Forrest M. Hoffman[1] and Martial Mancip[2]

1. Computational Earth Sciences Group, Oak Ridge National Laboratory, Oak Ridge, Tennessee, USA
2. Institut Pierre Simon Laplace (IPSL), Paris, France

Working Group report on

Terrestrial biosphere model evaluation

About 15 of the iLEAPS-Marie Curie workshop attendees from Africa, Europe, and the United States participated in a lively discussion focussed on the evaluation of terrestrial models typically run within general circulation models (GCMs). The task was to recommend the best methods for thoroughly evaluating scientific model performance. Initially, group members identified a variety of difficulties that impede such evaluations, including

1) the mismatch between the spatial and temporal scales of measurements and models,
2) limits of model assumptions, and
3) the dangers of tuning models for specific geographic regions / forcing (input) data / execution modes (offline or coupled).

The group formulated a list of elements important for organised and methodical model-data comparisons. These elements are

- an experimental protocol designed to elucidate model performance under past, present, and future climates across all relevant space and time scales;
- metadata standards to simplify manipulation and analysis of model results, including standardising biome and carbon pool types;
- evaluation metrics based on comparison of model results with best available satellite- and ground-based observational data sets;
- standardised diagnostics supporting all metric comparisons;
- a scoring methodology based on a community-developed weighting of model performance on metrics, taking into account importance and data uncertainty; and
- open distribution of model results, supporting related research by the wider community.

The group thought it important that models should be evaluated based, as much as possible, on our understanding of individual processes. Therefore, performance metrics should be based on comparison against measurements of processes such as photosynthesis and phenology instead of, for instance, global CO_2 fluxes with multiple error sources.

Moreover, because there are many ways to get "the right answer for the wrong reason," a comprehensive evaluation of model processes must include comparisons of a wide array of model variables. Also discussed was the importance of combining many data sets of similar observations for comparison with model results and of processing data sets in a consistent manner.

The group constructed lists of forcing and evaluation data sets, and group members described many of the strengths and weaknesses of these data. Commonly used meteorological forcing data sets include NCEP/NCAR reanalysis (National Center for Environmental Prediction/National Center for Atmospheric Research, 1948–2004), CRU (Climate Research Unit of the University of East Anglia, 1850–present), NCC (NCEP Corrected with CRU, 1949–2000), and ERA-interim (European Centre for Medium-Range Weather Forecasts reanalysis, 1989–2007).

Sources of observational data identified were the FLUXNET and AmeriFlux sites for surface energy and carbon flux measurements, Free-Air Carbon Dioxide Enrichment (FACE) sites for vegetation response to increases in CO_2, river gauge and GRACE (Gravity Recovery and Climate Experiment) satellite observations for hydrological measurements, NOAA (National Oceanic and Atmospheric Administration) flasks for records of the CO_2 seasonal cycle, National Aeronautics and Space Administration (NASA) MODIS (Moderate Resolution Imaging Spectroradiometer) and other satellite products for phenology and carbon fluxes, and tree rings and other proxies for climate and disturbance.

In addition, an effort was made to characterise the spatial (small to large) and temporal (short to long) scales of a variety of individual processes and variables/characteristics. The group felt it was important to develop metrics that would consider model performance across all relevant scales.

Recommendations from the group discussion were to

- write a review paper on the current state of best available data sets for model evaluation;
- encourage the development and sharing of "best" data sets by the community;
- better document model processes to improve understanding of evaluation results;
- encourage closer collaboration between modelling groups;
- encourage closer collaboration between measurement and modelling communities; and
- establish a mailing list to continue the model evaluation discussion and invite others in the research community to participate. Interested researchers can subscribe to this mailing list at: www.climatemodeling.org/mailman/listinfo/land-eval

In conclusion, the group reiterated the importance of confronting models with observations and that this should be done early and often. Models must be tested and evaluated in offline, partially coupled, and fully coupled modes over short and long time scales and over small and large spatial scales. Experiments should include historical, present-day, and future time periods.

Finally, everyone agreed that these are challenging and time-consuming tasks, but that we should work together to take advantage of one another's efforts and expertise. An international model evaluation effort focusing on models to be used in the upcoming IPCC (Intergovernmental Panel on Climate Change) Fifth Assessment Report could be the first step in building such a wide collaboration. ■

forrest@climatemodeling.org
Martial.Mancip@ipsl.jussieu.fr