Looking beyond general metrics for model comparison - lessons from an international model intercomparison study

Tanja de Boer-Euser[1], Laurène Bouaziz[2], Jan De Niel[3], Claudia Brauer[4], Benjamin Dewals[5], Gilles Drogue[6], Fabrizio Fenicia[7], Benjamin Grelier[6], Jiri Nossent[8,9], Fernando Pereira[8], Hubert Savenije[1], Guillaume Thirel[10], and Patrick Willems[3,9]

[1]Water Resources Section, Faculty of Civil Engineering and Geosciences, Delft University of Technology, P.O. Box 5048, NL-2600 GA Delft, The Netherlands
[2]Department Catchment and Urban Hydrology, Deltares, Boussinesqweg 1, 2629 HV Delft, The Netherlands
[3]Hydraulics division, Department of Civil Engineering, KU Leuven, Kasteelpark Arenberg 40, BE-3001 Leuven, Belgium
[4]Hydrology and Quantitative Water Management Group, Wageningen University and Research, P.O. Box 47, 6700 AA Wageningen, The Netherlands
[5]University of Liège, Place du 20-Août 7, 4000 Liège, Belgium
[6]LOTERR- UFR SHS, Université de Lorraine, Île du Saulcy, 57045 Metz Cedex 1, France
[7]Eawag, Überlandstrasse 133, CH-8600 Dübendorf, Switzerland
[8]Flanders Hydraulics Research, Berchemlei 115, B-2140 Antwerp, Belgium
[9]Vrije Universiteit Brussel (VUB), Department of Hydrology and Hydraulic Engineering, Boulevard de la Plaine 2, 1050 Brussels, Belgium
[10]Irstea, Hydrosystems and Bioprocesses Research Unit (HBAN), 1, rue Pierre-Gilles de Gennes, CS 10030, 92761 Antony Cedex, France

Correspondence to: Tanja de Boer-Euser (t.euser@tudelft.nl)

Abstract. International collaboration between research institutes and universities is a promising way to reach consensus on hydrological model development. Although model comparison studies are very valuable for international cooperation, they do often not lead to very clear new insights regarding the relevance of the modelled processes. We hypothesise that this is partly caused by model complexity and the comparison methods used, which focus too much on a good overall performance instead of focusing on a variety of specific events. In this study, we use an approach that focuses on the evaluation of specific events and characteristics. Eight international research groups calibrated their hourly model on the Ourthe catchment in Belgium and carried out a validation in time for the Ourthe catchment and a validation in space for nested and neighbouring catchments. The same protocol was followed for each model and an ensemble of best performing parameter sets was selected. Although the models showed similar performances based on general metrics (i.e. Nash-Sutcliffe Efficiency), clear differences could be observed for specific events. We analysed the hydrographs of these specific events and conducted three types of statistical analyses on the entire time series: cumulative discharges, empirical extreme value distribution of the peak flows and flow duration curves for low flows. The results illustrate the relevance of including a very quick flow reservoir preceding the root zone storage to model peaks during low flows and including a slow reservoir in parallel with the fast reservoir to model the recession for the studied catchments. This intercomparison enhanced the understanding of the hydrological functioning of the catchment in particular for low flows and enabled to identify present knowledge gaps for other parts of the hydrograph. Above all, it helped to evaluate each model against a set of alternative models.

1 Introduction

Large efforts of the hydrological community go into the development of a large variety of hydrological models that are able to filter and reproduce relevant hydrological processes and are preferably applicable in a range of catchments (e.g., Kumar et al., 2013). The outflow from catchments is a combination of different runoff processes, occurring in different parts of the catchment and at different moments throughout the year (e.g., Berghuijs et al., 2014; Nippgen et al., 2015; Penna et al., 2015). Threshold behaviour (e.g., Spence, 2010; McMillan, 2012) and heterogeneity of moisture states (e.g., Detty and McGuire, 2010; Rinderer et al., 2014) create complex systems from which it is difficult to filter the relevant time scales and processes. Overall, hydrological models vary in process representation (conceptual vs. physically based), in the degree of spatial distribution (lumped, semi-distributed and fully distributed) and in the actual runoff process being modelled (e.g., Fenicia et al., 2016). The disadvantage of this abundance of models is that new insights and developments are very scattered and difficult to combine (e.g., Weiler and Beven, 2015). However, a large advantage of having all these different models is their possible use as multiple working hypotheses (e.g., Clark et al., 2011) in a model comparison study to investigate which processes, process representations and spatial distributions are suitable for a set of catchments.

Comparison studies are common in hydrological science and each study has its own twist. While some studies may focus on simulations in a large variety of catchments with widely different characteristics (e.g., Gupta et al., 2014; Duan et al., 2006; Gudmundsson et al., 2012), others focus on a variety of model structures in a limited number of catchments (e.g., Breuer et al., 2009; Holländer et al., 2009; Nicolle et al., 2014; Vansteenkiste et al., 2014; Koch et al., 2016). Many of them rely on international collaboration between several institutes and universities to tackle important open hydrological research questions. Large sample studies enable rigorously testing of alternative model hypotheses and deriving ranges for which model structures are applicable in specific catchments (e.g., Gupta et al., 2014; Thirel et al., 2015a, b). A lesson learned from comparative hydrology in a small number of catchments is the importance of soft data (modeller's system understanding) as well as hard data (data and model), as among others described by Winsemius et al. (2009) and Holländer et al. (2013). In the first and second distributed model intercomparison projects, Reed et al. (2004) and Smith et al. (2012) assessed the performance of lumped versus distributed models and calibrated versus uncalibrated models. They recommended to look in more detail at differences in model structures to increase our understanding of cause and effect. Ceola et al. (2015) pointed out that previous intercomparison studies have contributed little to deriving the causes of performance differences between various model structures. They state that this could be attributed to the complexity and the large differences of model structures, and to the difficulty to link the presence of a model feature to a better or worse performance. Nevertheless, comparison experiments with different model structures should be encouraged to maintain the dialogue between different research groups and agree on adequate modelling concepts (Weiler and Beven, 2015).

During the last decade, model comparison studies have become much easier to carry out due to the large amount of freely available data and the increasing options for sharing data, tools and models. However, a solid model comparison study requires both a clear protocol, and a fair comparison method for the model results (Ceola et al., 2015; Hutton et al.). Protocols can, among others, contain information regarding pre-processing of data, calibration techniques or guidelines for transferring pa-

rameter sets. Very strict protocols do not always line up with the experience of the modeller and the different requirements for each model. Therefore, protocols should be clear, but can never be all-embracing. On the other hand, assessing the performance of the different model realisations should be identical. Standard performance measures (i.e., Nash-Sutcliffe Efficiency, Root Mean Squared Error, Mean Absolute Error) give a general overview, but are unable to point out small differences between model realisations (e.g., Schaefli and Gupta, 2007; Euser et al., 2015). The small differences can possibly be visualised by focussing on specific events and by using more specific performance indicators like hydrological signatures (e.g., Nijzink et al., 2016), or statistics of selected storm events (e.g., Reed et al., 2004; Smith et al., 2012). Using additional data sources for model comparison can further discriminate between model conceptualisations (e.g., Rakovec et al., 2016).

Thus, model comparison studies can be a powerful tool to maintain the scientific dialogue and may contribute to increasing catchment understanding. In this study, different universities and institutes working in and studying the transboundary Meuse basin, in Western Europe, applied their rainfall-runoff model to a set of subcatchments of the Meuse basin using the same meteorological forcing. Modelled fluxes were analysed to gain insight in the behaviour of a set of hydrological models. Our objectives are as follows: (i) set a clear calibration protocol for the participating modellers (ii) propose an evaluation protocol focused on the assessment of specific events instead of general metrics, and discuss the challenges associated to a general and objective approach to model evaluation, (iii) apply the evaluation protocol to various hydrological models proposed by various international institutions, and (iv) relate differences in the simulated hydrographs to model components, and to their associated processes representations.

2 Study areas and data

This study focusses on three subcatchments of the Meuse basin in the Belgian Ardennes: Ourthe, Lesse and Semois catchments and on the two main subcatchments of the Ourthe: Ourthe Orientale (eastern side) and Ourthe Occidentale (western side) catchments (Figure 1 and Table 1).

The Ourthe catchment at Tabreux was selected for calibration because of the limited influence of artificial reservoirs and its meso-scale which enables to focus mainly on hydrology instead of hydraulics. One large reservoir is located in the Ourthe catchment at the confluence of the Ourthe Orientale and Ourthe Occidentale; a short study showed that the influence on the downstream discharge is relatively small (see Section 10 of the Supplement for more explanation). The Ourthe is a typical rain-fed river with a fast response to rainfall due to shallow soils and steep slopes (Driessen et al., 2010) and has a strong seasonal behaviour (Euser et al., 2015).

Many studies have already been carried out in the Ourthe catchment (e.g., Driessen et al., 2010; Rakovec et al., 2012; Euser et al., 2015) because of its significant contribution of flow volumes in the Meuse during floods (de Wit et al., 2007). The catchment of the Ourthe at Tabreux has a total area of 1607 km^2 with an elevation ranging between 107 and 663 meters. Mean annual precipitation and potential evaporation are 1000 $\mathrm{mm\ y}^{-1}$ and 730 $\mathrm{mm\ y}^{-1}$ respectively. The main land use is agriculture (28$\%$ crops and 28$\%$ pasture), followed by forestry (46$\%$) and only 6$\%$ of the catchment has an urban cover (CORINE Land use map, European Environment Agency, 2006).

The neighbouring Lesse and Semois catchments and the nested Ourthe Occidentale and Ourthe Orientale catchments were selected for validation. The Lesse and the Semois catchments are about 25% smaller than the Ourthe catchment, and their forest cover is slightly higher than in the Ourthe catchment (Table 1). Annual mean precipitation is similar in the Lesse catchment while it is 25% higher in the Semois catchment. The upstream parts of the Semois catchment and the nested catchments within the Ourthe (Occidentale and Orientale) are relatively flat, while the Lesse catchment and downstream parts of the Ourthe and Semois catchments have steeper slopes. The hourly specific discharge of the Ourthe at Tabreux is most similar to that of the Lesse (on average 3% lower than Ourthe, with R^2 of 0.91) and least similar to that of the Semois (on average 50% higher than Ourthe, with R^2 of 0.78). The hourly specific discharges in the Ourthe Orientale and Occidentale are on average 7% and 5% higher than in the Ourthe at Tabreux, with R^2-values of 0.92 and 0.88 respectively.

Data preparation involved interpolation of hourly precipitation station data based on Thiessen polygons. The station data is collected and made available for this study by the Service Public de Wallonie[1]. Daily minimum and maximum temperatures from the freely available gridded E-OBS dataset (0.25° x 0.25° resolution; Haylock et al., 2008) were disaggregated to hourly values using radiation data at Maastricht (Royal Netherlands Meteorological Institute[2]) and a sine function. Daily potential evaporation was derived with the Hargreaves formula (Hargreaves and Samani, 1985) and disaggregated to hourly values using a sine function during the day and no evaporation at night. Precipitation and temperature data were available for the period from 1 January 2000 to 31 December 2010. The data (distributed, semi-distributed or lumped) was made available to the researchers through an FTP server. Figures of hourly observed discharge, precipitation, potential evaporation and temperature for each catchment can be found in Section 1 of the supplement.

3 Methods

This comparison study roughly consists of three elements: the modelling protocol followed by each participant, the models used by each participant and the tools used for comparing the individual model results.

3.1 Modelling protocol

Eight international research groups participated in this model comparison study using one or several rainfall-runoff models. A total of eleven models were used, consisting of seven independent models and four models from the SUPERFLEX framework (Fenicia et al., 2011). A modelling protocol was prescribed to enable a comparison of the results. The protocol for the modelling involved a split-sample calibration and validation for pre-defined periods using a common dataset (Klemeš, 1986) for the Ourthe catchment. The validation consisted of a blind validation in time (same catchment, but a different period) for the Ourthe catchment and a blind validation in space (same period, but different catchments) for the nested Ourthe Orientale and Ourthe

[1]Service Publique de Wallonie, Direction générale opérationnelle de la Mobilité et des Voies hydrauliques, Département des Etudes et de l'Appui à la Gestion, Direction de la Gestion hydrologique intégrée, Boulevard du Nord 8 - 5000 Namur

[2]http://www.knmi.nl/nederland-nu/klimatologie/uurgegevens, visited 14-12-2012

Occidentale catchments and for the neighbouring Lesse and Semois catchments. Blind validation implies that only time series of forcing data (and no discharge observations) were given to the participants.

Calibration was carried out for the Ourthe at Tabreux for the period: 1 January 2004 to 31 December 2007, using 2003 as a spin-up year. Nash-Sutcliffe Efficiency (E_{NSE}; Nash and Sutcliffe, 1970) and E_{NSE} of the log of the flows (E_{NSElog}; Collischonn et al., 2008) were used as objective functions for calibration. E_{NSE} was chosen as an objective function for calibration because it is a common metric in hydrology to assess model performance with regard to high flows. E_{NSElog} was chosen as a second objective function to take low-flow performance into account as well. Participants were free to use a calibration method of their choice to estimate parameter values as long as they used the prescribed objective functions. Although it makes the comparison of the model results less straightforward, a free calibration method does account for the experience a modeller has with a specific model. Using the Pareto front between E_{NSE} and E_{NSElog}, the best 20 realisations were selected for each model to account for a range in model realisations (2a).

Validation in time was carried out for the Ourthe at Tabreux for the period from 1 January 2001 to 31 December 2003, using 2000 as a spin-up year. This period was selected for validation as it includes some relevant hydrological events such as the drought in the summer of 2003 and high flows during the winters. The validation period has relatively high flows compared to the calibration period. An additional validation in time was carried out for the period 2008-2010 for the Ourthe at Tabreux, using the calibration period as a spin-up. For the latter period, participants only received forcing data (2b).

Validation in space was carried out for two nested catchments of the Ourthe: Ourthe Orientale (at Mabompré) and Ourthe Occidentale (at Ortho) for the period from 1 January 2001 to 31 December 2010, using 2000 as a spin-up year. Additionally, the derived parameter sets for the Ourthe at Tabreux were used to model the neighbouring Lesse (at Gendron) and Semois (at Membre) catchments for the same period. Only the forcing data was provided to the participants for this validation in space (Figure 2b).

3.2 Descriptions of models

Each modelling group provided results as described above. A variety of models was used, including lumped, semi-distributed and fully distributed models. All models are conceptual, but their degree of complexity varies and they are used by institutes or universities working in the Meuse basin. Figure 3 depicts the main fluxes and storages of the applied models. Table 2 shows for each model the used forcing, the calibration method, and whether parameters were regionalised. Below a short description is given for each model: the term 'root zone storage' is used for the reservoir from which transpiration is modelled. Further, the term 'very quick runoff' is used for a faster process than 'fast runoff', these terms can be compared with 'overland flow' and 'interflow' respectively. The response times for the very quick runoff, the fast runoff and the groundwater runoff are for most models in the order of 1 day, 8 days and 80 days respectively.

GR4H-CemaNeige (Mathevet, 2005) is a combination of the CemaNeige snow module (Valéry et al., 2014) and an hourly version of GR4J (Perrin et al., 2003). GR4H is an empirical four parameters hourly model with a root zone storage and two routing routines: one for very quick and one for fast runoff. The division of water between the two routines is fixed at

a 0.1-0.9 ratio; both reservoirs interact with the groundwater. Interception is taken into account by substracting potential evaporation from precipitation to obtain net precipitation. GR4H was developed for high flows rather than for low flows, as low flows are rarely studied at an hourly time step.

PREvision et Simulations pour l'Annonce et la Gestion des Etiages Sévères, PRESAGES (Lang et al., 2006), is a daily tool for low-flow forecasting and evaluation, and it was slightly modified to run on an hourly time scale. It is derived from GR4H with differences being no incorporation of snow and a separated groundwater reservoir connected in series with the fast runoff reservoir. There is no longer interaction between the very quick runoff reservoir and the groundwater.

Wageningen Lowland Runoff Simulator, WALRUS (Brauer et al., 2014a, b), is a lumped conceptual model for application in lowland areas with shallow groundwater tables. The model consists of three reservoirs: a combined root zone and groundwater reservoir, a combined very quick and fast runoff reservoir and a surface water reservoir. Snow accumulation and melt are simulated in a pre-processing step and interception was not taken into account. Note that the Ourthe catchment is not located in lowlands; we included WALRUS in the comparison to investigate where a model designed for lowlands would succeed and fail in a hilly catchment.

M2-M5 models of the SUPERFLEX framework (Fenicia et al., 2011) are four lumped conceptual models with an increasing degree of complexity. M2 consists of a root zone storage and a non-linear fast runoff reservoir. M3-M5 extend the M2 conceptualisation by adding a lag function (M3-M5), a snow routine (M4-M5) and a groundwater reservoir (M5). Interception is not taken into account by all four models.

NAM is an adapted version of the hydrological model which is coupled to MIKE11 (Nielsen and Hansen, 1973). It consists of a snow module, interception reservoir, root zone storage and a groundwater reservoir; the latter are configured in parallel. Fast and very quick runoff are generated from the interception reservoir, but depend on the saturation of the root zone storage.

FLEX-Topo (Savenije, 2010; Euser et al., 2015) is a conceptual model with three parallel model elements that represent the three dominant hydrological response units (HRUs) of this area: wetlands, hillslopes and agricultural fields. The model elements only interact with each other via the groundwater reservoir and the stream network. Each HRU consists of a snow module, interception reservoir, root zone storage and fast runoff reservoir. The agricultural area has an additional very quick flow reservoir.

VHM (Willems, 2014; Willems et al., 2014) involves a step-wise and data-based procedure to identify a parsimonious lumped conceptual model. For the Ourthe catchment, a model was identified which consists of a root zone storage and three runoff components: very quick runoff, fast runoff and groundwater runoff, which are configured in parallel; interception was not taken into account.

wflow_hbv is a completely distributed version of the conceptual HBV model (Lindström et al., 1997) in the wflow framework[3] with a kinematic wave as routing instead of the original triangular routing function. The model has an interception reservoir, snow module, root zone storage, fast runoff reservoir and a groundwater reservoir. The parameter values are uniform for the entire catchment, except for maximum interception capacity, which is related to land use.

3.3 Evaluation methods

The results of the eleven models and five catchments were compared in multiple ways. First, the scores obtained for the objective functions (E_{NSE} and E_{NSElog}) were compared. This step enabled to determine the overall calibration performance of the models. We expected that this analysis would not reveal much difference; so, two additional analyses were carried out: a statistical analysis and a hydrograph comparison for specific periods (Figure 2c). These additional analyses focused on aspects that were not specifically taken into account during the calibration procedure, as to investigate the full range of a model's capabilities.

Three types of statistical analyses and comparisons of simulation results and observations were conducted: cumulative discharges, empirical extreme value distribution of the peak flows and flow duration curves for low flows. The cumulative discharges were plotted for the entire modelled period and used to investigate the overall water balance. The empirical extreme value distributions were constructed from independent peak discharges, following Willems (2009); the return period was calculated as the mean time interval between the exceedance of given runoff amounts. This analysis of peak flows was carried out to investigate if models were able to simulate the full range of peak discharges observed in the catchments. In addition, the empirical extreme value distribution can provide information on the usefulness of models for flood (frequency) studies and extrapolations to more extreme events, by assessing the shape of the distribution, as well as the tendency of the difference between higher modelled and observed peak flows. The flow duration curves were constructed for the lowest 20% of the discharges. Low flows are important in the Meuse basin, especially from a user's perspective; comparing observed and simulated flow duration curves helped to assess how well models were able to reproduce low flows.

Finally, specific periods were selected to compare modelled and observed hydrographs. By looking at specific events, more detailed differences can be observed between models. Four different periods outside the calibration period were selected for this analysis: a summer period, a transition from low to high flows and two winter periods. The summer of 2008 was selected, because many high intensity precipitation events occurred during this period; during the summers in the calibration period, these events did not occur very frequently, making this a benchmark period. The autumn of 2003 was selected as a low to high flow transition period, as 2003 was a very dry summer, so problems with re-saturation were likely to be largest during this year. The two analysed winter periods were 2002-2003 and 2010. In the studied catchments, winter runoff can consist of rainfall (in 2002-2003) in case of higher temperatures or of snow melt (2010) in case of lower temperatures. By comparing these two winter periods, the model's ability to reproduce both conditions was investigated. It should be noted that not all models contain a snow routine, thus the winter of 2010 was also used to investigate how important a snow routine is for simulating discharges.

[3] http://wflow.readthedocs.io/en/latest/

The statistical analyses and specific periods of the hydrographs were first compared visually, additionally the relative error between a set of observed and modelled signatures were assessed. The modelled signature values were calculated based on the best performing model realisation and are shown in the specific plots. The best performing model realisation was selected for each signature to reflect the best achievable model performance and to minimize the effect of the different band widths in model realisations between the different models.

4 Results

The analyses of metrics, statistics and hydrographs for the eleven model structures, run for the five catchments for the period 1 January 2001 until 31 December 2010, showed different model performances. All analysed figures can be found in the supplement (Sections 3 to 9); overall they confirm that all models perform well (maximum E_{NSE} varying between 0.85 and 0.91 and maximum E_{NSElog} between 0.85 and 0.93 for the entire modelled period for Tabreux; supplement, Section 2). In all figures, the results for the 20 selected realisations per model are shown and their band width is closely related to the calibration method applied.

It was found that even very simple lumped models (M2) could perform as well as very complex (semi-) distributed models (FLEX-Topo and wflow_hbv) under wet conditions. Most models had higher performances during the validation period than during calibration and blind validation periods in terms of E_{NSE} and E_{NSElog}, probably due to the wetter conditions during the validation period. The hydrographs and the cumulated discharges over the entire period showed that most models slightly underestimated observed flows, except for FLEX-Topo. A number of relevant differences between models and catchments are highlighted below. For each section, we explain our findings by showing the results for the most illustrative catchment.

4.1 Internal averaging within the Ourthe catchment

Yearly simulated and observed flows in the Ourthe and its two nested catchments (Ourthe Orientale and Ourthe Occidentale) possibly show the effect of internal averaging, as depicted in Figure 4. While discharged volumes are underestimated by all models in the Ourthe Occidentale, they are overestimated by most models in the Ourthe Orientale and this seems to average out for the Ourthe at Tabreux. Topography, land cover and geology are comparable for the Ourthe Orientale and Ourthe Occidentale catchments, with the Ourthe Orientale catchment being a little steeper and having slightly more forest cover. However, the topography of both catchments differs from that of the Ourthe catchment at Tabreux, indicating that parameters may not be directly regionalised.

Another difference between the Ourthe catchment and its subcatchments is the volume of precipitation and runoff; the Ourthe Orientale catchment receives more precipitation and produces less runoff than the Ourthe Occidentale catchment. Previous studies (e.g., Kleidon and Heimann, 1998; Gao et al., 2014; de Boer-Euser et al., 2016) showed a link between climate (i.e., precipitation and evaporation volumes) and root zone storage capacity. Following their argument, the root zone storage capacity should indeed be larger for the Ourthe Orientale catchment and smaller for the Ourthe Occidentale catchment compared to the Ourthe catchment to meet the evaporative demand of the Ourthe Orientale catchment. Using the root zone storage capacity of

the Ourthe catchment for the Ourthe Orientale catchment could lead to too high modelled discharges, using it for the Ourthe Occidentale catchment could lead to too low modelled discharges. On the other hand, it is also possible that precipitation is underestimated in the Ourthe catchment, as all models are underestimating the runoff volume for the Ourthe Occidentale.

4.2 Modelling flood peaks

Figure 5 shows the independent peak flows versus the empirical return period for both observed and modelled discharges for the Lesse. The signatures used for the flood peaks are the slopes of the upper and lower part of the distribution; the break point between the upper and lower slope is set at a return period of 1.5 years. Out of all studied catchments, high flows extremes are the most difficult to capture by the models for the Lesse catchment: most models underestimate the flood peaks for this catchment, while they can capture them relatively well for the other catchments. For the Lesse catchment, all models are able to model the lower peaks, but they underestimate the higher peaks. Although these higher peaks are difficult to simulate, Figure 5 shows that it is not impossible as at least one model (GR4H-CemaNeige) is able to reproduce the steeper increase in peak flow for high return periods and capture the highest peaks. The other models have a varying behaviour: some capture a part of the higher slope, while others have a poor performance for this signature. The fact that some models are able to capture the highest peaks reduces the probability that data errors and handling are the cause of underestimating the highest peaks in these catchments, as one might have concluded if all models had failed.

What is striking about the results for all catchments is that the simplest models, consisting of only two reservoirs like M2, perform as good or sometimes even better than more complex models. This indicates that during these very wet events, in the entire catchment fast flow paths were activated and all water is drained towards the stream. With a parsimonious model structure it is relatively easy to calibrate the limited number of parameters to fit the peak flows. When a model is more complex, including a splitting component between the fast runoff and another runoff reservoir, it might be more difficult to calibrate the model and peak flows might be over or underestimated. Model M5 also contains such a splitter of about 20% going to the groundwater reservoir, but this does not seem to be high enough to influence the performance negatively. The importance of a groundwater and an interception reservoir during these events is limited as they represent only a very limited fraction of the peak flows, as can be seen from the difference between M4 and M5.

4.3 Modelling low flows

Low flows were analysed by plotting the lowest 20% of the observed and modelled flow duration curves. The slope and mean of this part of the flow duration curve were used as signatures. Discharges during the summer recession periods are generally low (ranging between 0.004 $\mathrm{mm\ h^{-1}}$ and 0.015 $\mathrm{mm\ h^{-1}}$ for the lowest 20%) compared to the average discharge (0.05 $\mathrm{mm\ h^{-1}}$). The influence of a groundwater reservoir on the modelled discharge is significant as the flow duration curves illustrate, for example for the Ourthe at Tabreux (Figure 6). Adding a groundwater reservoir improves the simulation of the low flows, as illustrated by the difference in performance between models M4 and M5. The only difference between the two models is the presence of a groundwater reservoir and where M4 underestimates low flows, they are properly simulated by M5. This indicates that water is stored during the high flow period in winter and released again during the low flow period in summer.

The configuration of the groundwater reservoir is also important: model structures with a groundwater reservoir parallel to the fast runoff reservoir (M5, NAM, FLEX-Topo, VHM) generally give the best results. Model structures without a groundwater reservoir (M2-M4) underestimate the low flows, while models with a serial or interactive groundwater reservoir (GR4H, PRESAGES, WALRUS, wflow_hbv) overestimate the low flows or model the recessions too steep. On one hand, this indicates the importance of preferential recharge in the catchment, on the other hand it indicates the existence of runoff processes with different time scales. With a parallel groundwater reservoir, the time scales for runoff generation are decoupled, with a serial or interactive groundwater reservoir they are connected. These differences in results between models indicate that the processes related to fast and slow runoff generation occur relatively independent of each other in the studied catchments. The described results are clearly visible in the flow duration curves of the Ourthe at Tabreux and the Lesse at Gendron; for the other catchments the same patterns can be found, but slightly shifted upwards or downwards.

4.4 The effect of a very quick runoff component

In the summer of 2008, precipitation intensities were higher than in other years, although total precipitation amounts were similar. This resulted in a flashy response of summer peaks which is clearly shown for the catchment of the Ourthe at Tabreux in Figure 7. The performance for this peaky response was assessed with signatures for the average slope of the declining limbs and the total discharged volume. The antecedent root zone storages before the events can be expected to be low, due to high transpiration rates in summer. While most models are not able to capture the summer peaks, VHM and FLEX-Topo are able to simulate the dynamics well, although FLEX-Topo overestimates the summer peaks. As shown in Figure 3, VHM and FLEX-Topo are the only models that contain a very quick runoff component preceding the root zone storage and independent of the root zone storage. Hence, it illustrates that this component is essential for simulating short, intense summer events, which are likely to cause infiltration excess overland flow (i.e., precipitation intensity being higher than infiltration capacity of the soil). Under dry conditions the infiltration capacity of the soil is assumed to be disconnected to the saturation of the soil. Therefore, the very quick flow component should be independent of the root zone storage and should precede it; otherwise these short intense summer rainfall events are stored in the soil instead of discharging directly to the river. Models with a very quick runoff component, which is affected by the root zone storage (WALRUS and NAM) and models with a very quick runoff component following the root zone storage (GR4H, wflow_hbv, PRESAGES) do perform better than models where the very quick runoff component is entirely lacking (M2 to M5), but they miss the sharpness of the response, due to damping of the generated peaks. These findings are consistent for all studied catchments.

4.5 Transition from low to high flows

The largest differences in model results between the modelled catchments occur for the transition from low to high flows. The signatures used for this period are the ratio between the first and the highest peak of this specific period and the total discharged volume. For the transition period in 2003, runoff is overestimated for all models in the Ourthe Orientale (Figure 8), while only to a minor extent in the other catchments. In the Lesse catchment, the performance during this transition period is the highest from the four selected periods for almost all models. In addition, the performance in the Lesse catchment is also higher than

that for the calibrated Ourthe catchment for almost all models. The variability in performance between models and between subcatchments for this event prevented pinpointing model components that explain the differences in performance.

Although all models overestimate the discharge of the Ourthe Orientale (Figure 8), their response is different. They especially differ in simulating the two highest peaks: PRESAGES and WALRUS simulate the first one relatively well, but underestimate the second. The other models overestimate the first peak and vary in how they simulate the second one. As the transition period is controlled by the modelled rate of infiltration and its dependence to soil saturation state, one reason could be explained by differences in modelled evaporation in the antecedent period; however, the model with the lowest evaporation (PRESAGES) is not the model with the highest overestimation of discharge. FLEX-Topo strongly overestimates the discharges; this can partly be caused by the root zone storage capacity. This model has a climate derived root zone storage capacity (de Boer-Euser et al., 2016), which is significantly higher for the Ourthe Orientale catchment than for the Ourthe catchment (see also section 'Internal averaging within the Ourthe catchment'). The difficulty the models encounter to model this transition may be linked to the hysteretic behaviour in dry-to-wet transition periods (autumn) and in wet-to-dry transition periods (spring). This finding illustrates remaining knowledge gaps with regard to the rewetting of catchments after dry periods, which seems to work differently than what is currently assumed in our models.

4.6 The effect of a snow routine

The models that include a snow routine (GR4H, WALRUS, M4, M5, NAM, FLEX-Topo and wflow_hbv) did not perform significantly better than the others during snow events. Figure 9 shows the winter of 2010 for the Semois catchment: this is the catchment and period with the largest differences between models with and without snow. For this period the timing and the discharged volume of the snow melt peak at the beginning of March 2010 were used as signatures. The models with a snow routine can reproduce the snow melt peak slightly better. The differences are, however, rather small. Although it could be expected that having a snow module improves the performance during a snow event, it was not clearly found in this study. Possible causes for the limited effect of the snow module are that snow cover mainly occurs for short periods of time and that the influence of snowmelt on the discharge is limited, but that some snow does occur every winter. In addition, the discharges corresponding to snow melt periods are similar in magnitude to those originating from liquid precipitation. These aspects, in combination with the use of general metrics for calibration, lead to the possibility that (small) influences of snow on the discharge are compensated by other parameters when a model does not have a snow module.

4.7 Results for all catchments

Figures 4 to 9 9 show plots for specific catchments; Figure 10 additionally shows the relative error for eight signatures, calculated for the periods shown in the plots for all models and all catchments. A red symbol indicates that the modelled value is too low, a blue symbol that the modelled value is too high; darker colours indicate larger errors and light or white colours indicate that the modelled signature is very close to the observed signature. The figure shows that the cumulative flows can be reproduced well by all models for all catchments. The higher slope of the peak distribution is difficult to reproduce by most models for all catchments. This contrast between average performance and modelling peak flows was found by Donnelly et al.

(2016) as well. Furthermore, it can be seen that in case of larger errors, the signatures are generally underestimated and not overestimated, except for the slope of the lower flow duration curve and the slope of the falling limbs in case of FLEX-Topo.

5 Discussion

5.1 Findings about the Meuse basin

The results of this study show first of all differences and similarities between catchments and models. In addition, the analysis of model behaviour under relatively dry conditions (Figures 6 and 7) shows which model configurations are more suitable for these catchments than others: the conceptualisations of the very quick runoff component and the groundwater reservoir. The very quick runoff component is necessary and should precede infiltration into the root zone storage and not be affected by it. The groundwater reservoir is necessary as well and can best be implemented in parallel to fast runoff generation. The effect of a very fast runoff component is directly visible in the hydrographs and consistent for all catchments. The effect of a (different conceptualisation of the) groundwater reservoir is best visible in the lower parts of the flow duration curve and the strength of the effect varies per catchment. The results thus indicate that both components are important for a conceptual model of these catchments, especially when the model is aimed to be applicable for analysis of peak and low flows. High flows are best predicted when the root zone storage directly flows to the fast runoff store, with only limited splitting towards other reservoirs.

These findings show that in summer during intense rainfall events, the infiltration capacity of the soil is exceeded by the rainfall intensity and leads to a very rapid response, for at least part of the catchment. The results regarding the recession illustrate the preferential pathways that exist between the unsaturated soil and the groundwater and their importance during low-flow periods. Regarding peak flows, we found that during these events, fast flowpaths in the entire catchment are activated and that all water is rapidly drained to the river. Very simple model structures (which may contain only two reservoirs) are then sufficient to model peak flows well in the Meuse. However, these simple model structures are not able to capture the full range of regimes, especially during low flows. These results highlight how difficult it is to develop an all-round model structure which is able to capture all different regimes that occur in the studied basins.

We were able to generalise the importance of two components to improve low flow predictions and we pointed out an important component to model peak flows better; other results were too variable in space and between model structures and could therefore not be linked to specific model structural components. The comparison consists of eleven model structures, each with specific details. Therefore, other differences and similarities in the modelled discharge could not be easily related to differences and similarities in model conceptualisations. This highlights remaining knowledge gaps with regard to important processes that occur during the transition from low to high flows (Figure 10), which are not well understood and implemented in our models. In these periods, we believe that vegetation plays a crucial role, influencing infiltration and evaporation, but these dynamics seem to be lacking in our models (cf., Seibert et al., 2016).

Another general result of this comparison study is the higher performance of the normal (non-blind) validation period compared to the calibration and blind validation periods. Although performance generally decreases during validation, some studies show an increase in performance (e.g., Hrachowitz et al., 2014) for the validation period. Often, this indicates that

hydro-meteorological conditions in the validation period were easier to model. The same holds here: the validation period is the wettest period, and most conceptual models yield the best performance under wet conditions. The higher performance during validation and the hydro-meteorological differences between the calibration and validation periods show that the models are transferable in time and space within our testing periods.

5.2 Benefits of an intercomparison study

Intercomparison studies can provide a more detailed overview of a model's potential than single model studies. In that sense, they enable individual modellers to reflect on familiar model structures, through comparative identification of lacking or relevant components. In a single model study, a poor performance may be easily blamed on data shortcomings or model structural errors. In an intercomparison study, it is less likely that the poor performance of a certain model is due to data errors if there is at least one model that performs well when forced with the same input data.

Comparing model performance following a fixed protocol and linking results to model components provides a strong basis for improved model development for all modellers. Dominant runoff processes and their model representations can be derived and added to the various model structures. New experiments and hypotheses on catchment understanding can be formulated and tested by all modellers in their specific model. This may ultimately lead to the development of all-round model structures that are applicable for different hydrological regimes. As such, even though it is a time-consuming process, it is worthwhile to increase our understanding, to get to know other model structures and to stimulate the dialogue between different institutes and universities.

Preliminary results of the model comparison study were sent to the modellers with the question to evaluate the model results and speculate on how their model structure could be improved. One of the responses was that processes were not (or only recently) specifically included in the model (e.g., fast runoff caused by infiltration excess overland flow, snow), because they were not necessary for earlier applications. In addition, the prescribed calibration objectives and lumped precipitation forcing used for most models were brought up as reasons for inferior model performance: the method used to calibrate VHM for this experiment differed from the normal calibration method applied (Willems, 2009). This may have played an important role in the underestimations of peak flows. As explained by Willems (2009) and others, the E_{NSE} objective function applied to the flows at all time steps gives weight to the high flows but less to extremely high flows because they are typically of shorter duration. When comparing the automatic calibration applied for this study to the manual calibration normally applied, focussing on the hydrological extremes (Willems, 2014), improved results for peak flows are obtained for the manual calibration, as illustrated in Figure 11.

This intercomparison study shows that the assessed models have different strengths in capturing specific characteristics of the runoff response. Single models may have been developed to perform better on a specific aspect at the expense of another one, as explained by Duan et al. (2007). Applying a multimodel ensemble instead of relying on a single model outcome provides more information on model structure uncertainty. This helps hydrologists to better understand the catchment functioning and improve uncertainty estimations. In an operational context, multimodel ensemble are useful to make more informed decisions.

5.3 Comparison of models

The choice of calibration method was left to the individual modellers, with the only constraint that E_{NSE} and E_{NSElog} had to be used as objective functions. This resulted in some modellers using a search algorithm, while others applied uniform sampling of the parameter space. In addition, the (width of the) parameter space before sampling varied per model. This freedom in calibration probably has affected the results; on the other hand, we considered that the calibration algorithm chosen is strongly linked to the model and modeller's experience. As some methods used a search algorithm while others applied uniform sampling, the range of the model realisations varied considerably between models: for some models the 20 realisations were almost identical, while for others there were large differences. This added an additional source of variability to the comparison, but this variability did not alter the conclusions.

After the calibration on E_{NSE} and E_{NSElog}, the models were compared focussing on specific periods and statistics. Although the general metrics showed a high performance for all models, (large) differences are observed when focusing on the specific periods or statistics. This is especially true when modelling events under drier conditions, which are the conditions when different model behaviours were most visible. A model evaluation based on visual inspection of the hydrographs during specific events may sound subjective, but because it focuses on very specific events, the human eye easily detects patterns that reflect model performance. Combining the visual inspection with the relative error for specific signatures enabled us to further identify similarities between models and catchments, as shown in Figure 10. This emphasises again the importance of a broad but specific model evaluation, especially for a model comparison study.

The majority of the models considered in this study is lumped and used lumped forcing. Only two models, FLEX-Topo and wflow_hbv, used the semi and completely distributed forcing respectively. The distribution of the forcing and the model did not seem to have a significant impact on model performance compared to the other models. The differences in model structure affected model performance more than the differences in distribution of forcing. This is in line with earlier studies (e.g., Euser et al., 2015; Vansteenkiste et al., 2014), which showed that distribution of forcing data has a smaller effect on performance than the selection of model structure.

The varying degree of experience of the modellers with both their model and calibration technique and with the studied areas is likely to influence the reproducibility of this experiment. However, the similar forcing data used and the defined protocol enabled to reduce the degrees of freedom of the modellers and enabled the comparison of the results.

This study is a large step forward in the international cooperation between universities and institutes working in the Meuse basin. Sharing data and model results in this set-up has never been realised before, but it is fundamental to open up the dialogue and advance hydrological understanding of the studied catchment in a more coherent way.

5.4 Future intercomparison studies

We think that international model intercomparison studies are very important and are definitely valuable in future research programs. First of all, they are a good opportunity to increase cooperation and discussion between different institutes. In addition, it is a good means for young scientists to get to know the models used in neighbouring universities and institutes.

To increase the possibility to draw strong conclusions about the hydrological functioning of a catchment, a different set-up may be useful. If all modellers would select a very strong element of their model, this could be incorporated in all the other models. By doing this, in a controlled sequence and actually creating a virtual laboratory, probably more insight could be obtained regarding hydrological functioning and suitable model conceptualisations. In addition, more independent data sources, besides discharge, would probably strongly increase the possibility to obtain insight about the hydrological functioning of the studied catchments.

6 Conclusions

For this study we compared eleven models for five subcatchments of the Meuse basin. All models were calibrated on the Ourthe at Tabreux; they were then evaluated for two different periods and five different catchments. E_{NSE} values for all models and all catchments were comparable, with in some cases even higher performances during the validation period. Although E_{NSE} values were comparable, a more detailed analysis, focussing on specific events through hydrograph inspection and statistics, revealed clear differences between the models, especially for drier conditions. We found that a very quick runoff component preceding and not affected by the unsaturated store was relevant to model the hydrological response after short and intense summer precipitation events. This conceptualisation ensures that water is not stored in the soil but quickly flows to the river. Also a groundwater reservoir implemented in parallel to the fast runoff generation, representing preferential pathways for groundwater infiltration, seemed necessary to model the recession best. For high flows, we found that very simple and lumped model structures with only an unsaturated store and a fast runoff component performed better than complex models. This highlights the difficulty to develop model structures which are able to cope with different hydrological regimes (high and low flows). The presence of knowledge gaps was further revealed by the inability of our models to predict the transition from a low-flow period to high flows well, probably related to the lack of vegetation dynamics included in our models. Thus, from this study we can conclude that often more detailed analyses are required to relate differences in the hydrograph to model structure components. A model intercomparison study is a valuable approach to draw conclusions about hydrological functioning of a system, and most of all, it is a great opportunity to reflect on your model structure by comparing it with other models. This leads to the question: "What is my model doing well in comparison to other models and why?". This points out the model structure components to keep and in the end, focussing on this question will improve our hydrological understanding.

Acknowledgements. The authors would like to thank the Service Public de Wallonie, Direction générale opérationnelle de la Mobilité et des Voies hydrauliques, Département des Etudes et de l'Appui à la Gestion, Direction de la Gestion hydrologique intégrée (Bld du Nord 8-5000 Namur, Belgium) for providing the precipitation and discharge data. We would like to thank Bernhard Becker for organising the Meuse symposia that have led to this fruitful cooperation and thank Frederiek Sperna Weiland for the valuable discussions and her contribution in data preparation. Further we would like to thank Michel Pirotton for his review and Pierre Archambeau, Sébastien Erpicum, Michel Pirotton for the analysis of the impact of the Nisramont dam presented in the supplement. We would like to thank Rohini Kumar and two anonymous

reviewers for their valuable comments which helped us to improve the manuscript.

References

Berghuijs, W. R., Sivapalan, M., Woods, R. A., and Savenije, H. H. G.: Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales, Water Resour. Res., 50, 5638–5661, doi:10.1002/2014WR015692, 2014.

Brauer, C. C., Teuling, A. J., Torfs, P. J. J. F., and Uijlenhoet, R.: The Wageningen Lowland Runoff Simulator (WALRUS): a lumped rainfall-runoff model for catchments with shallow groundwater, Geosci. Model Dev., 7, 2313–2332, doi:10.5194/gmd-7-2313-2014, 2014a.

Brauer, C. C., Torfs, P. J. J. F., Teuling, A. J., and Uijlenhoet, R.: The Wageningen Lowland Runoff Simulator (WALRUS): application to the Hupsel Brook catchment and the Cabauw polder, Hydrol. Earth Syst. Sci., 18, 4007–4028, doi:10.5194/hess-18-4007-2014, 2014b.

Breuer, L., Huisman, J., Willems, P., Bormann, H., Bronstert, A., Croke, B., Frede, H.-G., Gräff, T., Hubrechts, L., Jakeman, A., Kite, G., Lanini, J., Leavesley, G., Lettenmaier, D., Lindström, G., Seibert, J., Sivapalan, M., and Viney, N.: Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use, Adv Water Resour, 32, 129–146, doi:10.1016/j.advwatres.2008.10.003, 2009.

Ceola, S., Arheimer, B., Baratti, E., Blöschl, G., Capell, R., Castellarin, A., Freer, J., Han, D., Hrachowitz, M., Hundecha, Y., Hutton, C., Lindström, G., Montanari, A., Nijzink, R., Parajka, J., Toth, E., Viglione, A., and Wagener, T.: Virtual laboratories: new opportunities for collaborative water science, Hydrol. Earth Syst. Sci., 19, 2101–2117, doi:10.5194/hess-19-2101-2015, 2015.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resour. Res., 47, W09 301, doi:10.1029/2010WR009827, 2011.

Collischonn, B., Collischonn, W., and Morelli Tucci, C.: Daily hydrological modeling in the Amazon basin using TRMM rainfall estimates, J Hydrol., 360, 207–216, doi:10.1016/j.jhydrol.2008.07.032, 2008.

de Boer-Euser, T., McMillan, H. K., Hrachowitz, M., Winsemius, H. C., and Savenije, H. H. G.: Influence of soil and climate on root zone storage capacity, Water Resour. Res., 52, 2009–2024, doi:10.1002/2015WR018115, 2016.

de Wit, M., Peeters, H., Gastaud, P., Dewil, P., Maeghe, K., and Baumgart, J.: Floods in the Meuse basin: Event descriptions and an international view on ongoing measures, International Journal of River Basin Management, 5, 279–292, doi:10.1080/15715124.2007.9635327, 2007.

Detty, J. M. and McGuire, K. J.: Topographic controls on shallow groundwater dynamics: implications of hydrologic connectivity between hillslopes and riparian zones in a till mantled catchment, Hydrol. Process., 24, 2222–2236, doi:10.1002/hyp.7656, 2010.

Donnelly, C., Andersson, J. C., and Arheimer, B.: Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe, Hydrological Sciences Journal, 61, 255–273, 2016.

Driessen, T. L. A., Hurkmans, R. T. W. L., Terink, W., Hazenberg, P., Torfs, P. J. J. F., and Uijlenhoet, R.: The hydrological response of the Ourthe catchment to climate change as modelled by the HBV model, Hydrol. Earth Syst. Sci., 14, 651–665, doi:10.5194/hess-14-651-2010, 2010.

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H., Gusev, Y., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E.: Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops, J Hydrol., 320, 3–17, doi:10.1016/j.jhydrol.2005.07.031, 2006.

Duan, Q., Ajami, N., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Advances in Water Resources, 30, 1371–1386, doi:10.1016/j.advwatres.2006.11.014, 2007.

Euser, T., Hrachowitz, M., Winsemius, H., and Savenije, H.: The effect of forcing and landscape distribution on performance and consistency of model structures, Hydrol. Process., 29, 3727–3743, doi:DOI: 10.1002/hyp.10445, 2015.

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, W11 510, doi:10.1029/2010WR010174, 2011.

Fenicia, F., Kavetski, D., Savenije, H. H. G., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, Water Resources Research, 52, 954–989, doi:10.1002/2015WR017398, 2016.

Gao, H., Hrachowitz, M., Schymanski, S. J., Fenicia, F., Sriwongsitanon, N., and Savenije, H. H. G.: Climate controls how ecosystems size the root zone storage capacity at catchment scale: Root zone storage capacity in catchments, Geophys. Res. Lett., 41, 7916–7923, doi:10.1002/2014GL061668, 2014.

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., and Savenije, H. H. G.: Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration, Hydrology and Earth System Sciences, 18, 4839–4859, 2014.

Gudmundsson, L., Tallaksen, L. M., Stahl, K., Clark, D. B., Dumont, E., Hagemann, S., Bertrand, N., Gerten, D., Heinke, J., Hanasaki, N., Voss, F., and Koirala, S.: Comparing Large-Scale Hydrological Model Simulations to Observed Runoff Percentiles in Europe, Journal of Hydrometeorology, 13, 604–620, doi:10.1175/JHM-D-11-083.1, 2012.

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., and Andréassian, V.: Large-sample hydrology: a need to balance depth with breadth, Hydrol. Earth Syst. Sci., 18, 463–477, doi:10.5194/hess-18-463-2014, 2014.

Hargreaves, G. and Samani, Z.: Reference Crop Evapotranspiration from Temperature, Appl. Eng. Agric., 1, 96–99, doi:10.13031/2013.26773, 1985.

Haylock, M. R., Hofstra, N., Klein Tank, A. M. G., Klok, E. J., Jones, P. D., and New, M.: A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006, J Geophys Res, 113, doi:10.1029/2008JD010201, 2008.

Holländer, H. M., Blume, T., Bormann, H., Buytaert, W., Chirico, G., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Kraft, P., Stamm, C., Stoll, S., Blöschl, G., and Flühler, H.: Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data, Hydrol. Earth Syst. Sci., 13, 2069–2094, doi:10.5194/hess-13-2069-2009, 2009.

Holländer, H. M., Bormann, H., Blume, T., Buytaert, W., Chirico, G. B., Exbrayat, J.-F., Gustafsson, D., Hölzel, H., Krauße, T., Kraft, P., Stoll, S., Blöschl, G., and Flühler, H.: Impact of modellers' decisions on hydrological a priori predictions, Hydrol. Earth Syst. Sci., 10, 8875–8944, doi:10.5194/hessd-10-8875-2013, 2013.

Hrachowitz, M., Fovet, O., Ruiz, L., Euser, T., Gharari, S., Nijzink, R., Freer, J., Savenije, H. H. G., and Gascuel-Odoux, C.: Process consistency in models: The importance of system signatures, expert knowledge, and process complexity, Water Resour. Res., 50, 7445–7469, doi:10.1002/2014WR015484, 2014.

Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C., and Arheimer, B.: Most computational hydrology is not reproducible, so is it really science?, Water Resources Research, pp. n/a–n/a, doi:10.1002/2016WR019285, http://dx.doi.org/k10.1002/2016WR019285.

Kleidon, A. and Heimann, M.: A method of determining rooting depth from a terrestrial biosphere model and its impacts on the global water and carbon cycle, Glob. Chang. Biol., 4, 275–286, doi:10.1046/j.1365-2486.1998.00152.x, 1998.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, doi:10.1080/02626668609491024, 1986.

Koch, J., Cornelissen, T., Fang, Z., Bogena, H., Diekkrüger, B., Kollet, S., and Stisen, S.: Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment, J Hydrol., 533, 234–249, doi:10.1016/j.jhydrol.2015.12.002, 2016.

Laloy, E. and Vrugt, J. A.: High-dimensional posterior exploration of hydrologic models using multiple-try DREAM $_{(ZS)}$ and high-performance computing, Water Resour. Res., 48, doi:10.1029/2011WR010608, 2012.

Lang, C., Freyermuth, A., Gille, E., and François, D.: Le dispositif PRESAGES (PREvisions et Simulations pour l'Annonce et la Gestion des Etiages Sévères) : des outils pour évaluer et prévoir les étiages, Géocarrefour, 81, 15–24, doi:10.4000/geocarrefour.1715, 2006.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J Hydrol., 201, 272–288, doi:10.1016/S0022-1694(97)00041-3, 1997.

Mathevet, T.: Which rainfall-runoff model at the hourly time-step? Empirical development and intercomparison of rainfall runoff model on a large sample of watersheds., Ph.D. thesis, ENGREF University, Paris, France, 2005.

McMillan, H.: Effect of spatial variability and seasonality in soil moisture on drainage thresholds and fluxes in a conceptual hydrological model, Hydrol. Process., 26, 2838–2844, doi:10.1002/hyp.9396, 2012.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I — A discussion of principles, J Hydrol., 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Nicolle, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., Le Lay, M., Besson, F., Soubeyroux, J.-M., Viel, C., Regimbeau, F., Andréassian, V., Maugis, P., Augeard, B., and Morice, E.: Benchmarking hydrological models for low-flow simulation and forecasting on French catchments, Hydrol. Earth Syst. Sci., 18, 2829–2857, doi:10.5194/hess-18-2829-2014, 2014.

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

Nijzink, R. C., Samaniego, L., Mai, J., Kumar, R., Thober, S., Zink, M., Schäfer, D., Savenije, H. H. G., and Hrachowitz, M.: The importance of topography-controlled sub-grid process heterogeneity and semi-quantitative prior constraints in distributed hydrological models, Hydrol. Earth Syst. Sci., 20, 1151–1176, doi:10.5194/hess-20-1151-2016, 2016.

Nippgen, F., McGlynn, B. L., and Emanuel, R. E.: The spatial and temporal evolution of contributing areas, Water Resour. Res., 51, 4550–4573, doi:10.1002/2014WR016719, 2015.

Penna, D., van Meerveld, H. J., Oliviero, O., Zuecco, G., Assendelft, R. S., Dalla Fontana, G., and Borga, M.: Seasonal changes in runoff generation in a small forested mountain catchment, Hydrol. Process., 29, 2027–2042, doi:10.1002/hyp.10347, 2015.

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.

Rakovec, O., Weerts, A. H., Hazenberg, P., Torfs, P. J. J. F., and Uijlenhoet, R.: State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy, Hydrol. Earth Syst. Sci., 16, 3435–3449, doi:10.5194/hess-16-3435-2012, 2012.

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., Attinger, S., Schäfer, D., Schrön, M., and Samaniego, L.: Multiscale and Multivariate Evaluation of Water Fluxes and States over European River Basins, J Hydrometeorol, 17, 287–307, doi:10.1175/JHM-D-15-0054.1, 2016.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., and Participants, D.: Overall distributed model intercomparison project results, Journal of Hydrology, 298, 27–60, 2004.

Rinderer, M., van Meerveld, H. J., and Seibert, J.: Topographic controls on shallow groundwater levels in a steep, prealpine catchment: When are the TWI assumptions valid?, Water Resour. Res., 50, 6067–6080, doi:10.1002/2013WR015009, 2014.

Savenije, H. H. G.: HESS Opinions "Topography driven conceptual modelling (FLEX-Topo)", Hydrol. Earth Syst. Sci., 14, 2681–2692, doi:10.5194/hess-14-2681-2010, 2010.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, doi:10.1002/hyp.6825, 2007.

Seibert, S. P., Jackisch, C., Ehret, U., Pfister, L., and Zehe, E.: Exploring the interplay between state, structure and runoff behaviour of lower mesoscale catchments, Hydrol. Earth Syst. Sci. Discuss., pp. 1–51, doi:10.5194/hess-2016-109, 2016.

Smith, M. B., Koren, V., Zhang, Z., Zhang, Y., Reed, S. M., Cui, Z., Moreda, F., Cosgrove, B. A., Mizukami, N., Anderson, E. A., et al.: Results of the DMIP 2 Oklahoma experiments, Journal of Hydrology, 418, 17–48, 2012.

Spence, C.: A Paradigm Shift in Hydrology: Storage Thresholds Across Scales Influence Catchment Runoff Generation, Geography Compass, 4, 819–833, doi:10.1111/j.1749-8198.2010.00341.x, 2010.

Thirel, G., Andréassian, V., and Perrin, C.: On the need to test hydrological models under changing conditions, Hydrological Sciences Journal, 60, 1165–1173, doi:DOI:10.1080/02626667.2015.1050027, 2015a.

Thirel, G., Andréassian, V., Perrin, C., Audouy, J.-N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindström, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., and Vaze, J.: Hydrology under change: An evaluation protocol to investigate how hydrological models deal with changing catchments, Hydrological Sciences Journal, 60, 1184–1199, doi:DOI:10.1080/02626667.2014.967248, 2015b.

Valéry, A., Andréassian, V., and Perrin, C.: 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2 – Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments, J Hydrol., 517, 1176–1187, doi:10.1016/j.jhydrol.2014.04.058, 2014.

Vansteenkiste, T., Tavakoli, M., Van Steenbergen, N., De Smedt, F., Batelaan, O., Pereira, F., and Willems, P.: Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation, J Hydrol., 511, 335–349, doi:10.1016/j.jhydrol.2014.01.050, 2014.

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., and Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models, Water Resour. Res., 39, 1214, doi:10.1029/2002WR001746, 2003.

Weiler, M. and Beven, K.: Do we need a Community Hydrological Model?, Water Resour. Res., pp. 7777–7784, doi:10.1002/2014WR016731, 2015.

Willems, P.: A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models, Environmental Modelling & Software, 24, 311–321, doi:10.1016/j.envsoft.2008.09.005, 2009.

Willems, P.: Parsimonious rainfall–runoff model construction supported by time series processing and validation of hydrological extremes – Part 1: Step-wise model-structure identification and calibration approach, Journal of Hydrology, 510, 578–590, doi:10.1016/j.jhydrol.2014.01.017, 2014.

Willems, P., Mora, D., Vansteenkiste, T., Taye, M. T., and Van Steenbergen, N.: Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes – Part 2: Intercomparison of models and calibration approaches, Journal of Hydrology, 510, 591–609, doi:10.1016/j.jhydrol.2014.01.028, 2014.

Winsemius, H. C., Schaefli, B., Montanari, A., and Savenije, H. H. G.: On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information, Water Resour. Res., 45, W12 422, doi:10.1029/2009WR007706, 2009.

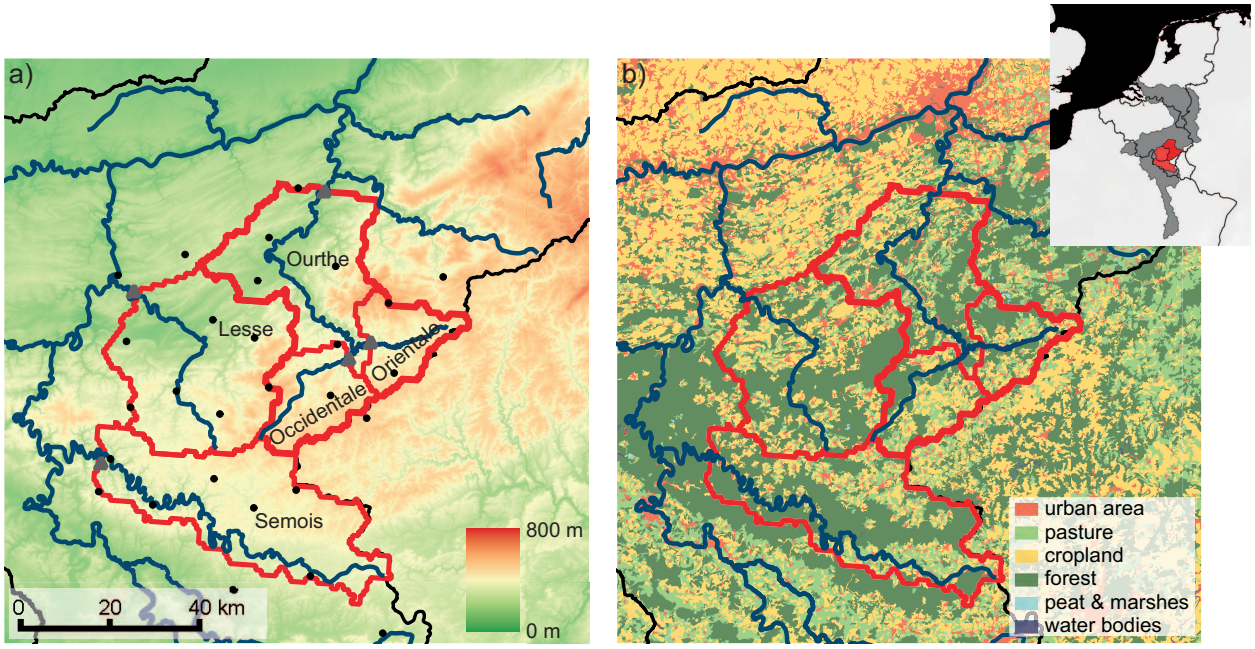

Figure 1. a) Studied catchments (red outline) with DEM and used rain (black dots) and discharge (grey triangles) gauges; b) studied catchments with land cover. The thin black line indicates the catchment boundary of the Meuse River. DEM is obtained from http://hydrosheds.cr.usgs.gov, on 05-06-2013; land use data is derived from CORINE Land Cover (European Environmental Agency, 2006).

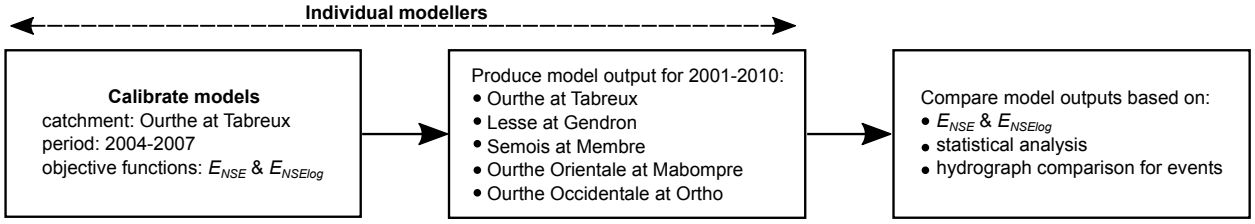

Figure 2. Different steps of the modelling protocol: a schematisation of the model structures used by the individual modellers is presented in Figure 3.

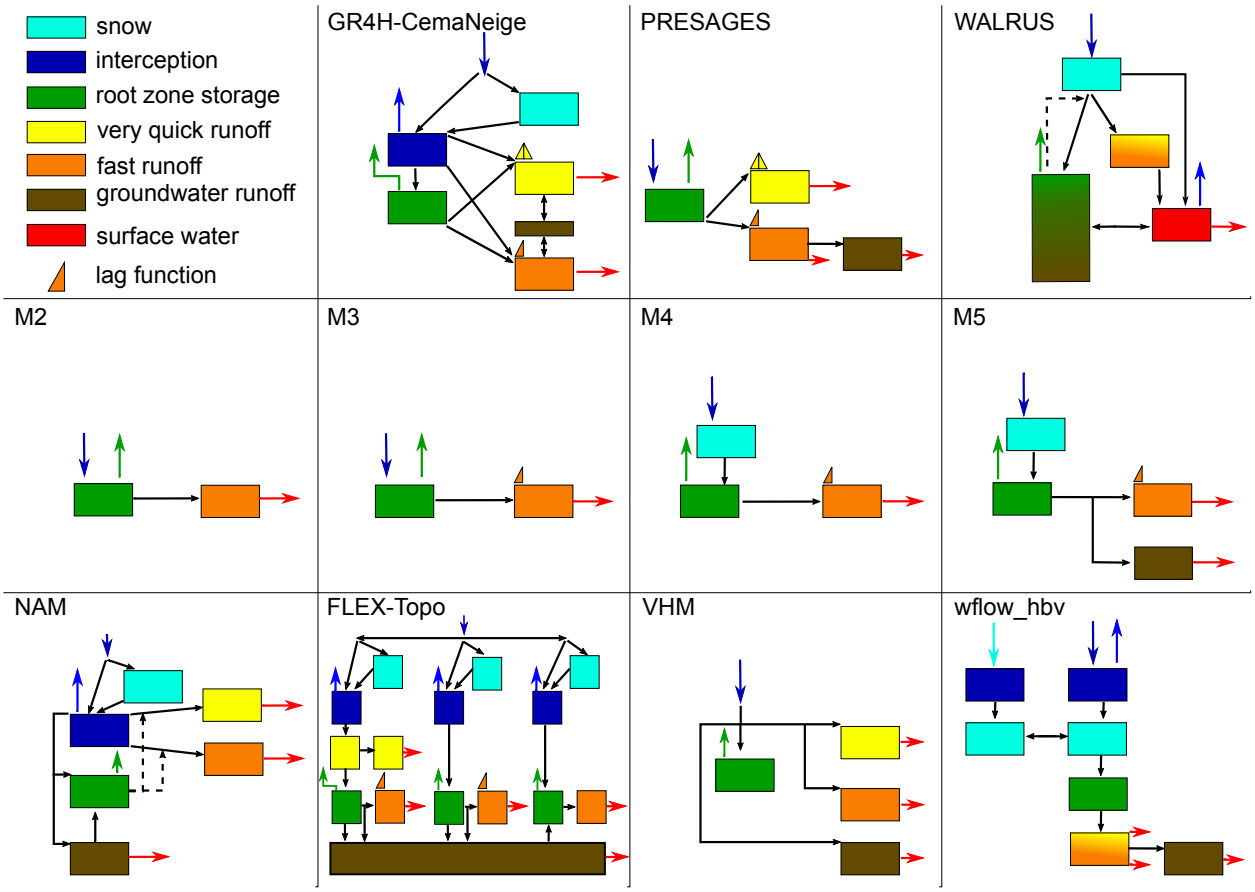

Figure 3. Overview of the eleven used model structures: the schematisation of the model structures is slightly simplified, with the aim to highlight the similarities and differences between the models. The solid lines indicate fluxes between model storages; the dashed lines indicate the influence of the state in a model storage to a flux.

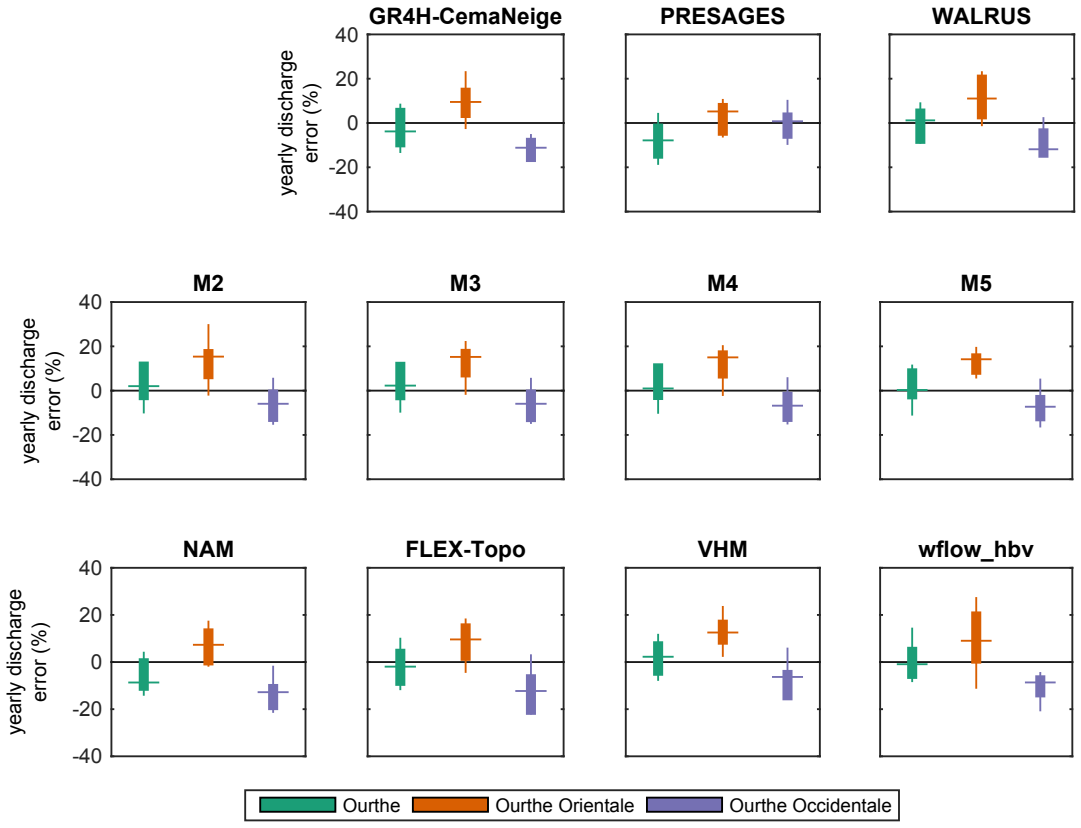

Figure 4. Difference between observed and modelled yearly discharge for Ourthe (green bars), Ourthe Orientale (orange bars) and Ourthe Occidentale (purple bars). Note: to make the graphs more readable, outliers were not plotted.

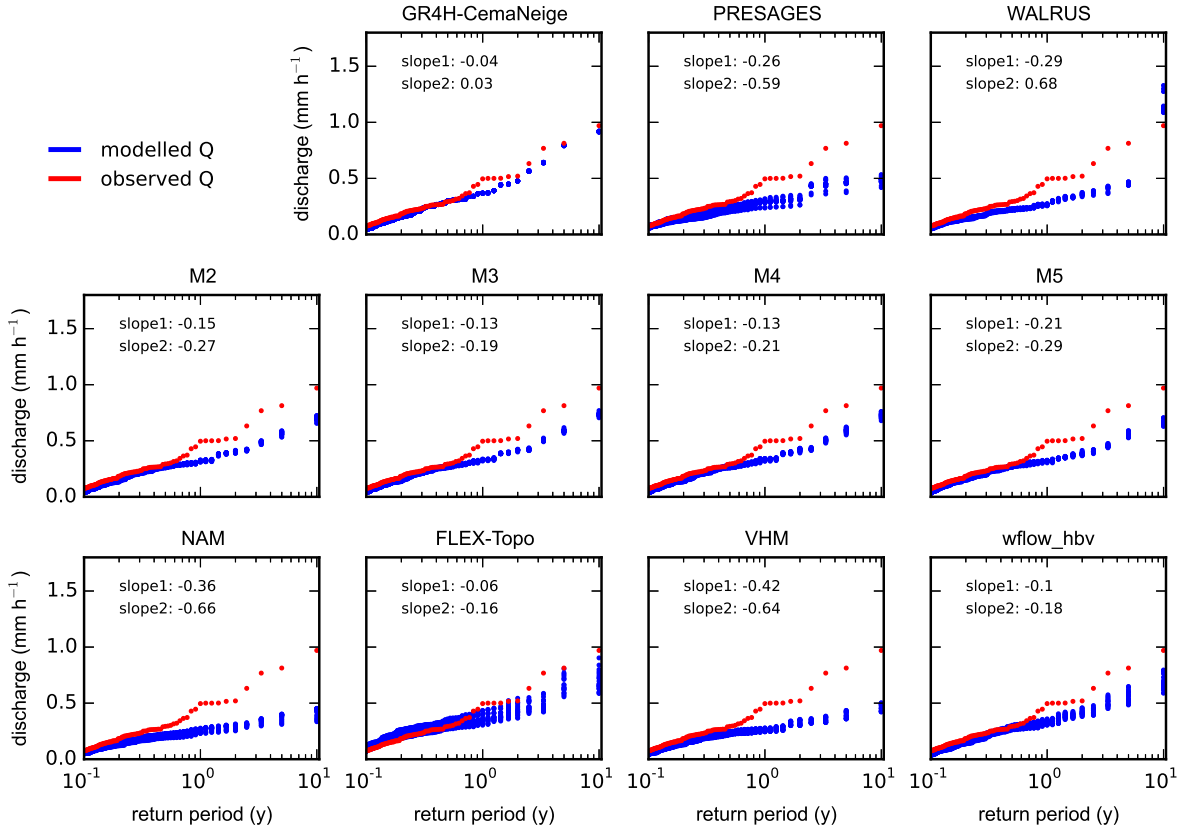

Figure 5. Empirical extreme value analysis for peak flows for the Lesse at Gendron for the total modelled period (2001-2010) (red dots = observed, blue dots = modelled, the spread in the blue dots shows the different realisations). 'slope1' presents the relative error in the slope of the distribution with $T_r < 1.5$ years; 'slope2' presents the relative error in the slope of the distribution with $T_r > 1.5$ years.

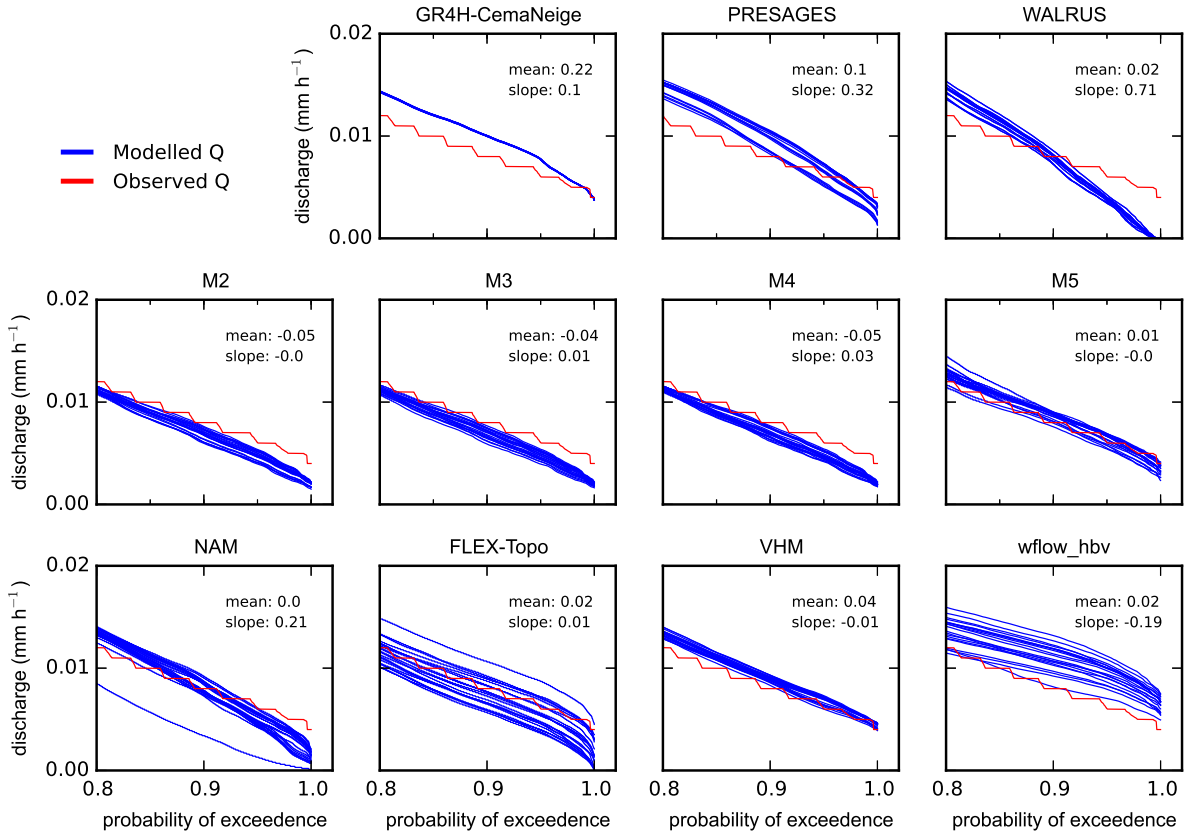

Figure 6. Lowest 20% of the flow duration curves for the Ourthe at Tabreux for all models (red line = observed, blue lines = modelled).

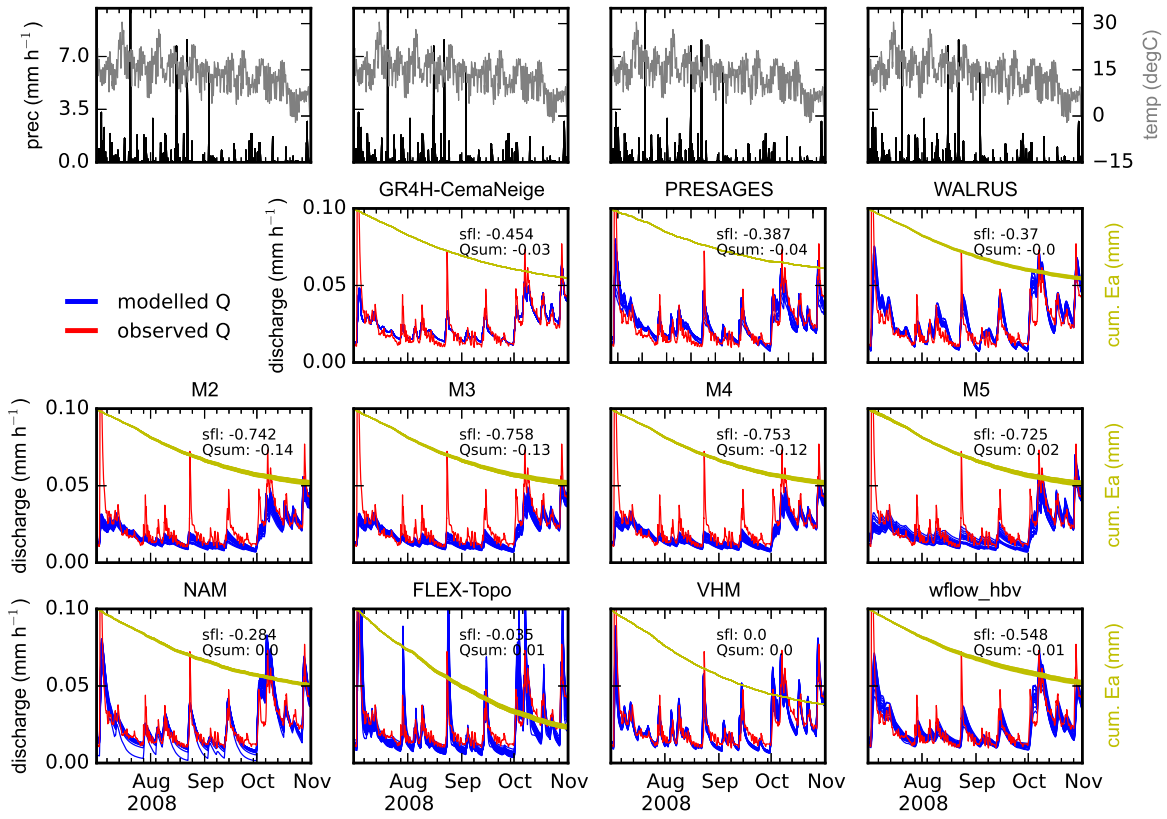

Figure 7. Modelled (blue) and observed (red) discharges for summer 2008 for the Ourthe at Tabreux. The green line shows the cumulative actual evaporation for the plotted period. Note: the four graphs with precipitation and temperature on top are the same. 'sfl' presents the relative error in the average slope of the falling limbs; 'Qsum' presents the relative error in the total modelled discharge for the presented period.

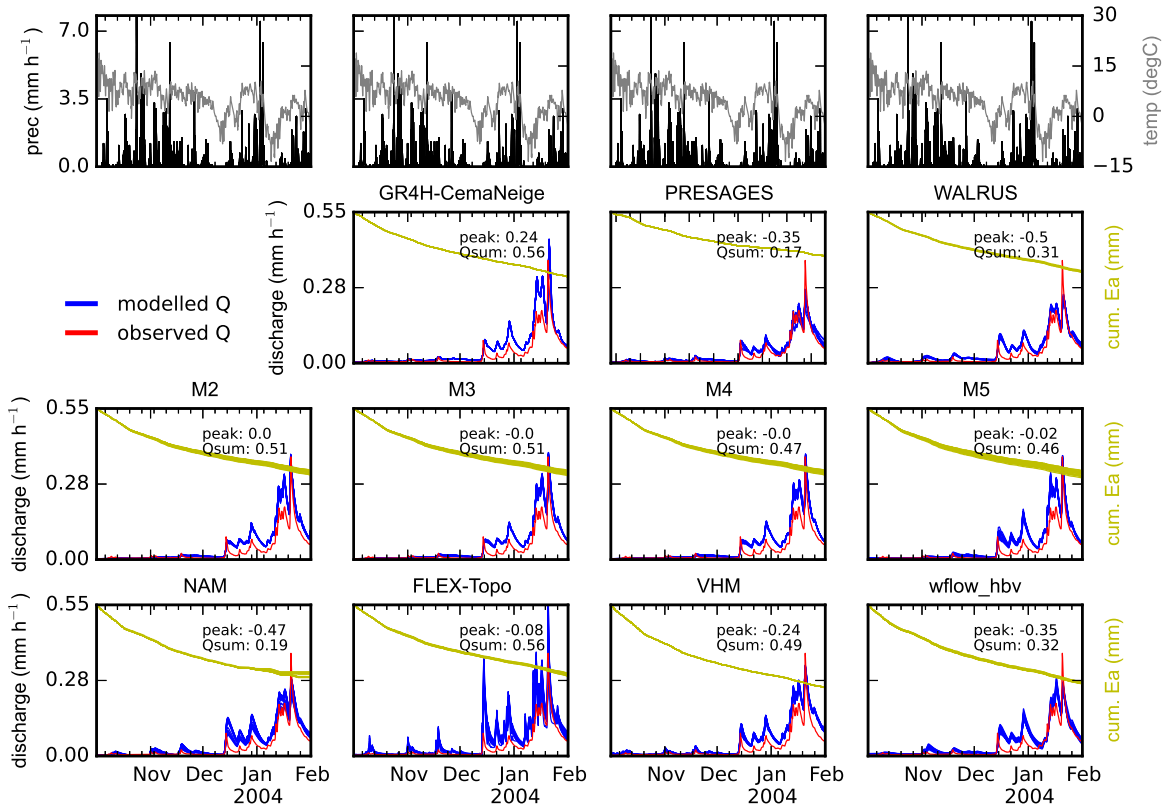

Figure 8. Modelled (blue) and observed (red) discharges for autumn 2003 for the Ourthe Orientale at Mabompré. The green line shows the cumulative actual evaporation for the plotted period. Note: the four graphs with precipitation and potential evaporation on top are the same. 'peak' presents the relative error in the ratio between the first and the highest peak; 'Qsum' presents the relative error in the total modelled disharge for the presented period.

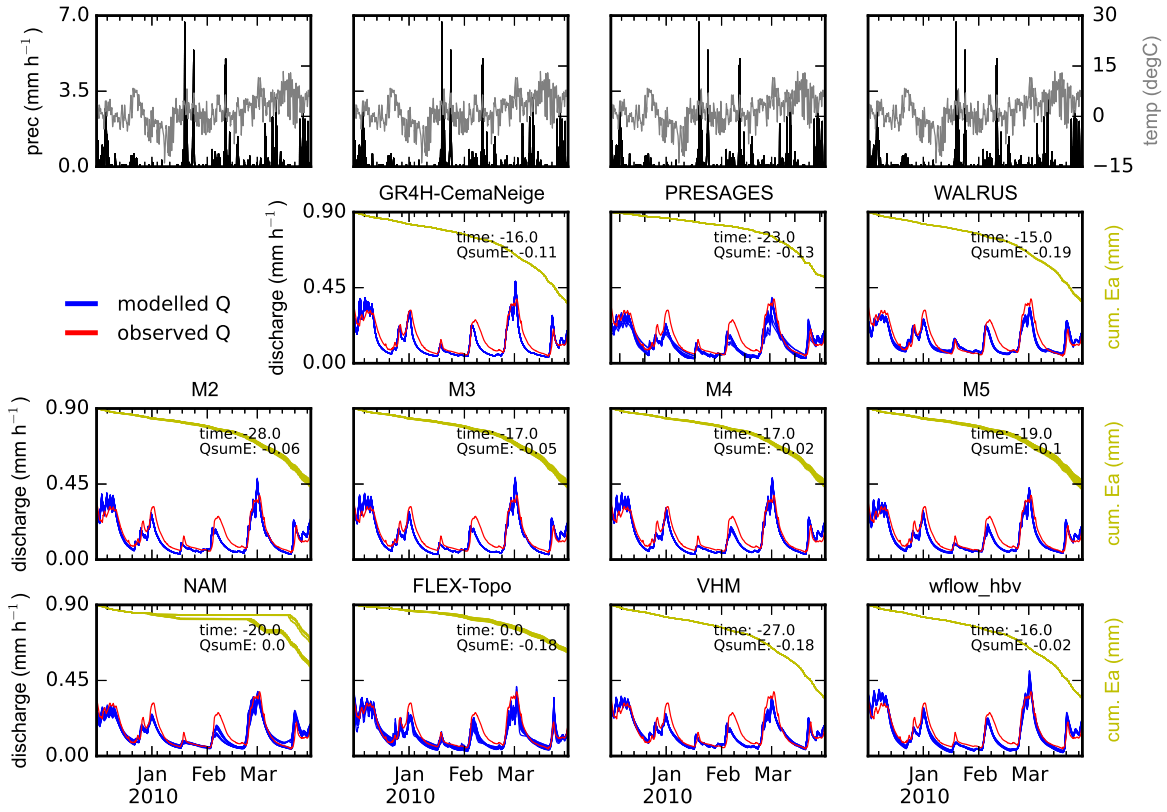

Figure 9. Modelled (blue) and observed (red) discharge for the Semois at Membre for the 2010 winter period. The green line shows the cumulative actual evaporation for the plotted period. Note that the plots with precipitation and temperature on top are the same. 'time' presents the offset in hours of the timing of the highest peak; 'QsumE' presents the relative error in the modelled discharge of the snow melt peak in March.

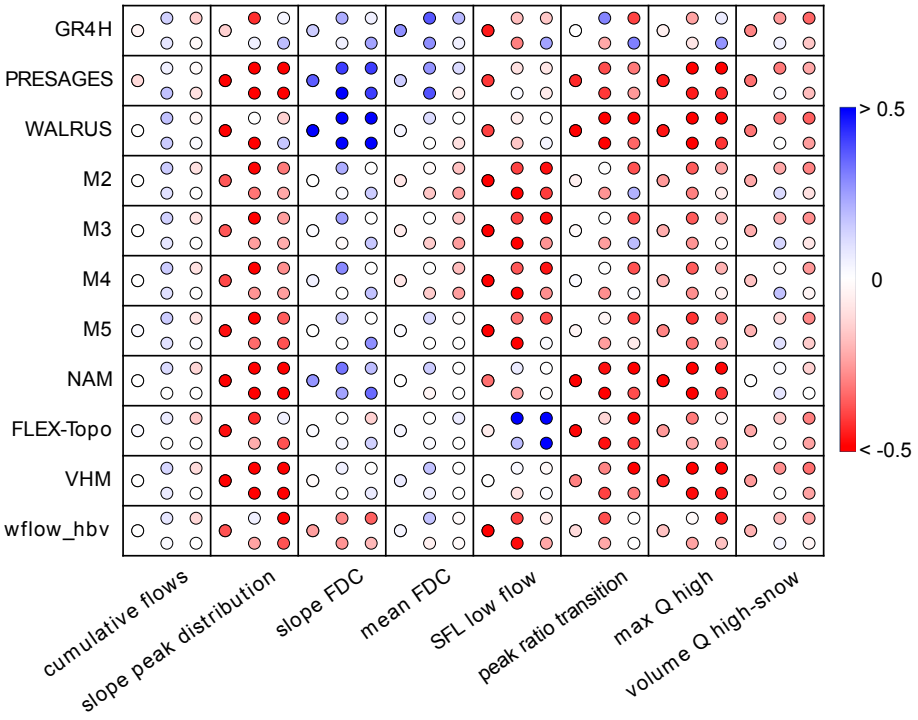

Figure 10. Summary of the performance of a set of signatures for all models and all catchments. The five dots in each square represent the different catchments: left is Ourthe, top middle is Ourthe Orientale, top right is Ourthe Occidentale, bottom middle is Lesse and bottom right is Semois. A red symbol indicates that the modelled value is below the observed value, a blue symbol that the modelled value is higher than the observed value; darker colors indicate larger differences and light or white colours indicate that the modelled signature is very close to the observed signature.

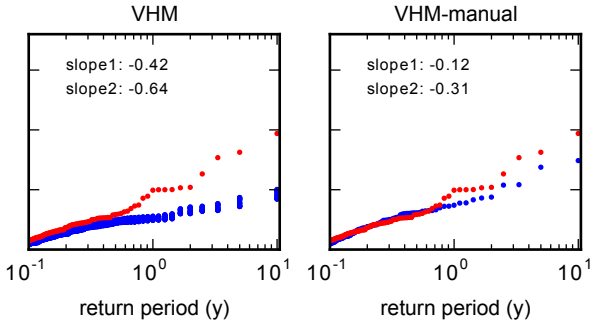

Figure 11. Difference between observed (red) and modelled (blue) empirical frequency distribution of peak flows for the Lesse at Gendron for the entire modelled period (2001-2010) for the automatic (left) and the manual (right) VHM calibrations. 'slope1' presents the relative error in the slope of the distribution with $T_r < 1.5$ years; 'slope2' presents the relative error in the slope of the distribution with $T_r > 1.5$ years.

Table 1. Catchment characteristics

	Ourthe	Orientale	Occidentale	Semois	Lesse
Catchment area (km^2)	1607	317	379	1226	1286
Max elevation (m)	662	662	596	569	586
Min elevation (m)	108	294	303	176	116
Elevation range (m)	554	368	293	393	482
Mean slope	0.090	0.081	0.077	0.087	0.086
Max slope	0.75	0.62	0.58	0.94	0.79
Max flow distance (km)	144	32	44	174	83
Forest cover (%)	46	48	40	56	55
Pasture cover (%)	21	20	23	18	11
Urban cover (%)	6	5	4	5	5
Crop cover (%)	27	27	33	21	29
Mean annual precipitation (mm y^{-1})	1000	1080	1010	1250	1000
Mean annual runoff (mm y^{-1})	460	480	500	670	420
Mean annual temperature ($^\circ$C)	9.6	9.1	9.3	9.6	9.8
Mean annual pot evaporation (mm y^{-1})	730	710	720	750	740
Rising Limb Density	0.72	0.84	0.82	0.61	0.69
Coefficient of autocorrelation (lag = 24h)	0.91	0.91	0.91	0.93	0.86

Table 2. Characteristics of the configuration of the different models

Model	Forcing	Calibration	Parameters [a]	Regionalisation	Group
GR4H	Lumped	Pre-filtering of parameter space using three quantiles for each of the four parameters, followed by stepwise calibration to optimum	4	No	IRSTEA
PRESAGES	Lumped	Optimization with 100 starting points within the parameter space that converge to local minima, which results in more than 2000 parameter sets	6	River routing based on catchment area	Université de Lorraine
WALRUS	Lumped	Manual narrowing of parameter space 500 samples with latin hypercube, 10 best ones for Levenberg-Marquardt optimisation	3	No	Wageningen University and Research
M2	Lumped	MOSCEM-UA (Vrugt et al., 2003)	5	No	Eawag
M3	Lumped	MOSCEM-UA	6	No	Eawag
M4	Lumped	MOSCEM-UA	7	No	Eawag
M5	Lumped	MOSCEM-UA	9	No	Eawag
NAM	Lumped	DREAM_ZS (Laloy and Vrugt, 2012)	12	No	Flanders Hydraulics Research
FLEX-Topo	Semi-distributed	Manual narrowing of parameter space, 2000 uniform samples	20[b]	Percentages HRUs; hydraulic length	Delft University of Technology
VHM	Lumped	MOSCEM-UA	12	No	University of Leuven
Wflow_hbv	Distributed	Manual narrowing of parameter space, 2000 uniform samples	9	Interception capacity	Deltares

[a]Number of calibrated parameters; [b] 11 of the parameters were linked to other parameters based on parameter constraints (e.g., Gharari et al., 2014)