Hydrol. Earth Syst. Sci. Discuss., 7, C1359–C1361, 2010
www.hydrol-earth-syst-sci-discuss.net/7/C1359/2010/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments" *by* J. A. Velázquez et al.

M. Clark (Referee)

mclark@ucar.edu

Received and published: 7 July 2010

This paper on the performance and reliability of multi-model hydrological simulations is one of the best papers that I have read in some time. The use of a relatively large number of distinct models and a large number of catchments allows the authors to draw some strong conclusions on the merits and performance of multi-model systems. I have just a few comments that may improve the quality of the manuscript.

1. The authors are missing some existing papers on multi-model combination techniques, especially the papers by Lucy Marshall on hierarchical mixture of experts (e.g., Marshall et al., WRR 2005; 2006; Marshall et al., HP 2007), and the papers on bayesian model averaging (e.g., Vrugt and Robinson, WRR 2007; Vrugt et al., Env. Fluid Mech., 2008; Wohling and Vrugt, WRR 2008; as well as the seminal papers on BMA by Raftery). It may also be worthwhile to cite the papers by Neuman in groundwater, e.g., Neuman (SERRA, 2003), as well as some recent multi-model papers in meteorology (e.g., from the DEMETER project).

2. In the discussion of the rank histogram, it may be useful to cite the recent paper by Thyer et al., published in WRR in 2009. They present a cumulative version of the rank histogram, which may facilitate comparisons among multiple catchments. Note that the departure from a uniform distribution can be quantified using the KS statistic.

3. Construction of the reliability diagram requires specifying a threshold, but this threshold is never defined. Also, if this threshold represents an extreme event (e.g., a flood), it is likely that the reliability diagram is subject to substantial sampling uncertainty, especially at high probability levels. The authors may wish to consider placing confidence limits on the reliability diagram, for background see Bradley et al., 2003, published in Weather and Forecasting, and Bradley et al., 2004, published in Journal of Hydromet. and for an example see Clark and Slater 2006, also published in the Journal of Hydromet.

4. It would be interesting at some point to provide some explanation as to why the uncalibrated multi-model forecasts have poor reliability.. e.g., poor sampling from the model space, all models wrong for the same reasons, impact of uncertainty in model inputs that affects all models in the same way, etc. I understand that it is difficult to pinpoint the causes for poor reliability, but some speculation may be warranted, especially if it helps define areas for future research.

5. In terms of calibrating ensembles, it would be good to cite Wood and Schaake (Jour-

nal of Hydromet., 2008, and Johnson and Bowler, Monthly Weather Review, 2009, as well as the BMA papers (Raftery for theory and Vrugt et al. for hydrological applications).

6. This comment is motivated by curiosity: To what extent does including "bad" models degrade the probabilistic predictions? Can the authors rank the models in each basin, and present statistics on probabilistic performance when including the best (1,2,3,...,17) models? This may be of interest to other people as well.

Again, a very nice piece of work!

Cheers, Martyn.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 7, 4023, 2010.