Hydrol. Earth Syst. Sci. Discuss., 8, C6311–C6318, 2012
www.hydrol-earth-syst-sci-discuss.net/8/C6311/2012/

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 thank the Anonymous Referee #1 for his constructive comments on our manuscript and work. In the following lines we are providing detailed answers to Referee #1 comments, taking into account their added value to the manuscript. First of all, we need to clarify some points that guide our answers. This study can be seen as a three-step project, evaluating more and more into details the interest of individual models and collective ensemble regarding performance, robustness and diversity (for the ensemble approach). We define transposability (in time) as the combined qualitative synthesis of performance (adequacy of observed and simulated discharge) and robustness (model's ability to provide a similar level of good performance for different tests) un-

der contrasted conditions. The first objective is to identify the level of appropriateness of each individual model under contrasted conditions through 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 also tested. Results show that this ensemble leads to good results, as it is the best overall in terms of transposability (i.e. performance and robustness) for both catchment and all DSST. After this analysis, and pushing further the ensemble philosophy, model combinations have been explored. This analysis shows that identifying sub-selections of models that perform better than the simple twenty-model ensemble is possible, but need specific and detailed analysis and some arbitrary decisions (linked with the application objective). Identification of a sub-selection is thus not the main goal of the study and the Coefficient of Variation (CV) is used as an indication of the level of variability in the simulated discharge time series.

- Referee #1 comment concerning the use of CV

> The multimodel approach asks for the definition of several criteria for interpreting the results. We based our work on the combined analysis of various metrics: on one hand the Nash-Sutcliffe efficiency on transformed streamflows (NSEsqrt), looking at the performances of the multimodel combinations i.e. the efficiency between simulated combined streamflows and the observed streamflow; on the other hand the Coefficient of variation (CV) looking at the hydrological responses diversity for the models combinations. It seems that the use of the Coefficient of Variation (CV) is not clear enough in our manuscript. Accordingly, in the revised version of the manuscript, we will keep as potential sub-selections all the combinations with better performances than the twenty-member ensemble and using CV as a descriptive criteria of simulations variability, commenting both performance and diversity but only using performance (NSEsqrt) and transposability (DSST) as sub-selection tools. We do not advocate that a higher CV produces better performance of the ensemble simulation (see Figures 8 & 9). But we think that if a user wants to go beyond the simple use of the twenty-member en-

semble (providing yet good results, better than individual models), a better diversity (higher CV) may be required for avoiding considering, as good sub-selection, a group of models with similar simulated streamflow time series, especially if number of combinations with better performance than twenty-member multimodel is high. Diversity combined with performance offer added possibility to encompass observed streamflow and to reinforce prior statements found in the literature that an ensemble should not just be a collective of "best" models (see e.g., Velázquez et al., 2010). Comments in this way will be added, giving user the choice to keep all CV or only the ones better than twenty-member ensemble, depending on the project context and the number of combinations with better performances than twenty-member ensemble. Here are the parts of manuscript that will be changed, in response to this comment: P10906, L1 to 6 will be changed as follow: "In addition to the performance and transposability calculations, the collective diversity of the models is of concern for the multimodel approach. By seeking for diversity in the simulated time series, we aim at avoiding redundancy between the components of the ensemble model. This diversity is assessed through the mean coefficient of variation (CV) calculated on the simulated discharges (Kottegoda and Rosso, 2009; Brochero et al., 2011)" P10915, L5. Following lines will be inserted: "Kottegoda, N. T. and Rosso, R., 2009. Applied Statistics for Civil Environnemental Engineers. Electronic version Wiley, Chichester." P10908, L24 to 26 will be changed as follow: "As mentioned in section 2.4, considering CV as a complementary criterion aims at measuring the hydrological range of the model responses (i.e. structural variability), and hence avoiding model redundancy." Paragraphs concerning sub-selections are also modified in that sense.

- Referee #1 comment concerning guidance on how one would ensure a "well-chosen sub-selection"

> The notion of a "well-chosen sub-selection" depends of course on the objectives of the user. One could give more importance on the best performance combination (research or one-time project), other would favour a lesser number of models among the

best performing combinations (operational project). It must be reminded that the main goal of this study is not to focus on the best sub-selection ensemble. It is a three-step study. First, we define and apply methodology to test potential transposability in time of the twenty individual models. Second, we analyse if a simple twenty-member ensemble can provide better performance and transposability. Third, we test if sub-selections can be found. The user can conclude that twenty-model ensemble provide good enough results or look at further improved performances and diversity (as evocated in 10911 lines 14 to 17). Moreover, we highlight in Figure 12 that sub-selection offers small performance gain over the twenty-model ensemble. Here are the parts of manuscript that will be changed, corresponding to this comment : P10910, L19 to 21 will be changed as follow: "This sub-selection will be identified accordingly to the user's objectives; one may prefer a lower number of models, best performance, or a mix of performance and diversity."

- Referee #1 comment concerning aspect of the hydrograph

> For this study, we focused on hydrograph simulation (calibration on RMSEsqrt, performance evaluated with NSEsqrt) and volume error (PB, now PVE), having dam management applications in mind. Consequently, less emphasis has been given to hydrological peaks, not that they are necessarily wrong. It is often advocated that climate models still have difficulties to reproduce extremes precipitation, smoothing the data at a catchment scale, not to mention the uncertainties linked to climate modelling (emission scenarios, GCM, downscaling and bias correction). Researchers focusing on extreme hydrology have specific tools (mainly statistical) and methodology dealing with these specificities. Here, the aim is to concentrate on overall simulated hydrographs. Of course, if one is interested with extreme hydrology (flood peaks, low flows) choice of calibration's cost function and criteria for evaluation of performances could be selected accordingly, as well as the ensemble "averaging' method. The use of complementary criteria (NSE and NSElog) in the revised version will help having a more general overview on model efficiency, especially in high and low flow conditions. Here are the

parts of manuscript that will be changed, corresponding to this comment : P10905, L5 to 8 will be changed as follow: “RMSEsqrt can be considered a multi-purpose criterion focusing on the simulated hydrograph. It puts less weight on high flows than the standard RMSE (on non-transformed discharge) (Chiew and McMahon, 1994; Oudin et al., 2006).” P10905, L16, following lines will be added: “To give more emphasis on high and low flow conditions, we also used the Nash-Sutcliffe Efficiency on non-transformed streamflows (NSE) that gives more weight to large errors generally associated with peak flows, and the Nash-Sutcliffe Efficiency on logarithmic-transformed streamflows (NSElog) that puts more weight on low flows.” Other comments will be done in the entire manuscript when relevant, as well as adds in Table 3 and new figures 7 and 9 (presenting PB, NSE and NSElog results for each catchment).

- Referee #1 minor comments

> As required by the referee, we will now on use “H” for Humid instead of “W” for Wet in the manuscript. > As required, we will replace absolute percentage bias by percentage volume error (PVE) > The mistake in Eq.4 will be corrected > A Column indentifying the model numbers will be added in Table 2. > On this manuscript, it is very hard to totally separate results and discussion because it describes a three-step analysis (as evocated earlier) where individual performance is evaluated first, twenty-member ensemble is evaluated second, and sub-selection is evaluated third. Manuscript modifications will be done in this direction. Subtitles have been added in the results and discussion part. > The paragraph P10900, l5ff will be modified > P10904l15ff, we explain that we selected non-continuous hydrologic years for the DSST. For example, Dry/Warm periods correspond to years 1979, 1982, 1994, 2000 and 2001 for Au Saumon catchment. Of course this selection offers the possibility to work on the most contrasted years, but calibration cannot be done only pasting these years next to next because the catchment hydrological continuity will be lost. We opted to calibrate and validate the models simulating the entire dataset but only the selected years (example Dry/Warm years for calibration and Wet/Cold years for validation in the first case)

are used to compute the cost function (RMSEsqrt) for calibration and performance (NSEsqrt, PB) for validation years. > About comment on P10908l15ff : Of course, identification of the best-compromise models is partially subjective and arbitrary. So, for more objectivity, we will now on identify the best-compromises models as those with the best three mean ranks of performance. Parts of manuscript will be changed to account for that comment. > As highlighted by referee #1, there are 220 combinations minus 6196 combinations (because all combinations of less than five models are not used in this graph for diversity evaluation). The correct number of combinations is 1042380. Parts of manuscript will be changed to account for that comment. > A rank analysis was chosen, as evocated P10906,L22, because using NSEsqrt may include some bias due to a different denominator value for each DSST (contrasted hydrological periods). Using ranks allow circumventing this issue. Of course, ranks may sometimes be misleading too, especially when the performance of the models is quite close to one another. Overall we opted to graph both means of evaluation. We now acknowledge the possible limitations of using ranks in section 2.4. > We agree with the reviewer. Figures 10 & 11 will be removed, as well as all links to them.

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

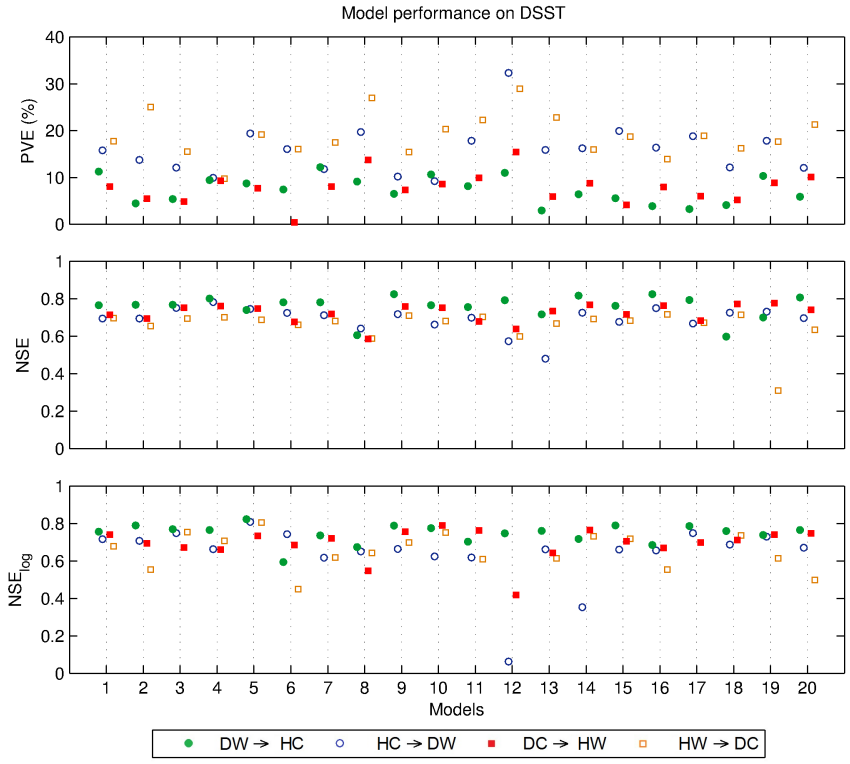

Fig. 1. Figure 7

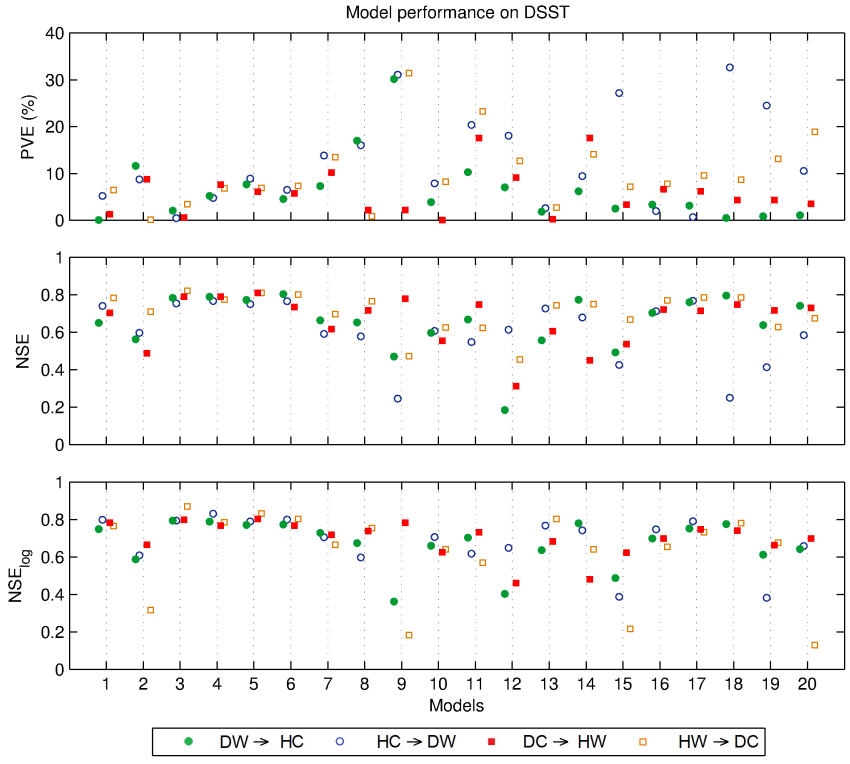

Fig. 2. Figure 9