https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences Discussions

Lack of robustness of hydrological models: A large-sample diagnosis and an attempt to identify the hydrological and climatic drivers

Léonard Santos[1], Vazken Andréassian[1], Torben O. Sonnenborg[2], Göran Lindström[3], Alban de Lavenne[1,3], Charles Perrin[1], Lila Collet[1,4], Guillaume Thirel[1]

[1]Université Paris-Saclay, INRAE, UR HYCAR, Antony, France
[2]GEUS, Copenhagen, Denmark
[3]SMHI, Norrköping, Sweden
[4]Now at EDF R&D, OSIRIS Department, 7 boulevard Gaspard Monge, 91120 Palaiseau, France

Correspondence to: Vazken Andréassian (vazken.andreassian@inrae.fr)

Abstract. The transferability of hydrological models over contrasted climate conditions, also identified as model robustness, has been the subject of much research in last decades. The occasional lack of robustness identified in such models is not only an operational challenge – since it affects the confidence that can be placed in projections of climate change impact – but it also hints at possible deficiencies in the structure of these models. This paper presents a large-scale application of the robustness assessment test (RAT) for three hydrological models with different levels of complexity: GR6J, HYPE and MIKE SHE. The dataset comprises 352 catchments located in Denmark, France and Sweden. Our aim is to evaluate how robustness varies over the dataset and between models and whether the lack of robustness can be linked to some hydrological and/or climatic characteristics of the catchments (thus providing a clue on where to focus model improvement efforts). We show that although the tested models are very different, they encounter similar robustness issues over the dataset. However, models do not necessarily lack robustness on the same catchments and are not sensitive to the same hydrological characteristics. This work highlights the applicability of the RAT regardless of model type and its ability to provide a detailed diagnostic evaluation of model robustness issues.

1 Introduction

1.1 Hydrological modelling under climate change

Several recent international initiatives have raised concern about the issue of model robustness in hydrology. By *model robustness* we mean the ability of a hydrological model to adapt to contrasting climate conditions. For example, the Panta Rhei decade of the International Association of Hydrological Sciences (IAHS) (Montanari et al., 2013) and the Unsolved Problems in Hydrology (UPH) initiative of Blöschl et al. (2019) (see e.g. UPH no. 19: *How can hydrological models be adapted to be able to extrapolate to changing conditions?*) questioned the applicability of hydrological models in the context of global change. In parallel, a large number of hydrological modelling studies have been carried out to understand how

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

climate change impacts hydrology (see e.g. Intergovernmental Panel on Climate Change, IPCC, Pachauri et al., 2014), and it seems essential to verify that the models used for this purpose withstand non-stationary climate conditions.

Over the past decade, several publications (e.g. Refsgaard et al., 2014; Thirel et al., 2015; among others) highlighted that hydrological models are not as independent of climate conditions as was expected. Indeed, models can be sensitive to the climate conditions of the period on which they were set up or calibrated (see e.g. Vaze et al., 2010; Coron et al., 2011). This dependency can be revealed using the split-sample testing (SST) approach proposed by Klemeš (1986), which consists in testing the model on different time periods for in calibration or evaluation (see Sect. 1.2). In split-sample experiments, model performance commonly decreases when switching from the calibration period to the evaluation period, and it has been shown that this decrease is intensified as the difference in climate conditions between periods increases (Brigode et al., 2013; Westra et al., 2014).

Different ad hoc solutions have been proposed to address this symptom. Varying the parameter values according to climate conditions is one such solution. For example, Gharari et al. (2013) proposed a method to calibrate time-consistent parameters based on the distance to Pareto optimum, while other studies focused on time-variant parameters linked to climate conditions (Stephens et al., 2019; Zeng et al., 2019; Lan et al., 2020). Although these methods make it possible to improve robustness, they are not curative, i.e. they serve as a 'patch' for models that need to withstand changes in climate. They do not explain the reasons behind the symptoms: Why do model parameters exhibit this kind of unwanted dependence on climate? And why does this occur on some catchments and not on others?

1.2 Assessing model robustness from the perspective of a changing climate

In hydrological practice, model robustness has traditionally been assessed using the SST (Klemeš, 1986). Klemeš (1986) introduced four levels of the SST, in which the third one, called the differential SST (DSST), aimed at evaluating a model over a period where climate conditions differ from those of the calibration period. After a few early attempts to apply the DSST scheme (e.g. Refsgaard and Knudsen, 1996; Donnelly-Makowecki and Moore, 1999; Xu, 1999; Seibert, 2003), this test was more extensively used over the past decade to check the robustness of rainfall–runoff models under a changing climate (Vaze et al., 2010; Broderick et al., 2016; Dakhlaoui et al., 2017; Rau et al., 2019).

In addition, some authors proposed improvements to the DSST: Coron et al. (2012) suggested a generalized version of the SST (GSST) designed to evaluate models over all possible combinations of time periods; Gelfan and Millionshchikova (2018) introduced in the DSST a component that depends on model performance to avoid selecting apparently robust models with poor performances; Dakhlaoui et al. (2019) proposed a generalized differential SST (GDSST) by adding a bootstrap selection tool to create a number of contrasting climatic sub-periods; Gelfan et al. (2020) proposed a more complex evaluation strategy that uses DSST in one step of the analysis. All of the aforementioned methods remain linked to SST and include either one or several calibration steps. However, the use of calibration and evaluation periods is not always suitable for assessing the robustness of models that are calibrated manually or that have complex calibration procedures, or even no calibration at all.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

When searching for a more widely applicable methodology, Nicolle et al. (2021) proposed a test inspired by the GSST of Coron et al. (2012) and by the subsequent work of Coron et al. (2014): the robustness assessment test (RAT). The RAT is designed to highlight unwanted correlations between climatic conditions and model performances, as these may represent an issue in modelling the hydrological cycle under a changing climate. The proposed RAT was found to give results similar to GSST for catchments in France. In addition, the RAT has the major advantage of requiring only a single simulated flow time series (and an observed one for comparison): there is no need to resort to multiple calibration experiments. Therefore, the RAT can be used to compare the robustness of different models with minimal effort.

However, detecting cases where a model lacks robustness is not sufficient: we also need to understand the underlying reasons for this flaw. For example, Sleziak et al. (2018) used DSST in Austria and identified an influence of land cover and catchment wetness on robustness. Birhanu et al. (2018) compared the model robustness of four models in order to evaluate how model complexity influences the robustness. They concluded that catchment characteristics play a more important role in the lack of robustness than model complexity. However, it is often difficult to link the lack of robustness to model characteristics or to specific hydrological processes.

1.3 Scope of the paper

This paper aims at moving a step forward in our understanding of what makes a model occasionally sensitive to climate change. The RAT (Nicolle et al., 2021) is applied to a large set of catchments spanning various climate conditions in three European countries, in order to evaluate the robustness of three rainfall–runoff models with various process representations and parameter estimation approaches. The large test set is used to evaluate how model robustness varies over a wide range of conditions and to characterize catchments where models lack robustness. The use of three different models will provide more general conclusions for characterizing catchments that raise robustness issues in hydrological modelling.

2 Evaluation method

2.1 The robustness assessment test

The robustness assessment test (RAT, Nicolle et al., 2021) is chosen since it can be applied without controlling the model calibration process. Indeed, the three models used for this experiment were calibrated once and separately at the three institutes involved in this study. The RAT only requires observed climatic variables (to be used as a potential predictor for the model bias), as well as simulated and observed flows covering a sufficiently long time period (at least 30 years, as shown in the study by Nicolle et al., 2021).

Figure 1 summarizes the three steps of the RAT procedure: (i) the time series of the climatic predictors and flow are aggregated by hydrological year, (ii) a score assessing the difference between observed and simulated flows (here bias) is computed for each year, (iii) the correlation between this annual score and the annual values of a chosen predictor is analysed. A significant correlation between the score and the predictor will reveal suspicious dependencies that may affect

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

the model extrapolation capacity. In this case, we will say for the sake of simplicity that the model "reacts" to the RAT for the catchment in question. Similarly, the catchment on which the model reacts will be termed a "reactive catchment". Behind these terms, let us stress that the "reaction" is an unhealthy sign (it is definitely not what modellers aim for), and this does not tell us the causes of the behaviour: i.e. whether the issues are due to the model structure or parameters, or to the presence of a trend in the observed data.

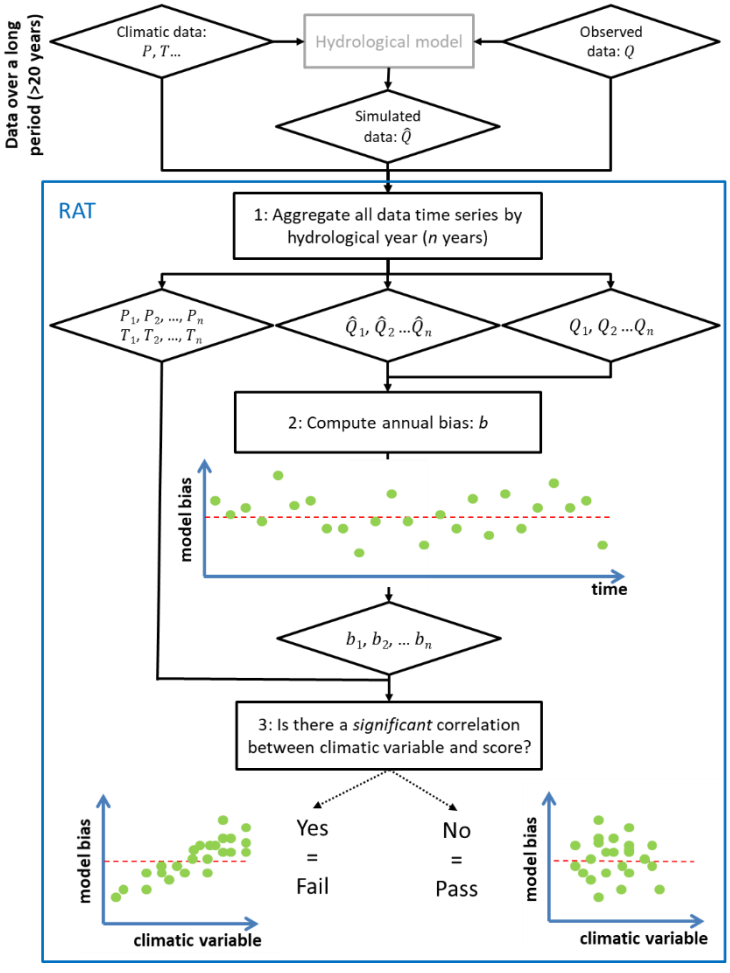


Figure 1: Flow chart of the robustness assessment test (RAT), with the three steps necessary to evaluate the robustness of a hydrological model (from Nicolle et al., 2021)

In this study, we consider hydrological years to be between 1 October and 30 September. The relative bias is computed every year between the observed and simulated flows (see Eq. (1) in Nicolle et al., 2021). Three climatic variables are used as potential predictor and compared with the bias: the annual mean air temperature [°C], the annual precipitation [mm y^{-1}] and the annual value of the humidity index, which is the ratio between the annual precipitation and the annual potential

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

evaporation [-]. The correlation test is based on the Spearman correlation, so as to handle non-linear relationships. The significance threshold is set at a p-value of 0.05.

2.2 Catchment set

The RAT is applied to a large catchment set over western and northern Europe (Figure 2) to test the method and evaluate robustness over a variety of catchments. The dataset comprises a total of 352 catchments, in which 146 are located in France, 43 in Denmark and 163 in Sweden. The dataset was set up by partners that collaborated in this work (INRAE in France, GEUS in Denmark and SMHI in Sweden). The catchment area varies from approx. 1 to 27,000 km^2 with a median of 530 km^2. The catchments cover a wide range of hydrological regimes (including contrasted or non–contrasted pluvial regimes, nival regimes, and mixed regimes) and four Köppen–Geiger classes (temperate with no dry season and warm summer: Cfb; temperate with dry and warm summer: Csb; continental with dry and warm summer: Dfb; and continental with dry and cold summer: Dfc).

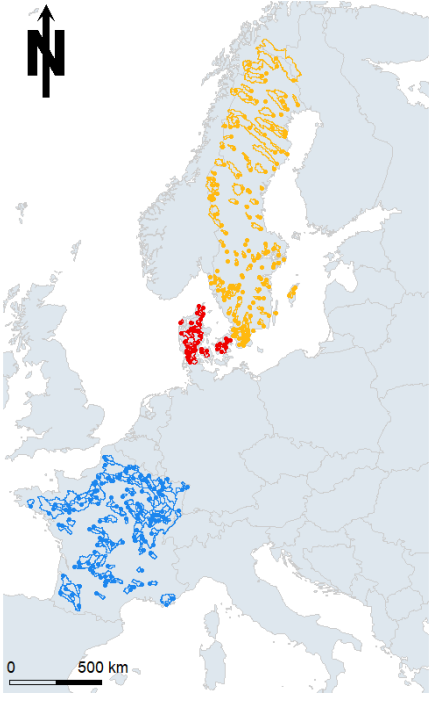


Figure 2: Location and boundaries of the catchments used for this study.

The hydrology of French rivers is under a double influence: geology and climate. The catchments located on the sedimentary deposits in the north and south-west are strongly buffered by the role of connected aquifers, and often strongly karstified in Jurassic plateaux. By comparison, the Hercynian granitic massifs (in central and western France) show a more classic

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

hydrology, typical of superficial catchments. In the Pyrenees, the Alps, the Jura and the Vosges mountain ranges, hydrology can be heavily influenced by snowmelt. Around the Mediterranean Sea, and especially in the highlands, very heavy precipitation causes flash floods almost every autumn. The rest of the French territory has a rather mild (temperate) climate.

Swedish hydrology is characterized by decreasing air temperature from south to north, and decreasing wetness from west to east. The highest runoff occurs in the mountain range along the western border with Norway, where the largest rivers originate, and also on the south-western coast. The south-east is rather dry. Most of the large rivers are developed for hydropower production, and water is stored in lakes and reservoirs for hydroelectricity production in the winter. Sweden also has many lakes, which act as natural reservoirs.

In Denmark, hydrology varies from west to east. Geologically speaking, the western part of the Jutland peninsula (continental part of Denmark) is dominated by glacial outwash sand and gravel formations where precipitation easily infiltrates and that often form large inter-connected aquifers. The eastern part of Denmark is characterized by till and moraine deposits with high clay content that are drained by tile drains and numerous smaller streams. As a result, less surface runoff and other fast flow components (drain flow) are generated in the western part compared to the eastern part. Therefore, streamflow in the western part of the country is dominated by baseflow while overland flow is rarely an important flow component (van Roosmalen et al., 2007). By contrast, the catchments in the eastern part of Denmark are more responsive with more variable flow (Henriksen et al., 2021).

Several climate characteristics, named climatic signatures, were calculated for each catchment (see statistics in Table 1). Repartition maps and distributions of these climate characteristics at a national scale are provided in Supplementary materials 1 and 2, respectively. Annual precipitation and potential evaporation show a wide variability over the dataset. Catchments with the highest amount of precipitation are located in southern and eastern France, southern Denmark and western Sweden while catchments with the lowest precipitation amount are found in eastern Sweden mostly. Regarding the humidity index, the catchments are all relatively humid with the driest catchments in south-eastern Sweden and in south-eastern and northern France. The indexes of precipitation variability and intensity are higher in eastern Sweden and south-eastern France and lower in western Sweden. The fraction of days without precipitation varies between 7% and 64% over the dataset; in half of the catchments the percentage of dry days is between 35% and 45%. This fraction is higher in south-eastern France and lower in northern Sweden. The seasonality index (de Lavenne and Andréassian, 2018) characterizing the synchronicity between precipitation and potential evaporation varies from 0.18 to 0.51. The lowest seasonality index values (mainly found in north-western Sweden) mean that runoff is favoured over potential evaporation, because precipitation mainly occurs when evaporative demand is low. High seasonality index values, found in northern France and south-eastern Sweden, mean that potential evaporation is favoured. Snowfall fraction varies between 0 and 57% with a south–north gradient, but more than half of the catchments have less than 10% of snowfall. The distribution of climatic characteristics by country provided in Supplementary material 2 also shows that these characteristics vary strongly across France and Sweden. In Denmark, however, the distribution shows less spatial variability with values around the average of the dataset.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

Table 1: Distribution of climatic signatures over the catchment set (all three countries)

Signature description	Abbreviation in this paper	Quantile (%)				
		Min	25	50	75	Max
Mean annual precipitation [mm y^{-1}]	P_{MA}	408	676	826	960	1502
Mean annual potential evaporation [mm y^{-1}]	E_{MA}	222	478	563	668	843
Humidity index (P_{MA}/E_{MA}) [-]	I_{HUM}	0.81	1.24	1.44	1.73	5.47
Precipitation variability (coefficient of variation) [mm d^{-1}]	P_{CV}	1.40	1.79	1.88	1.99	3.34
Precipitation intensity index (daily precipitation percentile 99 divided by daily mean precipitation) [-]	P_{int}	6.38	8.23	8.72	9.46	17.11
Mean annual ratio of days without precipitation [-]	D_{WoP}	0.07	0.36	0.39	0.43	0.64
Seasonality index (synchronicity between precipitation and potential evaporation occurrence) [-]	I_{Seaso}	0.18	0.40	0.42	0.44	0.51
Solid precipitation fraction (precipitation that occurs when daily temperature is below 0°C) [-]	S_{Frac}	0.00	0.03	0.06	0.15	0.57

Statistics on flow signatures are compiled in Table 2, where most of the flow signatures are calculated following Westerberg and McMillan (2015). The repartition maps and distributions at national scale of these flow characteristics are provided in Supplementary materials 3 and 4, respectively. Mean flow varies from 95 to 1344 mm y^{-1} with low values in northern France and south-west Sweden and high values in western Sweden. Regarding the runoff ratio, values vary between 15% and 124%, with the highest values in northern Sweden and the lowest values in central France. Five catchments, located in the mountains of north-west Sweden, have values greater than 100%. These values may be the result of an underestimation of the precipitation measurement due to orographic effects that are not captured by the interpolation method used.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences Discussions

Low flows are characterized by several descriptors: the low percentiles (0.01 to 5), the frequency and duration of low-flow events, the baseflow index (I_{BF}, from Pelletier and Andréassian, 2020) and the variability of low flows. Catchments with very low flows are located in southern and south-eastern Sweden and in central France. These catchments also have a low variability of low flows and a high frequency and duration of low-flow events. By contrast, catchments in continental Denmark (Jutland peninsula) and northern France are characterized by higher values of low flows that are more variable. The occurrence and duration of low-flow events are lower in these regions, and high I_{BF} values show that aquifers play a key role in the hydrology of these regions.

High flows are determined by the high quantiles (85 to 99) as well as the frequency and duration of high-flow events and their variability. The values of high quantiles vary considerably over the dataset (e.g. Q_{99} varies from 0.67 to 27 mm d^{-1}) and the highest values are located in western Sweden. Regarding the frequency and duration of high-flow events, no clear geographic pattern emerges. Flow variability is higher in France and lower in Denmark.

Finally, three signatures are computed to measure flow dynamics: the slope of the flow duration curve that evaluates flashiness, the overall flow variability and the 1-day autocorrelation. The slowest catchments are located in Denmark and northern France while the fastest-responding catchments are found in south-eastern Sweden and south-eastern France.

Table 2: Description and distribution of flow signatures over the catchment set (from Table 2 in Westerberg and McMillan, 2015). The abbreviations in column 2 are used in Sect. 3 and 4.

Signature description	**Abbreviation in this paper**	**Quantile (%)**				
		Min	**25**	**50**	**75**	**Max**
Mean flow [mm y^{-1}]	Q_{mean}	95	248	347	447	1344
Flow percentiles (0.01, 0.1, 1, 5, 50, 85, 95 and 99%) [mm d^{-1}]	$Q_{0.01}$	0.00	0.01	0.03	0.10	0.76
	$Q_{0.1}$	0.00	0.01	0.05	0.11	0.85
	Q_{1}	0.00	0.03	0.07	0.16	1.02
	Q_{5}	0.00	0.06	0.11	0.23	1.10
	Q_{50}	0.10	0.38	0.54	0.80	1.91
	Q_{85}	0.41	1.23	1.67	2.17	7.58
	Q_{95}	0.54	2.15	3.07	4.14	14.36
	Q_{99}	0.67	3.60	5.47	7.80	27.01
High-flow event frequency (mean number of days with flow over 9 times the median flow) [d y^{-1}]	Q_{hffreq}	0.0	0.6	5.9	13.2	53.3
High-flow event mean duration	Q_{hfdur}	1.0	2.6	3.8	5.7	17.0

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

[d]						
Low-flow event frequency (mean number of days with flow below 0.2 times the mean flow) [d y^{-1}]	Q_{lffreq}	0.0	24.9	62.4	99.0	249.8
Low-flow event mean duration [d]	Q_{lfdur}	1.0	11.6	17.1	27.8	130.5
Baseflow index [-]	I_{BF}	0.01	0.13	0.25	0.43	0.90
Slope of flow duration curve (from 33% to 66% exceedance values) [-]	S_{FDC}	0.48	1.38	1.67	1.89	2.74
Overall flow variability (daily flow coefficient of variation) [-]	Q_{CV}	0.24	0.95	1.20	1.43	2.70
Low-flow variability (mean annual minimum flow above median flow) [-]	Q_{LV}	0.00	0.12	0.21	0.35	1.74
High-flow variability (mean annual maximum flow above median flow) [-]	Q_{HV}	1.6	6.4	11.7	19.4	88.1
One-day autocorrelation of flow [-]	Q_{AC}	0.37	0.88	0.94	0.98	1.00
Runoff ratio (flow divided by precipitation) [-]	R_R	0.15	0.34	0.43	0.55	1.24

The signatures listed in Table 1 and Table 2 are used in order to investigate potential factors affecting the robustness of the three models tested. Catchments on which each model reacts to RAT are compared with catchments where the model does not react. We use a Mann–Whitney *U* test (Wilcoxon, 1945; Mann and Whitney, 1947) to identify whether the distributions of the two signatures are significantly different (note that the same method was used, e.g. by Fowler et al. (2016) to compare catchment characteristics). The Mann–Whitney *U* test evaluates the probability that two groups originate from the same distribution by focusing on the relative rank of the groups. We use a classic (but nonetheless arbitrary) threshold for the p-value: 0.05. These tests will allow us to target the robustness issues within the models and to better understand the RAT results.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

2.3 Used data

For each catchment, daily precipitation, daily mean air temperature (referred to as "temperature" in this paper) and daily potential evaporation are available to run the models and to apply the RAT. For French catchments, temperature and precipitation are extracted from the SAFRAN reanalysis (Vidal et al., 2010). SAFRAN covers France on an 8 km grid and climatic data are aggregated by catchments (Delaigue et al., 2022). Potential evaporation is calculated using the formula proposed by Oudin et al. (2005). These data are available over 61 calendar years between 1958 and 2018. It should be noted that for the interpretation of the results, the location of ground stations used by SAFRAN to build the reanalysis can change over the available period and therefore can have an impact on the model robustness. River flow data are available for each catchment outlet from the Banque HYDRO database (Leleu et al., 2014). Periods of flow data availability vary for each catchment: from 27 to 61 years between 1958 and 2018 with an average close to 50 years.

For Sweden, daily temperature, precipitation and observed flow are available for the same 35 calendar years between 1981 and 2016. Potential evaporation is also calculated for each catchment using the Oudin formula. Precipitation and temperature data are extracted from the PTHBV database (Johansson, 2002). This database covers Sweden on a 4 km grid and is based on extrapolation from measurement station data. River flow data for the 163 gauged stations are extracted from the official database of SMHI gauging stations. Meteorological data are available at a sub-catchment scale of an average size of 13 km^2 and are aggregated at catchment scale.

For Danish catchments, data on precipitation, temperature, potential evaporation and flow are available for the same 30 calendar years between 1989 and 2019. A dynamic gauge catch correction (Stisen et al., 2011) is applied to precipitation and the results are subsequently interpolated to a 10 km grid (Scharling and Kern-Hansen, 2012). Potential evaporation is calculated using the Makkink equation adjusted for Danish conditions (Scharling, 1999). The Makkink equation is a global radiation-based simplification of the Penman equation. Both temperature and potential evaporation are available on a 20 km grid resolution. Daily data on river flow are available from the national database ODA (Surface water database; https://odaforalle.au.dk/main.aspx). To minimize correlation between discharge time series, there are no nested catchments in the Danish dataset.

2.4 Hydrological models

The robustness of three models is evaluated in this work. The models were set up, calibrated and run by the three contributing groups of this work, according to their own expertise. Table 3 presents a brief description of the three models.

GR6J (Pushpalatha et al., 2011) is a lumped bucket-type model that simulates catchment runoff response to rainfall using six free parameters. This model derives from the GR4J model (Perrin et al., 2003) and is run using the "airGR" R package (Coron et al., 2017, 2021). Snow accumulation and melt are calculated using the CemaNeige routine (Valéry et al., 2014) that splits the catchment into five elevation bands and simulates snow processes with two additional parameters. The GR6J model is calibrated against observed flow for each catchment using the KGE criterion (Gupta et al., 2009) calculated on

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences Discussions

square root transformed flows as objective function. The calibration is done automatically using a mixed global–local search optimization algorithm presented by Coron et al. (2017). The period used for calibration covers all the available flow data minus a 4–year warm-up period to initialize the internal state variables (store levels). GR6J is the only model that we are able to apply on all the catchments of the dataset.

HYPE (Lindström et al., 2010) is a process-based semi-distributed model that was designed for both quantity and quality modelling. Here, we use on Swedish catchments only, the Sweden-scale version (S-HYPE, Strömqvist et al., 2012). S-HYPE has been developed continuously since the first version described by Strömqvist et al. (2012). In the version used here (S-HYPE-2016b) the whole country is divided into sub-catchments of an average size of ca. 13 km^2. These sub-catchments are divided into hydrological response units (HRUs) that depend on soil type and land uses. A large number of parameters are used to adapt the model, which are spatialized by sub-catchments, land use and soil types. Local super parameters, i.e. deviations in key characteristics (see Lindström, 2016), are also calibrated for parameter regions in S-HYPE. The S-HYPE model was calibrated manually. Since the model is used (among other things) operationally for flood warning at the SMHI, calibration was focused primarily on the timing of discharge, and secondly on volume errors. The NSE (Nash and Sutcliffe, 1970) is very sensitive to timing errors and was therefore used as the main numerical criterion in the calibration process. Results are available for all Swedish catchments in the dataset for the entire period of flow data availability at https://www.smhi.se/data/hydrologi/vattenwebb.

The MIKE SHE/MIKE 11 modelling system (Graham and Butts, 2005), only used for the Danish catchments, has a physically–based and fully distributed description of the terrestrial hydrological cycle. It is based on a three-dimensional description of the saturated zone that is parameterized according to a geological model. Drainage flow is conceptualized as a linear reservoir assumed to occur when the water table is above the position of the drains. The unsaturated zone is described by a simple water balance module termed the “two-layer” method. Evaporation is described by a simple method accounting for the water balance in the root zone. Two-dimensional overland flow is simulated using a diffusive wave approximation. Flow is simulated as a one-dimensional process by MIKE 11 using the kinematic routing approach. The model is discretized into a 500 m horizontal grid with 11 computational layers and is run with daily inputs on climatic forcing. More information on the model is found in the manual (DHI, *MIKE SHE, User Guide and Reference Manual*). For this work, the MIKE SHE version set up and applied by the National Water Resources Model (Højberg et al., 2013) is used. The model is calibrated using auto-calibration provided by PEST (Højberg et al., 2013). Based on a sensitivity analysis, the most sensitive parameters are selected as free parameters including hydraulic conductivities of the geological units, drainage time constant, river–aquifer exchange coefficient and root depth of the dominant soil type. Several less sensitive parameters are tied to the free parameters.

As shown in Table 3, the three models have different process representations. They also have different spatial resolutions and different methods for parameter estimation. Since these three models cover various modelling approaches, they potentially have differences in robustness and this work analyses how their structure influences their robustness.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences Discussions

Table 3: Main characteristics of the three models used

	GR6J	**HYPE**	**MIKE SHE**
Spatialization	Lumped	Semi-distributed (sub-catchments + HRUs)	Distributed (grid to hillslope)
Model time step	Daily	Daily	Daily
Parameter estimation procedure	Automatic (OF = KGE on square root transformed flows)	Manual (aided by NSE value)	Automatic (OF= 8 metrics including NSE, water balance, and ME, etc.)
Number of estimated parameters (and their spatialization)	6 parameters for the rainfall–runoff model + 2 parameters for the snow-accounting routines	About 100 (soil and land-use dependent) + local tuning	9 free parameters and several tied parameters. Homogeneity assumed to a relatively high degree
Process complexity (as stated by Hrachowitz and Clark, 2017)	+	++	+++

3 Lessons learnt from a single model applied to the entire dataset

Out of the three models used in this paper, we were able to apply only one model, GR6J, to the entire dataset (because of its relative simplicity of calibration). The results of GR6J are therefore used to evaluate how robustness varies over the three countries studied. A geographic analysis is first carried out, followed by an analysis to link the occasional lack of robustness to catchment characteristics.

3.1 Overall evaluation

Figure 3 shows the location and the number of catchments where GR6J reacts to the RAT (i.e. a significant correlation exists between the bias and a given predictor), for the three predictors used (temperature, precipitation and humidity). When temperature is used as a predictor for the RAT, GR6J fails the robustness test over 99 catchments (28% of the total). When precipitation and humidity index are considered, GR6J fails the robustness test on 16% and 18% of the catchment set, respectively. Note that these numbers are above the 5% threshold that we would expect to observe if only chance was playing a role. This shows that the model has a significant robustness issue over the dataset.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

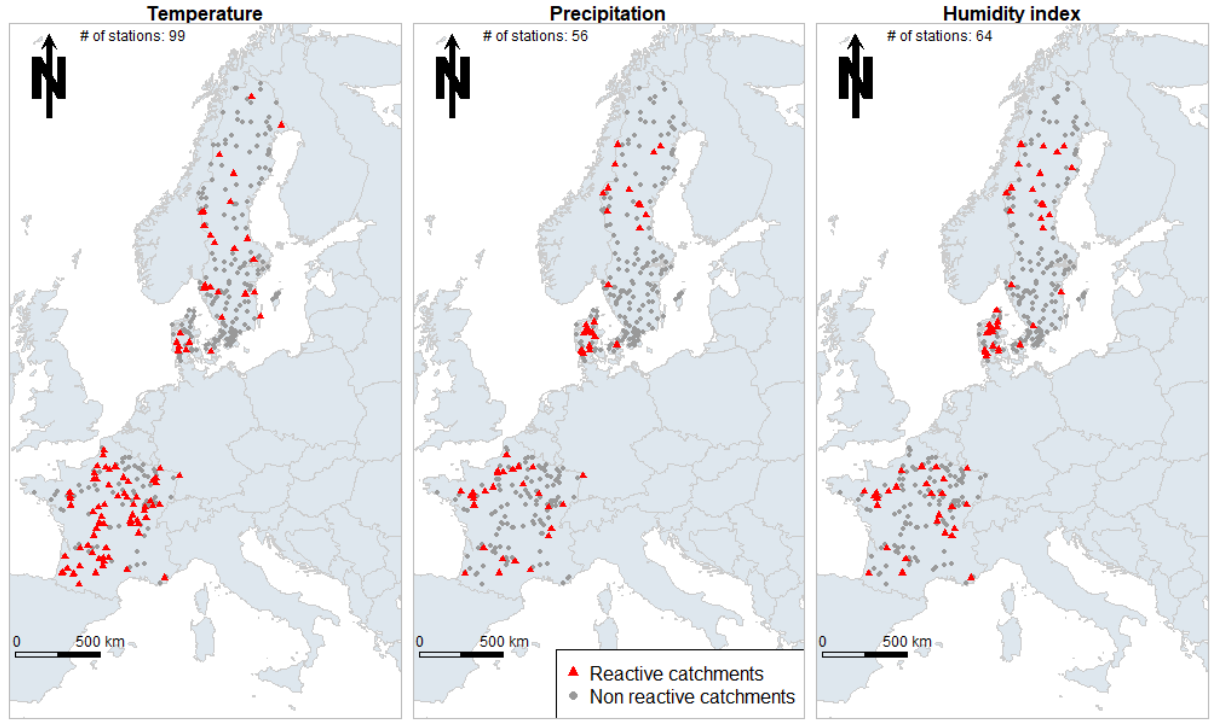


Figure 3: Location of the catchments where GR6J reacts to the RAT (in red) using temperature (left), precipitation (centre) and humidity index (right) as predictors. Numbers at the top of the maps represent the numbers of reactive catchments out of the total of 352 catchments

280

The spatial distribution of the reactive catchments follows different patterns when temperature or precipitation is used as a predictor in the RAT: (i) when temperature is used as a predictor, reactive catchments are especially numerous in France; (ii) when sensitivity to precipitation is considered, there are fewer reactive catchments in France but more in Denmark; (iii) results obtained with humidity index and precipitation are very similar (this was expected because the humidity index is 285 calculated as the ratio of the precipitation amount to the potential evaporation amount: since the annual variability of precipitation is much higher than the variability of the potential evaporation, it is logical to observe similar results when the two variables are used as predictors).

Overall, the reactive catchments where GR6J is identified as lacking robustness are often grouped together geographically, which indicates that some common (regional) hydrological features cause this problem. For example, catchments react more 290 often in the Jutland peninsula and in northern Sweden when precipitation or humidity index is used as a predictor.

However, we cannot identify any obvious reason for the spatial pattern of the reactive catchments. For example, it is not clear why so many reactive catchments are located in France when temperature is used as a predictor. An example of these is given in Figure 4, in which the temperature is clearly correlated with the bias (bottom left panel) while no clear correlation appears for precipitation and humidity index time series (bottom centre and right panels). A hypothesis could be the higher

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

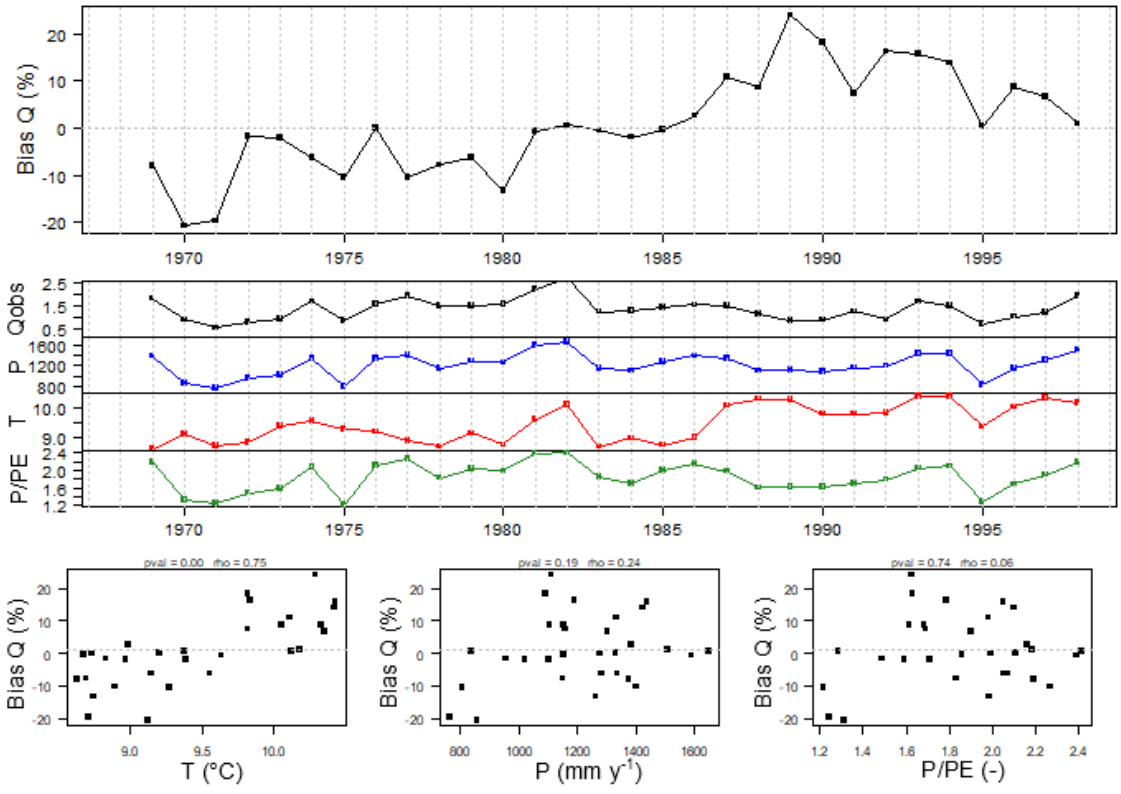


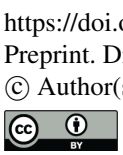
Figure 4: Illustration of the RAT results for the GR6J model applied to the Ognon River at Chavigny-sur-l'Ognon (north-east France): top plot represents the annual streamflow bias time series, the four middle plots represent the time series of the annual flow values and of the three climatic predictors, and bottom scatter plots represent the correlation between bias and annual temperature (left), precipitation (centre) and humidity index (right)

The conclusion of this series of tests on GR6J is that the model seems to have robustness issues over the dataset but that, at this point, the RAT results cannot be explained by the location of the reactive catchments alone. Thus, catchment characteristics are included in the analysis to evaluate whether robustness issues could possibly be explained by the specificities of local hydrology and whether this could be linked to the structure of the models.

3.2 Link to catchment hydro-climatic characteristics

In order to investigate potential factors affecting the robustness of the GR6J model, we analyse catchment characteristics. Catchments on which GR6J reacts to the RAT are compared with those where GR6J does not react to the RAT. Figure 5

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024



Hydrology and Earth System Sciences
Discussions

shows an example of the methodology for mean annual precipitation over the catchment. The boxplot represents the distribution of mean annual precipitation, on the left for catchments where GR6J reacts to the RAT and on the right for catchments where GR6J does not react to the RAT. This shows that GR6J is less robust on the drier catchments (with temperature used as a predictor). For precipitation and humidity index, no significant differences in mean annual precipitation exist between reactive catchments and non-reactive ones.

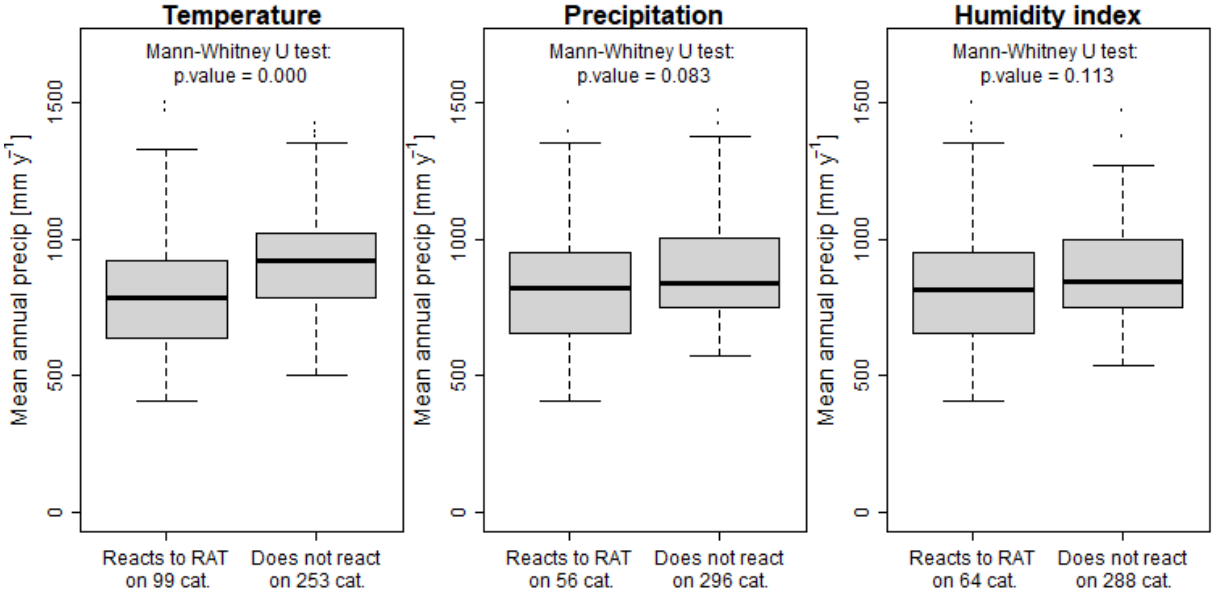


Figure 5: Comparison of the catchment area distribution for catchments where GR6J reacts or does not react to the RAT using temperature (left), precipitation (centre) and humidity index (right) as predictors.

Following the same methodology, Figure 6 shows the results of the Man–Whitney *U* test described above for the climatic signatures listed in Table 1. It indicates those signatures for which the difference between reactive and non-reactive catchments is significant. If the colour is red (blue), the Mann–Whitney *U* test indicates that reactive catchments have lower (higher) values of the signature than non-reactive catchments. The shade of the dot colour indicates how significant the difference is: if it is grey no significant difference exists (p-value higher than 0.05); if it is dark red or blue, the difference is highly significant (p-value lower than 0.01).

When temperature is the predictor, Figure 6 shows that the catchments on which GR6J reacts to the RAT have higher precipitation and potential evaporation amounts and a higher number of dry days. The higher seasonality index indicates that precipitation mainly occurs during the low potential evaporation season (low synchronicity between high precipitation and high potential evaporation). The amount of precipitation that falls as snow is also lower than the dataset average.

These results are not straightforward to interpret. The low synchronicity between precipitation and potential evaporation emphasized by the seasonality index values reveals that the reactive catchments have drier warm seasons (high potential

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

evaporation and low precipitation season). The reactive catchments are also mainly located in France where potential evaporation is the highest. The link between these two signatures may lead to dry seasons on which potential evaporation has a major impact. Given that potential evaporation is directly calculated from temperature, changes in temperature may influence hydrology during the warm season and it is possible that GR6J has difficulties in handling these inter-annual changes in potential evaporation.

If either precipitation or humidity index is used as a predictor, the difference between the two distributions does not show a similar pattern. Lower differences in climatic signature are evident between reactive and non-reactive catchments. The only discernible result is that, when precipitation is the predictor, catchments on which GR6J reacts to the RAT have less solid precipitation and/or a higher potential evaporation amount.

Consequently, it is difficult to find an explanation in terms of model representation based on climatic considerations and thus we now address flow signatures. We can only stress that the snow module is not the source of lack of robustness here, since the snow fraction is lower for reactive catchments.

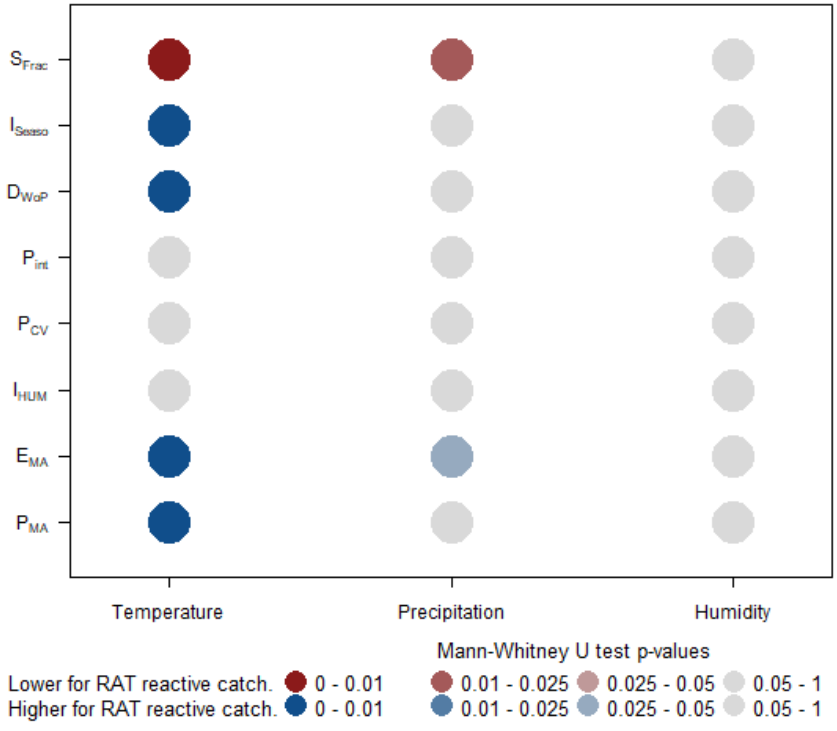


Figure 6: Results of the Mann–Whitney U test to evaluate the difference in climatic signatures (see Table 1) between catchments on which GR6J reacts to the RAT and catchments on which GR6J does not. The number of catchments in each subset can be found in Figure 5. Blue (red) circles mean that the signature is significantly higher (lower) for reactive catchments. P_{MA}: mean annual precipitation, E_{MA}: mean annual evaporation, I_{HUM}: humidity index, P_{CV}: precipitation variability, P_{int}: precipitation intensity index, D_{WoP}: days without precipitation ratio, I_{Seaso}: seasonality index, S_{Frac}: snow fraction.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

We now look at flow signatures to interpret robustness failures. Figure 7 complements the description of catchments on which GR6J reacts to the RAT. When temperature is used as a predictor, the reactive catchments are characterized by a low runoff ratio. The low autocorrelation and the short duration of low-flow and high-flow events suggest that the reactive catchments are more responsive than the dataset average.

Similarly to what was explained for Figure 6, the low runoff ratio for reactive catchments indicates that potential evaporation may have more influence on these catchments.

When precipitation and humidity index are used as predictors, low flows also seem to have relatively high values ($Q_{0.01}$ to Q_5 and high I_{BF}) and high variability on reactive catchments. Regarding the slope of the flow duration curves, the catchments that react to the RAT seem slower than average. Only when precipitation is the predictor, the total flow variability and high-flow variability are also below normal.

The catchments with potential robustness issues are characterized by slow response with high baseflow. Similar observations were made by Sleziak et al. (2018), who showed that the lack of robustness in Austrian catchments was higher for catchments with slow response ("dominant soil moisture regime"). In the present work, this can be explained since, in this kind of catchment, conditions of precipitation and humidity of a given year may influence flow during several subsequent years (possibly due to groundwater storage). It is known that GR6J has difficulties in representing this behaviour, described by de Lavenne et al. (2022) as the "catchment memory". The RAT results suggest that this flaw in the model may lead to robustness issues.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

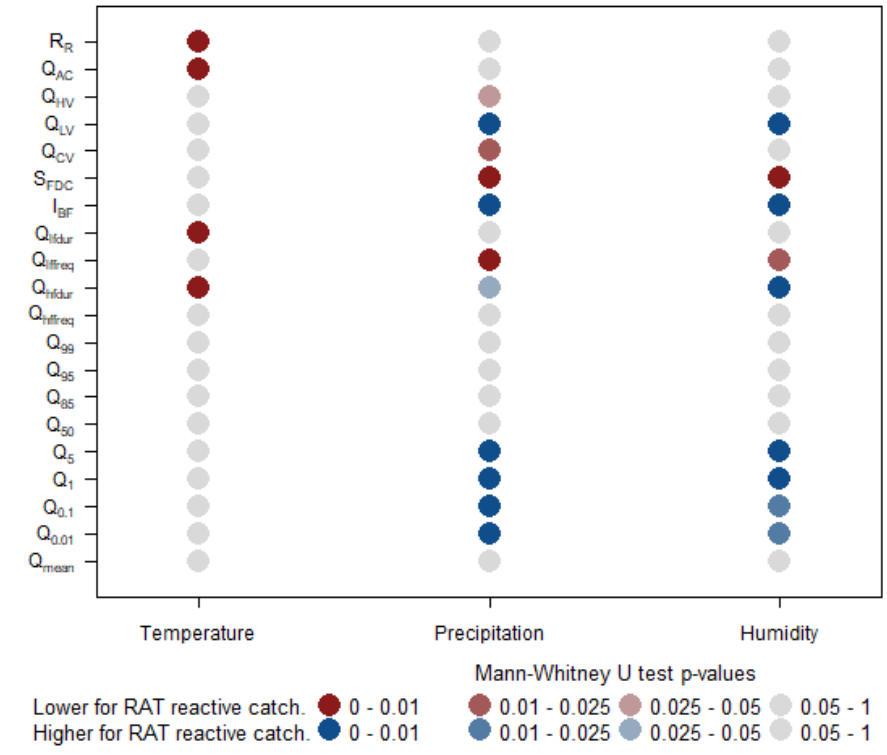


Figure 7: Results of the Mann–Whitney U test to evaluate the difference in flow signature between catchments on which GR6J reacts to the RAT and catchments on which GR6J does not. The number of catchments in each subset can be found in Figure 5. Blue (red) squares mean that the signature is significantly higher (lower) for reactive catchments. Q_{mean}: annual mean flow, $Q_{[0.01–99]}$: flow percentiles, $Q_{[hf-lf]freq}$: frequency of [high-low]flow events, $Q_{[hf-lf]dur}$: duration of [high-low] flow events, I_{BF}: baseflow index, S_{FDC}: slope of the flow duration curve, $Q_{[C-L-H]V}$: [total-low-high]flow variability, Q_{AC}: flow 1 day autocorrelation, R_R: runoff ratio.

To summarize, these evaluations do not lead to clear explanations on the lack of robustness of GR6J. However, two paths can be explored to improve its robustness: (i) when temperature changes over the catchment, the robustness of GR6J could be increased by improving its capability to handle inter-annual changes in potential evaporation, (ii) when a precipitation trend impacts the catchment, the robustness of GR6J could be improved by a better consideration of the catchment memory within the model.

4 Comparing model robustness in Denmark and Sweden

Here, we compare the robustness of the three models presented in Sect. 0. By applying the RAT to these models, our goal is to understand whether the catchments detected by the RAT as reactive are model-specific. In addition to this, we aim at highlighting the differences between the models and try to interpret these differences.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

4.1 S-HYPE vs GR6J in Sweden

Figure 8 compares the catchments on which GR6J and S-HYPE react to the RAT in Sweden. The numbers of reactive catchments are similar for the two models but their location varies, even if some catchments are common to the two models. When temperature is used as the predictor, catchments on which S-HYPE reacts to the RAT are mainly located in the Scania region (extreme south of Sweden). Catchments on which GR6J reacts to the RAT are scattered over Sweden, but the large number observed for S-HYPE in Scania is not observed for GR6J. It is, however, interesting to note that catchments in mid-western Sweden seem to present a robustness issue for both models. When precipitation and humidity index are taken as predictors, GR6J reacts in northern regions while S-HYPE reacts more in central and south-eastern Sweden. Overall, Figure 8 shows that S-HYPE and GR6J react differently to the RAT and indicates that their lack of robustness probably has different origins.

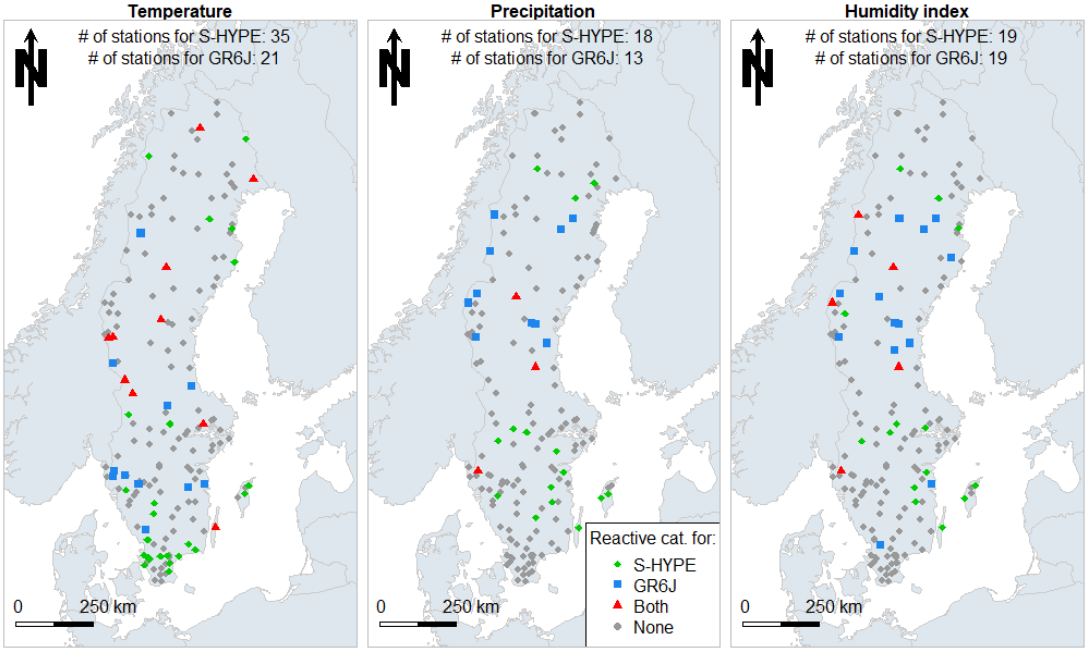


Figure 8: Location and number of Swedish catchments on which S-HYPE and GR6J react to the RAT using temperature (left), precipitation (centre) and humidity index (right) as predictors. Numbers at the top of the maps represent the numbers of reactive catchments out of 163 for each model.

Figure 9 compares how robustness is linked to catchment climate characteristics. The figure shows that there is a large difference in the catchment climatic characteristics between HYPE and GR6J. GR6J reacts to the RAT for humid catchments with a higher number of rainy days, less aridity and lower potential evaporation. It is interesting to note that GR6J responds differently for Sweden than for the rest of the dataset, probably because of the specificity of Swedish hydrology (e.g. the influence of snow). Northern catchments seem to cause more robustness issues for GR6J. In these catchments, streamflow is

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

regulated by hydroelectric power stations. Since regulation is not explicitly represented in the GR6J model, it is possible that this aspect of the catchment hydrology may lead to flaws in the models. Snow that strongly influences hydrology in northern Sweden may also be a reason for the issues in GR6J. The snow module that adds two parameters to be calibrated may then create robustness issues (even if this is not the case over the whole dataset).

In the case of S-HYPE, when temperature is used as a predictor, reactive catchments seem to have less snow fraction than average and more potential evaporation. This is possibly due to the fact that latitude is not taken into account in the evaporation calculation in the HYPE model (Oudin formula is not used in the model). This may lead to robustness issues in the catchments where evaporation has an impact. However, the significance is relatively weak, and no clear difference exists between reactive and non-reactive catchments: model robustness cannot really be linked to catchment characteristics.

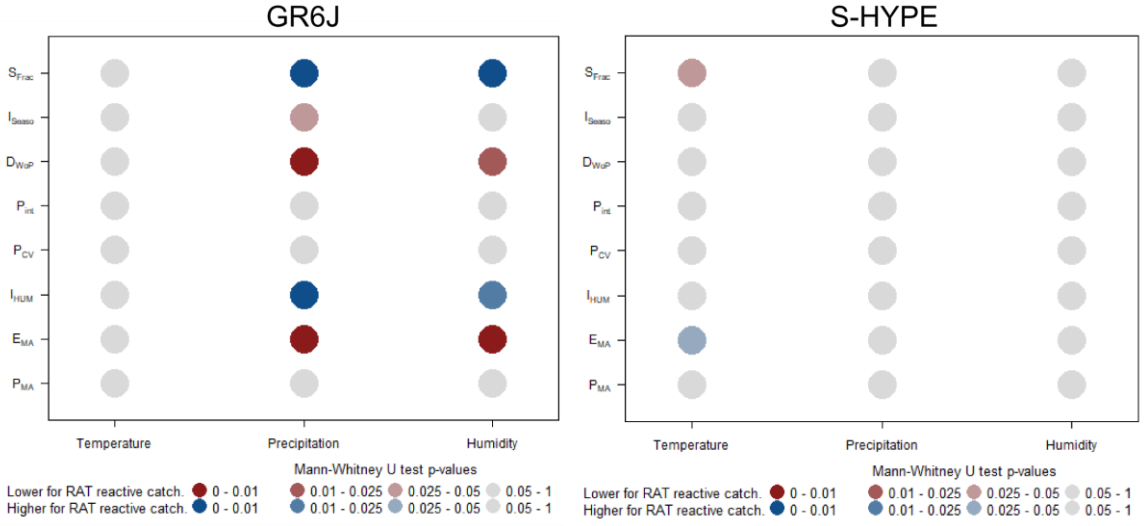


Figure 9: Results of the Mann–Whitney U test to evaluate the difference in climatic signatures in Sweden. The left plot represents differences between Swedish catchments on which GR6J reacts and catchments on which it does not and the right plot represents differences between catchments on which S-HYPE reacts to the RAT and catchments on which it does not. The number of catchments in each subset can be found in Figure 8. Blue (red) squares mean that the signature is significantly higher (lower) for reactive catchments. P_{MA}: mean annual precipitation, E_{MA}: mean annual evaporation, I_{HUM}: humidity index, P_{CV}: precipitation variability, P_{int}: precipitation intensity index, D_{WoP}: days without precipitation ratio, I_{Seaso}: seasonality index, S_{Frac}: snow fraction.

Similarly to Figure 9, Figure 10 compares how the RAT results are linked to flow signatures for GR6J and S-HYPE. It also shows large differences in behaviour between the two models. When precipitation and humidity index are taken as predictors, GR6J reacts for wet catchments with high flow and high runoff ratio. This confirms the results from Figure 9. It seems that GR6J reacts to RAT on large river catchments where the flow is higher than the average. In these catchments, streamflows are more often regulated by human activities and, since there is no regulation module in GR6J (unlike in HYPE), it can create robustness issues in the model.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

For S-HYPE, again, no clear difference exists between reactive and non-reactive catchments. This result suggests that the HYPE model has robustness issues on random catchments (at least regarding the signatures evaluated here). One possible hypothesis to explain this can be that it is calibrated manually, often using super parameters, and this may lead to different robustness issues for different locations. The choice of objective function (namely the NSE) and the focus on flood forecasting also led the modeller to place more focus on timing than on water balance, which can explain why the bias error is less significant for S-HYPE. The calibration procedure may lead to additional issues in terms of robustness that do not depend on catchment location or regime.

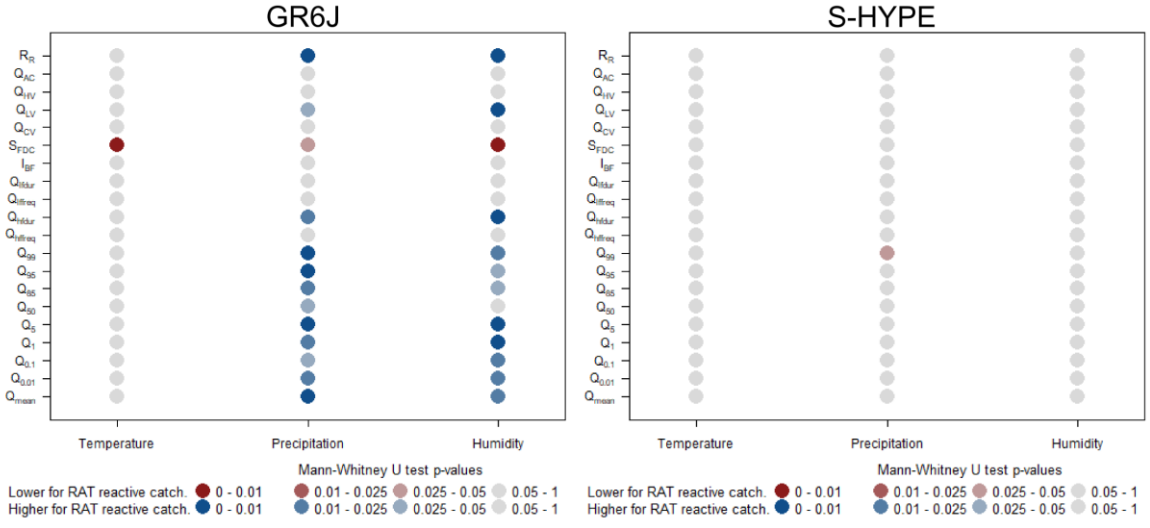


Figure 10: Results of the Mann–Whitney U test to evaluate the difference in flow signatures in Sweden. The left plot represents differences between Swedish catchments on which GR6J reacts and catchments on which it does not and the right plot represents differences between catchments on which S-HYPE reacts to the RAT and catchments on which it does not. The number of catchments in each subset can be found in Figure 8. Blue (red) squares mean that the signature is significantly higher (lower) for reactive catchments. Q_{mean}: annual mean flow, $Q_{[0.01-99]}$: flow percentiles, $Q_{[hf-lf]freq}$: frequency of [high-low]flow events, $Q_{[hf-lf]dur}$: duration of high-low flow events, I_{BF}: baseflow index, S_{FDC}: slope of the flow duration curve, $Q_{[C-L-H]V}$: [total-low-high]flow variability, Q_{AC}: flow 1 day autocorrelation, R_R: runoff ratio.

In summary, S-HYPE and GR6J have equivalent numbers of reactive catchments. Also, GR6J seems to behave differently in Sweden compared to France and Denmark, perhaps due to river regulations and higher snow fractions. It is very difficult to understand the issues found for S-HYPE since the reactive catchments do not differ significantly from the non-reactive ones. This can be due to the different calibration treatment of the model that was calibrated manually and for a flood forecasting purpose.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

4.2 MIKE SHE vs GR6J in Denmark

Figure 11 presents the catchments on which GR6J and MIKE SHE react to the RAT in Denmark. Overall, reactive catchments are mainly located in the Jutland peninsula (which corresponds to continental Denmark). In Denmark, unlike in Sweden and France, there are more reactive catchments when precipitation and humidity index are used as predictors than when temperature is the predictor. However, as for the rest of the dataset, reactive catchments are almost the same when precipitation and humidity index are the predictors. The fact that MIKE SHE reacts to the RAT on fewer catchments than GR6J (13 vs 22, respectively) shows that MIKE SHE is more robust than GR6J in Denmark. Despites this, there are several common reactive catchments between MIKE SHE and GR6J: 57% of the catchments on which MIKE SHE reacts were also reactive with GR6J. This shows that GR6J and MIKE SHE have common causes that may explain their lack of robustness.

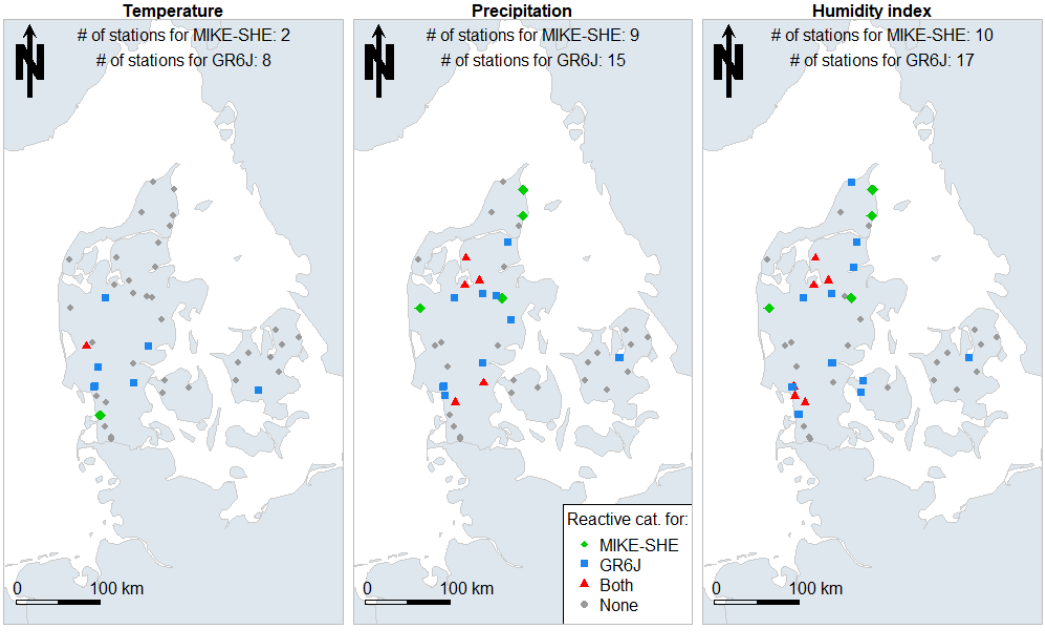


Figure 11: Location and number of Danish catchments on which MIKE SHE and GR6J react to the RAT using temperature (left), precipitation (centre) and humidity index (right) as predictors. Numbers at the top of the maps represent the number of reactive catchments out of 43 for each model.

To confirm the relationship between the robustness of MIKE SHE and GR6J, Figure 12 shows the differences in climatic characteristics between reactive and non-reactive catchments. Here, as for Sweden, the profile of catchments on which GR6J reacts to RAT differs from the rest of the dataset. Reactive catchments are more humid with more regular precipitation and a lower seasonality index.

In the case of MIKE SHE, very few differences seem to exist between the catchments on which the model reacts and the catchments on which it does not (Figure 12). If temperature is the predictor, the catchments on which the model reacts are

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Hydrology and Earth System Sciences
Discussions

less snowy than the average. If precipitation or humidity index is the predictor, the reactive catchments are characterized by less potential evaporation and fewer days without rainfall (for humidity index as the predictor).

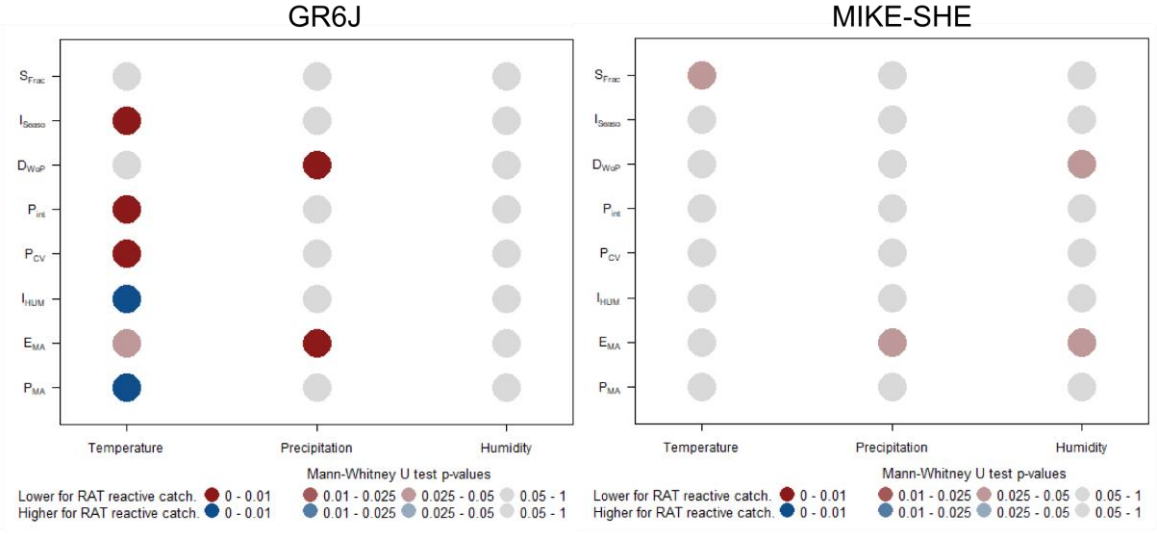


Figure 12: Results of the Mann–Whitney U test to evaluate the difference in climatic signatures in Denmark. The left plot represents differences between Danish catchments on which GR6J reacts and catchments on which it does not and the right plot represents differences between catchments on which MIKE SHE reacts to the RAT and catchments on which it does not. The number of catchments in each subset can be found in Figure 11. Blue (red) squares mean that the signature is significantly higher (lower) for reactive catchments. P_{MA}: mean annual precipitation, E_{MA}: mean annual evaporation, I_{HUM}: humidity index, P_{CV}: precipitation variability, P_{int}: precipitation intensity index, D_{WoP}: days without precipitation ratio, I_{Seaso}: seasonality index, S_{Frac}: snow fraction.

Regarding flow signatures (Figure 13), GR6J shows similar results in Denmark than in the entire catchment set (Figure 7). Reactive catchments are characterized by high baseflow and slow response (precipitation and humidity index as predictors). The reason may be the same as for the whole dataset (Sect. 3.2).

MIKE SHE shows some similarities with GR6J regarding the characteristics of the catchments on which it reacts to the RAT when humidity index is taken as a predictor. The reactive catchments for MIKE SHE have higher baseflow and slower response than the average, similarly to GR6J. Surprisingly, this is not the case when precipitation is taken as a predictor, even if the reactive catchments are almost the same.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

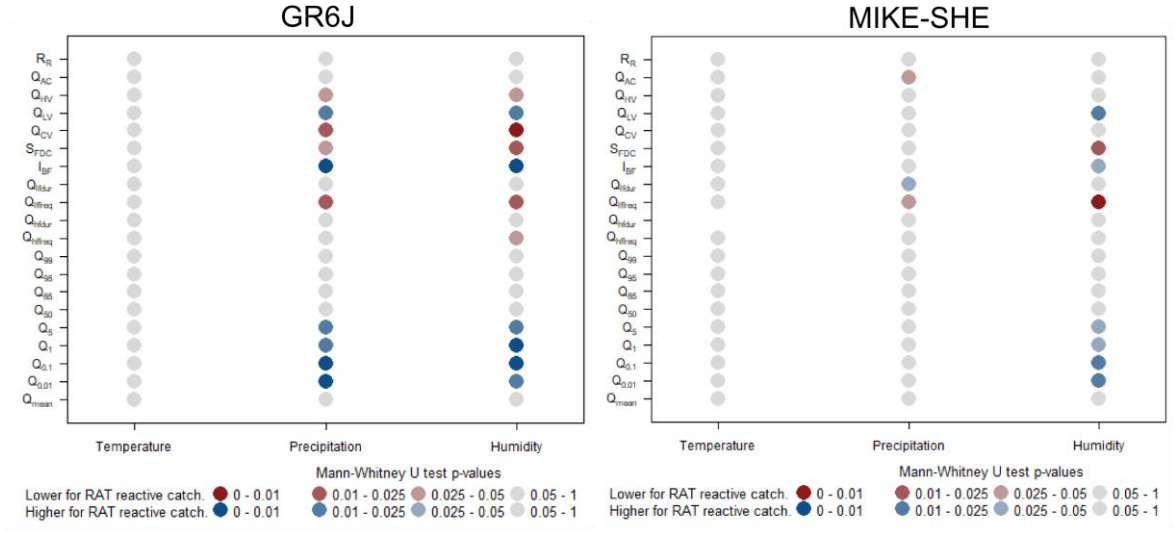


Figure 13: Results of the Mann–Whitney U test to evaluate the difference in flow signatures in Denmark. The left plot represents differences between Danish catchments on which GR6J reacts and catchments on which it does not and the right plot represents differences between catchments on which MIKE SHE reacts to the RAT and catchments on which it does not. The number of catchments in each subset can be found in Figure 11. Blue (red) squares mean that the signature is significantly higher (lower) for reactive catchments. Q_{mean}: annual mean flow, $Q_{[0.01-99]}$: flow percentiles, $Q_{[hf-lf]freq}$: frequency of [high-low]flow events, $Q_{[hf-lf]dur}$: duration of [high-low] flow events, I_{BF}: baseflow index, S_{FDC}: slope of the flow duration curve, $Q_{[C-L-H]V}$: [total-low-high]flow variability, Q_{AC}: flow 1 day autocorrelation, R_R: runoff ratio.

To summarize, the GR6J model shows robustness issues for the same type of catchment in Denmark as for the whole dataset. Comparing the models, fewer catchments react for MIKE SHE than for GR6J, even if some similarities exist between the catchments that react for the two models. It is, however, difficult to characterize these catchments for the MIKE SHE model due to their low number.

4.3 Summary and discussion on the model comparison

The RAT was used to compare the robustness of GR6J and S-HYPE in Sweden and of GR6J and MIKE SHE in Denmark. Overall, the number of RAT-reactive catchments (Table 4) can be seen as a rough indicator of model robustness. The results show that GR6J is slightly more robust than S-HYPE in Sweden and MIKE SHE is slightly more robust than GR6J in Denmark. However, these numbers should not be the only indicator of model robustness since their use does not facilitate our understanding of the robustness issues.

Table 4: Number of reactive catchments for each country and model and proportion in terms of the total number of catchments (N = 352: 163, 43 and 146 for Sweden, Denmark and France, respectively).

	Model

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

Country	Predictor	GR6J	S-HYPE	MIKE SHE
Sweden	Temperature	21 (13%)	35 (21%)	-
	Precipitation	13 (8%)	18 (11%)	
	Humidity index	19 (12%)	19 (12%)	
Denmark	Temperature	8 (19%)	-	2 (5%)
	Precipitation	15 (35%)		9 (21%)
	Humidity index	17 (40%)		10 (23%)
France	Temperature	70 (48%)	-	-
	Precipitation	26 (18%)		
	Humidity index	28 (19%)		
All three	Temperature	99 (28%)	-	-
	Precipitation	56 (16%)		
	Humidity index	64 (18%)		

To improve this understanding, characterization of the reactive catchments shows that MIKE SHE and GR6J both react to the RAT on catchments with high baseflow, which indicates that both models have difficulties in representing long-term groundwater evolution. This seems to be a critical issue for model robustness (and thus a possible priority topic for model improvement). The characterization also shows that GR6J and S-HYPE robustness is sensitive to potential evaporation. The calculation of potential evaporation for the models may also lead to robustness issues (this was also shown by e.g. Birhanu et al., 2018). To confirm this, we tested GR6J in the French catchments using the Penman–Monteith evaporation formula (that is less dependent on temperature). This test showed that, even if the number of reactive catchments decreases when temperature is the indicator, the number of reactive catchments increases when both precipitation and humidity index are the indicators, which shows that the choice of formula is not straightforward. The last observation from this catchment characterization is that GR6J seems to have robustness issues on catchments in which streamflow is regulated by dams. This is probably due to the fact that the model is not built to take this into account and that calibration may have led to distorted values of parameters.

The choice made in this paper was essentially to try to explain the model robustness flaws on the basis of issues in the model structure (e.g. the water balance function of GR6J). However, the model comparison cannot be fully understood without

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

taking into account the difference in the calibration process. In particular, Figure 10 shows that catchments on which S-HYPE presents robustness issues are difficult to characterize. The manual calibration with local tuning may provide a potential explanation for this. In addition, it is important to note that S-HYPE was calibrated only on a sub-period (between 1999 and 2008), which may consequently affect the robustness of the model, compared to GR6J that was calibrated over the whole period. The objective function is also an important factor for explaining the results of the RAT. Indeed, GR6J and MIKE SHE were calibrated by taking into account the water balance bias (within the KGE for GR6J and as one of the objective functions for MIKE SHE). S-HYPE was calibrated only in regard to the NSE with a focus on flood forecasting, which does not include an explicit water balance component. Because of the way RAT is designed (using the water balance bias as a metric), this has probably also affected the results of the S-HYPE model. Consequently, although it is most likely not the only factor, calibration choices may explain why S-HYPE appears slightly less robust than GR6J and why the reactive catchments are so difficult to characterize.

4.4 While all that glitters is not gold, all that is dull is not worthless

The meaning of a reaction to the RAT needs to be discussed. By itself, it only indicates that the annual model bias is correlated with a given climate indicator. Although it is a bad omen regarding the capacity of extrapolation of the model, its interpretation is not straightforward: it is a “yes or no” test that requires interpretation. The slope of the relationship between bias and indicator may also be interesting to examine, since a low slope is certainly not as problematic as a high one.

If a model reacts to the RAT, it could also be for “good” reasons, i.e. because of a time-dependent bias in the forcing data or because of a drift in the measured streamflow. Even a robust model will be affected by a trend in input data, yielding the impression that the hydrological model lacks robustness. Such an erroneous conclusion could also be due to widespread changes in land use, construction of an unaccounted storage reservoir or the evolution of water uses.

If a model does not react to the RAT, it does not mean that it has no robustness issue at all; indeed the RAT is designed to only give an initial diagnosis about model health. However, the large-sample analysis carried out in this paper gave an overall idea of the robustness of the models by using a large dataset. It allowed us to find patterns in the model robustness issues that served as a diagnosis to improve these issues in the future without having to deploy a complex experimental set-up.

In the same vein, it is interesting to evaluate how much the results of the RAT are influenced by the performance of the models. Indeed, the performance can have two possible effects: if it is too low the model may react to the RAT because it does not represent correctly the hydrological processes in the catchments, but if the performances are very high it can be that the model is over-adapted to the calibration period and will react to RAT. However, if the model does not show a high performance over the observed period, it is likely that the performance under future climate will remain low leading to high uncertainties in flow projections. It is thus important to add a performance check to the RAT. For example, Gelfan et al. (2020) proposed such a method in which the model is not seen as robust if it remains under a certain performance threshold.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

In the case of our study, the three models have a good performance in the majority of the catchments (the KGE value is greater than 0.7 on almost all the catchments).

Although we are confident that the RAT is useful, it is not a universal panacea for hydrological models.

5 Conclusion

5.1 Synthesis

This paper presented a large-sample analysis of the robustness of three models to a changing climate. The RAT allowed us to compare these three different models without controlling their calibration process. The analysis of the hydrological signatures of the catchments that react to the RAT allowed us to detect some critical points to start working on in order to improve the robustness of the models.

The GR6J model is the only model that was set up and calibrated over the whole dataset. Overall, it reacts to the RAT on 148 catchments (with all predictors together), which represent 42% of the catchments in the dataset. The analysis of these catchments shows that potential evaporation accounting and water balance calculation are possibly not robust when air temperature changes over time. It can also indicate that the groundwater changes and the regulation due to human activities are difficult to handle by the model when the amount of precipitation and the humidity index change over time. Consequently, this work gives potential clues to detect the model components that should receive priority attention in order to make it more robust in the context of climate change.

The MIKE SHE model reacts to the RAT on 13 catchments (with all predictors together), which represent 38% of the Danish catchments tested. The analysis of the catchment signatures only shows that MIKE SHE has robustness issues in humid catchments with high baseflow when the humidity index changes over time. This can indicate that the model has robustness issues when the groundwater regime changes over time.

The model reacts to RAT on 55 catchments (with all predictors together), which represent 33% of the tested catchments in Sweden. The analysis shows a potential issue in evaporation calculation when temperature changes over time. However, the differences between the flow signatures of the catchments that react to the RAT, and those of the catchments that do not, are low, which did not enable detection of other issues in HYPE. We assume that it is due to the calibration process that was manual and spatialized by HRUs.

The comparison between the models in Denmark shows that MIKE SHE appears slightly more robust than GR6J but overall reacts on the same type of catchments. The two models seem to have difficulties in representing groundwater variations even if it is probably not for the same reasons. In Sweden, S-HYPE seems slightly less robust than GR6J and reacts on a different kind of catchment. These differences can be explained by differences in the calibration processes. Parameters in S-HYPE are regionalized over the territory while the GR6J parameters are specific to every catchment. The calibration periods and the objective functions are different for the two models and can explain the differences in terms of robustness.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

595 **5.2 Perspectives**

Our analysis pointed out flaws in the models in terms of robustness to changing climate.

First, the climatic and flow signatures used in the paper do not seem to be sufficient to explain the robustness issues of the models (especially in the case of S-HYPE). In Sweden and Denmark, more snow signatures may help to refine the analysis regarding snow processes and to better understand potential issues in the model snow modules. S-HYPE may also be more
600 sensitive to land use or soil cover since the model parameters are regionalized by HRUs (soil and land use combination). This analysis would be useful for pointing out any region or parameter on which robustness issues exist. The evolution of land use in time may also be interesting to examine, since it is also an indicator of changing climate and can induce some errors in models that are parameterized by HRUs like S-HYPE.

The analysis also highlighted some issues that are due to potential evaporation calculation. It would thus be interesting to test
605 several formulas for the calculation of potential evaporation so as to check whether it is possible to optimize model robustness. Birhanu et al. (2018) tested the robustness of different formulas and concluded that the simplest of them do not necessarily decrease the robustness. However, these conclusions were made using an SST and it may be interesting to test them using the RAT. We ran such a test using the Penman–Monteith equation and GR6J. The test yielded conflating results, which are difficult to interpret (fewer reactive catchments when temperature is the indicator but more catchments when
610 precipitation is the indicator).

More systematic tests are needed to better understand the influence of the calibration set-up. The RAT could, for example, be used to evaluate the effect of objective functions by using several types of criteria and flow transformations. It could also be interesting to test the influence of the period used for calibration and how period selection can be optimized to better satisfy the RAT. In the same vein, most systematic evaluations can be made in combination with progressive changes in model
615 structure to test the robustness issues attributed to model structure and optimize model robustness.

Acknowledgements

This work was funded by the project AQUACLEW, which is part of ERA4CS, an ERA-NET initiated by JPI Climate, and funded by FORMAS (SE), DLR (DE), BMWFW (AT), IFD (DK), MINECO (ES), ANR (FR) with co-funding by the European Commission [Grant 690462].

620 The gridded SAFRAN climate reanalysis data can be ordered from Météo-France. Observed flow data are available from the French HYDRO database (http://www.hydro.eaufrance.fr/).

The GR models, including GR6J, are available from the airGR R package.

Competing interests

The contact author declared that none of the authors has any competing interests.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024


Hydrology and Earth System Sciences
Discussions

625

References

Birhanu, D., Kim, H., Jang, C., and Park, S.: Does the Complexity of Evapotranspiration and Hydrological Models Enhance Robustness?, Sustainability, 10, 2837, https://doi.org/10.3390/su10082837, 2018.

630 Blöschl, G., Bierkens, M. F. P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J. W., McDonnell, J. J., Savenije, H. H. G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S. T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S. K., Baker, V., Bardsley, E., Barendrecht, M. H., Bartosova, A., Batelaan, O., Berghuijs, W. R., Beven, K., Blume, T., Bogaard, T., Borges de Amorim, P., Böttcher, M. E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Y., Chen, Y.,

635 Chifflard, P., Claps, P., Clark, M. P., Collins, A. L., Croke, B., Dathe, A., David, P. C., de Barros, F. P. J., de Rooij, G., Di Baldassarre, G., Driscoll, J. M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W. H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Gonzalez Bevacqua, A., González-Dugo, M. P., Grimaldi, S., Gupta, A. B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H.,

640 Hohmann, C., Holko, L., Hopkinson, C., Hrachowitz, M., Illangasekare, T. H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., et al.: Twenty-three unsolved problems in hydrology (UPH) – a community perspective, Hydrological Sciences Journal, 64, 1141–1158, https://doi.org/10.1080/02626667.2019.1620507, 2019.

Brigode, P., Oudin, L., and Perrin, C.: Hydrological model parameter instability: A source of additional uncertainty in estimating the hydrological impacts of climate change?, Journal of Hydrology, 476, 410–425,

645 https://doi.org/10.1016/j.jhydrol.2012.11.012, 2013.

Broderick, C., Matthews, T., Wilby, R. L., Bastola, S., and Murphy, C.: Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods, Water Resources Research, 52, 8343–8373, https://doi.org/10.1002/2016WR018850, 2016.

Coron, L., Andréassian, V., Bourqui, M., Perrin, C., and Hendrickx, F.: Pathologies of hydrological models used in changing

650 climatic conditions: A review, IAHS-AISH Publication, 344, 39–44, 2011.

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., and Hendrickx, F.: Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments, Water Resources Research, 48, W05552, https://doi.org/10.1029/2011WR011721, 2012.

Coron, L., Andréassian, V., Perrin, C., Bourqui, M., and Hendrickx, F.: On the lack of robustness of hydrologic models

655 regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments, Hydrol. Earth Syst. Sci., 18, 727–746, https://doi.org/10.5194/hess-18-727-2014, 2014.

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

Coron, L., Delaigue, O., Thirel, G., Dorchies, D., Perrin, C., and Michel, C.: airGR: Suite of GR Hydrological Models for Precipitation-Runoff Modelling., 2021.

Dakhlaoui, H., Ruelland, D., Tramblay, Y., and Bargaoui, Z.: Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia, Journal of Hydrology, 550, 201–217, https://doi.org/10.1016/j.jhydrol.2017.04.032, 2017.

Dakhlaoui, H., Ruelland, D., and Tramblay, Y.: A bootstrap-based differential split-sample test to assess the transferability of conceptual rainfall-runoff models under past and future climate variability, Journal of Hydrology, 575, 470–486, https://doi.org/10.1016/j.jhydrol.2019.05.056, 2019.

Delaigue, O., Génot, B., Mendoza Guimarães, G., Lebecherel, L., Brigode, P., and Bourgin, P. Y.: Database of watershed-scale hydroclimatic observations in France, 2022.

Donnelly-Makowecki, L. M. and Moore, R. D.: Hierarchical testing of three rainfall–runoff models in small forested catchments, Journal of Hydrology, 219, 136–152, https://doi.org/10.1016/S0022-1694(99)00056-6, 1999.

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., and Peterson, T. J.: Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, Water Resources Research, 52, 1820–1846, https://doi.org/10.1002/2015WR018068, 2016.

Gelfan, A., Kalugin, A., Krylenko, I., Nasonova, O., Gusev, Y., and Kovalev, E.: Does a successful comprehensive evaluation increase confidence in a hydrological model intended for climate impact assessment?, Climatic Change, 163, 1165–1185, https://doi.org/10.1007/s10584-020-02930-z, 2020.

Gelfan, A. N. and Millionshchikova, T. D.: Validation of a Hydrological Model Intended for Impact Study: Problem Statement and Solution Example for Selenga River Basin, Water Resources, 45, 90–101, https://doi.org/10.1134/S0097807818050354, 2018.

Gharari, S., Hrachowitz, M., Fenicia, F., and Savenije, H. H. G.: An approach to identify time consistent model parameters: sub-period calibration, Hydrol. Earth Syst. Sci., 17, 149–161, https://doi.org/10.5194/hess-17-149-2013, 2013.

Graham, D. and Butts, M.: Flexible, integrated watershed modelling with MIKE SHE, in: Watershed Models, 245–272, 2005.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Henriksen, H. J., Jakobsen, A., Pasten-Zapata, E., Troldborg, L., and Sonnenborg, T. O.: Assessing the impacts of climate change on hydrological regimes and fish EQR in two Danish catchments, Journal of Hydrology: Regional Studies, 34, 100798, https://doi.org/10.1016/j.ejrh.2021.100798, 2021.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

690 Højberg, A. L., Troldborg, L., Stisen, S., Christensen, B. B. S., and Henriksen, H. J.: Stakeholder driven update and improvement of a national water resources model, Environmental Modelling & Software, 40, 202–213, https://doi.org/10.1016/j.envsoft.2012.09.010, 2013.

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

695 Johansson, B.: Estimation of areal precipitation for hydrological modelling in Sweden, Göteborg : Göteborg university, 2002.

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

Lan, T., Lin, K., Xu, C.-Y., Tan, X., and Chen, X.: Dynamics of hydrological-model parameters: mechanisms, problems and
700 solutions, Hydrol. Earth Syst. Sci., 24, 1347–1366, https://doi.org/10.5194/hess-24-1347-2020, 2020.

de Lavenne, A. and Andréassian, V.: Impact of climate seasonality on catchment yield: A parameterization for commonly-used water balance formulas, Journal of Hydrology, 558, 266–274, https://doi.org/10.1016/j.jhydrol.2018.01.009, 2018.

de Lavenne, A., Andréassian, V., Crochemore, L., Lindström, G., and Arheimer, B.: Quantifying multi-year hydrological memory with Catchment Forgetting Curves, Hydrol. Earth Syst. Sci., 26, 2715–2732, https://doi.org/10.5194/hess-26-2715-
705 2022, 2022.

Leleu, I., Tonnelier, I., Puechberty, R., Gouin, P., Viquendi, I., Cobos, L., Foray, A., Baillon, M., and Ndima, P.-O.: La refonte du système d'information national pour la gestion et la mise à disposition des données hydrométriques, La Houille Blanche, 100, 25–32, https://doi.org/10.1051/lhb/2014004, 2014.

Lindström, G.: Lake water levels for calibration of the S-HYPE model, Hydrology Research, 47, 672–682,
710 https://doi.org/10.2166/nh.2016.019, 2016.

Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., and Arheimer, B.: Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales, Hydrology Research, 41, 295–319, https://doi.org/10.2166/nh.2010.007, 2010.

Mann, H. B. and Whitney, D. R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the
715 Other, The Annals of Mathematical Statistics, 18, 50–60, https://doi.org/10.1214/aoms/1177730491, 1947.

Montanari, A., Young, G., Savenije, H. H. G., Hughes, D., Wagener, T., Ren, L. L., Koutsoyiannis, D., Cudennec, C., Toth, E., Grimaldi, S., Blöschl, G., Sivapalan, M., Beven, K., Gupta, H., Hipsey, M., Schaefli, B., Arheimer, B., Boegh, E., Schymanski, S. J., Di Baldassarre, G., Yu, B., Hubert, P., Huang, Y., Schumann, A., Post, D. A., Srinivasan, V., Harman, C., Thompson, S., Rogger, M., Viglione, A., McMillan, H., Characklis, G., Pang, Z., and Belyaev, V.: "Panta Rhei—Everything
720 Flows": Change in hydrology and society—The IAHS Scientific Decade 2013–2022, Hydrological Sciences Journal, 58, 1256–1275, https://doi.org/10.1080/02626667.2013.809088, 2013.

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

Nicolle, P., Andréassian, V., Royer-Gaspard, P., Perrin, C., Thirel, G., Coron, L., and Santos, L.: Technical Note – RAT: a Robustness Assessment Test for calibrated and uncalibrated hydrological models, Hydrol. Earth Syst. Sci. Discuss., 2021, 1–22, https://doi.org/10.5194/hess-2021-147, 2021.

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., and Loumagne, C.: Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling, Journal of Hydrology, 303, 290–306, https://doi.org/10.1016/j.jhydrol.2004.08.026, 2005.

Pachauri, R. K., Allen, M. R., Barros, V. R., Broome, J., Cramer, W., Christ, R., Church, J. A., Clarke, L., Dahe, Q., and Dasgupta, P.: Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change, L.A. Meyer (eds.)., Ipcc, 151 pp., 2014.

Pelletier, A. and Andréassian, V.: Hydrograph separation: an impartial parametrisation for an imperfect method, Hydrol. Earth Syst. Sci., 24, 1171–1187, https://doi.org/10.5194/hess-24-1171-2020, 2020.

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

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., and Andréassian, V.: A downward structural sensitivity analysis of hydrological models to improve low-flow simulation, Journal of Hydrology, 411, 66–76, https://doi.org/10.1016/j.jhydrol.2011.09.034, 2011.

Rau, P., Bourrel, L., Labat, D., Ruelland, D., Frappart, F., Lavado, W., Dewitte, B., and Felipe, O.: Assessing multidecadal runoff (1970–2010) using regional hydrological modelling under data and water scarcity conditions in Peruvian Pacific catchments, Hydrological Processes, 33, 20–35, https://doi.org/10.1002/hyp.13318, 2019.

Refsgaard, J. C. and Knudsen, J.: Operational Validation and Intercomparison of Different Types of Hydrological Models, Water Resources Research, 32, 2189–2202, https://doi.org/10.1029/96WR00896, 1996.

Refsgaard, J. C., Madsen, H., Andréassian, V., Arnbjerg-Nielsen, K., Davidson, T. A., Drews, M., Hamilton, D. P., Jeppesen, E., Kjellström, E., Olesen, J. E., Sonnenborg, T. O., Trolle, D., Willems, P., and Christensen, J. H.: A framework for testing the ability of models to project climate change and its impacts, Climatic Change, 122, 271–282, https://doi.org/10.1007/s10584-013-0990-2, 2014.

van Roosmalen, L., Christensen, B. S. B., and Sonnenborg, T. O.: Regional Differences in Climate Change Impacts on Groundwater and Stream Discharge in Denmark, Vadose Zone Journal, 6, 554–571, https://doi.org/10.2136/vzj2006.0093, 2007.

Scharling, M.: Climate Grid Denmark: Precipitation, air temperature and potential evapotranspiration 20×20 and 40×40 km., Danish Meteorological Institute, 1999.

Scharling, M. and Kern-Hansen, C.: Climate Grid Denmark, Dataset for use in research and education, Daily and monthly values 1989-2010, 10x10 km precipitation sum, 20x20 km average temperature, accumulated potential evaporation (Makkink), average wind speed, accumulated global radiation., Danish Meteorological Institute, 2012.

Seibert, J.: Reliability of Model Predictions Outside Calibration Conditions: Paper presented at the Nordic Hydrological Conference (Røros, Norway 4-7 August 2002), Hydrology Research, 34, 477–492, https://doi.org/10.2166/nh.2003.0019, 2003.

Sleziak, P., Szolgay, J., Hlavčová, K., Duethmann, D., Parajka, J., and Danko, M.: Factors controlling alterations in the performance of a runoff model in changing climate conditions, Journal of Hydrology and Hydromechanics, 66, 381–392, https://doi.org/10.2478/johh-2018-0031, 2018.

Stephens, C. M., Marshall, L. A., and Johnson, F. M.: Investigating strategies to improve hydrologic model performance in a changing climate, Journal of Hydrology, 579, 124219, https://doi.org/10.1016/j.jhydrol.2019.124219, 2019.

Stisen, S., Sonnenborg, T. O., Højberg, A. L., Troldborg, L., and Refsgaard, J. C.: Evaluation of Climate Input Biases and Water Balance Issues Using a Coupled Surface–Subsurface Model, Vadose Zone Journal, 10, 37–53, https://doi.org/10.2136/vzj2010.0001, 2011.

Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C., and Lindström, G.: Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale, Hydrological Sciences Journal, 57, 229–247, https://doi.org/10.1080/02626667.2011.637497, 2012.

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

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.

Vaze, J., Post, D. A., Chiew, F. H. S., Perraud, J.-M., Viney, N. R., and Teng, J.: Climate non-stationarity – Validity of calibrated rainfall–runoff models for use in climate change studies, Journal of Hydrology, 394, 447–457, https://doi.org/10.1016/j.jhydrol.2010.09.018, 2010.

Vidal, J.-P., Martin, E., Franchistéguy, L., Baillon, M., and Soubeyroux, J.-M.: A 50-year high-resolution atmospheric reanalysis over France with the Safran system, International Journal of Climatology, 30, 1627–1644, https://doi.org/10.1002/joc.2003, 2010.

Westerberg, I. K. and McMillan, H. K.: Uncertainty in hydrological signatures, Hydrol. Earth Syst. Sci., 19, 3951–3968, https://doi.org/10.5194/hess-19-3951-2015, 2015.

Westra, S., Thyer, M., Leonard, M., Kavetski, D., and Lambert, M.: A strategy for diagnosing and interpreting hydrological model nonstationarity, Water Resources Research, 50, 5090–5113, https://doi.org/10.1002/2013WR014719, 2014.

Wilcoxon, F.: Individual Comparisons by Ranking Methods, Biometrics Bulletin, 1, 80–83, https://doi.org/10.2307/3001968, 1945.

https://doi.org/10.5194/hess-2024-80
Preprint. Discussion started: 21 May 2024

Xu, C.: Operational testing of a water balance model for predicting climate change impacts, Agricultural and Forest Meteorology, 98–99, 295–304, https://doi.org/10.1016/S0168-1923(99)00106-9, 1999.

Zeng, L., Xiong, L., Liu, D., Chen, J., and Kim, J.-S.: Improving Parameter Transferability of GR4J Model under Changing Environments Considering Nonstationarity, Water, 11, 2029, https://doi.org/10.3390/w11102029, 2019.

795