Hydrol. Earth Syst. Sci. Discuss., 8, C5982–C5984, 2012
www.hydrol-earth-syst-sci-discuss.net/8/C5982/2012/

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 #1

Received and published: 2 February 2012

This manuscript presents an interesting study where the performance of a number of conceptual runoff models and ensembles of these models is evaluated. The authors use a differential split sample approach, which allows them to address the important question of how well model perform when applied outside the calibration conditions. This makes this study a useful contribution, especially because there are not too many studies using such a test in literature (the authors list all I am aware of).

However, besides several minor issues, I have one major objection with the study. This issue is the use of the coefficient of variation to evaluate ensemble simulations. It is not clear at all why a large CV should be good. The provided reference to Brochero et al. (2011) does not provide any help either (it can also be noted that Brochero et al. use a

similar measure, but did not propose its use, this was apparently be done by (Kottegoda and Rosso, 2009)). It might be argued that a high diversity of model simulations (i.e. large CV) may provide a better performance of the ensemble simulation, but a large CV in itself does not indicate a good/suitable simulation. Actually the argument that a large CV could result in better simulations is contradicted by the results shown in figures 8 & 9, where there is no tendency for an increase of NSE with CV, but for the best models actually an opposite effect. Using different criteria certainly is valuable, but the use of CV seems unjustified. This has obvious impacts, for instance, it seems unreasonable to throw away half or more of the best ensembles just because their CV was smaller. I encourage the authors to re-do part of their analyses and to use some other, better motivated criteria!

Two other important questions could be better discussed:

Can you provide any idea/guidance on how one would ensure a "well-chosen sub-selection" (P 10910: l20)?

How do your conclusions depend on the aspect of the hydrograph you are looking at? One might argue that an ensemble is suitable for average behavior as it averages out errors of the models. For extreme flow, however, just this averaging might result in less good simulations.

Minor comments:

Using W for both wet and warm is confusing, please consider different abbreviations
Please better define the term robustness. How is it evaluated in quantitative terms?

Eq 3: the term absolute percentage bias might be misleading, rather use percentage volume error

Eq 4, exponent 2 is missing for the stdev

Table 2: add a column with the model numbers, as it is now it is not clear whether M05 refers to the fith model in the table or not.

Please separate results and discussion, mixing these parts makes it sometimes difficult to understand where the interpretation starts.

P10900, l5ff: this paragraph is not needed

P10904, l15ff: I think I understand what is meant here, but some clarification would be useful to avoid a misunderstanding of how the simulations were done for calibration/validation. If I understand it correctly, initial conditions are implicitly considered, but it is not fully clear to me, what effect the sequence of years may have. Can you comment on how long memory effects in these two catchments might be?

P10908l15ff: it is difficult to see why just these models were chosen as best-compromise solutions, M04, for instance, seems similar good as M05 for Au Saumon.

P10908, l22: actually there are 'only' 2ˆ20 -1 ensembles as one combination would be no model at all. Furthermore, it seems as single model ensembles were not considered here (no CV=0 in Figs 8&9), which further reduces the number of ensembles.

Figures 5&6: the rank figure provides basically the same information as the NSE figure, but magnifies differences. In this way, sometimes tiny differences might be overemphasized. Some information on which differences in NSE are significant would be useful. The focus on ranks seems less reasonable

Figures 10 & 11 are not really helpful. A figure of no of models against model performance (NSE) would be more interesting

---

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