Hydrol. Earth Syst. Sci. Discuss., 7, C2734–C2746, 2010
www.hydrol-earth-syst-sci-discuss.net/7/C2734/2010/

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.

J. A. Velázquez et al.

juan-alberto.velazquez.1@ulaval.ca

Received and published: 6 October 2010

We deeply thank Dr. Clark his objective review and important comments. Our answers follow:

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)."

The following paragraphs will be added in Page 4025 Line 3:

The selection of the best model for a given application is a complex task. For instance, Marshall et al. (2005) proposed a method in which hydrological models may be compared in a Bayesian framework accounting for model and parameter uncertainty, while Clark et al. (2008) proposed a Framework for Understanding Structural Errors (FUSE) in order to diagnose differences in hydrological model structures. The latter approach allowed the elaboration of 79 different model structures combining components of 4 existing hydrological models. Results lead the authors concluding that it is unlikely that a single model structure may provide the best streamflow simulation for basins of different climate regimes.

The following text will be added in Page 4025, Line 28:

The Ensemble Bayesian Model Averaging (BMA) has been proposed for multimodel combination (Raftery et al., 2003, 2005). In such framework, the probability density function (pdf) of the quantity of interest predicted by the BMA is essentially a weighted average of individual pdf's predicted by a set of individual models that are centered around their forecasts. The weights assigned to each of the models reflect their contribution to the forecast skill over the training period. Typically, the ensemble mean outperforms all or most of the individual members of the ensemble (Raftery et al., 2005). BMA has been successfully applied in streamflow prediction (Duan et al., 2007), groundwater hydrology (Neuman, 2003), soil hydraulic (Wöhling and Vrugt, 2008) and surface temperature, and sea level pressure (Vrugt et al., 2008). However, Vrugt et al. (2007) report no advantage when comparing multimodel BMA and Ensemble Kalman filtering (Evensen, 1994).

In meteorology, the DEMETER project aimed developing a multi-model ensemble-based system for seasonal to interannual prediction, which relies on seven global atmosphere–ocean coupled models, each running from an ensemble of initial conditions. The evaluation demonstrates the enhanced reliability and skill of the multimodel ensemble over a more conventional single-model ensemble approach. (Palmer et al., 2004; Hagerdon et al., 2005). Output from the DEMETER multimodel system has been also applied to malaria prediction models (Jones et al., 2010).

And in Page 4026, Line 14:

Marshall et al. (2006) and Marshall et al. (2007) used a hierarchical mixture of experts (HME) allowing changes in the model structure, depending on the state of the catchment. The framework, tested on Australian catchments, combines results from two models structures in the first case and two parameterizations of a conceptual model in the second case. Results showed that the HME improves performance over any model taken alone.

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."

The following text will be added in page 4032, Line 26:

Alternatives to the rank histogram exist, such as the QQplot (e.g. Thyer et al. 2009). They remained unexplored here.

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."

The following text will be added in Page 4033, Line 9:

Verification results can be quite sensitive to sampling variability in some cases (Bradley et al. 2003, Clark et al., 2006). To assess this situation, we assigned confidence limits to the reliability diagram using a bootstrap technique.

In Page 4036, Line 10:

The reliability diagrams are presented for three discharge thresholds, larger than 0 (Fig 6), than quantile 50 (Fig a) and than quantile 75 (Fig b) of the observation time series. It is noted that, for some catchments (e.g. watershed 224 and 292), there is an improvement in the reliability of the ensembles for larger discharges.

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."

We theorize that the lack in reliability in multimodel ensembles simulation (and also in forecasting) is due to many different sources of uncertainty, namely initial conditions, meteorological data, and model structure and parameterization.

The following text will be added in the conclusion of the paper (Page 4040, Line 2):

Some scientific questions rest unanswered and need to be investigated in the future. 1) How much model selection influences multimodel performance and reliability? We suggest constructing multimodel ensembles by using different type of models (e.g. distributed, lumped and even neuronal network models). We theorize that such variety may also improve multimodel ensembles, as the results obtained with different lumped model structures of this study. 2) How uncertainty in initial conditions, meteorological data, and model structure propagates during hydrological forecasting? More research assessing all sources of uncertainty should be carried and emergent tools like particulate filtering (e.g. Moradkhani et al., 2005) may help identify the uncertainty sources that should be dealt with in priority.

5."In terms of calibrating ensembles, it would be good to cite Wood and Schaake (Journal 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)."

The following text will be added in page 4031, Line 14:

Many methods for post processing the probabilistic forecasts from ensembles have been proposed, such as the ensemble dressing (i.e., kernel density) approaches (Roulston and Smith, 2003; Wang and Bishop, 2005; Fortin et al., 2006), Bayesian model averaging (Raftery et al., 2005), nonhomogeneous Gaussian regression (Gneiting et al., 2005), logistic regression techniques (Hamill et al., 2004; Hamill and Whitaker et al, 2006), analog techniques (Hamill and Whitaker et al., 2006), forecasting assimilation (Stephenson et al., 2005), statistical postprocess calibration approach (Wood and Schaake, 2008), variance inflation method (Johnson and Bowler, 2009), and several others.

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".

In Table A, we present the median CRPS obtained with the best ranked models in each basin. Best results are obtained using 11 models (0.1959) and 10 models (0.1962).

However, these results are larger than the median CRPS obtained after the optimization procedure (0.1850).

References:

Bradley, A. A., Hashino, T., and Schwartz, S. S.: Distributions-oriented verification of probability forecasts for small data samples, Weather Forecast., 18, 903–917, 2003.

Clark, M. P. and Slater, A. G.: Probabilistic quantitative precipitation estimation in complex terrain, J. Hydrometeor., 7, 3–22, 2006.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resour. Res., 44, W00B02, doi:10.1029/2007WR006735, 2008.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-Model Ensemble Hydrologic Prediction Using Bayesian Model Averaging, Adv. Water Resour., 30 , 1371–1386, 2007.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99(C5), 10,143–10,162, doi:10.1029/94JC00572, 1994.

Fortin, V., Favre, A. C., and Said, M.: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member, Q. J. Roy. Meteor. Soc., B132, 1349–1369, 2006.

Gneiting, T., Raftery, A.E., Westveld III, A.H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133(5), 1098-1118, 2005.

Hagerdorn, R., Doblas-Reyes, F., Palmer, T.: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept., Tellus A, 57(3), 219-233, 2005.

Hamill, T.M., Whitaker, J. S., and Wei X.: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts, Mon. Weather Rev., 132, 1434–1447, 2004.

Hamill, T.M., Whitaker, J. S., and Mullen S. L.: Reforecasts: An important new dataset for improving weather predictions, B. A. Meteorol. Soc., 87, 33–46, 2006.

Johnson, C. and Bowler, N.: On the Reliability and Calibration of Ensemble Forecasts, Mon. Wea. Rev., 137, 1717–1720, 2009.

Jones, A. E., and Morse, A. P.: Application and Validation of a Seasonal Ensemble Prediction System Using a Dynamic Malaria Model, J. Climate, 23, 4202–4215, doi: 10.1175/2010JCLI3208.1 , 2010:

Marshall, L., Nott D., and Sharma, A.: Hydrological model selection: A Bayesian alternative, Water Resour. Res., 41, W10422, doi:10.1029/2004WR003719, 2005.

Marshall, L., Sharma, A., and Nott D.: Modeling the catchment via mixtures: Issues of model specification and validation, Water Resour. Res., 42, W11409, doi:10.1029/2005WR004613, 2006.

Marshall, L., Nott D., and Sharma, A.: Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework, Hydrol. Process., 21(7), 847-861, 2007.

Moradkhani, H., Hsu, K.-L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the Particle Filter, Water Resour. Res., 41(5), 1–17, 2005.

Neuman, S.P.: Maximun likelihood Bayesian averaging of uncertain model predictions, Stoch. Env. Res. Risk A., 17, 291-305, 2003.

Palmer, T.N., Alessandri, A., Andersen, U., et al.: Development of a European multi-

model ensemble system for seasonal to inter-annual prediction (DEMETER)., B. Am. Meteorol. Soc., 85, 853-872, 2004.

Raftery, A.E., Balabdaoui, F., Gneiting, T., and Polakowski, M.: Using Bayesian model averaging to calibrate forecast ensembles, Tech. Rep. 440, Dep. of Stat., Univ. of Wash., Seattle, 2003.

Raftery, A.E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev., 133, 1155-1174, 2005.

Roulston, M. S. and Smith, L. A.: Combining dynamical and statistical ensembles, Tellus, 55A, 16-30, 2003.

Stephenson, D. B., Coelho, C. A. S., Balmaseda, M., and Doblas- Reyes, F. J.: Forecast assimilation: A unified framework for the combination of multi-model weather and climate predictions, Tellus, 57A, 253-264, 2005.

Thyer, M., Renard, B., Kavetski, D., Kuczera, G., Franks, S. W., and Srikanthan, S.: Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, Water Resour. Res., 45, W00B14, doi:10.1029/2008WR006825, 2009.

Vrugt, J. A. and Robinson, B. A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resour. Res., 43, W01411, doi:10.1029/2005WR004838, 2007.

Vrugt, J. A., ter Braak, C. J. F., Clark, M. P., Hyman, J. M., and Robinson B. A.: Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, Water Resour. Res., 44, W00B09, doi:10.1029/2007WR006720, 2008.

Wang, X., and Bishop, C. H.: Improvement of ensemble reliability with a new dressing kernel, Q. J. Roy. Meteor. Soc., 31, 965-986, 2005.

Wöhling, T. and Vrugt, J. A.: Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models, Water Resour. Res., 44, W12432, doi:10.1029/2008WR007154, 2008.

Wood, A.W., and Schaake, J.C.: Correcting errors in streamflow forecast ensemble mean and spread, J. Hydrometeorol., 9(1), 132-148, 2008.

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

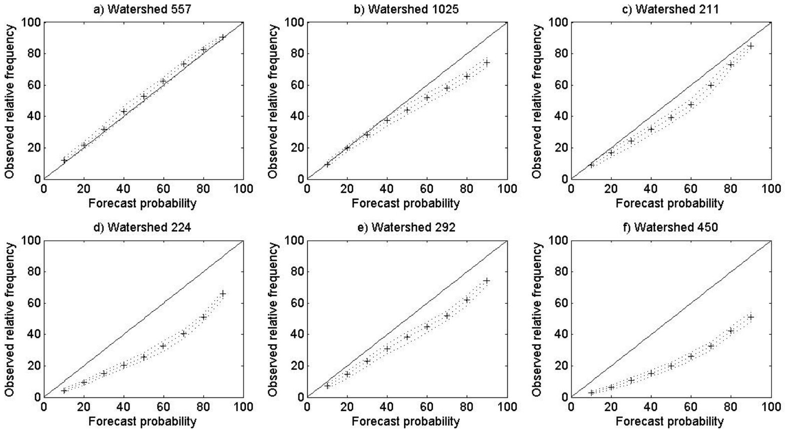


Fig. 1. (Fig.6.) Reliability plots for the same catchments as in Figure 5, for all time series. Dashed lines depict the 95% confidence interval.

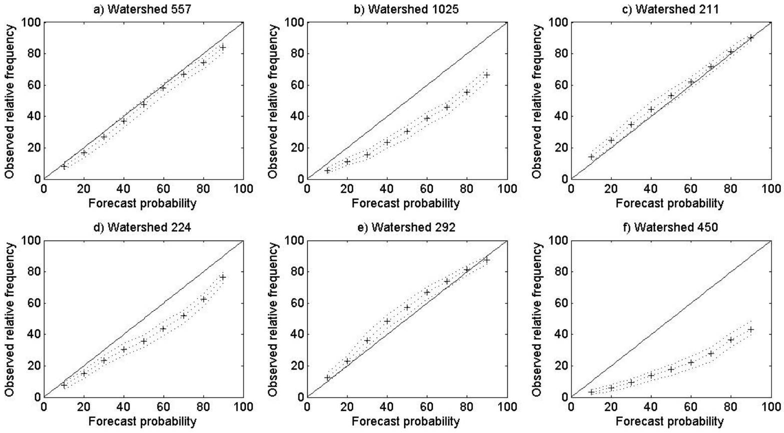


Fig. 2. (Fig.a.) Reliability plots for the same catchments as in Figure 5, for discharge larger than quantile 50 of the observation time series. Dashed lines depict the 95% confidence interval.

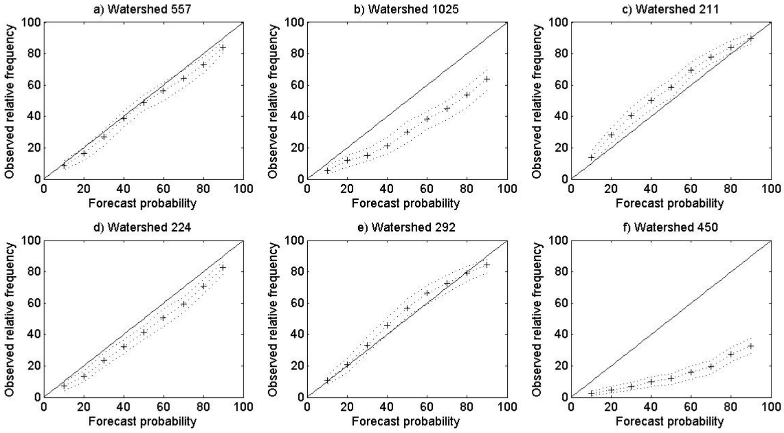
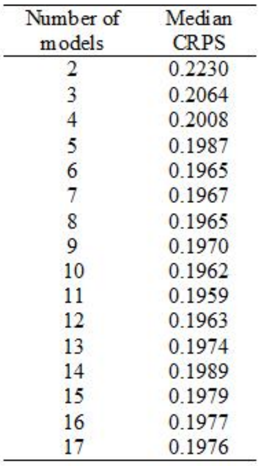


Fig. 3. (Fig.b.) Reliability plots for the same catchments as in Figure 5, for discharge larger than quantile 75 of the observation time series. Dashed lines depict the 95% confidence interval.



Number of models	Median CRPS
2	0.2230
3	0.2064
4	0.2008
5	0.1987
6	0.1965
7	0.1967
8	0.1965
9	0.1970
10	0.1962
11	0.1959
12	0.1963
13	0.1974
14	0.1989
15	0.1979
16	0.1977
17	0.1976

Fig. 4. (Table A.) Median CRPS obtained with the best ranked models in each basin.