Article type : Technical Paper

# Evaluating Global Climate Models for Hydrological Studies of the Upper Colorado River Basin

*David W. Pierce, Daniel R. Cayan, Jordan Goodrich, Tapash Das, Armin Munévar*

Division of Climate, Atmospheric Sciences, and Physical Oceanography (Pierce, Cayan), Scripps Institution of Oceanography, La Jolla, California, USA; School of Science (Goodrich), University of Waikato, Hillcrest, Hamilton, New Zealand; Water Resources and Resilience (Das, Munévar), Jacobs, San Diego, California, USA (Correspondence to Pierce: dpierce@ucsd.edu).

**Research Impact Statement**: The latest CMIP6 generation of climate models still have biases in the Upper Colorado River Basin but show clear improvements over previous generations after a simple bias correction is performed.

**ABSTRACT:** Three generations of Global Climate Models (GCMs), CMIP3, CMIP5, and CMIP6, are evaluated for performance simulating seasonal mean and annual-to-decadal variability of temperature and precipitation in the Upper Colorado River Basin. Low-frequency precipitation variability associated with drought is a particular focus and found to be a significant model shortcoming. The evaluation includes remote teleconnected atmospheric responses to the Pacific Ocean, including the El Niño/Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO). GCMs have improved their simulation of the Upper Basin over model

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.xxxx/JAWR.12974

This article is protected by copyright. All rights reserved

generations, but primarily in atmospheric circulation metrics. Persistent winter precipitation biases have changed little, including in multiyear precipitation variability. Users generally bias-correct GCM data before use; evaluation using a simple spatially and temporally averaged bias correction shows that the CMIP6 models outperform earlier generations after the bias correction, although more complex precipitation biases remain even after the simple bias correction. These model rankings will be useful when selecting GCMs for a variety of hydrological and ecological climate studies in the Upper Basin.

(KEYWORDS: Global Climate Models; Colorado River; Upper Colorado River Basin; model evaluation; winter precipitation bias, regional climate model evaluation)

# INTRODUCTION

The Upper Colorado River Basin (UCRB hereafter) drains a multi-state area in the Southwestern United States, stretching from southwest Wyoming through western Colorado and eastern Utah into portions of northern Arizona and New Mexico. The UCRB is a vital source of water in this largely arid region, supplying water to nearly 40 million inhabitants, irrigation to 5.5 million acres of farmland, and water flow to numerous wildlife refuges and national parks (USBR, 2012). The Colorado River also is an important source of hydropower, capable of supplying 4,200 megawatts of electricity generation to the region (USBR, 2012). Simulating multi-year precipitation variability, drought, and future climate changes in the UCRB is therefore of substantial societal and economic importance.

A significant amount of research has examined how the UCRB's annual discharge, typically measured at Lees Ferry, might respond to warming, either historical or projected future changes (e.g., McCabe et al. 2017; Udall and Overpeck 2017; Xiao et al. 2018; Hoerling et al. 2019). The propensity for drought and long-term reliability of the water supply are other important concerns. Different methods have been used to examine these questions. For example, some approaches use estimates of the Upper Basin's flow sensitivity to regional temperature and precipitation variations, then apply projected climate changes to estimate the Basin's response (e.g., Barnett and Pierce, 2008; Rajagopalan et al. 2009; Vano et al. 2012; for a review see Vano et al. 2014). This approach can use runoff sensitivity estimates from observational studies or land surface models to consider possible future water shortfalls (e.g., Bennett et al. 2018), and may use future temperature and precipitation trends indicted by global climate model (GCMs)

This article is protected by copyright. All rights reserved

projections (e.g., Dettinger et al. 2015). However, many regional impact studies use data from one or more GCM projections as the basis for analysis, with the GCMs often being statistically or dynamically downscaled to the Upper Basin to better capture important details of the regional topography (e.g., Barnett et al. 2004; Christensen et al. 2004; Christensen and Lettenmaier, 2007; Cayan et al. 2010; Dawadi and Ahmad, 2012; Seager et al. 2012; Ficklin et al. 2013; Tillman et al. 2017). In such cases it is best to select GCMs that do a credible job of simulating the historical climate and its variability in the Upper Basin, as poorly performing GCMs could misrepresent processes that are important to how Upper Basin drought and discharge could change in the future.

The purpose of this work is to evaluate the performance of GCMs in reproducing the historical mean climate and variability of temperature and precipitation in the Upper Basin, with an emphasis on studies of hydrology and water management in the region. Climate measures in the immediate region of the Upper Basin are examined as well as remote teleconnected signals associated with Upper Basin climate fluctuations, such as those originating from the tropical Pacific Ocean through the El Nino/Southern Oscillation (ENSO). Seasonal, annual, and multi-year timescales are considered, as they all have important implications for regional ecosystems and the existing water management infrastructure.

There has been substantial previous work on the evaluation of GCMs using a wide variety of metrics for both global and regional applications, although few that have focused specifically on the Upper Basin (cf. Tamaddun et al. 2019). For example, Gleckler et al. 2008 evaluated global measures of performance using GCMs from the Coupled Model Intercomparison Project version 3 (CMIP3) archive. A similar global analysis for the subsequent generation of GCMs, CMIP5, appears in Flato et al. 2013. These global evaluations consider such aspects as the Earth's radiation fields, surface precipitation and temperature, and winds, pressures, and temperatures at key vertical pressure levels in the atmosphere. Pierce et al. 2009 performed a similar evaluation using the CMIP3 models but focusing on the western United States; the procedure used here is based on the approach developed in that work. Rupp et al. (2013) performed a CMIP5 GCM evaluation for the Pacific Northwest, and subsequently for the Southwest United States as reported by California Department of Water Resources (CA DWR 2015). Knutti et al. (2017) examined how GCMs can be weighted to take into account model

This article is protected by copyright. All rights reserved

quality scores, such as developed here, when analyzing a multi-model ensemble. Lorenz et al. explore weighted multimodel ensemble predictions of summer maximum temperature over North America, and Brunner et al. (2019) examine temperature and precipitation projections over Europe using a model weighting scheme that incorporates model performance and independence.

The evaluation shown here differs from those previous efforts in several key areas. First, it is focused on the UCRB specifically. Second, the analysis includes both the older-generation CMIP3 and CMIP5 models as well as the newer CMIP6 models. Third, we compare the performance of different variables in an absolute sense, an effort that previously has generally been avoided in favor of relative measures of performance. This point will be explained in more detail below.

Since one of our key purposes is to evaluate hydroclimatic features, we include evaluations of multi-year precipitation variability important to drought processes. Numerous other studies that have examined such processes in various regions, such as Rupp et al. 2013 and Abatzoglou and Rupp 2017 in the Pacific Northwest, and Moon et al. (2018) and Ukkola et al. (2018) for global evaluations. Global climate teleconnections are also included since they are of first-order importance to climate variability in the region. Such teleconnected responses to the western U.S. have previously been considered by Pierce et al. (2009) and Rupp et al. (2013), for example.

One question we examine is whether the performance and quality metrics indicate that previous generation models (such as from the CMIP3 archive) should be discarded from consideration. This question is relevant because the CMIP3 models show, on average, drier future conditions in the UCRB than the more recent models (Ficklin et al. 2015). Either arbitrarily excluding or unjustifiably including the CMIP3 models could bias understanding of future drought in the Upper Basin. Evaluations of model skill improvement over the CMIP generations that examine global measures, rather than the UCRB-specific metrics considered here, can be found in Bock et al. (2020) and Fasullo et al. (2020).

The current work focuses on the GCMs’ representation of temperature and precipitation in the UCRB and teleconnected responses to the tropical and North Pacific. Land surface models (LSMs) are a key part of GCMs and have evolved considerably over the model generations. Because of their importance, LSMs and their responses to climate change in the CMIP models

This article is protected by copyright. All rights reserved

have been examined in their own right (e.g., Boone et al. 2009; Dirmeyer et al. 2013, Li et al. 2018; Li et al. 2021). LSM fields such as runoff and streamflow are not examined here but were addressed as part of this project and will be reported at a later date.

Results from this analysis can be used to inform model selection for a variety of climate impact studies in the Upper Basin. Although our focus is on hydrology and drought, ecosystems are also strongly affected by local temperature and precipitation so GCM selection is important for ecological application as well. Likewise, human health and regional energy demand will be impacted by future temperature changes, so GCM-based studies in those fields could employ the evaluation developed here.

# DATA AND METHODS

### *Variable Selection*

We obtained GCMs data from three generations of GCMs from the Climate Model Intercomparison Project (CMIP), referred to as CMIP3 (Meehl et al., 2007), CMIP5 (Taylor et al., 2012), and CMIP6 (Eyring et al. 2016). GCMs produce a wide variety of variables describing the state of the atmosphere, ocean, land surface, and cryosphere, although it is not feasible to save all variables in the CMIP archives and variable coverage is smaller in the earlier CMIP generations. The current work focuses on the GCMs' performance in temperature, precipitation, and teleconnections associated with temperature and precipitation. The relatively coarse resolution of CMIP GCMs yield a poor simulation of land surface processes such as snowpack and soil moisture in a topographically diverse region such as the UCRB. Dynamically or statistically downscaled data, not examined here, is generally better suited to examining such surface fields in a geographically limited, rugged region.

### *Global Climate Models*

We evaluate data from 82 GCMs: 16 CMIP3, 35 CMIP5, and 31 CMIP6 models, as shown in Table 1. The last column shows the approximate spatial resolution of the model's atmospheric data files as they appear in the CMIP archive. The North American CMIP3 and CMIP5 data were obtained from the U.S. Bureau of Reclamation (Reclamation hereafter) archive of climate model output available from Lawrence Livermore National Laboratory's Green Data Oasis archive (https://gdo-dcp.ucllnl.org/downscaled_cmip_projections/dcpInterface.html).

This article is protected by copyright. All rights reserved

Additional CMIP3 and CMIP5 data, and all the CMIP6 data, were downloaded from the Earth System Grid (e.g., https://esgf-node.llnl.gov/search/cmip6/) in mid-to-late 2020. We only include models that provide daily fields of minimum and maximum temperature (Tmin and Tmax) and precipitation, required for hydrological modeling work not described here. Additionally, we only include models that have data for both a historical and future climate change simulation. Several CMIP6 models have historical data available but no future shared socioeconomic pathway (SSP; Raihi et al. 2017) simulation, and so were excluded. Other models lack daily data over the historical or future period, and likewise were not analyzed (in particular, at the time of writing the CESM2 family of models do not provide daily Tmin/Tmax over the historical period).

The CMIP3, CMIP5, and CMIP6 generations used different historical periods, ending in 1999, 2005, and 2014, respectively. As a compromise between excluding recent data and using a different analysis period for all 3 generations, we used a historical period of 1950-1999 for CMIP3 and 1950-2005 for the CMIP5 and CMIP6 models. Monthly mean daily average temperature (Tavg) was formed as the mean of monthly averaged Tmin and Tmax and is the temperature quantity analyzed here.

The ensemble members used are shown in Table 1. Sea level pressure from only the first realization was available for the CMIP3 models. Only historical realizations are shown in the table since the model/observations comparison only uses data over the historical period. Each ensemble member was evaluated on all metrics, and then the final metric for each model was taken as the mean of the values for all the ensemble members. This approach prevents models with many ensemble members from having undue influence on the results. Additionally, the spread across the ensemble members was used to quantify uncertainty.

[TABLE 1 GOES HERE]

Data in Reclamation's archive had been re-gridded to a common 2-by-2-degree latitude-longitude grid for the CMIP3 models and a common 1-by-1-degree grid for the CMIP5 models. To examine the models on the same grid and explore the effect of spatial resolution on our results, we interpolated the CMIP6 and CMIP3 models to the same 1-by-1-degree grid via bilinear interpolation and aggregated the CMIP5 and CMIP6 data to the 2-by-2-degree grid. We found that whether the 1x1 or 2-by-2-degree grid is used makes only a minor difference in the

This article is protected by copyright. All rights reserved

final ranked model quality results, so most of the results here will be shown using the 1x1-degree gridded data.

*Observations*

Daily temperature and precipitation over North America were obtained from Livneh et al. 2015 (Livneh hereafter), a gridded product based on airport and cooperative weather stations. The data cover central Mexico through southern Canada at a 16th-degree latitude-longitude resolution over the period 1950-2013, which was trimmed to 1950-2005 to match the CMIP5 and CMIP6 model historical periods. Values were aggregated to the same common 1x1 and 2x2 degree grids as the models. Daily minimum and maximum temperature were averaged to produce daily average temperature, then averaged to monthly values to match the GCM data. Massmann (2020) shows that Livneh does well in representing temperature and precipitation across the CONUS for the purposes of hydrological modeling. Pierce et al. (2021) find that Livneh precipitation extremes on a daily timescale are distorted by the data processing methodology, but this does not affect the monthly-averaged analysis performed here.

For global observations of monthly sea level pressure (SLP) and temperature we used the ERA5 reanalysis (Hersbach et al. 2018) over the period of 1950-2005. Although historical station-based estimates of monthly temperature exist they are not spatially complete, so the reanalysis data was used in preference. In comparisons with the older NCEP reanalysis product (Kalney et al., 1996), some minor differences in model ranking were found in the metrics sensitive to global SLP when using the ERA5 vs. NCEP reanalysis. This shows that observational uncertainty can affect model ranking, but this aspect of uncertainty is not explored in the current work (cf. Lorenz et al. 2018).

## CULLING OF GCMs BASED ON GLOBAL METRICS

GCMs have a wide range of performance, and it is not consistent which model performs best on which metric (e.g., Gleckler et al., 2008). However, some models perform systematically worse than the other GCMs across a range of global metrics. To alleviate the concern that a highly targeted, Upper Basin-centric analysis might select models that do well in this small region but poorly in simulating the overall Earth's climate, an initial culling was performed using published hemispheric to global scale metrics to eliminate the bottom-performing 25 percent of

This article is protected by copyright. All rights reserved

models. This was done separately for the CMIP3, CMIP5, and CMIP6 GCMs. Most of the figures in the main text use this culled set of models. For key figures (called out below), the supplementary information contains figures made using the full, un-culled set of models for comparison.

The CMIP3 culling was based on Gleckler et al. (2008), specifically the model performance in the Northern Hemisphere extra-tropics (their Figure 3d). This resulted in the elimination of the following four models: ipsl_cm4, giss_model_e_r, ncar_pcm, and imncm3_0. The CMIP5 culling was based on Flato et al. (2013), which eliminated the following nine models: GISS-E2-R, IPSL-CM5A-MR, inmcm4, FGOALS-g2, bcc-csm1-1-m, MIROC-ESM-CHEM, MIROC-ESM, IPSL-CM5A-LR, and IPSL-CM5B-LR. The CMIP6 culling was based on three sources: Brunner et al. 2020, who evaluated the GCMs for performance and independence from each other; Tokarska et al. 2020, who used emergent constraints to identify models that have historical warming inconsistent with observations; and the online analysis by the Program for Climate Model Diagnosis and Intercomparison (PCMDI), available at https://cmec.llnl.gov/results/physical.html (accessed Jan 27, 2021). A subjective evaluation of the results of those three studies resulted in the following 8 models being eliminated: CanESM5, HadGEM3-GC31-LL, NESM3, UKESM1-0-LL, FGOALS-g3, INM-CM4-8, NorCPM1, and BCC-ESM1. After the culling, 61 GCMs remained for the subsequent analysis (12 CMIP3, 26 CMIP5, and 23 CMIP6). Several Supporting Information figures show key results for the entire, un-culled set of models for interested readers.

As mentioned previously, the CMIP3 model projections tend to show drier end-of-century conditions in the Upper Basin than the CMIP5 and CMIP6 models. Does the culling, which is based only on historical data, affect this outcome? This is examined in Figure 1, which shows histograms of GCM-projected temperature (red) and precipitation (green) changes in the Upper Basin, both before (top row) and after (bottom row) the global culling. Models with multiple ensemble members use the ensemble-mean result, so that different models are weighted equally in the figure even if they have different numbers of ensemble members. The distribution of projected temperature increases becomes narrower after the global culling, with both the greatest and least warming models culled. The central peak also becomes notably more pronounced. The precipitation distribution, by contrast, is less affected although some of the

This article is protected by copyright. All rights reserved

extreme wet models are culled. A similar pattern is seen in the seasonal results (Supporting Information, Figure S1), with the culling affecting the standard deviation of the projected temperature change considerably more than the projected precipitation trend. The seasonal results also show the most warming in summer as the surface dries, while the greatest precipitation trend is in winter. Natural variability affects the spread of projected trends, especially in a region as small (in a global sense) as the UCRB. Future work using models with large-ensemble simulations could help distinguish between spread based on natural variability and spread due to differences between models.

[FIGURE 1 GOES HERE]

## Model Metrics

Metrics were developed to evaluate model historical performance in simulating regional precipitation and temperature characteristics and teleconnections of Pacific surface temperature and large scale atmospheric sea level pressure that relate to Upper Basin precipitation variations. The metrics used for these evaluations are based on spatial fields of z-scores, i.e., a spatial field of differences between the model and observations normalized by a measure of variation in the observed value. The normalization allows for model-observed differences to be sensibly evaluated—are they large or small compared to observed variability?

Specifically, for seasonal-mean quantities, such as DJF seasonal mean precipitation, the z-scores comparing model to observations are calculated as follows:

$$z_{score}(x,y) = \frac{( < model\,(x,y,t) > - < obs(x,y,t) > )}{\mathrm{stddev}(obs(x,y,t))} \qquad \text{Eq. 1}$$

Where the angle brackets <> indicate averaging the time sequence of seasonal means over time, stddev() indicates the standard deviation over time, and *model(x,y,t)* and *obs(x,y,t)* are the time series of seasonal mean data in the model and observations, respectively. This approach non-dimensionalizes the errors, so that measures with different units (e.g., temperature and precipitation measures) can be sensibly compared.

The z-score is calculated at each point in the domain, yielding a 2-dimensional map of z-scores. The final overall skill score ss, or metric value, is then calculated as:

$$ss = 1 - \mathrm{RMS}(z_{score})$$

This article is protected by copyright. All rights reserved

where RMS indicates the root mean square spatial average over the domain. We do not include area weighting of the grid cells in this work because of the limited domain size, but it could be included if analyzing a more extensive region. This formulation follows the traditional skill score scaling with 1 being a perfect skill, and more negative values indicating less agreement with the observations, measured by the yardstick of observational variability. A skill score of zero means that the model and observations have an RMS mean difference of 1 standard deviation over the domain.

As a detailed example, consider JJA mean precipitation over the region evaluated for a CMIP5 GCM. There are 56 seasons of observations for this quantity (since the historical period is 1950 to 2005). At each point in the domain, the model and observed mean are calculated over the 56 seasons. The sample standard deviation of the observations is then calculated at each point from the 56 seasonal values. The difference between the model and observed mean at each point, divided by the sample standard deviation at that point, yields the z-score at that point. The final skill score is 1 minus the RMS of the z-scores over the domain.

Measures of variability require a slightly modified approach to calculate the observed standard deviation. For example, consider the metric of standard deviation of winter (December, January, February [DJF]) precipitation averaged into 10-year blocks, which has a direct relationship to drought. The monthly data are first averaged over consecutive, non-overlapping, 10-year blocks and the standard deviation computed at each point. This value is easily obtained from the observations and models, but then the denominator in the z-score needs to be calculated, i.e., what is the typical spread in the estimate of the standard deviation of 10-year blocks of precipitation? This was estimated using a block bootstrap method where 100 random, 50-year long sequences of 10-year blocks were constructed (with replacement) from the actual sequence of years, with the 10 years in each block being sequential calendar years. The sample standard deviation of 10-year blocks is computed for each random sequence, and the distribution of standard deviations examined. We use a similar bootstrap method to estimate the spread in observed quantities whenever that quantity was needed and not available by direct computation. Note that some of the metrics relating to teleconnected climate responses (for example, from Pacific Ocean sea surface temperatures to UCRB precipitation) are multivariate. In these cases,

This article is protected by copyright. All rights reserved

for each random trial the same randomly-determined temporal ordering was used for all relevant variables to preserve the temporal coherence of the fields.

With only a limited time span of data available to analyze (56 years), the sampling errors of the low frequency (5- and 10-yr averaged) metrics are higher than the seasonal metrics. This reduces the reliability of the low-frequency metrics compared to the seasonal metrics. The current analysis does not attempt to down-weight the low frequency metrics to account for their reduced reliability, although such an approach could be useful. For example, Rupp et al. (2013) discussed the issue of metric reliability and accomplished this by effectively setting the weight of unreliable metrics to zero. Metric reliability and how to combine it with the other measures of metric and model ranking uncertainty shown below are not considered here, although it would be a useful direction for further research.

The use of z-scores as the basis for our metrics is ultimately why we can compare dissimilar variables, allowing us to do absolute comparisons of the error in different metrics rather than only relative errors. For example, the question of whether a model does a better job simulating mean winter precipitation or summer temperature variability can be quantitatively answered by noting that (as a hypothetical example) a model's mean winter precipitation field may be, on average, 2 standard deviations away from the observed value, while the summer temperature variability errs by only 1 standard deviation. It is in this sense that we can say that the simulation of the precipitation field is worse than the simulation of temperature variability.

The absolute approach to metrics developed here is a departure from most previous efforts at evaluating model quality. More commonly, relative measures of error are used, such that error measures in different variables are normalized to have the same range. However, the relative approach has the disadvantage that a metric with a wide range of values, spanning a range from a very good to a very poor simulation, has as much influence on the model ranking as a metric that is well simulated by all models. Therefore, the relative approach discards useful information that could better discriminate between models.

Uncertainty in estimating metrics from the limited available time period of observations can contribute to a large sample standard deviation in the observations, yielding a lower z-score (all else being equal). Therefore, a smaller error in a well-known or stable quantity can

This article is protected by copyright. All rights reserved

potentially yield a poorer (larger) z-score than a larger error in a poorly known or highly variable quantity.

### *Seasonal Means and Variability in the Upper Basin*

Of fundamental importance to the evaluation performed here is the ability of the GCMs to simulate the mean climate and variability in the Upper Basin. Accordingly, 32 metrics are used to evaluate the seasonal (DJF, March, April, May [MAM], JJA, September, October, November [SON]) mean temperature and precipitation in the region, and the standard deviation of those seasonal quantities evaluated in 1-, 5-, and 10-year blocks. These values are compared to observations at each point in the domain, which is shown in Figure 2.

[FIGURE 2 GOES HERE]

An example of the metric fields for winter (DJF) precipitation is shown in Figure 3. The upper left shows the observed winter precipitation, in millimeters per day, while the lower left shows the standard deviation of the observations, used to form the z-score. The middle column shows a model that does well on this measure, CanESM2, with the precipitation field in the top row and the z-score in the bottom row. By contrast, the right column shows a model that does poorly on this measure, FIO-ESM. In this and the following examples (including in the Supporting Information) we do not always select the "best" and "worst" models to show, which indeed our results indicate are subject to uncertainty, but rather show results from a variety of well- and poorly-performing models rather than repeating the same models. Although even CanESM2 does not capture much of the spatial pattern, at least the values are reasonably close to the observations. FIO-ESM, on the other hand, is like many other GCMs in simulating far more winter precipitation than observed. The occurrence of relatively large errors in mean precipitation across many of the GCMs is likely due to the poor representation of topography. An analogous example for summer temperature variability is shown in the Supporting Information, Figure S2.

[FIGURE 3 GOES HERE]

### *Amplitude and Phase of the Seasonal Cycle*

The amplitude and phase of the seasonal cycle are calculated for monthly temperature and precipitation, yielding another four metrics. These are calculated from the best-fit annual

This article is protected by copyright. All rights reserved

sinusoid. However, it should be noted that precipitation in the Upper Basin has a relatively weak seasonal cycle. The seasonal cycle of precipitation is a useful metric even though the observed seasonal cycle is weak since a model that had a pronounced seasonal cycle (unlike the observations) would yield a poor simulation of the region. The observed spread in the amplitude and phase estimates were formed by the bootstrap method described above.

### *El Nino/Southern Oscillation and the Pacific Decadal Oscillation*

Various modes of climate variability are associated with temperature and precipitation variations in the Upper Basin, notably the El Nino/Southern Oscillation (ENSO; e.g., Hidalgo and Dracup, 2003) and the Pacific Decadal Oscillation (PDO; e.g., McCabe and Wolock 2020). The fidelity of the models' depiction of these modes was evaluated in three aspects: their mean expression in sea surface temperature (SST) anomalies in the Pacific Ocean; the variance spectrum of the SST pattern; and the associated response in temperature and precipitation over the western United States. SST is computed from near-surface air temperature by including air temperature in ocean regions only and limiting the lowest allowable temperature to the freezing point of seawater, −1.8 °C. Lower values generally indicate the presence of sea ice. We chose this route rather than trying to download ocean model SSTs directly because of the substantial gain in efficiency gained by only downloading one set of temperature files, combined with fact that GCMs have very similar SST and 2-meter temperature fields when evaluated on a common 1x1 degree grid.

The observed and model ENSO indices are taken as the principal component associated with the leading empirical orthogonal function (EOF) of monthly SST anomalies over the region 135 E to 80 W, 10 S to 10 N. The EOF is taken in preference to a construction based on the so-called Nino regions, such as Nino 3.4 (170 W to 120 W, 5 S to 5 N), because models do not necessarily capture the correct spatial pattern of ENSO. Using the EOF means that the model's own representation of ENSO SST patterns is used as the basis of that model's ENSO index. The observed SST pattern, as well as examples from an ensemble member of models that have a relatively good (CESM1-CAM5) and bad (CSIRO-Mk3-6-0) simulations of the ENSO pattern are shown in Figure 4. Both models extend the variability too far to the west, a common problem with GCMs, but CSIRO-Mk3-6-0 does this much more than CESM1-CAM5. The z-scores for

This article is protected by copyright. All rights reserved

CSIRO-Mk3-6-0 in the far western tropical Pacific exceed 5 standard deviations, compared to < 3 for CESM1-CAM5.

[FIGURE 4 GOES HERE]

The variance spectrum of the principal component associated with the leading EOF is used as a metric in addition to the SST anomaly pattern. This is a useful metric of model quality because some GCMs simulate a very regular 2-year ENSO cycle that is unlike the observed irregular cycle with enhanced variability at periods between 2 and 7 years. Rather than a spatial pattern of differences between the model and observed value, as done with the measures of model quality described previously, the difference between the logarithms of the model and observed power at frequencies between 2 and 7 years per cycle is computed. The logarithms are used so that the model having twice the power as the observations gives the same error as the observations having twice the power as the model.

The teleconnected precipitation and temperature response associated with the SST pattern is taken over the entire North American domain west of 105 W (25.5 N to 52.5 N, 150 W to the coast). This broader domain was used in preference to only the Upper Basin domain because inspection showed significant structure in the teleconnected fields over this wider region, while the Upper Basin tends to straddle the zero line of the response. These teleconnections are evaluated over the cold season only (ONDJFM), when the teleconnected precipitation and temperature signal to North America is strongest. Teleconnections have ramifications over the warm season as well but are poorly simulated during this season in current models (Jong et al. 2021). The teleconnected response pattern over the Upper Basin is determined by linear regression between the leading principal component and the response field of interest over the Upper Basin (precipitation or temperature). This approach assumes that the teleconnected response is linear in the associated SST pattern (e.g., the El Nino response is the opposite of the La Nina response), which is likely untrue, but our attempt at using composites to capture this non-linearity resulted in a low signal to noise ratio due to splitting the data into three pieces. An example for the teleconnected response in precipitation to ENSO variability is shown in the Supporting Information, Figure S3.

The PDO is evaluated in the same way as ENSO, with the leading EOF of SST anomalies taken as the PDO index, but in this case the domain is 145 E to 110 W, 20 N to 55 N. The

This article is protected by copyright. All rights reserved

observed SST pattern and examples of well and poorly performing models are shown in the Supporting Information, Figure S4. An example for the teleconnected response in precipitation to PDO variability is shown in the Supporting Information, Figure S5.

Each of the climate modes (ENSO, PDO) contributes four metrics (mean SST pattern, spectrum, teleconnected cold season response over the western United States in temperature and precipitation), for a total of eight metrics.

*Remote Correlations with Upper Basin Precipitation*

There are two ways to evaluate the connection of Upper Basin climate variability to wider hemispheric or global fluctuations. One method, described in the last section, is to examine the effect of known climate modes of variability (such as ENSO and the PDO) on the Upper Basin. The other way is to start with precipitation fluctuations in the Upper Basin and examine how other fields correlate with those fluctuations.

The latter approach was implemented by forming the time series of cold (ONDJFM) and warm (AMJJAS) season precipitation in the Upper Basin, then correlating those time series with temperature and sea level pressure fluctuations elsewhere around the globe. Examination of the observed correlation maps suggested that a suitable domain to evaluate the correspondence between model and observations is 100 E to 60 W, 10 S to 60 N. The variability was evaluated by the bootstrap method. This field is an example where relatively large sampling variability with respect to the signal leads to low z-scores with a comparatively weak ability to distinguish between models. Examples of the patterns for well and poorly performing models are shown in the Supporting Information, Figures S6 through S9.

## Results

Before describing the results, we emphasize a few key points on interpreting skill scores. 1) There is no absolute guide as to which metrics to pick to describe diverse aspects of the climate system. This must be guided by experience with the study domain and the aspects of climate relevant to the problem of interest. 2) Skill score values are 1 minus the RMS average z-score of the model error, averaged over the Upper Basin domain. Loosely, positive skill scores indicate that the model biases are no larger than typical fluctuations due to natural variability, while negative skill scores mean that biases are larger than typical variability.

This article is protected by copyright. All rights reserved

The overall portrait plot of model skill scores for each metric is shown in Figure 5. A summary model skill score across all the metrics, indicated by the number in the parenthesis after the model name, is constructed as the Euclidian distance between the perfect model skill point (1, 1, 1, …, 1) and the model's skill scores in all the metrics. Lower values therefore signify better models. This is referred to as Dss, signifying the distance in skill score space. The models are ordered in Figure 5 such at that the best models are at the bottom of the plot (smallest distance to the perfect skill point, therefore lowest Dss), and the worst models at the top of the plot (largest Dss). For models with more than one ensemble member, the ensemble mean value is shown. Uncertainty in the model rankings estimated by spread across the ensemble members will be shown below (Figure 8 and Figure 11).

It was previously noted that metrics with large observational uncertainty (particularly the poorly sampled 10-year average metrics) give a large denominator in Eq. 1, yielding skill scores near zero. In other words, if the observational uncertainty is large, it cannot be definitively concluded that model results are inconsistent with the observations, leading to skill scores that are near zero. By contrast, it can be easier to conclude that model-observational differences are large in well-observed quantities with low observational uncertainty, leading to negative skill scores (i.e., it is known that the models are inconsistent with the observations). This can be seen in Figure 5, where the skill scores in 1-yr variability tend to be more negative than the skill scores in the 5- and 10-yr averages but is an outcome of larger uncertainty in the poorly sampled low-frequency metrics. For example, Abatzoglou and Rupp (2017) found that GCM fidelity was generally lower-on multi-year timescales than seasonal or annual timescales when evaluating CMIP5 GCM simulations of drought in the Pacific Northwest.

An analogous figure including all models (no culling) is given in the Supporting Information, Figure S10. The culling eliminates some models that would otherwise score well in the UCRB. For example, HadGEM3-GC31-LL is culled on the basis of poor global performance in the evaluation of Brunner et al. (2020), their Figure 4. This finding indicates the importance of considering global metrics even for regional GCM applications, as models that perform poorly on global metrics may be doing well in the UCRB but for the wrong physical reasons.

[FIGURE 5 GOES HERE]

This article is protected by copyright. All rights reserved

One striking aspect of Figure 5 is that some metrics have consistently low skill across many models, seen as dark blue vertical columns, particularly metrics associated with winter precipitation variability. Somewhat ironically, the models that stand out for being much better than normal on the winter precipitation variability metric, EC-Earth3 and EC-Earth3-Veg, do unusually poorly on winter and spring temperature variability. The uneven range of skill score variability is illustrated in Figure 6, which shows the distribution of metric values, sorted by the mean metric value, with better simulated metrics having higher means (closer to the perfect value of 1).

[FIGURE 6 GOES HERE]

The best simulated metric is the annual phase of Upper Basin temperature (tas phase), which is not surprising since the phase is largely controlled by solar insolation and the tilt of the earth's surface with respect to the sun, quantities that are specified in the models. Of the 8 worst metrics, 6 are associated with precipitation variability. Winter precipitation variability on the 1-yr time scale is by far the worst simulated quantity, with a mean metric value less than -5. This likely is influenced by poor model treatment of topography in the Upper Basin. As noted earlier, some metrics with a relatively large spread in the observations, such as the estimated spectral power in ENSO and the PDO, do relatively well in the sense that the model values cannot be shown to be outside the wide range of uncertainty.

*Estimating Uncertainty using Ensemble Members*

Models with multiple ensemble members can be used to explore how sampling and model-simulated natural climate variability affect the metric scores and overall model ranking. Supporting Information Figure S11 shows the estimated standard deviation for each metric, ranked from most to least certain. This was estimated in two ways: 1) by calculating the standard deviation of each metric from every model with at least three ensemble members, then taking the mean of the model estimates as the final standard deviation (referred to as the mean of the model values); 2) by forming, for each metric, the anomalies of each model's skill scores with respect to that model's mean skill score, then pooling anomalies from all the models and calculating the standard deviation of the result (referred to as the pooled method). The difference between these approaches is minor.

This article is protected by copyright. All rights reserved

There are some similarities between the uncertainty in each metric and the mean value of each metric (Figure 6). For example, the phase of the annual cycle of temperature in the Upper Basin (tas phase) is both the best-simulated and least uncertain metric, while the precipitation variability metrics tend to be the worst simulated and most uncertain. However, there are interesting differences as well. For example, the precipitation variability metrics tend to be the least well simulated (Figure 6), but the uncertainties (Figure S11) are substantially influenced by the averaging period in the metric, with longer averaging periods yielding fewer independent samples and more uncertainty.

The spread of values in the overall metric skill score, Dss, for each model with at least 3 ensemble members is shown in the Supporting Information, Figure S12. Individual models exhibit a range of spreads, including a standard deviation of 2.67 for CNRM-CM5 (n=5 ensemble members), and 0.28 for cccma_cgcm3_1 (n=5). A Monte-Carlo simulation indicates that this nearly order-of-magnitude discrepancy would happen only about 2.5% of the time by chance under the null hypothesis that all the models have the same standard deviation of Dss values.

Later figures that display uncertainty in the model's Dss scores are based on this analysis. The uncertainty in models that have less than 3 ensemble members is estimated as the multi-model mean from models with at least three ensemble members. Given that the evidence suggests different models have different levels of variability, this should be considered a rough estimate.

### *Redundancy in the Metrics*

The skill scores presented up to now have been exhaustive, often measuring similar aspects of model performance (for example, the variability of precipitation averaged into 1-, 5-, and 10-year blocks). This redundancy of information can be addressed by forming the EOFs of the skill score matrix (Pierce et al. 2009; Rupp et al. 2013). Computing the EOFs forms optimal combinations of metrics that best describe the model variability, taking covariability between the metrics into account. The number of EOFs to retain is usually chosen so that only modes above the noise floor are kept (Wilks, 2011). Here we choose six modes, which account for 89.1% of the variance.

This article is protected by copyright. All rights reserved

The leading two EOFs (which describe the weighting of each metric) and associated principal components (PCs, which describe the weighting of each model) of the skill score matrix are shown in Figure 7. The EOF weightings show that the leading mode describes poor model performance in simulating seasonal precipitation variability (large negative peaks for DJF, MAM, and JJA precipitation standard deviation), and a large number of models have this problem, led by GFDL-ESM2M and FIO-ESM. Unsurprisingly, these models do not do well in the metric evaluation (Figure 5). The second mode shows co-varying behavior in the quality of model simulations of summer precipitation and temperature variability, and winter precipitation variability. The associated PC shows that 4 models express this behavior strongly: GFDL-ESM2M, GFDL-ESM2G, gfdl_cm2_1, and gfdl_cm2_0. Since this represents two model generations from the same institution, the PC suggests a common physical parameterization or coding approach gives rise to this behavior (c.f. Knutti et al., 2013). However, the CMIP6 models from GFDL (GFDL-ESM4 and GFDL-CM4) do not express this relationship, suggesting that a recent change in the model physics or microphysics has altered this behavior.

[FIGURE 7 GOES HERE]

The overall model rankings (Dss) after the EOF processing are shown in Figure 8. Given the uncertainty in Dss calculated from the ensemble members (indicated by the horizontal red bars), the model with the best overall ranking, EC-EARTH, is not significantly different from any of the other 5 best models. We evaluate the significance of the difference in means using the method of Lanzante 2005, which properly accounts for the joint or pooled uncertainties when estimating the statistical significance of the difference in means of two uncertain quantities. The first model that EC-EARTH is significantly better than is cnrm_cm3, which is rank 6. The curve bends at higher Dss values, indicating that there is a broad and indistinguishable range of relatively good models, but the poorly performing models are more distinct. A version of Figure 8 with no culling is shown in Supporting Information Figure S13.

[FIGURE 8 GOES HERE]

### *The Dominance of Precipitation Errors*

The results up to now indicate that precipitation errors, especially in winter, are the most problematic aspect of Upper Basin simulations. Since the z-scores normalize by observed natural

This article is protected by copyright. All rights reserved

variability, this is not an artifact of the large variability in this quantity. Given the coarse resolution of most GCMs and the importance of topography to generating precipitation in the Upper Basin, it is not surprising that simulated precipitation in the region often has significant biases. Many applications use bias corrected precipitation in their modeling of the basin to address this. Discarding otherwise high-performing models based on a precipitation bias is questionable when the bias will be removed before the data are used.

The following two approaches were explored to address this issue:

1) Forming separate precipitation, temperature, and atmospheric circulation indices from the relevant metrics, then weighting those three indices equally to form the final model ranking. This prevents the precipitation biases from dominating the overall model quality ranking, while still allowing absolute rankings amongst the metrics in each class. We term this the “Index-3” approach since it forms an overall index made up of three equally weighted subclasses of indices (temperature, precipitation, and circulation).

2) Using a simple bias correction that removes the annual mean bias averaged over the entire Upper Basin region (i.e., a single value) before calculating the metrics. Because only a single annual value is removed for the entire region, this retains the models’ simulation of the annual cycle and spatial variability. However, it substantially reduces discrepancies between the model fields and observations.

### *Index-3*

The Index-3 method forms an index that equally weights temperature, precipitation, and circulation metrics. The temperature class includes all seasonal metrics of Upper Basin mean temperature, the standard deviation of temperature averaged into 1-, 5-, and 10-year blocks, and the amplitude and phase of the annual cycle of temperature. The precipitation class was formed similarly. The circulation class includes all metrics based on ENSO and the PDO, including the teleconnected responses of temperature and precipitation in the Upper Basin region, and the metrics based on the wider-scale correlation maps of surface temperature and sea level pressure with warm and cold season precipitation fluctuations in the Upper Basin. This equal-weighting by class (temperature, precipitation, circulation) gives increased weight to the circulation-based metrics since there are fewer circulation metrics than temperature or precipitation metrics.

This article is protected by copyright. All rights reserved

After each metric was assigned to one of the classes, the values of all metrics that fell into each class were averaged by model. This aggregation yields three quality values per model, one each for temperature, precipitation, and circulation. For each class, the range of values across all models was normalized to the range 0 (best) to 1 (worst), so that the three classes have the same range, in keeping with the purpose of this exercise. The final Index-3 value for a model is the average of the three normalized class values for that model.

The model quality scores using Index-3 are shown in the Supporting Information, Figure S14, and show a substantial rearrangement of model rankings (cf. Figure 8). For instance, four CMIP3 (black text) do relatively well in the regular rankings (low Dss), while in the Index-3 result the best CMIP3 model appears at rank 17.

Some insight into this behavior can be gained by examining changes over model generations in the overall Index-3 and the individual temperature, precipitation, and circulation indices (Figure 9). The difference between the means of pairs of CMIP distributions was evaluated by a two-sample t-test, which indicates that the CMIP generation means are significantly different for the circulation index, but not for the temperature, precipitation, or overall Index-3 indices. In other words, progress across model generations has been dominated by better depictions of large-scale atmospheric circulation, while regional biases (especially in winter precipitation) have not fared as well. Bock et al. (2020) compared a variety of global GCM fields to observations across the CMIP3, 5, and 6 generations, and generally found improvement in the representation of global surface temperature and precipitation fields. However their Figures 3 (temperature) and 4 (precipitation) show that biases across the Western U.S., the focus of interest here but a small part of their global evaluation, show little improvement across model generations. Fasullo (2020) likewise evaluated GCM simulations across CMIP3, 5, and 6, and found that some of the biggest generational improvements were found in aspects of the circulation and ENSO, generally being greater than the improvements found for climatology or on seasonal timescales.

[FIGURE 9 GOES HERE]

This article is protected by copyright. All rights reserved

### *Simple bias correction*

The index-3 approach is useful in that it elucidates that the circulation metrics have been the main improvement over model generations, but ultimately most applications bias correct model data before using it. Often, this is done as part of a downscaling process. Accordingly, metrics based on a simple bias correction scheme rather will now be explored, as it better reflects how GCM data is generally used in UCRB studies.

Exactly how to evaluate a bias corrected model is still a research question. Common bias correction approaches based on quantile mapping remap the entire model distribution to the observed distribution at every point, which would obviate any comparison with observations in terms of means or variability.

The approach taken here is to implement a very simple bias correction rather than a full quantile mapping. The intent is to eliminate the mean biases but still evaluate the model's simulation of the spatial variability and annual cycle of temperature and precipitation in the Upper Basin. To do this, the mean model bias over all spatial points and times is removed, either additively (for temperature) or multiplicatively (for precipitation). Following this simple bias correction, the metrics are recalculated and the result analyzed as shown previously.

The portrait plot of metric values after simple bias correction is shown in Figure 10. Comparing to the same result without the simple bias correction (Figure 5), it is clear that when bias correction is added there is a significant overall improvement, as might be expected. A version including the culled models is given in the Supporting Information, Figure S15.

Interestingly, some metric values become at least 1 standard deviation worse after simple bias correction (Supporting information, Figure S16). The fact that some metrics degrade might seem counterintuitive, but it happens due to offsetting errors. A model that has lower than observed mean annual precipitation will be bias corrected by multiplying the precipitation fields by a value greater than 1, so that the annual mean matches observations. If that model already has too much precipitation variability, the variability will increase even more, and the skill score will go down as a result.

[FIGURE 10 GOES HERE]

This article is protected by copyright. All rights reserved

Figure 11 shows overall model quality rankings after the simple bias correction is applied, with redundant information removed by forming the EOFs as described previously. One striking aspect of Figure 11 is the high performance of the CMIP6 models (red text), which take 10 of the top 12 places (along with 2 CMIP5 models). Before the simple bias correction, the top 12 places had 3 CMIP3 models, 4 CMIP5 models, and 5 CMIP6 models—a much more equal distribution (Figure 8). Like the Index-3 results, this again illustrates that while biases persist across the model generations, correcting with even a single number (the annual and Upper Basin regional average) reveals that the newer CMIP6 models, as a group, are clearly preferable. Indeed, Figure 11 shows a strong preponderance of CMIP6 models in the top quarter of all models. A similar plot but including all models (no culling) is given in the Supporting Information Figure S17.

[FIGURE 11 GOES HERE]

The change in model quality from simple bias correction is quantified in Figure 12. Before bias correction a two-sample t-test indicates no significant difference between the CMIP3 and CMIP6 means, but after the bias correction the difference is significant at the $p=0.01$ level. Combined with our previous finding that the Index-3 circulation index shows significantly less error in CMIP5 and CMIP6 models than CMIP3 models, this implies that the newer models still struggle with systematic biases, but as a group they do a significantly better job than the older CMIP3 generation in simulating spatial and temporal variability associated with atmospheric circulation patterns.

[FIGURE 12 GOES HERE]

# DISCUSSION

## *Spatial Resolution in the Depiction of Climate Fields*

An important component of the metrics are spatial patterns of mean temperature, precipitation, and variability, and the CMIP5 and CMIP6 models as a group have improved spatial resolution compared with the older CMIP3 generation. This raises the question of whether better results obtained from the CMIP5/6 models are simply due to a more resolved spatial depiction of the fields. We can begin by examining the effect that degrading the spatial resolution of the CMIP5/6 models to match the CMIP3 models has on the model scores. If the

This article is protected by copyright. All rights reserved

primary reason CMIP5/6 models perform better is because they do not smear the spatial fields as much as the lower-resolution CMIP3 models, then degrading the spatial resolution of the CMIP5 models might show less difference between the model generations than seen in Figure 12.

This is tested in Supporting Information Figure S18, which is the same as Figure 12 except that the 1-by-1-degree CMIP5 and CMIP6 data have been aggregated to the 2-by-2-degree grid used by the CMIP3 models. The superiority of the CMIP5 and CMIP6 models remains, and at the same level of significance. The better performance of the CMIP5/6 models in the metrics is not due exclusively to a better resolved depiction of the surface temperature and precipitation fields.

Another way to test the effects of model resolution is to stratify model performance by the spatial resolution of the model, as shown in Figure 13. The final model performance is the Dss value from Figure 11, and the spatial resolution is taken as the average of the latitudinal and longitudinal resolutions from Table 1. The relationship between Dss scores and model resolution is shown for each model generation individually (black, blue, and red least-squares best fit trend lines for CMIP3, 5, and 6, respectively), and for all models taken together (purple trend lines). Results both before (left panel) and after (right panel) the simple bias correction are shown. Only one relationship between model quality and spatial resolution is significantly at the 95% confidence interval: the *decrease* in model performance with higher resolution in CMIP3 with no bias correction (left panel, black line). Otherwise, no statistically significant relationships between final model score and the model resolution are found, either when all models are taken together or when each model generation is considered individually, although the trends for the CMIP5 models comes close. We do not argue that model resolution is immaterial to simulations of the UCRB, but these results show that differences in spatial resolution are not the major factor driving differences in performance across GCMs.

[FIGURE 13 GOES HERE]

### *Relation of Model Quality to Projected Climate Changes*

It is natural to wonder whether the better-performing models have a systematically different representation of future climate change than the worse-performing models. The regression between model quality and model-projected precipitation trend for the SSP585

This article is protected by copyright. All rights reserved

(CMIP6), RCP 8.5 (CMIP5), and SRES A2 (CMIP3) scenarios is shown in the Supporting Information, Figure S19. No consistent relation between model quality and precipitation change is found. In addition, few individual metrics were found to have any significant relationship with the projected precipitation change. No combination of metrics identified in this way explained more than 20 percent of the variability in projected precipitation change. By contrast, Rupp et al. (2017) found that better performing models showed larger positive winter precipitation projections in the Pacific Northwest. The difference here may be due to the Pacific Norwest falling in the region where GCMs more consistently predict wetter conditions, while the UCRB is close to the zero line where GCMs predict positive precipitation trends to the north and negative trends to the south.

*Relation to Global Model Evaluations*

Our model evaluation has focused on the UCRB. How do our model rankings compare to published global model rankings? Although a complete evaluation is beyond the scope of this work, some interesting features are evident from the comparison. Many models are reasonably consistent in their ranking across the evaluations, especially considering the uncertainties involved (Figure 11). For example, GFDL-ESM4, FGOALS-f3-L, MPI-ESM1-2-HR, and ACCESS-CM2 score well both here and in Brunner et al. 2020 (their Figure 4). However, there are exceptions. For example, MIROC6 is one of the best models in Fasullo (2021, their Table 1), average in Brunner et al. 2020, and one of the lowest-ranking models in this work. Conversely, IPSL-CM6A-LR does poorly in Fasullo (2021), average in Brunner et al. 2020, and well here. The existence of such models shows that both global and regional metrics should be consulted before selecting a GCM to use in a regional study. Doing well on global metrics is not sufficient to guarantee good performance in a regional setting, and doing well on the regional metrics is not sufficient to guarantee good performance on the global metrics.

*Model Genealogy*

In this work we have not considered the commonality between models due to shared code or parameterizations (e.g., Knutti et al. 2013, Brunner et al. 2020). However, this may be a consideration when selecting GCMs for applications if a diverse set of models is desired. Our purpose is to evaluate the models that met our data requirements and were available when the

This article is protected by copyright. All rights reserved

analysis was undertaken; if desired, the model rankings develop here can be used to select which one of a family of related set of models is best suited for a user's application.

# CONCLUSIONS

We have evaluated the ability of CMIP3, CMIP5, and CMIP6 GCMs to reproduce the mean and variability of climate in the Upper Colorado River Basin, including at multi-year time scales for drought applications (5- and 10-year averaging intervals) and teleconnections with the wider hemispheric region. Using a set of 48 metrics, we have ranked 62 GCMs by overall quality of their simulations of seasonal and annual temperature and precipitation, for both the mean and variability in the region. The ranking included an initial culling of the GCMs, with 25% of each generation of models discarded based on poor performance on global metrics. This reduces the chance that a model does well in the limited region of the Upper Colorado River Basin for the wrong dynamical reasons.

A key aspect of our approach is to evaluate the models after a simple bias correction has been applied. This is motivated by the fact that stakeholders and impact studies in the region generally use bias-corrected fields. However additional information was obtained from the original (non-bias corrected) GCM fields, with the main finding that the CMIP3 models do systematically worse than the CMIP5 and CMIP6 models on the metrics relating to global atmospheric circulation. The CMIP6 models also do significantly better than the CMIP3 models after the simple bias correction is applied. Nonetheless, it is worth noting that even after the simple bias correction, the GCMs show appreciable residual biases in their depiction of the climate in the Upper Basin, particularly in the interannual variability of winter precipitation. Although GCMs are currently the best tools available for projecting future climate change over broad regions, they have problems simulating relatively small regions with significant topography, such as the Upper Basin. Our results show that these biases have persisted across model generations, even as performance on metrics of atmospheric circulation has improved.

# SUPPORTING INFORMATION

Additional supporting information may be found online under the Supporting Information tab for this article: Additional figures and illustrations.

## DATA AVAILABILITY

This article is protected by copyright. All rights reserved

Metric values and model quality results are available for download at http://cirrus.ucsd.edu/~pierce/pierce_et_al_2021_UCRB_GCM_selection. This will allow individual practitioners to weight individual metrics or model results as needed for their applications.

## ACKNOWLEDGEMENTS

Funding for this work was provided by the U.S. Bureau of Reclamation through subcontract CAPDC87 from CH2M (now Jacobs), the California Energy Commission contract EPC-16-093, and the California and Nevada Applications Program award NA11OAR43101. This study contributes to DOI's Southwest Climate Adaptation Science Center activities. Hersbach, H. et al. (2018) was downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store. This work benefitted from comments by three anonymous reviewers, which are gratefully acknowledged.

## LITERATURE CITED

Abatzoglou, J. T., and D. E. Rupp. 2017. "Evaluating climate model simulations of drought for the northwestern United States." *Internat. J. of Climatol.* 37: 910-920.

Barnett, T., R. Malone, W. Pennell, D. Stammer, B. Semtner, and W. Washington. 2004. "The Effects of Climate Change on Water Resources in the West: Introduction and Overview." *Climatic Change* 62: 1-11.

Barnett, T. P., and D. W. Pierce. 2008. "Sustainable Water Deliveries from the Colorado River in a Changing Climate." *Proceedings of the National Academy of Sciences* (PNAS) 106 (18): 7334-7338. Doi/10.1073/pnas.0812762106

Bennett, K., E., J. R. Urrego Blanco, A. Jonko, T. J. Bohn, A. L. Atchley, N. M. Urban, and R. S. Middleton. 2018. "Global Sensitivity of Simulated Water Balance Indicators Under Future Climate Change in the Colorado Basin." *Water Resources Research* (WRR) 54: 132-149. https://doi.org/10.1002/2017WR020471

Bock, L, A. Lauer, M. Schlund, M. Barreiro, N. Bellouin, C. Jones, G. A. Meehl, V. Predoi, M. J. Roberts, and V. Eyring. 2020. "Quantifying Progress Across Different CMIP

This article is protected by copyright. All rights reserved

Phases with the ESMValTool." *J. Geophys. Res. Atmos*. 125, 28 pp., https://doi.org/10.1029/2019JD032321

Boone, A., P. de Rosnay, G. Balsamo, A. Beljaars, F. Chopin, et al. 2009. "The AMMA Land Surface Model intercomparison project (ALMIP)." *Bull. Am. Met. Soc.* 90 (12): 1865-1880. Doi: 10.1175/2009BAMS2786.1

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

Brunner, L., A. G. Pendergrass, F. Lehner, A. L. Merrifield, R. Lorenz, and R. Knutti. 2020. "Reduced Global Warming from CMIP6 Projections when Weighting Models by Performance and Independence." *Earth Systems Dynamics* 11: 995-1012. https://doi.org/10.5194/esd-11-995-2020

CA-DWR, California Department of Resources Climate Change Technical Advisory Group. 2015. "Perspectives and Guidance for Climate Change Analysis." https://www.water.ca.gov/LegacyFiles/climatechange/docs/2015/Perspectives_Guidance_Climate_Change_Analysis.pdf.

Cayan, D. R., T. Das, D. W. Pierce, T. P. Barnett, M. Tyree, and A. Gershunov. 2010. "Future Dryness in the Southwest US and the Hydrology of the Early 21st Century Drought." *Proceedings of the National Academy of Sciences* (PNAS) 107: 21271-21276. Doi/10.1073/pnas.0912391107

Christensen, N. S., A. W. Wood, N. Voisin, D. P. Lettenmaier, and R. N. Palmer. 2004. "The Effects of Climate Change on the Hydrology and Water Resources of the Colorado River Basin." *Climatic Change* 62: 337-363.

Christensen, N. S., and D. P. Lettenmaier. 2007. "A Multimodel Ensemble Approach to Assessment of Climate Change Impacts on the Hydrology and Water Resources of the Colorado River Basin." *Hydrology and Earth System Sciences* (HESS) 11: 1417-1434.

This article is protected by copyright. All rights reserved

Dawadi, S., and S. Ahmad. 2012. “Changing Climatic Conditions in the Colorado River Basin: Implications for Water Resources Management.” *J. Hydrology* 430-431: 127-141. Doi:10.1016/j.jhydrol.2012.02.010

Dettinger, M., B. Udall, and A. Georgakakos. 2015. “Western Water and Climate Change.” *Ecological Applications* 25 (8): 2069-2093.

Dirmeyer, P. A., Y Jin, B. Singh, and XQ Yan. 2013. “Trends in land-atmosphere interactions from CMIP5 simulations.” J. *Hydromet*. 14(3): 829-849. Doi: 10.1175/JHM-D-12-0107.1

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor. 2016. “Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization.” *Geosci. Model Dev.,* 9 (5): 1937-1958. DOI: 10.5194/gmd-9-1937-2016.

Fasullo, J. T. 2020. “Evaluating simulated climate patterns from the CMIP archives using satellite and reanalysis datasets using the Climate Model Assessment Tool (CMATv1).” *Geosci. Model Dev*. 13: 3627. https://doi.org/10.5194/gmd-13-3627-2020.

Ficklin, D. L., S. L. Letsinger, I. T. Stewart, and E. P. Maurer. 2016. “Assessing Differences in Snowmelt-Dependent Hydrologic Projections using CMIP3 and CMIP5 Climate Forcing Data for the Western United States.” *Hydrology Research* 47 (2) 483-500. Doi: 10.2166/nh.2015.101

Ficklin, D. L., I. T. Stewart, and E. P. Maurer. 2013. “Climate Change Impacts on Streamflow and Subbasin-Scale Hydrology in the Upper Colorado River Basin.” *PLOS One* 8 (8): e71297. Doi:10.137/journal.pone.0071297

Flato, G., J. Marotzke, B. Abiodun, P. Braconnot, S.C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, C. Forest, P. Gleckler, E. Guilyardi, C. Jakob, V. Kattsov, C. Reason, and M. Rummukainen. 2013. “Evaluation of climate models.” In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Edited by T.F. Stocker, D. Qin, G.-

This article is protected by copyright. All rights reserved

K. Plattner, M. Tignor, S.K. Allen, J. Doschung, A. Nauels, Y. Xia, V. Bex, and P.M. Midgley, Cambridge University Press, pp. 741-882, doi:10.1017/CBO9781107415324.020.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux. 2008. "Performance metrics for climate models." *Journal of Geophysical Research Atmospheres* 113: D06104. 10.1029/2007JD008972

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., Thépaut, J-N. 2018. "ERA5 hourly data on single levels from 1979 to present." *Copernicus Climate Change Service (C3S) Climate Data Store (CDS)*. (Accessed on 22-JAN-2021), 10.24381/cds.adbb2d47.

Hidalgo, H. G., and J. A. Dracup, 2003. "ENSO and PDO Effects on Hydroclimatic Variations of the Upper Colorado River Basin." *Journal of Hydrometeorlogy* 4: 5-23.

Hoerling, M., J. Barsugli, B. Livneh, J. Eischeid, X. Quan, and A. Badger. 2019. "Causes for the Century-Long Decline in Colorado River Flow." *Journal of Climate* 32: 8181-8203. Doi:10.1175/JCLI-D-19-0207.1

IPCC, 2013. "*Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*." Cambridge: Cambridge University Press.

Jenkins, G. M., and Watts, D. G. 1968. *Spectral Analysis and its Applications*. Emerson-Adams Press, Inc.

Jong, B-T., J. Ting, and R. Seager. 2021. "Assessing ENSO Summer Teleconnections, Impacts, and Predictability in North America." *J. Climate* 34: 3629-3643. https://doi.org/10.1175/JCLI-D-20-0761.1

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, et al. 1996. "The NCEP/NCAR 40-year Reanalysis project." *Bulletin of the American Meteorological Society* (BAMS) 77: 437-471.

Knutti, R., D. Masson, and A. Gettelman. 2013. "Climate model genealogy: Generation CMIP5 and how we got there." *Geophysical Research Letters* 40: 1194-99. doi:10.1002/grl.50256

This article is protected by copyright. All rights reserved

Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring, 2017: “A climate model projection weighting scheme accounting for performance and interdependence.” *Geophys. Res. Lett*. 44(4): 1909-1918. https://doi.org/10.1002/2016GL072012

Lanzante, J. R. 2005. “A cautionary note on the use of error bars.” Journal of Climate 18: 3699-3703. https://doi.org/10.1175/JCLI3499.1

Li, JD., CY Miao, W. Wei, G. Zhang, LJ Hua, YL Chen, and XX Wang. 2021. “Evaluation of CMIP6 Global Climate Models for Simulating Land Surface Energy and Water Fluxes During 1979-2014.” J. *Advances Modeling Earth Sys*. 13(6): art. e2021MS002515. Doi 10.1029/2021MS002515

Li, L, Y. Wang, V. K. Arora, D. Eamus, H. Shi, J. Li, L. Cheng et al. 2018. “Evaluating Global Land Surface Models in CMIP5: Analysis of Ecosystem Water- and Light-Use Efficiencies and Rainfall Partitioning.” *J. Clim*. 31(8): 2995-3008. https://doi.org/10.1175/JCLI-D-16-0177.1

Livneh, B., T. J. Bohn, D. W. Pierce, F. Munoz-Arriola, B. Nijssen, R. Vose, D. R. Cayan, and L. Brekke. 2015. “A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950-2013.” *Scientific Data* 2: article 150042 (2015). doi:10.1038/sdata.2015.42.

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

Massmann, C. 2020. “Evaluating the Suitability of Century-Long Gridded Meteorological Datasets for Hydrological Modeling.” *J. Hydrometeorology* 21(11): 2565-2580. Doi: https://doi.org/10.1175/JHM-D-19-0113.1

McCabe, G. J, D. M. Wolock, G. T. Pederson, C. A. Woodhouse, and S. Mcafee. 2017. “Evidence that Recent Warming is Reducing Upper Colorado River Flows.” *Earth Interactions* (21) 1-14.

This article is protected by copyright. All rights reserved

McCabe, G. J., and D. M. Wolock. 2020. “The Water-Year Balance of the Colorado River Basin.” *Journal of the American Water Resources Association* (JAWRA) 56 (4): 724-737. https://doi.org/10.1111/1752-1688.12848

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor. 2007. “The WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research.” *Bulletin of the American Meteorological Society* (BAMS) Sept. 2007: 1383-1394. https://doi.org /10.1175/BAMS-88-9-1383

Moon, H., Gudmundsson, L., & Seneviratne, S. I., 2018. “Drought persistence errors in global climate models.” *J. Geophys. Res. Atmospheres* 123(7): 3483-3496.

Pierce, D. W., T. P. Barnett, B. D. Santer, and P. J. Gleckler. 2009. “Selecting global climate models for regional climate change studies.” *Proceedings of the National Academy of Sciences* 106 (21): 8441-8446. doi:10.1073/pnas.0900094106

Pierce, D. W., L. Su, D. R. Cayan, M. D. Risser, B. Livneh, and D. P. Lettenmaier. 2021. “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

Rajagopalan, B., K. Nowak, J. Prairie, M. Hoerling, B. Harding, J. Barsugli, A. Ray, and B. Udall. 2009. “Water Supply Risk on the Colorado River: Can Management Mitigate?” *Water Resources Research* (WRR) 45: W08201. Doi:10.1029/2008WR007652

Riahi, K., D. P. van Vuuren, E. Kriegler, J. Edmonds, B. C. O’Neill, et al. 2017. “The Shared Socioeconomic Pathways and their Energy, Land Use, and Greenhouse Gas Emissions Implications: An Overview.” *Global Environmental Change* 42: 153-168. DOI: 10.1016/j.gloenvcha.2016.05.009

Rupp, D. W., J. T. Abatzoglou, K. C. Hegewisch, and P. W. Mote. 2013. “Evaluation of CMIP5 20th century simulations for the Pacific Northwest USA.” *Journal of Geophysical Research Atmospheres* 118: 10,884-10,906. Doi:10.1002/jgrd.50843

Rupp, D. E., J. T. Abatzoglou, and P. W. Mote. 2017. “Projections of 21st century climate of the Columbia River Basin.” *Climate Dynamics* (49), 1783-1799.

This article is protected by copyright. All rights reserved

Sanderson, B. M. and M. F. Wehner. 2017. “Model weighting strategy.” In: *Climate Science Special Report: Fourth National Climate Assessment, Volume I*, edited by D. J. Wuebbles, D.W. Fahey, K.A. Hibbard, D.J. Dokken, B.C. Stewart, and T.K. Maycock, 436-442. Washington D. C.: U.S. Global Change Research Program. doi: 10.7930/J06T0JS3.

Seager, R., M. Ting, C. Li, N. Naik, B. Cook, J. Nakamura, and H. Liu. 2012. “Projections of Declining Surface-Water Availability for the Southwestern United States.” *Nature Climate Change* 3 (5) 482-486. Doi:10.1038/NCLIMATE1787

Tamaddun, K. A., A. Kalra, S. Kumar, and S. Ahmad. 2019. “CMIP5 Models’ Ability to Capture Observed Trends under the Influence of Shifts and Persistence: An In-Depth Study on the Colorado River Basin.” *Journal of Applied Meteorology and Climatology* (56) 1677-1688. Doi: 10.1175/JAMC-D-18-0251.1

Taylor, K. E., R. J. Stouffer, and G. A. Meehl. 2012. “An Overview of CMIP5 and the Experiment Design.” *Bulletin of the American Meteorological Society* (BAMS) Apr. 2012: 485-498. DOI: 10.1175/BAMS-D-11-00094.1

Tillman, F. D., S. Gangopadhyay, and T. Pruitt. “Understanding the Past to Interpret the Future: Comparison of Simulated Groundwater Recharge in the Upper Colorado River Basin (USA) Using Observed and General-Circulation-Model Historical Climate Data.” *Hydrogeological Journal* 25: 347-358. Doi: 10.1007/s100040-016-1481-0

Tokarska, K. B., M. B. Stolpe, S. Sippel, E. M. Fischer, C. J. Smith, F. Lehner, and R. Knutti. 2020. “Past Warming Trend Constrains Future Warming in CMIP6 Models.” *Science Advances* 6 (12): eaaz9549. DOI: 10.1126/sciadv.aaz9549

Udall, B., and J. Overpeck. 2017. “The Twenty-First Century Colorado River Hot Drought and Implications for the Future.” *Water Resources Research* (WRR) 53: 2404-2418. Doi:10.1002/2016WR019638

Ukkola, A. M., Pitman, A. J., De Kauwe, M. G., Abramowitz, G., Herger, N., Evans, J. P., and Decker, M. (2018). “Evaluating CMIP5 model agreement for multiple drought metrics.” *Journal of Hydrometeorology* 19(6): 969-988.

This article is protected by copyright. All rights reserved

USBR, 2012. “Colorado River Basin Water supply and demand study: Executive summary.” U.S. Dept. Interior, Bureau of Reclamation. https://www.usbr.gov/watersmart/bsp/docs/finalreport/ColoradoRiver/CRBS_Executive_Summary_FINAL.pdf (accessed 2021-02-22).

Vano, J. A., T. Das, and D. P. Lettenmaier. 2012. “Hydrologic Sensitivities of Colorado River Runoff to Changes in Precipitation and Temperature.” *Journal of Hydrometeorology* 13: 932-949. Doi: 10.1175/JHM-D-11-069.1

Vano, J. A., B. Udall, D. R. Cayan, J. T. Overpeck, et al. 2014. “Understanding Uncertainties in Future Colorado River Streamflow.” *Bulletin of the American Meteorological Society* (BAMS) Jan. 2014: 59-78. Doi: 10.1175/BAMS-D-12-00228.1

Wilks, D. S. 2011. *Statistical Methods in the Atmospheric Sciences, Third Edition*. Oxford: Academic Press.

Xiao, M., B. Udall, D. P. Lettenmaier. 2018. “On the Causes of Declining Colorado River Streamflows.” *Water Resources Research* (WRR) 54: 6739-6756. https://doi.org/10.1029/2018WR023153

This article is protected by copyright. All rights reserved

Table 1. Models used in this analysis, whether they belong to the CMIP-3, 5, or 6 generation, the number of historical ensemble members analyzed, and the approximate resolution of the atmospheric model, in degrees (longitude x latitude).

| Model | CMIP generation | Num. ensemble members | Approx. atmo resolution (deg) |
|---|---|---|---|
| 1. ACCESS-CM2 | 6 | 3 | 1.875 x 1.25 |
| 2. ACCESS-ESM1-5 | 6 | 5 | 1.875 x 1.25 |
| 3. AWI-CM-1-1-MR | 6 | 5 | 0.938 x 0.935 |
| 4. BCC-CSM2-MR | 6 | 1 | 1.125 x 1.121 |
| 5. BCC-ESM1 | 6 | 3 | 2.8 x 2.8 |
| 6. CNRM-CM6-1 | 6 | 1 | 1.4 x 1.4 |
| 7. CNRM-CM6-1-HR | 6 | 1 | 0.5 x 0.5 |
| 8. CNRM-ESM2-1 | 6 | 5 | 1.4 x 1.4 |
| 9. CanESM5 | 6 | 7 | 2.8 x 2.8 |
| 10. EC-Earth3 | 6 | 4 | 0.7 x 0.7 |
| 11. EC-Earth3-Veg | 6 | 5 | 0.7 x 0.7 |
| 12. FGOALS-f3-L | 6 | 3 | 1.25 x 1.0 |
| 13. FGOALS-g3 | 6 | 4 | 2.0 x 2.278 |
| 14. GFDL-CM4 | 6 | 1 | 1.25 x 1.0 |
| 15. GFDL-ESM4 | 6 | 1 | 1.25 x 1.0 |
| 16. HadGEM3-GC31-LL | 6 | 4 | 1.875 x 1.25 |
| 17. IPSL-CM6A-LR | 6 | 21 | 2.5 x 1.27 |
| 18. INM-CM4-8 | 6 | 1 | 2.0 x 1.5 |
| 19. INM-CM5-0 | 6 | 10 | 2.0 x 1.5 |
| 20. MIROC6 | 6 | 7 | 1.4 x 1.4 |
| 21. KACE-1-0-G | 6 | 3 | 1.875 x 1.25 |
| 22. MPI-ESM1-2-HR | 6 | 10 | 0.938 x 0.935 |
| 23. MPI-ESM-1-2-HAM | 6 | 2 | 1.875 x 1.865 |

This article is protected by copyright. All rights reserved

| Model | CMIP generation | Num. ensemble members | Approx. atmo resolution (deg) |
|---|---|---|---|
| 24. MPI-ESM1-2-LR | 6 | 9 | 1.875 x 1.865 |
| 25. MRI-ESM2-0 | 6 | 6 | 1.125 x 1.121 |
| 26. NESM3 | 6 | 3 | 1.875 x 1.865 |
| 27. NorCPM1 | 6 | 30 | 2.5 x 1.9 |
| 28. NorESM2-LM | 6 | 3 | 2.5 x 1.9 |
| 29. NorESM2-MM | 6 | 3 | 1.25 x 0.942 |
| 30. TaiESM1 | 6 | 1 | 1.25 x 0.942 |
| 31. UKESM1-0-LL | 6 | 5 | 1.875 x 1.25 |
| 32. ACCESS1-0 | 5 | 1 | 1.875 x 1.25 |
| 33. ACCESS1-3 | 5 | 1 | 1.875 x 1.25 |
| 34. CanESM2 | 5 | 5 | 2.8 x 2.8 |
| 35. CCSM4 | 5 | 5 | 1.25 x 0.942 |
| 36. CESM1-BGC | 5 | 1 | 1.25 x 0.942 |
| 37. CESM1-CAM5 | 5 | 3 | 1.25 x 0.942 |
| 38. CMCC-CM | 5 | 1 | 0.75 x 0.75 |
| 39. CNRM-CM5 | 5 | 5 | 1.4 x 1.4 |
| 40. CSIRO-Mk3-6-0 | 5 | 10 | 1.875 x 1.865 |
| 41. EC-EARTH | 5 | 4 | 1.125 x 1.12 |
| 42. FGOALS-g2 | 5 | 1 | 2.8 x 3.0 |
| 43. FGOALS-s2 | 5 | 2 | 2.812 x 1.659 |
| 44. FIO-ESM | 5 | 3 | 2.8 x 2.8 |
| 45. GFDL-CM3 | 5 | 1 | 2.5 x 2.0 |
| 46. GFDL-ESM2G | 5 | 1 | 2.5 x 2.0 |
| 47. GFDL-ESM2M | 5 | 1 | 2.5 x 2.0 |
| 48. GISS-E2-R | 5 | 5 | 2.5 x 2.0 |
| 49. GISS-E2-R-CC | 5 | 1 | 2.5 x 2.0 |

This article is protected by copyright. All rights reserved

| Model | CMIP generation | Num. ensemble members | Approx. atmo resolution (deg) |
|---|---|---|---|
| 50. HadGEM2-AO | 5 | 1 | 1.875 x 1.25 |
| 51. HadGEM2-CC | 5 | 1 | 1.875 x 1.25 |
| 52. HadGEM2-ES | 5 | 4 | 1.875 x 1.25 |
| 53. IPSL-CM5A-LR | 5 | 4 | 3.75 x 1.9 |
| 54. IPSL-CM5A-MR | 5 | 1 | 2.5 x 1.268 |
| 55. IPSL-CM5B-LR | 5 | 1 | 3.75 x 1.9 |
| 56. MIROC-ESM | 5 | 1 | 2.8 x 2.8 |
| 57. MIROC-ESM-CHEM | 5 | 1 | 2.8 x 2.8 |
| 58. MIROC5 | 5 | 1 | 1.4 x 1.4 |
| 59. MPI-ESM-LR | 5 | 3 | 1.875 x 1.865 |
| 60. MPI-ESM-MR | 5 | 1 | 1.875 x 1.865 |
| 61. MRI-CGCM3 | 5 | 1 | 1.125 x 1.121 |
| 62. NorESM1-M | 5 | 1 | 2.5 x 1.9 |
| 63. NorESM1-ME | 5 | 1 | 2.5 x 1.9 |
| 64. bcc-csm1-1 | 5 | 1 | 2.8 x 2.8 |
| 65. bcc-csm1-1-m | 5 | 1 | 1.125 x 1.121 |
| 66. inmcm4 | 5 | 1 | 2.0 x 1.5 |
| 67. bccr_bcm2_0 | 3 | 1 | 2.8 x 2.8 |
| 68. cccma_cgcm3_1 | 3 | 5 | 3.75 x 3.7 |
| 69. cnrm_cm3 | 3 | 1 | 2.8 x 2.8 |
| 70. csiro_mk3_0 | 3 | 1 | 1.875 x 1.865 |
| 71. gfdl_cm2_0 | 3 | 1 | 2.5 x 2.0 |
| 72. gfdl_cm2_1 | 3 | 1 | 2.5 x 2.0 |
| 73. giss_model_e_r | 3 | 1 | 5.0 x 3.95 |
| 74. inmcm3_0 | 3 | 1 | 5.0 x 4.0 |
| 75. ipsl_cm4 | 3 | 1 | 3.75 x 2.5 |

This article is protected by copyright. All rights reserved

| Model | CMIP generation | Num. ensemble members | Approx. atmo resolution (deg) |
|---|---|---|---|
| 76. miroc3_2_medres | 3 | 3 | 2.8 x 2.8 |
| 77. miub_echo_g | 3 | 3 | 3.75 x 3.7 |
| 78. mpi_echam5 | 3 | 3 | 1.875 x 1.865 |
| 79. mri_cgcm2_3_2a | 3 | 5 | 2.8 x 2.8 |
| 80. ncar_ccsm3_0 | 3 | 7 | 1.4 x 1.4 |
| 81. ncar_pcm1 | 3 | 4 | 2.8 x 2.8 |
| 82. ukmo_hadcm3 | 3 | 1 | 3.75 x 2.5 |

This article is protected by copyright. All rights reserved

## Figure Captions

Figure 1. Histograms of model-projected surface temperature changes (°C; tas; red bars/left column) and precipitation trends (mm/day per century; pr; green bars/right column) in the Upper Colorado River Basin. Top row: for all models before the global culling. For models that have multiple ensemble members, values are averaged across ensemble members before plotting. Bottom row: after the global culling. Changes are for the sresa2, RCP 8.5, and SSP85 scenarios for the CMIP3, CMIP5, and CMIP6 models, respectively. Temperature changes are evaluated as the 2070-2099 mean minus the 1950-2005 mean. Precipitation changes are evaluated as a best-fit least squares linear trend (millimeters per day per century) over the period 1950-2099. For reference, the mean observed annual precipitation over the Upper Basin is approximately 1.09 millimeters per day.

Figure 2. The Upper Colorado River Basin (brown outline) and the evaluation domain used in this work, as indicated by the centers of the 1-by-1 gridcells (black dots). Colors show elevation in meters.

Figure 3. Example showing the fields for the calculation of winter (DJF) precipitation metric. Top left: observed field (mm day$^{-1}$). Bottom left: Standard deviation of observations (mm day$^{-1}$). Top center: A model that performs well, CanESM2. Bottom middle: the z-score for CanESM2, i.e., the difference between the model and observations, divided by the observed standard deviation. Right column: Same as the middle column, but for FIO-ESM, which performs poorly on this metric.

Figure 4. Top left: the observed SST pattern (°C) associated with ENSO. Top right: the standard deviation (°C) of the observed pattern. Middle row: for CESM1-CAM5, the model's observed pattern of SST for ENSO (left) and the model's z-score (right, dimensionless). Bottom row: same, for CSIRO-Mk3-6-0.

Figure 5. Portrait plot of the model skill scores. The metrics are along the X axis (orange/red shows good skill, blues show poor skill), and the models along the Y axis. Metric labels, from left to right, are: seasonal (DJF, MAM, JJA, SON) mean ($\bar{\boldsymbol{x}}$) and standard deviation in 1-, 5-, and 10-year blocks ($\sigma_1$, $\sigma_5$, $\sigma_{10}$) of temperature (T) and precipitation (P); the seasonal cycle

This article is protected by copyright. All rights reserved

evaluated via the amplitude (A) and phase (ϕ) of temperature and precipitation; ENSO and the PDO evaluated via the mean SST pattern in the tropical Pacific ($\overline{\boldsymbol{x}}$), the spectrum (S), and the teleconnected response over the western U.S. in temperature (T) and precipitation (P); and the teleconnected (TCON) correlation maps of temperature (T) and sea level pressure (S) with Upper Basin precipitation variability during the warm (Wrm; AMJJAS) and cold (Cld; ONDJFM) seasons. Names of CMIP3 models are shown in black, CMIP5 models in blue, and CMIP6 in red.

Figure 6. Distribution of skill scores sorted by mean metric value. Better simulated metrics have higher skill scores (closer to the perfect value of 1) and are plotted at the top. Worse simulated metrics are at the bottom. The whiskers and dots show the mean of the metric (center line), interquartile range (box), 90 percent range (bars), and extreme values (dots).

Figure 7. The leading two EOFs (red) and associated PCs (blue) of the skill score matrix. CMIP3, 5, and 6 model names are in black, blue, and red, respectively. The first two EOFs explain 68.1% and 9.4% of the variance, respectively.

Figure 8. Model quality rankings after the EOF process has been applied. Best models (lowest Dss values) are at the bottom, worst models at the top. The 95% confidence intervals (red lines) are estimates derived from an analysis of models with multiple ensemble members – see text for details. The number of realizations is shown by n along the right hand side. Black, blue, and red names indicate CMIP3, CMIP5, and CMIP6, respectively. The vertical black bars with diamonds illustrate, for a few example models chosen to span the results, the range of models whose rankings are statistically indistinguishable from the base model (indicated by the diamond) given the uncertainty in Dss. For example, the Dss value of TaiESM1 is not statistically distinct from the Dss values of models ranging from GFDL-ESM4 to GISS-E2-R-CC.

Figure 9. Performance of the CMIP3, CMIP5, and CMIP6 (black, blue, and red, respectively) models on Index-3 (top left), and the individual components of index-3, the temperature (top right), precipitation (bottom left), and circulation (bottom right) indices. The diamonds show the mean of each CMIP's distribution; the dots show individual model values. The P value shown in the panels is the chance that the CMIP distributions have means that differ only due to sampling fluctuations.

This article is protected by copyright. All rights reserved

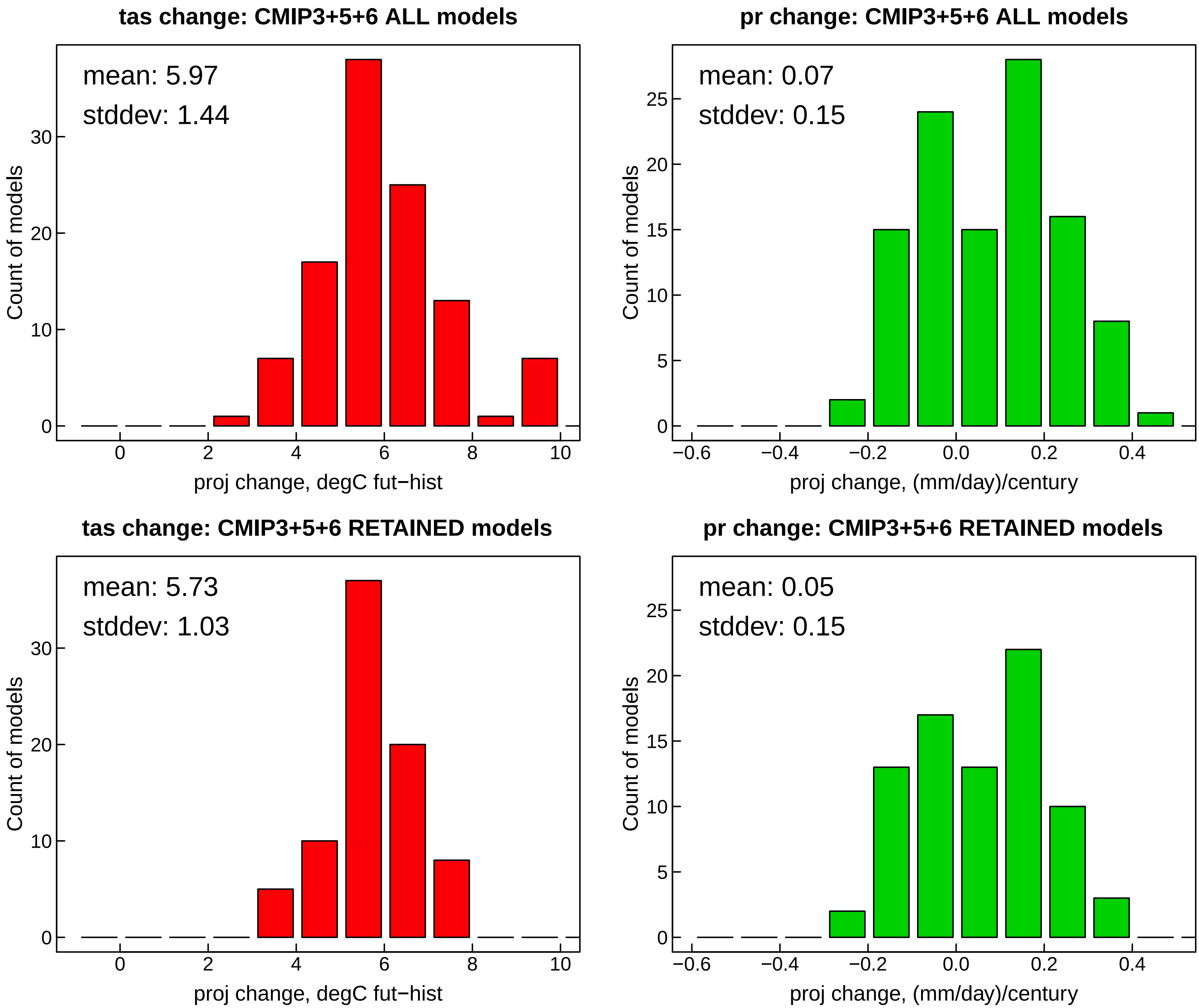
Figure 10. Portrait plot of the model skill scores for the case with simple bias correction. The format is the same as Figure 8; see that caption for figure details.

Figure 11. Model quality rankings after the simple bias correction has been applied and redundant information removed via an EOF approach. The format is the same as Figure 8.

Figure 12. Model errors (lower is better) for the CMIP3, CMIP5, and CMIP6 model generations, both before (light dots) and after (dark dots) the simple bias correction. The diamonds indicate the mean of the distributions. The P values shown in the lower left are the chance that the means of the CMIP3 and CMIP6 distributions differ only due to sampling fluctuations, as estimated by a two-sample t-test.

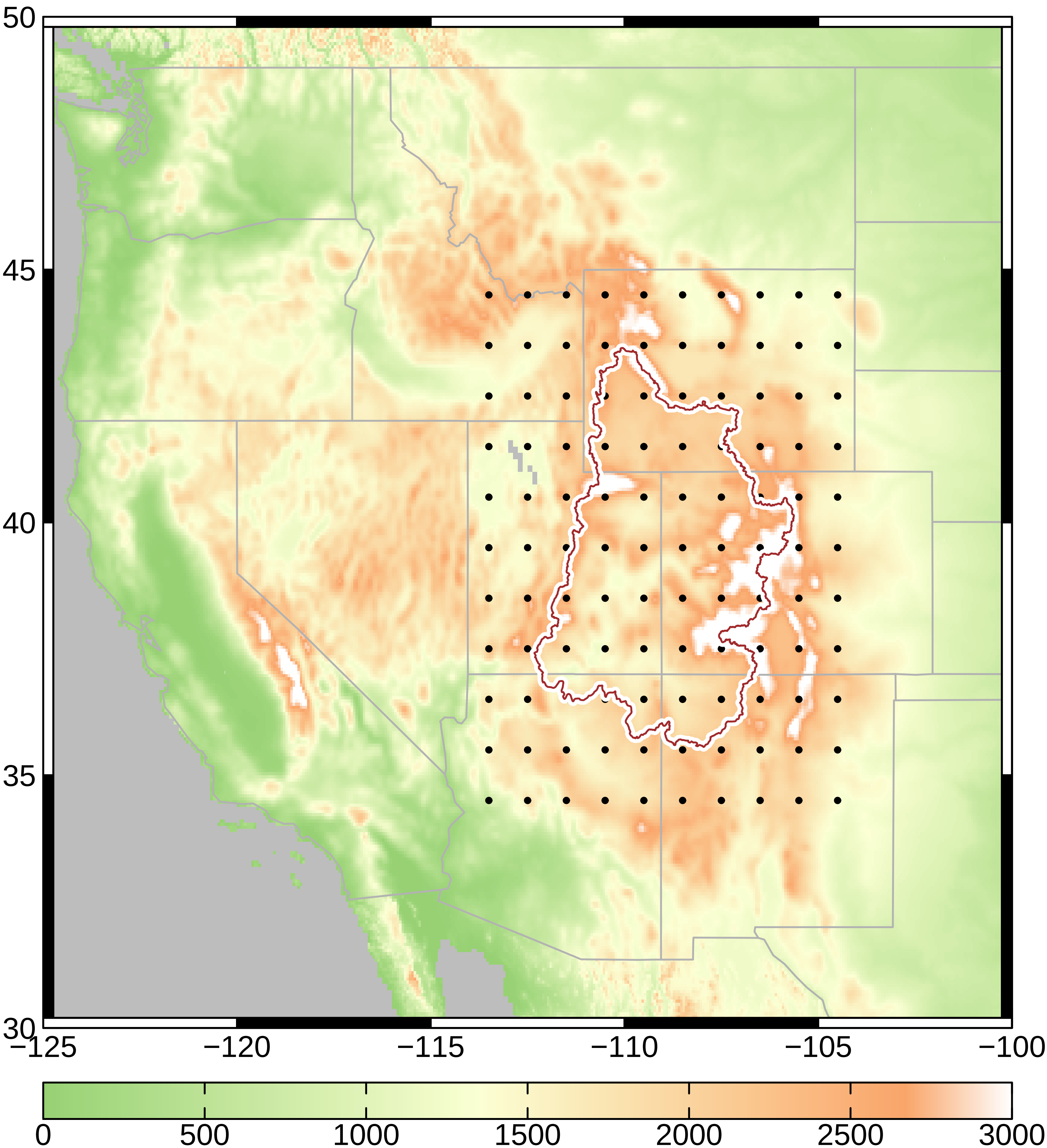
Figure 13. Final model performance (Dss from Figure 11) as a function of model spatial resolution (column 4 of Table 1). Left: no bias correction. Right: With simple bias correction. The spatial resolution is taken as the average of the longitudinal and latitudinal resolutions, in degrees. Solid purple line: least-squares best-fit line using all models. Black, blue, and red lines: least-squares best fit lines using only CMIP 3, 5, and 6 models, respectively. None of the trends are significantly different from zero at the 95% confidence level except the CMIP3 trend in the No-BC case.

This article is protected by copyright. All rights reserved

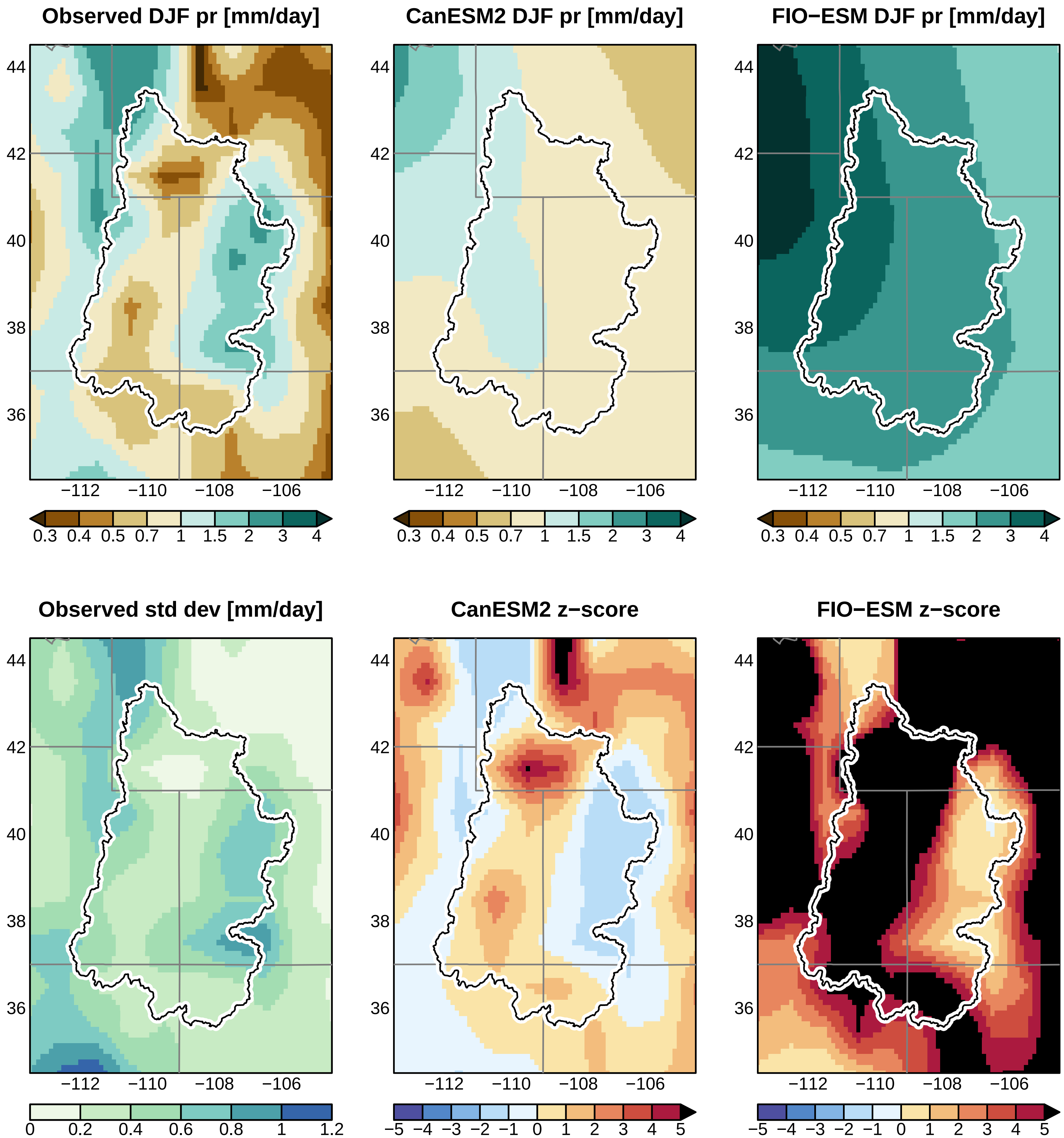



jawr_12974_f1.eps

This article is protected by copyright. All rights reserved

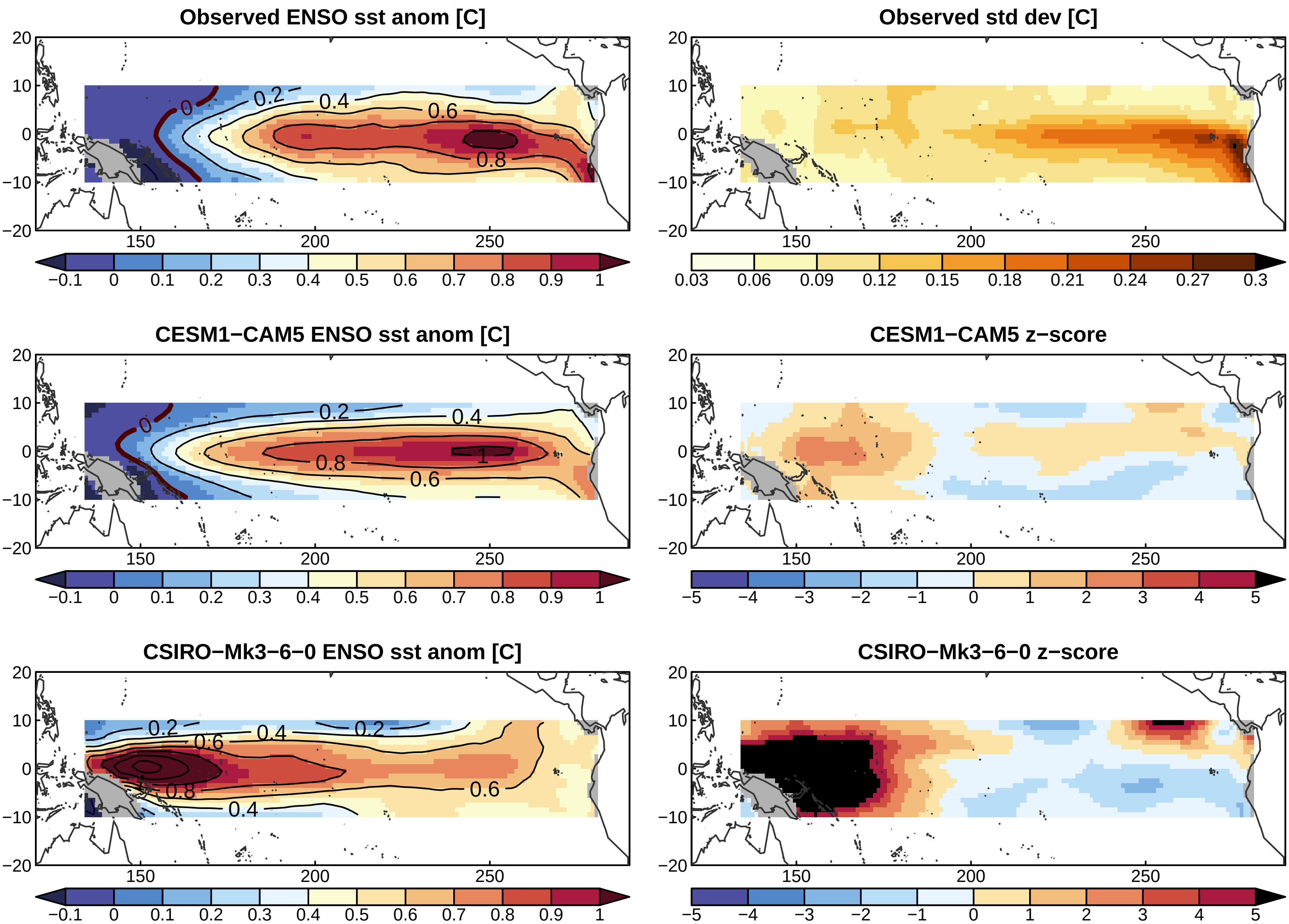



jawr_12974_f2.eps
This article is protected by copyright. All rights reserved



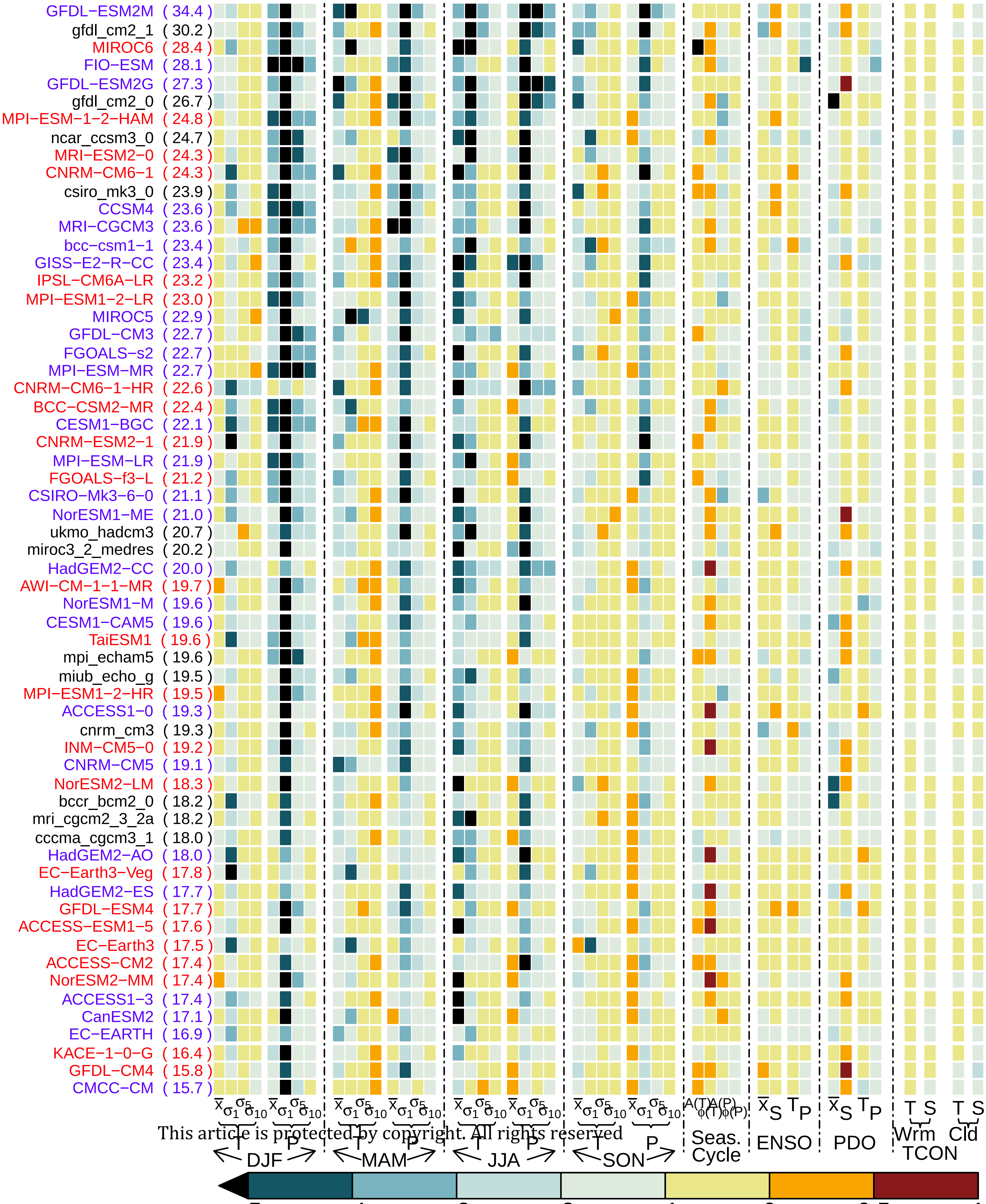

jawr_12974_f3.eps

This article is protected by copyright. All rights reserved

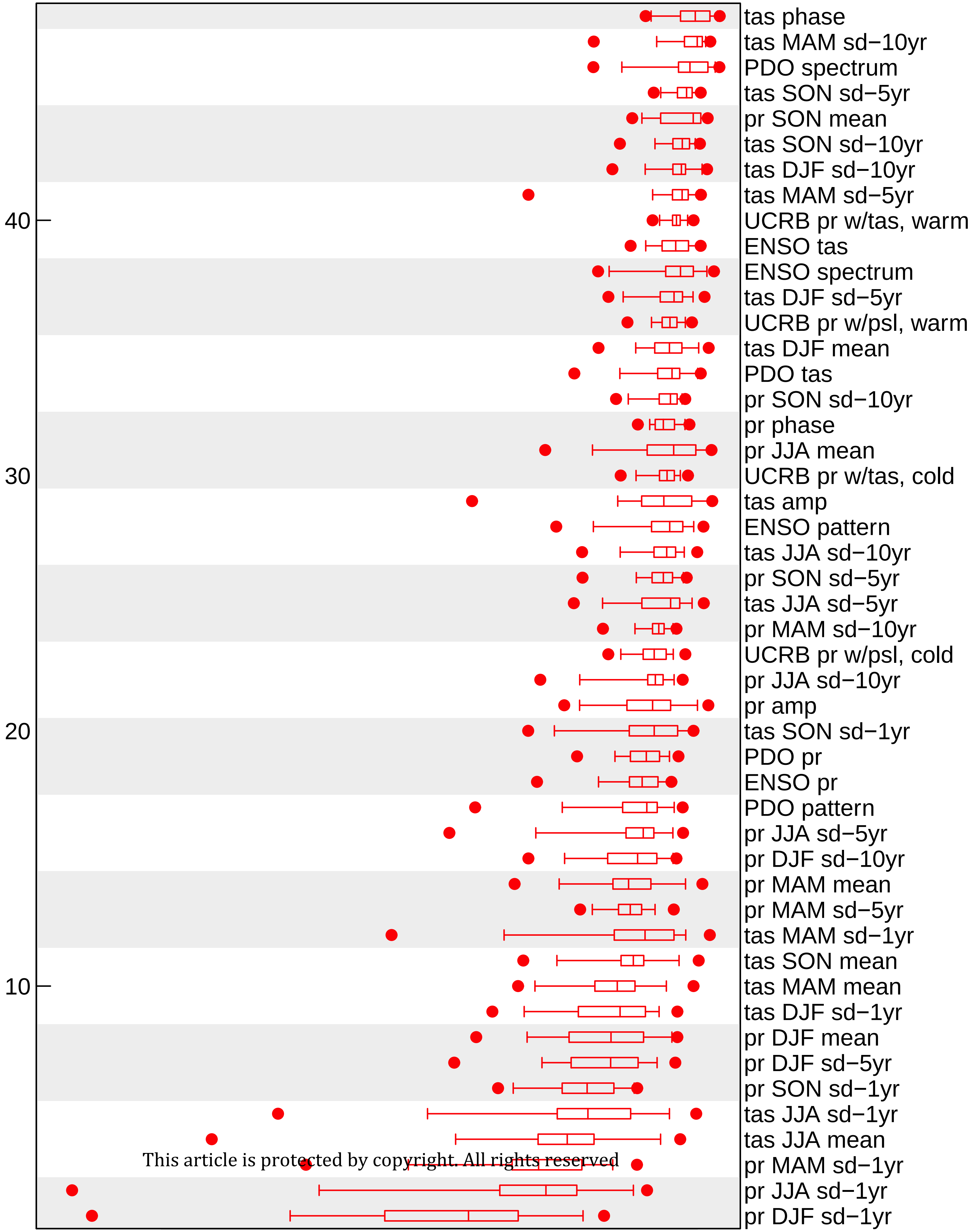



jawr_12974_f4.eps

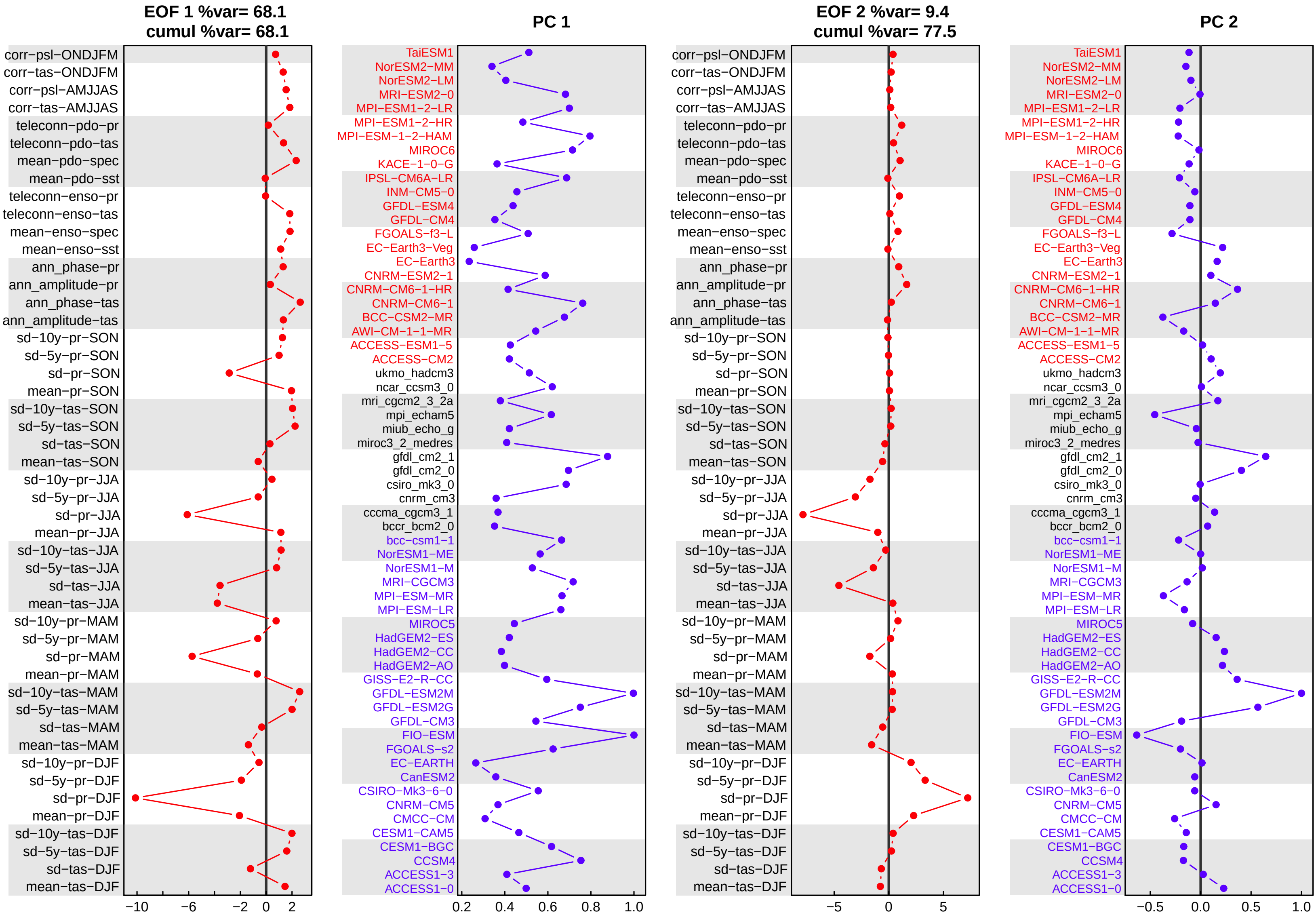
This article is protected by copyright. All rights reserved



This article is protected by copyright. All rights reserved

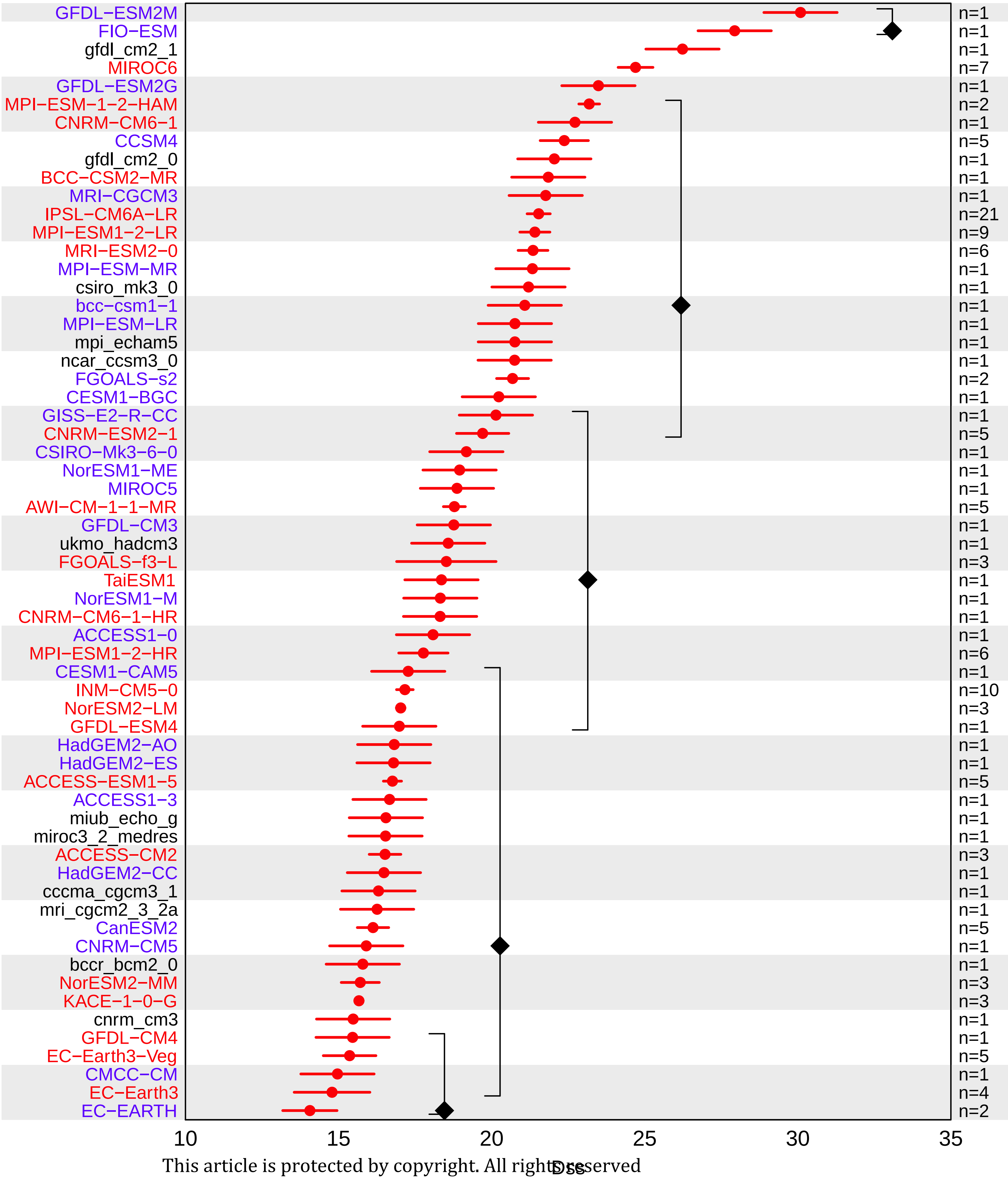



This article is protected by copyright. All rights reserved



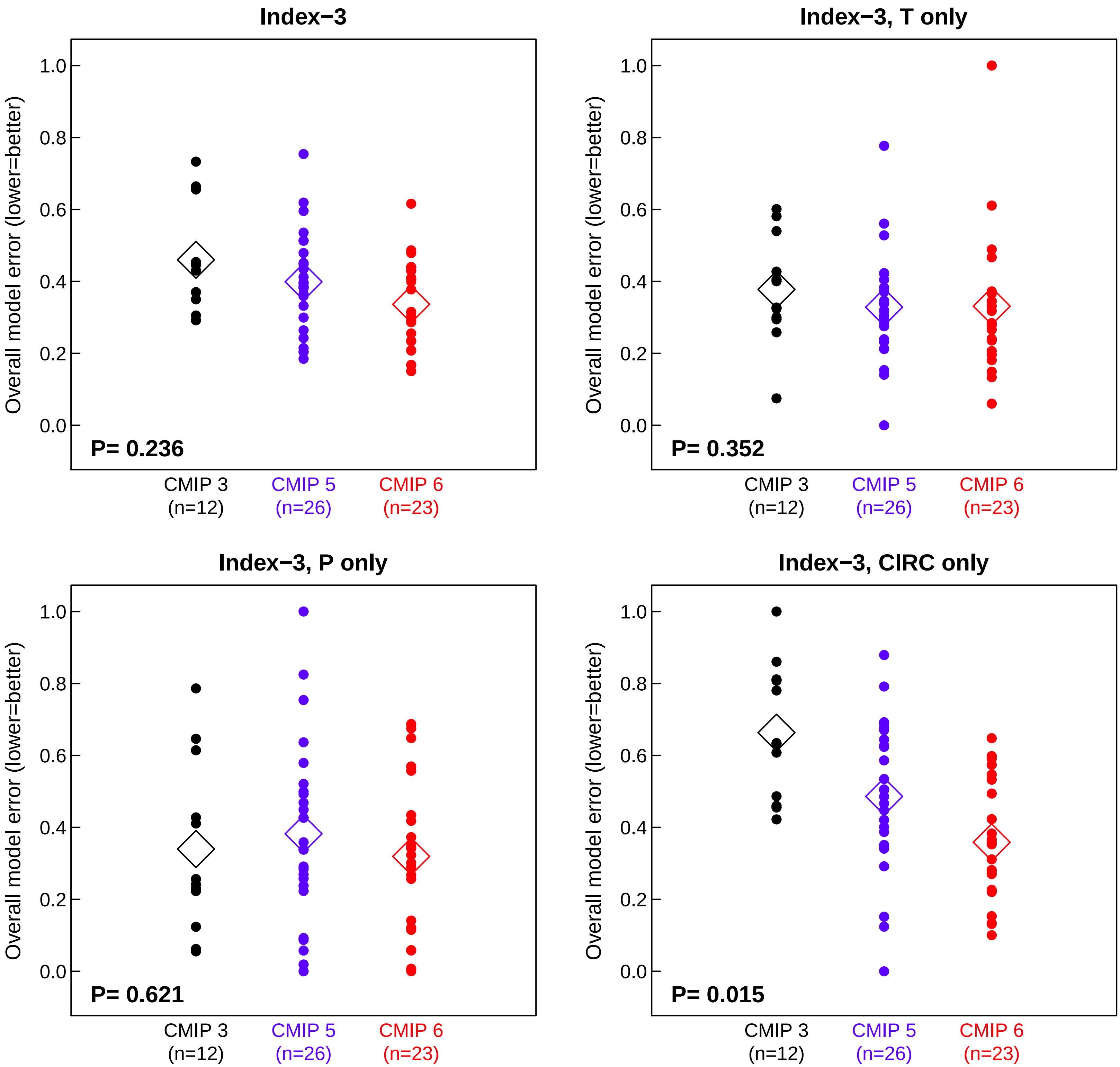

jawr_12974_f7.eps

This article is protected by copyright. All rights reserved



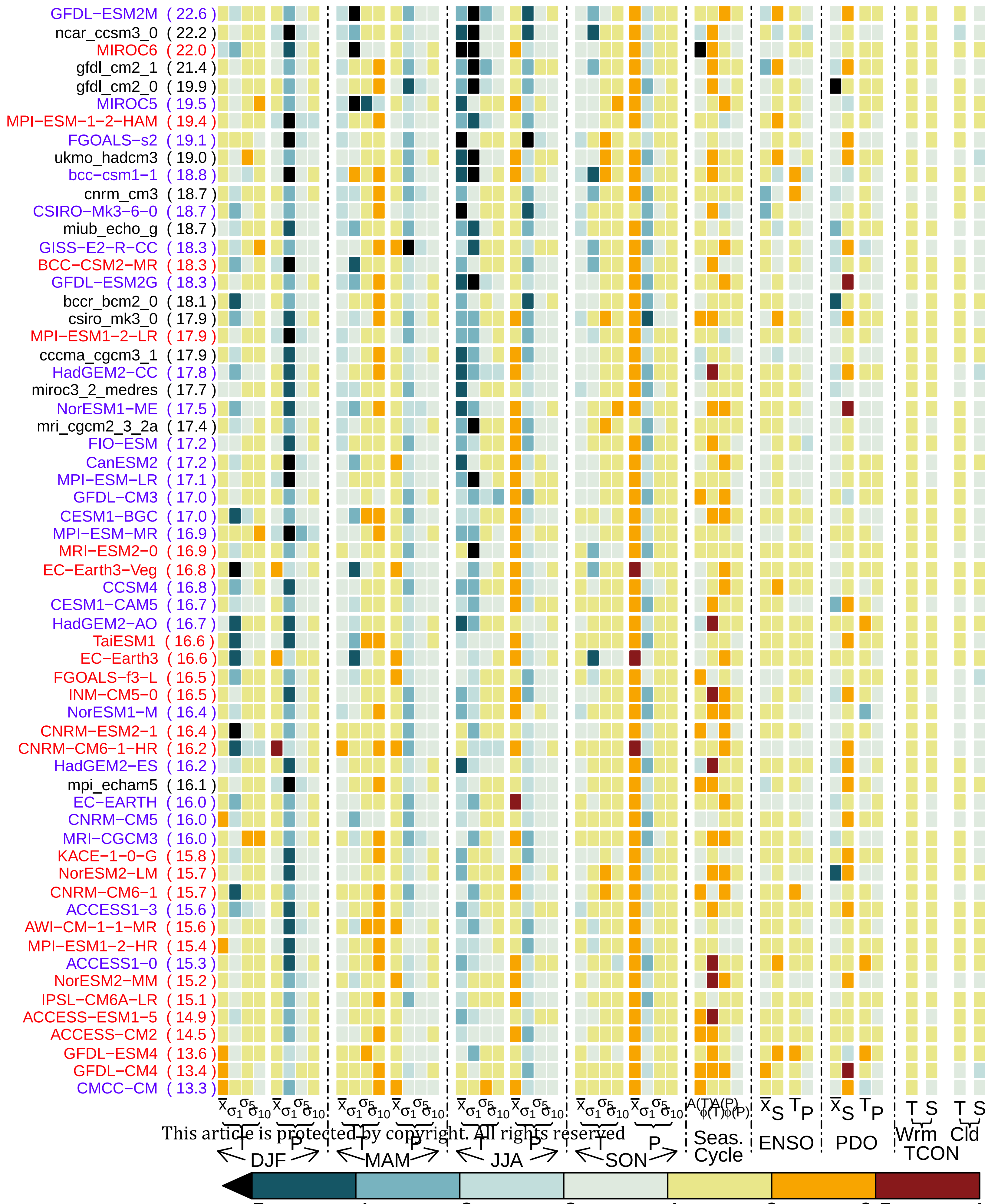

This article is protected by copyright. All rights reserved

jawr_12974_f8.eps

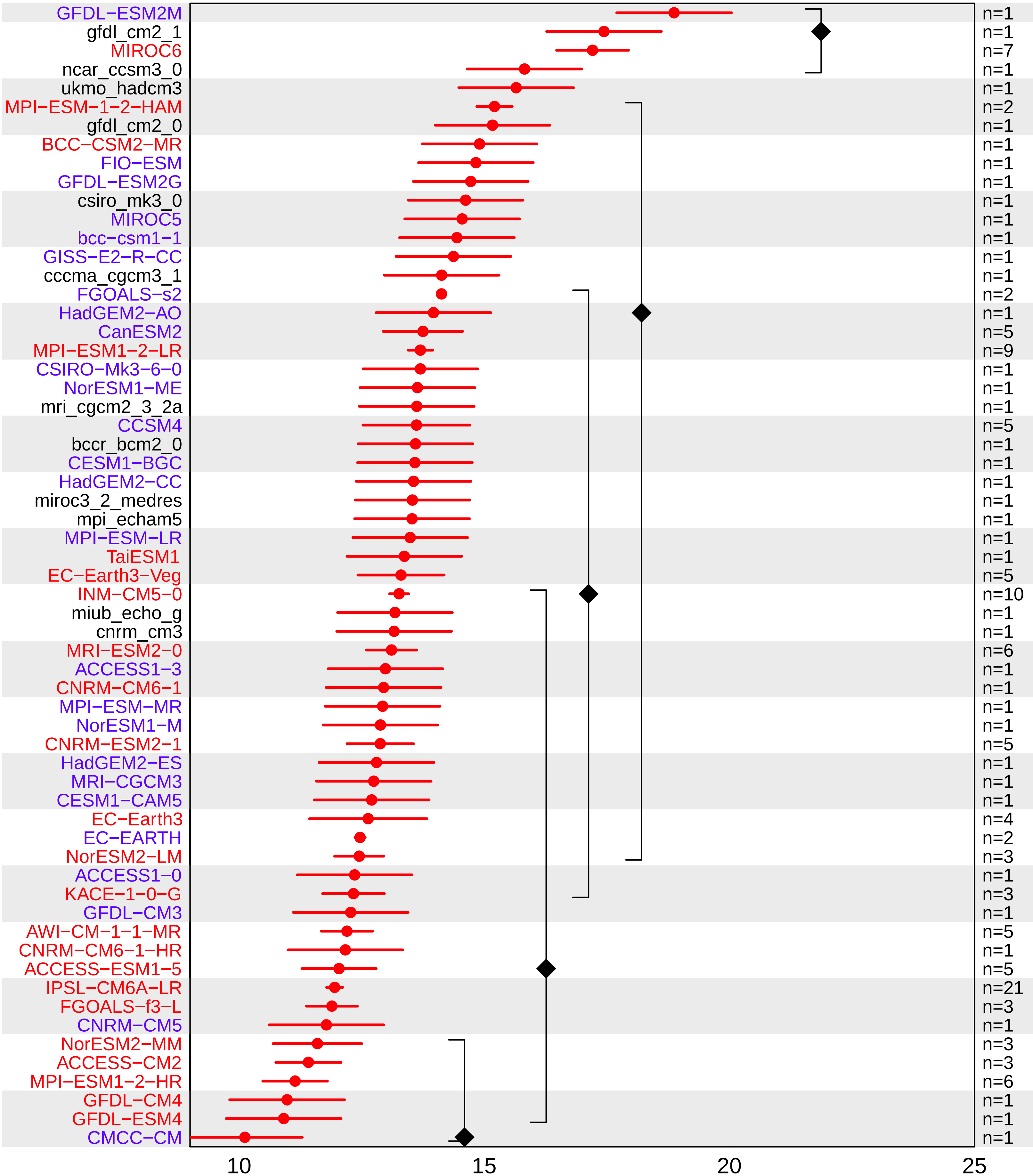



jawr_12974_f9.eps

This article is protected by copyright. All rights reserved

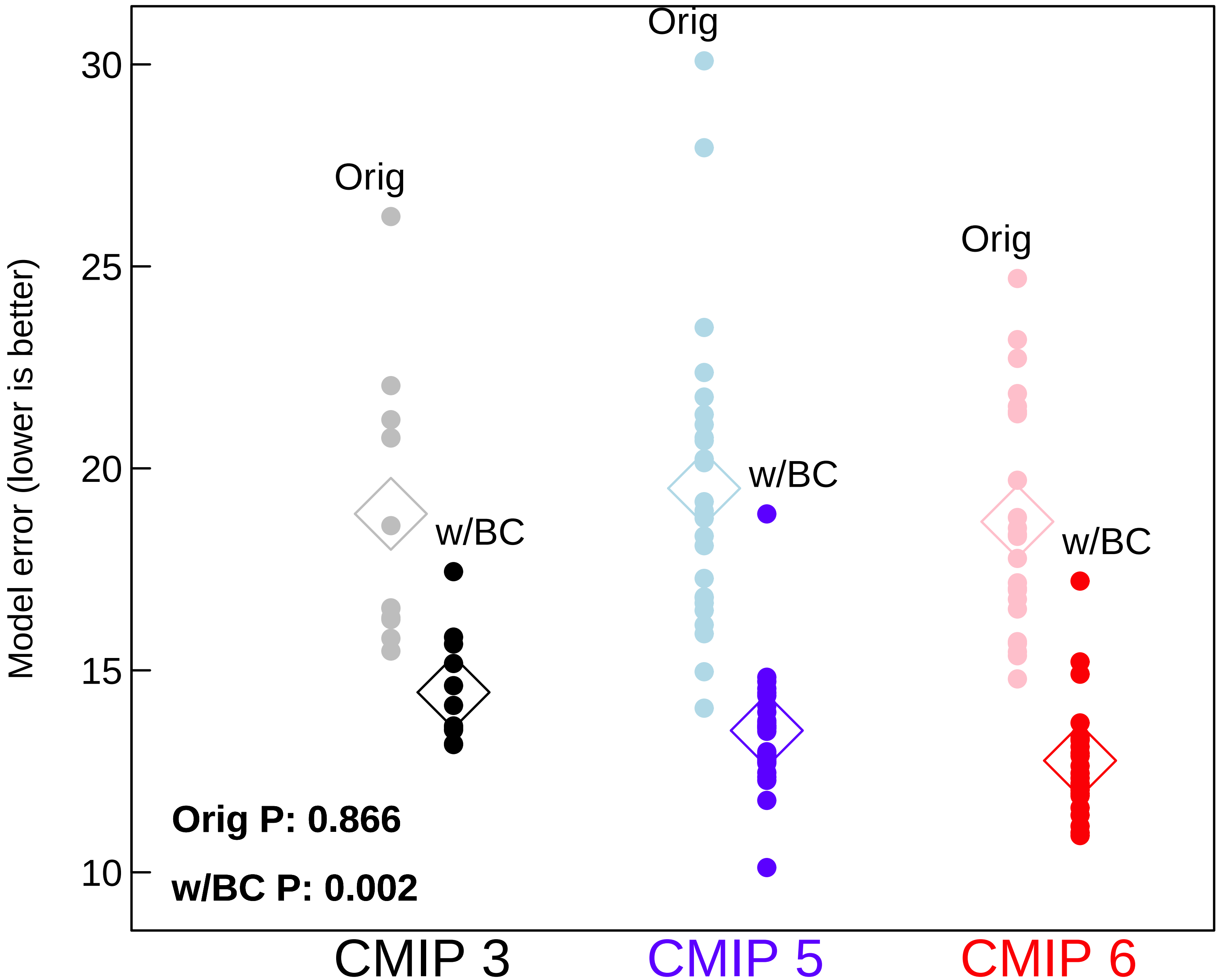


This article is protected by copyright. All rights reserved



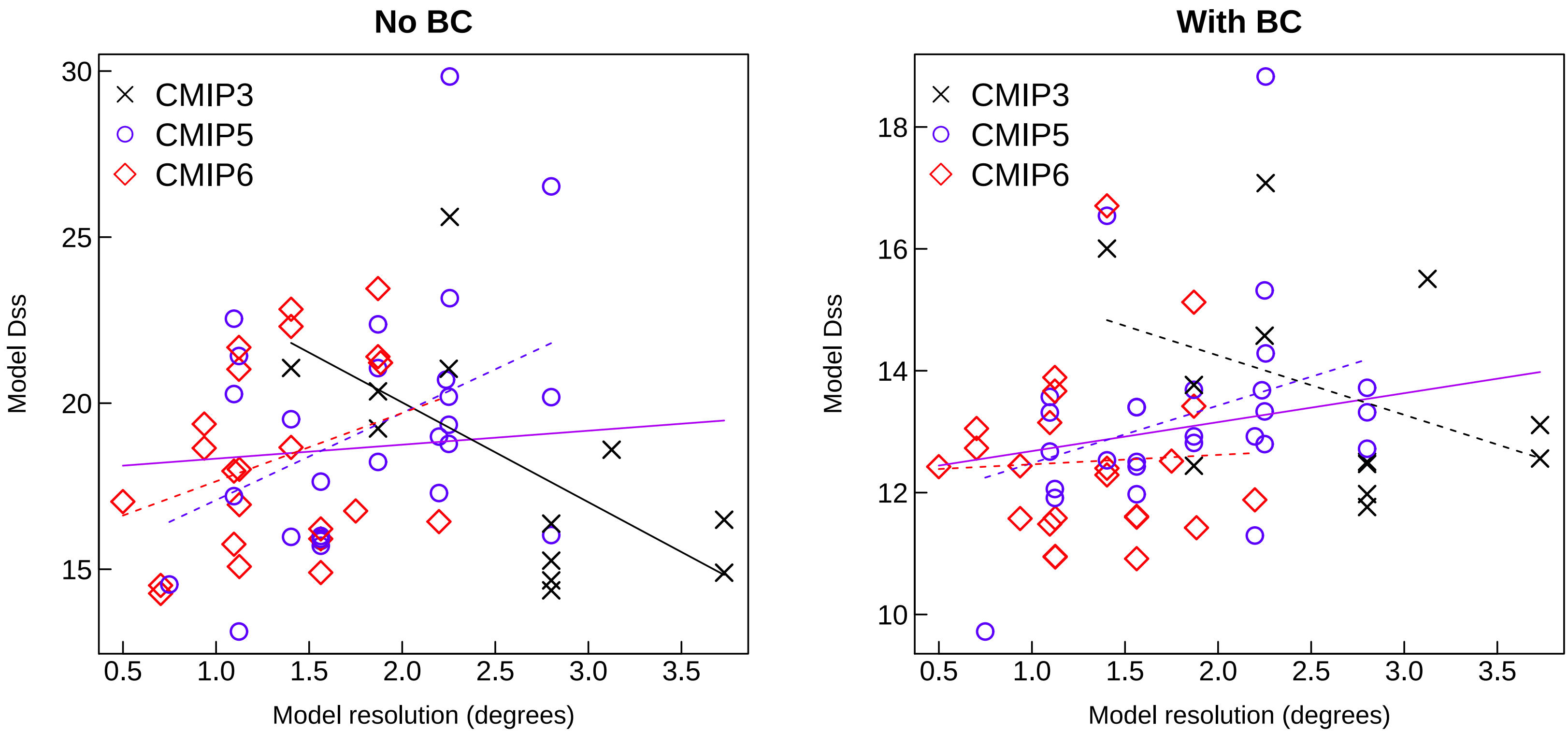

This article is protected by copyright. All rights reserved

jawr_12974_f11.eps




jawr_12974_f12.eps

This article is protected by copyright. All rights reserved




jawr_12974_f13.eps

This article is protected by copyright. All rights reserved