Report from HydroCLM Working Group Meeting

CUAHSI-NCAR Working Group Meeting - Benchmarking Hydrologic Model Performance

Report to NSF, December 2014

Participants

Jennifer Adam (Washington State University)
Ed Beighley (Northeastern University)
Diogo Bolster (University of Notre Dame, CUAHSI Board of Directors)
Martyn Clark (NCAR)
Cedric David (JPL NASA)
Alejandro Flores (Boise State University)
David Gochis (NCAR)
Rick Hooper (CUAHSI)
Mukesh Kumar (Duke University)
David Lawrence (NCAR)
Ruby Leung (Pacific Northeast National Laboratory)
Scott Mackay (CUAHSI Board of Directors, University at Buffalo, *SUNY*)
Reed Maxwell (Colorado School of Mines)
Scott Peckham (University of Colorado, Boulder)
Yadu Pokhrel (Michigan State University)
Ying Fan Reinfelder (Rutgers, *The State University of New Jersey*)
Chaopeng Shen (Pennsylvania State University)
David Tarboton (Utah State University)
Audrey Sawyer (Ohio State University)
Sean Swenson (NCAR)
Xubin Zeng (University of Arizona)

Summary

A second working group meeting was held on Dec 16th during AGU, as a part of the CUAHSI-NCAR collaborative effort to advance hydrologic process representations in Earth System Models (ESMs), with 4 short presentations to stimulate discussions. Reed Maxwell shared lessons from past benchmarking projects. David Gochis presented a multi-scale multivariable framework beyond the traditional scores (e.g., correlation) to include, in time low-to-high frequency and extreme events, and in space patterns of heterogeneity and organization, giving us a unified framework to evaluate the many models or versions. David Lawrence introduced ILAMB, a community benchmarking system for global land models in ESMs with emphasis on breadth of performance (e.g., carbon cycle, hydrology), an open-source system, and links with observation and experimental communities. Martyn Clark presented a broad overview of the benchmarking problem, including using synthetic test cases from theories and lab experiments, multi-scale multivariate observations from CZOs, and advancing the “science” of benchmarking, e.g., by applying information theory to evaluate how models use available data.

The discussions highlighted the rich set of data available and many possible approaches, posing challenges to defining a single path forward. But some common threads emerged: (1) an initial focus on multivariate and multi-scale model evaluation making use of the data from CZOs and other watersheds; (2) focus on natural and well-constrained managed systems, and (3) focus on ET, the most critical flux at the land-atmosphere interface. The group agreed that an immediate task is to apply CLM to simulate ET

at a small collection of research catchments. A postdoc will be recruited to assemble the test cases with forcing and evaluation datasets that can be used to evaluate any hydrology model at a range of scales, facilitating an open community process to model development. The focus on using data from CZOs and other experimental watersheds will bring in the rich set of subsurface observations not yet tapped for testing large-scale models, and will help illuminate the "dark" side of the land-atmosphere interface.

1. Background

As part of a collaborative effort between CUAHSI and NCAR to improve hydrologic representations in weather, climate and earth system models, the first working group meeting was held in September 2014 at NCAR which outlined a long-term vision and short-term tasks. The specific recommendations from the September 2014 meeting were to (1) hold a second working group meeting at AGU to define model performance metrics, (2) position the currently funded CUAHSI postdoc to develop specific test cases and assemble the dataset needed to evaluate CLM hydrologic performance, and (3) seek support for a Hydrology Process Team (HPT) to conduct synthesis to recommend best ways to represent hydrology in ESMs, and to implement and test the new model components.

Here we report the discussions and conclusions of the second working group meeting at AGU and the next steps for model benchmarking.

2. Second Working Group Meeting at AGU

The meeting was held on December 16th in San Francisco. The objective was to develop a framework and the datasets for evaluating hydrologic model performance against observations. Although this is part of the CUAHSI-NCAR collaboration on large-scale modeling, we defined the scope more broadly as to include hydrologic models of all scales, so that the resulting methods and datasets are useful for the broader hydrologic modeling community. Key questions are: how do we measure progress in hydrologic model developments? Specifically, what are the essential hydrologic variables and the scales that are both observed and simulated by the models, that allow us to best diagnose model deficiencies, understand model sensitivities, inter-compare among different process representations, and establish milestones of modeling advances?

Four short presentations helped stimulate the discussions. Reed Maxwell shared lessons from past hydrologic benchmarking projects, emphasizing openness, transparency, and benchmarking as a mechanism to build the community rather than divisive competitions.

David Gochis presented a multi-scale and multi-variable framework, starting from streamflow data. It takes us beyond the traditional skill scores (e.g., absolute error, correlation, root mean square error, at arbitrary scales) to include, in time, low-to-high frequency hydrologic responses and extreme events, and in space, fundamental patterns of heterogeneity and organization. The spatial and temporal scales and the variables (e.g., streamflow, ET, soil moisture, surface skin temperature, snowpack, inundation area and depth, and groundwater level) together form a multi-axis space in which each model can be placed in relation to others, giving us a unified framework to evaluate the many models and their many versions. As Jennifer Adam pointed out, this framework can be used as a test-bed for model development, and allows for a user to know which model is best for specific processes and at specific spatial-temporal scales, and to decide what modules to enable or disable.

David Lawrence introduced ILAMB, a community benchmarking system for global land models in ESMs with emphasis on the breadth of performance (e.g., in carbon cycle, ecosystem, surface energy, and hydrological processes), an open-source software system, and linkages with the observational and experimental communities. Hydrologic benchmarking, particularly at the large-scale end, can learn from and potentially contribute to the ILAMB system.

Martyn Clark presented a broader overview of the benchmarking problem, including (i) the use of synthetic test cases from theories and laboratory experiments at column and hillslope scales; (ii) multi-scale and multivariate observations from CZOs, and (iii) advancing the "science" of benchmarking, e.g., by applying concepts of information theory to evaluate how models use available data on meteorology, vegetation, soils and topography.

The discussions, involving members from the hydrologic, climate and earth system modeling communities with interests in a wide range of variables and scales, highlighted the rich set of data that can be used and the many possible approaches, which also posed challenges to defining a single path forward. But some common threads emerged.

- Of all possible model evaluation directions, the group focused discussion on multivariate and multi-scale model evaluation making widespread use of the data from CZOs and other research watersheds. This focus defers work on use of synthetic test cases for model benchmarking [*Maxwell et al.*, 2014] and also defers work on the use of information theory for model benchmarking [*Abramowitz et al.*, 2008; *Nearing and Gupta*, 2014]. This simple start will allow us to make rapid initial progress within the funding and time constraints.
- There was some discussion of the appropriate spatial scale(s) for model analysis, and the extent to which analyses at larger spatial scales are complicated by the effects of water management which may be poorly represented in some models. The consensus is to focus on natural processes as much as possible, analyzing multi-scale behavior to the extent that it is possible to do so in unregulated river basins. Regulated river systems (e.g. ones including a small reservoir with known operating characteristics) would be considered provided the impacts of management operations could be quantified and readily integrated into the model and/or analyses. While focusing on small, unregulated basins can help isolate natural vs. anthropogenic forcings, we also need to proceed at larger basin scales where, in some cases, management impacts or reconstructed natural flows are available. Additionally, CLM is already moving towards including such water management activities, so including well constrained large river basins is in-line with the long term CLM development goals.
- There was extensive discussion on the myriad of data available for model evaluation, and the need to define a tractable and targeted research strategy. The group converged on focusing on ET processes, as ET is perhaps the most critical process in land-surface models from both the atmosphere and the hydrologic perspective. The group also recognized that evaluating ET processes requires evaluating multiple hydrologic states and fluxes, available in CZOs and other research watersheds. Furthermore understanding total ET behavior in models necessitates understanding of the various components comprising total ET such as transpiration, soil and canopy evaporation, snow sublimation and open and ponded water evaporation processes.
- There was a substantial amount of discussion on the need for a controlled and systematic approach to model evaluation, obtained through both a hierarchal approach to model perturbations (i.e., as pointed out by Ruby and Mukesh; changing one thing at a time) and through multivariate and multi-scale model evaluation which provides scope to successively focus attention on a subset of model processes.

3. Next steps

A subset of the group discussed the path forward after the meeting. The group concluded that a good initial focus is to apply CLM to simulate biophysical and hydrologic processes for a small collection of research catchments (e.g., < 5 catchments) ranging in size from a few tens of sq. km to a few thousand square km, and evaluate CLM using the suite of data available in those catchments. There is a desire to not "over-prescribe" the analysis strategy; rather, to provide the postdoc with the intellectual freedom to demine meaningful benchmarks for land models. The postdoc will be supervised by both CUAHSI and

NCAR scientists, who will provide guidance on different benchmarking approaches (e.g., the “science” of benchmarking), different data analysis strategies (e.g., focusing on multivariate signals at different space-time scales), the limitations of certain evaluation datasets and the limitations of current model evaluation metrics to diagnose specific model weaknesses (e.g., the limitations of sum-of-squared error metrics). The postdoc can immediately begin an inventory and assemble the datasets, and if necessary, will receive training and support to configure and run CLM for the selected research catchments.

The outcome from this effort will be a set of benchmark test cases including associated forcing, evaluation and geophysical characterization datasets that can be used to evaluate any hydrologic and land-surface model, facilitating an open community process to land-surface model development. For example, anyone can run their model for the same set of experiments used in the CLM evaluation, compare their results to CLM, identify reasons for better/worse model performance, and make tangible suggestions to improve hydrologic process representation in land models. Additionally, model analysis tools and scripts produced by this benchmarking effort will also be made available to the community as part of a broader effort to improve the use of community-vetted tools for hydrological model evaluation and validation. The intent is to foster greater community engagement in evaluating and improving the representation of hydrologic processes in weather, climate and Earth System models.

References cited

Abramowitz, G., R. Leuning, M. Clark, and A. Pitman (2008), Evaluating the performance of land surface models, *Journal of Climate*, 21 (21), 5468-5481.

Maxwell, R. M., M. Putti, S. Meyerhoff, J. O. Delfs, I. M. Ferguson, V. Ivanov, J. Kim, O. Kolditz, S. J. Kollet, and M. Kumar (2014), Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks, *Water Resources Research*, 50 (2), 1531-1549.

Nearing, G. S., and H. V. Gupta (2014), The quantity and quality of information in hydrologic models, *Water Resources Research*, n/a-n/a, doi: 10.1002/2014wr015895.