Selecting Models and Assessing Credibility

Cal-Adapt
Published: 2024-01-01
Last modified: 2026-10-11

For attribution, please cite this work as:

Cal-Adapt. (2026). *Selecting Models and Assessing Credibility*. Cal-Adapt. https://analytics.cal-adapt.org/scientific-guidance/model-selection.html

The Analytics Engine hosts multiple models and ensemble members across its two downscaled datasets. This page covers how to choose among them, how the pre-selected General Use Projections can be used, and how model validation and credibility can inform those choices. Background on ensemble members and model ensembles is in About climate projections. Guidance on choosing between the WRF and LOCA2-Hybrid datasets is on the WRF vs. LOCA2-Hybrid page.

How should users choose among models and ensemble members?

The Analytics Engine Data Catalog shows the total number of models and ensemble members from both WRF and LOCA2-Hybrid datasets that are available to users. Broadly, both datasets include several models, and LOCA2-Hybrid has multiple ensemble members for a select few models. Both come from GCMs that have been pre-selected based on a rigorous skill evaluation of how well they capture key characteristics of California's climate (Krantz et al., 2021).

Once a dataset has been selected, it is generally recommended that users examine as many models and ensemble members as possible to develop the most complete picture of the range of plausible future outcomes. The Analytics Engine has tools and functionalities that allow users to more easily work with a larger number of models than was previously feasible for the energy sector.

However, there may be certain instances where selecting a smaller number of model runs is necessary (e.g. time series data applications). How a user undertakes this type of sub-selection will depend on the type of climate data application (including the variable, location, and scale of analysis) and the user's ability to process large amounts of data. In cases where it is necessary to select a smaller set of model runs, **users must first perform a preliminary exploratory analysis on the full suite of available data within the Analytics Engine to determine the appropriate set of models and ensemble members for their specific context**. During the exploratory analysis, it might be useful for users to determine if, for their specific application, they care more about **(1) sampling across a range of potential outcomes, or (2) gaining a statistical understanding of (changing) extreme weather events**. While choosing between these two different options is not ideal and is always limiting, there are cases where such difficult trade-offs might need to be made based on practical considerations.

Sampling across a range of outcomes

For cases where it is determined that a range of outcomes are needed, the users' preliminary analysis could entail a **ranking exercise** to characterize all the available model runs for the range of outcomes for the region, timescale, and primary variable(s) of interest (e.g. characterizing model

runs from hottest to coolest, or wettest to driest, or based on the change signal in the models). Users can then sub-select a set of model runs to represent the median, mean, and/or either tail of the distribution, depending on which aspect of the distribution is most relevant to their application. If the goal is to select a plausible range of outcomes, then it may not be necessary to distinguish among models and ensembles. If calculating a mean, median, or some other statistical property of the model distribution, though, users should be mindful of how models with different numbers of ensembles are weighted.

Understanding changing extremes

For cases where it is determined that a statistically robust understanding of changing extreme weather events is required, **it is recommended that users avoid sub-sampling to the extent possible**. Extreme event analyses require large ensembles of climate data because these events are often poorly sampled within a single climate model run (a 1-in-100 year event may only occur once in a given ensemble run). Properly understanding the statistics of such extremes requires large ensembles so that there are enough instances of rare events to understand their properties and frequencies. See the Extreme Value Analysis page for more.

A user's preliminary analysis should then entail an assessment of both the internal variability and model uncertainty of the entire dataset. This is because both these types of variability are important for extreme event analyses. *Internal variability* can be examined in the Analytics Engine as the differences among ensemble members from each individual model (i.e. as intra-model variability), and *model uncertainty* can be examined as the differences between various models in the dataset (i.e. as model-to-model variability or inter-model variability). The availability of multiple ensemble members for some models hosted in the Analytics Engine makes it possible to conduct extreme event analyses for each individual model, and to distinguish how different models represent changing characteristics of extreme events. After comparing across models to see how similarly or differently they represent the internal variability, users can then choose models and ensemble members accordingly.

If a user still finds it necessary to sub-sample the models, they should perform the aforementioned analyses of the two types of variability for all the Analytics Engine models that have multiple ensemble members. Users should then cross-check the inter-model spread of their chosen sub-sample with the inter-model spread of the broader dataset to check if the full range of outcomes is captured in their sub-sample. For more on the two sources of variability, see the Uncertainty page.

The default setting in many Analytics Engine notebooks is to provide outputs that span all models and ensembles. Users can use the tools in the Analytics Engine to examine both the internal variability and model uncertainty. Model sub-selections can also be done through the Cal-Adapt Data Download Tool or other data access methods.

What are General Use Projections (GUPs) and can they be used?

While it is always recommended for users to examine as many models and ensemble members as possible, Analytics Engine also hosts a set of "General Use Projections". These five LOCA2-Hybrid model runs have been pre-selected to reasonably capture a range of future outcomes for a limited number of commonly used climate variables, scales, and SSPs. - `ACCESS-CM2 r1i1p1f1` - `MPI-ESM1-2-HR r3i1p1f1` - `EC-Earth r1i1p1f1` - `FGOALS-g3 r1i1p1f1` - `MIROC6 r1i1p1f1`

If users are unable to perform context-specific preliminary analyses (see How should users choose among models and ensemble members?) or are less comfortable working with large amounts of data, **the GUPs can provide a minimum set of model runs to start with before more sophisticated analyses can be done**.

Users should review this short memo for details on how the GUPs were identified and what ranges of outcomes they span (and do not span). Broadly, the GUPs were evaluated for SSP 370 and 2045-2074 (mid-century) changes relative to the 1950-2014 historical period. This was done for selected variables including temperature (change in statewide average annual daily Tmax, degree F), precipitation (change in statewide average annual precipitation, % change), and wind (change in statewide average annual wind speed, % change). In the assessments, the five selected GUPs were able to reasonably capture the range of projected changes for the above-mentioned precipitation and temperature metrics. Future changes in wind speed projected by these GCMs varied quite a lot by location in the state, therefore users who require changes in wind at specific locations should evaluate a broader set of models for their specific location of interest rather than relying solely on the reduced set of general use projections.

Limitations of GUPs

Since GUPs only represent a minimum set of starter projections, they are limited in their scope, and users should exercise caution while drawing conclusions from this sub-selection. While the GUPs attempt to cover a range of variability for a limited set of pre-determined metrics, it cannot be guaranteed that these model runs will necessarily cover the appropriate range for every user-relevant outcome, i.e. for different time periods, SSPs, regions, and/or variables/metrics. Particularly for applications which involve examining extreme events, the GUPs may not be appropriate, as they may not be able to reasonably capture the full range of extremes.

Therefore to the extent possible, it is recommended that users rely upon a larger number of models and ensembles, or conduct their own preliminary analyses to determine which models are needed to capture the appropriate range of plausible outcomes for their specific contexts (see How should users choose among models and ensemble members?). It is worthwhile to reiterate that existing and forthcoming tools of the Analytics Engine will make working with larger amounts of model runs much easier for users, and therefore users will be able to make use of the full suite of data rather than relying on limited sets of projections.

How are climate models validated?

To get a sense of how reliable a model is, climate data users may want to know how well a given climate model has performed. This exercise is commonly referred to as “model validation.” Climate models are typically validated by modeling the historical climate and then comparing this modeled historical climate to the observed historical climate. Observations of the historical climate come from a variety of sources: weather stations, satellites, and aircraft. If the modeled historical climate accurately recreates the observed historical climate record, the climate model is often considered valid.

Climate models can also be put through a series of standardized diagnostic tests to see how skillfully they perform as compared to other validated models, especially with regard to how they generate phenomena observed in paleoclimate records. This practice is referred to as model intercomparison. A common intercomparison test inputs an instantaneous increase in carbon dioxide to four times

pre-industrial levels and diagnoses the results of each model. Another common test is a gradual, steady increase in carbon dioxide concentrations until four times pre-industrial levels are reached.

To learn more about the diagnostic tests that were used to determine which climate models to downscale for California's Fifth Climate Change Assessment, please refer to Krantz et al. 2021.

How can users assess the credibility of the Analytics Engine data?

All of the downscaled data available in the Analytics Engine comes from GCMs that have been pre-selected based on a rigorous skill evaluation of how well the models capture several relevant characteristics of California's climate. These GCMs were shown to perform well for both process-based as well as local climate metrics. The models were able to simulate the key physical processes and patterns that strongly influence the hydrological cycle and extreme weather in California, and they were also able to capture local climatic patterns such as annual and seasonal patterns of temperature and precipitation. This sort of state-level GCM assessment is unique, and lends broad credibility to the data that is being hosted in the Analytics Engine. More details on how the GCMs were evaluated and selected are available in Krantz et al., 2021.

Although all the models in the Analytics Engine have been shown to perform well at capturing California's climate broadly, this does not necessarily mean that all models perform equally well for any specific sub-region within California or for any specific metric (particularly ones that were not evaluated within the GCM selection process). In such cases, users may want to additionally understand whether and to what extent the data being used is credible for their specific region, metric, or application. Here, an additional level of credibility analysis could be of interest to a user. However, the approach for such additional user-driven GCM evaluations is still an evolving area of research and experimentation that scientists and practitioners are collectively trying to figure out.

For additional credibility analyses, users may first want to qualitatively assess whether and to what extent the GCMs represent the key underlying physical processes that are relevant to the phenomenon or region of interest. This can be done using existing studies or literature on the GCMs. Then they may want to quantitatively examine the historical performance of the models or dataset, for the metric that is of interest. Upon such analyses, users might consider three key questions:

1. Are the class of models available suitable/appropriate for the specific use case, climate phenomenon, or variable of interest?
2. Is there a need to differentiate between the various models based on their historical performance?
3. Should some models be avoided (or culled) for specific regions, variables, or applications?

However, these are nuanced questions that require advanced analyses that cannot be predicted or easily identified. **Overall, it is worthwhile to reiterate that user-centric credibility evaluations are still an evolving matter in both science and practice.** Sub-selections or culling of models based on very specific credibility assessments need to be done carefully and in consultation with climate science experts on the best course of action.

Credibility evaluation when using wind data

One instance where an additional credibility evaluation might be useful to users, is for cases where specific wind variables are needed. Wind is often a tricky variable to capture well in climate models due to its extremely local characteristics, which can be affected by local buildings, trees, hills, etc. A user may want to validate that the data they are using reasonably captures the distribution of wind extremes in a given location of interest before proceeding to analyze changes in that metric in the future. (It can be noted that such a validation analysis might need access to specific observed data sets to compare to, which can be difficult to obtain for variables like wind).

When such fine-scale features are being examined, it is possible that only a subset or even none of the models represent the phenomenon of interest very well. In this case users may want to rule out certain models or might determine that the available data is not appropriate for their application. Ruling out models should be balanced with considerations about sampling (see How should users choose among models and ensemble members?), so that users ensure that they do not overscreen models and miss out on plausible and realistic outcomes.