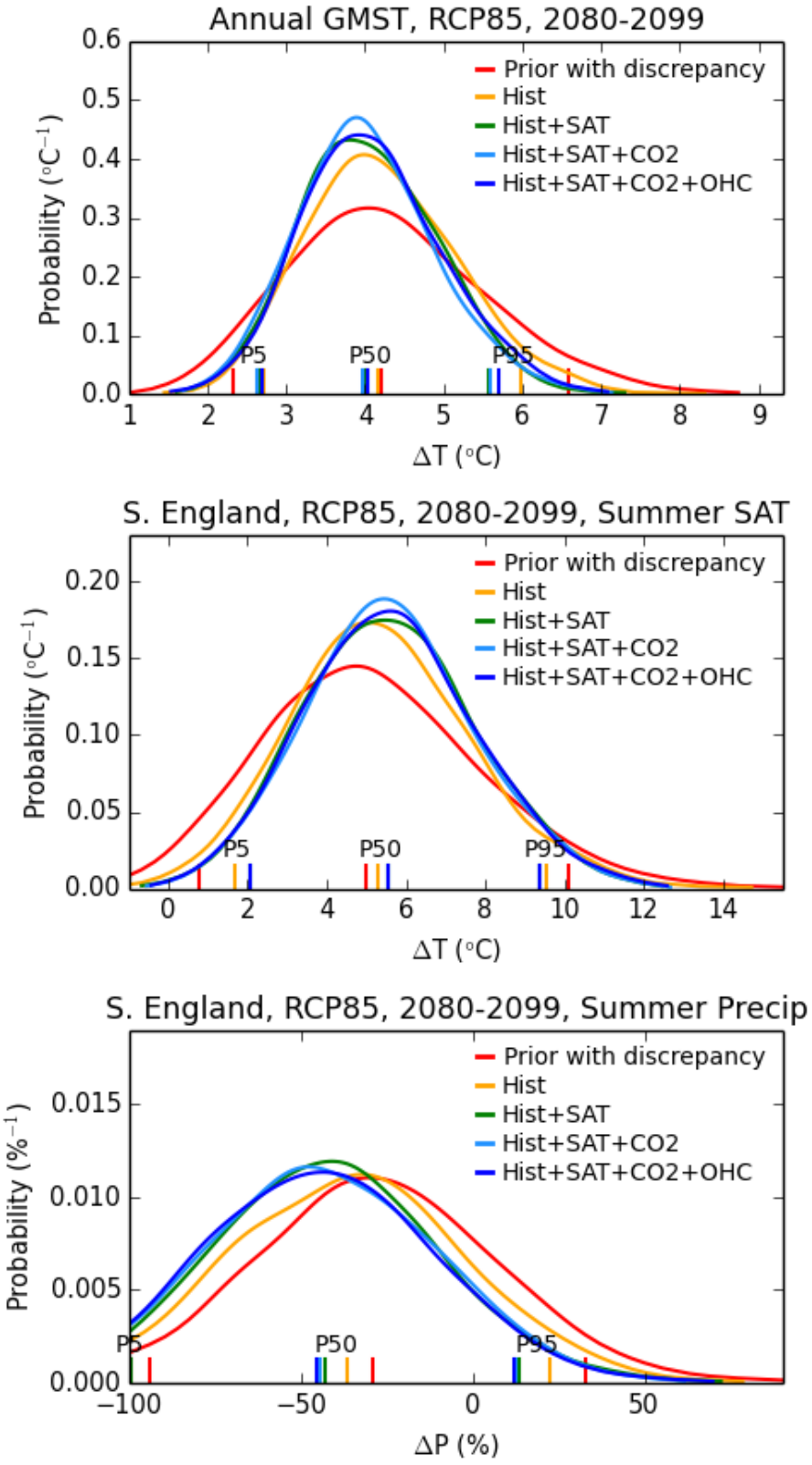

Figure 2: *Examples of the impact of constraints derived from historical climatology (Hist), historical global surface air temperature trends (Hist+SAT), historical trends in atmospheric CO_2 concentration (Hist+SAT+CO2), and upper ocean heat content (Hist+SAT+CO2+OHC), in modifying the prior distribution to form the posterior. The 5th, 50th (median) and 95th percentiles are plotted, along with the pdfs.*

Figure 2 illustrates the impact of observational constraints on UKCP18 results for global mean temperature, and summer temperature and precipitation for Southern England; 2080s relative to 1981-2000, under RCP8.5. The red curves show the prior distributions. The blue curves show the posterior distributions, in which all constraints are applied simultaneously, accounting for relationships between them. The other curves show sensitivity tests in which individual constraints are applied successively, in order to assess their influence. Results show that as well as narrowing the range, specific constraints can also alter the relative likelihood assigned to different parts of the distribution. As an example, the chance of a summer drying is upweighted in the posterior, compared with the prior distribution.

Experiences with use of observational constraints in UKCP18 illustrate that:

- Combining multiple constraints provides more powerful estimates of the relative likelihood for alternative future outcomes, since each individual constraint contributes new information that is partially independent.
- When considering the effects of individual constraints in our sensitivity tests, the first two constraints applied (climatology and historical temperature changes) have the largest impact.
- A caveat is that the diagnosed impact of specific constraints can depend on the order in which they are applied. Here, for, example, the effects of historic changes in carbon dioxide concentration and ocean heat content are correlated with the impact of past changes in surface temperature, and appear small because they are applied after the surface temperature constraint (see figure 2). If the order was reversed, the carbon dioxide and heat content constraints might appear larger.
- There is plenty of scope to refine such constraint methods in the future. For example, metrics of climate variability are not yet considered in the set of historical climatology constraints.

3.4 Pathways to selecting/combining observational constraints for non-initialised projections (UEdin with ETH and IPSL)

We have piloted an example of combining different observational constraint methods, in order to test a potential approach that might harness the various strengths of each paradigm, and as an opportunity to discuss the challenges and limitations involved in such an endeavour.

3.4.1 The ClimWIP approach: model weighting based on historical performance

The latest version of the Climate model weighting by Independence and Performance (ClimWIP) method used here is described in Brunner et al. (2020b, in press) and is based on earlier work by Brunner et al. (2019), Lorenz et al. (2018), and Knutti et al. (2017). ClimWIP applies weights to an ensemble of climate models based on their historical performance in a range of diagnostics and also provides the possibility of accounting for model inter-dependence.

In this first pilot study, we only use the performance weighting part of ClimWIP and disregard the dependence weights. Performance is based on each model's generalised distance to observations

(ERA5 and MERRA2) in 5 diagnostics evaluated from 1980 to 2014: global, spatially resolved fields of climatology and variability of near-surface air temperature and sea level pressure as well as global, spatially resolved fields of near-surface air temperature trend (Brunner et al 2020b). It is important to note that the weights used here are optimised to constrain global mean temperature change in the second half of the 21st century for the full CMIP6 ensemble. However, to test the general applicability of our approach we apply them to a subset of 9 models and regional temperature change for which DAMIP simulations (Gillett et al., 2016) are available. Hence, we stress that in a more thorough study the calibration parameters of ClimWIP as well as the diagnostics should be selected based on the target region and variable. This could, for example, include physical understanding of the most important processes relevant for the target as well as statistical relationships between the diagnostics and the target.

We use two different combinations of diagnostics to calculate the performance weights: the optimised combination described in Brunner et al. (2020b) including temperature trends, and a version without temperature trend in order to avoid accounting for it twice when applying constraints subsequently, as it is also used by ASK (see table 1), and the ASK method is strongly driven by temperature trends. Note that this modification of ClimWIP will most likely reduce its performance as a constraint.

The distribution of weights assigned to the CMIP6 ensemble when using all five diagnostics, and when not using the temperature trend, is shown in figure 3 for independence, performance, and the combined case. Figure 4 shows the correlation between both cases for the subset of models that provide DAMIP runs. Both figures highlight the crucial role trend plays for some models (such as HadGEM3) with their weights changing considerably between the cases with and without trend information included in the weights. As discussed in Brunner et al. (2020b) this is attributable to a large degree to several models which show strong warming (orange labels in figure 3) receiving low weights as their trend differs from the observations.

Table 1: *List of the CMIP6 models used in the ASK-ClimWIP constraining intercomparison pilot study.*

CMIP6 Model Name	*Number of ensemble members included*	**ClimWIP weighting** *(no trend information)*	**ClimWIP weighting** *(with trend information)*
ACCESS-ESM1-5	3	0.1627	0.1381
BCC-CSM2-MR	1	0.0132	0.0792
CNRM-CM6-1	5	0.0772	0.0762
CanESM5	10	0.0216	0.0049
GFDL-ESM4	1	0.4051	0.5047
HadGEM3-GC31-LL	4	0.2582	0.0070
IPSL-CM6A-LR	5	0.0313	0.0639
MIROC6	3	0.0052	0.0627
MRI-ESM2-0	1	0.0256	0.0633
Total	33	1.0	1.0

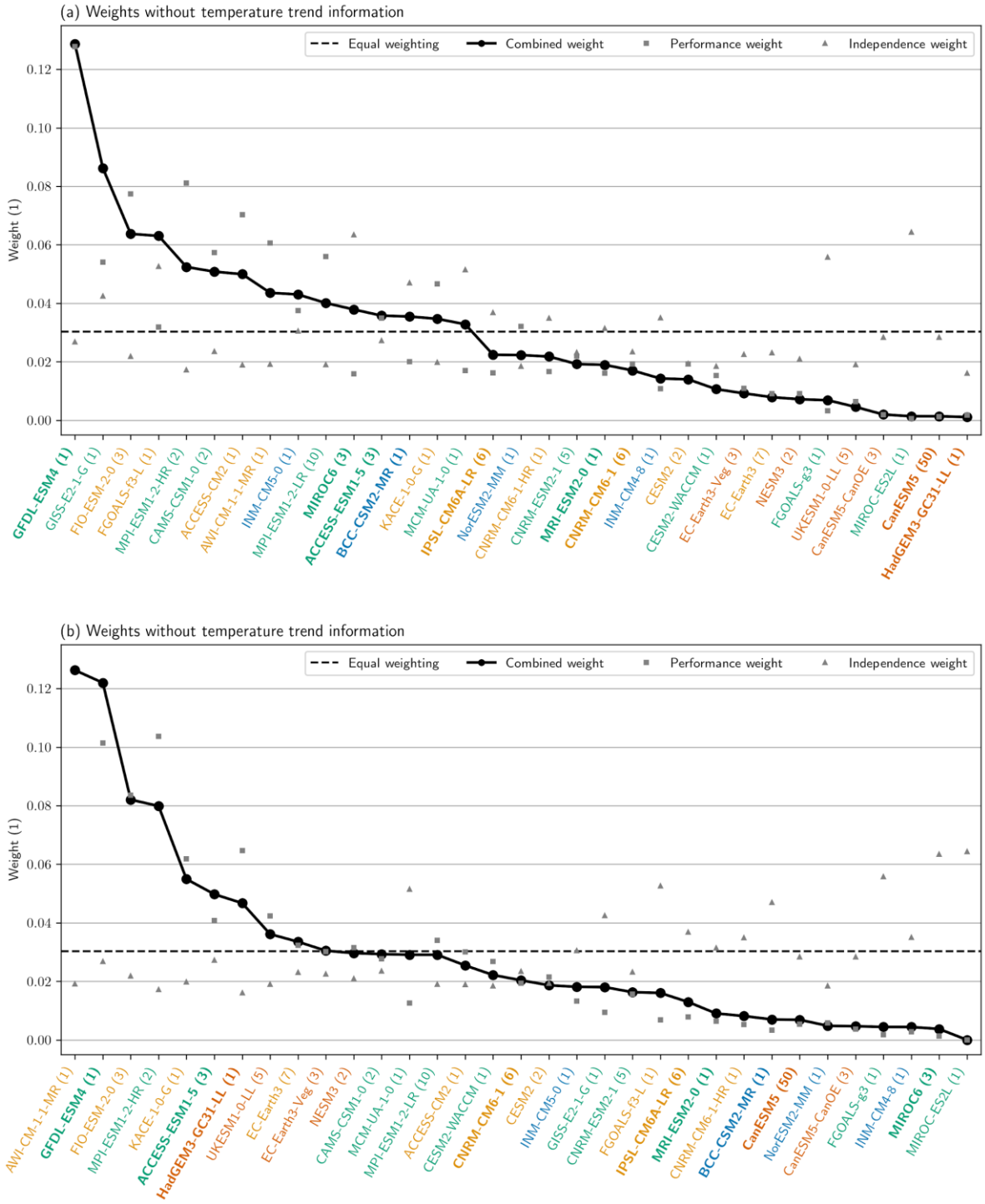

Figure 3: *Combined independence-performance weights for each CMIP6 model (line with dots) as well as pure performance weights (squares) and pure independence weights (triangles). All three cases are individually normalised and the equal weighting each model would receive in a normal arithmetic mean is shown for reference (dashed line). Labels of the models with DAMIP runs (used in this study, see Table 1) are highlighted in bold font. Labels are coloured by each model's Transient Climate Response value: > 2.5 °C - red, > 2 °C - yellow, > 1.5 °C - green, and ≤ 1.5 °C - blue. The number of ensemble members per model is shown in brackets after the model name. (a) Weights including temperature trend information and (b) weights without temperature trend information. Adapted from Brunner et al. (2020b, revised).*

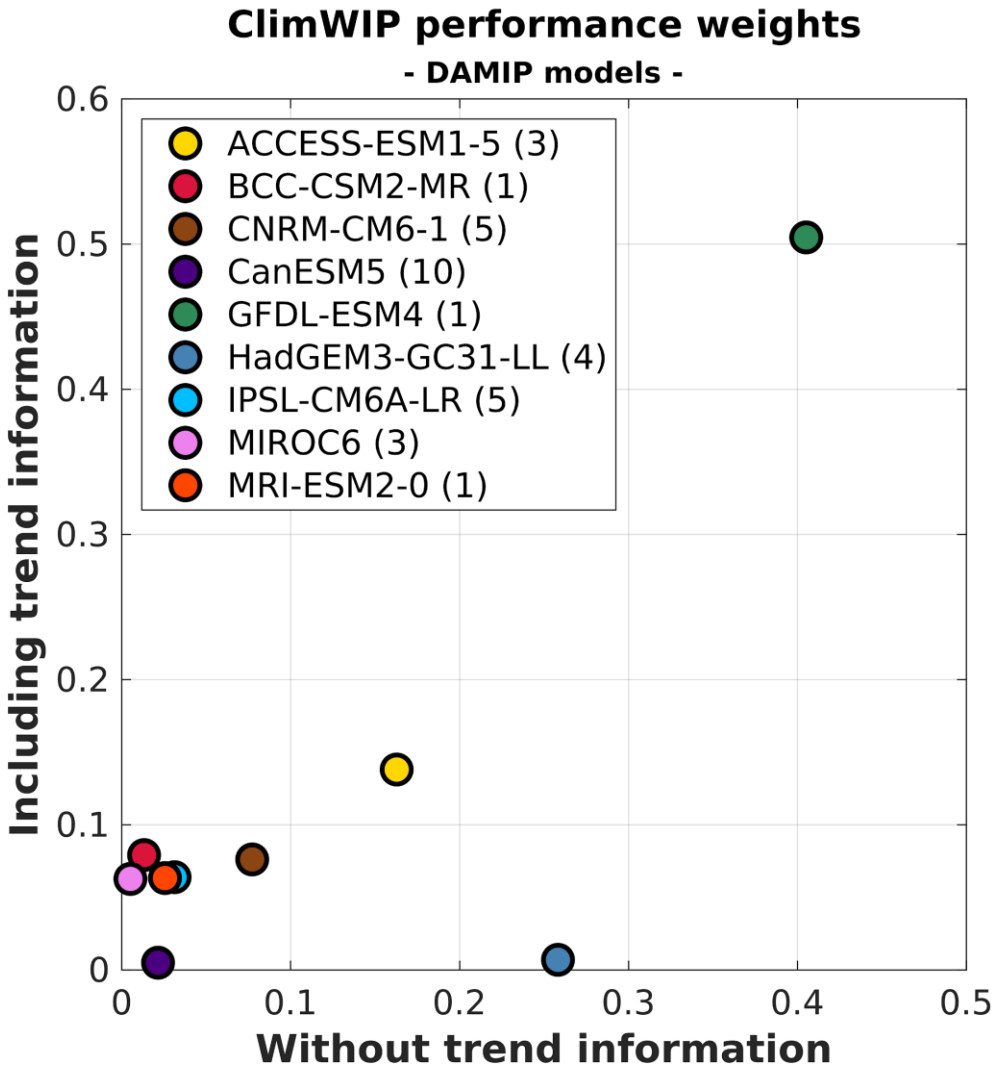

Figure 4: *The relationship between the model-specific relative weights assigned by ClimWIP, computed without using temperature trend information (along the x-axis), and including a temperature trend metric (along the y-axis). The values are shown in Table*

3.4.2. The ASK approach: constraint based on estimation of the forced signal

The ASK method (Allen et al., 2000; Stott & Kettleborough, 2002; Kettleborough et al., 2007, Shiogama et al., 2016) utilises regression-based detection and attribution techniques to derive an observational constraint on the best estimate and uncertainty range of future climate projections. The underlying approach seeks to detect a signal in observations that can be related to the climate model's forced response, thus the ASK method is most readily applicable over regions and seasons where the spatio-temporal patterns have a relatively high signal-to-noise ratio. Given estimates of the noise (internal variability of the model), the method computes a range of factors by which the simulated model response can be scaled and remain consistent with the observed signal. This set of scaling factors can then be applied as a scaling to future climate projections, providing an observational constraint based on the detection of the forced response in the historical simulations.

ASK Background

The method assumes that the true observed climate response, y_{obs}, to historical forcing is a simple linear combination of one or more (*n)* individual forcing fingerprints, X_j. These are usually obtained by averaging across the multimodel ensemble, or in some cases shown below, by applying a weighted average across results from individual models. Fingerprints are scaled by their respective scaling factors, β_j, accounting for noise in both the observations, ε_{obs}, and in the modelled response to each of the forcings, ε_j, as expressed:

$$y_{obs} = \sum_{j=1}^{n} \beta_j (X_j - \varepsilon_j) + \varepsilon_{obs} \qquad (1)$$

In this study we explore four different linear combinations (Eq. 1) of model fingerprints, as follows:

a) $y_{obs} = \beta_{a1}(X_{Historical} - \varepsilon_{Historical}) + \varepsilon_{obs}$, (2)
where $\widehat{\beta}_{ALL} = \beta_{a1}$.

b) $y_{obs} = \beta_{b1}(X_{Historical} - \varepsilon_{Historical}) + \beta_{b2}(X_{GHG} - \varepsilon_{GHG}) + \varepsilon_{obs}$,
where $\widehat{\beta}_{GHG} = \beta_{b1} + \beta_{b2}$; $\widehat{\beta}_{OTH} = \beta_{b1}$.

c) $y_{obs} = \beta_{c1}(X_{Historical} - \varepsilon_{Historical}) + \beta_{c2}(X_{Nat} - \varepsilon_{Nat}) + \varepsilon_{obs}$,
where $\widehat{\beta}_{ANT} = \beta_{c1}$; $\widehat{\beta}_{NAT} = \beta_{c1} + \beta_{c2}$.

d) $y_{obs} = \beta_{d1}(X_{Historical} - \varepsilon_{Historical}) + \beta_{d2}(X_{GHG} - \varepsilon_{GHG}) + \beta_{d3}(X_{Nat} - \varepsilon_{Nat}) + \varepsilon_{obs}$,
where $\widehat{\beta}_{GHG} = \beta_{d1} + \beta_{d2}$; $\widehat{\beta}_{NAT} = \beta_{d1} + \beta_{d3}$; $\widehat{\beta}_{OTH} = \beta_{d1}$.

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. 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).

Two different approaches were used for estimating the confidence intervals on scaling factors:

1. Noise sampling: Confidence intervals are estimated by adding samples from the piControl simulations (of the same length) to the noise-reduced fingerprints and observations, and recomputing the TLS regression (10,000 times) in order to build a distribution of scaling factors, from which the 5th-95th percentile range can be computed.
2. Bootstrapping: Following DelSole et al. (2019), new fingerprints are determined by randomly sampling, with replacement, corresponding values from the observations and model fingerprints to form new arrays the same length as the original. A scaling factor is calculated by regressing the resampled model fingerprint(s) onto the resampled observations, and the process is repeated (10,000 times) in order to build a distribution from which the 5th-95th percentile range can be computed.

For the results that follow, we have chosen to display confidence intervals from the noise sampling approach. The bootstrap confidence intervals have also been checked, and while there are slight differences in the spread, the two measures generally provide consistent and robust agreement.

Datasets used in the study

Observations come from the gridded E-OBS v19.0e dataset (Haylock et al., 2008), with monthly values computed from the daily data. The study uses CMIP6 model simulations (Eyring et al., 2016) run with historical forcings, and Detection and Attribution MIP (DAMIP) single-forcing simulations (Gillett et al., 2016) over the same period. For the future projections, historical simulations are extended with CMIP6 Scenario-MIP Shared Socioeconomic Pathway (SSP) (Gidden et al., 2019) simulations. This analysis uses the set of 9 models with 33 total ensemble members (Table 1), that were readily available in CEDA (retrieved through JASMIN in September 2020), common to the required set of simulations.

Note that a smaller number of DAMIP single-forcing simulations are available compared to the Scenario-MIP runs, which limits (the ASK approach) from being able to be applied (as a constraint) to the full set of CMIP6 future projections. Herein the 'unconstrained' or raw spread of future CMIP6 models displayed for comparison to the constrained distributions, is the common limited set of models (Table 1) from which the historical detection and attribution has been performed.

The monthly surface air temperature fields from the observations and each of the CMIP6 model ensemble members were spatially regridded to a regular 2.5° × 2.5° latitude-longitude grid, with only the gridboxes over land (with no missing data throughout time) being retained in the analysis. The resulting masked fields (from observations and all individual model ensemble members) were spatially averaged over a European domain (EUR) and three sub-domains (NEU, CEU and MED; as described in Brunner et al., 2020a).

3.4.3 Comparing and combining the ClimWIP and ASK constraints

As an initial step for this pilot study, we here applied a very simple and straightforward approach to combining the observational constraints from ASK and ClimWIP. We explored the use of model performance weighting (3.4.1) in constructing each of the multi-model mean fingerprints that are subsequently used in the derivation of the detection and attribution constraint (3.4.2).

Thus, two sets of multi-model mean fingerprints are computed in each instance (for each region, season, etc.). Firstly, an equal-weighted set of multi-model fingerprints, reflecting the standard ASK approach used heretofore, and for comparison, a second set of multi-model fingerprints computed using the ClimWIP performance weighting. When combining the constraints in this way, we choose to restrict ourselves to the ClimWIP performance weights that were derived without temperature trend information (recall 3.4.1), given this would otherwise incorporate a measure of the model's climate sensitivity into the weighting, which might impact the justification for applying subsequent detection and attribution techniques.

Annual surface temperature anomalies from 1950 to 2014 are displayed in Figure 5 with the upper left panel showing the equal-weighted time series, and the lower left panel showing the time series after applying the ClimWIP weights (without trend). The same observed annual time series (E-OBS, black line) has been plotted in each panel, along with the CMIP6 multi-model mean (of ensemble means) of the all-forcing historical simulations (brown line), the greenhouse gas single-forcing historical simulations (red line), and the natural single-forcing historical simulations (green line). A measure of the internal variability of the CMIP6 models is estimated by averaging the standard

deviation (65-yr samples) of the associated piControl simulations, and is indicated by the background-shaded region.

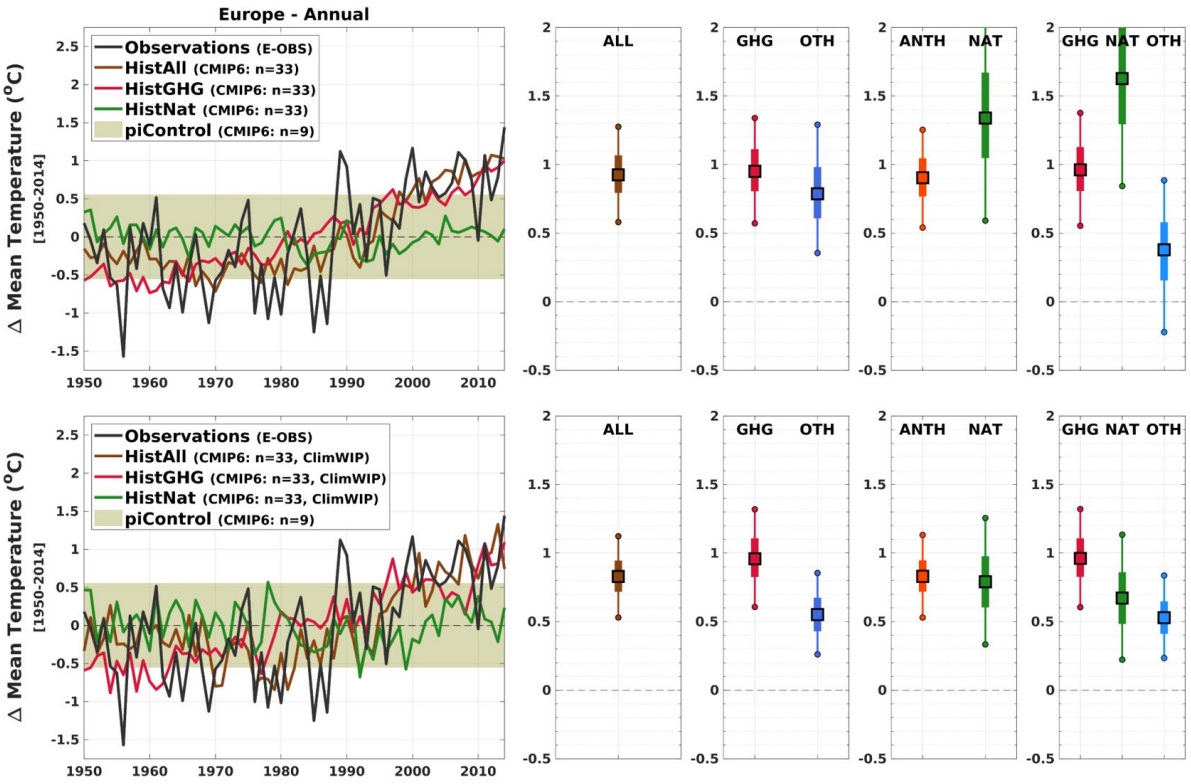

Figure 5: *Annual time series (left panels) of European surface air temperature anomalies (relative to 1950-2014) from observations (E-OBS v19, black line) and CMIP6 historical simulations (all forcings, brown line; GHG-only forcing, red line; and NAT-only forcing, green line), displaying the multi-model mean of ensemble means (9 models, 33 total ensemble members), with the shaded region denoting the multi-model mean variability (±1 standard deviation) of the associated piControl simulations. The scaling factors (right panels) are derived from TLS regressions of the CMIP6 model fingerprints and the observations and indicate to what extent the multi-model mean fingerprint needs to be scaled to best match observations (centre) and can be scaled to still be consistent with observations (5-95% range). Results show the 1-signal (ALL), 2-signal (GHG & OTH; ANTH & NAT), and 3-signal (GHG, NAT & OTH) scaling factors. The observations and model fingerprints are the conjoined annual time series of three spatially-averaged regions: NEU, CEU, and MED. The upper panels display the time-series fingerprints and associated scaling factors using an equal-weighted multi-model mean; the lower panels use ClimWIP performance (without trend information) weighting to weight the multi-model mean fingerprints for deriving the scaling factors. Confidence intervals show the 5th-95th (thin bars) and 25th-75th (thicker bars) percentile ranges of the resulting scaling factors.*

The scaling factors were derived through a total least squares regression of the multi-model mean fingerprints onto the observations. The scaling factors that are shown in Figure 5 were derived using fingerprints comprising the conjoined annual time series of the three spatially-averaged European subregions (NEU, CEU, and MED; 3 x 65 = 195 years). The analysis was also performed using a single European average fingerprint, and separate single-subregion fingerprints (NEU/CEU/MED). As expected, the 3-region fingerprint generally provides a tighter constraint on the scaling factor

because of the additional (spatial) information included in the fingerprint that strengthens the signal to noise. However, the subsequent qualitative differences in the derived scaling factors (to be shown in the results that follow), when comparing the impact of using an equal-weighted or ClimWIP-weighted fingerprint, are generally found to be robust irrespective of the particular fingerprint formulation.

Multiple total least square regressions have been performed with the different model fingerprints, including a 1-signal all-forced (ALL), two different 2-signal approaches using the greenhouse gas (GHG), other anthropogenic (OTH), combined anthropogenic (ANTH), and natural (NAT) forcings (GHG & OTH; ANTH & NAT), and a 3-signal (GHG, NAT & OTH) detection and attribution analysis, as expressed in Eq. 2(a-d). The resulting sets of scaling factors (for annual European temperature projections) are shown in the right-hand panels of Figure 5, with the upper and lower panels showing the scaling factors following regressions using the equal-weighted and ClimWIP-weighted model fingerprints, respectively.

When comparing between the different constraints, one notes an overall narrowing of the uncertainty range in the scaling factors (providing a slightly tighter constraint) when using the ClimWIP-weighted model fingerprints, particularly the natural scaling factor. The best-estimate magnitudes of the leading signal (ALL, GHG, ANTH) scaling factors also remain robust. Results suggest that the response to aerosols is larger (than in the observations) in both the weighted and unweighted cases. This provides motivation for using the ASK method with the single forced future simulations, as it suggests that methods which do not account for the disparity between the greenhouse gas and aerosol forcings could lead to biased projections. Overall, the sensitivity of the estimated amplitude of natural and aerosol response probably reflects model differences in emphasis between ClimWIP weighted and unweighted cases.

To explore differences between seasons, Figure 6 shows the summer and winter (upper two and lower two panels, respectively) time series of European temperature anomalies, along with the associated scaling factors; once again (as in Figure 5) showing the results of using either equal-weighted multi-model mean fingerprints or ClimWIP-weighted fingerprints (the upper and lower panels of each season, respectively). Note that the combination of results yields stronger constraints when separating out the greenhouse gas signal (which is particularly useful for constraints) in particular, as the contribution by natural forcing, other anthropogenic and greenhouse gas forcing to winter temperature change is far more separable (given noise uncertainty) in the ClimWIP constrained case. Results also appear to be less degenerate (no longer getting the very large error bars in the other and natural signals).

The different constraints are summarised in Figure 7, where they have been applied to the CMIP6 future projections for the mid-century (2041-2060), displayed here as the annual and seasonal European surface air temperature difference (°C) relative to a recent reference period (1995-2014). The black bars show the unconstrained CMIP6 distribution (5th-95th percentile spread by the thin lines, 25th-75th percentile spread by the thicker lines, and the 50th percentile by the square box), which can be compared with the ClimWIP-only constraints (blue bars), the ASK-only constraints (solid red and orange bars), and the ASK-ClimWIP combined constraints (red and orange bars, white stripe).

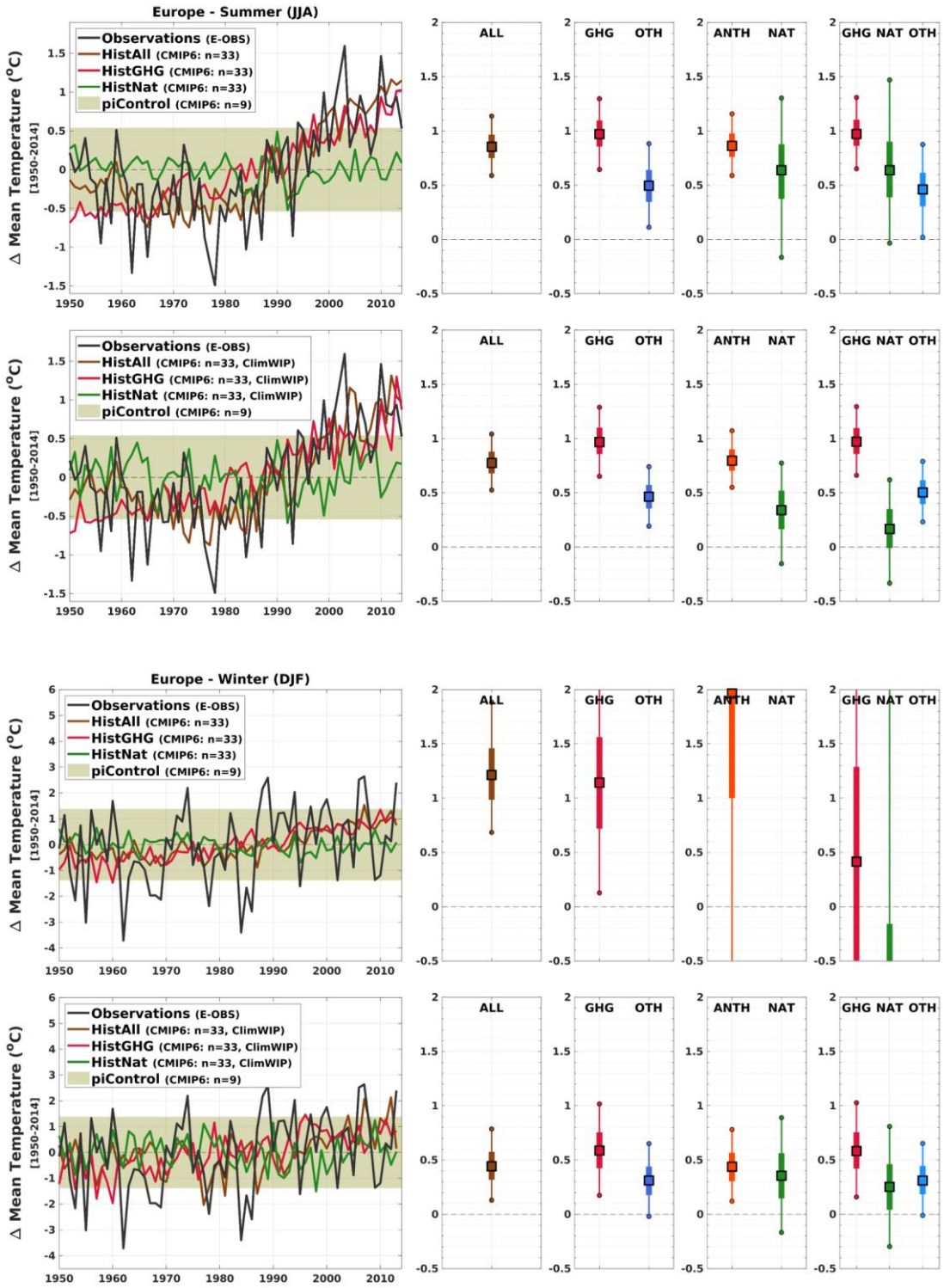

Figure 6: *Same as Figure 5, but separately showing both European summer (JJA, upper two panels) and winter (DJF, lower two panels) temperature anomalies (relative to 1950-2014), and associated scaling factors. Note that the annual time series of temperature anomalies have been plotted on a scale spanning 3x that shown in the summer panels, due to larger natural variability. The missing scaling factors in the equal-weighted winter panels (third row) are located outside of the displayed axis range, and suggest degeneracy of the natural and other anthropogenic forcing response with internal variability in this case.*

This study yields three different constraints to compare with the unconstrained CMIP6 distribution:

1. **ClimWIP-only** (both without-trend and including-trend variants)
2. **ASK-only** (focussing on the ASK-GHG and ASK-ANT variants)
3. **ASK-ClimWIP combined** (using the without-trend variant of ClimWIP)

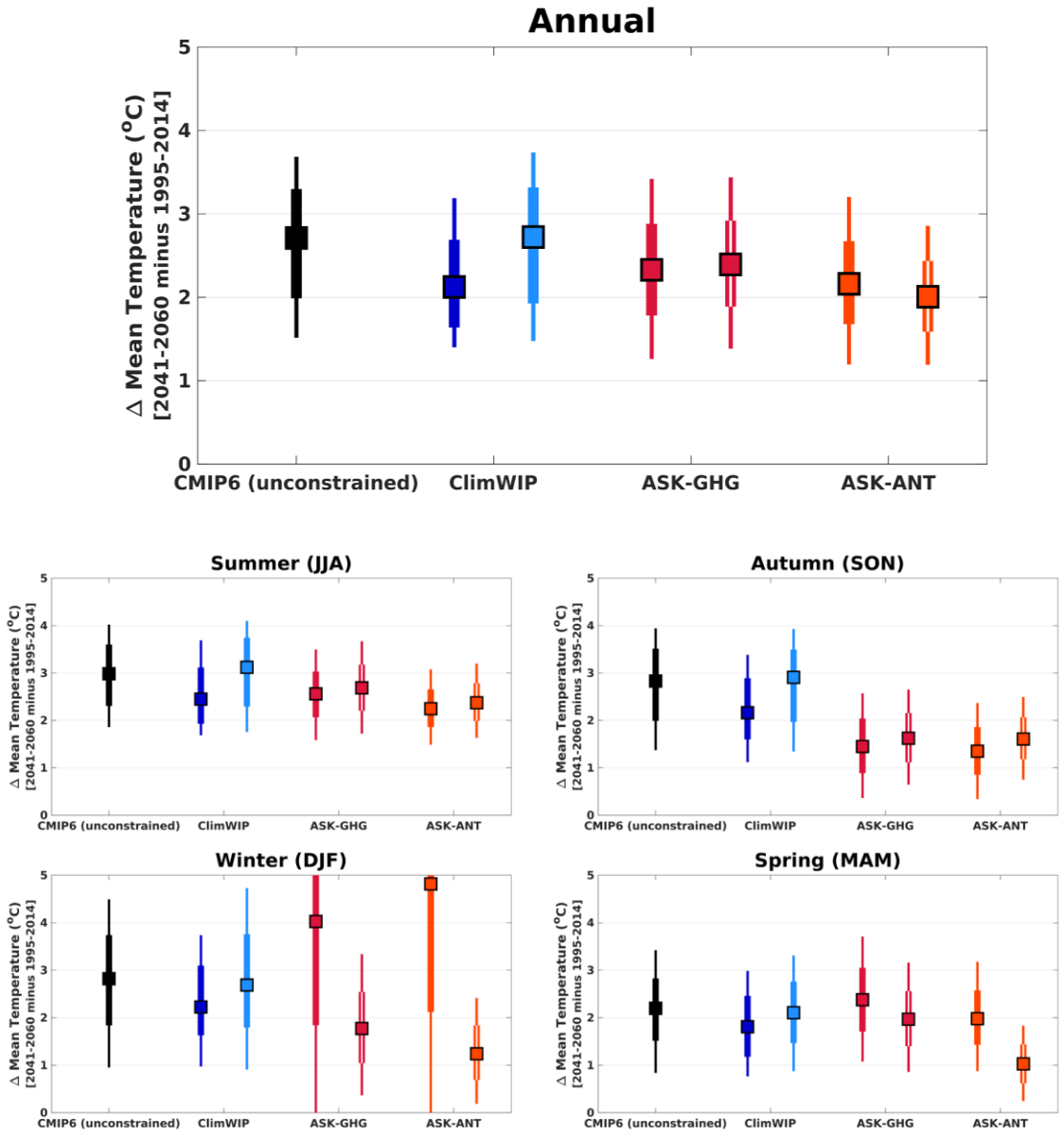

Figure 7: *Distributions of projected European annual (top panel) and seasonal (lower four panels) mean temperature anomalies for the period 2041-2060, relative to 1995-2014. Line colours denote the CMIP6 (SSP5-8.5) unconstrained model range (black); the ClimWIP performance-based weighted distributions, derived without the global temperature trend (light blue), and including the temperature trend (dark blue); the GHG- (red) and ANT- (orange) constrained distributions derived using the ASK method, using equal-weighted model fingerprints (solid bars) and the ClimWIP-weighted model fingerprints (bars with white stripes). The median (square marker) of each distribution, along with the 5th-95th (thin bars) and 25th-75th (thicker bars) percentile range of the distributions is shown.*

The projected temperature ranges under the ASK constraints (ASK-only and ASK-ClimWIP combined) are calculated by simply scaling the future CMIP6 (SSP5-85) multi-model mean temperature anomalies (unweighted and ClimWIP-weighted, respectively) by the associated scaling factors (with their uncertainty ranges, as shown in Figures 5 and 6). The constraints on projections of future annual European temperature (Figure 7, top panel) all indicate a slight reduction in the magnitude of warming, except for the ClimWIP-only variant without trend information (light blue) which for discussion purposes here we will disregard because its primary purpose in this study was to provide the weighted fingerprints utilised in the ASK-ClimWIP combined constraints (recall section 3.4.1). Both the primary ClimWIP-only (dark blue) and ASK-only (red and orange) constraints show a reduction in warming, and a slightly tighter spread compared to the unconstrained, especially the inner 25th-75th percentile range.

When applied to various seasons (Figure 7, lower panels), some differences emerge in the projected temperature ranges shown by the different constraints. While summer and autumn indicate a consistent reduction in the projected warming, both when applying ClimWIP-only and ASK-only constraints, in winter and spring there is less agreement, in particular winter which shows a large uncertainty range in the ASK-only constraint. This re-emphasises the limits of being able to apply an ASK constraint in periods with particular low signal-to-noise; winter, for example, where as previously shown (Figure 6) the scaling factors have a huge uncertainty range. As shown with the scaling factors discussed earlier, there is some promise in the combined ASK-ClimWIP constraints in being able to provide tighter constraints during these seasons of reduced signal-to-noise, and our future research efforts will evaluate to what extent this level of improvement is robust.

3.4.4 Out-of-sample testing of the observational constraints using CMIP5 as pseudo-observations

In this section we employ an out-of-sample framework for testing the performance of the different approaches outlined above in 3.4.3 (ClimWIP-only, ASK-only, ASK-ClimWIP combined). This is done by drawing a set of pseudo *'observations'* (pseudo-obs) from the ensemble of CMIP5 simulations. That is, we treat a randomly drawn simulation from the CMIP5 archive as a test observed case, which is used both over the historical period for deriving the various observational constraints (i.e. as the observations onto which we regress the model fingerprints), and then also into the future period in order to test the reliability of the constraints. This process is then repeated for several pseudo-obs samples (we use 63 in this analysis) to build statistical measures of the constrained distributions, including the spread and reliability.

This was feasible in-part because the various ClimWIP weighting vectors for the CMIP6 models had already been computed using CMIP5 as pseudo-obs (Brunner et al., 2020b), and thus provided a relatively straightforward means of testing the different constraints. In contrast to the "real" observations (where a combination of two observational datasets was used), the pseudo-obs approach always only uses one model at a time. In addition, there are certain features of the CMIP5 and CMIP6 suite of models that need to be kept in mind, for example, that some CMIP6 models are "next generation" CMIP5 models and as such might be structurally quite similar. A more detailed discussion can be found in Brunner et al. (2020b).

A limitation to this testing framework is that it assumes the same radiative forcing. The CMIP5 future simulations were run under the RCP8.5 emissions scenario, whereas the CMIP6 simulations utilise the newer SSP5-8.5. While designed to follow a similar radiative forcing change, the SSP can follow slightly different pathways to the RCP8.5 scenario (Forster et al., 2020). While these differences may have some impact on the interpretation of our results, and thus will require further investigation, it is thought that these are likely to be relatively minor. In any case, the ASK method itself (as implemented here, scaling the all-forced future runs with the GHG-only or ANT-only scaling factors) already makes underlying assumptions about the future radiative forcing agents.

A benefit of using CMIP5 simulations here is that the same prediction tool can be used without having to withhold a model that is used for the perfect model test (as had to be done in Schurer et al., 2018). A further benefit of using a separate set of CMIP5 simulations for testing, rather than just testing from within the CMIP6 distribution (e.g. a leave-one-out sample test) is that it provides a stress-test for the situation when the 'actual truth' lies outside of the ensemble, i.e. can the constraint capture a future that may be outside of the unconstrained envelope?

Notwithstanding the assumption of similar radiative forcing, the benefit of testing whether the constraints can pick-up (through weighting and/or forced-signal scaling factors) the different climate sensitivities of CMIP5 vs CMIP6 makes this out-of-sample testing an interesting pursuit.

The results of the out-of-sample testing are summarised in Figure 8, for both the unconstrained CMIP6 distribution (black line) and the different constrained distributions (coloured lines), for the projections of European summer temperature. The left panels display the average spread or "sharpness" of the 63 sets of distributions (one set of constrained distributions derived using each of the pseudo-obs members), while the right panels display the percentage of the future pseudo-obs trajectories that are found to lie within those various distributions, or in other words, the "reliability" of the constraining approach. Every marker (plotted in 5-yr increments) denotes the 20-yr moving average of the distributions' spread and reliability, thus the marker at 2050, for example, reflects the average spread and reliability of the constraints, averaged from 2041-2060.

The upper panels depict the 5th-95th percentile *"outer"* spread (on the left) and the associated reliability of the distributions (on the right), thus the 90% marker (dotted black horizontal line) indicates the ideal reliability for a well-performing constraint (i.e. that 90% of pseudo-obs trajectories were found to lie between the 5th-95th percentile of the projected distribution). In a similar way, the 25th-75th percentile *"inner"* spread and reliability (ideally capturing 50%) is shown in the lower panels. A reliability measure that is significantly lower than 90% (or 50%) implies that the constrained outer (or inner) distribution is inadequate for capturing the future time series (in this pseudo-obs testing framework), whereas a reliability that is significantly higher than 90% (or 50%) would suggest that the outer (or inner) distribution is under-confident, meaning that it might not be providing an adequately tight constraint. While a constrained distribution might be more reliable in capturing the future pseudo-obs when compared with another constraint, it might be doing so because it is a much wider (less sharp) constraint. Thus, it is important to look at reliability in conjunction with the sharpness in order to explore the practical utility or benefit of a particular constraint compared with another.

The results for the unconstrained CMIP6 distribution (black lines in Figure 8) sets a point of comparison for the various constrained distributions. While the average outer spread of the

constrained distributions are about the same or wider than the CMIP6 unconstrained distribution, some of the constraints do narrow the width of the inner distribution (25th-75th percentile). The CMIP6 unconstrained distribution is unreliable at capturing the CMIP5 pseudo-obs, as one might expect with the differing climate sensitivities (between CMIP6 and CMIP5). At mid-century (2050; 2041-2060), for example, only 50% of the pseudo-obs members were tracked within the unconstrained CMIP6 distribution's 5th-95th percentile range.

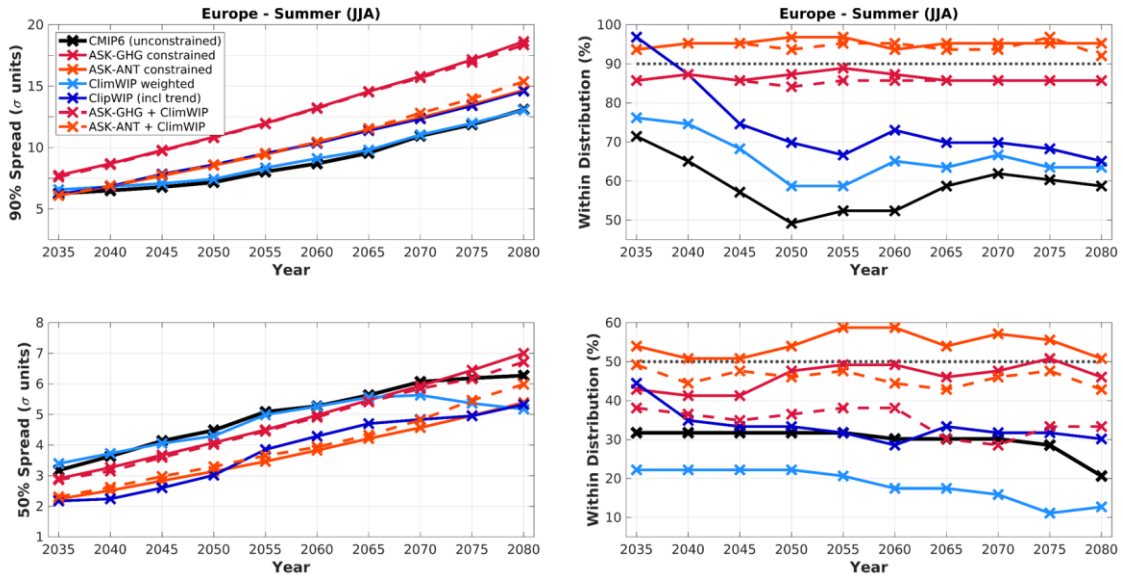

Figure 8: *Summary plots showing the sharpness of the spread (left panels) of the CMIP6 (9 models, 33 ensemble members, SSP5-8.5 emissions scenario) unconstrained and constrained distributions, and the percentage (right panels) of the 'pseudo-obs' members (CMIP5 RCP8.5, n=63) that lie within each of these distributions, for future projections (20-yr moving window) of European summer (JJA) temperature; depicting the central 90% (5th-95th percentile, upper panels), and 50% (25th-75th percentile, lower panels) of the distributions. Line colours denote the CMIP6 unconstrained model range (black); the GHG- (red) and ANT- (orange) constrained distributions derived using the ASK method; the ClimWIP performance-based weighted distributions, derived without the global temperature trend (light blue), and including the temperature trend (dark blue); and the combined ASK+ClimWIP constraints (dashed lines). The spread (left panels) of the projected distributions is displayed in units relative to the average standard deviation of the associated piControl simulations.*

The various constrained distributions are nearly all seen to be more reliable distributions of future projections than the unconstrained CMIP6 reference (right panels of FIgure 8). That is, a higher percentage of the pseudo-obs members track through the constrained outer distributions. The ASK-only constraints (ASK-GHG in red, and ASK-ANT in orange) are both very reliable throughout the projected 21st century, whereas the ClimWIP-only constraints are reliable early on in the period, but decline over subsequent decades. While this changing reliability with time is of interest here, further ongoing work will be required to better understand the causes (and to contextualise the results across different seasons and regions). The ClimWIP weighting that includes trend information (dark blue line) provides a more reliable constraint than the ClimWIP variant without trend information (light blue), whose inner spread is actually less reliable than the unconstrained. The combined ASK +

ClimWIP constraints (dashed red and orange lines) do not have a big impact on reliability (compared to ASK-only), except to reduce it somewhat with respect to the inner distribution (lower right panel).

Along with the reliability results, we now also take note of the differing sharpness of the various constraints (left panels of Figure 8). The two ASK constraints differ in their average sharpness, with ASK-GHG seen to be wider than ASK-ANT. This is consistent with the scaling factors based on true observations, discussed in the previous section (3.4.3), where ASK-ANT was found to also provide a tighter constraint (Figs 5-7). Combining both the reliability and sharpness information, ASK-ANT provides a good observational constraint, demonstrating reliability along with a relatively narrow distribution. The ClimWIP-only (including trend) constraint exhibits a sharpness comparable to the ASK-ANT constraint, while the ClimWIP without-trend variant is wider in the inner distribution, and slightly narrower in the outer distribution. The combined ASK + ClimWIP constraints yield an average spread that is very similar to that of the ASK-only constraints.

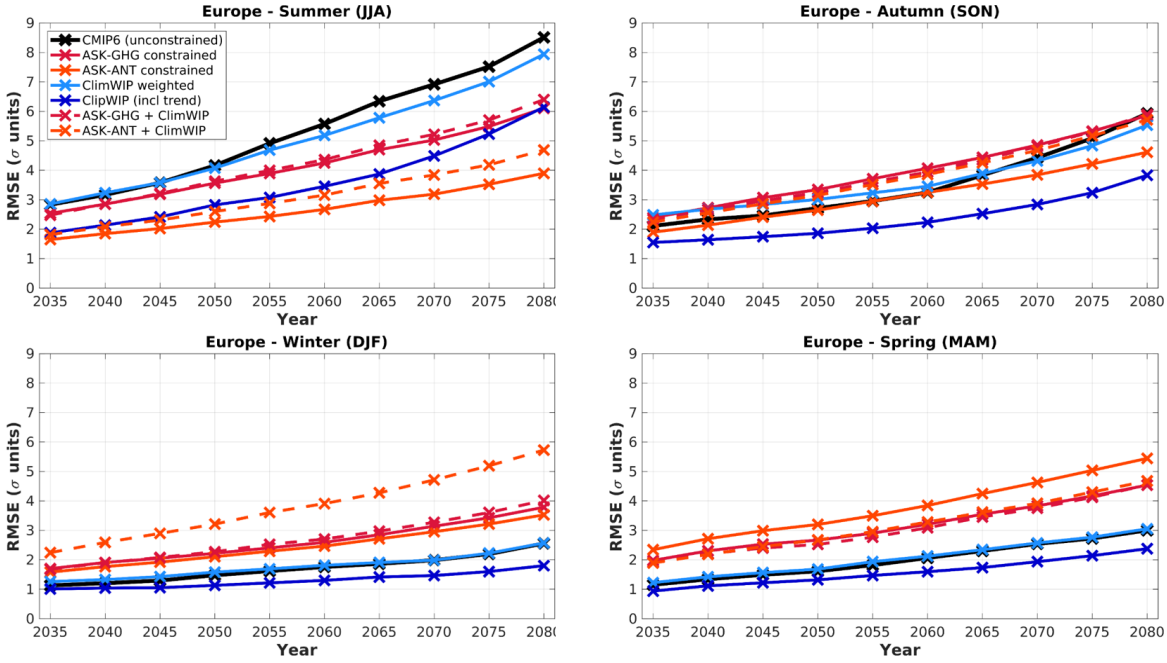

Figure 9: *Seasonal dependence of results: The RMSE difference between each of the 'pseudo-obs' (63 members, taken from CMIP5) and the centre (median) of each of the CMIP6 raw and constrained distributions (different coloured lines), for future projections (20-yr moving window) of European temperature, for summer (JJA, top left), autumn (SON, top right), winter (DJF, bottom left), and spring (MAM, bottom right). Units are the average standard deviation of the associated piControl simulations (different for each season).*

Along with the sharpness and reliability already discussed, we also show an additional metric, the average RMSE (Figure 9; the top-left panel displays European summer, as in Figure 8), that is computed between the pseudo-obs members and the median of the unconstrained and various constrained distributions. The RMSE metric also shows the improvement gained by using a constrained distribution (to predict the location of future pseudo-obs) compared with the unconstrained reference (black line). The ClimWIP weighting constraint without trend information (light blue line) offers very little improvement; however, the ClimWIP constraint with trend information (dark blue) significantly reduces the average error in the predictions. The ASK-ANT

constraint is seen to be the best by this metric, significantly reducing the RMSE (by almost half) throughout the century.

While the combined ASK + ClimWIP constraint shows substantial improvement compared to the unconstrained CMIP6 reference in summer, the RMSE remains larger compared to the equivalent ASK-only (i.e. with equal model weighting) constraints. Hence, this out-of-sample testing using CMIP5 pseudo-obs has not yet clearly demonstrated, in either reliability, sharpness, or error metrics, a benefit that can be gained through the combined ASK + ClimWIP constraint. However, the results are intriguing in various ways, and more needs to be done to fully investigate these differences. Reflections on how to avoid double counting in observational constraints (e.g. Annan and Hargreaves 2011) has been included in this work and will be further refined.

The performance of these different constraints (ClimWIP-only, ASK-only, ASK+ClimWIP combined) varies markedly when considering other seasons (Figure 9) and subregions (Figure 10). While the ASK-only constraints perform well in summer (as discussed with Figure 8), in other seasons they do quite poorly (at least by the RMSE metric shown in Figure 9), even when just compared to the unconstrained distribution. This is not surprising, however, given the differing signal-to-noise ratio seen across seasons (as was evident in Figure 5). The ClimWIP-only constraint, however, particularly the variant that includes a model's trend performance, provides a constraint that is better than the CMIP6 reference across seasons (and subregions, see Figure 10).

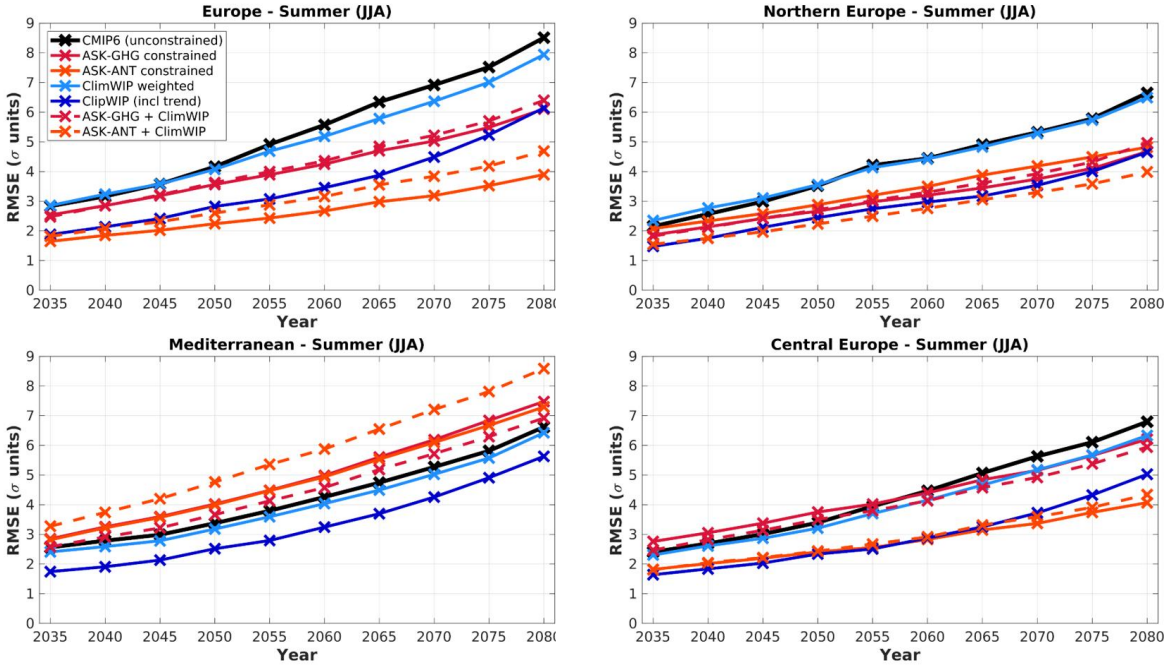

Figure 10: *Regional dependence of results: The RMSE difference between each of the 'pseudo-obs' (63 members, from CMIP5) and the centre (median) of each of the CMIP6 raw and constrained distributions (different coloured lines), for future projections (20-yr moving window) of summer (JJA) temperature, for several regions: Europe (top left), Northern Europe (top right), Mediterranean (bottom left), and Central Europe (bottom right). Units are the average standard deviation of the associated piControl simulations.*

In summary, using ASK as an observational constraint for projections of European summer temperature is shown to be reliable in this framework. The results of out-of-sample testing produce constrained distributions (5th-95th and 25th-75th percentile ranges) that are found to contain the future pseudo-obs members around 90% and 50% of the time, respectively. While ClimWIP provides a similarly sharp constraint (especially when including trend information) and has good reliability in the near term, it appears to be less successful in predictions (of CMIP5, in this testing framework) beyond the first couple of decades. We note again that the ClimWIP weights used here are tailored to constrain the global mean, annual mean temperatures for the entire CMIP6 ensemble, so we would likely expect additional skill if the ClimWIP weights were calibrated to the specifics of this study (region, season, etc.), something that we are interested in exploring further. We have also begun exploring results for precipitation, which shows some early promise but requires more work to be explored prior to publication in the scientific literature.

There is an indication that the use of model weighting can potentially provide sharper constraints on projections, but not in all cases and some of these systematic effects are yet to be understood. Gaining a better understanding of the seasonal and subregional variation in the performance of the different observational constraints as well as their performance across other climate variables is an ongoing task (crosscutting between WP2 and WP5).

3.4.5 Using observational skill scores for predictions as model weighting

Lastly, we experimented with using lead-year correlations from Subpolar gyre predictions (Table 2; see below), instead of the ClimWIP performance weights, for constructing the multi-model mean fingerprints used in the ASK method and deriving scaling factors (Figure 11). While the estimate of the GHG signal against other forcings remained fairly robust, the 3-signal regression degenerated sharply. This may either reflect that weights inferred from the SPG region are not relevant for European land region projections, or that near-term prediction skill, at least in that case, does not indicate improved performance for projections. We examine the influence of SPG on European summer surface air temperature in more detail in section 3.5.3, while leaving a detailed analysis of the effect of decadal prediction skill on projections to be analyzed in the future, evaluating if skill scores for the same target of prediction can add value to a constraint.

Table 2: *Anomaly correlation coefficient of SST in the North Atlantic subpolar gyre in initialized hindcasts for lead times 1-3 years (1st column), 5-7 years (second column) and non-initialized hindcasts (historical simulations, third column) with HadISST sea surface temperature.*

Model Name	Anomaly Correlation Coefficient (1970-2014)		
	LY 1-3	LY 7-9	HIST
CanESM5	-0.45	0.71	0.85
CESM1-DPLE	0.88	0.83	0.81
HadGEM-GC3.1-MM	0.93	0.80	0.26
IPSL-CM6A-LR	0.73	0.81	0.84
MIROC6	0.91	0.63	0.92
MPI-ESM1.2-HR	0.91	0.68	0.84

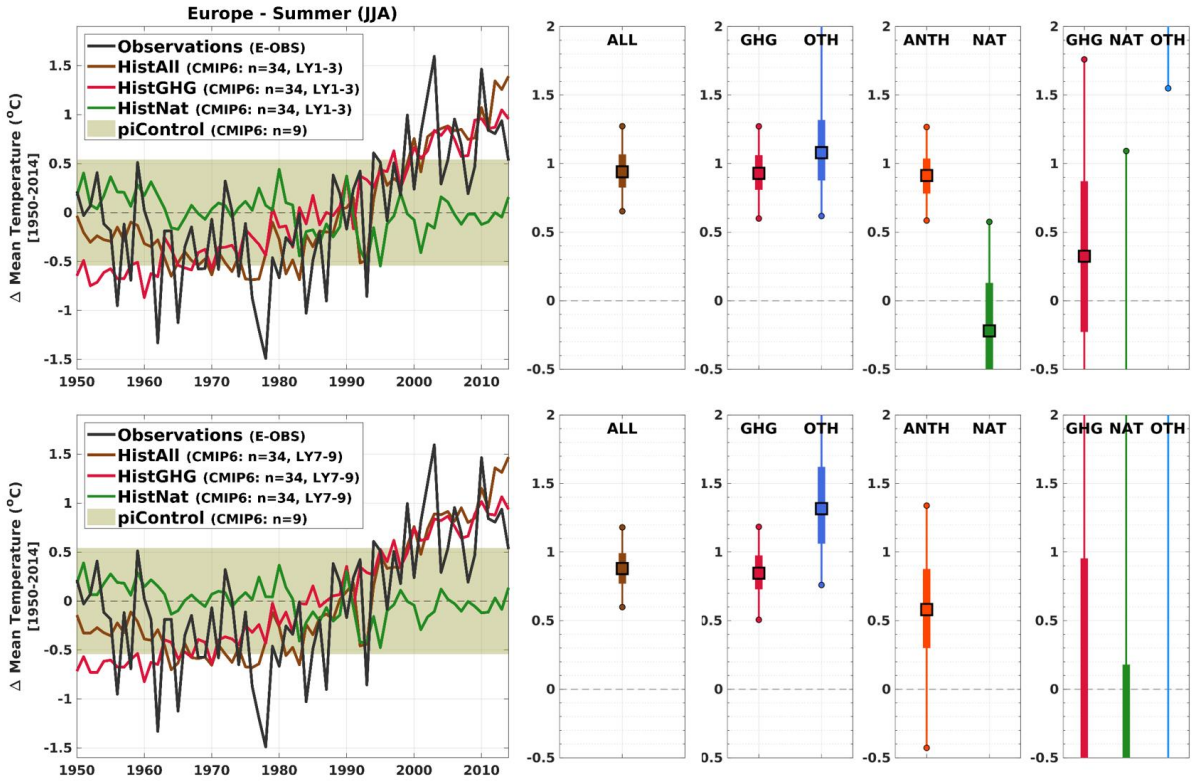

Figure 11: *Same as in Figure 5, but here showing the time series of European summer temperature anomalies weighted by the lead-year correlation weights as shown in Table 1, along with the scaling factors derived from the weighted model fingerprints regressed onto the observations. Weights are based on the correlation for lead-years 1-3 (LY1-3, upper panels) and lead-years 7-9 (LY7-9, lower panels).*

3.5 Examples of Observational constraints used in initialized predictions (IPSL)

Several examples of observational constraints in initialized decadal climate prediction simulations are presented as well as tested for their potential. These examples include: constraining the amount of agreement between model simulations and observations (*predictive skill*) that arises from the initialization process (e.g. Doblas-Reyes et al., 2013) as well as different forcings in the climate system over time (as in Brune et al., 2018; Borchert et al., 2019a, Christensen et al., 2020) (section 3.5.1); constraining the predictive skill found in different initialized model systems using the models' inherent characteristics (e.g. Menary et al., 2015; Menary & Hermanson, 2018) (Section 3.5.2); and constraining predictive skill over Europe using predictions of North Atlantic sea surface temperature (SST) (as suggested by Sgubin et al., under review) (section 3.5.3). We thus test decadal prediction experiments for observational constraints on the time dimension, the model dimension, and the spatial dimension, respectively. Among these assessments, the analysis of observational constraints on the time (exploring the changing importance of various forcings and internal variability over time) and model (a first step towards weighing initialized climate prediction ensembles) dimensions is closest to the approach used for climate projections (section 3.4), while a constraint on the spatial dimension lies beyond what is currently proposed for projections. All of these explorative

investigations will pave the way to merging initialized and non-initialized climate predictions in order to tailor near-term climate prediction to individual users' needs.

The analyses we present rely on the following methods. We consider SST and surface air temperature (SAT) for the period 1960-2014 in our analyses. Our analyses are based on simulations from the CMIP6 archive. We analyze initialized decadal hindcasts from the DCPP project (HC; Boer et al., 2016), non-initialized historical simulations that are driven with reconstructed forcing (HIST; Eyring et al., 2016), as well as simulations from the detection and attribution MIP (DAMIP; Gillett et al., 2016) that isolate the effects of different forcings on climate. For comparison and to constrain predictions, SST from HadISST (Rayner et al., 2003) and SAT from the HadCRUTv4 gridded observational data set (Morice et al., 2012) are used. Agreement between model simulations and observations (*prediction* or *hindcast skill*) is here quantified as pearson correlation between simulation and observation (*Anomaly Correlation Coefficient*, *ACC*; Jolliffe and Stephenson, 2003) and mean square skill score (*MSSS*; Smith et al., 2020). ACC tests whether a linear relationship exists between prediction and observation, whereas MSSS quantifies the absolute difference between simulation and observation.

We do not subtract any trend from the data sets prior to analysis. In all cases, anomalies against the mean state over the period 1970-2005 of the respective data set are formed; this equates to a lead-time dependent mean bias correction in initialized hindcasts. When focusing on temperature in the subpolar gyre (SPG), we analyze area-weighted average SST in the region 45-60N, 10-50W. Surface temperature over Europe is represented by land grid-points in the SREX regions Northern Europe (N-EU), Central Europe (C-EU) and Mediterranean (MED). Whenever appropriate, we analyze how prediction skill changes over time (so-called *windows of opportunity*; Borchert et al., 2019a, Mariotti et al., 2020) to attribute changes in skill to specific climatic boundary conditions.

3.5.1 Constraining the sources of decadal prediction skill for North Atlantic SST

Focusing on the North Atlantic subpolar gyre (SPG) region, we first study the role that different forcings play in modulating North Atlantic SST. A paper summarizing these findings was recently accepted for publication in Geophysical Research Letters (Borchert et al., 2020). By showing how well simulated North Atlantic SST variations from HC, HIST, and DAMIP align with observations, we illustrate the added value of climate model initialization (linking to T5.1) as well as how and when forcing impacts decadal variations of North Atlantic SST. This analysis reflects an observational constraint on the credibility of North Atlantic SST predictions in the context of different forcings and internal variability in the climate system.

Multi-model means of initialized decadal prediction simulations as well as historical simulations from CMIP5 (based on 6 models; Table 3) and CMIP6 (based on 7 models; Table 3) show that CMIP6-simulated SST generally correlate better with observations than that simulated by CMIP5 models (Fig. 12). The improvement from CMIP5 to CMIP6 is mainly explained by improved representation of North Atlantic SST in historical simulations in CMIP6 compared to CMIP5 (Fig. 12e-h), indicating a reduced impact of initialization on decadal prediction skill in CMIP6 (as will be investigated explicitly in T5.1, D5.2). The relatively small influence of initialization on North Atlantic SST prediction skill in CMIP6 is particularly surprising in the subpolar gyre (SPG) area (indicated in Fig. 12e,f), which has in the past been strongly implicated to show large initialization-related skill increase (e.g. Yeager & Robson, 2018; Brune & Baehr, 2020). This implies a stronger role for forcing in

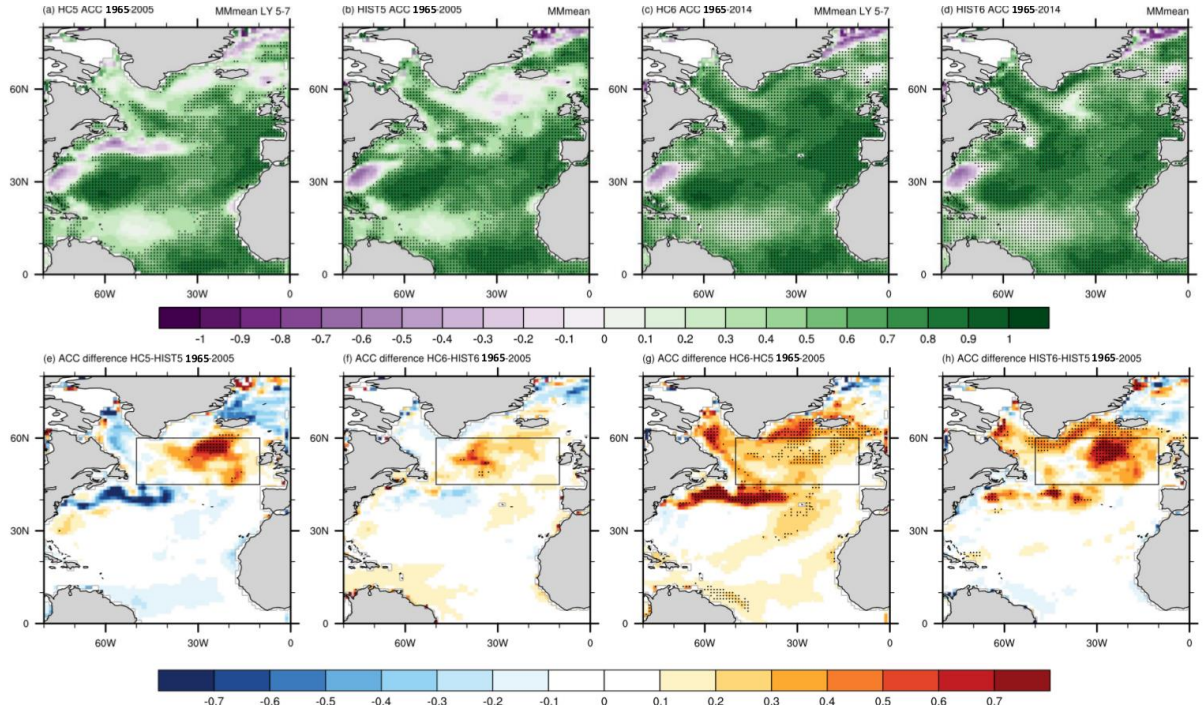

Figure 12: *Anomaly correlation (ACC) between different multi-model ensemble means and observed SST from HadISST. (a,b) show ACC of CMIP5 initialized hindcasts (HC5; lead years 5-7) and historical simulations (HIST5) for the period 1965-2005, respectively. The multi-model means are based on the same 6 models (Table 3). (c,d) show ACC of CMIP6 initialized hindcasts (HC6; lead years 5-7) and historical simulations (HIST6) for the period 1965-2014, respectively. Again, these multi-model means are based on the same models, but the model set is not coherent between CMIP5 and CMIP6. (e-h) show the difference between HC5 and HIST5, HC6 and HIST6, HC6 and HC5, and HIST6 and HIST5 for the period 1965-2005, respectively. Stippling indicates significance at the 95% confidence level. The black outline marks the subpolar gyre (SPG) region. From Borchert et al. (2020), their figure 1.*

Table 3: *Models used in the analysis presented in section 3.5.*

Modelling Centre	Model	Ensemble Size				
		CMIP5		CMIP6		
		Historical	Decadal Hindcasts	Historical	Decadal Hindcasts	DAMIP
BCC, China	BCC-CSM1.1 (Wu et al., 2013)	3	3			
	BCC-CSM2-MR (Wu et al., 2019)			3		3
CCCma, Canada	CanCM4 (Merryfield et al., 2013)	10	5			
	CanESM5 (Swart et al., 2019)			20	10	10
CNRM, France	CNRM-CM6-1 (Voldoire et al., 2019)			30		10
IPSL, France	IPSL-CM5A-LR (Dufresne et al., 2013)	6	3			
	IPSL-CM6A-LR (Boucher et al., 2020)			30	10	10
JAMSTEC, Japan	MIROC5 (Watanabe et al., 2010)	5	6			
	MIROC6 (Tatebe et al., 2019)			10	10	10
MOHC, UK	HadCM3 (Smith et al., 2013)	10	10			
	HadGEM3-GC31-LL (Kuhlbrodt et al., 2018)			4		4
	HadGEM3-GC31-MM (Andrews et al., 2020)			2	10	
MPI-M, Germany	MPI-ESM1.0-LR (Giorgetta et al., 2013)	3	3			
	MPI-ESM1.2-HR (Muller et al., 2018)			10	10	
MRI, Japan	MRI-ESM2.0 (Kawai et al., 2019)			5		5
NASA, USA	GISS-E2.1-G (Kelley et al., 2019)			10		5
NCAR, USA	CESM1.1-CAM5 (Yeager et al., 2018)				20	
	CESM2 (Danabasoglu et al., 2020)			10		
NCC, Norway	NorESM2-LM (Seland et al., 2020)			3		10
	NorCPM1 (Bethke et al, under review)			30	10	
Total models (members)		6 (47)	6 (30)	13 (167)	7 (80)	9 (67)

decadal variations of North Atlantic SST than previously thought, revealed by CMIP6 simulations. Understanding which forcing dominates SST variations at which time can therefore help constrain prediction skill for North Atlantic SST. We approach this problem using the forcing partitioning in DAMIP simulations, focusing on SPG SST.

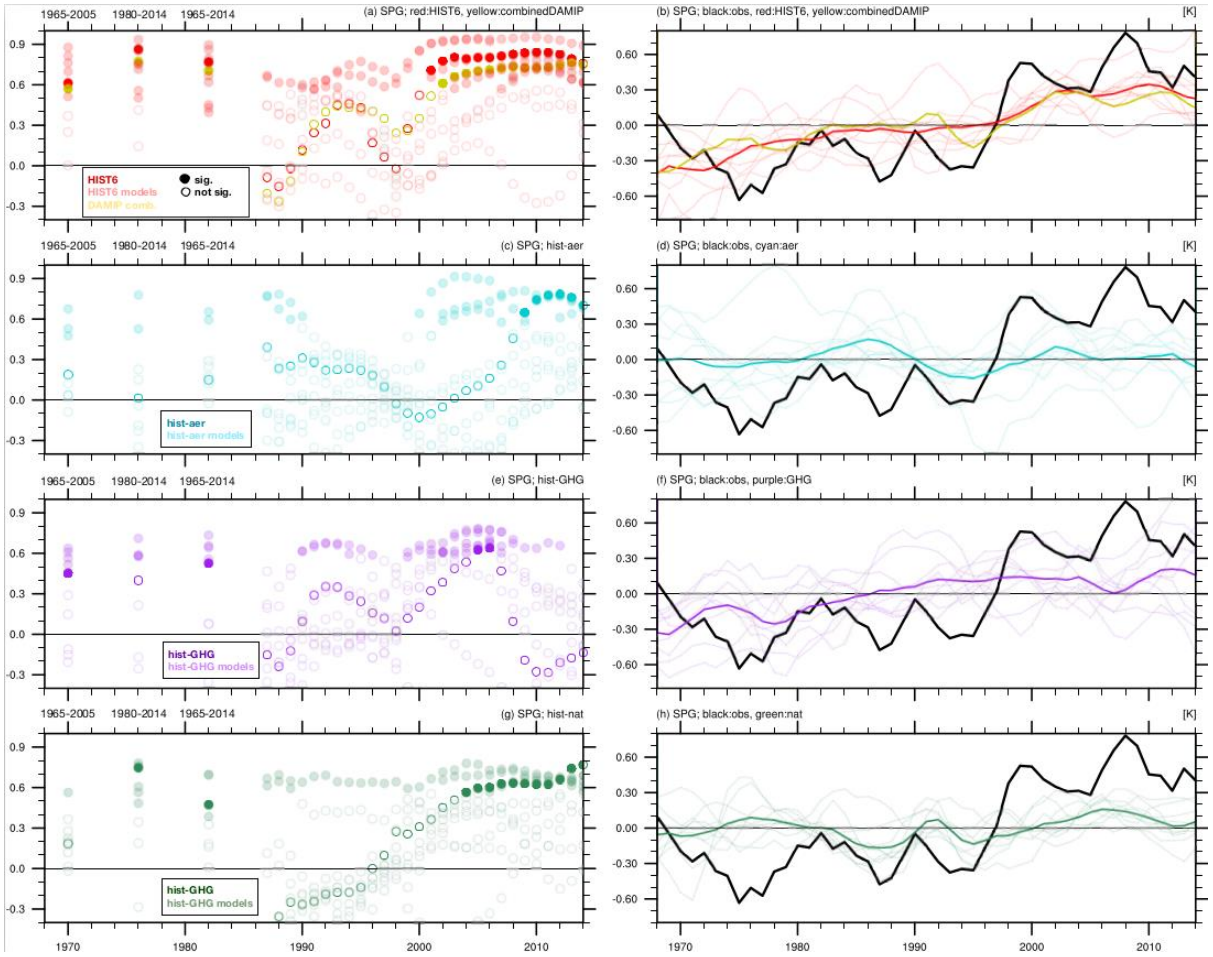

Figure 13 : *SPG SST ACC skill and time series in a 9-model multi-model ensemble for which DAMIP simulations are available (see Table 3) from (a,b) HIST6 (red), and for simulations from the DAMIP scenarios (c,d) hist-aer (cyan), (e,f) hist-GHG (purple), and (g,h) hist-nat (green). The linear sum of the multi model ensemble means of hist-aer, hist-GHG and hist-nat is shown in (a,b) in yellow. (a,c,e,g) ACC for a rolling 20-year window (dots/circles), positioned over the last year of the respective 20-year period. ACC is examined for the multi-model mean (solid colors) and the individual models (shading). Circles/dots on the left show skill for the full CMIP5 and CMIP6 periods as well as the period since 1980, specified above. Full circles indicate significant skill, empty circles indicate that skill is not significantly different from 0 (95% confidence, see Methods). (b,d,f,h) Time series of SPG SST anomalies in observations (thick black) as well as the multi-model ensemble mean (solid colors) and the individual models (shading). From Borchert et al. (2020), their figure 3.*

The contribution of external forcing to observed SPG SST variations is particularly strong in the period after around 1980 (Fig. 13a). Using a 9-member multi-model ensemble mean from DAMIP simulations (Table 3), we decompose the different forcing contributions to the total SPG SST correlation to observations over time. The DAMIP-based decomposition into aerosol (hist-aer), greenhouse gas (hist-GHG) and natural (hist-nat) forcings is appropriate because their linear sum

resembles the total signal simulated by the historical simulation (Fig. 13a,b), an assumption also used in the projection-focused ASK method.

Correlations of the simulations with isolated forcings to observations indicate the degree to which the individual forcings explain observed variability. We find a number of somewhat consistent signals among models. Several of the single models agree that anthropogenic aerosols only explain an insignificant amount of observed SPG SST variations until very recently (fig. 13c); greenhouse gas forcing explains a significant amount of SPG SST variations during a very short time window around the 1990s (fig. 13 e); and natural forcing explains much of the SPG SST variations observed since the 1980s, but not before that (fig. 13g). Due to the substantial amount of model spread found in this analysis, however, these findings are sensitive to the set of models used in the analysis. After 1980, hist-aer, hist-GHG, and hist-nat explain 0% (two standard deviation spread of resampling the individual models 10,000 times with replacement: -6 to 10%), 16% (-1 to 35%) and 55% (34 to 59%) of the observed SST variance, respectively. These findings implicate natural forcing over the other examined forcings as an important driver of the high skill in HIST6 after 1980.

This work indicates that at times of strong forcing, predictions and projections of North Atlantic SST with CMIP6 models averaged together can be expected to be credible. Natural forcing, particularly major volcanic eruptions (Hermanson et al., 2020; Borchert et al., 2020), plays a particularly prominent role here. In absence of strong forcing trends, initialization is needed to generate skill in decadal predictions of North Atlantic SST (Borchert et al., 2020; their figure 2). Constraining predictions of the past with different forcing scenarios and internal variability thus reflects a promising approach to understanding the credibility of decadal climate forecasts.

3.5.2 Constraining decadal hindcast skill between different CMIP6 models

The previously presented observational constraint works on the time dimension for the multi-model mean similar to that used in ASK constraints (which, however, focus on a different timescale). We now consider an approach more similar to model-related weights presented in section 3.3 and 3.4. Instead of multi-model means, we here assess the 7 individual CMIP6 DCPP decadal prediction systems with the aim of linking the skill in model systems to their inherent properties. Again, we focus this analysis on North Atlantic subpolar gyre SST due to its previously demonstrated ties to European SAT (Gastineau & Frankignoul, 2015; Borchert et al., 2019b).

Initialized predictions from CMIP6 show general agreement on ACC skill for SPG SST (fig. 14a). The only prominent outlier to this is the CanESM5 model, which displays a strong initialization shock until approximately lead year 7 due to issues in the North Atlantic region with the direct initialization from ORAS5 observations (Sospedra-Alfonso, pers. comm.). For this reason, we will discuss CanESM as a special case whenever appropriate. The other 6 models generally agree on high initial skill which degrades over lead time (fig. 14a), showing some degree of spread in ACC skill that could be linked in model or prediction system properties. This spread in skill is found for both ACC and MSSS (fig. 14b), indicating the robustness of this result. Moreover, skill degradation occurs at different rates in the different model systems, representing another possible angle at which to try to explain the skill differences.

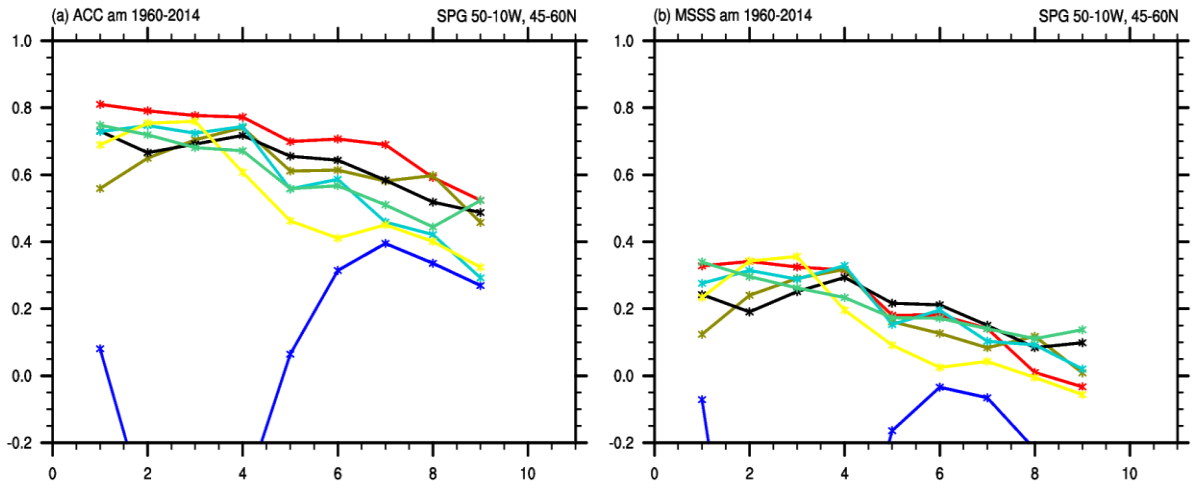

Figure 14 *(a) ACC for SPG SST in 7 initialized CMIP6 model systems (y-axis) against lead time in years (x-axis). (b) as (a), but showing MSSS on the y-axis. Colors indicate the individual models: blue = CanESM5, red = CESM-DPLE, brown = IPSL-CM6A-LR, black = HadGEM3-GC31-MM, cyan = MIROC6, yellow = MPI-ESM1.2-HR, green = NorCPM1.*

Sgubin et al. (2017) showed that the representation of stratification in the upper 2000 meters of the North Atlantic subpolar gyre in different models is a promising constraint on climate projections, impacting among other things the likelihood with which a sudden AMOC collapse is projected to happen in the future. Ocean stratification impacts North Atlantic climate variability not only on multidecadal time scales, but also locally on the (sub-)decadal time scale. It is therefore apposite to explore this as an observational constraint on the model dimension and test whether models that show comparatively realistic SPG stratification also show higher SPG SST prediction skill. To this end, we calculate a stratification indicator similar to Sgubin et al. (2017) by integrating SPG density from the surface to 2000m depth for the period 1960-2014 in the different CMIP6 HIST models and EN4 reanalysis (Ingleby and Huddleston, 2007), and calculating the root-mean square difference between modeled and observed stratification. This difference will be examined for a possible linear relationship to SPG SST prediction skill.

For short lead times of up to 5 years, we find a strong negative linear relationship between SPG SST as well as AMV prediction skill in the 7 prediction systems analyzed here, and SPG stratification bias in the corresponding historical simulation (fig. 15). Note that due to the initialization issue in CanESM5 discussed above, that model does not behave in line with the other models at short lead times. These findings indicate that models that simulate a realistic SPG stratification tend to predict SPG surface temperature skilfully. While inspiring hope that SPG mean stratification state might be a good indicator of SPG SST hindcast skill, this result is based on a regression over 6 data points and therefore lacks robustness. More models will be added to this analysis as DCPP simulations become available on ESGF.

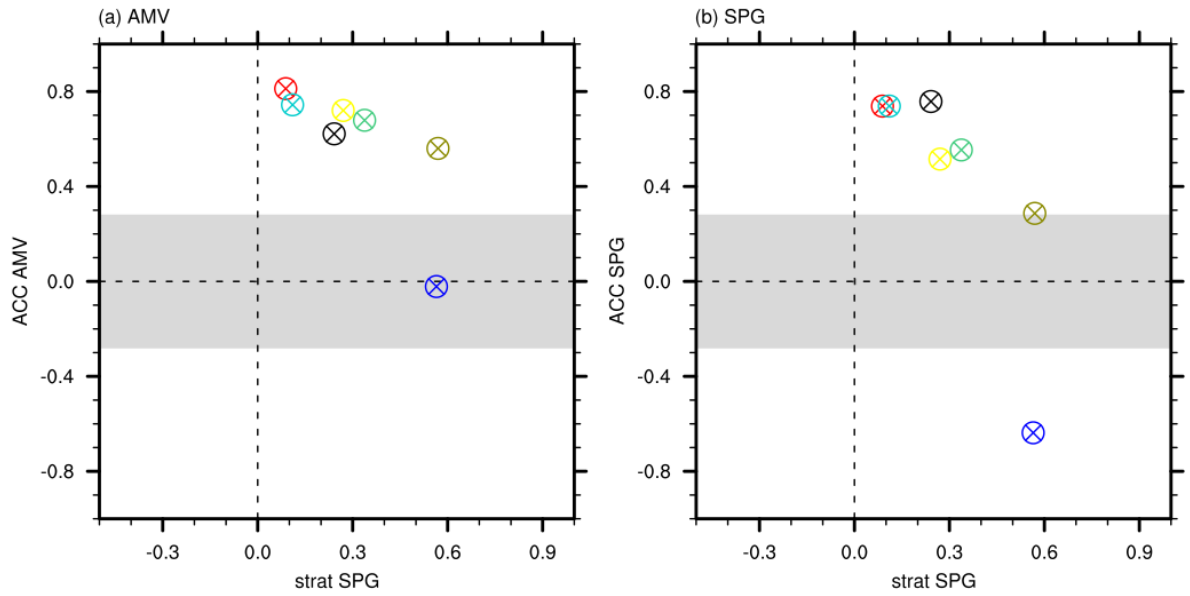

Figure 15 : *ACC for Atlantic Multidecadal Variability (AMV; defined as North Atlantic SST in the North Atlantic 0-70N minus global mean SST) (left) and SPG SST (right) in 7 initialized CMIP6 model systems at lead years 1-5 (y-axis) against RMSE between observed and simulated SPG stratification in the period 1960-2014 from the corresponding historical simulations (x-axis). Colors indicate the individual models: blue = CanESM5, red = CESM-DPLE, brown = IPSL-CM6A-LR, black = HadGEM3-GC31-MM, cyan = MIROC6, yellow = MPI-ESM1.2-HR, green = NorCPM1.*

A possible inherent property of climate models that explains the skill to expect from them is the amount of climate variability produced by the model by itself. This variability is here represented by standard deviation of SPG SST over 500 years in the pre-industrial control (piC) simulations of the 7 different CMIP6 models (*sigma SPG*). The assumption here is that models that produce pronounced SPG SST variability by themselves reproduce decadal SPG SST changes more accurately than those that do not. At long lead times of 6-10 years, decadal North Atlantic SST hindcast skill shows reasonable linear increase with sigma SPG in the respective control simulations (fig. 16 a,b). A possible cause for increased skill in models with higher sigma is the linear relationship of sigma SPG to SPG variability persistence in the piControl simulations (figure 16c): higher variability implies larger persistence. And one could expect in turn that longer decorrelation time scale leads to longer predictability. The linear correlation between sigma SPG and SPG hindcast skill, however, is not robust across lead times (not shown); at shorter lead times of up to 5 years, there is no linear relationship between the two. This is probably because the initialization process masks some of the influence of inherent model features at short lead times. Again, this analysis is limited by the small amount of models for which decadal hindcast simulations are currently available. Adding more models to this analysis could point towards other conclusions, or strengthen the results presented here. Additionally, extending this analysis to other regions in the piControl simulations (as in Menary & Hermanson, 2018) would provide valuable insights into the way that the representation of underlying dynamics in different models preconditions their skill of prediction SPG SST.

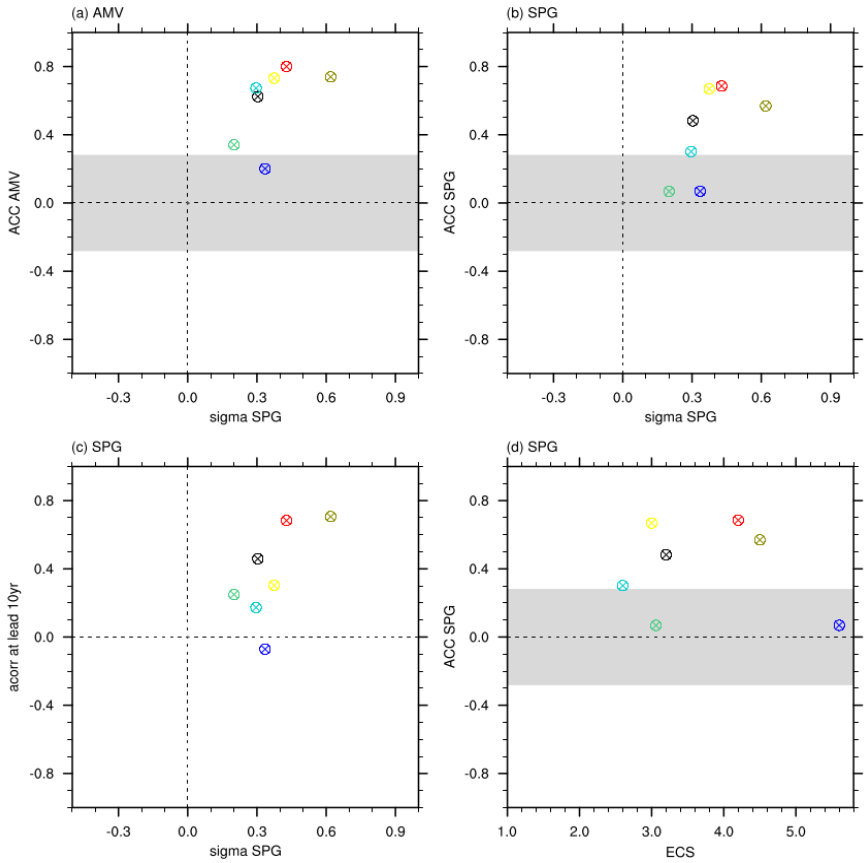

Figure 16 : *ACC for North Atlantic (AMV and SPG) SST at lead years 6-10 connected to inherent model characteristics: (a) AMV SST ACC in initialized hindcasts from 7 CMIP6 models (y-axis) against standard deviation of SPG SST in the corresponding models' piControl simulations, 10 year low-pass filtered. (b) as (a), but for SPG instead of AMV ACC. (c) shows decorrelation time scale (i.e. autocorrelation at a 10 year lag) of SPG SST in piControl simulations (y-axis) against standard deviation of SPG SST in the corresponding models' piControl simulations. (d) as (b) but comparing SPG ACC with the models' equilibrium climate sensitivity. Grey shading represents insignificant hindcast skill for North Atlantic SST. Colors indicate the individual models: blue = CanESM5, red = CESM-DPLE, brown = IPSL-CM6A-LR, black = HadGEM3-GC31-MM, cyan = MIROC6, yellow = MPI-ESM1.2-HR, green = NorCPM1.*

Equilibrium climate sensitivity (ECS) is found to show a reasonable linear relationship to SPG SST hindcast skill at long lead times (fig. 16d), representing another avenue to explore for understanding why models show higher or lower predictive skill for SPG SST. A physical reason for this high agreement between a model's ECS and its capability of predicting SPG SST at long lead times, however, remains to be assessed. It is noteworthy that this linear relationship only holds when the CanESM model is not included in the regression, which is reasonable due to the aforementioned issues with these initialized hindcasts. Other possible discriminant factors for decadal SPG SST prediction skill are summarized in table 4, including model initialization strategy and resolution of the ocean model in the respective model. Ocean resolution is clearly not a distinctive criterion to identify skillful models. Firm conclusions are difficult to draw concerning initialization procedure, as initialization procedures are diverse and do not cluster in an obvious way. These hypotheses will be pursued further during the remainder of the project.

Table 4: *Model characteristics for CMIP6 initialized hindcast models used in this analysis (under construction).*

		#hist	#hind	Init	Resolution OCE	ECS / TCR	
CanESM	CanESM5	20	10	F 3DORAS5 T&S nudging + SST (ERSST &OISST) + Sea ice HadISST & h + atm 2DT, u, v, h nudged to ERA40 and ERA-INTERMI,+ land & ocean BGC	ORCA1	5.6 / 2.7	Sospedra-Alfonso and Boer 2020
CESM	CESM1.1	20	40 (20)	F CORE*-FOSI	1deg	4.2 (CESM2: 5.2 / 2.0)	Yeager et al. 2018
IPSL	IPSL-CM6A-LR	30	10	A SST+SSS nudging	eORCA1	4.5 / 2.3	Boucher et al. 2020 Estella-Perez et al 2020
Metoffice	HadGEM3-GC2	4	10	F nudging to 3DT+S analysis (10days)+ sea ice (1day)+ atm T,u,v(6hrs)	eORCA1	3.2 (GC1.3 5.5/2.6)	Dunstone et al. 2018
MIROC	MIROC-6	10	10	A 3DT,S sea ice, atmT,v,v	tripolar ca 1°	2.6 / 1.6	Tatebe et al. 2019, Watanabe et al 2020
MPI	MPI-ESM1.2	10	10	A Brut force. 3DT+S ORAS4 Anomalies + sea ice concentration anomalies, 3DT,vortcity,divergence,P (full)	0.4, 40 levels	3.0 / 1.7	Pohlmann et al. 2019
NorCPM	NorCPM1	30	10	A Kalman filter 3DT&S EN4 + SST (HadISST & OISST)	ocean bipolar 1°, atm 1.9°x2.5°	3.06 + 1.6 (NorES2-LM ECS2.56)	Bethke et al., in prep
EC-Earth	EC-Earth3?			F ORAS4 T-3 nudging as in Sanchez-Gomez et al. 2016.	eORCA1	4.3 (4.1 Ec-Earth-Veg3) / ?	
GFDL	GFDL-CM4?			ECDA Ensemble Kalman Filter 3DT+s+AtmT+u+v		GFDL-CM4 ECS 3.89	

3.5.3 Constraining decadal predictions of seasonal European SAT using North Atlantic SST

Finally, we examine observational constraints on the space dimension of initialized prediction experiments. These investigations shed light on processes that influence and might therefore lead to decadal prediction skill in certain regions. Here, we examine predictable North Atlantic SST as a possible source of prediction skill for European surface air temperature, which is famously difficult to predict due to small signal-to-noise ratio (e.g. Hanlon et al., 2013; Borchert et al., 2019b; Smith et al., 2020). Over Europe, we use the SREX regions as a first step of homogenizing boundary conditions with the ultimate aim of merging predictions and projections.

We analyze the decadal prediction skill for SAT in SREX regions in CMIP6 models during summer and winter (fig. 17) after subtracting the linear trend, which is stronger in the SAT indices than for SPG SST. Temperature in SREX regions shows generally low hindcast skill. Winter SAT hindcast skill (fig. 17a,c,e) is generally lower than summer SAT hindcast skill (fig. 17b,d,f). We find some differences between the different regions, with a tendency for higher skill towards the South. Because hindcast skill for SREX SAT shows more inter-model spread than for SPG SST, attempts to connect hindcast skill in individual models to inherent properties of the model (as above) is promising interesting insights. The generally low (and possibly insignificant) level of skill for all models in the SREX regions indicates, however, that discriminatory features of skill between models would not enable the identification of skillful models without further treatment. An analysis of windows of opportunity (as in section 3.5.1), for example, might reveal times of high prediction skill in the SREX regions.

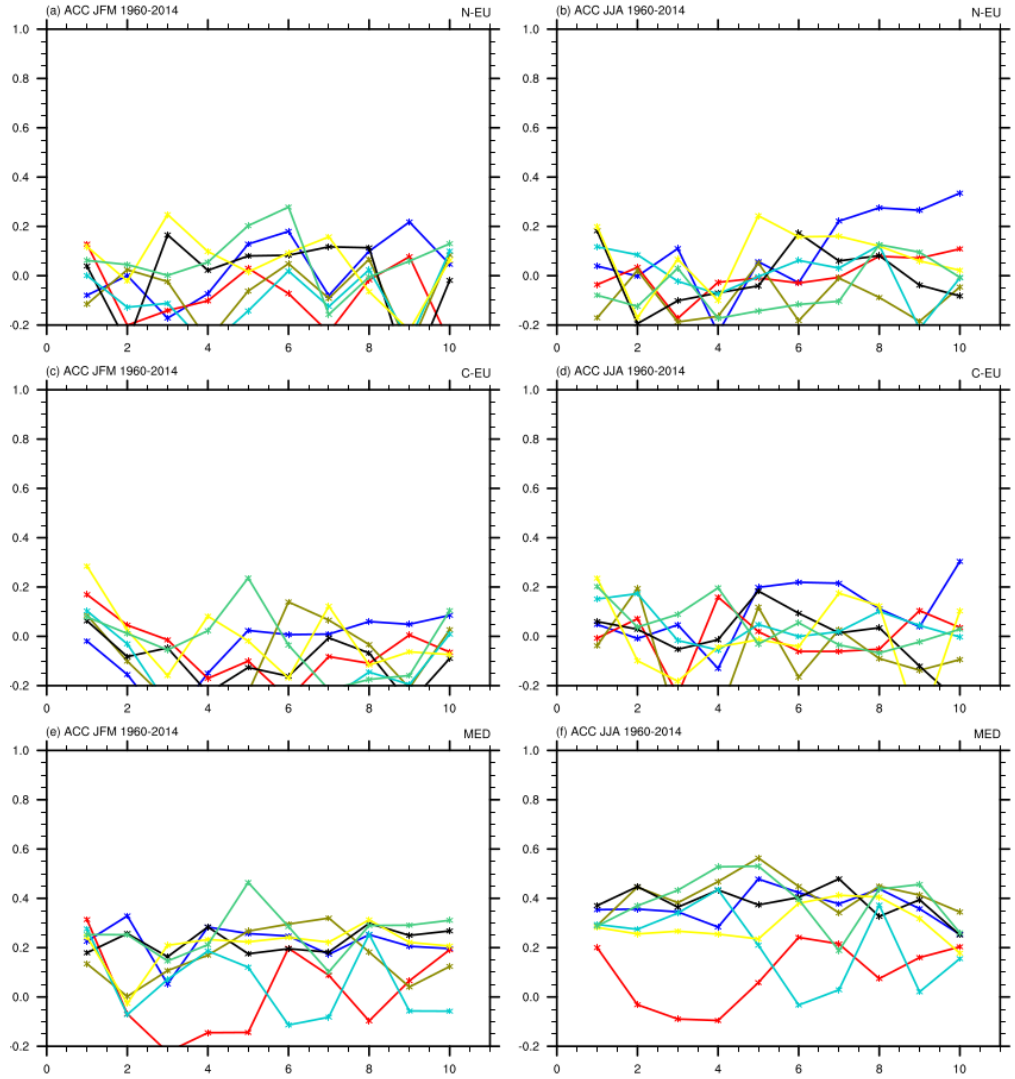

Figure 17 : *ACC of detrended SAT in the SREX regions (a,b) Northern Europe, (c,d) Central Europe, and (e,f) Mediterranean (y-axis) against lead time in years (x-axis). The left column shows skill for winter (JFM), the right column for summer (JJA). Colors indicate the individual models: blue = CanESM5, red = CESM-DPLE, brown = IPSL-CM6A-LR, black = HadGEM3-GC31-MM, cyan = MIROC6, yellow = MPI-ESM1.2-HR, green = NorCPM1.*

Windows of opportunity for North Atlantic SST (AMV: average global SST between -70-70N subtracted from average SST 0-70N in the North Atlantic; Trenberth & Shea, 2006) as well as SAT in the SREX regions across all lead times and all models are presented in figure 18 . This figure enables two assessments: identifying model differences of prediction skill in all regions and across all lead times, and linking windows of high prediction skill found in North Atlantic SST (here interpreted as opportunity for prediction) to those found in European SAT inferring links between SST and SAT predictability (as in Sgubin et al., under review; their figure 5).

We find that there is general agreement among models on both the magnitude and timing of windows of opportunity for AMV and SPG and these windows have been interpreted by Borchert et al (2018) and Borchert et al. (2019a) as a reaction to changes in oceanic heat transport. An exception is SPG SST in the CanESM5 decadal prediction system, which is known to show erratic behaviour in this region due to problems with model initialization (Sospedra-Alfonso, pers. comm). Windows of

opportunity for SREX SAT also show wide-spread coherence between models (fig. 18 c-h), albeit at a much lower level. Summer SAT in the NEU and CEU regions is not predictable in any model and any time window, and we find no prediction skill for winter SAT in the CEU and MED regions.

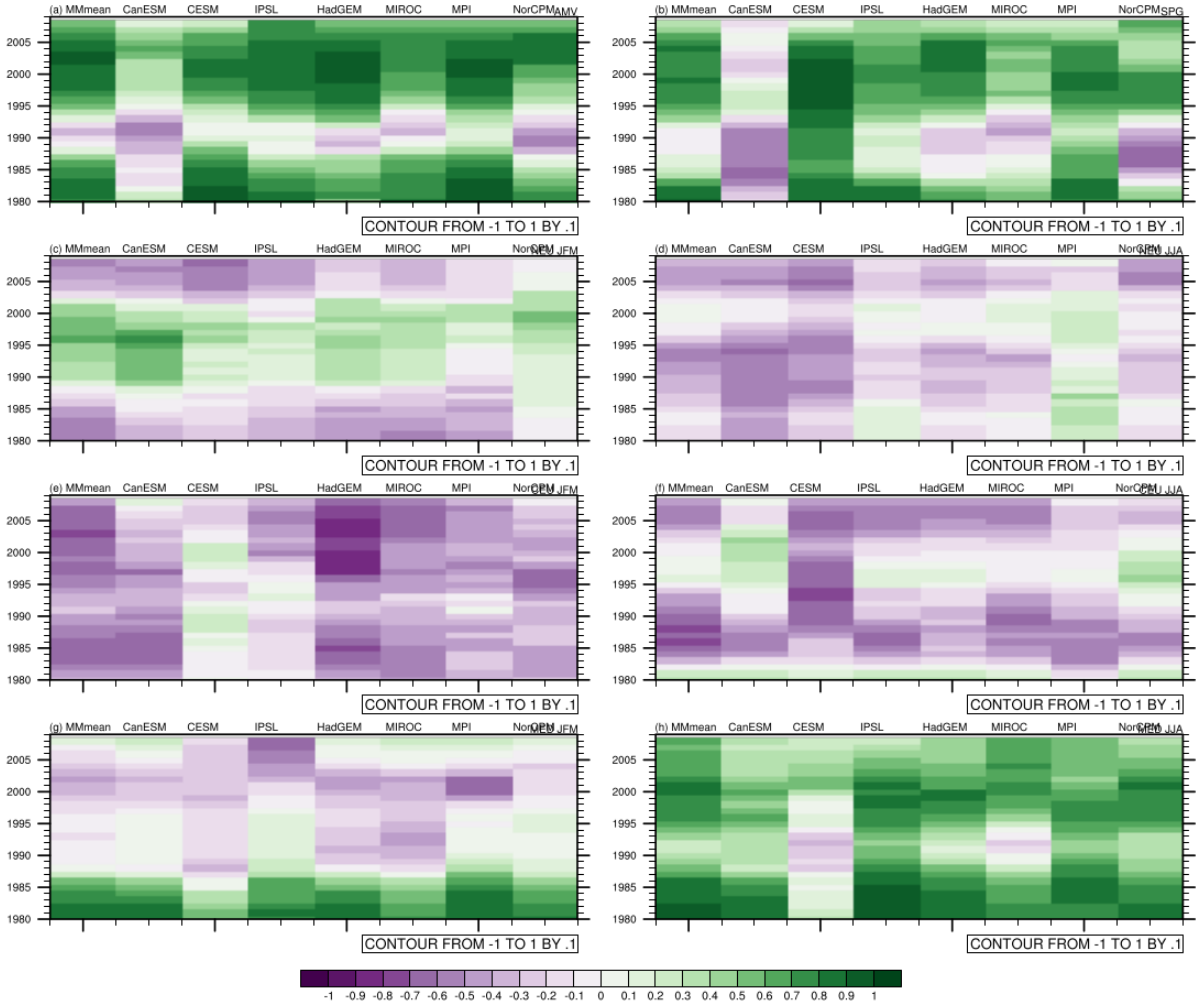

Figure 18 *Comparison of decadal ACC hindcast skill for detrended annual mean North Atlantic SST ((a) AMV and (b) SPG) and detrended winter (JFM) and summer (JJA) mean SAT in the SREX regions ((c,d) NEU, (e,f) CEU, (g,h) MED) for a sliding 20-year long window (the end year of which is shown on the y-axis) across multi-model mean and individual model means (x-axis). Colors indicate ACC (low to high = purple to green). Inspired by Sgubin et al. (under review), their figure 5.*

Interestingly, winter SAT in the NEU region is skillfully predicted in all models around the 1990s, potentially connected to the strong 1990s shift in North Atlantic SST (fig. 18c). All models show high skill in the MED region during summer as well as similar windows of opportunity as North Atlantic SST (fig. 18h), pointing towards a possible connection of North Atlantic SST predictions to decadal predictions of summer SAT in this region. These findings highlight the complexities in predicting European SAT on the decadal time scale, and the need for advancing our knowledge on, among other things, links between predictable North Atlantic SST and unpredictable European SAT to identify models, seasons, time periods and lead time horizons for which predictions can be relied upon. The presented analysis should thus be extended and elaborated on to produce actionable predictions for society.

3.6: Observational constraining subsets of large ensembles (BSC)

This work explores the possibility to constrain large ensembles of transient climate simulations based on their agreement with observed climate anomaly patterns, with the aim to obtain improved information for predicting the climate of the following seasons and years. Sub-selecting ensemble members that more closely resemble the observed climate state (e.g. Ding et al., 2018), aligns the internal climate variability of the sub-selected ensemble with the observed climate variability, similar to initialised climate prediction. We therefore also refer to these constraints relative to the observed anomalies as 'pseudo-initialisation'.

Here we use NCAR's Large Ensemble (LENS; Kay et al. 2015) of historical climate simulations, extended with the RCP8.5 scenario after 2005, and in each year select the 10 ensemble members that most closely resemble the observed state of global SST anomaly patterns. We then evaluate the skill of the constrained and sub-selected ensembles in predicting the observed climate in the following months, years and decade. We also compare the skill of 'un-initialised' (LENS40, ensemble of all 40 LENS simulations) and 'pseudo-initialised' (LENS10, ensemble of the best 10 ensemble members identified in each year) simulations against 'initialised' decadal predictions with NCAR's Decadal Prediction Large Ensemble consisting of 40 initialised ensemble members (DPLE; Yeager et al. 2018).

In this explorative study, the best 10 members of the LENS simulations are selected based on their pattern correlation of global sea surface temperature anomalies with observed (i.e HadISST) anomalies. These pattern correlations are calculated using the average anomalies of the 5 months prior to 1 November of each year, for consistency of the 'pseudo-initialisation' with the initialized prediction system (DPLE40) predictions, which are also initialised on 1 November of each year. We also tested ensemble selection based on the pattern correlation of different time periods (up to 10 years) prior to the 1 November initialisation date, to better phase in low-frequency variability, but these tests did not provide clearly improved skill over the 5-months selection.

Figure 19 compares the skill of different SST indices for the constrained pseudo-initialised ensemble (LENS10), the full LENS40, and the initialised prediction system. All three ensembles show very high skill (R>0.9) in predicting global mean SSTs on inter-annual to decadal time-scales, primarily due to the warming trend. Larger differences in the prediction skill between the three ensembles are apparent for indices of Pacific and Atlantic SST variability. The constrained LENS10 ensemble shows significant skill in predicting the ENSO and IPO indices in the first ~6-7 months after initialisation, with correlations only about ~0.1 lower compared to the initialised DPLE40 ensemble. LENS10 shows improved skill over LENS40 during the first 2 forecast years for ENSO and IPO. For the AMV index, LENS10 shows increased skill over LENS40 for up to 7 forecast years, while DPLE40 shows high skill (R>0.7) for all forecast times up to one decade.

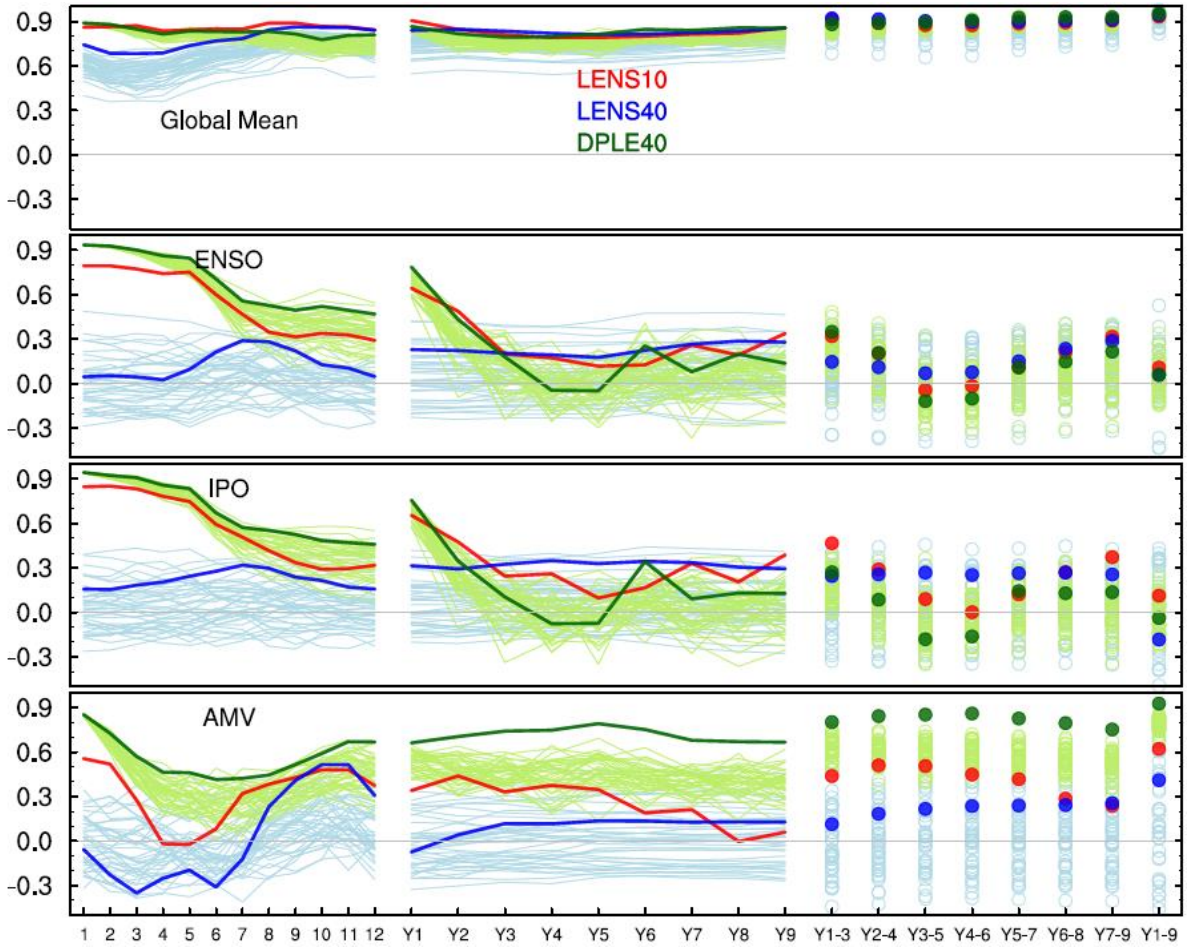

Figure 19: *Correlation skill for different forecast times: (left) lines represent skill for first 12 forecast months; (center) lines represent skill for first 9 forecast years; dots represent skill for multi-annual mean forecasts. Correlations from LENS10 are shown by red lines and dots, LEN40 by blue lines and dots and DPLE by green lines and dots. IPO is calculated as a tripole index (Henley et al., 2015) from SST anomalies, ENSO is based on area-weighted mean of SST anomalies at Nino3.4 region (i.e. 5°S-5°N, 170°W-120°W), and AMV is calculated as a weighted area average SST anomalies for 0-60°N of the North Atlantic ocean with global mean (60°S-60°N) SST removed.*

Figure 20 shows that the spatial distribution of forecast skill of the LENS10 ensemble is often comparable to that of the DPLE40 for seasonal and annual mean forecasts. The skill of LENS40 is relatively lower than both the pseudo-initialized and the initialized predictions at least for the first few forecast months and the first forecast year. On longer time scales, LENS10 has some added skill in the North Atlantic, but decreased skill in other regions such as parts of the Pacific.

These analyses demonstrate the value of constraining large ensembles of climate simulation according to the phases of internal variability for predictions of the real-world climate. We find added value in comparison to the large (un-constrained) ensemble for up to 7 years in the Atlantic, and up to 2 years in indices of Pacific variability. In our ongoing work we will explore the sensitivity to specific selection criteria with the aim to optimise the skill of the constrained ensemble, and extend the selection to make use of initialised decadal predictions for constraining the ensemble to enable predictions beyond one decade (contributing to Task 5.2).

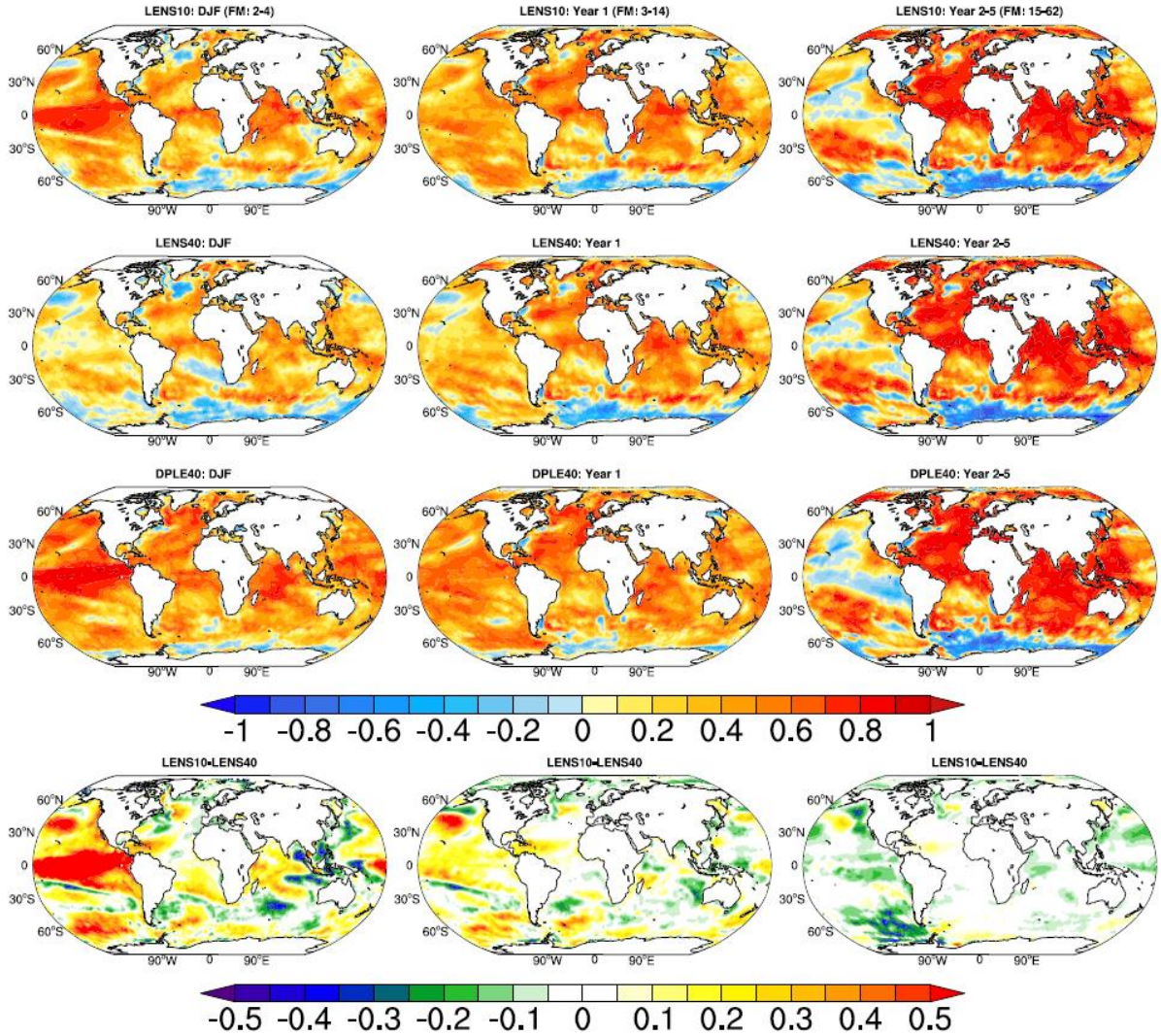

Figure 20: *Forecast skill, measured as anomaly correlation against HadISST, for LENS10 (top row), LENS40 (second row), DPLE40 (third row) and skill difference between LENS10 and LENS40 (bottom row). Left, center and right columns represent forecast skill of the mean of forecast months 2-4, forecast months 3-14 and forecast months 15-62 respectively.*

4. Prospects and considerations for a framework of applying constraints seamlessly across both initialised and non-initialised projections.

With observational constraints, we refer both to weighting schemes that weight according to performance (Brunner et al., 2020b), as well as methods that decide which models are within an observational constraint and which outside (model selection methods; see e.g. discussion in Tokarska et al., 2020; drawing on the ASK method). Projections use skill scores or probabilistic comparisons to observations that also can be used. At present, we are not employing emergent constraints that are based on statistical relationships between specific observable variables and future predictions only (see Hall, 2018). The literature has been fairly sceptical about them unless based on clear physical principles (Hall, 2018; Sherwood et al., 2020). Initialized predictions also use skill scores that are natural for model weighting, and that depend on lead time. The skill in initialized

predictions may also vary strongly with climate state (Borchert et al., 2018; Yeager, 2020; Christensen et al., 2020) and initialization procedure for which there is to date no consensus.

The question arises: if applied to similar model data, will these different metrics for model performance or weights favour similar or different climate model traits and with it similar models for different timescales? This is important for the merged product: if the choice of timescale strongly influences the model weights or leads to selection of different models, the merged predictions might be inconsistent over time by having a discontinuous underlying climate change signal, for example. It remains to be explored how detrimental such a discontinuity might be, as the signal during the initialized time horizon is still small compared to noise on all but global scales. Merging methods that weight initialized and non-initialized predictions differently with forecast horizon (to exclusively use predictions early on and transition to exclusive use of projections after a transition point) may moderate the effect of such hidden discrepancies, yet it would be preferable to not let them arise.

Questions we have discussed and need to further evaluate include:

- To what extent do weighting or model selection criteria used across projections and predictions favour similar model traits and to what extent are they uncorrelated or anticorrelated?
- Do any of the weighting schemes preferentially select or highly weight models with stronger or weaker response, e.g. with high ECS or TCR or with low values (e.g. Table 4)?
- Does the combination of multiple observational constraints, both for projections or across prediction and projections improve the forecast and projection?
- Do any of the weighting schemes reward or penalize climate models with high or low internal climate variability and with that, with low or high signal-to-noise ratio for external forcing or predictable signals? (see figure 16)?
- Are there any other factors where different weighting schemes select differently, such as the response to natural forcings?
- Which dimensions beyond the model dimension (e.g. the temporal and spatial dimensions) show promise in the application of observational constraints?

This will need to be considered when merging initialized and uninitialized simulations in task 5.2, particularly, if using observational constraints.

Conclusions from this deliverable show that:

- Overall, observational constraints have the potential to narrow down current uncertainty in predicted or projected changes, and in some cases can flag the possibility of change outside the model range (e.g., ASK). There is evidence that combining observations from different methodologies can add value and represent a step towards a more robust approach. This is a lesson also learned from UKCP18. In all these approaches, observations are used as a benchmark leading to skill scores and weights across predictions and projections.
- When developing and applying observational constraints, performance of the constraint should be evaluated and compared across the prediction/projection timeline
- Model skill/accuracy can originate from different physical mechanisms, for example from strong climatological performance of models, from the representation of physical

mechanisms in specific models, from realistic mechanisms for internal climate variability, their climate sensitivity, or from trends, or combinations of all.

- When applying multiple constraints in sequence, overfitting should be avoided. When applying only non-overlapping observational constraints by removing trends from ClimWIP, the two different constraints in ClimWIP and ASK, for example, can pull the multi-model mean in different directions. In other cases, notably winter, the use of multiple constraints looks extremely promising and can make the difference between failing to constrain based on the signal and arriving at a constraint after weighting. We are in the process of exploring the reasons for this and robustness of it to evaluate if the improvement in our pilot study is expected to carry through to larger applications.
- In initialized predictions, models can show various behaviours in terms of evolution of skill with prediction horizon. It will be interesting to explore more clearly what controls this evolution. Sequential application of skill weights with future projection constraints (ASK) did not appear to show consistent improvements, possibly due to different factors influencing skill in different applications, regions, and timelines
- Initialized models show different skills that vary with hindcast horizon and over time. These may again pull in different directions from those arising for projections, which may lead to different model families being favoured for different timelines if using weighting. We require carefully designing experiments when aiming to assess or merge initialized and non-initialized simulations together for future climate information, for example with regards to different model versions used for predictions and projections, and careful evaluation of the underlying forced signal and magnitude of variability
- Based on our pilot study here, it is not clear if there is a common strand of behaviour of different observational metrics.
- Sub-selecting data from large ensemble simulations that match the observed conditions prior to a hypothetical hindcast step, provides a good testbed for merging, as skills are fairly comparable between both, particularly early in the decade.

5. Discussion

5.1 Lessons learnt

These considerations manifest in the following lessons learned that will be considered in EUCP merging work across task 5 and beyond. We recommend that:

- Where observational constraints are used, it needs to be clear what main model characteristics explain these constraints and that some perfect/imperfect model tests have been conducted to evaluate reliability of constraints both individually and in sequence (3.4.4, fig. 9., 10). The performance of the constraint over time should also be considered, as some are particularly valuable early and others later, and as some conditions seem to lead to improved hindcast skills compared to other climatological conditions (cf. e.g. section 3.5.3, fig. 18; also 3.4.4, fig. 8).

- Consistent application of the constraint across a merging boundary (see deliverable 5.2) must be ensured, both in terms of the constrained forced signal and internal variability (cf. e.g. section 3.5.1, fig. 13; note also role of climate transient response in 3.4.2, e.g. fig. 5))
- The selective availability of climate models between initialized and non-initialized simulations will be a challenge when merging in time - this challenge also makes the consistent application of observational constraints difficult between those two simulation types (cf. e.g. section 3.4.5, fig. 11). This difficulty needs to be well managed.
- Sub-selection of ensemble members from large ensembles is an excellent testbed for merging exercises 3.6, figures 19, 20).
- Observational constraints should also be considered to support spatial merging. For example, they may help to inform choices where high resolution and coarse resolution models show discrepancies - this might be helpful for task 5.4

5.2 Reflection on the role of this deliverable in WP5 progress

The team on this deliverable draws on D2.1 and D2.2, and has further evaluated methods as well as developed new approaches. It also sets the stage for D5.2 and D5.3. The team has worked closely with WPs 1 and 2 particularly, but there is also scope for a possible future link with WP3 through comparing high resolution projections with the observational constraints discussed here.

The shift of several planned workshops online (e.g. the Oxford workshops on merging, summer 2020) has helped delivery on WP5 despite the pandemic, but has limited interaction between project partners and has not fully enabled the detection of synergies and common interests, and development of creative ideas, that is so much easier done during in-person informal interactions. We have done our best to overcome this difficulty but hope that some more interaction is possible later in the project.

6. Planned future publications

A publication arising from the entire deliverable is planned at a later date, when results are slightly more mature and weighting/observational constraints in initialized simulations are available from more groups. The publications will focus on use of observational constraints in both predictions and projections (the combination of both will make it attractive), opportunities to combine multiple constraints, pitfalls in that, and influence of common metrics such as signal to noise ratio, climate sensitivity/response strength, response to natural forcings and magnitude of natural variability on constraints across the board. Furthermore, a short introductory publication on the use of observations as model evaluation and weighting tool is planned for an invited contribution to Frontiers in Climate edited by Matt Collins, with authors here, however, focusing on methodological considerations rather than results, as specialized papers first need to be published.

Several publications, many reaching across project partners, are planned based on work reported in this deliverable, including:

Hegerl, Ballinger, Borchert, Brunner, Donat, Mignot (2021): Use of observational constraints from predictions to projections: opportunities and challenges. Submission invited for special issue in frontiers in research.

Ballinger, Brunner, Schurer, Undorf, Hegerl, 2021: Application of ASK constraint on European regions, in prep.

Borchert, L.F., M. Payo, M.B. Menary, D. Swingedouw, G. Sgubin, J. Mignot: Contrasting hindcast skill over Europe in different CMIP6 decadal prediction systems, in prep.

Sgubin, G., D. Swingedouw, L.F. Borchert, M.B. Menary, T. Noël, H. Loukos, J. Mignot: A systematic investigation of the skill in air temperature prediction over Europe for potential applications to climate services. Under review.

7. References

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.

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

Annan, J.D., and J.C. Hargreaves (2011). On the generation and interpretation of probabilistic estimates of climate sensitivity. Climatic Change 104, 423–436 . https://doi.org/10.1007/s10584-009-9715-y

Bethke, I., Y. Wang, F. Counillon, M. Kimmritz, F. Fransner, A. Samuelsen, et al. (under review). NorCPM and its contribution to CMIP6 DCPP.

Boer, G. J., Smith, D. M., Cassou, C., Doblas-Reyes, F., Danabasoglu, G., Kirtman, B., Kushnir, Y., Kimoto, M., Meehl, G. A., Msadek, R., Müller, 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., W. A. Müller, and J. Baehr (2018): Atlantic Ocean Heat Transport Influences Interannual-to-Decadal Surface Temperature Predictability in the North Atlantic Region. J. Climate, 31, 6763–6782, https://doi.org/10.1175/JCLI-D-17-0734.1.

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., M.B. Menary, D. Swingedouw, G. Sgubin, L. Hermanson, J. Mignot (2020). Improved decadal predictions of North Atlantic subpolar gyre SST in CMIP6. in press at Geophysical Research Letters.

Booth BBB, Harris GR, Murphy JM, House JI, Jones CD, Sexton DMH, Sitch S (2017), Narrowing the range of future climate projections using historical observations of atmospheric CO2. J. Clim. 30, 3039-3053, https://doi.org/10.1175/jcli-d-16-0178.1

Boucher, O., Servonnat, J., Albright, A. L., Aumont, O., Balkanski, Y., & Bastrikov, V., et al. (2020). Presentation and evaluation of the IPSL-CM6A-LR climate model. Journal of Advances in Modeling Earth Systems, 12, e2019MS002010. https://doi.org/10.1029/2019MS002010

Brune, S., Düsterhus, A., Pohlmann, H. *et al.* Time dependency of the prediction skill for the North Atlantic subpolar gyre in initialized decadal hindcasts. *Clim Dyn* **51,** 1947–1970 (2018). https://doi.org/10.1007/s00382-017-3991-4

Brune, S, Baehr, J. (2020).Preserving the coupled atmosphere–ocean feedback in initializations of decadal climate predictions. WIREs Clim Change. 11:e637. https://doi.org/10.1002/wcc.637

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

Brunner, L., 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. (2020a). 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

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., and Knutti, R.: Reduced global warming from CMIP6 projections when weighting models by performance and independence, Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2020-23, in press, 2020b.

Christensen, H. M., J. Berner, and S. Yeager, 2020: The Value of Initialization on Decadal Timescales: State-Dependent Predictability in the CESM Decadal Prediction Large Ensemble. J. Climate, 33, 7353–7370, https://doi.org/10.1175/JCLI-D-19-0571.1.

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., et al. (2020). The Community Earth System Model Version 2 (CESM2). Journal of Advances in Modeling Earth Systems, 12, e2019MS001916. https://doi.org/10.1029/2019MS001916

DelSole, T., Trenary, L., Yan, X. et al. (2019). Confidence intervals in optimal fingerprinting. Clim Dyn 52, 4111–4126. https://doi.org/10.1007/s00382-018-4356-3

Ding, H., Newman, M., Alexander, M. A., and Wittenberg, A. T. (2018). Skillful climate forecasts of the tropical Indo-Pacific Ocean using model-analogs. Journal of Climate, 31(14), 5437–5459. https://doi.org/10.1175/JCLI-D-17-0661.1.

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

Dufresne, J., Foujols, M., Denvil, S. et al. (2013). Climate change projections using the IPSL-CM5 Earth System Model: from CMIP3 to CMIP5. Clim Dyn 40, 2123–2165. https://doi.org/10.1007/s00382-012-1636-1

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Fereday, D., O'Reilly, C., et al. (2018). Skilful seasonal predictions of summer European rainfall. Geophysical Research Letters, 45, 3246– 3254. https://doi.org/10.1002/2017GL076337

Estella-Perez, V., Mignot, J., Guilyardi, E. et al. Advances in reconstructing the AMOC using sea surface observations of salinity. Clim Dyn 55, 975–992 (2020). https://doi.org/10.1007/s00382-020-05304-4

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

Forster, P. M., Maycock, A. C., McKenna, C. M., and Smith, C. J.: Latest climate models confirm need for urgent mitigation, Nature Climate
Change, 10, 7–10, https://doi.org/10.1038/s41558-019-0660-0, 2020.

Gastineau, G., and C. Frankignoul, 2015: Influence of the North Atlantic SST Variability on the Atmospheric Circulation during the Twentieth Century. *J. Climate*, **28**, 1396–1416, https://doi.org/10.1175/JCLI-D-14-00424.1.

Gidden M, Riahi K, Smith S, Fujimori S, Luderer G, Kriegler E, van Vuuren DP, van den Berg M, et al. (2019). Global emissions pathways under different socioeconomic scenarios for use in CMIP6: a dataset of harmonized emissions trajectories through the end of the century. Geoscientific Model Development Discussions 12 (4): 1443-1475. DOI:10.5194/gmd-2018-266.

Gillett, N. P., Shiogama, H., Funke, B., Hegerl, G., Knutti, R., Matthes, K., Santer, B. D., Stone, D., and Tebaldi, C. (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

Giorgetta, M. A., et al. (2013), Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, J. Adv. Model. Earth Syst., 5, 572– 597, doi:10.1002/jame.20038.

Hall, A., Cox, P., Huntingford, C. *et al.* Progressing emergent constraints on future climate change. *Nat. Clim. Chang.* **9,** 269–278 (2019). https://doi.org/10.1038/s41558-019-0436-6

Hanlon, H. M., Morak, S., and Hegerl, G. C. (2013), Detection and prediction of mean and extreme European summer temperatures with a multimodel ensemble, J. Geophys. Res. Atmos., 118, 9631–9641, doi:10.1002/jgrd.50703.

Harris GR, Sexton DMH, Booth BBB, Collins M, Murphy JM (2013), Probabilistic Projections of Transient Climate Change, Clim Dyn 40: 2937. https://doi.org/10.1007/s00382-012-1647-y

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.

Henley, B.J., Gergis, J., Karoly, D.J. et al. A Tripole Index for the Interdecadal Pacific Oscillation. Clim. Dyn., 45, 3077–3090 (2015). https://doi.org/10.1007/s00382-015-2525-1

Hermanson, L., Bilbao, R., Dunstone, N., Ménégoz, M., Ortega, P., Pohlmann, H., et al. (2020). Robust multiyear climate impacts of volcanic eruptions in decadal prediction systems. Journal of Geophysical Research: Atmospheres, 125, e2019JD031739. https://doi.org/10.1029/2019JD031739

Ingleby, B., and M. Huddleston, 2007: Quality control of ocean temperature and salinity profiles - historical and realtime data, J. Mar. Syst., 65, 158-175

Jolliffe, I.T. and Stephenson, D.B. (2003) Forecast Verification. A Practitioner's Guide in Atmospheric Science. John Wiley & Sons Ltd., Hoboken, 240 p.

Kawai, H., Yukimoto, S., Koshiro, T., Oshima, N., Tanaka, T., Yoshimura, H., and Nagasawa, R. (2019). Significant improvement of cloud representation in the global climate model MRI-ESM2, Geosci. Model Dev., 12, 2875–2897, https://doi.org/10.5194/gmd-12-2875-2019

Kay, J. E., and Coauthors, 2015: The Community Earth System Model (CESM) Large Ensemble project: A community resource for studying climate change in the presence of internal climate variability. Bull. Amer. Meteor. Soc., 96, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1.

Kelley, M., Schmidt, G. A., Nazarenko, L. S., Bauer, S. E., Ruedy, R., & Russell, G. L., et al. (2020). GISS-E2.1: Configurations and climatology. Journal of Advances in Modeling Earth Systems, 12, e2019MS002025. https://doi.org/10.1029/2019MS002025

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.

Knutti R., Masson D & Gettelman, A. (2013): Climate model genealogy: CMIP5 and how we got there. GRL, 40, 1194-1199.

Knutti, R., Sedláček, J., Sanderson, B. M., Lorenz, R., Fischer, E. M., & Eyring, V. (2017). A climate model projection weighting scheme accounting for performance and interdependence. Geophysical Research Letters, 44(4), 1909–1918. https://doi.org/10.1002/2016GL072012

Kuhlbrodt, T., Jones, C. G., Sellar, A., Storkey, D., Blockley, E., Stringer, M., et al. (2018). The low-resolution version of HadGEM3 GC3.1: Development and evaluation for global climate. Journal of Advances in Modeling Earth Systems, 10, 2865– 2888. https://doi.org/10.1029/2018MS001370

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., & Knutti, R. (2018). Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America. Journal of Geophysical Research: Atmospheres, 123(9), 4509–4526. https://doi.org/10.1029/2017JD027992

Mariotti, A., C. Baggett, E.A. Barnes, E. Becker, A. Butler, D.C. Collins, P.A. Dirmeyer, L. Ferranti, N.C. Johnson, J. Jones, B.P. Kirtman, A.L. Lang, A. Molod, M. Newman, A.W. Robertson, S. Schubert, D.E. Waliser, and J. Albers, (2020). Windows of Opportunity for Skillful Forecasts Subseasonal to Seasonal and Beyond. Bull. Amer. Meteor. Soc., https://doi.org/10.1175/BAMS-D-18-0326.1

Menary, M. B., Hodson, D. L. R., Robson, J. I., Sutton, R. T., Wood, R. A., and Hunt, J. A. (2015), Exploring the impact of CMIP5 model biases on the simulation of North Atlantic decadal variability, Geophys. Res. Lett., 42, 5926– 5934, doi:10.1002/2015GL064360.

Menary, M.B., Hermanson, L. Limits on determining the skill of North Atlantic Ocean decadal predictions. Nat Commun 9, 1694 (2018). https://doi.org/10.1038/s41467-018-04043-9

Merryfield, W.J., W. Lee, G.J. Boer, V.V. Kharin, J.F. Scinocca, G.M. Flato, R.S. Ajayamohan, J.C. Fyfe, Y. Tang, and S. Polavarapu (2013). The Canadian Seasonal to Interannual Prediction System. Part I: Models and Initialization. Mon. Wea. Rev., 141, 2910–2945, https://doi.org/10.1175/MWR-D-12-00216.1

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D. (2012), Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set, J. Geophys. Res., 117, D08101, doi:10.1029/2011JD017187.

Müller, W. A., Jungclaus, J. H., Mauritsen, T., Baehr, J., Bittner, M., Budich, R., et al. (2018). A higher-resolution version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). Journal of Advances in Modeling Earth Systems, 10, 1383– 1413. https://doi.org/10.1029/2017MS001217

Murphy, J.M., Sexton, D.M.H., Jenkins, G.J., Boorman, P.M., Booth, B.B.B., Brown, C.C., Clark, R.T., Collins, M., Harris, G.R., Kendon, E.J., Betts, R.A., Brown, S.J., Howard T.P., Humphrey, K.A., McCarthy, M.P., McDonald, R.E., Stephens, A., Wallace, C., Warren, R., Wilby, R., Wood, R.A. (2009). UK Climate Projections Science Report: Climate change projections. Met Office Hadley Centre, Exeter, U.K., https://webarchive.nationalarchives.gov.uk/20181204111026/http://ukclimateprojections-ukcp09.metoffice.gov.uk/22530#projections

Murphy, J.M., Harris, G.R., Sexton, D.M.H., Kendon, E.J., Bett, P.E., Clark, R.T., Eagle, K.E., Fosser, G., Fung, F., Lowe, J.A., McDonald, R.E., McInnes, R.N., McSweeney, C.F., Mitchell, J.F.B., Rostron, J.W., Thornton, H.E., Tucker, S., Yamazaki, K. (2018) UKCP18 Land Projections: Science Report, Met Office Hadley Centre, Exeter, U.K., https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/UKCP18-Land-report.pdf

Pohlmann, H., Botzet, M., Latif, M., Roesch, A., Wild, M., & Tschuck, P. (2005). Estimating the decadal predictability of a coupled AOGCM. Journal of Climate, 17(22), 4463-4472.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century, J. Geophys. Res., 108, 4407, doi:10.1029/2002JD002670, D14

Seland, Ø., Bentsen, M., Seland Graff, L., Olivié, D., Toniazzo, T., Gjermundsen, A., Debernard, J. B., Gupta, A. K., He, Y., Kirkevåg, A., Schwinger, J., Tjiputra, J., Schancke Aas, K., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Hafsahl Karset, I. H., Landgren, O., Liakka, J., Onsum Moseid, K., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iverson, T., and Schulz, M.: The Norwegian Earth System Model, NorESM2 – Evaluation of theCMIP6 DECK and historical simulations, Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2019-378, in review, 2020.

Sexton DMH, Murphy JM, Collins M, Webb MJ (2012), Multivariate probabilistic projections using imperfect climate models, Part I: Outline of methodology, Clim Dyn, 38: 2513, https://doi.org/10.1007/s00382 011 1208 9

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

Sherwood, S. C., Webb, M. J., Annan, J. D., Armour, K. C., Forster, P. M., Hargreaves, J. C., et al. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, 58, e2019RG000678. https://doi.org/10.1029/2019RG000678

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

Smith, D., Eade, R. & Pohlmann, H. (2013). A comparison of full-field and anomaly initialization for seasonal to decadal climate prediction. Clim. Dyn. 41, 3325–3338

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

Sospedra-Alfonso, R., & Boer, G. J. (2020). Assessing the impact of initialization on decadal prediction skill. Geophysical Research Letters, 47, e2019GL086361. https://doi.org/10.1029/2019GL086361

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.

Swart, N. C., Cole, J. N. S., Kharin, V. V., Lazare, M., Scinocca, J. F., Gillett, N. P., Anstey, J., Arora, V., Christian, J. R., Hanna, S., Jiao, Y., Lee, W. G., Majaess, F., Saenko, O. A., Seiler, C., Seinen, C., Shao, A., Sigmond, M., Solheim, L., von Salzen, K., Yang, D., and Winter, B. (2019). The Canadian Earth System Model version 5 (CanESM5.0.3), Geosci. Model Dev., 12, 4823–4873, https://doi.org/10.5194/gmd-12-4823-2019

Tatebe, H., Ogura, T., Nitta, T., Komuro, Y., Ogochi, K., Takemura, T., Sudo, K., Sekiguchi, M., Abe, M., Saito, F., Chikira, M., Watanabe, S., Mori, M., Hirota, N., Kawatani, Y., Mochizuki, T., Yoshimura, K., Takata, K., O'ishi, R., Yamazaki, D., Suzuki, T., Kurogi, M., Kataoka, T., Watanabe, M., and Kimoto, M. (2019). Description and basic evaluation of simulated mean state, internal variability, and climate sensitivity in MIROC6, Geosci. Model Dev., 12, 2727–2765, https://doi.org/10.5194/gmd-12-2727-2019

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

Trenberth, K. E., and Shea, D. J. (2006), Atlantic hurricanes and natural variability in 2005, Geophys. Res. Lett., 33, L12704, doi:10.1029/2006GL026894.

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 DECK experiments with CNRM-CM6-1. Journal of Advances in Modeling Earth Systems, 11, 2177– 2213. https://doi.org/10.1029/2019MS001683

Watanabe, M., T. Suzuki, R. O'ishi, Y. Komuro, S. Watanabe, S. Emori, T. Takemura, M. Chikira, T. Ogura, M. Sekiguchi, K. Takata, D. Yamazaki, T. Yokohata, T. Nozawa, H. Hasumi, H. Tatebe, and M. Kimoto (2010). Improved Climate Simulation by MIROC5: Mean States, Variability, and Climate Sensitivity. J. Climate, 23, 6312–6335, https://doi.org/10.1175/2010JCLI3679.1

Wu, T., Song, L., Li, W. et al. (2014). An overview of BCC climate system model development and application for climate change studies. Acta Meteorol Sin 28, 34–56. https://doi.org/10.1007/s13351-014-3041-7

Wu, T., Lu, Y., Fang, Y., Xin, X., Li, L., Li, W., Jie, W., Zhang, J., Liu, Y., Zhang, L., Zhang, F., Zhang, Y., Wu, F., Li, J., Chu, M., Wang, Z., Shi, X., Liu, X., Wei, M., Huang, A., Zhang, Y., and Liu, X. (2019). The

Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6 , Geosci. Model Dev., 12, 1573–1600, https://doi.org/10.5194/gmd-12-1573-2019

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

Yeager S. G., G. Danabasoglu, N. A. Rosenbloom, W. Strand, S. C. Bates, G. A. Meehl, A. R. Karspeck, K. Lindsay, M. C. Long, H. Teng, and N. S. Lovenduski, Predicting Near-Term Changes in the Earth System (2018): A Large Ensemble of Initialized Decadal Prediction Simulations Using the Community Earth System Model, Bull. Amer. Meteor. Soc., 99, 1867-1886. DOI:10.1175/BAMS-D-17-0098.2.

Yeager, S. The abyssal origins of North Atlantic decadal predictability. *Clim Dyn* **55,** 2253–2271 (2020). https://doi.org/10.1007/s00382-020-05382-4