Memorandum on Evaluating Global Climate Models for Studying Regional Climate Change in California

Interim Deliverable for EPC-20-006, Prepared by:
Will Krantz[1], David Pierce[2], Naomi Goldenson[1], Daniel Cayan[2]
November 29, 2021

- This research is funded by the California Energy Commission (CEC) through its Electric Program Investment Charge (EPIC) Program, which invests in scientific and technological research to accelerate the transformation of the electricity sector to meet the state's energy and climate goals.
- The applied research grant, EPC-20-006, will integrate the latest downscaling approaches applied to the recently produced global climate models (GCMs) with an engagement process to develop a robust, usable, set of climate projections applicable for California.
- This memo is being shared to support transparent and timely consideration of interim deliverables that are relevant for energy stakeholders and all those interested in California's next generation of climate projections. The memo includes data that were not produced through CEC funding.

This memorandum is submitted to the CEC by UC San Diego's Scripps Institution of Oceanography. The memo meets deliverable requirements under Task 3 of the California Energy Commission's applied research grant EPC-20-006: Development of Climate Projections for California and Identification of Priority Projections.

[1] Department of Atmospheric and Oceanic Sciences, University of California, Los Angeles, California
[2] Scripps Institute of Oceanography, La Jolla, California

Synopsis

With a new generation of global climate models (GCMs) in the Coupled Model Intercomparison Project-phase 6 (CMIP6), localized projections for California will also be updated under *Development of Climate Projections for California and Identification of Priority Projections*, EPC-20-006, (hereafter *Climate Projections for California*). This document describes the processes used to rank GCMs that best capture the relevant characteristics of the climate in California.

In addition to advances in GCMs themselves, advances in methods to localize their projections, or regional downscaling, have also been made. In *Climate Projections for California*, global climate simulations over the California region will be downscaled with a new hybrid approach, combining physics-based simulations with an updated version of the Localized Constructed Analogs (LOCA; Pierce et al., 2014) statistical downscaling method. High-resolution physics-based simulations will first be used to downscale a selection of GCMs over a California domain, driven by the GCM behavior in key variables at the boundaries (dynamical downscaling). These simulations will then be used to train a new version of LOCA that takes advantage of the information in the dynamical downscaling about the future climate to downscale a broader set of GCMs than would be computationally feasible with dynamical downscaling alone.

These new developments, both in the GCMs (described in the Introduction, below) and the downscaling methods just described, necessitate a new approach to selection of GCMs that best capture the unique characteristics of California's climate. Previously, the fidelity of a GCM at representing the large-scale climate beyond California was used as a first step to rule out GCMs that were unlikely to provide reliable climate projections under future warming scenarios, followed by the elimination of further GCMs in a series of steps down to local performance metrics (DWR, 2015; Pierce et al., 2018). Instead of a hierarchical culling of GCMs, we now rank them in one unified procedure in which performance across metric categories can be easily compared, and greater weight given to those of greater relevance to a given application. A key change is in how the hemispheric scale climate model evaluation is performed. In the previous effort, existing metrics for the entire Northern Hemisphere were used as an initial GCM quality filter, after which California-region specific climate model evaluations were applied. In the new approach, broad scale (covering not only California but much of the Pacific Ocean) atmospheric circulation metrics that are key to weather and climate variability over the state are used alongside regional metrics that evaluate the fidelity of simulated temperature and precipitation over the state. This gives a more targeted evaluation of processes that affect California's climate.

A variety of metrics were designed to assess the realism of the processes that drive weather and climate. These processes include global-scale modes of variability (like El Nino, which was considered before) and California-specific processes such as the conditions that drive Santa Ana winds, a major factor in regional wildfire risk. These "process-based" metrics are combined with a set of metrics that assess GCM errors in the resulting local climate conditions ("local" metrics). It is important to minimize both types of errors to produce the most reliable set of localized projections through our downscaling.

The context for the GCM evaluations is provided below, followed by details about all of the metrics employed in each category ("process-based" in Section 2, and "local" in Section 3). Then a methodology section describes how all of these metrics are combined and weighted to produce a unified assessment: the overlap between the top performers from each ranking produces a set of priority models for downscaling. These priority models are the GCMs with the strongest overall performance. Some other GCMs with intermediate performance are better used with caution. There are two potential categories we identify for which caution should be applied: Some GCMs might be "right for the wrong reasons", performing well locally despite poor process representation, while others might harbor inadequacies in some of their parameterizations, causing high local errors despite otherwise good representation of atmospheric processes. The latter can still be used for dynamical downscaling, for which the regional model will produce its own local solution but would not perform as well for statistical downscaling.

Another benefit of our approach, unifying these categories of metrics, is that we can identify sets of GCMs that perform indistinguishably well in a statistical sense without necessarily specifying an a priori number of GCMs to prioritize. Ongoing work, informed by stakeholder engagement, will inform the final selection of priority GCMs. The considerations that might enter into the selection process of a representative set of simulations is also discussed in the concluding section of this document.

1. Introduction

This memo presents a process for evaluating the performance of GCMs across California and the western US. The goal is using this evaluation to identify a prioritized set of models from the latest generation of GCMs (CMIP6) for subsequent dynamical and statistical downscaling to support climate change analysis and applications relevant for California.

The CMIP6 generation of models includes GCMs that improve the representation of continental-scale atmospheric circulation patterns that produce realistic weather and climate in California in both an average sense and in terms of variability compared to CMIP5 (Cannon, 2020; Simpson et al., 2020). GCMs represent global and regional historical climate to varying degrees of fidelity with historical observations, and we assume global models that better represent key aspects of climate during recent decades will perform better in representing global and regional climate in future decades. Despite these improvements in CMIP6, some GCMs exhibit a higher climate sensitivity (the amount of warming for a given increase in greenhouse gas concentration) than seen in the last generation of GCMs (CMIP5), resulting in modeled warming over the historical period that is unrealistically large compared to observations (e.g., Tokarska et al. 2020) along with projected warming in future decades that is extreme. The latest report from the IPCC (AR6) concludes that the highest climate sensitivities in CMIP6 are very unlikely to be a correct depiction of the Earth's climate sensitivity, but that these models still contain useful information about how the climate system will evolve as it warms (IPCC, 2021). We adopt a similar approach by evaluating models based only on their

performance representing the regional climate of California and not eliminating GCMs because of an overly high climate sensitivity. By doing so we take advantage of the improved representation of climate processes across the entire CMIP6 ensemble and can explore possible outcomes across a wide range of potential future warming. However, this choice will necessitate communication of the subsets of GCMs best suited to answer particular types of questions of relevance for California, which will be done in collaboration with our colleagues in the CEC-funded project: Analytics and Data Platform to Facilitate Electricity Sector Adaptation (EPC-20-007), i.e., GCMs with an unrealistically high climate sensitivity should not be used to estimate when warming thresholds or climate impacts directly linked to temperature will be reached in coming decades.

This work is aimed at model evaluation for both statistical and dynamical downscaling, two distinct methodologies that benefit from different GCM evaluation strategies. In dynamical downscaling, meteorological variables along the boundary of the domain are the primary source of information passed to the regional climate model, although interior nudging to the GCM fields is often used as well. For the dynamical downscaling component of the work localizing GCM projections for California, it is essential that the GCMs chosen to provide boundary conditions for the simulations are skilled at producing the large-scale patterns of circulation that control the location and frequency of mesoscale events that drive regional weather and climate in California.

Evaluations of GCMs aimed at statistical downscaling, on the other hand, can place more emphasis on the performance of each model on the representation of climate in the region being downscaled, such as the spatial patterns of seasonal and annual mean and variance of temperature and precipitation. However, it cannot be assumed that a GCM with low regional biases is appropriately simulating the physically-based drivers of regional climate. To avoid models that are “right for the wrong reasons”, the evaluation presented below combines an evaluation of local climate metrics with process-based metrics to recommend a set of GCMs best suited for downscaling to study future climate.

Evaluating a large set of GCMs with different strategies affords us the opportunity to compare ranked model results from the two approaches, lending additional insight into the strengths and weaknesses of the evaluation methods and further illuminating how GCMs should be selected for regional downscaling.

It is worth noting that this separation of evaluation methodologies into the analysis of boundary information (appropriate to dynamical downscaling) and regional information (appropriate to some forms of statistical downscaling) is a simplification, albeit a useful one. For instance, regional patterns of interannual temperature and precipitation variability are unlikely to be correct in a region influenced by El Niño-Southern Oscillation (ENSO) unless the GCM has a reasonable simulation of ENSO in the first place, i.e., a local evaluation of variability influenced by tele-connected climate responses provides an indirect evaluation of the (dynamical) generating mechanisms.

2. Stakeholder Input into GCM Evaluation Process

Preliminary results from Freitas, Jagannathan, Jones (in prep), based on interviews with 20 key stakeholders across California's energy sector, indicate that organizations pursue a variety of approaches for GCM selection based on the decision or metric of interest to them. Often their perceptions and needs for GCM performance evaluation are embedded within conversations on GCM selection. Overall, we found that stakeholders relied on modeling community experts to lead and guide the first level of model performance evaluation for regional processes. They suggested that climate modeling experts were the ones who would know the best approach for identification of the top models that are "good" for California and should hence be downscaled.

With regard to which metrics needs to be used for GCM evaluation (and then further downscaling), stakeholders from IOUs and consulting groups expressed a need for experts (scientists, regulatory bodies, other technical experts) to identify these models based on scientific metrics that are most relevant for the regional context and processes. To illustrate: an IOU representative said, "We really have to rely on climate scientists to say this is a good set of models". Several stakeholders also stated that they looked at the DWR's Perspectives and Guidance for Climate Change Analysis for understanding model credibility for regional processes - further suggesting that they are looking to experts and regulators for this question. Although stakeholders acknowledge that they relied on expert opinion for the evaluation process, they did indicate that they think it was important for them to be part of the conversation on how to determine the appropriate spread between models and were looking for guidance on why one model might be better than another for a given decision or in a specific scenario.

3. Process-based climate metrics

To evaluate the performance of GCMs at simulating the physical processes that strongly influence the hydrological cycle and extreme weather in California, one set of metrics is constructed to capture large-scale patterns of circulation, pressure, and moisture transport in the historical period. Because the regional climate model used for downscaling is primarily driven by data from the GCM at the lateral boundaries, particular attention is paid to metrics that capture the behavior of the jet stream and moisture transport over the eastern Pacific Ocean. A brief description of the metrics used is given below.

The model data for all of these metrics was taken from the Earth System Grid archive for CMIP5 and CMIP6 models. Although the goal is to identify CMIP6 models for downscaling, the previous generation of GCMs is included to compare performance across model generations. The performance of the GCMs is evaluated against European Center for Medium Range Weather Forecasting (ECMWF) Reanalysis, version 5 (ERA5) reanalysis, with both regridded to a common 1˚ or 2˚ grid depending on the metric. Except where otherwise specified in the description of metrics below, the performance of models is evaluated and compared to ERA5 over the historical reference period of 1979-2014.

Normalized Mean Square Error

To compute an overall skill score for each GCM that weights dissimilar metrics and variables in a fair way, it is important that each metric be unitless, computed consistently, and span a similar numerical range. This could be accomplished by normalizing or re-scaling each metric, but such normalization would exaggerate small differences in metrics where all models perform similarly well. Instead, we standardize the computation of each metric using the Normalized Mean Square Error (NMSE). The NMSE was originally proposed by Williamson (1995), and has recently been used by Simpson et al. (2020) to evaluate the performance of Community Earth System Model 2 (CESM2) against the rest of the CMIP6 models. For a model field X_m, the NMSE is given by:

$$NMSE(X_m) = \frac{(X_m - X_o)^2}{\overline{(X'_o)^2}}$$

Where X_o is the observed data field (from ERA5 in this case), the overbar represents an area weighted spatial average, and the prime represents the deviation from the spatial average. The NMSE is chosen for the process-based metrics because it is designed to capture biases or differences in the phase or amplitude of the overall spatial pattern in a particular field. Computing all of our process-based metrics as NMSEs of various model fields ensures that the metrics are on a consistent scale that can be directly compared, but still preserves information about which metrics capture larger meaningful differences between models.

Overview of process-based metrics

Northern hemisphere circulation

One of the highest priorities for the process-based evaluation is capturing the large-scale circulation patterns across the entire northern hemisphere, with 16 of the 30 process focused metrics in our evaluation measuring signatures of seasonal circulation. Calculations of NMSE from the 300-hPa eddy stream function ($\psi*$), 850-hPa zonal winds, and 10-day high-pass-filtered eddy meridional wind variance ($v'v'$) are all adapted from Simpson et al. (2020) and included in our evaluation.

Blocking

A northern hemisphere blocking index captures the frequency of days where a significant circulation anomaly blocks the westerly mid-latitude winds. These blocking events can produce heat waves in the summer and extreme cold events in the winter. The blocking metric used is originally described in Masato et al. (2013) and here we use the specific implementation calculated for the CMIP6 ensemble in Simpson et al. (2020). Blocking is calculated between 25°N and 75°N for the historical period of 1979-2014, with separate metrics calculated for the winter and summer months. The metric originally calculated for a subset of the CMIP6 ensemble by Simpson et al. (2020) has been calculated for additional CMIP6 models as well as CMIP5 models for this evaluation.

A supplementary blocking metric developed by Scaife et al. (2010) and also computed in Simpson et al. (2020) normalizes for biases in the mean flow to remove the contribution of these biases to the blocking calculation, to represent the portion of the error that is due to errors in synoptic variability alone. This metric is also included as “mean fixed” blocking.

Wind shear

Early dynamical downscaling experiments as part of this project identified a poor representation of off-shore wind shear as a cause of precipitation biases in the California region. In this work a wind-shear metric was added to identify those GCMs that realistically capture the off-shore vertical structure of the winds at the boundaries. The monthly difference is taken between the zonal winds at two levels in the atmosphere (250 hPa minus 850 hPa) over the northern Pacific Ocean (between 150-230˚E and 20-66˚N). The NMSE of this field averaged over seasons is computed relative to the ERA5 reanalysis, after regridding all data via bilinear interpolation to a common 2˚ grid.

California extreme precipitation

Three sets of metrics focused on capturing processes related to extreme precipitation in California are calculated as described below:

Metrics are used to evaluate the GCM’s representation of average vertically integrated column water vapor (IWV), sea level pressure (SLP), and 250-hPa zonal wind (u250) on days of extreme precipitation in California. Following the method used in Norris et al. (2021), extreme wet days are identified as those with average California precipitation above the 95th percentile of all wet season days (November-April) for years 1979-2014. The NMSE is computed comparing each GCM to ERA5 over this same reference period within a domain bounded by 20˚N-60˚N and 150˚W-100˚W. For total column water vapor (CWV), only ocean points are considered due to the difficulty of vertically integrating over land with data on limited pressure levels.

The California precipitation mode (CPM) is a distinct mode of atmospheric pressure over the North Pacific that strongly influences extreme precipitation and dry days in California. The mode is identified as the 3rd empirical orthogonal functions (EOF) of 500 hPa geopotential height (Z500) anomalies (20-75N, 90-170W), with positive anomalies in this mode strongly associated with extreme (>99th percentile) precipitation days and negative anomalies strongly associated with dry (<10th percentile) days (Chen et al., 2021). An NMSE is computed comparing the pattern of the 3rd Z500 EOF of each GCM to ERA5 over the reference period of 1982-2014, which captures how well the GCMs re-produce the location and strength of this mode.

Southern California faces unique patterns of change to precipitation variability and extreme precipitation that may not be well captured by metrics that focus on the entire state (Swain et al., 2018)**.** A set of metrics examining the meteorological circulations that cause heavy precipitation in Los Angeles County help ensure these patterns are included in the evaluation. Data from 553 heavy precipitation days in winter (October to April) from 1950 to 2019 is clustered using a self-

organizing map (SOM; Lin et al., in prep). The days were selected when any 24-hour rain gauge measurement in Los Angeles Country was greater than its 2-year return value. The SOM was applied to the combination of standardized 250-hPa streamfunction (ST250) anomaly and integrated water vapor transport (IVT) anomaly over the domain of 195W-110W and 20N-50N, based on the National Center for Environmental Prediction (NCEP) 20th century reanalysis (Compo et al., 2011). Four modes from the SOM with the highest cross-model variance that are associated with heavy precipitation days are used to compute metrics. An NMSE is calculated for each mode by comparing the 2d map of the average value of IVT or ST250 across all the days associated with the mode between the GCM and ERA5.

Santa Ana winds

The Santa Ana winds in southern California are among the strongest drivers of fire risk in the region. The strong gradient in sea level pressure (SLP) across southwestern California that is associated with the Santa Ana winds is a crucial feature for a GCM studying climate change in California to model accurately. The method described in Abatzoglou et al. (2013) is used to identify Santa Ana events between 1979 and 2004[3] in both the GCMs and ERA5 reanalysis, and an NMSE is calculated from the mean SLP across the western US and eastern Pacific (130W-100W, 30N-50N) on days with strong Santa Ana events. While the empirical relationship used for this metric (from Abatzoglou 2013) was developed relative to Los Angeles. We suspect that models that capture this well would behave similarly well in terms of the large-scale boundary conditions controlling the development of comparable winds in nearby Santa Barbara.

El Niño Southern Oscillation

Large scale modes of natural climate variability can influence temperature and precipitation in California via teleconnections, for example the ENSO. We identify the sea surface temperature (SST) pattern associated with ENSO variability by calculating the first empirical orthogonal function (EOF, discussed in another context below), of the sea surface temperatures over the Pacific Ocean basin, north of 60˚S. The pattern is compared via the NMSE with that derived via the same procedure from the NOAA Extended Reconstructed SST V3b dataset over the period 1854-2005.

4. Local climate metrics

The metrics described in section 2 evaluate boundary conditions for a regional dynamical downscaling effort. Another approach, suited to statistical downscaling methods that use the

[3] As with the local metrics described in Section 3, here the end year was chosen to allow the maintenance of the same period for CMIP5 and CMIP6 models to better compare the two generations. CMIP5 models only considered a historical period through 2004. Elsewhere we've extended through the end of the CMIP6 historical period (2014) by adding on an extra 10 years to the CMIP5 models that come from a projection instead of the historical simulation. Both approaches are acceptable as long as the same period is used consistently within a given metric's definition for both generations of GCMs.

spatial field in the region of interest to produce a downscaled result, is to evaluate local climate metrics in the region being downscaled. For example, California's seasonal and annual patterns of temperature and precipitation, both in the mean and in variability. The process for calculating these local climate metrics is described below.

Z-score / skill score

The local climate metrics are constructed and evaluated following the methods in Pierce et al. 2021. Briefly, the GCM data are first bilinearly interpolated to a common 1˚ latitude/longitude grid. Then the seasonal (DJF, MAM, JJA, SON) means of the variables of interest are computed at every point in the domain 32 to 42 North, 125 to 114 West (a box covering California and Nevada). The period of evaluation is 1950-2005, taken to be consistent with the CMIP5 historical period. This discards the last 9 years of the CMIP6 historical period (which ends in 2014), but provides a consistent evaluation period for the two CMIP generations. During the previous (4th) California Climate Assessment we examined whether an analogous shift in climatological period relative to the 3rd California Assessment affected the quality ranking of the models and found that it did not, given the uncertainty in model rankings due to natural climate variability and a limited observational time period. We use Livneh et al. 2015 for the observed data set, aggregated to the same 1˚ spatial grid as used for the models and taken over the same time span. In this work we examine the seasonal means of daily average temperature and precipitation.

After computing the seasonal means of the models and observations (obs), we form the z score at every point. In essence, this compares the model-obs difference using as a yardstick the variability of the observations. The key idea is that model-obs differences can be evaluated by seeing how large they are compared to natural year-to-year climate variability. This is evaluated at every point in the domain, forming a spatial map of z scores:

$$z(x,y) = \frac{<model(x,y,t)> - <obs(x,y,t)>}{stddev(obs(x,y,t))}$$

Where the angle brackets <> indicate averaging the sequence of seasonal mean values over time, and stddev() indicates the standard deviation over time. The resultant spatial pattern of z scores is then spatially averaged to form a skill score, which is traditionally formulated so that 1 is perfect skill and progressively more negative numbers indicate less skill:

$$ss = 1 - RMS(z)$$

where RMS indicates the root mean square of the z score field.

Simple bias correction

Coarse resolution GCM solutions tend to produce biased results in regions where the topography cannot be spatially resolved, for example, over the Sierra Nevada. Because of this, bias correction is a standard part of statistically downscaling a GCM. In recognition of this we

apply a simple bias correction to the GCM results before they are evaluated using the metrics. The purpose is to avoid penalizing a model for a bias that would be removed before use anyway. As described in Pierce et al. 2021, the simple bias correction subtracts off the time- and space-averaged error of the GCM with respect to the observations. Since the simple bias correction uses only a single number for all times and locations, GCMs can still sensibly be evaluated for their ability to reproduce spatial and temporal variability in comparison to the observations, unlike when full bias correction is used. The full LOCA downscaling process uses a sophisticated bias correction method (Pierce et al. 2015), but GCMs cannot be evaluated after the full bias correction because they are then statistically indistinguishable from the observations in measures the bias correction is designed to correct, rendering any such comparison moot. In practice, we find that the simple bias correction makes the biggest difference to winter (DJF) precipitation, which tends to be poorly represented in GCMs due to their inability to resolve California's varied topography.

Overview of local climate metrics

The regional metrics consist of seasonal means of temperature (T) and precipitation (P); the standard deviation of T and P averaged into 1-, 5-, and 10-year blocks; and amplitude and phase of the annual harmonic of T and P, used to evaluate the representation of the seasonal cycle; and the standard deviation of monthly values of T and P taken in January and July, used to evaluate a shorter timescale than the seasonal means and variability. Altogether that yields 40 regional metrics covering T and P.

5. Methodology

After computing both sets of metrics described in sections 2 and 3 above for all available GCMs (see the discussion about data availability in section 5), a ranking is separately produced for the process-based metrics and the local climate metrics. The methodology for computing a total score and ranking is nearly the same for both sets of metrics, and is described below.

Reducing redundancy with Empirical Orthogonal Functions

In both the processed-based and local climate metrics used in this evaluation, priority is given to ensuring coverage of a wide range of phenomena rather than ensuring each metric is independent. A significant amount of overlap and redundancy is expected across each set of metrics. To prevent any particular climate variable or process from being over-represented in the evaluation, the redundant information captured by the metrics must first be eliminated. This is done by computing a set of EOFs from the metrics, and only retaining a subset of EOFs that capture most of the variation between models. The result is a reduced set of linear combinations of metrics that efficiently captures nearly all of the variance across GCMs.

The method of using EOFs to eliminate redundancy and re-weight model metrics has been used previously to evaluate regional performance of CMIP5 GCMs using a similar set of climate metrics (Pierce et al., 2009; Rupp et al., 2013; Pierce et al. 2021). These studies demonstrated

that the first few leading EOFs captured a significant fraction of the variance across models, indicating a significant redundancy in the coverage of metrics included. Additionally, the EOF analysis provides insight into patterns of covariance among the metrics and processes that can be useful when interpreting model performance.

The process of calculating EOFs and using a truncated subset to re-weight the metrics is carried out separately for both the process based and local climate metrics, with the two rankings combined afterwards, as described below.

Total error score

For both the process-based and local climate metrics, an overall skill score is computed for each model using the metrics after EOF decomposition. This overall score is computed as the euclidean distance between the point representing perfect model skill and the point represented by the model's score on each metric. For the process-based metrics, perfect model skill would be an NMSE value of zero for all metrics and any larger value represents decreased performance. For the local metrics the perfect skill score is 1 and any value less than that represents decreased performance. The fact that one of these is defined as a measurement of error and the other a measurement of skill leads to these different numerical conventions. One of the advantages of defining a total error score as a euclidean distance is that it rectifies this difference.

In both cases a smaller euclidean distance, referred to as the Dss or total error score, represents better overall model performance. Because the Dss and total error score are computed in different skill spaces for the local metrics and process-based metrics respectively, the scores can not be compared directly, but can be used to rank the model performance in each of the two domains of performance. Rather than compute a single score that combines the local and process-based scores, the ranked lists are compared to identify models that perform well across both sets of metrics. A group of top performing models is selected from each list, and the overlap between the top performers from each ranking produces a set of priority models for downscaling.

Uncertainty and grouping of models

Each climate model expresses some degree of natural variability, resulting in a range of possible scores in this evaluation. For GCMs that have data available from several realizations, a separate skill score was computed for each ensemble member with the standard deviation of these scores representing an uncertainty around the model's mean performance. This uncertainty analysis is used to determine when differences in model performance are statistically significant and to create groups of models with statistically equivalent performance.

Due to limited data availability, the full uncertainty analysis described here can not be completed for all models and an approximation must be used. For models that have fewer than three ensemble members, we cannot calculate an uncertainty due to natural variability. In this case

we approximate it as the average uncertainty from all other models. This approximation works well for the local climate metrics where only a few models are lacking multiple ensemble members. For the process-based metrics, a further simplification is used due to the limited availability of data. In this case one model (CESM2) with the largest available ensemble (10 members) is used to calculate an estimated uncertainty that is used for all other models.

6. Results

Process-based metric scores

The NMSE values for the process based metrics are shown in Figure 1. Groups of metrics focused on similar processes are color coded and grouped together, models are colored according to their CMIP generation (CMIP5 = blue, CMIP6 = red), and lower NMSE scores represent better performance on a given metric.

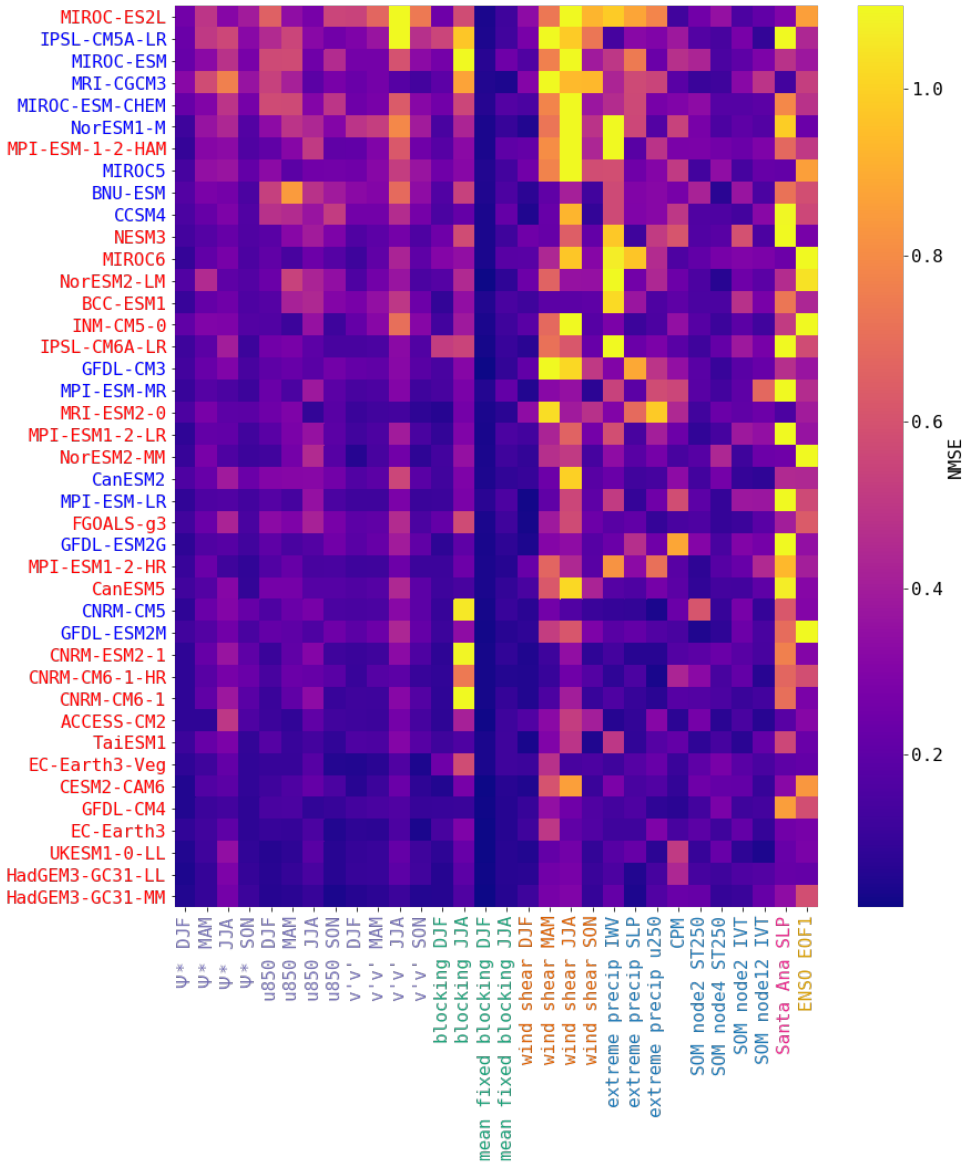

Figure 1. *Normalized mean square error (NMSE) values across all process-based climate metrics (x-axis) for each GCM (y-axis). The metrics are color coded to group measurements of similar processes. The GCMs are color coded by model generation, with CMIP5 models in blue and CMIP6 models in red. Lower NMSE scores indicate better model performance.*

The NMSE values in Figure 1 include the 40 models that had sufficient data to be included in the evaluation: Out of the 169 CMIP5 and CMIP6 models listed on Earth System Grid Federation (ESGF), 111 had enough data available to calculate at least some metrics, and only 27 had the necessary data available to calculate all 30 metrics used in this evaluation. An additional 13 models that were missing fewer than 5 metrics were included in this evaluation by replacing the missing value with the mean NMSE value for that metric across other models. There is a tradeoff between choosing a set of metrics with thorough coverage of physical processes and the number of models that will have enough data available to be included. For

example, the eddy meridional wind variance and blocking metrics each require daily data that is only available for 43 models. Including these metrics significantly restricts the set of models that can be evaluated. Nevertheless, we kept these metrics in the evaluation because both provide unique and useful information for differentiating model performance, as shown in the discussion of the EOF process below. In this case none of the GCMs excluded had sufficient data available to be candidates for dynamical downscaling.

Since our goal is to identify meaningful differences between models, we draw special attention to the metrics shown in Figure 1 that have a wide range of NMSE values across the GCMs. Notably, the first ENSO EOF, summer blocking, sea level pressure during Santa Ana events, and seasonal wind shear measurements in the spring and summer all show a wide range of NMSE values. These metrics are particularly helpful in differentiating model performance. On other metrics like winter blocking and the winter eddy stream function, the model scores show little variation. These metrics still represent processes that are important for California climate, but they do not capture significant differences for model performance.

Ranking computed with EOF process

A truncated set of EOFs is generated from the NMSE values shown in Figure 1 to reduce the redundancy of information collected by the metrics. The choice of how many EOFs to retain is based on Kaiser's rule for eigenvalue significance, along with the rule of thumb from North et al. (1982) to ensure degenerate multiplets of EOFs are not split. In our case, the first six EOFs are determined significant and together capture 91% of the total variance across models. The NMSE scores for all models are then re-constructed from just the retained EOFs. Summing the squares of the components of these six retained EOFs produces a summary of the total weight that each metric contributes to the re-constructed error scores, shown in Figure 2. Without truncation, each metric has an equal weight of 1. The distribution of weights after truncation shows that a small set of metrics is responsible for most of the variation between models, and most of the metrics in our evaluation do not meaningfully separate the performance of the models. This weighting does not reflect the importance of each metric in simulating California climate, but rather how much variation there is in each metric across GCMs. The metrics with low weight indicate that all GCMs in our evaluation perform equally well for those measures. In this sense, it is not surprising that many of the metrics representing basic circulation patterns have low weights, while metrics aimed at patterns of variability and extremes that are more difficult for GCMs to model have higher weights.

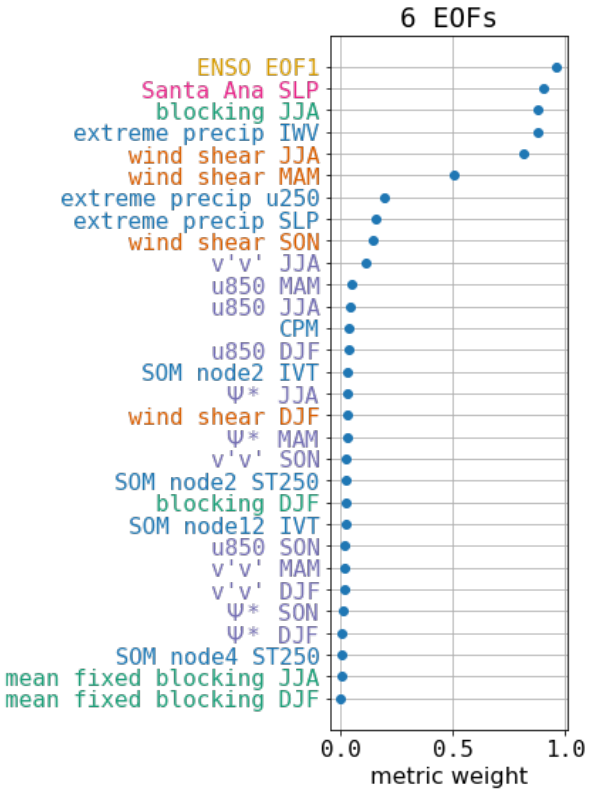

Figure 2. *The relative weight of each metric in the retained set of EOFs. The metrics are color coded as in Figure 1.*

The final ranking of models from the scores computed after EOF truncation is shown in Figure 3. Each model is shown with error bars representing the uncertainty due to natural variability across ensemble members as described in an earlier section. Groups of models that are statistically indistinguishable from a particular target model are also indicated with brackets. Overall CMIP6 models (in red text) represent a higher portion of the best performing models, and different versions of models from the same modeling center tend to appear close to each other in the ranking.

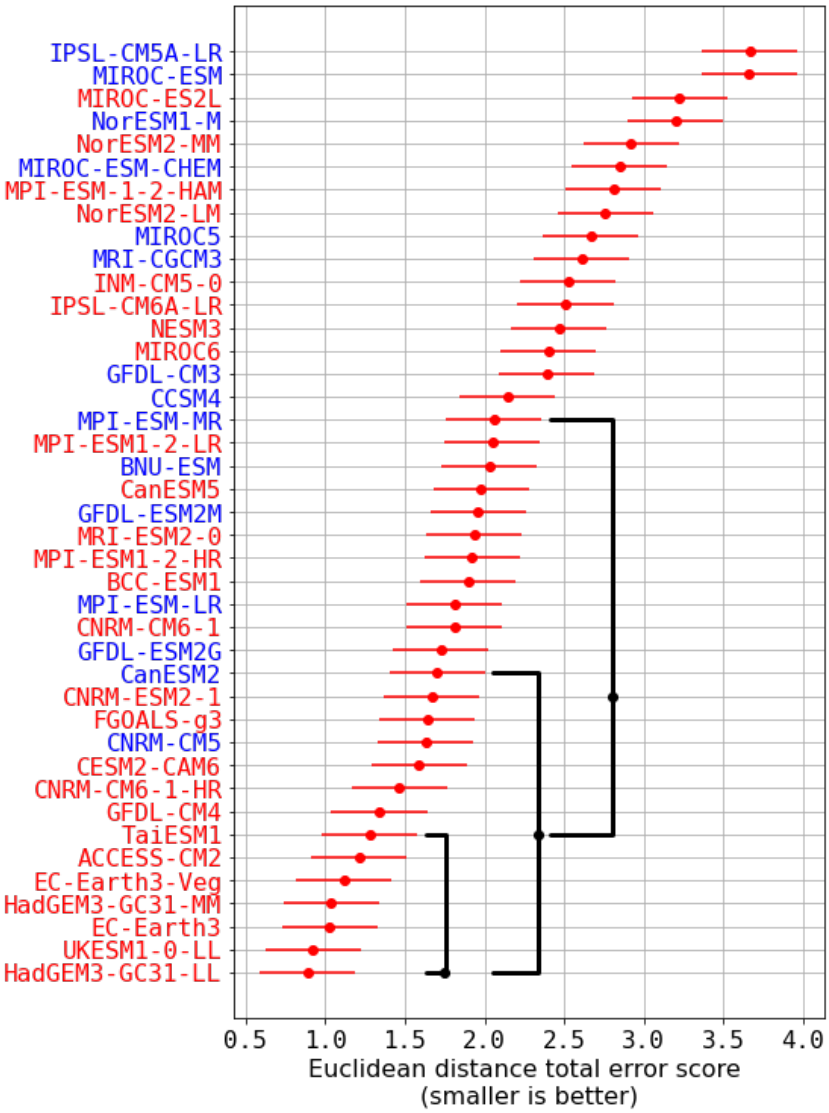

Figure 3. *Ranking of the GCMs by their total error score from the process-based metrics. GCMs are colored by model generation with CMIP5 models in blue and CMIP6 models in red. Error bars represent estimated uncertainty due to natural variability. Groups of models that are statistically indistinguishable from the model indicated by each black dot are denoted by black brackets. Lower values of the total error score indicate better model performance. Local climate metric scores.*

Results from the local metric scores are shown in Figure 4. More positive values indicate better model performance in comparison to the observations. All results are shown after the simple bias correction described above. Three generations of models are included; CMIP6 (red), CMIP5 (blue), and CMIP3 (black). The older CMIP3 models do noticeably worse than the more recent CMIP5 and CMIP6 models. This is perhaps unsurprising given the advancement in physical parameterizations and spatial resolution as model generations have progressed, but nonetheless is important to confirm for a particular region influenced by a particular set of dynamical processes.

The simulation of wintertime (DJF) precipitation variability is a weak point in many of the model simulations. This is likely due to California's rather dramatic and varied topography just inland from the coast, falling in a climate regime subject to prevailing westerly storm tracks during the winter. The GCMs do not resolve either the coastal mountain ranges or the Sierra Nevada, leading to large misrepresentation of statewide DJF precipitation processes in the coarse-resolution GCMs. By contrast summer (JJA) precipitation patterns are well simulated, which is not surprising since California's Mediterranean climate has very little precipitation in summer. Likewise, the phase of the season cycle of temperature is represented well, another easy to simulate metric since it is largely determined by the geographical location of California and the Earth's orbit around the sun, both of which are specified to the GCMs. Some models from the same institution (irrespective of CMIP generations) have similar scores, for example the EC-EARTH models are grouped at the bottom of the figure (tending to do well), while the GISS models are grouped at the top of the figure (tending to do relatively poorly).

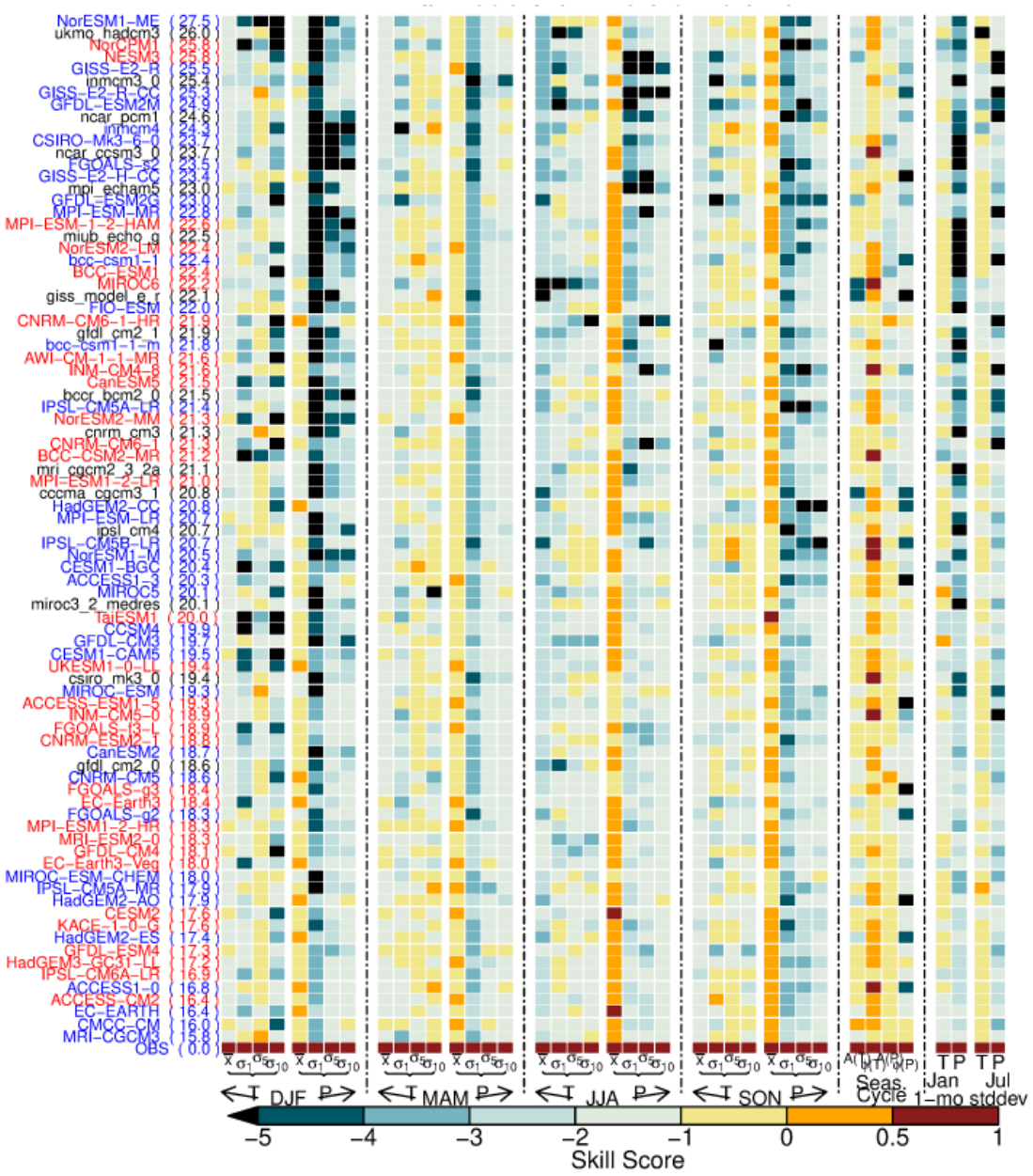

Figure 4. *Skill score values for the local climate metrics (x-axis) for each GCM (y-axis). GCMs from CMIP6 (red), CMIP5 (blue), and CMIP3 (black) are included. Each model's total error score (Dss) is included next to the model name. Higher values of the skill score indicate better model performance.*

Local climate metric ranking

The final model skill scores after the EOFs have been used to reduce the metric redundancy are shown in Figure 5. Seven EOF modes have been retained, which together account for 87.7% of the variance. The uncertainty in rankings (horizontal red bars) are calculated from the spread in Dss across ensemble members for each model that has three or more ensemble members, as shown along the right hand edge of the panel. For those models with less than three ensemble members, the average uncertainty in models that have >= 3 ensemble members is used. Given

the uncertainties, a wide range of models are indistinguishable from the “best” model (vertical bars with diamonds).

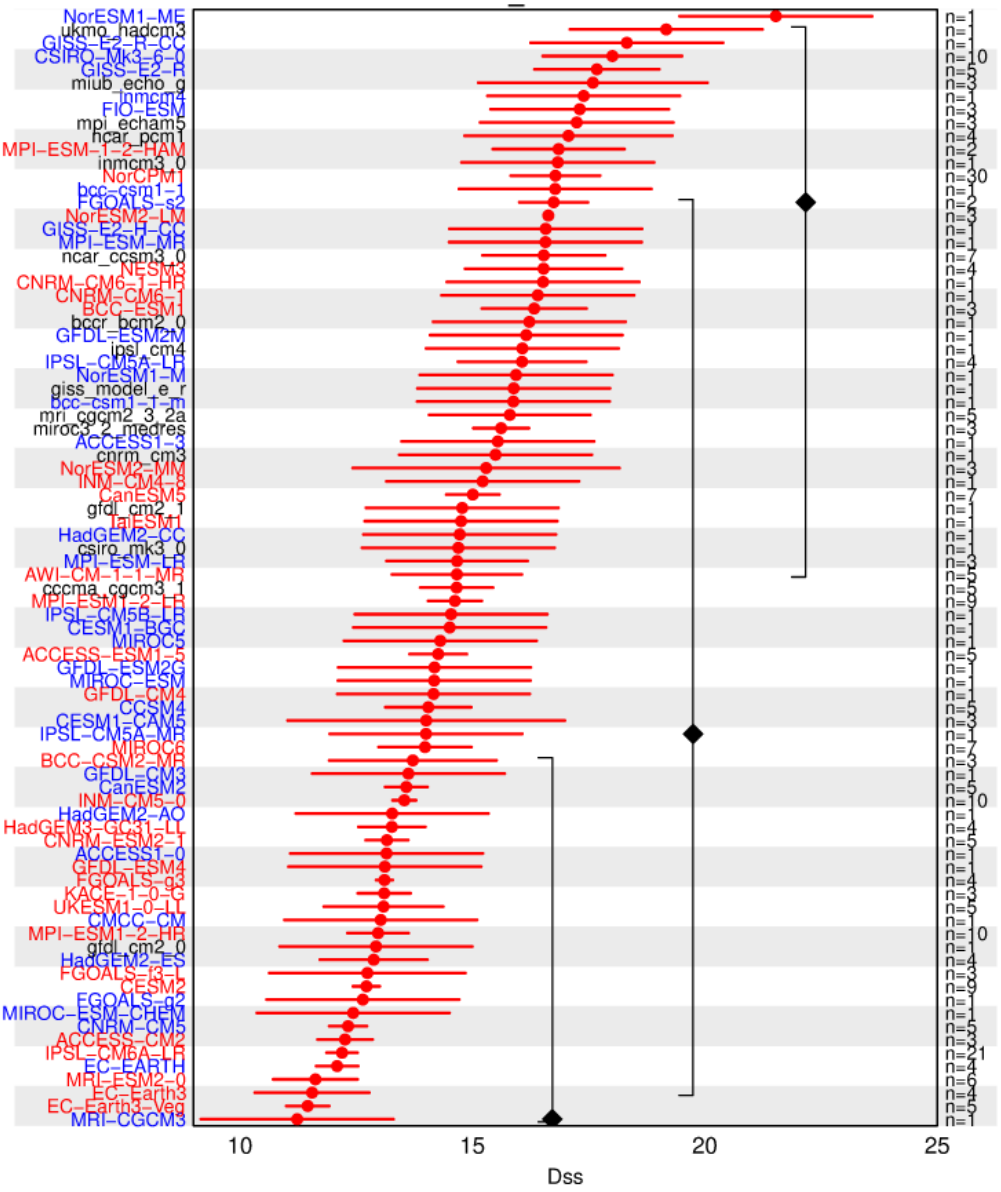

Figure 5. *Model ranking by total error score for local climate metrics (Dss). Error bars represent the uncertainty across multiple ensemble members. The number of ensemble members included from each model is shown on the right. Groups of models that are statistically indistinguishable from the indicated model are denoted by black brackets. GCMs are colored by their model generation, as above. Lower values of Dss indicate better model performance.*

Combined evaluation from process-based and local metrics

Each of the two evaluation methodologies applied here produces an independently ranked list of GCMs, along with groupings of the best performing models. These two rankings are combined

to produce an overall set of models recommended for downscaling to support climate change analysis and applications for California. A scatterplot showing the relationship between the two sets of model ranks is shown in Figure 6. For simplicity in comparing the ranked lists, only models that appear in both rankings are included. The models that perform well in both evaluations (the lower left of Figure 6) demonstrate an accurate representation of California climate (they perform well in the local climate metrics), and we have confidence that they do so because of a good representation of large-scale climate processes (they perform well in the process-based metrics, and are “right for the right reasons”). These models are highlighted as the top choices for dynamical downscaling. This method of identifying the set of best performing models allows us to examine the relationship between the two ranking methods, and also avoids allowing good performance in one ranking from compensating for poor performance in another. The top statistical group of both rankings includes EC-Earth3, EC-Earth3-Veg, ACCESS-CM2, UKESM1-0-LL, and HadGEM3-GC31-LL.

While there is some consistency about top performing models across the two sets of evaluation methodologies, a number of models that perform well in one ranking do poorly in the other. Within the top grouping of models from the process-based ranking, five of the seven models also fall in the top group of the local climate based ranking (one of these seven models, HadGEM3-GC31-MM, is not included in the local climate ranking due to unavailable data). Among the seventeen models from the top group of the local climate ranking, there is a much wider range of performance on the process-based ranking. Notably there are two CMIP5 models (MRI-CGM3 and MIROC-ESM-CHEM) and one CMIP6 model (IPSL-CM6A-LR) that are very highly ranked by the local metrics but among the worst in the process-based metrics. These models that rank much lower in the local climate metrics than in the process-based metrics seem to be in line with the hypothesis that regional climate performance in GCMs can appear accurate despite poor representation of the driving processes due to offsetting biases or statistical coincidence.

Model similarities play a varied role in determining when models will fall close to each other in the rankings. Several modeling centers have multiple versions of their GCMs included in the ranking, and while some tend to appear near each other in the rankings (EC-Earth3 variants perform very well, all NorESM2 variants perform relatively poorly), others fall across a wide range in both rankings (MPI, CNRM). In most cases models from the same modeling center are more closely ranked in the process-based evaluation than the local climate evaluation, suggesting that the process-based metrics are capturing similarities in the models’ physical simulations and the additional variation across the local metrics is due to either changes in parameterizations or natural variability between runs.

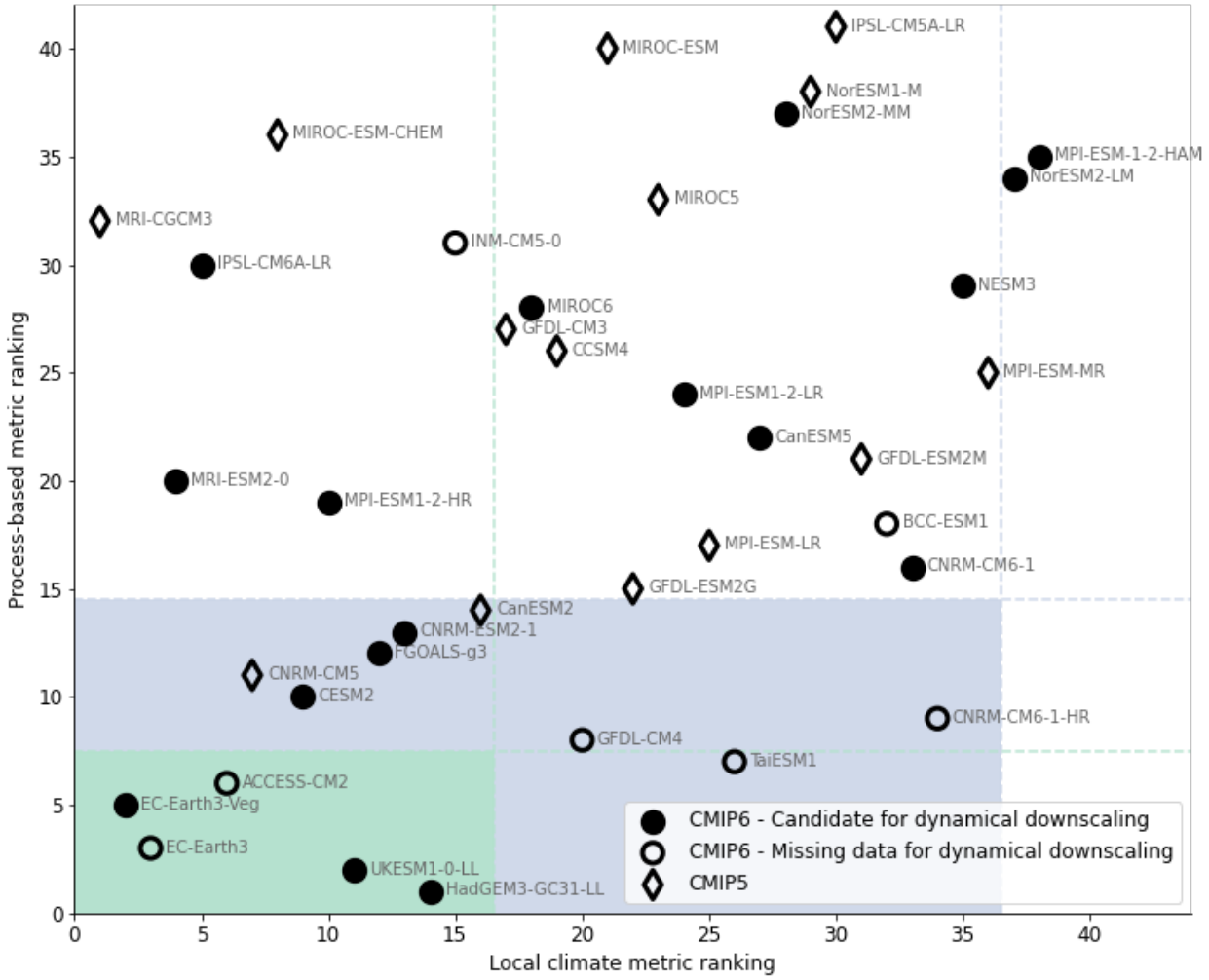

Figure 6. *Comparison of the GCM rankings by local climate metric performance and process-based metric performance. Models that appear in one ranking but not the other are not included. CMIP6 models with sufficient data for dynamical downscaling are indicated by solid circles. The top two statistical groupings of models are shown in green and blue.*

7. Summary and conclusion

This memo outlines a process for identifying GCMs for downscaling over California by ranking their performance across both local climate variables and regional to hemispheric scale processes. By examining model performance along these two axes, we identify models that are able to produce accurately simulated conditions in California with confidence they are doing so with a correct representation of the circulation patterns that drive these conditions. Using the uncertainty associated with internal variability across model ensembles, models are grouped into statistically indistinguishable sets of the best performing models. Although all the models are individually ranked, we emphasize that within a statistical set the models are considered to perform equally well for that category of metrics and there is no strong preference between

them for an application like dynamical or statistical downscaling. This allows us to take a simple approach when combining the rankings to identify the best performing models overall by simply taking the overlap of models that are in the top statistical sets of each ranking. In general, our evaluation finds that CMIP6 models outperform CMIP5 models, but some individual GCMs from the previous generation perform quite well, particularly on the local climate metrics. Additionally, there is a high degree of overlap between the best performing models in both ranking methodologies, but outside of the top performers there is low correlation between overall process-based and local climate performance.

It should be noted that this is a targeted evaluation of model performance specific to the California region. The metrics used here were developed and chosen to provide thorough coverage of the most important aspects of California climate for stakeholders. There is a particular focus on the hydrological cycle and extreme weather given their importance for adaptation planning. The result is a recommendation for models that are best suited for downscaling over California, but not a universal evaluation of GCM performance for all applications. The approach taken here is meant to illustrate that the most appropriate GCM to use depends on the application being considered, and a stakeholder with specific interests may choose a different collection of metrics and come to different conclusions about the set of models suited for future projects.

This memo has outlined a process for identifying the GCMs that are best suited for downscaling over California. Some guidance and discussion of the next steps in the downscaling process is offered here. The next step is selecting a set of individual model realizations from these GCMs for the computationally intensive task of dynamical downscaling. The output from these dynamically downscaled realizations will be used to train the next generation of LOCA, which can then be used to statistically downscale a larger set of model realizations at much lower computational cost. This necessitates a careful strategy for selecting the model realizations that are dynamically downscaled, as they will define the range of climate data that LOCA is trained on. Although all the GCMs identified in this memo are skilled at modeling California climate, there is still a considerable diversity of outcomes that they produce in simulations of future climate change. Because we make no assumptions about which of these future changes is more likely, the goal for selecting model realizations to dynamically downscale is to representatively span the range of possible future changes. Among the important dimensions of future change to consider are the mean and variability of temperature and precipitation, along with changes to extreme rainfall events and drought duration and frequency. Finally, given the goal of selecting a diverse set of future predictions, model genealogy should be considered to avoid over-representing GCMs from the same modeling center or that share extensive physical similarities. Incorporating these recommendations will help ensure that the set of downscaled models provides the most complete picture about possible changes California will face and the challenges of adaptation that stakeholders must prepare for.

References

Abatzoglou, J. T., Barbero, R., & Nauslar, N. J. (2013). Diagnosing Santa Ana Winds in Southern California with Synoptic-Scale Analysis. *Weather and Forecasting*, *28*(3), 704–710. https://doi.org/10.1175/WAF-D-13-00002.1

Cannon, A. J. (2020). Reductions in daily continental-scale atmospheric circulation biases between generations of global climate models: CMIP5 to CMIP6. *Environmental Research Letters*, *15*(6), 064006. https://doi.org/10.1088/1748-9326/ab7e4f

Chen, D., Norris, J., Goldenson, N., Thackeray, C., & Hall, A. (2021). A Distinct Atmospheric Mode for California Precipitation. *Journal of Geophysical Research: Atmospheres*, *126*(12), e2020JD034403. https://doi.org/10.1029/2020JD034403

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., … Worley, S. J. (2011). The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654), 1–28. https://doi.org/10.1002/qj.776

DWR. (2015). *Perspectives and Guidance for Climate Change Analysis* (p. 142). https://water.ca.gov/-/media/DWR-Website/Web-Pages/Programs/All-Programs/Climate-Change-Program/Climate-Program-Activities/Files/Reports/Perspectives-Guidance-Climate-Change-Analysis.pdf

IPCC, 2021: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press. In Press.

Norris, J., Hall, A., Chen, D., Thackeray, C. W., & Madakumbura, G. D. (2021). Assessing the Representation of Synoptic Variability Associated With California Extreme Precipitation in CMIP6 Models. *Journal of Geophysical Research: Atmospheres*, *126*(6), e2020JD033938. https://doi.org/10.1029/2020JD033938

North, G. R., Bell, T. L., Cahalan, R. F., & Moeng, F. J. (1982). Sampling Errors in the Estimation of Empirical Orthogonal Functions. *Monthly Weather Review*, *110*(7), 699–706. https://doi.org/10.1175/1520-0493(1982)110<0699:SEITEO>2.0.CO;2

Pierce, D. W., D. R. Cayan, J. Goodrich, T. Das, and A. Menevar, 2021: Evaluating Global Climate Models for hydrological studies of the Upper Colorado River Basin. J. American Water Association, in press .

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

Pierce, D. W., D. R. Cayan, E. P. Maurer, J. T. Abatzoglou, and K. C. Hegewisch, 2015: Improved bias correction techniques for hydrological simulations of climate change. J. Hydrometeorology, v. 16, p. 2421-2442. DOI: http://dx.doi.org/10.1175/JHM-D-14-0236.1

Pierce, D. W., D. R. Cayan, and B. L. Thrasher, 2014: Statistical downscaling using localized constructed analogs (LOCA). J. Hydrometeorology, v. 15, p. 2558, doi:10.1175/JFM-D-14-0082.1

Pierce, D. W., Kalansky, J. F., & Cayan, D. R. (2018). Climate, drought, and sea level rise scenarios for California's fourth climate change assessment. *California Energy Commission and California Natural Resources Agency*.

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
Simpson, I. R., Bacmeister, J., Neale, R. B., Hannay, C., Gettelman, A., Garcia, R. R., Lauritzen, P. H., Marsh, D. R., Mills, M. J., Medeiros, B., & Richter, J. H. (2020). An Evaluation of the Large-Scale Atmospheric Circulation and Its Variability in CESM2 and Other CMIP Models. *Journal of Geophysical Research: Atmospheres*, *125*(13), e2020JD032835. https://doi.org/10.1029/2020JD032835
Swain, D. L., Langenbrunner, B., Neelin, J. D., & Hall, A. (2018). Increasing precipitation volatility in twenty-first-century California. *Nature Climate Change*, *8*(5), 427–433. https://doi.org/10.1038/s41558-018-0140-y
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: eaaz9549.
Williamson, D. L. (1995). Skill scores from the AMIP simulations. *World Meteorological Organization-Publications*, 253–258.