Hydrol. Earth Syst. Sci. Discuss., 8, 10895–10933, 2011
www.hydrol-earth-syst-sci-discuss.net/8/10895/2011/
doi:10.5194/hessd-8-10895-2011

Hydrology and Earth System Sciences Discussions

Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions

G. Seiller[1], **F. Anctil**[1], **and C. Perrin**[2]

[1]Chaire de recherche EDS en prévisions et actions hydrologiques, Département de génie civil et de génie des eaux, 1065, avenue de la Médecine, Québec, Qc, G1V 0A6, Canada
[2]Cemagref, Hydrosystems and Bioprocesses Research Unit (HBAN), 1, rue Pierre-Gilles de Gennes, 92761 Antony Cedex, France

Received: 17 November 2011 – Accepted: 24 November 2011 – Published: 9 December 2011

Correspondence to: G. Seiller (gregory.seiller.1@ulaval.ca)

Published by Copernicus Publications on behalf of the European Geosciences Union.

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. This is achieved by applying the Differential Split Sample Test procedure on twenty lumped conceptual models, for two different catchments, in the Province of Québec (Canada) and in the State of Bavaria (Germany). First, a calibration/validation procedure was applied on four historical non-continuous periods with contrasted climate conditions. Then, model efficiency was quantified individually (for each model) and collectively (for the model ensemble). The individual analysis evaluated model performance and robustness. The ensemble investigation, based on the average of simulated discharges, focused on the twenty-member ensemble and all possible model subsets. Results showed that using a single model without performing a Differential Split Sample Test may provide hazardous results in terms of climate transposability. Overall, some models turned out as a good compromise in terms of performance and robustness, but never as much as the twenty-model ensemble. Model subsets offered yet improved performance and structural diversity, but at the expanse of spatial transposability.

1 Introduction

There is a large consensus that the bulk of the adaptation strategies to climate change will be driven by water issues. Already, some components of the water cycle are of concern, such as precipitation frequency and intensity, snow cover, soil moisture, surface runoff, atmospheric water pressure, evapotranspiration, and others (Bates et al., 2008). These findings stress the importance of quantifying the impacts of climate change on the hydrologic cycle and evaluating related uncertainties.

The most common way assessing the impact of climate change on water resources combines the use of climate projections and hydrological modelling (see e.g.,

HESSD
8, 10895–10933, 2011

Prudhomme et al., 2003; Merritt et al., 2006; Maurer, 2007; Minville et al., 2008; Ludwig et al., 2009; Görgen et al., 2010; Bae et al., 2011). Four main steps must be considered in such impact studies (Boé et al., 2009): (1) constructing gas emission/concentration scenarios, (2) modelling global climate, (3) downscaling and bias correcting the meteorological projections, and (4) estimating impact with hydrological models. All these chained steps have associated uncertainties whose relative importance may differ between climate conditions and catchment characteristics.

1.1 Hydrological modelling in a climate change context

Hydrological modelling in a climate change context is a major challenge for the scientific community. Its associated uncertainties mainly emerge from structural and stochastic issues (Breuer et al., 2009). Structural uncertainties result from the simplified, incomplete, sometimes incorrect, description of the hydrological processes. They originate from the choice of the equations embedded in the model structure or from the way the model is coded (see e.g., Beven, 2000). On the other hand, stochastic uncertainties are generated by errors in input (e.g. precipitation, temperature) and output data (discharge), which are caused by difficulties and limitations in measurement and spatialization techniques. Various studies already analyzed the propagation of data errors in the modelling process (Andréassian et al., 2001, 2004; Oudin et al., 2006a, b; Perrin et al., 2007). Yet stochastic uncertainty is also linked to parameter identification since the model parameters are often determined through a calibration procedure exploiting one or more objective functions. This commonly used procedure may face equifinality issues (Beven and Freer, 2001). Model validation strategies, which should help confirming the applicability and the accuracy of the calibrated model, are also a source of uncertainty in the way they are performed: less demanding model testing may result in underestimating uncertainty.

Another difficulty in using hydrological models in climate change impact studies arise from the need of identifying model parameters that are suitable for both current and future conditions. This difficulty stems from the non-stationary nature of climate.

HESSD
8, 10895–10933, 2011

Common practice usually assumes that parameters associated to the hydro-climatic conditions of the calibration data set remain valid in other test periods, making implicit the assumption of the stationarity of the rainfall-runoff transformation. This assumption generally holds when application conditions are not much different from the calibration ones. However, in a climate change context, the contrasts of climate conditions between the calibration and projection periods are important, thus questioning the stationarity hypothesis. Hence model transposability in time under contrasted conditions must be analyzed in details and could even become a criterion for the selection of modelling tools to be used in impact studies.

To this end, demanding validation methods must be designed. Several authors proposed, adapted, or applied testing schemes to evaluate models' ability to perform well under contrasted climate conditions (Refsgaard and Knudsen, 1996; Xu, 1999; Donnelly-Makowecki and Moore, 1999; Seibert, 2003; Xu et al., 2005; Refsgaard et al., 2006; Görgen et al., 2010; Vaze et al., 2010; Merz et al., 2011). All are inspired by the "Hierarchical scheme for systematic testing of hydrological simulation models" formulated by Klemeš (1986), which identified four levels of model tests, among which is the Differential Split-Sample Test (DSST). The principle of DSST is to calibrate the model on data prior to a change (pre-change) and validate it on post-change data. In the context of climate change projections, present and future conditions must then be confronted. Since by definition, future observations are not yet available, the identification of post-change data is impossible and so the actual model evaluation. As a surrogate, one may use existing observations to calibrate and validate models on time periods with dissimilar climatic characteristics, thus mimicking the contrast between present and projected future conditions (even if the contrast may in fact be smaller). According to Refsgaard and Knudsen (1996), "a model is said to be validated if its accuracy and predictive capability in the validation period have been proven to lie within acceptable limits or errors". The application of DSST in this perspective may help evaluating the limits of hydrological models for climate change impact studies and their associated uncertainties.

HESSD
8, 10895–10933, 2011

1.2 Model intercomparison and multimodel ensemble

Because models are abstractions of real systems, it cannot be anticipated which one offers more accuracy and predictive capability for specific catchments and hydrologic conditions. Model intercomparison has been identified as a convenient mean approaching this issue (e.g., Chiew et al., 1993; Refsgaard and Knudsen, 1996; Perrin et al., 2001; Reed et al., 2004; Breuer et al., 2009; Görgen et al., 2010; Bae et al., 2011). The main goal of an intercomparison study is evaluating multiple representations of the hydrological behaviour, beyond a single deemed "appropriate" model. Moreover, it offers the possibility of quantifying uncertainty (probabilistic approach) for different conceptualizations of the reality.

Beside model intercomparison, an ensemble perspective can be evaluated through multimodel combination. Multimodel aims at extracting as much information as possible from the existing models. The rationale behind ensembles is that simulations from a single model contains errors from several sources, but that the combination of several models with different concepts and aims of development may compensate each other and provide better results than the deterministic approach (Ajami et al., 2006). For instance, Shamseldin et al. (1997) combined five hydrological models. Their results indicate that the multimodel combination performs generally better than the use of any single model. Similar conclusions were drawn by Loumagne et al. (1995), Georgakakos et al. (2004), Butts et al. (2004), Ajami et al. (2006), Kim et al. (2006), Duan et al. (2007), Viney et al. (2009), and Velázquez et al. (2010).

1.3 Objectives

Hydrological models used in climate change studies are subject to similar stochastic uncertainties, which arise from the climatology, but dissimilar structural uncertainties. The confrontation of a selection of hydrological models is an appropriate way to address the latter uncertainties. However, the lack of evaluation of the hydrological uncertainty under a contrasted forcing (i.e. "risky conditions") is detrimental to our capacity of

HESSD
8, 10895–10933, 2011

interpreting projections. Unfortunately, this step is often ignored. This paper explores the structural uncertainties of a selection of twenty lumped conceptual models through DSST. The main idea is to quantify their suitability under climate change, following two points of view: individual and collective (ensemble).

The material and methods section presents the catchments, the twenty lumped conceptual hydrological models, the adaptation made to the DSST, and the performance criteria. The third section, Results and discussion, mainly addresses the following two questions: What is the level of appropriateness of each selected model, in terms of transposability in time (i.e. performance and robustness), under contrasted conditions? Is there any added-value using all these models together, or a subset of them based on their diversity, performance and transposability in time? Finally, conclusions are drawn in the fourth section.

2 Material and methods

2.1 Studied catchments

Two basins are studied here: the Haut-Saint-François River in the Province of Québec (Canada) and the Isar River in the State of Bavaria (Germany). The Canadian study site is representative of water management for hydroelectric production, flood protection and recreational activities, while the German one is typical of catchments with strong anthropogenic impacts (i.e. soil sealing, stream realignment/channelization, dam construction, etc.). The Haut-Saint-François River is subject to a snow-melt maximum in spring and high discharges in fall. The Isar runoff regime is characterized mainly by alpine snow-melt in spring and a strong summer precipitation maximum.

A single natural sub-catchment for each respective system is studied in order to avoid additional complexities linked to dam management: the Au Saumon (SAU) catchment in Canada and the Schlehdorf (SLD) catchment in Germany.

HESSD
8, 10895–10933, 2011

The Au Saumon catchment (Fig. 1) drains 738 km^2 of land, which altitude ranges between 277 and 1092 m, for a mean annual air temperature of 4.5 °C. Its mean annual precipitations reaches 1284 mm (1975–2003), of which 355 mm is snow, leading to a mean annual discharge of 771 mm (see Table 1). Its land use mostly consists of mixed coniferous and deciduous forests and some croplands. Geology corresponds to Ordovician, Silurian and Devonian sedimentary rocks resulting in limestone, sandstone and shale type of soils (silt-loam soils). The Schlehdorf catchment (Fig. 2) drains 708 km^2. Its altitude ranges from 603 to 2562 m, for a mean annual air temperature of 5.2 °C. Mean annual precipitation attains 1420 mm (1970–2000), of which 347 mm is snow, for a mean annual discharge of 983 mm. Land use is defined essentially as coniferous and deciduous forests and rocks, while geology is pre-Alps Trias and Jurassic limestone and dolomite (sandy-loam, loam). The two catchments are influenced by snow and are thus possibly impacted by changes in both precipitation and temperature.

Although a larger number of catchments is necessary for drawing general conclusions (see e.g., Andréassian et al., 2006, 2009), we limited our investigations to these two study catchments in order to present results in details.

2.2 Lumped conceptual hydrological models

Twenty lumped conceptual hydrological models were selected in this study. They are all based on commonly available hydrological models, but some were modified so that they can all be employed in a similar framework. The choice of these models is mainly based on known performance and structural diversity, i.e. 4 to 10 free parameters, and 2 to 7 storage units.

They all correspond to various conceptualizations of the rainfall-runoff modelling process applied in a lumped mode. They all are designed to take into account soil moisture, a range of contributions to total flow, depending on stores, interconnections, and routing. The soil moisture accounting procedure has various formulations (linear and non-linear, with one or several layers) and the routing components include linear and

HESSD
8, 10895–10933, 2011

non-linear formulations, various unit hydrographs or simple time delays. Most of these model versions originate from the work by Perrin et al. (2001) and Mathevet (2005), and were used by Velázquez et al. (2010).

Table 2 and Fig. 3 illustrate the characteristics and structural diversity of the selected models. Because the aim of this study is not identifying the best model, they will be named M_{01} to M_{20} from here on. A majority of models have 6 or 7 free parameters. Almost all models have soil storage except M_{18} which compensates with a more detailed and specific surface storage. Only two structures, M_{01} and M_{05}, do not include a slow routing storage (often considered as the groundwater storage). They compensate with overland flow routing storage and unit-hydrograph-based routing. Only M_{12} exploits an interflow (delayed) routing storage.

All models were applied in exactly the same conditions: they were run at the a daily time step and fed with identical inputs of areal catchment precipitation and potential evapotranspiration estimated by the McGuinness formulation (McGuinness and Bordne, 1972). Oudin et al. (2005) showed that, on four of the models used here and a set of 308 catchments, this latter formulation exploiting extraterrestrial radiation and mean daily temperature is as efficient as more complex evapotranspiration formulations, for rainfall-runoff modelling objectives.

Snow accumulation and melt are simulated with the CemaNeige snow accounting module (Valéry, 2010). This two-parameter module is based on a degree-day approach. CemaNeige includes an altitudinal distribution into five zones of equal areas. Available temperature and precipitation data are extrapolated over the catchment using altitudinal gradients, which provides inputs for each zone (Valéry et al., 2010). The distinction between liquid and solid precipitations then relies on the air temperature at each altitudinal zone. Two internal states of the snowpack for each zone are also defined: the thermal state of the snowpack and the melting potential. The development of CemaNeige was based on 380 catchments from France, Switzerland, Sweden and Canada, showing various levels of snow influence on flows.

HESSD
8, 10895–10933, 2011

One main advantage of using this snow accounting module here lays in its parsimony (only two free parameters) that does not add undue extra complexity to the hydrological models. Investigating the sensitivity of hydrological simulations to snow modelling is out of the scope of this article, but remains an obvious source of uncertainty in the modelling process.

To evaluate the usefulness of the multimodel approach, the models were combined in a deterministic way: the output of the multimodel was calculated as the average of the outputs of individual models (e.g., Shamseldin et al., 1997). As discussed later in Sect. 3.2, all possible model combinations were tested to try to identify the best performing ones.

2.3 Differential split sample testing

As highlighted in the introduction, in a climate change context, the transposability in time of hydrological models should be assessed and used as a criterion for the selection of appropriate projection tools. The common hypothesis that the models are forced with stationary time series does not apply to a changing climate. So, there is no guarantee that the parameters optimized for the actual time series will still be appropriate for future conditions. This is why hydrological tests on much contrasted climatic conditions are sought here, following the Differential Split Sample Test (DSST) concept detailed by Klemeš (1986). The idea is to calibrate the model on a time series with selected characteristics (e.g. wet and cold) and to validate it on a contrasted time series (e.g. dry and warm), placing the model in a demanding situation in order to evaluate its transposability.

We propose the three-step testing procedure below:

– Select five non-continuous hydrologic years (1 October to 30 September) for four contrasted climate conditions: dry/warm (DW), dry/cold (DC), wet/warm (WW), and wet/cold (WC), based on annual precipitation and temperature – see illustration in Fig. 4 for the Au Saumon catchment (SAU). The selection maximizes

HESSD
8, 10895–10933, 2011

the distance between the yearly average and the median value of the time series, both in terms of precipitation and temperature, which are believed to have the largest impact on streamflow – mean yearly values are important in a water resources perspective. We acknowledge that other precipitation and temperature characteristics, such as the yearly maximum daily values, could have been considered, but were found more appropriate for studies focusing on flood or low-flow events.

- Calibrate and validate on contrasted time series: DW → WC (calibration on DW and validation on WC), WC → DW, DC → WW, WW → DC. This corresponds to test configurations along the diagonals in Fig. 4. Contrasts between calibration and validation, both in terms of precipitation and temperature, should produce the most differentiated flow responses.

- Compare models performance and rank, obtained in the various configurations: DW → WC, WC → DW, DC → WW, WW → DC.

The choice of non-continuous periods provides more contrasted conditions than continuous periods. Obviously, we kept the continuous logic of the tested models by running the models on the entire time series, from the first to the last selected year (in calibration and validation), but only the selected years were next considered for computing the efficiency criteria. The selection of non-continuous periods is valid here because the studied catchments do not show flow components with strong interannual characteristics. Table 1 presents the main characteristics of the selected periods for each catchment.

2.4 Model calibration and performance criteria

The Shuffled Complex Evolution (SCE) (Duan and Gupta, 1992; Duan et al., 1994) automatic optimization algorithm is used for model parameter calibration.

HESSD
8, 10895–10933, 2011

The objective function is the Root Mean Square Error applied to the root-squared transformed streamflow ($RMSE_{sqrt}$):

$$RMSE_{sqrt} = \sqrt{\frac{\sum_{i=1}^{N}\left(\sqrt{Q_{sim,i}} - \sqrt{Q_{obs,i}}\right)^2}{N}} \quad (1)$$

where $Q_{obs,i}$ and $Q_{sim,i}$ are the observed and simulated streamflows at time step i, and N is the total number of observations. $RMSE_{sqrt}$ yields a more multi-purpose criterion than the standard RMSE (on non-transformed discharge), which emphasises the large errors that generally occur during flood events (Chiew and McMahon, 1994; Oudin et al., 2006a, b).

Two other criteria were used for evaluation. The first one is the Nash-Sutcliffe Efficiency criterion (Nash and Sutcliffe, 1970), calculated on root-squared transformed streamflows for the same reason:

$$NSE_{sqrt} = 1 - \frac{\sum_{i=1}^{N}\left(\sqrt{Q_{sim,i}} - \sqrt{Q_{obs,i}}\right)^2}{\sum_{i=1}^{N}\left(\sqrt{Q_{obs,i}} - \overline{\sqrt{Q_{obs}}}\right)^2} \quad (2)$$

in which $\overline{\sqrt{Q_{obs}}}$ is the mean of observed square root transformed flows on the test period. NSE_{sqrt} values range from negative infinity to 1, a value of 1 indicating a perfect model simulation. NSE_{sqrt} provides information on the overall agreement between observed and simulated discharge.

The second criterion is the absolute percent bias (PB) (Moriasi et al., 2007) and corresponds to the calculation of total volume differences between observed and simulated discharge:

$$PB = \frac{\left|\sum_{i=1}^{N}\left(Q_{sim,i} - Q_{obs,i}\right)\right|}{\sum_{i=1}^{N} Q_{obs,i}} \times 100 \quad (3)$$

HESSD
8, 10895–10933, 2011

Additionally to the performance and transposability calculations, the collective diversity of the models is of concern to the multimodel approach. The search for diversity, which is a reflection of the models structural variety, is needed to take into account the range of the hydrologic response in the multimodel approach. This diversity is assessed with the mean coefficient of variation (CV) calculated on the simulated discharges (Brochero et al., 2011):

$$\mathrm{CV} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\sigma_i}{\mu_i}\right) \quad (4)$$

with $\sigma_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(Q_{\mathrm{sim},i,m} - \overline{Q_{\mathrm{sim},i}}\right)}$ and $\mu_i = \frac{1}{M}\sum_{m=1}^{M} Q_{\mathrm{sim},i,m}$, where m is the model, and M is the total number of models.

3 Results and discussion

3.1 Individual performance of each model

The appraisal of the individual worth of the models is based on a performance and rank analysis in validation, for all Differential Split Sample Tests (i.e. DW → WC, WC → DW, DC → WW, and WW → DC). The $\mathrm{NSE}_{\mathrm{sqrt}}$ and PB results, for every models and tests on the Au Saumon time series, are compiled in Table 3 and illustrated in Fig. 5 (for ranks and $\mathrm{NSE}_{\mathrm{sqrt}}$), while results for the Schlehdorf catchment are shown in Table 3 and Fig. 6. In each case, the four DSSTs are identified by a specific color and shape; while the grey bars stress the rank of performance range for each hydrological model, and the black horizontal lines the mean individual rank. In other words, a longer grey bar reflects lower robustness and high black line (close to one) a better mean rank of performance, both components of the climate transposability.

For the analysis, it must be kept in mind that comparison of performances between DSST may be biased by the selection of the $\mathrm{NSE}_{\mathrm{sqrt}}$ criterion, because the variance

HESSD
8, 10895–10933, 2011

used as the denominator is different for each selected period (Martinec and Rango, 1989). To circumvent this possible bias, the analysis will primarily be based on the NSE_{sqrt} performance ranks.

For the Au Saumon catchment (Fig. 5), calibration on dry/warm years and validation on wet/cold years (DW → WC) yields the best performance on average. But the level of performance differs between models. M_{19} behaves badly in the WW → DC test ($NSE_{sqrt} = 0.57$) and M_{12} in the WC → DW test ($NSE_{sqrt} = 0.65$), while performance is generally inferior for M_{08} (from 0.72 to 0.60). It is also noteworthy that some models, such as M_{01}, may not provide the best mean rank but show better robustness than other models, with identical rank in the four DSSTs (i.e. seventh in all cases). Similar statements could be made for M_{15}, M_{07} and M_{08}, for example. It can also be shown that some models perform badly for validation on dry years and well for validation on wet years (e.g. M_{20}), or the exact opposite (e.g. M_{03}). Best mean rank models are M_{09}, M_{05} and M_{04}. On the other way, M_{08}, M_{12} and M_{13} show a poor performance with mean rank varying respectively from 18.75 to 15.75. Results in terms of water balance seem quite sensitive to the type of test, as shown by PB values (Table 3 and Fig. 7). Several models tend to under-evaluate water volumes. This is expected for the tests with calibration on wet years and validation on dry years but it sometimes also occurs for the opposite situation. The DW → WC (PB values from 2.92 to 12.17 %) and DC → WW (from 0.43 to 15.46 %) tests yield the best general results. In the two other cases, PB values are worse (from 9.17 to 32.29 % for WC → DW; from 9.72 to 28.92 % for WW → DC). This statement is linked to the under-evaluation of water volume, more penalising for these two tests as illustrated in Fig. 7.

Results for the Schlehdorf catchment (Fig. 6) highlight different models than for the Au Saumon catchment. For instance, M_{09}, M_{14}, and M_{18} show low robustness, while M_{02} and M_{15} are robust in that respect but offer low performance (mean ranks respectively 15.25 and 17.25). M_{03}, M_{04} and M_{06} give good climate transposability with a mean performance rank from 2.5 to 6 and also with a good robustness. In general, performance is more contrasted from one DSST to the other and from one model to the

HESSD
8, 10895–10933, 2011

other, than for the Au Saumon catchment. Overall, M_{03}, M_{04}, M_{05} and M_{06} are the most appealing models, both in terms of robustness and rank of performance. As for the Au Saumon, PB performance (Table 3) shows contrasted results. It can be noted that M_{09} is probably the worst model with PB exceeding 30 % for three of the DSSTs. As illustrated in Fig. 7, statements concerning water balance for the Schlehdorf catchment are closer to what could be expected. Most models have a tendency to overestimate water balance for tests with calibration on dry years and validation on wet years while they under-estimate water quantities for the opposite situation. The range of performance for water balance is however larger for this catchment.

These results illustrate the difficulty in identifying a single lumped model that could behave well in terms of performance and robustness, when tested under contrasted geology, topography, and climatology of the Province of Québec, Canada, and the State of Bavaria, Germany. This remains one of the main challenges of hydrological projection studies under climate change. Nevertheless, this methodology allows identifying best-compromise individual models for each catchment based on results illustrated in Figs. 5 and 6. For Au Saumon catchment, models M_{05} and M_{09} are the best-compromise, whereas for Schlehdorf M_{03}, M_{04} and M_{06} can be underlined.

3.2 Collective performance

Multimodel combination (ensemble) is often recognized as a promising mean for improving performance beyond the best single model. A deterministic multimodel ensemble analysis, taking the average of simulated streamflow series as output, is next performed here. We explored all possible models combinations (2^{20} possibilities i.e. 1 048 576 combinations) and calculated performance (NSE_{sqrt}) and diversity criteria (CV). Consideration of CV aims at measuring the level of diversity of each ensemble, a reflection of the hydrological range of responses (i.e. structural variability).

Results for the Au Saumon and Schlehdorf catchments are illustrated in Figs. 8 and 9, respectively. The red lines and circle represent the performance and the diversity of the twenty-member ensembles, while the blue vertical line is the

HESSD
8, 10895–10933, 2011

performance of the best individual model. Table 3 and Fig. 7 also illustrate the multimodel results.

The twenty-member ensemble gives better results than the best individual model for all DSSTs on the Au Saumon catchment (Fig. 8): a NSE_{sqrt} of 0.86 for DW → WC, 0.81 for WC → DW, 0.80 for DC → WW and 0.79 for WW → DC. This is not true, for three of the Schlehdorf DSSTs. However, the multimodel approach is still valuable since the best model is different for each DSST, a sign of a lack of climate transposability (Table 3): M_{04} is the best single model in WC → DW (NSE_{sqrt} of 0.81), M_{05} in DC → WW (0.83), and M_{03} in WW → DC (0.86). In each case, no other single model surpasses the twenty-model performance.

Figures 8 and 9 also show that for DW → WC and WC → DW on the Au Saumon catchment, the search for diversity is compatible with the best performance (NSE_{sqrt}), while the opposite is true for the other two tests. For Schlehdorf catchment, diversity seems indifferent to performance for WC→DW and WW → DC, while there is no real tendency in the other two situations.

Concerning water balance, Fig. 7 also draws the multimodel cumulative error between observed and simulated discharge. Ensembles (mean simulation) reduce variance and synthesize the structural model variability. For cases where water balance is over and under-estimated on the same test, the ensemble approach is the most efficient (e.g. DW → WC for Schlehdorf catchment).

Results also reveal many other model combinations (sub-selections) that provide better performance and diversity than the twenty-member ensemble. They are located in the upper right portion of the DSST plots in Figs. 8 and 9. Figure 10 proposes a more detailed exploration of them for Au Saumon catchment. They include 7.3 % of the possible combinations of the DW → WC test, 18.5 % of the WC → DW test, 10.2 % of the DC → WW test, and 9.2 % of the WW → DC test. The same holds for the Schlehdorf catchment (not shown here), for which they encompass 15.1 % of the possible combinations of the DW → WC test, 13.8 % of WC → DW test, 16.2 % of DC → WW test, of 10.9 % for WW → DC test.

HESSD
8, 10895–10933, 2011

Because one needs to work on performance, diversity and robustness, combinations accurate for all four DSSTs are sought, separately for both catchments. We identified model combinations that not only lead to better performance and diversity than the twenty-member ensemble, but that also provide enhanced robustness relative to the DSST, a feature that is deemed important in a climate change context. Figure 11 illustrates such selected ensembles for the Au Saumon catchment. They represent only 0.94 % of the possible combinations (9896 ensembles). For the Schlehdorf catchment, only 181 ensembles were identified (0.02 %). With these efficient, diversified and robust ensembles, we can evaluate the collective interest of each model, in other words, the added-value of the structure for an ensemble approach in a climate change context for each catchment. Moreover, we can emphasize the better performance offered by smaller combinations (e.g. 5 or 6 members), as also depicted in Table 3.

To evaluate the benefit of the above selected model ensembles, they were confronted to the individual models and to the twenty-model ensemble. Figure 12 illustrates this comparison for both catchments, where the boxplots give performance range of the ensembles, black diamonds, the twenty-model ensembles performance (by definition it is the minimal range of the selected ensembles), and the coloured circles and squares, the individual performance. Results show that only the multimodel offers good performance and robustness. In short, the twenty-model ensemble is a good option for contrasted conditions, but a well-chosen sub-selection has a potential for increased performance.

As a final analysis, Fig. 13 illustrates the ranking of the individual models, in terms of occurrence count in the selected ensembles and the mean individual rank, for the Au Saumon and Schlehdorf catchments. Note that all models participate to the ensembles, but not in a uniform way. For the Au Saumon catchment, M_{05} is the most frequently selected model with 9811 appearances in 9896 combinations, whereas M_{08} is used only 1034 times. Interestingly, M_{05} is one of the best models in terms of climate transposability, based on the DSSTs, while M_{08} is the worst ones (see Fig. 5). On the other hand, M_{07} and M_{15}, which have shown great robustness and correct

HESSD
8, 10895–10933, 2011

performance, are also not frequently used. This is the same for the best-compromise model M_{09}. It is clear that, comparing selection counts and mean individual rank, no link can be identified.

The same analysis differs considerably in the case of the Schlehdorf catchment. M_{02} and M_{08} are present in all 181 combinations, and M_{10} is almost absent (4 selections). Interestingly, M_{02} showed a poor range of performance but high robustness, while M_{08} performance and robustness were close to average (a situation even less favourable for M_{09}, the third most frequently selected model). As for Au Saumon catchment, no link can be highlighted between selection counts and mean individual rank.

The DSST collective evaluation of the models stresses one more time the interest of ensembles over the use of a single model, especially in terms of climate transposability, which is of paramount importance for climate change applications, but also in terms of catchment transposability, since only the twenty-model ensemble provides an interesting modelling option for both catchments. Then, if one wants to increase further the performance, it has also been shown that many pertinent ensembles exist (i.e. sub-selections) but need specific and detailed analysis unlike the simple use of the twenty member ensemble.

4 Conclusions

Evaluating hydrological model behaviour under contrasted conditions for calibration and validation is, in our opinion, a pre-requisite to climate change applications. The aim of this study was to assess the relevance of twenty lumped conceptual hydrological models in a climate change context, based on Differential Split Sample Tests. Two case studies were used: the Au Saumon and Schlehdorf catchments (natural), located in the Province of Québec (Canada), and the State of Bavaria (Germany), respectively. This approach allowed climate transposability evaluation of all twenty individual models, along with their collective qualities.

HESSD
8, 10895–10933, 2011

The analysis of the individual value of each lumped model was carried out by looking at their performance in simulating streamflows under contrasted validation and calibration conditions, to highlight their relevance for climate impact studies. 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. Furthermore, this is the only approach that was successful for both catchments, indicating a strong potential for catchment transposability (a point that would need to be tested further on many other catchments). In some cases, individual models surpassed the twenty-model ensemble in performance, but the fact that no individual model achieved this under more than one contrasted forcing (out of four) only stresses further the higher climate transposability of the ensemble.

Pushing further the ensemble philosophy, all possible model combinations (2^{20} possibilities) have been explored. Many combinations were found to provide increased performance and diversity over the twenty-member ensemble, leaving an operational hydrologist with the option of fine tuning ensembles for each specific catchment (at the potential expanse of spatial transposability) or of exploiting the more general twenty-ensemble. Of course, the twenty-ensemble gathered here may not be the only general option under contrasted forcing (such as climate change), but it seems that a large number of models have better chance to be appropriate for many catchments. It is also noteworthy that even worse-performing individual models were successfully contributing to an ensemble, reinforcing 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). Model diversity was thus confirmed as a sought quality of hydrological ensembles, especially under contrasted forcing.

HESSD
8, 10895–10933, 2011

Acknowledgements. The authors acknowledge NSERC, Ouranos, and Hydro-Québec for financial support, as well as the other partners in the QBIC[3] project.

References

Ajami, N. K., Duan, Q., Gao, X., and Sorooshian, S.: Multimodel combination techniques for analysis of hydrological simulations: application to distributed model intercomparison project results, J. Hydrometeorol., 7(4), 755–768, 2006.

Andréassian, V., Perrin, C., Michel, C., Usartsanchez, I., and Lavabre, J.: Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models, J. Hydrol., 250(1–4), 206–223, 2001.

Andréassian, V., Perrin, C., and Michel, C.: Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models, J. Hydrol., 286(1–4), 19–35, 2004.

Andréassian, V., Hall, A., Chahinian, N., and Schaake, J.: Introduction and synthesis: why should hydrologists work on a large number of basin data sets?, IAHS Publication, 307, 1–5, 2006.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., and Valéry, A.: *HESS Opinions* "Crash tests for a standardized evaluation of hydrological models", Hydrol. Earth Syst. Sci., 13, 1757–1764, doi:10.5194/hess-13-1757-2009, 2009.

Bae, D. H., Jung, I. W., and Lettenmaier, D. P.: Hydrologic uncertainties in climate change from IPCC AR4 GCM simulations of the Chungju Basin, Korean J. Hydrol., 401(1–2), 90–105, 2011.

Bates, B., Kundzewicz, Z. W., Wu, S., and Palutikof, J.: Le changement climatique et l'eau – Rapport du Groupe d'Experts Intergouvernemental sur l'Évolution du Climat, IPCC Secretariat, Geneva, 237 pp., 2008.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrol. Earth Syst. Sci., 4, 203–213, doi:10.5194/hess-4-203-2000, 2000.

Beven, K. J. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249(1–4), 11–29, 2001.

HESSD

8, 10895–10933, 2011

Boé, J., Terray, L., Martin, E., and Habets, F.: Projected changes in components of the hydrological cycle in French river basins during the 21st century, Water Resour. Res., 45(8), 1–15, 2009.

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A. J., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N. R.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Adv. Water Resour., 32(2), 129–146, 2009.

Brochero, D., Anctil, F., and Gagné, C.: Simplifying a hydrological ensemble prediction system with a backward greedy selection of members – Part 1: Optimization criteria, Hydrol. Earth Syst. Sci., 15, 3307–3325, doi:10.5194/hess-15-3307-2011, 2011.

Butts, M., Payne, J., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, J. Hydrol., 298(1–4), 242–266, 2004.

Chiew, F. and McMahon, T.: Application of the daily rainfall-runoff model MODHYDROLOG to 28 Australian catchments, J. Hydrol., 153(1–4), 383–416, 1994.

Chiew, F., Stewardson, M., and McMahon, T.: Comparison of six rainfall-runoff modelling approaches, J. Hydrol., 147(1–4), 1–36, 1993.

Donnelly-Makowecki, L. and Moore, R.: Hierarchical testing of three rainfall-runoff models in small forested catchments, J. Hydrol., 219(3–4), 136–152, 1999.

Duan, Q. and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour., 28(4), 1015–1031, 1992.

Duan, Q., Ajami, N., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Adv. Water Resour., 30(5), 1371–1386, 2007.

Duan, Q., Sorooshian, S., and Gupta, V.: Optimal use of the SCE-UA global optimization method for calibrating watershed models, J. Hydrol., 158, 265–284, 1994.

Georgakakos, K., Seo, D., Gupta, H., Schaake, J., and Butts, M.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298(1–4), 222–241, 2004.

Görgen, K., Beersma, J., Brahmer, G., Buiteveld, H., Carambia, M., de Keizer, O., Krahe, P., Nilson, E., Lammersen, R., Perrin, C., and Volken, D.: Assessment of climate change impacts on discharge in the Rhine river basin?, results of the RheinBlick2050 project, International Commission for the Hydrology of the Rhine Basin Secretariat, Lelystad, 211 pp., 2010.

HESSD

8, 10895–10933, 2011

Kim, Y.-O., Jeong, D., and Ko, I. H.: Combining rainfall-runoff model outputs for improving ensemble streamflow prediction, J. Hydrol. Eng., 11(6), 578–588, 2006.
Klemeš, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J., 31(1), 13–24, 1986.
Loumagne, C., Vidal, J., Feliu, C., Torterotot, J., and Roche, P.: Procédure de décision multi-modèle pour une prévision des crues en temps réel, application au bassin supérieur de la Garonne, Rev. Sci. Eau, 8(4), 539–561, 1995.
Ludwig, R., May, I., Turcotte, R., Vescovi, L., Braun, M., Cyr, J.-F., Fortin, L.-G., Chaumont, D., Biner, S., Chartier, I., Caya, D., and Mauser, W.: The role of hydrological model complexity and uncertainty in climate change impact assessment, Adv. Geosci., 21, 63–71, doi:10.5194/adgeo-21-63-2009, 2009.
Martinec, J. and Rango, A.: Merits of statistical criteria for the performance of hydrological models, Wiley Online Library, Water Resour. Bull., 25(2), 421–432, 1989.
Mathevet, T.: Quels modèles pluie-débit globaux au pas de temps horaire?, École Nationale du Génie Rural, des Eaux et des Forêts, Paris, 2005.
Maurer, E. P.: Uncertainty in hydrologic impacts of climate change in the Sierra Nevada, California, under two emissions scenarios, Climatic Change, 82(3–4), 309–325, 2007.
McGuinness, J. L. and Bordne, E. F.: A comparison of lysimeter-derived potential evapotranspiration with computed values, Search, Alderman, 71 pp., 1972.
Merritt, W., Alila, Y., Barton, M., Taylor, B., Cohen, S., and Neilsen, D.: Hydrologic response to scenarios of climate change in sub watersheds of the Okanagan Basin, British Columbia, J. Hydrol., 326(1–4), 79–108, 2006.
Merz, R., Parajka, J., and Blöschl, G.: Time stability of catchment model parameters: implications for climate impact analyses, Water Resour. Res., 47(2), 1–17, 2011.
Minville, M., Brissette, F., and Leconte, R.: Uncertainty of the impact of climate change on the hydrology of a nordic watershed, J. Hydrol., 358(1–2), 70–83, 2008.
Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., and Veith, T. L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations, Am. Soc. Agr. Biol. Eng., 50(3), 885–900, 2007.
Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models, Part I – A discussion of principles, J. Hydrol., 10, 282–290, 1970.
Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andreassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2 – To-

HESSD
8, 10895–10933, 2011

wards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, J. Hydrol., 303(1–4), 290–306, 2005.
Oudin, L., Andréassian, V., Mathevet, T., Perrin, C., and Michel, C.: Dynamic averaging of rainfall-runoff model simulations from complementary model parameterizations, Water Resour. Res., 42(7), 1–10, 2006a.
Oudin, L., Perrin, C., Mathevet, T., Andreassian, V., and Michel, C.: Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, J. Hydrol., 320(1–2), 62–83, 2006b.
Perrin, C., Michel, C., and Andreassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, J. Hydrol., 242(3–4), 275–301, 2001.
Perrin, C., Oudin, L., and Andreassian, V.: Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, Hydrolog. Sci. J., 52(1), 131–151, 2007.
Prudhomme, C., Jakob, D., and Svensson, C.: Uncertainty and climate change impact on the flood regime of small UK catchments, J. Hydrol., 277(1–2), 1–23, 2003.
Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D., and Dmip participants: Overall distributed model intercomparison project results, J. Hydrol., 298(1–4), 27–60, 2004.
Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32(7), 2189, 1996.
Refsgaard, J. C., Vandersluijs, J., Brown, J., and Vanderkeur, P.: A framework for dealing with uncertainty due to model structure error, Adv. Water Resour., 29(11), 1586–1597, 2006.
Seibert, J.: Reliability of model predictions outside calibration conditions, Nord. Hydrol., 34(5), 477–492, 2003.
Shamseldin, A. Y., O'Connor, K. M., and Liang, G.: Methods for combining the outputs of different rainfall-runoff models, J. Hydrol., 197(1–4), 203–229, 1997.
Valéry, A.: Modélisation précipitations-débit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants. Thèse de Doctorat, Ecole Doctorale 398 Géosciences et Ressources Naturelles, Agro-Paris Tech, 2010
Valéry, A., Andréassian, V., and Perrin, C.: Regionalization of precipitation and air temperature over high-altitude catchments – learning from outliers, Hydrolog. Sci. J., 55(6), 928–940, 2010.
Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J.-M., Viney, N. R., and Teng, J.: Climate non-stationarity – validity of calibrated rainfall-runoff models for use in climate change studies, J.

HESSD
8, 10895–10933, 2011

Hydrol., 394(3–4), 447–457, 2010.
Velázquez, J. A., Anctil, F., and Perrin, C.: Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, Hydrol. Earth Syst. Sci., 14, 2303–2317, doi:10.5194/hess-14-2303-2010, 2010.
Viney, N. R., Bormann, H., Breuer, L., Bronstert, A., Croke, B. F. W., Frede, H., Gräff, T., Hubrechts, L., Huisman, J. A., Jakeman, A. J., Kite, G .W., Lanini, J., Leavesley, G., Lettenmaier, D. P., Lindström, G., Seibert, J., Sivapalan, M. and Willems, P.: Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions, Adv. Water Resour., 32(2), 147–158, 2009.
Xu, C.-Y.: Operational testing of a water balance model for predicting climate change impacts, Agr. Forest Meteorol., 98–99(1), 295–304, 1999.
Xu, C.-Y., Widén, E., and Halldin, S.: Modelling hydrological consequences of climate change – progress and challenges, Adv. Atmos. Sci., 22(6), 789–797, 2005.

HESSD
8, 10895–10933, 2011

Table 1. Main characteristics of the periods selected for the DSST on the Au Saumon and Schlehdorf catchments (DW: dry/warm; DC: dry/cold; WW: wet/warm; WC: wet/cold).

	Au Saumon					Schlehdorf				
	DW	DC	WW	WC	1975–2003	DW	DC	WW	WC	1970–2000
Average annual total precipitation ($mm\,yr^{-1}$)	1126	1158	1421	1431	1284	1296	1229	1613	1517	1420
Average daily mean temperature (°C)	5.22	3.87	5.28	3.86	4.50	5.94	4.68	5.70	4.78	5.21
Average daily min temperature (°C)	0.11	−1.11	0.06	−1.29	−0.59	2.01	1.06	2.02	1.28	1.55
Average daily max temperature (°C)	10.33	8.85	10.49	9.00	9.58	9.88	8.29	9.38	8.27	8.88
Average annual total discharge ($mm\,yr^{-1}$)	677	765	883	874	771	870	834	1106	1054	983
Daily min discharge ($mm\,d^{-1}$)	0.06	0.09	0.06	0.13	0.06	0.69	0.87	0.82	0.73	0.69
Daily max discharge ($mm\,d^{-1}$)	30.33	30.24	47.44	43.79	47.44	15.24	15.81	32.69	26.77	32.69

HESSD
8, 10895–10933, 2011

Table 2. Main characteristics of the 20 model versions used in the study.

Model acronym	Number of optimized parameters	Number of storages	Derived from
BUCK	6	3	BUCKET (Thorthwaite and Mather, 1955)
CEQU	9	2	CEQUEAU (Girard et al., 1972)
CREC	6	3	CREC (Cormary and Guilbot, 1973)
GARD	6	3	GARDENIA (Thiery, 1982)
GR4J	4	3	GR4J (Perrin et al., 2003)
HBV0	9	3	HBV (Bergström et al., 1973)
HYMO	6	5	HYMOD (Wagener et al., 2001)
IHAC	7	3	IHACRES (Jakeman et al., 1990)
MART	7	4	MARTINE (Mazenc et al., 1984)
MOHY	7	3	MOHYSE (Fortin et al., 2006)
MORD	6	4	MORDOR (Garçon, 1999)
NAM0	10	7	NAM (Nielsen et al., 1973)
PDM0	8	4	PDM (Moore et al., 1981)
SACR	9	5	SACRAMENTO (Burnash et al., 1973)
SIMH	8	4	SIMHYD (Chiew et al., 2002)
SMAR	8	4	SMAR (O'Connell et al., 1981)
TANK	7	4	TANK (Sugarawa, 1979)
TOPM	7	3	TOPMODEL (Beven and Kirkby, 1979)
WAGE	8	3	WAGENINGEN (Warmerdam et al., 1997)
XINA	8	5	XINANJIANG (Zhao et al., 1980)

HESSD
8, 10895–10933, 2011

Table 3. Validation performance (DSST) for individual models and multimodel for the Au Saumon and Schlehdorf catchments.

Criteria	DSST	Best model	Median	Worst model	Multimodel (twenty-members)	Multimodel (best sub-selection)
Au Saumon						
NSE_{sqrt} [–]	DW → WC	0.83 (M_{09})	0.81	0.67 (M_{08})	0.86	0.87 (6 mod)
	WC → DW	0.80 (M_{03})	0.75	0.65 (M_{12})	0.81	0.84 (5 mod)
	DC → WW	0.79 (M_{10})	0.75	0.60 (M_{08})	0.80	0.81 (6 mod)
	WW → DC	0.77 (M_{05})	0.74	0.57 (M_{19})	0.79	0.81 (5 mod)
PB [%]	DW → WC	2.92 (M_{13})	6.94	12.17 (M_{07})	2.16	0.19
	WC → DW	9.17 (M_{10})	15.94	32.29 (M_{12})	15.83	14.68
	DC → WW	0.43 (M_{06})	8.01	15.46 (M_{12})	2.89	4.12
	WW → DC	9.72 (M_{04})	18.19	28.92 (M_{12})	19.00	17.08
Schlehdorf						
NSE_{sqrt} [–]	DW → WC	0.80 (M_{04})	0.71	0.31 (M_{12})	0.83	0.87 (6 mod)
	WC → DW	0.81 (M_{04})	0.66	0.05 (M_{18})	0.79	0.85 (6 mod)
	DC → WW	0.83 (M_{05})	0.73	0.43 (M_{12})	0.81	0.86 (5 mod)
	WW → DC	0.86 (M_{03})	0.74	0.38 (M_{09})	0.85	0.88 (6 mod)
PB [%]	DW → WC	0.02 (M_{01})	4.17	30.11 (M_{09})	1.94	3.34
	WC → DW	0.42 (M_{03})	9.12	32.61 (M_{12})	11.62	4.52
	DC → WW	0.08 (M_{10})	5.04	17.55 (M_{11})	1.50	1.56
	WW → DC	0.17 (M_{02})	7.99	31.41 (M_{09})	10.04	3.85

HESSD
8, 10895–10933, 2011

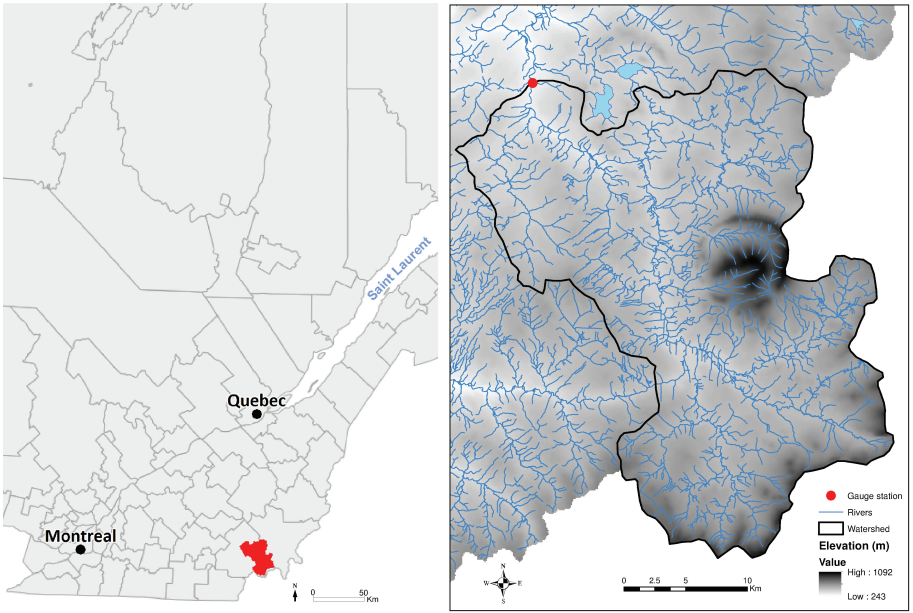

Fig. 1. Location of the Au Saumon catchment (738 km^2; Canada).

HESSD
8, 10895–10933, 2011

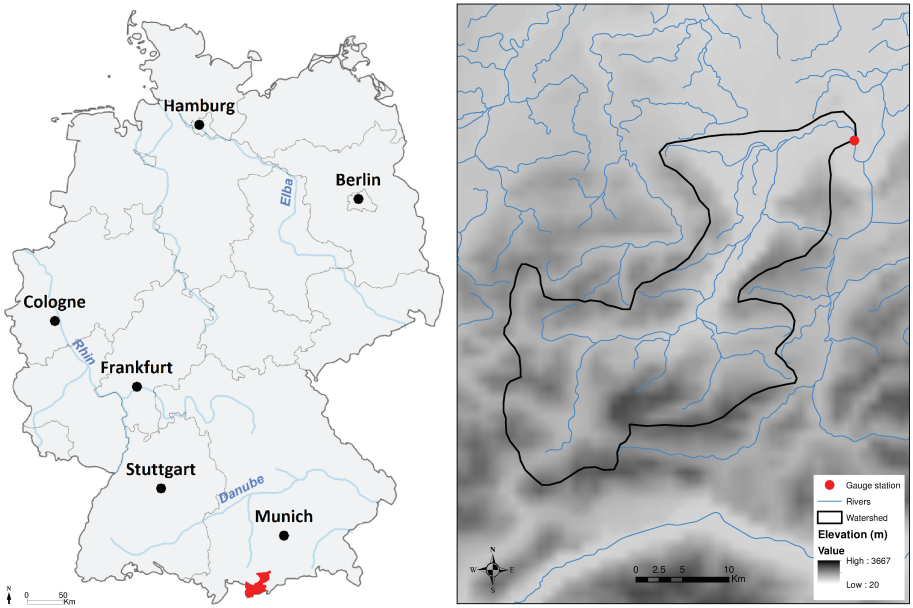

Fig. 2. Location of the Schlehdorf catchment (708 km^2; Germany).

HESSD
8, 10895–10933, 2011

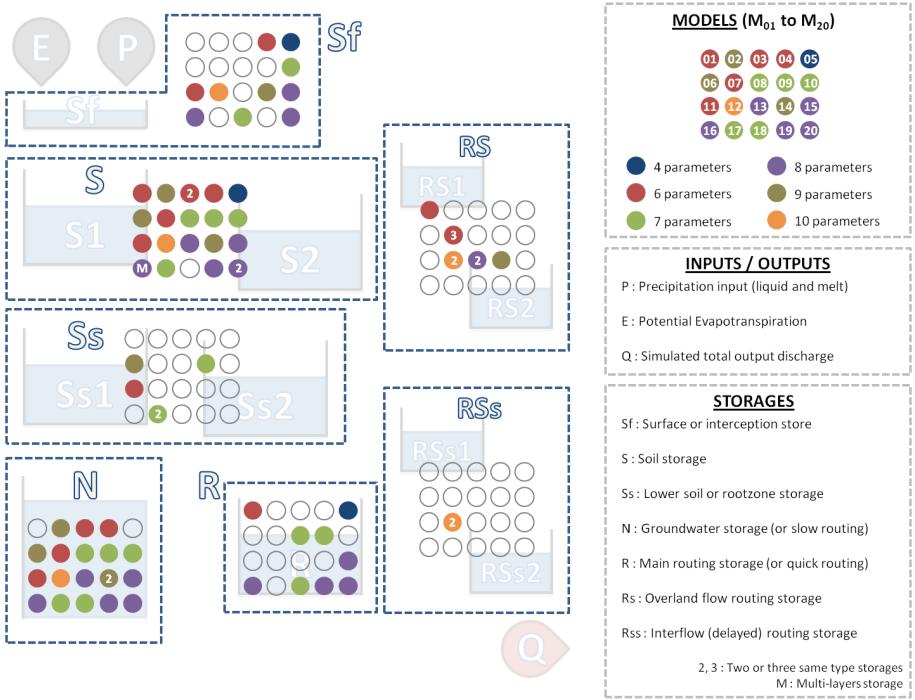

Fig. 3. Illustration of model structural diversity (all models are put in the same frame).

HESSD
8, 10895–10933, 2011

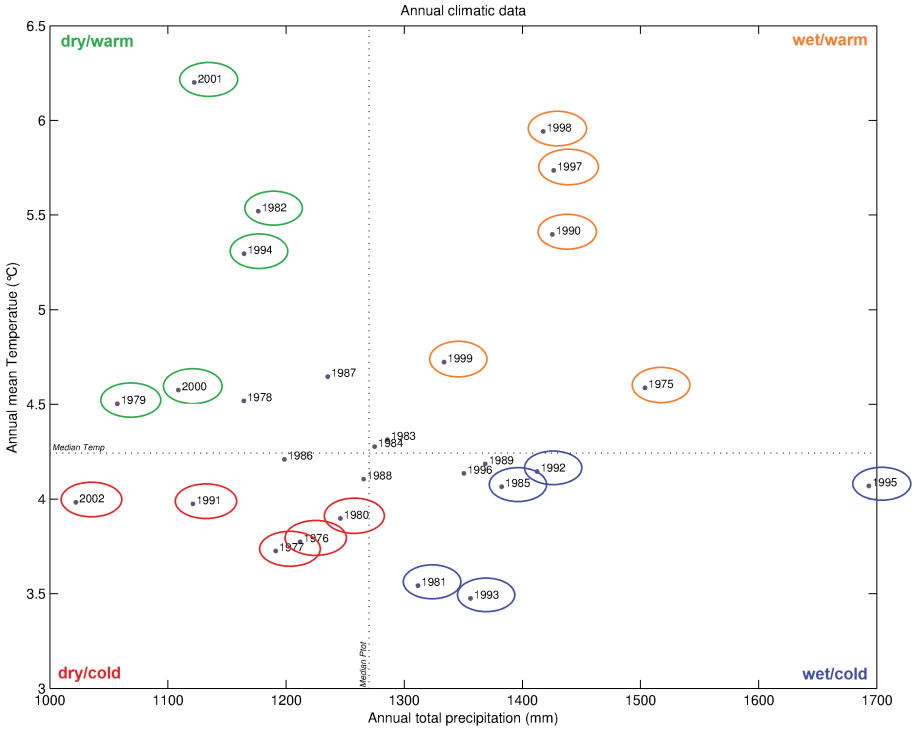

Fig. 4. Time series clustering results for the Au Saumon catchment (SAU).

HESSD
8, 10895–10933, 2011

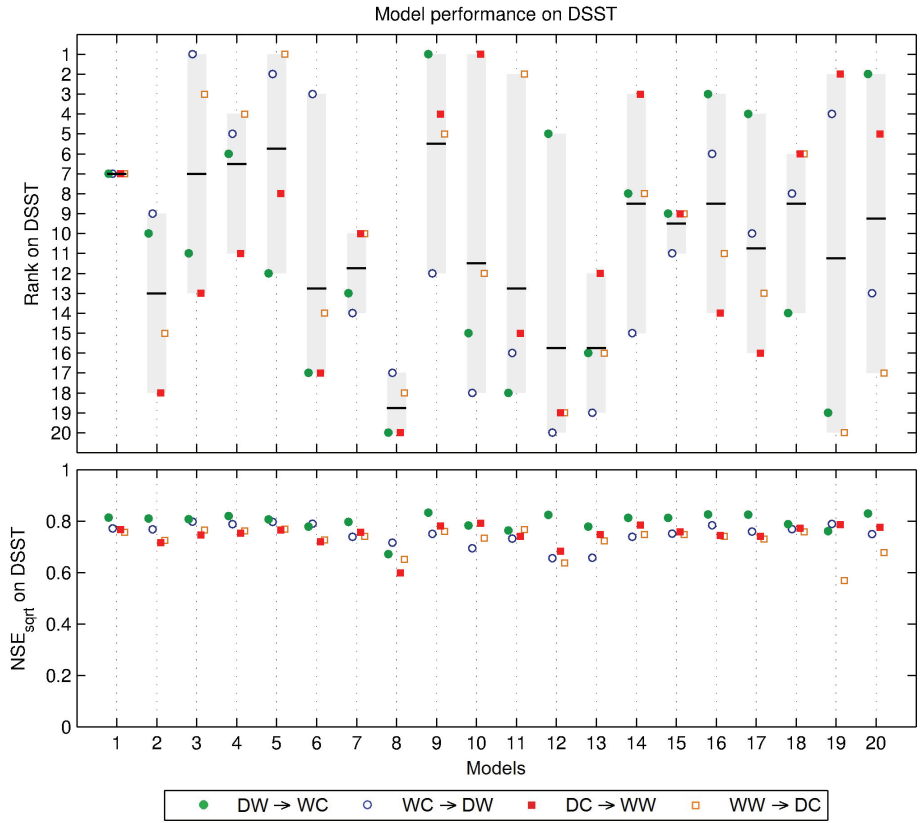

Fig. 5. Validation performance and robustness (DSST) for the Au Saumon catchment (SAU).

HESSD
8, 10895–10933, 2011

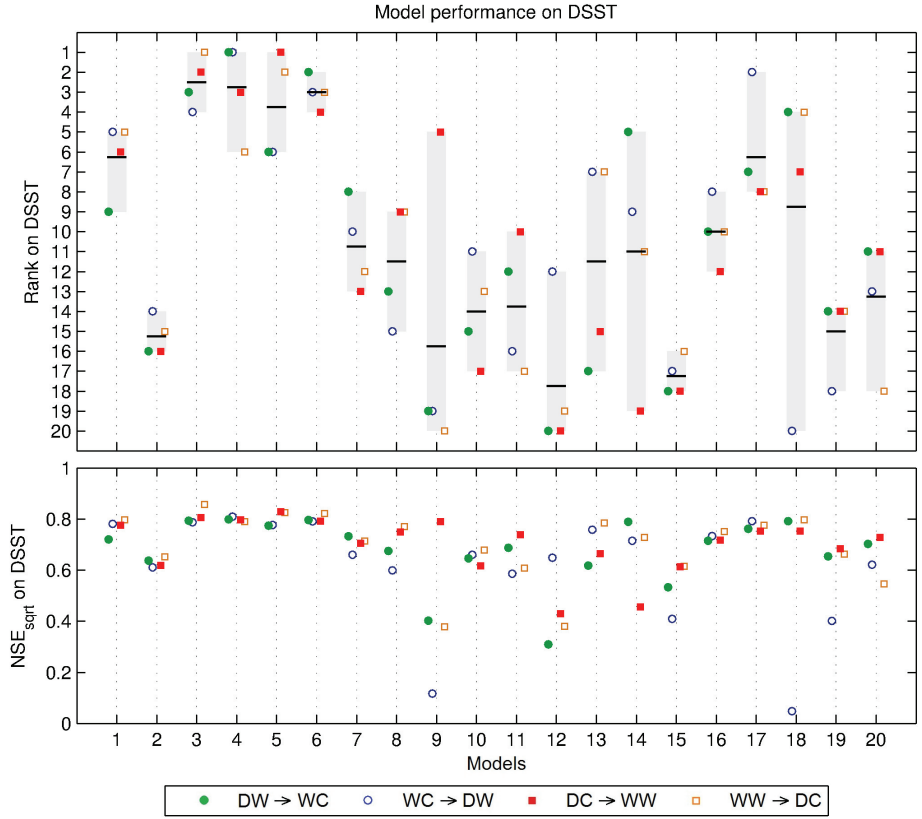

Fig. 6. Validation performance and robustness (DSST) for the Schlehdorf catchment (SLD).

HESSD
8, 10895–10933, 2011

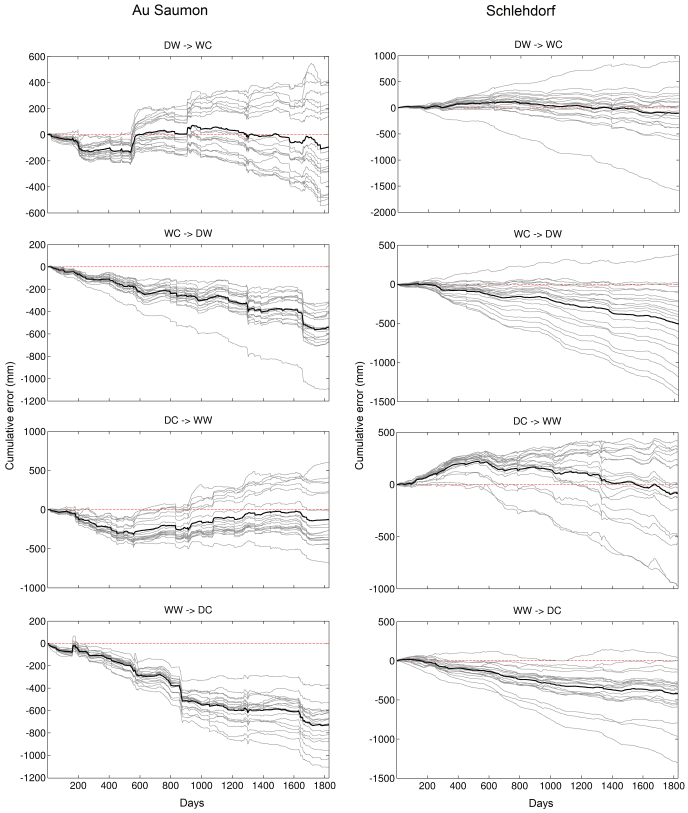

Fig. 7. Cumulative error between observed and simulated discharges for all the DSS tests in validation, for the Au Saumon and Schlehdorf catchments. Grey lines are the twenty individual models, large black line is the twenty-member ensemble and the horizontal dashed line indicates the optimal value.

HESSD
8, 10895–10933, 2011

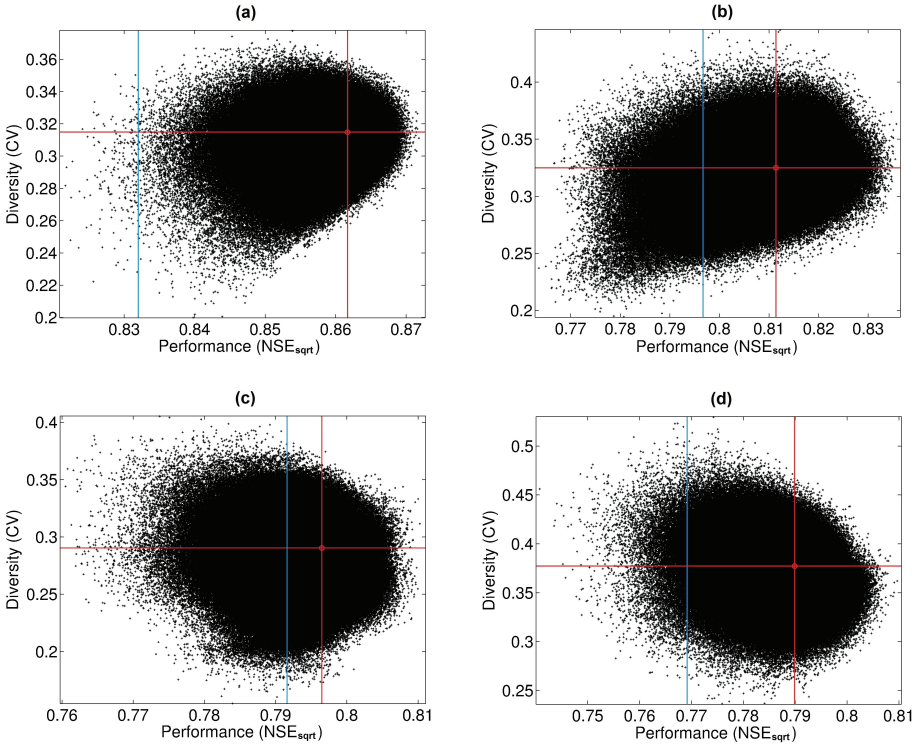

Fig. 8. Validation performance (NSE_{sqrt}) and diversity (CV) for all model combinations (2^{20} points) and Differential Split Sample Tests for the Au Saumon catchment (SAU): **(a)** calibration on DW years (dry/warm) and validation on WC years (wet/cold); **(b)** calibration on WC years (wet/cold) and validation on DW years (dry/warm); **(c)** calibration on DC years (dry/cold) and validation on WW years (wet/warm); **(d)** calibration on WW years (wet/warm) and validation on DC years (dry/cold). Red lines and circle illustrate performance and diversity of the twenty-member ensembles and blue lines, of the best individual model for each test.

HESSD
8, 10895–10933, 2011

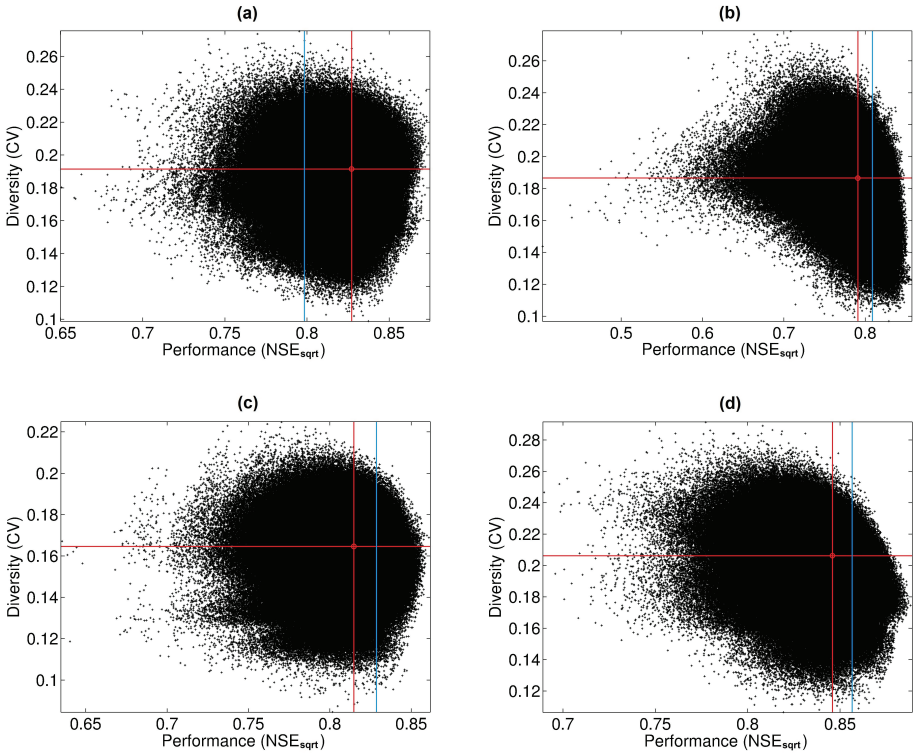

Fig. 9. Validation performance (NSE_{sqrt}) and diversity (CV) for all model combinations (2^{20} points) and Differential Split Sample Tests for the Schlehdorf catchment (SLD): **(a)** calibration on DW years (dry/warm) and validation on WC years (wet/cold); **(b)** calibration on WC years (wet/cold) and validation on DW years (dry/warm); **(c)** calibration on DC years (dry/cold) and validation on WW years (wet/warm); **(d)** calibration on WW years (wet/warm) and validation on DC years (dry/cold). Red lines and circle illustrate performance and diversity of the twenty-member ensembles and blue lines, of the best individual model for each test.

HESSD
8, 10895–10933, 2011

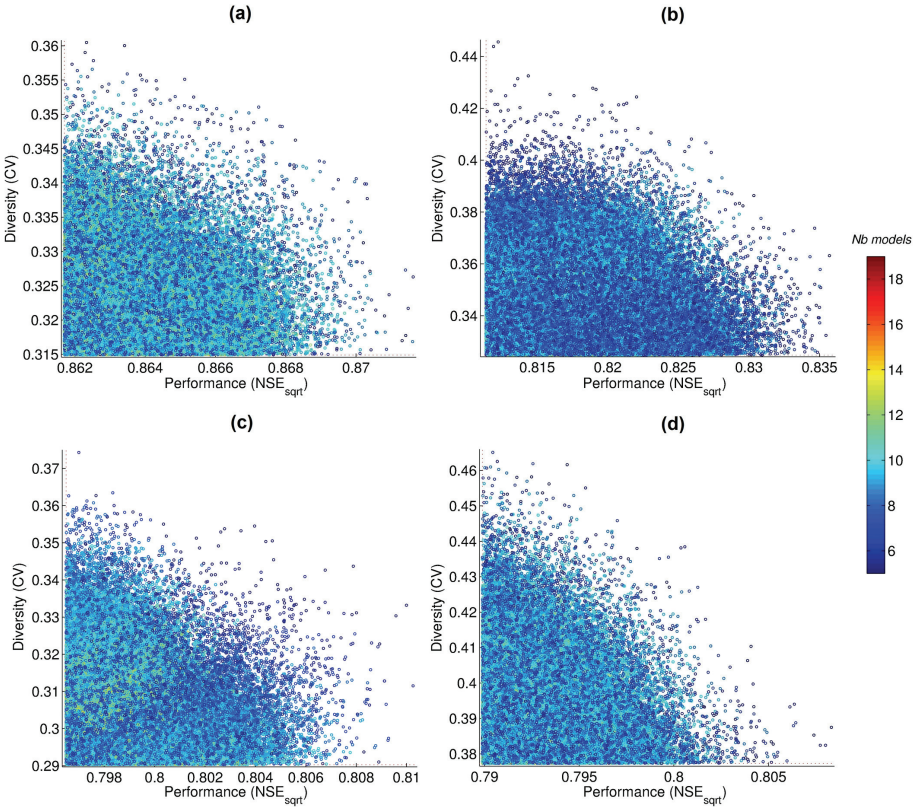

Fig. 10. Validation performance (NSE_{sqrt}) and diversity (CV) for the Au Saumon catchment (SAU) for the model combinations that surpass the twenty-model ensemble in individual Differential Split Sample Test: **(a)** calibration DW (dry/warm) and validation WC (wet/cold); **(b)** calibration WC (wet/cold) and validation DW (dry/warm); **(c)** calibration DC (dry/cold) and validation WW (wet/warm); **(d)** calibration WW (wet/warm) and validation DC (dry/cold).

HESSD
8, 10895–10933, 2011

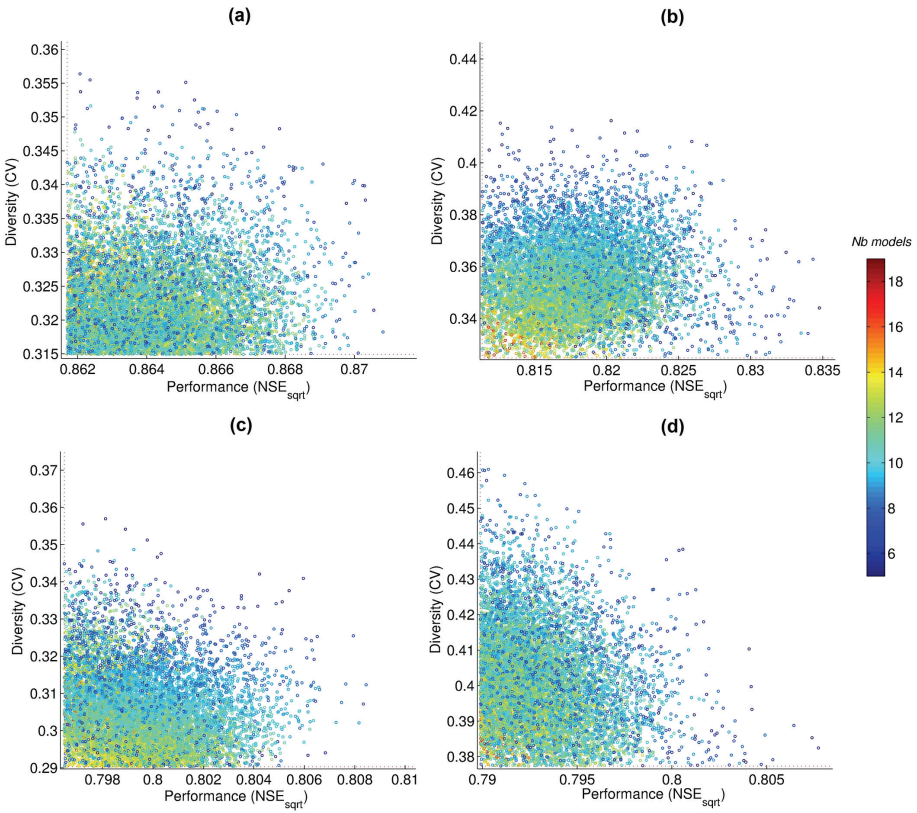

Fig. 11. Validation performance (NSE_{sqrt}) and diversity (CV) for the Au Saumon catchment (SAU) for the model combinations that surpass the twenty-model ensemble in all four Differential Split Sample Tests: **(a)** calibration DW (dry/warm) and validation WC (wet/cold); **(b)** calibration WC (wet/cold) and validation DW (dry/warm); **(c)** calibration DC (dry/cold) and validation WW (wet/warm); **(d)** calibration WW (wet/warm) and validation DC (dry/cold).

HESSD
8, 10895–10933, 2011

Multimodel evaluation under contrasted conditions

G. Seiller et al.

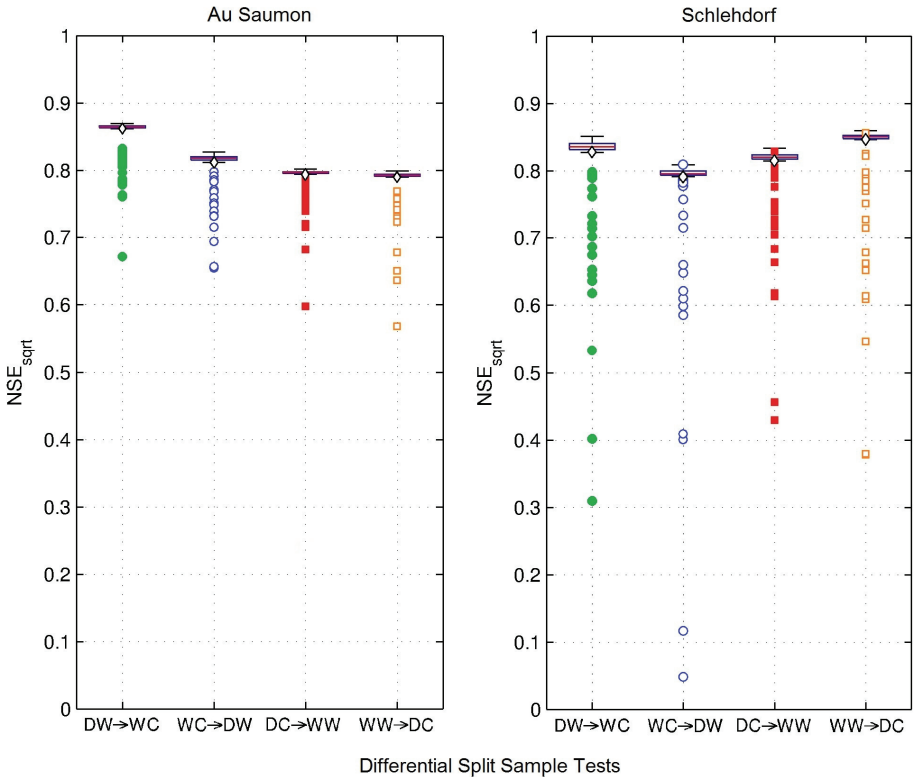

Fig. 12. Individual and multimodel DSST validation performance. Boxplots depict the range of the multimodel combinations, diamonds represent the twenty-model ensemble, and the circles and squares, the individual models, for the Au Saumon and Schlehdorf catchments.

HESSD
8, 10895–10933, 2011

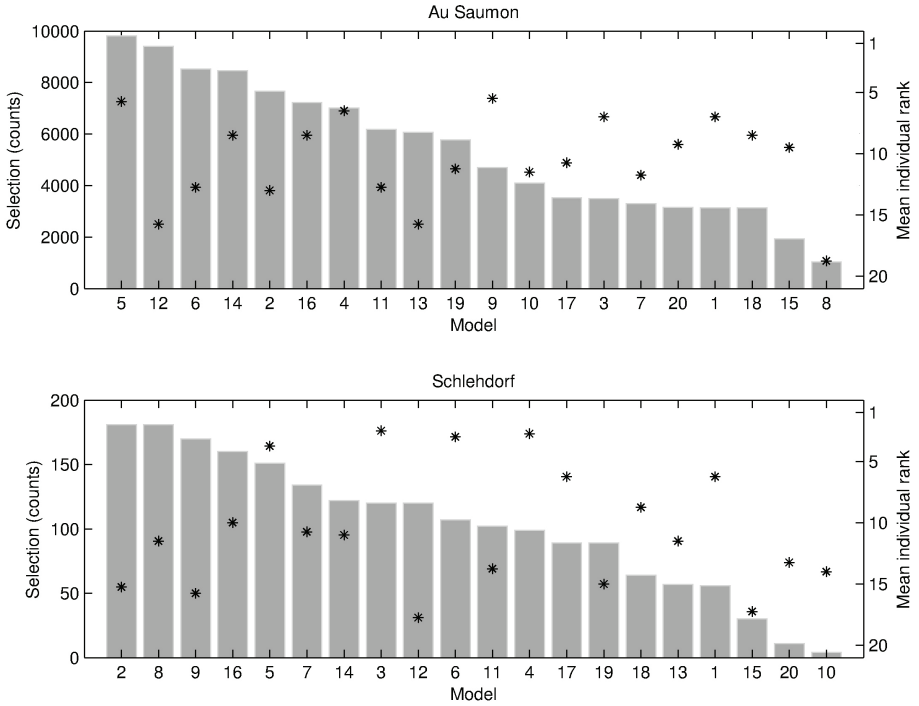

Fig. 13. Occurrence for each model in the selected ensembles for the Au Saumon (top) and Schlehdorf (bottom) catchments (grey bars), and mean individual rank (black stars).

HESSD
8, 10895–10933, 2011