Hydrol. Earth Syst. Sci. Discuss., 8, C6326–C6332, 2012
www.hydrol-earth-syst-sci-discuss.net/8/C6326/2012/

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.

G. Seiller et al.

gregory.seiller.1@ulaval.ca

Received and published: 8 March 2012

We want to thank Dr. Olga Semenova for her comments on our manuscript and work.

- Referee #3 comment concerning definitions and approach

In our study, we define model suitability as their ability to reproduce (simulate) the catchment outlet discharge (performance) in a context analogous to climate change (transposability in time). The latter is achieved deploying models under validation conditions different than the calibration ones, leading to performance and robustness evaluation under Differential Split Sample Tests. To clarify the terminology, we removed the term "suitability" and only used "performance" as a generic term to indicate the level of

accuracy of the model, and "transposability" or "robustness" to qualify the capacity of the model to perform equally well in conditions different from the calibration ones. We now provide in section 2.3 a short definition of these terms. At this point, it must be clarified that this project is not concerned with "model development" nor "hydrological modelling in ungauged basins". Further, we do not advocate that our study lead to an analysis of the physical adequacy of models structure and estimated parameters, because lumped hydrological models are not adapted for this kind of analysis and because our objectives are others. We think that a deeper analysis of the reasons why models perform well or not on the studied catchments would require more systematic testing of various model options, and complementary information on the hydrological behaviour of the catchments (see e.g. the study by Fenicia et al. 2011 and Kavetski et al. 2011 on some experimental catchments). Our first objective was to identify the level of appropriateness of each individual model under contrasted conditions (under Differential Split Sample Tests). Then, because individual results illustrate the difficulty in identifying a single lumped model that could behave well under contrasted conditions, the twenty-member multimodel is tested. Results show that this ensemble provides improved results in terms of transposability (i.e. performance and robustness) for both catchment and DSST. Pushing further the ensemble philosophy, model combinations have also been explored, using performance and robustness as evaluation tools. This latter analysis showed that identifying better sub-selections than the simple twenty-model ensemble is achievable, but need specific and detailed analysis and the consideration of some arbitrary decisions (linked to the targeted usage).

- Referee #3 comment concerning general conclusions and choice of models

> Model selection is always complex. It is thus open to criticism. Efforts have been given here in first identifying 35 models, from which 20 have been selected based on their diversified structures. Moreover, it must be kept in mind that, whatever the application context, multimodel gives echo to the "overproduce and select" philosophy dominant in machine learning. We acknowledge that this model selection is only rep-

resentative of one type of models (lumped conceptual). Selecting other model types, like distributed physically-based models, would be useful to introduce more diversity. However, given the selected evaluation framework (large number of tests using automatic calibration, putting all models in the same framework and feeding them with the same data), we chose to limit these tests to parsimonious lumped models to emphasize structural uncertainty (and not other sources like spatial variability). These limitations are now acknowledged in section 2.2.

- Referee #3 comment concerning physical adequacy

> The physical interpretation of the components of conceptual models is always difficult, since it is often impossible (or at least very difficult) to validate the internal consistency of models (problems in data availability, spatial representativeness, etc.). We think this comment also apply for other types of models like physically-based models. We had tried to put the models into a general conceptual framework describing flow components, but we acknowledge that this can create some confusion. So this paragraph was rewritten to avoid misunderstanding.

- Referee #3 comment concerning conclusions and evaluation of model performance

> Actually, in our conclusions we never argue that “a single model performance always loses in front of the ensemble” and that no “model is rejected”. Please look at our conclusions, P10912 Lines 3 to 10: “ This investigation showed that it is unsafe to rely on a single model, unless it is handpicked for each specific catchment as highlighted by best-compromise models. In particular, many models exhibited low transposability between contrasted climate conditions, whereas it is a much needed (yet seldom checked) quality for climate change applications. Taken together, the twenty models offered better climate transposability; as if the many model structure compensate for one another’s weaknesses, as illustrated by several results.”

- Referee #3 comment concerning evaluation criteria

For completeness, all validation results will be provided, including PB (now PVE), NSE (on non-transformed discharge), as well as NSElog (on log-transformed discharge) values. We will add to the text that, as highlighted by the criteria used for the analysis, models are evaluated and ranked and that different criteria could lead to somehow different interpretations. Parts of manuscript will be changed to account for that comment (see Referee #1 answers), as well as adds in Table 3 and new figures 7 and 9 (presenting PB, NSE and NSElog results for each catchment).

- Referee #3 comment concerning multimodel

> The selected models are all quite simple (it is often considered as a quality, especially in operational context) and are mainly developed based on a top-down approach. However, they often provide good simulations, as highlighted in this study based on Differential Split Sample Tests on very contrasted conditions. In the same way, multimodel ensembles are not physically justified, but often lead to performance improvement exploiting the community of experts paradigm. Note also that we think it is very difficult to a priori ("initially") select a model solely based on the perceptual analysis of dominant processes. It also often happens that the functions introduced in the model play another role, which is very difficult to identify if no additional data is available. So we think that is very complicated to judge the concept by themselves and that it is more reliable to rely on testing for model selection. Obviously, some improvements in performance may be gained by chance and other functions may be better suited. We think that the testing framework proposed here helps discriminating between models (and identifying those models that are less wrong than the others). Last, the fact that some apparently poor models are selected in multi-model combination just indicates that they still provide some useful information (probably on a limited number of events or on specific conditions).

- Referee #3 comment concerning statements and conclusions

> As said earlier, please note that this project does not aim analyzing the “model struc-

tures on their adequacy to physical processes" but to look at the level of appropriateness of each selected model in terms of transposability in time (i.e. performance and robustness) under contrasted conditions and evaluate if there is any added-value using all these models together, or a subset of them based on their performance and transposability in time. However, we introduced a discussion in the conclusion on the link of these results with previous works that investigated the issue of physical relevance of model structures (see recent work by Fenicia et al. 2011 and Kavetski et al. 2011).

- Referee #3 comment concerning technical notes and corrections

> Other criteria could have been used, but we choose to employ PB (now PVE : Percentage Volume Error) as a balance criteria, in addition to NSE, a coefficient of efficiency. > The mistake P10902 L12 will be rectified

References :

Fenicia, F., Kavetski, D. and Savenije, H.H.G., 2011. Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. Water Resources Research, 47: W11510.

Kavetski, D. and Fenicia, F., 2011. Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights. Water Resources Research, 47: W11511.

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

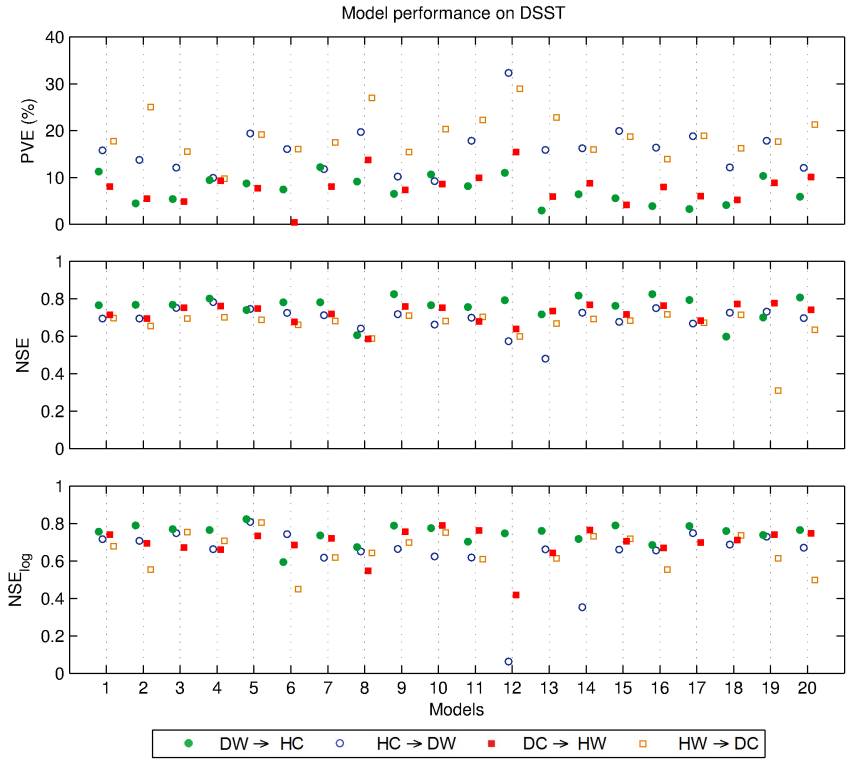

Fig. 1. Figure 7 - Au Saumon catchment

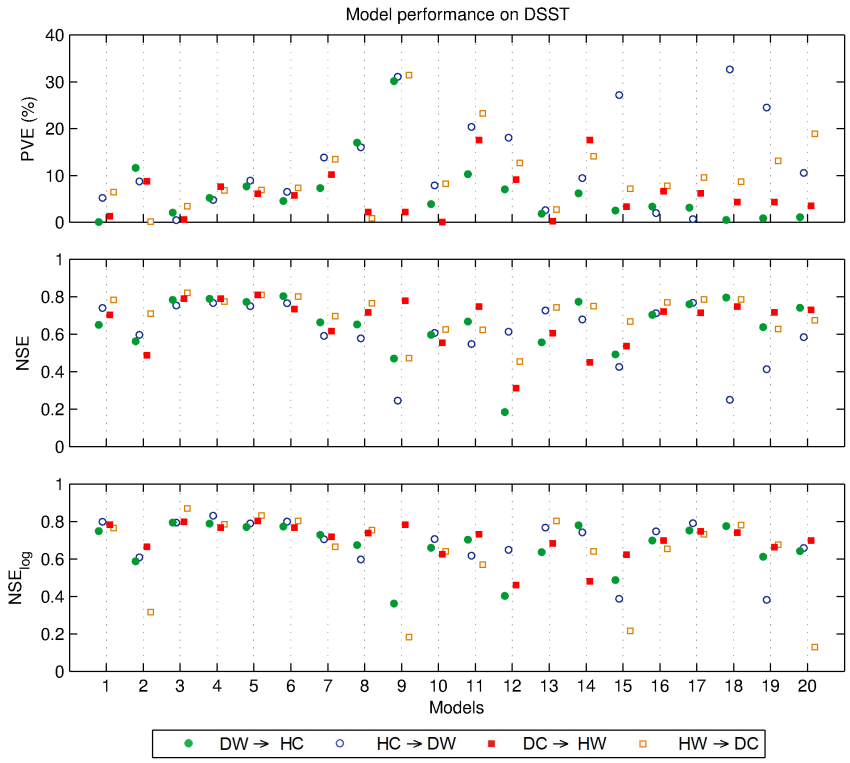

Fig. 2. Figure 9 - Schlehdorf catchment