Hydrol. Earth Syst. Sci. Discuss., 8, C6319–C6325, 2012
www.hydrol-earth-syst-sci-discuss.net/8/C6319/2012/
© Author(s) 2012. This work is distributed under
the Creative Commons Attribute 3.0 License.

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 also thank Anonymous Referee's #2 for his comments, which help improving the manuscript.

- Referee #2 comment concerning NSEsqrt (point 1)

> As mentioned in the reply to Referee #1, we have chosen to focus on the hydrograph simulation (calibration on RMSEsqrt, performance evaluated with NSEsqrt) and volume error (PB, now PVE), having dam management applications in mind. The use of transformed flow values was found more robust to estimate model parameters than the use of non-transformed ones in some earlier work cited in the manuscript (Oudin

et al., 2006) with the type of models applied in our study. This provides some kind of "all-purpose" models, but we acknowledge that other objective function might be more suitable for specific objectives. Obviously, the NSEsqrt and NSE criteria are not fully independent, but our experience shows that one gets complementary assessment using these two criteria on some catchments. 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. It may be reminded that uncertainties linked with climate modelling (emission scenarios, GCM, downscaling and bias correction) are numerous and are still of concern. Researchers focusing on extremes in hydrology have specific tools (mainly statistical) and methodology dealing with these specificities. Of course, if one is more interested with flood peaks or low flows choice of the cost function for calibration and of the criteria for evaluation should be selected accordingly, as well as the ensemble averaging' method. The use of complementary criteria (NSE and NSElog) will help having a more general overview on model efficiency, especially in high and low flow conditions. 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 #2 comment concerning mean hydrograph for DSST periods (point 2)

> As required, the mean daily hydrographs will be provided in the revised manuscript (new Figure 5). These hydrographs are mean daily interannual values for each DSST on validation periods for observed, simulated and twenty-member ensemble(e.g. for Au Saumon, test DW -> HC, graph show daily mean interannual hydrographs of years 1981, 1985, 1992, 1993 and 1995). Figure 5 is added and commented in the text.

- Referee #2 comment concerning evaluation of model performance (point 3)

> The fact that NSE leads to lower values under drier conditions is exactly the reason why we use ranks on performance combined with direct NSEsqrt in the same figure.

HESSD

8, C6319–C6325, 2012

CC BY

Due to a denominator in Eq.2 that differ from one DSST to another, we know that an evaluation based on NSEsqrt and PB is strictly valid one DSST at the time, as explained P10906 line 22 to P10907 line 3 of the manuscript. However, ranks allow comparing all models and DSST results. We thus took advantage of both situations: commenting NS-Esqrt and PB for comparison between models for each DSST and rank for comparison between DDST for each model and between models. For completeness, all PB (now PVE) results will now be presented and commented in details, as well as NSE results (on non-transformed discharge) and NSElog ones (on log-transformed discharge). We also tried to comment on the possible reasons for good/poor performance by analyzing model structures, but it is difficult to be conclusive based on the sole tests shown in the article and further analysis was beyond the scope of this paper. Figures 7 and 9 are added as well as NSElog and NSE results (see Referee #1 answers). > Mean daily interannual hydrographs for each DSST on validation period for observed, simulated and twenty-member ensemble will be commented. But it must be reminded that lumped conceptual models are empirical in nature. Here are the parts of manuscript that will be changed, corresponding to these comments: P10904, L22 add the following lines: "This results in maximum differences between periods of about 27% in mean flow, as also illustrated in Figure 5 that show the mean daily regime curve for each selected period (thick lines). In the Au Saumon catchment, strong differences appear in the spring snowmelt flood as well as in low flows. In the Schlehdorf catchment, base flow as well as summer high flows show important variations between periods." P10908, L17 add the following lines: "Figure 5 also points out the larger variability of individual models (in grey) for the Schlehdorf catchment than for Au Saumon catchment. Note that in a few cases, some models showed an outlier behaviour (e.g. M12 for the Au Saumon catchment in the HC→DW case and M09 for the Schlehdorf catchment in the DW→HC case strongly underestimate streamflows). This indicates the identification of poorly robust parameter sets in some cases, a limitation that may not appear when applying SST under similar conditions."

- Referee #2 comment concerning analysis of collective performance (point 4)

HESSD

8, C6319–C6325, 2012

Interactive Comment

> A similar answer has been provided to Referee #1. We based our analysis on different tools: on one hand the Nash-Sutcliffe efficiency on transformed streamflows (NSEsqrt), looking at the performance 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 diversity in model combination. We think that if a user wants to go beyond the simple use of the twenty-member ensemble (providing yet good results, better than individual models), a better diversity (higher CV) may be required to avoid considering, as a good sub-selection, a group of model with more similar simulated streamflows. Diversity combined with performance offer added possibility to encompass the observed streamflow. 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 only as a descriptive criteria of simulations variability, commenting both performance and diversity but only using performance (NSEsqrt) and transposability (DSST) as sub-selection tools. Here are the parts of manuscript that will be changed, corresponding to this comment: See Referee #1 answers.

- Referee #2 comment concerning calibration (point 5)

> This is a very interesting idea, which however goes beyond the objectives of the present study that focus on Differential Split Sample Tests. Here we preferred to focus specifically on the extrapolation capacity of models. However, it is likely that the suggested option should improve model robustness.

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

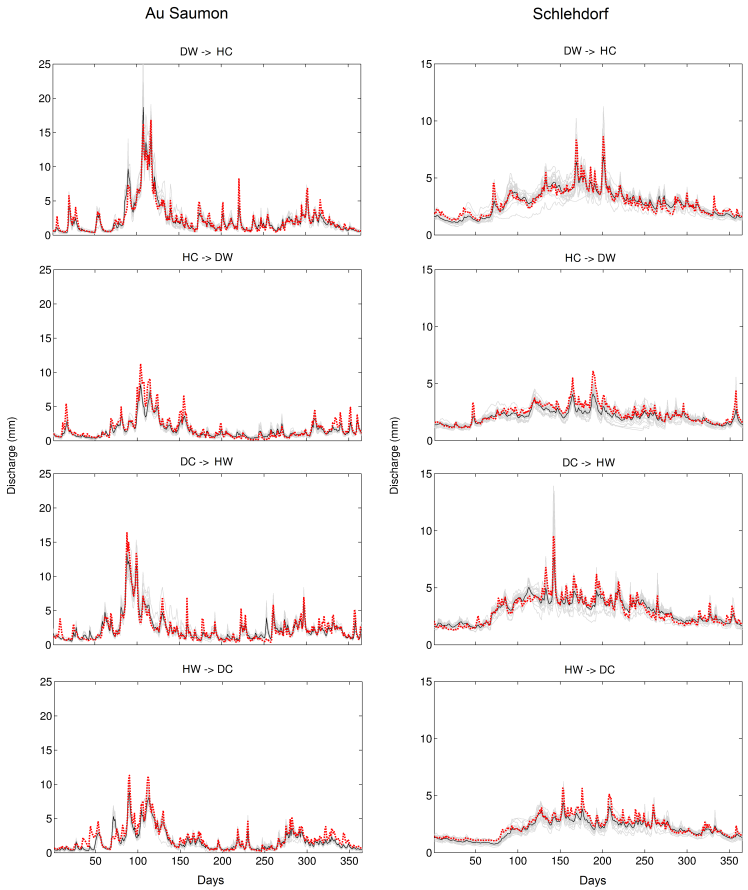

Fig. 1. Figure 5

HESSD
8, C6319–C6325, 2012

Interactive Comment

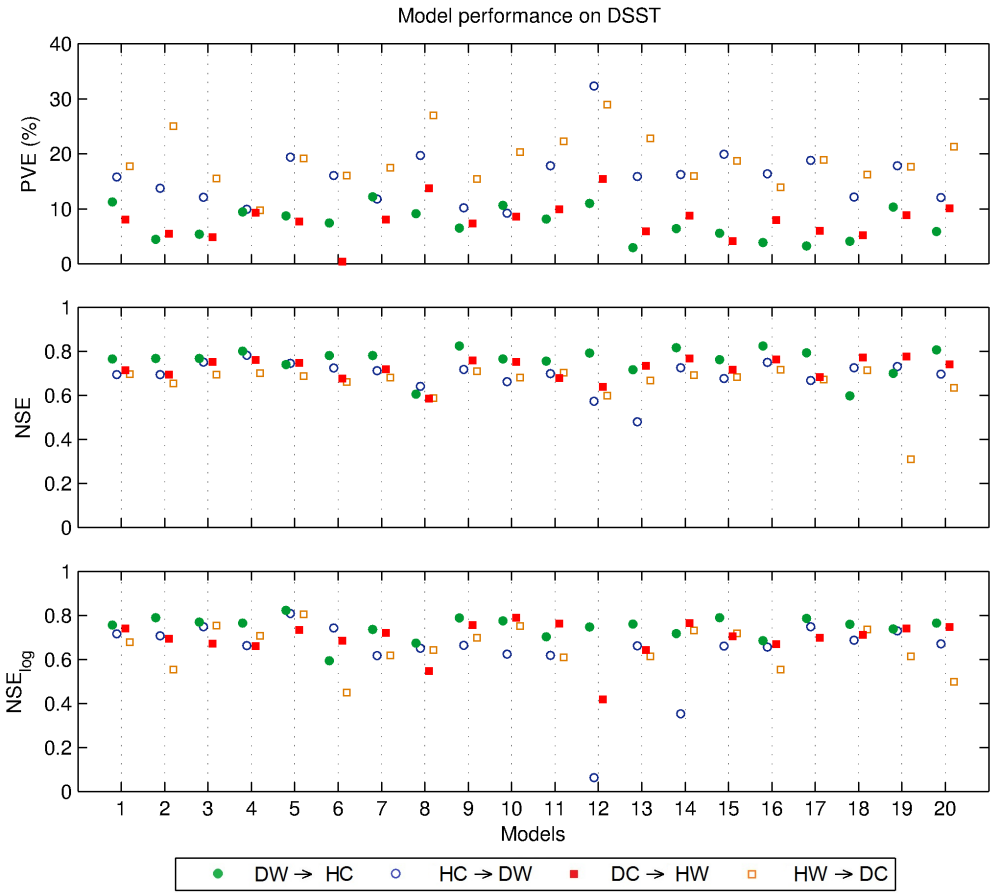

Fig. 2. Figure 7

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper

HESSD
8, C6319–C6325, 2012

Interactive Comment

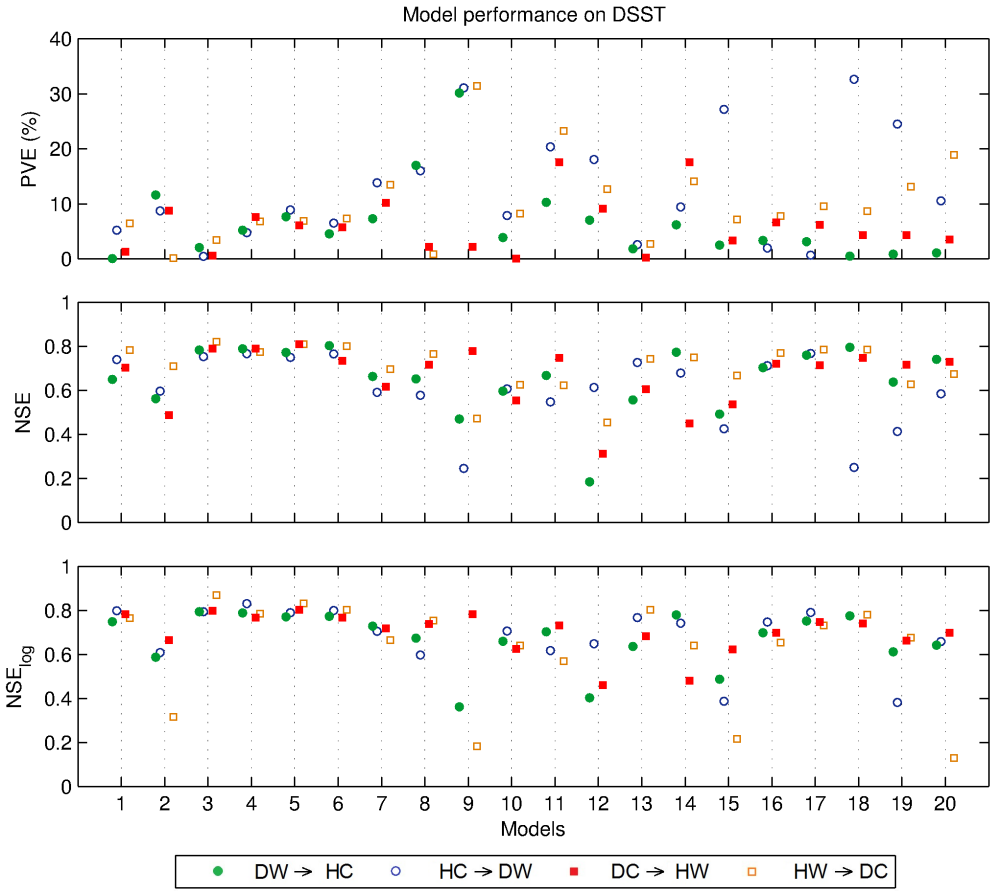

Fig. 3. Figure 9

Full Screen / Esc
Printer-friendly Version
Interactive Discussion
Discussion Paper