Improving Earth System Model Selection Methodologies for Projecting Hydroclimatic Change: Case Study in the Pacific Northwest

Nicholas D. Lybarger[a], Abigail Smith[a], Andrew J. Newman[a], Ethan D. Gutmann[a], Andrew W. Wood[a,e], Christopher Frans[d], Michael Warner[b], Jeffrey R. Arnold[c]

a. U.S. National Science Foundation National Center for Atmospheric Research, Boulder, Colorado

b. U.S. Army Corps of Engineers, Seattle, Washington

c. MITRE Corporation, McLean, Virginia

d. US Bureau of Reclamation, Lakewood, Colorado

e. Colorado School of Mines, Golden, Colorado

Corresponding author: Nicholas D. Lybarger (nlybarger@ucar.edu)

Key Points:

- A method for evaluation of climate models over the PNW using a combination of global and regional metrics has been developed.
- This allows for a reduced envelope of ESMs for impact applications without significantly affecting the future trend projections.

Abstract

The rapid expansion of Earth system model (ESM) data available from the Coupled Model Intercomparison Project Phase 6 (CMIP6) necessitates new methods to evaluate the performance and suitability of ESMs used for hydroclimate applications as these extremely large data volumes complicate stakeholder efforts to use new ESM outputs in updated climate vulnerability and impact assessments. We develop an analysis framework to inform ESM sub-selection based on process-oriented considerations and demonstrate its performance for a regional application in the US Pacific Northwest. First, a suite of global and regional metrics is calculated, using multiple historical observation datasets to assess ESM performance. These metrics are then used to rank CMIP6 models, and a culled ensemble of models is selected using a trend-related diagnostics approach. This culling strategy does not dramatically change climate scenario trend projections in this region, despite retaining only 20% of the CMIP6 ESMs in the final model ensemble. The reliability of the culled trend projection envelope and model response similarity is also assessed using a perfect model framework. The absolute difference in temperature trend projections is reduced relative to the full ensemble compared to the model for each SSP scenario, while precipitation trend errors are largely unaffected. In addition, we find that the spread of the culled ensemble temperature and precipitation trends includes the trend of the "truth" model ~83-92% of the time. This analysis demonstrates a reliable method to reduce ESM ensemble size that can ease use of ESMs for creating and understanding climate vulnerability and impact assessments.

Plain Language Summary

This study provides an updated and rigorously tested method for evaluating the performance of climate models for applications relevant to water managers and other stakeholders. Using traditional metrics of climate model performance, both regional and global, as well as newly developed metrics based on processes important for the simulation of precipitation, we have created a generalizable, systematic, and succinct method for reducing the number of models to be considered for climate change impact applications. By reducing the number of relevant models to around 20% of the total models and not having a significant impact on future temperature and precipitation trend projections in doing so, we strongly reduce the computational effort needed to gain a realistic simulation of the future climate for a given regional impact. This method is tested with a variety of statistical tests and found to be reliable for our application over the Pacific Northwest United States.

1. Introduction

Understanding future changes in regional hydroclimates is a key priority for water security and resource climate change impact analyses. For example, increasing air temperatures are driving changes in the accumulation of snowpack, shifts in the timing of snowmelt runoff and in the fraction of precipitation falling as snow (Serreze et al., 1999; Barnett et al., 2008; Easterling et al., 2017; Mote et al., 2018; Musselman et al., 2021). Annual precipitation has decreased over much of the Western U.S. (Prein et al., 2017; Henn et al., 2018), yet there is substantial variability both regionally and seasonally in future projections of precipitation, including the frequency and magnitude of heavy precipitation events (e.g., Easterling et al., 2017; Lopez-Cantu et al., 2020; Kim et al., 2020). In order for water managers to incorporate changes in risk over time, reliable future projections of precipitation and air temperature that can be developed with minimal cost to the partner organizations are needed.

Earth system models (ESMs) are a valuable tool for creating future projections of the large-scale climate processes that in part govern precipitation and temperature patterns. The Coupled Model Intercomparison Project phase 6 (CMIP6; Eyring et al., 2016) provides ESM projection output from more than 70 models (https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/), including some models with many projection ensemble members. This historically large data volume presents a challenge for intended users who apply ESM projections in climate vulnerability assessments, which tend to involve bias correction and spatial downscaling as well as impact models such as hydrologic and water management systems models (e.g. Brekke et al., 2008, 2009; Mote et al., 2011; Rupp et al., 2013; Clark et al., 2016; Newman et al., 2022). This results in a chain of models and simulations (*X* ESMs * *Y* downscaling methods * *Z* impact models), and

subsequently it is challenging to apply to more than a subset of CMIP6 models. Thus water managers and other users are almost universally required to select a subset of ESMs for their analyses -- in some cases moderately (e.g. Brekke et al., 2008), and in other cases severely, as when just a few models are chosen for a storylines, scenario narrative approach (e.g. Basharin et al., 2015; Najafi et al., 2011), or “four corners” style approach wherein only a few models projecting the most extreme precipitation and temperature change are chosen (e.g. Hosseinizadeh et al., 2015).

Ideally, evaluations of the reliability (i.e., the ability of an ESM to credibly represent the observed historical climate, e.g. Giorgi, 2020) of ESMs for climate impact applications would focus on assessing confidence in the ESM change signals of user defined key hydroclimatic variables across global to regional scales (Doblas-Reyes et al., 2021; Goldenson et al., 2023). In practice this is very challenging and a variety of methods and tools have been developed. Most simply, one can assume all ESMs are equally plausible (e.g. Meehl et al., 2007). In this case, one could pick an arbitrary number of ESMs based on the multi-model mean (e.g. Pierce et al., 2009), the full ensemble response (e.g., Sanderson et al., 2017), or pick ESMs that span the vulnerability range of the application (e.g. Weaver et al., 2017). Beyond equal plausibility, methods have generally focused on variables and metrics related to the specific region, such as regional interannual variability, regional trends and seasonality, or daily extremes (Mote and Salathe, 2010; Rupp et al., 2013; Sanderson et al., 2017; McSweeney et al., 2015).

Global-scale metrics have primarily been included via global trends or regional teleconnection metrics or specific oscillation indices. For example, the El Nino-Southern Oscillation (ENSO) sea surface temperature (SST) pattern via the Nino3 (or Nino3.4), North Pacific Index, North Atlantic Oscillation, and their corresponding regional precipitation and

temperature correlations are included to represent global scale processes (Brekke et al., 2008; Pierce et al., 2009; Rupp et al., 2013; Snover et al., 2013). Model response or genealogical (e.g., code) similarity has also been used to evaluate or select ESMs (Masson and Knutti, 2011; Knutti et al., 2013; Sanderson et al., 2017; Brunner et al., 2020). Other novel approaches using the concepts of emergent constraints over the globe or a region, or global trend constraints have been used to assess ESMs and develop relationships to scale (or constrain) responses for particular variables or regions (Hausfather et al., 2020; Simpson et al., 2021; Lyu et al., 2021; Tokarska et al., 2020; Ribes et al., 2022). However, focusing on global evaluation only for regional water security impact studies may be problematic as ESM performance and subsequent hydrologic response varies across regions (Melsen et al., 2018; Asenjan et al., 2023), thus the emphasis on regional metrics in regional studies.

There has also been a concerted effort to increase understanding, reproducibility, and access to ESM evaluation tools and results (Phillips et al., 2014; Righi et al., 2020; Maloney et al., 2019; Eyring et al., 2020; Parding et al., 2020; Schlund et al., 2023; Merrifield et al., 2023). These tools allow users to develop their own evaluations with varying levels of complexity moving from direct manipulation of ESM data (e.g., Schlund et al., 2023) to pre-processed CMIP diagnostics (Phillps et al., 2014) to web-based user platforms with simpler evaluations and accessible documentation focused on climate services (Parding et al., 2020), to more advanced offline multi-metric, flexible tools such as ClimSIPS (Merrifield et al., 2023) that require a relatively higher level of ESM familiarity by the user.

Here we report our work to synthesize key aspects of the aforementioned research through combining global evaluation (e.g., global temperature trends), physically based teleconnection metrics, and regional metrics to explore the effects of ESM evaluation, selection

via culling, and the resulting impacts on the range of future projections for the Northwestern United States and SW Canada domain (the Pacific Northwest or PNW). We include temporal split-sample evaluation where possible as a form of cross-validation, which is a critical tool for model prediction skill evaluation (e.g. Klemes, 1986; Wilks, 2019) that may provide a test of trend fidelity. We also explore perfect model comparisons for ESM evaluation, response similarity, and projection reliability (Sanderson et al., 2017; Liang et al., 2020). The core aim of this work is to increase confidence in ESM model selection and the consequent projections for water-resource applications by producing an integrated assessment of ESMs and incorporating different tests and metrics focused on model trends and processes. This evaluation methodology was co-designed with the US Army Corps of Engineers (USACE) Climate Preparedness and Resilience Program and is available in an open-source code base.

2. Data

2.1 CMIP6 models

Total monthly precipitation, monthly average temperature, and monthly average tropical sea surface temperature data from 63 CMIP6 models from the Earth System Grid Federation (ESGF, Cinquini et al., 2014) archive are evaluated here. This collection is intended to be as comprehensive as possible while acknowledging that differences in data availability exist between modeling centers. Occasionally, spatiotemporal inconsistencies between ensemble members for a given model were found that limited the ability to include those members, including differing grid definitions, inconsistent temporal coverage, and missing data. Table A1 summarizes specifications for the models included in this analysis.

2.2 Verification Datasets

We used a collection of gridded observations and reanalysis data as the basis of our evaluation. Capturing observational uncertainty is an important aspect of ESM verification; observations may be sparse for many regions and time periods, and even where adequate coverage exists, the observational uncertainty can be considerable (Rupp et al., 2013; Henn et al., 2018). To address this, six different sources of verification data are considered here in order to assess CMIP6 model performance. Three of these verification datasets have global coverage, while the other three include data only over the contiguous United States (CONUS). Each of these verification datasets is compared to the observation ensemble mean in an effort to assess observational agreement.

Gridded monthly observational and reanalysis precipitation and 2-meter air temperature datasets are considered to facilitate grid-to-grid comparisons with the ESMs. The European Centre for Medium-Range Weather Forecasting (ECMWF) Reanalysis Version 5 (ERA5), provides output from 1950 to the present at 0.25°x0.25° resolution (Hersbach et al., 2020). Two global observation-based gridded interpolation products are used as well. The Climatic Research Unit (CRU) gridded time series data includes all land areas except Antarctica at 0.5°x0.5° resolution from 1901 to 2021 (Harris, I. et al., 2020) while the University of Delaware (UDel) provides global monthly terrestrial time series of temperature and precipitation over land at 0.5°x0.5° resolution from 1901 to 2017 (Willmott and Matsuura, 2001). Due to the extensive temporal coverage of the CRU and UDel datasets, these sources are sufficient for the verification of global annual trends of temperature and precipitation. ERA5, CRU, and UDel act as sources of verification for the global metrics in this study.

The following three verification sources provide data only over CONUS and are used for the evaluation of CMIP6 models over the northwestern US. The Livneh et al. (2015)

hydrometeorological dataset provides daily maximum and minimum temperature, and daily precipitation at 1/16°x1/16° resolution from 1950-2011. Here, the daily average temperature is derived from the average of daily maximum and minimum temperatures, then these are temporally aggregated to provide monthly values. Oregon State University's Parameter-elevation Regressions on Independent Slopes Model (PRISM) Climate Group hosts monthly precipitation and temperature data at 4km x 4km resolution from 1981 to present (Daly et al., 2008). The Gridded Meteorological Ensemble Tool (GMET) is the final source of evaluation data applied in this analysis (Newman et al., 2015). The GMET dataset used here is conceptually similar to PRISM (in using terrain features to aid interpolation) but provides an ensemble in contrast to the deterministic datasets of Livneh or PRISM. These daily, 1/16°x1/16° data are aggregated to monthly means or totals, then the ensemble mean is taken prior to application as an evaluation data source. The temporal coverage of this dataset is 1970-2021. For all metrics, all data sources and ESM data are interpolated to a common 1°x1° grid and ocean points are masked so that all data sources are consistent. The Nino3.4 index is taken from the NOAA Physical Science Laboratory dataset using the HadISST1 historical reconstruction of SST (Rayner et al., 2003), while the ENSO Longitude Index (ELI, ° longitude, Patricola et al., 2020), discussed in Section 3, is taken from the National Energy Research Scientific Computing Center archive, computed from the ERSSTv5 reconstruction of historical SST (Huang et al., 2017). Table 1 summarizes and describes the verification datasets used in this analysis.

Table 1: List of verification datasets used in this analysis and their properties. Note that ERSST is used only for calculation of ELI, and HadISST is used only for calculation of the Nino3.4 index.

Dataset	Spatial Coverage	Resolution (Lon x Lat)	Temporal Coverage	Reference
Climatic Research Unit (CRU) gridded time series	Global (land only)	0.5° x 0.5°	1901 - 2021 (monthly)	Harris, I. et al., 2020
University of Delaware (UDel) terrestrial air temperature and precipitation gridded monthly time series	Global (land only)	0.5° x 0.5°	1900 - 2017 (monthly)	Willmott and Matsuura, 2001
ECMWF Reanalysis version 5 (ERA5)	Global	0.25° x 0.25°	1959 - present (daily)	Hersbach et al., 2020
Parameter-elevation Regression Independent Slopes Model (PRISM)	CONUS	0.04° x 0.04°	1981 - 2021 (daily)	Daly et al., 2008
Gridded Meteorological Ensemble Tool (GMET)	CONUS	0.0625° x 0.0625°	1970 - 2021 (daily)	Newman et al., 2015 (updated version)
Livneh near-surface gridded meteorological and derived hydrometeorological data (Livneh)	CONUS (land only)	0.0625° x 0.0625°	1915 - 2013 (daily)	Livneh et al., 2015
Extended Reconstructed Sea Surface Temperature (ERSST)	Global (ocean only)	2° x 2°	1854 - present (monthly)	Huang et al., 2017
Hadley Centre Sea Ice and Sea Surface Temperature (HadISST)	Global (ocean only)	1° x 1°	1870 - present (monthly)	Rayner et al., 2003

3. Methods

3.1 Model evaluation metrics

Table 2 lists the evaluation metrics considered in this analysis. Twenty-two of the twenty-eight metrics are drawn from the metrics used by Rupp et al. (2013) (R13). All of these metrics are categorized as “Highest” or “Higher” confidence in R13 except the linear trends of precipitation and temperature, which are included due to the importance of these quantities to users projecting future impacts of climate change and the risks associated with those projections. The use of such trends is challenging due to the influence of unforced low frequency climate variability on trends in multi-decadal to century-scale projections, but the lengthening observational record is helping to enable this strategy. Throughout this paper, temperature and precipitation trends are calculated using the least-squares method to find the linear regression of

region-averaged annual mean temperature or annual total precipitation for each model ensemble member, verified against CRU and UDel for the historical period, 1901-2014.

Table 2: List of metrics used in this analysis. For Trend-T and Trend-P, only CRU and UDel are used for verification over the period 1901-2014 to match the CMIP6 temporal coverage. For all other metrics, the full extent of the given verification data is used. Application domain refers to the spatial extent of the metric calculation, either global (G) or regional (R).

Metric	Application Domain	Description
Mean-T	G, R	Mean annual temperature across temporal extent of data
Mean-P	G, R	Mean annual precipitation across temporal extent of data
SeasAmp-T	R	Seasonal amplitude of temperature as the average difference between the hottest and coldest month of each year in a given dataset
SeasAmp-P	R	Seasonal amplitude of precipitation as the average difference between the driest and wettest month of each year in a given dataset
Trend-T	G, R	Linear trend of annual average temperature over the period 1901-2014
Trend-P	G, R	Linear trend of annual total precipitation over the period 1901-2014
DJF ELI Median	G	Median DJF value of ENSO Longitude Index
DJF ELI LevStat	G	Levene's statistic computed from comparison of DJF ELI time series against ERSSTv5
Nino3.4-pr r	G, R	Spatial correlation between temporal correlation maps of Nino3.4 index and precipitation
ELI-pr r	G, R	Spatial correlation between temporal correlation maps of ELI and precipitation
Nino3.4-T r	G, R	Spatial correlation between temporal correlation maps of DJF Nino3.4 index and temperature
ELI-T r	G, R	Spatial correlation between temporal correlation maps of DJF ELI and temperature

SpaceCorr MMM-T	G, R	Spatial correlation of mean seasonal temperature maps
SpaceCorr MMM-P	G, R	Spatial correlation of mean seasonal precipitation maps
SpaceSD MMM-T	G, R	Spatial standard deviation of mean seasonal temperature maps (Normalized by mean of verification data standard deviation)
SpaceSD MMM-P	G, R	Spatial standard deviation of mean seasonal precipitation maps (Normalized by mean of verification data standard deviation)

In addition to those metrics, we included 6 new metrics meant to probe different aspects of the models' representation of ENSO which is crucial for accurately simulating PNW (and global) seasonal precipitation and temperature (e.g. Tziperman et al., 1998; Schonher and Nicholson, 1989; Hoell et al., 2016). Using the Niño3.4 index time series averaged over DJF of each simulated year, a temporal correlation map is computed between grid cell's DJF temperature and precipitation for each model ensemble member and verification dataset. Then, the spatial correlation between each of these temporal correlation maps is computed. The process is then repeated, instead using the ELI DJF time series to produce the temporal correlation maps. The ELI is included here due to its demonstrated skill in capturing ENSO diversity (Patricola et al., 2020), an important nuance for PNW hydrometeorological impacts. Representations of ENSO variability using both the canonical Nino3.4 index and the newly developed ELI are both included in this analysis due to the differing process representation necessary to capture both. While Nino3.4 simply captures the average SST in a static box in the eastern Pacific, the ELI gives the average longitude of deep convection-permitting (>28°C) SSTs in the Pacific.

ENSO diversity is defined here as the representation of El Niño events with sea surface temperature anomalies centered on the central Pacific (CP El Niños) and those with SST anomalies centered on the eastern Pacific (EP El Niños). In general, CP El Niños have a smaller impact on PNW precipitation and temperature than do EP El Niños (Patricola et al., 2020). Two more ELI-based metrics are included to more directly assess how well ENSO is represented in

each model. The DJF average ELI for each year is used for these two metrics. For one of these metrics, the distribution of DJF ELI for each model ensemble member is compared to the distribution from observations by using Levene's test to determine the likelihood that that modeled distribution is drawn from the same distribution as observed. The other metric is simply the median value of the DJF ELI time series. We find that all CMIP6 models share an eastward bias in ELI, meaning ENSO diversity is skewed toward EP El Niños in the CMIP6 ESMs.

The 20 global and 22 regional R13 metrics are combined with the 6 global and 4 regional additional ENSO metrics to form the 52 metric combined suite used in this evaluation. Once these metrics are computed, they are combined using the relative error formulas as in R13, recreated here, slightly modified in order to clearly denote the comparisons between each ensemble member. For many of the metrics, the mean of the verification data metric can be directly compared to each ensemble member of a given model. However, for the ENSO teleconnection spatial correlation metrics (Nino3.4-pr r, Nino3.4-T r, ELI-pr r, and ELI-T r), the seasonal spatial correlation metrics (SpaceCorr MMM-pr, SpaceCorr MMM-T), and the seasonal spatial standard deviation metrics (SpaceSD MMM-pr, SpaceSD-MMM-T), each model and ensemble member must be compared to each verification dataset, in turn, to avoid washing out the spatial variability by taking the mean of the verification data prior to those comparisons and potentially favoring models with a larger number of ensemble members (or a number similar to the number of verification datasets).

The error for *i* metric, *j* model, and *k* ensemble member for all metrics except those specified in the previous paragraph is given by Eq. 1a:

$$E_{i,j,k} = \left| x_i - y_{i,j,k} \right| \qquad (1a)$$

where x_i is the mean observed metric value and y is the model metric value. For the metrics specified in the previous paragraphs, Eq. 1a for each l verification dataset takes on the form:

$$E_{i,j,k} = \frac{1}{L}\sum_{l=1}^{L}\left|1 - y_{i,j,k,l}\right| \qquad (1b)$$

for L verification datasets. In the case of the spatial correlation metrics, y is first computed against each verification dataset independently, while for the spatial standard deviation metrics, y is normalized by the standard deviation of each verification dataset, in turn. The relative error for each metric, model, and ensemble member is then:

$$E^{*}_{i,j,k} = \frac{E_{i,j,k} - min\,(E_{i,j,k})}{(E_{i,j,k}) - min\,(E_{i,j,k})} \qquad (2)$$

with minima and maxima determined independently for each metric across all models and ensemble members. The total relative error for a given model is then computed as the sum of the ensemble mean relative error:

$$E_j^{tot} = \sum_{i=1}^{M}\frac{1}{K}\sum_{k=1}^{K}E^{*}_{i,j,k} \qquad (3)$$

for K ensemble members and M metrics. This summed relative error is then normalized to give the normalized error score, which ranges from 0 to 1:

$$E_j^{norm} = \frac{E_j^{tot} - min\,(E_j^{tot})}{max\,(E_j^{tot} - \left(E_j^{tot}\right))} \qquad (4)$$

A comparable spread of ensemble members for each model can be computed by summing the relative error from Eq. 2 over all metrics and applying the same normalization as in Eq. 4. Note that this allows for individual ensemble members to acquire normalized error scores less than 0 or greater than 1, as this distribution is normalized by the ensemble mean summed relative error values. This framework is designed to be flexible and allow a regional evaluation to be performed for any region on the globe, or even for component-based metric definitions (if some

form of dimension reduction is used). For this paper application, the evaluation focuses on a climate change projection application in the US PNW. The domain over which the regional metrics are computed and the culling based on regional trends is applied and shown in Fig. 1. While the PNW is evaluated here, the core methodology of this evaluation could be used for any region. Error scores are calculated from both PNW regional metrics and global metrics in the final evaluation. This is done to ensure that ESMs selected for regional performance have also met a minimum threshold of performance at the global scale, and it is recommended to include global metrics for any regional analysis to ensure that physical processes are being correctly represented at multiple spatial scales.

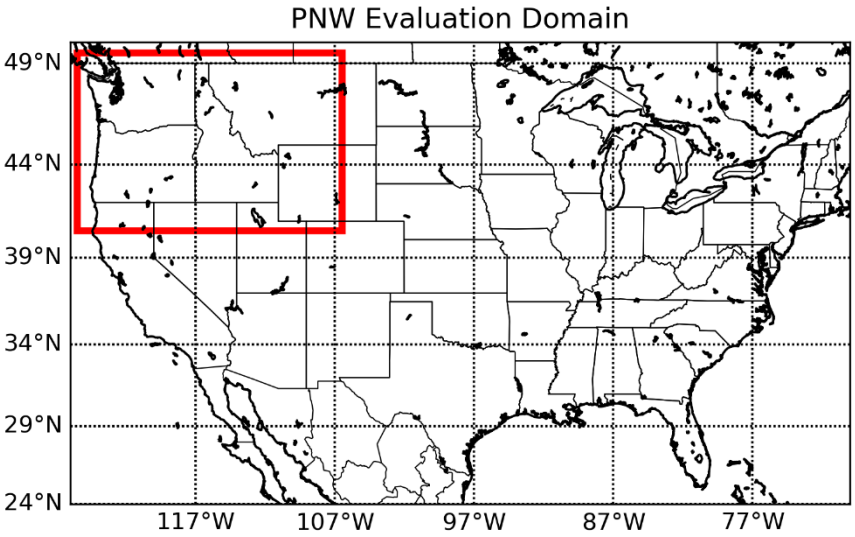

Figure 1: The PNW domain (red box) used for the computation of regional metrics and for the culling criteria based on historical temperature and precipitation trends.

In addition, we perform split-sample and perfect model evaluations of the ESM and projection data to increase our confidence in our model selection and culled projections. In the split-sample analysis, the PNW regional analysis and the global analysis is repeated using only

the period 1901-1950, with 1950-2014 serving as the period over which the trend envelope of the culled ensemble is verified. Each of these periods (50 and 65 years respectively) is long enough to ameliorate to some extent the influence of unforced variability, though more so for global than regional metrics. For the perfect model framework (e.g., Sanderson et al., 2017, Liang et al., 2020; Anderson 1996), each CMIP6 model, in turn, serves as the verification dataset. In this case, individual ensemble members of that model take the role of an individual observational dataset. The perfect model experiment allows a check on whether our evaluation framework is shown to have skill in selecting a subset of the ESMs that reliably predicts the trends that the "perfect model" expresses in the SSP scenarios, and assesses ESM response similarity. It is predicated on the idea that in all the models, the metrics relate similarly to the model's climate sensitivity.

3.2 Culling Method

We consider the case of model culling, or binary weighting, because one of our aims is to provide an objective framework for reducing the number of ESMs considered for a given regional hydrometeorological application, and the many nuances of model weighting are beyond the scope of this analysis. To evaluate the impacts of culling methods, we analyze projected future temperature and precipitation trends, which are crucial to hydrometeorological applications and an appropriate use for ESMs. Fig. 2 shows the projected trends over the years 2015-2100 (1901-2014 for the historical trends) in the PNW, and for global land gridpoints for all CMIP6 models included in this study for the historical, SSP2-4.5, SSP3-7.0, and SSP5-8.5 runs. There is a wide spread in the projected trends, especially for precipitation, due not only to the uncertainty across the SSPs, but also the model spread within each SSP. Culling therefore

runs the risk of misrepresenting the uncertainty of possible futures, particularly extremes, thus any criteria applied should represent this uncertainty in a scientifically defensible way.

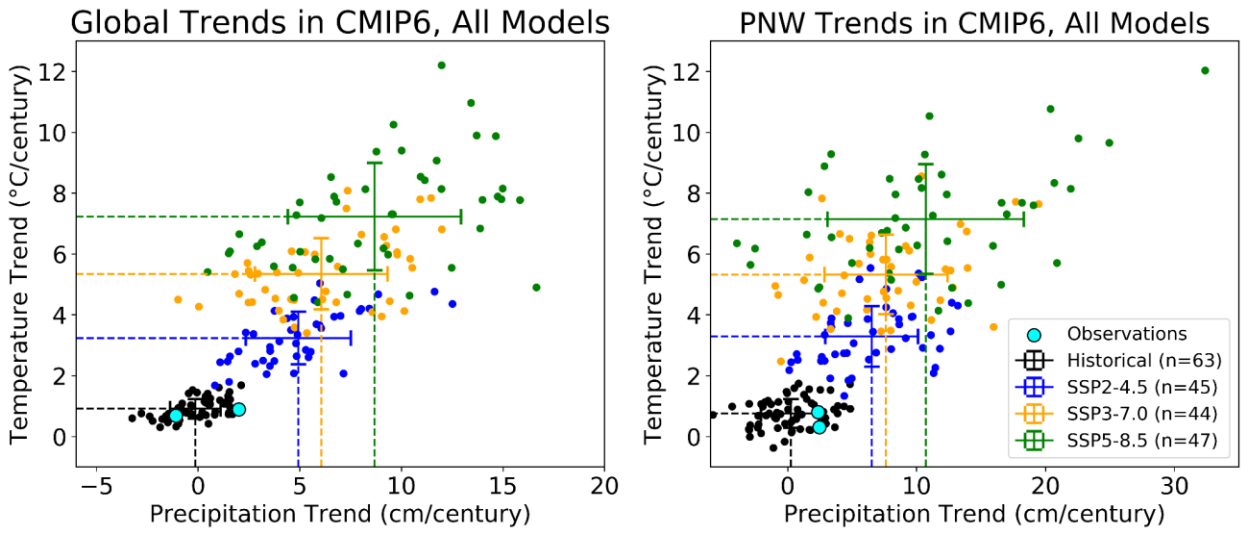

Figure 2: Global (left) and PNW (right) temperature and precipitation trend projection envelopes for historical and SSP CMIP6 runs (each denoted by a plotting point) over land gridpoints within the respective domains. Each point in this figure represents the ensemble mean projection for a given CMIP6 model. The multi-model mean trends are highlighted by dashed lines extending to the respective axis for clarity. Error bars represent ±1 standard deviation of the multi-model ensemble about the multi-model mean. The historical trends are calculated over the period 1901-2014, while the SSP trends are calculated over the period 2015-2100. CRU and UDel observations are included as cyan points coinciding with the historical CMIP6 envelope. Note that the range of the x-axes differ.

Due to the importance of accurate trend representation for ESM future climate applications, regional precipitation and temperature trends during the historical runs of the CMIP6 models are used to develop a novel criterion for reducing ESM ensemble size. Once

model rankings are computed, the CMIP6 model average trends are computed as a function of the model ensemble size. That is, beginning with the best performing ESM, models are added to the average trend calculation in order of their ranking until all CMIP6 models are included in the ensemble. For each ESM added to the ensemble, the precipitation and temperature trend error is calculated as the absolute difference between the verification dataset mean trend and the ESM ensemble trend, normalized by the standard deviation of those error values. These normalized errors are then added together to determine the total historical trend error as a function of ensemble size. The optimal ensemble size is determined by the minimum of this total historical trend error, with an additional requirement that the culled envelope exceed 10% of the total ensemble size due to the higher volatility of trend projections for envelopes smaller than this threshold.

4. Results

Here we present each component of our evaluation in order of their complexity, moving from individual metric plots through our full perfect model projection reliability evaluation. First, the results of the metric suite as applied globally and to the PNW region (Fig. 1) are shown alongside the performance of each verification dataset with respect to the observational mean in order to demonstrate the variety of responses seen in the CMIP6 ensemble and the model uncertainty as compared to the observational uncertainty. Next, this metric suite performance is aggregated as described in Section 3, using the relative error methodology of R13 (Section 4.1). The CMIP6 model ensemble is then culled based on the historical precipitation and temperature trends, and the effect of this culling on the projection envelope of future trends is evaluated (Section 4.2). This same methodology is then applied using only the period 1901-1950, using 1950-2014 for verification, in order to demonstrate the efficacy of this method in retaining

important features of the projection envelope using a fraction of the CMIP6 models (Section 4.3). Finally, a perfect model framework is applied to assess the reliability of this method in choosing models with projected trends similar to the "perfect" model, as well as assessing the similarity of response for each model as compared to each other model in the CMIP6 ensemble (Section 4.4).

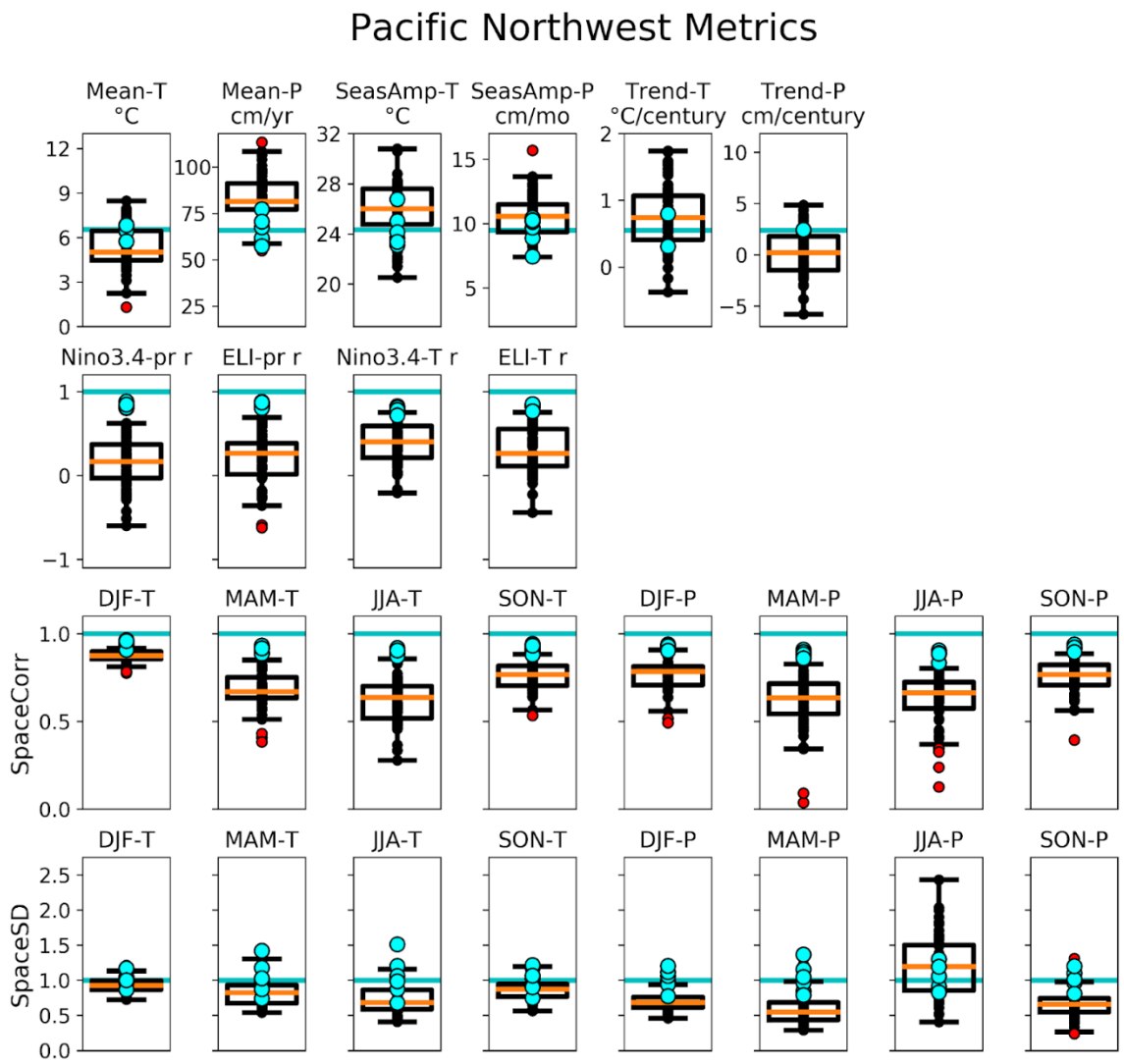

Figure 3: Box plots representing the 25th and 75th percentile of ensemble mean PNW metric performance for all CMIP6 models considered. The median of CMIP6 performance for each metric is shown by the horizontal orange line. The lower (upper) whiskers correspond to metric

values representing the 25th (75th) percentile minus (plus) 1.5x the interquartile range, with red points demarcating outlier model ensemble mean values that fall beyond this range. The cyan points represent the six verification datasets (two for the precipitation and temperature trends, as discussed in Section 2.2), while the black points represent each model's ensemble mean metric performance. The "target" value representing the observational mean or perfect correlation is shown as a cyan line.

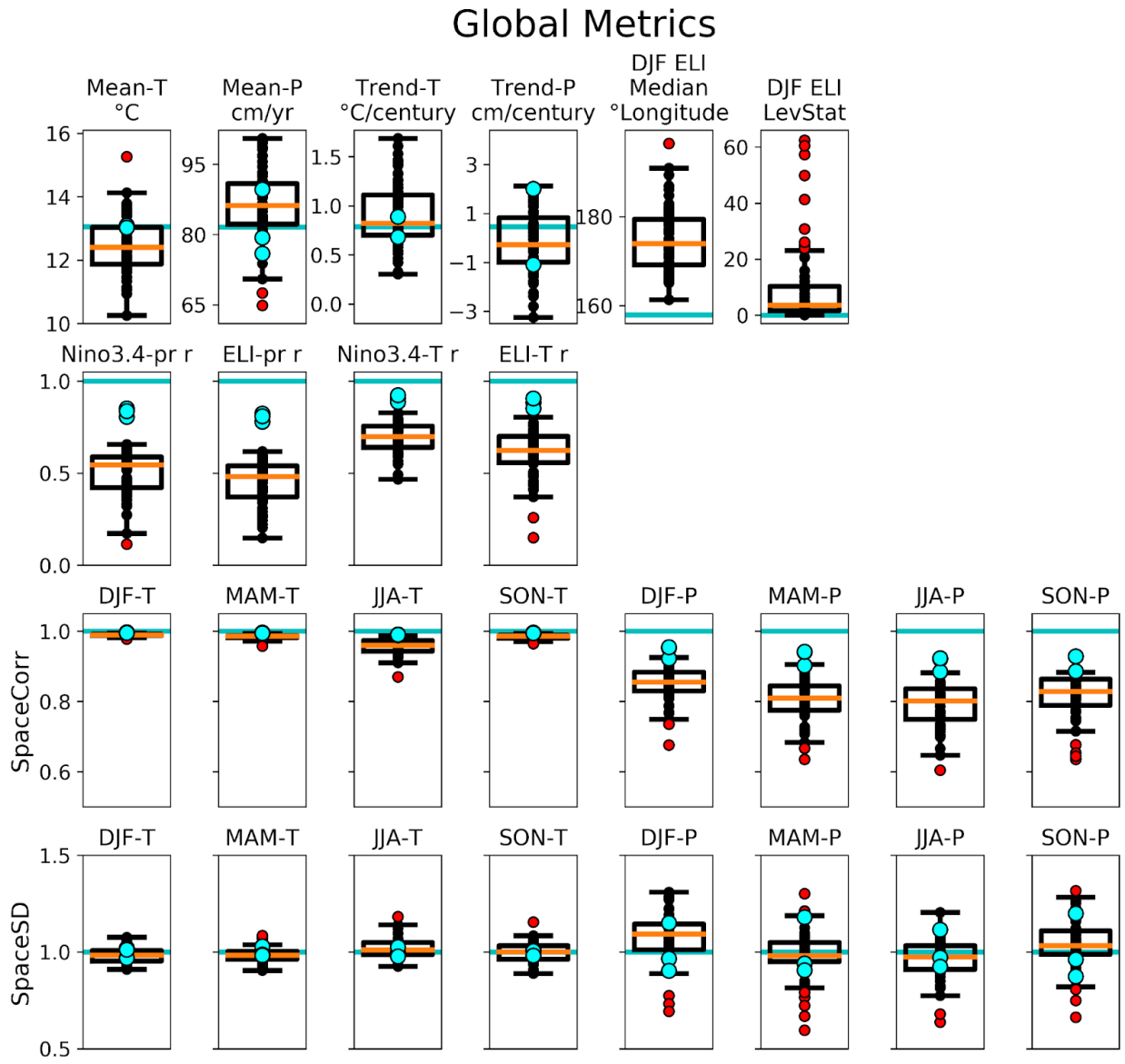

Figure 4: As with Figure 3, but for global metrics. Verification datasets here include only CRU, UDel, and ERA5 due to the global coverage of these data. Global trends, like regional trends, are verified only against CRU and UDel, as discussed in Section 2.2. ELI is computed from ERSSTv5 and Nino3.4 is computed from HadISST1, as discussed in Section 2.2.

4.1 Global and Regional Metric Performance

The distribution of ensemble average performance for each model for the period 1900-2014 over the PNW is shown in Fig. 3. Also shown in cyan is the performance of each verification dataset relative to the ensemble mean of these datasets, represented by the horizontal cyan line. It is clear from Fig. 3 that the spread of the verification datasets is generally smaller than the spread of the model performance. As expected, models tend to perform better for temperature metrics compared to precipitation metrics. The annual mean temperature and linear trend of temperature lie quite close to the observed distribution, but the models tend to overestimate the mean annual precipitation while underestimating the linear trend of precipitation. Fig. 4 shows the performance of CMIP6 ESMs over the globally applied metric suite. Due to the wide variance of the seasonal cycles of precipitation and temperature at different locations across the globe, these metrics are not included in the global metric suite. The global metric box plots show once again that CMIP6 models generally capture temperature metrics much better than precipitation metrics. On a global scale, the mean annual temperature tends to be underestimated, while the mean annual precipitation is slightly overestimated by the ensemble mean. On the other hand, modeled precipitation trend uncertainty is much closer to the observational uncertainty than for the global temperature trend, with a slight bias toward more warming than the observational mean.

The newly developed ENSO teleconnection metrics demonstrate that many models are flawed in their representation of ENSO. Only a few models at the tail of the distribution in the PNW lie within the range of verification spread for temperature, while none do for precipitation, with some even showing negative spatial correlation values compared with the observed teleconnection pattern, while the verification datasets are very consistent with each other (Fig. 3). Globally, despite strong agreement between the verification data, no ESMs approach this performance for either temperature or precipitation (Fig. 4). Finally, normalized error scores for the combined global+PNW metric suite are shown in Fig. 5 for each model, with the ensemble spread represented by the colored points above and below each model's ensemble average error score.

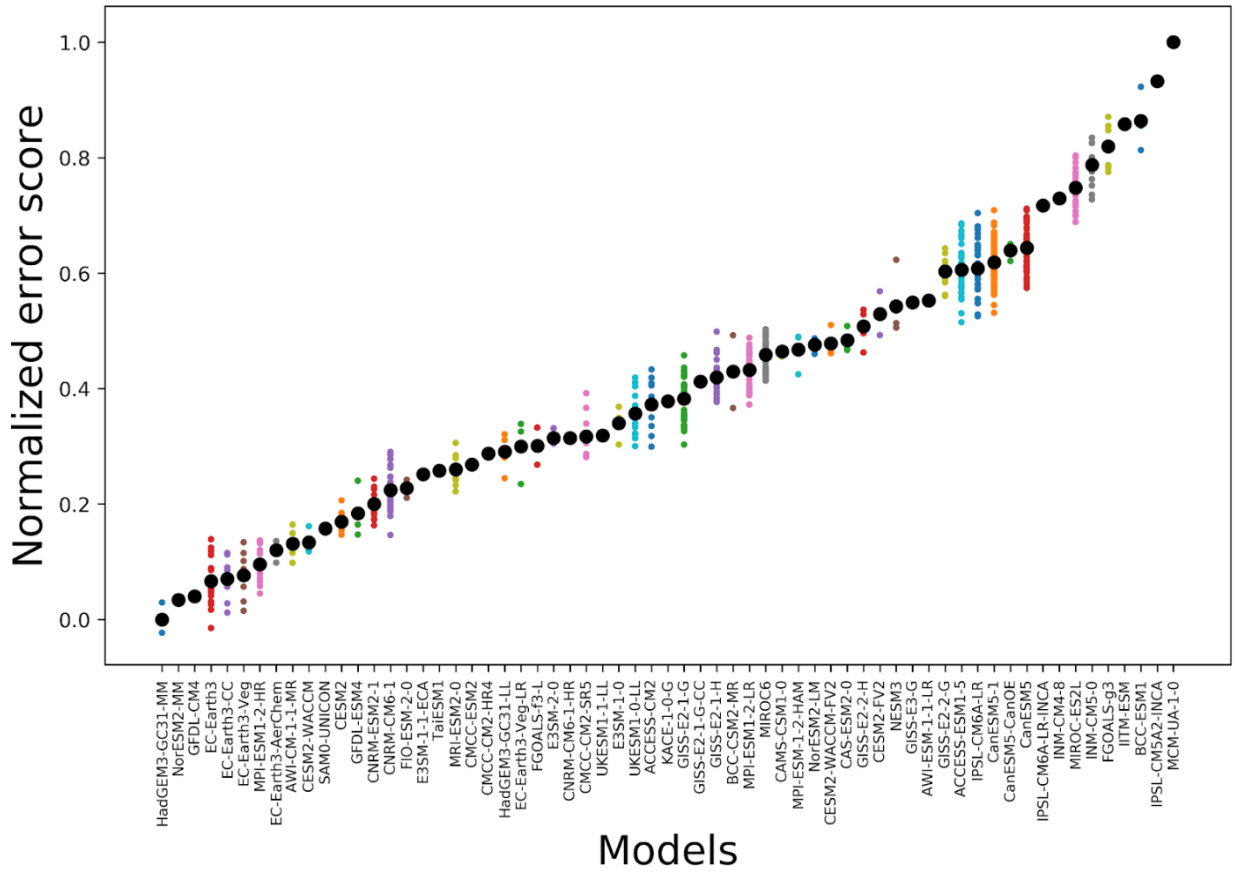

Figure 5: Model rankings based on the normalized error score of the 52 combined global and PNW metrics. The ensemble mean value for each model is shown as a bold black point, while multicolored points represent each ensemble member from that model.

4.2. Future projections from a culled ensemble

The historical trends over the PNW as a function of ensemble size are shown in Fig. 6, with the optimal ensemble size highlighted. Note that as the ensemble size grows, the ensemble average trend becomes less and less sensitive to the inclusion of additional models. Because not all modeling centers include every SSP in their model runs (see Table A1), the optimal envelope size is determined three more times using historical data only from models that include SSP2-4.5, SSP3-7.0, and SSP5-8.5 runs, respectively. These are shown alongside the full historical ensemble in Fig. 6. In each case, an optimal envelope of similar size (12-14 selected models out of 44-63 total models) is found. While this figure shows that even the best performing models show considerable differences in trend projections, and the combined precipitation/temperature trend error shows considerable variation, it also shows that for this application, a smaller number of high performing models reproduces observed trends more accurately than the full CMIP6 suite, particularly for precipitation trends.

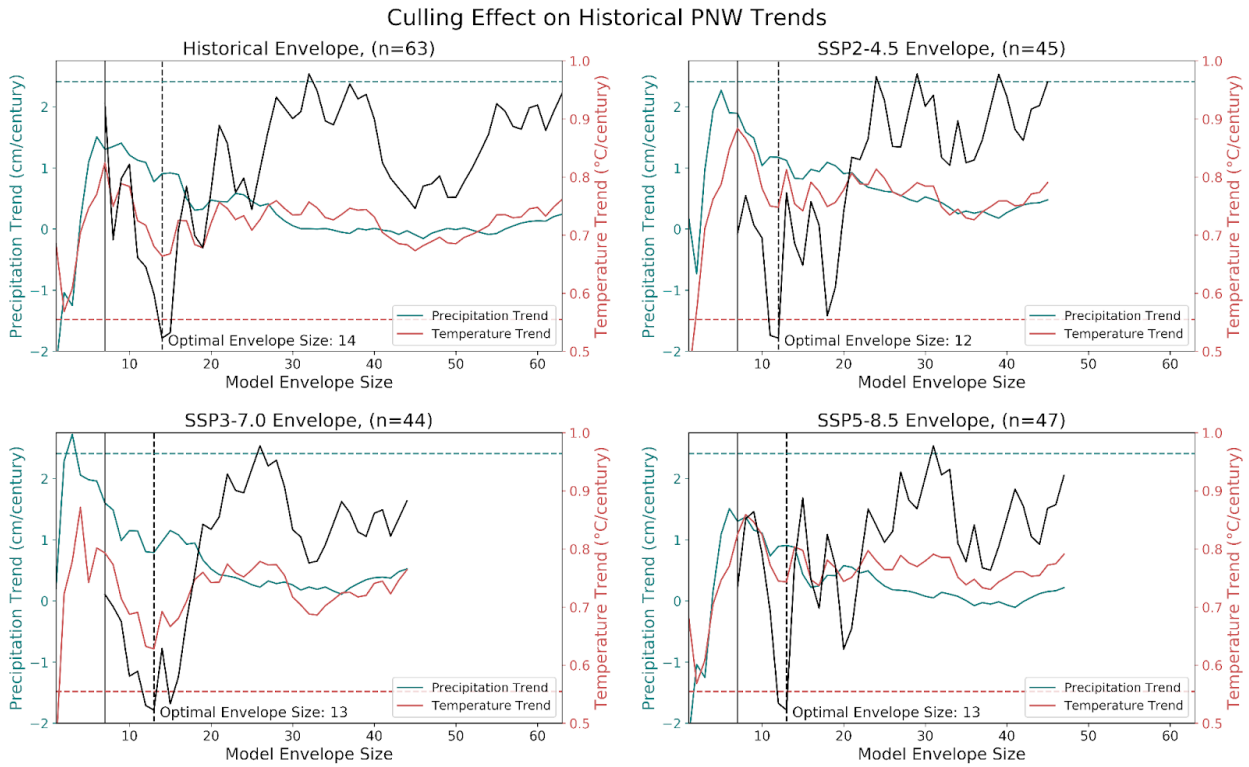

Figure 6: Historical PNW precipitation (blue, left y-axis) and temperature (red, right y-axis) trends as a function of CMIP6 ESM ensemble size. In black is the combined error (axis not shown), minimized to find the optimal envelope size. The solid vertical black line marks the minimum size for the optimal envelope, defined at 10% of the total ensemble. The leftmost points include only the top performing models based on the normalized error scores shown in Fig. 5. The mean of the verification dataset trends are shown as horizontal dashed lines. The black vertical dashed line represents the optimal envelope size determined by minimizing the difference between the ensemble average trend and the observational average trend. The top left plot uses all models in the historical ensemble, while the other three include only models that provided data for the respective SSP run shown.

The effect of this culling criterion on the precipitation and temperature trend projections for the PNW as applied to the PNW+global ESM metric performance ranking is shown in Fig. 7.

In each SSP, the ESMs with the most extreme trend projections tend to be culled, especially in SSP5-8.5, where the culled ensemble precipitation trend is reduced substantially. While the culled ensemble mean temperature trend is barely affected compared to the full ensemble for any scenario, the standard deviation of the culled ensemble temperature trend projections is reduced in each scenario. For the precipitation trends, differing behavior is seen in the culled ensemble depending on the scenario considered, with a slight increase in the culled ensemble mean projection in SSP2-4.5 and SSP3-7.0, and a decrease in the culled ensemble mean projection in SSP5-8.5. This method as applied here tends to selectively cull ESMs with the most extreme wetting trends, especially in SSP5-8.5, while retaining several ESMs with the least extreme wetting trends or drying trends. Such an asymmetry is not seen for the temperature trends, where the culling method tends to remove models with the coolest warming trends and the hottest warming trends. This result demonstrates that the culling method does not greatly affect the features of the central tendency of the distribution of projected trends, especially for temperature, in turn giving confidence to decision-makers that the center of mass of projected trends is well represented in the culled ensemble, despite the culled ensemble being composed of only 12-13 ESMs. Thus using the culled ensemble for hydrological impact studies would greatly reduce the sample size and remove outlier models without greatly affecting the central tendency of the trends from the full CMIP6 ensemble. Whether the reproduction of the central tendency is enough for a given impact application would have to be assessed on a case-by-case basis.

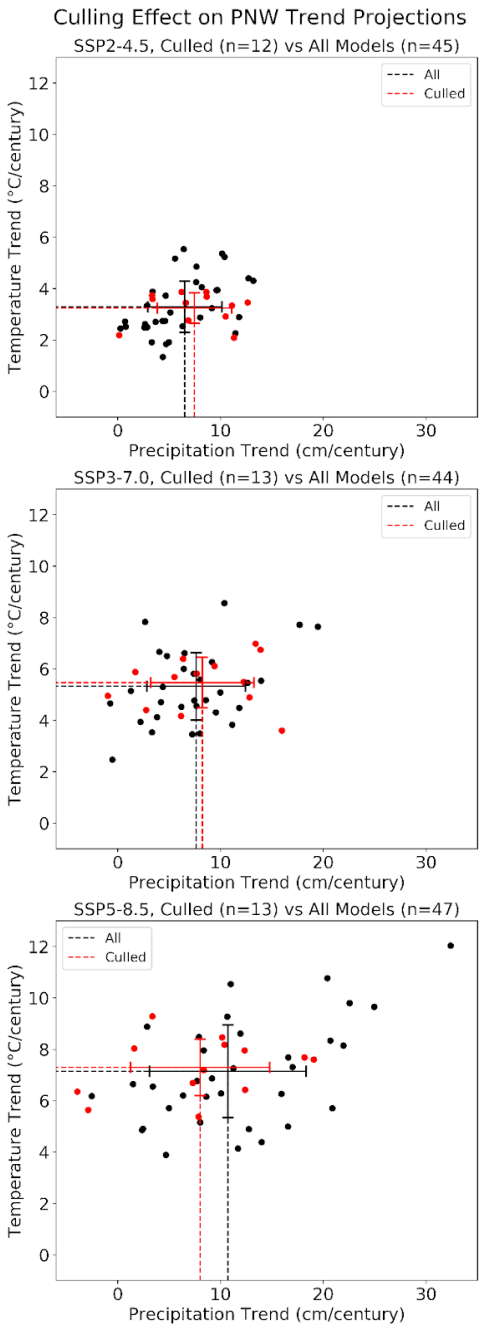

Figure 7: Projected precipitation vs temperature linear trends over the PNW for SSP2-4.5 (top), SSP3-7.0 (middle), and SSP5-8.5 (bottom) for the full model ensemble (black) and the culled model ensemble (red), which includes only the top performing models (based on the combined global-PNW normalized error scores) selected by the optimal envelope size criteria. Trends are computed over the period 2014 - 2100.

4.3. Split Sample Analysis of Culling Methodology

Because SSP projections of future trends cannot be directly verified, a split sample method is used here, separating the historical period into a "training" period (1901-1950) and "verification" period (1950-2014). In this case, the ESM evaluation is performed using the same PNW+global metric suite, but using only the CRU and UDel datasets for the error calculations, as these are the only verification datasets with the required temporal coverage for this analysis. This method allows determination of the fidelity of the culled ensemble trend "predictions". As done with the full historical period, the culling effect on ensemble mean trends as a function of culled sample size is computed, with the optimal ensemble size being 14 in this case as well (Fig. 8). The effect of the selection method on the "projected" trends during the verification period as compared to the full ensemble is shown in Fig. 9. While it is found that the culled sample results in a slight deterioration of the ensemble mean temperature and precipitation trends as compared to the observed CRU and UDel trends over this period, likely due at least in part to the very small observed trends, we do find that this method again captures the center of mass of the full ensemble, and in this case retains some extreme behavior as well, particularly for precipitation. Given the uncertainty in the observations of even the direction of precipitation trends during this period, it should not be surprising that the model uncertainty in precipitation trend projections is quite wide. The mean of the projected precipitation trend is increased to 0.82 cm/century from 0.53 cm/century using the full ensemble, while the standard deviation of the culled distribution is reduced by only 13%. The mean of the projected temperature trend is more strongly affected with the culled ensemble mean being ~2.2 °C/century and the full ensemble mean being ~2.0 °C/century, with a 34% reduced standard deviation as well. Still, we do find that even with a limited subset of the ESMs (n=14), this method gives similar precipitation trend predictions

using only ~20% of the model ensemble, while retaining representation of the center of mass of temperature trends, albeit with a bias toward models with stronger warming than observed. In addition, 12 of the 14 ESMs selected by the 1901-1950 PNW+global metric evaluation are found in the 1901-2014 PNW+global culled ensemble. These model selection criteria are therefore found to be relatively insensitive to the time period considered, even though several verification datasets are based only on the last few decades of data.

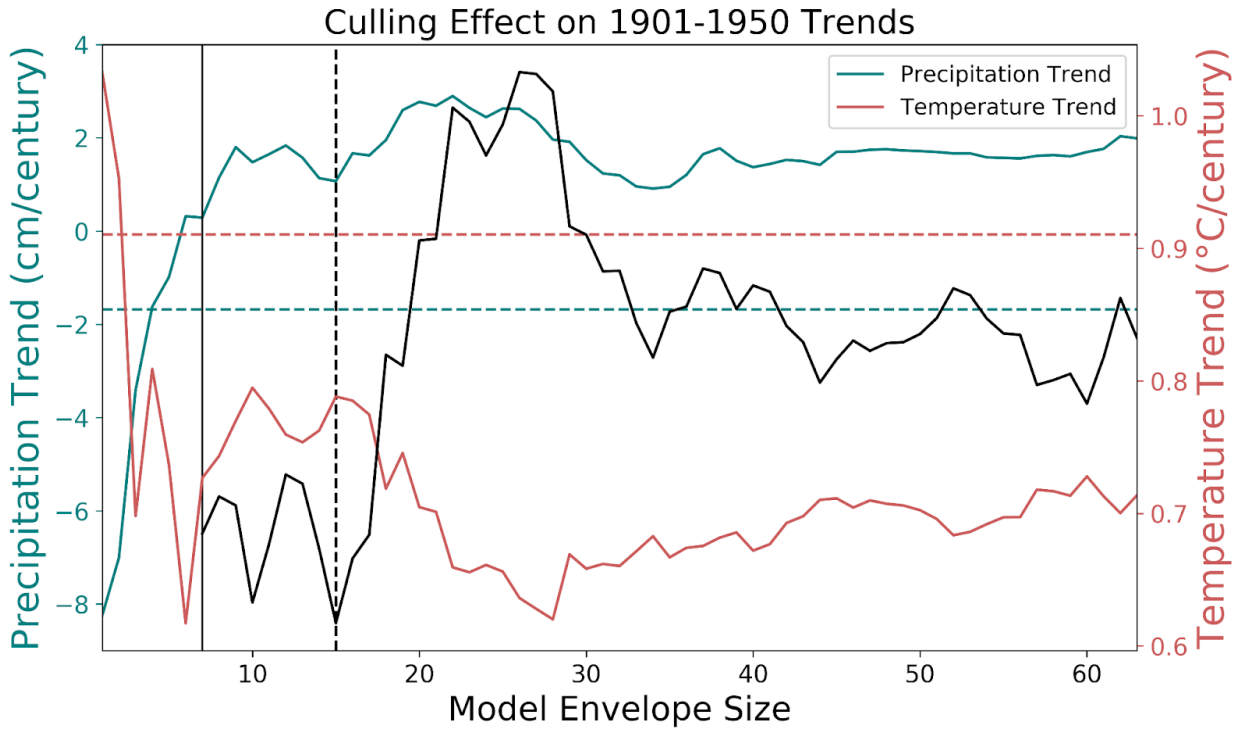

Figure 8: As with Fig. 6, but with the global+PNW metrics applied only over the period 1901-1950.

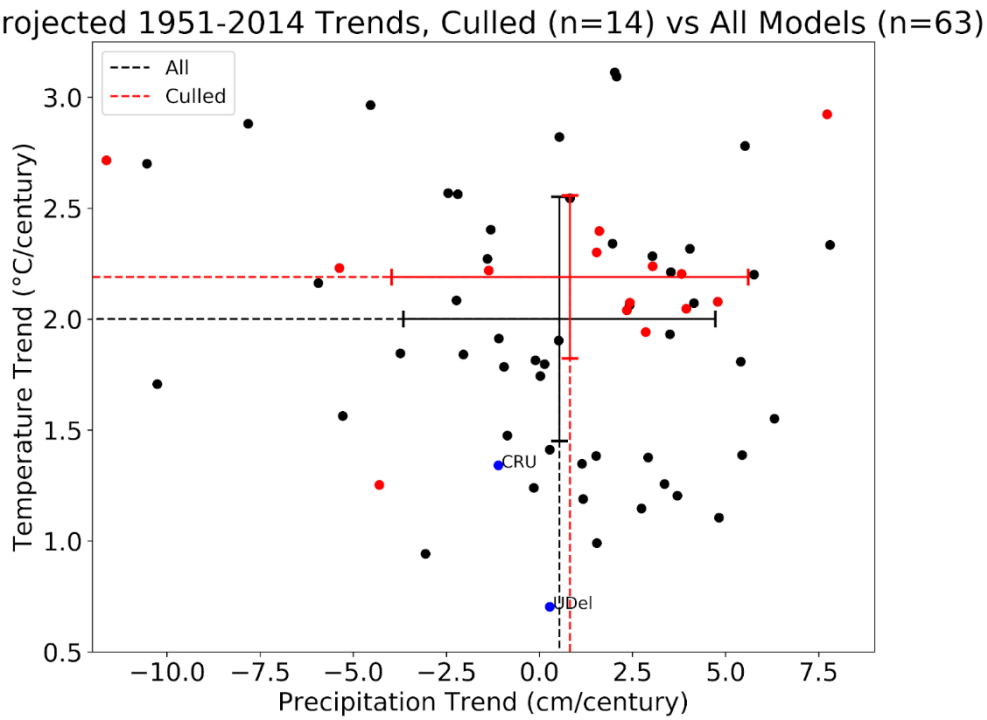

Figure 9: As with Fig. 7, but for the split sample analysis as applied to the combined PNW+global metric suite and culled using the PNW precipitation and temperature trend criteria over the period 1901-1950.

4.4 Perfect model evaluation

The metric suite evaluation is then applied in a perfect model scenario, wherein each model, in turn, is considered to be truth, with each ensemble member of a given true model being treated as an individual observational dataset (e.g. Liang et al., 2020; Suarez-Gutierrez et al., 2021; Lenderink et al., 2023). This framework allows, in an overarching sense, a test of the metric suite's ability to select for models with realistic representation of processes important to temperature and precipitation trends by allowing the "verification" of trends in SSP runs of the truth model. In addition, this analysis acts as a test of the similarity between the perfect model and all other models, selectively choosing for models with PNW+global metric performance similar to the other models in the CMIP6 ensemble. For each perfect model, the ensemble

members of that model act as though they were each a different dataset representing observations. Normalized relative error calculations for the other 62 models are then computed by comparing the perfect model's ensemble members to the ensemble members of each of the other models, in turn. This outputs a model ranking based on the ensemble mean relative error score for each other model in the CMIP6 suite, ultimately resulting in 63 different sets of model rankings of the other 62 models.

The distribution of these rankings, organized by the mean ranking for a given model compared to every perfect model, is shown in Fig. 10. For each perfect model, the mean absolute error between the projected trends for the three SSP runs in the perfect model and those in each evaluated model is computed. The distribution of these mean absolute errors for all perfect models in each SSP is shown in Fig. 11. Also shown in Fig. 11 is the mean absolute error distribution between the culled ensembles and the perfect model, using the optimal envelope size computed for the PNW comparison to observations. This figure demonstrates a tendency for the culling method to select for models that better match the projected temperature trend of the perfect model while maintaining a similar spread of projections as the full ensemble. For the precipitation trends, the absolute errors in the distribution of culled ensembles are largely unaffected. From these data, the containing ratio is calculated: that is, the ratio of the perfect models that lie within the spread of the respective culled ensemble. This is a measure of the reliability of this method to select for a culled envelope which includes the "truth" in its projection spread. For SSP2-4.5, this ratio is 0.84 for temperature trends and 0.87 for precipitation trends. For SSP3-7.0, the ratio is 0.90 for temperature trends and 0.86 for precipitation trends. Finally, for SSP5-8.5, the ratio becomes 0.83 for temperature trends and 0.92 for precipitation trends. These ratios indicate skill of this method at selecting an appropriate

subset of models that include the “truth” in its projection envelope. Along with the reduction in temperature trend error and the spread of the culled ensemble seen in Fig. 11, this method is shown to have skill in selecting for models with projected trends that match the projections of the perfect model.

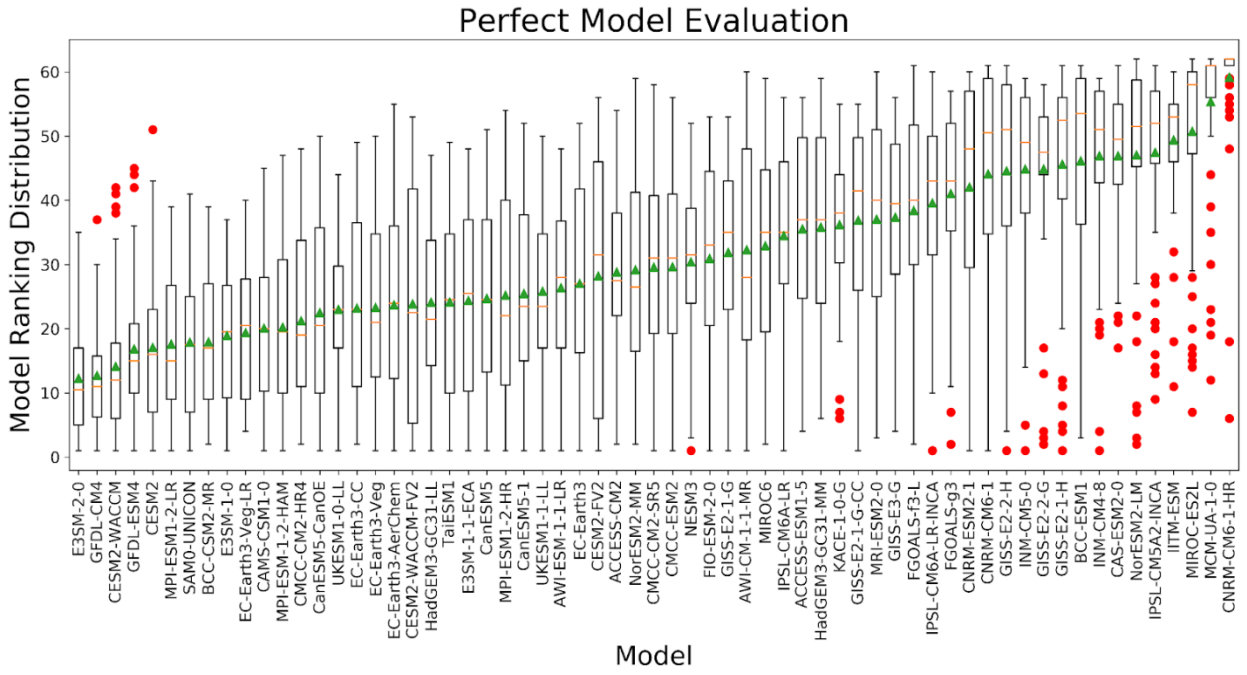

Figure 10: Distribution of rankings given to each model by evaluation against all other models. The models are organized by the mean of the ranking given from the evaluation against all other models (green triangle). The 25th and 75th percentiles are shown by the boxes, while the whiskers represent those percentiles ± 1.5x the interquartile range. The median value is shown by the orange bar. Red dots represent rankings outside the 25th and 75th percentiles ± 1.5x the interquartile range. For each model on the x-axis, the distribution shown consists of 62 data points, representing the ranking of that model as compared to every other “perfect” model.

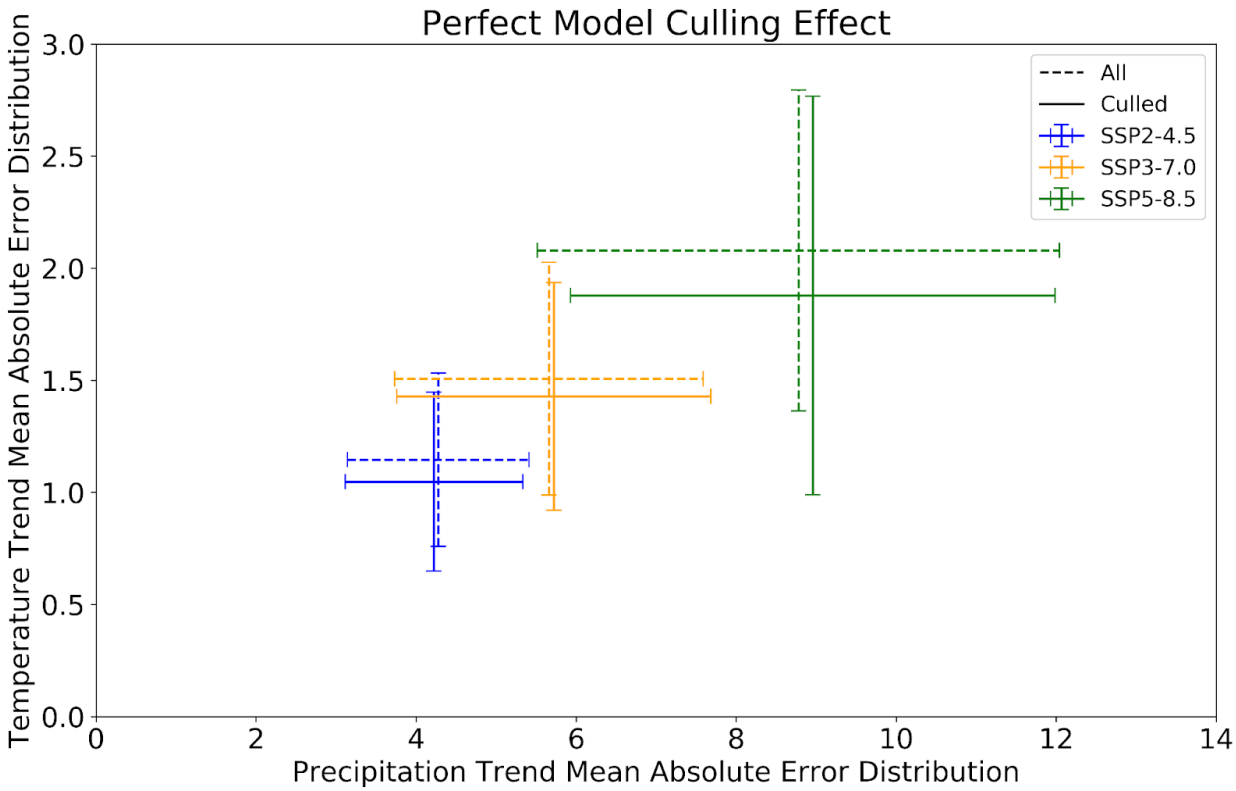

Figure 11: Distribution of mean absolute errors in projected temperature trends (y-axis) and precipitation trends (x-axis). Each data point in a given distribution represents the mean absolute error of the projected trends of all models with respect to a given perfect model ensemble average. The dotted lines represent the full model ensemble, while the solid lines represent the culled ensemble.

5. Summary

A modified version of the R13 higher confidence metric suite is used as a base to develop a flexible framework for ESM evaluation that was co-designed with the USACE Climate Resilience and Preparedness Program, other users, and initial input from the community (Newman et al., 2022). We incorporate several new ENSO metrics using the canonical Nino3.4 index as well as the newly developed ELI, which represents processes relating to ENSO diversity and important to CONUS teleconnections differently than Nino3.4 index. This framework can be

easily modified to be applied to any region of interest across the globe. We develop a new, potentially useful criteria for model culling based on applying thresholds of historical precipitation and temperature trend errors to pre-ranked model scenarios. When this method is applied to the PNW CONUS using a joint PNW+global metric suite, it is found that the culled ensemble retains the mean and standard deviation of the full CMIP6 ensemble despite being composed of only ~20% of the total number of CMIP6 models. We applied the method to the PNW over the period 1901-1950, using 1950-2014 for verification. The culled ensemble exhibits a stronger warming trend than both the full ensemble and the verification datasets, while the precipitation trends of the culled ensemble are very similar to the full ensemble. The split sample analysis evaluation contained 12 of the 14 models found in the culled ensemble as applied to the full historical period, demonstrating insensitivity of the culled ensemble to the historical time period chosen, in turn giving confidence in our metric suites' insensitivity to internal variability.

We also applied our evaluation method within a perfect model scenario experiment, treating each CMIP6 model in turn as the verification dataset. This provides another way to verify the SSP projections and thereby assess the metric suite's skill in selecting for models that represent processes similarly to the verification, and whether that skill is reflected in climate change projections. In this case, the culling method tends to reduce the error in projected temperature trends for all SSP runs, while having less effect on the projected precipitation trends. However, as in other applications, the distribution of the projected trends is maintained with a much smaller envelope of models considered. This perfect model evaluation can be used to inform certain impact applications as to the uniqueness of a given model response, and should be used jointly with the model rankings as compared to observations depending on the desired hydrometeorological impact application as model response and genealogical similarity is

generally recognized as an important criteria (Knutti et al., 2013; Merrifield et al., 2023). Comparing the ranking distributions of Fig. 10 with the rankings determined from evaluation with respect to observations shown in Fig. 5 yields some interesting information. For instance, CNRM-CM6-1-HR tends to rank poorly in the perfect model evaluation, despite being near the center of the rankings in Fig. 5. This indicates that despite being generally dissimilar to other models, it still ranks relatively highly as compared to observations, suggesting more value for impact applications than would be expected based on its observational ranking as it is most dissimilar from the other ESMs. On the other hand, E3SM-2-0 is ranked almost exactly the same as CNRM-CM6-1-HR in the observational comparison, despite being the most similar to the other models in the CMIP6 ensemble, suggesting it is providing less unique information. These examples serve to demonstrate that, for a given application, users should consider using the information contained in Fig. 5 and Fig. 10 jointly depending on the range of model response in which they may be interested and how they may want to incorporate model response into their selections. Future studies could further explore model uniqueness impacts on projection selection.

6. Discussion

Similar to the ESM evaluation tools available, our method and code is easily extensible to include other observational datasets and metrics. For example, oceanic heat content (OHC) datasets from both observations and for CMIP6 models are becoming available (e.g. Lyu et al., 2021). We did not include OHC here as the Lyu et al. (2021) open-source dataset only included 28 of the CMIP6 models. However, of the 12 models with either excessively high or low OHC trends as defined by +/- 1 std deviation of observed OHC trend uncertainty, only five would be

retained in our analysis depending on user decisions related to model representativeness (e.g., CESM2 and CESM2-WACCM are retained in our rankings).

By definition our culling method removes poor performing (outlier) ESMs, which can be seen by cross-referencing Figures 5 and 6. Further, examination of the precipitation and temperature trends across SSPs (Fig. 7) highlights that our method removes many (but not all) of the outlier models for end of century change signal, tends to preserve many models in the 'center of mass' of the CMIP6 full model ensemble and retains similar spread characteristics to the full ensemble, which may also be expected (e.g., Sanderson et al., 2017). This is a positive characteristic as noted above, retaining the mean projection and spread with a fraction of the models implies a potentially significant cost savings for impact projection generation. However, some of the most extreme projections are removed, which may be detrimental to particular types of risk assessments, such as full system stress tests designed to identify potential futures with vulnerabilities (Brown et al., 2012; Steinschneider et al., 2015). Therefore we again stress that users of this method be mindful of their specific application needs and how that meshes with the assumptions and behavior of this (and any) evaluation methodology, so they may supplement or modify their workflows appropriately. For example, one could use the culled ensemble and then re-introduce particular outlier projections to fit any known or explore unknown specific installation vulnerabilities.

Another metric, or culling decision, could be the equilibrium climate sensitivity (ECS). There has been much discussion that ESMs with ECS values above roughly 4.5 °C are too sensitive to climate forcings and many ESMs may be overestimating the recent observed warming since 1980 (e.g. Nijsse et al., 2020; Zelinka et al., 2020; Meehl et al., 2020; Tokarska et al., 2020, Scafetta, 2022). However, it is unclear if high ECS should be a disqualifying

characteristic for regional applications. Exclusion or inclusion of high ECS models is particularly complicated for water security applications. Asenjan et al. (2023) found that including high ECS models for hydrologic change studies significantly changed the projections in only some of the regions they examined. Here, six of our top twenty models are from only two distinct modeling systems (CESM/E3SM and CNRM-CM6/ESM2) (Fig. 5) that have an ECS greater than 4.5 °C, (CESM2, CESM2-WACCM, E3SM-1-1-ECA, CNRM-ESM2-1, CNRM-CM6-1). We include observed global temperature trends as a metric where the high ECS models do relatively poorly (not shown), but they generally perform well for regional metrics across the PNW, highlighting the complexities of regional evaluations using ESMs. Note that model response, using perfect model or other response similarity metrics (e.g. Sanderson et al., 2017) and genealogical similarity could further reduce the hot models retained as a second culling step as needed.

Daily metrics could be included if found to be necessary for a specific application. However, in this study we did not include daily metrics for two primary reasons. The first is pragmatic; we desire to be as inclusive as possible in the number of CMIP6 models and ensemble members in our evaluation, and many modeling groups provide daily data for only a few simulations. Second, very few if any water security climate change impact assessments use ESM output directly, the ESM data are statistically bias corrected and downscaled, or dynamically downscaled (and often then statistically bias corrected) because of the substantial errors in ESM data from this perspective. Additional inclusion of non-trend metrics also does not test ESM change projection fidelity, and it is unclear if there would be any added discriminatory power to identify additional poor performing models. For example, Wehner et al. (2020) and Wehner (2020) evaluated CMIP5 and 6 models for historical and future changes of daily precipitation and temperature extremes and found no significant differences between the two

generations of models, which could indicate a lack of discriminatory power. It would be worthwhile to investigate the additional information content of daily data, including changes in daily fields, above the metric set used here in future work. Finally, this ESM evaluation effort is part of a broader multi-project effort to provide quantitative guidance on the fidelity of core aspects of the climate impacts modeling chain (ESM, downscaling, hydrology) for water resource applications. The common objective is to co-develop verification-oriented strategies and approaches for designing or selecting models and methods based on their ability to robustly and reliably project future change -- which remains a challenge for the community. This builds off of initial efforts in the community to quantify the breadth of uncertainty in this impact modeling chain (e.g. Gutmann et al., 2012, 2014, 2022; Mendoza et al., 2015; Mizukami et al., 2016; Clark et al., 2016; Kao et al., 2022). Our co-designed evaluations also build on our well-developed researcher-agency-user relationships and falls within the broader literature finally recognizing the need for more 'fit-for-purpose' evaluations of ESMs, among other modeling systems (e.g. Parker, 2020; Briley et al., 2020; Findlater et al., 2021). Future work will explore the interplay between selection of ESMs, downscaling schemes, and hydrology models and assess subsequent projection spread and fidelity.

Acknowledgements

This work was funded by the US Army Corps of Engineers Climate Preparedness and Resilience Program and also supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977. We would like to acknowledge all participants of a small workshop discussing the important topic of ESM evaluation at the start of this work, as well as Chanel Mueller and

William Veatch of the US Army Corps of Engineers for their management of this project and input to enable more usable outcomes from this study.

Data Availability Statement

Climate model data from the Earth System Grid Federation (ESGF) CMIP6 data holdings (Cinquini et al., 2014) were used in the creation of this manuscript. In addition, the observational datasets listed in Table 1 are used for verification of the climate model data. Figures were made using Matplotlib version 3.1.3 (Caswell et al., 2020). Data analysis is supported by Xarray version 2022.3.0 (Hoyer and Hamman, 2017) , xESMF regridder version 0.6.2 (Zhuang et al., 2022), xSkillScore version 0.0.24 (Bell et al., 2021), SciPy version 1.8.0 (Virtanen et al., 2020), and NumPy version 1.21.5 (Harris, C. et al., 2020). Software used for plotting and analysis has been made available on Github at https://github.com/nlybarger/ESM_regional_evaluation (Lybarger, 2023).

References

Anderson, J. L. (1996). Selection of Initial Conditions for Ensemble Forecasts in a Simple Perfect Model Framework. *Journal of the Atmospheric Sciences*, *53*(1), 22–36. https://doi.org/10.1175/1520-0469(1996)053<0022:SOICFE>2.0.CO;2

Asenjan, M.R., Brissette, F., Martel, J.-L., & Arsenault, R. (2023). The Dilemma of Including “Hot” Models in Climate Impact Studies: A Hydrological Study (preprint). Hydrometeorology/Modelling approaches. https://doi.org/10.5194/hess-2023-47

Barnett, T. P., Pierce, D. W., Hidalgo, H. G., Bonfils, C., Santer, B. D., Das, T., et al. (2008). Human-Induced Changes in the Hydrology of the Western United States. *Science*, *319*(5866), 1080–1083. https://doi.org/10.1126/science.1152538

Basharin, D., Polonsky, A., & Stankūnavičius, G. (2015). Projected precipitation and air temperature over Europe using a performance-based selection method of CMIP5 GCMs. *Journal of Water and Climate Change*, *7*(1), 103–113. https://doi.org/10.2166/wcc.2015.081

Bell, R., A. Spring, R. Brady, D. Squire, Z. Blackwood, M.C. Sitter, & T. Chegini. (2021). xarray-contrib/xskillscore: Release v0.0.23 (v0.0.23) [Software]. Zenodo. https://doi.org/10.5281/zenodo.5173153

Brekke, L. D., Dettinger, M. D., Maurer, E. P., & Anderson, M. (2008). Significance of model credibility in estimating climate projection distributions for regional hydroclimatological risk assessments. *Climatic Change*, *89*(3), 371–394. https://doi.org/10.1007/s10584-007-9388-3

Brekke, L. D., Maurer, E. P., Anderson, J. D., Dettinger, M. D., Townsley, E. S., Harrison, A., & Pruitt, T. (2009). Assessing reservoir operations risk under climate change. *Water Resources Research*, *45*(4). https://doi.org/10.1029/2008WR006941

Briley, L., Kelly, R., Blackmer, E. D., Troncoso, A. V., Rood, R. B., Andresen, J., & Lemos, M. C. (2020). Increasing the Usability of Climate Models through the Use of Consumer-Report-Style Resources for Decision-Making. *Bulletin of the American Meteorological Society*, *101*(10), E1709–E1717. https://doi.org/10.1175/BAMS-D-19-0099.1

Brown, C., Ghile, Y., Laverty, M., & Li, K. (2012). Decision scaling: Linking bottom-up vulnerability analysis with climate projections in the water sector. Water Resources Research, 48(9). https://doi.org/10.1029/2011WR011212

Brunner, L., Pendergrass, A. G., Lehner, F., Merrifield, A. L., Lorenz, R., & Knutti, R. (2020). Reduced global warming from CMIP6 projections when weighting models by performance and independence. Earth System Dynamics, 11(4), 995–1012. https://doi.org/10.5194/esd-11-995-2020

Caswell, Thomas A., Michael Droettboom, Antony Lee, John Hunter, Eric Firing, David Stansby, Jody Klymak, Tim Hoffmann, Elliott Sales de Andrade, Nelle Varoquaux, Jens Hedegaard Nielsen, Benjamin Root, Phil Elson, Ryan May, Darren Dale, Jae-Joon Lee, Jouni K. Seppänen, Damon McDougall, Andrew Straw, … Jan Katins. (2020). matplotlib/matplotlib v3.1.3 (v3.1.3) [Software]. Zenodo. https://doi.org/10.5281/zenodo.3633844

Cinquini, L., Crichton, D., Mattmann, C., Harney, J., Shipman, G., Wang, F., et al. (2014). The Earth System Grid Federation: An open infrastructure for access to distributed geospatial data. *Future Generation Computer Systems*, *36*, 400–417. https://doi.org/10.1016/j.future.2013.07.002

Clark, M. P., Wilby, R. L., Gutmann, E. D., Vano, J. A., Gangopadhyay, S., Wood, A. W., et al. (2016). Characterizing Uncertainty of the Hydrologic Impacts of Climate Change. *Current Climate Change Reports*, *2*(2), 55–64. https://doi.org/10.1007/s40641-016-0034-x

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., et al. (2008). Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States [Dataset]. *International Journal of Climatology*, *28*(15), 2031–2064. https://doi.org/10.1002/joc.1688

Doblas-Reyes, F., Sörensson, A., Almazroui, M., Dosio, A., Gutowski, W., Haarsma, R., et al. (2021). IPCC AR6 WGI Chapter 10: Linking global to regional climate change (pp. 1363–1512). https://doi.org/10.1017/9781009157896.012

Easterling, D.R., K.E. Kunkel, J.R. Arnold, T. Knutson, A.N. LeGrande, L.R. Leung, et al., (2017). Precipitation change in the United States. *Climate Science Special Report: Fourth National Climate Assessment*, *1*, 207-230. U.S. Global Change Research Program, Washington, DC, USA, doi: 10.7930/J0H993CC.

ESGF CMIP6 Data Holdings. (June 8, 2023) [Collection]. Retrieved February 5, 2023, from https://pcmdi.llnl.gov/CMIP6/ArchiveStatistics/esgf_data_holdings/

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Bock, L., Lauer, A., Righi, M., Schlund, M., Andela, B., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – an extended set of large-scale diagnostics for quasi-operational and comprehensive evaluation of Earth system models in CMIP. *Geoscientific Model Development*, *13*(7), 3383–3438. https://doi.org/10.5194/gmd-13-3383-2020

Findlater, K., Webber, S., Kandlikar, M., & Donner, S. (2021). Climate services promise better decisions but mainly focus on better data. *Nature Climate Change*, *11*(9), 731–737. https://doi.org/10.1038/s41558-021-01125-3

Giorgi, F. Producing actionable climate change information for regions: the distillation paradigm and the 3R framework. Eur. Phys. J. Plus 135, 435 (2020). https://doi.org/10.1140/epjp/s13360-020-00453-1

Goldenson, N., Leung, L. R., Mearns, L. O., Pierce, D. W., Reed, K. A., Simpson, I. R., et al. (2023). Use-Inspired, Process-Oriented GCM Selection: Prioritizing Models for Regional Dynamical Downscaling. Bulletin of the American Meteorological Society, 104(9), E1619–E1629. https://doi.org/10.1175/BAMS-D-23-0100.1

Gutmann, E., Rasmussen, M., Liu, C., Ikeda, K., Gochis, J., Clark, M., et al. (2012). A comparison of statistical and dynamical downscaling of winter precipitation over complex terrain. *Journal of Climate*, 262–281. https://doi.org/10.1175/2011JCLI4109.1.

Gutmann, E., Pruitt, T., Clark, M. P., Brekke, L., Arnold, J. R., Raff, D. A., & Rasmussen, R. M. (2014). An intercomparison of statistical downscaling methods used for water resource assessments in the United States. *Water Resources Research*, *50*(9), 7167–7186. https://doi.org/10.1002/2014WR015559

Gutmann, E. D., Hamman, J. J., Clark, M. P., Eidhammer, T., Wood, A. W., & Arnold, J. R. (2022). En-GARD: A Statistical Downscaling Framework to Produce and Test Large Ensembles of Climate Projections. *Journal of Hydrometeorology*, *23*(10), 1545–1561. https://doi.org/10.1175/JHM-D-21-0142.1

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy [Software]. Nature, 585(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset [Dataset]. *Scientific Data*, *7*(1), 109. https://doi.org/10.1038/s41597-020-0453-3

Hausfather, Z., Drake, H. F., Abbott, T., & Schmidt, G. A. (2020). Evaluating the Performance of Past Climate Model Projections. Geophysical Research Letters, 47(1), e2019GL085378. https://doi.org/10.1029/2019GL085378

Henn, B., Clark, M. P., Kavetski, D., Newman, A. J., Hughes, M., McGurk, B., & Lundquist, J. D. (2018). Spatiotemporal patterns of precipitation inferred from streamflow observations across the Sierra Nevada mountain range. *Journal of Hydrology*, *556*, 993–1012. https://doi.org/10.1016/j.jhydrol.2016.08.009

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis [Dataset]. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Hoell, A., Hoerling, M., Eischeid, J., Wolter, K., Dole, R., Perlwitz, J., et al. (2016). Does El Niño intensity matter for California precipitation? *Geophysical Research Letters*, *43*(2), 819–825. https://doi.org/10.1002/2015GL067102

Hosseinizadeh, A., SeyedKaboli, H., Zareie, H., Akhondali, A., & Farjad, B. (2015). Impact of climate change on the severity, duration, and frequency of drought in a semi-arid agricultural basin. *Geoenvironmental Disasters*, *2*(1), 23. https://doi.org/10.1186/s40677-015-0031-8

Hoyer, S. & Hamman, J., (2017). xarray: N-D labeled Arrays and Datasets in Python [Software]. Journal of Open Research Software. 5(1), p.10. DOI: https://doi.org/10.5334/jors.148

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., et al. (2017). Extended Reconstructed Sea Surface Temperature, Version 5 (ERSSTv5): Upgrades, Validations, and Intercomparisons [Dataset]. *Journal of Climate*, *30*(20), 8179–8205. https://doi.org/10.1175/JCLI-D-16-0836.1

Kao, S.-C., Ashfaq, M., Rastogi, D., Gangrade, S., Uria Martinez, R., Fernandez, A., et al. (2022). *The Third Assessment of the Effects of Climate Change on Federal Hydropower* (No. ORNL/TM-2021/2278). Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States). https://doi.org/10.2172/1887712

Kim, Y.-H., Min, S.-K., Zhang, X., Sillmann, J., & Sandstad, M. (2020). Evaluation of the CMIP6 multi-model ensemble for climate extreme indices. *Weather and Climate Extremes*, *29*, 100269. https://doi.org/10.1016/j.wace.2020.100269

Klemes, V. (1986). Operational testing of hydrological simulation models. *Hydrological Sciences Journal*, *31*(1), 13–24. https://doi.org/10.1080/02626668609491024

Knutti, R., Masson, D., & Gettelman, A. (2013). Climate model genealogy: Generation CMIP5 and how we got there. *Geophysical Research Letters*, *40*(6), 1194–1199. https://doi.org/10.1002/grl.50256

Lenderink, G., de Vries, H., van Meijgaard, E., van der Wiel, K., & Selten, F. (2023). A perfect model study on the reliability of the added small-scale information in regional climate change projections. *Climate Dynamics*, *60*(9), 2563–2579. https://doi.org/10.1007/s00382-022-06451-6

Liang, Y., Gillett, N. P., & Monahan, A. H. (2020). Climate Model Projections of 21st Century Global Warming Constrained Using the Observed Warming Trend. Geophysical Research Letters, 47(12), e2019GL086757. https://doi.org/10.1029/2019GL086757

Livneh, B., Bohn, T. J., Pierce, D. W., Munoz-Arriola, F., Nijssen, B., Vose, R., et al. (2015). A spatially comprehensive, hydrometeorological data set for Mexico, the U.S., and Southern Canada 1950–2013 [Dataset]. *Scientific Data*, *2*(1), 150042. https://doi.org/10.1038/sdata.2015.42

Lopez-Cantu, T., Prein, A. F., & Samaras, C. (2020). Uncertainties in Future U.S. Extreme Precipitation From Downscaled Climate Projections. *Geophysical Research Letters*, *47*(9), e2019GL086797. https://doi.org/10.1029/2019GL086797

Lybarger, N.D. (2023). nlybarger/ESM_regional_evaluation: Initial release (v1.0.0). Zenodo. https://doi.org/10.5281/zenodo.8231348

Lyu, K., Zhang, X., & Church, J. A. (2021). Projected ocean warming constrained by the ocean observational record. *Nature Climate Change*, *11*(10), 834–839. https://doi.org/10.1038/s41558-021-01151-1

Maloney, E. D., Gettelman, A., Ming, Y., Neelin, J. D., Barrie, D., Mariotti, A., et al. (2019). Process-oriented evaluation of climate and weather forecasting models. Bull. Amer. Meteorol. Soc., 100(9), 1665–1686. https://doi.org/10.1175/BAMS-D-18-0042.1

Masson, D., & Knutti, R. (2011). Climate model genealogy. *Geophysical Research Letters*, *38*(8). https://doi.org/10.1029/2011GL046864

McSweeney, C.F., Jones, R.G., Lee, R.W. et al. Selecting CMIP5 GCMs for downscaling over multiple regions. Clim Dyn 44, 3237–3260 (2015). https://doi.org/10.1007/s00382-014-2418-8.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., et al. (2007). THE WCRP CMIP3 Multimodel Dataset: A New Era in Climate Change Research. *Bulletin*

of the American Meteorological Society, *88*(9), 1383–1394. https://doi.org/10.1175/BAMS-88-9-1383

Meehl, G. A., Senior, C. A., Eyring, V., Flato, G., Lamarque, J.-F., Stouffer, R. J., et al. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26), eaba1981. https://doi.org/10.1126/sciadv.aba1981

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J. J. F., Clark, M. P., et al. (2018). Mapping (dis)agreement in hydrologic projections. Hydrology and Earth System Sciences, 22(3), 1775–1791. https://doi.org/10.5194/hess-22-1775-2018

Mendoza, P. A., Clark, M. P., Mizukami, N., Newman, A. J., Barlage, M., Gutmann, E. D., et al. (2015). Effects of Hydrologic Model Choice and Calibration on the Portrayal of Climate Change Impacts. *Journal of Hydrometeorology*, *16*(2), 762–780. https://doi.org/10.1175/JHM-D-14-0104.1

Merrifield, A. L., Brunner, L., Lorenz, R., Humphrey, V., & Knutti, R. (2023). Climate model Selection by Independence, Performance, and Spread (ClimSIPS v1.0.1) for regional applications. Geoscientific Model Development, 16(16), 4715–4747. https://doi.org/10.5194/gmd-16-4715-2023.

Mizukami, N., Clark, M. P., Gutmann, E. D., Mendoza, P. A., Newman, A. J., Nijssen, B., et al. (2016). Implications of the Methodological Choices for Hydrologic Portrayals of Climate Change over the Contiguous United States: Statistically Downscaled Forcing Data and Hydrologic Models. *Journal of Hydrometeorology*, *17*(1), 73–98. https://doi.org/10.1175/JHM-D-14-0187.1

Mote, P., Brekke, L., Duffy, P. B., & Maurer, E. (2011). Guidelines for constructing climate scenarios. *Eos, Transactions American Geophysical Union*, *92*(31), 257–258. https://doi.org/10.1029/2011EO310001

Mote, P. W., & Salathé, E. P. (2010). Future climate in the Pacific Northwest. *Climatic Change*, *102*(1), 29–50. https://doi.org/10.1007/s10584-010-9848-z

Mote, P. W., Li, S., Lettenmaier, D. P., Xiao, M., & Engel, R. (2018). Dramatic declines in snowpack in the western US. *Npj Climate and Atmospheric Science*, *1*(1), 1–6. https://doi.org/10.1038/s41612-018-0012-1

Musselman, K. N., Addor, N., Vano, J. A., & Molotch, N. P. (2021). Winter melt trends portend widespread declines in snow water resources. *Nature Climate Change*, *11*(5), 418–424. https://doi.org/10.1038/s41558-021-01014-9

Najafi, M. R., Moradkhani, H., & Jung, I. W. (2011). Assessing the uncertainties of hydrologic model selection in climate change impact studies. *Hydrological Processes*, *25*(18), 2814–2826. https://doi.org/10.1002/hyp.8043

Newman, A. J., Clark, M. P., Craig, J., Nijssen, B., Wood, A., Gutmann, E., et al. (2015). Gridded Ensemble Precipitation and Temperature Estimates for the Contiguous United States [Dataset]. *Journal of Hydrometeorology*, *16*(6), 2481–2500. https://doi.org/10.1175/JHM-D-15-0026.1

Newman, A. J., Arnold, J. R., Wood, A. W., & Gutmann, E. D. (2022). A Workshop on Improving Our Methodologies of Selecting Earth System Models for Climate Change Impact Applications. *Bulletin of the American Meteorological Society*, *103*(4), E1213–E1219. https://doi.org/10.1175/BAMS-D-21-0316.1

Nijsse, F. J. M. M., Cox, P. M., & Williamson, M. S. (2020). Emergent constraints on transient climate response (TCR) and equilibrium climate sensitivity (ECS) from historical warming in CMIP5 and CMIP6 models. *Earth System Dynamics*, *11*(3), 737–750. https://doi.org/10.5194/esd-11-737-2020

Parding, K. M., Dobler, A., McSweeney, C. F., Landgren, O. A., Benestad, R., Erlandsen, H. B., et al. (2020). GCMeval – An interactive tool for evaluation and selection of climate model ensembles. *Climate Services*, *18*, 100167. https://doi.org/10.1016/j.cliser.2020.100167

Parker, W. S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science*, *87*(3), 457–477. https://doi.org/10.1086/708691

Patricola, C. M., O'Brien, J. P., Risser, M. D., Rhoades, A. M., O'Brien, T. A., Ullrich, P. A., et al. (2020). Maximizing ENSO as a source of western US hydroclimate predictability. *Climate Dynamics*, *54*(1), 351–372. https://doi.org/10.1007/s00382-019-05004-8

Phillips, A. S., Deser, C., & Fasullo, J. (2014). Evaluating Modes of Variability in Climate Models. *Eos, Transactions American Geophysical Union*, *95*(49), 453–455. https://doi.org/10.1002/2014EO490002

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

Prein, A. F., Rasmussen, R. M., Ikeda, K., Liu, C., Clark, M. P., & Holland, G. J. (2017). The future intensification of hourly precipitation extremes. *Nature Climate Change*, *7*(1), 48–52. https://doi.org/10.1038/nclimate3168

Rayner, N. A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century [Dataset]. *Journal of Geophysical Research, 108*(D14), 4407. https://doi.org/10.1029/2002JD002670

Ribes, A., Boé, J., Qasmi, S., Dubuisson, B., Douville, H., & Terray, L. (2022). An updated assessment of past and future warming over France based on a regional observational constraint. Earth System Dynamics, 13(4), 1397–1415. https://doi.org/10.5194/esd-13-1397-2022

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., et al. (2020). Earth System Model Evaluation Tool (ESMValTool) v2.0 – technical overview. *Geoscientific Model Development*, *13*(3), 1179–1199. https://doi.org/10.5194/gmd-13-1179-2020

Rupp, D. E., Abatzoglou, J. T., Hegewisch, K. C., & Mote, P. W. (2013). Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA. *Journal of Geophysical Research: Atmospheres*, *118*(19), 10,884-10,906. https://doi.org/10.1002/jgrd.50843

Sanderson, B. M., Xu, Y., Tebaldi, C., Wehner, M., O'Neill, B., Jahn, A., et al. (2017). Community climate simulations to assess avoided impacts in 1.5 and 2°C futures. *Earth System Dynamics*, *8*(3), 827–847. https://doi.org/10.5194/esd-8-827-2017

Scafetta, N. (2022). Advanced Testing of Low, Medium, and High ECS CMIP6 GCM Simulations Versus ERA5-T2m. *Geophysical Research Letters*, *49*(6), e2022GL097716. https://doi.org/10.1029/2022GL097716

Schlund, M., Hassler, B., Lauer, A., Andela, B., Jöckel, P., Kazeroni, R., et al. (2023). Evaluation of native Earth system model output with ESMValTool v2.6.0. *Geoscientific Model Development*, *16*(1), 315–333. https://doi.org/10.5194/gmd-16-315-2023

Schonher, T., & Nicholson, S. E. (1989). The Relationship between California Rainfall and ENSO Events. *Journal of Climate*, *2*(11), 1258–1269. https://doi.org/10.1175/1520-0442(1989)002<1258:TRBCRA>2.0.CO;2

Serreze, M. C., Clark, M. P., Armstrong, R. L., McGinnis, D. A., & Pulwarty, R. S. (1999). Characteristics of the western United States snowpack from snowpack telemetry (SNO^{TEL}) data. *Water Resources Research*, *35*(7), 2145–2160. https://doi.org/10.1029/1999WR900090

Simpson, I. R., K. A. McKinnon, F. V. Davenport, M. Tingley, F. Lehner, A. Al Fahad, and D. Chen, 2021: Emergent Constraints on the Large-Scale Atmospheric Circulation and Regional Hydroclimate: Do They Still Work in CMIP6 and How Much Can They Actually Constrain the Future?. J. Climate, 34, 6355–6377, https://doi.org/10.1175/JCLI-D-21-0055.1.

Snover, A. K., Mantua, N. J., Littell, J. S., Alexander, M. A., Mcclure, M. M., & Nye, J. (2013). Choosing and Using Climate-Change Scenarios for Ecological-Impact Assessments and Conservation Decisions: Choosing and Using Climate-Change Scenarios. *Conservation Biology*, *27*(6), 1147–1157. https://doi.org/10.1111/cobi.12163

Steinschneider, S., McCrary, R., Wi, S., Mulligan, K., Mearns, L. O., & Brown, C. (2015). Expanded Decision-Scaling Framework to Select Robust Long-Term Water-System Plans under Hydroclimatic Uncertainties. Journal of Water Resources Planning and Management, 141(11), 04015023. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000536

Suarez-Gutierrez, L., Milinski, S., & Maher, N. (2021). Exploiting large ensembles for a better yet simpler climate model evaluation. *Climate Dynamics*, *57*(9), 2557–2580. https://doi.org/10.1007/s00382-021-05821-w

Tokarska, K. B., Stolpe, M. B., Sippel, S., Fischer, E. M., Smith, C. J., Lehner, F., & Knutti, R. (2020). Past warming trend constrains future warming in CMIP6 models. *Science Advances*, *6*(12), 9549. https://doi.org/10.1126/sciadv.aaz9549

Tziperman, E., Cane, M. A., Zebiak, S. E., Xue, Y., & Blumenthal, B. (1998). Locking of El Niño's Peak Time to the End of the Calendar Year in the Delayed Oscillator Picture of ENSO. *Journal of Climate*, *11*(9), 2191–2199. https://doi.org/10.1175/1520-0442(1998)011<2191:LOENOS>2.0.CO;2

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python [Software]. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Weaver, C. P., Moss, R. H., Ebi, K. L., Gleick, P. H., Stern, P. C., Tebaldi, C., et al. (2017). Reframing climate change assessments around risk: recommendations for the US National Climate Assessment. *Environmental Research Letters*, *12*(8), 080201. https://doi.org/10.1088/1748-9326/aa7494

Wehner, M., Gleckler, P., & Lee, J. (2020). Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 1, model evaluation. Weather and Climate Extremes, 30, 100283. https://doi.org/10.1016/j.wace.2020.100283

Wehner, M. F. (2020). Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: Part 2, projections of future change. Weather and Climate Extremes, 30, 100284. https://doi.org/10.1016/j.wace.2020.100284

Wilks, D. S. (2019). *Statistical Methods in the Atmospheric Sciences*. Elsevier Science. Retrieved from https://books.google.com/books?id=apTzwQEACAAJ

Willmott, C. J., & Matsuura, K. (2001). Terrestrial Air Temperature and Precipitation: Monthly and Annual Time Series (1950 - 2017) [Dataset]. Retrieved January 19, 2023, from http://climate.geog.udel.edu/~climate/html_pages/README.ghcn_ts2.html

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., et al. (2020). Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019GL085782

Zhuang, J., R. Dussin, D. Huard, P. Bourgault, A. Banihirwe, S. Raynaud, B. Malevich, M. Schupfner, J. Hamman, S. Levang, A. Jüling, M. Almansi, F.G. Rondeau, S. Rasp, R. Bell, T.J. Smith, & X. Li. (2022). pangeo-data/xESMF: v0.6.3 (v0.6.3) [Software]. Zenodo. https://doi.org/10.5281/zenodo.6780573

Appendix

Table A1: CMIP6 models included in this study, with numbers of ensemble members for each run considered, resolution, and development center.

Model	Number of Historical Ensemble Members	Number of SSP2-4.5 Ensemble Members	Number of SSP3-7.0 Ensemble Members	Number of SSP5-8.5 Ensemble Members	Atmospheric Resolution (Lon x Lat)	Center
ACCESS-CM2	10	5	5	5	1.88 x 1.25	Commonweath Scientific and Industrial Research Organization, Australia
ACCESS-ESM1-5	40	40	40	40	1.88 x 1.25	Commonweath Scientific and Industrial Research Organization, Australia

AWI-CM-1-1-MR	5	1	5	1	0.94 x 0.93	Alfred Wegener Institute for Polar and Marine Research, Germany
AWI-ESM-1-1-LR	1	-	-	-	1.88 x 1.85	Alfred Wegener Institute for Polar and Marine Research, Germany
BCC-CSM2-MR	3	1	1	1	1.12 x 1.11	Beijing Climate Center, China Meteorological Administration
BCC-ESM1	3	-	3	-	2.81 x 2.77	Beijing Climate Center, China Meteorological Administration
CAMS-CSM1-0	3	2	2	2	1.12 x 1.11	Chinese Academy of Meteorological Sciences
CAS-ESM2-0	4	2	2	2	1.41 x 1.42	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences
CESM2	11	6	8	5	1.25 x 0.94	Community Earth System Model Contributors
CESM2-FV2	3	-	-	-	2.5 x 1.89	Community Earth System Model Contributors
CESM2-WACCM	3	5	1	5	1.25 x 0.94	Community Earth System Model Contributors
CESM2-WACCM-FV2	3	-	-	-	2.5 x 1.89	Community Earth System Model Contributors
CMCC-CM2-HR4	1	-	-	-	1.25 x 0.94	Centro euro-Mediterraneo sui Cambiamenti Climatici, Italy
CMCC-CM2-SR5	11	1	1	1	1.25 x 0.94	Centro euro-Mediterraneo sui Cambiamenti Climatici, Italy
CMCC-ESM2	1	1	1	1	1.25 x 0.94	Centro euro-Mediterraneo sui Cambiamenti Climatici, Italy
CNRM-CM6-1	29	10	6	6	1.41 x 1.39	National Centre of Meteorological Research, France
CNRM-CM6-1-HR	1	1	1	1	0.5 x 0.5	National Centre of Meteorological Research, France
CNRM-ESM2-1	11	10	5	5	1.41 x 1.39	National Centre of Meteorological Research, France
CanESM5	65	50	50	50	2.81 x 2.77	Canadian Centre for Climate Modeling and Analysis
CanESM5-1	72	-	-	-	2.81 x 2.77	Canadian Centre for Climate Modeling and Analysis
CanESM5-CanOE	3	3	3	3	2.81 x 2.77	Canadian Centre for Climate Modeling and Analysis
E3SM-1-0	5	-	-	5	1.0 x 1.0	Department of Energy, USA

E3SM-1-1-ECA	1	-	-	1	1.0 x 1.0	Department of Energy, USA
E3SM-2-0	5	-	-	-	1.0 x 1.0	Department of Energy, USA
EC-Earth3	22	69	57	58	0.7 x 0.7	EC-EARTH Consortium
EC-Earth3-AerChem	3	-	1	-	0.7 x 0.7	EC-EARTH Consortium
EC-Earth3-CC	10	9	-	1	0.7 x 0.7	EC-EARTH Consortium
EC-Earth3-Veg	8	7	6	8	0.7 x 0.7	EC-EARTH Consortium
EC-Earth3-Veg-LR	3	3	3	3	1.12 x 1.11	EC-EARTH Consortium
FGOALS-f3-L	3	1	1	1	1.25 x 1.0	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences
FGOALS-g3	6	4	5	4	2.0 x 5.18	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences
FIO-ESM-2-0	3	3	-	3	1.25 x 0.94	The First Institute of Oceanography, SOA, China
GFDL-CM4	1	1	-	1	1.25 x 1.0	NOAA Geophysical Fluid Dynamics Laboratory, USA
GFDL-ESM4	3	3	1	1	1.25 x 1.0	NOAA Geophysical Fluid Dynamics Laboratory, USA
GISS-E2-1-G	47	36	27	15	2.5 x 2.0	NASA Goddard Institute for Space Studies, USA
GISS-E2-1-G-CC	1	1	-	-	2.5 x 2.0	NASA Goddard Institute for Space Studies, USA
GISS-E2-1-H	25	10	6	10	2.5 x 2.0	NASA Goddard Institute for Space Studies, USA
GISS-E2-2-G	11	5	5	5	2.5 x 2.0	NASA Goddard Institute for Space Studies, USA
GISS-E2-2-H	5	-	-	-	2.5 x 2.0	NASA Goddard Institute for Space Studies, USA
GISS-E3-G	1	-	-	-	1.25 x 1.0	NASA Goddard Institute for Space Studies, USA
HadGEM3-GC31-LL	5	5	-	4	1.88 x 1.25	Met Office Hadley Center, UK
HadGEM3-GC31-MM	4	-	-	4	0.83 x 0.56	Met Office Hadley Center, UK
IITM-ESM	1	1	1	1	1.88 x 1.89	Indian Institute of Tropical Meteorology
INM-CM4-8	1	1	1	1	2.0 x 1.5	Institute for Numerical Mathematics, Russia
INM-CM5-0	10	1	5	1	2.0 x 1.5	Institute for Numerical Mathematics, Russia

IPSL-CM5A2-INCA	1	-	1	-	3.75 x 1.89	Institut Pierre Simon Laplace, France
IPSL-CM6A-LR	33	11	11	7	2.5 x 1.27	Institut Pierre Simon Laplace, France
IPSL-CM6A-LR-INCA	1	-	-	-	2.5 x 1.27	Institut Pierre Simon Laplace, France
KACE-1-0-G	3	3	3	3	1.88 x 1.25	National Institute of Meteorological Sciences, Korea Meteorological Administration
MCM-UA-1-0	1	1	1	1	3.75 x 2.22	University of Arizona, USA
MIROC-ES2L	31	30	10	10	2.81 x 2.77	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies
MIROC6	50	50	50	50	1.41 x 1.39	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology
MPI-ESM-1-2-HAM	3	-	3	-	1.88 x 1.85	Max Planck Institute for Meteorology, Germany
MPI-ESM1-2-HR	10	2	10	2	0.94 x 0.93	Max Planck Institute for Meteorology, Germany
MPI-ESM1-2-LR	31	30	30	30	1.88 x 1.85	Max Planck Institute for Meteorology, Germany
MRI-ESM2-0	12	10	5	6	1.12 x 1.11	Meteorological Research Institute, Japan
NESM3	5	2	-	2	1.88 x 1.85	Nanjing University of Information Science and Technology, China
NorESM2-LM	3	13	1	1	2.5 x 1.89	Norweigian Climate Center, Norway
NorESM2-MM	3	2	1	1	1.25 x 0.94	Norweigian Climate Center, Norway
SAM0-UNICON	1	-	-	-	1.25 x 0.94	Seoul National University, Korea
TaiESM1	2	1	1	1	1.25 x 0.94	Research Center for Environmental Changes, Academia Sinica, Taiwan
UKESM1-0-LL	19	17	16	5	1.88 x 1.25	National Environmental Research Council, Met Office Hadley Center, UK
UKESM1-1-LL	1	-	1	-	1.88 x 1.25	National Environmental Research Council, Met Office Hadley Center, UK