Creating "Intelligent" Ensemble Averages Using a Process-Based Framework

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

Introduction and Motivation

- The CMIP5 archive contains climate projections from over 50 models provided by 20 modeling centers from around the world.
- Individual model projections are subject to biases created by structural model uncertainties, and some models reproduce certain climate processes better than others.
- Ensemble averaging of multiple models is used to add value to individual model projections by constructing a consensus projection.

Question: Should models be weighted based on performance?

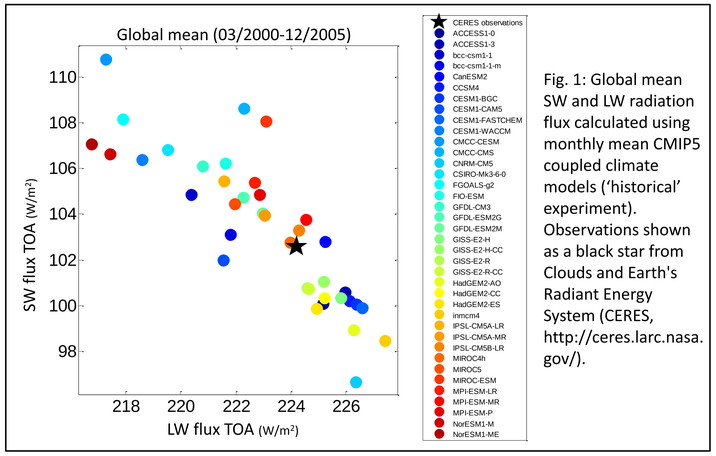


Fig. 1: Global mean SW and LW radiation flux calculated using monthly mean CMIP5 coupled climate models ('historical' experiment). Observations shown as a black star from Clouds and Earth's Radiant Energy System (CERES, http://ceres.larc.nasa.gov/).

Project Goal

- Develop a framework for systematically testing metrics to identify optimal metrics for unequal weighting multi-model ensembles. The intention is to produce improved ("intelligent") unequal-weight ensemble averages.
- A unique aspect of this project is the construction and testing of climate process-based model evaluation metrics, defined as a metric based on the relationship between two or more physically related climate variables.

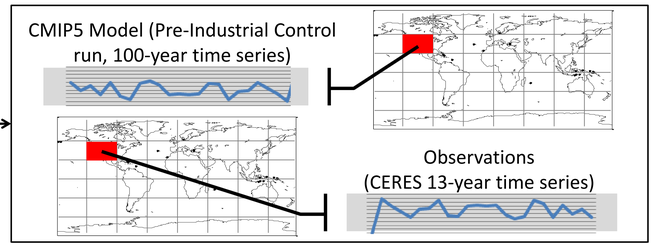
Approach: create and compare performance of model weights using:

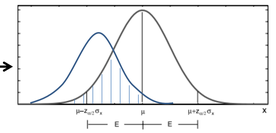
1) Mean state, variability, and frequency distributions
2) Process-based metrics (e.g. components of climate sensitivity, Nino3.4 index, MJO)

The goal of the framework is to advise better methods for ensemble averaging climate models and create better climate predictions.

Example Calculation of Weighting Metric

1. Select resolution (per grid point, regional, global)
2. Calculate anomaly time series for the model (control run: Pre-Industrial Control) and observations (detrended)
3. Use statistical test to compare time series (F-test for equal variances, Kolmogorov-Smirnov test for distribution similarity)
4. Calculate p-value of test to determine "amount of overlap" between distributions (large p-value: high similarity): $0 \le p \le 1$
5. Repeat for each grid point (or region): obtain map of weights for each model
6. Apply weights to chosen experiment (AMIP, 'historical', RCP future scenarios) for selected variable
7. Calculate weighted ensemble mean





.1	.6	.7	.3	.9	.5	.1	.4
.2	.4	.1	.6	.2	.3	.1	.5
.6	.7	.7	.8	.7	.8	.3	.4
.4	.3	.4	.7	.6	.2	.1	.4
.6	.4	.6	.4	.5	.8	.7	.3
.5	.6	.7	.4	.8	.7	.6	.4

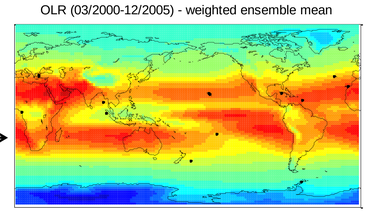


Fig. 2: Weighted ensemble mean OLR (W/m²): CMIP5 historical experiment: all models. Weighting metric determined per grid point, per model by comparing variability of 100-year model (control experiment) time series to 13-year observed (CERES) detrended time series monthly variability (calculated as p-value of statistical F-test for equal variances).

Results and Future Work

Tested process metrics:

Longwave component of climate sensitivity

$$\frac{OLR\ anomaly}{surface\ temperature\ anomaly} = \frac{\delta OLR}{\delta Ts}$$

$$\frac{(longwave\ cloud\ component)\ anomaly}{surface\ temperature\ anomaly} = \frac{\delta LWcf}{\delta Ts}$$

Hypothesis: models better able to simulate the process metric produce better overall simulations.

Initial results: weighted means can show significant differences from equal weight mean

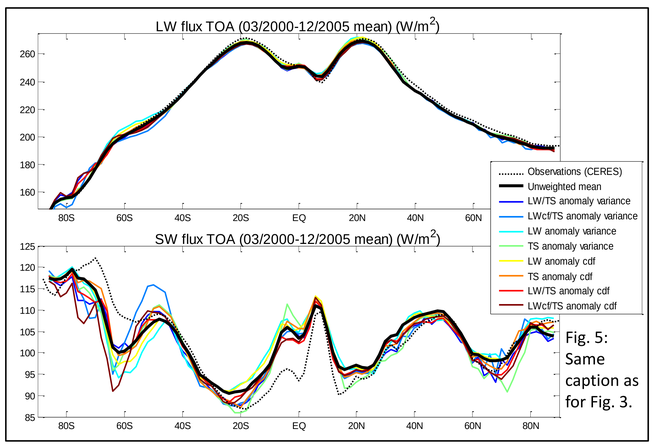


Fig. 5: Same caption as for Fig. 3.

Some weighting metrics improve model mean for certain variables and regions

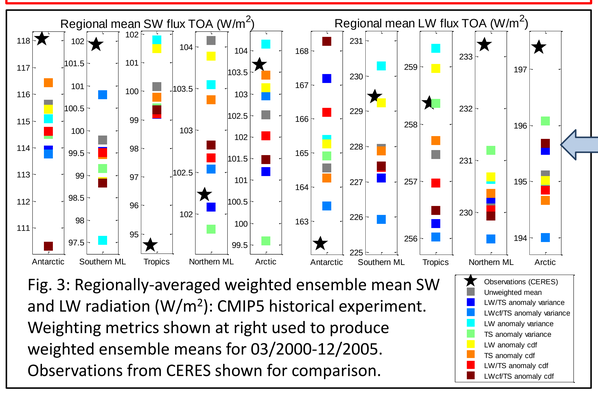
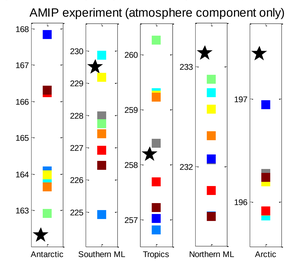


Fig. 3: Regionally-averaged weighted ensemble mean SW and LW radiation (W/m²): CMIP5 historical experiment. Weighting metrics shown at right used to produce weighted ensemble means for 03/2000-12/2005. Observations from CERES shown for comparison.

Fig. 4: Same as Fig. 3 but ensemble means were produced using simulation data AMIP present-day experiment (LW flux, W/m²).



Question: Does the weight computed for each model depend on the chosen state?

Future work and exploration questions:

- Test a variety of weighting metrics for different quantities and regions (Which process metrics are most important for constraining different quantities? Are some metrics better for certain regions?)
- Test scale- and state-dependency of weights (i.e. Do the same models perform consistently better or worse?)
- Apply weights to future projections to construct new ("intelligent") climate predictions

Metrics will be tested on 21st century trends using a "perfect model" approach

- In lieu of having future observations, one model can be treated as the "perfect model" to create weights
- Approach will be repeated choosing different models as the "perfect model" (way to test sensitivity/robustness of metrics)