Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Prioritizing the Selection of CMIP6 Model Ensemble Members for DownscalingProjections of CONUS Temperature and Precipitation

Julia M. Longmate (jmlongmate@berkeley.edu)
University of California Berkeley https://orcid.org/0000-0003-2058-5969
Mark D. Risser
Lawrence Berkeley Laboratory: E O Lawrence Berkeley National Laboratory
Daniel R. Feldman
Lawrence Berkeley Laboratory: E O Lawrence Berkeley National Laboratory

Research Article

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

Posted Date: March 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1428854/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Noname manuscript No.
(will be inserted by the editor)

Prioritizing the Selection of CMIP6 Model Ensemble Members for Downscaling Projections of CONUS Temperature and Precipitation

Julia M. Longmate · Mark D. Risser · Daniel R. Feldman

Received: date / Accepted: date

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, 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 models over the conterminous United States (CONUS) and CONUS sub-regions. We calculate the pattern-centered root mean square difference of IC ensemble members relative to multi-model ensemble averages 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 much greater than intra-model variability at both the scales of the CONUS as a whole and CONUS sub-regions, indicating that selecting a single IC ensemble member per model is sufficient to sample the range of projected hydroclimate states in the 21st Century. Regionally-resolved Taylor diagrams identify where more IC ensemble member downscaling efforts should be focused. Our results suggest that, with parsimonious sampling, the cost of downscaling temperature and precipitation fields over the CONUS for CMIP6 may have increased only marginally over previous CMIP activities in spite of greatly increased data volumes.

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 ($\sim$ 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

J. Longmate
University of California, Berkeley, CA, United States
E-mail: jmlongmate@berkeley.edu

M. Risser
Lawrence Berkeley National Laboratory, Berkeley, CA, United States

D. Feldman
Lawrence Berkeley National Laboratory, Berkeley, CA, United States

modeling at large spatial scales and long time-scales instead of at small time and short spatial scales (Clark et al., 2015) , even though the 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 analogue, they are blunt tools for developing projections at the spatial scales of interest for developing local infrastructure and operations plans. 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 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 there are many variables that 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). In light of the volume of data available in the CMIP6 archive, 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. In spite of 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 number of ensemble members per model could over- or under-weight a given model. The idiosyncracies of CMIP models have not diminished, and are unlikely to diminish, over time, so the costs of downscaling must be addressed.

As of this writing, 57 CMIP6 models that have reported results to the Earth System Grid Federation (Cinquini et al., 2014) 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 yet been established. 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.

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 purposes of downscaling the CMIP6 multi-model ensemble, a parsimonious approach must be taken to sample the archive.

This paper is organized as follows: first, in Section 2 we present an overview of the CMIP6 archive, our use of Taylor diagrams and their dispositivity for concisely summarizing multi-model ensembles, and our methodology for quantifying the relative differences between inter- and intra-model variability among the multi-model ensemble. We present the results of our analysis in Section 3 and conclude with a set of recommendations for prioritizing downscaling routines in Section 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 contains contributions from 44 modeling centers and 113 models. While the CMIP6 archive continues to grow as additional outputs and ensemble members are added, 61 models have historical model outputs and 49 have future projections for the variables 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 comparable initialization, physics, and forcing variants and are meant to capture both the climate change and 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 a 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

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).

	# of IC ensemble members				# of IC ensemble members		
Model	*pr*	*tasmax*	*tasmin*	**Model**	*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

broadly relevant to investigations of societal impacts due to changing temperature and precipitation patterns (Wu et al., 2022). These three SSPs are among the scenarios for which the archive contains the most model-ensembles (O'Neill et al., 2021), and the divergent boundary conditions of the scenarios lead to substantial differences in end-of-century hydroclimate (Hawkins and Sutton, 2009). On average, a total of $\sim$ 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, and which overlaps with the time frame of available historical observation data, see Section 2.2) is larger, and contains $\sim$ 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 Observational data product

While most of our explorations evaluate variability around a multi-model average, for reference we also utilize the Livneh et al. (2015a,b) gridded hydrometeorological observational dataset (henceforth L15) for comparison with historical simulations. L15 contains daily precipitation, minimum daily temperature, and maximum daily temperature on a $1/16^{\circ}$ or $\sim$ 6km high-resolution grid beginning in 1950. The L15 data product is largely an extension of Livneh et al. (2013) (henceforth L13) to a larger domain (North America versus CONUS) and L15 extends to 2013 (as opposed to 2011 in L13). The L15 product is used here due to its wide utilization, such as training data for the LOCA statistical downscaling method (Pierce et al., 2014).

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).

	Precip.			**Max. Temp.**			**Min. Temp.**		
Model	*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

2.3 Methods

2.3.1 Regridding model output

Because our analysis of the multi-model ensemble generally must occur on a common grid, we conservatively remap (Jones, 1999) all future projections to the coarsest model resolution in the archive, with roughly 250 km grid cells. Historical simulations however can be compared against the L15 product at the resolutions at which models provide them, as the resolution of L15 is much higher than model output and can be remapped to each individual model resolution.

2.3.2 Summarizing model climatologies

We calculated 30-year averages of monthly data in 30-year moving windows across 2015-2100 for the future projections, and in 30-year windows across 1900-2015 for historical simulations. This length of time 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). Individual ensemble members are compared with a multi-model ensemble average, weighted inversely by the number of ensemble members per each model in order to give models equal weight regardless of the number of IC ensemble members per model. (We also explored the use of unweighted averages and found negligible differences in our results; see Appendix A.) In addition to annual averages, we constructed three-month seasonal averages, focusing on December/January/February (DJF) for precipitation and daily minimum temperature and June/July/August (JJA) for daily maximum temperature.

Historical simulations, compared against both the historical record using the L15 product and the multi-model ensemble mean, are also analyzed to compare and contrast with the analysis of future projections. The similarities between historical analysis relative to L15 and the analysis relative to the multi-model ensemble mean give confidence that our approach of comparing conservatively remapped future projections to the multi-model ensemble mean produces a reasonable measure of the range of the multi-model ensemble, rather than producing a measure of biases in the multi-model ensemble.

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 4th 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 ($\sim$10 in the Northeast). The three custom regions are thus useful for examining different 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, East 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 Section 2.3.4) make use of the NCA4 regions in order to bring regional climatology information to bear on their interpretation, while regional variability ratios (see Section 2.3.5) are calculated across these three larger regions in order to avoid small geospatial sample sizes.

2.3.4 Taylor diagrams

Taylor diagrams (Taylor, 2001) are useful visualizations of the performance of multiple model outputs relative to a reference dataset, most often from observations, and use this comparison to provide quantitative performance metrics for each model. 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. When examining future projections, however, we have no corresponding observations, so a different reference data set must be adopted. The reference dataset that we use for Taylor diagrams of future projections is the multi-model ensemble mean across all models and ensembles, weighting models (rather than ensembles) equally. This approach emphasizes differences between models and ensemble members.

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 with the reference data set (either the multi-model ensemble average or the L15 product). 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 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. Furthermore, it should be noted that the RMS differences do not account for overall mean differences in the two fields. In the following, RMS differences generically refer to a specific variable (precipitation or temperature) in a specific time period (historical or one of the future scenarios) for a specific region (CONUS or one of the subregions).

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 or the L15 product. The model-ensemble average is normalized and ensemble members are normalized by the model-ensemble average.

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.

In order 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 its 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

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

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

$$\overline{D}_i \overset{\text{iid}}{\sim} N(\overline{\overline{D}}, \omega^2), \quad i = 1, \ldots, N. \tag{3}$$

In general this framework is robust to the specific random effects distribution in Eq. 3 (McCulloch and Neuhaus, 2011); we further explored other distributions for Eqs. 2 and 3 (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

$$R = \frac{\omega}{\tau}, \tag{4}$$

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. 2 and 3, standard statistical software can be used to yield maximum likelihood estimates of the ratio of variances in Eq. 4 and the Delta method can be used to quantify uncertainty in this ratio (for more information, see Appendix D).

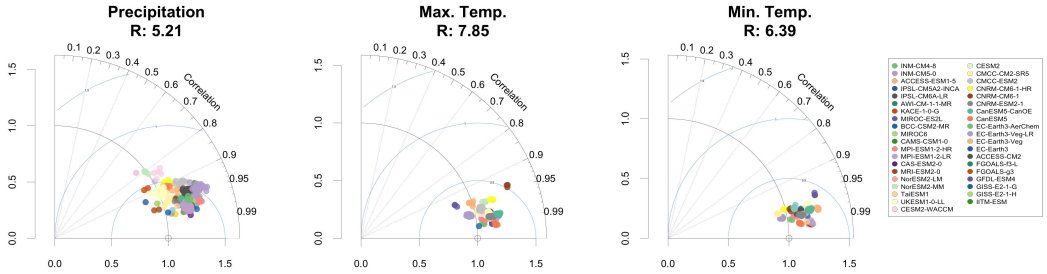
3 Results

3.1 Case study: SSP370 end-of-century

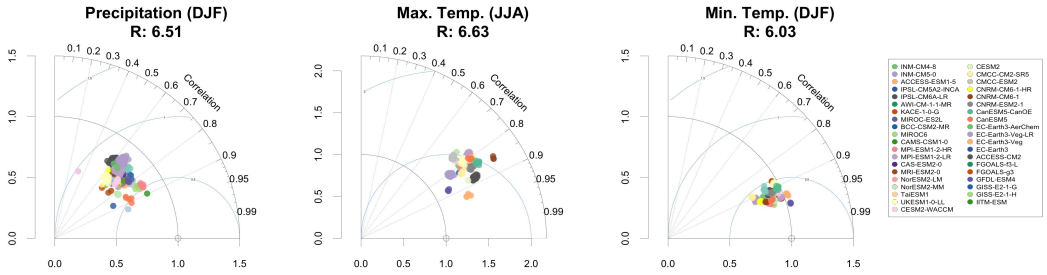
Using Taylor diagrams, we compared between- and within-model ensemble behavior across different shared socioeconomic pathway scenarios (SSP245, SSP370, and SSP585) and regions (all of CONUS and each of the seven different NCA4 regions) for the 30-year end-of-century period spanning 2070-2100. To illustrate the information contained in the large number of analyses that we performed, we selected the SSP370 end-of-century projections as a case study since it represents a mid-range forcing scenario among the SSPs. 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 B and C; 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, region, season and weighting are similar across scenarios.

CONUS-wide Taylor diagrams for precipitation rate, minimum temperature, and maximum temperature are shown in Figure 1. For each variable, the ensemble members of each model have very small RMS differences (generally < 0.5 units) and relative standard deviations close to one, indicating a high degree of correspondence between each ensemble member and the model-ensemble average. The ensemble members have a slight tendency towards under-dispersion (i.e., relative standard deviations less than one) for seasonal precipitation and minimum temperature, and over-dispersion (i.e., relative standard deviations greater than one) for annual estimates and seasonal maximum temperature. This behavior is more prominent for precipitation as precipitation has much larger natural variability than surface temperature, which is reflected in small RMS differences for minimum temperature and relatively larger RMS differences for precipitation.

These Taylor diagrams also provide a useful demonstration of how the variability ratio (R, shown in Figure 1 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 the board, R is at least 5, which tells us that the difference between models is roughly five times larger than the differences in the ensemble members of an individual model. Visually, large values of R like we see here reflect the general behavior of the TD statistics: the ensemble members of a specific model are clustered together such that across models these clusters are separated from one another. For this case study, R for annual precipitation is smallest and R for maximum temperature is the largest (though the



(a) Annual CONUS-wide Taylor diagrams

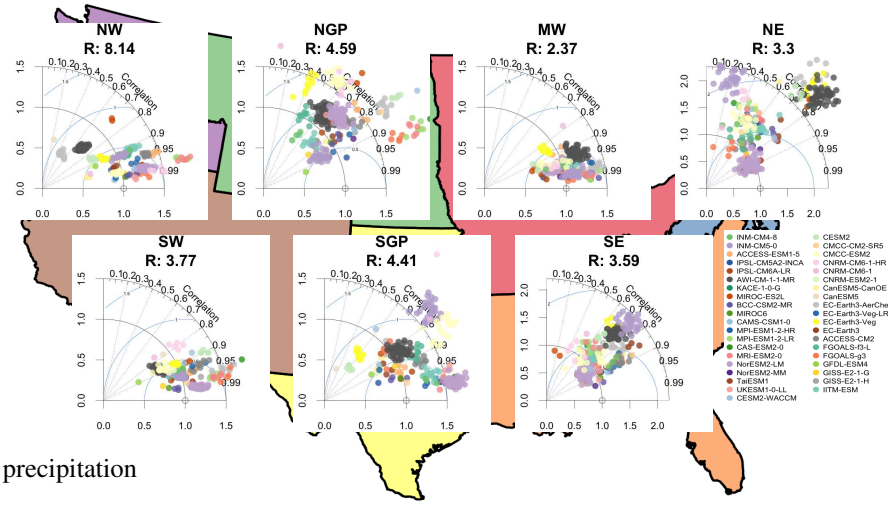


(b) Seasonal CONUS-wide Taylor diagrams

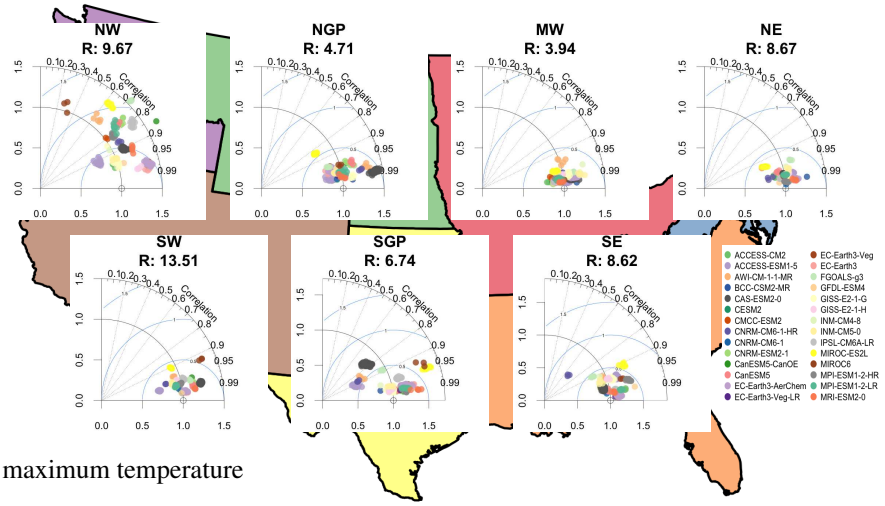
Fig. 1: Taylor diagrams for CONUS-wide 30-year averages of daily precipitation (left), maximum temperature (middle), and minimum temperature (right) from the 2070-2100 period of SSP370. Here, the reference data set is the multi-model ensemble average, and the points shown represent individual ensemble members. Each model is shown with a different color. Panel (a) shows annual estimates and panel (b) shows seasonal estimates.

magnitude of the confidence intervals on these ratios renders them all similar.) Across all regions and variables, with large and small RMS differences, 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.

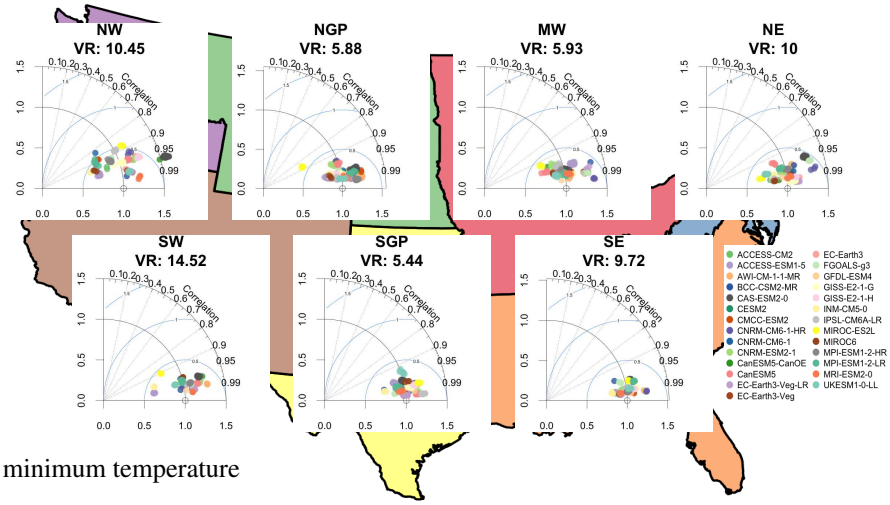
Moving from a CONUS-wide assessment to a regional focus, in Figures 2 and 3 we now show Taylor diagrams for annual and seasonal summaries, respectively, for each variable in each of the seven climate regions (in the CONUS) used by NCA4. Differences in ensemble distribution across Taylor diagrams between regions are in some cases quite large. As in the CONUS-wide Taylor diagrams, ensemble members in regional Taylor diagrams tend to be clustered around model-specific "means," but, as expected, the similarity between individual ensemble members and the model-ensemble average varies more at the regional scale than when considering all of CONUS. Furthermore, these differences between ensemble members relative to the model-ensemble average are generally larger. For the NCA4 regions, precipitation shows largest variability, which is consistent with previous findings that regional variability of precipitation in climate model ensembles grows at smaller spatial scales (Rasmussen et al., 2012; Nguyen et al., 2018). There is one slight peculiarity in the Taylor diagrams, most noticeable for precipitation in the Northwest, wherein models and ensemble members tend have similar pattern correlations but a wider range of relative standard deviations (i.e., the Taylor diagram spread for a given model is predominantly clustered around a single radial value). For Northwest precipitation, this seems to be attributable to large topographic features in this region (the Cascade mountain range) which



(a) Annual precipitation

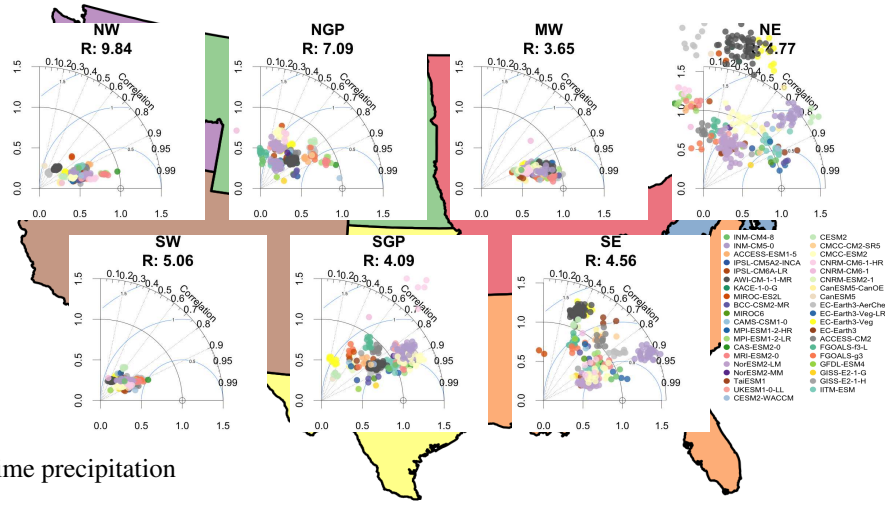


(b) Annual maximum temperature

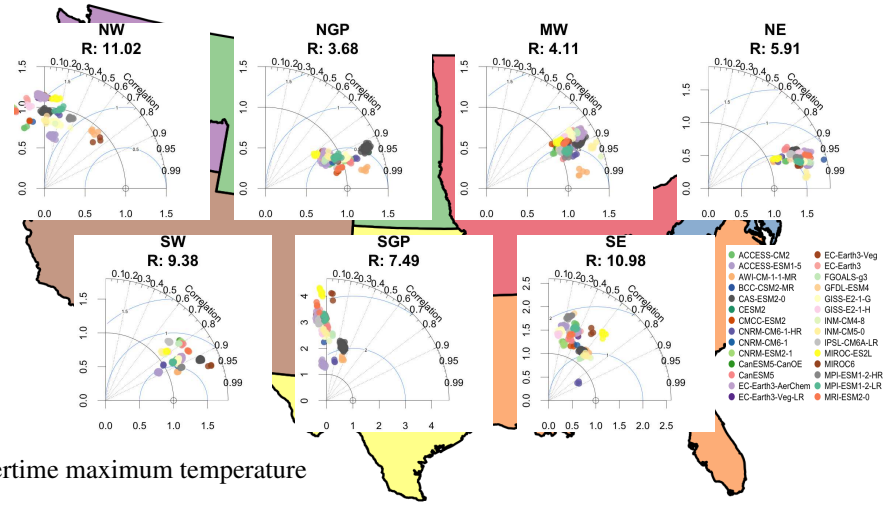


(c) Annual minimum temperature

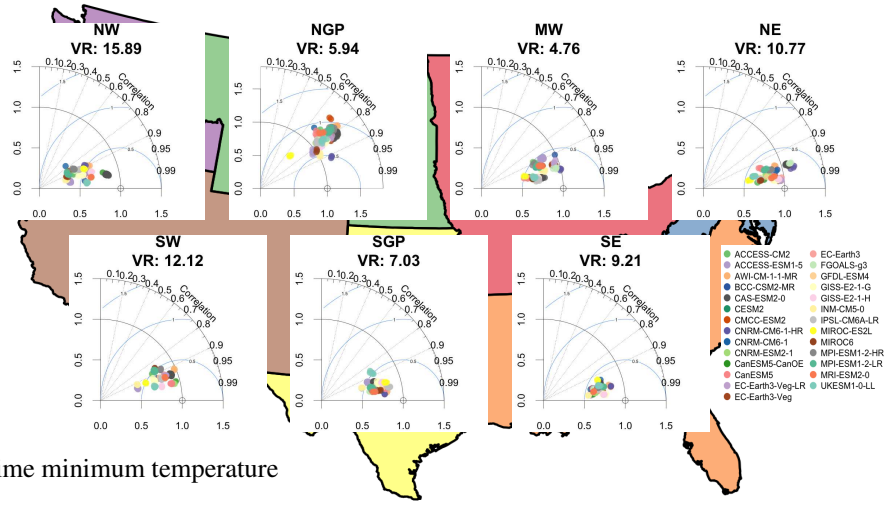
Fig. 2: Regional Taylor diagrams for the 30-year annual average of SSP370 over 2070-2010, split across the seven NCA4 regions, for precipitation (panel a.), maximum temperature (panel b.), and minimum temperature (panel c.). The variability ratio R for each variable/region is shown in the strip text. As in Figure 1, the reference data set is the multi-model ensemble average; points shown represent individual ensemble members with each model is shown using a different color.



(a) Wintertime precipitation



(b) Summertime maximum temperature



(c) Wintertime minimum temperature

Fig. 3: Regional Taylor diagrams for the 30-year seasonal average of SSP370 over 2070-2010, split across the seven NCA4 regions, for wintertime precipitation (panel a.), summertime maximum temperature (panel b.), and wintertime minimum temperature (panel c.). The variability ratio R for each variable/region is shown in the strip text. As in Figure 1, the reference data set is the multi-model ensemble average; points shown represent individual ensemble members with each model is shown using a different color.

constrain the spatial pattern of precipitation across models, as well as to overall greater values for precipitation in this region.

In this case study, regional precipitation and seasonal daily maximum temperature overall exhibit the most distinct ensemble-model clustering patterns between regions, in that the appearance of the Taylor diagram differs greatly between regions. Regional minimum temperature estimates are however clustered tightly near the reference point (i.e., a RMS difference of zero and relative standard deviation of one), indicating a high degree of similarity between all models and ensemble members across all regions. Regional maximum temperature estimates similarly tend to cluster near the reference point, although there is greater variation in the Northwest and Southern Great Plains than in other regions. These comparisons show qualitatively that along these metric dimensions, the differences between models and ensemble members tend to be larger when summarized regionally than when averaged across all of CONUS, and differ greatly between regions.

Regional variability ratios are generally smaller for annual calculations than seasonal. We expect to observe smaller variability ratios for Taylor diagrams in which we observe large differences between ensemble members but not distinct differences between ensemble members based on their model. We do for example see this in the differences between Northern Great Plains and Northwest annual precipitation, in Figure 2; ensembles of a given model in Northwest are more distinctly clustered around a model-specific mean and separated from different models, indicating higher between-model variability, and, as we observe, likely a higher R than Northern Great Plains, which has less distinct separation between ensembles by model. For variables and regions in which all ensembles are highly similar to the multi-model ensemble average, values of R still range between $\sim 4 - 14$, as we observe in the difference between the low R in the Midwest compared with the relatively high R for the Southwest in annual maximum temperature in Figure 2.

3.2 Variability ratios across variable, SSP, and regions

To distill the large amount of information summarized by Taylor diagrams across variables, regions, projections, and time period, we calculated CONUS-wide and regional variability ratios R for moving window 30-year averages across the SSP projections from 2015-2100. Figure 4 shows best estimates of R along with a 95% confidence interval, plotted as a function 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). For reference, the variability ratio for the historical simulations and its 95% confidence interval is shown at the minimum mean GMT, relative both to L15 historical observations as well as the multi-model ensemble average across historical simulation ensembles. As discussed in Section 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 across three larger custom subregions.

Most notably, all estimates of CONUS-wide and regional variability ratios R are larger than one, indicating that the inter-model variability is larger than intra-model variability, and distinct realizations of a given model are more similar to each other than to realizations of a different model. This remains true even when considering the uncertainty (i.e., the confidence intervals for R). CONUS-wide, R ranges between five and ten, indicating that the inter-model variability over these estimates is approximately 5-10 times greater than the intra-model variability. In other words, the differences between ensemble members of the same model are much more similar than the typical behavior of an arbitrary ensemble member of a different model.

In these regional assessments, estimates of R do approach 1, particularly over 30-year time periods in future scenarios corresponding to higher mean GMT, and particularly for precipitation in the Central and East regions, and as well as minimum temperature in the East region. Larger estimates of R are generally observed in the western regions. Notably R for minimum temperature in the West is quite large, between-model variance is ~ 13 times larger than within-model variance in the West, though all ensemble members plotted in Taylor

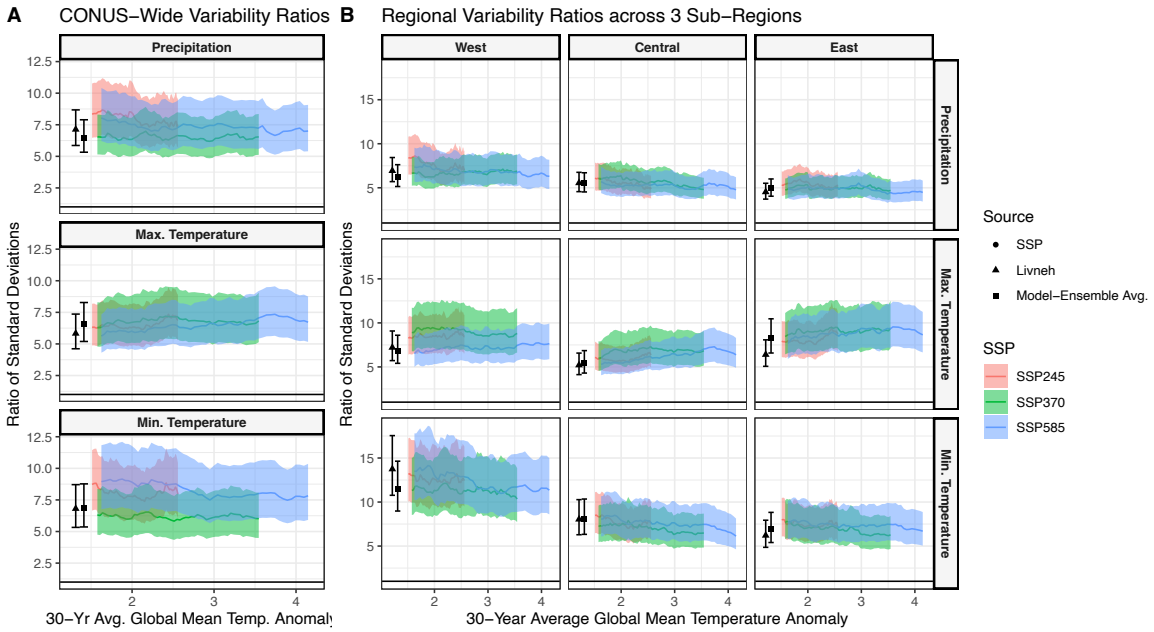


Fig. 4: The variability ratio of between-model to within-model standard deviations across rolling 30-year windows from 2015 to 2100, plotted versus the average change in global mean temperature during that 30-year period for shared socioeconomic pathway (SSP) 245 (red), 370 (green), and 585 (blue). Panel 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. The corresponding variability ratio for the historical simulations (1980-2010) is shown with a black error bar, and a reference line at $R = 1$ is shown in all panels.

diagrams for minimum temperature in Figures 2 and 3 are clustered quite close to the model-ensemble average. While we might expect the relative magnitude of between-model variability and within-model variability to change under different levels of warming, the ratio does not exhibit any clear relationship with global mean temperature. Even towards the end of the century for SSP585, the CONUS-wide R remains larger than one across all three variables. Our observation that between-model variability is consistently greater than within-model variability across variables and scenarios holds as scenarios project into the future.

Historical variability ratios constructed from the 30-year period at the end of the historical simulations (1985-2015, shown as black points, Figure 4) are similar to those estimated for each SSP in the 30-year period at the beginning of the future projections (2015-2045). Confidence intervals are also similar, if slightly narrower. While there are differences between analyzing the historical simulations and future projections (in particular that the historical simulations have more models and ensembles, and have a historical record against which they can be compared), all historical R estimates for the end of the historical period lie within the confidence intervals of R estimates at the beginning of the SSPs. This indicates that there are likely sufficient models and ensembles for the estimation of R to be robust to the inclusion or exclusion of individual models or ensembles.

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, we find that this effect 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 we have shown through a set of analyses that survey the CMIP6 multi-model ensemble 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 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 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.

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.

In spite of these caveats, the findings of this analysis clearly 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 5th National Climate Assessment (NCA5). 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.

Acknowledgments

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.

References

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

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, doi: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" URL 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, doi: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, doi:https://doi: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, doi:10.5194/gmd-9-1937-2016

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

Giorgi F, Gutowski WJJ (2015) Regional dynamical downscaling and the CORDEX initiative. Annual Review of Environment and Resources 40:467–490, doi: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), doi: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, doi: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, doi: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, doi: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, doi: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, doi:https://doi.org/10.1016/j.wace.2020.100269

Livneh B, Rosenberg EA, Lin C, Nijssen B, Mishra V, Andreadis KM, Maurer EP, Lettenmaier DP (2013) A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states: Update and extensions. Journal of Climate 26(23):9384–9392

Livneh B, Bohn TJ, Pierce DW, Munoz-Arriola F, Nijssen B, Vose R, Cayan DR, Brekke L (2015a) 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 (2015b) 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) doi:https://doi.org/10.7289/v5x34vf6, accessed April 13, 2020.

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, doi: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), doi: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, URL 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, doi:10.1038/nature02771

Nguyen P, Thorstensen A, Sorooshian S, Hsu K, Aghakouchak A, Ashouri H, Tran H, Braithwaite D (2018) Global precipitation trends across spatial scales using satellite observations. Bulletin of the American Meteorological Society 99(4):689 – 697, doi:10.1175/BAMS-D-17-0065.1

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, doi: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, doi: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, doi:https://doi.org/10.1175/JHM-D-14-0082.158

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

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

Rasmussen SH, Christensen JH, Drews M, Gochis DJ, Refsgaard JC (2012) Spatial-scale characteristics of precipitation simulated by regional climate models and the implications for hydrological modeling. Journal of Hydrometeorology 13(6):1817 – 1835, doi:10.1175/JHM-D-12-07.1

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, doi: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, volume II doi: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, doi: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, doi: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, doi: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, doi: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, doi:https://doi.org/10.1002/joc.3603

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

USGCRP (2021) Department of defense climate risk analysis. Department of Defense URL 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, doi:https://doi.org/10.1038/nclimate2118

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, doi: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, doi:https://doi.org/10.1088/2515-7620/ac4bab

Statements and declarations

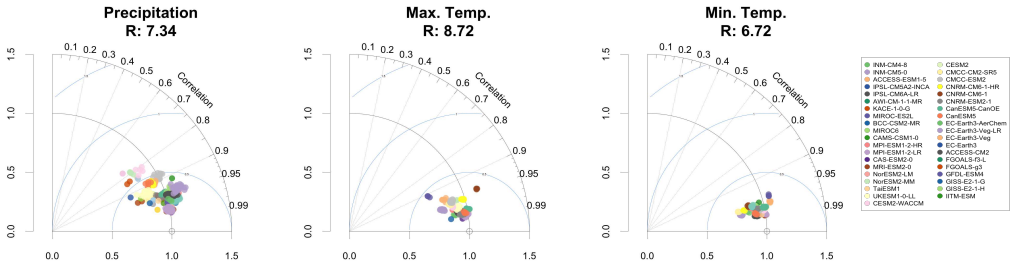
Funding This project was funded by a contract from the Strategic Environmental Research and Development Program (SERDP) under Project RC19-1391. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science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

Competing interests The authors declare that they have no conflict of interest.

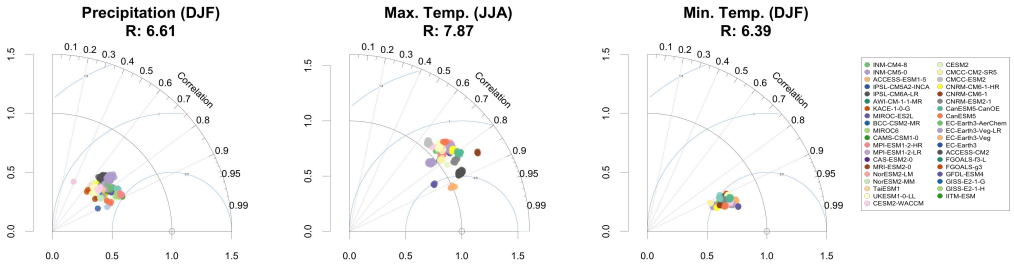
Author contributions All authors contributed to the study conception and design. Data were downloaded by Mark Risser, analysis was performed by Julia Longmate, and figures were created by Julia Longmate. All authors contributed to the interpretation of results. The manuscript was written by Julia Longmate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

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/. The Livneh et al. (2015a,b)) historical hydrometeorological dataset (L15) used here is made available by the authors and can be accessed on the National Oceanographic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI) database at https://www.ncei.noaa.gov/access/metadata/landing-page/bin/iso?id=gov.noaa.nodc:0129374.

A Additional future projections: SSP370 Taylor diagrams, unweighted

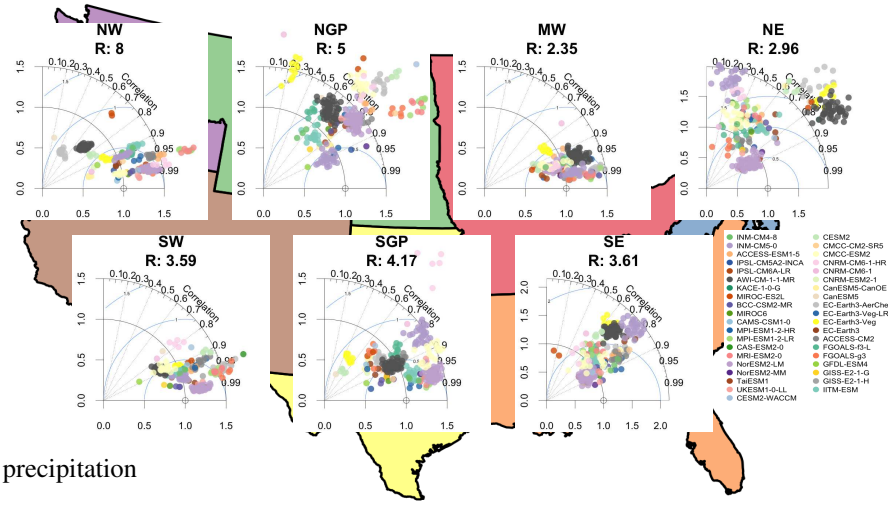


(a) CONUS annual, 2070-2100

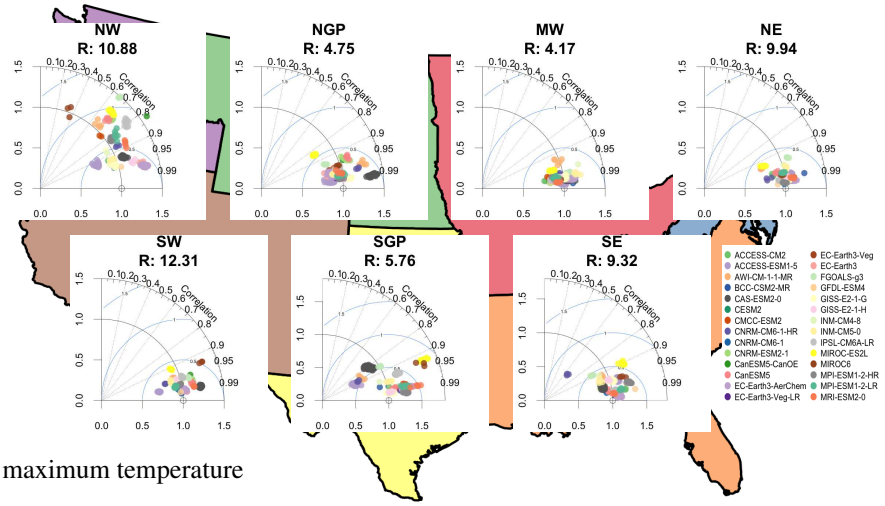


(b) CONUS seasonal, 2070-2100

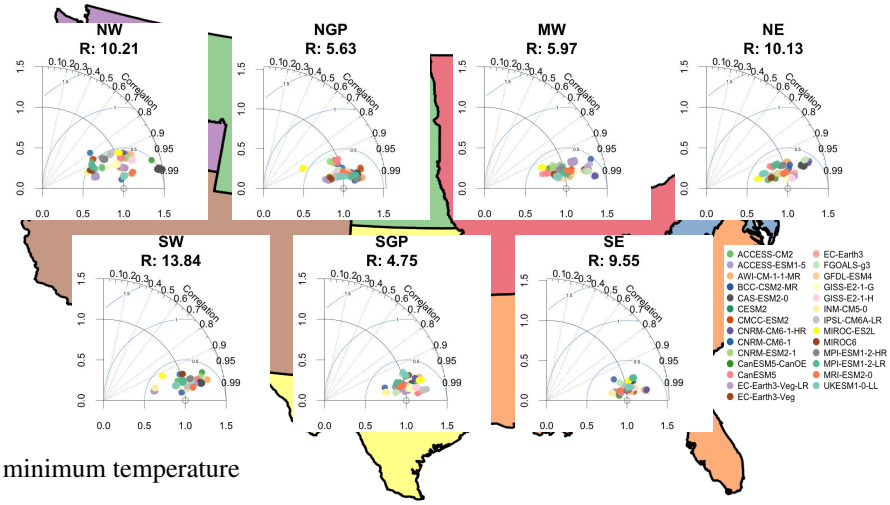
Fig. A.1: 30-year CONUS-wide averages of daily precipitation (left), maximum temperature (middle), and minimum temperature (right) from the 2070-2100 period of SSP370. The reference dataset here is the unweighted multi-model ensemble average, so that ensembles are given equivalent weight rather than models. 30-year averages of daily precipitation (left), maximum temperature(middle), and minimum temperature (right) from the 2070-2100 period of SSP370. Each model is shown with a different color. Panel (a) shows annual estimates and panel (b) shows seasonal estimates.



(a) Annual precipitation

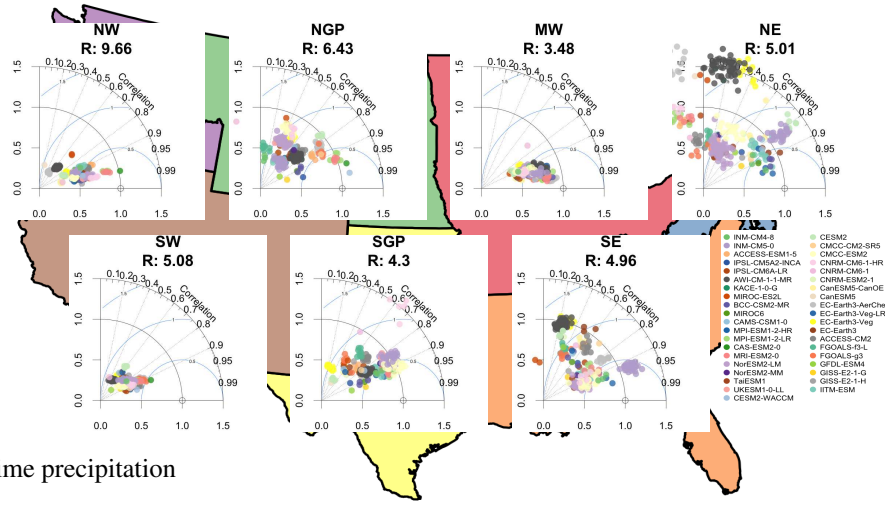


(b) Annual maximum temperature

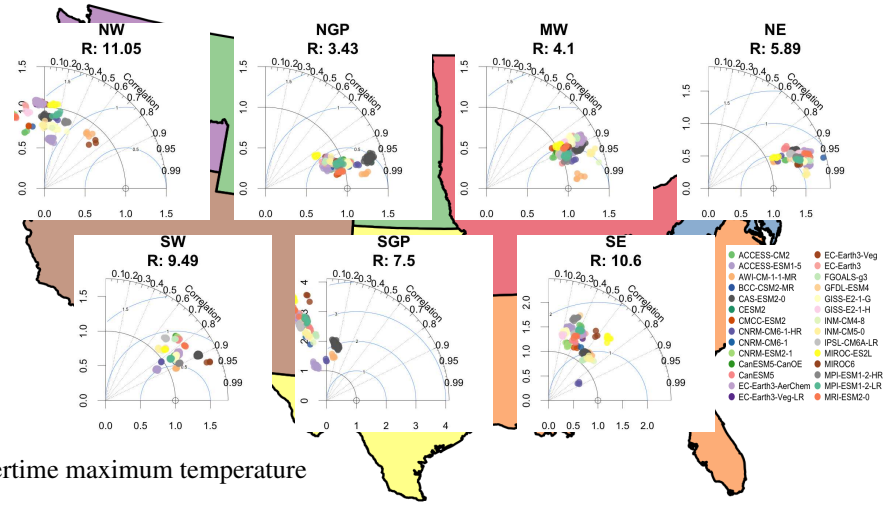


(c) Annual minimum temperature

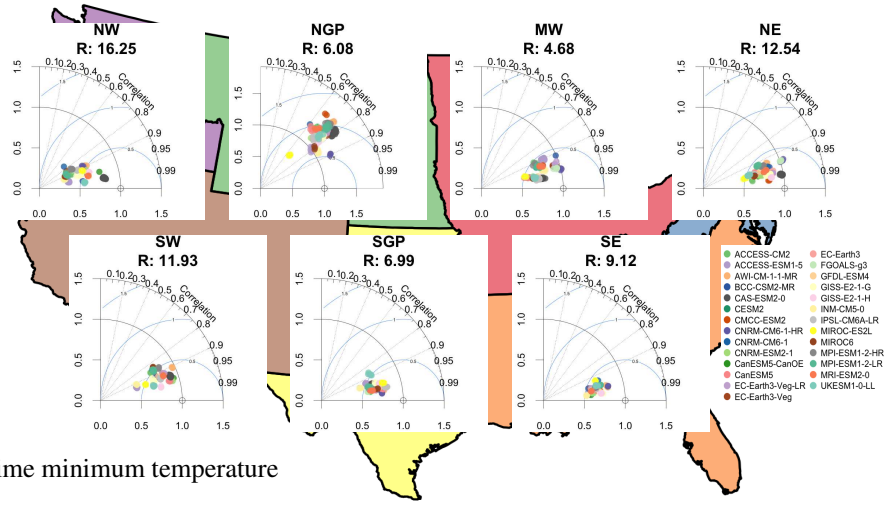
Fig. A.2: Taylor diagrams for regional (NCA4 regions) 30-year annual average of SSP370 over 2070-2010, compared against the unweighted multi-model ensemble average.



(a) Wintertime precipitation



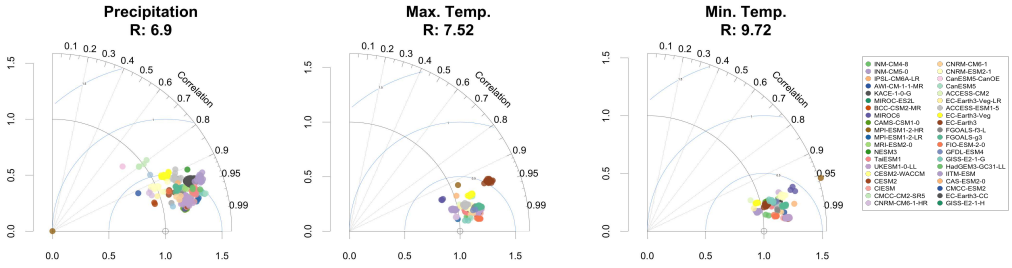
(b) Summertime maximum temperature



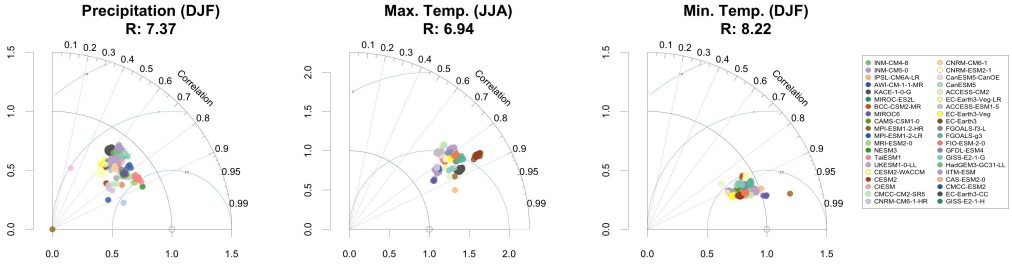
(c) Wintertime minimum temperature

Fig. A.3: Regional Taylor diagrams for the 30-year seasonal average of SSP370 over 2070-2010, compared against the unweighted multi-model ensemble average.

B Additional future projections: SSP245 Taylor diagrams

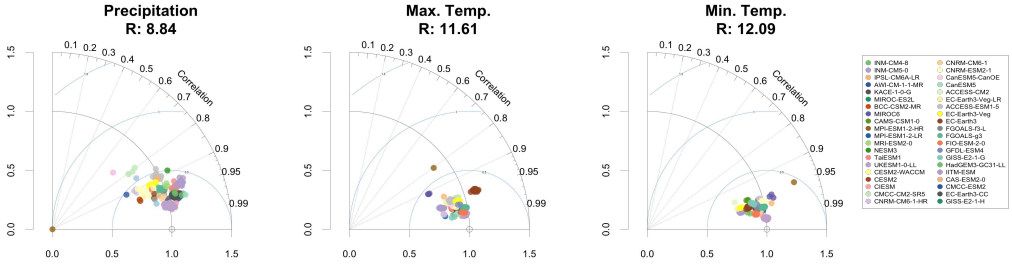


(a) CONUS annual, 2070-2100

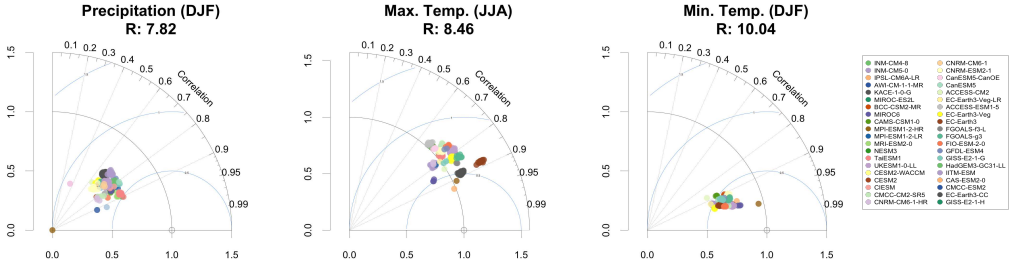


(b) CONUS seasonal, 2070-2100

Fig. B.1: Taylor diagrams for CONUS-wide 30-year averages from the 2070-2100 period of SSP245, compared against the weighted multi-model ensemble average.

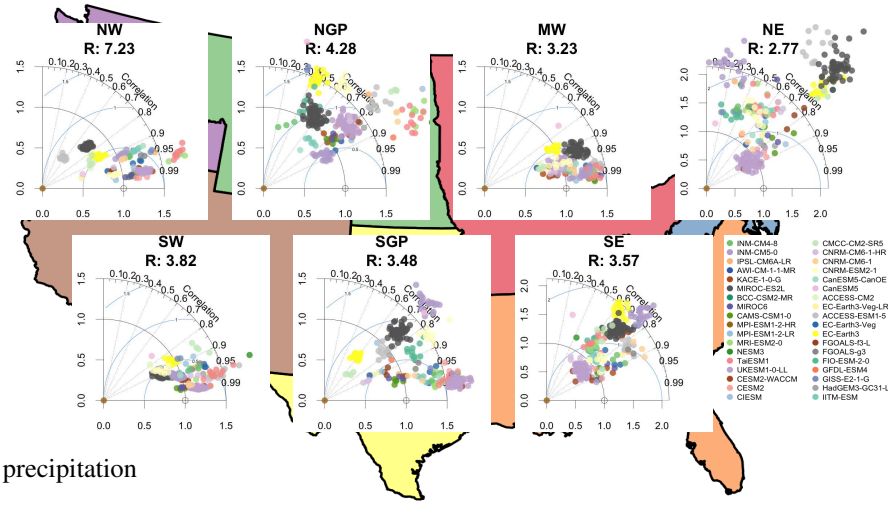


(a) CONUS annual, 1985-2015

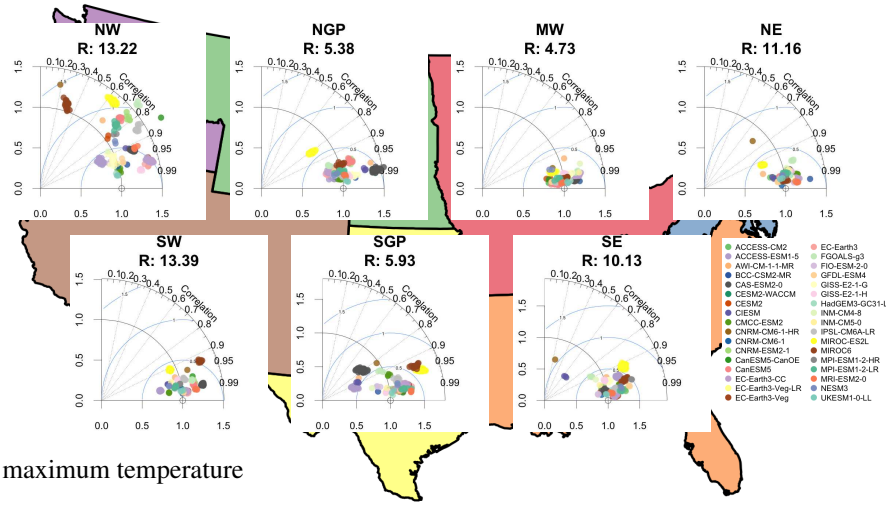


(b) CONUS seasonal, 1985-2015

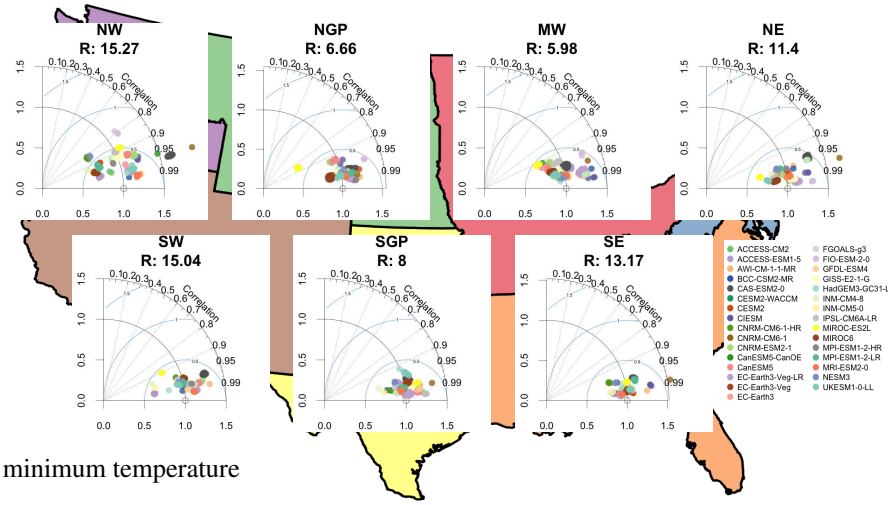
Fig. B.2: Taylor diagrams for CONUS-wide 30-year averages from the 2070-2100 period of SSP245, compared against the unweighted multi-model ensemble average.



(a) Annual precipitation

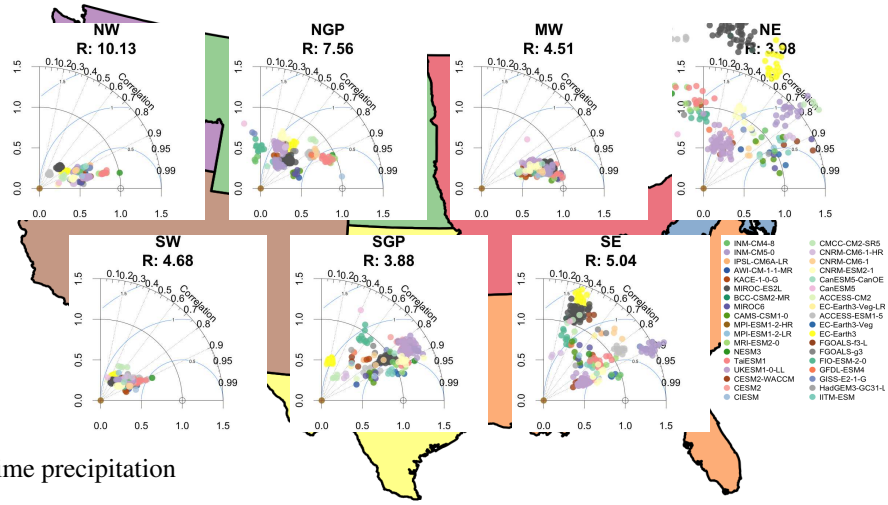


(b) Annual maximum temperature

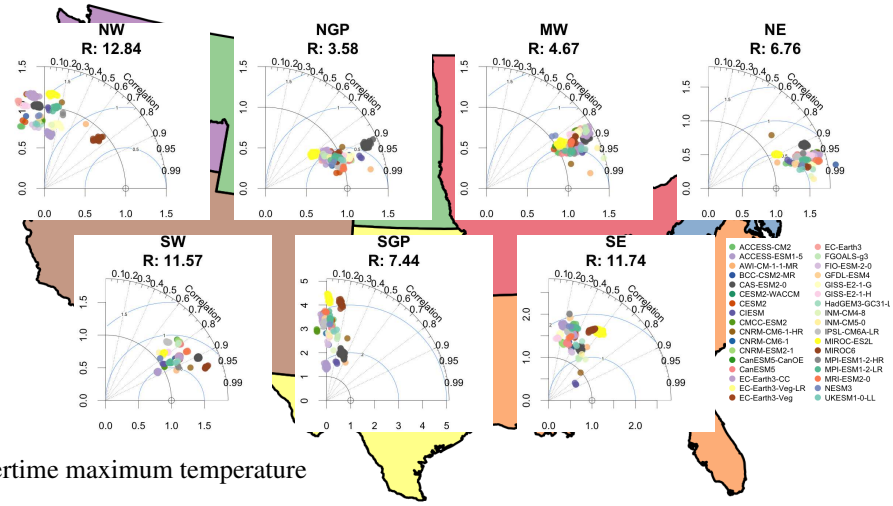


(c) Annual minimum temperature

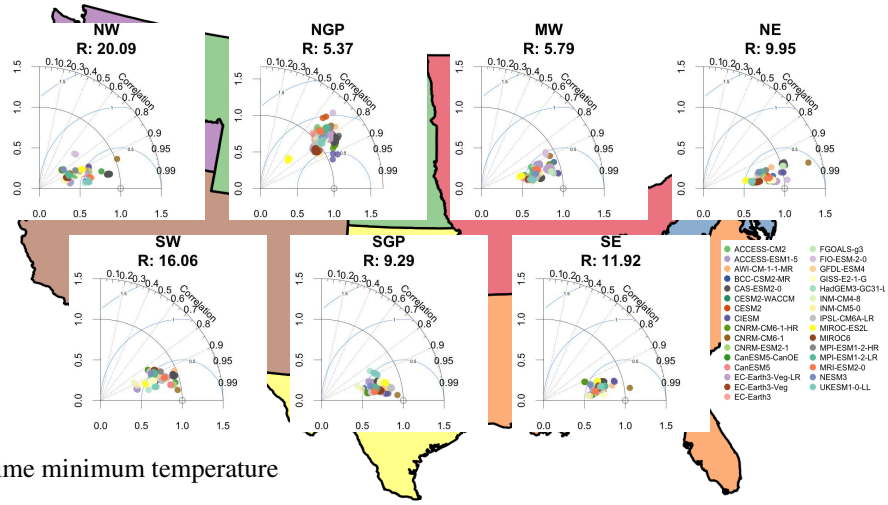
Fig. B.3: Regional Taylor diagrams for the 30-year annual average of SSP245 over 2070-2010, compared against the weighted multi-model ensemble average.



(a) Wintertime precipitation

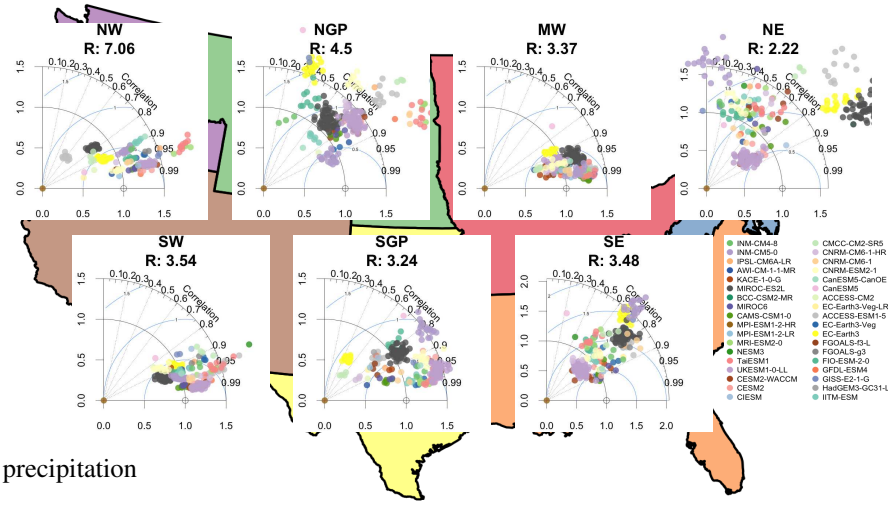


(b) Summertime maximum temperature

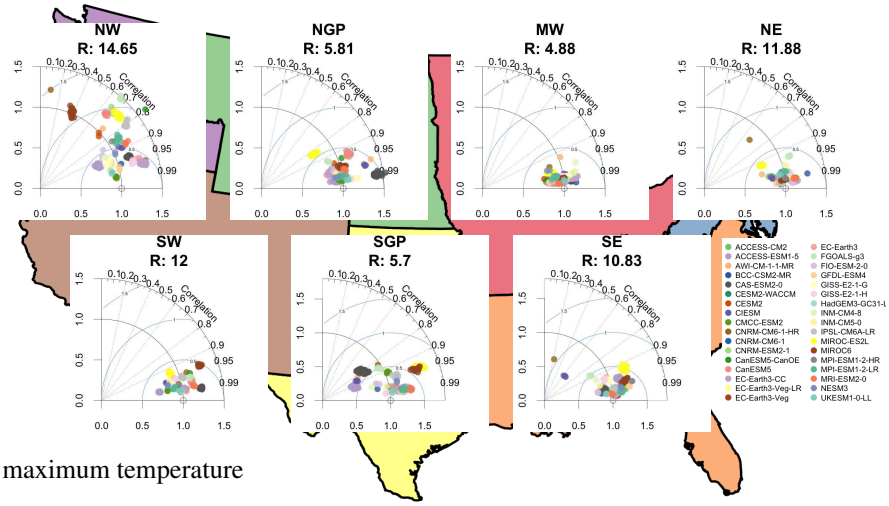


(c) Wintertime minimum temperature

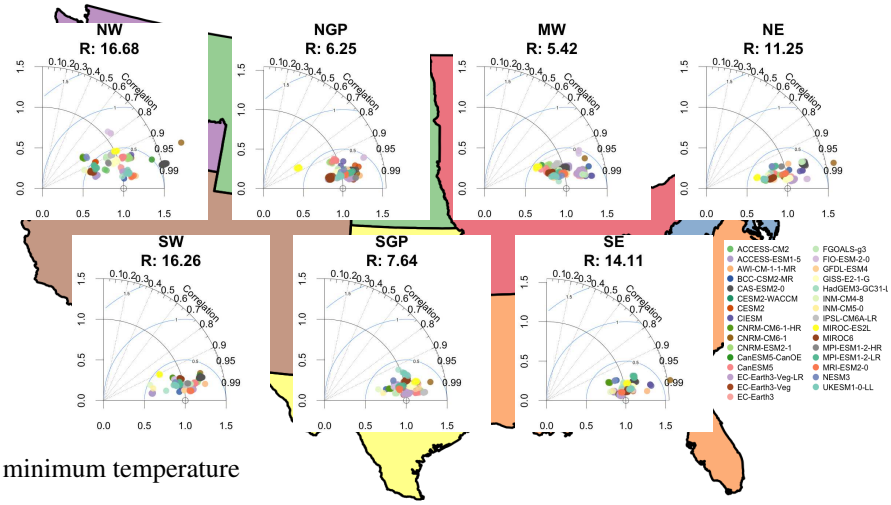
Fig. B.4: Regional Taylor diagrams for the 30-year seasonal average of SSP245 over 2070-2010, compared against the weighted multi-model ensemble average.



(a) Annual precipitation

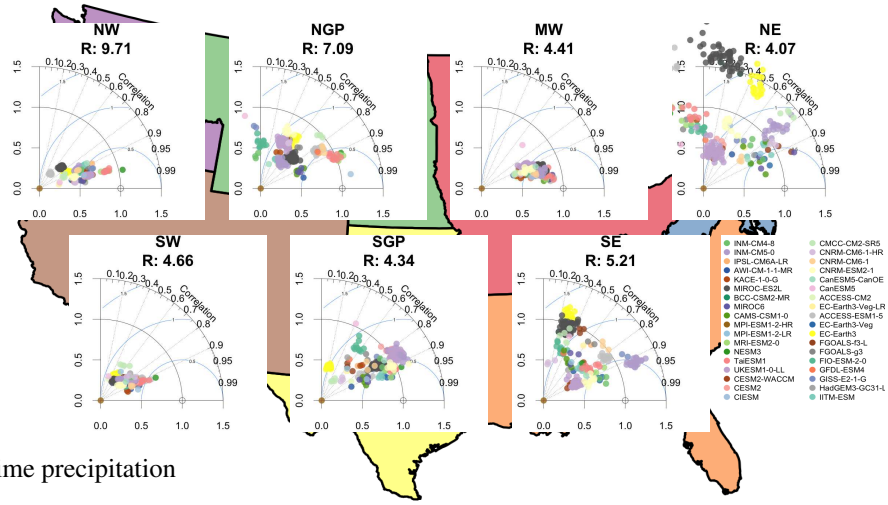


(b) Annual maximum temperature

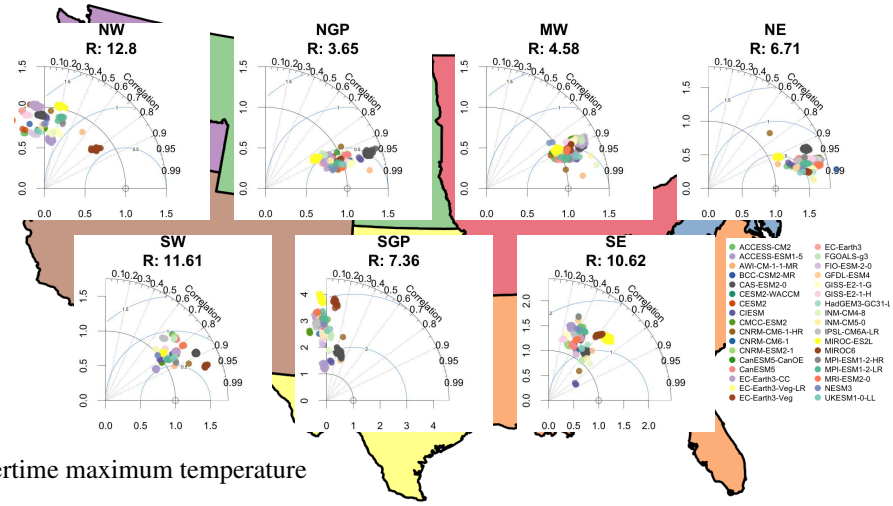


(c) Annual minimum temperature

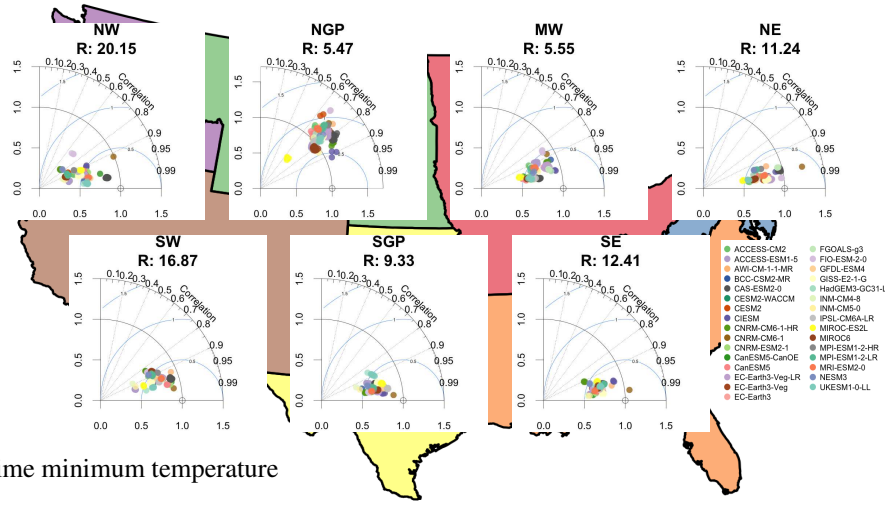
Fig. B.5: Regional Taylor diagrams for the 30-year annual average of SSP245 over 2070-2010, compared against the unweighted multi-model ensemble average.



(a) Wintertime precipitation



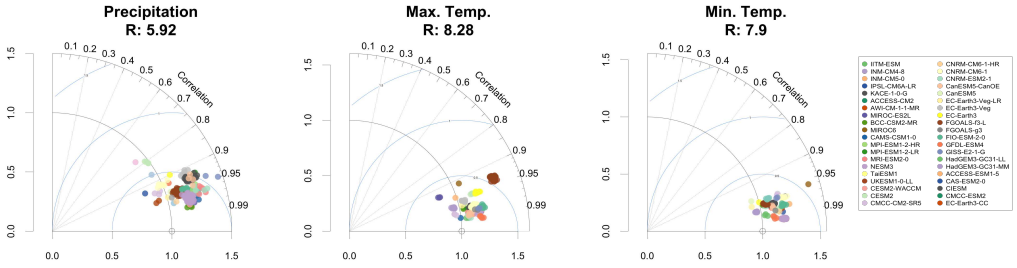
(b) Summertime maximum temperature



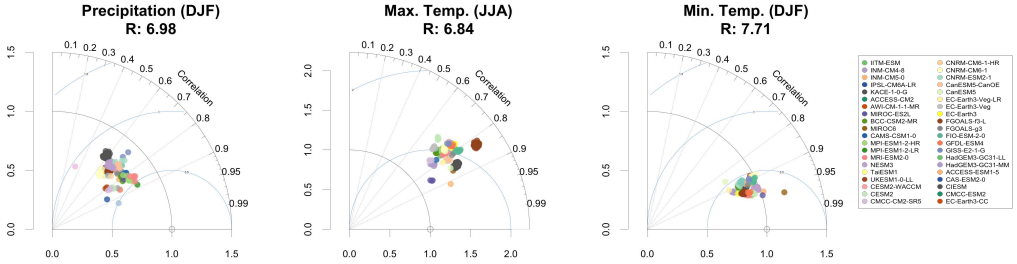
(c) Wintertime minimum temperature

Fig. B.6: Regional Taylor diagrams for the 30-year seasonal average of SSP245 over 2070-2010, compared against the unweighted multi-model ensemble average.

C Additional future projections: SSP585 Taylor diagrams

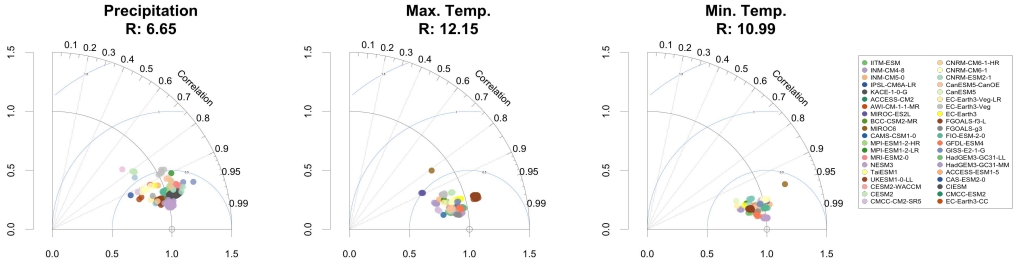


(a) CONUS annual, 2070-2100

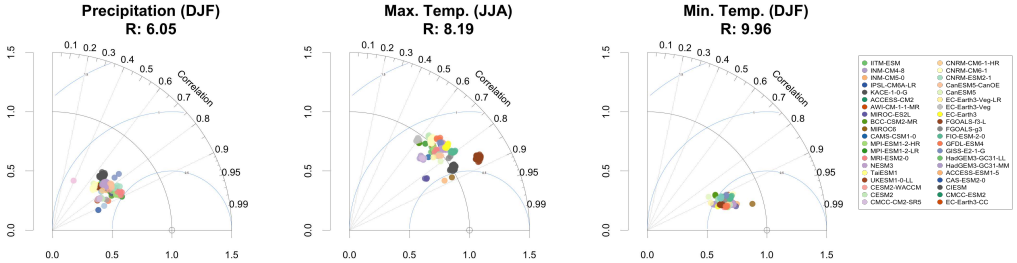


(b) CONUS seasonal, 2070-2100

Fig. C.1: Taylor diagrams for CONUS-wide 30-year averages from the 2070-2100 period of SSP585, compared against the weighted multi-model ensemble average.

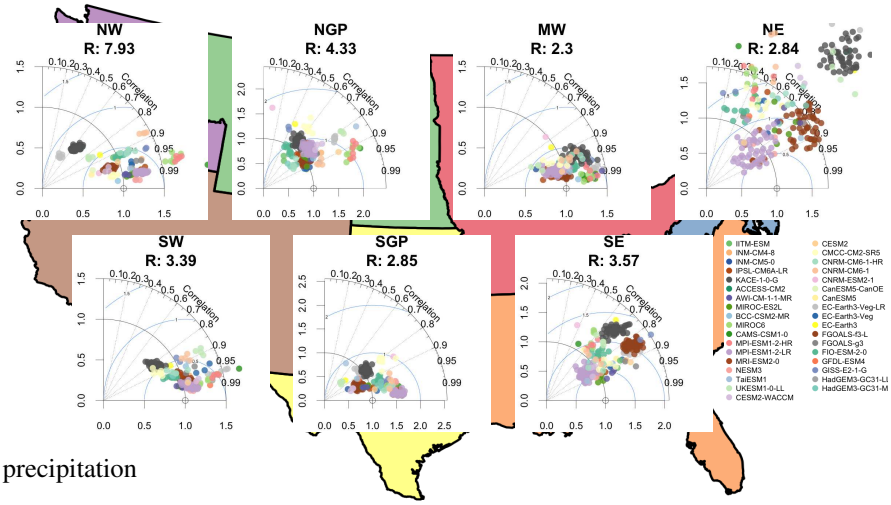


(a) CONUS annual, 2070-2100

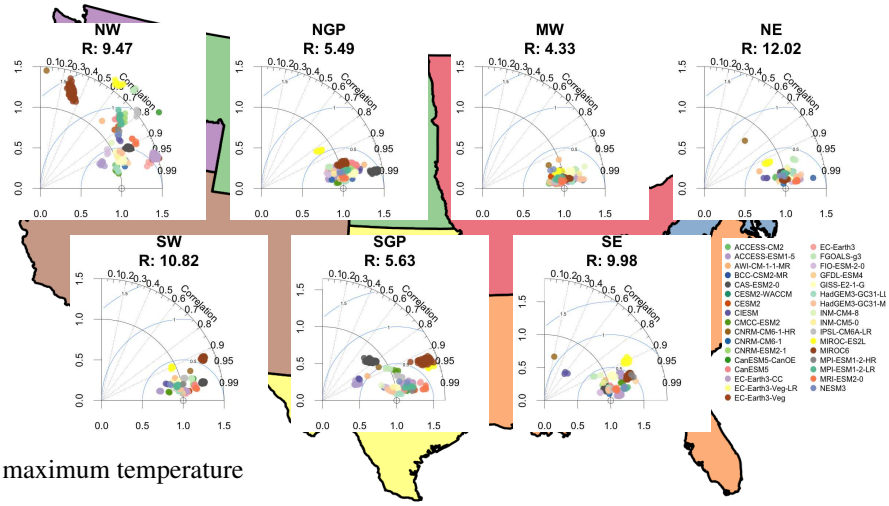


(b) CONUS seasonal, 2070-2100

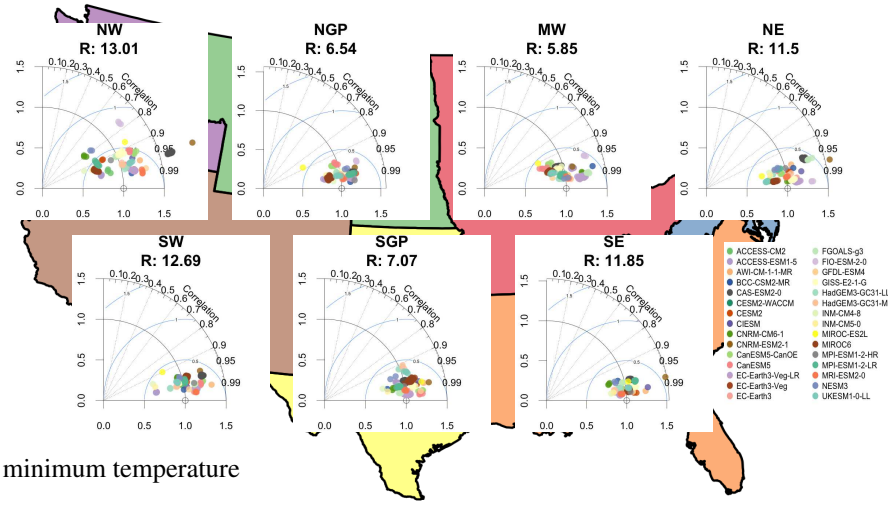
Fig. C.2: Taylor diagrams for CONUS-wide 30-year averages from the 2070-2100 period of SSP585, compared against the unweighted multi-model ensemble average.



(a) Annual precipitation

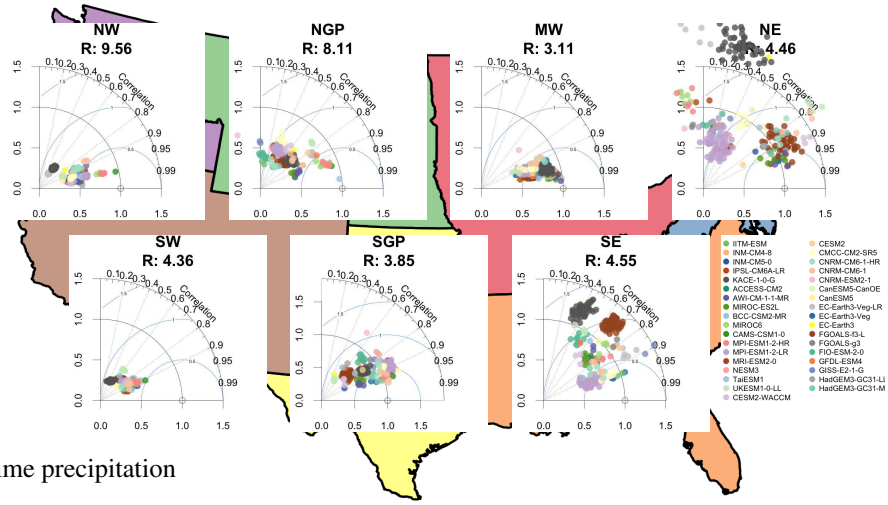


(b) Annual maximum temperature

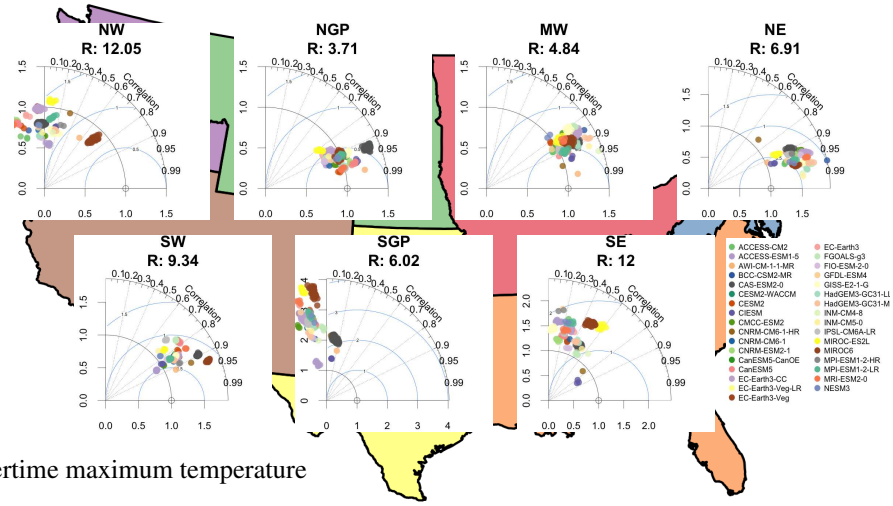


(c) Annual minimum temperature

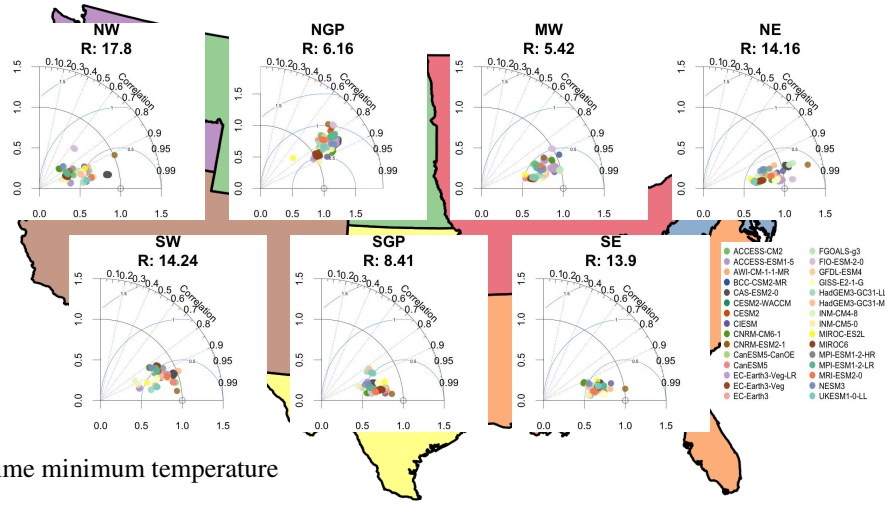
Fig. C.3: Regional Taylor diagrams for the 30-year annual average of SSP585 over 2070-2010, compared against the weighted multi-model ensemble average.



(a) Wintertime precipitation

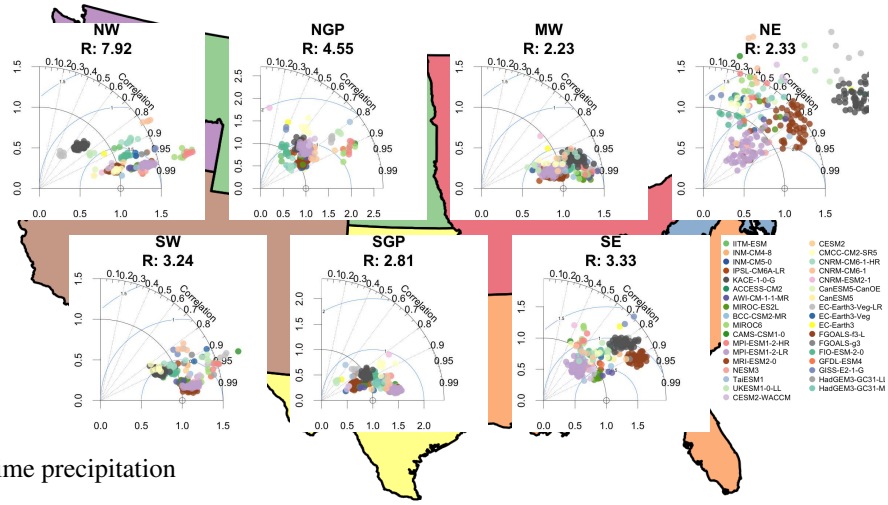


(b) Summertime maximum temperature

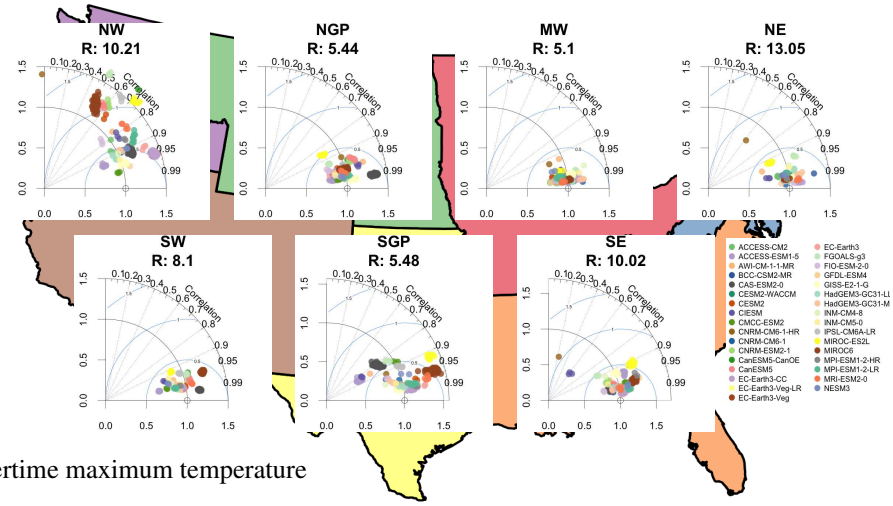


(c) Wintertime minimum temperature

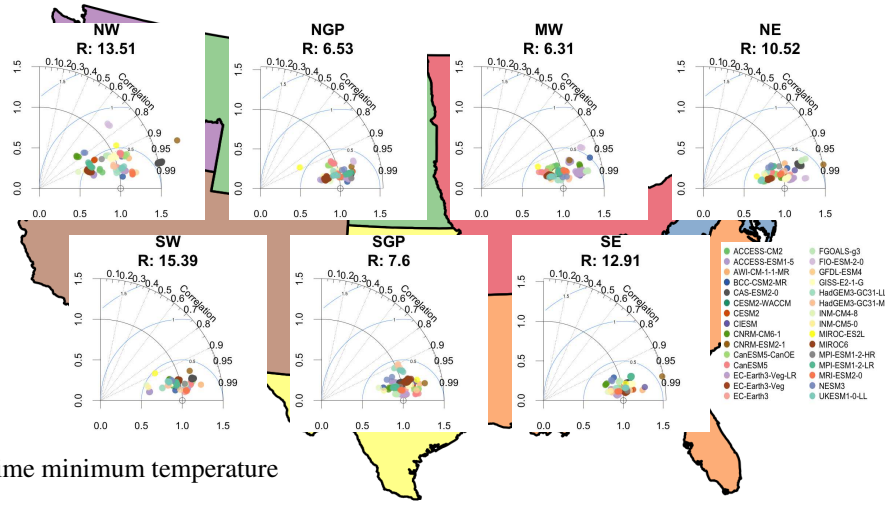
Fig. C.4: Regional Taylor diagrams for the 30-year seasonal average of SSP585 over 2070-2010, compared against the weighted multi-model ensemble average.



(a) Wintertime precipitation

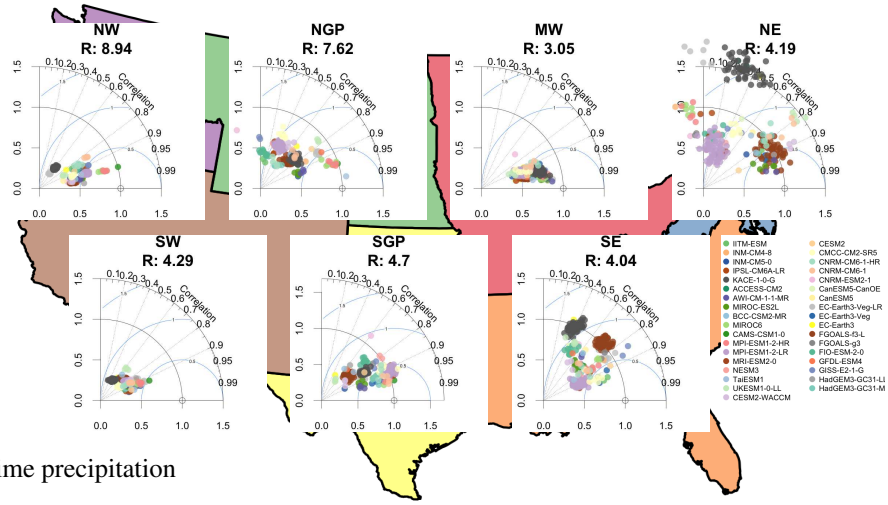


(b) Summertime maximum temperature

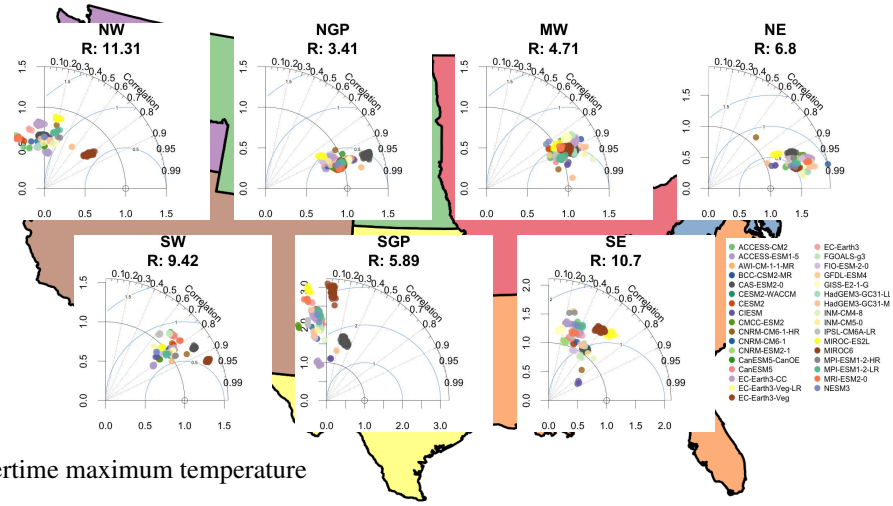


(c) Wintertime minimum temperature

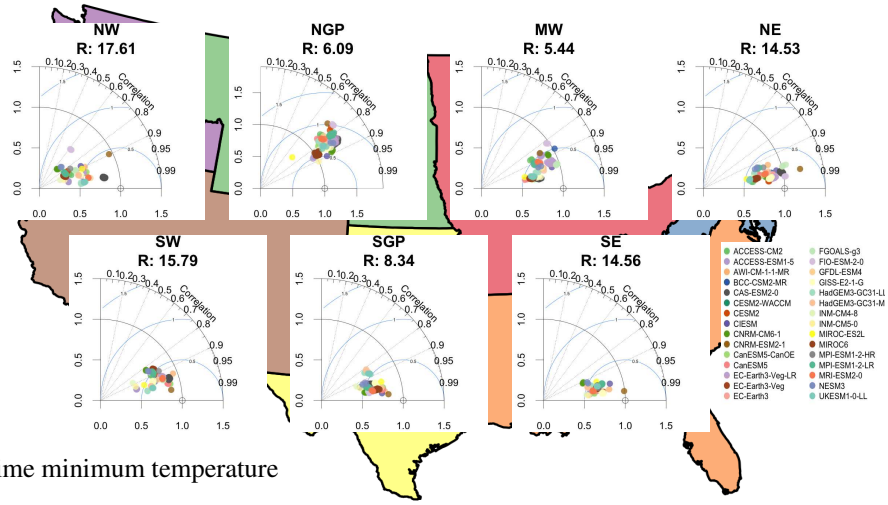
Fig. C.5: Regional Taylor diagrams for the 30-year annual average of SSP585 over 2070-2010, compared against the unweighted multi-model ensemble average.



(a) Wintertime precipitation



(b) Summertime maximum temperature



(c) Wintertime minimum temperature

Fig. C.6: Regional Taylor diagrams for the 30-year seasonal average of SSP585 over 2070-2010, compared against the unweighted multi-model ensemble average.

D 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.,

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

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

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

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

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

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

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