Hydrol. Earth Syst. Sci. Discuss., 8, C6048–C6051, 2012
www.hydrol-earth-syst-sci-discuss.net/8/C6048/2012/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions" *by* G. Seiller et al.

Anonymous Referee #2

Received and published: 8 February 2012

As it is stated in Abstract, "This paper proposes a methodology to interpret hydrological projections in a climate change context and to quantify model suitability as well as their potential transposability in time." A behavior of 20 hydrological models and their ensemble was studied in contrasting meteorological conditions using Differential Split Sample Test procedure. Conclusions about performance, robustness and temporal transposability of individual models and their different combinations were drawn from the analysis of the obtained results. The problem is very actual, important and interesting. It is really important to prove that a hydrological model which has been calibrated using current meteorological forcings can (or cannot) be applied under changed climatic conditions. An interesting approach is suggested, but many things need improvement or revision.

1. Model performance is evaluated by NSE_{sqrt} (calculated on root-squared transformed streamflows) and systematic error PB. NSE_{sqrt} evaluates how the model can capture overall dynamics of streamflow, while PB evaluates volume error. However, in the climate change context it is also important to know how a model can reproduce the shape of river hydrographs, whether it is able to capture flood peaks, low flows, timing, etc. These are more important than overall dynamics in the climate change context. Low PB and high NSE_{sqrt} values don't guarantee good reproduction of different compartments of hydrograph. The selection of criteria for model evaluation should reflect the intended use of the model. Which hydrological projections are you keeping in mind?

   Besides that I am not sure that application of NSE_{sqrt} for model evaluation is much better than NSE (non-transformed). My own experience has shown that since low-flow events have much more frequency than high peaks, the peaks don't prevent appropriate calibration and validation.

2. In Section 2.3. four samples of contracted climate conditions are presented. Could you please provide mean hydrograph for each case to show how different they are (annual discharge values given in Table 1 are useful, but not very informative).

3. Section 3.1. Individual performance of each model is evaluated very formal.

   First, it is known that in catchments with drier conditions (with lower streamflow dynamics) NSE values are lower than in catchments with wet conditions (with higher dynamics) due to smaller standard deviation of observations in the denominator in Eq.2.

   Second, the drier conditions the higher PB-values, because of lower observed streamflow (which is in the denominator in Eq.3). This doesn't mean that the model performance is poor, absolute volume errors may be small and only relative (divided by observed streamflow) errors become high.

Thus, lower values of NSE and higher PB-values under dry conditions don't inevitably mean worse model performance.

Third, analysis of the results is performed separately for NSE and PB (Table 3, Figs. 5, 6, 7). However, high NSE doesn't guarantee low PB and vice versa. Thus, model M_{09} is the best in terms of NSE for DW→WC, however, its performance cannot be classified as good if it has a high bias (PB-value isn't provided).

Fourth, I think it is necessary to compare and analyze mean simulated hydrographs. To evaluate the performance of the model it is necessary to reveal whether the model is able to reproduce different compartments of hydrograph.

In general, I think that a multi-criterian evaluation of model performance should be done to draw conclusions about applicability of any model for hydrological projections.

I am not sure that application of ranks is justified. This is also rather formal (models with slightly different criteria values get different ranks, while their hydrographs may be very similar).

It would be interesting to find some comments concerning the behavior of the best and the worst models. What features make these models to be the best or the worst?

4. Section 3.2. Analysis of the collective performance is also very formal. The above comments are also related to this section. Besides that, I cannot understand the value of CV. Why high CV-values are good, i.e. why a large scatter among the models is good? I think the smaller the scatter the better.

5. In the climate context, one more test would be interesting. It is known that calibration can be more effective if it is performed on a variety of meteorological conditions. So, it would be interesting to perform calibration on contrasting conditions (by selection of several extreme years) and than to validate it as you do.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 8, 10895, 2011.