Journal of Hydrology 511 (2014) 335–349

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation

Thomas Vansteenkiste [a,b], Mohsen Tavakoli [c,d], Niels Van Steenbergen [a,b], Florimond De Smedt [c], Okke Batelaan [c], Fernando Pereira [b], Patrick Willems [a,c,*]

[a] KU Leuven, Hydraulics Division, Kasteelpark Arenberg 40, BE-3001 Leuven, Belgium
[b] Flanders Hydraulics Research, Berchemlei 115, BE-2140 Antwerp, Belgium
[c] Vrije Universiteit Brussel, Department of Hydrology and Hydraulic Engineering, Pleinlaan 2, BE-1050 Brussels, Belgium
[d] Ilam University, Department of Natural Resources, Pajohesh Blv., 69315-516 Ilam, Iran

ARTICLE INFO

Article history:
Received 10 July 2013
Received in revised form 16 January 2014
Accepted 22 January 2014
Available online 31 January 2014
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Vazken Andréassian, Associate Editor

Keywords:
Extreme flows
Rainfall–runoff model
Predictive uncertainty
Model structure
Multi-objective calibration

SUMMARY

Five hydrological models with different spatial resolutions and process descriptions were applied to a medium sized catchment in Belgium in order to assess the accuracy and differences of simulated hydrological variables, including peak and low flow extremes and quick and slow runoff subflows. The models varied from the lumped conceptual NAM, PDM and VHM models over the intermediate detailed and distributed WetSpa model to the highly detailed and fully distributed MIKE SHE model. The latter model accounts for the 3D groundwater processes and interacts bi-directionally with a full hydrodynamic MIKE 11 river model. A consistent protocol to model calibration was applied to all models. This protocol uses information on the response behavior of the catchment extracted from the river flow and input time series and explicitly focuses on reproducing the quick and slow runoff subflows, and the extreme high and low flows next to testing the conventional model performance statistics. Also the model predictive capacity under high rainfall intensities, which might become more extreme under future climate change was explicitly verified for the models. The tail behavior of the extreme flow distributions was graphically evaluated as well as the changes in runoff coefficients in relation to the changing rainfall intensities.

After such calibration, all tested models succeed to produce high performance for the total runoff and quick and slow runoff subflow dynamics and volumes, peak and low flow extremes and their frequency distributions. Calibration of the lumped parameter models is much less time consuming and produced higher overall model performance in comparison to the more complex distributed models.

1. Introduction

Hydrological models are widely applied in water engineering for design and scenario impact investigations. Depending on the type of application, the catchment characteristics and the data availability, different spatial and temporal scales, different model conceptualizations and parameterizations are considered. In some cases, the most appropriate model is selected based on these criteria. However, in many applications the model selection seems subject to the common practice of the modeller. Rarely an objective model selection seems conducted (Najafi et al., 2011). Moreover, hydrological studies are often based on one particular hydrological model. The selected model structure might, however, strongly affect the study results, as was shown before by Breuer et al. (2009), Viney et al. (2009), Huisman et al. (2009), Ludwig et al. (2009), Maurer et al. (2010), Bae et al. (2011), Gosling et al. (2011), Smith et al. (2012), Van Steenbergen and Willems (2012) and Velázquez et al. (2012), among others. However, these studies did not draw much attention to the performance of the models under extreme conditions. The calibration was based on statistics evaluating the overall runoff performance, whereas it is known that this does not necessarily lead to good model performance for high and low flow extremes (Westerberg et al., 2011). It is more appropriate to consider multiple objectives that focus on the different aspects of the fit between simulated and observed discharges. Freer et al. (1996) used several performance measures in their Generalised Likelihood Uncertainty Estimation (GLUE) framework. Boyle et al. (2000), Madsen (2000), Yu and Yang (2000), Wagener et al. (2001, 2003), Ferket et al. (2010) and Zhang et al. (2011) applied performance measures on the subflow components of the

* Corresponding author at: KU Leuven, Hydraulics Division, Kasteelpark Arenberg 40, BE-3001 Leuven, Belgium. Tel.: +32 16 3 21658; fax: +32 16 3 21989.
E-mail address: Patrick.Willems@bwk.kuleuven.be (P. Willems).

http://dx.doi.org/10.1016/j.jhydrol.2014.01.050
0022-1694/

runoff discharges or on periods covering different catchment response modes, e.g. wet periods, draining periods, dry periods; or high and low flows above or below a threshold. Westerberg et al. (2011) developed a calibration method including flow-duration curves. Model calibration based on multiple objectives is qualitatively more balanced but does not necessarily statistically perform the best (Westerberg et al., 2011; Willems, 2014). This might raise concern that uncertainty in the impact predictions is additionally induced by the calibration of the models.

Within this paper the influence of the model structure on the model performance for catchment runoff, including high and low flow conditions, is investigated by an ensemble of five hydrological models with different spatial resolutions and process descriptions. In order to obtain consistent and reliable models for use in water engineering (design) applications or scenario-based impact assessment, all models are consistently calibrated by a given systematic but time demanding calibration protocol. The protocol relies on information of runoff subflows and various types of runoff responses derived from the observed river flow, rainfall and potential evapotranspiration (ETo) time series. Explicit focus is given to the high and low flow extremes. It is analyzed whether the models produce reliable estimates of the flow regimes under the current climate and how well they simulate the changes in quick runoff coefficient under changing rainfall intensities. To cover a wide set of model complexities, the selected models in this study vary from the lumped conceptual models NAM, PDM and VHM, over the intermediate detailed and distributed model WetSpa, to the highly detailed and fully distributed model MIKE-SHE. The latter model simulates next to the catchment runoff also internal discharges and groundwater heads.

The Grote Nete catchment in Belgium is taken as case study. It is recognized that next to testing different model structures also different catchments with different meteorological and hydrological characteristics should be studied. Practical barriers, however, prevented us from repeating the approach on other catchments. High quality data and good knowledge on the case study processes and particularities are indeed required to make exhaustive studies on model structures behavior.

2. Study area and models

2.1. Study area

The Grote Nete catchment is located in the northeast of Belgium, with an area of 385 km^2 at the outlet limnigraphic station of Geel-Zammel (Fig. 1). The long term mean annual precipitation in the catchment ranges from about 600 to 1100 mm with an areal average of 828 mm based on the years 2002–2008. The precipitation is almost equally distributed during the winter and summer periods. The long term average annual ETo is about 670 mm. The topography is flat, ranging from 12 m in the west to 69 m in the east with an average value of 22 m. It has a shallow phreatic surface. Catchment slopes are in the range of 0–5%, with an average value of 0.3%. The soils are predominantly composed of sand, sandy loam in the southern and valley areas, and silt. The Grote Nete

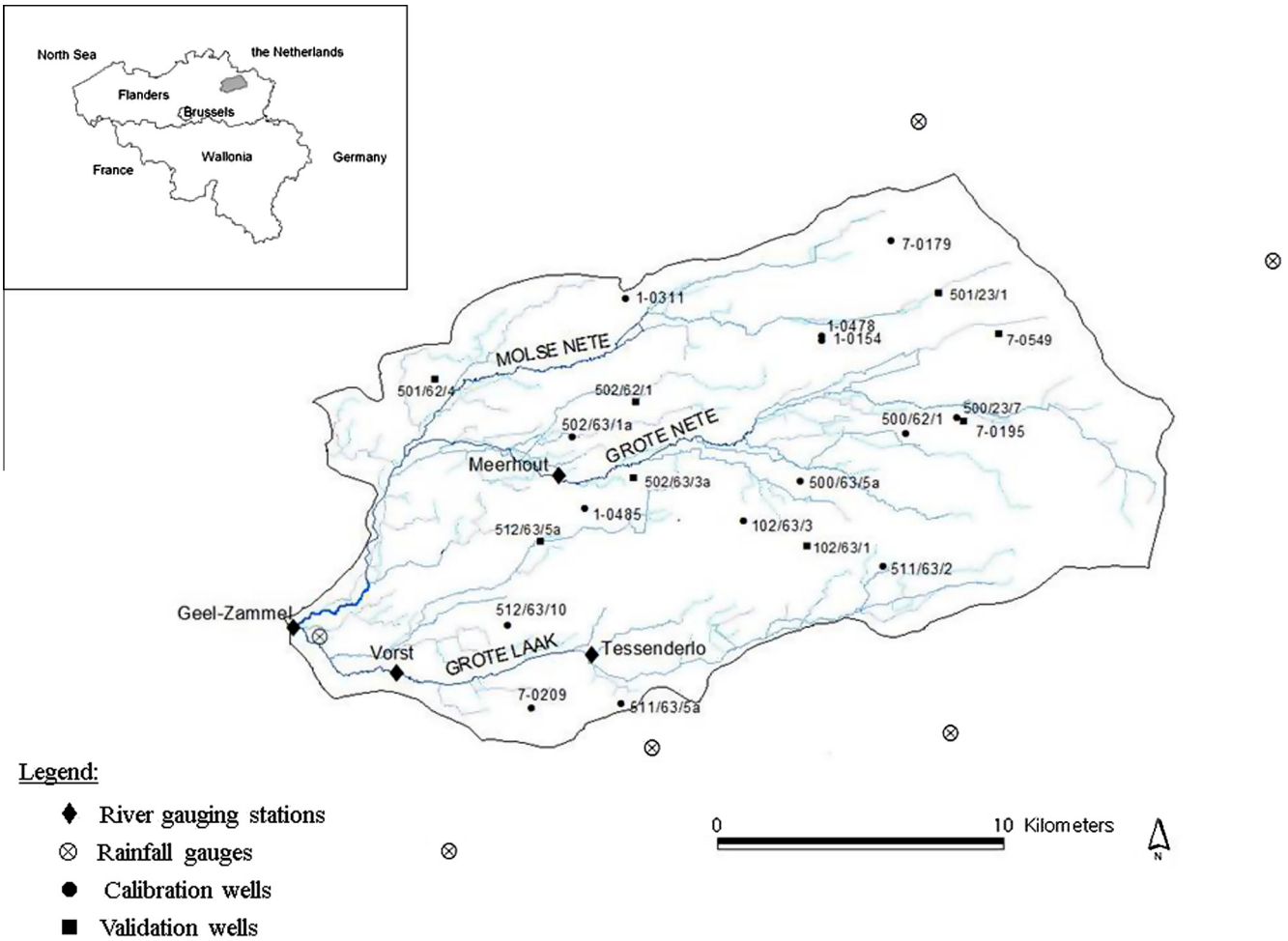

Fig. 1. Location of the Grote Nete catchment in Belgium, and position of the rain gauges, flow gauging station and groundwater wells.

catchment contains numerous river tributaries, and a dense network of ditches and subsurface pipe drains that feed into the main Grote Nete, Molse Nete, and Grote Laak rivers (Batelaan et al., 2003; Woldeamlak et al., 2007).

Consistent hourly data for rainfall, ETo and river flow available for this study ranged from September 2002 to December 2008. The catchment rainfall was derived by spatial interpolation of hourly rainfall observations at six rain gauges inside or close to the catchment (Fig. 1), applying the Thiessen interpolation method after correcting the cumulative rainfall volumes of some rain gauges by a factor. This correction factor was needed for the rain gauges without wind shield or influenced by wind disturbances, and derived by matching the cumulative rainfall volume of a nearby gauge with wind shield and without major wind losses. ETo values were obtained by the method of Bultot et al. (1983), which is a modified Penman–Monteith method.

Groundwater head observations were available for the three main aquifers (three geological units) covering the catchment: the Quaternary (HCOV 0100), the Pleistocene and Pliocene (HCOV 0230), and the Miocene (HCOV 0250) aquifers, which are mainly composed of sand. A total of 30 piezometers in 23 wells were selected (Fig. 1), with head observations between 2 and 12 times a year.

2.2. Hydrological models

Among the five selected hydrological models, three models are frequently applied worldwide, namely MIKE SHE (Abbott et al., 1986a, 1986b), NAM (Nielsen and Hansen, 1973), and PDM (Moore, 1985). The two other models have been developed locally, namely VHM (Willems, 2014) and WetSpa (Wang et al., 1996; Liu and De Smedt, 2004). All five models have been employed successfully to describe the hydrological behavior of rivers in Flanders in the past, but also abroad.

The NAM, PDM and VHM model codes are of the lumped conceptual type. They simulate the rainfall–runoff processes on a (sub)basin scale by means of rather simple concepts and parameters that need calibration. Their model structures are semi-empirical, and have a catchment lumped physical basis. They consider the hydrological system by a combination of storage elements and transport routing units. They include a soil storage accounting procedure, but with different formulations (linear or non-linear), and elements describing the surface and groundwater storage. Their routing modules includes one or more linear or non-linear stores. The models are of moderate complexity and their number of calibration parameters varies depending on the number of processes and the process descriptions.

2.2.1. NAM

For the application within this study the basic model formulations were applied. The basic NAM model considers four reservoirs: snow storage, surface storage, soil storage and groundwater storage and generates three subflows: overland flow, interflow and base flow. The mode of operation of this NAM model is characterized as a system that continuously accounts for the water contents in the storages. The NAM model applies a linear dependency on the soil storage for the quantification of its subflows, which are then further routed by linear reservoirs. Overland flow is routed by a combination of two reservoirs, interflow and baseflow by one reservoir. This basic NAM configuration involves 9 calibration parameters. More details on the NAM model structure and equations can be found in DHI (2008).

2.2.2. PDM

The Probability Distributed Model (PDM) considers the catchment as a collection of soil storage reservoirs, each with a different storage capacity. The spatial variation of these storage reservoirs within a catchment is described by a probability distribution (Moore, 1985). The basic PDM model, used in this study, consists of three main components: (1) a probability distributed soil storage for separation of direct runoff and subsurface runoff, (2) a surface storage component for transforming direct runoff to surface runoff, (3) a groundwater storage which receives drainage water from the soil storage and releases water to the groundwater runoff component (Moore, 2007). River runoff is generated from precipitation falling on the portion of the cell that is saturated, and by drainage from water stored in the soil. In most cases – also in this study – the Pareto distribution is used to describe the variation in storage capacity, which is equivalent to the generation of surface runoff as a power relation of the relative soil storage. Given that PDM only has surface and groundwater storages, only two subflows are simulated. The surface or quick runoff is routed, common to NAM, by two linear reservoir is series; the baseflow is formed after a non-linear cubic transformation of the recharge water released from the groundwater storage. These different processes modelled by PDM are conceptualized by 14 calibration parameters.

2.2.3. VHM

VHM starts form a generalized lumped conceptual model structure with storage elements, representing (1) the groundwater storage, (2) the surface storage, and (3) the unsaturated zone storage. These different storage elements are driven by precipitation contributions, which are transformed into three subflows by reservoirs models. A soil water storage model is the central component in the model and controls the time-variability of the different rainfall fractions, contributing to the subflows. Simple linear functions as well as a more complex exponential relations of the relative soil storage can be applied for describing these rainfall fractions. The case-specific model structure and related equations are identified from the rainfall and ETo input series and river flow series (see also Section 3). This is done through a number of time series processing techniques (Willems, 2009, 2014). The number of parameters to be calibrated depends on the chosen model structure, but the basic application needs 12–13 parameters. The number of routing reservoirs is flexible but for the application within this study it is matched with the NAM and PDM model configuration in order to maximize the model configuration correspondences, i.e. 2 linear reservoirs in series for overland flow routing.

2.2.4. WetSpa

WetSpa is a distributed model code, which considers for each grid cell four layers in the vertical direction: a vegetation zone, root zone, transmission zone and saturated zone. A mixture of physical and empirical relationships is used to describe the hydrological processes. Evapotranspiration is calculated based on the relationship developed by Thornthwaite and Mather (1955). The surface runoff and infiltration are calculated using a soil-storage modified rational method with a potential runoff coefficient depending on land cover, soil type, slope, magnitude of rainfall and antecedent soil storage. Percolation and interflow are assumed to be gravity driven. Interflow is determined by Darcy's law and a kinematic wave approximation in function of hydraulic conductivity, soil storage content, slope angle and root depth. Groundwater storage is simplified using a lumped linear reservoir and the groundwater discharge is proportional to the groundwater storage and the recession coefficient. The total river discharge at the catchment outlet is obtained by superimposing the surface runoff contributions from all grid cells, and the groundwater outflow generated for the entire catchment in lumped form.

2.2.5. MIKE SHE

This MIKE SHE hydrological model (Abbott et al., 1986a, 1986b) simulates the terrestrial water cycle including evapotranspiration, overland flow, surface and root zone soil water storages, and groundwater runoff. For the generation of the runoff flows, different representations and solution techniques can be applied in the model. The choice of method is based on the objectives of the study and the availability of data. Within this study the model applies the complete, physics-based flow description for all processes: evapotranspiration is calculated using the Kristensen and Jensen (1975) method, overland flow is described using the diffusive wave approximation of de Saint–Venant equations, movement of water in the unsaturated zones is modelled as gravity flows and the groundwater flow is simulated by a 3-D groundwater flow model based on by the 3-D Darcy equation. All these processes were in this study model using a grid of 250 m. The MIKE SHE model was developed in combination with its hydrodynamic MIKE 11 river component, which enables simulation of the groundwater–surface water interactions on a physical basis, incorporating river streamflow dynamics and flow regulation mechanisms along the river. The MIKE 11 model is a full hydrodynamic model based on de Saint–Venant equations. The model considered in this study includes implementation of all weirs, culverts, bridges and other hydraulic structures that significantly affect the river flow and water level simulation.

3. Calibration methodology and evaluation tools

3.1. Presentation of the calibration protocol

Given that this study explicitly focuses on the high and low flow extremes, the model calibration explicitly focused on these extremes rather than only on the simulation of the general runoff characteristics and response modes of the hydrographs. This requires multiple objectives to be considered. Multi-objective calibration strategies have been investigated and automated for conceptual rainfall–runoff models (Gupta et al., 1998; Madsen, 2000, 2003; Wagener et al., 2001; Madsen et al., 2002; Vrugt et al., 2003) as well as distributed ones (Madsen, 2003; Ajami et al., 2004; Muleta and Nicklow, 2005; Bekele and Nicklow, 2007; Laloy and Bielders, 2009). These calibration strategies are based on the optimization of a multi-objective function. However, for the spatially distributed hydrological models automatic calibration still remains a challenging problem because the distributed hydrologic models have more complex structures and significantly larger parameter sets that must be specified. These distributed models are also computationally expensive, causing automatic calibration to be subject to severe computational time constraints. One solution is to balance different performance criteria to come up with a single "optimal" parameter set after applying an automatic global multi-objective optimization. Such approach aims at determining a Pareto-optimal set of tradeoff solutions with respect to the different objectives, and the modeller can then select the tradeoff solution(s) that is preferred (e.g. Vrugt et al., 2003; Laloy and Bielders, 2009). This approach, however, typically encounters problems when used with more than say three different objective functions as the higher dimensional the objective space, the more likely that any set of parameter values within the bounds set is a Pareto solution (though 5-dimensional objective spaces can be found in the literature, see e.g. Li and Zhang, 2009).

Because of the large parameter sets of the distributed models and because a handful different performance criteria aimed to be considered in this study (see next), including criteria to evaluate peak and low flows, a step-wise calibration protocol was selected. This moreover allowed – for the conceptual models, but also partly for the distributed models – to assess the different parameter values based on their primary function, in a transparent step-wise way, identifying and calibrating parts of the model structure based on multiple and non-commensurable information derived from the observed data. Although there are options to automate this step-wise calibration approach, in this study a non-automatic approach was followed. As is the case for any manual method, this has limitations in terms of subjectivity (dependence of the model parameter values to the hydrologist performing the calibration). However, the step-wise calibration protocol also aimed to reduce this subjectivity. The protocol relies on physical knowledge of the runoff behavior of the catchment and was consistently applied to all five hydrological models, as discussed in Section 3.3. It uses subflow information and requires a preliminary analysis of the hydrograph recession limbs and separation of the observed river flow series into its runoff subflow components. The protocol also makes use of independent peak and low flow extremes extracted from the river flow series, and separation of this series in events. The events are different for the quick and slow flow components. This requires prior application of a number of time series processing steps, which are first presented in next section.

3.2. Prior time series processing

The total catchment runoff is split conceptually into its overland flow, interflow and baseflow components, representing, respectively, the quick, intermediate and slow response modes of the hydrographs. These subflows are used to calibrate the outlet of the different storage reservoirs and related submodels of the conceptual models. The subflows are determined using a recursive digital filter implemented in the WETSPRO – Water Engineering Time Series PROcessing – tool (Willems, 2009), which is based on the extended Chapman filter for exponential recessions. The filter procedure is based on the clear difference in the order of magnitude of the recession constants for the different runoff subflows. These recession constants describe the catchment response time for the specific subflows and can be identified as the inverse of the decreasing slope of the flow values during long recession periods (dry weather periods) in a $\ln(q)$ time series plot, as described in Willems (2009). Another parameter of the filter is the average fraction of the subflow volumes over the total runoff volumes. This fraction is assessed by a trial-and-error procedure matching the subflow filter results with the total flow results solely during the periods identified as subflow recession periods, again applying the method discussed by Willems (2009). The filter is applied in two steps: in a first step the baseflow is obtained, and in a second step the interflow from the remaining quick flow components. The recession constant of overland flow is finally estimated from the resulting overland flow results.

The flow extremes are extracted from the flow series by applying a revised peak over threshold approach (Willems, 2009), where the extreme peak flows are selected as the highest flows during independent quick flow events and the extreme low flows as the minimum flows during independent slow flow periods. The river flow series is separated in (approximately) independent quick and slow flow events based on criteria for the inter-event time, the inter-event low flow discharge and the peak height. More specifically, two quick flow events are considered independent when the minimum time length between the events exceeds the quick subflow recession constant and the difference between the inter-event minimum flow and the baseflow is small, e.g. lower than a given fraction of the event maximum flow. For the separation of the slow flow events, the baseflow recession constant is taken as the minimum time length between independent events and two slow flow events are considered independent when the inter-event minimum flow is lower than a given fraction of the event maxi-

mum flow. Moreover, only events are selected of which the maximum flow value is higher than a given threshold.

3.3. Calibration steps

3.3.1. General

Based on the rainfall, ETo, total flow and subflow information per quick and slow flow event, as obtained from the time series processing results, the conceptual models were calibrated in a step-wise way. The methodology for this is based on the model structure identification and calibration procedure underlying the VHM modelling approach, developed and improved by Willems (2014).

The approach aims to identify and calibrate individual submodels and their parameters in a transparent, sequential way, based on their primary function and based on multiple and non-commensurable information derived from the observed data (Willems, 2014). It has previously been applied and discussed also for rainfall–runoff applications in different regions of the world by Taye et al. (2011), Liu et al. (2011), Van Steenbergen and Willems (2012) and Taye and Willems (2013). Whereas in VHM the model equations are identified is a case-specific way, the model structure of NAM and PDM is fixed. It is shown in Willems et al. (2014) that the step-wise model calibration protocol has clear advantages. When this method is compared with a conventional automatic search algorithm (which uses a multi-criteria objective function that is solely computed on total runoff at the catchment outlet), the subflows simulated by the models calibrated via the conventional method highly differ from the ones obtained by the time series pre-processing. Only when the step-wise calibration protocol is applied (manual calibration, or automatic calibration for each step), good results are obtained for the subflows. Willems et al. (2014) moreover shows that the proposed protocol brings superiority with respect to the model predictive power on extreme flows.

3.3.2. Lumped models

In a first step of the calibration approach (and model structure identification approach for VHM), the subflow recession constants obtained during the subflow filtering were applied directly as flow routing parameters to the model components that apply linear reservoir routing. For NAM and VHM, these components drive the routing of the baseflow, interflow and overland flow. These three subflows are indeed modelled by linear reservoirs in NAM. In VHM, more complex (non-linear, cascade) routing models can be applied but linear reservoirs are applied by preference as the most parsimonious routing models, hence tested first and applied in this case. Since PDM considers only 2 subflows and applies a non-linear routing for its groundwater flow, only the identified overland flow recession constant was directly applied in that model. The PDM groundwater flow routing parameters were obtained indirectly by visually matching the shape of the baseflow filter results. It is important to be aware that the filter results themselves are based on a filter model, and do not necessarily match the real subflows. The filter is, however, based on assumptions that are consistent with the typical parameterization of the subflows in the conceptual models (Willems, 2014).

In the second step of the calibration approach (and model structure identification approach for VHM), the model water balance is closed by fixing the catchment outflow to the observed flow, and optimizing the rest (rainfall minus total flow) water dynamics. This is done by optimizing the soil storage submodel and the submodel that transfers ETo to actual evapotranspiration. Per slow flow event, the rest rainfall depth (rainfall minus total flow) is computed from the observed rainfall and flow, cumulated in time, and the actual evapotranspiration assumed by the evapotranspiration submodel subtracted. This leads to a time series of soil storage variations, hence information on the soil storage dynamics. A first assessment of the evapotranspiration submodel/parameters is made to obtain realistic estimates of the actual/potential evapotranspiration ratio for the study region, and realistic variations of the soil storage dynamics (close to saturation during most wet winter periods, low saturation during dry summer periods). The soil storage and evapotranspiration submodels/parameters are afterwards optimized by maximizing the agreement of the modelled versus observations-based soil storage depths for all slow flow events. This is done by first visualizing and evaluating the model versus observations based soil storage depths in a scatterplot, and afterwards minimizing the mean squared residual for these events. In this scatterplot and optimization, the event-based depths are transformed by a Box–Cox transformation (Box and Cox, 1964; following the approach by Willems, 2009) in order to reach homoscedastic residuals, hence to give equal weight the high and low depths. This is to avoid that more weight is given to the high depths in the optimization (see Willems, 2009). Indeed, note that this study focuses both on high and low flows.

Based on the temporal variation in relative soil storage levels, for VHM the relationship equation is identified between the overland and interflow runoff coefficient and the relative soil saturation. The parameters of that relationship, idem for the submodel parameters controlling the quick subflows in NAM and PDM, are optimized by comparing the quick runoff depths for all quick flow events. This is again first done through an initial assessment by means of a scatterplot, and afterwards by optimization of the mean squared residual after Box–Cox transformation.

After this identification/calibration of the individual submodels/parameters, the entire model is simulated and the results evaluated for the flow extremes and more general performance criteria. This includes time series plots of total flow, statistical performance measures, hydrographs shape during extreme events in summer and winter periods, and comparison of subflow and the total flow volumes (more details will follow in Section 3.4). In case of systematic differences in any of these plots, model fine-tuning was done. This required some iterations, but experience of the authors learned that the step-wise calibration procedure saved time in that process. The consideration of the step-wise approach during the calibration, which includes different model performance criteria, moreover limited the impact of the calibration strategy choice on the results from the models' comparison.

3.3.3. Distributed models

For the calibration of the MIKE-SHE and WetSpa models, sets of parameters controlling the event-based soil storage, evapotranspiration, quick and slow flow volumes and hydrograph shapes were optimized in a similar way as discussed above for the conceptual models. The same model evaluation plots were made. However, because of the large sets of model parameters in the distributed models, only the most sensitive parameters were involved in the calibration. For the grid-based parameters, spatial variations in these parameters were set depending on topographical, land use and soil type information, as described in the earlier research by Vansteenkiste et al. (2013). Only (sub)catchment averaged scaling factors to these spatial parameter maps were involved in the calibration process. Next to the total catchment runoff and its subflows, the MIKE SHE model is additionally calibrated against internal river discharges within and groundwater heads across the catchment. The WetSpa model was only calibrated based on the observed flow downstream the catchment, as was the case for the lumped models.

Table 1 summarizes the main steps of the model calibration protocol and how they were applied to the five models.

Table 1
Summary of the model calibration protocol applied to the different models.

Model component/ step in the protocol	VHM (13 calibration parameters)	NAM (9 calibration parameters)	PDM (14 calibration parameters)	WetSpa (14 calibration parameters per grid cell)	MIKE SHE (30 calibration parameters per grid cell)
Routing models subflows: matching hydrograph shapes	For baseflow, interflow and overland flow: test linear reservoir routing and use recession constants filter	For baseflow, interflow and overland flow linear reservoir routing: use recession constants filter	– For overland flow linear reservoir routing: use recession constant filter – For recharge and baseflow routing: calibrate parameters controlling baseflow shape	– For overland flow: calibrate scaling factor to initial spatial map of potential runoff coefficient – For interflow: idem spatial parameter map for hydraulic conductivity – For baseflow linear reservoir routing: use recession constants filter	– For overland flow: calibrate scaling factor to initial spatial map of overland Manning coefficient and Manning coefficient river network – For baseflow: idem spatial parameter maps controlling groundwater runoff speeds
Soil water storage: closing water balance	Identify and calibrate soil storage capacity and submodel to convert ETo to actual evapotranspiration	Calibrate soil storage parameters and parameters controlling actual evapotranspiration	Calibrate soil storage parameters	Calibrate scaling factors to initial spatial maps of soil and surface water storage capacity parameters	Calibrate root zone soil and surface water storage capacity parameters
Runoff coefficients to subflows and soil storage: matching event-based subflow volumes	For overland flow, interflow and soil storage: identify and calibrate relation of runoff coefficients to soil water storage (saturation excess) and antecedent rainfall (infiltration excess)	For overland flow, interflow and soil storage: calibrate parameters controlling runoff coefficients and soil storage (saturation excess)	For quick flow and soil storage: calibrate parameters controlling runoff coefficients and soil storage (saturation excess)	– For overland flow: adjust scaling factor for spatial map of potential runoff coefficient – For interflow: idem spatial parameter map for hydraulic conductivity	– For overland flow: calibrate scaling factors for spatial parameter maps controlling overland flow volumes – For baseflow: idem spatial parameter maps controlling groundwater runoff volumes
Spatial variation in river flows and subflows: total runoff flow evaluation based on internal flow gauging stations	–	–	–	–	Fine-tune scaling factors for different spatial parameter maps controlling overland flow, surface and soil water storages, and groundwater runoff
Groundwater heads	–	–	–	–	Fine-tune scaling factors for spatial parameter maps controlling groundwater runoff
Extreme events evaluation: model fine-tuning	Fine-tune submodels runoff coefficients for overland flow and interflow	Fine-tune submodels runoff coefficients for overland flow and interflow	Fine-tune submodels runoff coefficients for quick flow	Fine-tune, mainly of parameter maps controlling overland flow	Fine-tune, mainly of parameter maps controlling overland flow

3.4. Tools for model performance evaluation

3.4.1. Conventional performance statistics

To evaluate the simulated total runoff discharges, following conventional model performance statistics were considered: the Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970), the squared correlation coefficient (R^2) of total simulated versus observed series, the percentage root mean square error (RMSE) and mean absolute error (MAE) on the three highest flow events in the period. They were applied to the hourly discharges.

The total water balance of the models was tested by plotting simulated versus observed cumulative runoff volumes over the simulation period. The difference was quantified by means of the water balance discrepancy (WDB), calculated as the percentage difference between modelled and observed total runoff volumes.

Obviously next to these quantitative measures, also qualitative (visual) flow time series comparison is conducted, with focus on the shape of the hydrographs and the hydrograph peaks. For the MIKE SHE model, also the temporal and spatial variations in groundwater heads for the three main aquifers are evaluated. This is done for 18 wells during model calibration, and 12 other wells at which results are evaluated for spatial validation purposes.

3.4.2. Performance statistics of subflows

The same statistics were computed for the subflows, comparing the simulated quick, intermediate and slow subflows with the subflow filter results. Whereas the NAM, PDM and VHM models simulate three subflows, including interflow, the MIKE SHE and PDM models only produce overland or quick flow and base- or slow flow as output. They do not simulate interflow. For the evaluation of the

baseflow simulation, the output of the lower layer storage of the models is considered. All models, except MIKE SHE, make use of the reservoir method for simulation of the baseflow, and produce direct baseflow series. Given that the MIKE SHE model has a detailed 3-D groundwater flow description that generates groundwater runoff spatially variable along the river network, the baseflow downstream the catchment cannot directly be extracted from the model. Therefore, the MIKE SHE subflow components were derived from the total modelled runoff discharges using the numerical filter (same filter as applied to the observed series) to evaluate the subflow model performance.

Note that the subflow separation does not necessarily provide reliable estimates of the real overland flow, interflow and groundwater flow components. However, because for MIKE-SHE the subflow separation method was applied to both the observed and simulated runoff series and because the conceptual models produce the same type and number of subflows as the subflow filter does, relative comparison of the accuracy of the quick, intermediate and slow catchment sub-responses were considered very informative (independent of the physical attributes of these subflows) next to the comparison of total flows.

Remember that the subflow separation involves identification of the subflow recession periods. By graphically comparing the shape of the recession limbs – or the recession constants – during these recession periods, the response times of the runoff subflows and the related routing submodels are tested.

3.4.3. Performance statistics on high and low flows

The observed and simulated high and low flow extremes are evaluated (i) by their match in a scatterplot of simulated versus observed values of the extremes after Box–Cox transformation, and (ii) by comparison of the empirical frequency distributions of the extremes. Next to the mean squared model residual, which was already optimized in the step-wise model identification/calibration, the mean and standard deviation of the model residuals are computed and evaluated. The empirical frequency distribution evaluation involves plotting of the extreme flows versus their empirical return period, calculated as the total length of the time series in years divided by the rank number after sorting of the extremes from high to low (rank number 1 given to the highest extreme value). The tail of this empirical frequency or extreme value distribution is evaluated for its shape, which represents the tendency of the increase in extremes versus increase in return period. The models are considered accurate in their simulation of higher extremes and their performance in making extrapolations when the observed and simulated extremes show similar tendency (Willems, 2009). To allow easy visual checking of this similarity, the peak flow extremes are plotted versus the ln-transformed empirical return period. This leads, following the extreme value theory based on peak over threshold values, in many cases – also in the case – to a linear tail (e.g. Madsen et al., 1997; Willems et al., 2007). The same applies to the low flow extremes but after plotting the inverse of the low flow extremes versus the ln-transformed empirical return period.

3.4.4. Measure of the model predictive power on extreme flow changes

As explained in Section 3.3, during the calibration some components of the model structure were evaluated, mainly the ones that are directly related to the generation of extremes. This was done by plotting the runoff coefficients versus the relative soil storage state for the different models, and compare these with the coefficients derived from the time series processing results. Another method that was applied to evaluate the predictive power of the models for simulating changes in extremes, has been developed by Van Steenbergen and Willems (2012). The method studies changes in the quick runoff coefficient under changing rainfall intensities. This is done by plotting the model based changes in quantiles of runoff coefficients for different rainfall intensity classes (e.g. 5–15%, 15–25%, etc.) and comparing these with to the corresponding changes derived from the observations. In this way the accuracy of the model based flow changes under changing climatic conditions was empirically tested.

4. Results

The five hydrological models in this study were applied under the same meteorological and catchment conditions. They were run at an hourly time step, using the same precipitation and ETo inputs, and were calibrated against the same measured river flow data at the catchment outlet. The calibration data cover the period from September 2002 to the end of 2005 which covers a wide range of meteorological and hydrological conditions, including a very wet winter with several high peak flows and flooding (December 2002–January 2003) and a very dry summer (2005). The validation is then carried out for another 3 year period (January 2006 to December 2008). All simulations were carried out with a warming up period from January to August 2002.

The prior time series processing steps – subflow filtering, event separation and extraction of flow extremes – were applied to the full available hourly river flow series for 2002–2008. The identified values of the recession constants are 2100 h for baseflow and 120 h for interflow. The estimation of the overland flow recession constant was more difficult: upper and lower limits of this constant were identified as being [15, 30] hours. The average fraction of the subflow volumes over the total runoff volumes were identified to be 0.3 and 0.4 for respectively the baseflow and interflow components.

For the separation of the flow series in independent quick and slow flow events, the quick subflow recession constant of 120 h for interflow was considered as the minimum time length between the quick flow events. The threshold for the inter-event minimum flow for defining quick flow events was taken a fraction 0.3 of the event maximum. For the separation of the slow flow periods, the minimum time length between the events was increased to 4200 h and the threshold for the inter-event minimum flow taken to be a fraction of 0.5 of the event maximum flow. For the minimum peak flow height, a value of 3 m^3/s was selected.

4.1. Conventional model performance statistics

4.1.1. Discharge

The evaluation of the conventional model performance statistics (Table 2) for the simulated total runoff discharges show that all five models capture generally well the overall basin runoff and streamflow dynamics. For the calibration period, NSE values are high for all models (>0.70) with the VHM model having the

Table 2
Statistical performance values for the calibration period (09/2002–2005) and validation period (2006–2008).

	NAM	PDM	VHM	MIKE SHE	WetSpa
Calibration period					
NSE (–)	0.77	0.79	0.71	0.79	0.76
R^2 (–)	0.88	0.89	0.86	0.89	0.89
MAE (%)	7.86	13.26	9.18	14.38	10.06
RMSE (%)	9.18	17.19	13.14	19.51	13.48
Validation period					
NSE (–)	0.75	0.74	0.57	0.66	0.63
R^2 (–)	0.89	0.88	0.81	0.84	0.85
MAE (%)	11.71	12.81	14.08	13.62	12.86
RMSE (%)	14.50	14.41	14.81	17.97	16.54

lowest efficiency (0.71). For the validation period, the model performance is less good for VHM (NSE = 0.57) and the distributed models (NSE between 0.6 and 0.7), whereas they are as high as 0.75 and 0.74 for the PDM and NAM models respectively. The R^2 of simulated versus observed total runoff flows does not differ much for the five models. The R^2 values are high and vary between 0.8 and 0.9 for all models.

4.1.2. Winter and summer events

Qualitative inspection of the hydrographs during winter (Fig. 2) shows that the simulated hydrographs are in good agreement with the observed ones. The recession limb of the simulated hydrographs matches the recession limb of the observed hydrographs very well and the value and timing of the peak discharges are very well estimated by all models during this season.

For summer periods (Fig. 3) the distributed models tend to overestimate the hydrographs. This is particularly the case for the WetSpa model. Next to the flow overestimations, quicker recessions are noticed in the MIKE SHE simulated hydrographs. No clear explanations were found for these anomalies. The lumped models perform better for these summer events, though the hydrograph volumes are in some periods underestimated. This is mainly

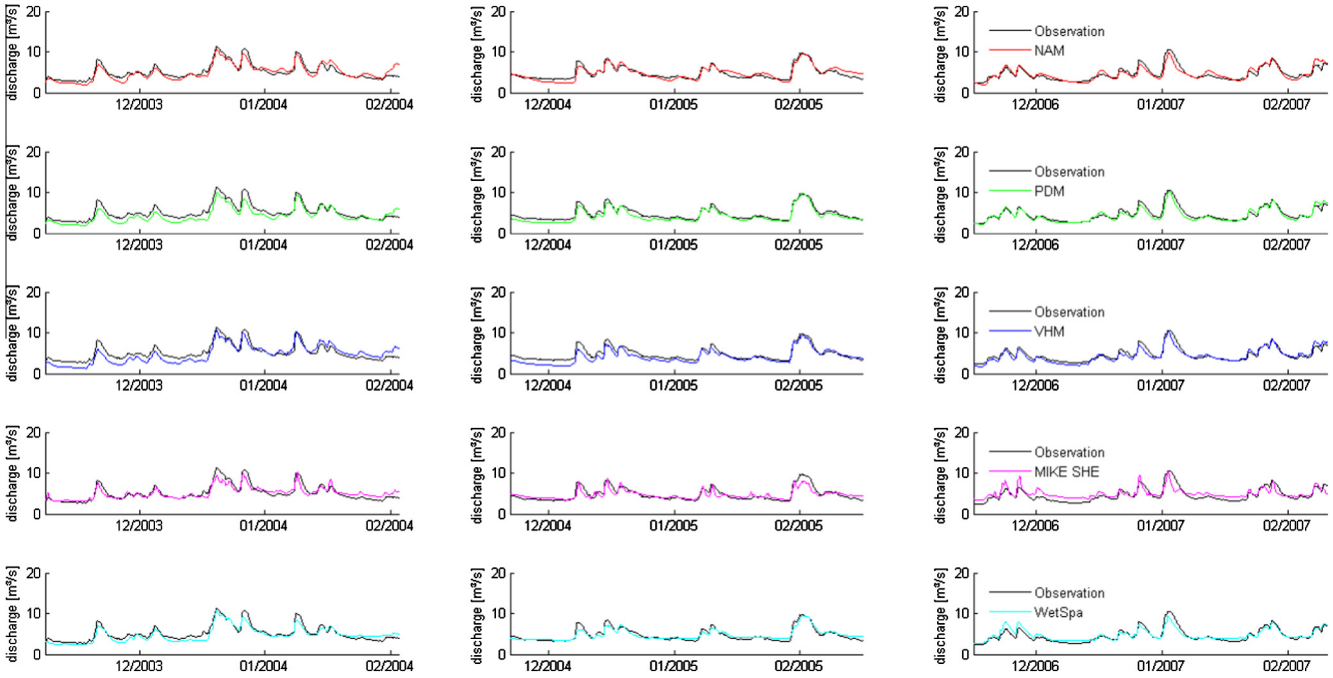

Fig. 2. Observed and simulated runoff discharges for winter events downstream the catchment.

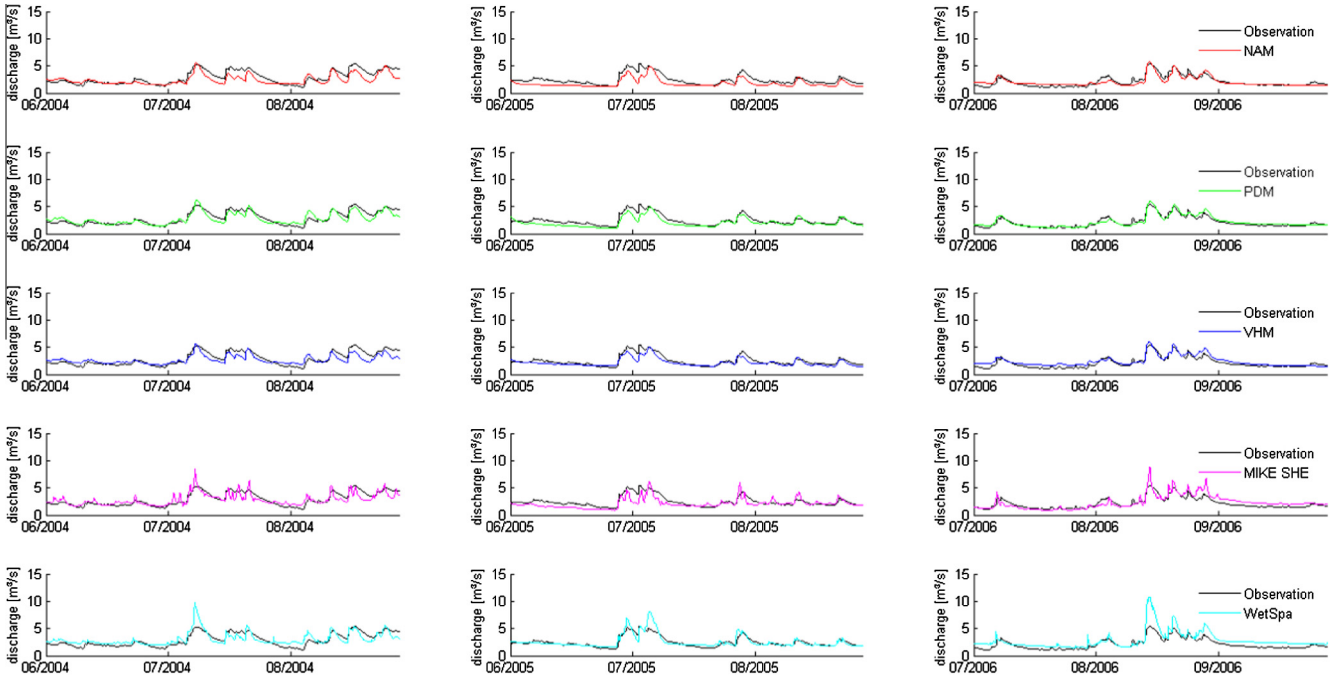

Fig. 3. Observed and simulated runoff discharges for summer events downstream the catchment.

seen in the NAM model results, where considerable underestimations are noted in the flows after the peak moments. The lower discharges between the summer hydrograph events seem to be well estimated by all models. No consistent over- or underestimations are observed in these baseflow results.

4.1.3. Water balance

To compare the models' capabilities in simulating the total water balance, the runoff volumes downstream the catchment are compared in Fig. 4. The WBD was calculated at the end of the calibration and validation periods. A good match with the observations was found for the calibration period and for the lumped models (WDB < 1.79%), whereas for the distributed models the WBD values are 5% for MIKE SHE and 10% for WetSpa. During the validation period, the results are less good with higher WBD, particularly for the distributed models (WDB of 10% for MIKE SHE and 13% for WetSpa). Similar to the simulations of the summer and winter events, the distributed models perform less accurate in comparison to the lumped models.

4.1.4. Runoff subflows

Evaluation of the modelled baseflow series shows that all models reproduce very well the observations-based baseflow filter results (Fig. 5). The models simulate the seasonal variation of the baseflow regime very well. Only for the period around September 2008, the NAM and VHM baseflow results do not match well the filter results. For this period, the NAM and VHM models underestimate the baseflow, while the PDM and the distributed models respond more accurately. For the distributed models, it is noticed that the baseflow simulated by the WetSpa model follows the observed seasonal variation very well, but with some overestimations during winter and spring periods. For the validation period these

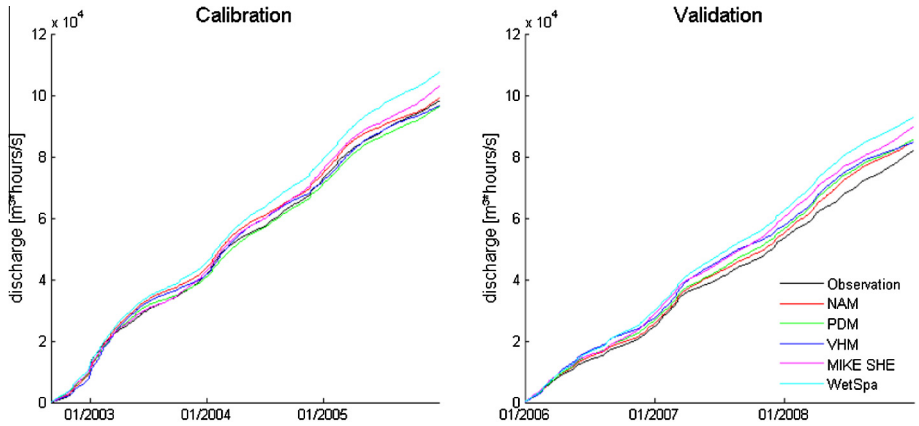

Fig. 4. Observed and simulated cumulative discharge downstream the catchment.

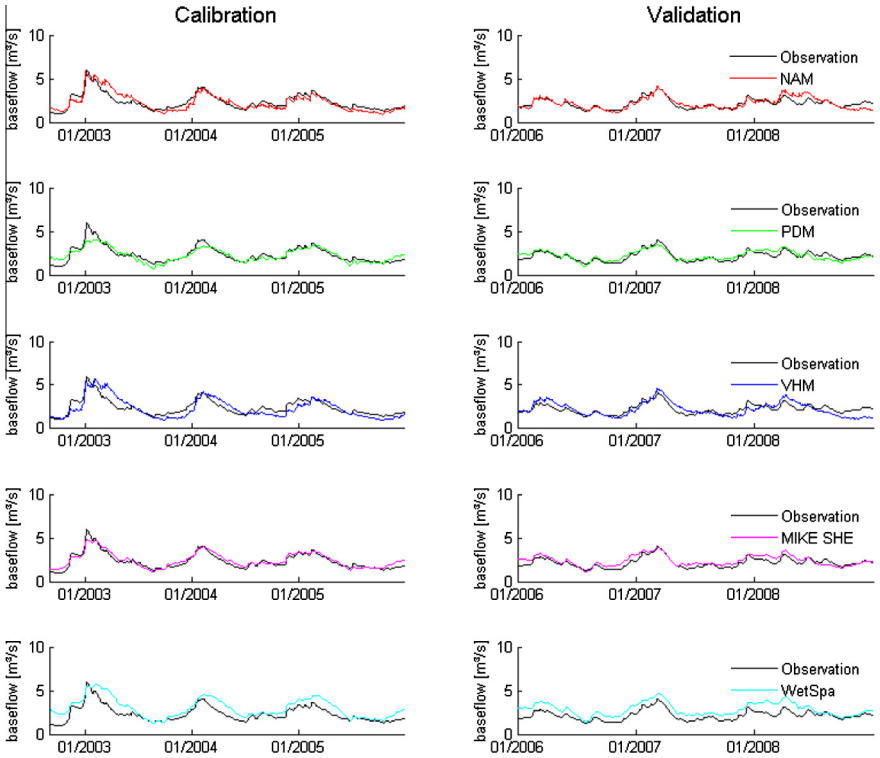

Fig. 5. Observed and simulated baseflow downstream the catchment.

overestimations are more pronounced. They lead to relatively high overestimations of the cumulative volumes (WBD > 25%). Similar overestimations in the baseflow series are seen in the NAM and VHM results, followed by small underestimations during autumns. These under- and overestimations during consecutive seasons compensate each other and lead to very good overall cumulative volumes (small WBDs, between −4.69% and 1.49%). The NAM, VHM and WetSpa models all apply a linear reservoir for the baseflow simulation, which explains this similar baseflow behavior. The PDM and MIKE SHE models apply other baseflow model concepts, which is reflected in their baseflow results. The PDM model simulates well the lower baseflows but does not capture the higher baseflows during the winter and early spring periods (winter 2002–2003, winter 2003–2004). The PDM cumulative baseflow volumes have a good match with the observed ones (−3.15% < WBD < 2.30%). This is also the case for the MIKE SHE modeled baseflow, whereas this model shows clear overestimations for the validation period (WDB = 10.29%). Given that the total and baseflow volumes have a high accuracy, it is expected that the simulated quick flow volumes match the filter results well too (−6.14% < WBD 4.75%), except for the WetSpa model. The WetSpa model underestimates the quick flow volumes (WDB around 35%).

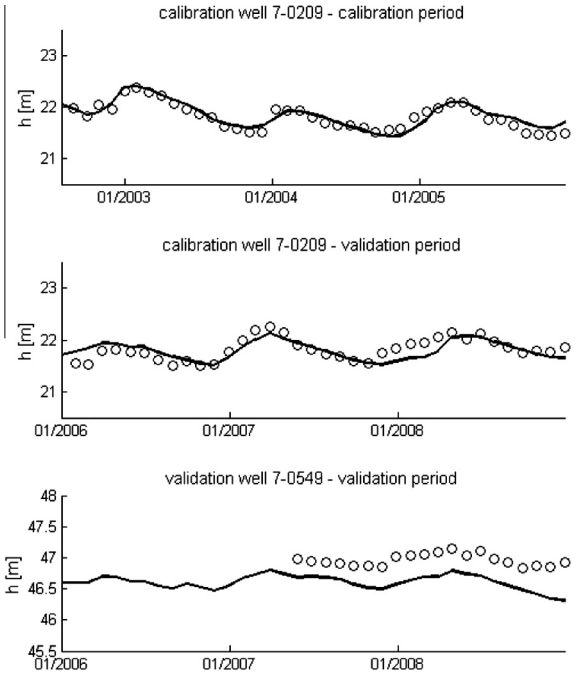

Fig. 6. MIKE SHE head levels for selected calibration well 7-0209 and validation well 7-0549.

For the MIKE SHE model the groundwater related flows could also be validated for the heads in the three main aquifers at the different groundwater wells. Fig. 6 shows the results at a selected well for the calibration period, validation period as well as a validation well not considered during the calibration. It is noted that the groundwater heads and their seasonal variations simulated by the MIKE SHE model have high correspondences with the observations. RMSE values of 0.14 m and 0.15 m are found for calibration well 7-0209 for respectively the calibration and validation period, and 0.18 m for the validation well. Investigation of the results for all calibration and validation wells shows that the model reproduces the historical spatial and seasonal distribution of the observed heads very well for the different geological units. Systematic differences in absolute head levels or seasonal head variations are found for some wells (example of well 7-0549 in Fig. 6), but in general the groundwater dynamics are well captured. The RMSEs vary between 0.10 m and 0.90 m for the calibration wells and between 0.10 m and 0.50 m for the validation wells.

4.2. Model performance for high and low flows

The model performance for the peak and low flow extremes was evaluated based on the entire simulation period, given the limited data series for the calibration and validation periods separately.

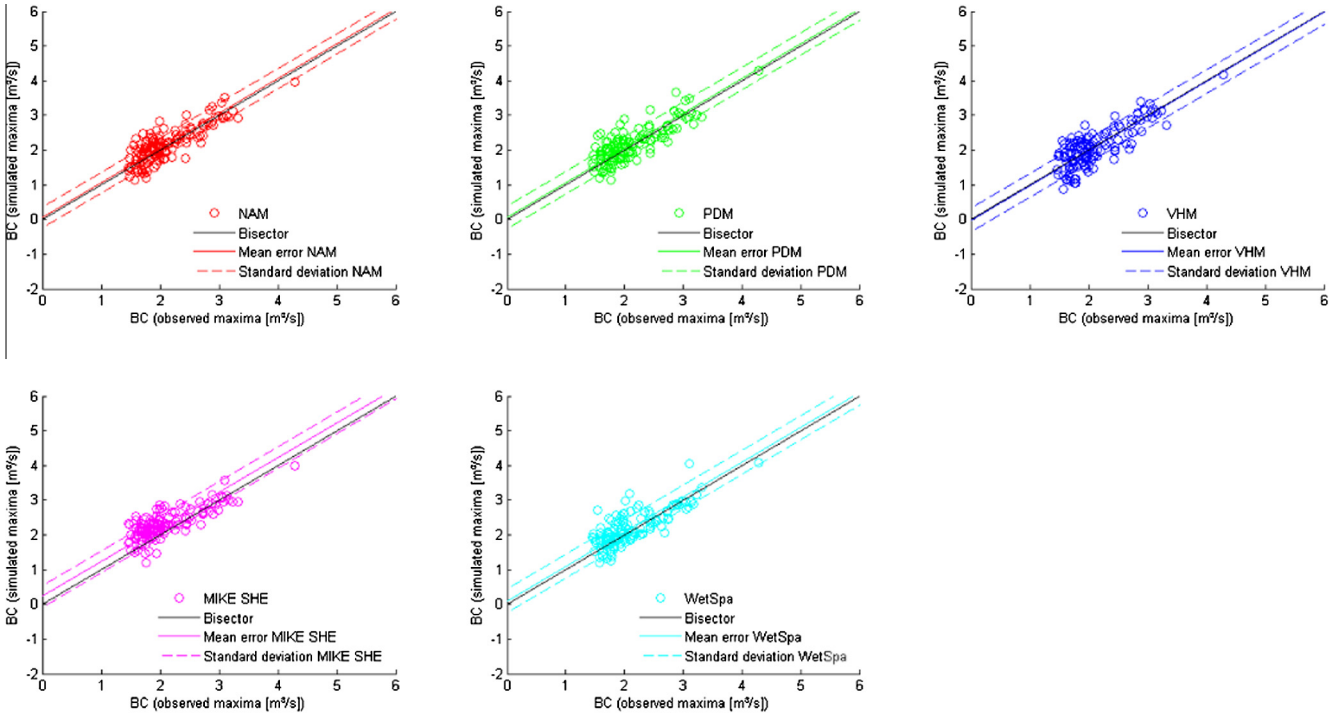

Fig. 7. Simulated versus observed hourly peak flows downstream the catchment after Box–Cox transformation ($\lambda = 0.25$).

4.2.1. Peak flows

Fig. 7 shows the comparison of the simulated versus observed peak flows after Box–Cox transformations in a scatter plot. The mean and standard deviation of the model residuals are indicated in the figure by the solid and dotted lines. The standard deviation is assumed constant (independent on the flow magnitude) after Box–Cox transformation (homoscedastic model residuals). All plots show a good agreement between the observed and simulated peak flows, which is seen by the small scatter and the low standard deviations. Only the MIKE SHE model tends to slightly overestimate the peak flows with a small positive mean deviation (0.22 m^3/s) identified in the plot. For the other models this mean deviation is much smaller (<0.1 m^3/s). This is also confirmed in the frequency distribution of the empirical extremes in Fig. 8, where mainly the lower observed peak flows are slightly overestimated by MIKE SHE. These lower peak flow extremes are also slightly overestimated in the NAM, PDM and WetSpa model results, but well estimated by the VHM model. At the other end of the distribution – the highest peak flows – the PDM model shows a steeper upper tail. The tail's shape is indeed heavier, which means that extrapolations towards higher extremes than the empirical ones are expected to be significantly higher. This might have consequences when applying the models in climate change impact research, where the changes in flow extremes are essential. This divergence in the peak flow distribution's tail for the PDM model was also identified by Van Steenbergen and Willems (2012). They explained it by the model structure in which the surface runoff directly depends on the soil storage capacity, described by the Pareto distribution. This distribution induces a power relation between the soil storage and the surface runoff coefficient (Van Steenbergen and Willems, 2012), whereas the NAM and VHM models consider a linear relation.

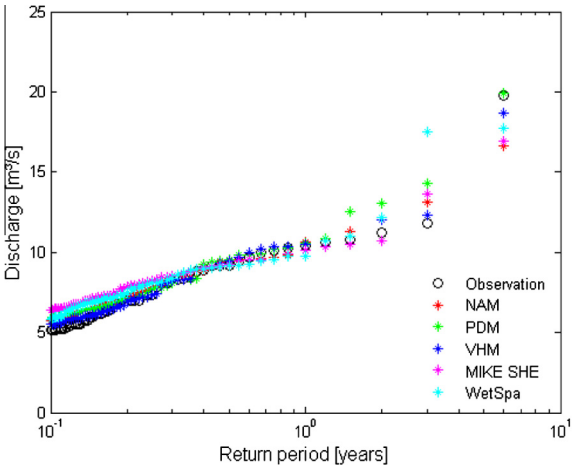

Fig. 8. Empirical extreme value distributions of simulated versus observed hourly peak flows downstream the catchment.

4.2.2. Low flows

The model performance of the low flows was evaluated on a daily basis in a similar way as the peak flows. All models simulate the minima very well, except the WetSpa model which is overestimating the low flows slightly (Fig. 9). The low flow empirical frequency distribution plot (Fig. 10) confirms these results. The latter plot at the same time shows that the models tend to overestimate the least extreme low flows, whereas the most extreme low flows are slightly underestimated. These under- and overestimations compensate each other and lead to very small mean low flow errors in the validation plot (Fig. 9). The underestimations of the most extreme low flows might be of high importance given that they point towards stronger underestimations for drier conditions. This is of particular relevance if the model would be applied for climate change impact investigations, given that the climate change signal over Belgium points towards drier summer conditions (Baguis et al., 2009; Bauwens et al., 2011). The deviations are particularly clear for the PDM model.

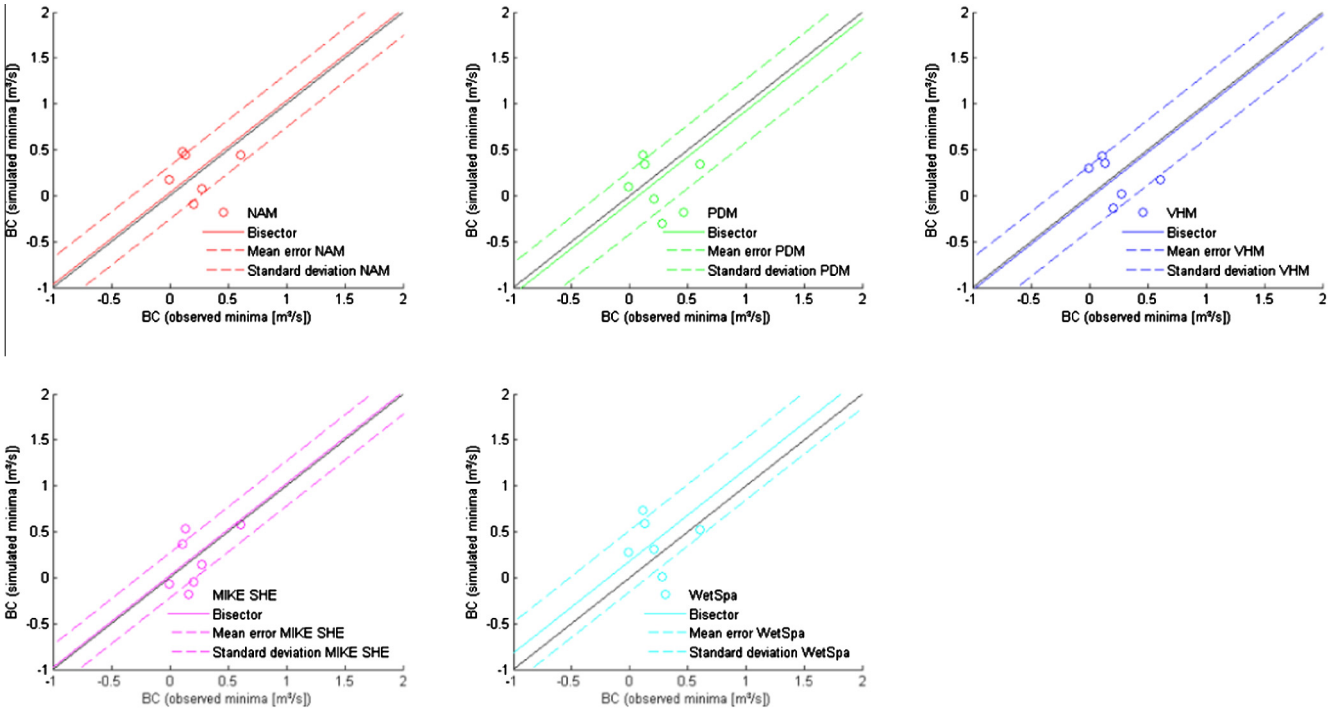

Fig. 9. Simulated versus observed daily low flows downstream the catchment after Box–Cox transformation ($\lambda = 0.25$).

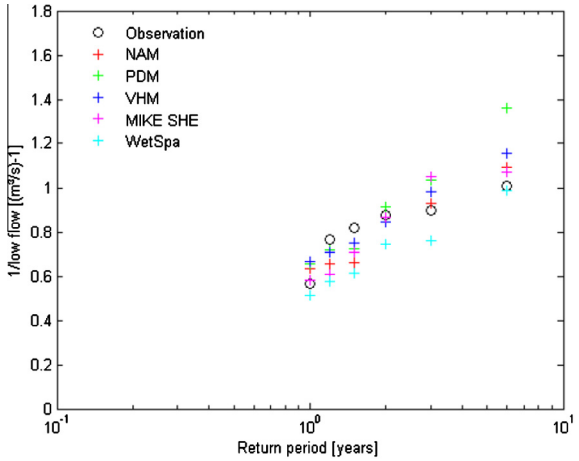

Fig. 10. Empirical extreme value distributions of simulated versus observed daily low flows downstream the catchment.

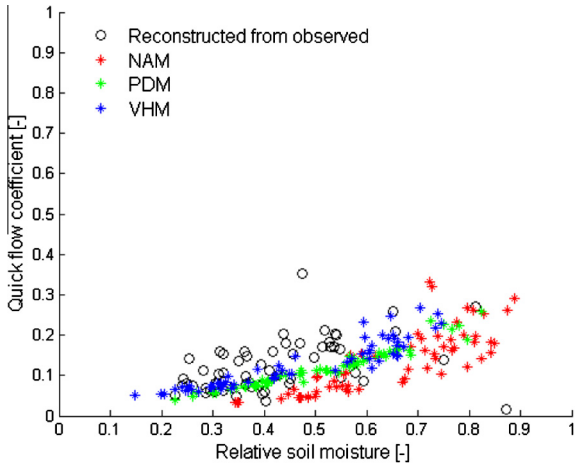

Fig. 11. Quick runoff coefficient for the NAM, PDM and VHM models.

4.3. Model structure for quick flow generating processes

Analysis of the quick runoff coefficients, estimated as the ratio of quick runoff volume versus total precipitation volume during the quick flow events, in function of the soil storage for the different models (Fig. 11) is informative on the differences in model structure for the lumped conceptual models. The slope of the quick runoff coefficient versus the soil storage shows a convex curve for the PDM model (higher increase in runoff coefficient for higher soil saturation), whereas the slope is more constant for the NAM and VHM models albeit the scatter in runoff coefficients is higher for these models. Fig. 11 moreover shows that the NAM model requires a higher soil saturation to produce similar quick runoff as the VHM and PDM models. Given that no direct measurements are available of the soil storage, the baseflow filter result was used as an indirect indicator of the relative soil saturation level. A linear relation is assumed between the baseflow and the relative soil storage (Van Steenbergen and Willems, 2012). Based on the model structural analysis in Fig. 11, it is noted that the VHM model runoff coefficient suits best the coefficients derived from the observations. The NAM results show a similar approx. linear relation between the runoff coefficient and the relative soil moisture, with higher soil moisture results but runoff coefficient values in the same range. The PDM model shows a slight convex relation with a tendency to higher runoff coefficients for the higher soil saturation levels. At low soil storage levels, the observation based coefficients are higher than the model based runoff coefficients.

4.4. Predictive power on extreme flow changes under changing rainfall conditions

Figs. 12 and 13 show the results of the evaluation of the model predictive power (Van Steenbergen and Willems, 2012) for projecting changes in peak flow extremes under changing rainfall-climate conditions.

4.4.1. Overland flow changes

For all considered classes of change in rainfall intensity, the NAM and VHM models simulate similar changes in overland flow

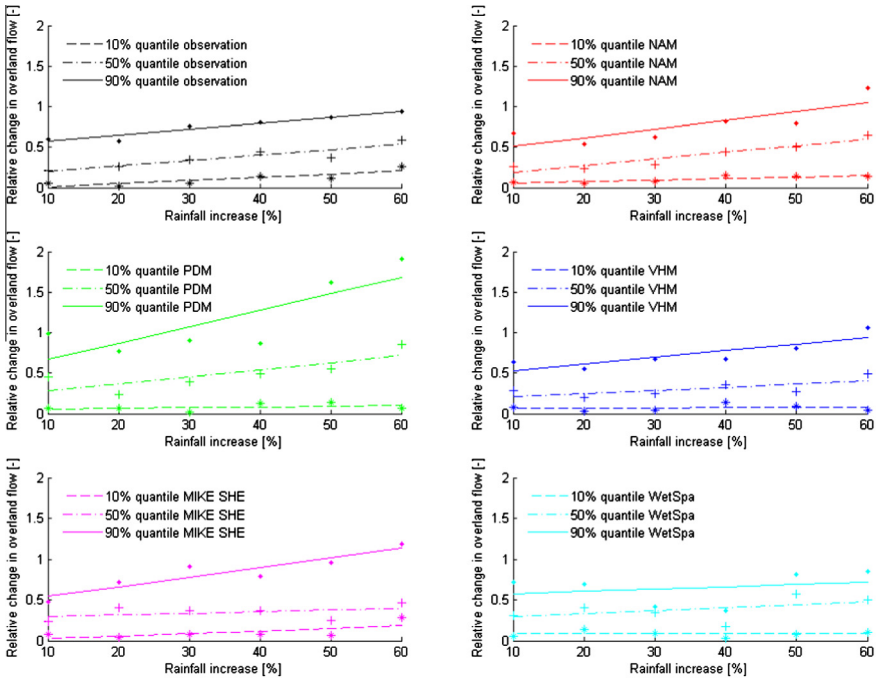

Fig. 12. Quantile analysis of the change in overland flow event volumes per rainfall intensity class.

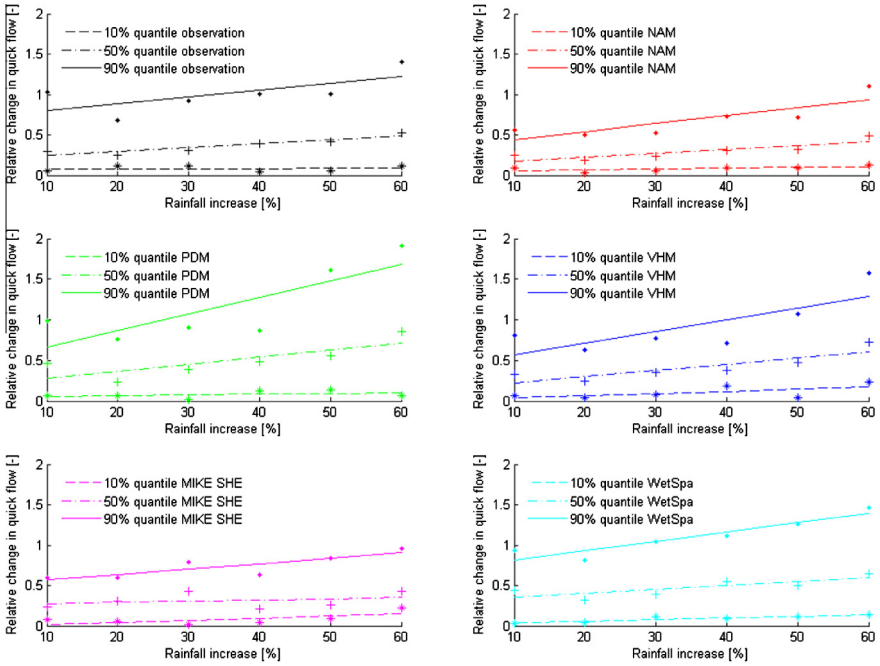

Fig. 13. Quantile analysis of the change in quick flow event volumes per rainfall intensity class.

event volumes (Fig. 12), which highly correspond with the observations. The lowest (10% quantiles) as well as the highest (90% quantiles) changes are accurately estimated. The overland flow changes by the PDM model show high differences for most of the rainfall intensity classes. Among the distributed models, the MIKE SHE simulations are generally more consistent with the observations than the WetSpa model. The MIKE SHE 50% and 90% quantile changes follow the observed changes closely with a small overestimation towards the high rainfall changes, indicating that the MIKE SHE model slightly overestimates the runoff coefficients under high rainfall change conditions. The pattern in the WetSpa quantile changes is less clear.

4.4.2. Quick flow changes

In order to make a meaningfull comparison with the PDM modelled overland flow, the above analyses on the overland flow changes are repeated for the combined interflow and overland flow in Fig. 13. In this figure the PDM modelled changes are equal to Fig. 12, but the observed and simulated changes by the other models differ after including the interflow. The comparison in the plots shows that the PDM model simulates quick flow changes in correspondence to the observations for the lower rainfall changes. For the higher rainfall changes (>45%) the 90% quantiles still differ largely from the observations and the results from the other models. The 10% and 50% quantiles by PDM show less difference. This finding supports the suggestion that the PDM model is overestimating the magnitude of the highest flows in response to high rainfall intensities. The VHM model results approximate the observed quick flow changes very well. However, for the highest quick flow changes (90% quantiles) underestimations are noticed for the lower rainfall changes (<30%). The 10% and 50% quantiles of quick flow changes are better simulated for the VHM model. The NAM model tends to underestimate the quick flow changes when interflow is incorporated in the analysis. For almost all rainfall change classes the 50% and 90% quantiles indicate lower changes in comparison to the observations. These underestimations of the interflow in NAM may be explained by the dependence of the interflow on the surface storage in the model structure. This is motivated by the observation that during autumn and summer almost no interflow is simulated because of high evapotranspiration, which depletes the surface storage reservoir. The opposite is seen in winter, but the interflow peaks have an upper limit due to the surface storage reservoir capacity. The MIKE SHE model also underestimates the quick flow changes. These underestimations are mainly the result of the interflow underestimations and are most clear for simulations of the highest changes (90% quantiles). For the WetSpa model quick flow changes are well modelled. However, in combination with the good estimation of the interflow volumes and the underestimation of the overland flow, this good quick flow result is obtained thanks to a bias in the interflow routing by the WetSpa model.

5. Conclusions

Although the structure complexity (spatial resolution and process details) play a role on the model efficiency, all five models appear to be fairly adequate for the catchment. The overall runoff processes and streamflow dynamics are well captured by the models. Winter events are generally better estimated than the summer ones, and the lumped models perform better than the distributed ones. Also in terms of overall water balance and the subflows the lumped models are slightly superior with very low discrepancies in the subflow volumes in comparison to the distributed models. Peak and low flows are very well simulated with small mean deviations. In general, the lumped models show a higher accuracy than the distributed ones. In addition, because of their smaller number of parameters, the lumped models could be calibrated more accurately. The lumped models also have a much smaller computational time (CPU time of about 1 min for the lumped models versus 17 h with MIKE SHE for a single 4-year calibration run with a personal computer equipped with a Intel Core i5 processor and having a CPU frequency of 2.60 GHz and RAM memory of 4 GB).

Other studies on the comparison between hydrological models including lumped and distributed models, confirmed that lumped and distributed models may lead to similar accuracy (e.g. Ajami et al., 2004; Reed et al., 2004; Breuer et al., 2009; Smith et al.,

2012; Apip et al., 2012; Lobligeois et al., 2013). Lobligeois et al. (2013) show this for western France where catchments are under oceanic climate conditions with quite spatially uniform precipitation fields. They conclude that the accuracy of distributed models increases in southern France for catchments in which precipitation fields are highly variable in space. That lumped models provide a valuable integrated view of the basin outlet response was also concluded in the Distributed Model Intercomparison Project for the Oklahoma region by Reed et al. (2004) and Smith et al. (2012). However, this study shows that the lower complexity of lumped models might lead to inconsistencies in the catchment (sub-)responses as identified from the observations. Examples of such inconsistencies for the present case study are given next. The NAM conceptualization of interflow underestimates the interflow peaks in winter. The peak flow frequency distributions revealed that the PDM model overestimates the highest peak flows, which was confirmed by the analysis on the predictive power of the models to simulate overland and quick flow changes under increasing rainfall conditions. Higher changes in runoff coefficients were found for the PDM model for the high rainfall intensity changes.

Pokhrel and Gupta (2011) and Smith et al. (2012) clarified that the benefit of distributed models mainly lies in the broader range of applications, including studies involving spatial scenarios and by providing estimates of hydrologic variables at interior catchment locations. This is also the case for the distributed WetSpa and MIKE SHE models in this study; the MIKE SHE model captures the spatial and temporal groundwater dynamics well, which is very useful for many applications such as agricultural and ecological impact studies.

Note that this paper did not had the ambition to draw definite conclusions on the applied modelling systems. The model calibration approach may affect such conclusions, and the study was limited to one case study. We therefore recommend evaluation of more calibration approaches, more model structures and for more case studies. For the catchment considered here, this research has shown that all five modelling systems are able to produce accurate results for a wide range of runoff properties: daily total flows, daily quick and slow subflows, cumulative volumes, peak flows, low flows, frequency distributions of peak and low flows, changes in overland and quick flows for given changes in rainfall. This was the case after a careful calibration, which involves constraining of the model to reproduce all these runoff properties.

One limitation of the current work is the manual nature of the applied calibration protocol, which inherently introduces "subjectivity" in the calibration, albeit more limited than in traditional manual calibration because of the step-wise protocol. In this respect, it would be interesting to investigate whether the same calibration protocol can be applied in a fully-automatic way.

The results in this paper show that the differences in the model ability to reproduce the various tested runoff properties are very limited, despite the large range of model complexity considered. It makes the models useful for a wide range of applications, such as flood studies, low flow and water scarcity investigations, reservoir design, and climate change impact assessment. It avoids that choices have to be made: whether the model calibration should focus primarily on the peak flows, or the low flows, or the cumulative volumes, . . . , depending on the specific application. The good performance of the models for extremes and changes in flow conditions for changes in rainfall gives some trust that the models are reliable in making extrapolations beyond the range of events considered during model calibration and validation. Further tests on this may include differential-split sample tests, as recommended by Refsgaard et al. (2014). It moreover would be useful to investigate how the different models – which have similar accuracy in current climate conditions – respond to climate change scenarios. It would be interesting to see how the differences in process descriptions and the complexity of these descriptions affect the runoff impact results of the climate scenarios.

Acknowledgement

This research was supported by Flanders Hydraulics Research, Ministry of Mobility and Public Works, Flanders, Belgium.

References

Abbott, M.B., Bathurst, J.C., Cunge, J.A., Oconnell, P.E., Rasmussen, J., 1986a. An introduction to the european hydrological system – Systeme Hydrologique Europeen, "She", 1. History and philosophy of a physically-based, distributed modeling system. J. Hydrol. 87 (1–2), 45–59.

Abbott, M.B., Bathurst, J.C., Cunge, J.A., Oconnell, P.E., Rasmussen, J., 1986b. An introduction to the european hydrological system – Systeme Hydrologique Europeen, "She", 2. Structure of a physically-based, distributed modeling system. J. Hydrol. 87 (1–2), 61–77.

Ajami, N.K., Gupta, H., Wagener, T., Sorooshian, S., 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. J. Hydrol. 298, 112–135.

Apip, Sayama, T., Tachikawa, Y., Takara, K., 2012. Spatial lumping of a distributed rainfall sediment-runoff model and its effective lumping scale. Hydrol. Process. 26, 855–871.

Bae, D.-H., Jung, I.-W., Lettenmaier, D.P., 2011. Hydrologic uncertainties in climate change from IPCC AR4 GCM simulations of the Chungju Basin, Korea. J. Hydrol. 401, 90–105.

Baguis, P., Roulin, E., Willems, P., Ntegeka, V., 2009. Climate change scenarios for precipitation and potential evapotranspiration over central Belgium. Theor. Appl. Climatol. 99 (3-4), 273–286.

Batelaan, O., De Smedt, F., Triest, L., 2003. Regional groundwater discharge: phreatophyte mapping, groundwater modelling and impact analysis of land-use change. J. Hydrol. 275, 86–108.

Bauwens, A., Sohier, C., Degré, A., 2011. Hydrological response to climate change in the Lesse and the Vesdre catchments: contribution of a physically based model (Wallonia, Belgium). Hydrol. Earth Syst. Sci. 15, 1745–1756.

Bekele, E.G., Nicklow, J.W., 2007. Multi-objective automatic calibration of SWAT using NSGA-II. J. Hydrol. 341 (3–4), 165–176.

Box, G.E.P., Cox, D.R., 1964. An analysis of transformations. J. Roy. Stat. Soc. 26, 211–243 (discussion 244–252).

Boyle, D.P., Gupta, H.V., Sorooshian, S., 2000. Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. 36, 3663–3674.

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., Viney, N.R., 2009. Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM) I: Model intercomparison with current land use. Adv. Water Resour. 32 (2), 129–146.

Bultot, F., Coppens, A., Dupriez, G.L., 1983. Estimation de l'évapotranspiration potentielle en Belgique (Procédure Révisée). Publication, Institute Royal Météorologique. Série A, No. 85, Uccle-Bruxelles, 28 pp.

DHI, 2008. MIKE SHE User Guide. DHI, Water and Environment, Hørsholm, Denmark.

Ferket, B.V.A., Samain, B., Pauwels, V.R.N., 2010. Internal validation of conceptual rainfall–runoff models using baseflow separation. J. Hydrol. 381, 158–173.

Freer, J., Beven, K., Ambroise, B., 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. Water Resour. Res. 32 (7), 2161–2173.

Gosling, S., Taylor, R.G., Arnell, N., Todd, M.C., 2011. A comparative analysis of projected impact of climate change on river runoff from global and catchment-scale hydrological models. Hydrol. Earth Syst. Sci. 15 (1), 279–294.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. 34, 751–763.

Huisman, J.A., Breuer, L., Bormann, H., Bronstert, A., Croke, B.F.W., Frede, H., Gräff, T., Hubrechts, L., Jakeman, A.J., Kite, G.W., Lanini, J., Leavesley, G., Lettenmaier, D.P., Lindström, G., Seibert, J., Sivapalan, M., Viney, N.R., Willems, P., 2009. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) III: Scenario analysis. Adv. Water Resour. 32 (2), 159–170.

Kristensen, K.H., Jensen, S.E., 1975. A model for estimating actual evapotranspiration from potential evapotranspiration. Nord. Hydrol. 6, 70–88.

Laloy, E., Bielders, C.L., 2009. Modelling intercrop management impact on runoff and erosion in a continuous maize cropping system: Part II. Model Pareto multi-objective calibration and long-term scenario analysis using disaggregated rainfall. Eur. J. Soil Sci. 60, 1022–1037.

Li, H., Zhang, Q., 2009. Multiobjective optimization problems with complicated Pareto sets, MOEA/D and NSGA-II. IEEE Trans. Evol. Comput. 13 (2), 284–302.

Liu, Y.B., De Smedt, F., 2004. WetSpa Extension, A GIS-ased Hydrologic Model for Flood Prediction and Watershed Management. Documentation and User Manual, Vrije Universiteit Brussel, Belgium.

Liu, T., Willems, P., Pan, X.L., Bao, A.M., Chen, X., Veroustraete, F., Dong, Q.H., 2011. Climate change impact on water resource extremes in a headwater region of the Tarim basin in China. Hydrol. Earth Syst. Sci. 15, 3511–3527.
Lobligeois, F., Andréassian, V., Perrin, C., Tabary, P., Loumagne, C., 2013. When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation on 3620 flood events. Hydrol. Earth Syst. Sci. Discuss. 10, 12485–12536.
Ludwig, R., May, I., Turcotte, R., Vescovi, L., Braun, M., Cyr, J.-F., Fortin, L.-G., Chaumont, D., Biner, S., Chartier, I., Caya, D., Mauser, W., 2009. The role of hydrological model complexity and uncertainty in climate change impact assessment. Adv. Geosci. 21, 63–71.
Madsen, H., 2000. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. J. Hydrol. 235, 276–288.
Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modeling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216.
Madsen, H., Rasmussen, P.F., Rosbjerg, D., 1997. Comparison of annual maximum series and partial duration series methods for modeling extreme hydrologic events. 1. At-site modeling. Water Resour. Res. 33 (4), 747–757.
Madsen, H., Wilson, G., Ammentorp, H.C., 2002. Comparison of different automated strategies for calibration of rainfall–runoff models. J. Hydrol. 261 (1–4), 48–59.
Maurer, E.P., Brekke, L.D., Pruitt, T., 2010. Contrasting lumped and distributed hydrology models for estimating climate change impacts on California watersheds. J. Am. Water Resour. Assoc. 46 (5), 1024–1035.
Moore, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. Hydrol. Sci. J. 30, 273–297.
Moore, R.J., 2007. The PDM rainfall–runoff model. Hydrol. Earth Syst. Sci. 11 (1), 483–499.
Muleta, M.K., Nicklow, J.W., 2005. Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. J. Hydrol. 306, 127–145.
Najafi, M.R., Moradkhani, H., Jung, I.W., 2011. Assessing the uncertainties of hydrologic model selection in climate change impact studies. Hydrol. Process. 25 (18), 2814–2826.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models, Part I – A discussion of principles. J. Hydrol. 10, 282–290.
Nielsen, S.A., Hansen, E., 1973. Numerical simulation of the rainfall–runoff process on a daily basis. Nord. Hydrol. 4, 171–190.
Pokhrel, P., Gupta, H.V., 2011. On the ability to infer spatial catchment variability using streamflow hydrographs. Water Resour. Res. 47 (8), W08534.
Reed, S., Koren, V., Smith, M.B., Zhang, Z., Moreda, F., Seo, D., Dmipparticipants, A., 2004. Overall distributed model intercomparison project results. J. Hydrol. 298, 27–60.
Refsgaard, J.C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T.A., Drews, M., Hamilton, D., Jeppesen, E., Kjellström, E., Olesen, J.E., Sonnenborg, T.O., Trolle, D., Willems, P., Christensen, J.H., 2014. A framework for testing the ability of models to project climate change and its impacts. Climatic Change 122, 271–282.
Smith, M.B., Koren, V., Zhang, Z., Zhang, Y., Reed, S.M., Cui, Z., Moreda, F., Cosgrove, B.A., Mizukami, N., Anderson, E.A., 2012. Results of the DMIP 2 Oklahoma experiments. J. Hydrol. 418–419, 17–48.
Taye, M.T., Willems, P., 2013. Identifying sources of temporal variability in hydrological extremes of the upper Blue Nile basin. J. Hydrol. 499, 61–70.
Taye, M.T., Ntegeka, V., Ogiramoi, N.P., Willems, P., 2011. Assessment of climate change impact on hydrological extremes in two source regions of the Nile River Basin. Hydrol. Earth Syst. Sci. 15, 209–222.
Thornthwaite, C.W., Mather, J.R., 1955. The Water Balance. Publication in Climatology, vol. 8, pp. 1–104.
Van Steenbergen, N., Willems, P., 2012. Method for testing the accuracy of rainfall–runoff models in predicting peak flow changes due to rainfall changes, in a climate changing context. J. Hydrol. 414–415, 425–434.
Vansteenkiste, T., Tavakoli, M., Ntegeka, V., Willems, P., De Smedt, F., Batelaan, O., 2013. Climate change impact on river flows and catchment hydrology: a comparison of two spatially distributed models. Hydrol. Process. 27 (25), 3649–3662.
Velázquez, J.A., Schmid, J., Ricard, S., Muerth, M.J., Gauvin St-Denis, B., Minville, M., Chaumont, D., Caya, D., Ludwig, R., Turcotte, R., 2012. An ensemble approach to assess hydrological models' contribution to uncertainties in the analysis of climate change impact on water resources. Hydrol. Earth Syst. Sci. 17, 565–578.
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., Willems, P., 2009. Assessing the impact of land use change on hydrology by ensemble modelling (LUCHEM) II: Ensemble combinations and predictions. Adv. Water Resour. 32 (2), 147–158.
Vrugt, J., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S., 2003. Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resour. Res. 39, 1214. http://dx.doi.org/10.1029/2002WR001746.
Wagener, T., Boyle, D.P., Lees, M.J., Wheater, H.S., Gupta, H.V., Sorooshian, S., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci. 5, 13–26.
Wagener, T., McIntyre, N., Lees, M.J., Wheather, H.S., Gupta, H.V., 2003. Towards reduced uncertainty in conceptual rainfall–runoff modeling: dynamic identifiability analysis. Hydrol. Process. 17, 455–476.
Wang, Z.-M., Batelaan, O., De Smedt, F., 1996. A distributed model for water and energy transfer between soil, plants and atmosphere (WetSpa). Phys. Chem. Earth 21 (3), 189–193.
Westerberg, I.K., Guerrero, J.-L., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S., Freer, J.E., Xu, C.-Y., 2011. Calibration of hydrological models using flow-duration curves. Hydrol. Earth Syst. Sci. 15, 2205–2227.
Willems, P., 2009. A time series tool to support the multi-criteria performance evaluation of rainfall–runoff models. Environ. Model. Softw. 24 (3), 311–321.
Willems, P., 2014. 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.. J. Hydrol. 510, 578–590.
Willems, P., Guillou, A., Beirlant, J., 2007. Bias correction in hydrologic GPD based extreme value analysis by means of a slowly varying function. J. Hydrol. 338, 221–236.
Willems, P., Mora, D., Vansteenkiste, Th., Teferi Taye, M., Van Steenbergen, N., 2014. Parsimonious rainfall–runoff model construction supported by time series processing and validation of hydrological extremes – Part 2: Intercomparison of models and calibration approaches. J. Hydrol. 510, 591–609.
Woldeamlak, S.T., Batelaan, O., De Smedt, F., 2007. Effects of climate change on the groundwater system of the Grote Nete catchment, Belgium. Hydrogeol. J. 15 (5), 891–901.
Yu, P.S., Yang, T.C., 2000. Fuzzy multi-objective function for rainfall–runoff model calibration. J. Hydrol. 238, 1–14.
Zhang, H., Huang, G.H., Wang, D., Zhang, X., 2011. Multi-period calibration of a semi-distributed hydrological model based on hydroclimatic clustering. Adv. Water Resour. 34, 1292–1303.