Evaluation of CMIP6 GCMs over the CONUS for downscaling studies

Moetasim Ashfaq[1], Deeksha Rastogi[2], Muhammad Adnan Abid[3], and Shih-Chieh Kao[2]

[1]Oak Ridge National Laboratory (DOE)
[2]Oak Ridge National Laboratory
[3]The Abdus Salam International Centre for Theoretical Physics

November 22, 2022

Abstract

Despite the necessity of Global Climate Models (GCMs) sub-selection in the dynamical downscaling experiments, an objective approach for their selection is currently lacking. Building on the previously established concepts in GCMs evaluation frameworks, we relatively rank 37 GCMs from the 6th phase of Coupled Models Intercomparison Project (CMIP6) over four regions representing the contiguous United States (CONUS). The ranking is based on their performance across 60 evaluation metrics in the historical period (1981–2014). To ensure that the outcome is not method-dependent, we employ two distinct approaches to remove the redundancy in the evaluation criteria. The first approach is a simple weighted averaging technique. Each GCM is ranked based on its weighted average performance across evaluation measures, after each metric is weighted between zero and one depending on its uniqueness. The second approach applies empirical orthogonal function analysis in which each GCM is ranked based on its sum of distances from the reference in the principal component space. The two methodologies work in contrasting ways to remove the metrics redundancy but eventually develop similar GCMs rankings. While the models from the same institute tend to display comparable skills, the high-resolution model versions distinctively perform better than their lower-resolution counterparts. The results from this study should be helpful in the selection of models for dynamical downscaling efforts, such as the COordinated Regional Downscaling Experiment (CORDEX), and in understanding the strengths and deficiencies of CMIP6 GCMs in the representation of various background climate characteristics across CONUS.

Evaluation of CMIP6 GCMs over the CONUS for downscaling studies

Moetasim Ashfaq*[1], Deeksha Rastogi[1], Muhammad Adnan Abid[2], Shih-Chieh Kao[3]
[1] Computational Sciences and Engineering Division (CSED), Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Earth System Physics, Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
[3] Environmental Science Division (ESD), Oak Ridge National Laboratory, Oak Ridge, TN, USA

Key Points

- A sub-selection of GCMs from the large CMIP ensemble is often necessary before downscaling due to several unavoidable constraints.
- We evaluate models for their objective sub-selection using two distinct approaches that remove the redundancy in 60 evaluation metrics.
- Two methods develop a similar ranking, placing the high-resolution models distinctively higher than their lower-resolution counterparts.

*Corresponding Author
Moetasim Ashfaq
mashfaq@ornl.gov

Abstract

Despite the necessity of Global Climate Models (GCMs) sub-selection in the dynamical downscaling experiments, an objective approach for their selection is currently lacking. Building on the previously established concepts in GCMs evaluation frameworks, we relatively rank 37 GCMs from the 6th phase of Coupled Models Intercomparison Project (CMIP6) over four regions representing the contiguous United States (CONUS). The ranking is based on their performance across 60 evaluation metrics in the historical period (1981–2014). To ensure that the outcome is not method-dependent, we employ two distinct approaches to remove the redundancy in the evaluation criteria. The first approach is a simple weighted averaging technique. Each GCM is ranked based on its weighted average performance across evaluation measures, after each metric is weighted between zero and one depending on its uniqueness. The second approach applies empirical orthogonal function analysis in which each GCM is ranked based on its sum of distances from the reference in the principal component space. The two methodologies work in contrasting ways to remove the metrics redundancy but eventually develop similar GCMs rankings. While the models from the same institute tend to display comparable skills, the high-resolution model versions distinctively perform better than their lower-resolution counterparts. The results from this study should be helpful in the selection of models for dynamical downscaling efforts, such as the COordinated Regional Downscaling Experiment (CORDEX), and in understanding the strengths and deficiencies of CMIP6 GCMs in the representation of various background climate characteristics across CONUS.

Plain Language Summary

Global Climate Models (GCMs) provide climate change projections at spatial scales that are much coarser than the scales at which regional and local planning decisions are made. Therefore, GCMs projections are spatially refined through various downscaling procedures. Often, a sub-selection of GCMs is needed before their downscaling due to issues related to their performance, data availability, and resources required for spatial refinement. Here we evaluate GCMs from the 6th phase of Coupled Models Intercomparison Project (CMIP6) over four regions representing the contiguous United States (CONUS) to guide the GCMs sub-selection decision-making objectively. We use two distinct approaches to relative rank the models using their performance across 60 evaluation metrics in the historical period. The two methodologies work in contrasting ways to remove the metrics redundancy but eventually develop similar GCMs rankings. These results should be helpful in the selection of models for dynamical downscaling efforts and understanding the strengths and deficiencies of GCMs in the representation of various background climate characteristics across CONUS.

1. Introduction

Global Climate Models (GCMs) are physics-based tools to study Earth system responses to natural climate variability and anthropogenically driven increases in greenhouse gas emissions and radiative forcing. Using a common set of future radiative pathways, the Coupled Model Intercomparison Projects (CMIP; Eyring et al., 2016) provide an extensive suite of GCM simulations through an international collaborative effort. Since its inception in 1995, not only have the number of GCMs participating in CMIP efforts increased, but they have also improved in terms of their physical complexity and spatial resolution. Every new iteration of CMIP is based on the premise that the more recent generations of GCMs will exhibit improvements over the previous ones as models progressively improve in terms of their computational efficiency, resolution, and representation of physical processes. Despite the significant advancements in GCMs, several challenges related to their horizontal grid spacing and inaccuracies in representing fine-scale land-atmosphere interactions remain unresolved, limiting the direct application of GCM-based climate projections in regional to local scale climate change impact assessments. The latest Phase 6 (CMIP6) includes over 50 GCMs. While the horizontal grid spacing for some of them is as fine as half a degree, the resolution of most CMIP6 GCMs is still insufficient (>1° horizontal grid spacing) to reliably assess the needs for mitigation or adaptation at policy-relevant regional and local scales. Therefore, it warrants the need for spatial refinement of projected climate change information through downscaling.

A sub-selection of GCMs from the large CMIP6 ensemble may be necessary before downscaling for several reasons, including the choice of downscaling framework, computational cost, and the need for better representation of critical climate processes relevant to the region of interest (McSweeney et al., 2015). This is the case in dynamical downscaling (also known as regional climate modeling), where not every GCM can/should be downscaled for several reasons. First and foremost, although GCM experiments are conducted at sub-hourly time scales, given the massive data flow, only a subset of variables at aggregated temporal scales are recorded (usually driven by the specific CMIP requirements). Therefore, not every GCM in the CMIP6 has archived sub-daily three-dimensional lateral boundary forcings fields needed for regional climate modeling. Second, the poor GCM skill over the domains of interest may propagate and result in the unreasonable fine-scale spatiotemporal distribution of downscaled prognostic variables, such as precipitation and temperature, in regional dynamical downscaling experiments (Giorgi, 2019). Therefore, dynamical downscaling of GCMs is limited to those models that exhibit *reasonable* skill. Third, several models participating in the CMIP6 share standard modeling components (e.g., same land, ocean, ice modules, or parametrization), meaning that these models may have similar systematic biases and do not necessarily represent independent realizations of future climate (Knutti et al., 2010 and 2013). Therefore, a downscaled ensemble of regional climate model experiments should consist of GCMs representing unique model developing institutes. However, such a strategy may not fully resolve this issue as modeling components or parametrization sharing is standard across the GCMs from different institutes (Boé, 2018; Knutti et al., 2013). Lastly, the number of downscaled GCMs also depends on the available capacity of the computational and data storage solutions.

There has been substantial progress in the mathematical art of identifying relatively better (or worse) performing models (e.g., Ahmadalipour et al., 2017; Ahmed et al., 2019; Chhin et al., 2018; Knutti et al. 2017; Lorenz et al. 2018; Overland et al. 2011; Parding et al. 2020; Pierce et al. 2009). However, there are no set criteria for the choice of evaluation metrics. Due to this reason, there is quite a disparity among studies on GCMs evaluation, as some are based on only a few climatological mean comparisons between simulations and observations (e.g., McSweeney et al., 2015; Mote and Salathé, 2010). In contrast, others use dozens of metrics covering various aspects of background climate (e.g., Chhin et al., 2018; Rupp et al., 2013). A lack of in-depth evaluation of GCMs in studies with a limited number of evaluation measures runs the risk of errors in their relative ranking in the CMIP ensemble. A model can yield reasonable climatological distribution of desired fields over a region while poorly simulating key Earth system processes (e.g., Beobide-Arsuaga et al. 2021; McBride et al. 2021; Mckenna et al. 2020). Alternatively, high covariance among the extensive suite of evaluation metrics used to investigate the relative skillfulness of models can also influence the GCMs ranking process. Despite these challenges, a large body of research towards developing GCMs evaluation frameworks provides valuable insight that requires seamless integration into the downscaling approaches. Unfortunately, to a large extent, the outcome of these efforts has not been systematically used in the choice of GCMs for downscaling studies, especially for international collaborative efforts such as the Coordinated Regional Downscaling Experiment (CORDEX; Giorgi et al. 2009). Given that the next phase of CORDEX experiments is still in planning, one of the primary aims of this study is to establish an objective GCMs selection approach as an essential part of the dynamical downscaling process.

As noted, the development of robust strategies to rank GCMs concerning their skillfulness has remained an active area of research during the last decade (Knutti et al., 2010; Rupp et al., 2013 and others). Instead of reinventing the wheel, our goal in this study is to use established concepts in this area to streamline the process of GCMs selection from the CMIP6 ensemble for the downscaling efforts. While this study focuses only on the contiguous United States (CONUS), the process can be repeated over any geographical area after modifications in the evaluation metrics as needed. To ensure that the outcome is not method-dependent, our GCMs evaluation employs two distinct approaches. The first approach is a simple weighted averaging technique. Each GCM is ranked based on its average performance across selected evaluation metrics after each metric is given a weight between zero and one depending on its uniqueness. The second approach is through the application of empirical orthogonal functions (EOFs) in which each GCM is ranked based on its distance from the reference (observations) in the principal component (PC) space (Chhin et al., 2018; Rupp et al., 2013; Sanderson et al., 2015). The PCs are further used to investigate the distinctiveness of the analyzed GCMs in the CMIP6 ensemble.

2. Methods

2.1 Data

The simulations data for 37 CMIP6 GCMs are obtained from Earth System Grid Federation (ESGF) archives (https://esgf-node.llnl.gov/search/cmip6) for the historical period (1980–2014) (Table 1), which include daily and monthly precipitation, mean, maximum, and minimum temperatures; monthly sea surface temperature; air pressure at sea level; and 500 mb geopotential height. Due to the unavailability of a complete set of variables required for evaluation at the time of analyses, some well-known models, such as the National Center for Atmospheric Research (NCAR) Community Earth System Model (CESM), are not included in this study. To support this evaluation, the gridded precipitation and temperature observations are obtained from three sources: 1) Daymet – maintained by the Distributed Active Archive Center at Oak Ridge National Laboratory (Thornton et al., 2021), 2) Livneh – initially produced by the University of Colorado at Boulder (UCB; Pierce et al., 2021), updated version available from the University of California Los Angeles, and 3) Parameter elevation Regression on Independent Slopes Model (PRISM) – the

United States Agriculture Department (USDA) official climatological data (Daly et al., 2018). Additionally, European Centre for Medium-Range Weather Forecasts Reanalysis 5 (ERA5; Hersbach et al. 2020) is used to reference sea surface temperature, air pressure at sea level, and 500 mb geopotential height. For comparisons, all the GCMs and reference datasets are remapped to a standard 1° latitude-longitude grid.

GCMs	Variant Label	Institute	Lon x Lat
ACCESS-CM2	r1i1p1f1	Commonwealth Scientific and Industrial Research Organization, Australia	192x144
ACCESS-ESM1-5	r1i1p1f1	Commonwealth Scientific and Industrial Research Organization, Australia	192x145
AWI-CM-1-1-MR	r1i1p1f1	Alfred Wegener Institute, Germany	384 ×192
AWI-ESM-1-1-LR	r1i1p1f1	Alfred Wegener Institute, Germany	192x96
BCC-CSM2-MR	r1i1p1f1	Beijing Climate Center, China Meteorological Administration, China	320x160
BCC-ESM1	r1i1p1f1	Beijing Climate Center, China Meteorological Administration, China	128x64
CanESM5	r1i1p1f1	Canadian Centre for Climate Modelling and Analysis, Canada	128×64
CMCC-CM2-SR5	r1i1p1f1	Euro-Mediterranean Centre on Climate Change, Italy	288×192
CNRM-CM6-1	r1i1p1f2	Centre National de Recherches Météorologiques, France	256x128
CNRM-CM6-1-HR	r1i1p1f2	Centre National de Recherches Météorologiques, France	720x360
CNRM-ESM2-1	r1i1p1f2	Centre National de Recherches Météorologiques, France	256x128
EC-Earth3	r1i1p1f1	European EC-Earth consortium	512x256
EC-Earth3-Veg	r1i1p1f1	European EC-Earth consortium	512x256
EC-Earth3-Veg-LR	r1i1p1f1	European EC-Earth consortium	320x160
FGOALS-f3-L	r1i1p1f1	Chinese Academy of Sciences, China	288x180
FGOALS-g3	r1i1p1f1	Chinese Academy of Sciences, China	180x80
GFDL-CM4	r1i1p1f1	Geophysical Fluid Dynamics Laboratory, USA	144x90
GFDL-ESM4	r1i1p1f1	Geophysical Fluid Dynamics Laboratory, USA	288x180
GISS-E2-1-G	r1i1p1f1	National Aeronautics and Space Administration (NASA), United States	144x90
HadGEM3-GC31-LL	r1i1p1f3	Met Office, United Kingdom	192x144
HadGEM3-GC31-MM	r1i1p1f3	Met Office, United Kingdom	432x324
INM-CM4-8	r1i1p1f1	Institute for Numerical Mathematics, Russia	180x120
INM-CM5-0	r1i1p1f1	Institute for Numerical Mathematics, Russia	180x120
IPSL-CM6A-LR	r1i1p1f1	Institut Pierre Simon Laplace, France	144x143
KACE-1-0-G	r1i1p1f1	National Institute of Meteorological Sciences, Republic of Korea	192 ×144
MIROC6	r1i1p1f1	Japan Agency for Marine-Earth Science and Technology, Japan	256x128
MIROC-ES2L	r1i1p1f2	Japan Agency for Marine-Earth Science and Technology, Japan	128x64

MPI-ESM-1-2-HAM	r1i1p1f1	Max Planck Institute for Meteorology, Germany	192x96
MPI-ESM1-2-HR	r1i1p1f1	Max Planck Institute for Meteorology, Germany	384x192
MPI-ESM1-2-LR	r1i1p1f1	Max Planck Institute for Meteorology, Germany	192x96
MRI-ESM2-0	r1i1p1f1	Meteorological Research Institute, Tsukuba, J+C34apan	320x160
NESM3	r1i1p1f1	Nanjing University of Information Science and Technology, China	192x96
NorCPM1	r1i1p1f1	Norwegian Climate Centre, Norway	144x96
NorESM2-LM	r1i1p1f1	Norwegian Climate Centre, Norway	144x96
NorESM2-MM	r1i1p1f1	Norwegian Climate Centre, Norway	288x192
SAM0-UNICON	r1i1p1f1	Seoul National University, South Korea	288x192
UKESM1-0-LL	r1i1p1f2	Met Office, United Kingdom	192 ×144

Table 1. List of the CMIP6 GCMs used in the evaluation. The variant label provides information about realization (*r*), initialization method (*i*), physics (*p*), and forcing (*f*).

2.2 Evaluation Metrics

For model evaluation, the entire CONUS is divided into four parts (North, East, West, and South) based on grouped 2-digit Hydrological Unit Codes (HUC2) regions (Figure 1), utilized by Naz et al. (2016). At the annual, seasonal, monthly, daily, and diurnal time scales, sixty metrics evaluate the CMIP6 GCMs. Table 2 describes the summary of these metrics. All metrics are calculated separately for each of the four regions, subsequently averaged to calculate disagreements at the CONUS scale for each model. The sixty evaluation criteria include both standalone and derived metrics. All metrics are calculated separately for the three observations (Daymet, Livneh, and PRISM), subsequently averaged to create a reference dataset. A model disagreement is calculated as a percent departure from the reference data for each standalone metric. Several derived metrics are based on the calculation of Taylor Stats (TS; Taylor, 2001) – a combination of root mean square error, bias, and pattern correlation (Table 2). For this purpose, model disagreements for each of the three statistical measures are calculated as percent departures from the reference data. Their averages represent the TS for that metric. The TS is calculated separately for the diurnal cycle metric for four seasons and then averaged to get the final measure. Similarly, TS for the metric representing precipitation from moderate to extreme events is also based on the average of individual TS for precipitation from events exceeding 75th, 90th, 95th, and 99th percentiles of precipitation. The combination of all seasons in a single metric for the diurnal cycle and four kinds of events ranging from moderate to extreme precipitation magnitudes in one metric is due to their relatively very high correlations across the CMIP6 GCMs ensemble. The dispersion metric averages the TS of 20 indices (Table 2), calculated after transforming the 3-dimensional (time, latitude, longitude) data into 1-dimension.

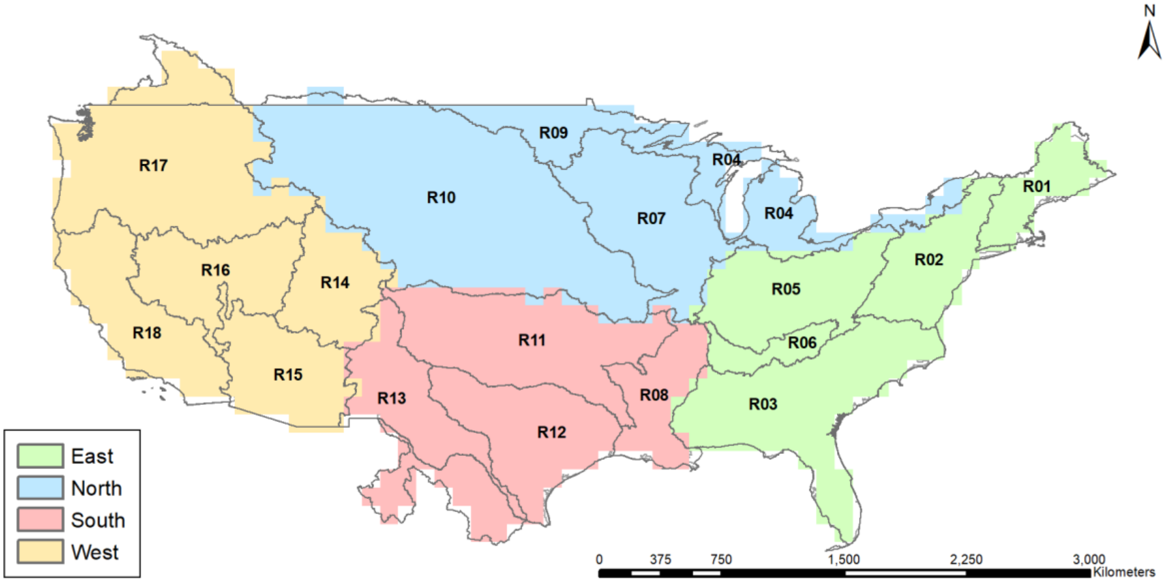

Figure 1. CONUS division in four HUC2 based regions for GCMs evaluations. The division was initially used by Naz et al. (2016). R01 to R18 represent 18 US HUC2s.

The GCMs evaluation also includes representation of three modes of natural climate variability, namely North Atlantic Oscillation (NAO), El Niño-Southern Oscillation (ENSO), and Pacific Decadal Oscillation (PDO), and their impacts on the distribution of winter (December–January–February, DJF) and summer (June–July–August, JJA) precipitation and temperature. The PDO index represents the first EOF of sea surface temperature over Northern Pacific (20°N–70°N, 110°E–260°E; Mantua et al. 1997; Newman et al. 2016). The ENSO index represents the sea surface temperature anomalies over the Nino3.4 region (5°S–5°N, 170°W–120°W; Trenberth, 1997). In both cases, the temporally varying global mean is removed from the sea surface temperatures to avoid any impact of global warming. The NAO index represents the first EOF of detrended sea level pressure over the Northern Atlantic (20°N–80°N, 90°W–40°E; Hurrell, 1995; Hurrell & Deser, 2009). The pattern correlation is used to measure GCMs' skills in representing these modes of variability. A more detailed background of these indices can be found in the NCAR climate data guide (https://climatedataguide.ucar.edu/).

GCMs Evaluation Metrics			
1. Amplitude[a] Mean P[1]	**2.** Amplitude Mean T[2]	**3.** Amplitude Mean Tmax[3]	**4.** Amplitude Mean Tmin[4]
5. Amplitude Standard Deviation P	**6.** Amplitude Standard Deviation T	**7.** Amplitude Standard Deviation Tmax	**8.** Amplitude Standard Deviation Tmin
9. Timing[b] of Peak P	**10.** Timing of Peak T	**11.** Timing of Peak Tmax	**12.** Timing of Peak Tmin
13. Annual Mean Standard Deviation of P	**14.** Annual Mean Standard Deviation of T	**15.** Annual Mean Standard Deviation of Tmax	**16.** Annual Mean Standard Deviation of Tmin
17. DJF[5] (Taylor Stats) P	**18.** DJF (Taylor Stats) T	**19.** DJF (Taylor Stats) Tmax	**20.** DJF (Taylor Stats) Tmin
21. MAM[6] (Taylor Stats) P	**22.** MAM (Taylor Stats) T	**23.** MAM (Taylor Stats) Tmax	**24.** MAM (Taylor Stats) Tmin
25. JJA[7] (Taylor Stats) P	**26.** JJA (Taylor Stats) T	**27.** JJA (Taylor Stats) Tmax	**28.** JJA (Taylor Stats) Tmin
29. SON[8] (Taylor Stats) P	**30.** SON (Taylor Stats) T	**31.** SON (Taylor Stats) Tmax	**32.** SON (Taylor Stats) Tmin
33. (Taylor Stats) Inter-quartile Range[c] P	**34.** (Taylor Stats) Inter-quartile Range Tmax	**35.** (Taylor Stats) Inter-quartile Range Tmin	**36.** (Taylor Stats) Diurnal T
37. (Taylor Stats) P from Moderate to Heavy Events	**38.** (Taylor Stats) Wet Days[d]	**39.** (Taylor Stats) P Intensity	**40.** (Taylor Stats) Summer Days[e]
41. (Taylor Stats) Ice Days[f]	**42.** (Taylor Stats) Tropical Nights[g]	**43.** (Taylor Stats) Frost Days[h]	**44.** Dispersion[i] P
45. Dispersion T	**46.** Dispersion Tmin	**47.** Dispersion Tmax	**48.** ENSO Amplitude
49. PDO Pattern	**50.** NAO Pattern	**51.** NAO Correlation with DJF P	**52.** NAO Correlation with DJF T
53. PDO Correlation with DJF P	**54.** PDO Correlation with DJF T	**55.** ENSO Correlation with DJF P	**56.** ENSO Correlation with DJF T
57. (Taylor Stats) 500mb Geopotential Height DJF	**58.** (Taylor Stats) 500mb Geopotential Height JJA	**59.** (Taylor Stats) Sea Level Pressure DJF	**60.** (Taylor Stats) Sea Level Pressure JJA
Taylor Stats			
Root Mean Square Error	Bias	Pattern Correlation	
Dispersion (based on 1-dimesnional time series of time x latitude x longitude)			
Lower Octile	Lower Sextile	Lower Quartile	Lower Tritile
Median	Upper Tritile	Upper Quartile	Upper Sextile
Upper Octile	Upper Dectile	Maximum	Range
0.1st Percentile	1st Percentile	5th Percentile	95th Percentile
99th Percentile	99.9th Percentile	Skewness	Kurtosis

[1]P = Precipitation, [2]T = Temperature, [3]Tmax = Maximum Temperature, [4]Tmin = Minimum Temperature, [5]DJF = December-January-February, [6]MAM = March-April-May, [7]JJA = June-July-August, [8]SON = September-October-November, [9]ENSO = El Niño-Southern Oscillation), [10]PDO = Pacific Decadal Oscillation, [11]NAO = North Atlantic Oscillation

[a]Amplitude = Difference between maximum and minimum in a monthly annual cycle

[b]Timing = Month Index with the maximum of the annual cycle

[c]Inter-quartile range = Difference between the 75th and 25th percentile of daily values in a year

[d]Wet days = Days with accumulated P ≥ 1.0 mm

[e]Summer days = Days with T ≥ 25 °C (77 °F)

[f]Ice days = Days with Tmax < 0 °C

[g]Tropical nights = Days with Tmin > 20 °C (68 °F)

[h]Frost days = Days with Tmin < 0 °C

[h]Dispersion = Spatiotemporal distribution of monthly data, calculated as an average of the Taylor Stats of 20 indices. The calculation of these indices is based on *stat_dispersion* function in the NCAR Command Language (NCL).

Table 2. Metrics used in GCMs evaluation.

2.3 Relative ranking methodology

Two approaches – a simple averaging technique based on the average performance across evaluation metrics and an EOF-based strategy that accounts for the distance of each simulated metric from the reference in the PC space – are used for model ranking. Although careful selections are made to use distinct criteria for GCMs evaluation, high correlations among the evaluation metrics are still possible given the interdependence of physical processes in the coupled Earth system, which could potentially bias the model ranking process when a simple averaging technique is employed. Therefore, following a method proposed by Sanderson et al. (2017) for assigning weights to GCMs based on their uniqueness, a weighting methodology is devised in which highly correlated metrics are down-weighted. First, percent departures from the reference data for all metrics are converted to normalized relative errors as follows:

$$RE_{G,i} = \frac{PD_{G,i} - min(PD_{G_{all},i})}{max(PD_{G_{all},i}) - min(PD_{G_{all},i})} \quad (1)$$

Where $RE_{G,i}$ and $PD_{G,i}$ represent the normalized relative error and percent departure from the reference data for GCM G in metric i, respectively. $PD_{G_{all},i}$ represents the array of percent departures from the reference data across all GCMs for that metric. Second, pairwise Pearson linear cross-correlations are calculated for all metrics, which are converted into a distance measure as follows:

$$C^*{}_{i,j} = 1 - abs(C_{i,j}) \quad (2)$$

Where $C_{i,j}$ and $C^*{}_{i,j}$ represent correlation and correlation-based distance between metric i and metric j, respectively. The small magnitude of $C^*{}_{i,j}$ reflects high correspondence between the metrics and vice versa. Furthermore, we calculate the Similarity Score (SS) for each pair of metrics as follows:

$$SS_{i,j} = e^{-\left(\frac{C^*{}_{i,j}}{D_x}\right)} \quad (3)$$

Where D_x is a tunable parameter representing the radius of similarity that determines the correlation-based distances over which a metric can be considered redundant. Note that some covariance between different spatiotemporal characteristics of prognostic variables or between the prognostic and diagnostic variables is acceptable and unavoidable in a coupled Earth system. Therefore, our goal is to target only those metrics that exhibit correlations to such an extent that those measures effectively become redundant. We use 0.2 for D_x as it only down-weights those metrics that exhibit very high correlations in the four regions (Figure 2). SS value ranges between 0 and 1, as a metric uniqueness decreases with $SS \rightarrow 1$. Next, for each metric, the effective redundancy (ER) is calculated as follows:

$$ER_i = 1 + \sum_{j \neq i}^{n} SS_{i,j} \quad (4)$$

The inverse of the ER_i provides the weight for that metric. Finally, the average weighted relative error for each GCM is calculated as follows:

$$RE^{*}{}_{G} = \sum_{i=1}^{m} (ER_i)^{-1} RE_{G,i} \quad (5)$$

These weighted relative errors ($RE^{*}{}_{G}$) are calculated separately for each of the four CONUS subregions. The regionally weighted relative errors are subsequently averaged to provide the CONUS-scale weighted relative error used in the simple averaging technique to calculate the relative ranks of each GCM. The GCM with the lowest weighted relative error ranks at the top, whereas the GCM with the highest weighted relative error ranks at the bottom.

On the other hand, in the multivariate EOF analyses, models' skill is evaluated using the sum of their Euclidean distances from the observations in the PC space, as follows:

$$D(O,G) = \sqrt{\sum_{i=1}^{n} (G_i - O_i)^2} \quad (6)$$

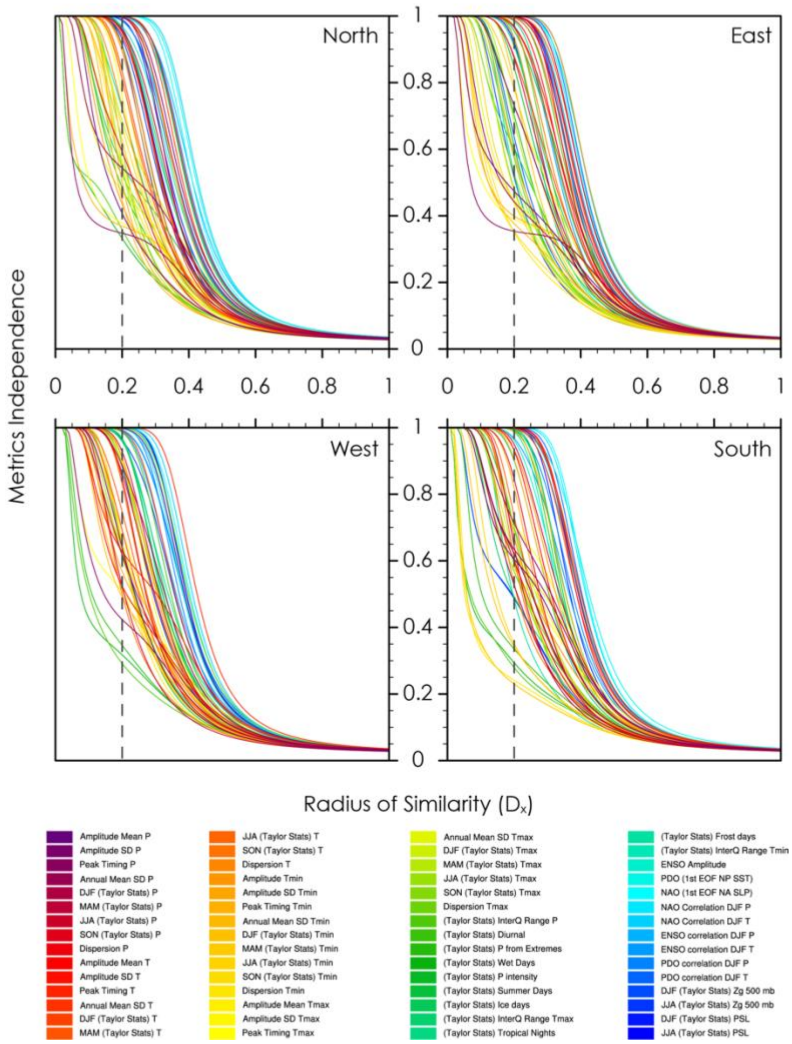

Figure 2. Metrics independence weights ($(ER_i)^{-1}$) as a function of the radius of their similarity (D_x). The grey vertical line represents the value of D_x used to calculate similarity scores.

Where $D(O, G)$ represents the Euclidean distance of GCM G from reference data O as a sum of the distances over n PCs, which in our case n =10. No strict criteria have been followed to select the number of PCs in calculating the sum of Euclidean distances through equation *6* in past studies. Some studies have used North's rule of thumb (North et al. 1982) to objectively sub-select statistically different numbers of PCs (e.g., Rupp et al. 2013), while others have made this selection subjectively (e.g., Chhin et al., 2018; Sanderson et al., 2015). However, they have acknowledged the difficulty of identifying each selected EOF's distinct characteristics (Rupp et al., 2013). This study tests the sensitivity of GCMs ranking to the number of PCs used in calculating Euclidean distances and notes that it substantially diminishes after the first ten modes (Figure 3). Therefore, distances between individual GCMs and observations are computed using the truncated set of the first ten modes. The GCM with the lowest total distance ranked at the top, whereas the GCM with the highest total distance ranked at the bottom.

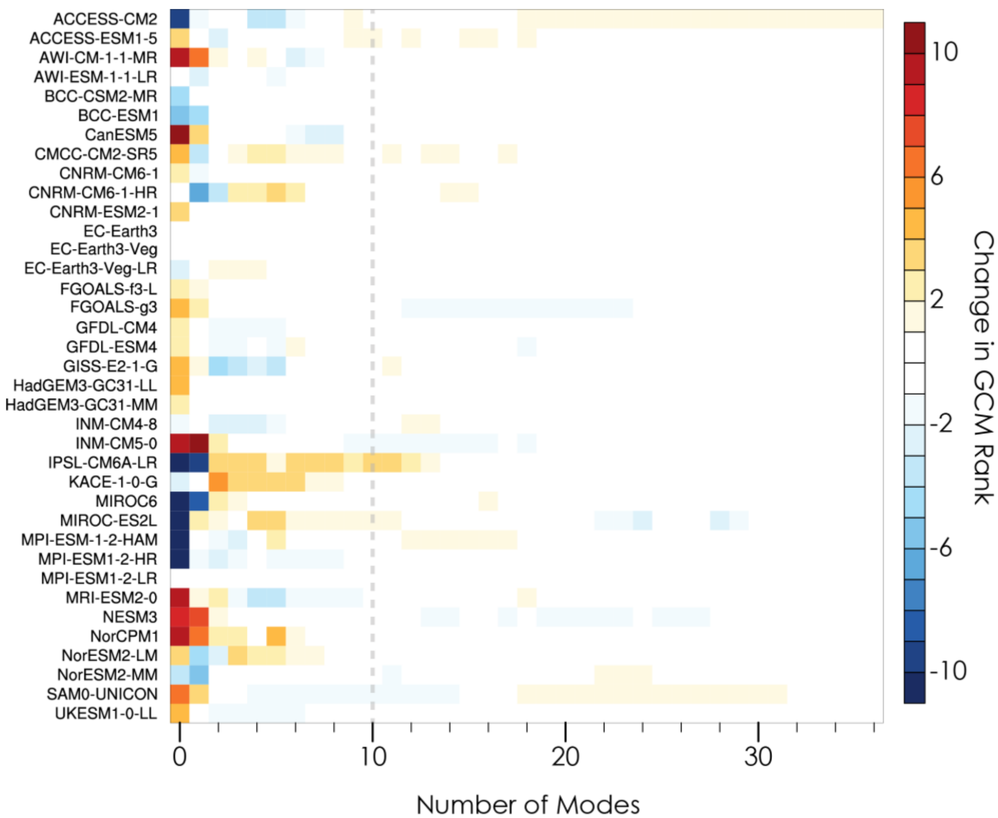

Figure 3. Deviation of GCMs ranking from the mean with the addition of PC modes. The grey line represents the number of modes used in this study for calculating the sum of the Euclidean distances.

3. Results and Discussion

3.1 The rationale for the choice of evaluation metrics

First and foremost, there may be questions regarding the rationale behind the choice of evaluation metrics used in this study. Note that our selection of metrics represents a wide range of spatiotemporal climate characteristics that are common across the CONUS and does not include those features that are unique to specific regions, such as integrated water vapor transport through atmospheric rivers in the western US, the monsoonal climate in the southwest and tornadic environment in the central and eastern United States. We have also avoided the inclusion of trends analyses in the metric suite, given that not all the observed regional trends are necessarily driven by the anthropogenic forcing, and the natural climate variability may influence some. Note that while greenhouse gas concentrations are aligned in the observations and historical CMIP6 GCMs simulations, the natural modes of climate variability, such as ENSO, PDO, NAO, are not. Therefore, lack of correspondence between regional-scale observed and simulated trends cannot be confidently used as a measure for model validation, as it is not straightforward to distinguish between the inconsistency arising from natural climate variability and that arising from model deficiencies. Irrespective of these choices, developing a well-defined universal set of metrics to assess modeling skill in climate models is relatively improbable, as it may vary depending on question framing, climate characteristics of the region of interest, and data availability. Nonetheless, metrics used in this study represent a wide range of stakeholders relevant climate characteristics over an area, including diurnal cycle, daily thresholds of temperature (e.g., frost days, summer days, ice days tropical nights), daily precipitation extremes, seasonal precipitation and temperature distributions, intra-annual variability (amplitudes, timing of peak magnitudes), the spatiotemporal characteristics of precipitation and temperature distributions (dispersion analyses), atmospheric dynamics and influences of relevant natural modes of climate variability. Therefore, not only this comprehensive evaluation should aid in decision-making when it comes to the selection of GCMs for downscaling studies, it is expected that the outcome of this evaluation would also be helpful for studies where spatial downscaling of GCMs is not intended. For studies with a more subregional focus, we expect that other metrics representing region-specific climate characteristics may be required for more informed model selection.

3.2 GCMs relative errors

The unweighted relative errors for each metric corresponding to all 37 GCMs are shown in Figure 4 for the North (see Figure 1 for regions definition) and in *Supplementary Figures* S1 to S3 for the remaining three regions. For ease of comparison, GCMs are sorted from left to right so that the GCM with the lowest average relative error is on the left and the one with the highest average relative error is on the right. Unlike the absolute error, the relative error is not a direct measure of modeling biases with respect to truth or observations, as it differentiates models from each other. Nonetheless, models with higher magnitudes of relative error would be further away from the observations than those with lower magnitudes. The line plot panel on the right displays

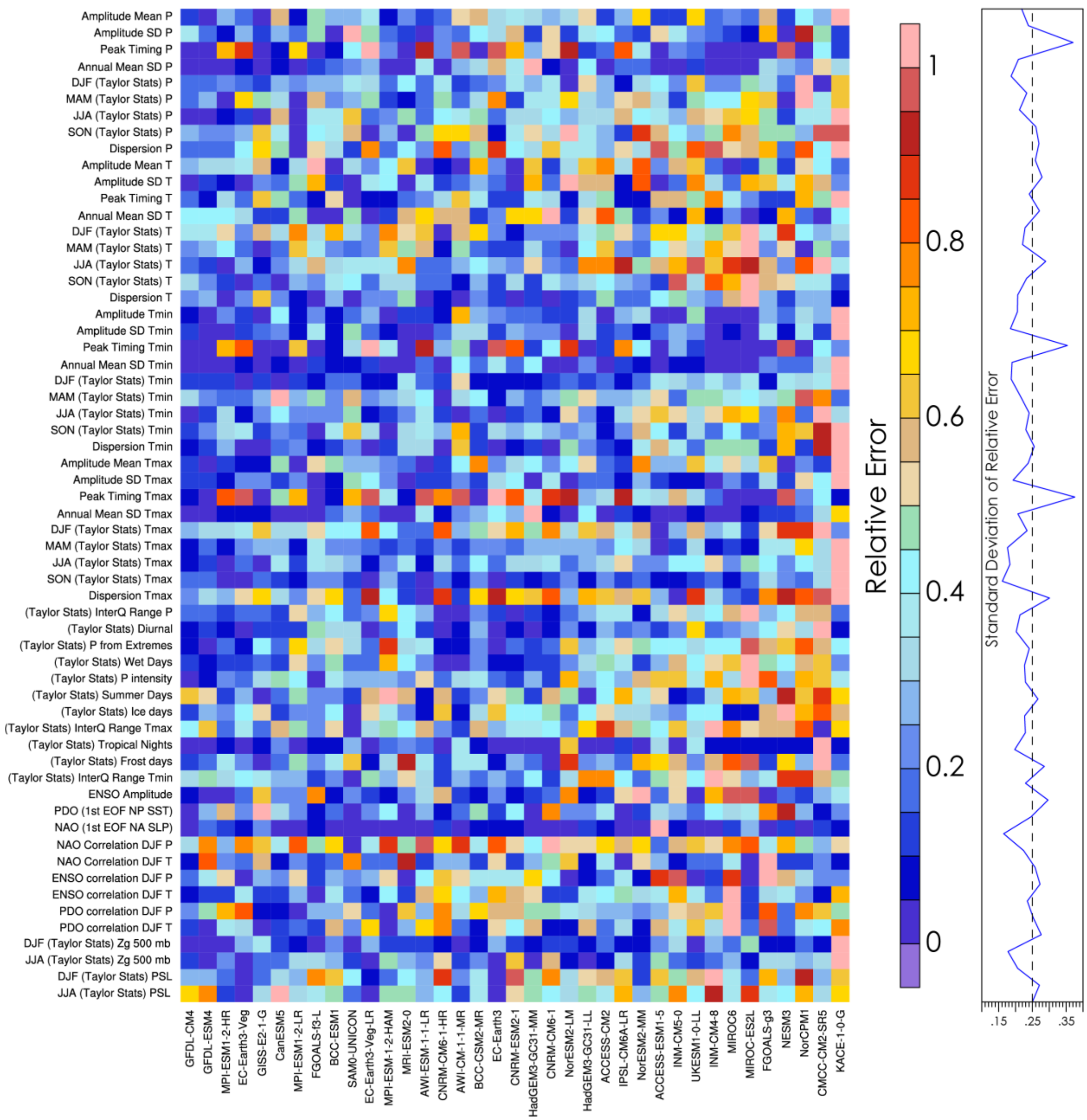

Figure 4. The unweighted relative errors of GCMs over the North. The left panel shows relative errors corresponding to each metric across all GCMs and the line plot on the right shows the standard deviation of the relative error for each metric across all GCMs.

the standard deviation of relative errors across GCMs for each metric. Note that if the performance of many models falls in a similar category, their relative errors display a similar range of colors. High standard deviation magnitudes represent substantial variation in modeling skills across the GCMs and vice versa.

Overall, many GCMs exhibit challenges in simulating key climate characteristics. For instance, while models are relatively skillful in representing oceanic and atmospheric patterns associated with natural forcing (ENSO, NAO, PDO), most show limited skill in simulating their influences on the distribution of seasonal mean precipitation and temperature over the South and West. Difficulties in reproducing the observed timing of peak magnitudes of precipitation, minimum temperature, and maximum temperature are also evident in the West and North, and metrics for precipitation characteristics are relatively poorly simulated in the South. One noticeable distinction between better and poor performing models is that the latter group is deficient in reproducing several daily-scale features of temperature and precipitation characteristics across all regions. Several models consistently display similar better performance across all four CONUS regions. For instance, KACE-1-0-G and NorCMP1 are always in the bottom three, while GFDL-CM4 and EC-EARTH3-Veg are mainly in the top three. Some models exhibit substantial variation in performance across regions. For instance, ACESS-ESM1-5 is near the bottom over the East and South but jumps to the top third in the West. Similarly, BCC-ESM1 falls in the fourth quarter over the North but remains at the average or below average over the rest of the regions. However, these relative unweighted rankings of the GCMs are inconclusive, given potential redundancy in the evaluation metrics.

3.3 Metrics redundancy

The pairwise absolute correlations, metrics similarity score, and overall metrics weight are shown in Figure 5 for the North and *Supplementary Figures* S4 to S6 for the remaining three regions. The correlation-based distance metric ($C^*_{i,j}$) shows that only ~0.8% of the total pairwise absolute correlations between any two metrics are > 0.8 ($C^*_{i,j} < 0.2$) in each region while 5–7% of $C^*_{i,j}$ are lower than 0.5 (absolute correlations > 0.5) across the four regions. These small numbers suggest that majority of the evaluation metrics are primarily independent of each other. Note that the primary intent for correlative analyses in this study is to minimize the possibility of unwanted spurious biases in the GCMs ranking process due to metric redundancy. Still, it also provides valuable insight into the spatiotemporal interplay of various characteristics of background climate over a region in GCM simulations. Over the CONUS, the strong positive associations among the evaluation metrics are relatively higher than the strong negative associations. To explain this point, if we only considered those cases where correlations are >±0.6 or stronger, there

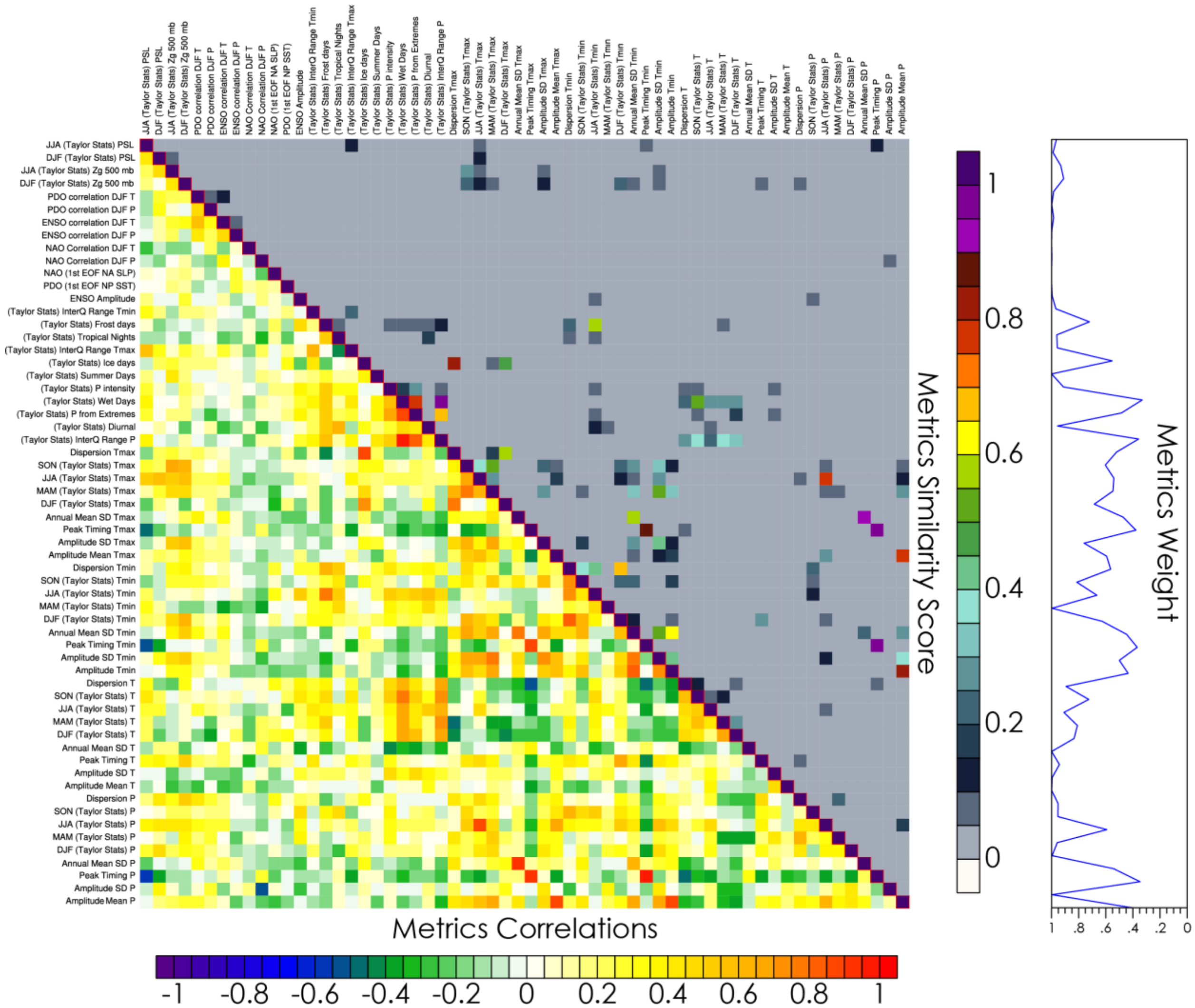

Figure 5. The correlation between the pairwise metrics (bottom triangle) and the corresponding similarity score (top triangle) over the North. Metrics with high correlations exhibit a high similarity score and are down-weighted. The line plot on the right shows the overall weight for each metric.

is only one instance over the South where the magnitude of negative correlation qualifies this threshold between any two metrics (Figure S6). The distribution of strong positive associations among the evaluation metrics is reasonably similar across four regions. Among them, the most notable and common cases across four regions include the covariance of modeling errors in metrics representing 1) the timing of peak magnitudes of precipitation, minimum temperature, and maximum temperature, 2) the wet days, precipitation from extremes, and interquartile precipitation range, and 3) the dispersion statistics of minimum temperature and its seasonal characteristics. Moreover, frost days metric strongly covary with metrics representing winter precipitation in the

South, and with metrics representing seasonal characteristics of minimum temperature and wet days in the East, while metric describing autumn (September–October–November, SON) mean temperature strongly correlates with those representing precipitation intensity and interquartile precipitation range in the South (≥ 0.8). Positive high correlations also exist between metrics for seasonal mean temperature characteristics with those for wet days and precipitation from extremes in the North (≥ 0.7). Most of these strong interdependencies require identifying systematic causative linkages for their physical explanation, which is neither the intent nor the focus of this study. Nonetheless, all such metrics with strong correlations are proportionally downweighed, as reflected in their corresponding similarity scores and overall weights.

The information redundancy in the evaluation metrics suite can also be taken care of using EOF analysis. It finds a subset of metrics that convey as much as original information by reducing the data dimensionality. One can examine individual loadings of PCs to identify metrics that provide maximum aid in distinguishing between better and poor-performing models. Note that more substantial loadings in our analyses do not necessarily mean that those associated variables are critical measures for a model to perform better; they imply a higher contribution of those metrics to a particular PC when EOF analysis is applied on the matrix of sixty measures across 37 CMIP6 GCMs. The list of significant contributors can potentially vary if the input data matrix is changed. Alternatively, metrics with weaker loadings may suggest that most models exhibit similar skills in simulating those characteristics. Therefore, such measurements provide little ability to identify models' distinctiveness.

When the first ten EOFs are considered, which represent approximately > 76% of the explained variance in each region, they reveal a regionally varying list of dominant metrics. Still, some interesting features are worth highlighting and explaining. Relatively fewer metrics, including the ones representing the timing of annual peaks for precipitation, minimum temperature, and the maximum temperature, noticeably contribute to the first few dominant modes over the North. Interestingly, this is the only region where these few modes distinctively exhibit higher variability across the GCMs (Figure 6). Therefore, it is understandable that these modes have a higher contribution to the first few PCs over the North. These metrics also exhibit strong loadings for several PCs in other regions. Moreover, South and East display the noticeable contribution from metrics representing the seasonal characteristics of minimum temperature to the first PC. In these cases, and many others not mentioned, the metrics contributing more to the first

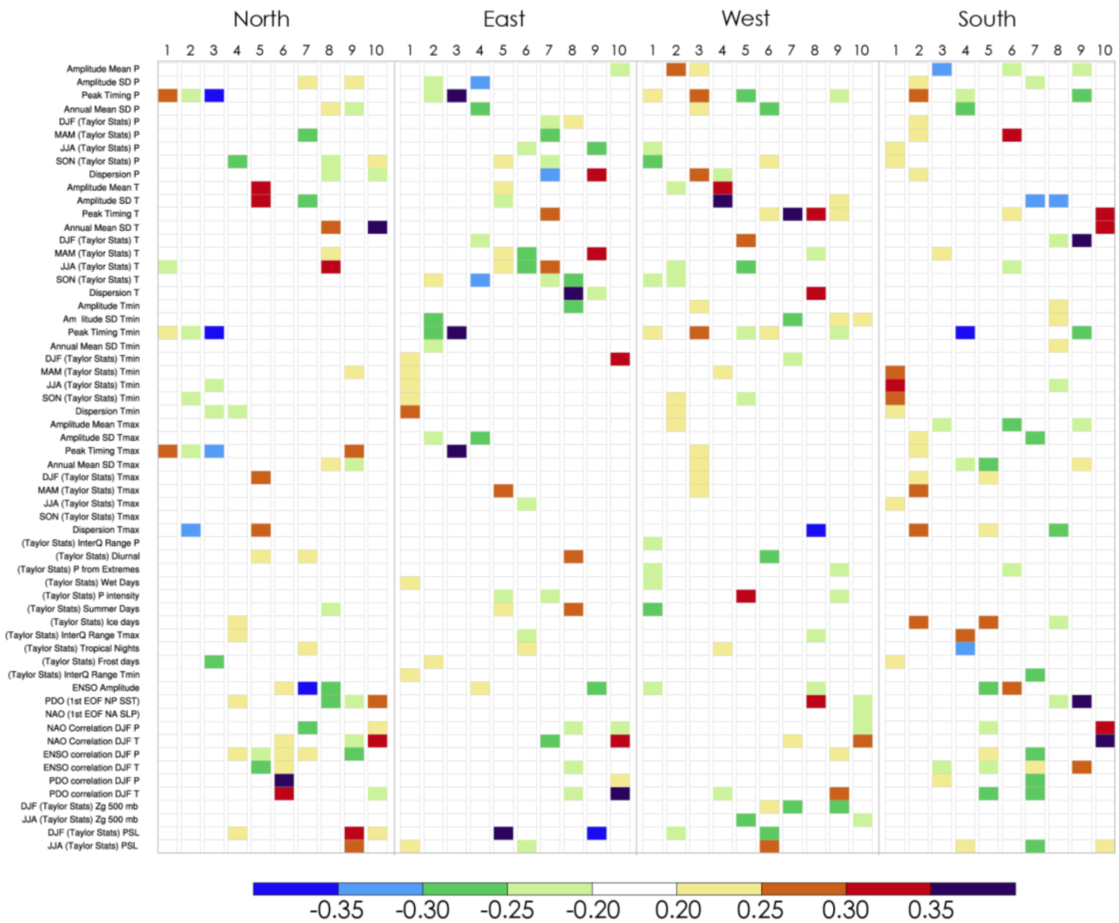

Figure 6. The loadings of metrics with a relatively substantial contribution to the first 10 EOFs over each region.

few PCs are likely the ones for which GCMs exhibit substantial variability in representing their characteristics. More interestingly though, these metrics are also the ones that display strong correlations with other evaluation measures. Recall that EOF analyses reduce data dimensionality while conserving the explained variance. Therefore, it should be intuitive that a single metric that exhibits strong correlations with several other metrics contributes more to the first few PCs. In principle, this approach contrasts with the first methodology. In the simple weighted averaging technique where weights are assigned to each metric before averaging, metrics with higher correlations are downweighed so that weights are distributed among the correlated set of metrics. In contrast, EOF analyses remove redundancy in data by assigning those metrics more weight that display correlations with several others, as the information in other metrics is already embedded in the selected set. However, this distinctiveness between the two approaches is not evident in the remaining PCs. For instance, metrics representing atmospheric teleconnections and dynamics make up the list with more substantial loadings for PCs 3–7 over the four regions. At the same time, most of them get very high weights in the simple averaging approach due to their relatively little to no correlations with other metrics.

3.4 GCMs relative ranking and independence

The regional and CONUS scale relative GCM rankings are shown in Figure 7 for the two methodologies. The two approaches yield reasonably similar results at the CONUS scale, as the same GCMs occupy not only the first and fourth quartiles in both techniques, the individual GCM placements within these quartiles are also very similar. For instance, the bottom five GCM rankings are identical in both cases, and the maximum difference in ranking in the fourth quartile ranges from 0 to 2. The commonality between the outcome of two approaches is also evident in regional rankings as identical models in the two approaches exhibit substantial deviation from their mean CONUS-scale relative measures (relative error or Euclidean distances), such as MRI-ESM2-0, CNRM-CM6-1, and MIROC6 over the South, GISS-E2-1-G and MIROC6 over the West, and ACCESS-CM2 and NorESM2-MM over the North. The remaining GCMs falling between the top and bottom quartiles tend to exhibit considerably minor differences in their weighted relative errors in the case of simple averaging and total Euclidean distance in the case of EOF analyses. The high-resolution model from several institutes distinctively performs better than the lower resolution version, with at least 5 level differences in their relative placement in both methodologies. For instance, MPI-ESM1-2-HR ranks higher than MPI-ESM1-2-LR, HdGEM3-GC31-MM ranks higher than HdGEM3-GC31-LL, while NorESM2-MM displays better performance than NorESM2-LM.

Several models in the CMIP6 share modeling components. The component sharing is more significant in the models from the same institute, such as models contributed by U.S. Geophysical Fluid Dynamics Laboratory (GFDL) or those contributed by the United Kingdom Met (UKMET) Office in the CMIP6. Components sharing across institutes are also standard. For instance, Australian Commonwealth Scientific and Industrial Research models (ACCESS-CM2, ACCESS-ESM1-5) share several components developed by GFDL and UKMET (https://research.csiro.au/access/about/). Similarly, the Norwegian Earth System Model (NorESM2) is based on the second version of CESM (CESM2) (Seland et al., 2020), while Seoul

National University Atmospheric Model Version 0 with a Unified Convection Scheme (SAM0-UNICON) is based on the first version of CESM (CESM1) (Park et al., 2019).

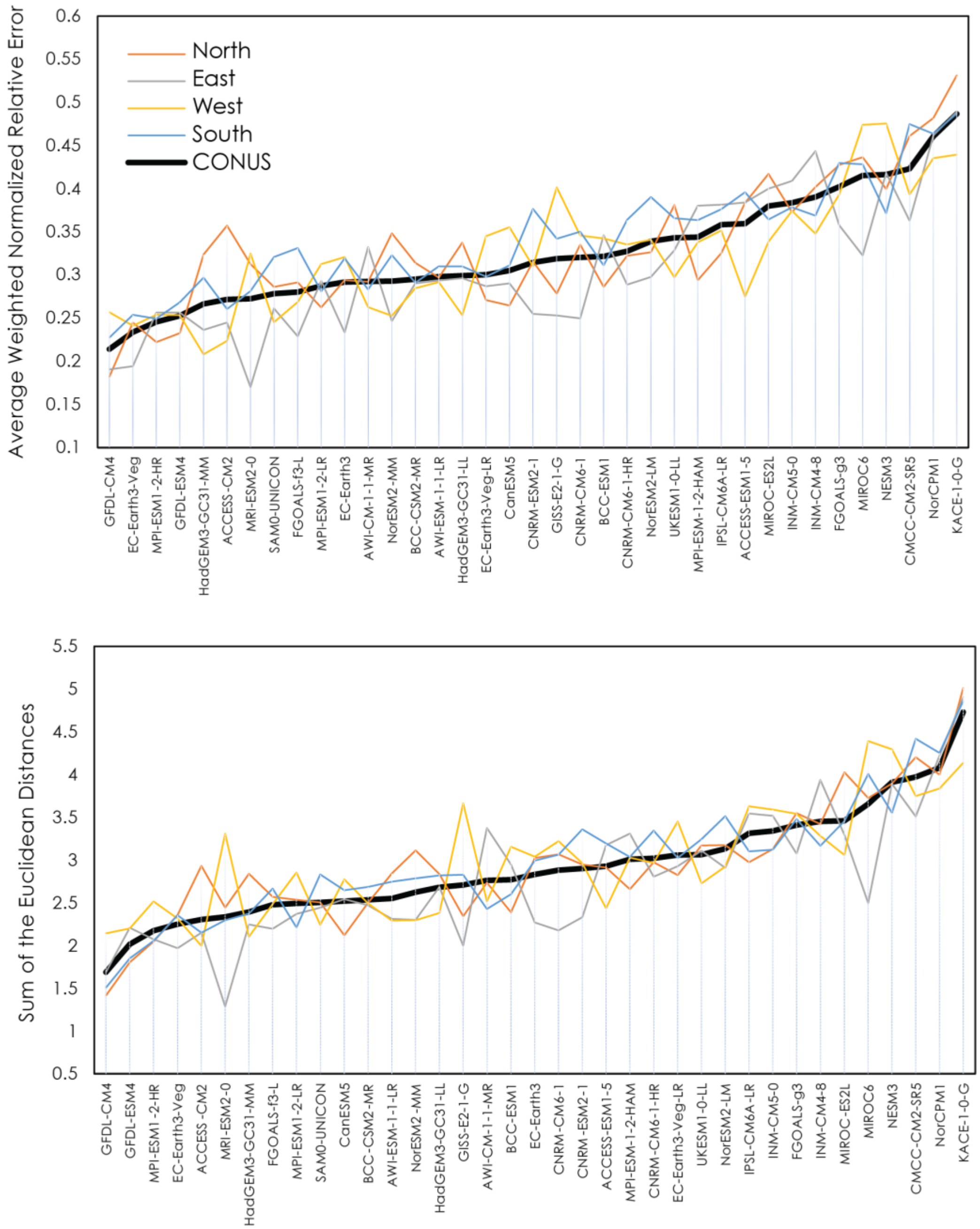

Figure 7. The ranking of GCMs using the simple weighted averaging (top) and EOF-based Euclidian distances. The thin lines represent models' relative ranking over four sub-regions, and the thick line represents the overall CONUS scale ranking.

Given the commonality of modeling components, it is quite possible that these models, particularly those from the same developing institute, exhibit similar biases. Other studies have used techniques to assign weights to models based on their independence, which is useful when various factors impacting the robustness of future climate change are in question (Knutti et al., 2017; Sanderson et al., 2015). However, this study intends to guide the sub-selection of GCMs for downscaling studies based on their performance in the historical period. Therefore, we restrict ourselves to the relatively less quantitative identification of models' interdependencies by comparing PCs from the EOF analysis – an approach quite commonly used in many earlier studies. When the loadings of the first two PCs from EOF analyses are compared, they show models from the same developing center clustering in the same PC space, highlighting the similarities among those models (Figure 8). Therefore, if a model selection is necessary for downscaling purposes, the selection of models should consider both the skill and the independence of the selected models. An easier choice in the case of many is to go for the higher resolution versions, as those display relatively better skill.

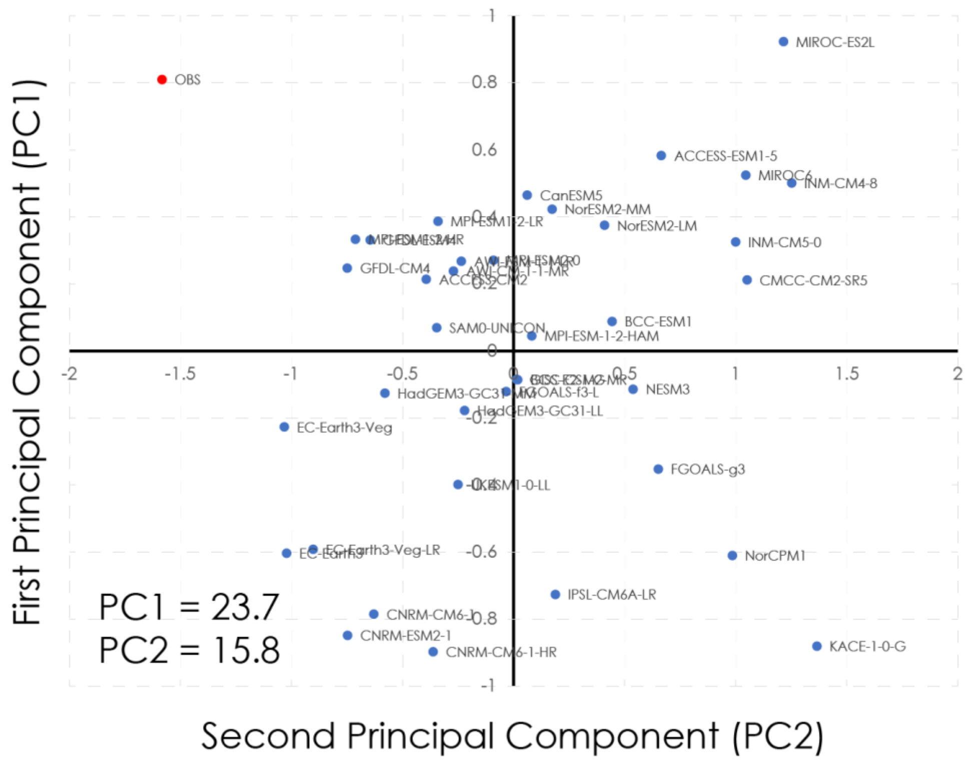

Figure 8. The loadings of PC1 versus PC2. The two PCs explain 39.5% of the total variance across the GCMs. OBS represents the observations.

4. Summary

We analyze the performance of CMIP6 GCMs across 60 evaluation metrics over four CONUS regions. The analysis is restricted to 37 models with complete data needed to calculate all evaluation metrics. Based on the performance of models across the evaluation measures, two methodologies are used to rank the models relative to each other while accounting for the redundancies in the metrics suite. The first methodology employs a simple weighted averaging technique where a GCM's relative errors across all evaluation metrics are averaged after each metric is assigned a weight based on its uniqueness. The second methodology employs EOF analysis to reduce the dimensionality of data where metrics that explain the variability across the GCMs ensemble receive higher loadings – the coefficients of the linear combination of the original metrics from which the PCs are constructed. The two methodologies work in contrasting ways to remove the metrics redundancy but eventually develop relatively similar GCMs rankings. The consistency in the model ranking between the two methods can also be partly due to an extensive suite of metrics used in analyses that perhaps reduce the possibility of substantial deviations in the outcome.

The evaluation in this study is intended for downscaling studies where GCMs sub-selection is necessary due to many unavoidable factors. Many of the evaluated models provide 6-hourly atmospheric fields. Therefore, the results from this study should be helpful in the selection of models for dynamical downscaling efforts, such as CORDEX. The results can also be beneficial in understanding the strengths and deficiencies of CMIP6 GCMs in representing various background climate characteristics if direct use of GCMs is intended. While we have used an extensive suite of evaluation metrics, this list is in no way comprehensive. It should be considered only as a guideline where a more in-depth understanding of GCMs performance is required, particularly of specific phenomena such as North American monsoon, Atmospheric rivers, and severe weather environments. Note that our study does not include any models from NCAR in the CMIP6 because their daily minimum and maximum temperatures data were not available at the time of this analysis. However, we would like to point out that NCAR models were among the better performing GCMs when fewer metrics were used (not shown). Lastly, note that only two methodologies are used for GCMs ranking. Therefore, results may not be entirely insensitive to the choice of the ranking process.

Acknowledgment

This research is supported by the US Department of Energy (DOE) Water Power Technologies Office as a part of the SECURE Water Act Section 9505 Assessment. We thank the Earth System Grid Federation (ESGF) for archiving and providing free access to the CMIP6 dataset. This research used the Oak Ridge Leadership Computing Facility resources at Oak Ridge National Laboratory, which is a DOE Office of Science User Facility. MA, DR, and SCK are employees of UT-Battelle LLC under contract DE-AC05-00OR22725 with the US DOE. Accordingly, the US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so, for US Government purposes.

Data Availability

All datasets used in this study are publicly available.

CMIP6: (from https://esgf-node.llnl.gov/projects/cmip6/)

Daymet: (from https://daymet.ornl.gov/)

ERA5: (from https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5)

Livneh: (from https://psl.noaa.gov/data/gridded/data.livneh.html)

PRISM: (from https://prism.oregonstate.edu/)

References

Ahmadalipour, A., Rana, A., Moradkhani, H., & Sharma, A. (2017). Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis. *Theoretical and Applied Climatology*, *128*(1), 71-87.

Ahmed, K., Sachindra, D. A., Shahid, S., Demirel, M. C., & Chung, E. S. (2019). Selection of multi-model ensemble of general circulation models for the simulation of precipitation and maximum and minimum temperature based on spatial assessment metrics. *Hydrology and Earth System Sciences*, *23*(11), 4803-4824.

Beobide-Arsuaga, G., Bayr, T., Reintges, A. et al. (2021). Uncertainty of ENSO-amplitude projections in CMIP5 and CMIP6 models. *Climate Dynamics,* **56,** 3875–3888. https://doi.org/10.1007/s00382-021-05673-4

Boé, J. Interdependency in multimodel climate projections: component replication and result similarity. *Geophy. Res. Lett.* **45**, 2771–2779 (2018).

Chhin R, Yoden S (2018) Ranking CMIP5 GCMs for model ensemble selection on regional scale: Case study of the Indochina Region. *Journal of Geophysical Research: Atmospheres* 123(17):8949–8974, DOI: https://doi.org/10.1029/2017JD028026

Danabasoglu et al. (2020). Journal of Advances in Modeling Earth Systems, 12(2), e2019MS001916, https://doi.org/10.1029/2019MS001916

Eyring V, Bony S, & Meehl GA et al. (2016). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development,* **9**, 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Gutowski et al. (2016). WCRP COORDINATED REGIONAL DOWNSCALING EXPERIMENT (CORDEX): A Diagnostic MIP for CMIP6; *Geoscientific Model Development* **9**., 4087-4095, doi:10.5194/gmd-2016-120.

Giorgi, F., & Gutowski, W.J. (2016). Coordinated Experiments for Projections of Regional Climate Change. *Current Climate Change Report,* **2**, 202–210. https://doi.org/10.1007/s40641-016-0046-6

Giorgi F, Jones, C., & Asrar, G. R. (2009). Addressing climate information needs at the regional level: the CORDEX framework. WMO Bulletin **58**(3):176–183

Hersbach H, et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, **146**(730), 1999-2049. https://doi.org/10.1002/qj.3803.

Hurrell, J.W. (1995). Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science,* **269**,676-679.

Hurrell, J. W., & Deser, C. (2009). North Atlantic climate variability: The role of the North Atlantic Oscillation. *Journal of Marine Systems*, **78**(1), 28-41.

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

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research letters,* **40**, *1194-1199*, https://doi.org/10.1002/grl.50256

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., & Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate,* **23**(10), *2739-2758.* https://doi.org/10.1175/2009JCLI3361.1

Lorenz et al. (2017). Prospects and Caveats of Weighting Climate Models for Summer Maximum Temperature Projections Over North America; *JGR Atmospheres*, **123**(9), 4509-4526, https://doi.org/10.1029/2017JD027992

McBride, L. A., Hope, A. P., Canty, T. P., Bennett, B. F., Tribett, W. R., & Salawitch, R. J. (2021).Comparison of CMIP6 Historical Climate Simulations and Future Projected Warming to an Empirical Model of Global Climate, Earth Syst. Dynam., **12**, 545–579. https://doi.org/10.5194/esd-12-545-2021

Mantua, N.J., Hare, S.R., Zhang, Y., Wallace, J. M. & Francis, R.C. (1997). A Pacific Interdecadal Climate Ooscillation with Impacts on Salmon Production. *Bulletin of the American Meteorological Society*, **78**, 1069-1079.

McKenna, S., Santoso, A., Gupta, A.S. et al. (2020). Indian Ocean Dipole in CMIP5 and CMIP6: characteristics, biases, and links to ENSO. *Scientific Reports*, **10,** 11500. https://doi.org/10.1038/s41598-020-68268-9

Newman, M., et al. (2016), The Pacific Decadal Oscillation, Revisited, *Journal of Climate*, **29**(12), 4399-4427.doi. 10.1175/JCLI-D-15-0508.1

Naz, B. S, Kao, S-C, Ashfaq, M., Rastogi, D., Mei, R. & Bowling, L.C. (2016). Regional hydrologic response to climate change in the conterminous United States using high-resolution hydroclimate simulation; *Global and Planetary changes*,**143**, 100-117. https://doi.org/10.1016/j.gloplacha.2016.06.003

North et al. (1982). Sampling Errors in the Estimation of Empirical Orthogonal Functions, *Monthly weather Review*, **110**(7), 699-706. doi:https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2

Overland et al. (2011). Considerations in the Selection of Global Climate Models for Regional Climate Projections: The Arctic as a Case Study; *Journal of Climate*, **24**, 6, 1583-1597. https://doi.org/10.1175/2010JCLI3462.1

Parding et al. (2020). GCMeval – An interactive tool for evaluation and selection of climate model ensembles, *Climate Services*, **18,** 100167. https://doi.org/10.1016/j.cliser.2020.100167

Park, S., J. Shin, S. Kim, E. Oh, and Y. Kim, 2019: Global climate simulated by the seoul national university atmosphere model version 0 with a unified convection scheme (SAM0-UNICON). *J. Climate*, **32**, 2917–2949, https://doi.org/10.1175/JCLI-D-18-0796.1.

Pierce et al. (2009). Selecting global climate models for regional climate change studies; *Proceedings of National Academy of Sciences, U S A.*, **106**(21), 8441–8446. doi: 10.1073/pnas.0900094106

Rupp, D. E, Abatzoglou, J. T., Hegewisch, K. C., & Mote, P. W. (2013). Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA, *JGR. Atmospheres*,**118**(19), 10888-10906. https://doi.org/10.1002/jgrd.50843

Sanderson, B., Wehner, M., & Knutti, R. (2017). Skill and independence weighting for multi-model assessments. *Geoscientific Model Development*, **10**, 2379-2396. https://doi.org/10.5194/gmd-10-2379-2017

Sanderson, B. M., Knutti, R., & Caldwell, P. (2015). A representative democracy to reduce interdependency in a multimodel ensemble, *Journal of Climate*, **28**, 5171–5194.

Seland, Ø., Bentsen, M., Olivié, D., Toniazzo, T., Gjermundsen, A., Graff, L. S., Debernard, J. B., Gupta, A. K., He, Y.-C., Kirkevåg, A., Schwinger, J., Tjiputra, J., Aas, K. S., Bethke, I., Fan, Y., Griesfeller, J., Grini, A., Guo, C., Ilicak, M., Karset, I. H. H., Landgren, O., Liakka, J., Moseid, K. O., Nummelin, A., Spensberger, C., Tang, H., Zhang, Z., Heinze, C., Iversen, T., and Schulz, M.: Overview of the Norwegian Earth System Model (NorESM2) and key climate response of CMIP6 DECK, historical, and scenario simulations, Geosci. Model Dev., 13, 6165–6200, https://doi.org/10.5194/gmd-13-6165-2020, 2020.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Trenberth, K. E. (1997) The Definition of El Niño. *Bulletin of the American Meteorological Society*, **78**, 2771-2777.

[Journal of Geophysical Research–Atmospheres]

Supporting Information for

Evaluation of CMIP6 GCMs over the CONUS for downscaling studies

Moetasim Ashfaq*[1], Deeksha Rastogi[1], Muhammad Adnan Abid[2], Shih-Chieh Kao[3]
[1] Computational Sciences and Engineering Division (CSED), Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Earth System Physics, Abdus Salam International Centre for Theoretical Physics, Trieste, Italy
[3] Environmental Science Division (ESD), Oak Ridge National Laboratory, Oak Ridge, TN, USA

Contents of this file

Figures S1 to S6

Introduction

The supplementary material provides the remaining regional analyses not shown in the main text.

Supplementary Figures

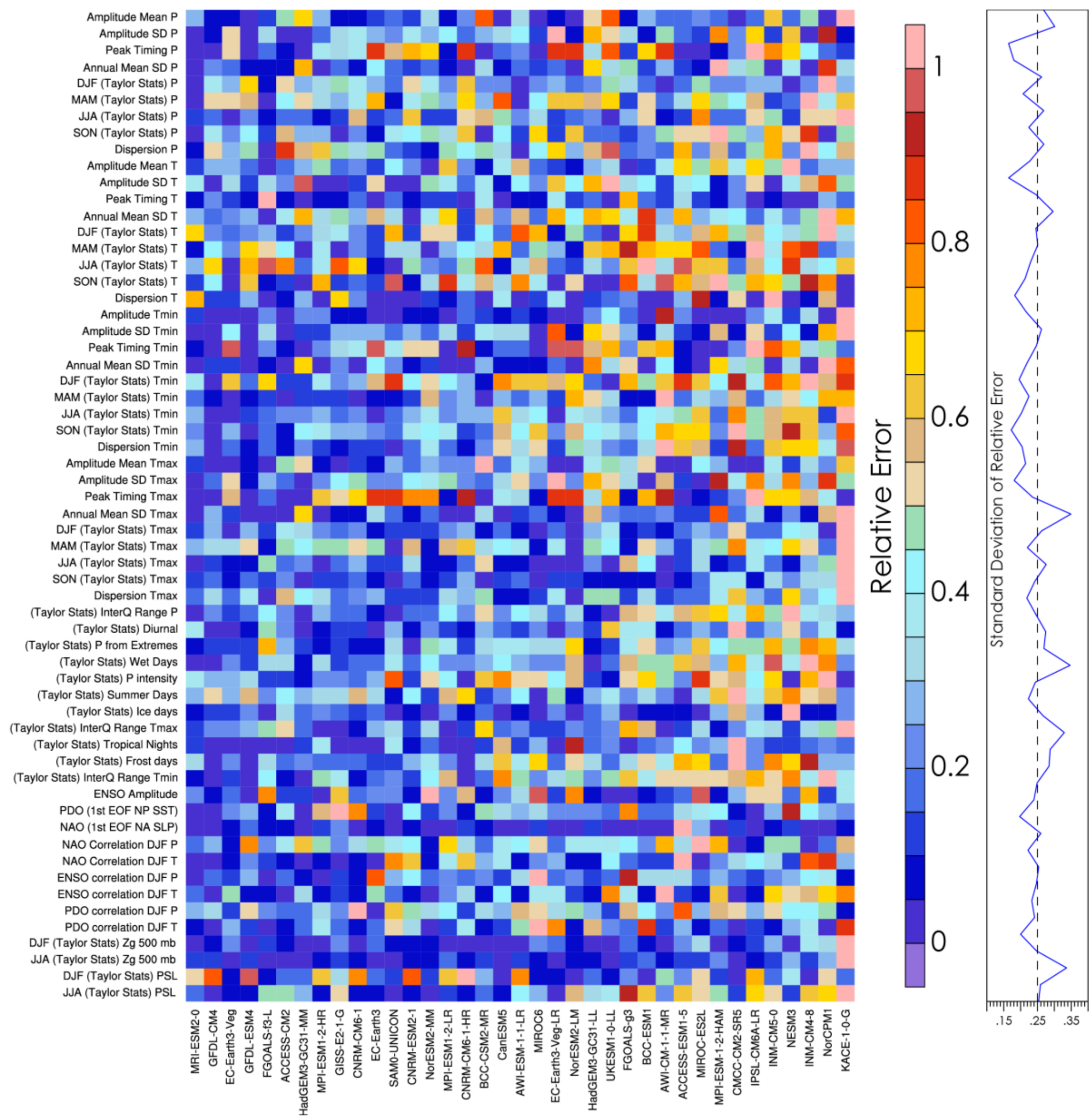

Figure S1. The unweighted relative errors of GCMs over the East. The left panel shows relative errors corresponding to each metric across all GCMs, and the line plot on the right shows the standard deviation of the relative error for each metric across all GCMs.

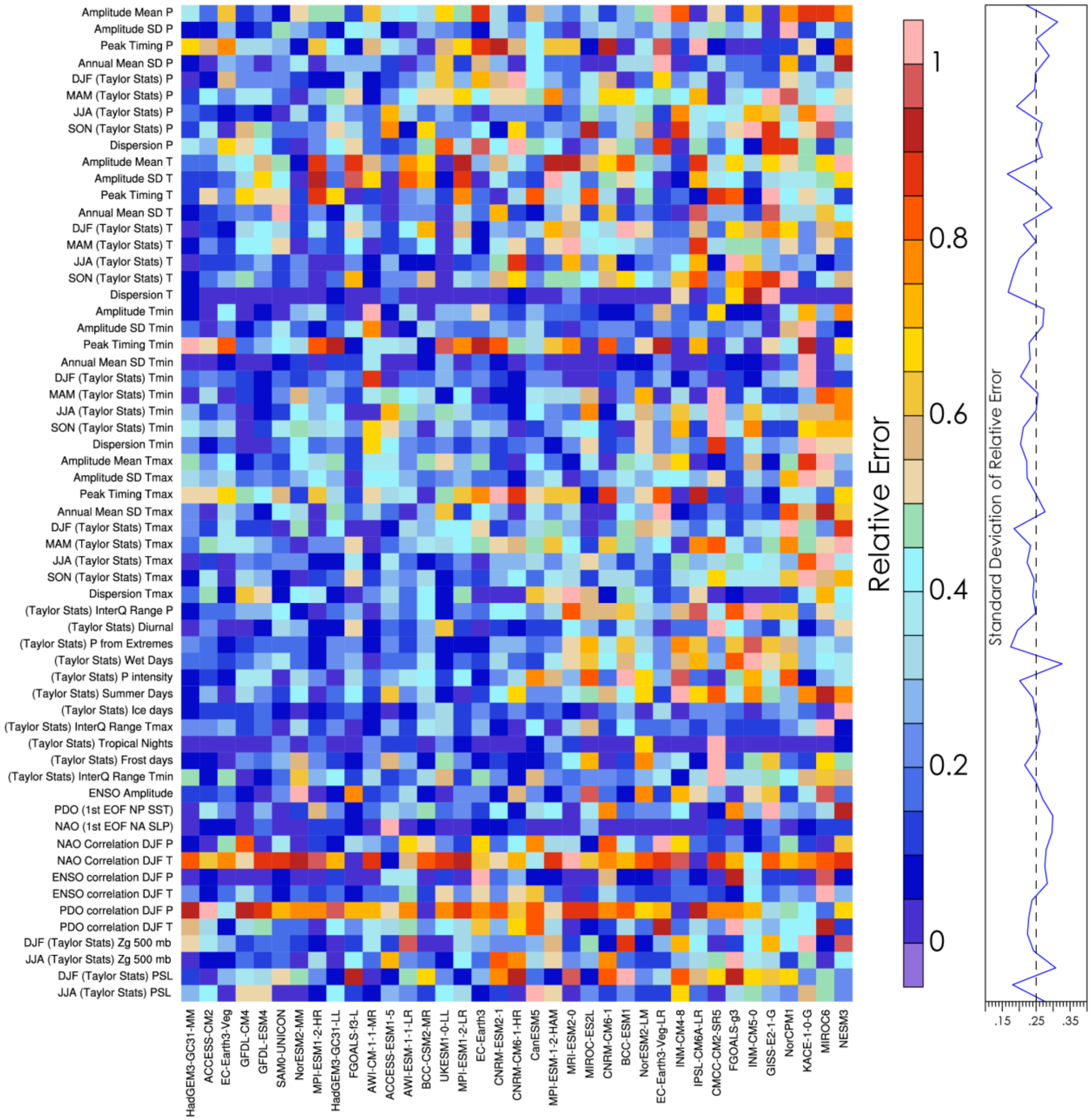

Figure S2. Same as in Figure S1 but over the West.

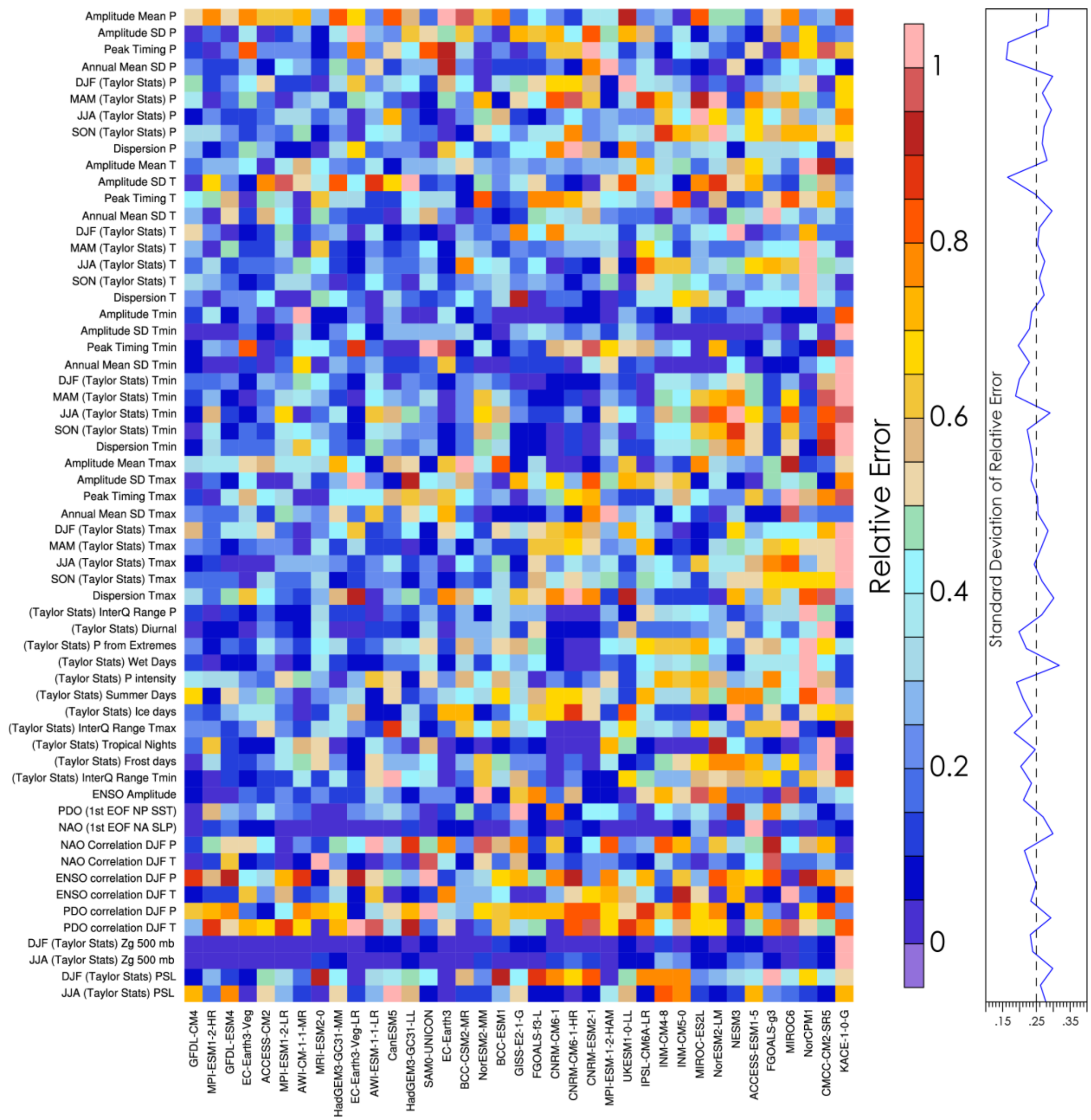

Figure S3. Same as in Figure S1 but over the South.

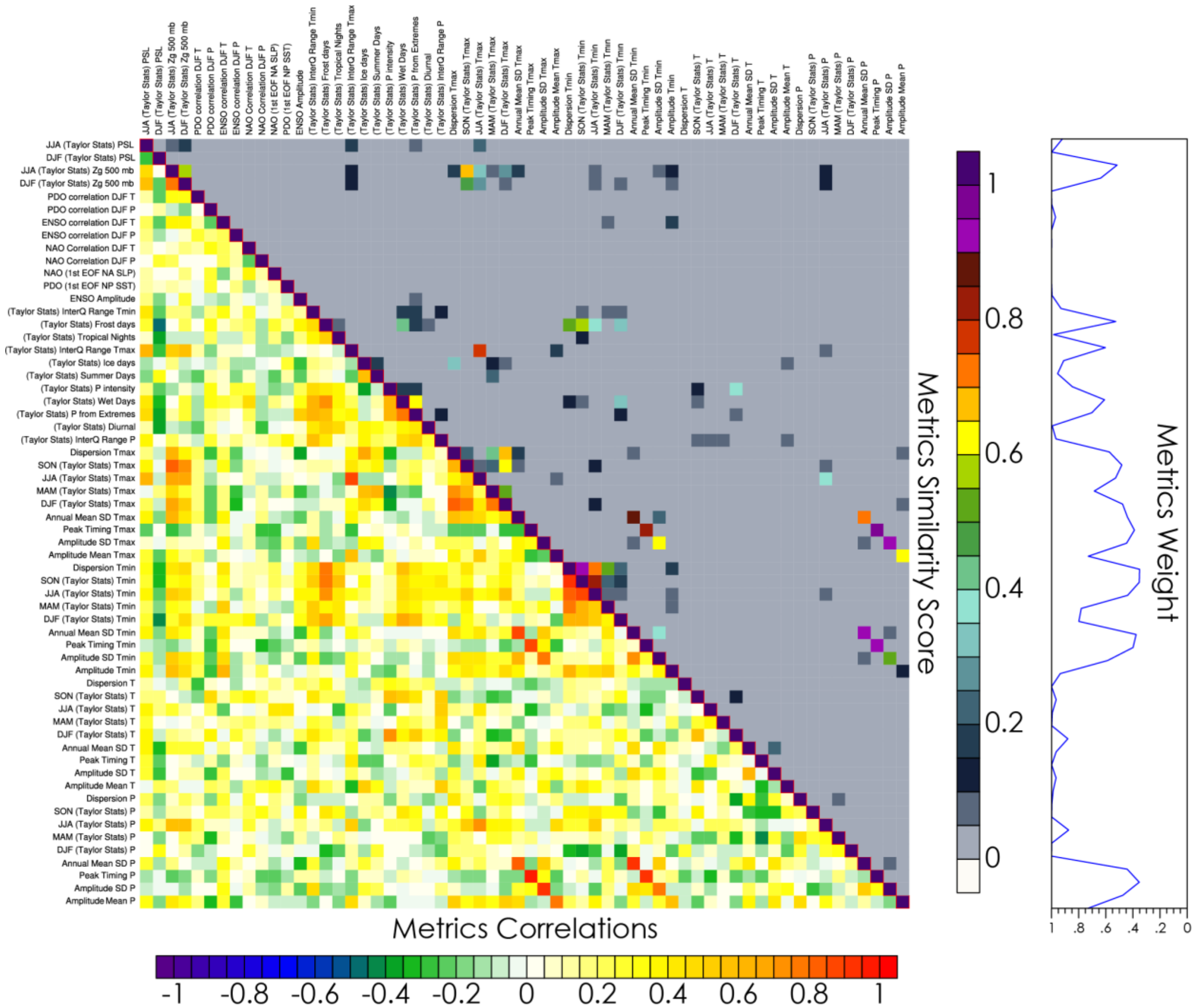

Figure S4. The correlation between the pairwise metrics (bottom triangle) and the corresponding similarity score (top triangle) over the East. Metrics with high correlations exhibit a high similarity score and are down-weighted. The line plot on the right shows the overall weight for each metric.

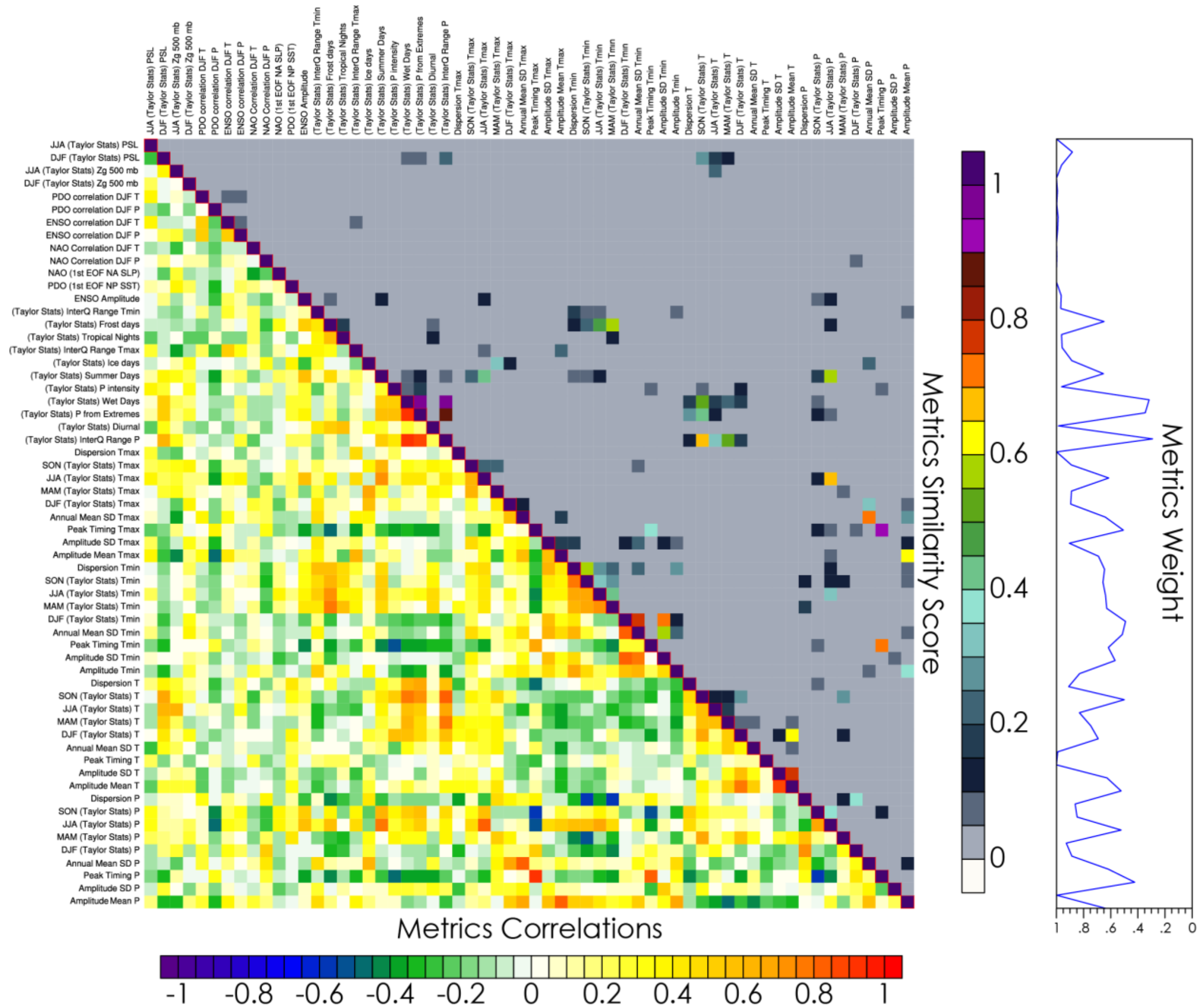

Figure S5. Same as in Figure S4 but over the West.

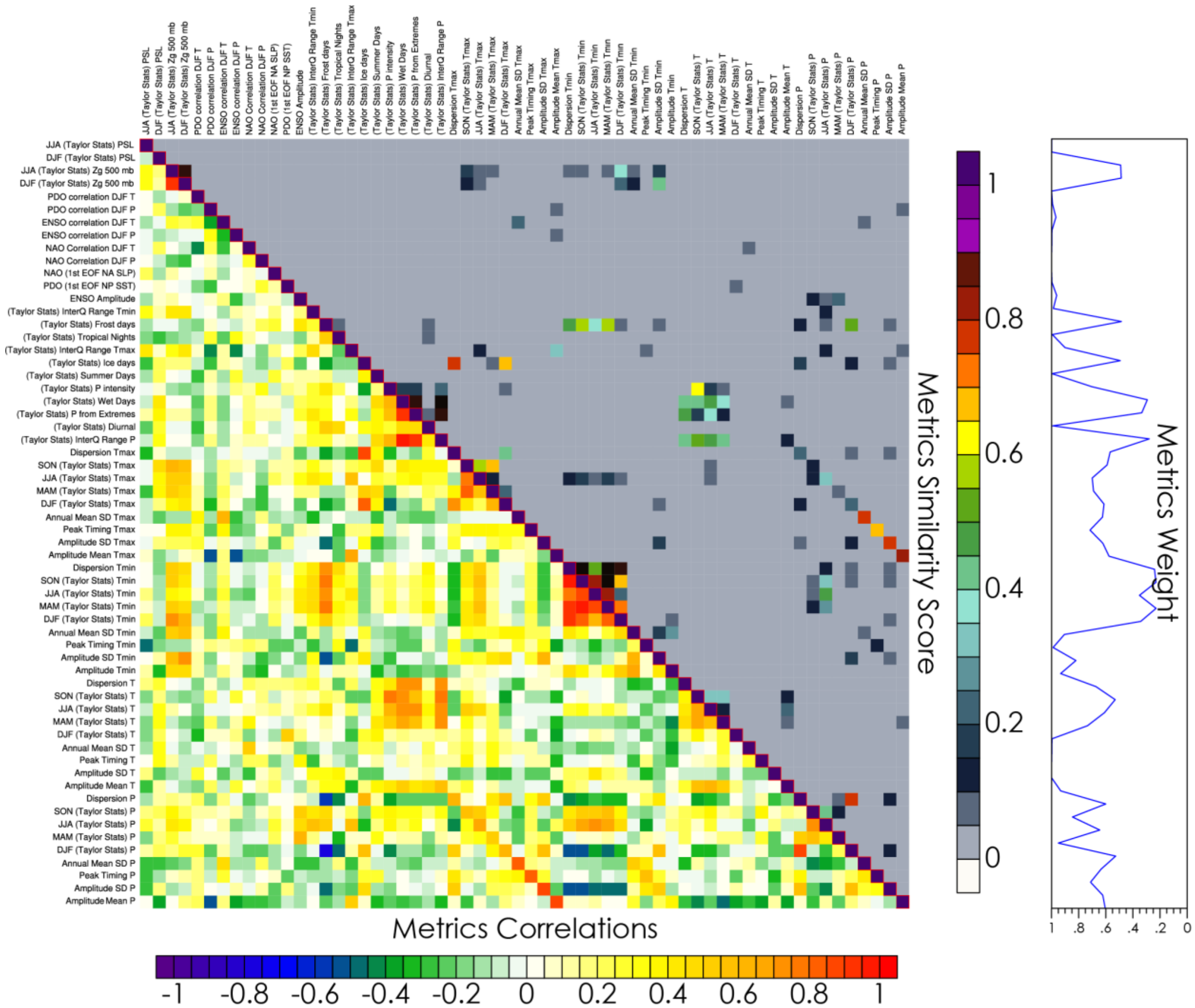

Figure S6. Same as in Figure S4 but over the South.