A framework for evaluating climate model performance metrics

Noël C. Baker and Patrick C. Taylor
NASA Langley Research Center, Hampton, VA. Contact: noel.c.baker@nasa.gov

Introduction and Motivation

- The widely-used Coupled Model Intercomparison Project 5 (CMIP5) archive serves as the basis of climate assessment for the United Nations Intergovernmental Panel on Climate Change (IPCC) report AR5.
- Ensemble averaging of multiple CMIP5 models is used to add value to projections by constructing a consensus projection.
- However, individual model projections are subject to biases created by structural model uncertainties, and some models reproduce certain climate processes better than others.

Can we use model performance metrics to improve ensemble-mean climate simulations?

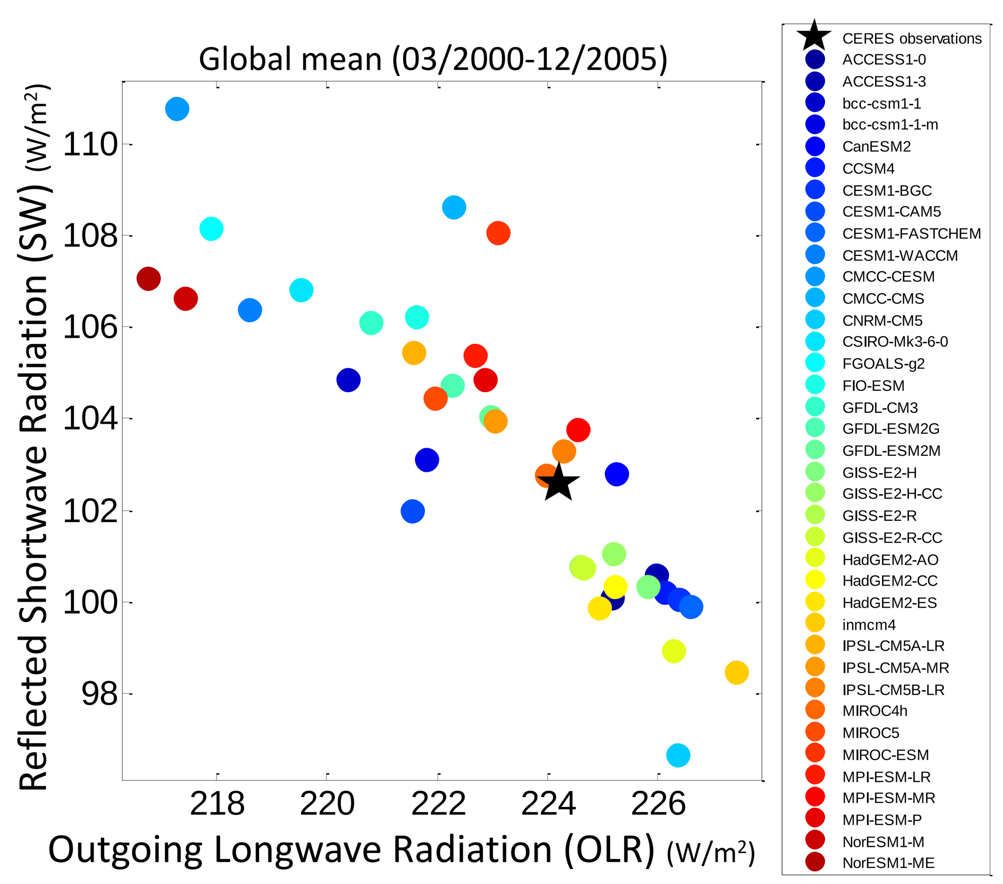


Fig. 1: Global mean top-of-atmosphere (TOA) OLR and SW radiation flux calculated using monthly mean CMIP5 coupled climate models ('historical' experiment). Observations shown as a black star from Clouds and Earth's Radiant Energy System (CERES, http://ceres.larc.nasa.gov/).

Calculation of Metrics

Hypothesis: models which are better able to simulate an important climate process produce better mean-state simulations.

The framework provides a way to test this hypothesis with many metrics. Current metrics are calculated by:

1. Using CMIP5 Pre-Industrial Control simulation data, calculate global mean detrended anomaly time series of metric for each CMIP5 model (last 100 years of simulation) and for observations (14 years of CERES data for radiation variables, 30 years of GISS data for temperature)
2. Use a statistical test to compare model and observations time series (F-test for equal variances, Kolmogorov-Smirnov test for distribution similarity)
3. Repeat the test for sequential slices of the model simulated time series (14-year slices for radiation metrics, 30-year slices for temperature metrics), add up the number of passed tests, and divide by the total number of tests (where $0 \le m \le 1$)
4. Using the metric value as a weight for each model, apply weights to the CMIP5 historical experiment and calculate weighted ensemble means for 20th-century climate quantities

The metrics represent model skill in reproducing observed climate processes.

28 metrics are used to test model performance related to radiation budget

All metrics are computed from a global mean detrended deseasonalized anomaly time series. Statistical tests for equal variances (F-test) and distribution similarity (K-S test) are performed on radiation and temperature time series.

Two statistical concepts are introduced as new performance metrics:

Local variance is defined as the variance of the first derivative of the time series (calculated using forward difference approximation).

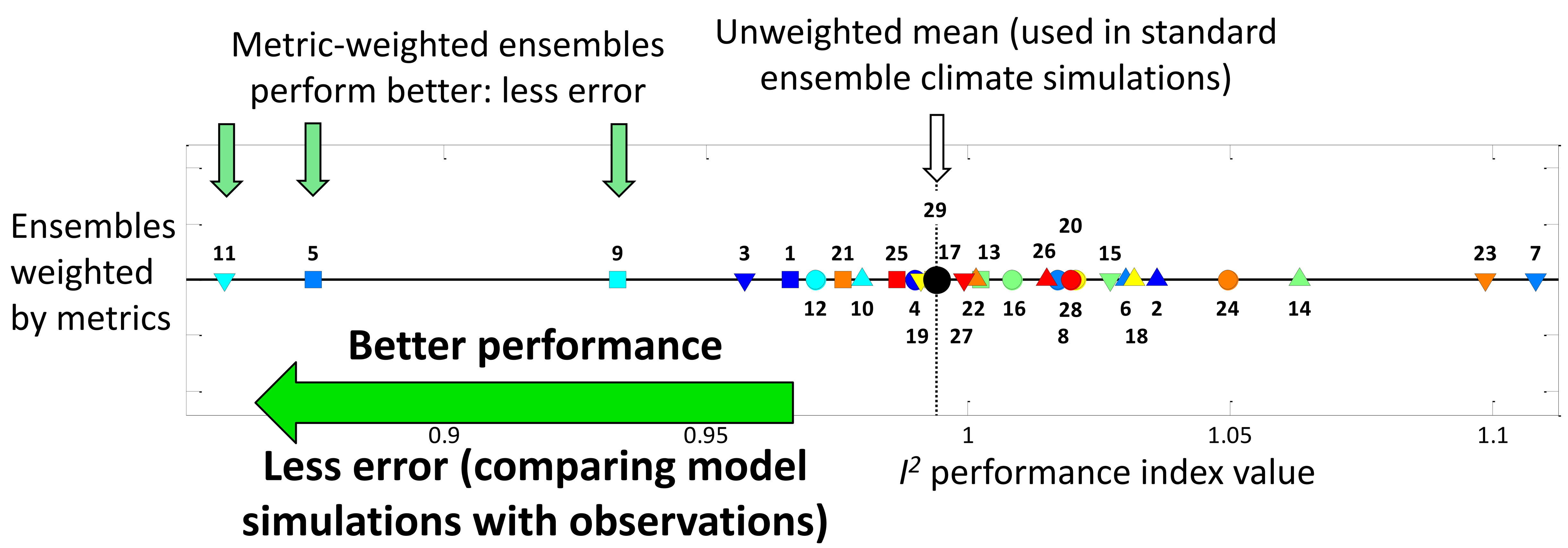
EMD, or Earth Mover's Distance, is defined as the measure of the distance between two probability distributions over a region.

Metric Performance and Results

Framework shows that some tested metrics produce better ensemble-mean simulations than standard methods

Metric performance is compared using the I^2 *performance index*.

Adapted from a model performance metric from Reichler and Kim (2008) by computing the normalized error for each weighted ensemble by squaring the grid-point differences between simulated and observed climate.



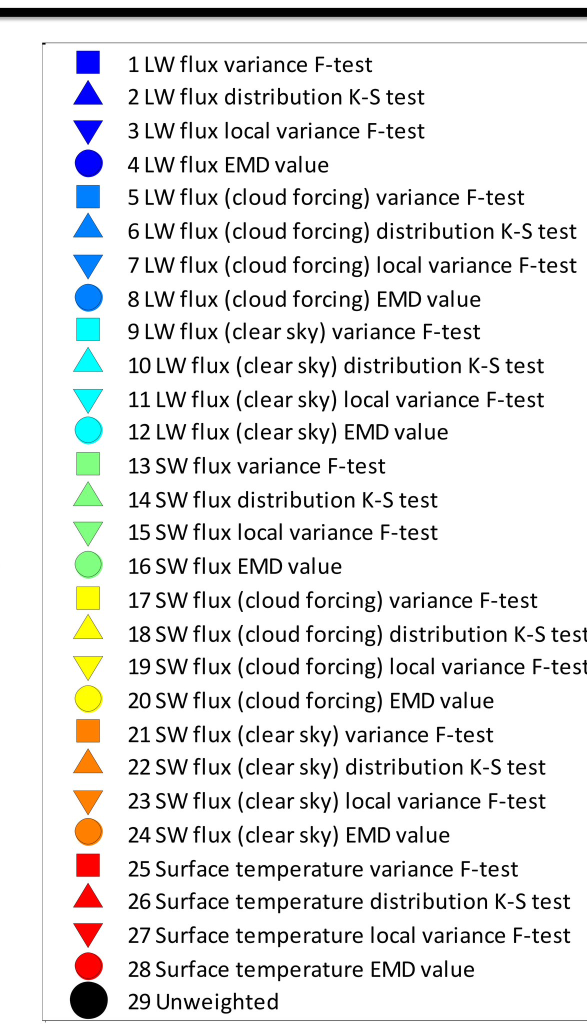


Fig. 2: Performance index I^2 of tested metrics. Colored markers indicate the value of I^2 for weighted ensemble means corresponding to numbers in legend. The black marker indicates the performance index value of the unweighted mean. Lower values (left side of plot) indicate less error relative to observations. Each I^2 value is the average of aggregated errors for 12 quantities related to the radiation budget.

Best-performing metrics are related to *OLR variance*

Fig. 3: Metrics values for all tested models, given for the top five performing metrics (listed in the legend in the order of highest performance). Tallest bars indicate models which perform well in all five metrics. Metric values are used as model weights to produce the weighted ensembles in Figure 1.