Hydrol. Earth Syst. Sci., 16, 1171–1189, 2012
www.hydrol-earth-syst-sci.net/16/1171/2012/
doi:10.5194/hess-16-1171-2012

Hydrology and Earth System Sciences

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, Université Laval, Département de génie civil et de génie des eaux, 1065, avenue de la Médecine, Québec, Qc, G1V 0A6, Canada
[2]Irstea, Hydrosystems and Bioprocesses Research Unit (HBAN), 1, rue Pierre-Gilles de Gennes, 92761 Antony Cedex, France

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

Received: 17 November 2011 – Published in Hydrol. Earth Syst. Sci. Discuss.: 9 December 2011
Revised: 15 March 2012 – Accepted: 29 March 2012 – Published: 11 April 2012

Abstract. This paper investigates the temporal transposability of hydrological models under contrasted climate conditions and evaluates the added value of using an ensemble of model structures for flow simulation. This is achieved by applying the Differential Split Sample Test procedure to twenty lumped conceptual models on a catchment in the Province of Québec (Canada) and another one 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 may provide hazardous results when the model is to be applied in contrasted conditions. Overall, some models turned out as a good compromise in terms of performance and robustness, but generally not as much as the twenty-model ensemble. Model subsets offered yet improved performance over the twenty-model ensemble, but at the expanse of spatial transposability (i.e. need of site-specific analysis).

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. 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 (5) 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 perspective

Building hydrological models suitable for investigating the impacts of climate change is a major challenge for the scientific community. The 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 outside calibration data, 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. 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.

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 structural uncertainty.

Model intercomparison may also provide information on model complementarity and thus open ways to create multimodel combinations with improved efficiency. 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 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 robustness when climate conditions strongly differ between calibration and validation, following two application modes: individual and collective (ensemble).

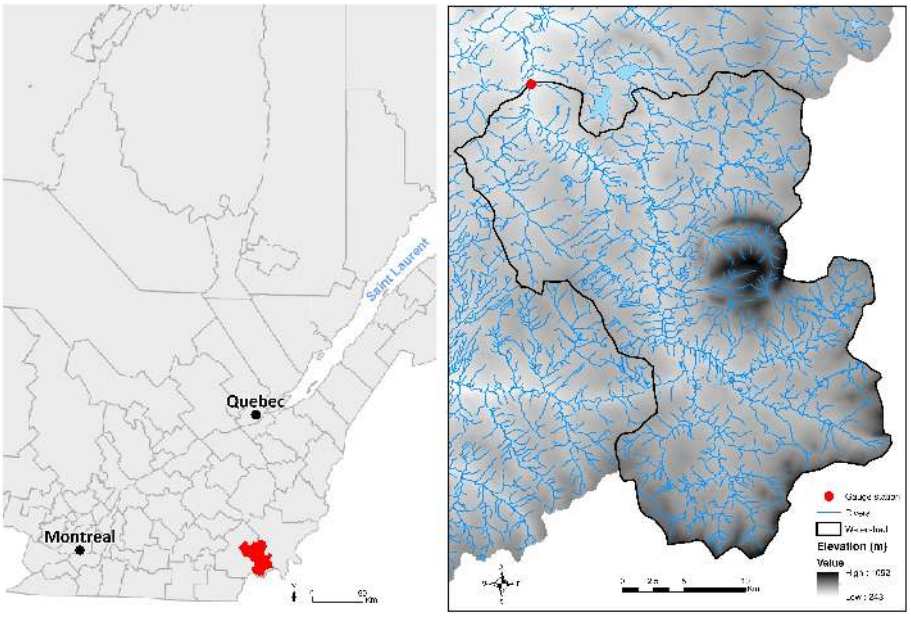

Fig. 1. Location of the Au Saumon catchment (738 km^2; Canada).

Our analysis 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 performance and transposability in time?

To answer these questions, the twenty hydrological models will be evaluated individually and collectively under the DSST framework on two catchments, in Canada and Germany.

The next section presents the catchments, data and models used, as well as the methodology and criteria selected to evaluate model performance. Then Sect. 3 details the results obtained by the models applied individually or as ensembles. Last we outline the main conclusions of this work.

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.

The Au Saumon catchment (Fig. 1) drains 738 km^2 of land. Its altitude ranges between 277 and 1092 m, for a mean annual air temperature of 4.5 °C. Its mean annual precipitation 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 reaches 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.

Table 1. Main characteristics of the periods selected for the DSST on the Au Saumon and Schlehdorf catchments (DW: dry/warm; DC: dry/cold; HW: humid/warm; HC: humid/cold) and relative maximum contrast between periods (computed as the ratio of the difference between maximum and minimum value over the four periods and the mean value over the whole record).

	Au Saumon						Schlehdorf					
	DW	DC	HW	HC	1975–2003	Relative max. contrast (%)	DW	DC	HW	HC	1970–2000	Relative max. contrast (%)
Average annual total precipitation (mm yr^{-1})	1126	1158	1421	1431	1284	23.8	1296	1229	1613	1517	1420	27.0
Average daily mean temperature (°C)	5.22	3.87	5.28	3.86	4.50	31.6	5.94	4.68	5.70	4.78	5.21	24.2
Average annual total discharge (mm yr^{-1})	677	765	883	874	771	26.7	870	834	1106	1054	983	27.7

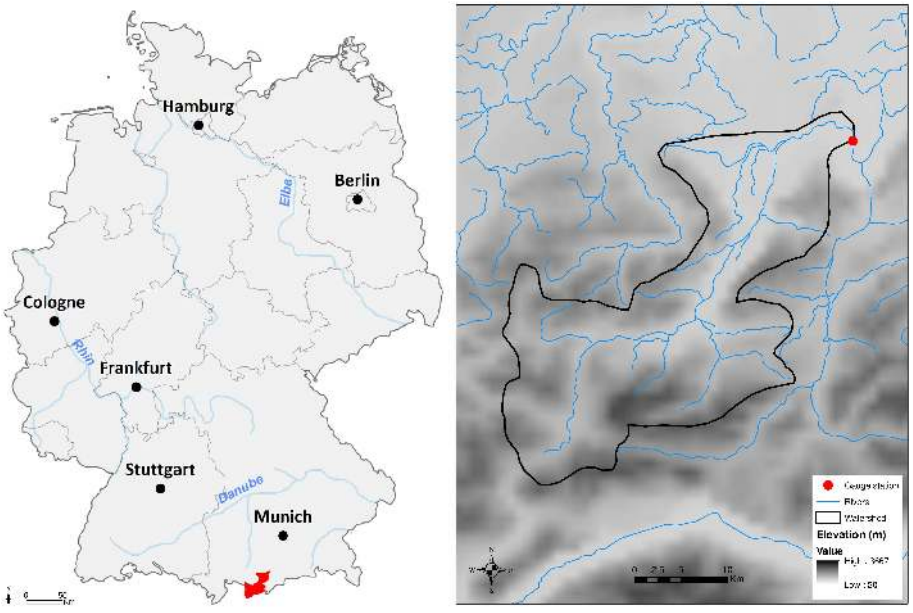

Fig. 2. Location of the Schlehdorf catchment (708 km^2; Germany).

2.2 Lumped conceptual hydrological models

Twenty lumped conceptual hydrological models were selected in this study, to get a wide variety of conceptualizations of the rainfall-runoff relationship. 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 are all 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 non-linear formulations, various unit hydrographs or simple time delays. Most of these model versions originate from the works by Perrin et al. (2001) and Mathevet (2005), and were used by Velázquez et al. (2010). Although these model structures represent a wide panel of how the rainfall-runoff relationship can be conceptualized, we acknowledge that this selection does not cover the whole spectrum of model types, e.g. not including distributed physically-based models. However, given the evaluation scheme adopted here and the amount of calculations needed, we limited this study to parsimonious models.

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. Some model structures (e.g. M_{01} and M_{05}) route one of the flow components simply using a unit hydrograph and not a routing store.

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

Model name	Model acronym	Number of optimized parameters	Number of storages	Derived from
M01	BUCK	6	3	BUCKET (Thorthwaite and Mather, 1955)
M02	CEQU	9	2	CEQUEAU (Girard et al., 1972)
M03	CREC	6	3	CREC (Cormary and Guilbot, 1973)
M04	GARD	6	3	GARDENIA (Thiery, 1982)
M05	GR4J	4	3	GR4J (Perrin et al., 2003)
M06	HBV0	9	3	HBV (Bergström et al., 1973)
M07	HYMO	6	5	HYMOD (Wagener et al., 2001)
M08	IHAC	7	3	IHACRES (Jakeman et al., 1990)
M09	MART	7	4	MARTINE (Mazenc et al., 1984)
M10	MOHY	7	3	MOHYSE (Fortin et al., 2006)
M11	MORD	6	4	MORDOR (Garçon, 1999)
M12	NAM0	10	7	NAM (Nielsen et al., 1973)
M13	PDM0	8	4	PDM (Moore et al., 1981)
M14	SACR	9	5	SACRAMENTO (Burnash et al., 1973)
M15	SIMH	8	4	SIMHYD (Chiew et al., 2002)
M16	SMAR	8	4	SMAR (O'Connell et al., 1981)
M17	TANK	7	4	TANK (Sugarawa, 1979)
M18	TOPM	7	4	TOPMODEL (Beven and Kirkby, 1979)
M19	WAGE	8	3	WAGENINGEN (Warmerdam et al., 1997)
M20	XINA	8	5	XINANJIANG (Zhao et al., 1980)

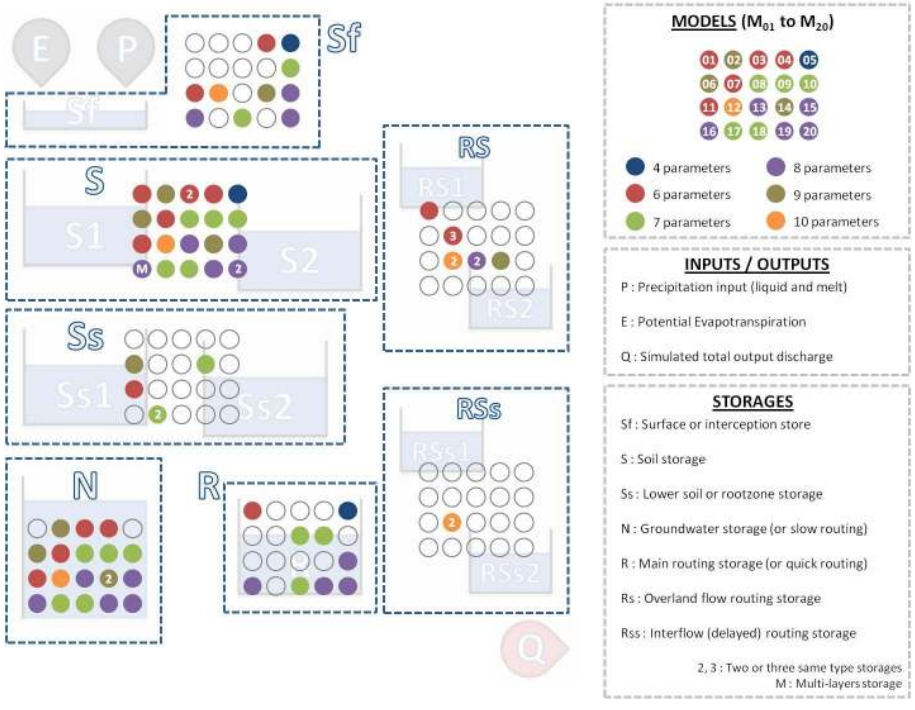

Fig. 3. Illustration of model structural diversity (all models are put in the same frame).

All models were applied in exactly the same conditions: they were run at the 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 state variables 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.

One main advantage of using this snow accounting module 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, almost 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. Temporal transposability can be understood as the capacity of the model to perform with the same level of accuracy under conditions different from the calibrations ones. This can be linked to robustness, a desired property of models whose parameters do not show oversensitivity to changes in data used for calibration. However, it is well known that model parameters depend on the information content of calibration series (see e.g. Wagener et al., 2003; Perrin et al., 2008). So, there is no guarantee that the parameters optimized for the current conditions will still be appropriate for the future ones. This is why hydrological tests on 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. humid 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 applied the three-step testing procedure below to our set of twenty models:

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

- Evaluate model performance using preselected criteria and comparatively assess the relative transposability of the tested models in the various configurations: DW $\rightarrow$ HC, HC $\rightarrow$ DW, DC $\rightarrow$ HW, HW $\rightarrow$ 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. Table 1 presents the mean characteristics of the selected periods for each catchment. Differences in mean precipitation or temperature between periods can range from 23.8 to 31.6 % of the mean value over the whole record, which represents significant contrasts. This results in maximum differences between periods of about 27 % in mean flow, as also illustrated in Fig. 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 flows as well as summer high flows show important variations between periods.

2.4 Model calibration and performance criteria

2.4.1 Optimization algorithm and objective function

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

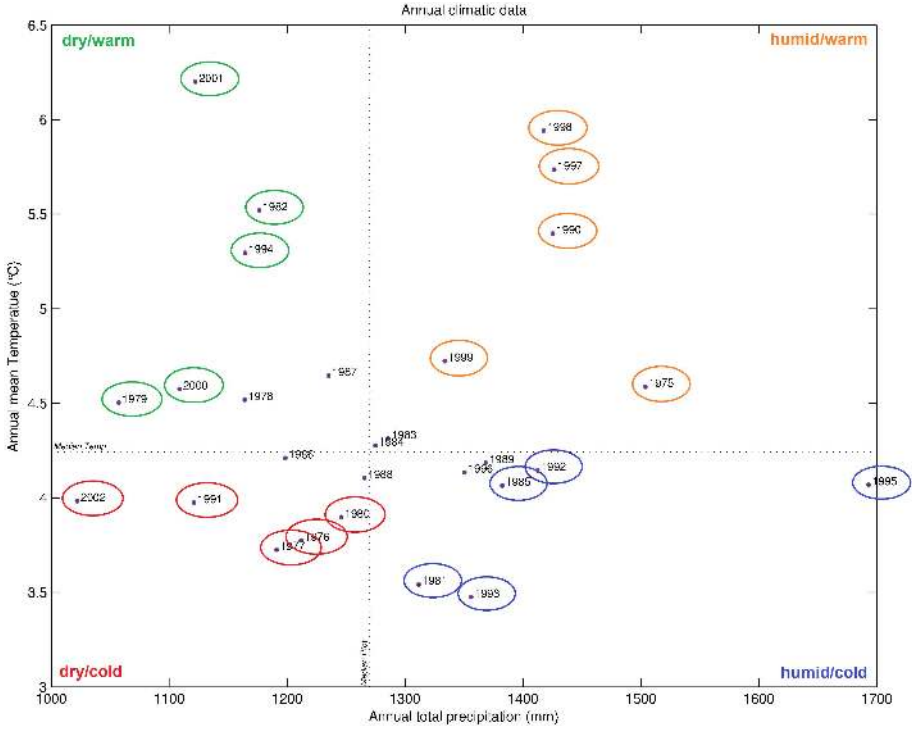

Fig. 4. Time series clustering results for the Au Saumon catchment (SAU).

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. $RMSE_{sqrt}$ 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., 2006a,b).

2.4.2 Efficiency criteria in validation

Several criteria were used for the evaluation of model performance in validation. 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. 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 (NSE_{log}) that puts more weight on low flows.

The percentage volume error (PVE) (Moriasi et al., 2007) was computed to give information on the agreement between observed and simulated total discharge over the test period:

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

A value of 0 indicates perfect agreement and larger values indicate increasing volume error (over- or underestimation).

Note that the comparison of performance in validation between DSST may be biased by the use of the NSE-type criteria, because the variance used as the denominator is different for each selected period (Martinec and Rango, 1989). To circumvent this possible bias, our analysis will primarily be performed on a relative basis, using the rank in model

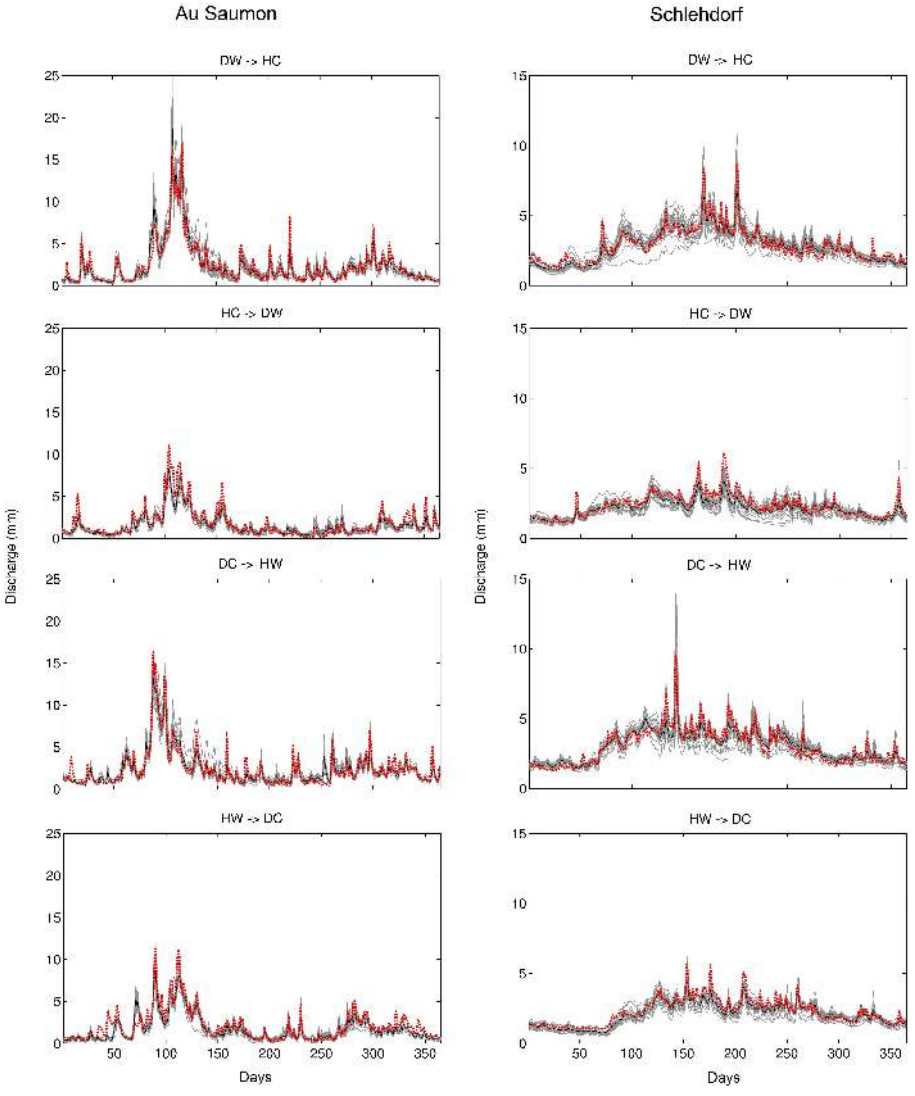

Fig. 5. Mean daily interannual discharges for all the DSS tests on validation periods, for the Au Saumon and Schlehdorf catchments. Grey lines are the twenty individual models, black line is the twenty-member ensemble and large red dotted line the observed discharge.

performance within the twenty-model set. We acknowledge that large difference in ranks may correspond to small differences in model performance or vice versa. But we think that this analysis by ranks makes the relative transposability more comparable between DSST. In the following, we will mainly analyze results based on ranks for the NSE_{sqrt} criterion.

In addition to the performance and transposability calculations, the collective diversity of the models is of concern for the multimodel approach. By analyzing diversity in the simulated time series, we aim at quantifying redundancy and/or complementarity 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):

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

with $\sigma_i = \sqrt{\frac{1}{M}\sum_{m=1}^{M}\left(Q_{\mathrm{sim},i,m} - \overline{Q_{\mathrm{sim},i}}\right)^2}$ 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. Here diversity will be used as a complementary criterion to actual performance to better understand what makes the strength of the multimodel approach.

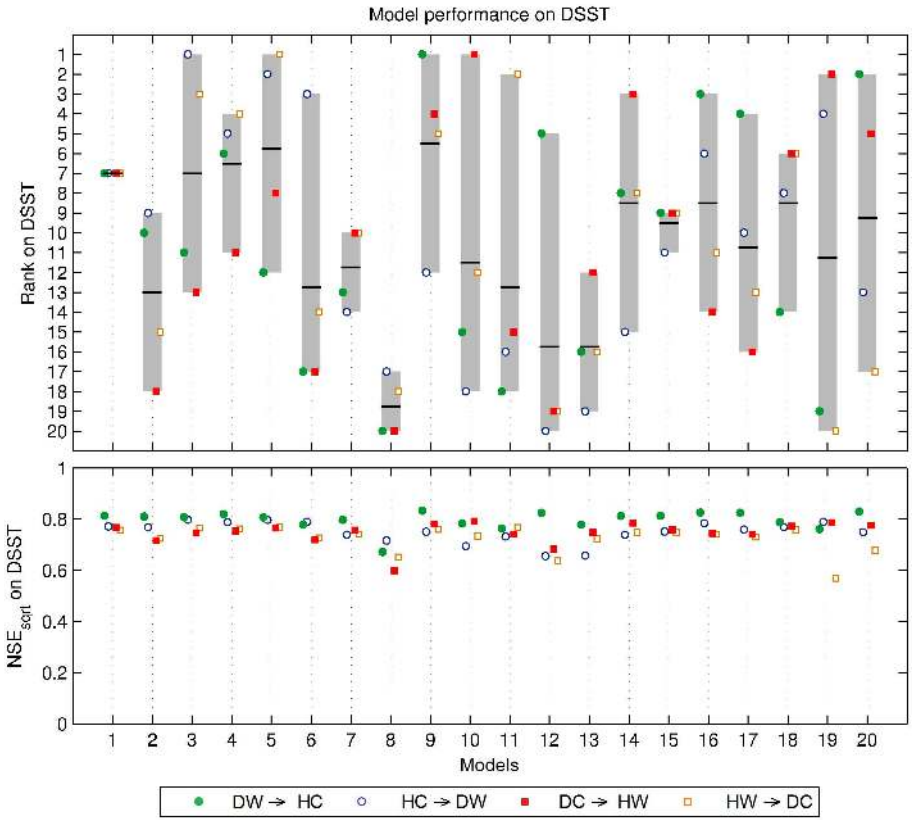

Fig. 6. Performance and rank in validation (NSE_{sqrt} criterion) for the four DSST on the Au Saumon catchment (SAU).

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.:

- validation on humid-cold period after calibration on dry-warm period (DW → HC),
- validation on dry-warm period after calibration on humid-cold period (HC → DW),
- validation on humid-warm period after calibration on dry-cold period (DC → HW),
- validation on dry-cold period after calibration on humid-warm period (HW → DC).

The NSE_{sqrt} and PVE results, for every models and tests on the Au Saumon time series, are compiled in Table 3 and illustrated in Figs. 6 and 7, while results for the Schlehdorf catchment are shown in Table 3, Figs. 8 and 9. 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. One should seek for models that have better performance than the others on average (better models obtain lower value of mean rank). For models with equivalent performance, one should reject those that are good on some DSST and bad on the others relatively to the other models (more robust/transposable models show shorter grey bars).

3.1.1 Au Saumon catchment

For the Au Saumon catchment (Figs. 6 and 7), M_{09}, M_{05} and M_{04} models produce the best mean ranks on NSE_{sqrt}. Interestingly, for these models, at least one DSST yields much less robust results than the others (e.g. HC → DW for M_{09}), showing that it is difficult for the best models to be robust in all test conditions. These three models seem also to perform differently between DSST: while M_{09} shows better robustness in validation on humid years after calibration on dry years, M_{05} and M_{04} are more robust in the reverse configuration. When looking at these three model structures, it is difficult to identify which key functions provide robustness. M_{05} and M_{04} differ from M_{09} in that they include a water balance correction function. All models have two flow components, and include at least one non linear routing store, but the number of routing stores varies from 1 to 3.

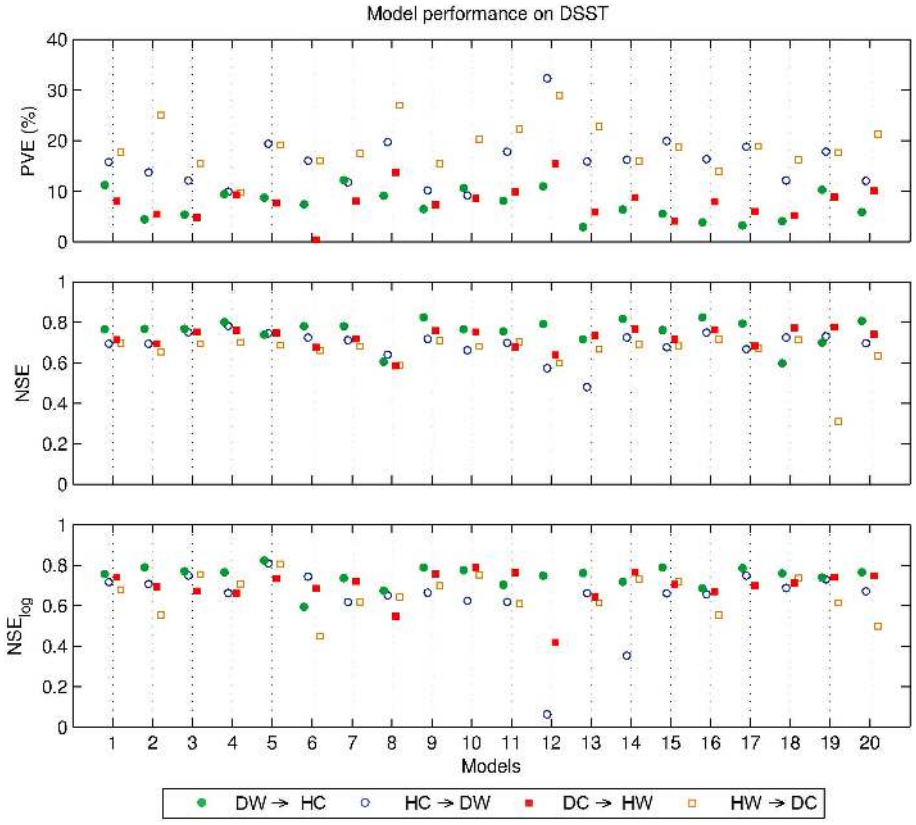

Fig. 7. Performance in validation (PVE, NSE and NSE_{log} criterion) for the four DSST on the Au Saumon catchment (SAU). Negative values for NSE not shown.

Conversely, M_{08}, M_{12} and M_{13} show a poor robustness with mean ranks varying respectively from 18.75 to 15.75. Although M_{08} appears poorly robust in all circumstances, M_{12} manages to get quite robust results in the DW → HC case. Like for the best models, it is difficult here to find what prevent these models from getting robust results. Their only common characteristic is to have only linear routing stores.

Some models can obtain similar ranks (e.g. M_{01} and M_{03}) but with different behaviours: M_{01} seems equally robust for all DSST while M_{03} shows much more contrasted results.

When looking at the other performance criteria (see Table 3 and Fig. 7), similar conclusions could be drawn that no single model could be the best on all DSST.

Results in terms of water balance seem quite sensitive to the type of test, as shown by PVE values (Table 3, Figs. 7 and 10). Several models tend to under-evaluate water volumes. This is expected for the tests with calibration on humid years and validation on dry years but it sometimes also occurs for the opposite situation. The DW→HC (PVE values from 2.92 % to 12.17 %) and DC → HW (from 0.43 % to 15.46 %) tests yield the best general results. In the two other cases, PVE values are worse (from 9.17 % to 32.29 % for HC → DW; from 9.72 % to 28.92 % for HW → DC). This statement is linked to the under-evaluation of water volume, more penalising for these two tests as illustrated in Fig. 10.

3.1.2 Schlehdorf catchment

Results for the Schlehdorf catchment (Figs. 8 and 9) highlight different models than for the Au Saumon catchment. For instance, M_{09}, M_{14}, and M_{15} show low robustness, while M_{03}, M_{04} and M_{06} give best climate transposability with mean ranks from 2.5 to 6. In general, for each DSST, differences in performance are larger between models than for the Au Saumon catchment. This also results in more contrasted robustness results, some models being robust in all DSST. Overall, M_{03}, M_{04}, M_{05} and M_{06} are the most appealing models, both in terms of robustness and performance on the various efficiency criteria. Like for the Au Saumon catchment, it is quite difficult to identify which common characteristics in the model structures make all of them quite equally satisfactory.

As for the Au Saumon, PVE performance (Table 3 and Fig. 9) shows contrasted results. It can be noted that M_{09} is probably the worst model with PVE exceeding 30 % for three of the DSSTs. As illustrated in Fig. 10, 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 humid years while they underestimate water quantities for the opposite situation. The

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 NSE_{sqrt} sub-selection)
			Au Saumon			
NSE_{sqrt} [−]	DW → HC	0.83 (M_{09})	0.81	0.67 (M_{08})	0.86	0.87 (6 mod)
	HC → DW	0.80 (M_{03})	0.75	0.65 (M_{12})	0.81	0.84 (5 mod)
	DC → HW	0.79 (M_{10})	0.75	0.60 (M_{08})	0.80	0.81 (7 mod)
	HW → DC	0.77 (M_{05})	0.74	0.57 (M_{19})	0.79	0.81 (5 mod)
NSE [−]	DW → HC	0.82 (M_{16})	0.77	0.60 (M_{18})	0.83	0.84
	HC → DW	0.78 (M_{04})	0.71	0.48 (M_{13})	0.73	0.77
	DC → HW	0.78 (M_{19})	0.74	0.59 (M_{08})	0.77	0.79
	HW → DC	0.72 (M_{16})	0.68	0.31 (M_{19})	0.71	0.73
NSE_{log} [−]	DW → HC	0.82 (M_{05})	0.76	0.59 (M_{06})	0.85	0.87
	HC → DW	0.81 (M_{05})	0.66	0.06 (M_{12})	0.78	0.83
	DC → HW	0.79 (M_{10})	0.71	0.42 (M_{12})	0.76	0.78
	HW → DC	0.80 (M_{05})	0.63	−5.24 (M_{17})	0.81	0.84
PVE [%]	DW → HC	2.92 (M_{13})	6.94	12.17 (M_{07})	2.2	0.2
	HC → DW	9.17 (M_{10})	15.94	32.29 (M_{12})	15.8	15.0
	DC → HW	0.43 (M_{06})	8.01	15.46 (M_{12})	2.9	2.4
	HW → DC	9.72 (M_{04})	18.19	28.92 (M_{12})	19.0	18.3
			Schlehdorf			
NSE_{sqrt} [−]	DW → HC	0.80 (M_{04})	0.71	0.31 (M_{12})	0.83	0.87 (5 mod)
	HC → DW	0.81 (M_{04})	0.66	0.05 (M_{18})	0.79	0.85 (5 mod)
	DC → HW	0.83 (M_{05})	0.73	0.43 (M_{12})	0.81	0.86 (7 mod)
	HW → DC	0.86 (M_{03})	0.74	0.38 (M_{09})	0.85	0.89 (8 mod)
NSE [−]	DW → HC	0.80 (M_{06})	0.66	0.18 (M_{12})	0.82	0.88
	HC → DW	0.77 (M_{17})	0.61	0.24 (M_{09})	0.74	0.82
	DC → HW	0.81 (M_{05})	0.72	0.31 (M_{12})	0.81	0.87
	HW → DC	0.82 (M_{03})	0.75	0.45 (M_{12})	0.83	0.86
NSE_{log} [−]	DW → HC	0.79 (M_{03})	0.70	0.36 (M_{09})	0.82	0.86
	HC → DW	0.83 (M_{04})	0.70	−0.28 (M_{18})	0.82	0.87
	DC → HW	0.80 (M_{05})	0.72	0.46 (M_{12})	0.80	0.84
	HW → DC	0.87 (M_{03})	0.67	−0.24 (M_{12})	0.83	0.89
PVE [%]	DW → HC	0.02 (M_{01})	4.17	30.11 (M_{09})	2.0	0.4
	HC → DW	0.42 (M_{03})	9.12	32.61 (M_{12})	11.6	3.5
	DC → HW	0.08 (M_{10})	5.04	17.55 (M_{11})	1.5	0.6
	HW → DC	0.17 (M_{02})	7.99	31.41 (M_{09})	10.0	4.5

range of performance for water balance is however larger for this catchment.

3.1.3 Synthesis on individual performance

These results illustrate the difficulty in identifying a single lumped model that could behave well in terms of performance and robustness, when tested under all possible contrasted conditions. This remains one of the main challenges of hydrological projection studies under climate change. Besides, model performance and robustness are clearly dependent on the test catchment, which corroborates previous findings obtained by applying the more usual SST. Here it seems more difficult to identify a generally robust model on the Au Saumon catchment than on the Schlehdorf catchment.

Nevertheless, our tests allow identifying best-compromise individual models for each catchment based on results illustrated in Figs. 6 and 8. For the Au Saumon catchment, models M_{04}, M_{05}, and M_{09} are the three best compromises, whereas for Schlehdorf M_{03}, M_{04}, and M_{06} are identified. This better robustness is quite difficult to explain solely based on the analysis of model structure components.

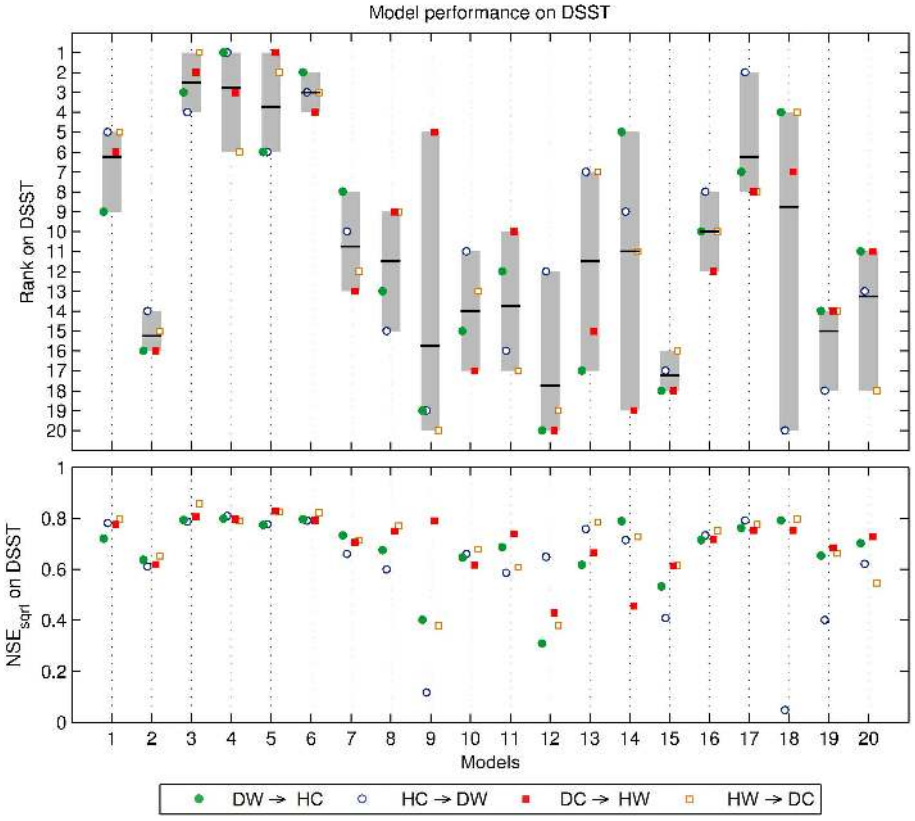

Fig. 8. Performance and rank in validation (NSE_{sqrt} criterion) for the four DSST on the Schlehdorf catchment (SLD).

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. M_{09} and M_{12} for the Schlehdorf catchment in the DW → HC case strongly underestimate streamflows). This indicates the identification of non robust parameter sets in some cases, a limitation that may not appear when applying SST under similar conditions.

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 almost all possible models combinations: 2^{20} possibilities (i.e. 1 048 576) minus all combinations of less than five models (i.e. 6196), which are excluded for the lack of a reliable evaluation of their diversity (CV). As mentioned in Sect. 2.4, considering CV is used to measure the hydrological range of the model responses (i.e. structural variability).

Results for the Au Saumon and Schlehdorf catchments are illustrated in Figs. 11 and 12, respectively. The red lines and circle represent the performance and the diversity of the twenty-member ensembles, while the blue vertical line is the performance of the best individual model. Table 3 and Figs. 5 and 10 also illustrate the multimodel results.

3.2.1 Twenty-member ensemble

The twenty-member ensemble gives better results than the best individual model for all DSSTs on the Au Saumon catchment, as shown in Fig. 11 and Table 3. Although the improvement is not large, it is substantial in all cases. This holds for only one of the four Schlehdorf DSSTs (Fig. 12 and Table 3). Nonetheless, the multimodel approach remains a valuable alternative 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 HC → DW (NSE_{sqrt} of 0.81), M_{05} in DC → HW (0.83), and M_{03} in HW → DC (0.86). In each case, no other single model surpasses the twenty-model performance. Table 3 also illustrate that for the three other evaluation criteria (NSE, NSE_{log} and PVE) some individual models overcome the twenty-member ensemble, showing that entire analysis based on different criteria could lead to somehow different interpretations.

Figures 11 and 12 illustrate the link between performance and diversity for the two catchments. For the DW → HC test, low diversity tends to limit model performance, while the

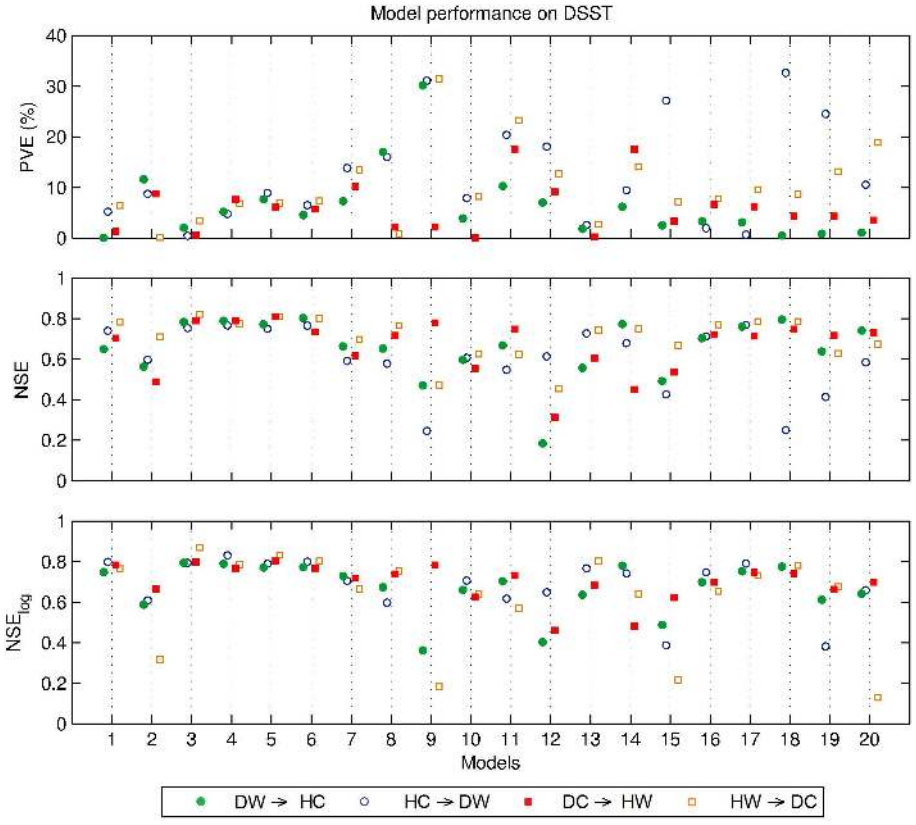

Fig. 9. Performance in validation (PVE, NSE and NSE_{log} criterion) for the four DSST on the Schlehdorf catchment (SLD). Negative values for NSE not shown.

opposite is true for the HW → DC test. For the HC → DW test, the two catchments show different behaviour, while for the HW → DC test, an intermediate diversity yields best performance.

Concerning water balance, Fig. 10 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 or under-estimated by the various models on the same test, the ensemble approach is the most efficient (e.g. DW → HC for Schlehdorf catchment). Figure 5 also illustrates these results and shows the good fit between observed (in large red dotted lines) and twenty-member-ensemble (black line) simulated series of mean daily discharge.

3.2.2 Sub-selections

Results also reveal that many other model combinations (sub-selections) provide better performance than the twenty-member ensemble. They are located in the right of the red lines portion of the DSST plots in Figs. 11 and 12. For the Au Saumon catchment (Fig. 11), they correspond to 19.9 % of the studied combinations for the DW → HC test, 36.5 % for the HC → DW, 28.4 % for the DC → HW, and 29.9 % for the HW → DC. The same holds for the Schlehdorf catchment (Fig. 12), for which they encompass 33.8 % of the combinations for DW → HC, 42.7 % for the HC → DW, 39.2 % for the DC → HW, and 34.3 % for the HW → DC test.

Because one needs to work on performance 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 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. They represent 5.80 % of the possible combinations (60 437 ensembles) for the Au Saumon catchment, and 6.58 % (68 627 ensembles) for the Schlehdorf catchment. With these efficient 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 to 8 members), as also depicted in Table 3.

3.3 Individual versus collective performance

To evaluate the benefit of the above selected model ensembles, they were confronted to the individual models and to

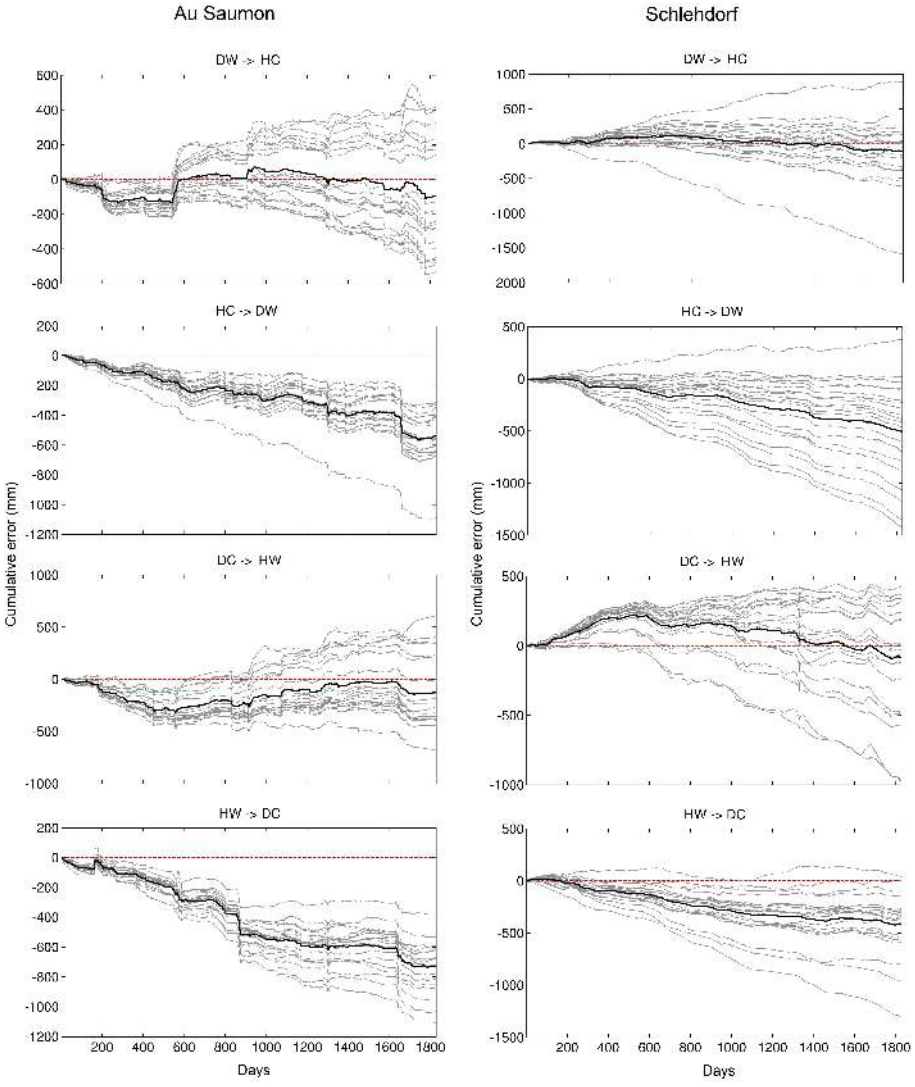

Fig. 10. Cumulative error between observed and simulated discharges for all the DSS tests in validation, for the Au Saumon and Schlehdorf catchments. Grey lines are the twenty individual models, large black line is the twenty-member ensemble and the horizontal dashed line indicates the optimal value.

the twenty-model ensemble. Figure 13 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 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, especially on Schlehdorf catchment where the gain in terms of NSE_{sqrt} is 0.05 on average (0.02 for Au Saumon catchment). This selected multimodel becomes better than the best individual models in all cases for NSE_{sqrt} criterion and almost all the other evaluation criteria. This sub-selection will be identified accordingly to the user's objectives; one may prefer a lower number of models, best performance in terms of NSE_{sqrt}, best performance on the overall criterion (NSE_{sqrt}, NSE, NSE_{log} and PVE), or a mix of performance and diversity.

As a final analysis, Fig. 14 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 59641 appearances in 60 437 combinations (i.e. 99 % of cases), whereas M_{08} is used only 2398 times (i.e. 4 % of cases). 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. 6). On the

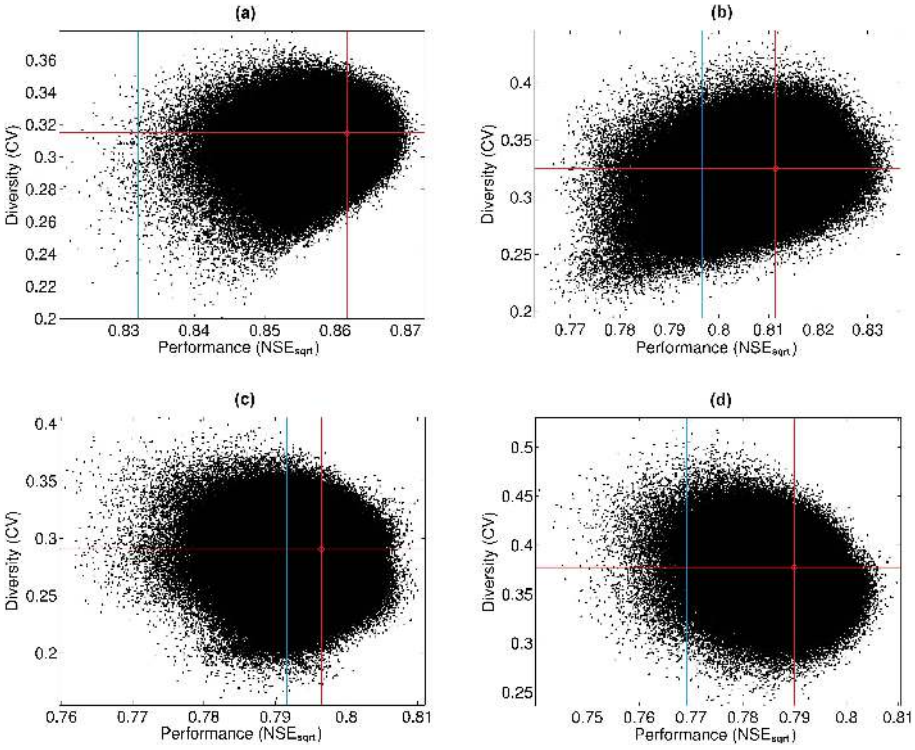

Fig. 11. Validation performance (NSE_{sqrt}) and diversity (CV) for all model combinations (2^{20} points) and Differential Split Sample Tests for the Au Saumon catchment (SAU): **(a)** calibration on DW years (dry/warm) and validation on HC years (humid/cold); **(b)** calibration on HC years (humid/cold) and validation on DW years (dry/warm); **(c)** calibration on DC years (dry/cold) and validation on HW years (humid/warm); **(d)** calibration on HW years (humid/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.

other hand, M_{07} and M_{15}, which have shown great robustness and correct performance, are also not frequently used. This is the same for the best-compromise model M_{09} (seventh commonly used model). Globally, comparing selection counts and mean individual rank, no link can be identified.

The same analysis differs in the case of the Schlehdorf catchment. M_{05} and M_{03} are present respectively in 54 788 (i.e. 80 %) and 52 136 (i.e. 76 %) combinations, and M_{15} is the lesser used (11 708 selections, i.e. 17 %). Interestingly, M_{05} and M_{03} showed a good range of performance and high robustness, while M_{15} lead to low performance and was systematically ranked among the poorest models. For Schlehdorf catchment, we can highlight some link between selection counts and mean individual rank. This link is clearer for this catchment probably because individual results were also more contrasted between models.

For both catchments, M_{05} is the most commonly used and also one of the best individual performances.

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.

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, assessing their relevance for climate impact

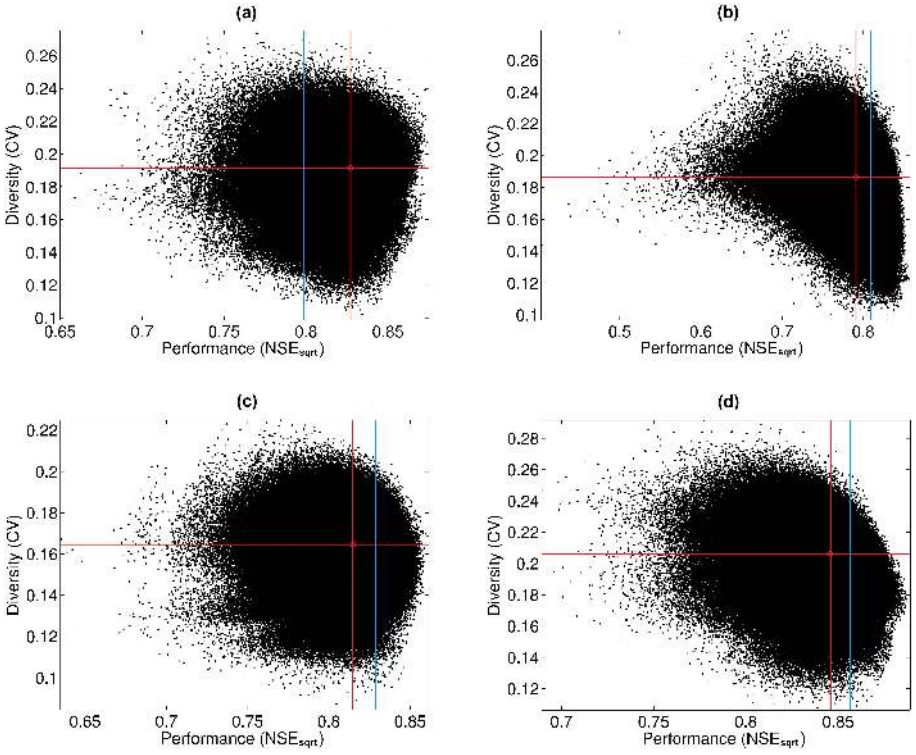

Fig. 12. Validation performance (NSE_{sqrt}) and diversity (CV) for all model combinations (2^{20} points) and Differential Split Sample Tests for the Schlehdorf catchment (SLD): **(a)** calibration on DW years (dry/warm) and validation on HC years (humid/cold); **(b)** calibration on HC years (humid/cold) and validation on DW years (dry/warm); **(c)** calibration on DC years (dry/cold) and validation on HW years (humid/warm); **(d)** calibration on HW years (humid/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.

studies. This investigation showed that it is unsafe to rely on a single lumped 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 structures 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, almost all possible model combinations (1 042 380 possibilities) have been explored. Many combinations were found to provide increased performance over the twenty-member ensemble, leaving an operational hydrologist with the option of fine

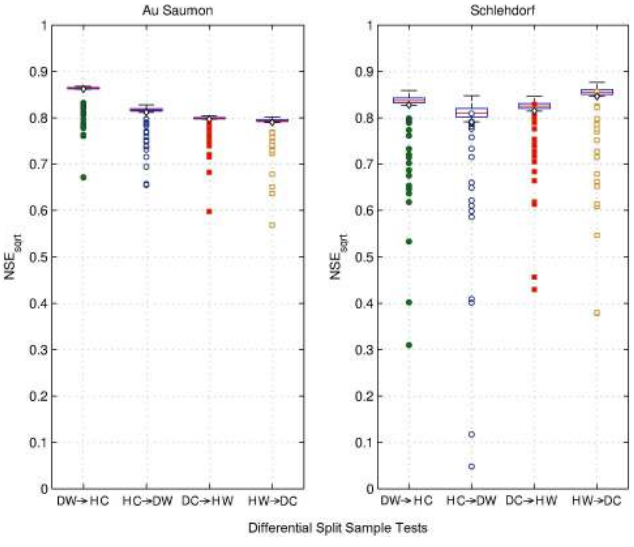

Fig. 13. Individual and multimodel DSST validation performance. Boxplots depict the range of the multimodel combinations, diamonds represent the twenty-model ensemble, and the circles and squares, the individual models, for the Au Saumon and Schlehdorf catchments.

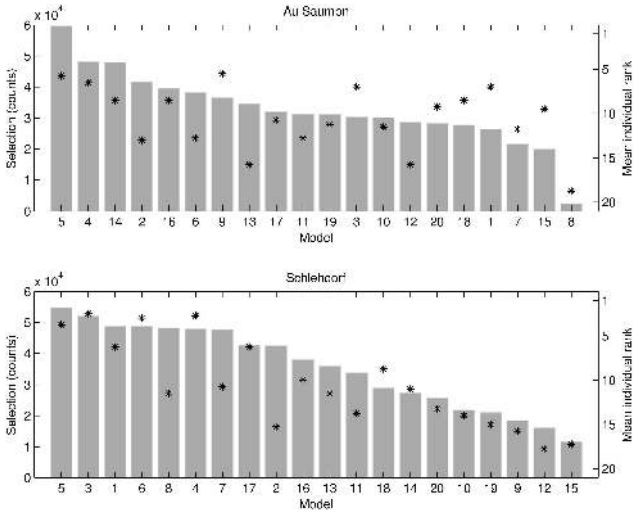

Fig. 14. Occurrence for each model in the selected ensembles for the Au Saumon (top panel) and Schlehdorf (bottom panel) catchments (grey bars), and mean individual rank (black stars).

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 if best performing models may more likely contribute to the ensemble, worse-performing individual models can successfully contribute to an ensemble (especially on Au Saumon catchment), 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). The role of diversity in the ensemble was also shown to have various influences on the ensemble performance, depending on the DSST.

This study does not provide an analysis of the physical adequacy of model structure and estimated parameters. 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 and Fenicia, 2011 on some experimental catchments).

Acknowledgements. The authors acknowledge NSERC, Ouranos, and Hydro-Québec for support, as well as the other partners in the QBIC^3 project. We also thank the three reviewers for their discussions, comments, and references.

Edited by: A. Gelfan

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, 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, 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, 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 Publ., 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, Korea. Journal of Hydrology, 401, 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, p. 237, 2008.

Bergström, S. and Forsman, A.: Development of a conceptual deterministic rainfall-runoff model, Nord. Hydrol., 4, 147–170, 1973.

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

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology, Hydrol. Sci. Bull., 24, 43–69, 1979.

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.

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, 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 Res., 32, 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.

Burnash, R. J. C., Ferral, R. L., and McGuire, R. A.: A generalized streamflow simulation system – Conceptual modelling for digital computers, US Department of Commerce, National Weather Service and State of California, Department of Water Resources, p. 204, 1973.

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, 242–266, 2004.

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

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

Chiew, F. H. S., Peel, M. C., and Western, A. W.: Application and testing of the simple rainfall-runoff model SIMHYD, in: Mathematical Models of Small Watershed Hydrology and Applications, edited by: Singh, V. P. and Frevert, D. K., Water Resources Publication, Littleton, Colorado, 335–367, 2002.

Cormary, Y. and Guilbot, A.: Étude des relations pluie-débit sur trois bassins versants d'investigation, IAHS Madrid Symposium, IAHS Publication no. 108, 265–279, 1973.

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

Duan, Q. and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 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, 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.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resour. Res., 47, W11510, doi:10.1029/2010WR010174, 2011.

Fortin, V. and Turcotte, R.: Le modèle hydrologique MOHYSE, Note de cours pour SCA7420, Département des sciences de la terre et de l'atmosphère, Université e du Québec à Montréal, 2006.

Garçon, R.: Modèle global Pluie-Débit pour la prévision et la prédétermination des crues, La Houille Blanche 7/8, 88–95, 1999.

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

Girard, G., Morin, G., and Charbonneau, R.: Modèle précipitations-débits à discrétisation spatiale, Cahiers ORSTOM, Série Hydrologie, IX, 35–52, 1972.

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, p. 211, 2010.

Jakeman, A. J., Littlewood, I. G., and Whitehead, P. G.: Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments, J. Hydrol., 117, 275–300, 1990.

Kavetski, D. and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resour. Res., 47, W11511, doi:10.1029/2011WR010748, 2011.

Kim, Y.-O., Jeong, D., and Ko, I. H.: Combining Rainfall-Runoff Model Outputs for Improving Ensemble Streamflow Prediction, J. Hydrol. Eng., 11, 578–588, 2006.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrolog. Sci. J. – des Sciences Hydrologiques, 31, 13–24, 1986.

Kottegoda, N. T. and Rosso, R.: Applied Statistics for Civil Environnemental Engineers, Electronic version, Wiley, Chichester, 2009.

Loumagne, C., Vidal, J., Feliu, C., Torterotot, J., and Roche, P.: Procédure de décision multimodèle pour une prévision des crues en temps réel, application au bassin supérieur de la Garonne, Revue des sciences de l'eau, 8, 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, 2009, http://www.adv-geosci.net/21/63/2009/.

Martinec, J. and Rango, A.: Merits of statistical criteria for the performance of hydrological models, Water resources Bulletin, Wiley Online Library, 25, 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, 2005.

Maurer, E. P.: Uncertainty in hydrologic impacts of climate change in the Sierra Nevada, California, under two emissions scenarios, Climatic Change, 82, 309–325, 2007.

Mazenc, B., Sanchez, M., and Thiery, D.: Analyse de l'influence de la physiographie d'un bassin versant sur les paramètres d'un modèle hydrologique global et sur les débits caractéristiques à l'exutoire, J. Hydrol., 69, 97–118, 1984.

McGuinness, J. L. and Bordne, E. F.: A comparison of lysimeter-derived potential evapotranspiration with computed values, Search, Alderman, p. 71, 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, 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, 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, 70–83, 2008.

Moore, R. J. and Clarke, R. T.: A distribution function approach to rainfall-runoff modeling, Water Resour. Res., 17, 1367–1382, 1981.

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, 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.

Nielsen, S. A. and Hansen, E.: Numerical simulation of the rainfall-runoff process on a daily basis, Nord. Hydrol., 4, 171–190, 1973.

O'Connell, P. E., Nash, J. E., and Farrell, J. P.: River flow forecasting through conceptual models, Part II – The Brosna catchment at Ferbane, J. Hydrol., 10, 317–329, 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 – Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling, J. Hydrol., 303, 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, 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, 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, 275–301, 2001.

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, J. Hydrol., 279, 275–289, doi:10.1016/S0022-1694(03)00225-7, 2003.

Perrin, C., Oudin, L., and Andreassian, V.: Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models, Hydrol. Sci., 52, 131–151, 2007.

Perrin, C., Andréassian, V., Rojas-Serna, C., Mathevet, T., and Le Moine, N.: Discrete parameterization of hydrological models: evaluating the use of parameter sets libraries over 900 catchments, Water Resour. Res., 44, W08447, doi:10.1029/2007WR006579, 2008.

Prudhomme, C., Jakob, D., and Svensson, C.: Uncertainty and climate change impact on the flood regime of small UK catchments, J. Hydrol., 277, 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, 27–60, 2004.

Refsgaard, J. C. and Knudsen, J.: Operational Validation and Intercomparison of Different Types of Hydrological Models, Water Resour. Res., 32, 2189, doi:10.1029/96WR00896, 1996.

Refsgaard, J., Vandersluijs, J., Brown, J., and Vanderkeur, P.: A framework for dealing with uncertainty due to model structure error. Adv. Water Resour., 29, 1586–1597, 2006.

Seibert, J.: Reliability of model predictions outside calibration conditions, Nord. Hydrol., 34, 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, 203–229, 1997.

Sugawara, M.: Automatic calibration of the tank model, Hydrolog. Sci. J., 24, 375–388, 1979.

Thiery, D.: Utilisation d'un modèle global pour identifier sur un niveau piézométrique des influences multiples dues à diverses activités humaines, IAHS Publication no. 136, 71-77, 1982.

Thornthwaite, C. W. and Mather, J. R.: The water balance, Publications in Climatology, Drexel Institute of Climatology, Centerton, NJ, 8, 1–104, 1955.

Valéry, A.: Modélisation précipitations – débit sous influence nivale, Élaboration d'un module neige et évaluation sur 380 bassins versants, 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, 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. Hydrol., 394, 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, 147–158, 2009.

Wagener, T., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V., and Sorooshian, S.: A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13–26, doi:10.5194/hess-5-13-2001, 2001.

Wagener, T., McIntyre, N., Lees, M. J., Wheater, H. S., and Gupta, H. V.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, Hydrol. Process., 17, 455–476, 2003.

Warmerdam, P. M. M., Kole, J., and Chormanski, J.: Modelling rainfall-runoff processes in the Hupselse Beek research basin, Ecohydrological processes in small basins, Proceedings of the Strasbourg Conference, 24-26 September 1996, IHP-V, Technical Documents in Hydrology no. 14, UNESCO, Paris, 155–160, 1997.

Xu, C.-Y.: Operational testing of a water balance model for predicting climate change impacts, Agr. Forest Meteorol., 98–99, 295–304, 1999.

Xu, C.-Y., Widén, E., and Halldin, S.: Modelling Hydrological Consequences of Climate Change – Progress and Challenges, Adv. Atmos. Sci., 22, 789–797, 2005.

Zhao, R. J., Zuang, Y. L., Fang, L. R., Liu, X. R., and Zhang, Q. S.: The Xinanjiang model, IAHS Publication no. 129, 351–356, 1980.