Climate Dynamics (2023) 61:5171–5197
https://doi.org/10.1007/s00382-023-06846-z

Prioritizing the selection of CMIP6 model ensemble members for downscaling projections of CONUS temperature and precipitation

Julia M. Longmate[1] · Mark D. Risser[2] · Daniel R. Feldman[2]

Received: 7 March 2022 / Accepted: 24 May 2023 / Published online: 17 July 2023

Abstract
Given the mismatch between the large volume of data archived for the sixth phase of the Coupled Model Intercomparison Project (CMIP6) and limited personnel and computational resources for downscaling, only a small fraction of the CMIP6 archive can be downscaled. In this work, we develop an approach to robustly sample projected hydroclimate states in CMIP6 for downscaling to test whether the selection of a single initial condition (IC) ensemble member from each CMIP6 model is sufficient to span the range of modeled hydroclimate over the conterminous United States (CONUS) and CONUS subregions. We calculate the pattern-centered root mean square difference of IC ensemble member anomalies relative to each model's historical climatology for shared socioeconomic pathway (SSP) projections over 30-year time periods and compare the ratio of inter-model to intra-model variability for this metric. Regardless of SSP, inter-model variability is generally much greater than intra-model variability at the scales of the CONUS as a whole, as well as for most CONUS sub-regions. However for some variables and scenarios, inter- and intra-model variability are similar at sub-CONUS scales, indicating that selecting a single IC ensemble member per model may be sufficient to sample the range of projected hydroclimate states in the 21st Century across CONUS, but for specific regions and variables, more careful selection of ensemble members may be necessary. Regionally-resolved Taylor diagrams identify where more IC ensemble member downscaling efforts should be focused if resources are available to do so. Our results suggest that, with parsimonious sampling, the requisite computational expense of downscaling temperature and precipitation fields over the CONUS for subsequent CMIP activities may increase only marginally despite the great increase in data volumes with each successive CMIP phase.

Keywords Initial condition ensemble · Shared socioeconomic pathways · Taylor Diagrams · Inter-model variability

1 Introduction

The simulations that comprise the Coupled Model Intercomparison Project version 6 (CMIP6) multi-model ensemble (Eyring et al. 2016) serve, among other purposes, to establish a plausible set of historical and future projections of the Earth system for a wide range of emissions scenarios (O'Neill et al. 2016). The utility of these projections for local planning in the 21st Century faces challenges, however, because each Earth System Model (ESM) in CMIP6 first focuses on ensuring model skill at planetary to continental spatial scales and interannual to centennial time scales through loose constraints, such as top-of-atmosphere energy balance and a large-scale circulation of the atmosphere and ocean, supported by theory (Mauritsen et al. 2012; Schmidt et al. 2017). The coarse resolution of each ESM (~ 100 km) either under-resolves or simply does not resolve the processes that contribute to local impacts. ESM skill is achieved through physical and parameterized process modeling at large spatial scales and long time scales instead of at small time and short spatial scales (Clark et al. 2015), even though only the latter scales are relevant to infrastructure and operations planning.

As a result, while ESMs physically model climatic conditions that are without an historical analog, they are blunt tools for developing projections at the spatial scales of interest for local infrastructure and operations planning. The mismatch between ESMs and local needs is highlighted by the existence of model biases relative to the historical observational record at regional and local levels, which may be large

✉ Julia M. Longmate
jmlongmate@berkeley.edu

[1] University of California, Berkeley, CA, USA

[2] Lawrence Berkeley National Laboratory, Berkeley, CA, USA

enough to preclude a model's adoption, even indirectly, by a user for planning purposes (Wang et al. 2014; Kim et al. 2020; Srivastava et al. 2020; Pierce et al. 2021a). For a wide range of applications that require localized information, the mismatch between the CMIP6 models and both process representation and the spatial resolution needs of the user necessitates downscaling solutions. There are a wide range of downscaling techniques ranging from statistical methods (Wood et al. 2004; Abatzoglou and Brown 2012; Stoner et al. 2013; Pierce et al. 2014) to hybrid methods (Gutmann et al. 2014) to fully dynamical methods that contain explicit and parameterized representations of atmospheric and surface physical processes (e.g., Giorgi and Gutowski 2015). These methods construct local projections from coarse GCM outputs at scales relevant to local-level infrastructure and operations planning (e.g., Wood et al. 2004; Stoner et al. 2013; Pierce et al. 2014; Giorgi and Gutowski 2015; Gutmann et al. 2016).

While many variables are potentially of interest to local infrastructure and operations planning, we focus here on the daily surface air temperature and precipitation in the Conterminous United States (CONUS). These variables are central to local planning and management (e.g., Moss et al. 2017) and there are established workflows for analyzing these variables for climate assessments performed by a wide range of United States federal agencies (e.g., Melillo and Yohe 2014; USGCRP 2021). However, even if the focus is limited to precipitation and temperature, the CMIP6 archive represents an extremely large volume of data, including dozens of separate contributions from different modeling centers as well as different ensemble members of each model produced through changing initial conditions (IC; e.g., Murphy et al. 2004; Deser et al. 2012; Kay et al. 2015; Deser et al. 2020) or perturbed physical parameterizations (e.g., Murphy et al. 2004; Rostron et al. 2020). IC ensembles are particularly useful for assessing a given model's internal variability, especially in the context of adaptation decision-making (Mankin et al. 2020). The development of localized solutions from that archive must contend with the volume of data available in the CMIP6 archive while capturing the range of not just inter-model but intra-model variability. Because of this, the question becomes: how can one adopt a judicious and parsimonious approach to downscaling CMIP6 which is suitable across the broad range of hydroclimates represented in the CONUS?

As a practical matter, there are significant personnel and computational costs to downscaling; the expenses incurred for dynamical downscaling solutions are well-known (Giorgi and Gutowski 2015) but are also non-negligible even for statistical downscaling. Despite efforts to homogenize model input for subsequent analysis, there remain idiosyncrasies in each contributing model (Pierce et al. 2021b). A significant amount of preparation is required for each model to account for, in addition to these, differences in grid scales, vertical coordinate convention, completeness of variables being reported, and calendaring systems. For example, the UKESM1 model has a 360-day year (Sellar et al. 2019), unlike other models, which must be accounted for in statistically downscaling an ensemble. Additionally, mismatches in the number of ensemble members per model could over- or under-weight a given model. The idiosyncrasies of CMIP models have not diminished, and are unlikely to diminish, over time. Finally, the data storage and data provision challenges for downscaling outputs are perennial. Solutions to these practical matters associated with ever-growing ensembles of climate models that need downscaling correspondingly need to scale with the growing data volumes, with sustainable levels of computational and personnel support.

As of this writing, 57 CMIP6 models have reported results to the Earth System Grid Federation (Cinquini et al. 2014) that fulfill these requirements. Together, these constitute $\sim$ 1770 total simulations for which downscaled solutions could be developed. However, it is highly impractical to develop downscaling solutions for the complete set of CMIP6 simulations, and there is limited guidance in the scientific literature for navigating a multi-model ensemble of ESMs where some models contribute multiple ensemble members (McSweeney et al. 2012). Additionally, the implications of intra-model uncertainty on the corresponding uncertainty of downscaled hydroclimate projections have been under-investigated.

One major previous effort to downscale the CMIP5 archive in North America, the Localized Constructed Analogs (LOCA), used a convention whereby a single IC ensemble member was chosen to produce one downscaled climate model per emissions scenario (Brekke et al. 2013). This assumed that IC ensemble members of a given model were all similar enough to each other that this sampling approach would not underestimate changes in the distributions of temperature and precipitation that were produced by model internal variability, as characterized by the range of IC ensemble members. The sufficiency of selecting a single ensemble member for developing downscaled solutions that do not underestimate the impact of model internal variability on hydroclimate projections has not been established and motivates the need to quantify the range spanned by a given model's ensemble members and all models. Here, we seek to understand how best to sample such a large and heterogeneous archive, and whether that sampling should expend resources to include different IC ensemble members or should favor a greater range of models.

The main blunt summary statistics we could be interested in using to compare the similarities between models are (1) mean biases, (2) mean state changes (or anomalies), and (3) the spatial pattern in responses. Mean biases are removed from GCMs prior to downscaling, so from a downscaling perspective these do not represent meaningful differences between models. Differences in mean state changes and the spatial pattern in responses between models and IC ensemble members however both characterize how different model results may yield distinct simulations of future climate; either by predicting very different rates of change (e.g. a warmer simulation versus a cooler one), or by simulating different spatial patterns in where changes occur. Pattern scaling has been used as a method to extrapolate climate simulations to time periods not covered by the simulation, and relies on the assumption that the pattern of state change is relatively stable over time (Fowler et al. 2007). Rather than looking at how spatial patterns of anomalies change (which we expect to be minimal based on the pattern scaling assumption), in this paper we look at mean state changes, or anomalies, relative to the historical baseline, and compare the anomalies of different ensemble members across and within models by constructing a multi-model ensemble mean of anomalies. The multi-model ensemble mean is used as a point of reference that indicates a “central” estimate of model behavior across the CMIP6 archive, for a given climate variable, scenario, and 30-year time period. By comparing the relative similarity of spatial patterns in anomalies with the multi-model ensemble mean, we can summarize several dimensions of model difference into a single metric, the variability ratio (later denoted R), to characterize differences between and within models.

In this paper, we compare the variability between models (inter-model variability) to the variability within each model (intra-model variability), and specifically ask: does a random sampling of ensemble members provide an unbiased sample of the multi-model ensemble, even if this yields only a single ensemble member from a given model? In answering this question, our objective is not to rank model and ensemble performance but rather to determine an optimal approach to sampling the archive for a wide range of subsequent analyses of hydroclimate fields while remaining agnostic to measures of skill. Specifically, we do not consider mean biases or metrics of model performance but instead, focus on the higher-order differences between models and between ensemble members. We discuss the magnitude of the ratio of inter- to intra-model variability across future projections, variables, and CONUS sub-regions. This work builds on the results of Mankin et al. (2020) by exploring the navigation of the CMIP6 multi-model ensemble with multiple downscaled solutions. The distinction in this paper is that for the purpose of downscaling the CMIP6 multi-model ensemble, a parsimonious approach must be taken to sample the archive, so we explore what parsimony entails.

This paper is organized as follows: first, in Sect. 2 we present an overview of the CMIP6 archive and its range of mean state changes over CONUS, our use of Taylor diagrams and their suitability for concisely summarizing ensembles of models each with different IC ensemble members, our methodology for quantifying the relative differences between inter- and intra-model variability among the multi-model ensemble, and an exploration of the robustness of this approach to the presence of outlier ensemble members. We present the results of our analysis in Sect. 3 and conclude with a set of recommendations for prioritizing downscaling routines in Sect. 4.

2 Data and methods

2.1 Overview of the CMIP6 archive

As of February 28, 2022, the CMIP6 archive on the Earth System Grid Federation contained contributions from 44 modeling centers and 113 models. While the CMIP6 archive continues to grow as additional outputs and ensemble members are added, as of this date 61 models have historical model outputs and 49 have future projections for the variables of precipitation rate (pr), minimum daily temperature (tasmin), and maximum daily temperature (tasmax); see Tables 1 and 2. Throughout this work we focus on IC ensemble members, which are distinct realizations of each model with identical physics and forcings and nearly identical initial conditions which were produced to capture climate change and assess Earth System internal variability in ESMs (e.g., Murphy et al. 2004; Deser et al. 2012; Kay et al. 2015; Deser et al. 2020). Of these, only IC ensemble member historical simulations and future projections added to the archive prior to October 1, 2021 which had complete (or near-complete, with a ± 12 month margin of error) monthly time series spanning a given 30-year time period are in our analysis.

We chose to look at three specific shared socioeconomic pathways (SSPs), namely SSP245, SSP370, and SSP585, since these span a wide range of end-of-century radiative forcings across all of the scenarios and are broadly relevant to investigations of societal impacts due to changing temperature and precipitation patterns (Wu et al. 2022). We acknowledge that these three SSPs are a subset of all possible scenarios which we could have analyzed. We chose to compare scenarios with similarly large numbers of ensemble members. These three SSPs are among the scenarios for which the archive contains the most model-ensembles (O’Neill et al. 2021) (at the time we began analysis in October 2021, SSP245 and SSP370 had the highest number of ensemble members provided by modeling centers, followed by SSP126 and SSP585), and the divergent boundary conditions of the scenarios lead to substantial differences in end-of-century hydroclimate (Hawkins and Sutton 2009; Lehner et al. 2020). We also wanted to ensure that this analysis is relevant to those scenarios that are being downscaled for

Table 1 Number of CMIP6 historical initial condition (IC) ensemble members that fulfill our restriction criteria in the CMIP6 archive for daily precipitation rate (pr), maximum daily temperature (tasmax), and minimum daily temperature (tasmin)

Model	# of IC ensemble members			Model	# of IC ensemble members		
	pr	tasmax	tasmin		pr	tasmax	tasmin
ACCESS-CM2	3	3	2	ACCESS-ESM1-5	40	40	28
AWI-CM-1-1-MR	5	4	4	AWI-ESM-1-1-LR	0	1	1
BCC-CSM2-MR	3	3	3	BCC-ESM1	3	3	3
CAMS-CSM1-0	2	0	0	CanESM5	25	25	24
CAS-ESM2-0	4	4	3	CESM2	11	0	0
CESM2-FV2	3	0	0	CESM2-WACCM	3	0	0
CESM2-WACCM-FV2	3	0	0	CIESM	3	3	3
CMCC-CM2-HR4	1	0	0	CMCC-CM2-SR5	1	0	0
CMCC-ESM2	1	1	0	E3SM-1-0	5	0	0
E3SM-1-1	1	0	0	E3SM-1-1-ECA	1	0	0
EC-Earth3	70	73	66	EC-Earth3-AerChem	2	2	2
EC-Earth3-CC	1	1	0	EC-Earth3-Veg	9	9	8
EC-Earth3-Veg-LR	3	3	3	FGOALS-f3-L	3	0	0
FGOALS-g3	6	6	3	FIO-ESM-2-0	3	3	3
GFDL-ESM4	3	3	1	GISS-E2-1-G	12	12	12
GISS-E2-1-G-CC	1	1	1	GISS-E2-1-H	10	10	10
GISS-E2-2-H	5	5	0	ICON-ESM-LR	5	0	0
INM-CM4-8	1	1	1	INM-CM5-0	2	2	1
IPSL-CM5A2-INCA	1	0	0	IPSL-CM6A-LR	32	32	32
IPSL-CM6A-LR-INCA	1	1	0	KACE-1-0-G	3	0	0
KIOST-ESM	1	0	0	MCM-UA-1-0	1	0	0
MIROC6	49	50	50	MPI-ESM-1-2-HAM	3	3	3
MPI-ESM1-2-HR	10	10	9	MPI-ESM1-2-LR	10	10	10
MRI-ESM2-0	10	10	9	NESM3	5	5	5
NorCPM1	15	0	0	NorESM2-LM	3	0	0
NorESM2-MM	2	0	0	SAM0-UNICON	1	1	1
TaiESM1	2	0	0				
				Total realizations:	403	340	301

widespread analysis: for the upcoming Fifth National Climate Assessment, only SSP245, SSP370, and SSP585 are being downscaled.

On average, a total of ~ 200 IC ensemble members from 38 models met our analysis requirements per scenario. The set of models and ensemble members with historical simulations for the 30-year period of 1980–2010 (the period of time for which we can maximize the number of IC ensemble members included) is larger, and contains ~ 350 total ensemble members across 53 models, averaging to 8 ensemble members per model, with a maximum of 72 ensemble members produced by a single model.

2.2 Mean state change

To display the range of projections of minimum temperature, maximum temperature and precipitation that the CMIP6 models and their ensemble members produce, we calculated the mean state change (or anomaly) for all ensemble members from all models, across variables, scenarios, regions, and 30-year time periods. In Fig. 1 we show box plots of mean state change for all ensemble members for a single region, the Southwest (boxplots for the remaining NCA4 regions can be found in Appendix 4). For all regions, scenarios, and variables, the mean state change of different ensemble members spans a broad range of values, and increases over time. In Fig. 1 the range of mean state change of ensemble members is shown without specifying from which model they originate. Despite the multi-model ensemble average and range of mean state change across all models and ensembles steadily increasing over time, the CONUS-wide and regional variability ratios in Fig. 7 show that between-model variability generally increases at a greater rate relative to within-model variability, with some regional exceptions. As context for Fig. 7, this suggests that differences in between- and within-model variability are driven by similarity between ensemble members from the same models, rather than by convergence of all ensemble

Table 2 Number of CMIP6 shared socioeconomic pathway (SSP) projection ensemble members for SSP245, SSP370, and SSP585 that fulfill our restriction criteria in the CMIP6 archive for daily precipitation rate (pr), maximum daily temperature (tasmax), and minimum daily temperature (tasmin)

Model	Precip.			Max. Temp.			Min. Temp.		
	SSP245	SSP370	SSP585	SSP245	SSP370	SSP585	SSP245	SSP370	SSP585
ACCESS-CM2	3	3	2	3	3	3	2	2	3
ACCESS-ESM1-5	19	30	0	30	30	10	24	27	6
AWI-CM-1-1-MR	1	5	1	1	5	1	1	5	1
BCC-CSM2-MR	1	1	1	1	1	1	1	1	1
BCC-ESM1	0	3	0	0	2	0	0	2	0
CAMS-CSM1-0	2	2	2	0	0	0	0	0	0
CanESM5	25	25	25	25	25	25	25	25	24
CAS-ESM2-0	0	2	0	2	2	2	2	2	2
CESM2	3	3	3	3	3	3	3	3	1
CESM2-WACCM	5	3	5	4	0	4	4	0	4
CIESM	1	0	0	1	0	1	1	0	1
CMCC-CM2-SR5	1	1	1	0	0	0	0	0	0
CMCC-ESM2	0	1	0	1	1	1	0	1	0
E3SM-1-1	10	0	0	0	0	0	0	0	0
EC-Earth3	50	57	53	72	57	58	66	5	45
EC-Earth3-AerChem	0	2	0	0	2	0	0	0	0
EC-Earth3-CC	0	0	0	1	0	1	0	0	0
EC-Earth3-Veg	6	6	6	8	6	8	7	6	5
EC-Earth3-Veg-LR	3	3	3	3	3	3	3	3	3
FGOALS-f3-L	1	1	1	0	0	0	0	0	0
FGOALS-g3	4	5	4	4	5	4	4	5	4
FIO-ESM-2-0	3	0	3	3	0	3	3	0	3
GFDL-ESM4	3	1	1	3	1	1	2	1	1
IITM-ESM	1	1	1	0	0	0	0	0	0
INM-CM4-8	1	1	1	1	1	1	1	0	1
INM-CM5-0	1	5	1	1	5	1	1	5	1
IPSL-CM5A2-INCA	0	1	0	0	0	0	0	0	0
IPSL-CM6A-LR	11	11	6	11	11	6	7	10	0
KACE-1-0-G	3	3	3	0	0	0	0	0	0
MIROC6	37	3	50	50	3	50	43	3	40
MPI-ESM-1-2-HAM	0	3	0	0	3	0	0	1	0
MPI-ESM1-2-HR	2	10	2	2	10	2	2	10	1
MPI-ESM1-2-LR	10	9	10	10	10	10	10	10	10
MRI-ESM2-0	1	5	0	5	5	4	5	5	3
NESM3	2	0	2	2	0	2	2	0	2
NorESM2-LM	0	3	0	0	0	0	0	0	0
NorESM2-MM	0	1	0	0	0	0	0	0	0
TaiESM1	1	1	1	0	0	0	0	0	0
Total	211	211	188	247	194	205	219	132	162

members towards greater agreement. The wide spread of values across ensemble members additionally showcases a small fraction of the differences across models and ensemble members that downscalers need to consider, and highlights the need to undertake prioritization for the selection of models and ensemble members.

2.3 Methods

2.3.1 Regridding model output

In order to provide tractable comparisons between model results, a common grid is required, so we conservatively

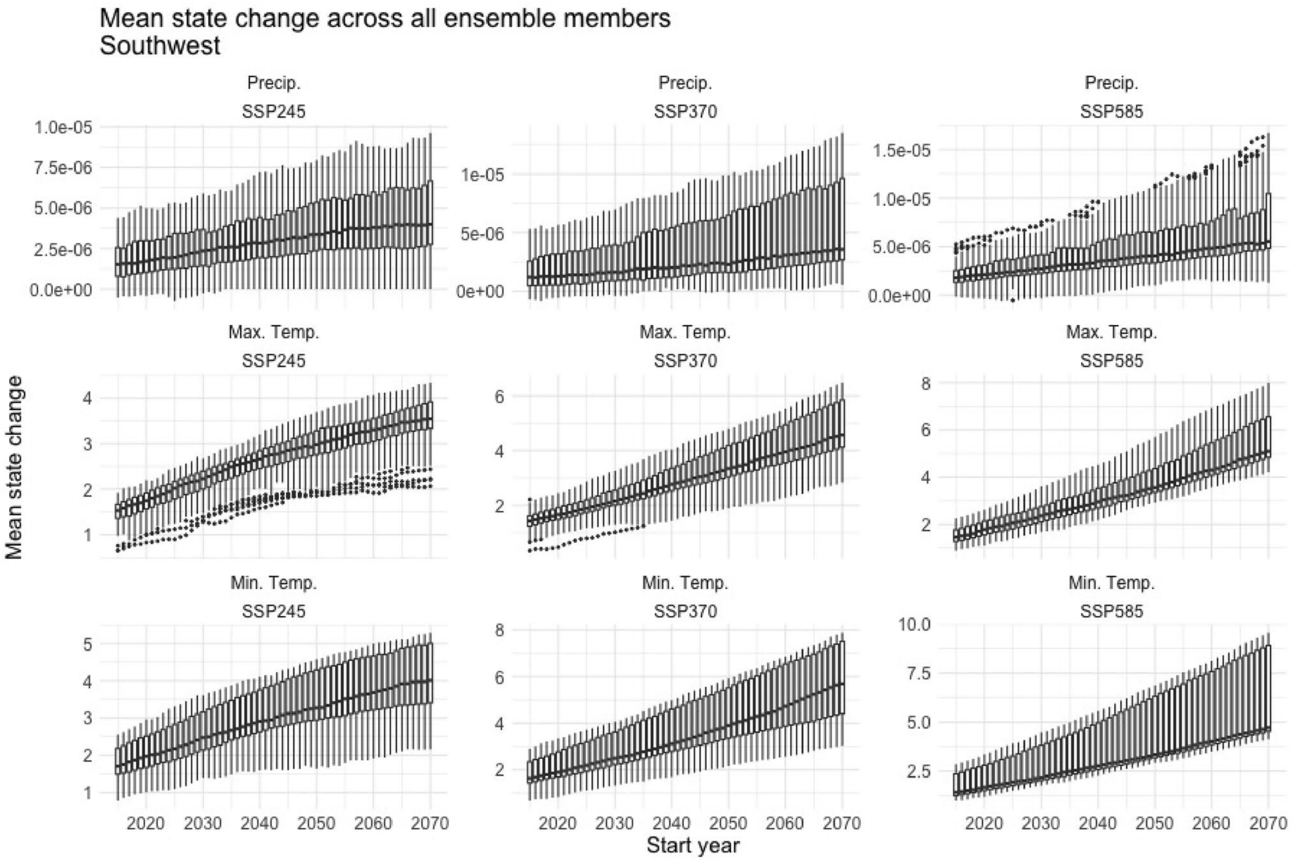

Fig. 1 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods are shown for each variable and scenario, for the NCA4 region of the Southwest. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

remap (Jones 1999) all future projections to the coarsest model resolution in the archive, with roughly 250 km grid cells.

2.3.2 Summarizing model climatologies

To calculate anomalies, we subtract the model-specific historical mean across the time period of 1980–2015 from the scenario- and ensemble member-specific annual mean of the corresponding climate variable. We calculate 30-year averages of these anomalies in moving windows across 2015–2100 for SSP245, SSP370, and SSP585 (e.g. the annual average value of maximum daily temperature for CESM2's IC ensemble member r1i1p1f1 in SSP370, averaged across 2070–2100 into a two-dimensional field of grid cells). The 30-year time period is standard for climate normal calculations (World Meteorological Organization 1989, 2007) because it is a sufficiently long period of time over which to average out annual to multi-decadal fluctuations (Arguez and Vose 2011).

Additionally, we construct a multi-model ensemble mean of these anomalies, weighted inversely by the number of ensemble members per each model to give models equal weight regardless of the number of IC ensemble members provided per model. Anomalies of individual ensemble members are compared with the multi-model ensemble average.

2.3.3 Regional focus

We look at three separate partitions of the CONUS: (1) all of CONUS, (2) the seven regions defined in the Fourth National Climate Assessment (NCA4; Reidmiller et al. 2017), and (3) three custom regions, “West”, “Central”, and “East.” At the coarsest resolution used, the smaller NCA4 regions contain very few grid cells (~10 in the Northeast). The three custom regions are thus useful for examining sub-CONUS regional differences while ensuring that each region has a sufficient number of grid cells to minimize small sample problems when calculating regional statistics. The three regions are defined as follows: the West region is everything west of 104° W, the East region is everything east of 95° W, and the Central region lies between 104° W and 95° W. The division of regions at these parallels preserves some rough boundaries in the regional climatology of CONUS (Regonda et al. 2016) while ensuring regions are both larger and more similar in size. Regional Taylor diagrams (see Sect. 2.3.4) make use of the NCA4 regions to bring regional climatology information to bear on their interpretation, while regional variability ratios (see Sect. 2.3.5) are calculated across these three larger subregions to avoid small geospatial sample sizes.

2.3.4 Taylor diagrams

Taylor diagrams (Taylor 2001) are useful visualizations of how multiple model outputs compare relative to a reference dataset, and we use this comparison to examine the relative differences between models and ensemble members within the CMIP6 archive. In this analysis, since we do not focus on traditional metrics of model performance such as mean bias, Taylor diagrams are well-suited to succinctly capture the higher-order statistics of the multiple models and their ensemble members across a range of CMIP6 experiments. Our metric of interest is the difference between the average of a field for an ensemble member across a 30-year period in a future projection and the average across the corresponding historical simulation from 1980-2015 for that ensemble member, which we refer to as "state change" or "anomaly". The reference dataset that we use for Taylor diagrams of future projections is the multi-model ensemble average of these anomalies, as discussed in Sect. 2.3.2. This approach emphasizes how the anomalies of each model's output from scenario experiments of the 21st Century differ between models and ensemble members across the CMIP6 archive.

The mean-centered statistics of a Taylor diagram (Taylor 2001) are designed to concisely summarize the degree of correspondence between two fields or "patterns," here comparing each ensemble member anomaly with the reference data set (here the multi-model ensemble average anomaly). The statistics of interest are the standard deviation of each ensemble member $\{\sigma_{ij} : i = 1, \ldots, N; j = 1, \ldots, n_i\}$ (where N is the total number of models and n_i is the number of ensemble members for model i) and of the reference data set σ_r as well as the correlation coefficient between each ensemble member and the reference data set, denoted $\{\rho_{ij} : i = 1, \ldots, N; j = 1, \ldots, n_i\}$. Both quantities are calculated across all grid cells in the region of interest. These statistics can be further aggregated into the centered pattern root mean square (RMS) difference D for each ensemble member, denoted $\{D_{ij} : i = 1, \ldots, N; j = 1, \ldots, n_i\}$, which can be written in terms of the standard deviations and correlation coefficient as

$$D_{ij} = \sqrt{\sigma_{ij}^2 + \sigma_r^2 - 2\sigma_{ij}\sigma_r\rho_{ij}}. \tag{1}$$

Note that D_{ij} approaches zero as the two fields or patterns become more alike. In the following, RMS differences generically refer to a specific variable (precipitation or temperature) in a specific time period (anomaly in one of the future scenarios) for a specific region (CONUS or one of the subregions). Furthermore, it should be noted that the RMS differences do not account for overall mean differences in the two fields. For the calculation of both centered pattern RMS differences and correlation coefficients, the mean of each field (i.e. the model-ensemble-specific mean anomaly) is subtracted out in the calculation of the correlation coefficient and root mean squared difference before these quantities are plotted. The equations for these two quantities (equations 1 and 2 respectively, from Taylor (2001)) highlight differences in spatial patterns of anomalies rather than mean differences.

The correlation coefficient R between two variables, f_n and r_n, defined at N discrete points in space, with mean values of $\bar{f}$ and $\bar{r}$ respectively is defined as:

$$R = \frac{\frac{1}{N}\sum_{n=1}^{N}(f_n - \bar{f})(r_n - \bar{r})}{\sigma_f \sigma_r} \tag{2}$$

The centered pattern RMS difference, E' is defined as:

$$E' = \frac{1}{N}\sum_{n=1}^{N}([(f_n - \bar{f}) - (r_n - \bar{r})]^2)^{\frac{1}{2}} \tag{3}$$

The Taylor diagram visualizes these three statistics (centered pattern RMS difference, standard deviation, and correlation coefficient) in a single plot. The reference data set used is either the multi-model ensemble average of anomalies. The model-ensemble average anomaly is normalized, and ensemble member anomalies are normalized by the model-ensemble average anomaly.

As discussed in Sect. 2.3.2, these fields are constructed as 30-year averages, and the field of reference is the multi-model ensemble mean, averaging the fields of each ensemble member anomaly for a given 30-year period and inversely weighted by number of ensembles per model. The field of reference is therefore time-dependent, rather than fixed to the present day or historical observation at a certain point in time (i.e. when we calculate the variability ratio for models and ensembles over 2015-2045 versus 2070-2100, the reference field used to calculate this ratio is also constructed over that same time period, rather than the present day). By looking at differences in patterns of anomalies across models at a given time period of a scenario, relative to the multi-model ensemble mean, mean state biases outside of these time horizons do not appear in our analysis.

2.3.5 Quantifying inter- and intra-model variability and their ratio

While Taylor diagrams are helpful for visualizing a large amount of information about IC ensemble behavior between models and within models, the units of distance between points are not uniform or easily interpretable (Gleckler et al. 2008). Even when restricting our scope to monthly climatologies over CONUS and its subregions, the amount of information in the CMIP6 multi-model ensemble across different variables, scenarios, spatial extents, and time periods is too large to summarize concisely. It is necessary to distill down the mean-centered statistics of a Taylor diagram into a single value to describe differences in inter- and intra- model variability. This analysis focuses on differences in spatial patterns of anomalies, and on the variability of these differences across the CMIP6 archive. Specifically, we are interested in how that variability emerges in different regions and at different spatial scales, and how characterizing this variability can assist in the parsimonious sampling of ensemble members for downscaling. Therefore the statistics we use differ from other metrics of model uncertainty and internal variability from other authors, with which the reader might be more familiar (e.g., the three sources of uncertainty discussed in Hawkins and Sutton 2009).

To assess whether a random selection of model ensemble members will produce an unbiased sample of the multi-model ensemble for downscaling temperature and precipitation in a particular region, we can use a standard statistical approach of random effects modeling that quantifies the magnitude of between-group variability relative to the within-group variability (Gelman 2005). In this case, "group" refers to a climate model, with the ensemble members comprising the items in each group; the quantity of interest is the centered RMS differences and their variability across and within models. The standard setup specifies that the ensemble members from each model represent a random sample of all possible IC ensembles, where each model has a model-specific mean centered RMS difference, say $\overline{D}_i$, and some variance τ^2 that does not depend on the model, which will indicate the "within-model" variability. For convenience, a Gaussian distribution is often assumed, wherein

$$D_{ij} \overset{\text{iid}}{\sim} N(\overline{D}_i, \tau^2), \quad i = 1, \dots, N; j = 1, \dots, n_i, \tag{4}$$

where $N(a, b)$ denotes a Normal distribution with mean a and variance b and "$\overset{\text{iid}}{\sim}$" denotes "independent and identically distributed as." The model-specific mean centered RMS difference values are furthermore assumed to arise from a super-population of all possible models that have an overall (across-model, or between-model) mean $\overline{\overline{D}}$ and variance ω^2, again following a Gaussian distribution

$$\overline{D}_i \overset{\text{iid}}{\sim} N(\overline{\overline{D}}, \omega^2), \quad i = 1, \dots, N. \tag{5}$$

In general this framework is robust to the specific random effects distribution in Eq. 5 (McCulloch and Neuhaus 2011); we further explored other distributions for Eqs. 4 and 5 (e.g., log-Normal) and found no difference in our results.

The quantity of interest is then the ratio of the between-model variability to the within-model variability, quantified in terms of the standard deviations ω and τ, denoted

$$R = \frac{\omega}{\tau}, \tag{6}$$

which we henceforth refer to as the "variability ratio," or "R." When $R > 1$ (i.e., the between-model variability is larger than the within-model variability), we can safely conclude that a random sampling of IC ensemble members will generally produce an unbiased sample of the multi-model ensemble. However, if $R < 1$, this indicates that the variability within the various models is larger than the differences between models and one must carefully choose ensemble members in order to sample the multi-model ensemble. Using the assumptions specified by Eqs. 4 and 5, standard statistical software can be used to yield maximum likelihood estimates of the ratio of variances in Eq. 6 and the Delta method can be used to quantify uncertainty in this ratio (for more information, see Appendix 1).

While the types of variability we describe here as "between-model variability" and "within-model variability" may at first glance seem similar to of the sources of uncertainty in Hawkins and Sutton (2009) termed "model uncertainty" and "internal variability," they differ in a few key ways. First, our points of reference differ significantly from that used in Hawkins and Sutton (2009); while the point of origin in their data space is the present day (where change relative to present day values starts at 0, and increases as time or forcing progresses), our data space is centered around a time-varying model-ensemble average. The variability ratio R is therefore not a ratio between the Hawkins and Sutton (2009) model uncertainty and internal variability, but rather a measure of between-model variability within the CMIP6 archive relative to the within-model variability across the models in the CMIP6 archive. It is time-varying and defined by the contents of the archive rather than a historical observational reference point. Second, the spatial correlation metrics we construct in this analysis highlight differences in patterns of anomalies rather than differences in means. As we will see in Sect. 3.2 and Fig. 7 of our results, the behavior of how spatial correlation metrics change as forcing increases will not follow the same trend seen in Hawkins and Sutton of model uncertainty increasing relative to internal variability. (Indeed, we see no significant changes in these patterns across the multi-model ensemble as forcing increases.)

2.3.6 Assessment of robustness to outlier ensemble members

In light of the fact that we plan to draw conclusions about an appropriate sampling scheme for ensemble members for the CMIP6 multi-model ensemble based on the variability ratio and its uncertainty, it is important to ensure that the R metric is robust to the presence of outlier or "extreme" ensemble members from a given model. Recall that if our estimate of R is significantly larger than 1, we conclude that a random sampling of IC ensemble members will produce an unbiased sample of the multi-model ensemble. Consider the following hypothetical scenario: suppose a small number of models (one or more) have a single ensemble member that is systematically different than the others, e.g., one ensemble member projects a decrease to precipitation over CONUS by end-of-century while all others project an increase to precipitation. In this case, random sampling of IC ensemble members may *not* yield an unbiased sample of the CMIP6 multi-model ensemble in the case that the "extreme" ensemble member is selected. The natural question becomes: is our metric R robust to such a scenario? In other words, will our estimate of R (and its uncertainty) reflect the presence of extreme ensemble members?

We conducted a synthetic data test to formally answer this question. Following the total number of models for which we have at least one ensemble member of daily precipitation rate ($N = 52$; see Table 1) and the corresponding number of ensemble members from these models (again see Table 1), we consider synthetic time slices of regionally-averaged anomalies at the end of the 21st Century. Using a Monte Carlo framework, we generate synthetic data and test our calculation of R as follows:

1. Each synthetic model is randomly assigned a mean, say m_i, and standard deviation, say s_i, where the model means range between 1 and 5 and the standard deviations range between 0.1 and 0.5 (note that these are the max possible ranges; individual synthetic data sets can have ranges that are smaller than these ranges).
2. The synthetic time slice y_{ij} for ensemble member $j = 1, \dots, n_i$ of model $i = 1, \dots, N$ (where the n_i are as in Table 1) are drawn from a Normal distribution with mean m_i and standard deviation s_i.
3. Using this synthetic data, we then estimate R and its uncertainty. In this case, none of the models have an "extreme" ensemble member; furthermore, we would expect $R > 1$ since the differences between models (model means range between 1 and 5) is larger than the within-model variability (which ranges between 0.1 and 0.5).
4. Next, we assess the effect of one more models having an extreme ensemble member. For $i = 1, \dots, N$, we
   - (a) Randomly sample i of the models, and
   - (b) For models that have more than one ensemble member, randomly sample one of its members and change the sign of the synthetic time slice y_{ij}.

   In each case we estimate R and its uncertainty from the "adjusted" synthetic data. See Fig. 2a for a sample synthetic data set.
5. Again for each $i = 1, \dots, N$, we calculate two probabilities:
   - (a) The probability of selecting the extreme ensemble member *from the models that have an extreme ensemble member*: on average this will be $\frac{\sum_{i:n_i>1} n_i^{-1}}{\sum_{i:n_i>1} 1} = 0.242$ for $i = 1$ (the average of one divided by the number of ensemble members from models with more than one ensemble member) and $\frac{\sum_{i:n_i>1} 1}{\sum_{i:n_i>1} n_i} = 0.1$ for $i = N$ (one over the average number of ensemble members from models that have more than one member).
   - (b) The probability of selecting the extreme ensemble member *from the entire ensemble*: this is simply $i/\sum_i n_i = i/403$.

We repeat this procedure for 100 synthetic data examples, and then average the best estimates of R and its confidence interval limits over these synthetic replicates. R code for replicating these results is provided in the Supplement.

Results are shown in Fig. 2b, which shows the estimated R and its uncertainty limits averaged over the 100 synthetic data examples. As expected, the variability ratio is significantly greater than 1 for the original data (i.e., when zero models have an extreme ensemble member); notably, R is significantly less than 1 when many of the models have an extreme ensemble member. When two, three, or four models have an extreme ensemble member, the variability ratio is significantly larger than one, with lower bounds of 1.26, 1.12, and 1.02, respectively; however, this represents only (at most) $4/403 = 0.01$ of the ensemble, and furthermore the probability of selecting one of these ensemble members from the "extreme" models is at most around 0.20.

The implications for sampling "extreme", or outlier, ensemble members for downscaling depend on both the existing differences between ensemble members present in the models for the metric of interest, as well as how a specific user of downscaled simulations might define an outlier ensemble member, which will depend on their intended application. In constructing R, we do not want the variability ratio to overemphasize the importance of one or a "small" number of outlier ensemble members,

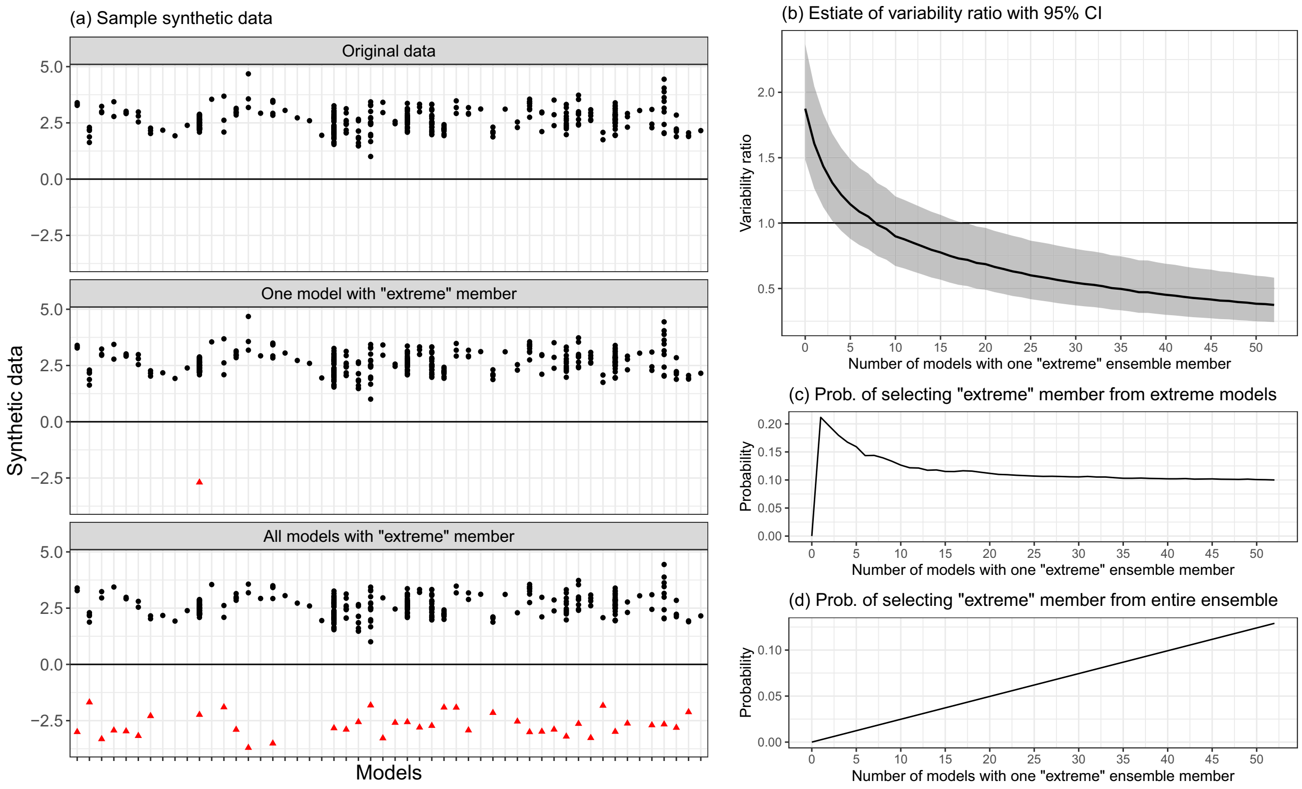

Fig. 2 Illustration of synthetic data (**a**) and corresponding assessment of estimating the variability ratio for sequentially more models that have an extreme ensemble member (**b**). **c**, **d** Show the probability of selecting an extreme ensemble member from the extreme models and selecting an extreme member from the entire ensemble, respectively

as that would render it overly sensitive to small shifts in the underlying set of ensemble members. In the synthetic example here, the variability ratio is not sensitive to small numbers (roughly five or fewer) of outlier ensemble members, as defined in the synthetic example here, and remains greater than one. We also want the variability ratio to be able to drop below one if enough outlier ensemble members are present, as we can see occurs in this synthetic example in Fig. 2b when the number of models with one outlier ensemble member rises above roughly 10. We therefore feel confident that the presence of one or two outlier ensemble members will not impact the variability ratio much, but that many significantly different ensemble members will drive the variability ratio down below one. As this example makes use of synthetic data emulating an unspecified metric, more specific conclusions about outlier ensemble members will be use-specific.

3 Results

3.1 Case study: SSP370 end-of-century

Using Taylor diagrams, we can visually compare between- and within-model ensemble behavior across different shared socioeconomic pathway scenarios (SSP245, SSP370, and SSP585) and regions (all of CONUS and each of the seven different NCA4 regions) for the 30-year end-of-century period spanning 2070–2100. Ensemble-specific anomalies are compared against the multi-model ensemble average of all anomalies for the same variable, scenario and time period, inversely weighted by the number of ensembles from each model. To illustrate the information contained in the large number of analyses that we performed, we selected the SSP370 end-of-century

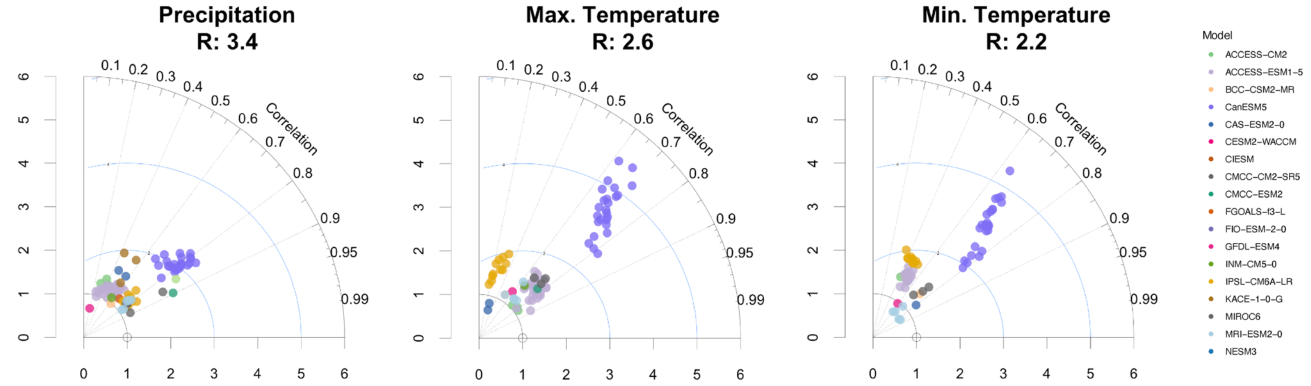

Fig. 3 CONUS-wide Taylor diagrams for the mean annual anomaly (from the historical period of 1980–2015) of daily maximum temperature in SSP370 over 2070–2100, relative to the multi-model ensemble average

projections as a case study. We use the corresponding Taylor diagrams to investigate CONUS-wide and regional patterns of variability, and to demonstrate the relationship between Taylor diagrams and variability ratios. Corresponding plots for the other SSP scenarios are included in Appendices 2 and 3; while the average global mean temperature (GMT) reached by the end of century differs between these scenarios, the end-of-century Taylor diagrams for a given variable and region are similar across scenarios.

These Taylor diagrams also provide a useful demonstration of how the variability ratio (*R*, shown in Fig. 3 in the subtitle of each panel) quantifies the between- to within-model variability in Taylor diagram statistics. The variability ratio is particularly useful in this case where there are a very large number of models and ensemble members and it is difficult to quickly compare the relative magnitude of between- to within-model variability by eye. Across variables and regions at the end of the century (2070–2100, Figs. 4, 5), the estimate of *R* ranges between 0.7 and 5.6, which tells us that the difference between models is roughly up to five and a half times larger than the differences in the ensemble members of an individual model for $R > 1$ (note that the minimum value of 0.7 occurs for SSP245 precipitation in the Northwest, see Appendix 3 Fig. 10). Visually, larger values of *R* reflect the general behavior of the

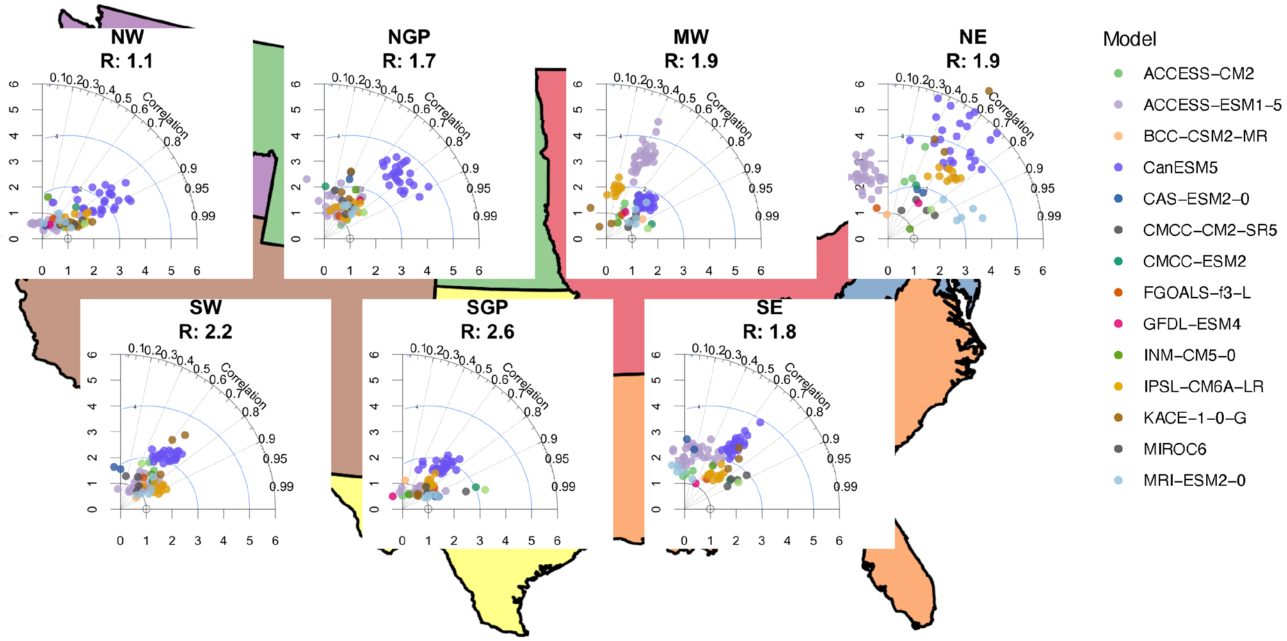

Fig. 4 Regional Taylor diagrams for the mean annual anomaly (from the historical period of 1980–2015) of precipitation in SSP370 over 2070–2100, relative to the multi-model ensemble average

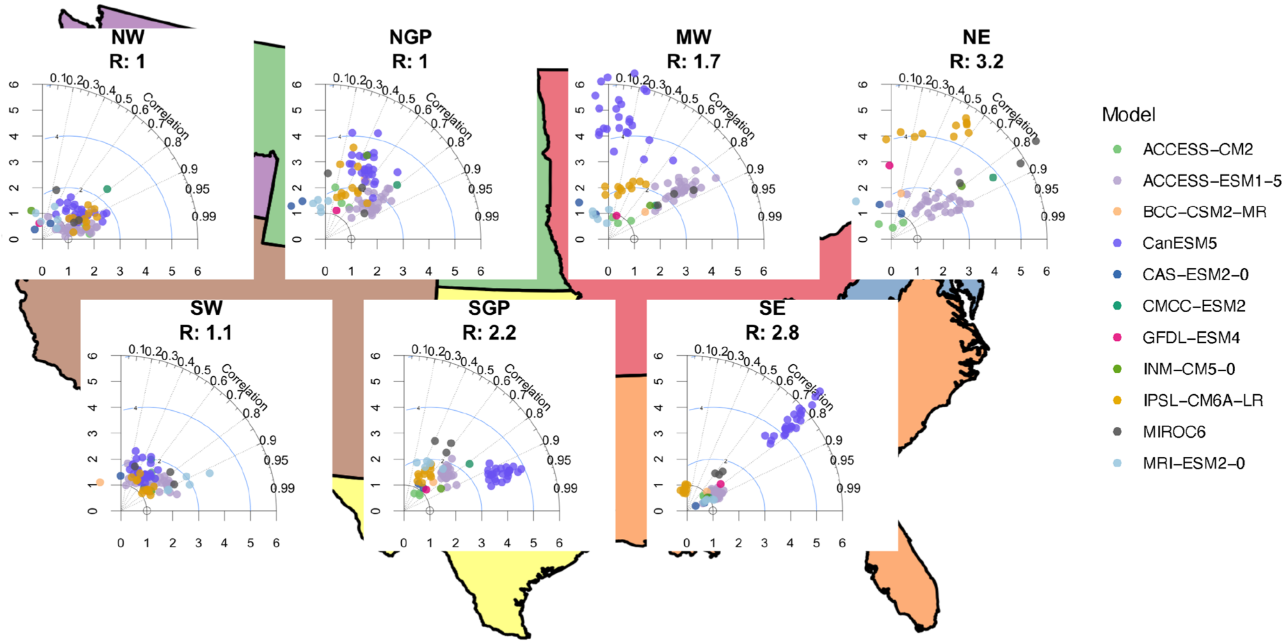

Fig. 5 Regional Taylor diagrams for the mean annual anomaly (from the historical period of 1980–2015) of daily maximum temperature in SSP370 over 2070–2100, relative to the multi-model ensemble average

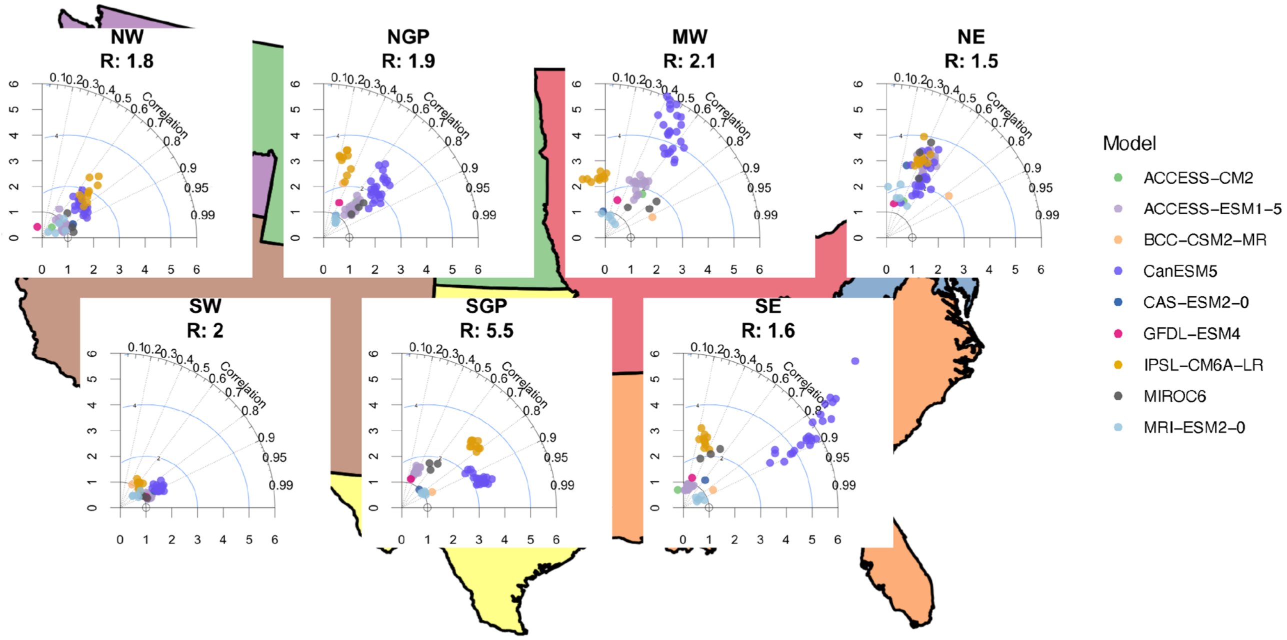

Fig. 6 Regional Taylor diagrams for the mean annual anomaly (from the historical period of 1980–2015) of daily minimum temperature in SSP370 over 2070–2100, relative to the multi-model ensemble average

TD statistics: the ensemble members of a specific model are clustered together such that across models these clusters are separated from one another, as we can observe in Fig. 6 in the Southern Great Plains, where R is relatively high. For this case study, R for maximum temperature in the Northwest and Northern Great Plains is the smallest, at $R = 1$, and R for minimum temperature is the largest (though the magnitude of the confidence intervals on these ratios renders them all similar, as we observe in Fig. 7). Across all regions and variables, with large and small RMS differences,

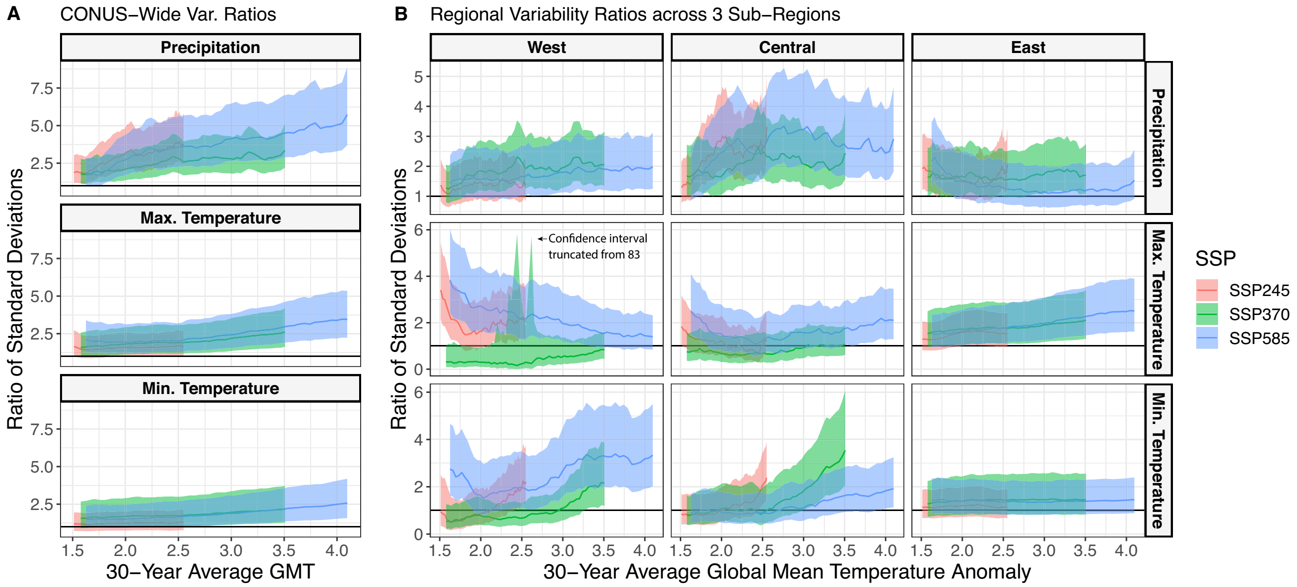

Fig. 7 Variability ratios for all regions and SSPs, for all 30-year time periods from 2015–2045 to 2070–2100, wherein the reference data used were the multi-model-ensemble average across all ensemble members for the same region, SSP, and 30-year time period (weighted inversely by number of ensembles per model, so that models are given equal weight rather than ensemble members). SSP245 is shown in red, SSP370 in green, and SSP585 in blue

individual ensemble members of the same model tend to cluster close to each other around a model-specific "mean;" *R* characterizes the distinctiveness of this grouping relative to the clustering of all ensembles. For annual CONUS-wide estimates, *R* is smaller for precipitation than for surface temperature despite a wider range of RMS differences among all ensembles. Note that *R* characterizes the ratio of between- to within-model differences and remains agnostic to the absolute magnitude of RMS differences.

In the CONUS-wide and regional Taylor diagrams, we observe a mixture of ensemble members clustering with other ensemble members from the same model (e.g. daily minimum temperature in the Southern Great Plains for SSP370, or precipitation in the Northeast and daily maximum temperature in the Southeast), as well as the absence of distinct clustering of ensemble members by models (e.g. precipitation in the Northern Great Plains). Across variables, we generally observe that the spatial patterns in anomalies across ensemble members agree more closely with the multi-model ensemble average in the Southwest, as well as the Northwest. This is plausibly attributable to the topographic constraints of the regions (the Cascade mountain range and the Sierra Nevadas).

3.2 Variability ratios across variable, SSP, and regions

To distill the large amount of information contained in Taylor diagrams of several variables, regions, projections, and time periods, we calculated CONUS-wide and regional variability ratios *R* for moving window 30-year averages across the SSP projections from 2015 to 2100, plotted versus the average change in global mean temperature during that 30-year period for shared socioeconomic pathway (SSP) 245 (red), 370 (green), and 585 (blue). Panel A shows CONUS-wide variability ratios for precipitation (top), maximum temperature (center), and minimum temperature (bottom), while panel B shows regional variability ratios across three regions (West, Central, and East) for the same three variables. Figure 7 shows best estimates of *R* along with a 95% confidence interval.

Variability ratios are plotted as a function not of time but rather of the average global mean temperature (GMT) anomaly during each 30-year period for each SSP (where the anomalies are calculated relative to the pre-industrial period). Two different scenarios with the same x-value thus represent approximately equivalent forcings and do not correspond cleanly to a future time period. As discussed in Sect. 2.3.3, to avoid the small sample size problems of the 7 NCA4 regions when comparing coarsely-regridded models, we calculate these statistics and compare them across three larger custom subregions.

For precipitation and minimum temperature, CONUS-wide and in some regions, variability ratios exhibit plausible increases as global mean temperature rises, indicating that as global mean temperature increases models become more different from other models than from their ensemble members. However the regional variability ratios also exhibit different trajectories between scenarios and variables as warming progresses. Most variability ratios and confidence intervals remain greater than one, but for some regions, variables, and scenarios R remains indistinguishable from one. Most notably, almost no estimates of CONUS-wide and regional variability ratios R have an upper bound confidence interval below one, indicating that the intra-model variability is not necessarily greater than inter-model variability, and distinct realizations of a given model are either as similar to each other or more similar to each other ($R = 1$) than to realizations of a different model ($R > 1$). The notable exception is maximum temperature in the West in the first half of the century for SSP370, for which $R < 1$. CONUS-wide, R ranges between 1.2 and 5.7, indicating that the inter-model variability over these estimates is approximately 1.2 to 5.7 times greater than the intra-model variability. In other words, ensemble members of the same model are generally as similar as or more similar than the typical behavior of an arbitrary ensemble member of a different model.

While we might expect the relative magnitude of between-model variability and within-model variability to change under different levels of warming, the ratio appears generally to increase with global mean temperature. Even towards the end of the century for SSP370 and SSP585, the CONUS-wide R remains larger than one across all three variables. Our observation that between-model variability is generally greater than within-model variability across variables, scenarios, and regions holds as scenarios project into the future, as R plausibly increases with the notable exception of maximum temperature in the West. For maximum temperature in the West we also observe different trajectories for R across scenarios, attributable to very small between- and within-model variabilities contributing to R (note that in 7 a mid-century confidence interval upper bound of 83 is truncated for ease of visualization). CONUS-wide variability ratios increase consistently, but for precipitation in the East, maximum temperature in the Central region, and minimum temperature in the West we observe R plausibly declining in the first half of the century. Confidence intervals remain large however, rendering claims about trajectories of variability ratios as average global mean temperature anomaly increases uncertain.

In summary, while we might expect R to decrease with increasing global mean surface temperature anomalies due to differences in model ensemble fields arising from model internal variability being amplified with increased radiative forcing (e.g., Andrews et al. 2015; Wills et al. 2020), we find that this effect generally remains small even with a high emissions scenario at the end of the 21st Century.

4 Discussion

The increased interest in the scientific community in Earth system modeling, along with the increased interest amongst a wide range of end-users in projects derived from earth system modeling, has led to rapid growth in CMIP modeled output. The size of the CMIP3 project was roughly 36 terabytes (TB), while the size of CMIP5 was roughly 1.8 petabytes (PB), and the CMIP6 archive is expected to be roughly 40 PB in size. At the same time, contributions from an increasing number of modeling centers containing multiple ensemble members and multiple experiments present a practical challenge to comprehensive efforts to downscale such a large and growing set of simulations. In this work, through a set of analyses that survey the CMIP6 multi-model ensemble, we have shown that efficient sampling of the ensemble for the purposes of subsequent downscaling and analysis is more tractable than the exponential growth of CMIP ensembles would suggest. This sampling can support assessments of model skill and weighting, and allow for parsimony in the production of a set of downscaling solutions that captures the range of how the ensemble of ESMs each parameterize atmospheric and surface processes that impact temperature and precipitation.

This is all the more important in the face of greatly-increasing numbers of ensemble members per model, especially since there are significant personnel, computational, and storage costs to downscaling each ensemble member for each model. Previous downscaling activities have operated under the untested assumption that a single ensemble member per model would be sufficient to sample the ensemble. We have tested this assumption by developing a variability ratio metric R to quantify between-to-within-model variability for temperature and precipitation and generally find $R > 1$ across the CONUS and also over CONUS sub-regions over the historical record and through the 21st Century, irrespective of emissions scenario. This finding indicates that the assumption of downscaling a single IC ensemble member is tenable, with caveats. Comparing models and ensembles regionally rather than CONUS-wide shows, perhaps unsurprisingly, a more complicated picture. By our criteria, the case for selecting a single, arbitrary ensemble member per model for downscaling will likely be tenable (because the confidence intervals of R include unity). Nevertheless, in order to ensure that the multi-model does not underestimate the range of mean-state changes, our results indicate that, at the regional level, some caution should be used in the selection of ensemble members and some thought put to which regional metrics and scenarios are of interest to end-users.

Given that one of the primary motivations for the development of multiple IC ensemble members per model is to estimate the internal climate variability of a model (e.g. Kay et al. 2015), it may be initially surprising that each IC ensemble member is close enough to its cohort to largely preclude the need to downscale any more than a small number of IC ensemble members. However, on closer inspection, the metric R shows us that the spatiotemporal correlations of temperature and precipitation in a model exhibit consistent patterns that are distinct for that model. That is why we used the R metric: those patterns are central to LOCA2 prediction. Moreover, LOCA2 and many other statistical methods typically make extensive use of bias correction (BC) (see Teutschbein and Seibert 2012 for a rationale), which removes most of the model internal variability signal that IC ensemble members are designed to encompass. Consequently, BC further diminishes the importance of downscaling many IC members to capture the range in relative changes in hydroclimate variables. At the same time, our findings show that the spatiotemporal correlation structures in different IC members may become more divergent where globally-averaged mean temperature anomalies exceed a 3 °C increase in some regions, such as the central United States, which could be the result of nonlinearities in coupled land-atmosphere processes (e.g., similar to the divergent simulations of the Great Plains Low-Level Jet found in Tang et al. (2017)). All in all, the efforts to develop multiple IC ensemble members as contributions to CMIP6 were not produced in the service of ensuring a realistic range of hydroclimate states across the CONUS; and yet that is precisely the purpose of downscaling a multi-model ensemble. The analysis here shows how to consider the different purposes of these modeling efforts.

There are several caveats to this analysis, however. First, this analysis looked only at IC ensemble members of temperature and precipitation over the CONUS. A similar analysis of other variables, in other regions, and/or perturbed-physics ensemble members (PPEs) may produce different results. Second, we have focused on 30-year averages of monthly climatologies because of the widespread use of climate normals, but analysis over longer or shorter periods could impact our conclusions. Third, we have compared models generating projections at different grid scales and we conservatively remap all models to the scale of the coarsest model. While remapping is necessary for intercomparisons, it also constrains the spatial resolution of any model to that of the coarsest model. We expect conservative remapping to have a small smoothing effect on models being remapped to similar but slightly coarser grid scales. Finally, we want to reiterate the separation between the analysis here and analyses of model skill and weight that often accompany the processing of downscaled solutions.

Despite these caveats, the findings of this analysis support the use of one, or at most a small number of IC ensemble members for downscaling temperature and precipitation in the 21st Century. Our method provides a framework for analyzing multiple models and ensemble members across a broad region, and is applicable to endeavors such as the Fifth National Climate Assessment. Additionally, it points to the sub-linear growth, to date, in personnel, computational, and storage costs for downscaling multi-model ensembles, which have grown by more than an order of magnitude in each successive CMIP phase. Such sub-linear scaling is critically important for the long-term sustainability of downscaling solution development in future CMIP activities. The question of parsimony for developing downscaling solutions is not likely to be made moot by increased computational or personnel resources: models and ensemble members, to date, have been growing in number and complexity with each phase of CMIP, which has tended to increase the potential computational and personnel resources required to develop downscaling solutions. Downscaling solutions remain specialized and have not been designed to scale with increased model and ensemble number. Our findings show that current approaches can reasonably support downscaling solution development for CMIP6, and if similar approaches to historical and scenario-based experiments are adopted for CMIP7 and beyond, the downscaling solutions that currently exist will continue to be able to provide the additional information needed to link coarse-resolution climate change effects as described by ESMs with the fine-scale change projections necessary to develop climate-risk-informed plans and operations at the local level.

Appendix 1: Quantifying uncertainty in the ratio of between- to within- model variability

While standard statistical software can provide maximum likelihood estimates of the between-model standard deviation ω and within-model standard deviation τ, and hence their ratio $R = \omega/\tau$, one can use the Delta method to quantify uncertainty in this ratio. Here, our goal is to construct a confidence interval for the $R = \omega/\tau$. In addition to maximum likelihood estimates of the variances, denoted $\widehat{\omega}$ and $\widehat{\tau}$, the `nlme` package for R (include citation) provides an approximate covariance matrix for the log standard deviations, i.e.,

$$X = \log \omega, \quad Y = \log \tau;$$

we then use these quantities to get at the uncertainty of the quantity of interest using the Delta method. Since the ratio of variances must be positive and a confidence interval should (1) not include negative values and (2) likely not be symmetric about the best estimate, we first construct our confidence interval on the log scale and then exponentiate the end points. The quantity we want the uncertainty of is the log of the ratio of the variances

$$\log \frac{\omega}{\tau} = \log \omega - \log \tau;$$

in terms of $X = \log \omega$ and $Y = \log \tau$, this can be written as

$$f(X, Y) = X - Y.$$

Denote $\widehat{\Sigma}$ = the approximate covariance matrix of the log standard deviations (obtained from the `nlme` package). The Delta method says that the estimated variance of $f(X, Y)$ is

$$\widehat{\mathrm{Var}} f(X, Y) = \nabla f(X, Y)^{\top} \cdot \widehat{\Sigma} \cdot \nabla f(X, Y),$$

where

$$\nabla f(X, Y) = \begin{bmatrix} \frac{\partial}{\partial X} f(X, Y) \\ \frac{\partial}{\partial Y} f(X, Y) \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

A 95% confidence interval for $f(X, Y)$ is then

$$(L_f, U_f) = \left(f(\widehat{X}, \widehat{Y}) - 1.96\sqrt{\widehat{\mathrm{Var}} f(X, Y)}, f(\widehat{X}, \widehat{Y}) + 1.96\sqrt{\widehat{\mathrm{Var}} f(X, Y)} \right)$$

and a corresponding confidence interval for $\exp\{f(X, Y)\} = \exp\{2 \log \omega - 2 \log \tau\} = \frac{\omega^2}{\tau^2}$ is

$$\left(\exp\{L_f\}, \exp\{U_f\} \right).$$

Appendix 2: CONUS-wide Taylor diagrams of SSP245 and SSP585

See Figs. 8 and 9.

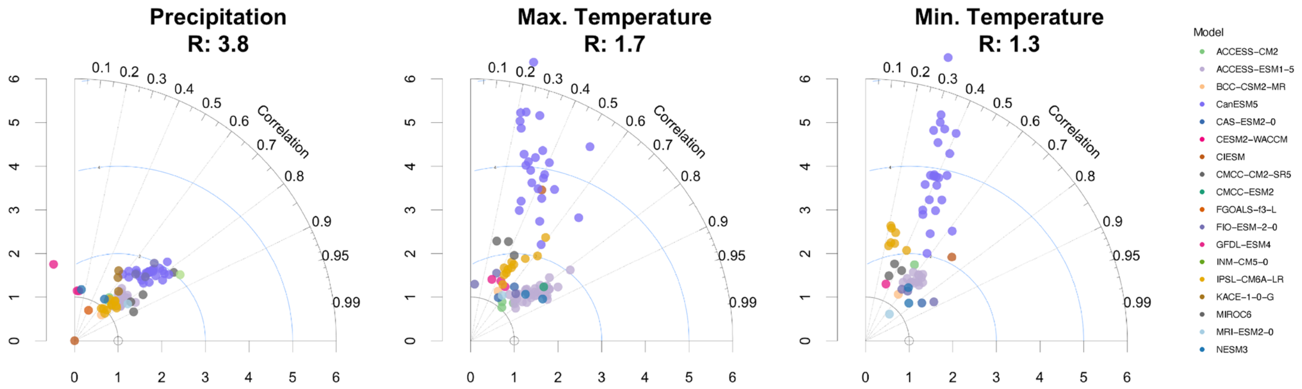

Fig. 8 CONUS-wide Taylor diagrams for the 30-year anomaly of SSP245 relative to the multi-model ensemble average of anomalies

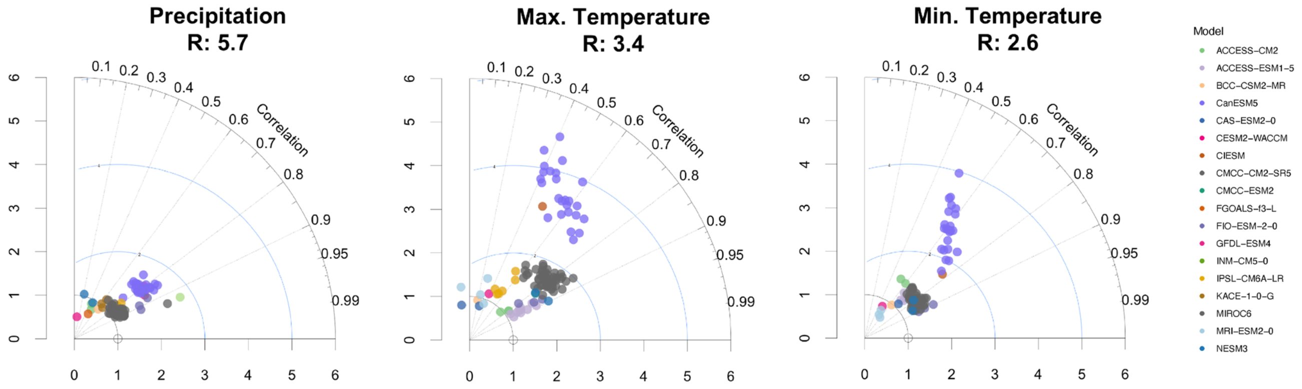

Fig. 9 CONUS-wide Taylor diagrams for the 30-year anomaly of SSP585 relative to the multi-model ensemble average of anomalies

Appendix 3: Regional Taylor diagrams of SSP245 and SSP585

See Figs. 10, 11, 12, 13, 14 and 15.

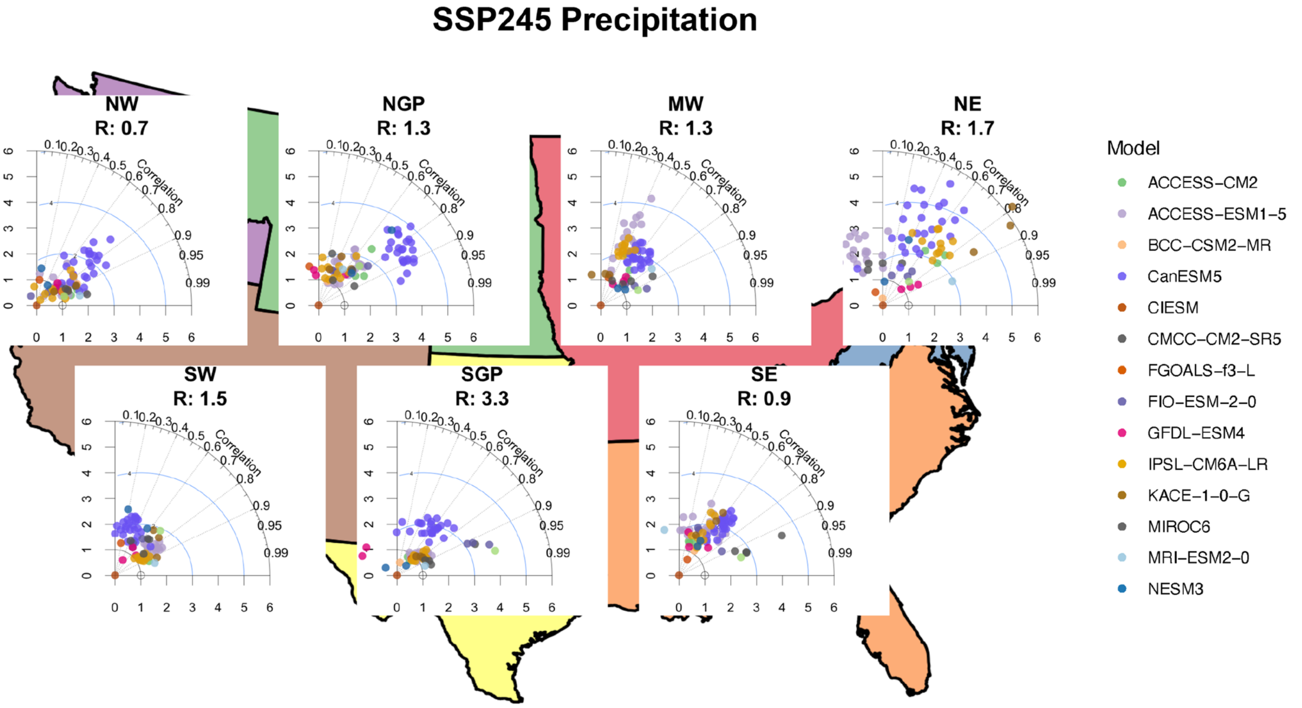

Fig. 10 Regional Taylor diagrams for the 30-year annual average anomalies of precipitation, showing models and ensembles for SSP245 over 2070–2100, compared against the multi-model ensemble average of these anomalies

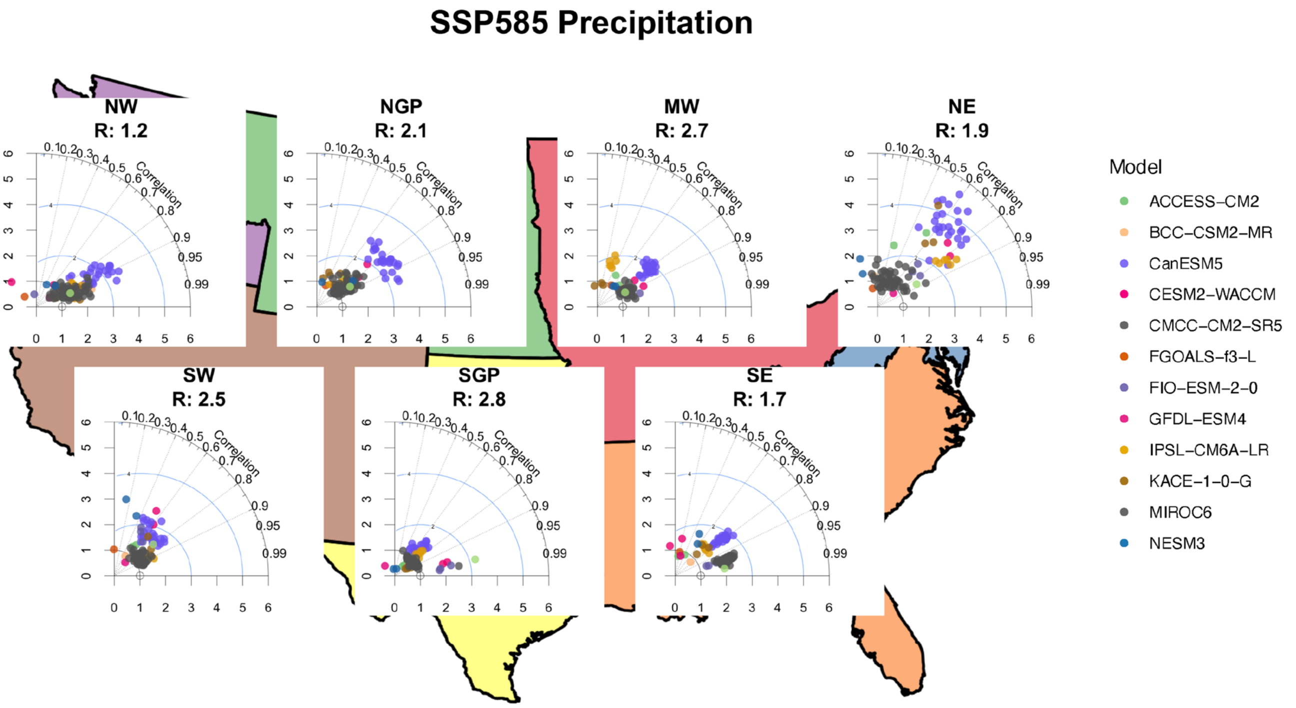

Fig. 11 Regional Taylor diagrams for the 30-year annual average anomalies of precipitation, showing models and ensembles for SSP585 over 2070–2100, compared against the multi-model ensemble average of these anomalies

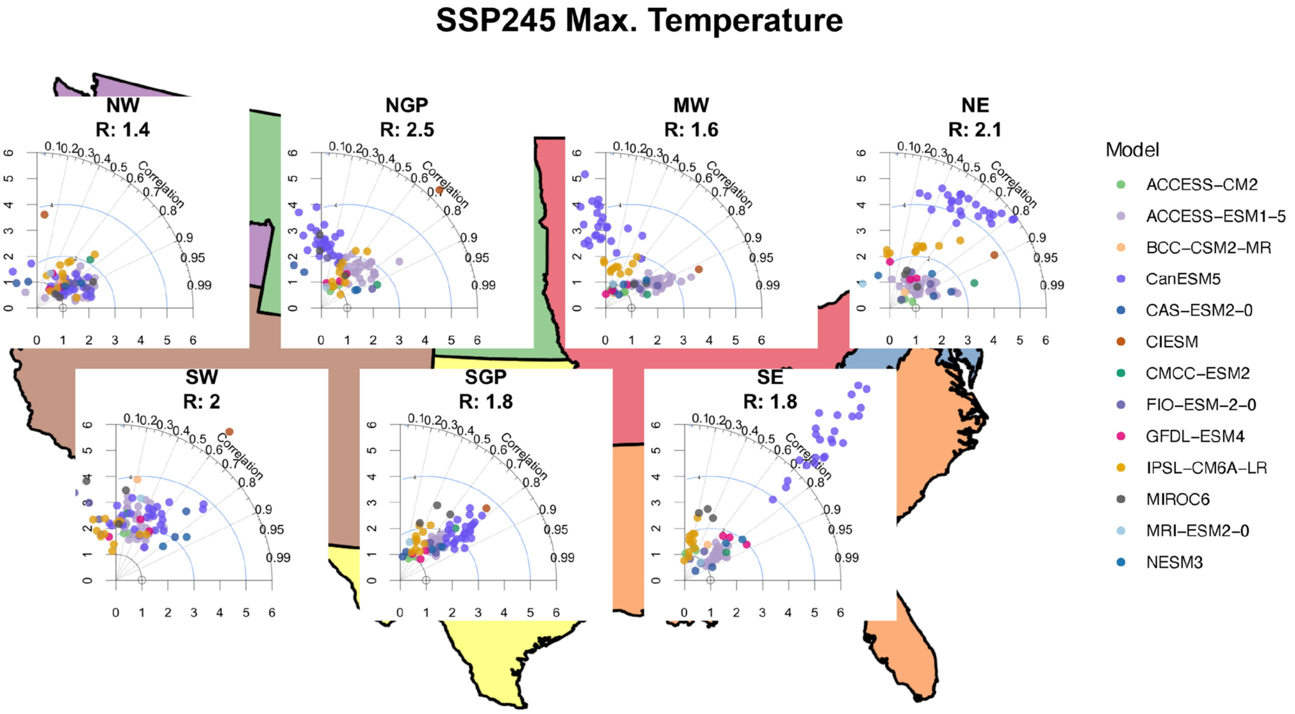

Fig. 12 Regional Taylor diagrams for the 30-year annual average anomalies of maximum temperature, showing models and ensembles for SSP245 over 2070-2100, compared against the multi-model ensemble average of these anomalies

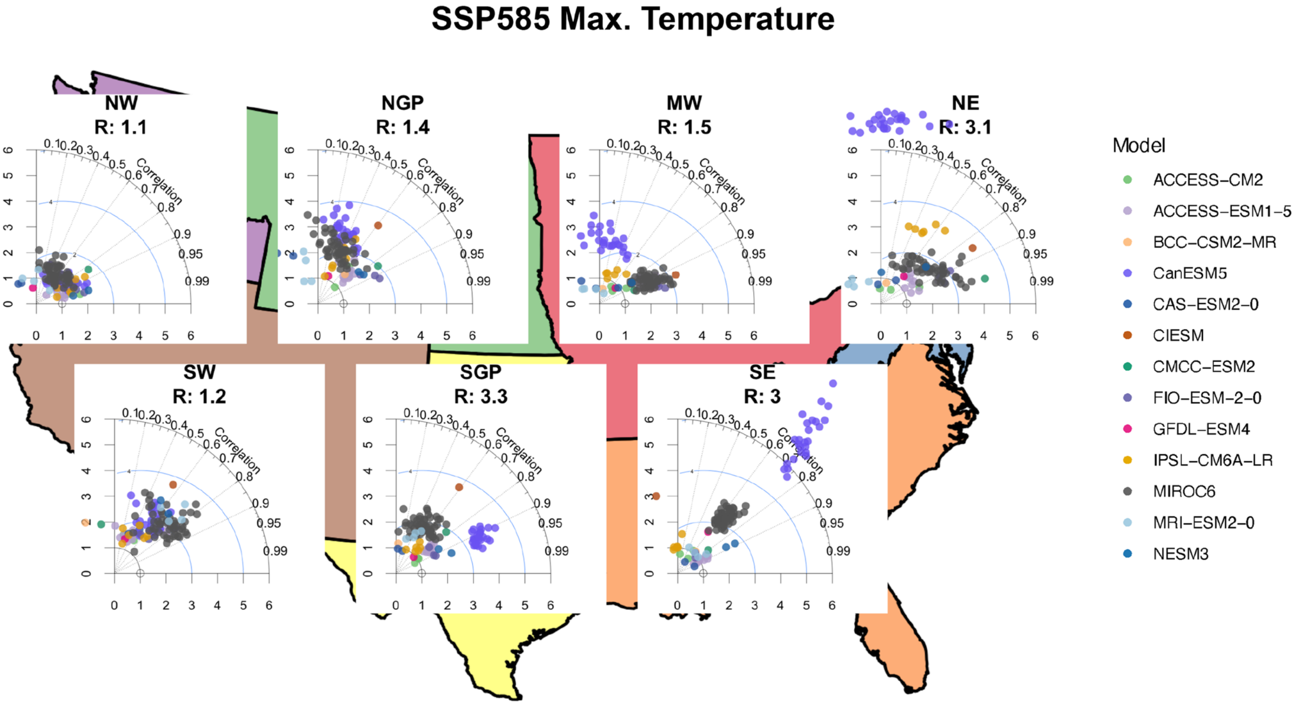

Fig. 13 Regional Taylor diagrams for the 30-year annual average anomalies of maximum temperature, showing models and ensembles for SSP585 over 2070-2100, compared against the multi-model ensemble average of these anomalies

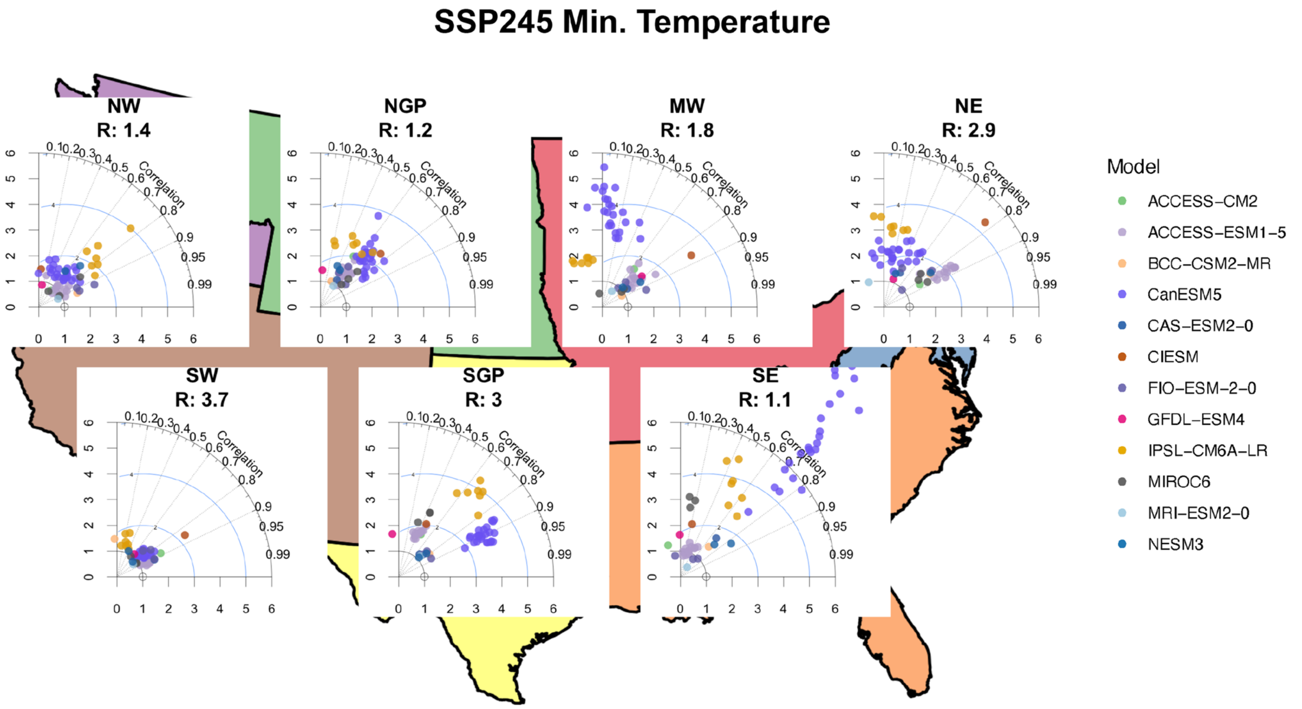

Fig. 14 Regional Taylor diagrams for the 30-year annual average anomalies of minimum temperature, showing models and ensembles for SSP245 over 2070-2100, compared against the multi-model ensemble average of these anomalies

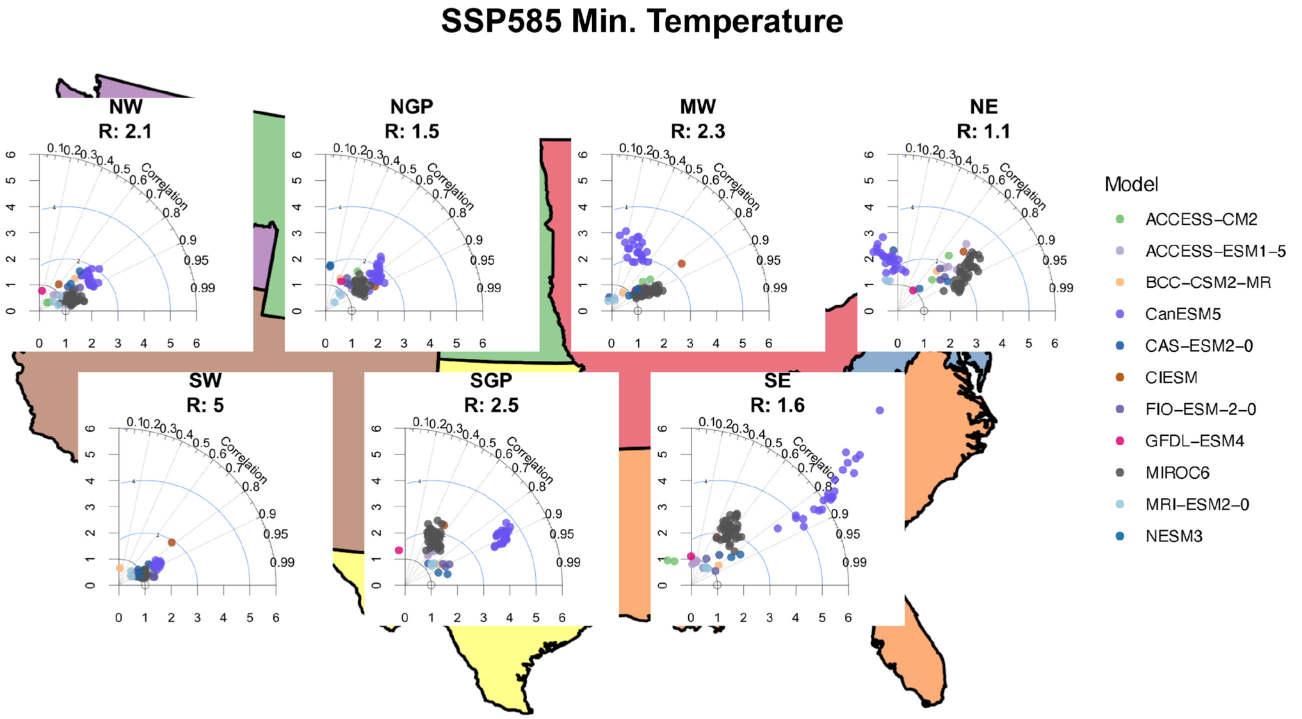

Fig. 15 Regional Taylor diagrams for the 30-year annual average anomalies of minimum temperature, showing models and ensembles for SSP585 over 2070-2100, compared against the multi-model ensemble average of these anomalies

Appendix 4: Mean state change boxplots for NCA4 regions

See Figs. 16, 17, 18, 19, 20 and 21.

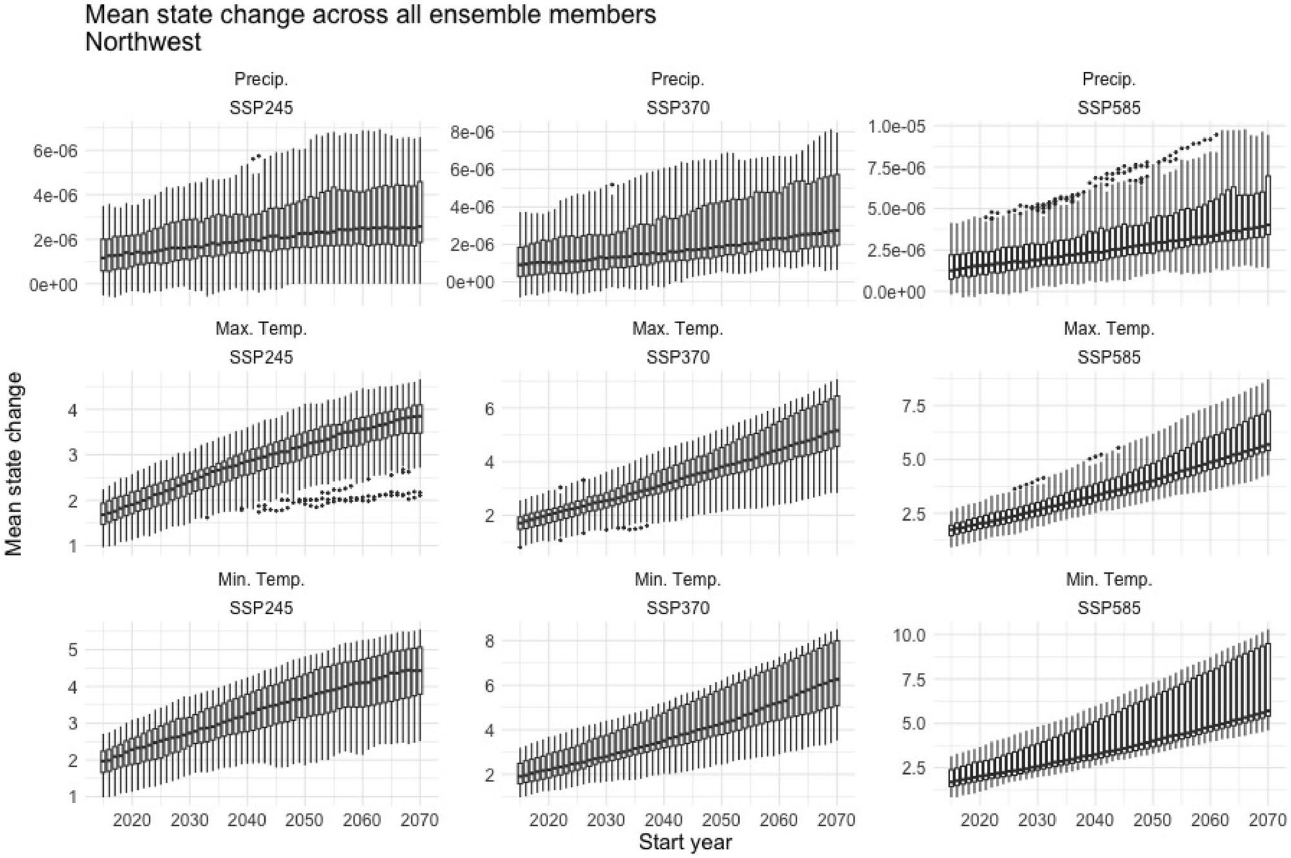

Fig. 16 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods is shown for each variable and scenario, for the NCA4 region of the Northwest. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

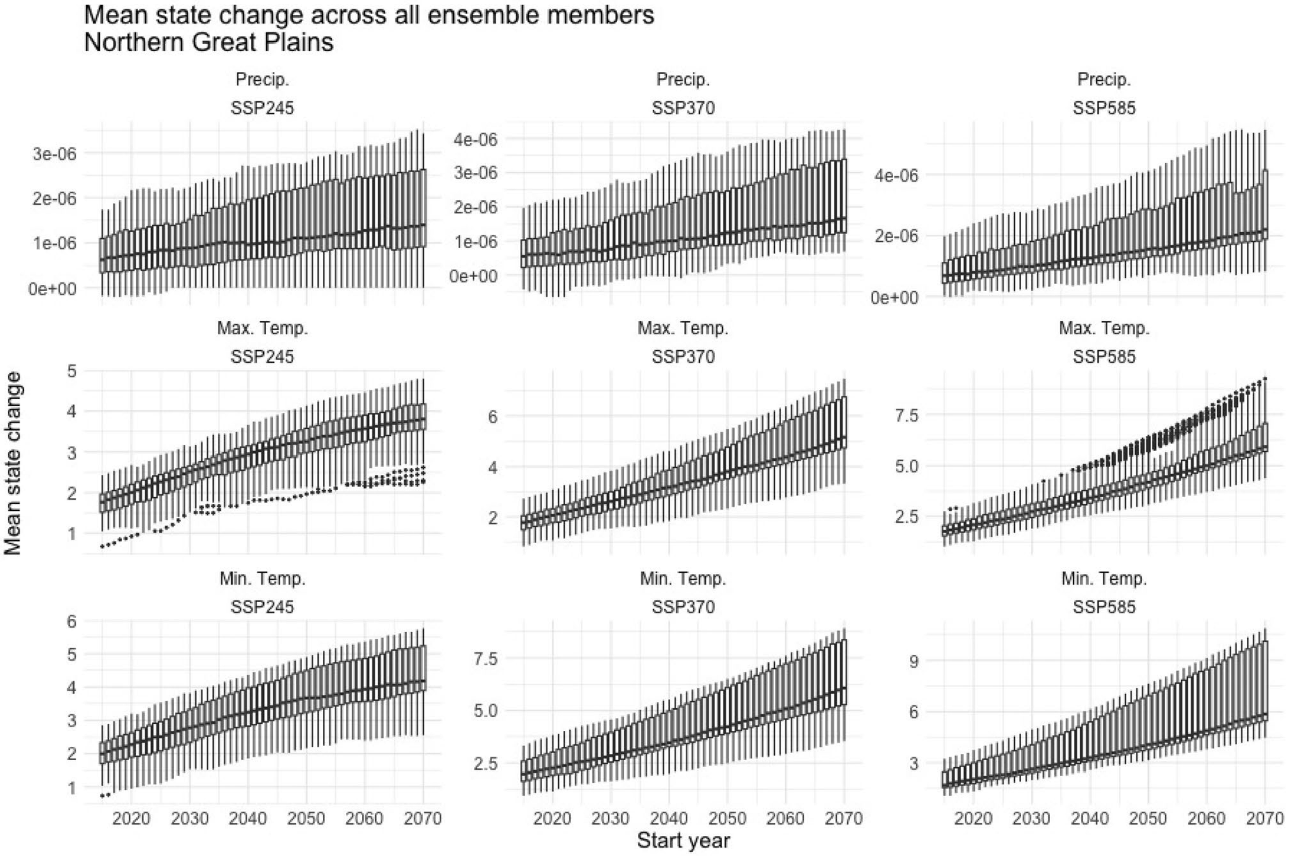

Fig. 17 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods is shown for each variable and scenario, for the NCA4 region of the Northern Great Plains. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

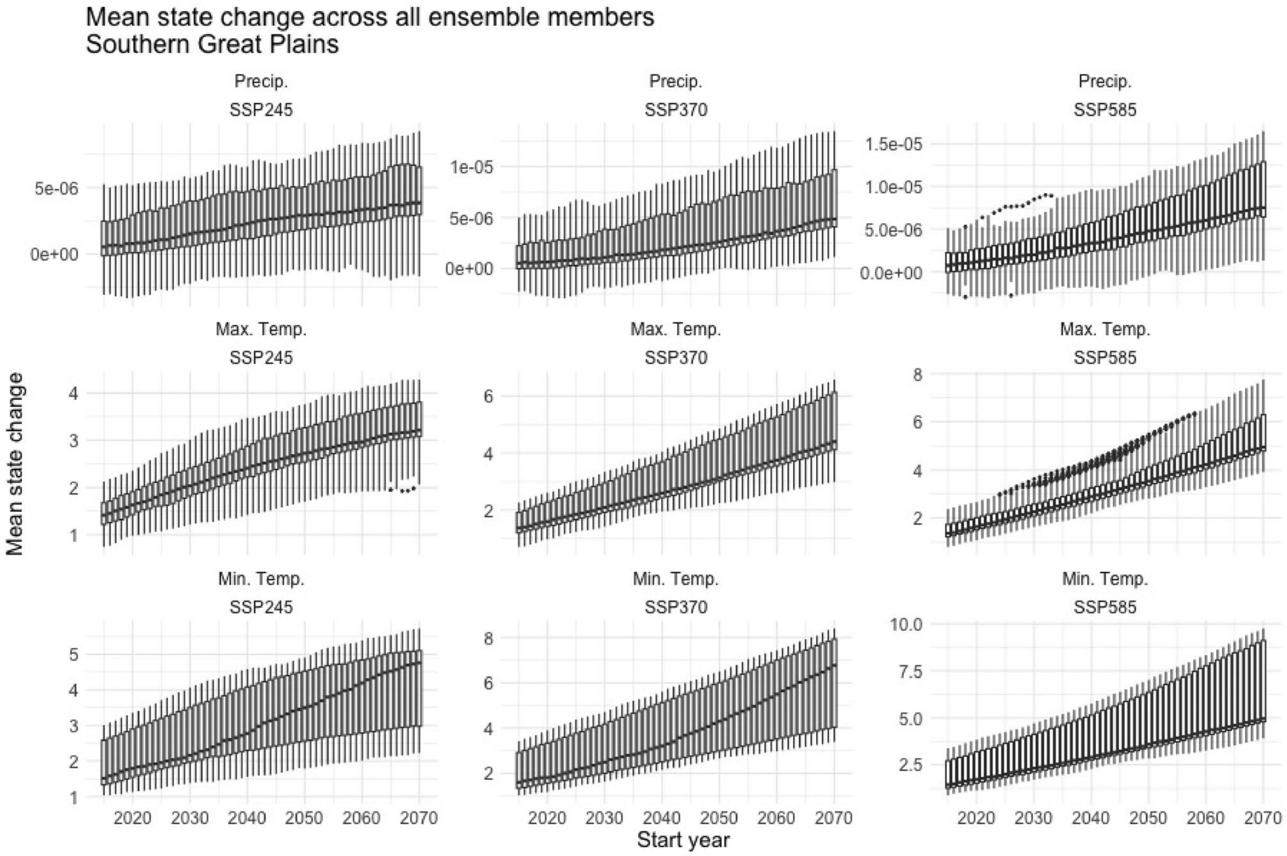

Fig. 18 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods is shown for each variable and scenario, for the NCA4 region of the Southern Great Plains. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

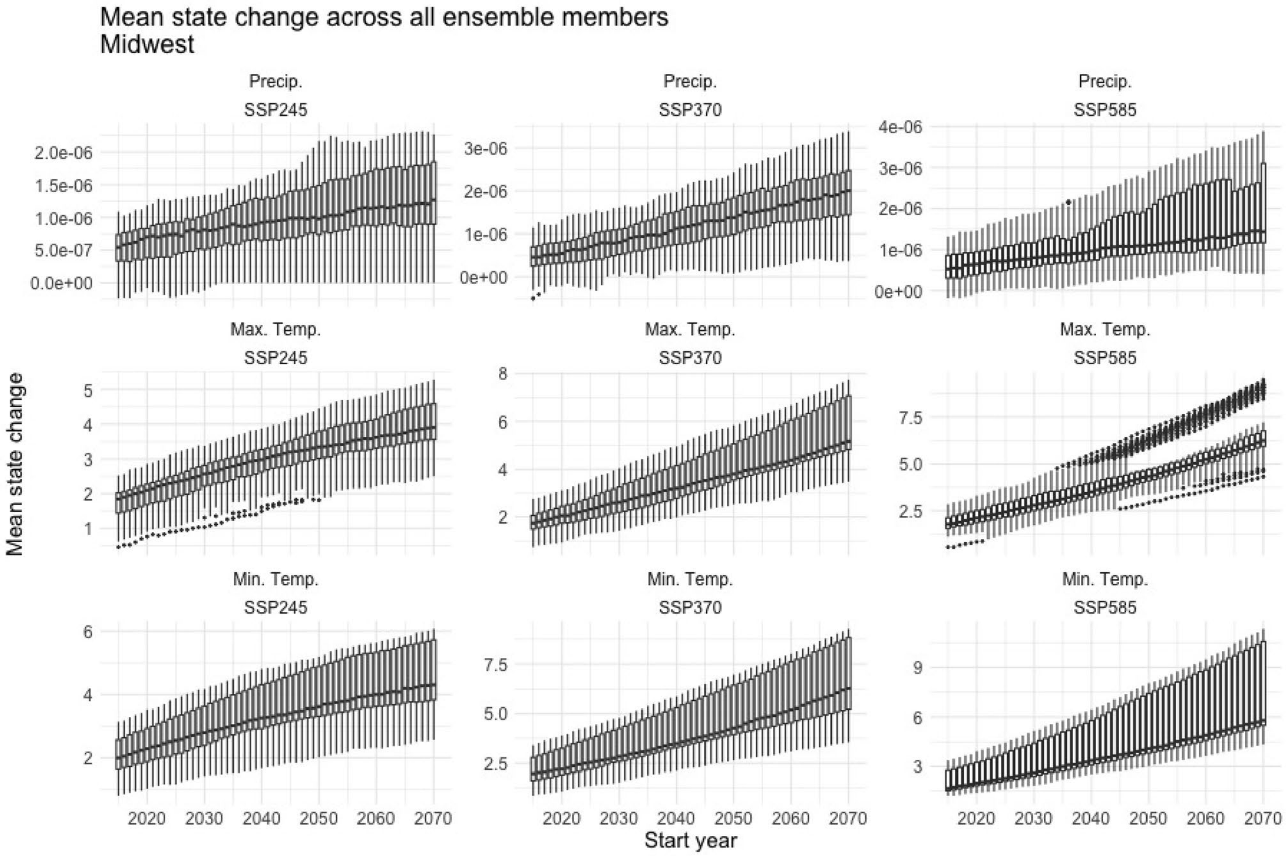

Fig. 19 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods is shown for each variable and scenario, for the NCA4 region of the Midwest. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

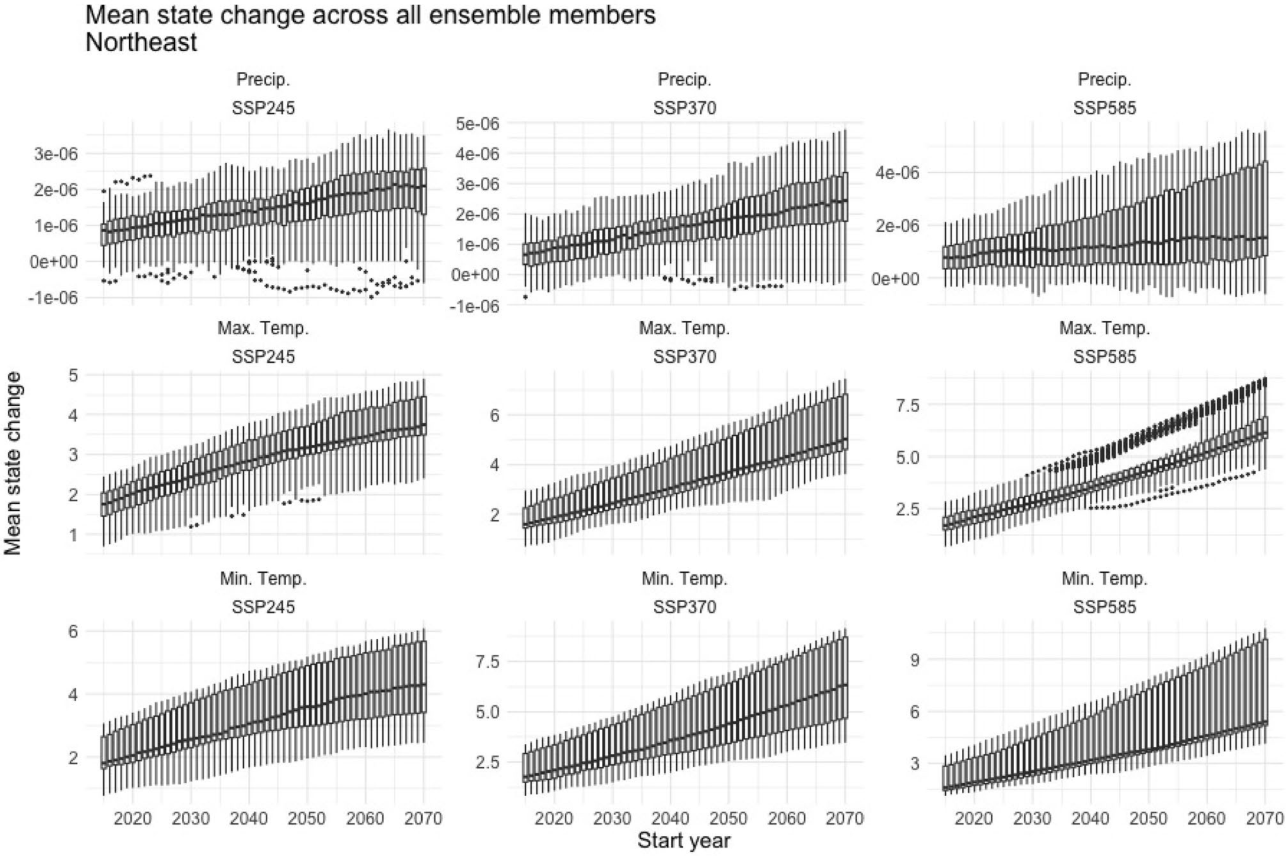

Fig. 20 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods is shown for each variable and scenario, for the NCA4 region of the Northeast. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

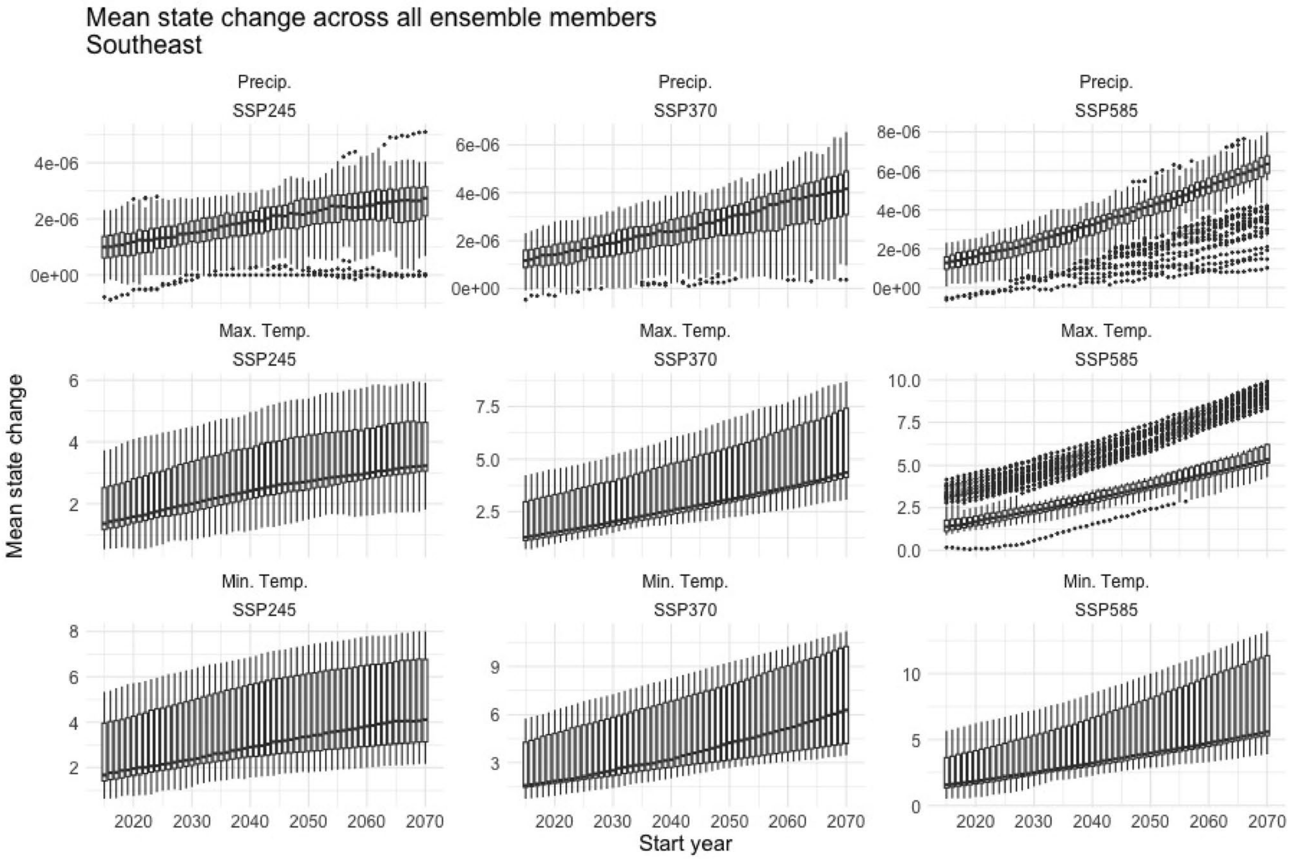

Fig. 21 Box and whisker plots of mean state change for all models and ensemble members for rolling 30-year time periods is shown for each variable and scenario, for the NCA4 region of the Southeast. The units of mean state change for temperature are degrees Celsius, and for precipitation are mm/day

Acknowledgements This project was funded by a contract from the Strategic Environmental Research and Development Program (SERDP) under Project RC19-1391. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Jeffrey Arnold of the U.S. Army Corps of Engineers, and David Pierce and Daniel Cayan of the University of California-San Diego provided guidance on directions for this research. This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Author Contributions All authors contributed to the study conception and design. Data were downloaded by MR, analysis was performed by JL, and figures were created by JL. All authors contributed to the interpretation of results. The manuscript was written by JL and edited by all authors. All authors read and approved the final manuscript.

Funding This project was funded by a contract from the Strategic Environmental Research and Development Program (SERDP) under Project RC19-1391. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Data availability The CMIP6 data analysed during the current study are available in the Earth System Grid Federation repository, accessible at https://esgf-node.llnl.gov/projects/esgf-llnl/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abatzoglou JT, Brown TJ (2012) A comparison of statistical downscaling methods suited for wildfire applications. International Journal of Climatology 32(5):772–780

Andrews T, Gregory JM, Webb MJ (2015) The dependence of radiative forcing and feedback on evolving patterns of surface temperature change in climate models. Journal of Climate 28(4):1630–1648. https://doi.org/10.1175/JCLI-D-14-00545.1, journals.ametsoc.org/view/journals/clim/28/4/jcli-d-14-00545.1.xml

Arguez A, Vose RS (2011) The definition of the standard wmo climate normal: The key to deriving alternative climate normals. Bulletin of the American Meteorological Society 92(6):699–704. https://doi.org/10.1175/2010BAMS2955.1

Brekke L, Thrasher BL, Maurer EP, Pruitt T (2013) Downscaled CMIP3 and CMIP5 climate projections: Release of downscaled cmip5 climate projections, comparison with preceding information, and summary of user needs. Prepared for: Users of the "Downscaled CMIP3 and CMIP5 Climate and Hydrology Projections: Release of Downscaled CMIP5 Climate Projections" http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/

Cinquini L, Crichton D, Mattmann C, Harney J, Shipman G, Wang F, Ananthakrishnan R, Miller N, Denvil S, Morgan M et al (2014) The earth system grid federation: An open infrastructure for access to distributed geospatial data. Future Generation Computer Systems 36:400–417. https://doi.org/10.1016/j.future.2013.07.002

Clark MP, Fan Y, Lawrence DM, Adam JC, Bolster D, Gochis DJ, Hooper RP, Kumar M, Leung LR, Mackay DS et al (2015) Improving the representation of hydrologic processes in earth system models. Water Resources Research 51(8):5929–5956. https://doi.org/10.1002/2015WR017096

Deser C, Phillips A, Bourdette V, Teng H (2012) Uncertainty in climate change projections: the role of internal variability. Climate dynamics 38(3):527–546

Deser C, Phillips AS, Simpson IR, Rosenbloom N, Coleman D, Lehner F, Pendergrass AG, DiNezio P, Stevenson S (2020) Isolating the evolving contributions of anthropogenic aerosols and greenhouse gases: A new cesm1 large ensemble community resource. Journal of climate 33(18):7835–7858

Eyring V, Bony S, Meehl GA, Senior CA, Stevens B, Stouffer RJ, Taylor KE (2016) Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. Geoscientific Model Development 9(5):1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Fowler H, Blenkinsop S, Tebaldi C (2007) Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. International Journal of Climatology 27:1547–1578. https://doi.org/10.1002/joc.1556

Gelman A (2005) Analysis of variance: why it is more important than ever. The Annals of Statistics 33(1):1–53. https://doi.org/10.1214/009053604000001048

Giorgi F, Gutowski WJJ (2015) Regional dynamical downscaling and the CORDEX initiative. Annual Review of Environment and Resources 40:467–490. https://doi.org/10.1146/annurev-environ-102014-021217

Gleckler PJ, Taylor KE, Doutriaux C (2008) Performance metrics for climate models. Journal of Geophysical Research: Atmospheres 113(D6), https://doi.org/10.1029/2007JD008972

Gutmann E, Pruitt T, Clark MP, Brekke L, Arnold JR, Raff DA, Rasmussen RM (2014) An intercomparison of statistical downscaling methods used for water resource assessments in the united states. Water Resources Research 50(9):7167–7186

Gutmann E, Barstad I, Clark M, Arnold J, Rasmussen R (2016) The intermediate complexity atmospheric research model (icar). Journal of Hydrometeorology 17(3):957–973. https://doi.org/10.1175/JHM-D-15-0155.1

Hawkins E, Sutton R (2009) The potential to narrow uncertainty in regional climate predictions. Bulletin of the American Meteorological Society 90(8):1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Jones PW (1999) First-and second-order conservative remapping schemes for grids in spherical coordinates. Monthly Weather Review 127(9):2204–2210. https://doi.org/10.1175/1520-0493(1999)127<2204:FASOCR>2.0.CO;2

Kay JE, Deser C, Phillips A, Mai A, Hannay C, Strand G, Arblaster JM, Bates S, Danabasoglu G, Edwards J et al (2015) The community earth system model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. Bulletin of the American Meteorological Society 96(8):1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

Kim YH, Min SK, Zhang X, Sillmann J, Sandstad M (2020) Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. Weather and Climate Extremes 29. https://doi.org/10.1016/j.wace.2020.100269

Lehner F, Deser C, Maher N, Marotzke J, Fischer EM, Brunner L, Knutti R, Hawkins E (2020) Partitioning climate projection uncertainty with multiple large ensembles and cmip5/6. Earth System Dynamics 11(2):491–508. https://doi.org/10.5194/esd-11-491-2020, esd.copernicus.org/articles/11/491/2020/

Livneh B, Bohn TJ, Pierce DW, Munoz-Arriola F, Nijssen B, Vose R, Cayan DR, Brekke L (2015) A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada 1950–2013. Scientific data 2(1):1–12

Livneh B, Bohn TJ, Pierce DW, Munoz-Arriola F, Nijssen B, Vose R, Cayan DR, Brekke L (2015) A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada (NCEI Accession 0129374). NOAA National Centers for Environmental Information Dataset (Daily precipitation) 2020. https://doi.org/10.7289/v5x34vf6, accessed April 13

Mankin JS, Lehner F, Coats S, McKinnon KA (2020) The value of initial condition large ensembles to robust adaptation decision-making. Earth's Future 8(10):e2012EF001610

Mauritsen T, Stevens B, Roeckner E, Crueger T, Esch M, Giorgetta M, Haak H, Jungclaus J, Klocke D, Matei D, et al. (2012) Tuning the climate of a global model. Journal of advances in modeling Earth systems 4(3)

McCulloch CE, Neuhaus JM (2011) Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. Statistical science 26(3):388–402

McSweeney CF, Jones RG, Booth BB (2012) Selecting ensemble members to provide regional climate change information. Journal of Climate 25(20):7100–7121. https://doi.org/10.1175/JCLI-D-11-00526.1

Melillo TR JM, Yohe G (2014) Climate change impacts in the United States: The third national climate assessment. US Global Change Research Program (841), https://doi.org/10.7930/J0Z31WJ2

Moss R, Kravitz B, Delgado A, Asrar G, Brandenberger J, Wigmosta M, Preston K, Buzan T, Gremillion M, Shaw P, et al. (2017)

Nonstationary weather patterns and extreme events: Informing design and planning for long-lived infrastructure. Tech. rep., ESTCP, https://www.serdp-estcp.org/News-and-Events/Blog/Nonstationary-Weather-Patterns-and-64Extreme-Events-Workshop-Report

Murphy JM, Sexton DM, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430(7001):768–772. https://doi.org/10.1038/nature02771

O'Neill BC, Tebaldi C, Vuuren DPv, Eyring V, Friedlingstein P, Hurtt G, Knutti R, Kriegler E, Lamarque JF, Lowe J, et al (2016) The scenario model intercomparison project (ScenarioMIP) for CMIP6. Geoscientific Model Development 9(9):3461–3482. https://doi.org/10.5194/gmd-9-3461-2016

O'Neill BC, Carter TR, Ebi K, Harrison PA, Kemp-Benedict E, Kok K, Kriegler E, Preston BL, Riahi K, Sillmann J et al (2021) Publisher correction: Achievements and needs for the climate change scenario framework. Nature Climate Change 11(3):274. https://doi.org/10.1038/s41558-020-00981-9

Pierce DW, Cayan DR, Thrasher BL (2014) Statistical downscaling using localized constructed analogs (loca). Journal of Hydrometeorology 15(6):2558–2585. https://doi.org/10.1175/JHM-D-14-0082.158

Pierce DW, Cayan DR, Goodrich J, Das T, Munavar A (2021) Evaluating global climate models for hydrological studies of the upper colorado river basin. JAWRA Journal of the American Water Resources Association. https://doi.org/10.1111/1752-491688.12974

Pierce DW, Su L, Cayan DR, Risser MD, Livneh B, Lettenmaier DP (2021) An extreme-preserving long-term gridded daily precipitation dataset for the conterminous united states. Journal of Hydrometeorology 22(7):1883–1895. https://doi.org/10.1175/JHM-D-20-0212.1

Regonda SK, Zaitchik BF, Badr HS, Rodell M (2016) Using climate regionalization to understand climate forecast system version 2 (cfsv2) precipitation performance for the conterminous united states (conus). Geophysical Research Letters 43(12):6485–6492. https://doi.org/10.1002/2016GL069150

Reidmiller DR, Avery CW, Easterling DR, Kunkel KE, Lewis KL, Maycock TK, Stewart BC (2017) Impacts, risks, and adaptation in the United States. Fourth national climate assessment II. https://doi.org/10.7930/NCA4.2018

Rostron JW, Sexton DM, McSweeney CF, Yamazaki K, Andrews T, Furtado K, Ringer MA, Tsushima Y (2020) The impact of performance filtering on climate feedbacks in a perturbed parameter ensemble. Climate Dynamics 55(3):521–551. https://doi.org/10.1007/s00382-020-05281-8

Schmidt GA, Bader D, Donner LJ, Elsaesser GS, Golaz JC, Hannay C, Molod A, Neale RB, Saha S (2017) Practice and philosophy of climate model tuning across six us modeling centers. Geoscientific Model Development 10(9):3207–3223. https://doi.org/10.5194/gmd-10-3207-2017

Sellar AA, Jones CG, Mulcahy JP, Tang Y, Yool A, Wiltshire A, O'connor FM, Stringer M, Hill R, Palmieri J et al (2019) Ukesm1: Description and evaluation of the uk earth system model. Journal of Advances in Modeling Earth Systems 11(12):4513–4558. https://doi.org/10.1029/2019MS001739

Srivastava A, Grotjahn R, Ullrich PA (2020) Evaluation of historical cmip6 model simulations of extreme precipitation over contiguous us regions. Weather and Climate Extremes 29. https://doi.org/10.1016/j.wace.2020.100268

Stoner AM, Hayhoe K, Yang X, Wuebbles DJ (2013) An asynchronous regional regression model for statistical downscaling of daily climate variables. International Journal of Climatology 33(11):2473–2494. https://doi.org/10.1002/joc.3603

Tang Y, Winkler J, Zhong S, Bian X, Doubler D, Yu L, Walters C (2017) Future changes in the climatology of the great plains low-level jet derived from fine resolution multi-model simulations. Scientific Reports 7(1):5029. https://doi.org/10.1038/s41598-017-05135-0

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres 106(D7):7183–7192. https://doi.org/10.1029/2000JD900719

Teutschbein C, Seibert J (2012) Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. Journal of hydrology 456:12–29

USGCRP (2021) Department of defense climate risk analysis. Department of Defense https://media.defense.gov/2021/Oct/21/2002877353/-671/-1/0/DOD-CLIMATE-RISK-ANALYSIS-FINAL.PDF

Wang C, Zhang L, Lee SK, Wu L, Mechoso CR (2014) A global perspective on cmip5 climate model biases. Nature Climate Change 4:201–205. https://doi.org/10.1038/nclimate2118

Wills RCJ, Battisti DS, Armour KC, Schneider T, Deser C (2020) Pattern recognition methods to separate forced responses from internal variability in climate model ensembles and observations. Journal of Climate 33(20):8693–8719. https://doi.org/10.1175/JCLI-D-19-0855.1, journals.ametsoc.org/view/journals/clim/33/20/jcliD190855.xml

Wood AW, Leung LR, Sridhar V, Lettenmaier D (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Climatic change 62(1):189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

World Meteorological Organization (1989) Calculation of monthly and annual 30-year standard normals. WCDP 10, WMO-TD 341

World Meteorological Organization (2007) The role of climatological normals in a changing climate. Tech. Rep. WCDMP-No. 61, WMO-TD/No. 1377

Wu L, Elshorbagy A, Alam MS (2022) Dynamics of water-energy-food nexus interactions with climate change and policy options. Environmental Research Communications 4. https://doi.org/10.1088/2515-7620/ac4bab