https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics
Discussions
Open Access
EGU

A weighting scheme to incorporate large ensembles in multi-model ensemble projections

Anna L. Merrifield[1], Lukas Brunner[1], Ruth Lorenz[1], and Reto Knutti[1]
[1]Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland
Correspondence: Anna L. Merrifield (anna.merrifield@env.ethz.ch)

Abstract. Multi-model ensembles can be used to estimate uncertainty in projections of regional climate, but this uncertainty often depends on the constituents of the ensemble. The dependence of uncertainty on ensemble composition is clear when single model initial condition large ensembles (SMILEs) are included within a multi-model ensemble. SMILEs introduce "new" information into a multi-model ensemble by representing region-scale internal variability, but also introduce redundant information, by virtue of a single model being represented by 50-100 outcomes. To preserve the contribution of internal variability and ensure redundancy does not overwhelm uncertainty estimates, a weighting approach is used to incorporate 50-members of the Community Earth System Model (CESM1.2.2), 50-members of the Canadian Earth System Model (CanESM2), and 100-members of the MPI Grand Ensemble (MPI-GE) into an 88-member Coupled Model Intercomparison Project Phase 5 (CMIP5) multi-model ensemble. The weight assigned to each multi-model ensemble member is based on the member's ability to reproduce observed climate (performance) and scaled by a measure of redundancy (dependence). Surface air temperature (SAT) and sea level pressure (SLP) diagnostics are used to determine the weights, and relationships between present and future diagnostic behavior are discussed. A new diagnostic, estimated forced trend, is proposed to replace a diagnostic with no clear emergent relationship, 50-year regional SAT trend.

The influence of the weighting is assessed in estimates of Northern European winter and Mediterranean summer end-of-century warming in the CMIP5 and combined SMILE-CMIP5 multi-model ensembles. The weighting is shown to recover uncertainty obscured by SMILE redundancy, notably in Mediterranean summer. For each SMILE, the independence weight of each ensemble member as a function of the number of SMILE members included in the CMIP5 ensemble is assessed. The independence weight increases linearly with added members with a slope that depends on SMILE, region, and season. Finally, it is shown that the weighting method can be used to guide SMILE member selection if a subsetted ensemble with one member per model is sought. The weight a SMILE receives within a subsetted ensemble depends on which member is used to represent it, reinforcing the advantage of weighting and incorporating all initial condition ensemble members in multi-model ensembles.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics Discussions
Open Access
EGU

1 Introduction

Projections of regional climate change are both key to climate adaptation policy and fundamentally uncertain due to the nature of the climate system (Deser et al., 2012; Kunreuther et al., 2013). In order to represent regional climate uncertainty to policy-makers, scientists often turn to multi-model ensembles to provide a range of plausible outcomes a region may experience (Tebaldi and Knutti, 2007). Uncertainty in a multi-model ensemble is commonly estimated from the ensemble spread, which can be represented e.g., as the 5-95% likely range of the distribution and is usually presented with respect to the arithmetic ensemble mean (e.g. Collins et al., 2013). This representation of uncertainty appears unambiguous, but is perhaps deceptively so. It is influenced by choices made in multi-model ensemble construction, choices that are often overlooked (Knutti et al., 2010a, b).

Multi-model ensembles, such as those constructed from Coupled Model Intercomparison Projects or CMIPs (Meehl et al., 2000), tend to be comprised of both different models and multiple members of the same model, subject to the same radiative forcing pathway intended to reflect plausible future emissions scenario (van Vuuren et al., 2011; O'Neill et al., 2014). This choice allows the multi-model ensemble to represent two types of regional-scale uncertainty: model uncertainty and internal variability (e.g. Hawkins and Sutton, 2009; Deser et al., 2012).

Model uncertainty accounts for differences in how models parameterize processes in the climate system that are not otherwise captured on the spatial and temporal resolution of global climate models. Subgrid scale processes are often the product of complex interactions and feedbacks between the land surface, ocean, cryosphere, and atmosphere, many of which can not be directly measured (e.g. Seneviratne et al., 2010; Deser et al., 2007). How models estimate these interactions can result in various advantages and limitations in how climate in different regions is represented, and thus affect regional uncertainty estimates. By considering differences in regional "performance", it becomes clear that uncertainty is affected by the assumption that each member of a multi-model ensemble is an equally plausible representation of observed climate. Known biases associated with cloud processes, land-atmosphere interactions, and sea surface temperature (e.g. Boberg and Christensen, 2012; Li and Xie, 2012; Pithan et al., 2014; Merrifield and Xie, 2016) may introduce more-than-representative uncertainty into projections of future climate. Using expert judgement to weight or select multi-model ensemble members based on process- or region-specific metrics of performance has been shown to justifiably constrain uncertainty (e.g. Abramowitz et al., 2008; Knutti et al., 2017; Lorenz et al., 2018).

The second type of uncertainty, internal variability, reflects the regional influence of the amalgamation of unpredictable fluctuations in the climate system (Deser et al., 2012; Knutti and Sedláček, 2013). Internal variability is ostensibly a feature of the climate system and therefore is sometimes referred to as irreducible (Hawkins and Sutton, 2009). However, the influence of internal variability on climate variables such as surface air temperature (SAT) can be quantified and accounted for in projections of future climate using dynamical adjustment methods (e.g. Deser et al., 2016; Sippel et al., 2019). Additionally, internal variability can be explicitly represented by sets of simulations from the same model, subject to identical forcing, in which members differ only by initial conditions (e.g. Kay et al., 2015; Maher et al., 2019). These single member initial condition

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions

large ensembles or SMILEs have become an indispensable tool to concisely represent uncertainty within a model, information that should be considered in a multi-model ensemble context (Rondeau-Genesse and Braun, 2019).

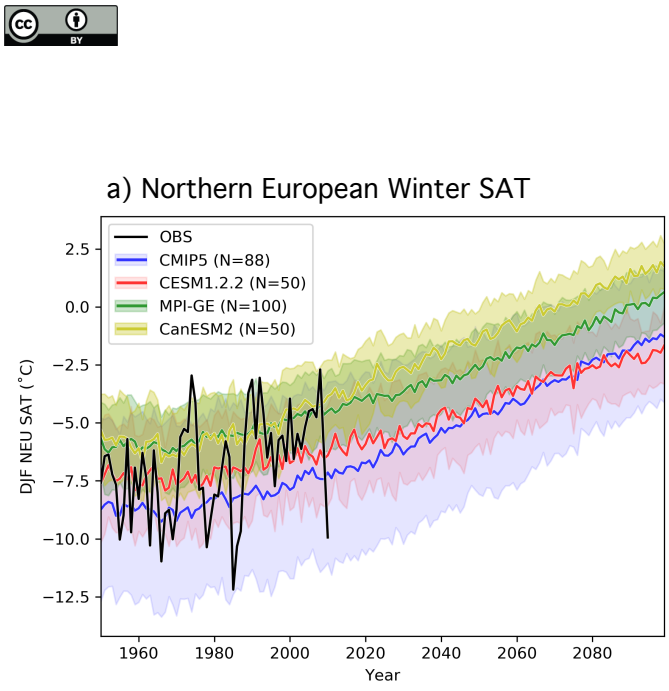
The prospect of including SMILE members into a multi-model ensemble highlights another tacit assumption made during multi-model ensemble construction: each member is an independent representation of climate. Though all members of a multi-model ensemble describe the same climate system, differences in performance tend to create a distribution of regional climate change estimates. If projections are too similar, the redundant information narrows this distribution and reduces uncertainty (Herger et al., 2018). There are several possible reasons for redundancy within a multi-model ensemble, first, different models can have similar biases with respect to observations. Models have historically shared code, from parametrization schemes to full components, and tend to have the same limitations associated with resolution (i.e., simplified topography) (Masson and Knutti, 2011; Knutti et al., 2013; Boé, 2018). Another contributor to redundancy is multiple initial condition ensemble members that project the same outcome, a situation made more likely with the 50 to 100 members of a SMILE. It therefore becomes important when incorporating SMILEs into a multi-model ensemble that uncertainty estimates reflect effective degrees of freedom in the ensemble (Pennell and Reichler, 2011). This can be achieved by down-weighting redundant information by a measure of independence (Abramowitz et al., 2019).

In this study, we evaluate if a performance and independence weighting scheme (Knutti et al., 2017; Lorenz et al., 2018; Brunner et al., 2019) can be used to include three SMILEs into a CMIP5 multi-model ensemble and provide a justifiably constrained estimate of European regional end-of-century warming uncertainty. Northern European winter and Mediterranean summer SAT changes between the 1990-2009 and 2080-2099 mean state are considered. We discuss details of the weighting method including emergent predictor relationships and optimal parameter choices for including "new" information associated with internal variability and mitigating the distributional constraint associated with redundancy. We highlight a new metric, estimated forced trend, which can be used as an alternative to trend-based metrics that are shown to not optimally reflect a model's performance on regional scales. We compare how the weighting shifts the CMIP5 distribution with and without the SMILEs included and explicitly compute independence weight as a function of the number of members for each SMILE. Finally, we use the weighting to demonstrate how subsetting, the practice of selecting one member from each model as a means of ensuring independence, affects a SMILE's representation within an ensemble. The SMILEs, CMIP5, and observational datasets used in the weighting are described in Section 2, while the weighting is detailed in Section 3. The influence of the weighting, the SMILE-specific independence weight evolution, and the drawbacks of representing a SMILE with a single member are discussed in Section 4. To close, conclusions and discussion is presented in Section 5.

2 Data

The multi-model ensemble used in this study is comprised of members from the CMIP5 ensemble and three SMILEs; each ensemble is shown in terms of their ensemble mean and spread (± 1 standard deviation) for Northern European (NEU) winter (December-January-February; DJF) SAT (Figure 1a) and Mediterranean (MED) summer (June-July-August; JJA) SAT (Fig.1b). The NEU and MED regions used are the SREX regions defined in Seneviratne (2012). All models are forced with

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019



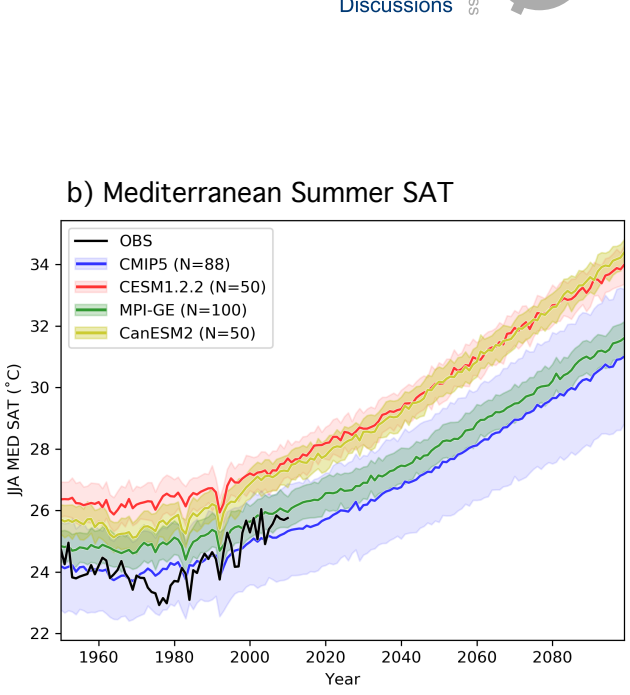


Figure 1. Observational estimate or OBS (ERA-20C; black) and the ensemble mean and spread ($\pm 1\sigma$) of a) DJF NEU and b) JJA MED SAT (°C) for the CMIP5 ensemble (blue), the CESM1.2.2 large ensemble (red), CanESM2 large ensemble (yellow), and the MPI grand ensemble (green). The ensemble mean of each ensemble is shown by a solid line and the ensemble spread is shown by the shading. The number of members, N, in each ensemble is indicated in the legend.

historical CMIP5 forcing from 1950-2005 followed by Representative Concentration Pathway 8.5 (RCP8.5) forcing from 2006–2099 (Meinshausen et al., 2011).

A global atmospheric reanalysis product, ERA-20C, is used to represent observed climate (Fig.1 black). Created by European Centre for Medium-Range Weather Forecasts (ECMWF), ERA-20C assimilates surface pressure and marine wind observations over the 20th century (1900-2010) into the IFS version Cy38r1 model (Poli et al., 2016). While the weighting can be based on several observational estimates to account for observational uncertainty, we chose to use a single observational estimate in order to have a simple and straight-forward definition of climate within which the sensitivity of the weighting scheme can be interrogated. ERA-20C reanalysis was chosen because it provides temporally and spatially complete SAT and SLP fields that extend back to 1950. Additionally, as reanalysis products are, after all, model-based, we selected a reanalysis product with both SLP and SAT available to ensure consistency in the relationship between the two fields. This was necessary because the SLP-SAT relationship is used to obtain the circulation-induced component of SAT, which is removed to obtain the estimated forced SAT trends (see Appendix A). Though ERA-20C is a reanalysis product, we henceforth refer to it as "observations" or "OBS" to distinguish it from members of the multi-model ensemble.

The basis multi-model ensemble to which SMILE members are added comes from the CMIP5 archive (Fig.1 blue). Members used are listed in Table 1. In total, 88 members from 40 model setups are used, including 13 initial condition ensembles ranging from 3 to 10 members. A similar CMIP5 multi-model ensemble has been used in Lorenz et al. (2018) and Brunner et al. (2019).

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics Discussions

Table 1. Summary of the CMIP5 Multi-Model Ensemble used in this study.

Model	Members Used	Model	Members Used
ACCESS1-0	r1i1p1	GISS-E2-H-CC	r1i1p1
ACCESS1-3	r1i1p1	GISS-E2-H	r(1-2)i1p1,r1i1p2,r(1-2)i1p3
bcc-csm1-1-m	r1i1p1	GISS-E2-R-CC	r1i1p1
bcc-csm1-1	r1i1p1	GISS-E2-R	r(1-2)i1p1,r1i1p2,r(1-2)i1p3
BNU-ESM	r1i1p1	HadGEM2-AO	r1i1p1
CCSM4	r(1-6)i1p1	HadGEM2-CC	r1i1p1
CESM1-BGC	r1i1p1	HadGEM2-ES	r(1-4)i1p1
CESM1-CAM5	r(1-3)i1p1	inmcm4	r1i1p1
CMCC-CESM	r1i1p1	IPSL-CM5A-LR	r(1-3)i1p1
CMCC-CMS	r1i1p1	IPSL-CM5A-MR	r1i1p1
CMCC-CM	r1i1p1	IPSL-CM5B-LR	r1i1p1
CNRM-CM5	r(1,2,4,6,10)i1p1	MIROC-ESM	r1i1p1
CSIRO-Mk3-6-0	r(1-10)i1p1	MIROC-ESM-CHEM	r1i1p1
CanESM2	r(1-5)i1p1	MIROC5	r(1-3)i1p1
EC-EARTH	r(1,2,8,9,12)i1p1	MPI-ESM-LR	r(1-3)i1p1
FGOALS-g2	r1i1p1	MPI-ESM-MR	r(1-3)i1p1
FIO-ESM	r(1-3)i1p1	MRI-CGCM3	r1i1p1
GFDL-CM3	r1i1p1	MRI-ESM1	r1i1p1
GFDL-ESM2G	r1i1p1	NorESM1-M	r1i1p1
GFDL-ESM2M	r1i1p1	NorESM1-ME	r1i1p1
Total			**88 members**

Three SMILEs are incorperated into the CMIP5-based multi-model ensemble: a 50-member ensemble generated using the Community Earth System Model version 1.2.2 (CESM1.2.2; Fig.1 red), the 50-member Canadian Earth System Model version 2 (CanESM2) large ensemble (Fig.1 yellow), and the 100-member Max Planck Institute for Meteorology Grand Ensemble (MPI-GE; Fig.1 green). Summarized in Table 2 and described in further detail below, the three SMILEs present a representative distribution of internal variability for each model in the two European regions and seasons considered.

The CESM1.2.2 large ensemble used in this study was derived from a 4700-yr CESM control simulation with constant preindustrial forcing generated at ETH Zürich (Sippel et al., 2019). CESM1.2.2 uses the Community Atmosphere Model, version 5.3 (CAM5.3) and has a horizontal atmospheric resolution of $1.9° \times 2.5°$ with 30 vertical levels (Hurrell et al., 2013). The preindustrial control run was branched at 20-year intervals, starting from the year 580, to create an ensemble with macro initial conditions, i.e., different coupled initial conditions picked from well separated start dates (Stainforth et al., 2007; Hawkins

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics Discussions

Table 2. Summary of the SMILEs used in this study.

Model	Members Used
CESM1.2.2	r(0-49)i1p1
MPI-GE	r(1-100)i1p3
CanESM2 historical-r(1-5)	r(1-10)i1p1
Total	**200 members**

et al., 2016). Members of the macro initial condition ensemble were run from 1850-1940 driven by historical CMIP5 forcing (Meinshausen et al., 2011). At year 1940, each macro initial condition member was branched into four different realizations, each subject to an atmospheric temperature perturbation of 10^{-13} to create "micro" initial condition ensembles (Hawkins et al., 2016). From these micro initial condition ensembles, 50 members were selected for the CESM1.2.2 large ensemble (specifically, 4 micro ensemble members from macro ensemble members 1 through 12 and 2 micro ensemble members from macro ensemble member 13).

The MPI Grand Ensemble was generated using the low resolution set up of the MPI Earth System Model (MPI-ESM1.1) (Giorgetta et al., 2013). The 100 member ensemble has macro initial conditions, a preindustiral control simulation was branched on the first of January for selected years between 1874 and 3524 to sample different states of a stationary and volcano-free 1850 climate (Maher et al., 2019). The MPI-GE uses ECHAM6.3 run in a T63L47 configuration (Stevens et al., 2013) as its atmospheric component model for a horizontal resolution of approximately 1.8°.

The CanESM2 (Arora et al., 2011) large ensemble was initiated from the 5 CanESM2 members contributed to CMIP5 (which are thus included in our CMIP5 basis multi-model ensemble). As with CESM1.2.2, the CanESM2 large ensemble has a combination of macro and micro initial conditions. Macro initial conditions were taken from year 1950 of the 5 original CanESM2 members. Each were then branched 10 times with micro initial conditions (a random permutation to the seed used in the random number generator for cloud physics) to give a total of 50 members (Swart et al., 2018). The CanESM2 large ensemble uses the CanAM4 atmosphere model run at a T63 spectral resolution.

3 Weighting Scheme

To constrain the multi-model ensemble uncertainty, multi-model ensemble members are weighted by a combination performance and independence weighting metric developed by Knutti et al. (2017), following on the work of Sanderson et al. (2015a, b). The basic principle is that a multi-model ensemble member will receive a performance weight based on how closely it resembles observed climate (based on chosen predictors; detailed in the following section). That performance weight will then be scaled by an independence weight which determines the degree to which a multi-model ensemble member is redundant, a duplicate of another member in the ensemble. Here, both the performance and independence weights are based on root-mean-

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics Discussions
Open Access
EGU

square-error (RMSE) distance metrics; D_i represents the distance between a multi-model ensemble member and observations and S_{ij} represents the distance between multi-model ensemble member i and multi-model ensemble member j.

It is particularly important to note the distinction between *multi-model ensemble member* and *model*, the language is intentionally chosen to highlight that each SMILE member becomes a multi-model ensemble member receiving a weight, though they come from a single model. The independence metric is based solely on S_{ij} and not on any prior knowledge of the multi-model ensemble member's origin. This is done to preserve the contribution of any member that adds independent information to the ensemble statistics, whether that information comes from internal variability within a SMILE or from inter-model differences.

The distance metrics are used in a weighting function w_i to determine the weight of each multi-model ensemble member (M):

$$w_i = \frac{e^{-\frac{D_i^2}{\sigma_D^2}}}{1 + \Sigma_{j \neq i}^{M} e^{-\frac{S_{ij}^2}{\sigma_S^2}}} \tag{1}$$

Upon computation of all the weights, each weight is then normalized by $\Sigma_i w_i$ such that they sum to 1.

The numerator of w_i serves as the performance weight, which decreases exponentially as members get further from observations ($D_i >> 0$). A shape parameter σ_D dictates the width of the performance weight Gaussian, determining how far apart a member and observations must be to be down-weighted. A larger σ_D results in more ensemble members receiving small performance weights and vice versa. Therefore, σ_D reflects how strongly one wishes to penalize a member for not resembling observations. Here, we select σ_D to be 0.4 (further discussion in Appendix B).

The denominator of w_i serves as the independence weight, which is based on how far a member is from all the other members in the ensemble. As with the performance weight, a shape parameter σ_S dictates the width of the Gaussian which is applied to each member pair. σ_S represents how close a member can be to another member before they are considered redundant. For a member with no close neighbors ($S_{ij} >> \sigma_S$), the independence weight tends to 1, preserving the member's overall weight. For a member with many close neighbors ($S_{ij} << \sigma_S$), the independence weight is greater than 1 and reduces its overall weight. The inclusion of SMILEs in a multi-model ensemble emphasizes the need for an independence weight; if SMILE members are simply redundant, they serve to bias ensemble statistics by virtue of being abundantly represented. It is therefore important to select a σ_S that is large enough such that members of a SMILE that are similar to each other are considered redundant members, but not so large that the majority of multi-model ensemble members are considered redundant. Here, we select DJF NEU σ_S to be 0.23 and JJA MED σ_S to be 0.26. Sensitivity to the choice of σ_S and further details on selection strategies are discussed in Appendix B.

3.1 Defining "Climate": Predictor Selection

Both the performance and the independence weight are based on a chosen definition of climate, a member's performance is based on its ability to reproduce observed climate and a member's independence is based on how much its climate differs from

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions

the climate in other members. When defining climate, the aim is to optimize the "fit for purpose". For example, in Knutti et al. (2017), aspects of climate relevant for September sea ice extent, such as the climatological mean and trend in hemispheric mean September Arctic sea ice extent, gridded climatological mean and standard deviation in SAT for each month, were chosen. Lorenz et al. (2018) further discusses strategy for choosing predictors for North American maximum temperature, and ultimately selected from a set of 24 predictors deemed relevant based on known physical relationships, predictor-target correlations, and variance inflation considerations.

Here "fit for purpose" is a relatively simple and straight-forward definition of climate within which the sensitivity of the weighting scheme can be interrogated. We base the weighting on 9 predictors: the climatology and interannual variability (represented by standard deviation) of SAT and SLP during the periods of 1950-1969 and 1990-2009 and a 50-year SAT trend for the period of 1960-2009. SAT and SLP are chosen as predictors because (i) they have been found to be highly relevant predictors by earlier studies (Brunner et al., 2019) and (ii) they are among the most comprehensively measured atmospheric fields prior to the satellite era (Trenberth and Paolino, 1980). In terms of spatial domain, SAT predictors are computed over their corresponding ocean-masked SREX regions (i.e. NEU for DJF and MED for JJA) and SLP predictors are computed over a larger domain which includes the North Atlantic ($25-90°$N and $60°$W$-100°$E).

To compute the aggregate distance metrics from 9 predictors, all predictor and observational fields are bilinearly interpolated to a shared $2.5° \times 2.5°$ latitude-longitude grid. For each predictor, RMSE distances are computed at each grid point within the predictor domain and then area-averaged to obtain predictor distance metric values for each member. The resulting distributions of predictor distance metrics are then normalized by their mid-range value ((maximum + minimum)/2). The normalization allows each member to have a D_i and S_{ij} that is equal part each of the 9 predictor distance metrics.

A final consideration in predictor selection is one of relationship between past and future predictor behavior. A member's performance weight is based on its ability to reproduce observed climate and this methodological choice follows from the concept of emergent constraints (e.g. Hall and Manabe, 1999; Allen and Ingram, 2002; Borodina and Knutti, 2017). The idea is that if a model accurately represents an aspect of historical climate, it is likely to realistically represent relevant physical processes and therefore is likely to provide a reliable future projection. For this to hold, a statistical relationship between the historical and future climate feature of interest must exist.

Statistical relationships between historical and future climate can be obscured by internal variability, and the inclusion of SMILEs in a multi-model ensemble highlights the need to understand the role of internal variability on the chosen predictors. In particular, internal variability has been shown to influence trends in regional SAT even on the 50-year predictor timescales we have selected (Deser et al., 2016). Because of this, a member may have a similar-to-observed SAT trend (and thus a higher performance weight) by chance, simply because it has similar-to-observed climate variability over the trend period (i.e. a similar set of El Niño and La Niña events or similar phasing of the Atlantic Multi-Decadal Oscillation). Because internal variability is inherently random in temporal phase (Deser et al., 2012), a member's match to observations over one trend period does not guarantee a match in the future. This issue is demonstrated in Figure 2ai, which shows that there is little discernible relationship (R^2 = 0.01) between the DJF NEU SAT trend from 1960-2009 and from 2050-2099 in CMIP5 with (black line) or without (blue line) the SMILEs. Few members have a similar-to-observed 1960-2009 DJF NEU SAT trend of 3.7 °C/50 years.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics
Discussions
Open Access
EGU

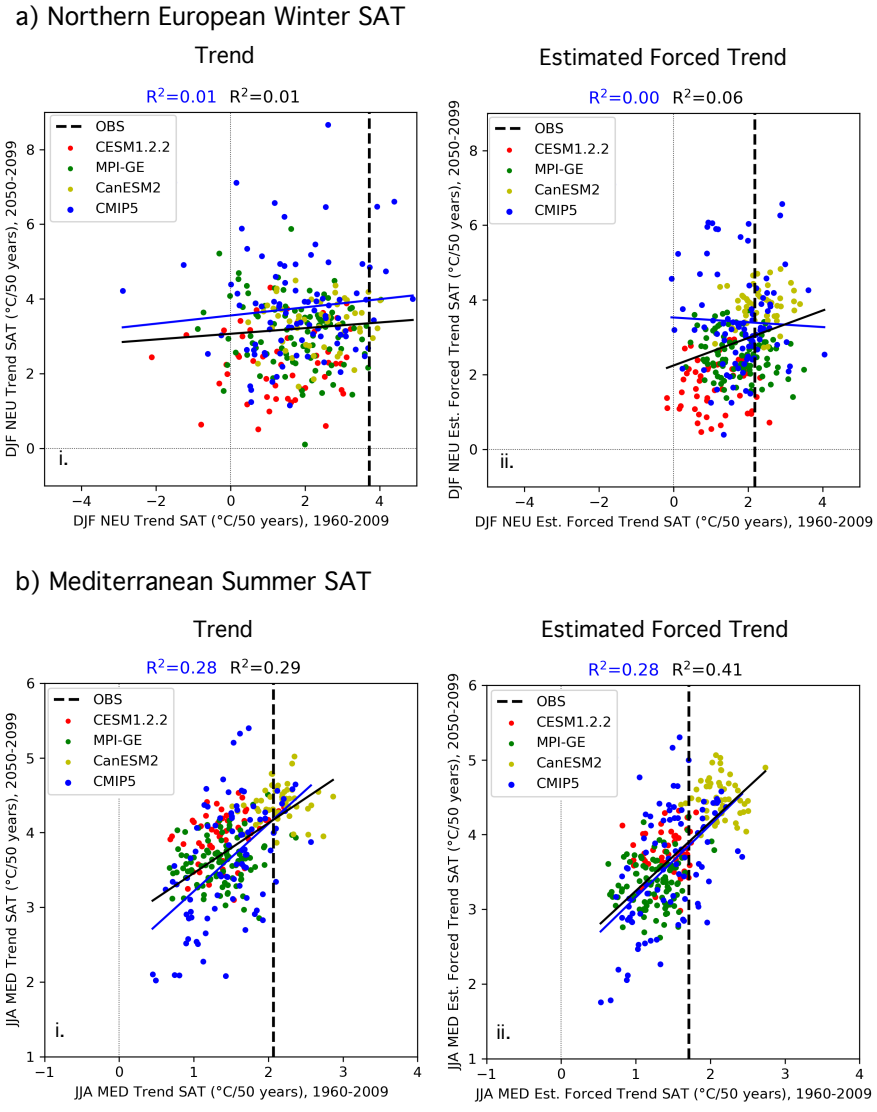


Figure 2. Predictor relationships of the domain-averaged 50-year trends of a) DJF NEU SAT and b) JJA MED SAT. 50-year raw trends are shown in panel i and 50-year estimated forced trends are shown in panel ii. In each panel, 1960-2009 is shown on the abscissa and 2050-2099 is shown on the ordinate. Observational estimates of the 1960-2009 trends are indicated with the dashed black line. Least-squares regression fits (solid lines) and R^2 values computed for solely the CMIP5 output are shown in blue and computed for all output (CMIP5 and the three SMILEs) are shown in black.

205 In contrast, a relationship emerges in summer between 1960-2009 and 2050-2099 Mediterranean SAT trends that is reinforced by the addition of the SMILEs (Fig.2bi).

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions

The removal of the influence of internal variability from regional SAT, however, provides an alternative performance metric on which observations and models can be compared. Using a method of dynamical adjustment (described in Appendix A and in further detail in Deser et al. (2016)), we obtain the estimated forced trend for 1960-2009 and 2050-2099. It is important to distinguish the *estimated* forced trend from the forced trend, which is defined as the average across ensemble members. With only one observed realization of climate, there is no observational equivalent to the ensemble mean. In contrast, the estimated forced trend can be computed in the same manner through dynamical adjustment in both observations and each multi-model ensemble member.

Internal variability serves to amplify the observed SAT trend in both seasons, by 1.5°C in Northern European winter (Fig.2a) and by 0.4°C in Mediterranean summer (Fig.2b). By removing the influence of internal variability, the resulting observed estimated forced SAT trends fall more centrally within the CMIP5 and SMILE distributions, as opposed to in the high side tail. In terms of weighting, this means more members will receive a non-zero performance weight and will ultimately contribute to the uncertainty estimate.

The estimated forced trend can also be thought of as a property of each model, a measure of response to the shared forcing analogous to climate sensitivity (Knutti et al., 2017). We find that SMILE members, which share both model setup and forcing with each other, also tend to have similar estimated forced trends (Fig.2a,bii). In winter, the clustering of SMILE estimated forced trends is striking in comparison with SMILE trends, CESM1.2.2 (CanESM2) members tend to have the least (most) NEU warming in both periods, with MPI-GE members in between. The addition of the SMILEs then introduces a positive relationship between past and future responses (Fig. 2aii, black line), while the relationship in CMIP5 is shown to be slightly negative (Fig. 2aii, blue line). In summer, the positive relationship seen even between past and future Mediterranean SAT trends (Fig.2bi) is bolstered further by the combination of removing internal variability and adding the SMILES (Fig.2bii). R^2 values increases from 0.28 to 0.41. CanESM2 has the most JJA MED warming in both the past and future periods, while MPI-GE has the least. Because estimated forced SAT trends in the regions of interest serve as a "fair" metric on which to compare models and observations, we use is as the ninth predictor in the definition of climate used in our weightings. Emergent relationships within the other eight predictors are discussed in Appendix C.

4 Results

To assess the influence of the weighting, we evaluate the magnitude of regional European end-of-century warming in terms of the SAT change (Δ) from 1990-2009 climatology to 2080-2099 climatology. Two ensembles are considered, one comprised solely of CMIP5 members (CMIP5; 88 members) and one comprised of all available members from CMIP5 and the three SMILE (ALL; 288 members). The SAT Δ distributions in both ensembles are shown as box-and-whiskers elements in Figure 3 a,bi, with the unweighted distributions shown in transparent colors in the background and the weighted distributions shown in solid colors in the foreground. Unweighted and weighted ensemble mean values are shown by solid horizontal lines within the box elements. Ensemble spread is illustrated by the box, which indicates the 25th and 75th percentile, and the whisker, which indicates the 5th and 95th percentile.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions

Three comparisons between the CMIP5 and ALL distributions help to elucidate (i) how the weighting constrains uncertainty in the magnitude of end-of-century regional European warming and (ii) why it is important to weight SMILE members within a multi-model ensemble. First, comparing distributional shifts between the unweighted and weighted CMIP5 distribution anchors how the weighting constrains uncertainty. In the absence of the weighting, the CMIP5 ensemble projects an ensemble mean end-of-century warming of 5.9°C and an interquartile spread of 2.2°C for Northern European winter (Fig.3ai). Applying weights to the CMIP5 ensemble shifts the DJF NEU SAT Δ ensemble mean downwards by 0.4°C, the 75th percentile downwards by 0.7°C, and 25th percentile downwards by 0.2°C. This distributional shift towards less end-of-century warming is a due, in part, to members with SAT Δ greater than 8°C receiving low weights, which are two orders of magnitude smaller than the average assigned weight. For Mediterranean summer, the CMIP5 ensemble projects an unweighted ensemble mean SAT Δ of 5.5°C and an interquartile spread of 1.5°C (Fig.3bi, transparent blue). The application of the weights results in a slight upward shift in the JJA MED SAT Δ ensemble mean (by 2%) and more substantial upward shifts in the 25th (by 5%) and 5th percentiles (by 10%). The weighting does not affect the 75th or 95th percentile of the JJA MED SAT Δ distribution; members which fall within the unweighted interquartile spread receive both the lowest and the highest weights. The contraction of the low end-of-century warming tail broadly consistent with the shift found for the same region and season in Brunner et al. (2019).

Secondly, a comparison of the unweighted CMIP5 (transparent blue) and ALL (transparent gray) distributions demonstrates why the weighting is necessary when incorporating SMILE members into a multi-model ensemble. The addition of 200 SMILE members to the 88 member CMIP5 ensemble shift the end-of-century warming distributions in both mean and interquartile spread. To help illustrate the mean shifts, maps of the difference between the unweighted ALL ensemble mean and the unweighted CMIP5 ensemble mean are shown for DJF NEU in Fig.3aii and for JJA MED in Fig.3abii. The added SMILE members shift the ALL SAT Δ distributions in the same direction as the weighting shifts the CMIP5 distributions, i.e. less DJF NEU end-of-century warming and more JJA MED end-of-century warming, but this is largely by chance. A SMILE with a different end-of-century warming tendency (for example, the 10-member CSIRO-Mk3-6-0 ensemble within CMIP5 has an average DJF NEU SAT Δ of 7.6°C) will shift the distribution accordingly. With the 200 new projections of end-of-the-century warming all equally contributing to distribution, the ALL ensemble has 25% less interquartile spread in both the DJF NEU and the JJA MED ensemble distributions; treating each SMILE member as an independent piece of information serves to artificially constrain uncertainty.

Thirdly, by comparing the weighted CMIP5 (solid blue) to the weighted ALL distribution (solid gray), we ascertain that the weighting can account for redundant information in the SMILEs. The ensemble mean of the weighted ALL ensemble is within 0.3°C of the weighted CMIP5 ensemble mean in DJF over the NEU domain (Fig.3aiii). The ALL weighting also recovers additional uncertainty in the 95th percentile and shifts interquartile spread upwards slightly towards that of the weighted CMIP5 ensemble. However, even after weighting, the addition of the SMILEs increases the positive skewness of the DJF NEU SAT Δ distribution, because no SMILE member warms more than 7°C between 1990-2009 and 2080-2099. As CMIP5 members that warm more than 8°C are down-weighted by the DJF NEU performance metric, the increased constraint on uncertainty brought by the SMILEs is consistent with the expected end-of-century warming range in CMIP5.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics Discussions

Strikingly, the weighted JJA MED ALL ensemble distribution is nearly identical to the JJA MED CMIP5 ensemble distribution (Fig.3bi,iii). This suggests that the SMILEs do not add much "new" information about JJA MED end-of-century warming to the CMIP5 ensemble, in part due to inter-member agreement within SMILEs and in part due to the SMILE projections falling centrally within the CMIP5 ensemble. Standard deviation in end-of-century JJA MED warming in each of the SMILEs is approximately $0.2°C$ (not shown), compared to a CMIP5 standard deviation of $1.1°C$. The recovery of uncertainty in the 5th and 95th percentiles achieved through weighting such that a 288-member ensemble has the same distribution as an 88-member ensemble is also a strong indicator that the weighting properly handles redundancy in JJA MED SAT Δ projections. Provided care is taken to select appropriate shape parameters (further discussion in Appendix B), we find the weighting approach introduced by Knutti et al. (2017) to be a suitable way to incorporate large initial condition ensembles into a multi-model ensemble.

4.1 Independence as a function of SMILE size

In addition to investigating the aggregate effect of weighting on multi-model ensemble distributions, we also explore how each SMILE is affected by the weighting individually. It is specifically of interest to know how the independence weight serves to scale a SMILE member's performance weight in the presence of other SMILE members. In Figure 4, we explore the independence weight as a function number of SMILE members included in the CMIP5 ensemble. This is done by adding CESM1.2.2, CanESM2, or MPI-GE members to the CMIP5 ensemble one at a time and evaluating the how the independence weight of the first member added evolves as a function of SMILE size. The add-a-member protocol is repeated such that every SMILE member is the first member added and subsequent members are added in numerical order, i.e., member 2 is added followed by member 1, 3, 4 and so on. In this setup, the performance weight of the first added member (not shown) stays approximately constant; small deviations only occur if a subsequently added member falls outside the distribution and shifts the predictor normalization. But the independence weight increases linearly with the number of SMILE members included in the CMIP5 ensemble, with slopes that depends on SMILE member, region, and season (Fig.4).

A simplified way to understand the evolution of a SMILE member's independence weight is in the case where every CMIP5 member and the first added SMILE member are all perfectly independent (from equation 1, $S_{ij} \rightarrow \infty, \Sigma_{j \neq i}^{M} e^{-\frac{S_{ij}^2}{\sigma_S^2}} \rightarrow 0$). In this limit, all ensemble members have an independence weight of 1. If a second SMILE member that happens to be identical to the first is then added to the ensemble (S_{ij} = 0), both SMILE members will have half their original weight, receiving an independence weight of 2. If N identical SMILE members are added, each would have a weight scaled by 1/N.

In reality, CMIP5 and SMILE ensemble members are not perfectly independent nor are SMILE members identical to each other. Because CMIP5 and SMILE ensemble members are not perfectly independent, the first SMILE member added to the CMIP5 ensemble can receive an independence weight above 1, a threshold shown with a dotted black line in the panels of Fig.4. Initial independence weights above 1 happen consistently in the CanESM2 large ensemble in both seasons (Fig.4a,bii) and in the MPI-GE in winter (Fig.4aiii), indicating that the first SMILE member added is already redundant within the CMIP5

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics
Discussions
Open Access
EGU

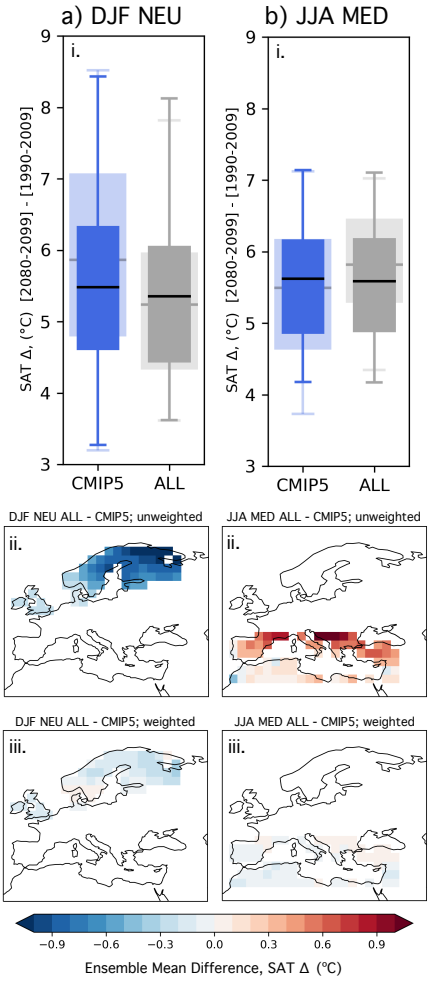


Figure 3. (i) Box-and-whiskers showing unweighted (transparent) and weighted (solid) distributions of (a) DJF NEU and (b) JJA MED SAT change (Δ, [2080-2099]-[1990-2009]) for the CMIP5 ensemble (blue) and ALL ensemble (CMIP5 with the 3 SMILEs; gray). The box element spans the 25th to 75th percentile of the distribution; mean SAT change is indicated by the horizontal line within the box. The whisker element spans the 5th to 95th percentile. (ii) Difference between the unweighted ALL ensemble mean and the unweighted CMIP5 ensemble mean at each grid-point for the (a) DJF NEU and (b) JJA MED. (iii) Difference between the weighted ALL ensemble mean and the weighted CMIP5 ensemble mean at each grid-point for the (a) DJF NEU and (b) JJA MED.

ensemble. Members of the CESM1.2.2 large ensemble tend to receive initial independence weights near 1 (Fig.4a,bi), as do MPI-GE members in summer (Fig.4biii).

Because SMILE members are not identical to each other, the first added SMILE member has an independence scaling of less than 1/N, (Fig. 4, solid black one-to-one line) in all cases. SMILE members with the maximum and minimum slopes of
310 independence weight as a function of members added are highlighted to demonstrate the range on independence scaling present

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions
Open Access
EGU

in each ensemble. The steeper the slope, the more similar a SMILE member is to other members of the SMILE. Overall, there is a larger range of slopes in the DJF NEU weighting than in the JJA MED weighting, consistent with larger inter-member spread in winter than in summer. The minimum slopes belong to the SMILE members that are relatively independent within the CMIP5 ensemble; initial independence weights range from 1.0 for CESM1.2.2 member 41 in the DJF NEU weighting (Fig.4ai) to 2.0 for CanESM2 member 20 in the JJA MED weighting (Fig.4ai). In winter, the independent CESM1.2.2 and CanESM2 members receive about a quarter of their performance weight when all 50 members are included in the CMIP5 ensemble. Over the 100 members of the MPI-GE, the independence weight of member 91 increases only from 1.4 to 7.5 (Fig.4aii). In Mediterranean summer, SMILE members are scaled by larger independence weights due to redundancy within the SMILEs. A notable example is MPI-GE member 56, which is relatively independent on its own in the CMIP5 ensemble, but ultimately has its performance weight scaled by an independence weight of 25.1 when included with its counterparts (Fig.4biii).

4.2 Should a SMILE be represented by a single member?

Another common practice used to address redundancy in a multi-model ensemble is to subset or select one ensemble member from each model. This follows from a different independence assumption than the distance-based one used in the weighting: models are independent from one another while initial condition ensemble members are not. Subsetting can also be thought of as a binary weighting scheme where a large portion of the multi-model ensemble information receives zero weight. For example, in our CMIP5 ensemble, subsetting by model corresponds to eliminating more than 50% of ensemble information (from 88 to 40 members). Subsetting by modelling center eliminates more than 75% of ensemble information (from 88 to 20 members).

It is, therefore, worthwhile to consider whether or not a SMILE can be effectively represented by a single member. We assess this by weighting a CMIP5 subset ensemble, which is comprised of the first member from each of the 40 models in our CMIP5 ensemble, with one SMILE member included for a total of 41 members. Each of the 200 SMILE members are added to the CMIP5 subset ensemble individually to serve as the SMILE's sole representative. In this configuration, the SMILE receives a weight by which it can be "ranked" in terms of its contribution to the region and season-specific weighted ensemble statistics. Weight of the SMILE, as represented by each SMILE member, is shown in Figure 5a,b; SMILE rank within the CMIP5 subset ensemble is shown in Figure 5c,d. As a guide, the weight and rank each SMILE receives when represented by member 1 are indicated with dashed line in each panel. We highlight member 1 because, often, when multiple initial condition members are available, the first member is selected (e.g. Liu et al., 2012; Karlsson and Svensson, 2013; Sillmann et al., 2013).

A SMILE's weight within the CMIP5 subset ensemble depends on which member is chosen to represent it. This reflects the dominant influence of internal variability on regional scales. In Northern European winter, the MPI-GE tends to receive the highest weight in the CMIP5 subset ensemble (w_i = 0.3), but can receive a weight of less than half that depending on which member is selected (Fig.5a, green). For the CanESM2 large ensemble, being represented by member 1 results in the SMILE receiving a relatively low rank of 20th out of 41 models within the DJF NEU ensemble (Fig.5c, yellow). If member 10 were to be used, the CanESM2 large ensemble would receive the 3rd highest weight. In Mediterranean summer, the CanESM2

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

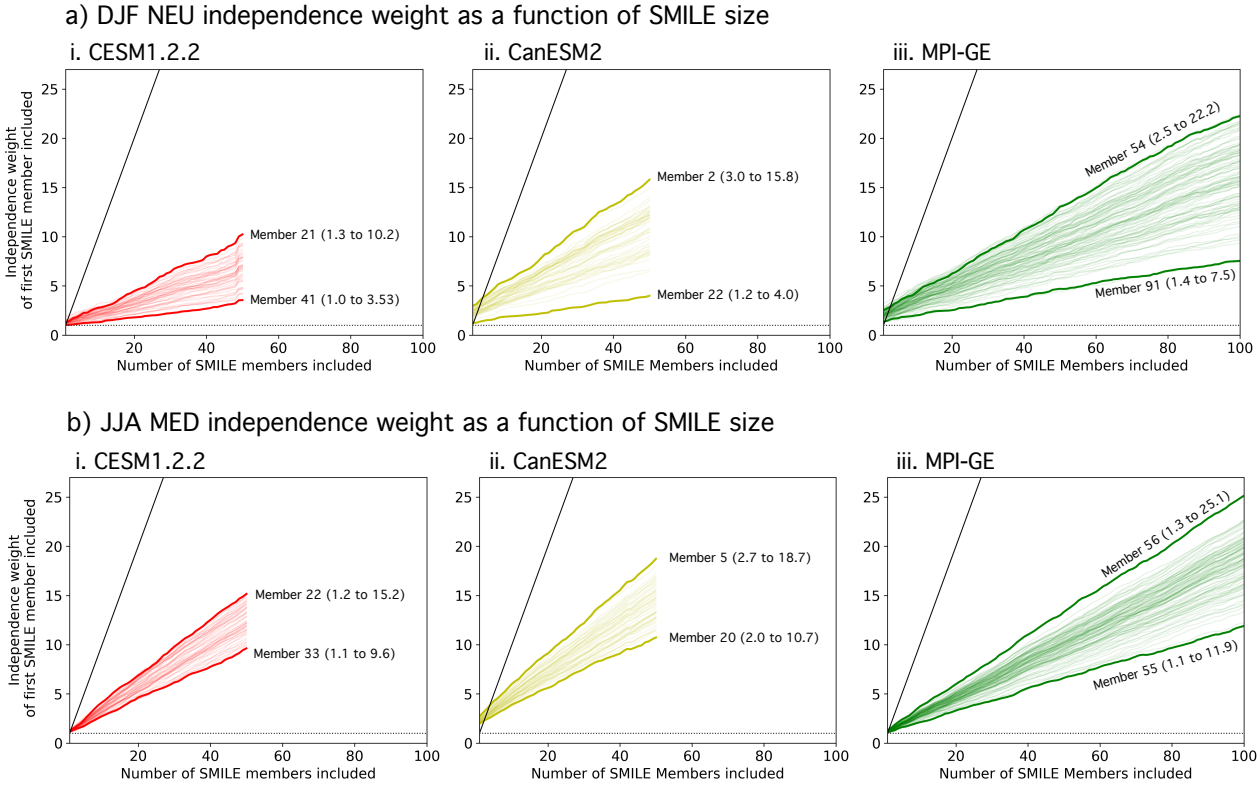


Figure 4. Evolution of the independence weight of a SMILE member as additional SMILE members are added one at a time to the CMIP5 ensemble. The CESM1.2.2 large ensemble (red; i), the CanESM2 large ensemble (yellow; ii), and the MPI-GE (green; iii) are added to the CMIP5 ensemble individually with each member having the opportunity to be included first. The independence weight when the number of SMILE members included is 1 indicates how redundant the SMILE member is within the CMIP5 ensemble. The slope of the independence weight indicates how similar other SMILE members are to the initial SMILE member, the maximum and minimum slopes in the SMILE are emphasized and labeled with the initial and final independence weight. DJF NEU evolution is shown in panel a and JJA MED evolution is shown in panel b.

large ensemble consistently receives a relatively low weight, while the CESM1.2.2 large ensemble and the MPI-GE tend to receive an average weight within the CMIP5 subset ensemble (Fig.5b). MPI-GE member 17 achieves the rank of 4th out of 41 models in the JJA MED ensemble, while member 1 assigns a rank of 22nd out of 41 models (Fig.5d, green). Ultimately, Figure 5 illustrates that uncertainty on regional-scales is likely to be over-constrained when information is eliminated through subsetting and that it is worthwhile to include the totality of information in SMILEs into uncertainty estimates through weighting.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

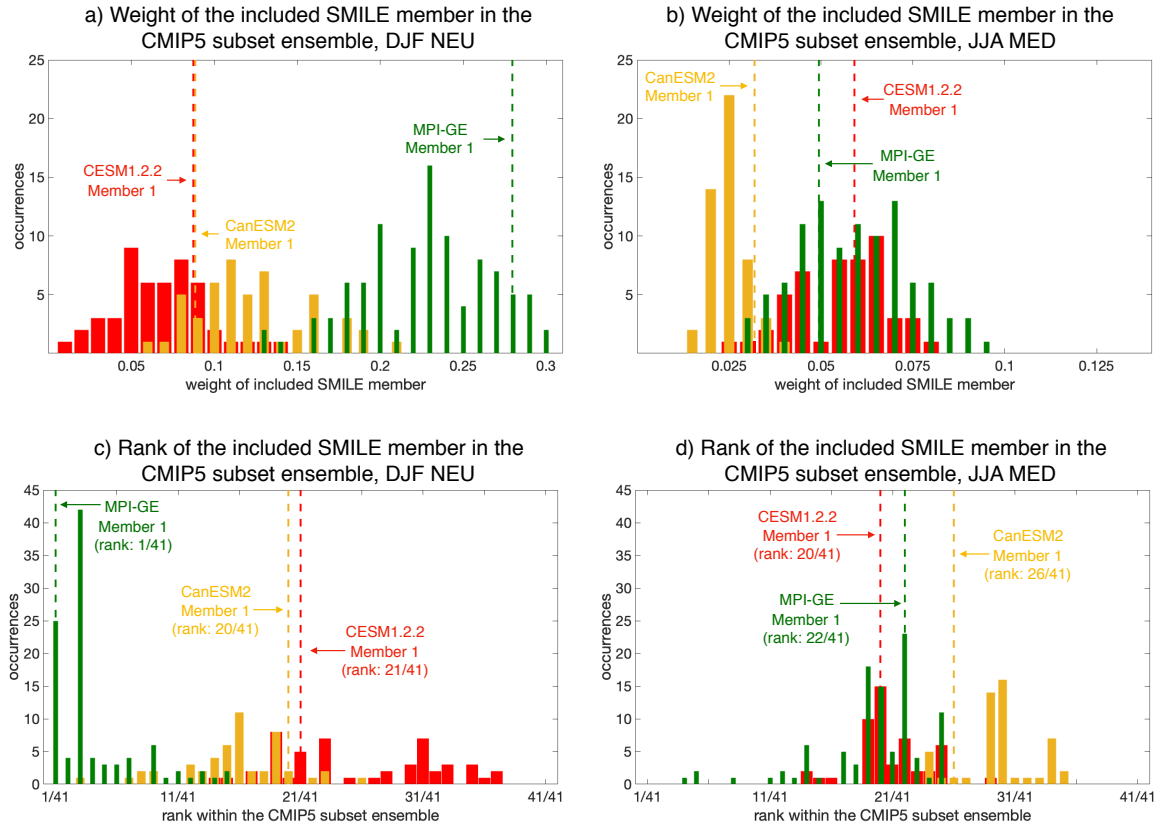


Figure 5. a) The DJF NEU weight of CESM1.2.2 large ensemble (red), CanESM2 large ensemble (yellow), and MPI-GE (green) as represented by each SMILE member. The weight of member 1 of each SMILE is indicated with a dashed line. b) As in a), but for JJA MED weights. c) The DJF NEU rank (i.e. position within the 41 member ensemble) of the included SMILE member. Rank 1/41 (41/41) is the model with the highest (lowest) weight. The rank of member 1 of each SMILE is indicated with a dashed line. d) As in c), but for JJA MED rank.

5 Conclusions

We find the performance and independence weighting scheme pioneered by Knutti et al. (2017) can be used to incorporate regional climate information from three single member initial condition large ensembles into a CMIP5 multi-model ensemble and return a justifiably constrained estimate of European regional end-of-century warming uncertainty. The weighting, which accounts for an ensemble member's ability to reproduce selected aspects of observed climate, is based on regional surface air temperature and sea level pressure climatology and interannual variability over two 20-year intervals during the historical period (1950-1969 and 1990-2009) and a 50-year estimated forced SAT trend computed using a method of dynamical adjust-

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions

ment (Deser et al., 2016). These predictors are shown to both differentiate ensemble members from one another and to bring in emergent relationships between past and future climate to the definition of performance. The principle of emergent constraints underpins the choice to use estimated forced SAT trend over SAT trend, as the former is an estimate of a model-specific property that can be compared with observations and the latter is influenced by internal variability even on 50-year timescales. Other specifics of the weighting scheme are also discussed. A new strategy to set the independence weight shape parameter at two standard deviations below the mean inter-member spread within the SMILEs, a CESM1.2.2 50-member ensemble, the CanESM2 large ensemble, and the MPI grand ensemble, is demonstrated to be effective in differentiating "new" information from redundant information that spuriously constrains uncertainty.

Uncertainty in Northern European winter and Mediterranean summer end-of-century warming is compared across combinations of a weighted and unweighted CMIP5 ensemble and a weighted and unweighted CMIP5 ensemble that includes the three SMILEs (ALL ensemble). "New" information associated with regional-scale internal variability within SMILEs contributes to narrowing weighted estimates of Northern European end-of-century winter warming uncertainty with respect to weighted CMIP5 estimates. With and without the SMILEs, the weighting of Northern European winter SAT change between 1990-2009 and 2080-2099 projects systematically less warming than the unweighted CMIP5 ensemble. The weighting is shown to recover Northern European end-of-century winter warming uncertainty in the ALL ensemble in comparison to the unweighted ALL ensemble, indicating the the independence weight effectively down-weights redundancy. The down-weighting of redundancy is even clearer in the case of Mediterranean summer end-of-century warming uncertainty, where the SMILE projections lack inter-member spread and fall centrally within the unweighted CMIP5 ensemble. The near-identical weighted CMIP5 and ALL distributions of Mediterranean summer SAT change suggest that the redundant SMILE information is properly handled by the independence weight.

To better understand the role of the independence weight in each SMILE, we computed independence weight as a function of number of SMILE members included in the CMIP5 ensemble for each SMILE individually. Independence weight increases linearly with the number of SMILE members included, with initial values indicating potential redundancy with respect to CMIP5 members and the slope indicating potential redundancy within the SMILE. SMILE members are more redundant in the JJA MED weighting than in the DJF NEU weighting; the most and least independent SMILE members are identified for the two cases. The maximum a SMILE member's performance weight is scaled is by an independence weight of 25 after the addition of 100 ensemble members. Finally, we show that it is worthwhile to include all SMILE members in multi-model ensembles in order to properly represent uncertainty due to regional-scale internal variability. When ensembles are subset to include one member of each model, uncertainty estimates are sensitive to which member is selected.

It is important to note that while the weighting has a relatively straightforward functional form, it requires an application-specific set of predictors and appropriate shape parameters. Strategies to select optimal shape parameters are discussed in Appendix B of this study and we advise that emergent predictor relationships are explored, as in Appendix C, to provide justification for the performance metric. We assess a relatively unconventional multi-model ensemble in this study, which is comprised of 200 members from 3 models and only 88 members from the remaining 40 models. This is a deliberate choice made to test the independence weight's ability to handle redundant information and to determine if internal variability is large

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions Open Access EGU

enough such that members are considered independent even though they come from the same model, as posited by Lorenz et al. (2018). We find both to be the case for regional European SAT change, but redundancy (internal variability) may contribute more (less) for other climate fields and on other spatial scales (i.e., globally). Determining best practices for representing uncertainty in a multi-model ensemble that includes initial condition ensemble members is necessary in advance of CMIP6, as modelling centers are slated to submit more ensemble members to the project than were submitted to CMIP5 (Eyring et al., 2016; Stouffer et al., 2017). A weighting scheme, such as the one assessed here, is thus ideal as it incorporates both model uncertainty and internal variability information into a justifiable estimate of CMIP6 uncertainty.

Appendix A: Dynamical Adjustment

To obtain estimated forced trends in SAT, a method of dynamical adjustment, based on constructed circulation analogues, is used (Deser et al., 2016; Lehner et al., 2017; Merrifield et al., 2017; Guo et al., 2019). Dynamical adjustment provides an empirically-derived estimate of the SAT trends induced by atmospheric circulation variability; removal of this circulation-driven component from a SAT record thus reveals an estimate of the radiatively-forced SAT trend. Dynamical adjustment relies on the ability to reconstruct a monthly mean circulation field, which we represent with sea level pressure (SLP) as in Deser et al. (2016), from a large set of analogues. SLP analogues are selected from 60 possible choices (from the period 1950-2010) in the observational record, excluding the target month, and the method is therefore referred to as the "leave-one-out" method of dynamical adjustment.

It is important to acknowledge that because of the paucity of analogue choices in leave-one-out dynamical adjustment, the term "analogue" is a bit of a misnomer. The term evokes the idea of a match, though in practice, analogues may not closely resemble the target which is discussed in more detail in the following paragraph. For convenience, we will continue to refer to the months used in target SLP construction as "analogues", but we do so with the understanding that target and analogue patterns may differ over the selection domain.

A month is determined to be an analogue of the target month if the Euclidean distance between target and analogue SLP is small. Euclidean distance is computed at each grid point and averaged over the domain 25-90°N, 60°W-100°E. This selection metric, therefore, does not require an analogue to match the target month spatially over the whole domain. This is necessary because, with 60 possible options, it is statistically unlikely that a "perfect" analogue will exist for a particular target month. van den Dool (1994) found that it would take on the order of 10^{30} years to find two Northern hemisphere circulation patterns that match within observational uncertainty. With this in mind, a smaller than hemispheric domain and an iterative averaging schemes are employed to make the most of "imperfect" analogues available (Wallace et al., 2012; Deser et al., 2014, 2016).

Once the Euclidean distances are determined, the N_a closest SLP analogues are chosen, and the iterative process of selecting N_s of N_a SLP analogues and optimally reconstructing target SLP commences. We use $N_a = 50$ and $N_s = 30$. The optimal reconstruction of target SLP is mathematically equivalent to multivariate linear regression; each analogue is assigned a weight (β_i) such that a weighted linear combination of analogues produces a least-squares estimate of the target SLP. The analogue

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019


Earth System Dynamics Discussions

weighting scheme ensures that analogues which are further from (closer to) the target, in a Euclidean distance sense, contribute less (more) to the constructed SLP field.

After SLP is constructed, the weights derived for each SLP analogue are applied to their corresponding monthly-averaged SAT fields. Prior to the application of weights, a quadratic trend representing anthropogenic warming is removed from the SAT record at each point in space. The purpose of this detrending is discussed in Deser et al. (2016). The weighted, detrended SAT fields are then used to construct a dynamic SAT anomaly field for the target month. SLP, which is a representative of low-level atmospheric circulation, and SAT are physically related; SLP-derived weights are applied to SAT to empirically construct that relationship. Conceptually, dynamic SAT anomalies are those that would occur given the attendant circulation pattern. The second through fifth steps of dynamical adjustment (selection of N_s of N_a SLP analogues, optimal reconstruction of target SLP, and construction of dynamic SAT) are then repeated N_r times. In this study, $N_r = 100$, following Lehner et al. (2017). The dynamic component of SAT in the target month is the average of the N_r constructions. It is then subtracted from SAT in the target month to find the residual component of SAT, used as an estimate of the radiatively-forced SAT trend.

Appendix B: Selecting σ_D and σ_S

Determining the shape parameters σ_D and σ_S is an important step in the weighting process. σ_D can be set using a perfect model test, as described in Lorenz et al. (2018). Here, the perfect model test is performed on an 43 member ensemble, which includes only the first initial condition member from the SMILEs and each of the 13 CMIP5 initial condition ensembles. This is done because having very similar members in the ensemble could bias the perfect model test, which is based on predicting one member using a weighted distribution of the rest. During the perfect model test, each member is assumed to be "truth" once and a weighting is performed using the remaining members to predict this "truth". A range of σ_D values are used for each truth prediction and the optimal σ_D is chosen as the smallest value such that the "truth" falls between the 10th and 90th percentile of the "truth" prediction in 80% of the cases. The 80% threshold is chosen so as to not be over-confident in the weighting.

The RMSE distances between multi-model ensemble members and observations (D_i) are shown in Figure B1. Members of the ALL ensemble are plotted alphabetically around the azimuth (see Table 1). In winter (Fig.B1a), distances between members and observations are distributed in a positively skewed fashion with the mode of the distribution at D_i = 0.41 and a tail of larger D_i values. In contrast, distances in summer (Fig.B1b) are approximately normally distributed about a mean of D_i = 0.66. The addition of the SMILEs to the distribution contribute to both of these distributional tendencies. With σ_D set to 0.4 in both cases, members are more strongly weighted by performance in winter than in summer.

σ_S can be determined using initial condition ensembles present in the multi-model ensemble, including SMILEs. The inclusion of SMILE members in a multi-model ensemble emphasizes the need for σ_S to be carefully selected, as SMILEs add redundant information and the purpose of σ_S is to reduce the influence of redundant information. However, not all information added by a SMILE is redundant. SMILEs also add "new" information to multi-model ensembles; inter-member distances in an initial condition ensemble could be as large as inter-model distances in the multi-model ensemble (Figure B2).

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

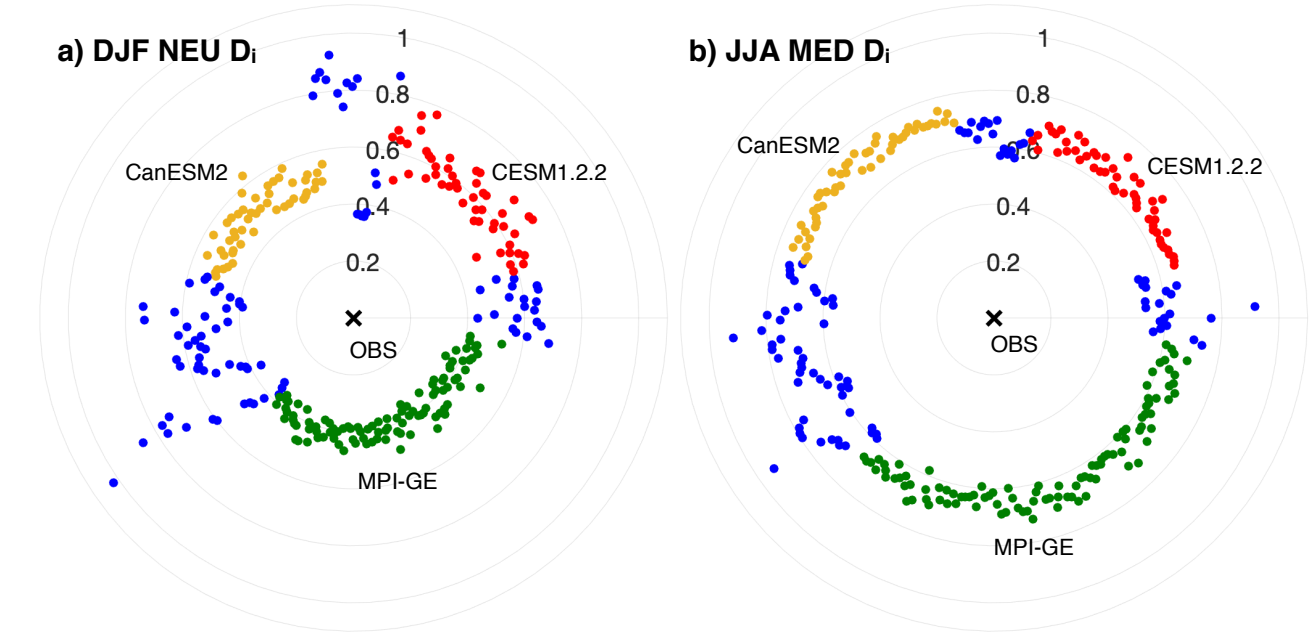


Figure B1. RMSE distance D_i, derived from the 9 climate-defining predictors, between observations (origin) and the 288 members of the ALL ensemble (CMIP5 (blue) + CESM1.2.2 (red) + CanESM2 (yellow) + MPI-GE (green)). Members are plotted in alphabetical order (see Table 1) around the azimuth. DJF NEU distances are shown in panel a and JJA MED distances are shown in panel b.

If σ_S is too small or too large, there are implications for the weighted ensemble mean and spread. This sensitivity to σ_S is shown in Figure B2. As in Figure 3, we compare the unweighted and weighted CMIP5 distributions to the unweighted and weighted ALL distribution. σ_S is varied from 0.05 to 0.8.

For small σ_S, all members are considered independent and redundant information reduces the ensemble spread in the ALL weighted distribution. The addition of the SMILE members shifts the unweighted mean of the CMIP5 distribution down by 11% in winter and up by 6% in summer. The subsequent application of what is largely the performance weight shifts the distribution down further in winter. This suggests that without the independence weight, redundant information from the SMILEs dominates the distribution and results in an underestimate of uncertainty in SAT change.

If σ_S is set on the order of the largest inter-member distances in a SMILE ($\sigma_S \geq \sim 0.4$), few members of the multi-model ensemble will be considered independent from each other. The systematic reduction of weights in the ensemble at large can also lead to an underestimate of uncertainty. Only members that are very far from other members will not have a reduction in weight, but these "independent" members tend to also be far from observations and therefore have little weight to begin with. The reduction of spread associated with a larger σ_S has more of an effect in DJF NEU SAT change distributions, consistent with the stronger performance weighting for the region and season.

In order to avoid an underestimate of uncertainty, either due to redundancy or from down-weighting independent information, we propose that σ_S should be set based on the S_{ij} distribution in initial condition ensembles present within the multi-model

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics
Discussions
Open Access
EGU

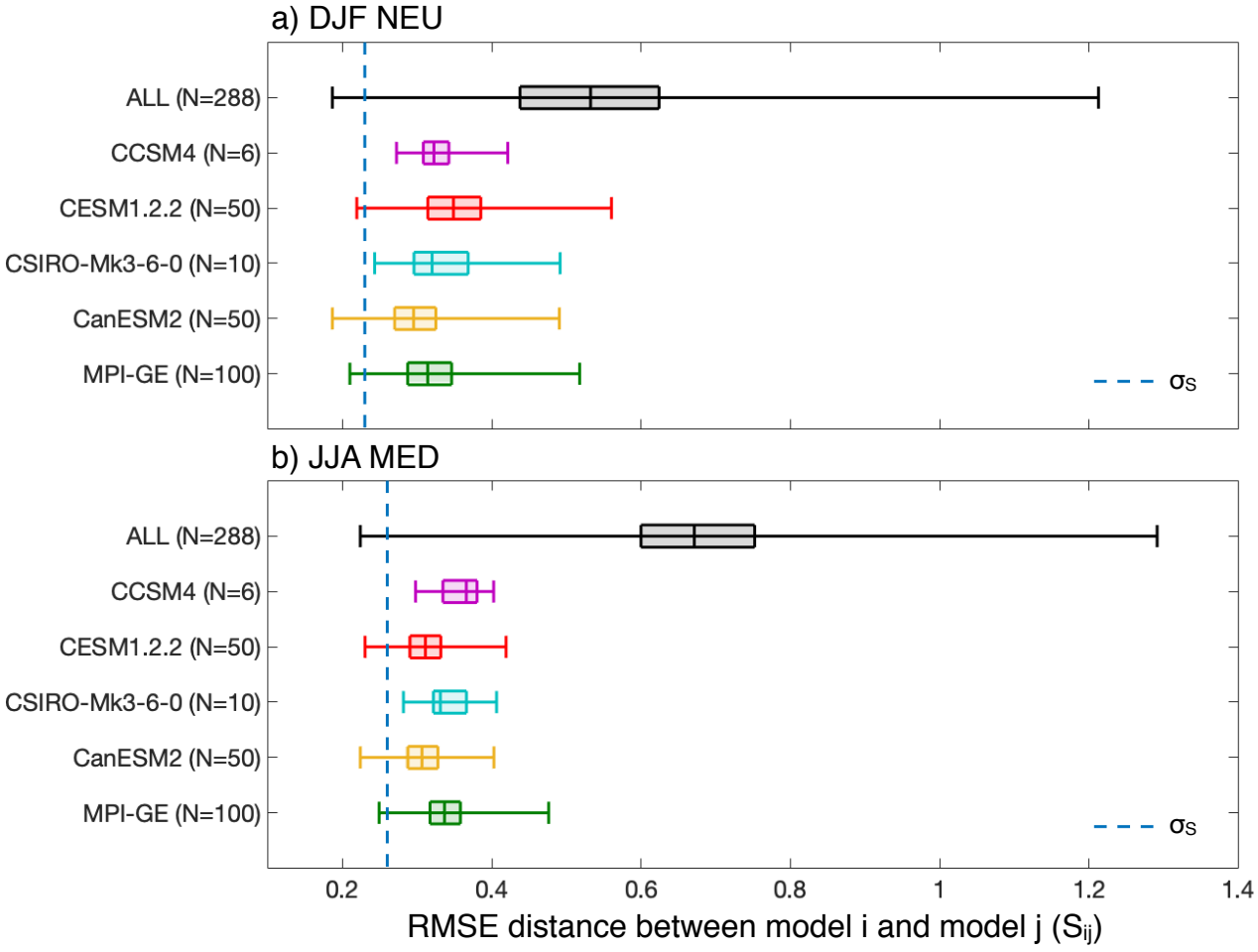
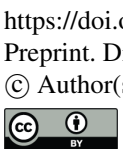


Figure B2. Distributions of RMSE distance (S_{ij}) within select CMIP5 initial condition ensembles, CCSM4 (magenta) and CSIRO-Mk3-6-0 (teal), and the SMILEs, CESM1.2.2 (red), CanESM2 (yellow), and MPI-GE (green). The box element spans the 25th to 75th percentile of the distribution; median S_{ij} is indicated by the horizontal line within the box. The whisker element spans the full range of the S_{ij} distribution. S_{ij} values are computed from a multi-model ensemble which includes all 288 members (black). The number of ensemble members in each ensemble is indicated by N. The value of σ_S used for the weighting is indicated with the dashed blue line. DJF NEU distances are shown in panel a and JJA MED distances are shown in panel b.

ensemble. We compute the S_{ij} with the three SMILEs and set σ_S at 2 standard deviations below the SMILE S_{ij} mean value (Fig. B2). The three values are then averaged. By this metric, DJF NEU σ_S is 0.23 and JJA MED σ_S is 0.26.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019



Earth System Dynamics
Discussions
Open Access
EGU

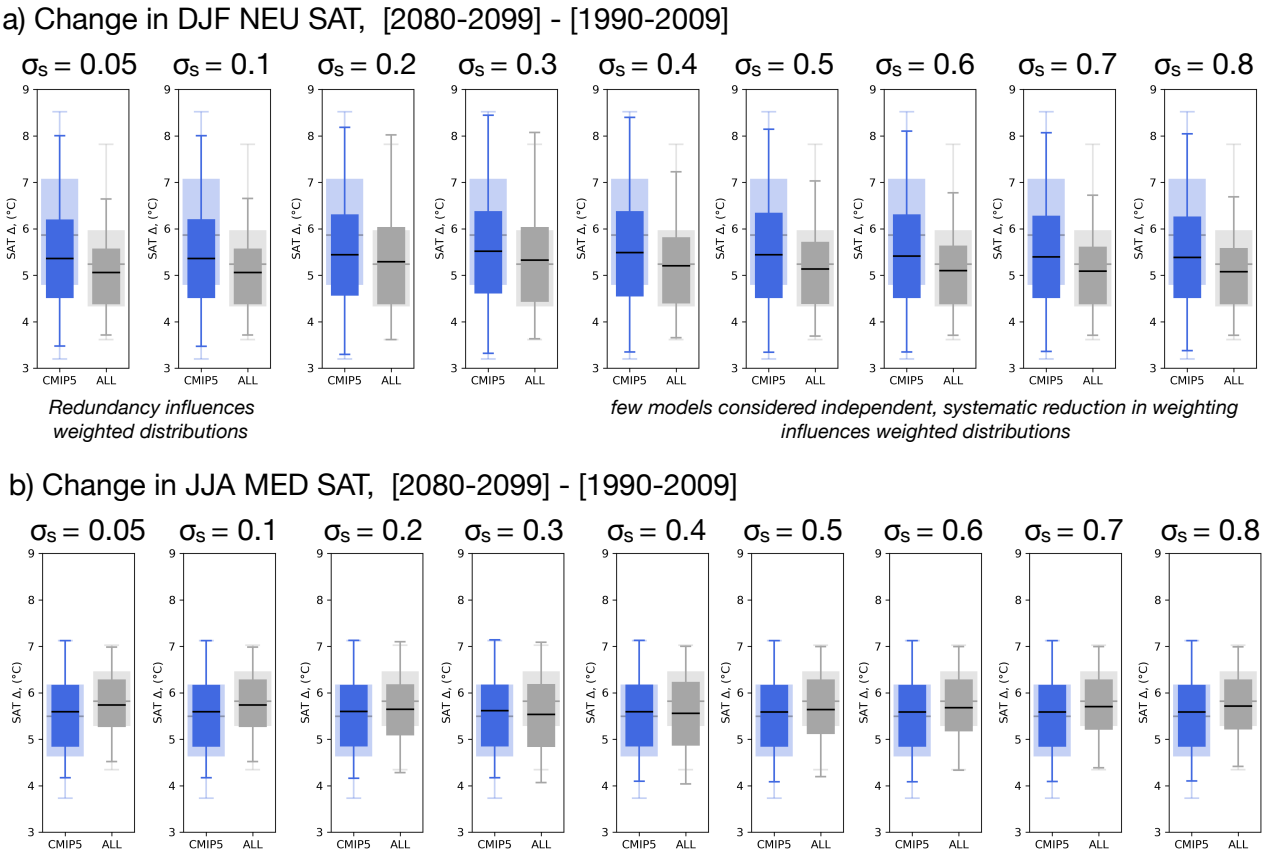


Figure B3. σ_S sensitivity of Figure 3. σ_S used for each weighting is indicated in the title of each panel. Box-and-whiskers showing the unweighted (transparent) and weighted (solid) distributions of SAT change (Δ, [2080-2099]-[1990-2009]) for the CMIP5 ensemble (blue) and ALL ensemble (CMIP5 with the 3 SMILEs; gray). The box element spans the 25th to 75th percentile of the distribution; mean SAT change is indicated by the horizontal line within the box. The whisker element spans the 5th to 95th percentile. DJF NEU SAT change is shown in a and JJA MED SAT change is shown b.

Appendix C: Emergent Predictor Relationships

In addition to relationships between past and future (estimated forced) trend (Fig.2), emergent relationships among the remaining predictors we use to represent climate are shown in Figure C1 and C2. Linear relationships are clear for climatological averages in both seasons; multi-model ensemble members with hotter mean state climate than other members during the historical period also tend to have hotter mean state climate than other members in the future. Similarly, the tendency of domain-averaged SLP values to be and remain lower or higher also persists into the future. This relationship is explored spatially in Figure C3 and C4. Mean states within SMILEs tend to cluster togehter. With the exception of JJA MED SLP climatology (Fig.C2c), the addition of the SMILEs does not change the linear relationship found in the CMIP5 multi-model ensemble.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

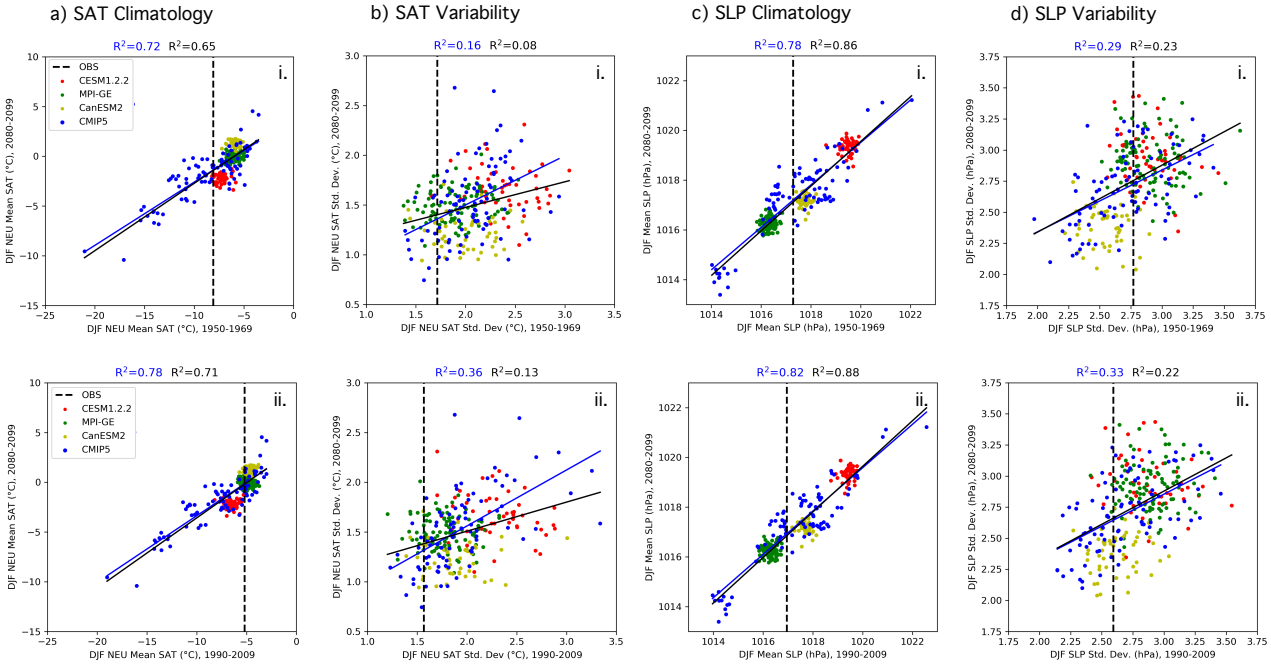


Figure C1. Predictor relationships in DJF comparing domain-averaged climate in two historical periods, (i) 1950-1969 and (ii) 1990-2009, to a future period, 2080-2099 in all panels. Observational estimates in the respective historical periods are indicated with a dashed black line in each panel. a) NEU SAT climatology (°C), b) NEU SAT standard deviation (°C), c) SLP climatology, averaged over the predictor region (hPa) and d) SLP standard deviation, averaged over the predictor region (hPa) are eight of the nine predictors used to determine member performance and independence. Least-squares regression fits (solid lines) and R^2 values computed for solely the CMIP5 output are shown in blue and computed for all output (CMIP5 and the three SMILEs) are shown in black.

For variability (standard deviation over the given period), members of SMILEs differ as much from each other as from other multi-model ensemble members in DJF (Fig.C1b,d). In JJA (Fig.C2b,d), several members of the CMIP5 multi-model ensemble have domain-averaged variability that falls outside the distribution of SMILE members. The addition of the SMILEs to the CMIP5 multi-model ensemble reduces correlations between historical and future variability for SAT and SLP in both seasons. This is particularly striking in JJA where the correlations tend to be due to the CMIP5 multi-model ensemble outliers.

Because the SLP predictor domain has a larger spatial extent than the SAT predictor domains, we also assess spatial patterns of climatological SLP which average to the lowest and highest domain-averages values in the 1990-2009 climatological period (Figures C3 and C4). The "end-members" illustrate the climatological emergent constraint relationship seen in Figures C1 and C2 in terms of pattern, which is important for a field like SLP which tends to feature dipoles on basin and continental scales.

In winter, multi-model ensemble members tend to feature similar-to-observed spatial patterns of climatological SLP in the predictor domain, with a low pressure center over the high latitude North Atlantic and a region high pressure over the Eurasian continent (Fig.C3). For the member with the lowest domain-average, the difference arises from a further extension of the

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

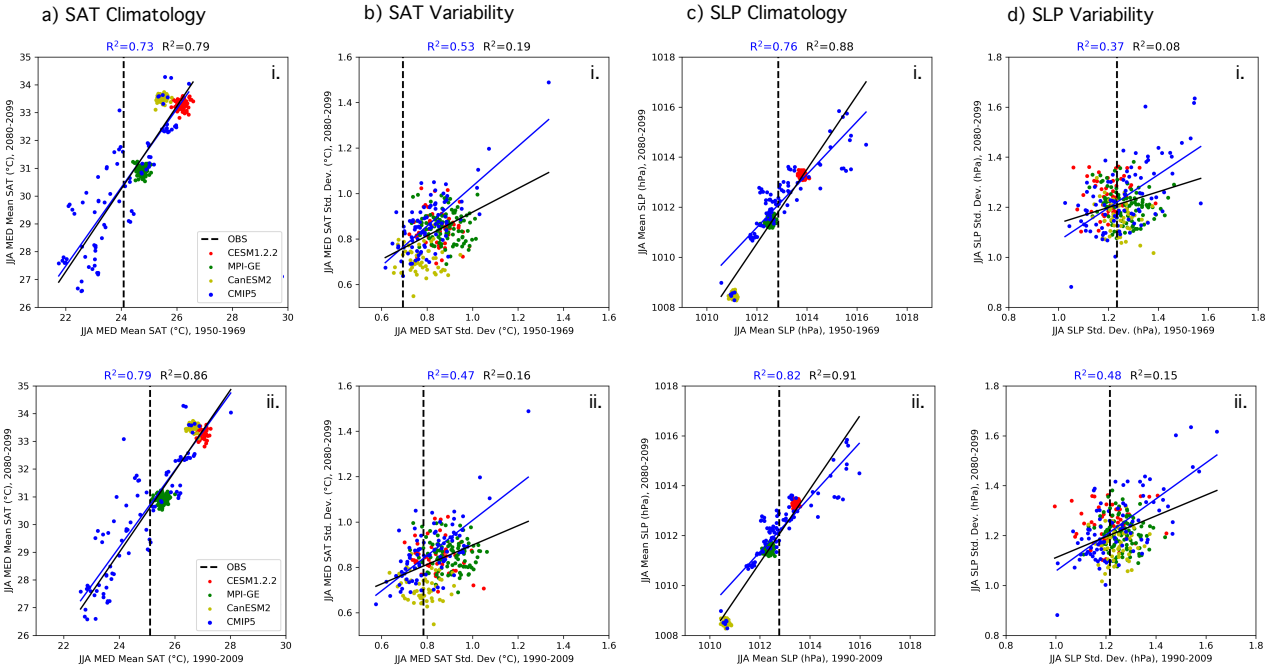


Figure C2. As is Figure C1, but for JJA. a) MED SAT climatology (°C), b) MED SAT standard deviation (°C), c) SLP climatology over the predictor region (hPa) and d) SLP standard deviation over the predictor region (hPa) are eight of the nine predictors used to determine member performance and independence.

low pressure center across Northern Europe than observed and a weaker high pressure center than observed, especially in the vicinity of the Tibetan plateau (Fig.C3 ii,v). For the member with the highest domain-average, the difference arises from high pressure features over high altitude regions, such as Greenland and the Tibetan plateau (Fig.C3 iii,vi).

In summer, members differ in spatial patterns of climatological SLP in the predictor domain, though most feature the high pressure center over the subtropical North Atlantic and lower pressure over the Eurasian continent seen in observations (Fig.C4). The member with the lowest domain-average features the afforementioned spatial pattern, but with a higher-than-observed amplitude i.e. both a higher North Atlantic subtropical high pressure center and a lower region of continental low pressure (Fig.C4 ii,v). In contrast, the member with the highest domain-average has high pressure over the entire Atlantic basin as well as over Greenland and the Tibetan plateau (Fig.C4 iii,vi). Most importantly, in all cases, the climatological behavior of the past continues into in the future which supports the primary tenet of an emergent constraint.

Author contributions. RK, RL, and LB conceived of and wrote the weighting scheme python package. ALM and LB implemented the weighting scheme with contributions from RL. ALM and LB analyzed the output. ALM wrote the paper with contributions from all co-authors.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics
Discussions
Open Access
EGU

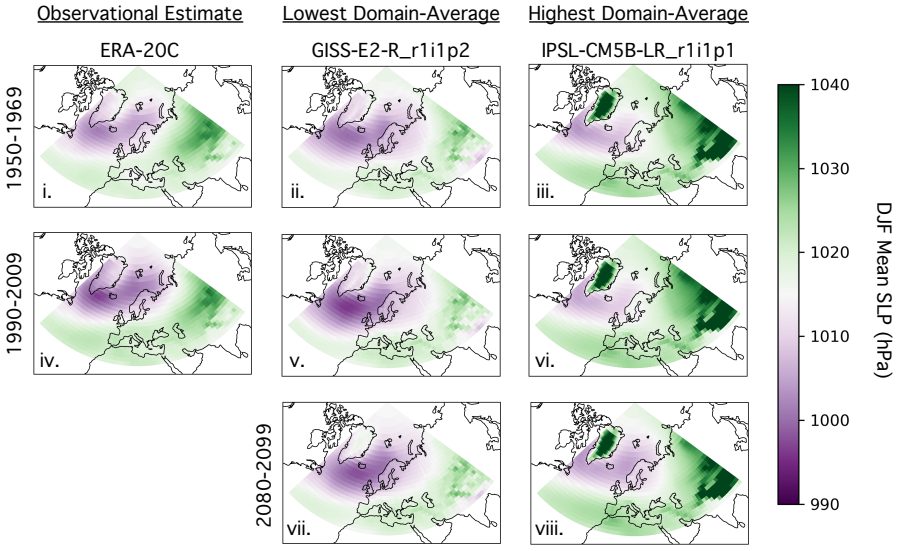


Figure C3. The spatial pattern of DJF SLP climatology for: 1950-1969 (i-iii), 1990-2009 (iv-vi), and 2080-2099 (vii-viii). The observational estimate of SLP climatology (ERA-20C) is shown in the left column (i,iv). The ensemble member with the lowest domain-average SLP climatology for the two historical periods (GISS-E2-Rr1i1p2) is shown in the center column (ii,v,vii). The ensemble member with the highest domain-average SLP climatology the two historical periods (IPSL-CM5B-LRr1i1p1) is shown in the right column (iii,vi,viii).

Competing interests. We declare that we have no conflict of interest.

Acknowledgements. We would like to thank Drs. Nicola Maher, Flavio Lehner, Iselin Medhaug, and Sebastian Sippel for their helpful comments on this manuscript. This project was funded by the European Union's Horizon 2020 research and innovation program under grant agreements 641816 (CRESCENDO) and 776613 (EUCP). We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups for producing and making available their model output. CMIP5 data was obtained from http://cmip-pcmdi.llnl.gov/cmip5/. The CESM1.2.2 large ensemble was generated at ETH Zürich and is available upon request. The CanESM2 large ensemble was generated by the Environment and Climate Change Canada's Canadian Centre for Climate Modelling and Analysis and is available at http://open.canada.ca/data/en/dataset/aa7b6823-fd1e-49ff-a6fb-68076a4a477c. The MPI Grand Ensemble was generated at the Max Planck Institute for Meteorology and is available esgf-data.dkrz.de/projects/mpi-ge/. ERA-20C data is provided by ECMWF and was obtained from https://apps.ecmwf.int/datasets/data/era20c-moda/levtype=sfc/type=an/. The weighting protocol is available as a python package and can be obtained via GitHub (https://github.com/lukasbrunner/ClimWIP) under a GPLv3.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Earth System Dynamics
Discussions
Open Access
EGU

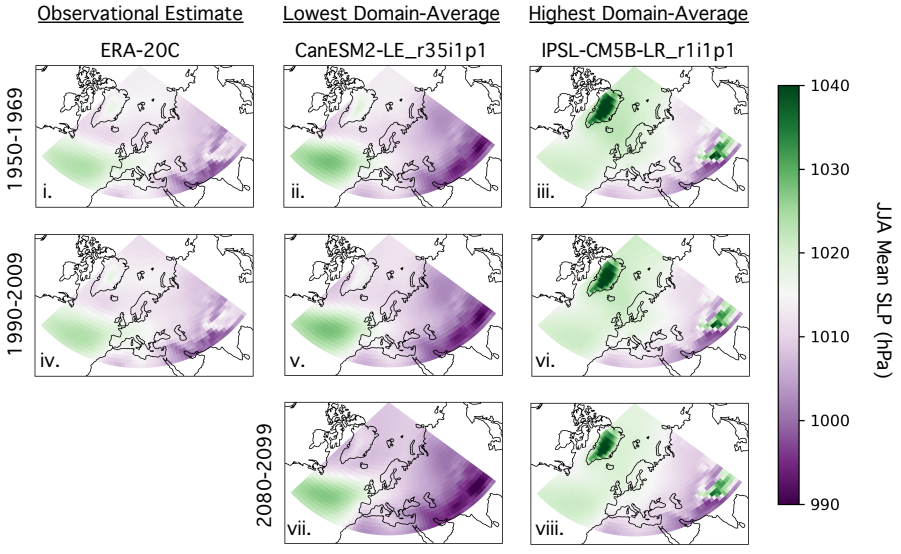


Figure C4. As is Figure A3, but for JJA SLP climatology. The observational estimate of SLP climatology (ERA-20C) is shown in the left column (i,iv). The ensemble member with the lowest domain-average SLP climatology for the two historical periods (CanESM2-LEr35i1p1) is shown in the center column (ii,v,vii). The ensemble member with the highest domain-average SLP climatology the two historical periods (IPSL-CM5B-LRr1i1p1) is shown in the right column (iii,vi,viii).

References

Abramowitz, G., Leuning, R., Clark, M., and Pitman, A.: Evaluating the performance of land surface models, J. Climate, 21, 5468–5481, https://doi.org/10.1175/2008JCLI2378.1, 2008.

Abramowitz, G., Herger, N., Gutmann, E., Hammerling, D., Knutti, R., Leduc, M., Lorenz, R., Pincus, R., and Schmidt, G. A.: ESD Reviews: Model dependence in multi-model climate ensembles: weighting, sub-selection and out-of-sample testing, Earth Syst. Dyn., 10, 2019.

Allen, M. R. and Ingram, W. J.: Constraints on future changes in climate and the hydrologic cycle, Nature, 419, 228–232, https://doi.org/10.1038/nature01092, 2002.

Arora, V. K., Scinocca, J. F., Boer, G. J., Christian, J. R., Denman, K. L., Flato, G. M., Kharin, V. V., Lee, W. G., and Merryfield, W. J.: Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases, Geophys. Res. Lett., 38, L05 805, https://doi.org/10.1029/2010GL046270, 2011.

Boberg, F. and Christensen, J.: Overestimation of Mediterranean summer temperature projections due to model deficiencies, Nature Clim Change, 2, 433–436, https://doi.org/10.1038/nclimate1454, 2012.

Boé, J.: Interdependency in multimodel climate projections: Component replication and result similarity, Geophysical Research Letters, 45, 2771–2779, https://doi.org/10.1002/2017GL076829, 2018.

Borodina, A., E. F. and Knutti, R.: Emergent Constraints in Climate Projections: A Case Study of Changes in High-Latitude Temperature Variability, J. Climate, 30, 3655–3670, https://doi.org/10.1175/JCLI-D-16-0662.1, 2017.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Brunner, L., Lorenz, R., Zumwald, M., and Knutti, R.: Quantifying uncertainty in European climate projections using combined performance-independence weighting, Environmental Research Letters, http://iopscience.iop.org/10.1088/1748-9326/ab492f, 2019.

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski, W., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A., and Wehner, M.: Long-term Climate Change: Projections, Commitments and Irreversibility, book section 12, p. 1029–1136, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, https://doi.org/10.1017/CBO9781107415324.024, www.climatechange2013.org, 2013.

Deser, C., Tomas, R., and Sun, L.: The Role of Ocean-Atmosphere Coupling in the Zonal-Mean Atmospheric Response to Arctic Sea Ice Loss, J. Climate, 28, 2168–2186, https://doi.org/10.1175/JCLI-D-14-00325.1, 2007.

Deser, C., Knutti, R., Solomon, S., and Phillips, A. S.: Communication of the role of natural variability in future North American climate, Nature Climate Change, 2, 775–779, https://doi.org/10.1038/NCLIMATE1562, 2012.

Deser, C., Phillips, A., Alexander, M. A., and Smoliak, B. V.: Projecting North American climate over the next 50 years: Uncertainty due to internal variability, J. Climate, 27, 2271–2296, https://doi.org/10.1175/JCLI-D-13-00451.1, 2014.

Deser, C. A., Terray, L., and Phillips, A. S.: Forced and internal components of winter air temperature trends over North America during the past 50 years: Mechanisms and implications, J. Climate, 29, 2237–2258, https://doi.org/10.1175/JCLI-D-15-0304.1., 2016.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: 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, 2016.

Giorgetta, M. A., Jungclaus, J., Reick, C. H., Legutke, S., Bader, J., Böttinger, M., Brovkin, V., Crueger, T., Esch, M., Fieg, K., Glushak, K., Gayler, V., Haak, H., Hollweg, H., Ilyina, T., Kinne, S., Kornblueh, L., Matei, D., Mauritsen, T., Mikolajewicz, U., Mueller, W., Notz, D., Pithan, F., Raddatz, T., Rast, S., Redler, R., Roeckner, E., Schmidt, H., Schnur, R., Segschneider, J., Six, K. D., Stockhause, M., Timmreck, C., Wegner, J., Widmann, H., Wieners, K., Claussen, M., Marotzke, J., and Stevens, B.: Climate and carbon cycle changes from 1850 to 2100 in MPI-ESM simulations for the Coupled Model Intercomparison Project phase 5, Journal of Advances in Modeling Earth Systems, 5, 572–597, https://doi.org/10.1002/jame.20038, 2013.

Guo, R., Deser, C., Terray, L., and Lehner, F.: Human influence on winter precipitation trends (1921–2015) over North America and Eurasia revealed by dynamical adjustment, Geophysical Research Letters, 46, 3426–3434, https://doi.org/10.1029/2018GL081316, 2019.

Hall, A. and Manabe, S.: The Role of Water Vapor Feedback in Unperturbed Climate Variability and Global Warming, Journal of Climate, 12, 2327–2346, https://doi.org/10.1175/1520-0442(1999)012<2327:TROWVF>2.0.CO;2, 1999.

Hawkins, E. and Sutton, R.: The Potential to Narrow Uncertainty in Regional Climate Predictions, Bull. Amer. Meteor. Soc., 90, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1, 2009.

Hawkins, E., Smith, R. S., Gregory, J. M., and Stainforth, D. A.: Irreducible uncertainty in near-term climate projections, Clim Dyn, 46, 3807–3819, https://doi.org/10.1007/s00382-015-2806-8, 2016.

Herger, N., Abramowitz, G., Knutti, R., Angélil, O., Lehmann, K., and Sanderson, B. M.: Selecting a climate model subset to optimise key ensemble properties, Earth System Dynamics, 9, 135–151, https://doi.org/10.5194/esd-9-135-2018, 2018.

Hurrell, J., Holland, M., Gent, P., Ghan, S., Kay, J., Kushner, P., Lamarque, J., Large, W., Lawrence, D., Lindsay, K., Lipscomb, W., Long, M., Mahowald, N., Marsh, D., Neale, R., Rasch, P., Vavrus, S., Vertenstein, M., Bader, D., Collins, W., Hack, J., Kiehl, J., and Marshall, S.: The Community Earth System Model: A Framework for Collaborative Research, Bull. Amer. Meteor. Soc., 94, 1339–1360, https://doi.org/10.1175/BAMS-D-12-00121.1, 2013.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Karlsson, J. and Svensson, G.: Consequences of poor representation of Arctic sea-ice albedo and cloud-radiation interactions in the CMIP5 model ensemble, Geophys. Res. Lett., 40, 4374–4379, https://doi.org/10.1002/grl.50768, 2013.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J., Bates, S., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J. F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., Polvani, L., and Vertenstein, M.: 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. Met. Soc., 96, 1333–1349, https://doi.org/10.1175/BAMS-D-13-00255.1, 2015.

Knutti, R. and Sedláček, J.: Robustness and Uncertainties in the New CMIP5 Climate Model Projections, Nature Climate Change, 3, 369–373, https://doi.org/10.1038/NCLIMATE1716, 2013.

Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P., Hewitson, B., and Mearns, L.: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections, in: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections, edited by Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., and Midgley, P., IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, 2010a.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G.: Challenges in Combining Projections from Multiple Climate Models, J. Climate, 23, 2739–2758, https://doi.org/10.1175/2009JCLI3361.1, 2010b.

Knutti, R., Masson, D., and Gettelman, A.: Climate model genealogy: Generation CMIP5 and how we got there, Geophys. Res. Lett., 40, 1194–1199, https://doi.org/10.1002/grl.50256, 2013.

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

Kunreuther, H., Heal, G., Allen, M., Edenhofer, O., Field, C. B., and Yohe, G.: Risk management and climate change, Nature Climate Change, 3, 447–450, https://doi.org/10.1038/NCLIMATE1740, 2013.

Lehner, F., Deser, C., and Terray, L.: Towards a new estimate of "time of emergence" of anthropogenic warming: insights from dynamical adjustment and a large initial-condition model ensemble, J. Climate, 109, 14 337–14 342, https://doi.org/10.1175/JCLI-D-16-0792.1, 2017.

Li, G. and Xie, S.: Origins of tropical-wide SST biases in CMIP multi-model ensembles, Geophys. Res. Lett., 39, L22 703, https://doi.org/10.1029/2012GL053777, 2012.

Liu, C., Allan, R. P., and Huffman, G. J.: Co-variation of temperature and precipitation in CMIP5 models and satellite observations, Geophys. Res. Lett., 39, L13 803, https://doi.org/10.1029/2012GL052093, 2012.

Lorenz, R., Herger, N., Sedláček, J., Eyring, V., Fischer, E. M., and Knutti, R.: Prospects and caveats of weighting climate models for summer maximum temperature projections over North America, Journal of Geophysical Research: Atmospheres, 123, 4509–4526, https://doi.org/10.1029/2017JD027992, 2018.

Maher, N., Milinski, S., Suarez-Gutierrez, L., Botzet, M., Dobrynin, M., Kornblueh, L., Kröger, J., Takano, Y., Ghosh, R., Hedemann, C., Li, C., Li, H., Manzini, E., Notz, N., Putrasahan, D., Boysen, L., Claussen, M., Ilyina, T., Olonscheck, D., Raddatz, T., Stevens, B., and Marotzke, J.: The Max Planck Institute Grand Ensemble: Enabling the Exploration of Climate System Variability, Journal of Advances in Modeling Earth Systems, 11, 1–21, https://doi.org/10.1029/2019MS001639, 2019.

Masson, D. and Knutti, R.: Climate model genealogy, Geophys. Res. Lett., 38, L08 703, https://doi.org/10.1029/2011GL046864, 2011.

Meehl, G. A., Boer, G. J., Covey, C., Latif, M., and Stouffer, R. J.: The Coupled Model Intercomparison Project (CMIP), Bulletin of the American Meteorological Society, 81, 313–318, https://doi.org/http://www.jstor.org/stable/26215108., 2000.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., Matsumoto, K., Montzka, S. A., Raper, S. C. B., Riahi, K., Thomson, A., Velders, G. J. M., and van Vuuren, D. P.: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300, Climatic Change, 109, 213, https://doi.org/10.1007/s10584-011-0156-z, 2011.

Merrifield, A. and Xie, S.: Summer U.S. Surface Air Temperature Variability: Controlling Factors and AMIP Simulation Biases, J. Climate, 29, 5123–5139, https://doi.org/10.1175/JCLI-D-15-0705.1, 2016.

Merrifield, A. L., Lehner, F., Xie, S.-P., and Deser, C.: Removing circulation effects to assess central U.S. land-atmosphere interactions in the CESM Large Ensemble, Geophys. Res. Lett., 44, 9938–9946, https://doi.org/10.1002/2017GL074831, 2017.

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

Pennell, C. and Reichler, T.: On the Effective Number of Climate Models, J. Climate, 24, 2358–2367, https://doi.org/10.1175/2010JCLI3814.1, 2011.

Pithan, F., Medeiros, B., and Mauritsen, T.: Mixed-phase clouds cause climate model biases in Arctic wintertime temperature inversions, Clim Dyn, 43, 289–303, https://doi.org/10.1007/s00382-013-1964-9, 2014.

Poli, P., Hersbach, H., Dee, D., Berrisford, P., Simmons, A., Vitart, F., Laloyaux, P., Tan, D., Peubey, C., Thépaut, J., Trémolet, Y., Hólm, E., Bonavita, M., Isaksen, L., and Fisher, M.: ERA-20C: An Atmospheric Reanalysis of the Twentieth Century, J. Climate, 29, 4083–4097, https://doi.org/10.1175/JCLI-D-15-0556.1, 2016.

Rondeau-Genesse, G. and Braun, M.: Impact of internal variability on climate change for the upcoming decades: analysis of the CanESM2-LE and CESM-LE large ensembles, Climatic Change, p. 1–16, https://doi.org/10.1007/s10584-019-02550-2, 2019.

Sanderson, B. M., Knutti, R., and Caldwell, P.: Addressing interdependency in a multimodel ensemble by interpolation of model properties, Journal of Climate, 28, 2015a.

Sanderson, B. M., Knutti, R., and Caldwell, P.: A representative democracy to reduce interdependency in a multimodel ensemble, Journal of Climate, 28, 2015b.

Seneviratne, S. I., e. a.: Changes in climate extremes and their impacts on the natural physical environment, in: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC), edited by Field, C. B., Barros, V., Stocker, T. F., Qin, D., Dokken, D. J., Ebi, K. L., Mastrandrea, M. D., Mach, K. J., Plattner, G. K., Allen, S. K., Tignor, M., and Midgley, P. M., pp. 109–230, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2012.

Seneviratne, S., Corti, T., Davin, E., Hirschi, M., Jaeger, E., Lehner, I., Orlowsky, B., and Teuling, A.: Investigating soil moisture-climate interactions in a changing climate: a review, Earth-Sci. Rev., 99, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004, 2010.

Sillmann, J., Kharin, V. V., Zhang, X., Zwiers, F. W., and Bronaugh, D.: Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate, J. Geophys. Res. Atmos., 118, 1716–1733, https://doi.org/10.1002/jgrd.50203, 2013.

Sippel, S., Meinshausen, N., Merrifield, A., Lehner, F., Pendergrass, A., Fischer, E., and Knutti, R.: Uncovering the Forced Climate Response from a Single Ensemble Member Using Statistical Learning, J. Climate, 32, 5677–5699, https://doi.org/10.1175/JCLI-D-18-0882.1, 2019.

Stainforth, D., Allen, M., Tredger, E., and Smith, L.: Confidence, uncertainty and decision-support relevance in climate predictions, Philos Trans A, 265, 2145–2161, https://doi.org/10.1098/rsta.2007.2074, 2007.

https://doi.org/10.5194/esd-2019-69
Preprint. Discussion started: 21 November 2019

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblueh, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component of the MPI-M Earth System Model: ECHAM6, Journal of Advances in Modeling Earth Systems, 5, 146–172, https://doi.org/10.1002/jame.20015, 2013.

Stouffer, R., Eyring, V., Meehl, G., Bony, S., Senior, C., Stevens, B., and Taylor, K.: CMIP5 Scientific Gaps and Recommendations for CMIP6, Bull. Amer. Meteor. Soc., 98, 95–105, https://doi.org/10.1175/BAMS-D-15-00013.1, 2017.

Swart, N. C., Gille, S. T., Fyfe, J. C., and Gillett, N. P.: Recent Southern Ocean warming and freshening driven by greenhouse gas emissions and ozone depletion, Nature Geoscience, 11, 836–841, https://doi.org/10.1038/s41561-018-0226-1, 2018.

Tebaldi, C. and Knutti, R.: The Use of the Multi-Model Ensemble in Probabilistic Climate Change Projections, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences, 365, 2053–2075, https://doi.org/10.1098/rsta.2007.2076, 2007.

Trenberth, K. and Paolino, D.: The Northern Hemisphere Sea-Level Pressure Data Set: Trends, Errors and Discontinuities, Mon. Wea. Rev., 108, 855–872, https://doi.org/10.1175/1520-0493(1980)108<0855:TNHSLP>2.0.CO;2, 1980.

van den Dool, H. M.: Searching for analogues, how long must we wait?, Tellus A, 46, 314–324, https://doi.org/10.1034/j.1600-0870.1994.t01-2-00006.x, 1994.

van Vuuren, D., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., Hurtt, G. C., Kram, T., Krey, V., Lamarque, J.-F., Masui, T., Meinshausen, M., Nakicenovic, N., Smith, S. J., and Rose, S. K.: The representative concentration pathways: an overview, Climatic Change, 109, 5–31, https://doi.org/10.1007/s10584-011-0148-z, 2011.

Wallace, J., Fu, Q., Smoliak, B. V., Lin, P., and Johanson, C. M.: Simulated versus observed patterns of warming over the extratropical Northern Hemisphere continents during the cold season, Proc. Natl. Acad. Sci. (USA), 109, 14 337—14 342, https://doi.org/10.1073/pnas.1204875109, 2012.