https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Behind the scenes of streamflow model performance

Laurène J. E. Bouaziz[1,2], Guillaume Thirel[3], Tanja de Boer-Euser[1], Lieke A. Melsen[4], Joost Buitink[4], Claudia C. Brauer[4], Jan De Niel[5], Sotirios Moustakas[5], Patrick Willems[5,9], Benjamin Grelier[6], Gilles Drogue[6], Fabrizio Fenicia[7], Jiri Nossent[8,9], Fernando Pereira[8], Eric Sprokkereef[10], Jasper Stam[10], Benjamin J. Dewals[11], Albrecht H. Weerts[2,4], Hubert H. G. Savenije[1], and Markus Hrachowitz[1]

[1]Department of Water Management, 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]Université Paris-Saclay, INRAE, UR HYCAR, 92160, Antony, France
[4]Hydrology and Quantitative Water Management Group, Wageningen University and Research, P.O. Box 47, 6700 AA Wageningen, The Netherlands
[5]Hydraulics division, Department of Civil Engineering, KU Leuven, Kasteelpark Arenberg 40, BE-3001 Leuven, Belgium
[6]Université de Lorraine, LOTERR, F-57000 Metz, 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, Pleinlaan 2, 1050 Brussels, Belgium
[10]Ministry of Infrastructure and Water Management, Zuiderwagenplein 2, 8224 AD Lelystad, The Netherlands
[11]Hydraulics in Environmental and Civil Engineering (HECE), University of Liege, Allée de la Découverte 9, 4000 Liege, Belgium

Correspondence: Laurène Bouaziz (L.J.E.Bouaziz@tudelft.nl)

Abstract. Streamflow is often the only variable used to constrain hydrological models. In a previous international comparison study, eight research groups followed an identical protocol to calibrate a total of twelve hydrological models using observed streamflow of catchments within the Meuse basin. In the current study, we hypothesize that these twelve process-based models with similar streamflow performance have similar representations of internal states and fluxes. We test our hypothesis by comparing internal states and fluxes between models and we assess their plausibility using remotely-sensed products of evaporation, snow cover, soil moisture and total storage anomalies. Our results indicate that models with similar streamflow performance represent internal states and fluxes differently. Substantial dissimilarities between models are found for annual and seasonal evaporation and interception rates, the number of days per year with water stored as snow, the mean annual maximum snow storage and the size of the root-zone storage capacity. Relatively small root-zone storage capacities for several models lead to drying-out of the root-zone storage and significant reduction of evaporative fluxes each summer, which is not suggested by remotely-sensed estimates of evaporation and root-zone soil moisture. These differences in internal process representation imply that these models cannot all simultaneously be close to reality. Using remotely-sensed products, we could evaluate the plausibility of model representations only to some extent, as many of these internal variables remain unknown, highlighting the need for experimental research. We also encourage modelers to rely on multi-model and multi-parameter studies to reveal to decision-makers the uncertainties inherent to the heterogeneity of catchments and the lack of evaluation data.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

1 Introduction

Hydrological models are valuable tools for short-term forecasting of river flows, long-term predictions for strategic water management planning but also to develop a better understanding of the complex interactions of water storage and release processes at the catchment-scale. In spite of the wide variety of existing hydrological models, they mostly include similar functionalities of storage, transmission and release of water to represent the dominant hydrological processes of a particular river basin (Fenicia et al., 2011), differing mostly only in the detail of their parametrizations (Gupta et al., 2012; Gupta and Nearing, 2014; Hrachowitz and Clark, 2017).

In all of these models, each individual model component constitutes a separate hypothesis of how water moves through that specific part of the system. Frequently, the individual hypotheses remain untested. Instead only the model output, i.e. the aggregated response of these multiple hypotheses, is confronted with data of the aggregated response of a catchment to atmospheric forcing. Countless applications of different hydrological models in many different regions across the world over the last decades have shown that these models often provide relatively robust estimates of streamflow dynamics, for both calibration and evaluation periods. However, various combinations of different untested individual hypotheses, can and do lead to similar aggregated outputs, i.e. model equifinality (Beven, 2006; Clark et al., 2016).

To be useful for any of the above applications, it is thus of critical importance that not only the aggregated but also the individual behaviors of the respective hypotheses are consistent with their real-world equivalents. Given the complexity and heterogeneity of natural systems together with the general lack of suitable observations, this remains a major challenge in hydrology (e.g., Jakeman and Hornberger, 1993; Beven, 2000; Gupta et al., 2008; Andréassian et al., 2012).

Studies have addressed the issue by constraining the parameters of specific models through the use of additional data sources besides streamflow. These sources include satellite-based total water storage anomalies (Winsemius et al., 2006; Werth and Güntner, 2010; Yassin et al., 2017), evaporation (Livneh and Lettenmaier, 2012; Rakovec et al., 2016a; Bouaziz et al., 2018; Demirel et al., 2018; Hulsman et al., 2019), near-surface soil moisture (Brocca et al., 2010; Sutanudjaja et al., 2014; Adnan et al., 2016; Kunnath-Poovakka et al., 2016; López López et al., 2017; Bouaziz et al., 2020), snow cover information (Gao et al., 2017; Bennett et al., 2019; Riboust et al., 2019), or a combination of these variables (Nijzink et al., 2018; Dembélé et al., 2020). Reflecting the results of many studies, Rakovec et al. (2016b) showed that streamflow alone is necessary but not sufficient to constrain model components to warrant partitioning of incoming precipitation to storage, evaporation and drainage.

Hydrological simulations are, however, not only affected by model parameter uncertainty, but also by the selection of a model structure and its parameterization (i.e. the choice of equations). Modeling efforts over the last four decades have led to a wide variety of hydrological models providing flexibility to test competing modeling philosophies, from spatially lumped model representations of the system to high-resolution small-scale processes numerically integrated to the catchment scale (Hrachowitz and Clark, 2017). Haddeland et al. (2011) and Schewe et al. (2014) compared global hydrological models and found that differences between models are a major source of uncertainty. Nonetheless, model selection is often driven by personal preference and experience of individual modelers rather than detailed model test procedures (Holländer et al., 2009; Clark et al., 2015; Addor and Melsen, 2019).

A suite of comparison experiments tested and explored differences between alternative modeling structures and parameterizations (Perrin et al., 2001; Reed et al., 2004; Duan et al., 2006; Holländer et al., 2009). However, these studies mostly restricted themselves to analyses of the models' skills to reproduce streamflow ("aggregated hypothesis"), with little consideration for the model internal processes ("individual hypotheses"). The Framework for Understanding Structural Errors (FUSE) was one of the first initiatives towards a more comprehensive assessment of model structural errors, with special consideration given to individual hypotheses (Clark et al., 2008).

Subsequent efforts towards more rigorous testing of competing model hypotheses, partially based on internal processes include Smith et al. (2012a, b) who tested multiple models for their ability to reproduce in-situ soil moisture observations as part of the Distributed Model Intercomparison Project 2 (DMIP2). They found that only two out of sixteen models provided reasonable estimates of soil moisture. In a similar effort, Koch et al. (2016) and Orth et al. (2015) also compared modeled soil moisture to in-situ observations of soil moisture for a range of hydrological models in different environments. In contrast, Fenicia et al. (2008) and Hrachowitz et al. (2014) used groundwater observations to test individual components of their models.

Here, in this model comparison study, we instead use globally available remotely-sensed products to evaluate four different model state and flux variables of twelve process-based models with similar overall streamflow performance, which are calibrated by several research groups following an identical protocol. The calibration on streamflow was conducted in our previous study (de Boer-Euser et al., 2017), in which eight research groups working on the Meuse basin applied their rainfall-runoff model(s) according to a defined protocol using the same forcing data to reduce the degrees of freedom and enable a fair comparison (Ceola et al., 2015). All models had a high overall streamflow performance based on commonly used metrics. We were able to attribute differences in performance to model structure components by focusing on specific hydrological events (de Boer-Euser et al., 2017). Our analyses were then limited to comparisons with hourly streamflow observations and the modeled response of internal processes remained unused.

In a direct follow-up of the above study, we here hypothesize that process-based models with similar overall streamflow performance rely on similar representations of their internal states and fluxes. We test our hypothesis by quantifying the differences in internal states and fluxes that occur between models and assess their plausibility using remotely-sensed estimates of evaporation, snow cover, soil moisture and total water storage anomalies.

2 Study area

We test our hypothesis using data from three catchments in the Belgian Ardennes; all of them are part of the Meuse River basin in North-West Europe: the Ourthe upstream of Tabreux (ID1), the nested Ourthe Orientale upstream of Mabompré (ID2) and the Semois upstream of Membre-Pont (ID3), as shown in Figure 1a,b. The Ardennes Massif and Plateau are underlain by relatively impermeable metamorphic Cambrian rock and Early Devonian sandstone. The pronounced streamflow seasonality of these catchments is driven by high summer and low winter evaporation, as precipitation is relatively constant throughout the year.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

The rain-fed Ourthe River at Tabreux (ID1) is fast-responding due to shallow soils and steep slopes in the catchment. Agriculture is the main land cover (27 % crops and 21 % pasture), followed by 46 % forestry and 6 % urban cover in an area of 1607 km^2 and an elevation ranging between 107 m and 663 m (de Boer-Euser et al., 2017). Mean annual precipitation, potential evaporation and streamflow are 979 mm yr^{-1}, 730 mm yr^{-1} and 433 mm yr^{-1} respectively for the period 2001–2017.

The nested Ourthe Orientale upstream of Mabompré (ID2) is characterized by a narrow elevation range from 294 m to 662 m, making this catchment suitable to analyze snow processes modeled by lumped models. The Ourthe Orientale upstream of Mabompré has an area of 317 km^2 which corresponds to 20 % of the Ourthe area upstream of Tabreux and has similar land cover fractions. Mean annual precipitation, potential evaporation and streamflow for the period 2001–2017 are also relatively similar with 1052 mm yr^{-1}, 720 mm yr^{-1} and 462 mm yr^{-1}, respectively. Snow is not a major component of the water balance, but occurs almost every year with mean annual snow days estimated between 35 and 40 days yr^{-1} (Royal Meteorological Institute Belgium, 2015).

Forest is the main land cover in the Semois upstream of Membre-Pont (ID3) with 56 %, followed by agriculture (18 % pasture and 21 % crop) and 5 % urban cover. The Semois upstream of Membre-Pont is 24 % smaller than the Ourthe upstream of Tabreux with 1226 km^2 and elevation ranges between 176 m and 569 m. Mean annual precipitation, potential evaporation and streamflow are respectively 38 %, 4 % and 46 % higher in the Semois at Membre-Pont with 1352 mm yr^{-1}, 759 mm yr^{-1} and 634 mm yr^{-1}.

3 Data

3.1 Hydrological and meteorological data

Hourly precipitation gauge data are provided by the Service Public de Wallonie (Service Public de Wallonie, 2018) and are spatially interpolated using Thiessen polygons for the period 2000-2017. Daily minimum and maximum temperatures are retrieved from the 0.25° resolution gridded E-OBS dataset (Haylock et al., 2008) and disaggregated to hourly values by linear interpolation using the timing of daily minimum and maximum radiation at Maastricht (Royal Netherlands Meteorological Institute, 2018). Daily potential evaporation is calculated from daily minimum and maximum temperatures using the Hargreaves formula (Hargreaves and Samani, 1985) and is disaggregated to hourly values using a sine function during the day and no evaporation at night. We use the same forcing for 2000–2010 as in the previous comparison study (de Boer-Euser et al., 2017) and follow the same approach to extend the dataset for the period 2011–2017. Observed hourly streamflow data for the Ourthe at Tabreux, Ourthe Orientale at Mabompré and Semois at Membre-Pont are provided by the Service Public de Wallonie for the period 2000–2017.

3.2 Remotely-sensed data

3.2.1 GLEAM evaporation

The Global Land Evaporation Amsterdam Model (GLEAM, Miralles et al., 2011; Martens et al., 2017) provides daily estimates of land evaporation by maximizing the information recovery on evaporation contained in climate and environmental satellite observations. The Priestley and Taylor (1972) equation is used to calculate potential evaporation for bare soil, short canopy and tall canopy land fractions. Actual evaporation is the sum of interception and potential evaporation reduced by a stress factor. This evaporative stress factor is based on microwave observations of vegetation optical depth and estimates of root-zone soil moisture in a multi-layer water-balance model. Interception evaporation is estimated separately using a Gash analytical model driven by observed rainfall. GLEAM v3.3a relies on reanalysis net radiation and air temperature from the European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5 data, satellite and gauge-based precipitation, satellite-based vegetation optical depth, soil moisture and snow water equivalent. The data are available at 0.25° resolution (Figure 1b) and account for subgrid heterogeneity by considering three land cover types. We spatially average the data over the Ourthe catchment upstream of Tabreux for the period 2001–2017.

3.2.2 MODIS Snow Cover

The Moderate Resolution Imaging Spectroradiometer (MODIS) AQUA (MYD10A1, version 6) and TERRA (MOD10A1, version 6) satellites provide daily maps of the areal fraction of snow cover per 500 m × 500 m cell (Figure 1b) based on the Normalized Difference Snow Index (Hall and Riggs, 2016a, b). For each day, AQUA and TERRA observations are merged into a single observation by taking the mean fraction of snow cover per day. The percentage of cells with a fractional snow cover larger than zero and fraction of cells without missing data (i.e. due to cloud cover) for the catchment of the Ourthe Orientale upstream of Mabompré is calculated for each day. For this study, we disregard observations during the summer months (JJA, when temperatures did not drop below 4°C) and only use daily observations in which at least 40 % of the catchment area has snow cover retrievals not affected by clouds, implying that we have 1463 daily observations of mean fractional snow cover over the Ourthe Orientale catchment upstream of Mabompré between 2001 and 2017.

3.2.3 SCATSAR-SWI1km Soil Water Index

SCATSAR-SWI1km is a daily product of soil water content relative to saturation at a 1 km × 1 km resolution (Figure 1b) obtained by fusing spatio-temporally complementary radar sensors (Bauer-Marschallinger et al., 2018). Estimates of the moisture content relative to saturation at various depths in the soil, referred to as Soil Water Index, are obtained through temporal filtering of the 25 km METOP ASCAT near-surface soil moisture (Wagner et al., 2013) and 1 km Sentinel-1 near-surface soil moisture (Bauer-Marschallinger et al., 2018). The Soil Water Index features as single parameter the characteristic time length T (Wagner et al., 1999; Albergel et al., 2008). The T-value is required to convert near-surface soil moisture observations to estimates of root-zone soil moisture. The T-value increases with increasing root-zone storage capacities (Bouaziz et al., 2020), resulting in more smoothing and delaying of the near-surface soil moisture signal. The Copernicus Global Land Service (2019) provides the Soil Water Index for T-values of 2, 5, 10, 15, 20, 40, 60 and 100 days. Since Sentinel-1 was launched in 2014, the Soil Water Index is available for the period 2015–2017 and is spatially averaged over the catchment of the Ourthe upstream of Tabreux.

3.2.4 GRACE Total Water Storage anomalies

The Gravity Recovery and Climate Experiment (GRACE, Swenson and Wahr, 2006; Swenson, 2012) twin satellites launched in March 2002 measure the Earth's gravity field changes by calculating the changes in the distance between the two satellites as they move one behind the other in the same orbital plane. Monthly total water storage anomalies (in cm) relative to the 2004–2009 time-mean baseline are provided at a spatial sampling of 1° (approximately 78 km x 110 km at the latitude of the study region, Figure 1b) by three centers: U. Texas / Center for Space Research (CSR), GeoForschungsZentrum Potsdam (GFZ) and Jet Propulsion Laboratory (JPL). These centers apply different processing strategies and tuning parameters which lead to variations in the gravity fields. We calculate the arithmetic mean of the three estimates to reduce noise in the gravity fields, as suggested by Sakumura et al. (2014). Next, we apply the scaling coefficients provided by NASA to restore some of the signal loss due to processing of GRACE observations (Landerer and Swenson, 2012). The data are then spatially averaged over the catchments of the Ourthe upstream of Tabreux and the Semois upstream of Membre-Pont for the period April 2002 to February 2017.

4 Methods

4.1 Models and Protocol

Eight research groups (Wageningen University, Université de Lorraine, Leuven University, Delft University of Technology, Deltares, Irstea (now INRAE), Eawag and Flanders Hydraulics Research) participated in the comparison experiment and applied one or several hydrological models (Figure 2). The models include WALRUS (Wageningen Lowland Runoff Simulator, Brauer et al., 2014a, b), PRESAGES (PREvision et Simulation pour l'Annonce et la Gestion des Etiages Sévères, Lang et al., 2006), VHM (Veralgemeend conceptueel Hydrologisch Model, Willems, 2014), FLEX-Topo (Savenije, 2010; Gao et al., 2014; Gharari et al., 2014; Euser et al., 2015), a distributed version of the HBV model (Hydrologiska Byråns Vattenbalansavdelning, Lindström et al., 1997), SUPERFLEX M2 to M5 models (Fenicia et al., 2011, 2014), dS2 (distributed simple dynamical systems, Buitink et al., 2019), GR4H (Génie Rural à 4 paramètres Horaire, Mathevet, 2005; Coron et al., 2017, 2019) combined with the CemaNeige snow module (Valéry et al., 2014) and NAM (NedborAfstrommings Model, Nielsen and Hansen, 1973). Main differences and similarities between models in terms of snow processes, root-zone storage, total storage and evaporation processes are summarized in Tables 1-3.

In our previous study (de Boer-Euser et al., 2017), we defined a modeling protocol to limit the degrees of freedom in the modeling decisions of the individual participants (Ceola et al., 2015), allowing us to meaningfully compare the model results.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

The protocol involved to force the models with the same input data and to calibrate them for the same time period, using the same objective functions. However, participants were free to choose a parameter search method, as we considered it to be part of the modelers experience with the model, even if this would make comparison less straightforward. The models were previously calibrated using streamflow of the Ourthe at Tabreux for the period 2004 to 2007, using 2003 as spin-up (de Boer-Euser et al., 2017). The Nash-Sutcliffe efficiencies of the streamflow and the logarithms of the streamflow were simultaneously used as objective functions to select an ensemble of feasible parameter sets and ensure a balance between the models' ability to reproduce both high and low flows. The models were subsequently tested and evaluated for the periods 2001 to 2003 and 2008 to 2010. In addition, by carrying out a proxy-basin differential split-sample test (Klemeš, 1986), not only the models' temporal but also their spatial transferability was tested by applying the calibrated model parameter sets to nested and neighboring catchments for the period 2001 to 2017, using 2000 as spin-up.

In the current study, we run the calibrated models for an additional period from 2011 to 2017 for the Ourthe at Tabreux (ID1), the Ourthe Orientale at Mabompré (ID2) and the Semois at Membre-Pont (ID3). The modeling groups have provided simulation results for each catchment in terms of streamflow, groundwater losses/gains, interception evaporation, root-zone transpiration, total actual evaporation, snow storage, root-zone storage and total storage as a sum of all model storage volumes (Table 2) at an hourly time step for the total period 2001–2017. We compare these modeled states and fluxes and evaluate them against their remotely-sensed equivalents as further explained in Sections 4.2 and 4.3.

4.2 Model evaluation: water balance

All models are evaluated in terms of the long-term water balance, which indicates the partitioning between drainage and evaporative fluxes and allows us to assess long-term conservation of water and energy. We compare mean annual streamflow with observations and mean annual actual evaporation and interception evaporation with GLEAM estimates for the Ourthe at Tabreux during the evaluation period 2008–2017. A detailed description of streamflow performance for specific events (low and high flows, snowmelt event, transition from dry to wet period) has been detailed in the previous paper (de Boer-Euser et al., 2017). In the current study, differences in streamflow dynamics are briefly summarized by assessing observed and modeled baseflow indices (I_{baseflow}, van Dijk, 2010) and flashiness indices ($I_{\mathrm{flashiness}}$, Fenicia et al., 2016), as these are representative of the partitioning of drainage into fast and slow responses. Seasonal dynamics of actual evaporation over potential evaporation and runoff coefficients during winter (Oct-Mar) and summer (Apr-Sep) are compared between models.

4.3 Model evaluation: internal states

We compare modeled snow storage, root-zone soil moisture and total storage between models and with remotely-sensed estimates of MODIS snow cover, SCATSAR-SWI1km Soil Water Index and GRACE total storage anomalies, respectively, as shown in Tables 2-3 and Figure 1c.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

4.3.1 Snow days

As most models used in this study are lumped, it is not possible to spatially evaluate modeled snow cover versus MODIS snow cover. However, we can classify each day in a binary way according to the occurrence of snow, based on a threshold for the percentage of cells in the catchment where snow cover is detected. MODIS snow cover observations are classified as days with and without snow using thresholds of both 5 and 10 % of snow-covered cells in the catchment to be counted as a day with snow, in a sensitivity analysis. For each model, snow days are distinguished from non-snow days whenever the water stored as snow is above 0.05 mm to account for numerical rounding. For each model (and each retained parameter set), we then compare if modeled snow coincides with 'truly' observed snow by MODIS, for each day with a MODIS observation. We create a confusion matrix with counts of true positives when observations and model results agree on the presence of snow (hits), false positives when the model indicates the presence of snow but this is not observed by MODIS (false alarms), false negatives when the model misses the presence of snow observed by MODIS (miss) and true negatives when observations and model results agree on the absence of snow (correct rejections). From this matrix, we calculate the recall as the ratio of hits over actual positives (number of days when snow is observed by MODIS) and the precision as the ratio of hits over predicted positives (number of days when snow is modeled). This allows us to identify, on the one hand, the ratio of days when snow observed by MODIS is correctly identified by the model and, on the other hand, the ratio of days when snow is modeled that are actually observed by MODIS. We therefore not only account for hits, but also for false alarms between model and remotely-sensed observations. We also compare annual maximum snow storage and number of days with snow between the seven models with a snow module (GR4H, M5, NAM, wflow_hbv, M4, FLEX-Topo, WALRUS). The snow analysis is performed in the catchment of the Ourthe Orientale upstream of Mabompré as it features the narrowest elevation range among the study catchments (i.e. 294-662 m a.s.l. versus 108-662 m for the Ourthe upstream of Tabreux) and thus plausibly permits a lumped representation of the snow component.

4.3.2 Root-zone soil moisture

We compare the range of relative root-zone soil moisture ($\overline{S}_{\mathrm{R}} = S_{\mathrm{R}}/S_{\mathrm{R,max}}$) between models for the period in which SCATSAR-SWI1km is available (2015–2017). Time series of catchment-scale root-zone soil moisture are available for all models except WALRUS and dS2 as these models have a combined soil reservoir (Figure 2). The dS2 model only relies on the sensitivity of streamflow to changes in total storage. In WALRUS, the state of the soil reservoir (which includes the root zone) is expressed as a storage deficit and is therefore not bound by an upper limit, allowing groundwater levels to drop infinitely (Table 2). Root-zone storage capacities ($S_{\mathrm{R,max}}$, mm) are available as calibration parameter for all other models. We relate the range in relative root-zone soil moisture to the maximum root-zone storage capacity $S_{\mathrm{R,max}}$, because we expect models with small root-zone storage capacities $S_{\mathrm{R,max}}$ to entirely utilize the available storage, through complete drying and saturation.

We then compare the similarity of the dynamics of modeled time series of the relative root-zone soil moisture with remotely sensed SCATSAR-SWI1km Soil Water Index for several values of the characteristic time length parameter (T in days). The T-value has previously been positively correlated with root-zone storage capacity, assuming a high temporal variability of root-

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

zone soil moisture and therefore a low T-value for small root-zone storage capacities $S_{\mathrm{R,max}}$ (Bouaziz et al., 2020). For each model and feasible realization, we identify the T-value that yields the highest Spearman rank correlation between modeled root-zone soil moisture and Soil Water Index. We then relate the optimal T-value to the root-zone storage capacity $S_{\mathrm{R,max}}$. This analysis enables us to identify potential differences in the representation and the dynamics of root-zone storage between models.

4.3.3 Total storage anomalies

For each model, we calculate time series of total storage (Table 2) and mean monthly total storage anomalies relative to the 2004-2009 time-mean baseline for comparison with GRACE estimates for the Ourthe upstream of Tabreux (ID1) and the Semois upstream of Membre-Pont (ID3). Both catchments coincide with two neighboring GRACE cells, allowing us to test how well the models reproduce the observed spatial variability. We further relate the modeled range of active storage (maximum minus minimum total storage over the time series) to Spearman rank correlation coefficients between modeled and GRACE estimates of total storage anomalies.

4.4 Interactions between storage and fluxes during dry periods

The impact of a relatively large (> 200 mm) versus relatively small (< 150 mm) root-zone storage capacity on actual evaporation, streamflow and total storage is assessed during a dry period in September 2016 by selecting two representative models with high streamflow model performance (GR4H and M5). The plausibility of the hydrological response of these two model representations is evaluated against remotely-sensed estimates of root-zone soil moisture and actual evaporation.

5 Results

5.1 Water balance

5.1.1 Streamflow

All models show high Nash-Sutcliffe Efficiencies of the streamflow and the logarithm of the streamflow ($E_{\mathrm{NS,Q}}$ and $E_{\mathrm{NS,logQ}}$) with median values of above 0.7 for the post-calibration evaluation period 2008–2017 (Figure 3a and Table 2 for the calculation of the Euclidean distances). The interannual variability of streamflow agrees strongly with observations for each model in the period 2008–2017 (Figure 3b). The difference between modeled and observed median streamflow varies between -5.6 % and 5.6 % and the difference in total range varies between -9.6 % and 20 %. This is in line with our results in the previous paper, in which we also showed that all models perform well in terms of commonly used metrics (de Boer-Euser et al., 2017). However, there are differences in the partitioning of fast and slow runoff, as shown by the flashiness and baseflow indices ($I_{\mathrm{flashiness}}$ and I_{baseflow}) in Figure 3c. Largest underestimation of the flashiness index occurs for M2 and dS2 ($\sim$20 %), while FLEX-Topo shows the highest overestimation (26 %). FLEX-Topo and WALRUS underestimate the baseflow index most (41 % and 70 % respectively), while GR4H and M5 show the highest overestimation (15 % and 21 % respectively). There is a strong similarity

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

between modeled and observed hydrographs for one of the best performing models M5, as quantified by its low Euclidean distance (Figure 3d and Table 2). The GR4H model is the only model which includes deep groundwater losses, but they are very limited and represent only 1.6 % of total modeled streamflow or approximately 7 mm yr^{-1} in the Ourthe at Tabreux.

5.1.2 Actual evaporation

Modeled median annual actual evaporation E_A (computed as the sum of transpiration and, if applicable, interception evaporation and sublimation, Table 3) for hydrological years between 2008–2017 varies between 507 and 707 mm yr^{-1} across models, with a median of 522 mm yr^{-1}, which is approximately 10 % lower than the GLEAM estimate of 578 mm yr^{-1}, as shown in Figure 4a. Annual actual evaporation of the VHM model is very high compared to the other models, with a median of 707 mm yr^{-1} and approximates potential evaporation (median of 732 mm yr^{-1}). Calibration of the VHM model is meant to follow a manual stepwise procedure including the closure of the water balance during the identification of soil moisture processes (Willems, 2014). However, in the automatic calibration prescribed by the current protocol, this step was not performed, which explains the unusual high actual evaporation in spite of relatively similar annual streamflow compared to the other models, as there is no closure of the water balance (Figure 3a).

Interception evaporation is included in four models, with GR4H showing the lowest annual interception evaporation of 100 mm yr^{-1} (19 % of E_A or 10 % of P), FLEX-Topo and wflow_hbv have relatively similar amounts of approximatively 250 mm yr^{-1} ($\sim$45 % of E_A or 26 % of P) and NAM has the highest annual interception evaporation of 340 mm yr^{-1} (65 % of E_A or 36 % of P), as shown in Figure 4a. Differences are related to the presence and maximum size of the interception storage (I_max), as shown in Table 3. GLEAM interception estimates of 189 mm yr^{-1} are almost twice as high as GR4H estimates, 25 % lower than FLEX-Topo and wflow_hbv, and 44 % lower than NAM values, suggesting a large uncertainty in the contribution of interception and transpiration to actual evaporation.

GLEAM estimates of actual evaporation show relatively high evaporation rates in winter and are never reduced to zero in summer, as opposed to modeled M5 estimates, as shown in Figure 4b. Deviations between GLEAM and modeled actual evaporation estimates are related to the different assumptions on which they rely. Potential evaporation used as input for our models has a median of 732 mm yr^{-1}, while GLEAM potential evaporation is only 416 mm yr^{-1}. GLEAM interception evaporation is not bounded by potential evaporation as opposed to our models in which $E_\mathrm{A} \leq E_\mathrm{P}$. GLEAM actual evaporation minus interception is 390 mm yr^{-1} or 94 % of potential evaporation, implying almost no water limited conditions, as opposed to our models in which actual evaporation in summer (Apr–Sep) is, due to water stress, reduced to approximatively 73 % of potential evaporation on average for all models except VHM (Figure 4c). Larger differences between models occur in the ratio $E_\mathrm{A}/E_\mathrm{P}$ during winter (Oct–Mar), when FLEX-Topo, wflow_hbv and VHM show $E_\mathrm{A}/E_\mathrm{P}$ ratios close to unity, and dS2 the lowest values of $E_\mathrm{A}/E_\mathrm{P} \sim 0.75$ as shown in Figure 4c. The dS2 model differs from all other models as it relies on a year-round constant water stress coefficient (C_cst), independent of water supply, while the stress coefficient depends on root-zone soil moisture content in all other models (Table 3).

Most models slightly overestimate summer runoff coefficients with values between 0.22 and 0.26 which are very close to the observed value of 0.22, as shown in Figure 4d. During winter, runoff coefficients vary between 0.55 and 0.71, which is close to the observed value of 0.66. This implies a relatively high level of agreement between models in reproducing the medium- to long-term partitioning of precipitation into evaporation and drainage and thus in approximating at least long-term conservation of energy (Hrachowitz and Clark, 2017).

5.2 Internal model states

5.2.1 Snow days

MODIS snow cover is detected over most of the catchment area for some time each year between November 2001 and November 2017, except for the periods of November 2006 to March 2007 and November 2007 to March 2008, when snow is detected in less than half of the catchment cells, as shown in Figure 5a. The number and magnitude of modeled snow storage events varies between models (Figure 5b). The modeled number of snow days per hydrological year varies from ∼28 days for FLEX-Topo, WALRUS and wflow_hbv to ∼62 days for GR4H and ∼90 days for NAM, M4 and M5, as shown in Figure 5c. The variability in median annual maximum snow storage varies from 3 mm for wflow_hbv and ∼5-6 mm for FLEX-Topo and WALRUS to ∼10 mm for GR4H, M4, M5 and 15 mm for NAM. We further evaluate the plausibility of these modeled snow processes by comparing modeled and observed snow cover, for days when a MODIS observation is available.

The presence of snow modeled by FLEX-Topo, wflow_hbv and WALRUS coincides for 94 % with the presence of snow observed by MODIS. However, these models fail to model snow for ∼69 % of days when MODIS reports the presence of snow, implying that these models miss many observed snow days, but when they predict snow, it was also observed (Figure 5d).

NAM, M4 and M5, on the other side, predict the presence of snow which coincides with observed snow by MODIS in ∼76 % of the positive predictions, implying a relatively high probability of false alarm snow prediction of ∼24 %. However, they miss only ∼38 % of actual positive snow days observed by MODIS (Figure 5d). This suggests that these models miss fewer observed snow days, but they also overpredict snow days numbers, which could be related to the use of a single temperature threshold to distinguish between snow and rain, as opposed to a temperature interval in the other models (Table 2).

GR4H is in between the two previously mentioned model categories, with a snow prediction which coincides with observed snow by MODIS in 83 % of the positive predictions and therefore only 17 % of false alarms. The model misses 51 % of actual positive snow days observed by MODIS. GR4H therefore shows a more balanced trade-off between the number of false alarms and the amount of observed snow events missed. This is illustrated in Figure 5d.

With an increased threshold to distinguish snow days in MODIS, from 0.05 to 0.10 as fraction of cells in the catchment with a detected snow cover (Figure 5d and Figure 5e respectively), we decrease the number of observed snow days. For all models, this leads to an increase in the ratio of false alarms over predicted snow days but also a decrease of the ratio of missed events over actual observed snow days by MODIS. However, as all models are similarly affected by the change in threshold, our findings on the differences in performance between models show little sensitivity to this threshold.

Despite the large variability in snow response between models, snow processes are represented by a degree-hour method in all models, suggesting a high sensitivity of the snow response to the snow process parametrization (Table 2).

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

5.2.2 Root-zone soil moisture

Vegetation accessible water volumes that can be stored in the root zone largely control the long-term partitioning of precipitation into evaporation and drainage. Most hydrological models include a representation of this root-zone storage capacity $S_{\mathrm{R,max}}$, which is estimated through calibration (Table 2). The calibrated root-zone storage capacities vary between 74 mm and 277 mm across studied models. The root-zone soil moisture content relative to saturation of models with relatively large root-zone storage capacities (here defined as $S_{\mathrm{R,max}} > 200$ mm) tends to never fully dry out (>0.20) and saturate (<0.94) as opposed to models with lower root-zone storage capacities ($S_{\mathrm{R,max}}$ <150 mm), in which the storage tends to either dry out completely and/or fully saturate (Figure 6a). If the vegetation accessible water storage dries out, this will lead to water stress and reduced transpiration. On the other hand, if the root-zone storage is saturated, no more water can be stored, resulting in fast drainage. The size of the root-zone storage capacity is therefore a key control of the hydrological response, allowing us to explain some of the observed variability between models.

We first compare the *ranges* of modeled and remotely-sensed estimates of relative root-zone soil moisture. The range of SCATSAR-SWI1km Soil Water Index (SWI) varies between 0.29 and 0.82 for a value of the characteristic time length (T-value) of 15 days and the range reduces as the T-value increases (Figure 6b). High T-values represent a low variability of soil moisture from one timestep to another and are associated with large root-zone storage capacities (Bouaziz et al., 2020). The range of relative root-zone soil moisture content of the SCATSAR-SWI1km data is smaller than obtained by the models. However, data matching techniques are often used to rescale satellite data to match the variability of modeled or observed data (Brocca et al., 2010). The need for this rescaling suggests the difficulty to meaningfully compare the range of modeled and remotely-sensed estimates of root-zone soil moisture content relative to saturation.

We then compare the *dynamics* of modeled and remotely-sensed estimates of root-zone soil moisture by calculating Spearman rank correlations between modeled root-zone soil moisture and remotely-sensed estimates of the Soil Water Index for the available T-values of 2, 5, 15, 20, 40, 60 and 100 days. As the T-value increases, the Soil Water Index is more smoothed and delayed. For each model realization, we identify the T-value which yields the highest Spearman rank correlation between Soil Water Index and modeled root-zone soil moisture (Figure 6c). The optimal T-value increases with the size of the calibrated root-zone storage capacity and varies between 15 and 60 days. A small root-zone storage capacity is indeed likely to fill through precipitation and empty through evaporation and drainage more rapidly than a large water storage capacity, leading to a higher temporal relative soil moisture variability. The mismatch between the relatively high root-zone storage capacities of VHM ($S_{\mathrm{R,max}} \sim 200$ mm) in relation to the relatively low optimal T-values of 20 days is likely related to the unclosed water balance (Section 5.1.2). The similarity between modeled root-zone soil moisture and Soil Water Index with optimal T-values is high, as implied by Spearman rank correlations varying between 0.88 and 0.90 across models. However, the disparity in optimal T-values between models underlines the different temporal representations of root-zone soil moisture content across models, implying that all these models cannot simultaneously provide a plausible representation of the catchment-scale vegetation accessible water content.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

5.2.3 Total storage anomalies

Total water storage anomalies obtained from GRACE are compared to the storage as simulated by the models, showing relatively similar seasonal patterns, as illustrated in Figure 7a for model M5. GRACE total storage anomalies of the Semois upstream of Membre-Pont and the Ourthe upstream of Tabreux are mainly represented by two neighboring cells (Figure 7b), allowing us to test how models represent the observed spatial variability. The range of anomalies in the Semois upstream of Membre-Pont (-90–117 mm) is larger than in the Ourthe upstream of Tabreux (-78–97 mm), implying 13 % less summer and 21 % more winter storage in the Semois upstream of Membre-Pont (Figure 7c). Median precipitation is also 37 % higher in the Semois upstream of Membre-Pont than in the Ourthe upstream of Tabreux during winter months (Oct-Mar), but relatively similar during summer months (Apr-Sep), as shown in Figure 7d. This difference in precipitation potentially leads to a wider range of modeled anomalies in the Semois upstream of Membre-Pont than in the Ourthe upstream of Tabreux for all models, as shown in Figure 7e. This implies that all models reproduce the spatial variability between both catchments observed by GRACE. As the models were calibrated for the Ourthe at Tabreux and parameter sets were transferred to the Semois upstream of Membre-Pont, the forcing data is the main difference to explain the modeled spatial variability.

The models are also able to represent the observed temporal dynamics of total storage anomalies, as suggested by Spearman rank correlation coefficients ranging between 0.62 and 0.80 for the Ourthe upstream of Tabreux (Figure 7f). There is, however, no relation between the Spearman rank correlations of the anomalies and the total modeled storage range (difference between maximum and minimum values), as shown in Figure 7f. PRESAGES, WALRUS, VHM and dS2 have the largest ranges of total modeled storage, varying between 260 and 280 mm and are also characterized by a relatively large root-zone storage capacities (PRESAGES, VHM) or no separate root zone (WALRUS dS2), while the total storage range of all other models is between 200 and 220 mm. The similarity in total storage range between most models is likely related to the identical forcing data and the similarity in the long-term partitioning of precipitation into drainage and evaporation (Section 5.1.2). However, the absolute values of total storage during a specific event or the partitioning in internal storage components may vary between models (Section 5.3).

5.3 Interactions between storage and fluxes during dry periods

As previously seen in Figure 6a, the relative root-zone soil moisture content of the GR4H model is always above 0.2 for the three years for which SCATSAR-SWI1km data are available, as opposed to M5 which fully dries out for some time during the summers of 2015–2017 (Figure 8a,b). Actual evaporation in M5 is also strongly reduced during these dry soil moisture periods unlike GR4H, as shown in Figure 8c,d. When zooming into the dry period around September 2016, Figure 8e,f shows median relative root-zone soil moisture in GR4H of ∼0.24 versus ∼0.01 for M5, while SCATSAR-SWI1km has a higher median value of ∼0.55 (for both optimal T-values of 20 and 40 days). The dryness of root-zone soil moisture in M5 leads to median daily evaporation of 0.8 mm d^{-1} against 1.3 mm d^{-1} for GR4H and prolonged periods of almost zero evaporation in M5 (e.g. 31/08–03/09, 09/09–15/09 and 22/09–30/09), while this neither occurs in GR4H nor in GLEAM actual evaporation, as shown in Figure 8g,h. The model M5 has one of the highest streamflow performance (Figure 3, Table 2), however, this yearly vegetation water stress is unlikely as ecosystems have adapted to overcome dry periods with a certain return period (Milly, 1994; Gentine et al., 2012; Gao et al., 2014; Wang-Erlandsson et al., 2016; de Boer-Euser et al., 2016; Nijzink et al., 2016). High streamflow performances, therefore, do not warrant the plausibility of internal process representation. Despite the dried-out root-zone storage in M5, there is still water available in the slow storage to sustain a baseflow close to observed values, as shown in Figure 8j,l. The streamflow responses of GR4H and M5 are both close to observations (Figure 8i,j) in spite of differences in storage and evaporation, suggesting different internal process representations for a similar aggregated streamflow response during a low flow period.

5.4 Summary of internal process representations

We summarize the differences between models in terms of their fluxes and states in Figure 9. Mean annual streamflow shows the highest degree of similarity between models, as this variable was used for calibration of the models. However, even if mean annual volumes are similar between models, the dynamics in terms of the partitioning into fast and slow runoff, as characterized by the flashiness and baseflow indices, already show a large variability. These differences suggest that calibration using Nash-Sutcliffe efficiencies of the streamflow and logarithm of the streamflow ensures a correct representation of the overall behavior but does not adequately constrain the partitioning between fast and slow runoff.

Mean annual actual evaporation is relatively similar for all models except VHM due to the issue of the unclosed water balance. Only one third of the models accounts for a separate interception module and mean annual interception rates largely differ between them, implying that root-zone transpiration compensates for over- or underestimation of interception to end up with a similar actual evaporation. Differences in transpiration rates are, therefore, partly explained by the presence or absence of an interception module.

A large variability also exists in modeled snow processes, which are represented in 60 % of the studied models. Maximum annual snow storage is likely underestimated by some models and overestimated by others, which underlines the high sensitivity of snow processes parametrization for snow accumulation and melt, despite similar forcing and degree-hour method (Table 2).

Models with relatively large ranges of simulated root-zone soil moisture content also show large ranges of total storage, which is related to the size of the root-zone storage capacity. The smaller root-zone storage capacity of model M5 compared to models M2-M4 leads to a smaller range of relative root-zone soil moisture, while the range of total storage is similar. This is likely due to the additional slow groundwater reservoir added in model M5 compared to M2-M4, which splits the available storage between the root-zone storage and the additional groundwater store. The smaller root-zone storage capacity further affects evaporation dynamics in summer and increases the baseflow index of M5 compared to M2-M4. This highlights the complex interactions in internal dynamics even in parsimonious lumped models with similar mean annual streamflow performance.

6 Discussion

6.1 Implications

While streamflow alone may be used to evaluate hydrological models, we subsequently use these models to understand internal states and fluxes in current and future conditions (Alcamo et al., 2003; Hagemann et al., 2013; Beck et al., 2017). Our findings show that similar streamflow responses obtained by models calibrated according to an identical protocol rely on different internal process representations. In other words, we might get the right answers but for the wrong reasons (Kirchner, 2006), as these models cannot at the same time all be right and different from each other (Beven, 2006).

Almost all models show a similar long-term partitioning of precipitation into drainage and evaporation, as they are forced and constrained by the same data, also leading to relatively similar volumes of total storage. However, the partitioning of total storage in several internal storage components differs between models, resulting in distinct runoff responses as expressed by the baseflow and flashiness indices.

Largest differences between models occur for snow processes, interception evaporation, reduction of winter evaporation and root-zone soil moisture. However, these processes either play a limited role on the overall water balance or can be compensated by other processes. Snow occurs every year but is not a major component of the streamflow regime (de Wit et al., 2007), interception evaporation can be compensated by root-zone evaporation, and very dry periods only occur for several weeks per year when streamflow is already very low.

The presence of interception or a slow storage (absent in M2-M4 but added in M5) affects the representation of other internal processes, including transpiration and/or root-zone soil moisture, implying that individual internal model components are altered by the presence/absence of other potentially compensating processes. Adding an additional internal model component changes the internal representation of water storage and fluxes through the system, which should be kept in mind if model parameters were to be fixed in alternative model structures. Furthermore, model improvements through additional process components and/or adapted parametrization should not only be evaluated in terms of the aggregated response, but also in the partitioning of fluxes and storages through the system (e.g. does the groundwater component improve the baseflow index at the expense of the availability of root-zone soil moisture during dry periods?). Models should be confronted with expert knowledge, e.g. on the occurrence of days with water stress or snow storage, to assess the realism of internal states and fluxes (Gharari et al., 2014; Hrachowitz et al., 2014; van Emmerik et al., 2015).

Applying these models to a future, more extreme climate in the same region might lead to contrasting insights regarding impacts of climate change, as also shown by studies of Hagemann et al. (2013), Melsen et al. (2018) and de Niel et al. (2019) in which model structures may lead to different signs of change of mean streamflow. Using one model or the other to assess the effect of rising temperatures on snow could lead to very different time scales of snow storage decline. Vegetation already experiences more intense water stress in some models compared to others and this would be exacerbated in more extreme drought scenarios (Melsen and Guse, 2019). More intense precipitation events could affect interception evaporation and therefore water availability in the root-zone differently from one model to another. Beyond model structure, the experience

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

each modeler has with its model and associated calibration procedure to constrain model parameters may also impact the simulation results (Melsen et al., 2019).

Our findings should, therefore, encourage modelers to use multiple data sources for model calibration and evaluation, as already suggested by many other studies (Samaniego et al., 2010; Rakovec et al., 2016a; Koch et al., 2018; Stisen et al., 2018; Nijzink et al., 2018; Veldkamp et al., 2018; Dembélé et al., 2020). Remotely-sensed estimates of soil moisture, evaporation and total storage anomalies are available at the global scale and in spite of potential biases with models, the temporal dynamics are useful to constrain our models (McCabe et al., 2017; Sheffield et al., 2018). Additionally, it seems essential to support decision-makers by studies relying on multi-model and multi-parameter systems, as also suggested by Haddeland et al. (2011) and Schewe et al. (2014), to reveal uncertainties inherent to the heterogeneous hydrological world (Beven, 2006; Savenije, 2010; Samaniego et al., 2010; Hrachowitz and Clark, 2017).

This study is the result of a joint research effort of institutes and universities gathering each year in Liège to exchange knowledge and work together on the Meuse basin. These international studies favor a close collaboration between scientists that can learn from each other to accelerate modeling advances (Archfield et al., 2015). Another advantage of comparing modeling results of several research groups is to quickly detect small mistakes in the modeling process, including shifts in the time series or using forcing data of one catchment to model another catchment.

6.2 Knowledge gaps

Many aspects of the hydrological response remain unknown and can hardly be evaluated against observations. While in-situ observations of snow, evaporation or soil moisture are rarely available at sufficient spatio-temporal scale, remotely-sensed estimates have the advantage of high spatial resolution, though they often rely on models themselves. The evaporation components estimated by GLEAM rely on different forcing data than used by our models, a different equation is used to calculate potential evaporation and interception is calculated separately using the Gash analytical model (Gash, 1979). Despite relatively similar mean annual actual evaporation between GLEAM and our models, the processes behind them rely on different internal representations. The ratio of actual over potential evaporation as a result of water stress at the catchment-scale, therefore, remains highly uncertain (Coenders-Gerrits et al., 2014; Mianabadi et al., 2019).

Measurements of the fraction of interception evaporation over precipitation depend on the site location with estimates of 37 % for a Douglas fir stand in the Netherlands (Cisneros Vaca et al., 2018), 27 %, 32 % and 42 % for three coniferous forests of Great Britain (Gash et al., 1980) and 50 % for a forest in Puerto Rico (Schellekens et al., 1999) and are difficult to extrapolate to other catchments due to the heterogeneity and complexity of natural systems.

The gravity fields seen by GRACE require smoothing of the noise induced by attenuated short wavelength. This spatial smoothing decreases the already coarse GRACE resolution even further and, therefore, signals in one location spread out into nearby regions (Bonin and Chambers, 2013). This leakage of signal from one region to the other implies that at least some of the observed differences between the Semois and Ourthe signals may originate from outside the two cells.

While areal fractions of snow cover can be estimated by MODIS, knowledge of catchment-scale snow water equivalent is lacking. If remotely-sensed estimates of near-surface soil moisture saturation are available, root-zone water content remains

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Open Access
Discussions

uncertain and while GRACE provides estimates of total storage anomalies, we lack knowledge on absolute total water storage. The spatial variability and the temporal dynamics of these remotely-sensed products provide useful, additional, independent information to understand the hydrological puzzle, but certainly not all the answers to evaluate the states typically included in process-based models. Measurements are, therefore, of crucial importance to increase our understanding of hydrological processes at the catchment-scale, which in turn will improve the quality of remotely-sensed products and model development (Vidon, 2015; Burt, T. P., McDonnell, 2015; van Emmerik et al., 2018).

7 Conclusions

Similar streamflow performance of process-based models, calibrated following an identical protocol, relies on different internal process representations. Most models are relatively similar in terms of the long-term partitioning of precipitation into drainage and evaporation. However, the partitioning between transpiration and interception, snow processes and the representation of root-zone soil moisture varies significantly between models, suggesting variability of water storage and release through the catchment. The comparison of modeled states and fluxes with remotely-sensed estimates of evaporation and root-zone soil moisture suggests that models with relatively small root-zone storage capacities lead to unrealistic drying-out of the root-zone storage and significant reduction of evaporative fluxes each summer. The dissimilarity in internal process representations between models implies that they are not necessarily providing the right answers for the right reasons, as they cannot simultaneously be close to reality and different from each other. While the consequences for streamflow may be limited for the historical data, the differences may exacerbate for more extreme conditions or climate change scenarios. Considering the uncertainty of process representation behind the scenes of streamflow performance and our lack of knowledge and observations on these internal processes, we invite modelers to evaluate their models using multiple variables, we encourage more experimental research, and highlight the value of multi-model multi-parameter studies to support decision making.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

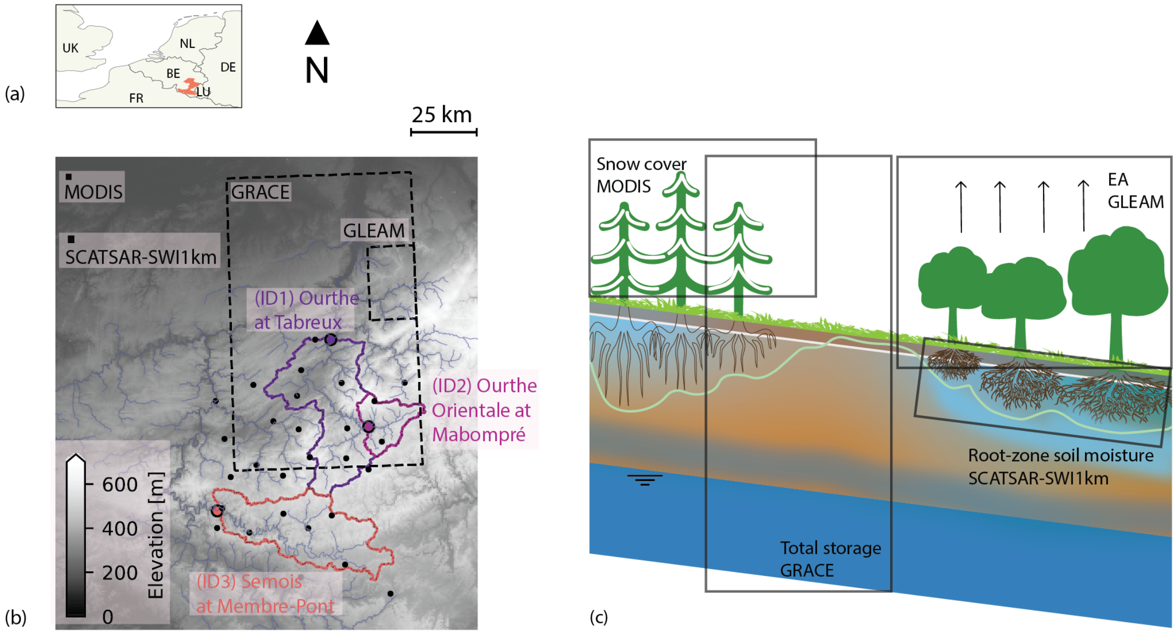

Figure 1. (a) Location of the study catchments in Belgium, North-West Europe. **(b)** Digital elevation model and catchments of the Ourthe upstream of Tabreux (ID1), Ourthe Orientale upstream of Mabompré (ID2) and Semois upstream of Membre-Pont (ID3). Pixel size of GRACE, GLEAM, MODIS and SCATSAR-SWI1km. Colored dots are the streamflow gauging locations and black dots are the precipitation stations. **(c)** Perceptual overview of the link between studied fluxes and states and remotely-sensed products.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

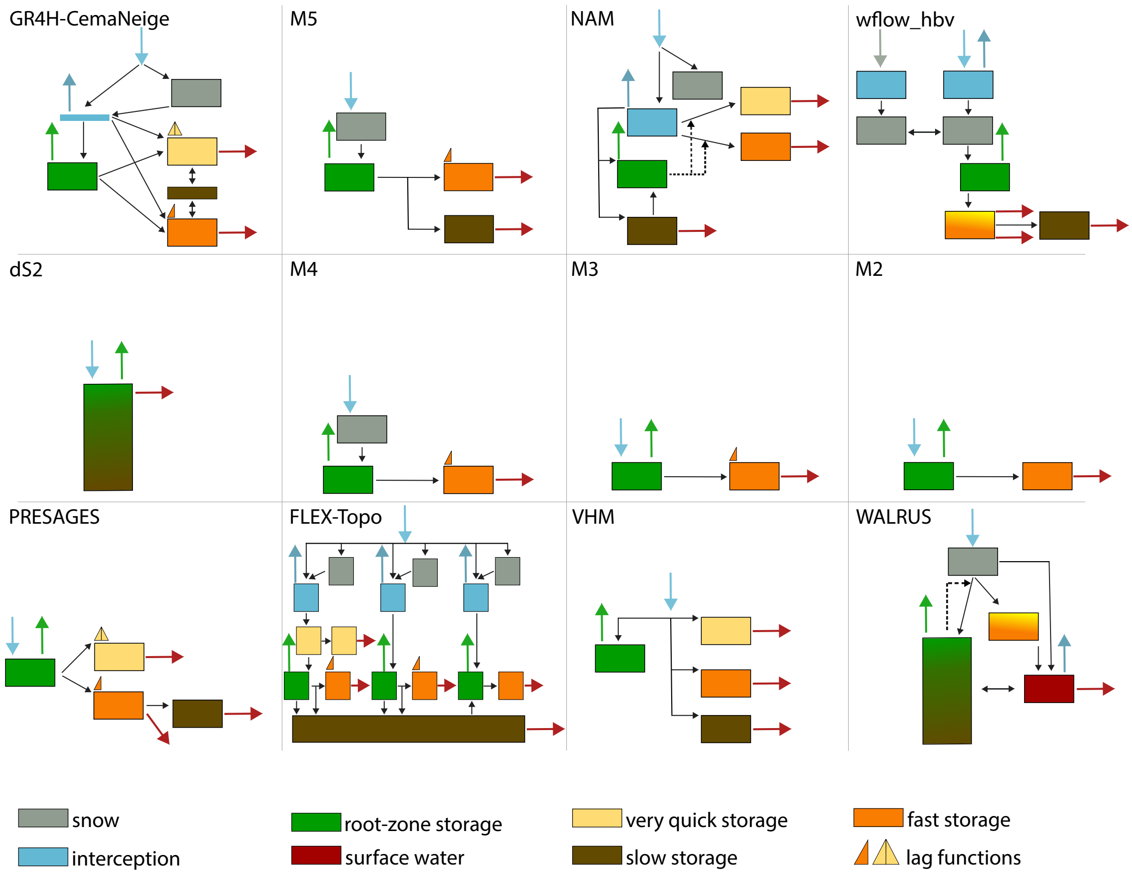

Figure 2. Simplified schematic overview of 12 model structures (adapted from de Boer-Euser et al., 2017) with the aim to highlight similarities and differences between the models. Solid arrows indicate fluxes between stores, while dashed arrows indicate the influence of a state to a flux. Colored arrows indicate incoming or outgoing fluxes, whereas black arrows indicate internal fluxes. The narrow blue rectangle in GR4H indicates the presence of an interception module without interception storage capacity (Table 3). Storages with a color gradient indicate the combination of several components in one reservoir.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

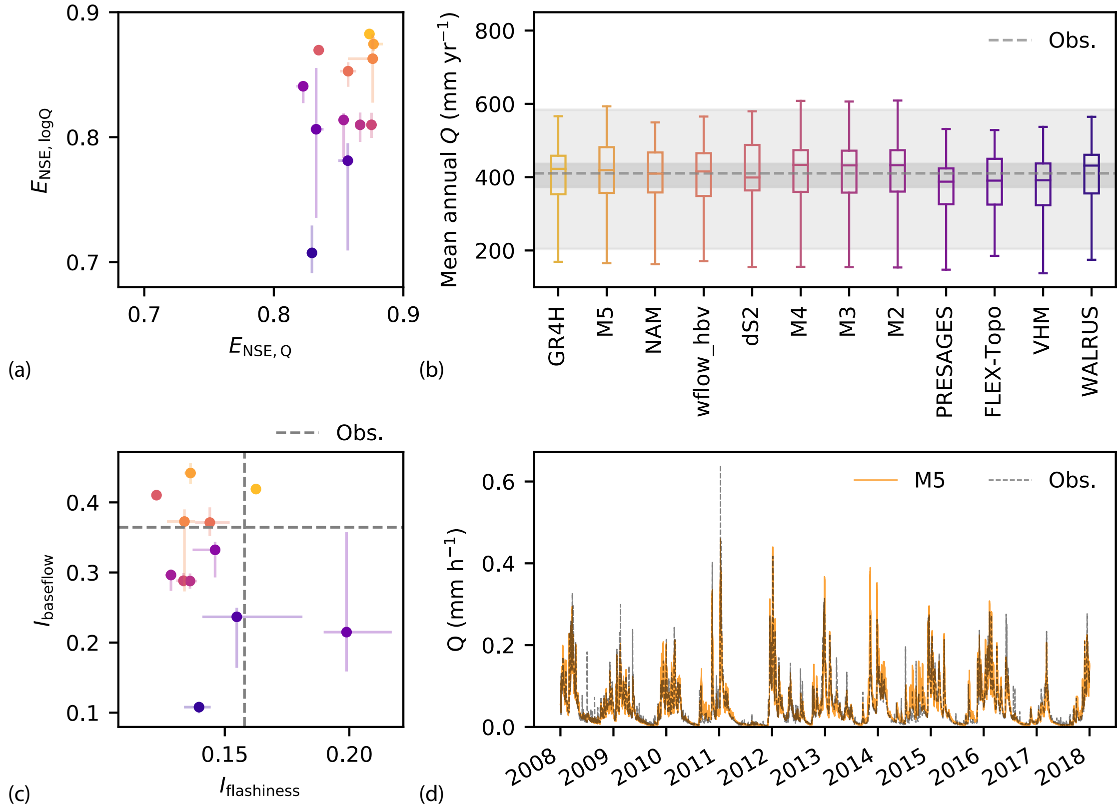

Figure 3. Evaluation of modeled streamflow performance for the Ourthe at Tabreux for the period 2008–2017. **(a)** Nash-Sutcliffe Efficiencies of the streamflow $E_{\mathrm{NS,Q}}$ and the logarithm of the streamflow $E_{\mathrm{NS,logQ}}$ (median, 25/75$^{\mathrm{th}}$ percentiles across parameter sets). **(b)** Modeled mean annual streamflow for hydrological years between 2008–2017 across feasible parameter sets. The models are ranked from the highest to the lowest performance according to the Euclidean distance of streamflow performance (see Table 2). The dashed line and grey shaded areas show median, 25/75$^{\mathrm{th}}$ and minimum-maximum range of observed mean annual streamflow. **(c)** Baseflow index I_{baseflow} as a function of the flashiness index $I_{\mathrm{flashiness}}$ (median, 25/75$^{\mathrm{th}}$ percentiles across parameter sets). Observed values are shown by the grey dashed lines. **(d)** Observed and modeled hydrographs of model M5 with high streamflow model performance (low Euclidean distance).

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

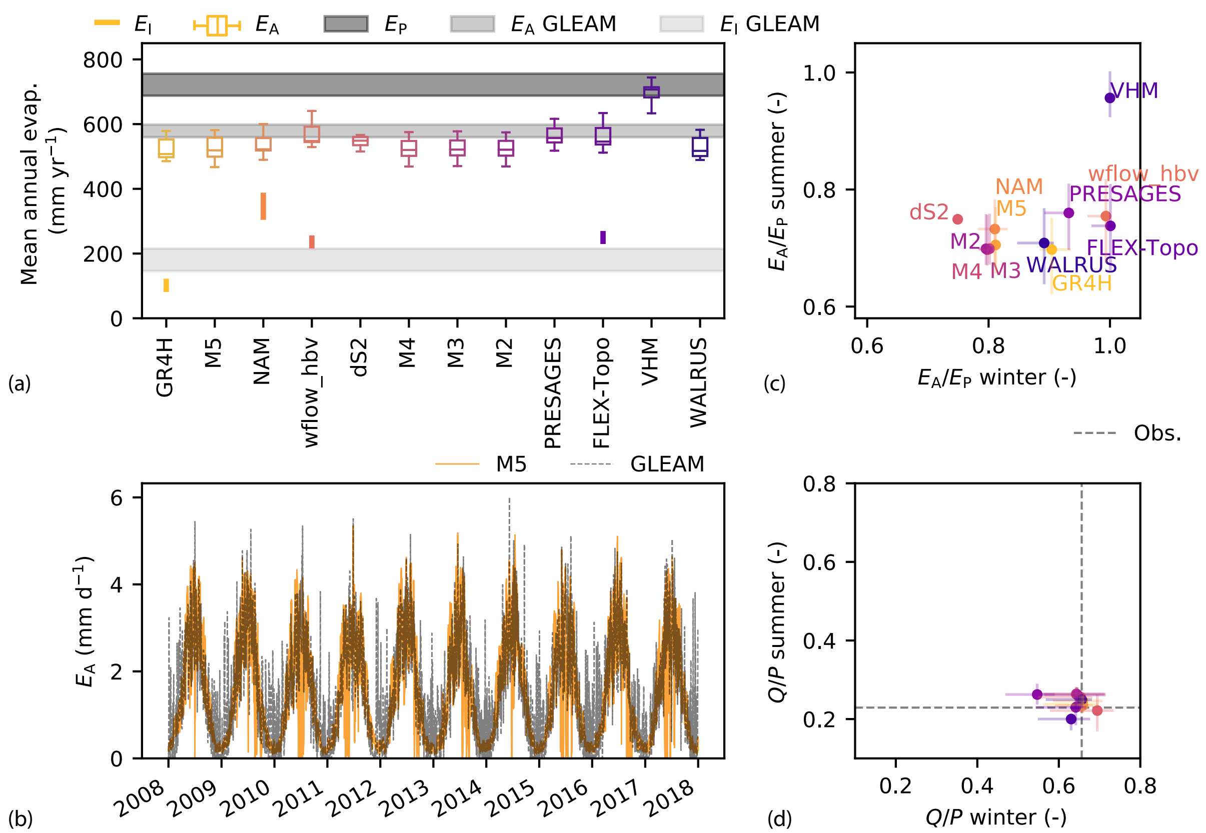

Figure 4. Evaluation of modeled evaporation for the Ourthe upstream of Tabreux for the period 2008–2017. **(a)** Modeled mean annual actual evaporation E_A and interception evaporation E_I for hydrological years between 2008–2017 across feasible parameter sets. The dark grey shaded area shows the range of potential evaporation E_P used as input for the models. The light grey shaded area shows GLEAM actual and interception evaporation. **(b)** Daily actual evaporation from GLEAM and modeled by the M5 model. **(c)** Summer against winter $E_\mathrm{A}/E_\mathrm{P}$ ratios for each model (median and 25/75$^{\mathrm{th}}$ percentiles across parameter sets). **(d)** Summer against winter runoff coefficient Q/P for each model (median and 25/75$^{\mathrm{th}}$ percentiles across parameter sets), plotted on the same scale. The dashed grey lines indicate the observed median runoff coefficients in summer and winter.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

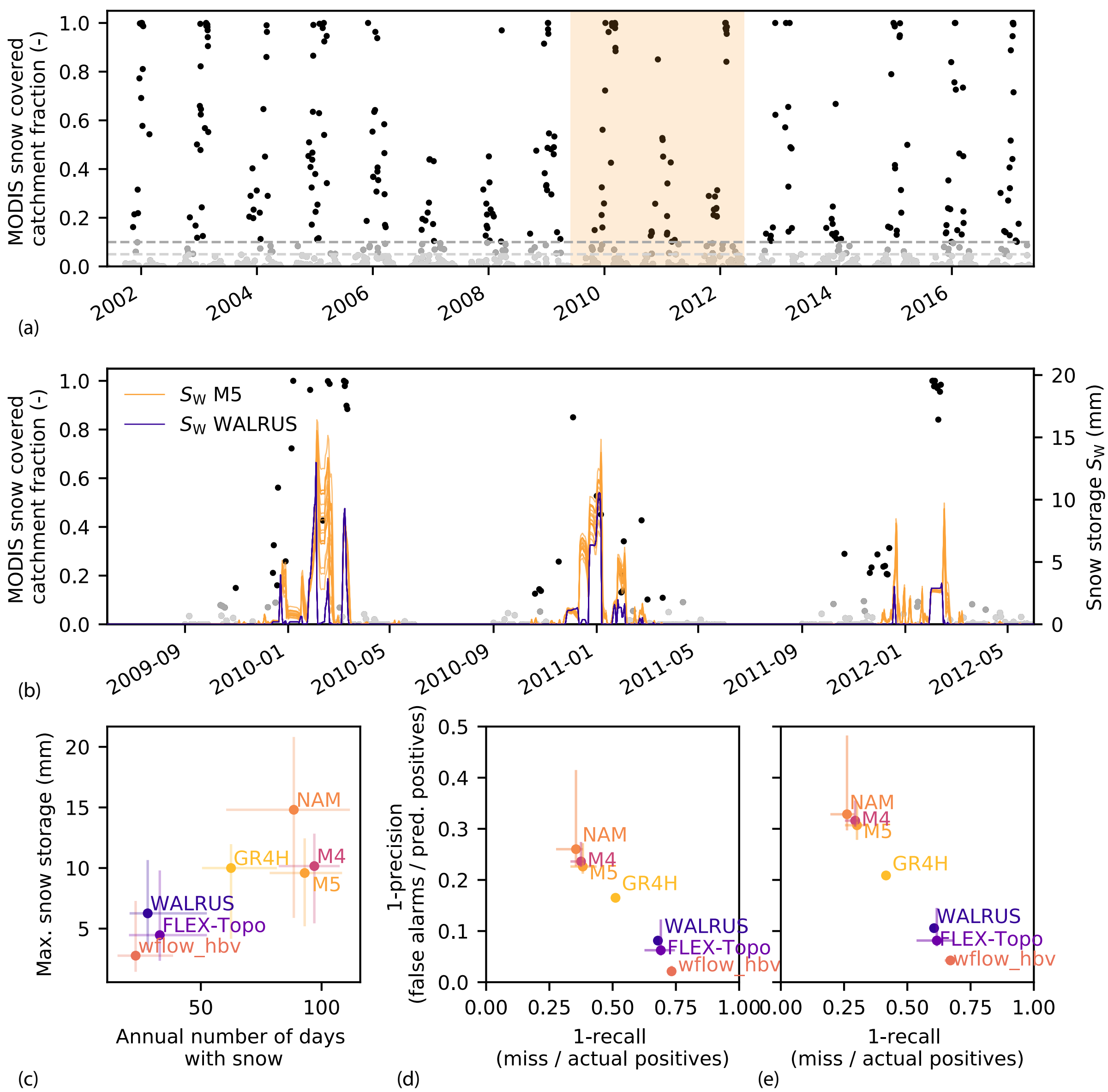

Figure 5. **(a)** Fraction of cells with a MODIS areal fraction snow cover greater than zero in the Ourthe Orientale upstream of Mabompré for the period 2001–2017. MODIS data are available once every three days on average. The dashed lines show the two thresholds of 0.05 and 0.10 selected to distinguish snow days. **(b)** Modeled snow storage for two contrasting models M5 and WALRUS for the light orange shaded period. **(c)** Median annual maximum snow storage as a function of number of days per year with snow. Light (yellowish) colors indicate models with higher performance (lower Euclidean distances). The vertical and horizontal error bars indicate the 25/75$^{\text{th}}$ percentiles of feasible parameter sets **(d,e)** Two-dimensional representation of the false alarm over predicted positives ratio (1-precision) as a function of miss over actual positives ratio (1-recall) when applying a threshold of **(d)** 0.05 and **(e)** 0.10 as fraction of cells with snow cover greater than zero. In this representation, the perfect model would be at the origin (100 % hits and 0 % false alarms). The vertical and horizontal error bars indicate the uncertainty within feasible parameter sets.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

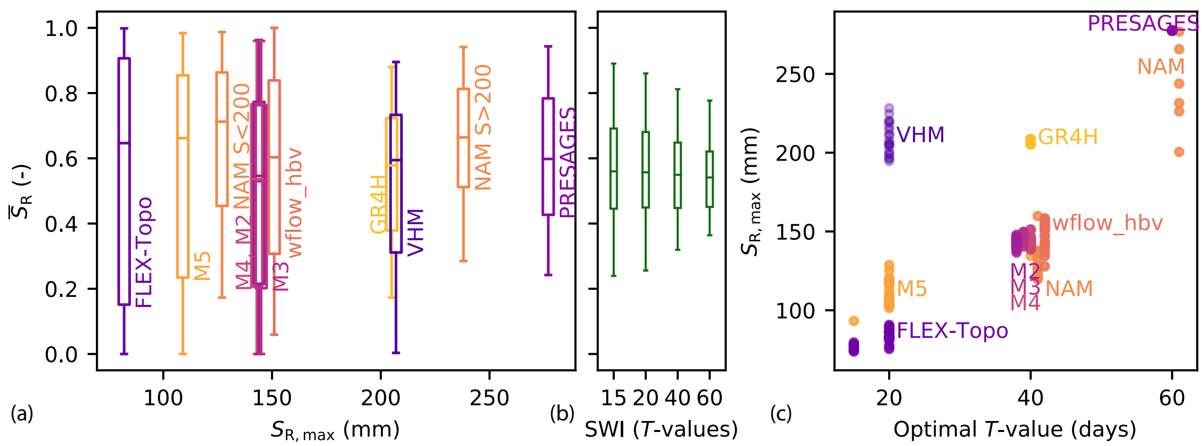

Figure 6. (a) Range of relative root-zone soil moisture $\overline{S}_{\mathrm{R}}$ in the Ourthe upstream of Tabreux for the period 2015–2017 as a function of the median root-zone storage capacity ($S_{\mathrm{R,max}}$) across parameter sets. The feasible parameters for NAM are split in two groups due to the large variability of $S_{\mathrm{R,max}}$ (subsets with $S_{\mathrm{R,max}}$ of ∼130 mm and ∼240 mm). **(b)** Range of the SCATSAR-SWI1km Soil Water Index for several values of the characteristic time length T (days) for the period when SCATSAR-SWI1km is available (2015–2017). **(c)** Root-zone storage capacity $S_{\mathrm{R,max}}$ as a function of the optimal T-value for each model realization. Optimal T-values are derived at the highest Spearman rank correlation between Soil Water Index and modeled root-zone soil moisture.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

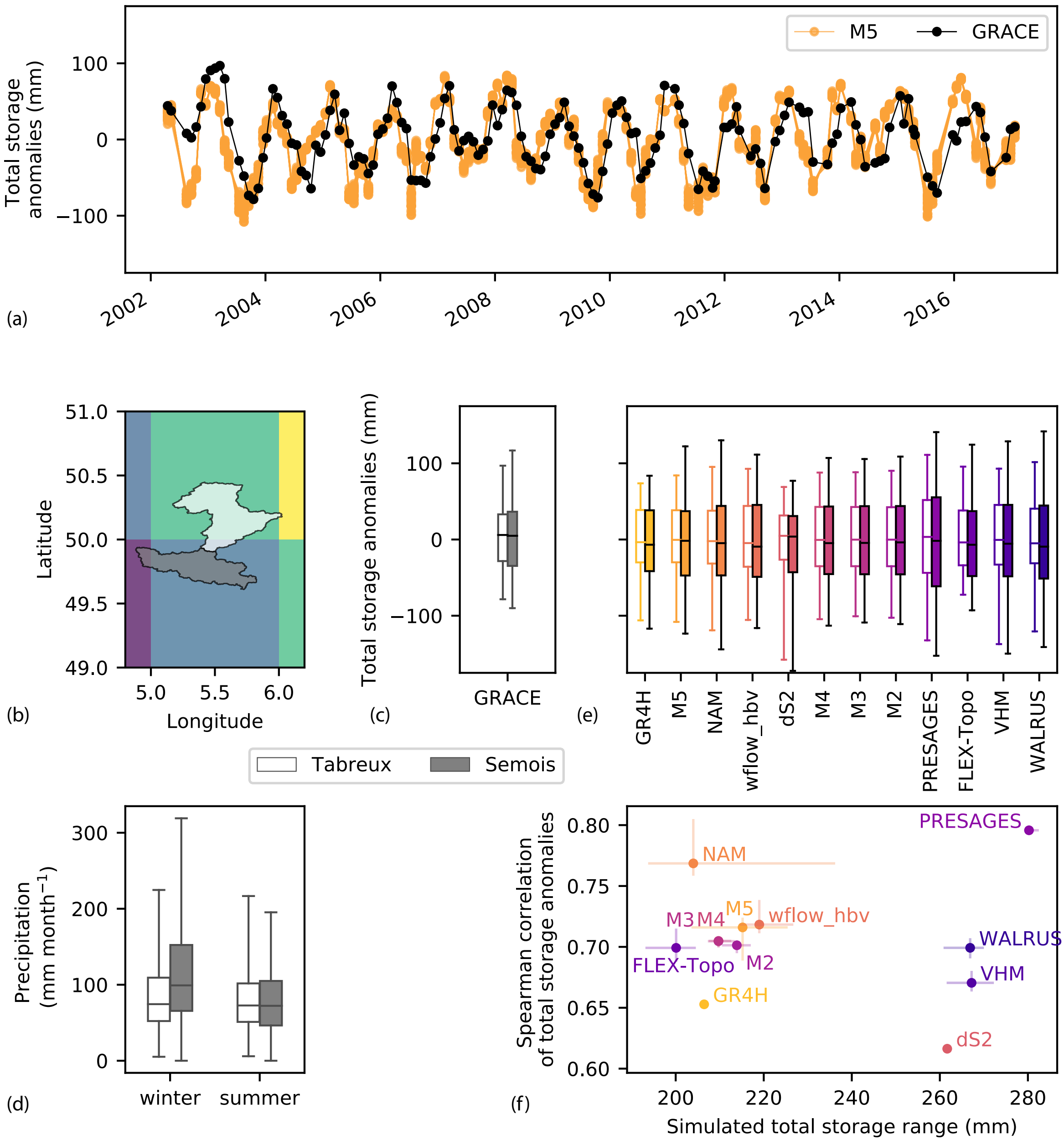

Figure 7. (a) Total storage anomalies modeled by M5 and compared to GRACE for the Ourthe upstream of Tabreux. **(b)** Catchments of the Ourthe upstream of Tabreux (light grey) and Semois upstream of Membre-Pont (dark grey) located in two neighboring GRACE cells. **(c)** Range of GRACE total storage anomalies for the Semois upstream of Membre-Pont compared to the Ourthe upstream of Tabreux for the period 2001–2017. **(d)** Mean monthly precipitation during winter and summer months in the Semois upstream of Membre-Pont compared to the Ourthe upstream of Tabreux. **(e)** Modeled total storage anomalies for both catchments. **(f)** Spearman rank correlations between GRACE and modeled total storage anomalies as a function of the range of modeled total storage for the Ourthe upstream of Tabreux.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

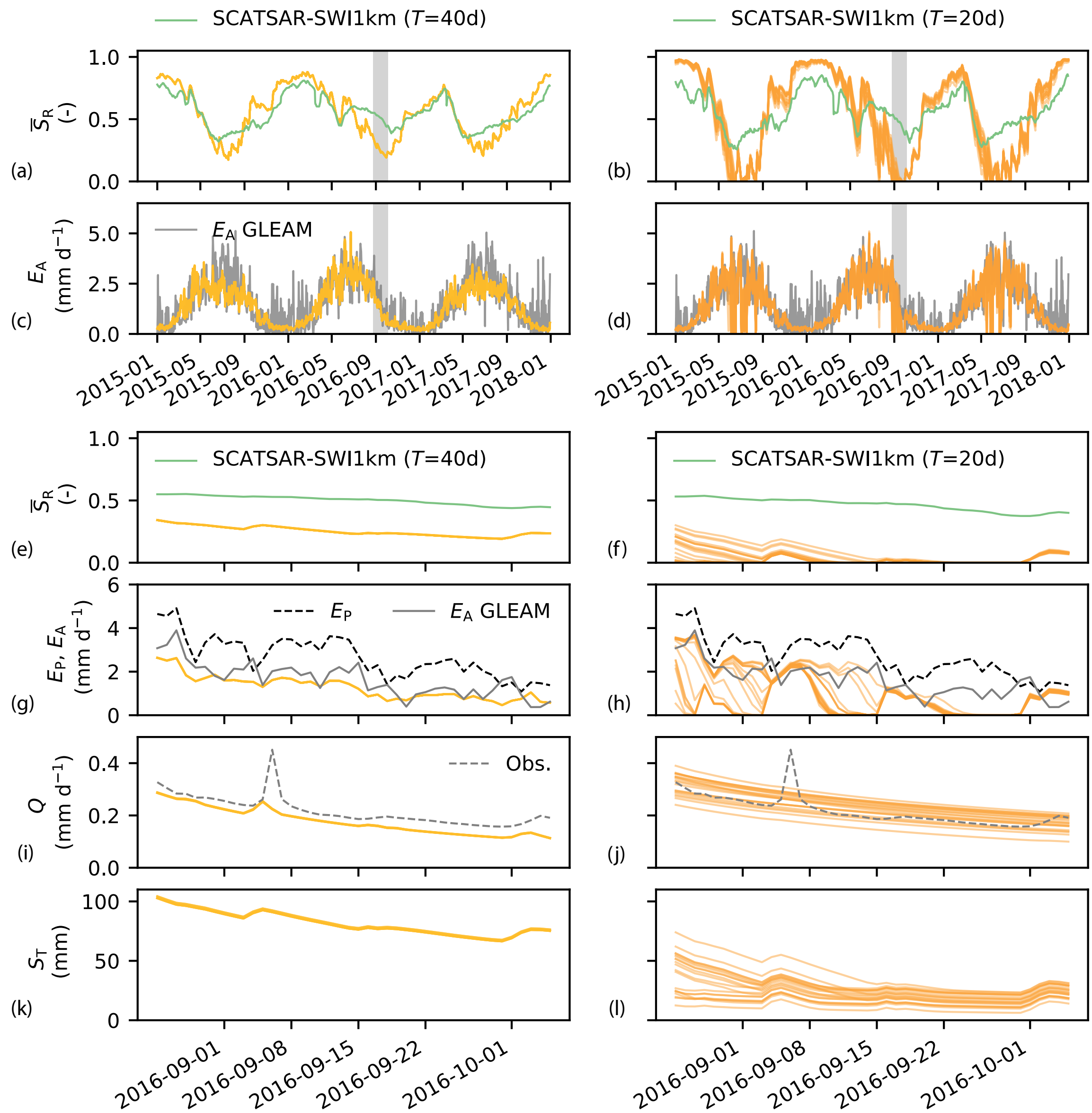

Figure 8. (a,b) Modeled relative root-zone soil moisture $\overline{S}_{\mathrm{R}}$ and SCATSAR-SWI1km Soil Water Index with optimal T-value for the period 2015–2017 for GR4H (yellow) and M5 (orange) respectively. **(c,d)** Actual evaporation E_{A} by GR4H and M5 for the period 2015–2017, showing a large reduction of evaporation during summer for M5 unlike GR4H and GLEAM actual evaporation **(e,f)** Zoomed-in modeled $\overline{S}_{\mathrm{R}}$ and SCATSAR-SWI1km root-zone soil moisture for the grey shaded period of September 2016 in (a,b,c,d). **(g,h)** Potential, modeled and GLEAM actual evaporation, **(i,j)** Modeled and observed streamflow Q, **(k,l)** Total storage S_{T} for the September 2016 dry period. The narrow uncertainty band of the GR4H model is related to its converging parameter search method.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

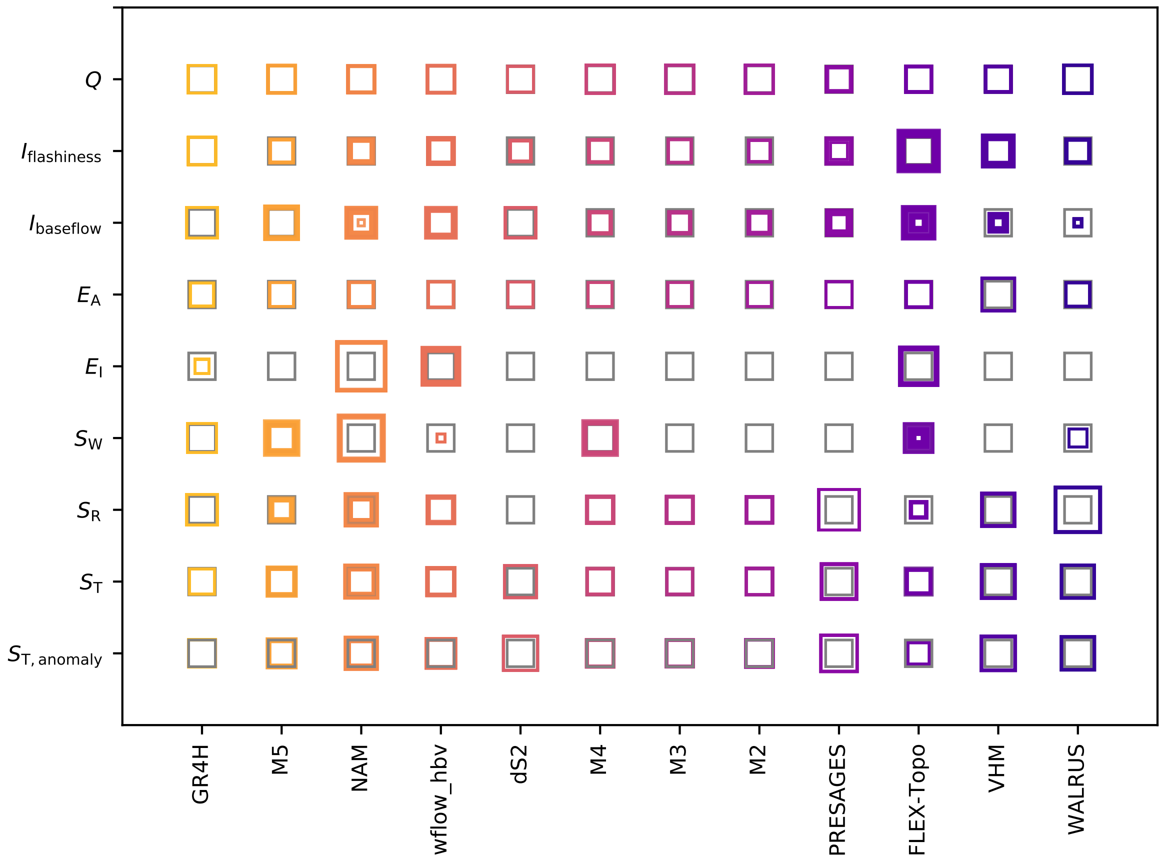

Figure 9. Summary of over- and underestimations of states and fluxes compared to a reference for the Ourthe upstream of Tabreux (grey squares). Mean annual modeled streamflow (denoted as Q), flashiness and baseflow indices ($I_{\mathrm{flashiness}}$ and I_{baseflow}) are compared with observed streamflow as a reference. Actual and interception evaporation (E_{A} and E_{I}) are compared with GLEAM estimates. Maximum annual snow water storage (S_{W}) is compared with the median of all models for the Ourthe Orientale upstream of Mabompré. The ranges of maximum and minimum root-zone soil moisture (S_{R}) and total storage (S_{T}) are compared with the median ranges of all models. The difference between maximum and minimum modeled total storage anomalies ($S_{\mathrm{T,anomaly}}$) is compared with GRACE estimates. The thickness of colored lines represents the realizations from the ensemble of parameter sets retained as feasible.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

Table 1. Description of symbols used for fluxes, storages and parameters in Tables 2 and 3

Symbol	unit	Description
Fluxes		
E_P	mm h^{-1}	Potential evaporation
E_I	mm h^{-1}	Interception evaporation
E_R	mm h^{-1}	Transpiration (and soil evaporation)
E_W	mm h^{-1}	Sublimation
E_A	mm h^{-1}	Total actual evaporation (sum of transpiration and if applicable interception and/or sublimation)
P	mm h^{-1}	Precipitation
P_R	mm h^{-1}	Precipitation entering the root-zone storage (after snow and/or interception if present or fraction/total precipitation)
Q	mm h^{-1}	Streamflow
Q_R	mm h^{-1}	Flux from root-zone to fast and/or slow runoff storage
Q_P	mm h^{-1}	Percolation flux from root-zone storage to slow runoff storage
Q_C	mm h^{-1}	Capillary flux from slow runoff storage to root-zone storage
Q_G	mm h^{-1}	Seepage (up/down) / extraction
Storages		
S_T	mm	Total storage
S_W	mm	Snow storage
S_I	mm	Interception storage
S_R	mm	Root-zone storage
$\overline{S}_\mathrm{R}$	-	Relative root-zone storage ($S_\mathrm{R}/S_\mathrm{R,max}$)
S_D	mm	Storage deficit
S_VQ	mm	Very quick runoff storage
S_F	mm	Fast runoff storage
S_S	mm	Slow runoff storage
S_SW	mm	Surface water storage
Parameters		
C_E	-	Correction factor for E_P
I_max	mm	Maximum interception capacity
$S_\mathrm{R,max}$	mm	Maximum root-zone storage capacity
S_thresh	mm	Threshold of root-zone storage above which $E_\mathrm{R} = E_\mathrm{P}$
L_P	-	Threshold of relative root-zone storage above which $E_\mathrm{R} = E_\mathrm{P}$
C_cst	-	Constant water stress coefficient to estimate E_R
a, b, S_0	-	Parameters describing the shape of the streamflow sensitivity
a_S	-	Fraction of land surface covered by surface water
a_G	-	Fraction of land surface not covered by surface water

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Table 2. Number of calibrated model parameters, model performance calculated for the period 2008–2017 with the Euclidean distance where a value of 0 would indicate a perfect model. Main characteristics describing snow storage, root-zone storage and total storage per model. Notations are defined in Table 1.

	GR4H	*M5*	*NAM*	*wflow_hbv*	*dS2*	*M4*	*M3*	*M2*	*PRESAGES*	*FLEX-Topo*	*VHM*	*WALRUS*
Number of calibrated parameters	4	9	12	9	4	7	6	5	6	20	12	3
Euclidean distance $\sqrt{(1-E_{\mathrm{NS,Q}})^2+(1-E_{\mathrm{NS,logQ}})^2}$	0.17	0.18	0.18	0.20	0.21	0.23	0.23	0.24	0.24	0.26	0.26	0.34
Snow storage S_{W} (compared to MODIS snow cover)												
Snow module	✓	✓	✓	✓	-	✓	-	-	-	✓	-	✓
Degree-hour method	✓	✓	✓	✓	-	✓	-	-	-	✓	-	✓
Elevation zones	✓	-	-	✓	-	-	-	-	-	✓	-	-
Temperature interval for rainfall and snow	✓	-	-	✓	-	-	-	-	-	✓	-	✓
Melt factor constant in time	-	✓	-	✓	-	✓	-	-	-	✓	-	✓
Melt factor $\sim$ snow storage	✓	-	✓	-	-	-	-	-	-	-	-	-
Refreezing of liquid water	-	-	✓	✓	-	-	-	-	-	-	-	-
Sublimation	-	-	-	-	-	-	-	-	-	✓	-	-
Calibration snow parameters	-	✓	✓	-	-	✓	-	-	-	✓	-	-
Root-zone storage S_{R} (compared to SCATSAR-SWI1km Soil Water Index)												
Separate root-zone module with capacity $S_{\mathrm{R,max}}$	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	-
$\frac{\mathrm{d}S_{\mathrm{R}}}{\mathrm{d}t}=P_{\mathrm{R}}-E_{\mathrm{R}}$	-	-	-	-	-	-	-	-	-	-	✓	-
$\frac{\mathrm{d}S_{\mathrm{R}}}{\mathrm{d}t}=P_{\mathrm{R}}-E_{\mathrm{R}}+Q_{\mathrm{C}}$	-	-	✓	-	-	-	-	-	-	-	-	-
$\frac{\mathrm{d}S_{\mathrm{R}}}{\mathrm{d}t}=P_{\mathrm{R}}-E_{\mathrm{R}}-Q_{\mathrm{R}}$	✓	✓	-	✓	-	✓	✓	✓	✓	-	-	-
$\frac{\mathrm{d}S_{\mathrm{R}}}{\mathrm{d}t}=P_{\mathrm{R}}-E_{\mathrm{R}}-Q_{\mathrm{R}}-Q_{\mathrm{P}}+Q_{\mathrm{C}}$	-	-	-	-	-	-	-	-	-	✓	-	-
Total storage S_{T} (anomalies are compared to GRACE total storage anomalies)												
$S_{\mathrm{T}}=-S_{\mathrm{D}}\cdot a_{\mathrm{G}}+S_{\mathrm{F}}\cdot a_{\mathrm{G}}+S_{\mathrm{SW}}\cdot a_{\mathrm{S}}$	-	-	-	-	-	-	-	-	-	-	-	✓
$S_{\mathrm{T}}(Q)=\frac{1}{a}\frac{1}{1-b}\cdot Q^{1-b}+S_0$	-	-	-	-	✓	-	-	-	-	-	-	-
$S_{\mathrm{T}}=S_{\mathrm{R}}+S_{\mathrm{F}}$	-	-	-	-	-	-	✓	✓	-	-	-	-
$S_{\mathrm{T}}=S_{\mathrm{W}}+S_{\mathrm{R}}+S_{\mathrm{F}}$	-	-	-	-	-	✓	-	-	-	-	-	-
$S_{\mathrm{T}}=S_{\mathrm{W}}+S_{\mathrm{R}}+S_{\mathrm{F}}+S_{\mathrm{S}}$	-	✓	-	-	-	-	-	-	-	-	-	-
$S_{\mathrm{T}}=S_{\mathrm{R}}+S_{\mathrm{VQ}}+S_{\mathrm{F}}+S_{\mathrm{S}}$	-	-	-	-	-	-	-	-	✓	-	✓	-
$S_{\mathrm{T}}=S_{\mathrm{W}}+S_{\mathrm{R}}+S_{\mathrm{VQ}}+S_{\mathrm{F}}+S_{\mathrm{S}}$	✓	-	-	-	-	-	-	-	-	-	-	-
$S_{\mathrm{T}}=S_{\mathrm{I}}+S_{\mathrm{W}}+S_{\mathrm{R}}+S_{\mathrm{F}}+S_{\mathrm{S}}$	-	-	-	✓	-	-	-	-	-	-	-	-
$S_{\mathrm{T}}=S_{\mathrm{I}}+S_{\mathrm{W}}+S_{\mathrm{R}}+S_{\mathrm{VQ}}+S_{\mathrm{F}}+S_{\mathrm{S}}$	-	-	✓	-	-	-	-	-	-	✓	-	-

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Table 3. Main characteristics describing evaporation processes per model, which are compared to GLEAM remotely-sensed-based evaporation (with $\checkmark^1$ indicates $L_\mathrm{P} = 1$ and $\checkmark^2$ indicates $E_\mathrm{I} = 0$). Notations are defined in Table 1.

	GR4H	M5	NAM	wflow_hbv	dS2	M4	M3	M2	PRESAGES	FLEX-Topo	VHM	WALRUS
Correction factor for potential evaporation	-	✓	-	-	-	✓	✓	✓	-	-	-	-
Interception evaporation E_I	✓	-	✓	✓	-	-	-	-	-	✓	-	-
Maximum interception storage I_max	-	-	✓	✓	-	-	-	-	-	✓	-	-
$I_\mathrm{max} \sim 1.1 - 3.4$ mm	-	-	-	-	-	-	-	-	-	✓	-	-
$I_\mathrm{max} \sim 1.4 - 2.9$ mm	-	-	-	✓	-	-	-	-	-	-	-	-
$I_\mathrm{max} \sim 5.3 - 6.9$ mm	-	-	✓	-	-	-	-	-	-	-	-	-
$E_\mathrm{I} = \begin{cases} E_\mathrm{P}, & \text{if } S_\mathrm{I} > 0. \\ 0, & \text{otherwise.} \end{cases}$	-	-	✓	✓	-	-	-	-	-	✓	-	-
$E_\mathrm{I} = \begin{cases} E_\mathrm{P}, & \text{if } P > E_\mathrm{P}. \\ P, & \text{otherwise.} \end{cases}$	✓	-	-	-	-	-	-	-	-	-	-	-
Transpiration E_R	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$E_\mathrm{R} = E_\mathrm{P} \cdot C_\mathrm{cst}$	-	-	-	-	✓	-	-	-	-	-	-	-
$E_\mathrm{R} = E_\mathrm{P} \cdot \frac{\overline{S}_\mathrm{R} \cdot (2-\overline{S}_\mathrm{R})}{1+E_\mathrm{P}/S_\mathrm{R,max} \cdot (2-\overline{S}_\mathrm{R})}$	✓	-	-	-	-	-	-	-	✓	-	-	-
$E_\mathrm{R} = E_\mathrm{P} \cdot C_\mathrm{E} \cdot \frac{\overline{S}_\mathrm{R} \cdot (1+m_1)}{\overline{S}_\mathrm{R}+m_1}$, with $m_1 = 10^{-2}$	-	✓	-	-	-	✓	✓	✓	-	-	-	-
$E_\mathrm{R} = \begin{cases} (E_\mathrm{P} - E_\mathrm{I}) \cdot \frac{\overline{S}_\mathrm{R}}{L_P}, & \text{if } \overline{S}_\mathrm{R} < L_P. \\ E_\mathrm{P} - E_\mathrm{I}, & \text{otherwise.} \end{cases}$	-	-	$\checkmark^1$	✓	-	-	-	-	-	✓	$\checkmark^2$	-
$E_\mathrm{R} = E_\mathrm{P} \cdot \mathrm{f}(S_\mathrm{d})$	-	-	-	-	-	-	-	-	-	-	-	✓
Actual evaporation E_A	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
$E_\mathrm{A} = E_\mathrm{R}$	-	✓	-	-	✓	✓	✓	✓	✓	-	✓	✓
$E_\mathrm{A} = E_\mathrm{R} + E_\mathrm{I}$	✓	-	✓	✓	-	-	-	-	-	-	-	-
$E_\mathrm{A} = E_\mathrm{R} + E_\mathrm{I} + E_\mathrm{W}$	-	-	-	-	-	-	-	-	-	✓	-	-

Data availability. Streamflow and precipitation data were provided by the Service Public de Wallonie in Belgium (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)). Hourly radiation data were retrieved from the portal of the Royal Netherlands Meteorological Institute (2018; http://www.knmi.nl/nederland-nu/klimatologie/uurgegevens). Daily temperature data were retrieved from the E-OBS OPeNDAP server http://opendap.knmi.nl/knmi/thredds/dodsC/e-obs_0.25regular/tg_0.25deg_reg_v17.0.nc (Haylock et al., 2008). Actual evaporation estimates from the Global Land Evaporation Amsterdam Model (GLEAM) are available through the SFTP server of GLEAM, (https://www.gleam.eu/, Miralles et al., 2011; Martens et al., 2017). MODIS Snow cover fractions (Hall and Riggs, 2016a, b) are available for download from the Earthdata portal https://earthdata.nasa.gov/. The Soil Water Index SCATSAR-SWI1km are available from the Copernicus Global Land Service at https://land.copernicus.eu/global/products/swi (Bauer-Marschallinger et al., 2018). GRACE land data (Swenson and Wahr, 2006; Swenson, 2012) are available at http://grace.jpl.nasa.gov, supported by the NASA MEaSUREs Program. The modeled states and fluxes of each model will be made available online in the 4TU data repository.

Author contributions. LJEB, GT, LAM, AHW and MH designed the study. GT, TdBE, JB, CCB, JdN, SM, BG, FF, JN and LJEB ran their models and provided simluated time series of all variables. LJEB conducted all the analyses and wrote the manuscript. All authors discussed the design, results and contributed to the final manuscript.

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. We thank Deltares and Rijkswaterstaat for the financial support to conduct this analysis. 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 streamflow data.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

References

Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, Water Resources Research, 55, 378–390, https://doi.org/10.1029/2018WR022958, 2019.

Adnan, M., Merwade, V., and Yu, Z.: Multi-objective calibration of a hydrologic model using spatially distributed remotely sensed / in-situ soil moisture, Journal of Hydrology, 536, 192–207, https://doi.org/10.1016/j.jhydrol.2016.02.037, 2016.

Albergel, C., Rüdiger, C., Pellarin, T., Calvet, J. C., Fritz, N., Froissard, F., Suquia, D., Petitpa, A., Piguet, B., and Martin, E.: From near-surface to root-zone soil moisture using an exponential filter: An assessment of the method based on in-situ observations and model simulations, Hydrology and Earth System Sciences, 12, 1323–1337, https://doi.org/10.5194/hess-12-1323-2008, 2008.

Alcamo, J., Döll, P., Henrichs, T., Kaspar, F., Lehner, B., Rösch, T., and Siebert, S.: Development and testing of the WaterGAP 2 global model of water use and availability, Hydrological Sciences Journal, 48, 317–338, https://doi.org/10.1623/hysj.48.3.317.45290, 2003.

Andréassian, V., Le Moine, N., Perrin, C., Ramos, M. H., Oudin, L., Mathevet, T., Lerat, J., and Berthet, L.: All that glitters is not gold: The case of calibrating hydrological models, Hydrological Processes, 26, 2206–2210, https://doi.org/10.1002/hyp.9264, 2012.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., Seibert, J., Hakala, K., Bock, A., Wagener, T., Farmer, W. H., Andréassian, V., Attinger, S., Viglione, A., Knight, R., Markstrom, S., and Over, T.: Accelerating advances in continental domain hydrologic modeling, Water Resources Research, 51, 10 078–10 091, https://doi.org/10.1002/2015WR017498, 2015.

Bauer-Marschallinger, B., Paulik, C., Hochstöger, S., Mistelbauer, T., Modanesi, S., Ciabatta, L., Massari, C., Brocca, L., and Wagner, W.: Soil moisture from fusion of scatterometer and SAR: Closing the scale gap with temporal filtering, Remote Sensing, 10, 1–26, https://doi.org/10.3390/rs10071030, 2018.

Beck, H. E., Van Dijk, A. I. J. M., De Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from ten state - of - the - art hydrological models, Hydrology and Earth System Sciences, 21, 2881–2903, https://doi.org/10.5194/hess-21-2881-2017, 2017.

Bennett, K. E., Cherry, J. E., Balk, B., and Lindsey, S.: Using MODIS estimates of fractional snow cover area to improve streamflow forecasts in interior Alaska, Hydrology and Earth System Sciences, 23, 2439–2459, https://doi.org/10.5194/hess-23-2439-2019, 2019.

Beven, K.: A manifesto for the equifinality thesis, Journal of Hydrology, 320, 18–36, https://doi.org/10.1016/j.jhydrol.2005.07.007, 2006.

Beven, K. J.: Uniqueness of place and process representations in hydrological modelling, Hydrology and Earth System Sciences, 4, 203–213, https://doi.org/10.5194/hess-4-203-2000, 2000.

Bonin, J. and Chambers, D.: Uncertainty estimates of a GRACE inversion modelling technique over greenland using a simulation, Geophysical Journal International, 194, 212–229, https://doi.org/10.1093/gji/ggt091, 2013.

Bouaziz, L., Weerts, A., Schellekens, J., Sprokkereef, E., Stam, J., Savenije, H., and Hrachowitz, M.: Redressing the balance: Quantifying net intercatchment groundwater flows, Hydrology and Earth System Sciences, 22, 6415–6434, https://doi.org/10.5194/hess-22-6415-2018, 2018.

Bouaziz, L. J., Steele-Dunne, S. C., Schellekens, J., Weerts, A. H., Stam, J., Sprokkereef, E., Winsemius, H. H., Savenije, H. H., and Hrachowitz, M.: Improved understanding of the link between catchment-scale vegetation accessible storage and satellite-derived Soil Water Index, Water Resources Research, https://doi.org/10.1029/2019WR026365, 2020.

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

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Brauer, C. C., Torfs, P. J., Teuling, A. J., and Uijlenhoet, R.: The Wageningen Lowland Runoff Simulator (WALRUS): Application to the Hupsel Brook catchment and the Cabauw polder, Hydrology and Earth System Sciences, 18, 4007–4028, https://doi.org/10.5194/hess-18-4007-2014, 2014b.

Brocca, L., Melone, F., Moramarco, T., Wagner, W., and Hasenauer, S.: ASCAT soil wetness index validation through in situ and modeled soil moisture data in central Italy, Remote Sensing of Environment, 114, 2745–2755, https://doi.org/10.1016/j.rse.2010.06.009, 2010.

Buitink, J., Melsen, L. A., Kirchner, J. W., and Teuling, A. J.: A distributed simple dynamical systems approach (dS2 v1.0) for computationally efficient hydrological modelling, Geoscientific Model Development Discussions, 20, 1–25, https://doi.org/10.5194/gmd-2019-150, 2019.

Burt, T. P., McDonnell, J. J.: Whither Field Hydrology?, Water Res. Research, 51, 5919–5928, https://doi.org/10.1002/2014WR016839, 2015.

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, Hydrology and Earth System Sciences, 19, 2101–2117, https://doi.org/10.5194/hess-19-2101-2015, 2015.

Cisneros Vaca, C., Van Der Tol, C., and Ghimire, C. P.: The influence of long-term changes in canopy structure on rainfall interception loss: A case study in Speulderbos, the Netherlands, Hydrology and Earth System Sciences, 22, 3701–3719, https://doi.org/10.5194/hess-22-3701-2018, 2018.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, 1–14, https://doi.org/10.1029/2007wr006735, 2008.

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. a., Freer, J. E., Gutmann, E. D., Wood, A. W., Brekke, L. D., Arnold, J. R., Gochis, D. J., and Rasmussen, R. M.: A unified approach for process-based hydrologicmodeling: 1. Modeling concept, Water Resources Research, 51, 1–17, https://doi.org/10.1002/2015WR017200.A, 2015.

Clark, M. P., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., Freer, J. E., Arnold, J. R., Moore, R. D., Istanbulluoglu, E., and Ceola, S.: Improving the theoretical underpinnings of process-based hydrologic models, Water Resources Research, 52, 2350–2365, https://doi.org/10.1002/2015WR017910, 2016.

Coenders-Gerrits, A. M., Van Der Ent, R. J., Bogaard, T. A., Wang-Erlandsson, L., Hrachowitz, M., and Savenije, H. H.: Uncertainties in transpiration estimates, Nature, 506, 2013–2015, https://doi.org/10.1038/nature12925, 2014.

Copernicus Global Land Service: Soil Water Index, available at: https://land.copernicus.eu/global/products/swi, last accessed: 2019-01-04, 2019.

Coron, L., Thirel, G., Delaigue, O., Perrin, C., and Andréassian, V.: The suite of lumped GR hydrological models in an R package, Environmental Modelling and Software, 94, 166–171, https://doi.org/10.1016/j.envsoft.2017.05.002, 2017.

Coron, L., Perrin, C., Delaigue, O., Thirel, G., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling, R package version 1.0.10.11, https://doi.org/10.15454/EX11NA, https://cran.r-project.org/package=airGR., 2019.

de Boer-Euser, T., McMillan, H. K., Hrachowitz, M., Winsemius, H. C., and Savenije, H. H.: Influence of soil and climate on root zone storage capacity, Water Resources Research, https://doi.org/10.1002/2015WR018115, 2016.

de Boer-Euser, T., Bouaziz, L., de Niel, J., Brauer, C., Dewals, B., Drogue, G., Fenicia, F., Grelier, B., Nossent, J., Pereira, F., Savenije, H., Thirel, G., and Willems, P.: Looking beyond general metrics for model comparison – Lessons from an international model intercomparison study, Hydrology and Earth System Sciences, 21, 423–440, https://doi.org/10.5194/hess-21-423-2017, 2017.

Hydrology and Earth System Sciences
Open Access
EGU
Discussions

de Niel, J., van Uytven, E., and Willems, P.: Uncertainty Analysis of Climate Change Impact on River Flow Extremes Based on a Large Multi-Model Ensemble, Water Resources Management, 33, 4319–4333, https://doi.org/10.1007/s11269-019-02370-0, 2019.

de Wit, M. J., van den Hurk, B., Warmerdam, P. M., Torfs, P. J., Roulin, E., and Van Deursen, W. P.: Impact of climate change on low-flows in the river Meuse, Climatic Change, 82, 351–372, https://doi.org/10.1007/s10584-006-9195-2, 2007.

Dembélé, M., Hrachowitz, M., Savenije, H. H., Mariéthoz, G., and Schaefli, B.: Improving the Predictive Skill of a Distributed Hydrological Model by Calibration on Spatial Patterns With Multiple Satellite Data Sets, Water Resources Research, 56, 0–3, https://doi.org/10.1029/2019WR026085, 2020.

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., and Stisen, S.: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, Hydrology and Earth System Sciences, 22, 1299–1315, https://doi.org/10.5194/hess-22-1299-2018, 2018.

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

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

Fenicia, F., McDonnell, J. J., and Savenije, H. H. G.: Learning from model improvement: On the contribution of complementary data to process understanding, Water Resources Research, 44, 1–13, https://doi.org/10.1029/2007WR006386, 2008.

Fenicia, F., Kavetski, D., and Savenije, H. H. G.: Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development, Water Resources Research, 47, 1–13, https://doi.org/10.1029/2010WR010174, 2011.

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., and Freer, J.: Catchment properties, function, and conceptual model representation: Is there a correspondence?, Hydrological Processes, 28, 2451–2467, https://doi.org/10.1002/hyp.9726, 2014.

Fenicia, F., Kavetski, D., Savenije, H. H., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, Water Resources Research, https://doi.org/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, Geophysical Research Letters, 41, 7916–7923, https://doi.org/10.1002/2014GL061668, 2014.

Gao, H., Ding, Y., Zhao, Q., Hrachowitz, M., and Savenije, H. H.: The importance of aspect for modelling the hydrological response in a glacier catchment in Central Asia, Hydrological Processes, 31, 2842–2859, https://doi.org/10.1002/hyp.11224, 2017.

Gash, J. H.: An analytical model of rainfall interception by forests, Quarterly Journal of the Royal Meteorological Society, 105, 43–55, https://doi.org/10.1002/qj.49710544304, 1979.

Gash, J. H., Wright, I. R., and Lloyd, C. R.: Comparative estimates of interception loss from three coniferous forests in Great Britain, Journal of Hydrology, 48, 89–105, https://doi.org/10.1016/0022-1694(80)90068-2, 1980.

Gentine, P., D'Odorico, P., Lintner, B. R., Sivandran, G., and Salvucci, G.: Interdependence of climate, soil, and vegetation as constrained by the Budyko curve, Geophysical Research Letters, 39, 2–7, https://doi.org/10.1029/2012GL053492, 2012.

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, https://doi.org/10.5194/hess-18-4839-2014, 2014.

Gupta, H. V. and Nearing, G. S.: Debates - The future of hydrological sciences: A (common) path forward? Using models and data to learn: A systems theoretic perspective on the future of hydrological science, Water Resources Research, 50, 5351–5359, https://doi.org/10.1002/2013WR015096, 2014.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, https://doi.org/10.1002/hyp.6989, 2008.

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., and Ye, M.: Towards a comprehensive assessment of model structural adequacy, Water Resources Research, 48, 1–16, https://doi.org/10.1029/2011WR011044, 2012.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Weedon, G. P., and Yeh, P.: Multimodel estimate of the global terrestrial water balance: Setup and first results, Journal of Hydrometeorology, 12, 869–884, https://doi.org/10.1175/2011JHM1324.1, 2011.

Hagemann, S., Chen, C., Clark, D. B., Folwell, S., Gosling, S. N., Haddeland, I., Hanasaki, N., Heinke, J., Ludwig, F., Voss, F., and Wiltshire, A. J.: Climate change impact on available water resources obtained using multiple global climate and hydrology models, Earth System Dynamics, 4, 129–144, https://doi.org/10.5194/esd-4-129-2013, 2013.

Hall, D. K. and Riggs, G. A.: MODIS/Terra Snow Cover Daily L3 Global 500m SIN Grid, Version 6. [MOD10A1], https://doi.org/https://doi.org/10.5067/MODIS/MOD10A1.006, https://nsidc.org/data/MOD10A1/versions/6, Accessed: 2019-06-10, 2016a.

Hall, D. K. and Riggs, G. A.: MODIS/Aqua Snow Cover Daily L3 Global 500m SIN Grid, Version 6. [MYD10A1], https://doi.org/https://doi.org/10.5067/MODIS/MYD10A1.006., https://nsidc.org/data/MYD10A1/versions/6, Accessed: 2019-06-10, 2016b.

Hargreaves, G. H. and Samani, Z. A.: Reference Crop Evapotranspiration from Temperature, Applied Engineering in Agriculture, 1, 96–99, https://doi.org/10.13031/2013.26773, 1985.

Haylock, M. R., Hofstra, N., Klein Tank, A. M., 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, Journal of Geophysical Research Atmospheres, 113, https://doi.org/10.1029/2008JD010201, 2008.

Holländer, H. M., Blume, T., Bormann, H., Buytaert, W., Chirico, G. B., 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, Hydrology and Earth System Sciences, 13, 2069–2094, https://doi.org/10.5194/hess-13-2069-2009, 2009.

Hrachowitz, M. and Clark, M. P.: HESS Opinions: The complementary merits of competing modelling philosophies in hydrology, Hydrology and Earth System Sciences, 21, 3953–3973, https://doi.org/10.5194/hess-21-3953-2017, 2017.

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

Hulsman, P., Winsemius, H. C., Michailovsky, C., Savenije, H. H. G., and Hrachowitz, M.: Using altimetry observations combined with GRACE to select parameter sets of a hydrological model in data scarce regions, Hydrology and Earth System Sciences Discussions, pp. 1–35, https://doi.org/10.5194/hess-2019-346, 2019.

Jakeman, A. J. and Hornberger, G. M.: How much complexity is warranted in a rainfall-runoff model?, Water Resources Research, 29, 2637–2649, https://doi.org/10.1029/93WR00877, 1993.

Kirchner, J. W.: Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, Water Resources Research, 42, 1–5, https://doi.org/10.1029/2005WR004362, 2006.

Klemeš, V.: Operational testing of hydrological simulation models, Hydrological Sciences Journal, 31, 13–24, https://doi.org/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, Journal of Hydrology, 533, 234–249, https://doi.org/10.1016/j.jhydrol.2015.12.002, 2016.

Koch, J., Demirel, M. C., and Stisen, S.: The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models, Geoscientific Model Development, 11, 1873–1886, https://doi.org/10.5194/gmd-11-1873-2018, 2018.

Kunnath-Poovakka, A., Ryu, D., Renzullo, L. J., and George, B.: The efficacy of calibrating hydrologic model using remotely sensed evapotranspiration and soil moisture for streamflow prediction, Journal of Hydrology, 535, 509–524, https://doi.org/10.1016/j.jhydrol.2016.02.018, 2016.

Landerer, F. W. and Swenson, S. C.: Accuracy of scaled GRACE terrestrial water storage estimates, Water Resources Research, 48, 1–11, https://doi.org/10.1029/2011WR011453, 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, https://doi.org/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, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Livneh, B. and Lettenmaier, D. P.: Multi-criteria parameter estimation for the Unified Land Model, Hydrology and Earth System Sciences, 16, 3029–3048, https://doi.org/10.5194/hess-16-3029-2012, 2012.

López López, P., Sutanudjaja, E. H., Schellekens, J., Sterk, G., and Bierkens, M. F.: Calibration of a large-scale hydrological model using satellite-based soil moisture and evapotranspiration products, Hydrology and Earth System Sciences, 21, 3125–3144, https://doi.org/10.5194/hess-21-3125-2017, 2017.

Martens, B., Miralles, D. G., Lievens, H., Van Der Schalie, R., De Jeu, R. A., Fernández-Prieto, D., Beck, H. E., Dorigo, W. A., and Verhoest, N. E.: GLEAM v3: Satellite-based land evaporation and root-zone soil moisture, Geoscientific Model Development, 10, 1903–1925, https://doi.org/10.5194/gmd-10-1903-2017, 2017.

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.

McCabe, M. F., Rodell, M., Alsdorf, D. E., Miralles, D. G., Uijlenhoet, R., Wagner, W., Lucieer, A., Houborg, R., Verhoest, N. E., Franz, T. E., Shi, J., Gao, H., and Wood, E. F.: The future of Earth observation in hydrology, Hydrology and Earth System Sciences, 21, 3879–3914, https://doi.org/10.5194/hess-21-3879-2017, 2017.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

Melsen, L. A. and Guse, B.: Hydrological Drought Simulations: How Climate and Model Structure Control Parameter Sensitivity, Water Resources Research, 55, 10 527–10 547, https://doi.org/10.1029/2019WR025230, 2019.

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J., Clark, M. P., Uijlenhoet, R., and Teuling, A. J.: Mapping (dis)agreement in hydrologic projections, Hydrology and Earth System Sciences, 22, 1775–1791, https://doi.org/10.5194/hess-22-1775-2018, 2018.

Melsen, L. A., Teuling, A. J., Torfs, P. J., Zappa, M., Mizukami, N., Mendoza, P. A., Clark, M. P., and Uijlenhoet, R.: Subjective modeling decisions can significantly impact the simulation of flood and drought events, Journal of Hydrology, 568, 1093–1104, https://doi.org/10.1016/j.jhydrol.2018.11.046, 2019.

Mianabadi, A., Coenders-Gerrits, M., Shirazi, P., Ghahraman, B., and Alizadeh, A.: A global Budyko model to partition evaporation into interception and transpiration, Hydrology and Earth System Sciences Discussions, pp. 1–32, https://doi.org/10.5194/hess-2018-638, 2019.

Milly, P. C.: Climate, interseasonal storage of soil water, and the annual water balance, Advances in Water Resources, 17, 19–24, https://doi.org/10.1016/0309-1708(94)90020-5, 1994.

Miralles, D. G., Holmes, T. R., De Jeu, R. A., Gash, J. H., Meesters, A. G., and Dolman, A. J.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453–469, https://doi.org/10.5194/hess-15-453-2011, 2011.

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

Nijzink, R., Hutton, C., Pechlivanidis, I., Capell, R., Arheimer, B., Freer, J., Han, D., Wagener, T., McGuire, K., Savenije, H., and Hrachowitz, M.: The evolution of root-zone moisture capacities after deforestation: A step towards hydrological predictions under change?, Hydrology and Earth System Sciences, 20, 4775–4799, https://doi.org/10.5194/hess-20-4775-2016, 2016.

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., Parajka, J., Freer, J., Han, D., Wagener, T., van Nooijen, R. R., Savenije, H. H., and Hrachowitz, M.: Constraining Conceptual Hydrological Models With Multiple Information Sources, Water Resources Research, 54, 8332–8362, https://doi.org/10.1029/2017WR021895, 2018.

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., and Zappa, M.: Does model performance improve with complexity ? A case study with three hydrological models, Journal of Hydrology, 523, 147–159, https://doi.org/10.1016/j.jhydrol.2015.01.044, 2015.

Perrin, C., Michel, C., and Andréassian, V.: Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, Journal of Hydrology, 242, 275–301, https://doi.org/10.1016/S0022-1694(00)00393-0, 2001.

Priestley, C. H. B. and Taylor, R. J.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, Monthly Weather Review, 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:otaosh>2.3.co;2, 1972.

Rakovec, O., Kumar, R., Attinger, S., and Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation, Water Resources Research, 52, 7779–7792, https://doi.org/10.1002/2016WR019430, 2016a.

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, Journal of Hydrometeorology, 17, 287–307, https://doi.org/10.1175/JHM-D-15-0054.1, 2016b.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., and Seo, D. J.: Overall distributed model intercomparison project results, in: Journal of Hydrology, vol. 298, pp. 27–60, https://doi.org/10.1016/j.jhydrol.2004.03.031, 2004.

Riboust, P., Thirel, G., Moine, N. L., and Ribstein, P.: Revisiting a simple degree-day model for integrating satellite data: implementation of swe-sca hystereses, Journal of Hydrology and Hydromechanics, 67, 70–81, https://doi.org/10.2478/johh-2018-0004, 2019.

Royal Meteorological Institute Belgium: Klimaatatlas, gemiddeld aantal dagen met sneeuw available at: https://www.meteo.be/nl/klimaat/klimaatatlas/klimaatkaarten/sneeuw, last access: 2020-03-26, 2015.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Royal Netherlands Meteorological Institute: Radiation data available at: http://www.knmi.nl/nederland-nu/klimatologie/uurgegevens, last accessed: 2018-04-30, 2018.

Sakumura, C., Bettadpur, S., and Bruinsma, S.: Ensemble prediction and intercomparison analysis of GRACE time-variable gravity field models, Geophysical Research Letters, 41, 1389–1397, https://doi.org/10.1002/2013GL058632, 2014.

Samaniego, L., Kumar, R., and Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale, Water Resources Research, 46, 1–25, https://doi.org/10.1029/2008WR007327, 2010.

Savenije, H. H.: HESS opinions "topography driven conceptual modelling (FLEX-Topo)", Hydrology and Earth System Sciences, 14, 2681–2692, https://doi.org/10.5194/hess-14-2681-2010, 2010.

Schellekens, J., Scatena, F. N., Bruijnzeel, L. A., and Wickel, A. J.: Modelling rainfall interception by a lowland tropical rain forest in northeastern Puerto Rico, Journal of Hydrology, 225, 168–184, https://doi.org/10.1016/S0022-1694(99)00157-2, 1999.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., Colón-González, F. J., Gosling, S. N., Kim, H., Liu, X., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., and Kabat, P.: Multimodel assessment of water scarcity under climate change, Proceedings of the National Academy of Sciences of the United States of America, 111, 3245–3250, https://doi.org/10.1073/pnas.1222460110, 2014.

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, 2018.

Sheffield, J., Wood, E. F., Pan, M., Beck, H., Coccia, G., Serrat-Capdevila, A., and Verbist, K.: Satellite Remote Sensing for Water Resources Management: Potential for Supporting Sustainable Development in Data-Poor Regions, Water Resources Research, 54, 9724–9758, https://doi.org/10.1029/2017WR022437, 2018.

Smith, M. B., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N., Anderson, E. A., and Cosgrove, B. A.: The distributed model intercomparison project – Phase 2 : Motivation and design of the Oklahoma experiments, Journal of Hydrology, 418-419, 3–16, https://doi.org/10.1016/j.jhydrol.2011.08.055, 2012a.

Smith, M. B., Koren, V., Zhang, Z., Zhang, Y., Reed, S. M., Cui, Z., Moreda, F., Cosgrove, B. A., Mizukami, N., Anderson, E. A., and Participants, D.: Results of the DMIP 2 Oklahoma experiments, Journal of Hydrology, 418-419, 17–48, https://doi.org/10.1016/j.jhydrol.2011.08.056, 2012b.

Stisen, S., Koch, J., Sonnenborg, T. O., Refsgaard, J. C., Bircher, S., Ringgaard, R., and Jensen, K. H.: Moving beyond run-off calibration—Multivariable optimization of a surface–subsurface–atmosphere model, Hydrological Processes, 32, 2654–2668, https://doi.org/10.1002/hyp.13177, 2018.

Sutanudjaja, E. H., Van Beek, L. P., De Jong, S. M., Van Geer, F. C., and Bierkens, M. F.: Calibrating a large-extent high-resolution coupled groundwater-land surface model using soil moisture and discharge data, Water Resources Research, 50, 687–705, https://doi.org/10.1002/2013WR013807, 2014.

Swenson, S.: GRACE monthly land water mass grids NETCDF RELEASE 5.0. Ver. 5.0. PO.DAAC, CA, USA, https://doi.org/http://dx.doi.org/10.5067/TELND-NC005, accessed: 2019-06-10, 2012.

Swenson, S. and Wahr, J.: Post-processing removal of correlated errors in GRACE data, Geophysical Research Letters, 33, 1–4, https://doi.org/10.1029/2005GL025285, 2006.

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, Journal of Hydrology, 517, 1176–1187, https://doi.org/10.1016/j.jhydrol.2014.04.058, 2014.

https://doi.org/10.5194/hess-2020-176
Preprint. Discussion started: 28 April 2020

Hydrology and Earth System Sciences
Discussions

van Dijk, A. I. J. M.: Climate and terrain factors explaining streamflow response and recession in Australian catchments, Hydrology and Earth System Sciences, 14, 159–169, https://doi.org/10.5194/hess-14-159-2010, 2010.

van Emmerik, T., Mulder, G., Eilander, D., Piet, M., and Savenije, H.: Predicting the ungauged basin: Model validation and realism assessment, Frontiers in Earth Science, 3, 1–11, https://doi.org/10.3389/feart.2015.00062, 2015.

van Emmerik, T., Popp, A., Solcerova, A., Müller, H., and Hut, R.: Reporting negative results to stimulate experimental hydrology: discussion of "The role of experimental work in hydrological sciences–insights from a community survey", Hydrological Sciences Journal, 63, 1269–1272, https://doi.org/10.1080/02626667.2018.1493203, 2018.

Veldkamp, T. I., Zhao, F., Ward, P. J., De Moel, H., Aerts, J. C., Schmied, H. M., Portmann, F. T., Masaki, Y., Pokhrel, Y., Liu, X., Satoh, Y., Gerten, D., Gosling, S. N., Zaherpour, J., and Wada, Y.: Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: A multi-model validation study, Environmental Research Letters, 13, https://doi.org/10.1088/1748-9326/aab96f, 2018.

Vidon, P. G.: Field hydrologists needed: A call for young hydrologists to (re)-focus on field studies, Hydrological Processes, 29, 5478–5480, https://doi.org/10.1002/hyp.10614, 2015.

Wagner, W., Lemoine, G., and Rott, H.: A method for estimating soil moisture from ERS Scatterometer and soil data, Remote Sensing of Environment, 70, 191–207, https://doi.org/10.1016/S0034-4257(99)00036-X, 1999.

Wagner, W., Hahn, S., Kidd, R., Melzer, T., Bartalis, Z., Hasenauer, S., Figa-Saldaña, J., De Rosnay, P., Jann, A., Schneider, S., Komma, J., Kubu, G., Brugger, K., Aubrecht, C., Züger, J., Gangkofner, U., Kienberger, S., Brocca, L., Wang, Y., Blöschl, G., Eitzinger, J., Steinnocher, K., Zeil, P., and Rubel, F.: The ASCAT soil moisture product: A review of its specifications, validation results, and emerging applications, Meteorologische Zeitschrift, 22, 5–33, https://doi.org/10.1127/0941-2948/2013/0399, 2013.

Wang-Erlandsson, L., Bastiaanssen, W. G., Gao, H., Jägermeyr, J., Senay, G. B., Van Dijk, A. I., Guerschman, J. P., Keys, P. W., Gordon, L. J., and Savenije, H. H.: Global root zone storage capacity from satellite-based evaporation, Hydrology and Earth System Sciences, 20, 1459–1481, https://doi.org/10.5194/hess-20-1459-2016, 2016.

Werth, S. and Güntner, A.: Calibration analysis for water storage variability of the global hydrological model WGHM, Hydrology and Earth System Sciences, 14, 59–78, https://doi.org/10.5194/hess-14-59-2010, 2010.

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, https://doi.org/10.1016/j.jhydrol.2014.01.017, 2014.

Winsemius, H. C., Savenije, H. H., Gerrits, A. M., Zapreeva, E. A., and Klees, R.: Comparison of two model approaches in the Zambezi river basin with regard to model reliability and identifiability, Hydrology and Earth System Sciences, 10, 339–352, https://doi.org/10.5194/hess-10-339-2006, 2006.

Yassin, F., Razavi, S., Wheater, H., Sapriza-Azuri, G., Davison, B., and Pietroniro, A.: Enhanced identification of a hydrologic model using streamflow and satellite water storage data: A multicriteria sensitivity analysis and optimization approach, Hydrological Processes, 31, 3320–3333, https://doi.org/10.1002/hyp.11267, 2017.