Comparing Methods to Constrain Future European Climate Projections Using a Consistent Framework

LUKAS BRUNNER,[a] CAROL MCSWEENEY,[b] ANDREW P. BALLINGER,[c] DANIEL J. BEFORT,[d]
MARIANNA BENASSI,[e] BEN BOOTH,[b] ERIKA COPPOLA,[f] HYLKE DE VRIES,[g] GLEN HARRIS,[b]
GABRIELE C. HEGERL,[c] RETO KNUTTI,[a] GEERT LENDERINK,[g] JASON LOWE,[b] RITA NOGHEROTTO,[f]
CHRIS O'REILLY,[d] SAÏD QASMI,[h] AURÉLIEN RIBES,[h] PAOLO STOCCHI,[f,i] AND SABINE UNDORF[c,j]

[a] *Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland*
[b] *Met Office Hadley Centre, Exeter, United Kingdom*
[c] *School of GeoSciences, University of Edinburgh, Edinburgh, United Kingdom*
[d] *Atmospheric, Oceanic and Planetary Physics, Department of Physics, University of Oxford, Oxford, United Kingdom*
[e] *Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici, Bologna, Italy*
[f] *The Abdus Salam International Centre for Theoretical Physics, Trieste, Italy*
[g] *Royal Netherlands Meteorological Institute, De Bilt, Netherlands*
[h] *CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France*

(Manuscript received 25 December 2019, in final form 18 June 2020)

ABSTRACT

Political decisions, adaptation planning, and impact assessments need reliable estimates of future climate change and related uncertainties. To provide these estimates, different approaches to constrain, filter, or weight climate model projections into probabilistic distributions have been proposed. However, an assessment of multiple such methods to, for example, expose cases of agreement or disagreement, is often hindered by a lack of coordination, with methods focusing on a variety of variables, time periods, regions, or model pools. Here, a consistent framework is developed to allow a quantitative comparison of eight different methods; focus is given to summer temperature and precipitation change in three spatial regimes in Europe in 2041–60 relative to 1995–2014. The analysis draws on projections from several large ensembles, the CMIP5 multimodel ensemble, and perturbed physics ensembles, all using the high-emission scenario RCP8.5. The methods' key features are summarized, assumptions are discussed, and resulting constrained distributions are presented. Method agreement is found to be dependent on the investigated region but is generally higher for median changes than for the uncertainty ranges. This study, therefore, highlights the importance of providing clear context about how different methods affect the assessed uncertainty—in particular, the upper and lower percentiles that are of interest to risk-averse stakeholders. The comparison also exposes cases in which diverse lines of evidence lead to diverging constraints; additional work is needed to understand how the underlying differences between methods lead to such disagreements and to provide clear guidance to users.

Corresponding author: Lukas Brunner, lukas.brunner@env.ethz.ch

1. Introduction

Human-induced climate change calls for rapid cuts in anthropogenic greenhouse gas emissions to avoid increasingly negative impacts. Even with such reductions, however, climate will continue to change over the next decades, requiring reliable information about regional future changes for assessing impacts, identifying risks, and making adaptation decisions. The typical way of providing this information is by making estimates of the most likely change and known uncertainties based on an ensemble of climate models, often expressed as a probability. The uncertainties are primarily driven by three sources: uncertain future emissions, model uncertainty, and internal variability. The development of future emissions involves global political decisions such as the Paris Agreement (UNFCCC 2015) and technological developments and is not discussed here—we focus on the response to a given concentration pathway. The uncertainties associated with climate model responses to external forcings and internal variability, in turn, have been widely explored using ensemble modeling approaches.

Multimodel ensembles (MMEs) such as the Coupled Model Intercomparison Projects CMIP5 (Taylor et al. 2012) and the ongoing CMIP6 (Eyring et al. 2016) allow the exploration of a range of plausible future climate outcomes that result from differences in the way that climate models represent the physical world. The CMIP datasets, hence, form the basis for assessments of model uncertainty in many global climate change assessments (notably the IPCC assessment reports) (IPCC 2013) and regional or local studies, and drive downstream higher-resolution modeling activities such as EURO-CORDEX (Jacob et al. 2014). While MMEs capture structural differences between models, such "ensembles of opportunity" (Tebaldi and Knutti 2007) are not designed to sample uncertainty comprehensively. Additional uncertainty is associated with, among other things, the complex interdependencies of models (Knutti et al. 2013) and the possible range of parameter settings within a given model, reflected in so-called perturbed parameter ensembles (PPEs) of a single model (Sanderson et al. 2008).

Internal variability refers to natural variations of climate on all spatial and temporal scales beyond those of individual weather events. The importance of internal variability in total uncertainty is strongly dependent on the lead time, time period, spatial scale, and variable. It typically plays a larger role in the nearer term, at local spatial scales, and in spatially and temporally heterogeneous variables such as precipitation (Hawkins and Sutton 2009). Internal variability can be isolated from

model uncertainty using large ensembles, which provide multiple realizations of the same model with slightly different initialisations. CMIP5 models are provided with anywhere between 1 and 10 realizations, but there have also been efforts to explore internal variability more fully with dedicated experiments such as the 40-member CESM1-based NCAR Large Ensemble (NCAR-LENS) (Kay et al. 2015) or the 100-member MPI Grand Ensemble (MPI-GE) (Maher et al. 2019).

When combining members of an MME into a coherent projection of future change, and associated uncertainty, often a "democratic" approach is taken: each model is considered as independent (Pennell and Reichler 2011; Knutti et al. 2013; Sanderson et al. 2015a) and equally plausible (Gleckler et al. 2008; Eyring et al. 2019) and therefore contributes equally to the MME distribution. While few climate scientists would argue that such a "one model–one vote" democracy is the optimal way to represent uncertainty in an MME, it often remains the default in absence of a consensus on a more sophisticated approach. Notably, model democracy has been used to summarize projection information in high-level assessments, including the series of IPCC assessment reports in all but a few isolated cases (e.g., for Arctic sea ice projections) (Collins et al. 2013).

In recent years, more sophisticated methods have been developed to address the challenge of combining MME projections in order to better quantify uncertainty and improve reliability. Efforts have been made to identify the aspects of historical climate that are relevant to projection confidence and to use them to exclude, down-weight, or rescale future projections based on model performance. Such performance-based methods often include the assumptions that 1) poor agreement between a certain model and observations in a given variable and region is an indication that the model should be trusted less (e.g., Giorgi and Mearns 2003; Sanderson et al. 2015b; Knutti et al. 2017b) or that 2) there are emergent relationships linking present-day behavior (e.g., realistic simulations of the observed mean climate, or a trend) to future changes (e.g., Hall and Qu 2006; DeAngelis et al. 2015; Knutti et al. 2017a; Caldwell et al. 2018; Selten et al. 2020; Tokarska et al. 2020).

The European Climate Prediction system (EUCP) project aims to produce such improved projections of future European climate on a time horizon from the present to the middle of the century (Hewitt and Lowe 2018). This work builds toward the EUCP goal by developing a common framework to compare different methods, by investigating underlying method properties, and by highlighting cases of high and low agreement

across methods in terms of their output distributions. We analyze eight methods from groups involved in EUCP that are used to represent the diverse set of existing approaches to constrain regional climate projections, including 1) methods that weight models based on their performance in reproducing observed mean, variability, or trend fields in one or more variables (e.g., Giorgi and Mearns 2003; Bishop and Abramowitz 2013; Knutti et al. 2017b; Sanderson et al. 2017; Merrifield et al. 2019; Amos et al. 2020); 2) detection and attribution-based methods, which scale models based on their representation of past forced changes from one or more sources such as anthropogenic CO_2 emissions (e.g., Allen et al. 2000; Stott and Kettleborough 2002; Kettleborough et al. 2007; Shiogama et al. 2016; Li et al. 2017; Tokarska et al. 2019); 3) Bayesian methods that update a prior distribution in light of new information provided by observations (Cressie 1991; Rougier et al. 2013; Renoult et al. 2020; Ribes et al. 2020, manuscript submitted to *Sci. Adv.*); and 4) single-model methods that focus on investigating internal variability not accounting for model uncertainty (Deser et al. 2012a,b, 2014; Martel et al. 2018; O'Reilly et al. 2020, manuscript submitted to *Earth Sys. Dyn.*).

Beyond that, there are other probabilistic methods available, which have been used both in academic studies and to produce climate projection data, that are not explored in this work—for example, more process-based emergent constraints that are often tailored to a specific application (e.g., Vogel et al. 2018; Hall et al. 2019; Eyring et al. 2019; Selten et al. 2020) and therefore harder to apply across a range of different settings. Still, with this study EUCP brings together a number of methods, provided by partners within the project, that represent a large and diverse ad hoc ensemble of opportunity to assess consistency of climate projection information.

Rigorously comparing different methods based on their results published in the literature alone is often very challenging or even impossible. Studies applying individual methods are typically not performed in a coordinated framework (as it exists for the model experiments themselves within CMIP, for example) and are not focused on enabling easy method intercomparison. Therefore, even when two methods investigate the same general target variable (such as temperature change) and region (such as Europe), a consistent comparison may be hindered by subtle differences in their 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.

Here we therefore develop a common experiment setup, including a defined set of European subregions at different spatial scales, a common time period, and set of variables to provide a level testing ground for different methods as far as possible. We then use this common setting to provide an evaluation of agreement (or a lack thereof) across the eight methods included. Exposing cases where the different lines of evidence, used by the methods, lead to diverging results is crucial as it highlights instances where any single method (or even the raw ensemble spread) is potentially overconfident. For cases with high agreement across multiple methods, in turn, we can have increased confidence in the robustness of the constrained distributions.

Ultimately, a framework to select or interpret distributions resulting from different lines of evidence is needed to provide clear guidance to users. However, developing such a framework is beyond the scope of this first intercomparison and we here limit ourselves to detailing the methods' fundamental differences in terms of underpinning assumptions, uncertainty sources captured, and applications of observational constraints to shed light on the reasons for method agreement or disagreement. In summary, we develop a common framework for consistently comparing a diverse ''method ensemble of opportunity'' to constrain European climate projections and to investigate method agreement (or a lack thereof) for temperature and precipitation change in several European regions in terms of median and uncertainty range. We summarize the underlying assumptions in the methods that can lead to differences in their constrained distributions and discuss possible ways forward.

2. Approaches to uncertainty quantification

In this section we describe the main properties of the eight methods to be compared. We will refer to the methods using their acronyms throughout the paper, a summary of these acronyms, full names, institutions, and reference publications can be found in Table 1. The methods' key features and assumptions are summarized in Table 2. There are a number of ways in which they might be categorized (e.g., Lopez et al. 2015); for the purpose of this study, we broadly divide them as follows:

1) weighting schemes: ClimWIP (Climate Model Weighting by Independence and Performance) and REA (reliability ensemble averaging),
2) detection and attribution-based methods: ASK (Allen–Stott–Kettleborough),

TABLE 1. Participating institutions, methods, and reference publications. Methods marked with an asterisk focus only on internal variability.

Institution name	Method acronym	Method name	References
ETH Zurich (Switzerland)	ClimWIP	Climate Model Weighting by Independence and Performance	Knutti et al. (2017b); Lorenz et al. (2018); Brunner et al. (2019)[a]
International Centre for Theoretical Physics (Italy)	REA	Reliability ensemble averaging	Giorgi and Mearns (2002, 2003)[b]
University of Edinburgh (United Kingdom)	ASK	Allen–Stott–Kettleborough	Allen et al. (2000); Stott and Kettleborough (2002); Kettleborough et al. (2007)
Centre National de Recherches Météorologiques (France)	HistC	Historically constrained probabilistic projections	Ribes et al. (2020, manuscript submitted to *Sci. Adv.*)[c]
Met Office (United Kingdom)	UKCP	U.K. Climate Projections (UKCP) Bayesian probabilistic projections method	Sexton et al. (2012); Harris et al. (2013); Sexton and Harris (2015); Murphy et al. (2018)
University of Oxford (United Kingdom)	CALL	Calibrated large ensemble projections	O'Reilly et al. (2020)
Royal Netherlands Meteorological Institute (Netherlands)	BNV[*]	Bootstrapped from natural variability	See the online supplemental material
Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (Italy)	ENA[*]	Ensemble analysis of probability distributions	See the online supplemental material

[a] Source code available online (https://github.com/lukasbrunner/ClimWIP).
[b] Source code available online (http://doi.org/10.5281/zenodo.3890966).
[c] Method tool available online (https://saidqasmi.shinyapps.io/bayesian).

3) Bayesian methods: HistC (historically constrained probabilistic projections) and UKCP (U.K. Climate Projections Bayesian probabilistic projections method), and

4) single-model methods: BNV (bootstrapped from natural variability), ENA (ensemble analysis of probability distributions), and CALL (calibrated large ensemble projections).

a. Weighting schemes (ClimWIP and REA)

The ClimWIP and REA approaches arrive at a probability distribution by applying a weighting scheme based on model performance and independence, but determine them by quite different metrics. For the model performance component, REA applies weights on a variable-by-variable basis (i.e., projections of precipitation are weighted according to local precipitation performance) while ClimWIP uses a set of six diagnostics based on a number of variables (temperature climatology, precipitation climatology, shortwave downward radiation climatology, shortwave upward radiation climatology and variance, and longwave downward radiation variance) (Brunner et al. 2019). The number and selection of these diagnostics follow Lorenz et al. (2018), who identify relevant diagnostics for projections of maximum temperature.

The two schemes also utilize different treatments of independence. ClimWIP determines model independence weights based on quantified distances between models (essentially downweighting models with high interdependence) (Knutti et al. 2017b). Recently Brunner et al. (2020, manuscript submitted to *Earth Syst. Dyn.*) have looked into the effect of this approach in more detail using ensemble members as separate models with (known) high interdependence. Conversely, REA treats model convergence as an indication of projection confidence, effectively downweighting outliers in projection space (Giorgi and Mearns 2002; Tegegne et al. 2019).

b. Detection and attribution-based methods (ASK)

ASK methods are based on a framework derived by Allen et al. (2000), Stott and Kettleborough (2002), and Kettleborough et al. (2007) in which fingerprint techniques used in detection and attribution (Allen and Stott 2003; Polson et al. 2013; Knutson 2017) are applied to the problem of uncertainty quantification in future projections. The space–time pattern of the response to a given external forcing as simulated by an ensemble mean is scaled to values consistent with observations. The underpinning assumption is that, because the magnitude of the response is uncertain due to uncertain feedbacks, estimating that magnitude from observations of change is important. The pattern of response, in turn, is assumed to be governed by the physics of the forcing response (i.e., the climate inertia) and further influence by aerosol response—and therefore correctly reflected in climate models. This method has, for example, been used to provide constrained projections of near-term

TABLE 2. Summary of data used by methods as well as most important features and limitations.

Acronym	Model data	Observational constraints	Treatment of model dependencies	Key assumptions
ClimWIP	CMIP5	Weighted based on historical performance of six diagnostics	Weighted based on historical independence from other models	Future model performance can be inferred from historical performance; interdependence of models can be inferred from model outputs
REA	CMIP5	Weighted based on historical performance of the target variable	Weighted based on the distance to the MME mean	Future model performance can be inferred from historical performance on a variable-by-variable basis; ensemble is truth centered
ASK	CMIP5	Scaled based on observed time–space change over the historical period.	—	Space–time pattern of climate response to forcing is governed by known physics and is correctly represented in models (which may not be true, e.g., for aerosols), whereas the amplitude is governed by uncertain feedbacks and is, hence, estimated from observations
HistC	CMIP5	Constrained based on historical warming trend	—	Real-world response to forcings is statistically indistinguishable from model responses; response to anthropogenic forcing is smooth over time
UKCP	CMIP5 and PPEs	Constrained based on the climatology of 12 variables and historical trends in surface temperature, upper ocean heat content, and CO_2 concentration	Uncertainties systematically sampled in a PPE framework	Future model performance can be inferred from historical performance; true climate lies within range of the sampled prior outcomes; patterns of equilibrium response are representative of the fully coupled response patterns; transient responses scale linearly with global temperature response
BNV	Single model	—	—	The spread obtained from single-model large ensembles is a measure of uncertainty due to internal variability
ENA	Single model	—	—	The spread obtained from single-model large ensembles is a measure of uncertainty due to internal variability
CALL	Single model	Ensemble projection distribution is scaled to optimize fit to observations	—	The relationship between the past evolution of the ensemble dataset and the observations contains meaningful information for the future evolution

(up to 2035) global temperatures in the IPCC's Fifth Assessment Report (Kirtman et al. 2013). While commonly used to constrain global temperature projections, the method has also been applied regionally (Stott et al. 2006). The ASK scaling factors are derived from the time pattern of change over the three European subregions and are thus not an overly strong constraint on the spatial pattern of the fingerprint. For temperature,

fingerprints are optimized in line with the literature, while nonoptimized data are used for precipitation where nonnormality is a serious concern (see also Polson et al. 2013; Schurer et al. 2020).

Here, ASK-ANT and ASK-GHG demonstrate the method when constraining the response to different external forcings, that is, the combined anthropogenic forcing (ANT) or that from greenhouse gases only

(GHG) as derived from single-forcing model experiments. Using the response to anthropogenic forcing to derive the scaling factor range does not account for variations of aerosol/GHG ratios in time (van Vuuren et al. 2011; Gidden et al. 2019), which limits the method's constraining power (Shiogama et al. 2016). However, using the response to GHG forcing only neglects the fact that future projections also include aerosols and other anthropogenic forcings. In addition to the uncertainty included in the results presented, there are potential sensitivities associated with the choices made in the application of the method such as the historical time period used or the noise sampling process (e.g., Allen and Tett 1999), the latter of which may be addressed by inflating the variance (Schurer et al. 2018).

The ASK methods give constraints on the estimated spread in the forced component of the future projections, to which uncertainty from internal variability is added to be consistent with the other methods. This is done by applying a Monte Carlo sampling approach to the scaling factor uncertainty and samples of temperature/precipitation change over two periods of same length and time distance as the baseline and future period used for this study from preindustrial control simulations. The ASK approach relies on a forced signal being detectable in observations, and is, therefore, applicable only where the forced signal has emerged. This is practically limited by the MME size, the availability of observations, and a large enough target region given internal variability. This requirement is unlikely to be met for the local scale; in order to be able to include this approach to the test case of local scales also investigated in this study, scaling factors are derived using larger-scale information and applied to the smaller scales.

c. Bayesian methods (HistC, UKCP)

The HistC approach proposed in Ribes et al (2020, manuscript submitted to *Sci. Adv.*) combines some of the principles of detection and attribution-based constraints (Ribes et al. 2017) and Bayesian probability theory. In HistC, 1) the forced response of each CMIP5 model is estimated in the historical period using a generalized additive model where the response to natural forcings is calculated using an energy balance model, and anthropogenic influence is assumed to be smooth in time; then 2) a multimodel distribution that characterizes the model uncertainty in this forced response is constructed (the "prior"); and finally 3) a historical constraint is applied, to subselect those trajectories that are consistent with available observations, given internal variability (the "posterior"). This approach accounts explicitly for the climate model uncertainty, which is challenging to account for in a regression-based

detection and attribution approach (ASK), while assuming that models are statistically indistinguishable from the truth. In cases where there is no detectable signal in the observations, the posterior will be equal to the prior, such that little or no constraint is applied.

UKCP also applies a Bayesian approach to produce probabilistic projections. In contrast to other methods that use the empirical spread of CMIP5 projections to represent prior model uncertainty, UKCP uses a statistical emulator trained on a single-model perturbed physics ensemble. This provides a more systematic and comprehensive sampling of climate responses by allowing a larger sample size in the emulated ensemble and structured sampling of uncertainties. Further, by basing the simulations on the emission-driven representative concentration pathway 8.5 (RCP8.5) scenario simulations (as opposed to the concentration-driven used in the other methods) and drawing from a second perturbed physic ensemble of Earth system model variants, this method samples additional uncertainties associated with the carbon cycle. This inclusion of additional uncertainties differentiates UKCP from the other methods described here. To further sample the structural error component associated with using a single perturbed physics model, CMIP5 Earth system model simulations are used to define an additional "discrepancy" term (Sexton et al. 2012). This methodology means that unconstrained distributions are wider than for the other methods.

Observational constraints are applied by weighting sampled outcomes by likelihood weights calculated from multivariate distances to observations. The observations comprise 12 climate variables reduced in dimensionality to 6 leading eigenvectors. In addition, historical trends for several climate indicators are also considered in the set of observational constraints, including the Braganza indices based on global mean surface temperature (Braganza et al. 2003), heat content change in the top 700 m of the oceans, and change in atmospheric CO_2 concentration over a recent 45-yr period (Booth et al. 2017). The separate components of the method are validated and the additional statistical uncertainties that arise are included at each stage. These include equilibrium response emulation error, error in converting from equilibrium to transient response, time-scaling error (including inherent model internal variability), and structural error estimates.

d. Single-model methods (BNV, CALL, and ENA)

The single-model methods make use of large ensembles, originally designed to explore the internal variability in the climate system. BNV and ENA are both intended to quantify and characterize the role of internal variability. They use different base models which have different estimates of the size of the forced response

and/or internal variability. BNV and ENA both represent the internal variability around a single-model ensemble mean without accounting for climate model uncertainty. ENA is a simple interpretation of the range of projections based on the NCAR-LENS (40 ensemble members) and the MPI-GE (100 ensemble members). BNV, based on a 16-member ensemble of EC-Earth (Aalbers et al. 2018), estimates internal variability using a bootstrapping approach, which consists of sampling a large number (10^4) of possible time series (bootstrap members) by drawing randomly from the entire ensemble (sampling with replacement). The internal variability is estimated from this bootstrap ensemble, rather than from the range of the raw ensemble projections. As a result, BNV gives more accurate results than simply using each member once, especially when the ensemble size is small.

The third method, CALL, uses a calibration approach to observed climate to extract an estimate of constrained climate change in the future (O'Reilly et al. 2020, manuscript submitted to *Earth Sys. Dyn.*). If the structural model error relative to observations is large this method can scale the future responses outside their original range; hence, it captures uncertainties in the future climate change response, as well as internal variability. In this sense, the output from this approach can be seen as more directly comparable with the results from the earlier multimodel methods. CALL also makes use of the NCAR-LENS model ensemble, rescaling the ensemble mean and spread to observations over a reference period in order to maximize the reliability over the observed period and provide a more reliable future projection range. The ensemble data are first decomposed into dynamical and residual components [following the method of Deser et al. (2016)] in order to avoid conflating forced response with variability and each component is then calibrated using homogeneous Gaussian regression before being combined to give the total projection [see O'Reilly et al. 2020, manuscript submitted to *Earth Sys. Dyn.*, section 2c(4) therein].

3. Introducing a consistent testing framework

A common set of variables, seasons, regions, and periods as well as a default processing order is introduced to allow for a quantitative and consistent comparison of results from the different methods. These settings are selected to 1) maximize the possible contributions by each method (not every method is able to deal with all variables, regions, etc.), 2) compare the methods at different spatial scales and for different variables, and 3) produce relevant results for the scientific community and policy-makers. Our aim is to maximize consistency

in the comparison from the raw data to the unconstrained distributions, and further to the constrained distributions of change. This necessarily means striking compromises between the preferred setup for a given method and the preferred setup for method intercomparison. Full adherence to a single standard is sometimes even impossible due to specific method requirements. Deviations stem, for example, from the need for single-forcing runs by the ASK approach, which restrict the model pool usable by this method. Indeed, the use of different subsets of the CMIP5 MME has been identified as a main source of deviations between the (unconstrained) distributions and we specifically address this in section 4d.

For the main comparison the methods use slightly differing model pools, with most of the results being based on multiple CMIP5 generation models using RCP8.5 forcing. ASK-ANT and REA use the same 10 models (29 runs), ASK-GHG uses one model less (9 models; 28 runs), and ClimWIP and HistC use 37 models (79 runs). In addition, ClimWIP and HistC both also use a subset of the same 10 models (29 runs) as ASK-ANT and REA, which allows a better comparison of results but potentially also limits optimal method performance. UKCP additionally uses perturbed physics ensembles as described in detail in section 2c. The remaining methods are based on large ensembles: 16 runs from EC-Earth (BNV), 100 runs from the MPI-GE (ENA MPI-GE), and 40 runs from NCAR-LENS (CALL and ENA CESM). A full list of models used by each method can be found in Table S1 in the online supplemental material. Note that while we refer to *multimodel methods* (ASK, ClimWIP, HistC, REA, and UKCP) and *single-model methods* (BNV, ENA, CALL), this only reflects the setup in this study. For example, the multimodel methods might also be applied to large ensembles like in Merrifield et al. (2019) where ClimWIP is applied to several large ensembles to explore model independence and the influence of internal variability.

Performance in the historical period is measured against a range of observational datasets in different methods. These include "direct" observations such as E-OBS, HadCRUT4, and CRU-TS3, as well as the reanalysis datasets ERA-Interim and MERRA2. For a detailed list of observations used by each method and their reference publications see Table 3.

All data are regridded to a regular 2.5° × 2.5° latitude–longitude grid using bilinear remapping. Then an ocean mask based on gridcell centers is applied and the regions are selected. We compare results for eight regions throughout Europe, representing three distinctively different spatial aggregation scales (Fig. 1). As bases we use the three European "SREX" regions (Field et al.

TABLE 3. Observational datasets and reference publications used by the different methods. Note that UKCP uses a large range of observations beyond this list that is detailed in Table B.1 of Murphy et al. (2018).

Dataset	Used by	Reference
CERES	ClimWIP; UKCP	Kato et al. (2013)
CRU TS 3	CALL; REA	Harris et al. (2014)
E-OBS v17e	ClimWIP	Cornes et al. (2018)
E-OBS v19e	ASK	Cornes et al. (2018)
ERA-Interim	ClimWIP	Dee et al. (2011)
GPCC	HistC	Schneider et al. (2017)
HadCRUT4	HistC; UKCP	Morice et al. (2012)
MERRA2	ClimWIP	Gelaro et al. (2017)

2012), which constitute well-established climatic regions and are defined here as medium-sized regions: "Northern Europe (NEU)," "Central Europe (CEU)," and the "Mediterranean (MED)." The combined European region (EUR: NEU + CEU + MED) is used as a large, continental-scale region. Last, the methods are applied to four "local" regions at the scale of a single grid cell, chosen to reflect different responses in summer temperature and precipitation throughout Europe: Falun, Sweden (FAL; in NEU); Dusseldorf, Germany (DUS; in the northwest of CEU); Sibiu, Romania (SIB; in the southeast of CEU); and Madrid, Spain (MAD; in MED). Our motivation in this is to test if the methods are able to produce robust results at such scales, driven by user needs which are often focused on more local scales.

We then apply the methods using the common setup and compare their results. From this analysis we draw initial information about the robustness of the results, which we infer from agreement on the median change and related uncertainties across methods. Our consistent framework allows tracing back cases in which the methods disagree to underlying differences in the methods (isolated from other sources such as differing regions, etc.) and we discuss these differences in the second part of the paper.

To compare the methods we show probability distributions of change in area-averaged summer (July–August) temperature and relative precipitation between reference (1995–2014) and future (2041–60) mean states. The distributions are based on the 10th, 25th, 50th, 75th, and 90th percentiles, which are calculated empirically (ClimWIP, ENA, REA), using bootstrap samples (BNV, CALL, HistC, UKCP), or using scaling factors applied to the multimodel mean (ASK). The CMIP5 distributions represent a combination of model uncertainty and internal variability; HistC and ASK isolate the forced response during processing but in order to allow a better comparison internal variability is added again at the end. UKCP includes additional
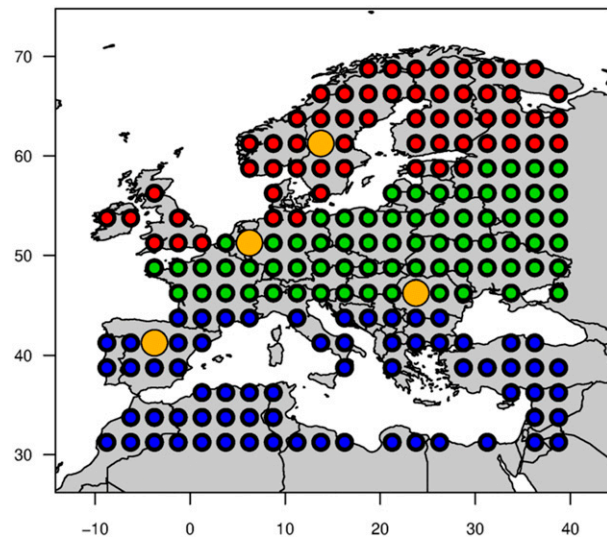

FIG. 1. Common grid and regions: Northern Europe (NEU; red dots), Central Europe (CEU; green dots), and the Mediterranean (MED; blue dots). The single grid cells are indicated by yellow dots and refer to (from the left): MAD, DUS, FAL, and SIB.

parameter and carbon cycle uncertainty while the single-model methods only sample internal variability.

To provide context for the projected changes, we also show an estimate of the 20-yr internal variability based on observations. The CMIP5 multimodel mean is calculated (based on the HistC model pool) and subtracted from the HadCRUT4 and GPCC time series for temperature and precipitation, respectively. The residuals are smoothed using a 20-yr running average, then the 10th, 25th, 50th, 75th, and 90th percentiles over the 1914–2013 period are calculated as decadal-scale internal variability estimates. We are aware that there are many different ways of providing such estimates for the internal variability, based on large ensembles, MMEs, and observations. However, since a discussion of internal variability is not the main focus of this study, the choice of the selected estimate is mainly based on its simplicity here.

4. Results

a. Temperature projections

From the multimodel methods in the SREX and combined European regions (Fig. 2), we see a reduction in the 25th–75th and the 10th–90th percentile ranges (jointly referred to as "spread" hereafter) by 20%–30% on average over all methods. This reduction can exceed 50% in individual regions and methods (e.g., ASK in NEU or REA in CEU) but in most cases the change is considerably smaller, and occasionally even an increase

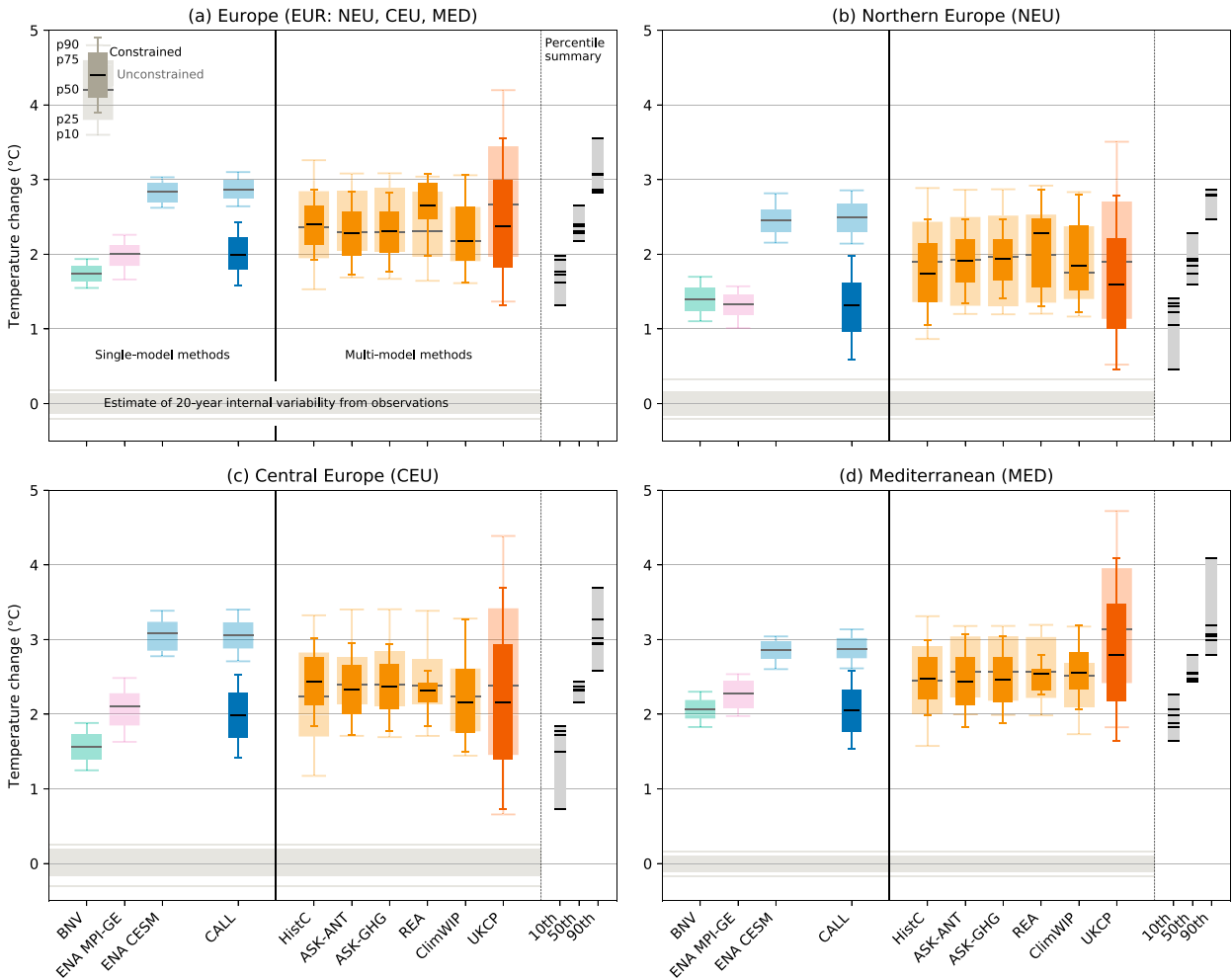## Summer (JJA) temperature change (2041-2060 minus 1995-2014)

FIG. 2. 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. The lighter boxes give the unconstrained distributions; the darker boxes give the constrained distributions. The colors indicate methods based on similar model pools: single-model ensembles (green: EC-Earth, magenta: MPI-GE, and blue: NCAR-LENS), CMIP5 (orange), and CMIP5 and PPE (red). The gray box and lines centered around zero show percentiles of 20-yr internal variability based on observations. A synthesis of all constrained multimodel distributions (excluding single-model methods) is shown on the rightmost side. The bars represent the 10th, 50th, and 90th percentiles of the methods, and the shading indicates the full spread.

in spread is found (e.g., ClimWIP and ASK-ANT in CEU). For the absolute values of the 10th and 90th percentiles (see percentile summary in Fig. 2) agreement between methods is low and the full range of values can exceed 1°C. This partly reflects the different components of uncertainty included in different methods, notably the additional treatment of carbon cycle uncertainty in UKCP, which contributes to a wider uncertainty estimate than any other method. Further differences in the underlying assumptions of the methods are discussed in section 5 and are summarized in Table 4, which is discussed in more detail in that section.

The median estimates agree better, in particular in CEU and MED where a range of 2.2°–2.4°C and of 2.4°–2.8°C is found, respectively. These results give additional confidence in the "best estimate" of change based on the different lines of evidence used by the methods. Similar considerations are true for the combined European domain, where again most methods agree with only REA showing slightly stronger warming widening the full range of medians to 2.2°–2.7°C. The largest disagreement in the median estimates is found for NEU with a full range of 1.6°–2.3°C.

For all four regions, the change in temperature by the middle of the century clearly emerges from the decadal-scale

internal variability estimated from HadCRUT4. In addition, the unconstrained distributions of the single-model methods also provide an estimate of remaining internal variability based on three large ensembles (EC-Earth, MPI-GE, NCAR-LENS). The unconstrained projections by the single-model methods are not discussed in further detail here and are mainly shown to provide context. BNV and ENA present unconstrained temperature differences based on three different large ensembles. CALL, in addition, presents a calibration approach, leading to the largest shifts of the distributions by about −1°C in combination with a doubling to tripling of the unconstrained spread (see also O'Reilly et al. 2020, manuscript submitted to *Earth Sys. Dyn.*, their Fig. 9). As a result the CALL calibration brings both median and spread closer to the respective mean values over the multimodel methods. Note that the two NCAR-LENS based unconstrained distributions (ENA CESM and CALL) differ slightly since ENA calculates percentiles directly, while CALL uses a bootstrapping approach.

b. Precipitation projections

The constrained precipitation distributions shown in Fig. 3 differ considerably throughout Europe, particularly in the tails. In NEU most methods lead to a considerably reduced spread and agree on a slight increase in the median precipitation estimate by the middle of the century, which lies within the range expected from internal variability. ClimWIP and UKCP, in contrast, revise the projected median precipitation change downward and also constrain only the upper percentiles, notably retaining projections of reduced rainfall. For CEU and MED, all methods agree on a median projection that points to a reduction in rainfall mostly exceeding present-day variability. In CEU, the magnitude of the median change is from approximately −5% to −10% except for HistC, which points toward no change. REA, in addition, strongly constrains the spread to less than half of the unconstrained one, while the reduction in spread for UKCP is below 10%. In MED there is little consensus on either the strength of the projected median change (ranging from −10% to −25%) or the uncertainty ranges, indicating considerable uncertainty across methods. Notably, the ASK-GHG constrained range exceeds the unconstrained one and both the UKCP and ASK-GHG methods indicate that drying signals in MED are stronger than those captured in the empirical CMIP5 range.

These wide range of results (in terms of median change as well as uncertainty) are challenging to interpret and clearly need additional research to disentangle. Some discussion of underlying method differences can be found in section 5. Here, we briefly mention two

characteristics that apply across all regions. First, the methods using multiple constraining metrics (ClimWIP and UKCP) exert considerably less impact on the projection range than those that depend on a single metric of precipitation. Discussions of multiple versus single metric approaches in existing literature suggest that single metrics lead to stronger constraints but might also offer overconfident projection ranges, in particular when they are not carefully selected (Sanderson et al. 2017; Lorenz et al. 2018; Brunner et al. 2020, manuscript submitted to *Earth Syst. Dyn.*). Second, these diverse projections still mostly remain within the considerably wider range of the UKCP projections, which as well as being a multimetric method, explicitly includes quantified estimates of more sources of uncertainty than other methods. Any definitive communication of results in such a scenario should ideally include some kind of recommendation on how to select or interpret the results to reconcile these differences, for example, based on a measure of method skill. However, in this first study we focus on identifying cases where the different lines of evidence drawn on from the methods lead to diverging results (having reduced the influence of other factors as much as possible) and stress the importance of carefully considering the choice of method based on the application (e.g., whether median changes are important versus a "worst-case" scenario).

c. Applying the methods to the gridcell scale

To test the applicability of the methods without any spatial aggregation, we also apply them to single grid cells (Figs. 4 and 5). Our aim is to check how the methods behave at local scales, without necessarily assuming that the results are physically meaningful. Unsurprisingly, we find that uncertainty is generally higher and the methods agree less on the potential for spread reduction and on the projected median change. The methods provide reasonably robust median temperature projections even on a gridcell scale. For precipitation, the methods mostly agree on the sign of the change and to a certain degree even on the magnitude of the projected change. Crucially, however, they do not agree on the spread, with several methods leading to hardly any reductions in spread, while, for example, REA shows spread reductions of more than 50% in many regions. The question of whether this is a realistically narrow uncertainty range or if there is little potential for spread reduction clearly needs further investigation.

Here we do not test method performance in detail, so a formal quantitative comparison of method performance over different spatial regimes remains beyond the scope of this study. However, establishing a performance metric using, for example, a perfect model test can estimate this, as done by
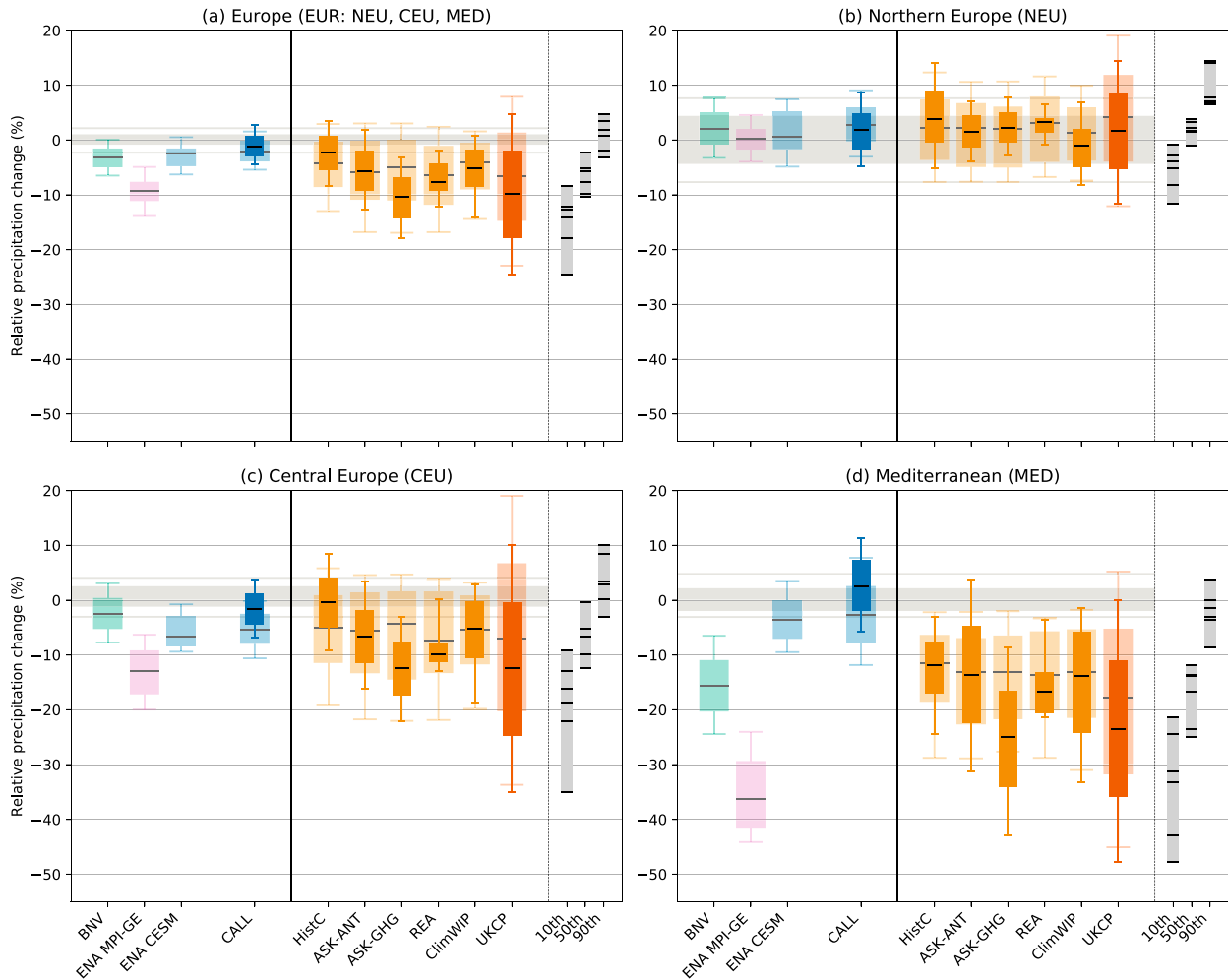
FIG. 3. As in Fig. 2, but for relative precipitation change.

Brunner et al. (2019) for ClimWIP. By looking into changes in the continuous ranked probability score (Hersbach 2000) and using a perfect model approach, Brunner et al. (2019) find that some skill can be gained by weighting or subselecting based on the ClimWIP method even on a gridcell level. However, they also highlight that there is a considerable risk of being overconfident when constraining change over regions that small.

d. The impact of working with different model subsets

Unconstrained temperature and relative precipitation distributions show considerable differences among the methods (light colored boxes in Figs. 2–5). Most of these differences can simply be attributed to the processes covered or to the fact that different subsets of CMIP5 are used (cf. Table S1). In fact, the unconstrained median change is remarkably consistent across all CMIP5 methods considering the different subsets. It has been

argued that the exact selection of models from an MME might not have a huge influence on the projections, given a large enough subset (Knutti 2010; Herger et al. 2018). Still, when producing a constrained distribution of change using any given method, it is reasonable to use as many models as possible to maximize method performance (as has been done for all results shown so far). In this section we control for the effect of different subsets and consider a case in which each of the CMIP5 methods uses a common pool of models even though this might not be the ideal setup of any given method. Doing so, the unconstrained distributions are identical by design except for HistC. The differences for the HistC distribution arise because 1) it uses a Gaussian fit to derive percentiles and not the MME itself and because 2) internal variability is estimated from preindustrial control runs and added to the extracted forced response for each model.
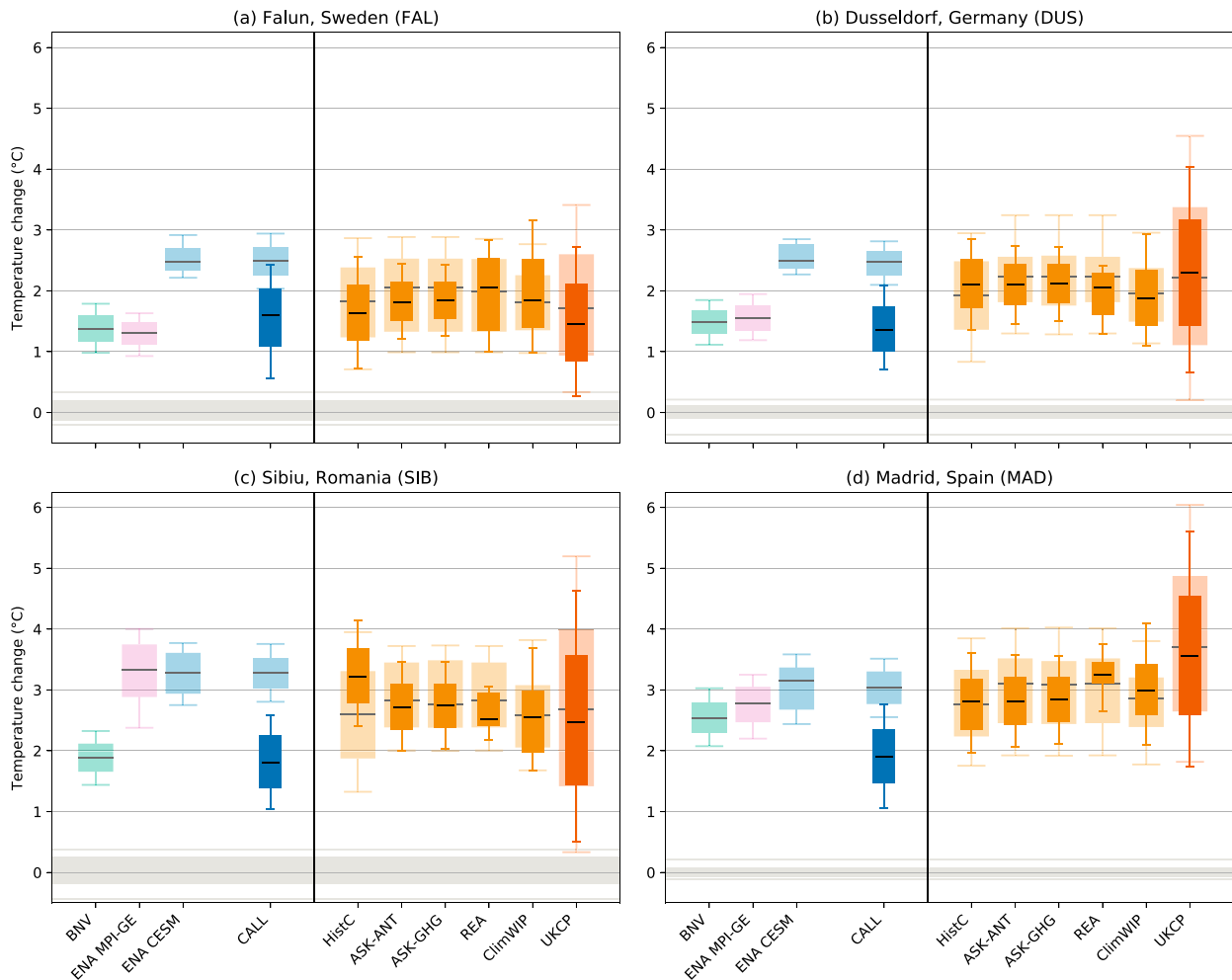
FIG. 4. As in Fig. 2, but for the four gridcell regions. Note that the *y* axis differs from that in Fig. 2.

Figure 6 shows selected results for temperature based on the same 29 simulations in the combined European and the three SREX regions. Similar to the case with differing model pools presented earlier, one common feature of the methods included in this comparison is a reduction in the spread from the unconstrained distributions by up to 50% or more. In general, the highest method agreement is found in CEU with the warmer ends of the distributions being reduced by about −0.3° and −0.6° on average for the 75th and 90th percentiles, respectively. Based on this consensus, it can be concluded that a warming of more than 3°C by the middle of the century is unlikely in CEU even under RCP8.5. Similar results are found for MED with all methods except ClimWIP strongly constraining the upper percentiles.

In NEU and the combined European region in turn, HistC leads to a slight downward shift of the unconstrained

median while REA shifts it upward and for ASK-ANT and ClimWIP hardly any change in the median is found. From this setup we can now attribute the differences in the constrained distributions solely to the application of different methods. The changes in the location of the median (between the unconstrained and constrained distributions) are clearly inconsistent, ranging from reduced warming (HistC), or no shift (ASK-ANT and ClimWIP), to enhanced warming (REA). Similarly for the estimate of uncertainly, three methods point to a reduction (ASK-ANT, HistC, and REA), while ClimWIP suggests little change from the unconstrained distribution.

All four methods discussed here are arguably based on observational constraints that can be physically justified and scientifically defended even if none of them is completely without caveats (Giorgi and Mearns 2002; Shiogama et al. 2016; Brunner et al. 2019; Ribes et al. 2020, manuscript submitted to *Sci. Adv.*). A

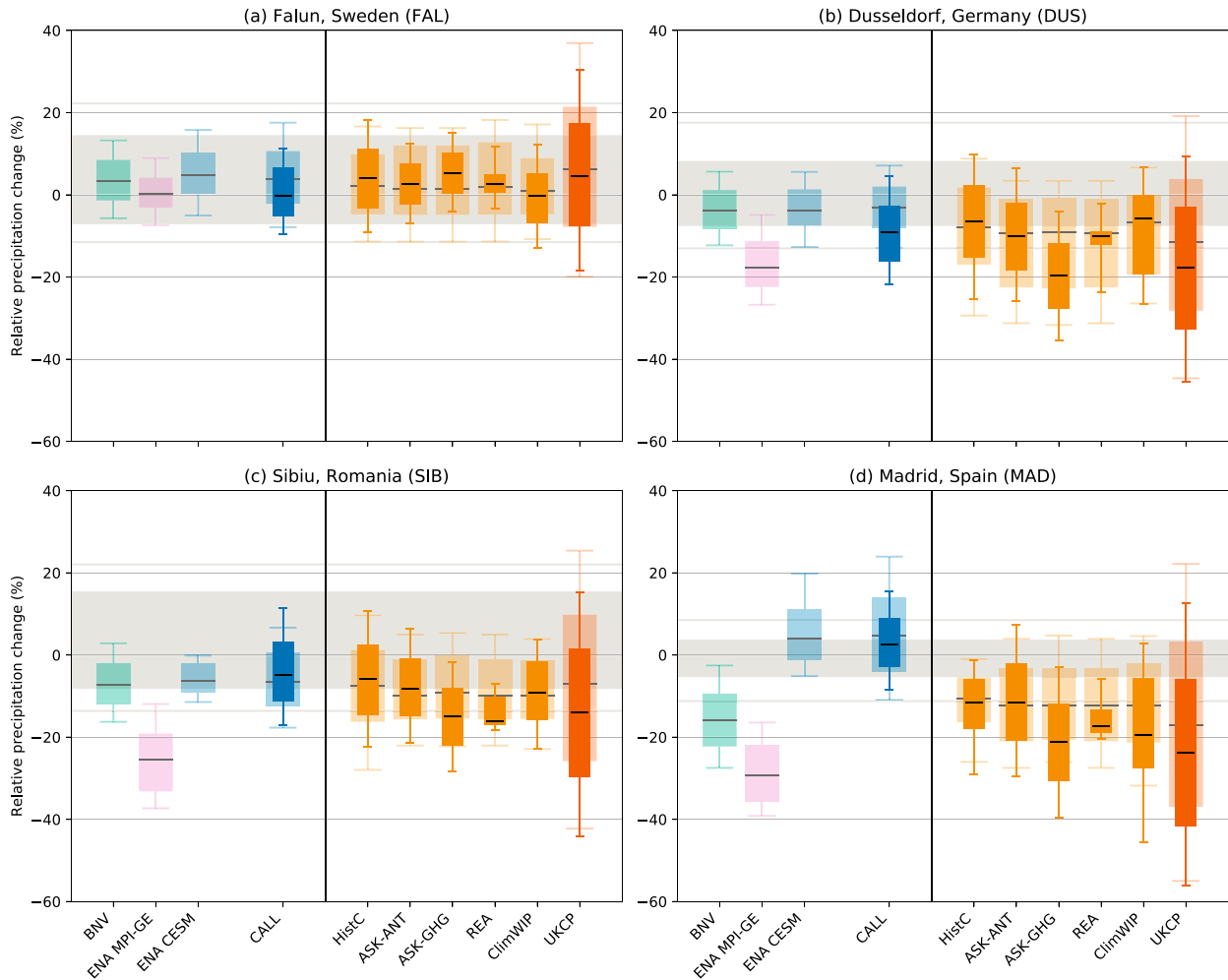Summer (JJA) precipitation change 2041-2060 relative to 1995-2014

FIG. 5. As in Fig. 3, but for the four gridcell regions. Note that the *y* axis differs from that in Fig. 3.

user could hence legitimately use any one of the constrained distributions and expect an improvement in skill over the raw model spread. Our results show that doing so has a high likelihood of success in two regions (CEU and MED), where all methods agree but, crucially, does not in NEU or in the combined European region. Despite not having a framework to resolve this discrepancy here, this finding nonetheless carries an important warning about only applying a given single method in the latter two regions.

e. The impact of model dependence in ClimWIP and REA

Here we use the examples of ClimWIP and REA to explore how the treatment of model dependence affects the resulting distributions. The convergence criterion

applied by REA is, more generally, often referred to as "truth centered"—that is, based on the assumption that model projections represent random samples from a distribution of plausible outcomes centered around the true climate. This is approach is often contrasted with the concept of "exchangeability," where the true climate is assumed to be drawn from the same distribution as the ensemble members, and therefore all members are exchangeable with the truth. These alternative interpretations are of direct relevance to the quantification of uncertainties in climate projections (Sanderson and Knutti 2012; Abramowitz et al. 2019). Under a truth-centered paradigm, estimated uncertainty decreases strongly as ensemble size increases because the uncertainty in the ensemble mean is estimated more precisely with more members (Lopez et al. 2006; Tebaldi and Sansó 2009; Annan and Hargreaves 2010; Knutti 2010). In contrast,

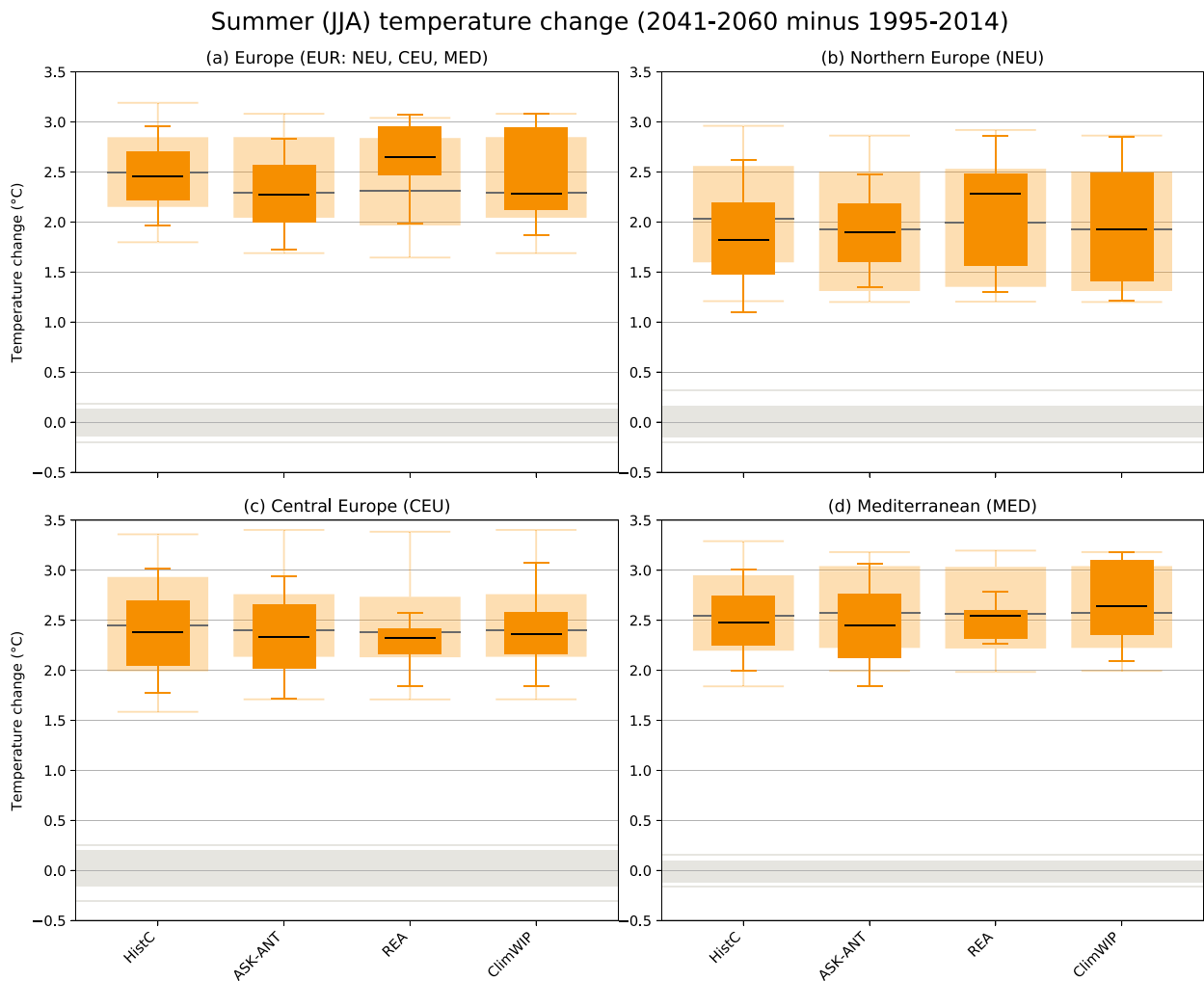## Summer (JJA) temperature change (2041-2060 minus 1995-2014)

FIG. 6. Similar to Fig. 2, but only for methods based on CMIP5. The distributions are based on a common subset of CMIP5 models to control for differences in the unconstrained distributions. Note that the *y* axis differs from that in Fig. 2.

in the exchangeability interpretation the uncertainty is characterized by the ensemble spread and is largely independent of the sample size (Annan and Hargreaves 2010). Annan and Hargreaves (2010) and Sanderson and Knutti (2012) suggest that the CMIP MMEs demonstrate some characteristics of both paradigms, and argue that these two seemingly contradictory viewpoints are actually complementary. This issue is, thus, far from resolved but we might expect differences in the estimates of uncertainty between methods that differ in these, and the two-stage weighting schemes in both REA and ClimWIP provide the opportunity to explore the respective impact of the dependence and convergence weighting independently of the performance weighting here.

Figure 7 shows such a decomposition of the ClimWIP and REA weighting into their performance and independence/convergence components. Generally, the

convergence part of the REA weighting leads to a reduction of spread without significantly shifting the distribution as expected. The REA performance weighting, in turn, can lead to both, spread reduction and a shift in the distribution. The combination of both components to the full REA weight is multiplicative (Giorgi and Mearns 2002) so that the total constraint applied can be considerably stronger than a mere linear combination. The independence weighting of ClimWIP does not have a strong effect on the unconstrained distributions and the total weight therefore mostly follows the performance weighting with small adjustments by the independence weighting. As shown in detail by Merrifield et al. (2019), the performance weighting is the dominating contribution for the CMIP5 MME (with only a few ensemble members per model). Adding three large ensembles with up to 100 members changes this behavior

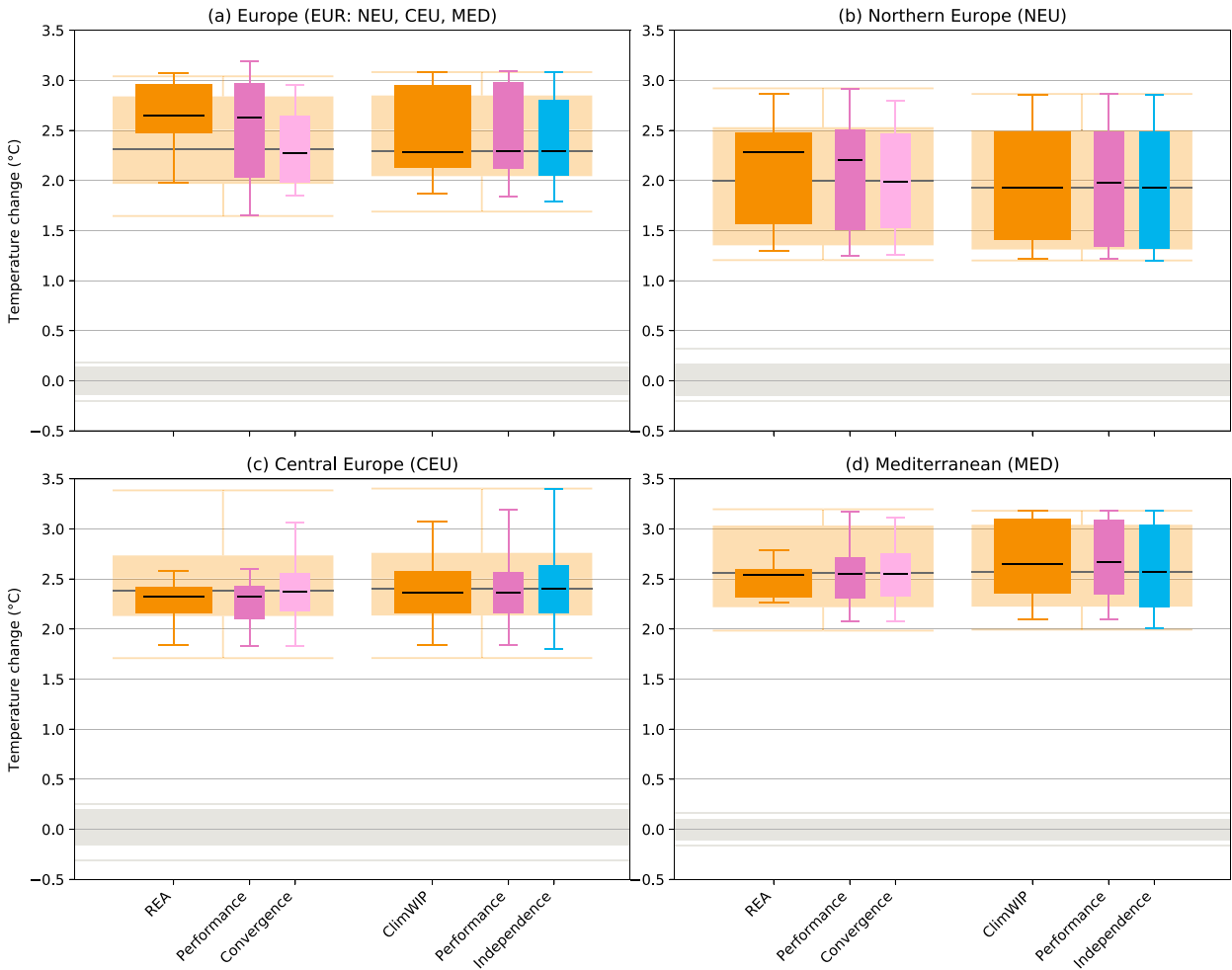Summer (JJA) temperature change (2041-2060 minus 1995-2014)

FIG. 7. Similar to Fig. 2, but only for REA and ClimWIP. The distributions are based on a common subset of CMIP5 models to control for differences in the unconstrained distributions, and the respective two components of the weighting are shown (model performance and dependence). Note that the *y* axis differs from that in Fig. 2.

and can lead to a considerably stronger contribution from the independence weighting.

These results clearly show that the truth centered and exchangeability assumptions underlying the independence and convergence components of ClimWIP and REA alone cannot explain the differences between the two methods. In section 5 below we will continue to explore the methods and their properties in more detail in order to identify further reasons for the differences in their constrained projections.

5. Discussion

We have shown that the different methods investigated generate some diversity in estimates of the median and considerable diversity in the range of projected changes depending on variable and region. The different

characteristics of the methods, such as their underlying assumptions, and the characteristics of the outputs might be used to support decisions on which method might be more appropriate in cases where results disagree. We summarize some of these key differences in Table 4, and in the following discussion we explore the implications of these differences in the context of the results above. Finally, we also look into possible ways forward by exploring avenues to providing clear recommendations for users.

a. Why do methods produce different projection ranges?

1) DIFFERENCES IN UNDERPINNING ASSUMPTIONS

Methods that assume that ensembles are ''truth centered'' have been demonstrated to result in narrower

TABLE 4. Key characteristics of the different methods.

	UKCP	ClimWIP	ASK	REA	HistC	CALL
Assumes truth centered				✓		
Constrained range can lie beyond unconstrained range			✓			✓
Multiple estimates of observations are used in weights/constraint	✓	✓				
Spatial scale at which constraint or performance weighting is calculated	Global + large scale	Same as target	Europe	Local	Global + local	Same as target
Multiple variables used to weight each target variables	✓	✓				
Observation uncertainty	✓	✓				
Includes estimate of internal variability	✓	✓		✓		✓
Carbon cycle	✓					
Model uncertainty (parameter)	✓					
Model uncertainty (structural)	✓	✓	✓	✓	✓	
Method error	✓	✓				
Outputs are spatially coherent	✓	✓	✓			
Outputs are physically coherent	✓	✓				

ranges of uncertainty than those that assume exchangeability (e.g., Annan and Hargreaves 2010). One method included explicitly assumes the underlying distribution to be truth centered (REA); however, this does not appear to lead to substantially narrower uncertainty ranges than for other methods here. Further, when the impact of the convergence weighting in REA is separated from the performance component, the convergence part (the truth-centered element) does not consistently have a large impact on the estimated uncertainty. This suggests that the impact of this assumption does not substantially affect the projection range in this case.

A further fundamental difference between methods is whether the constrained projection range can extend beyond the raw or unconstrained range. In ASK and CALL the systematic under- or overestimation of response to forcing is scaled, resulting in a range that, in some cases, extends beyond the envelope of raw projections indicating that a larger or smaller range than simulated would be consistent with the observed change to date. The other methods considered cannot result in a constrained distribution outside of the original spread and might thus under- or overestimate changes, while still drawing the projections closer to the "truth."

2) DIFFERENCES IN UNCERTAINTIES ACCOUNTED FOR

The size of the uncertainty range is, naturally, strongly affected by how comprehensively different sources of uncertainty are captured. The prime example for this is the UKCP method, which explicitly represents a larger number of sources of uncertainty than other methods, unsurprisingly leading to a substantially wider projection range. UKCP captures much of this additional uncertainty by drawing on perturbed physics ensemble projections which can be designed to sample uncertainties more strategically than CMIP multimodel "ensembles of opportunity." The UKCP method also explicitly accounts for carbon cycle uncertainty, which has been shown to be of comparable importance for the magnitude of the global temperature response as climate sensitivity (Booth et al. 2012). This feature of the UKCP method could be considered to address a shortcoming in the purely CMIP5-based methods.

Internal variability is treated with varying degrees of explicitness across the different methods. It is captured by the raw CMIP5 ensemble since the ensemble members capture different phases of variability. The sample size from each model can be improved by using multiple realizations from a model, thus increasing the number of internal variability realizations included in the ensemble. Some methods remove internal variability in the processing (ASK, HistC, UKCP) but to allow a consistent comparison the forced component is reinflated by adding an estimate of internal variability for the shown projection ranges.

Several methods use cross-validation frameworks, such as perfect model tests, to estimate the uncertainty of the methods themselves. In the cases of UKCP and ClimWIP, these estimates of method uncertainty are included in the constrained uncertainty range, while in others they are calculated as an external evaluation of the method and not included in the total uncertainty (CALL, HistC).

3) DIFFERENCES IN THE APPLICATION OF CONSTRAINTS

The methods draw on quite different characteristics of climate to measure performance and weight or constrain projections. Importantly, the methods use different

variables to constrain a given target variable—in REA, weighting is based on the local performance of the target variable (i.e., precipitation performance constrains precipitation) while other approaches use variables with known relationships to the target variable, or a basket of constraining variables (ClimWIP, UKCP). Further, different characteristics of those variables might be the basis of the constraint: while ASK, CALL, and HistC, in this application, mostly use characteristics of the time evolution of climate change, REA and ClimWIP here use spatial patterns. Combining several constraining metrics has been demonstrated to result in more conservative constraints, relative to the impacts of individual variables. Finally, different observational datasets, also based on certain assumptions and on different degrees of postprocessing (including reanalyses), are used by the methods.

There are also a number of more subtle differences that can be expected to affect the constraints calculated. These include whether the model constraints are based on more than one observational dataset, how methods treat multiple initial conditions members per model (are multiple realizations used—and if so, are members weighted individually, or the same weights applied to each realization of a model?), and whether constraints are calculated locally (i.e., over the same region as the target variable) or from a global or larger-scale region.

4) PHYSICAL AND SPATIAL CONSISTENCY CHARACTERISTICS OF THE OUTPUTS

Methods that calculate weights or constraints locally may not result in uncertainty estimates that are spatially coherent. These projections can be applied for each region independently, but estimates calculated separately for subregions may not sum to the same value as when calculated directly for a combined region, and calculations for neighboring regions may not have the appropriate relationship to each other. In practice, some methods use a combination so that the spatial coherency is partial.

Similarly, those methods where weights are calculated for more than one target variable using a common set of weights can be considered more physically consistent than those in which different target variables are weighted by different criteria. Again, in practice, this might be partial; for example, for ClimWIP the diagnostic variables are common for each target variable but the weights are not because of varying confidence estimates derived from a perfect model test based on the target variable. Physical consistency between variables offers the potential to provide joint probability estimates, which may be important where multivariate characteristics of projections are important.

b. How should the information be handled by users?

Our results raise a number of questions about how information from multiple methods can be communicated, combined, or applied, in particular for cases where constrained distributions disagree. A complex interplay between user needs, method properties, and output consistency needs to be considered in order to select the best possible information. Here we discuss several considerations, which provide a general perspective and might even serve as concrete guidelines for users, depending on their situation.

1) CONSIDERING THE DECISION CONTEXT

We have shown that in several cases the choice of method (across the multimodel methods, at least) has limited influence on the constrained median, but significant impact on the upper and lower percentiles. This provides the basis for some useful guidance based on the level of risk aversion the user has in a given context. Users with a relatively low level of risk aversion who wish to prepare for the most likely climate outcome could use results from any of these methods in such cases. However, those users with a higher level of risk aversion (i.e., those who are interested primarily in the lower and upper percentiles), may wish to consider carefully which method to draw on. Indeed, this intersection of communicating uncertainty in climate change projections and the needs of users will need increased future attention (Sutton 2019). Other method properties, such as spatial consistency or the inclusion of additional uncertainty, might provide additional arguments for or against certain methods (cf. section 5a).

2) USING AGREEING METHODS

To summarize our results we used a conservative "envelope approach" showing the full range spanned by the 10th, 50th, and 90th percentiles of the various methods (Figs. 2 and 3). In doing so we cannot claim any clear progress (e.g., a shift in the median, or a narrowing of uncertainty) over the conventional model democracy approach (which was used to derive unconstrained ranges) in most cases. However, for some cases, such as the temperature change from the CMIP5 methods in CEU and MED (Figs. 2 and 6), the majority of methods agree on the shift in median as well as the narrowing of the uncertainty range. This robustness in the results not only gives additional confidence in each of the individual estimates, but also indicates that, for such cases, each of the agreeing methods might be appropriate to use.

Naturally, this approach does not help for variables or regions where the methods disagree. However, our results still provide important information for these

cases as they highlight this disagreement based on the underlying method properties (given that we use a common setup to eliminate most other differences in processing). In this sense we here caution against using disagreeing methods without careful consideration and testing.

3) COMBINING METHOD OUTPUTS

Another way forward may be to average the method outputs into a combined probability distribution. This could be done by using a "method democracy" approach with equal weights for each method, or by weighting methods using some skill measure. For example, how well does the method predict an out-of-sample model projection? However, even if given a measure of skill, a number of other factors remain highly relevant for combining methods/distributions for a given application. For example, are the results physically/ spatially coherent? To what degree are the different sources of uncertainty captured? How comfortable are we with the validity of underpinning assumptions? How truly independent are the methods from each other (given that some, e.g., use the same models or observational datasets)? This means that such an approach also needs to be carefully considered and is definitely not applicable for all cases and methods. It is, therefore, not well suited as a general recommendation for users.

4) COMBINING METHODS BEFORE APPLYING THEM

The key information used to build the various constraints across methods (e.g., historical trend, regional model performance, or model independence) is not the same, and it is not entirely unexpected that these different lines of evidence can lead to different or even contradictory results in some cases. This highlights a way forward where future statistical methods could try to combine the various pieces of information together, rather than trying to combine the output. This seems to be a promising line of research: as several methods considered in this study report a clear added value (based on individual perfect model evaluation) while being based on different pieces of information as inputs, combining all these lines of evidence could lead to improved probabilistic projections.

It is obvious that it is not a priori clear that such a combination would be practically possible for all constraints and the question of the relative importance of different (and potentially contradictory) lines of evidence remains. In addition, such an approach effectively means the development of an entirely new method so that it can be seen as more of a long-term vision.

5) SELECTING METHODS BASED ON A CONSISTENT SKILL MEASURE

Probably the most promising way forward is the calculation of a skill measure to select between methods in cases where they disagree. One regularly used way of providing such an estimate of skill is a perfect model test. Such a test uses each of the models from the CMIP5 MME (or from an additional MME such as CMIP6) as "pseudo observations" in the historical period. The constrained distribution is then evaluated against the "truth" in the future, which is given by the same model from which the so-called pseudo observations were taken. The evaluation could be based on a selected skill metric such as the root-mean-square error or the continuous ranked probability score (Hersbach 2000).

Indeed, some form of perfect model test has already been applied to several methods included in this study in the past (Schurer et al. 2018; Brunner et al. 2019; O'Reilly et al. 2020, manuscript submitted to *Earth Sys. Dyn.*; Ribes et al. 2020, manuscript submitted to *Sci. Adv.*). However, such individual skill estimates are not necessarily comparable and should most probably not be used to decide between methods in the specific cases presented here. Combining the common settings introduced in section 3 with a testing framework standardising the "perfect models" used as well as the skill score and other settings therefore seems to be a promising approach for future work that can lead to clearer decision guidelines for users.

6. Summary, conclusions, and outlook

We have introduced a common framework to compare a diverse set of multimodel methods for quantifying uncertainties in projections of future European climate provided by groups within EUCP. The constrained median projections of temperature in 2041–60 are between 2° and 3°C warmer than the 1995–2014 average, which is well beyond the range of natural variability in present-day climate. While the median estimate is mostly robust across all methods, the spread is highly dependent on the method used. This partly reflects the fact that the constrained projections are based on different model pools and include different sources of uncertainty. Therefore, the choice of method has significant implications for users who are interested in the upper or lower ends of the distribution (i.e., those with a high level of risk aversion).

Constrained projections of median precipitation change show less consensus across methods, particularly in CEU and MED. All methods agree on a general drying in these two regions, but the median estimate of change varies from hardly any change to −25%. In NEU, the median

consistently stays within internal variability for all methods, with only a slight indication of a possible precipitation increase. The upper percentiles are constrained by most methods and in most regions so that future drying becomes increasingly likely relative to the unconstrained case. In general, the differences between the unconstrained and constrained projections can become considerably larger for precipitation and are less understood, so future work on their interpretation is required, along with careful consideration of method properties and applications, if constrained distributions are to be used.

In addition to four larger regions, we also tested the methods on four single grid cells to investigate the agreement at different spatial scales. The constrained median projections in these small regions mostly follow the behavior of the larger regions in which they are located and the higher internal variability reflected in the methods. Notably, most of the included methods have never been applied without spatial aggregation so that applying them to single grid cells was very much an experimental setup to check if physically meaningful results can be obtained. Clearly additional work, beyond these first encouraging results, will be required to further test and understand the methods' performance at such scales.

Three methods based on large ensembles of single models have also been included in this study to investigate estimates of internal variability and to provide context. Unsurprisingly, internal variability is found to consistently increase with decreasing region size also for these methods. Understanding the role of internal variability for weighting or constraining projections of future change is still very much an open topic, particularly for smaller regions and more heterogeneous variables such as precipitation. In addition, while this study has only looked at changes of the mean climate state, the role of internal variability becomes even more crucial when considering extremes.

A main part of this study is our discussion of the methods' properties, which can provide avenues to explain the differences in the constrained projections. These include fundamental underpinning assumptions (such as truth centered vs exchangeable), uncertainties considered (in models as well as observations), the treatment of model interdependencies (independence vs convergence), and the data used to constrain the projections (e.g., the variables considered or spatial vs time information). Ultimately all methods, including the unconstrained model democracy, are based on implicit and explicit assumptions that can be challenged. We, therefore, do not provide a single recommendation for selecting or combining methods in this study, but rather discuss several possible approaches.

One promising way to resolve this, which we discuss and that we currently pursue, is that of a coordinated perfect model test. Several of the methods investigated in this study have already been evaluated using such a test individually (Schurer et al. 2018; Brunner et al. 2019; O'Reilly et al. 2020, manuscript submitted to *Earth Sys. Dyn.*; Ribes et al. 2020, manuscript submitted to *Sci. Adv.*). Providing a consistent comparison across multiple methods could draw on the common settings developed in this study but needs to carefully consider additional questions such as the following: What models should be used as pseudo observations? How can overfitting be avoided (e.g., by providing anonymised models as pseudo observations)? What skill measure should we use? Do we focus on mean skill or on a "worst-case" scenario (or on a combination of both)? [See, e.g., Brunner et al. (2020, manuscript submitted to *Earth Syst. Dyn.*) for a discussion of this.] Are some methods systematically performing better for particular variables and/or predictions of extreme anomalies (e.g., very hot summers)? Future work in EUCP will, hence, 1) address whether such a verification framework could be used to provide a measure of method skill, 2) continue efforts to combine methods in more sophisticated ways to draw on their strengths, and 3) connect with decision makers to consider how the information in multimethod projections might be used and interpreted in relevant case studies.

REFERENCES

Aalbers, E. E., G. Lenderink, E. van Meijgaard, and B. J. van den Hurk, 2018: Local-scale changes in mean and heavy precipitation in Western Europe, climate change or internal variability? *Climate Dyn.*, **50**, 4745–4766, https://doi.org/10.1007/s00382-017-3901-9.

Abramowitz, G., and Coauthors, 2019: ESD reviews: Model dependence in multi-model climate ensembles: Weighting, sub-selection and out-of-sample testing. *Earth Syst. Dyn.*, **10**, 91–105, https://doi.org/10.5194/esd-10-91-2019.

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.*, **15**, 419–434, https://doi.org/10.1007/s003820050291.

——, and P. A. Stott, 2003: Estimating signal amplitudes in optimal fingerprinting, Part I: Theory. *Climate Dyn.*, **21**, 477–491, https://doi.org/10.1007/s00382-003-0313-9.

——, ——, 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.

Amos, M., and Coauthors, 2020: Projecting ozone hole recovery using an ensemble of chemistry-climate models weighted by model performance and independence. *Atmos. Chem. Phys.*, https://doi.org/10.5194/ACP-2020-86, in press.

Annan, J. D., and J. C. Hargreaves, 2010: Reliability of the CMIP3 ensemble. *Geophys. Res. Lett.*, **37**, L02703, https://doi.org/10.1029/2009GL041994.

Bishop, C. H., and G. Abramowitz, 2013: Climate model dependence and the replicate Earth paradigm. *Climate Dyn.*, **41**, 885–900, https://doi.org/10.1007/s00382-012-1610-y.

Booth, B. B., and Coauthors, 2012: High sensitivity of future global warming to land carbon cycle processes. *Environ. Res. Lett.*, **7**, 024002, https://doi.org/10.1088/1748-9326/7/2/024002.

——, G. R. Harris, J. M. Murphy, J. I. House, C. D. Jones, D. Sexton, and S. Sitch, 2017: Narrowing the range of future climate projections using historical observations of atmospheric CO_2. *J. Climate*, **30**, 3039–3053, https://doi.org/10.1175/JCLI-D-16-0178.1.

Braganza, K., D. J. Karoly, A. C. Hirst, M. E. Mann, P. Stott, R. J. Stouffer, and S. F. Tett, 2003: Simple indices of global climate variability and change: Part I—Variability and correlation structure. *Climate Dyn.*, **20**, 491–502, https://doi.org/10.1007/s00382-002-0286-0.

Brunner, L., R. Lorenz, M. Zumwald, and R. Knutti, 2019: Quantifying uncertainty in European climate projections using combined performance-independence weighting. *Environ. Res. Lett.*, **14**, 124010, https://doi.org/10.1088/1748-9326/ab492f.

Caldwell, P. M., M. D. Zelinka, and S. A. Klein, 2018: Evaluating emergent constraints on equilibrium climate sensitivity. *J. Climate*, **31**, 3921–3942, https://doi.org/10.1175/JCLI-D-17-0631.1.

Collins, M., and Coauthors, 2013: Long-term climate change: Projections, commitments and irreversibility. *Climate Change: 2013 Physical Science Basis*, T. Stocker et al., Eds., Cambridge University Press, 1029–1136, https://doi.org/10.1017/CBO9781107415324.024.

Cornes, R. C., G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones, 2018: An ensemble version of the E-OBS temperature and precipitation data sets. *J. Geophys. Res. Atmos.*, **123**, 9391–9409, https://doi.org/10.1029/2017JD028200.

Cressie, N., 1991: *Statistics for Spatial Data*. John Wiley and Sons, 928 pp.

DeAngelis, A. M., X. Qu, M. D. Zelinka, and A. Hall, 2015: An observational radiative constraint on hydrologic cycle intensification. *Nature*, **528**, 249–253, https://doi.org/10.1038/nature15770.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

Deser, C., R. Knutti, S. Solomon, and A. S. Phillips, 2012a: Communication of the role of natural variability in future North American climate. *Nat. Climate Change*, **2**, 775–779, https://doi.org/10.1038/nclimate1562.

——, A. Phillips, V. Bourdette, and H. Teng, 2012b: Uncertainty in climate change projections: The role of internal variability. *Climate Dyn.*, **38**, 527–546, https://doi.org/10.1007/s00382-010-0977-x.

——, ——, M. A. Alexander, and B. V. Smoliak, 2014: 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.

——, L. Terray, and A. S. Phillips, 2016: 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.

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

——, and Coauthors, 2019: Taking climate model evaluation to the next level. *Nat. Climate Change*, **9**, 102–110, https://doi.org/10.1038/s41558-018-0355-y.

Field, C. B., V. Barros, T. F. Stocker, and Q. Dahe, Eds., 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, 582 pp., https://doi.org/10.1017/CBO9781139177245.

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Gidden, M. J., and Coauthors, 2019: Global emissions pathways under different socioeconomic scenarios for use in CMIP6: A dataset of harmonized emissions trajectories through the end of the century. *Geosci. Model Dev.*, **12**, 1443–1475, https://doi.org/10.5194/gmd-12-1443-2019.

Giorgi, F., and L. O. Mearns, 2002: Calculation of average, uncertainty range, and reliability of regional climate changes from AOGCM simulations via the "reliability ensemble averaging" (REA) method. *J. Climate*, **15**, 1141–1158, https://doi.org/10.1175/1520-0442(2002)015<1141:COAURA>2.0.CO;2.

——, and ——, 2003: Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophys. Res. Lett.*, **30**, 1629, https://doi.org/10.1029/2003GL017130.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *J. Geophys. Res.*, **113**, D06104, https://doi.org/10.1029/2007JD008972.

Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophys. Res. Lett.*, **33**, L03502, https://doi.org/10.1029/2005GL025127.

——, P. Cox, C. Huntingford, and S. Klein, 2019: Progressing emergent constraints on future climate change. *Nat. Climate Change*, **9**, 269–278, https://doi.org/10.1038/s41558-019-0436-6.

Harris, G. R., D. M. Sexton, B. B. Booth, M. Collins, and J. M. Murphy, 2013: Probabilistic projections of transient climate change. *Climate Dyn.*, **40**, 2937–2972, https://doi.org/10.1007/s00382-012-1647-y.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations—The CRU TS3.10 dataset. *Int. J. Climatol.*, **34**, 623–642, https://doi.org/10.1002/joc.3711.

Hawkins, E., and R. Sutton, 2009: The potential to narrow uncertainty in regional climate predictions. *Bull. Amer. Meteor. Soc.*, **90**, 1095–1108, https://doi.org/10.1175/2009BAMS2607.1.

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

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hewitt, C. D., and J. A. Lowe, 2018: Toward a European climate prediction system. *Bull. Amer. Meteor. Soc.*, **99**, 1997–2001, https://doi.org/10.1175/BAMS-D-18-0022.1.

IPCC, 2013: *Climate Change 2013: The Physical Science Basis.* Cambridge University Press, 1535 pp., https://doi.org/10.1017/CBO9781107415324.

Jacob, D., and Coauthors, 2014: EURO-CORDEX: New high-resolution climate change projections for European impact research. *Reg. Environ. Change*, **14**, 563–578, https://doi.org/10.1007/s10113-013-0499-2.

Kato, S., N. G. Loeb, F. G. Rose, D. R. Doelling, D. A. Rutan, T. E. Caldwell, L. Yu, and R. A. Weller, 2013: Surface irradiances consistent with CERES-derived top-of-atmosphere shortwave and longwave irradiances. *J. Climate*, **26**, 2719–2740, https://doi.org/10.1175/JCLI-D-12-00436.1.

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.

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

Kirtman, B., and Coauthors, 2013: Near-term climate change: Projections and predictability. *Climate Change 2013: The Physical Science Basis*, T. Stocker et al., Eds., Cambridge University Press, 953–1028.

Knutson, T., 2017: Detection and attribution methodologies overview. *Climate Science Special Report: Fourth National Climate Assessment*, D. J. Wuebbles et al., Eds., U.S. Global Change Research Program, 114–132, https://doi.org/10.7930/J0319T2J.

Knutti, R., 2010: The end of model democracy? *Climatic Change*, **102**, 395–404, https://doi.org/10.1007/s10584-010-9800-2.

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

——, M. A. Rugenstein, and G. C. Hegerl, 2017a: Beyond equilibrium climate sensitivity. *Nat. Geosci.*, **10**, 727–736, https://doi.org/10.1038/ngeo3017.

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

Li, C., X. Zhang, F. Zwiers, Y. Fang, and A. M. Michalak, 2017: Recent very hot summers in Northern Hemispheric land areas measured by wet bulb globe temperature will be the norm within 20 years. *Earth's Future*, **5**, 1203–1216, https://doi.org/10.1002/2017EF000639.

Lopez, A., C. Tebaldi, M. New, D. Stainforth, M. Allen, and J. Kettleborough, 2006: Two approaches to quantifying uncertainty in global temperature changes. *J. Climate*, **19**, 4785–4796, https://doi.org/10.1175/JCLI3895.1.

——, E. B. Suckling, F. E. Otto, A. Lorenz, D. Rowlands, and M. R. Allen, 2015: Towards a typology for constrained climate model forecasts. *Climatic Change*, **132**, 15–29, https://doi.org/10.1007/s10584-014-1292-z.

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

Maher, N., and Coauthors, 2019: The Max Planck Institute Grand Ensemble: Enabling the exploration of climate system variability. *J. Adv. Model. Earth Syst.*, **11**, 2050–2069, https://doi.org/10.1029/2019MS001639.

Martel, J.-L., A. Mailhot, F. Brissette, and D. Caya, 2018: Role of natural climate variability in the detection of anthropogenic climate change signal for mean and extreme precipitation at local and regional scales. *J. Climate*, **31**, 4241–4263, https://doi.org/10.1175/JCLI-D-17-0282.1.

Merrifield, A. L., L. Brunner, R. Lorenz, and R. Knutti, 2019: A weighting scheme to incorporate large ensembles in multi-model ensemble projections. *Earth Syst. Dyn.*, https://doi.org/10.5194/esd-2019-69, in press.

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

Murphy, J. M., and Coauthors, 2018: UKCP18 Land Projections: Science Report, Met Office Rep., 191 pp., https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/UKCP18-Land-report.pdf.

Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Climate*, **24**, 2358–2367, https://doi.org/10.1175/2010JCLI3814.1.

Polson, D., G. C. Hegerl, X. Zhang, and T. J. Osborn, 2013: Causes of robust seasonal land precipitation changes. *J. Climate*, **26**, 6679–6697, https://doi.org/10.1175/JCLI-D-12-00474.1.

Renoult, M., and Coauthors, 2020: A Bayesian framework for emergent constraints: Case studies of climate sensitivity with PMIP. *Climate Past*, https://doi.org/10.5194/CP-2019-162, in press.

Ribes, A., F. W. Zwiers, J. M. Azaïs, and P. Naveau, 2017: A new statistical approach to climate change detection and attribution. *Climate Dyn.*, **48**, 367–386, https://doi.org/10.1007/s00382-016-3079-6.

Rougier, J., M. Goldstein, and L. House, 2013: Second-order exchangeability analysis for multimodel ensembles. *J. Amer. Stat. Assoc.*, **108**, 852–863, https://doi.org/10.1080/01621459.2013.802963.

Sanderson, B. M., and R. Knutti, 2012: On the interpretation of constrained climate model ensembles. *Geophys. Res. Lett.*, **39**, L16708, https://doi.org/10.1029/2012GL052665.

——, C. Piani, W. J. Ingram, D. A. Stone, and M. R. Allen, 2008: Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dyn.*, **30**, 175–190, https://doi.org/10.1007/s00382-007-0280-7.

——, R. Knutti, and P. Caldwell, 2015a: A representative democracy to reduce interdependency in a multimodel ensemble. *J. Climate*, **28**, 5171–5194, https://doi.org/10.1175/JCLI-D-14-00362.1.

——, ——, and ——, 2015b: Addressing interdependency in a multimodel ensemble by interpolation of model properties. *J. Climate*, **28**, 5150–5170, https://doi.org/10.1175/JCLI-D-14-00361.1.

——, M. Wehner, and R. Knutti, 2017: Skill and independence weighting for multi-model assessments. *Geosci. Model Dev.*, **10**, 2379–2395, https://doi.org/10.5194/gmd-10-2379-2017.

Schneider, U., P. Finger, A. Meyer-Christoffer, E. Rustemeier, M. Ziese, and A. Becker, 2017: Evaluating the hydrological cycle over land using the newly-corrected precipitation climatology

from the Global Precipitation Climatology Centre (GPCC). *Atmosphere*, **8**, 52, https://doi.org/10.3390/ATMOS8030052.

Schurer, A., G. Hegerl, A. Ribes, D. Polson, C. Morice, and S. Tett, 2018: Estimating the transient climate response from observed warming. *J. Climate*, **31**, 8645–8663, https://doi.org/10.1175/JCLI-D-17-0717.1.

——, A. P. Ballinger, A. R. Friedman, and G. Hegerl, 2020: Human influence strengthens the contrast between tropical wet and dry regions. *Environ. Res. Lett.*, **150**, 971–976, https://doi.org/10.1088/1748-9326/AB83AB.

Selten, F. M., R. Bintanja, R. Vautard, and B. J. van den Hurk, 2020: Future continental summer warming constrained by the present-day seasonal cycle of surface hydrology. *Sci. Rep.*, **10**, 4721, https://doi.org/10.1038/s41598-020-61721-9.

Sexton, D. M. H., and G. R. Harris, 2015: The importance of including variability in climate change projections used for adaptation. *Nat. Climate Change*, **5**, 931–936, https://doi.org/10.1038/nclimate2705.

——, J. M. Murphy, M. Collins, and M. J. Webb, 2012: Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Climate Dyn.*, **38**, 2513–2542, https://doi.org/10.1007/s00382-011-1208-9.

Shiogama, H., D. Stone, S. Emori, K. Takahashi, S. Mori, A. Maeda, Y. Ishizaki, and M. R. Allen, 2016: Predicting future uncertainty constraints on global warming projections. *Sci. Rep.*, **6**, 18903, https://doi.org/10.1038/srep18903.

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.

——, ——, and M. R. Allen, 2006: Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.*, **33**, L02708, https://doi.org/10.1029/2005GL024423.

Sutton, R. T., 2019: Climate science needs to take risk assessment much more seriously. *Bull. Amer. Meteor. Soc.*, **100**, 1637–1642, https://doi.org/10.1175/BAMS-D-18-0280.1.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An overview of CMIP5 and the experiment design. *Bull. Amer. Meteor. Soc.*, **93**, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc.*, **A365**, 2053–2075, https://doi.org/10.1098/rsta.2007.2076.

——, and B. Sansó, 2009: Joint projections of temperature and precipitation change from multiple climate models: A hierarchical Bayesian approach. *J. Roy. Stat. Soc.*, **A172**, 83–106, https://doi.org/10.1111/j.1467-985X.2008.00545.x.

Tegegne, G., Y.-O. Kim, and J.-K. Lee, 2019: Spatiotemporal reliability ensemble averaging of multi-model simulations. *Geophys. Res. Lett.*, **46**, 12 321–12 330, https://doi.org/10.1029/2019GL083053.

Tokarska, K. B., G. C. Hegerl, A. P. Schurer, A. Ribes, and J. T. Fasullo, 2019: Quantifying human contributions to past and future ocean warming and thermosteric sea level rise. *Environ. Res. Lett.*, **14**, 074020, https://doi.org/10.1088/1748-9326/ab23c1.

——, M. B. Stolpe, S. Sippel, E. M. Fischer, C. J. Smith, F. Lehner, and R. Knutti, 2020: Past warming trend constrains future warming in CMIP6 models. *Sci. Adv.*, **6**, eaaz9549, https://doi.org/10.1126/sciadv.aaz9549.

UNFCCC, 2015: Adoption of the Paris Agreement: Conference of the Parties, 21st Session (COP21). United Nations Framework Convention on Climate Change Rep., 32 pp., https://unfccc.int/resource/docs/2015/cop21/eng/l09r01.pdf.

van Vuuren, D. P., and Coauthors, 2011: The representative concentration pathways: An overview. *Climatic Change*, **109**, 5–31, https://doi.org/10.1007/s10584-011-0148-z.

Vogel, M. M., J. Zscheischler, and S. I. Seneviratne, 2018: Varying soil moisture–atmosphere feedbacks explain divergent temperature extremes and precipitation projections in central Europe. *Earth Syst. Dyn.*, **9**, 1107–1125, https://doi.org/10.5194/esd-9-1107-2018.