

Research papers

Application-based evaluation of multi-basin hydrological models

Yiheng Du^{a,*}, Jonas Olsson^{a,e}, Kristina Isberg^a, Johan Strömqvist^a, Yeshewatesfa Hundecha^a,
Benedito Cláudio da Silva^b, Sameh Adib Abou Rafee^b, Carlos Ruberto Fragoso Jr^c,
Stein Beldring^d, Anna Hansen^d, Cintia Bertacchi Uvo^{e,f}, Johanna Sörensen^e

^a Hydrology Research Unit, Swedish Meteorological and Hydrological Institute, Sweden

^b Institute of Natural Resources, Federal University of Itajubá, Brazil

^c Center for Technology, Federal University of Alagoas, Brazil

^d Hydrology Department, Norwegian Water Resources and Energy Directorate, Norway

^e Department of Building & Environmental Technology, Lund University, Sweden

^f Finnish Environment Institute, Finland

ARTICLE INFO

Keywords:

Application-based metrics

Model evaluation

Quantile extremes

Relative change

ABSTRACT

Hydrological models are generally calibrated and validated using a suite of well-known statistical metrics, which sometimes lack clear connection and tailoring to the local users' need and therefore limits the evaluation, especially in the case of global climate services. Therefore, in this study, two types of application-based evaluation metrics are introduced, addressing (i) temporal matching of quantile extremes and (ii) relative bias in flow signatures, which supplements commonly used model performance assessment metrics. The introduced metrics are compared to conventional statistical metrics, at seven case study areas across the world, with three model settings representing different datasets and calibrations, generated from the global hydrological model World-Wide HYPE (WW HYPE). The results suggest that different performance results can occur when comparing application-based metrics to conventional ones. This implies that different evaluation metrics reveal models' capability in various aspects, supporting their application under the corresponding circumstances. Finally, these metrics enable us to propose two model applicability scenarios: generally applicable models and conditionally applicable models. For instance, the WW HYPE with global dataset and local calibration can yield optimal estimates concerning the timing of quantile extremes and temporal variations in flow signatures, despite its sub-optimal performance in conventional evaluation metrics. Therefore, it may be considered as a conditionally applicable model which can be used in regions with limited local datasets, supplying reliable information for decision-makers in formulating strategic plans for water resources management.

1. Introduction

Climate services provide climate information to help individuals and organizations to make climate smart decisions. The consequence of climate change is global and will affect the shared future of mankind, which generates an essential demand for climate services in every corner of the world. However, conducting local modeling requires substantial resources in terms of time, data and budget, especially for water services, which are less attainable in regions with limited development, resources, or social stability. At the same time, the United Nations Framework Convention on Climate Change (UNFCCC) initiative on climate adaptation demands easy access to readily available hydro-climate data and information for assessments (Merks et al., 2020).

Therefore, global scale water services are crucial to guarantee equal access to hydrological information in resource sparse areas, to support local vulnerability, resilience and risk assessment. This raises the demand of evaluating their applicability at local scales.

Evaluation of hydrological models is generally conducted using a suite of well-known conventional statistical metrics (in the following called "conventional metrics"), which are conducted within time and signature domains. Time-domain metrics (Teegavarapu et al., 2022) include various error and performance measures, that fall into general categories such as scale-dependent (e.g. mean absolute error, MAE; Willmott & Matsuura, 2005), scale-independent (e.g. correlation coefficient, CC; Legates & McCabe, 1999), relative-based (e.g. relative error, RE; Törnqvist et al., 1985), and efficiency measures (e.g. Nash-Sutcliffe

* Corresponding author.

E-mail address: yiheng.du@smhi.se (Y. Du).

<https://doi.org/10.1016/j.jhydrol.2024.131727>

Received 20 March 2024; Accepted 13 July 2024

Available online 31 July 2024

0022-1694/© 2024 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Efficiency, NSE; Nash & Sutcliffe, 1970). Unlike these time-domain metrics based on evaluating the entire time series, signature-based metrics evaluate only specific features of the time series, for example, maximum daily flow, base flow index, interannual runoff variability and others (Kavetski et al., 2018; Westerberg et al., 2016; Yilmaz et al., 2008). However, to ensure local applicability of global climate services, corresponding model evaluation should be tailored to the local users' needs (Vincent et al., 2020; Hartmann et al., 2002; Nkiaka et al., 2019), together with efficiently communicating uncertainties and decision-making guidelines (Vaughan & Dessai, 2014; Crochemore et al., 2021). Therefore, besides the concise model evaluation provided by conventional metrics (Ewen, 2011; Pfannerstill et al., 2014), complementary evaluation metrics reflecting local model applications and the ways the services are used in practice need to be included, to ensure model's sufficiency under the corresponding conditions (Krysanova et al., 2018). In the following, we call such metrics "application-based metrics".

Application-based metrics should highlight specific aspects of hydrological processes that the operations in the targeted sectors heavily rely upon. For example, flood (or drought) forecasting is a main task in operational application, where a model is required to predict whether the streamflow will pass certain quantiles sometime in the future. In the European Flood Awareness System (EFAS), a threshold exceedance approach is used, where simulated discharge time series are transformed into dichotomous time series of 1 and 0 by comparing with model-based quantile thresholds (Thielen et al., 2009), allowing for qualitative estimation of event severity. Similarly, the operational system in Sweden uses indicators based on quantile thresholds (Arheimer et al., 2011). For example, a drought advisory threshold is defined as streamflow lower than quantile 0.95 (in the model climatology) for a number of consecutive weeks (SMHI, 2023). In Brazil, the State of Minas Gerais adopts a drought alert system with different levels, defined based on the minimum average flow of 7 days and 10 years of return time (Q7,10) and on the conditions of reservoirs, when they exist in the basin (CERH-MG, 2015). Another example of common climate service applications is climate change impact assessment, where a model is used to estimate the relative difference of some streamflow signature (e.g. average, seasonal or annual minimum streamflow) between a historical and a future time period, both in scientific analysis (Cao et al., 2021; Forzieri et al., 2014; Guo et al., 2017; Lee & Bae, 2016; Mello et al., 2021; Raulino et al., 2021) and operational application (Climate Information, 2020; Norwegian Centre for Climate Services, 2023). By highlighting these aspects of hydrological performance, the evaluation can better serve the needs of the targeted sectors and ensure the models are providing relevant information.

In light of the above common application scenarios, application-based evaluation can include e.g. metrics reflecting a model's ability to predict high and low quantile crossings (i.e. model suitability in a forecast context) or to predict differences in flow signatures between two periods (i.e. model suitability in a climate change context). Intuitively, a model that performs better in terms of conventional metrics can be assumed to perform better also in such practical applications, but this assumption is very seldom evaluated or confirmed (Filianoti et al., 2020). Although limitations of climate services are a reality when projecting an ultimately uncertain future (Donnelly et al., 2018), communications with users can be enhanced by outlining both recommendations and constraints.

Therefore, in this study we complement the conventional metrics with application-based metrics, that are designed to reflect model performance related to the above examples of application. In order to demonstrate, implement, and evaluate the new metrics, a set of model configurations with different, yet reasonably representative performances are needed. These configurations were derived from the GlobalHydroPressure project (GHP), where hydrological models were established in various potential applications. In GHP, seven different case study basins world-wide, covering a wide range of hydrological

pressures, stakeholder needs, socioeconomic conditions and climates, provided a good opportunity of demonstrating the use of application-based metrics as well as assessing the added value of local datasets and local calibration for adapting global hydrological models to different conditions.

In light of the above, the study has a number of connected and complementary objectives:

- to develop a framework for application-based evaluation;
- to apply and demonstrate the framework on three different configurations of a multi-basin hydrological model (World Wide HYPE) in seven case study locations, and assess the added value of local calibration and local forcing data;
- to investigate the connection between conventional metrics and application-based metrics; and
- to characterize the degree of confidence with which different signals from a global hydrological model can be interpreted in practical applications.

2. Application-based model evaluation

As described in the introduction, in the model evaluation we complement a suite of conventional metrics with two types of application-based metrics, based on (1) quantile extremes and (2) flow signatures, that reflect how hydrological models are commonly applied for practical purposes.

Regarding quantile extremes, there have been explorations of using quantiles (i.e. frequency-based thresholds) for model evaluation, e.g. in flash flood warning systems (Carpenter et al., 1999; Norbiato et al., 2009; Younis et al., 2008) and in terms of hydrological extremes (Candogan Yossef et al., 2012). The key underlying assumption is that the hydrological model has skill in ranking events even if the simulated extreme flows are biased relative to the observed data (Reed et al., 2007). Therefore, to evaluate model performance for events of different severity, on one hand we assess the model performance in terms of biases in extremes (described by quantile difference; Section 2.1.1). At the same time, we adopt frequency-based analysis to assess the model performance with a focus on temporal matching of extreme events (described by Peirce Skill Score; Section 2.1.2).

As for flow signatures, they have been employed as rigorous diagnostic-based model evaluation methods, to evaluate the model capability of reproducing functional behaviors of catchments (Donnelly et al., 2016; McMillan et al., 2017; Westerberg et al., 2011). The approach tests the ability of a hydrological model to reproduce the signatures by analyzing similarities and differences between observed and simulated signatures (e.g. long-term averages; Shafii & Tolson, 2015). Moreover, considering that climate conditions are non-stationary under climate change, evaluation performed on climatically contrasted discontinuous periods has been proposed, i.e. a split-sample analysis (Refsgaard et al., 2014; Thirel et al., 2015; Dakhlaoui et al., 2017, 2019), in order to improve evaluation of model performance under a changing climate. This is mostly done by dividing the evaluation period into distinctly different climate conditions (e.g. dry/wet or cold/warm; Refsgaard et al., 2014). Moreover, with the design described in the introduction, relative change is a useful way to exclude bias in amplitude (Forzieri et al., 2014; Guo et al., 2017). Therefore, besides assessing average bias in flow signatures (Section 2.2.1), we implement an extra evaluation metric, assessing models' performance in reproducing relative differences (RD; Section 2.2.2) between two subsets of streamflow conditions (i.e. high/low flow).

The application-based metrics are applied to weekly average runoff as this time step is common in operational application (World Meteorological Organization (WMO), 2021), especially for low flows. The weekly averages are calculated using calendar weeks (i.e. from Monday to Sunday) and for consistency the same weekly averages were adopted for evaluation of both high and low flows.

Four conventional statistical metrics (Maidment, 1993) are selected as benchmarks for the above application-based metrics: mean absolute error (MAE), correlation coefficient (CC), relative error (RE) and Nash-Sutcliffe Efficiency (NSE). MAE is a measure given in absolute terms, which presents errors in the same units as the measurement (m^3/s in our case). RE represents relative bias in the mean value, representing the models' performance on the long-term average annual runoff, while CC represents the temporal variability of streamflow. NSE is used to compare the ratio of the simulation error to an error obtained using a naïve simulation method (the average streamflow), where negative values indicate less skill compared to the average. Median values of each conventional metric were computed for each study area (Section 3.1), whereas absolute RE was used to avoid compensation of negative and positive errors during the calculation of medians.

2.1. Quantile extremes and temporal matching

2.1.1. Quantile differences

To assess the performance of models in terms of extremes, the bias of streamflow at different low (0.80, 0.9 and 0.95) and high (0.2, 0.1, 0.05) quantiles (i.e., exceedance probabilities) were evaluated, by the absolute quantile difference (QD, Fig. 1a) at each station:

$$QD(p) = QM(p) - QO(p) \quad (1)$$

where $QO(p)$ and $QM(p)$ denote the observed and modelled values at quantile probability p , respectively.

2.1.2. Temporal matching of quantile extremes

For evaluating temporal matching of hydrological extremes, a contingency table is used, as it is an effective way for analysis of extreme events that has been widely used for hydrometeorological studies (Aghakouchak & Mehran, 2013; Bartholmes et al., 2009). However, if categorical statistics in contingency tables are defined by amplitude thresholds, and distinct bias exists due to different distributions for each data series, interpretation across thresholds is difficult (Jenkner et al., 2008). To avoid this problem, the contingency table in our study is defined by frequency thresholds, which means the quantiles are computed for each of the model data series and observation data series independently. The quantiles are calculated by taking the weekly averaged streamflow data series for observation (QO) and models (QM) (Fig. 1.b1), rank all weeks during the study period in descending order and estimate the flow exceedance quantile (q_n) for week n as the ratio between the rank of the current week (r_n) and the total number of weeks (N):

$$q_n = r_n/N \quad (2)$$

After sample quantiles are computed for observation data (qO) and model data (qM) independently (Fig. 1.b2), the datasets can be compared using the same relative quantile thresholds within their own distribution. The same quantile thresholds as in section 2.1.1 are used for high flow extremes (0.05, 0.10 and 0.20) and low flow extremes (0.95, 0.90 and 0.80).

The standard contingency table (Table 1; Fig. 1.b3, b4) for dichotomous events (relating to the target threshold) contains the following entries: hits H, misses M, false alarms F and correct negatives Z (Mason, 1989; Gandin and Murphy, 1992). The four numbers may be used to compute several different categorical scores.

In this study, the Peirce Skill Score (PSS; Peirce, 1884) is adopted as it is ideally suited to measure the joint distribution, i.e. to display the model accuracy on its own (Jenkner et al., 2008), and it is calculated as:

$$PSS = \frac{H}{H+M} - \frac{F}{F+Z} \quad (3)$$

Positive PSS values indicate skill compared to a shuffled sample, since $PSS=0$ denotes the skill of a random forecast at all times, and $PSS=1$

indicates a perfect model.

2.2. Flow signatures and split-sample analysis

2.2.1. Flow signatures

Three flow signatures were selected to investigate different parts of the hydrograph: annual mean, maximum and minimum weekly streamflow. As a general evaluation, the model performance in amplitude terms is analyzed by the long-term averages of each flow signature for the entire study period. In a certain case study basin, let $QO_w^{y,b}$ denote the observed weekly flows ($1 \leq w \leq 52$) in year y ($1 \leq y \leq N_y$) and subbasin b ($1 \leq b \leq N_b$).

1 Denote by $SO_{type}^{y,b}$ the observed flow signatures for a certain year and subbasin, with subscripts representing one of three types: mean, maximum (max) or minimum (min) annual flow. The three types of signatures are thus calculated as

$$SO_{mean}^{y,b} = \sum_{w=1}^{52} \frac{QO_w^{y,b}}{w}, SO_{max}^{y,b} = \max(Q_w^{y,b}), SO_{min}^{y,b} = \min(Q_w^{y,b}) \quad (4)$$

2 Long-term averages of the observed signatures are denoted as SO_{type}^b and calculated as

$$SO_{type}^b = \sum_{y=1}^{N_y} \frac{SO_{type}^{y,b}}{y} \quad (5)$$

Steps 1–2 are repeated for the modelled weekly streamflows $QM_w^{y,b}$ resulting in signatures denoted as SM_{type}^b for each subbasin. When referring to SO_{type}^b and SM_{type}^b without need to distinguish between observed and modelled data, S_{type} (i.e. S_{max} , S_{mean} , S_{min}) is used.

2.2.2. Split-sample analysis

Split-sample analysis divides the study period into subsets based on different climate conditions, in our case, the high/low flow years from the observation records. The same high/low flow years will be applied to divide model simulations into subsets, therefore the evaluation metric generated from the subsets quantifies the agreement of higher/lower records and their occurrence years (Fig. 1c).

In a certain case study basin:

1. For each subbasin b , start from the observed flow signatures for each year y , i.e. $SO_{type}^{y,b}$, with types min, mean and max (Equation.4).

2. Sort the flow signatures in ascending order and divide the entire period into higher and lower subsets, separated by the median value of the signature.

3. Calculate averages of the flow signature in the higher (yh) and lower (yl) subsets:

$$SOH_{type}^b = \sum_{y=1}^{N_y/2} SO_{type}^{yh,b}, SOL_{type}^b = \sum_{y=1}^{N_y/2} SO_{type}^{yl,b} \quad (6)$$

4 Calculate the relative difference between the higher and lower observed flow signature averages, denoted as rdo_{type}^b :

$$rdo_{type}^b = (SOH_{type}^b - SOL_{type}^b) / SOL_{type}^b \quad (7)$$

5. Divide the modelled signatures $SM_{type}^{y,b}$ into the same observation-based subsets (yh and yl) and calculate SML_{type}^b , SMH_{type}^b and rdm_{type}^b as in steps 3–4 above.

6. Calculate the relative difference between rdo_{type}^b and rdm_{type}^b , denoted as RD_{type}^b (Fig. 1c) for each subbasin:

$$RD_{type}^b = (rdm_{type}^b - rdo_{type}^b) / rdo_{type}^b \quad (8)$$

In this context, a positive RD value indicates a greater discrepancy between the higher and lower subsets within the model-estimated flow signatures, compared with observations. On the other hand, a negative

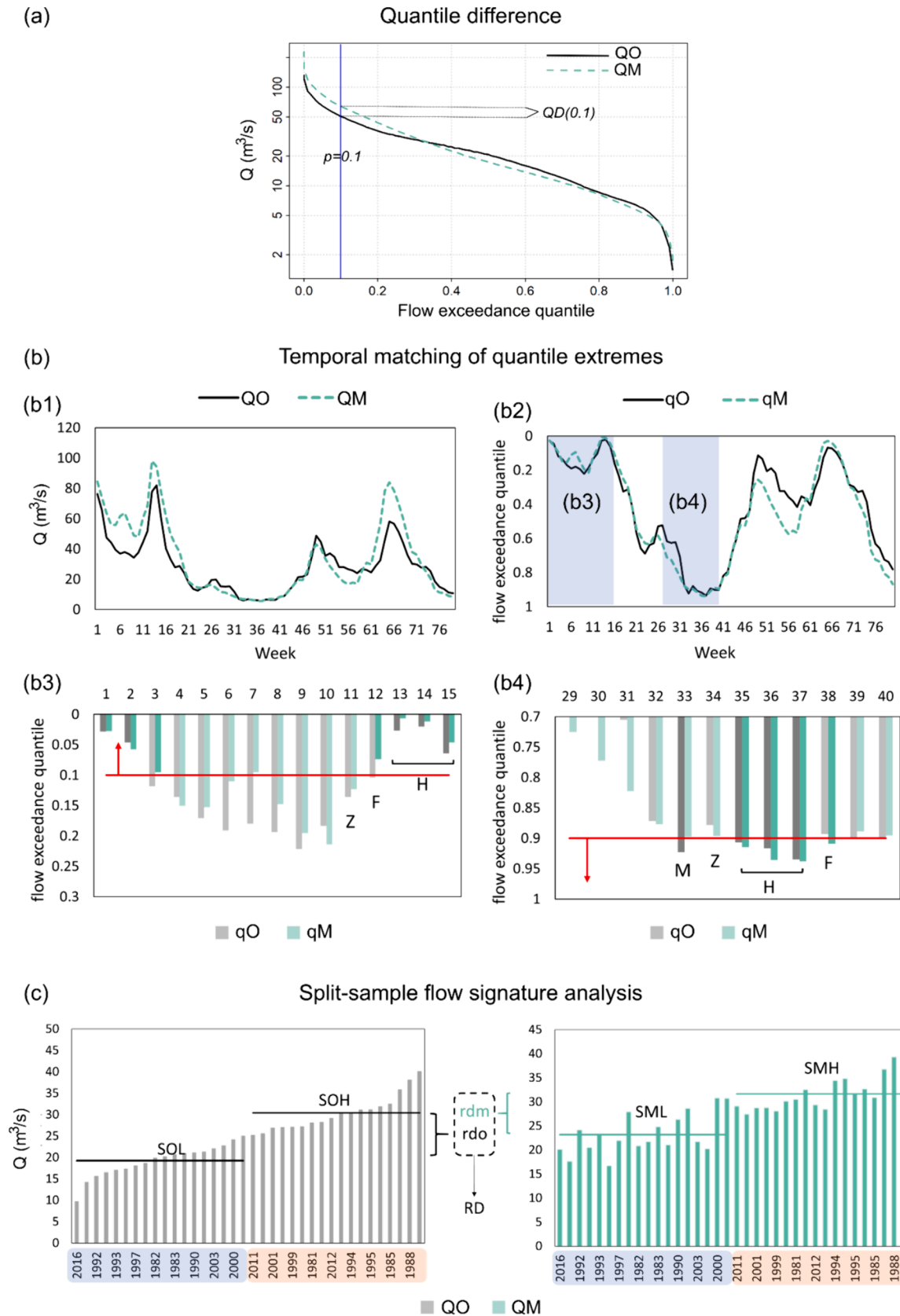


Fig. 1. Illustration of application-based metrics. (a) Observed (QO) and modelled (QM) streamflow at a defined exceedance probability for calculation of Quantile Difference (QD). (b) Example of quantile extreme events for calculation of temporal matching: observed (QO) and simulated (QM) streamflow in m^3/s (b1); observed (qO) and simulated (qM) streamflow in quantiles (b2); illustration of four types of events related to high flow (b3) and low flow (b4) thresholds (darker color denotes event crossing the corresponding thresholds). (c) Schematic of the calculation of RD for one flow signature type at one subbasin.

Table 1
Contingency table.

	Model: 1	Model: 0
Obs: 1	(1,1) H	(1,0) M
Obs: 0	(0,1) F	(0,0) Z

RD value suggests a lesser discrepancy, indicating a closer alignment between the higher and lower subsets. Consequently, the smaller the absolute value of RD, the better the agreement between the observed and modelled flow signature difference, respectively.

3. Case studies and model configurations

3.1. Study area and data

The locations of seven case study areas are shown in Figure A in the [Appendix](#), which represent an ensemble of opportunity based on the partners in the GlobalHydroPressure project, aiming to develop tools and methods for assessing hydrological pressure worldwide through coordinated case studies in Europe, South America and Asia. Information of case study characteristics and available datasets are listed in Table A in the [Appendix](#). The case study areas were classified into different Köppen climate systems ([Beck et al., 2018](#)), covering arid, tropical, temperate, and polar climate. The spatial extent of the different case studies ranges from a few to several hundred thousands of square kilometers.

Evaluation and comparison were done using output from the global hydrological model World Wide HYPE (WW-HYPE) ([Arheimer et al., 2020; Bartosova et al., 2021](#)). For each case study area, WW-HYPE submodels were extracted and forced with global and local inputs ([Section 3.2](#)). The global forcing data used for WW-HYPE and the case study submodels is HydroGFD version 3.2 ([Berg et al., 2021](#)). It is a near real-time data set of bias adjusted reanalysis data comprising daily precipitation as well as minimum, mean, and maximum daily temperature with a resolution of 25 km. In the WW-HYPE setup, the data are area weighted for each subbasin. Local streamflow and forcing data (mean precipitation and temperature) at daily time step are from different sources and for different time periods, provided by meteorological or water sector agencies in the case study countries (Table A in the [Appendix](#)). Local streamflow data are used for calibration and evaluation. The gridded observational forcing datasets (Sweden and Norway) are area weighted for each subbasin, as the global dataset, while for in situ observations, the point nearest the subbasin centre point was used to link precipitation and temperature data to the subbasins. Since the observations – streamflow, precipitation and temperature – start and end differently, the overlapping time period available for performance evaluation was determined for each case study separately.

We consider the local forcing data to represent the “first choice” in the different countries, when performing hydrological modelling in an arbitrary location. Any detailed quality control (and, if needed, improvement) of the local forcing data was not attainable within the resources of the study, but a simplified comparative assessment of the global and local forcing data in terms of mean values is shown in Table A in the [Appendix](#). Although the global and local mean values are generally similar, there is uncertainty in the local forcing datasets where the quality varies among the case studies. For example, in seNorge, there is cold bias in the coastal areas. Together with a warm bias in HydroGFD3.2, which is inherited from the original reanalysis datasets ([Berg et al., 2021; Lussana et al., 2019](#)), this gives a substantial difference between these two. However, the bias is less in the subbasins used for calibration which makes the dataset in general qualified for the purpose of this study. Overall, we assume and believe that quality of the forcing data is sufficient to meet the objectives of the study.

3.2. Configurations for World-Wide HYPE

In order to separate the influence from calibration and forcing data on model performance, three WW-HYPE model versions were developed in the seven case study areas, with different setting in terms of forcing data and calibration:

- Sim1: Global forcing data and global (i.e. generalized) calibration, extracted directly from WW-HYPE v. 1.3.7. The parameters have (in previous model development) been calibrated to reproduce the long-term water balance and the seasonal dynamics globally.
- Sim2: Global forcing data and local calibration. The parameters were (in this study) tuned to optimize the performance given the global forcing data (extracted for each basin).
- Sim3: Local forcing data and local calibration. The parameters were (in this study) tuned to optimize the performance given the local forcing data. Note that Sim3 is not performed in AKS, China, since the local meteorological data is restricted.

A stepwise parameter estimation method was used for the global calibration (Sim1) of WW-HYPE on monthly flows, where each step focuses on different processes ([Arheimer et al., 2020; Donnelly et al., 2016; Strömqvist et al., 2012](#)). This procedure allows the model to identify robust values for ungauged basins, as well as reliable process descriptions along the flow paths of the dominant flow-generation processes. Simultaneous calibration was conducted in numerous representative basins around the world to achieve the first goal. By separately calibrating basins that represented various climates, elevations, and land covers around the world, spatial variability was taken into consideration. The second goal was accomplished by employing a step-by-step methodology along the flow routes, following the HYPE process description, and only calibrating a small subset of the parameters driving a single process at a time. The calibration was a combination of automatic calibration toward the in-stream measurements and manual calibration, using NSE, RE and the Kling-Gupta Efficiency (KGE) as objective functions. More information of WW-HYPE model can be found in the [Appendix](#). Details of model performances for streamflow are available in [Arheimer et al. \(2020\)](#).

In the local calibration process (Sim2 and Sim3), specific model routines were added and optimized based on the dominant hydrological processes in the study areas. Automatic calibration was performed first with NSE and RE as objective functions, using the Differential Evolution (DE) algorithm for optimization. Afterwards, manual adjustment of a set of parameters was performed where visual judgment of the hydrographs was employed in addition to minimizing the objective functions, depending on the hydro-climatic characteristics of each study area:

- in VES (Norway) and AKS (China), soil, land use, snow and glacier modules were activated, with manual corrections on routing and missing outlet lakes;
- in Brazilian study areas, potential evapotranspiration and soil were the main processes, and in SF reservoir parameters were added;
- in EM (Sweden), soil, snow and potential evapotranspiration processes were used.

It deserves to be emphasized here that the evaluations for all model configurations are using the overlapping simulation period across the three configurations, as denoted in Table A in the [Appendix](#), without separating into calibration/validation subsets. This is in line with the objective of demonstrating the proposed new metrics, therefore we do not perform in-depth analysis into the parameter optimization. Moreover, the resources available for the calibration were limited and does not allow deep investigation for emerging issues, including the reasons behind certain cases of unsatisfactory model performance, e.g. quality of local observations. Nevertheless, as our primary focus is on the demonstration and evaluation of the proposed metrics, our methodology

requires a set of model simulations that represent a realistic range of performances (in a global modeling context), in terms of various metrics and signatures, therefore the chosen model configurations well fulfill this requirement.

4. Results

4.1. Conventional metrics

The results from evaluation by conventional statistical metrics are shown in Fig. 2. Medians within each study area were calculated for each statistical metric and plotted as points, besides the violin plot showing the density of metrics in all the evaluated stations regardless of study area. Progressive improvements when localizing model configurations were observed in MAE, as the medians in the violin plot decreased from 53.4 (Sim1), through 31.2 (Sim2), to 23.6 (Sim3), besides a gradually flatter distribution (Fig. 2a, right). The MAE values in different study areas showed a similar pattern that a larger improvement was achieved when introducing local calibration in Sim2, while less improvement was found when introducing local dataset in Sim3 (Fig. 2a,

left).

A similar progressive improvement was revealed in CC, with medians increasing through the localizing model setups: 0.657 for Sim1, 0.722 for Sim2 and 0.760 for Sim3 (Fig. 2b, right), which confirms a gradual improvement in the model performance with respect to temporal variability. In most of the study areas (Fig. 2b, left), considerable improvements occurred both in model configurations with local calibration and local dataset, while in EM and SF, larger improvement was achieved by adding local calibration. It is also interesting to note that the interquartile range of CC from Sim3 was larger than that in Sim2 (Fig. 2b, right), indicating a less stable simulation in terms of temporal variability by introducing local datasets.

The median absolute RE (%) showed a decrease from 55.8 % (Sim1), 12.6 % (Sim2), to 10.8 % (Sim3) (Fig. 2c, right). However, this improvement varied among case study areas (Fig. 2c, left): Sim2 showed a general superiority in most study areas (except SAP and SF) with values closer to zero, indicating that a better model performance of long-term averages was achieved with local calibration together with the global dataset.

Gradual improvements in NSE were observed across all study areas,

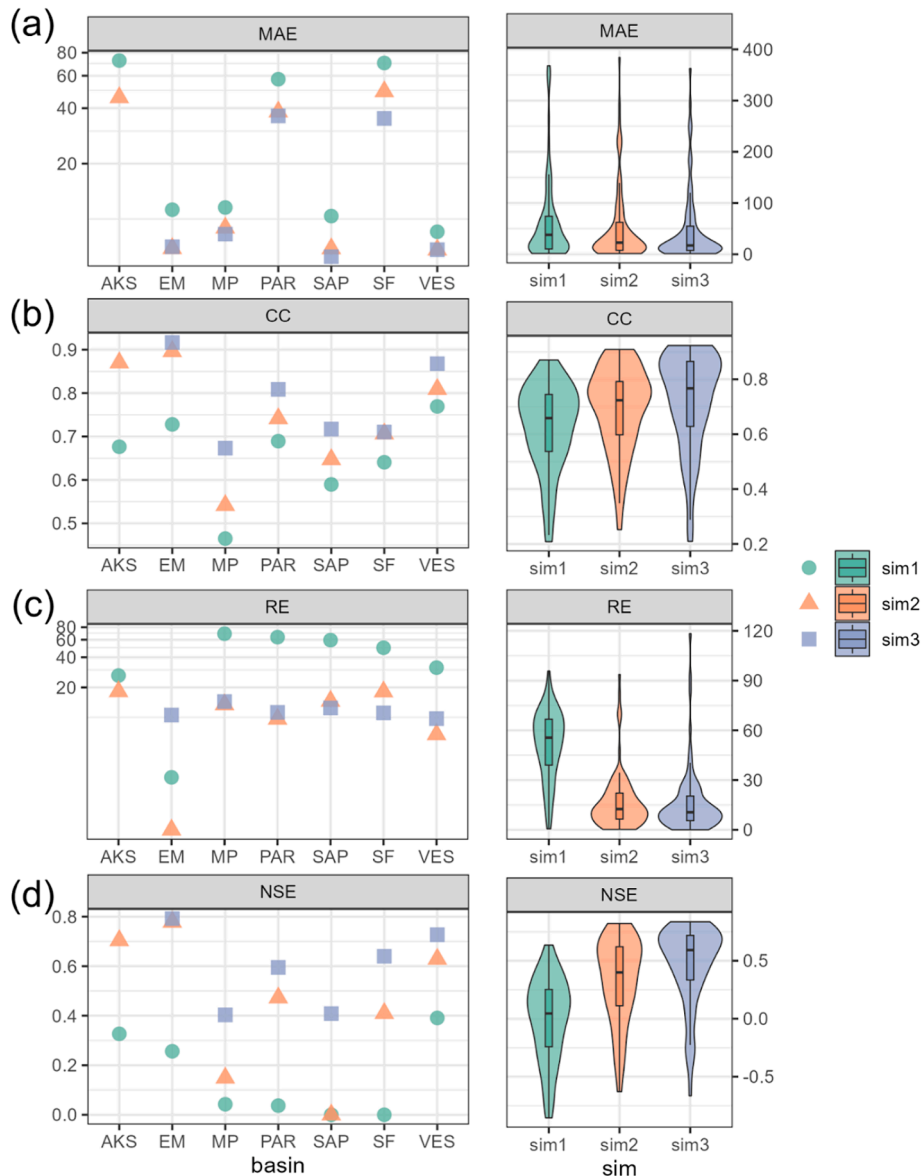


Fig. 2. Conventional evaluation metrics with median values for each study area (left panel) and full distribution for each model setup (right panel): (a) MAE; (b) CC; (c) RE (%); (d) NSE. Note the different scales on the y-axes.

from Sim1 to Sim2 and then to Sim3, as indicated by the increasing medians in each study area (Fig. 2d, left) and a shift towards higher NSE in the density (Fig. 2d, right). The median NSE for all study areas achieved increased from 0.042 (Sim1), through 0.314 (Sim2), to 0.507 (Sim3). Comparing different model setups (Fig. 2d, left), local calibration generally resulted in a significant improvement in NSE in most study areas, except for SAP where no obvious increase was observed in Sim2. Furthermore, the introduction of local datasets further improved NSE, except for EM with no clear improvement achieved in Sim3.

The above results from using conventional metrics to evaluate the performance of hydrological models serve as a benchmark for further analysis and comparison with application-based metrics. Generally, increasing localization of model configurations leads to an improvement in performance. However, the impact of localizing calibration or forcing data varied across the different case study areas. A significant enhancement in model performance was observed in cases of local calibration with global datasets, sometimes surpassing even the improvement achieved with local datasets. This may be attributed to the fact that the local observations used in Sim3, despite their accuracy, are limited in the aspect of spatial consistency, which can affect the model efficiency when local observations are used for parameter calibration.

Under most circumstances, the incorporation of local data can result in better performance, however, a possibility of over-fitting should be noted, where the model can excessively adapt to the specific characteristics of the local data at the expense of general applicability. Additionally, the influence of anthropogenic activities, e.g. the complex operational patterns of large reservoirs for hydropower generation and intense irrigation in PAR and SF, further complicates model simulation and sometimes leads to limited improvements when relying on local datasets.

4.2. Quantile extremes and time-matching

4.2.1. Flow quantiles

An assessment of high and low flow extremes was conducted for each model setup, evaluating QD at the selected thresholds (Fig. 3); high flow extremes are listed on the left side of each figure at thresholds 0.05, 0.1 and 0.2 and low flow extremes are on the right at thresholds 0.8, 0.9 and 0.95.

In the case of the AKS study area (Fig. 3a), Sim1 was found to underestimate both high and low flows at all evaluated thresholds. Sim2 also underestimated the extremes at most thresholds, although with a

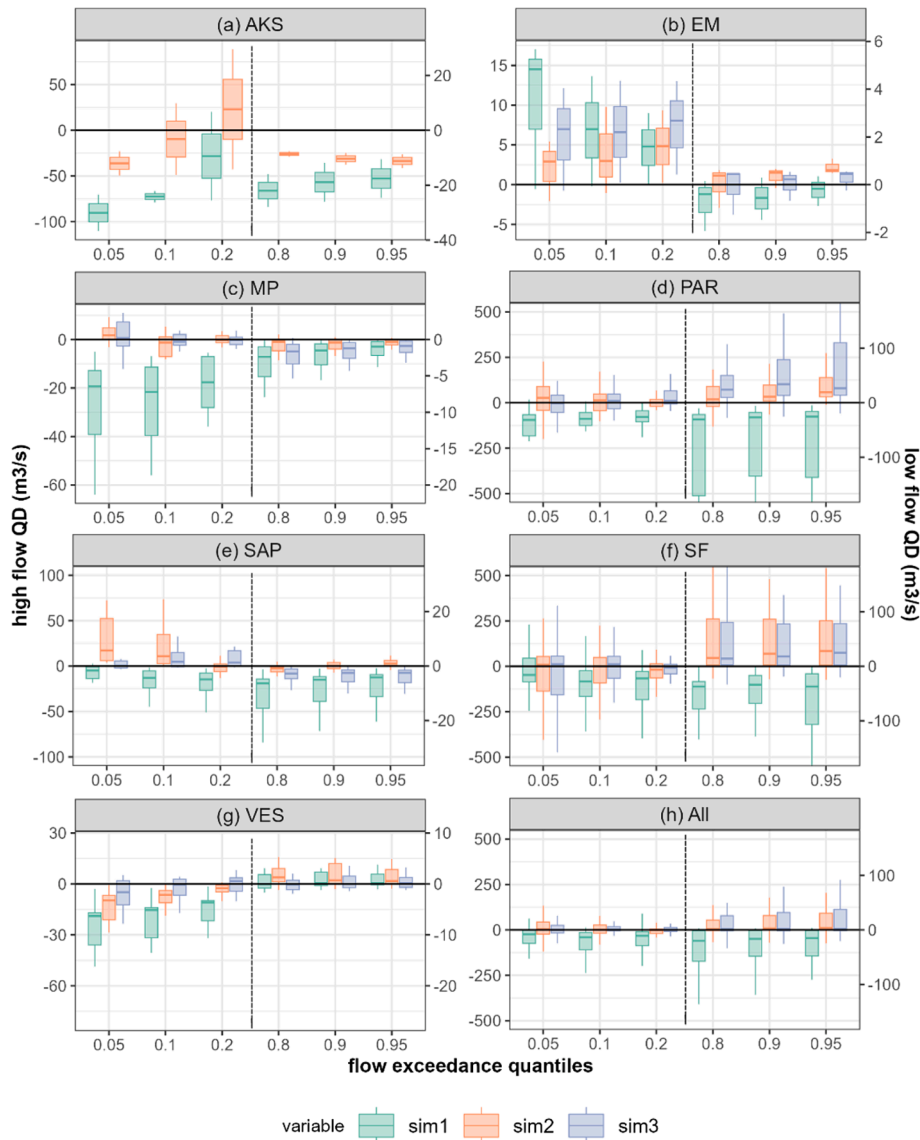


Fig. 3. Boxplot of QD at different quantile thresholds for each study area (a)-(g): high flow QD (0.05, 0.1, 0.2) at the left y-axis, low flow QD (0.8, 0.9, 0.95) at the right y-axis, with dotted line in between for separation. All (h) shows results from all stations in all case study areas. Note the different scales on the y-axes.

smaller amplitude of QD when compared with Sim1. This outcome suggests that the local calibration in Sim2 enhances the flow duration curve at most thresholds. For threshold 0.2, Sim1 and Sim2 provided comparable results concerning the amplitude of QD, with Sim1 underestimating and Sim2 overestimating the extremes.

In EM (Fig. 3b), all three model configurations displayed an overestimation of the target high flow extremes. Notably, Sim3 did not yield optimal performance in terms of QD for high flow extremes, but Sim2 consistently outperformed the other two model configurations. For low flow extremes, Sim2 and Sim3 produced slight overestimations, whereas Sim1 showed distinct underestimations.

In MP and PAR (Fig. 3c, d), by introducing local calibration to the global dataset (Sim2), the model performance was substantially improved with reduced QD in both high and low flows. With the local dataset in Sim3, not much improvement was achieved, especially at low flow thresholds. Sim3 showed inferior results to Sim2, indicating that

the global forcing dataset with local calibration is sufficient in terms of capturing quantile extremes. In SAP (Fig. 3e), Sim2 outperformed the other two model setups with a lower QD for low flows, while Sim3 proved superior for high flows. In SF (Fig. 3f), Sim2 and Sim3 brought certain improvements for both low and high flows regarding median QD, but sometimes with larger interquartile ranges, i.e. a larger spread among the subbasins.

In VES (Fig. 3g), Sim2 and Sim3 achieved considerable improvements in high flow QD. For low flows only Sim3 managed to outperform Sim1, and Sim2 gave less accurate results compared with Sim1, in terms of both median and spread.

In summary, there is generally a gradual improvement in the high flow quantiles from Sim1 to Sim3 (see also Fig. 3h) by adding local calibration and local forcing data. However, for the low flow quantiles the pattern is less clear with Sim1 often showing similar (and sometimes even better) results than Sim2 and Sim3, indicating the capability of the

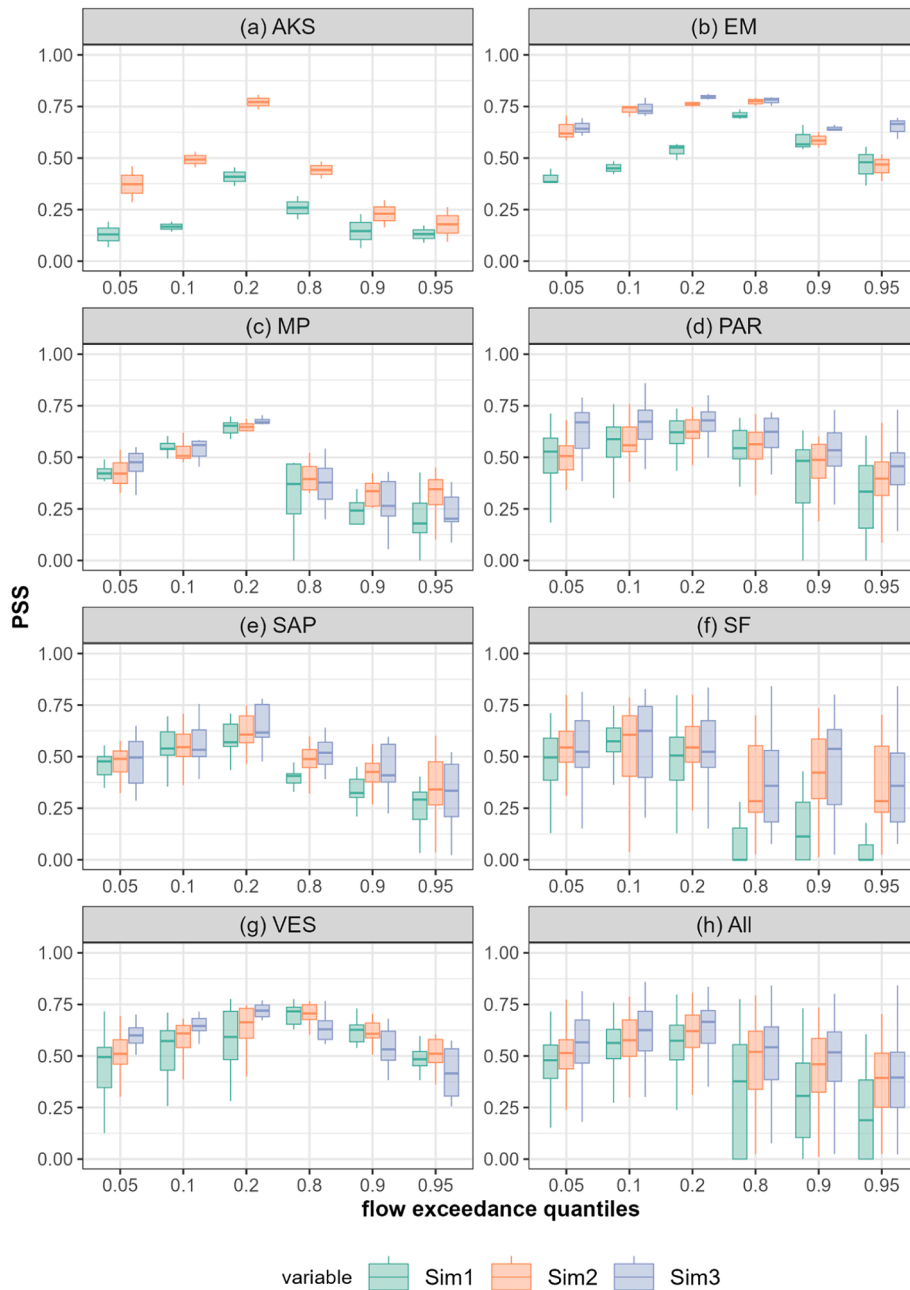


Fig. 4. Boxplots for PSS at defined thresholds in each case study area (a)-(g), All (h) denotes subbasins in all case study areas.

global configuration models in capturing prolonged dry periods. This can be attributed to the lower complexity and higher homogeneity in both space and time for low flow extremes, therefore the generalized configuration is effective in capturing their dynamics. The localized configurations, on the other hand, are tailored towards local hydrological characteristics, which are highly influenced by variability in high flow extremes and therefore sometimes not sufficient in low flow extremes. With different regional hydro-climatic regimes, the model performance in QD varies among different study areas, which sets the base for the application-based metrics of quantile time matching.

4.2.2. Temporal matching of quantile extremes

The analysis of time-matching quantile extremes was carried out using PSS for both high flow and low flow extremes, as depicted in Fig. 4. All case study areas display a more or less distinct inverted U shape, reflecting that the temporal matching is generally best for the least extreme quantiles (0.2 and 0.8) and worst for the most extreme ones (0.05 and 0.95). Generally, PSS values are somewhat higher for the high-flow quantiles than the low-flow quantiles, which is reflected in the combined diagram (Fig. 4h).

The results of PSS for high flow extremes are listed on the left side of each figure at thresholds of 0.05, 0.1 and 0.2 (Fig. 4). A general trend can be observed in which PSS values increase as the model configurations become more localized (Fig. 4h), although distinct patterns emerged when the case study areas were examined individually. For instance, a stepwise enhancement from Sim1 to Sim2 and subsequently to Sim3 was noted in the AKS, EM, and VES case study areas, as indicated by the rise in median PSS and/or the reduction in spread. Overall,

the differences among the three model configurations were minimal. According to median PSS values, Sim1 occasionally surpassed Sim2 (e.g., high flow thresholds in PAR) and Sim3 (e.g., threshold 0.1 in MP), while in other instances, Sim2 outperformed both Sim1 and Sim3 (e.g., thresholds 0.05 and 0.2 in SF).

The results of PSS for low flow extremes are listed on the right side of each figure at thresholds of 0.8, 0.9 and 0.95 in Fig. 4. Across all case study areas, PSS at low flow thresholds showed an increase from Sim1 to Sim2 and subsequently to Sim3 (Fig. 4h). However, distinctly different improvement patterns were observed among the various case study areas. In AKS, Sim2 showed higher PSS compared to Sim1 at low flow thresholds, although not to the same extent as observed for high flows. In EM, a remarkable enhancement was achieved by Sim3 with the local dataset at the lowest thresholds, whereas Sim2 had already significantly improved PSS for high flows. In the four Brazilian case study areas, PSS occasionally showed a gradual increase for low flow extremes as model configurations progressed from Sim1 to Sim3 (e.g., PAR and SF). Conversely, in MP and SAP, Sim2 outperformed the other two setups at specific thresholds. In VES, PSS at low flow thresholds displayed a behavior distinctly different from that at high flow thresholds, as performance is not increasing but even declining slightly upon incorporating local dataset. This finding suggests that Sim1 holds a potential advantage in capturing the timing of low flow extremes.

4.3. Flow signatures and split-sample analysis

4.3.1. Flow signatures

Fig. 5 presents a boxplot of flow signatures ($S_{\max}/S_{\text{mean}}/S_{\min}$) for the

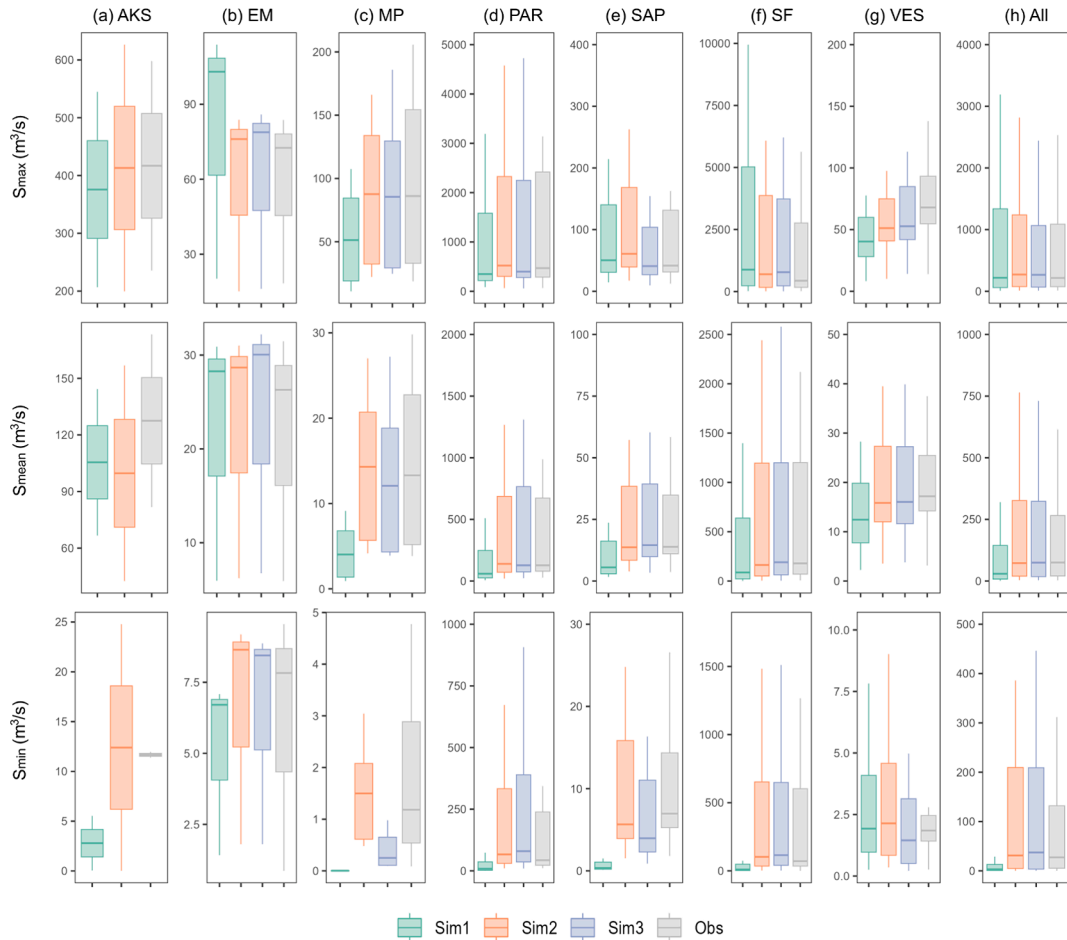


Fig. 5. Boxplot of flow signatures for each flow signature ($S_{\max}, S_{\text{mean}}, S_{\min}$) in case study area (a)-(g), and All (h) denote subbasins from all case study areas. Note the different scales on the y-axes.

various model configurations and corresponding observations in each study area. A general trend (Fig. 5h) indicates enhanced model performance with the incorporation of local calibration and local datasets, as demonstrated by the medians of Sim2 and Sim3 approaching the observations for the three flow signatures.

$S_{\max}$ showed improved performance from Sim1 to Sim2 and then to Sim3 (Fig. 5, $S_{\max}$), as the medians approached those of observations, particularly when local calibrations were included in the models. Notably, in EM, PAR, and SF, the best performance was achieved in Sim2, which employed local calibration with the global dataset. This aligns with the high flow QD results (Fig. 3), where Sim2 and Sim3 displayed a similar better performance in terms of QD in those case study areas.

Similarly to $S_{\max}$, generally clear improvements were observed in S_{mean} , when model configurations move from Sim1 to more localized ones, i.e. Sim2 and Sim3 (Fig. 5, S_{mean}). Overall, Sim1 tended to underestimate S_{mean} in most cases (except EM), while Sim2 and Sim3 gave results closer to observations. However, minimal improvement was seen from Sim2 to Sim3. For instance, the performance of Sim2 and Sim3 in MP, PAR, and SAP was comparable. This finding generally validates our confidence in the global dataset HydroGFD, as Sim2 demonstrated a strong capability when applying local calibration to the global dataset.

The results for $S_{\min}$ also showed variation among study areas and model setups (Fig. 5, $S_{\min}$), with Sim1 consistently underestimating $S_{\min}$, while Sim2 and Sim3 displayed improved but similar performance. Based on the performance of Sim1 with respect to $S_{\min}$, global calibration appears to be insufficient for capturing $S_{\min}$, whereas Sim2 and Sim3 achieved considerable enhancements when using local calibration. A similar pattern was observed in low flow QD (Fig. 3), where Sim2 and Sim3 showed substantial improvement in comparison to Sim1.

4.3.2. Split-sample analysis

As described in section 2.2.2, rdm and rdo were calculated for each flow signature. An illustrative scatter plot of rdo and rdm for the mean flow signature (S_{mean}) is shown in Fig. 6. Notably, rdm is consistently positive, i.e. SMH is always larger than SML (see Fig. 4.c). Points situated above the 1:1 line indicate a higher rdm value than rdo, suggesting that the model-estimated flow signatures gave a larger discrepancy between the higher and lower subsets, subsequently resulting in positive RD values (Equation (8)), and vice versa. Any underestimation represented by points below the 1:1 line can be attributed to either the flow signature

values or a mismatch in the years corresponding to the higher/lower subsets between models and observation. As illustrated in Fig. 6, for annual mean flow Sim1 showed a tendency for overestimating the differences between higher/lower subsets, with many points located above the 1:1 line. In contrast, Sim2 and Sim3 were situated closer to and mostly below the 1:1 line, indicating underestimation of the discrepancy between higher/lower subsets in these simulations.

Fig. 7 presents the computed RD between each model setup and observation. The general pattern across all case study areas (Fig. 7h) reveals that Sim2 and Sim3 tend to give negative RD for all three signatures, i.e. underestimates of the observed difference in the flow signatures, while Sim1 occasionally overestimates rdo with positive RD (as illustrated in Fig. 6).

Focusing on RD_{mean} (Fig. 7, second row, RD_{mean}) for each case study area individually, Sim2 and Sim3 show more robust results in terms of RD, with a smaller spread among the subbasins. In contrast, Sim1 displays larger RD_{mean} spreads in some study areas (e.g., MP, PAR). As local calibration and local datasets were incorporated, RD_{mean} gradually improved in the majority of case study domains (six case study areas, excluding VES), with the median RD_{mean} approaching zero in Sim2 and Sim3, though sometimes with a somewhat larger spread (e.g., EM, SF). However, for VES, Sim1 outperformed the other two model setups, and both Sim2 and Sim3 had negative RD_{mean} values.

Regarding $RD_{\min}$ and $RD_{\max}$, the results show greater uncertainty, with larger RD in both positive and negative directions (Fig. 7, $RD_{\max}$ and $RD_{\min}$). This is particularly true for $RD_{\min}$, where small discharge values (sometimes reaching zero) can generate considerable uncertainties. Consequently, RD for these two flow signatures should be interpreted in a general manner. For the annual maximum flow signatures (Fig. 7, $RD_{\max}$), Sim2 and Sim3 consistently underestimated the observed difference rdo, with negative $RD_{\max}$ for all case study areas. In contrast, Sim1 occasionally showed positive $RD_{\max}$ (e.g., PAR, SAP, and SF). For annual minimum flow signatures (Fig. 7, $RD_{\min}$), most model setups produced negative $RD_{\min}$ values. However, Sim1 displayed substantial uncertainty in its performance, with RD values reaching up to 500 % (e.g., PAR, SAP, and SF), due to the small magnitudes of minimum flow signatures.

To sum up, it should be noted that when a model has systematic bias in long term averages S, it may still offer added value in terms of RD. Sim2, with global forcing data and local calibration, can yield results comparable to Sim3, presenting an alternative for data-scarce regions.

5. Discussion

5.1. Evaluation cross the metrics

Based on the metrics introduced above, a more in-depth understanding can be achieved by a comprehensive evaluation across metrics.

A comparison of PSS values (Fig. 4) with QD values (Fig. 3) yields several noteworthy insights. First, certain model configurations show their superiority in reproducing extremes with both accurate amplitude and timing. For instance, in the case of low flow extremes in MP, Sim2 achieves the highest PSS and QD values closest to zero. Second, when model setups show similar QD values in terms of amplitude (i.e. deviation from zero), their performance can vary with respect to the timing of extreme events, as indicated by PSS. In AKS, at a high flow threshold of 0.2, Sim1 and Sim2 yield comparable QD amplitudes; however, Sim2 displays a PSS value exceeding 0.75, while the PSS of Sim1 is approximately 0.4. This suggests that Sim2 holds an advantage in simulating the timing of low flow periods which is not noticeable in metrics reflecting only the distribution (e.g. the flow duration curve). Furthermore, a lower QD value does not necessarily imply unqualified performance in terms of extreme event timing. This is evident in the low flow extremes in PAR, where Sim3, despite not outperforming Sim2 in terms of QD, achieves a higher PSS value. Similarly, in MP, Sim1 shows worse QD performance compared to the other two set-ups, yet its PSS performance

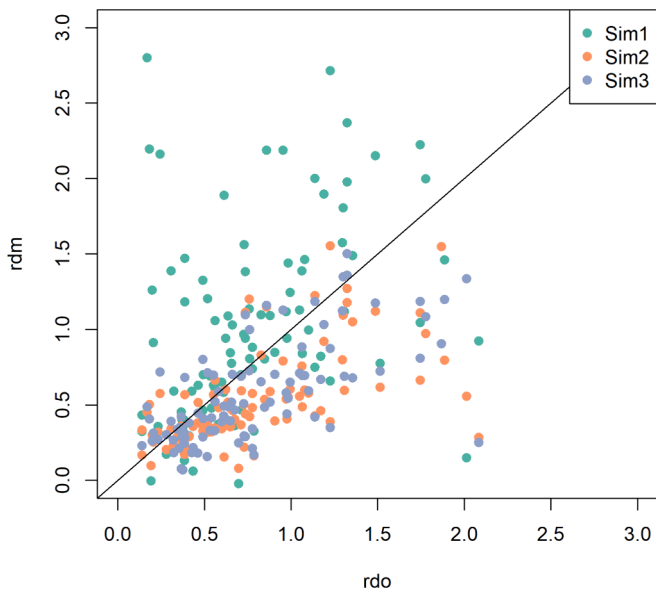


Fig. 6. Scatter plot of rdo and rdm for mean flow signatures at all stations in the seven case study areas.

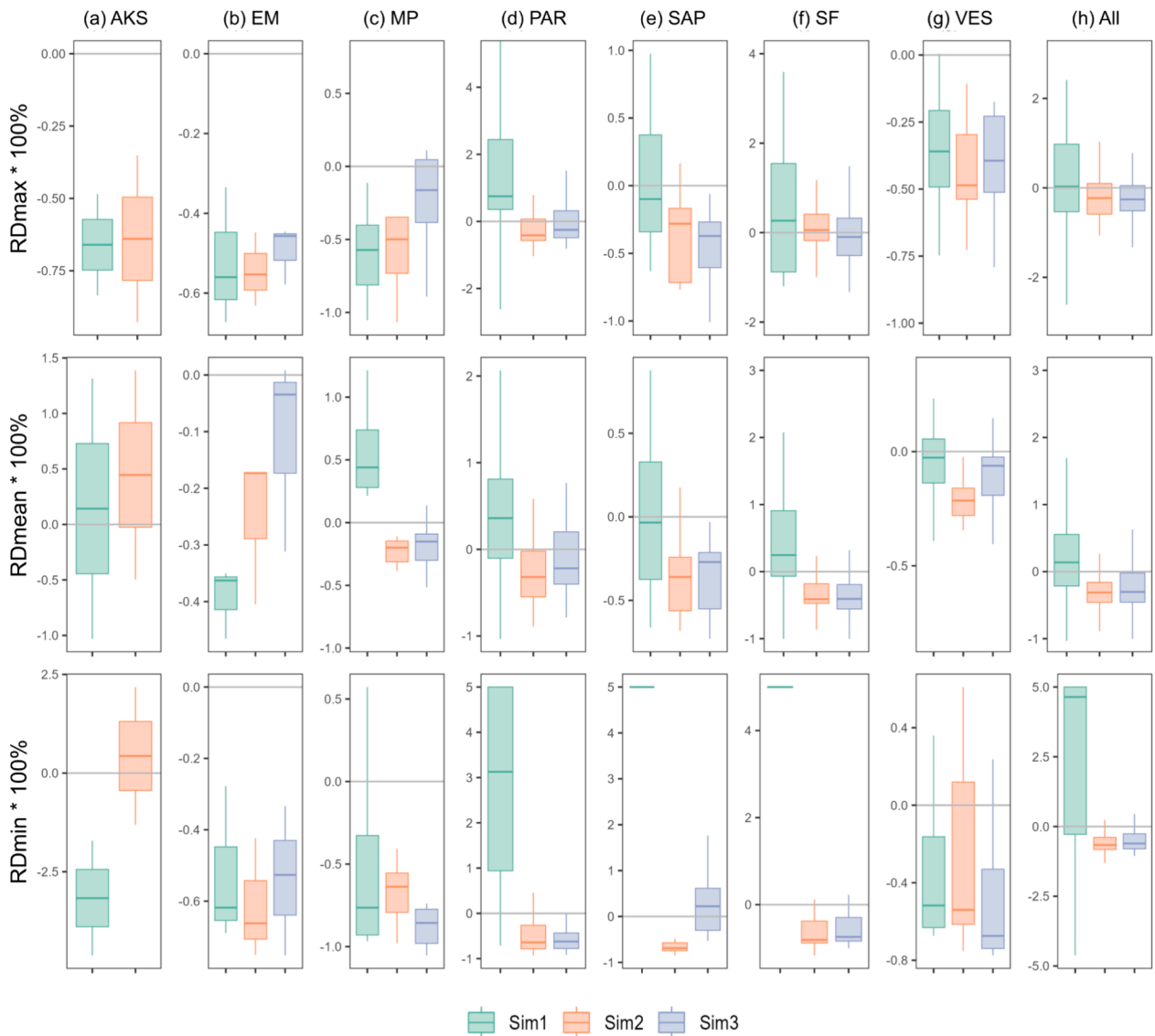


Fig. 7. Boxplot of RD_{max} , RD_{mean} and RD_{min} in flow signatures for each study area (a)-(g) and All (h) denotes subbasins from all case study areas. Note the different scales on the y-axes.

is comparable, indicating that Sim1 is better at simulating the temporal occurrence of extremes than the flow distribution.

Moreover, comparing the results from RD (Fig. 7) with long term averages of flow signatures S (Fig. 5), several interpretations can be made. In some study areas, the model setup that achieved the best performance in terms of RD also excelled in S , as seen for the maximum flow signatures in AKS, MP, and PAR. Furthermore, in some cases, the introduction of local calibration and local data in Sim2 and Sim3 led to improvements in both S and RD (e.g., MP and SF for S_{mean} and RD mean), indicating reduced bias in flow signatures and enhanced reproduction of the difference. Conversely, the results from RD and S occasionally yielded different conclusions. In AKS and VES, Sim1 outperformed or equaled Sim2 in terms of RDmean, while Fig. 5 demonstrated Sim1's reduced capability in accurately capturing S_{mean} . This suggests that when local calibration/local dataset in Sim2/Sim3 improves the model performance with regard to the long-term average of mean flow signatures, uncertainties in representing the relative difference between high- and low-flow years may arise as a trade-off. For minimum flow signatures, due to uncertainties stemming from small discharge values, the only conclusion that can be drawn is that model setups with advantages in simulating S do not always have the same

advantages in terms of RD.

5.2. Incorporation into hydrological services

Based on our results, it is clear that model performance can have either agreement or discrepancy when assessed by different evaluation metrics. Instances of agreement strengthen general confidence in the evaluated model configurations, while instances of discrepancy provide valuable insights for applying these models in relatively reliable contexts. We propose a schema for incorporating the application-based metrics into the hydrological service (Fig. 8), where the comprehensive evaluation can benefit the communication with users by providing an indicator of model applicability. Two “applicability categories” can be defined as follows, where suggestions of application scenarios can be provided to the users accordingly:

- **Generally Applicable Models:** These models have excellent performance in terms of both conventional and application-based metrics. They can be used confidently for various purposes across different time scales, including simulating discharge values for urban water supply, amplitude and timing of extreme events for flood or drought

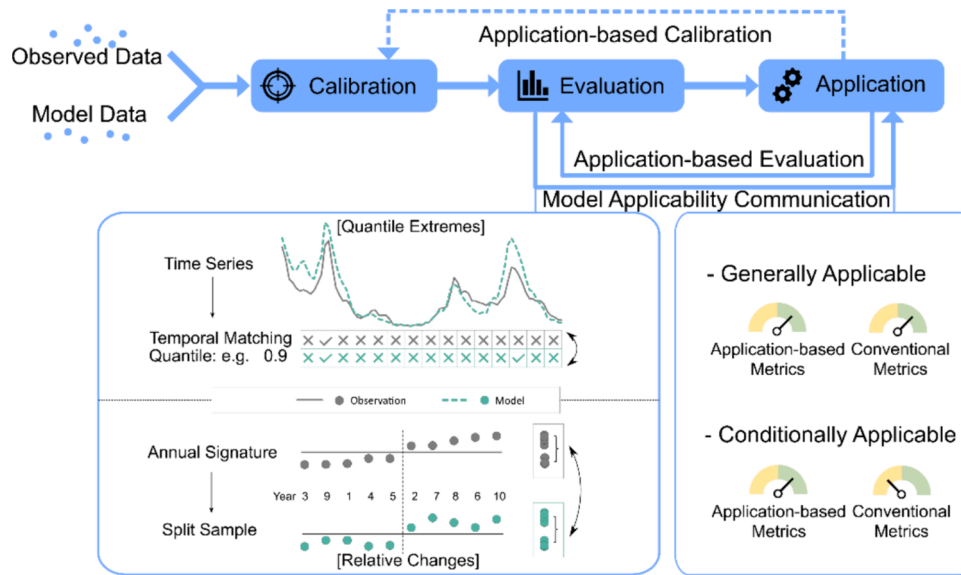


Fig. 8. Schema of incorporating comprehensive evaluation metrics into the hydrological service. Solid blue line denotes the process included in this study, dashed blue line denotes the next step.

risk monitoring, and decadal change of hydrological signatures. Examples of this type of models from this study include Sim2 in AKS and EM, as well as Sim3 in the other case study areas. This model configuration has near-optimal performance for most of the evaluated metrics, both conventional and application-based, and may therefore be regarded as an optimal model when the required forcing data and resources are available.

- **Conditionally Applicable Models:** These models show relatively good but not excellent performance in terms of conventional metrics, while the application-based metrics are highly qualified. These models can be applied under certain circumstances, acknowledging their potential limitations in terms of statistical accuracy. Their application should be limited to the context of the corresponding application-based metrics and not extended to value-based applications. For instance, Sim2 in VES and SAP show their accuracy in replicating the time matching of quantile extremes, while also having a satisfactory overall performance in conventional metrics. Thus, these models can have a value in a forecasting system based on model climatology quantiles, benefitting water resources planning activities, e.g. reservoir management. Moreover, Sim1 in SAP manages to reasonably well capture the relative difference in flow signatures between two periods. Given the challenges associated with the scarcity and collection of local forcing data in the Brazilian case study areas, the global model adds significant value by offering an accessible hydrological service for climate adaptation in these regions.

In summary, with applicability categories, corresponding scenarios for an available hydrological service can be communicated with the users, which not only expands possibilities for services with suboptimal performance but also builds up the overall reliability. Therefore, it is essential for hydrological service providers to furnish the information of model applicability. The next step of our work is to investigate a further implementation of the application-based metrics, to incorporate them into the model calibration process as objective functions (Fig. 8), therefore allowing the model to be tailored to certain application scenarios.

6. Conclusions

In conclusion, this research seeks to develop and test a set of

application-based evaluation metrics that focus on (1) temporal matching of discharge extremes and (2) differences in flow signatures between periods, complementing traditional metrics and offering additional insights for practical applications of multi-basin hydrological models. The proposed evaluation framework was compared to conventional statistical metrics, examining the performance of the World-Wide HYPE (WW-HYPE) global hydrological model in seven diverse case study areas. These assessments were conducted using three model configurations, which varied in terms of the localization (from global to local) of both the calibration and the forcing data, further illustrating the applicability of the developed metrics in hydrological modeling.

The main findings of this study can be summarized as follows:

1. The application of both conventional and application-based metrics to evaluate the models' capacity of simulating discharge from various perspectives generally indicates significant improvements when introducing local calibration and local forcing data, as compared with the original global model, where local calibration appears to be the main key to enhance model performance.
2. The comparison of conventional metrics and application-based metrics offers insights in particularly two respects. Firstly, in some cases, incorporating local calibration and local dataset can enhance model performance in both conventional and application-based metrics, thereby suggesting a clear advantage in localizing model setups as much as possible. Secondly, it is common for improvements to be observed in only specific aspects, reflected by specific evaluation metrics, but not in others. In such cases, it is worthwhile to assess the resources required for localizing model setups considering the expected improvements from the perspective of a particular user. As demonstrated in the results, local calibration (Sim2) often delivers relatively good outcomes, comparable to also using local forcing data (Sim3), but without demand for local forcing data.
3. Models can be classified according to the consistency between outcomes derived from statistical metrics and application-based metrics, with different application scenarios. Generally applicable models excel under most evaluation metrics, and therefore can be applied in various aspects, including urban water supply, flood or drought risk monitoring, and decadal change of water resources. Conditionally applicable models serve as a supplementary option to the generally applicable models, which can be useful especially in resource limited

regions, offering potentially reliable information within the context of the relevant application-based metrics.

Global multi-basin hydrological modelling is a growing field where much remains to be done. Even if performance varies for different regions/climates, basin sizes, dominant processes, etc., it is important to look for useful signals and applicability that may not be apparent in the results of conventional evaluation. This study was an attempt to tackle this challenge with the hope that it will spawn follow-up research along these lines.

CRedit authorship contribution statement

Yiheng Du: Formal analysis, Investigation, Visualization, Writing – original draft, Writing – review & editing. **Jonas Olsson:** Investigation, Methodology, Writing – original draft, Writing – review & editing. **Kristina Isberg:** Investigation, Validation, Visualization, Writing – review & editing. **Johan Strömqvist:** Investigation, Validation. **Yeshe-watesfa Hundecha:** Investigation, Validation. **Benedito Cláudio da Silva:** Data curation, Validation, Writing – review & editing. **Sameh Adib Abou Rafee:** Data curation, Validation. **Carlos Ruberto Fragoso Jr:** Data curation. **Stein Beldring:** Data curation, Validation. **Anna Hansen:** Data curation, Validation. **Cintia Bertacchi Uvo:** Writing – review & editing, Conceptualization, Project administration. **Johanna Sørensen:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the EC and Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG Water/JPI (grant APQ-04377-18), Formas (grant no 2018-02378), Research Council of Norway (grant no 297298), Fundação de Amparo à Pesquisa do Estado de Alagoas - FAPEAL (grant no 05/2018), in the frame of the collaborative international Consortium GlobalHydroPressure financed under the 2017 Joint call of the IC4Water - JPI Water, and CAPES (Grant no. 88881.178692/2018-01). C.B.Uvo was partially financed by the Academy of Finland (grant no.243040491) that supports the Freshwater Competence Centre.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jhydrol.2024.131727>.

References

- Aghakouchak, A., Mehran, A., 2013. Extended contingency table: Performance metrics for satellite observations and climate model simulations. *Water Resources Research* 49 (10). <https://doi.org/10.1002/wrcr.20498>.
- Arheimer, B., Lindström, G., & Olsson, J. (2011). A systematic review of sensitivities in the Swedish flood-forecasting system. In *Atmospheric Research* (Vol. 100, Issues 2–3). Doi: 10.1016/j.atmosres.2010.09.013.
- Arheimer, B., Pimentel, R., Isberg, K., Crochemore, L., Andersson, J.C.M., Hasan, A., Pineda, L., 2020. Global catchment modelling using World-Wide HYPE (WWH), open data, and stepwise parameter estimation. *Hydrology and Earth System Sciences* 24 (2). <https://doi.org/10.5194/hess-24-535-2020>.
- Bartholmes, J.C., Thielen, J., Ramos, M.H., Gentilini, S., 2009. The european flood alert system EFAS - Part 2: Statistical skill assessment of probabilistic and deterministic

- operational forecasts. *Hydrology and Earth System Sciences* 13 (2). <https://doi.org/10.5194/hess-13-141-2009>.
- Bartosova, A., Arheimer, B., de Lavenne, A., Capell, R., Strömqvist, J., 2021. Large-Scale Hydrological and Sediment Modeling in Nested Domains under Current and Changing Climate. *Journal of Hydrologic Engineering* 26 (5). [https://doi.org/10.1061/\(asce\)jhe.1943-5584.0002078](https://doi.org/10.1061/(asce)jhe.1943-5584.0002078).
- Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Data Descriptor: Present and future Köppen-Geiger climate classification maps at 1-km resolution. Doi: 10.1038/sdata.2018.214.
- Berg, P., Almén, F., Bozhinova, D., 2021. HydroGFD3.0 (Hydrological Global Forcing Data): A 25 km global precipitation and temperature data set updated in near-real time. *Earth System Science Data* 13 (4). <https://doi.org/10.5194/essd-13-1531-2021>.
- Candogan Yossef, N., Van Beek, L.P.H., Kwadijk, J.C.J., Bierkens, M.F.P., 2012. Assessment of the potential forecasting skill of a global hydrological model in reproducing the occurrence of monthly flow extremes. *Hydrology and Earth System Sciences* 16 (11). <https://doi.org/10.5194/hess-16-4233-2012>.
- Cao, W., Zhang, Z., Liu, Y., Band, L.E., Wang, S., Xu, H., 2021. Seasonal differences in future climate and streamflow variation in a watershed of Northern China. *Journal of Hydrology: Regional Studies* 38. <https://doi.org/10.1016/j.ejrh.2021.100959>.
- Carpenter, T.M., Sperflage, J.A., Georgakakos, K.P., Sweeney, T., Fread, D.L., 1999. National threshold runoff estimation utilizing GIS in support of operational flash flood warning systems. *Journal of Hydrology* 224 (1–2). [https://doi.org/10.1016/S0022-1694\(99\)00115-8](https://doi.org/10.1016/S0022-1694(99)00115-8).
- CERH-MG. (2015). Normative Deliberation No.49, 2015. www.igam.mg.gov.br/component/content/article/16/1553-escassez-hidrica. Accessed 16 May 2023.
- Climate Services, 2023. <https://klimaservicesenter.no/kss/om-oss/nccs?locale=en>.
- Crochemore, L., Cantone, C., Pechlivanidis, I.G., Photiadou, C.S., 2021. How Does Seasonal Forecast Performance Influence Decision-Making?: Insights from a Serious Game. *Bulletin of the American Meteorological Society* 102 (9). <https://doi.org/10.1175/BAMS-D-20-0169.1>.
- Dakhlaoui, H., Ruelland, D., Trambay, Y., Bargaoui, Z., 2017. Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. *Journal of Hydrology* 550. <https://doi.org/10.1016/j.jhydrol.2017.04.032>.
- Dakhlaoui, H., Ruelland, D., Trambay, Y., 2019. 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. <https://doi.org/10.1016/j.jhydrol.2019.05.056>.
- Donnelly, C., Andersson, J.C.M., Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate the E-HYPE multi-basin model across Europe. *Hydrological Sciences Journal* 61 (2). <https://doi.org/10.1080/02626667.2015.1027710>.
- Donnelly, C., Ernst, K., Arheimer, B., 2018. A comparison of hydrological climate services at different scales by users and scientists. *Climate Services* 11, 24–35. <https://doi.org/10.1016/j.cliser.2018.06.002>.
- Ewen, J., 2011. Hydrograph matching method for measuring model performance. *Journal of Hydrology* 408 (1–2). <https://doi.org/10.1016/j.jhydrol.2011.07.038>.
- Filianoti, P., Gurnari, L., Zema, D.A., Bombino, G., Sinagra, M., Tucciarelli, T., 2020. An evaluation matrix to compare computer hydrological models for flood predictions. *Hydrology* 7 (3). <https://doi.org/10.3390/hydrology7030042>.
- Forzieri, G., Feyen, L., Rojas, R., Flörke, M., Wimmer, F., Bianchi, A., 2014. Ensemble projections of future streamflow droughts in Europe. *Hydrology and Earth System Sciences* 18 (1). <https://doi.org/10.5194/hess-18-85-2014>.
- Gandin, L. S., & Murphy, A. H. (1992). Equitable skill scores for categorical forecasts. *Monthly Weather Review*, 120(2). Doi: 10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.
- Guo, J., Huang, G., Wang, X., Li, Y., Lin, Q., 2017. Investigating future precipitation changes over China through a high-resolution regional climate model ensemble. *Earth's Future*. <https://doi.org/10.1002/2016EF000433>.
- Hartmann, H. C., Pagano, T. C., Sorooshian, S., & Bales, R. (2002). Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bulletin of the American Meteorological Society*, 83(5). Doi: 10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.
- Climate Information. (2020). Providing Climate Information and Guidance for Adaptation and Mitigation Activities, 26 Oct. 2020. climateinformation.org/. Accessed 16 May 2023.
- Jenkner, J., Frei, C., Schwiertz, C., 2008. Quantile-based short-range QPF evaluation over Switzerland. *Meteorologische Zeitschrift* 17 (6). <https://doi.org/10.1127/0941-2948/2008/0344>.
- Kavetski, D., Fenicia, F., Reichert, P., Albert, C., 2018. Signature-Domain Calibration of Hydrological Models Using Approximate Bayesian Computation: Theory and Comparison to Existing Applications. *Water Resources Research* 54 (6). <https://doi.org/10.1002/2017WR020528>.
- Krysanova, V., Donnelly, C., Gelfan, A., Gerten, D., Arheimer, B., Hattermann, F., Kundzewicz, Z.W., 2018. How the performance of hydrological models relates to credibility of projections under climate change. *Hydrological Sciences Journal* 63 (5). <https://doi.org/10.1080/02626667.2018.1446214>.
- Lee, M.H., Bae, D.H., 2016. Uncertainty Assessment of Future High and Low Flow Projections According to Climate Downscaling and Hydrological Models. *Procedia Engineering* 154. <https://doi.org/10.1016/j.proeng.2016.07.560>.
- Legates, D.R., McCabe, G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resources Research* 35 (1). <https://doi.org/10.1029/1998WR900018>.
- Lussana, C., Einar Tveito, O., Dobler, A., Tunheim, K., 2019. SeNorge 2018, daily precipitation, and temperature datasets over Norway. *Earth System Science Data* 11 (4). <https://doi.org/10.5194/essd-11-1531-2019>.

- Maidment, D. R. (1993). Handbook of Hydrology.
- Mason, I., 1989. Dependence of the Critical Success Index on sample climate and threshold probability. *Australian Meteorological Magazine* 37 (2).
- McMillan, H., Westerberg, I., Branger, F., 2017. Five guidelines for selecting hydrological signatures. *Hydrological Processes* 31 (26). <https://doi.org/10.1002/hyp.11300>.
- Mello, C.R., Vieira, N.P.A., Guzman, J.A., Viola, M.R., Beskow, S., Alvarenga, L.A., 2021. Climate change impacts on water resources of the largest hydropower plant reservoir in Southeast Brazil. *Water (switzerland)* 13 (11). <https://doi.org/10.3390/w13111560>.
- Merks, J., Photiadou, C., Ludwig, F., Arheimer, B., 2020. Comparison of open access global climate services for hydrological data. *Hydrological Sciences Journal*. <https://doi.org/10.1080/02626667.2020.1820012>.
- Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I - A discussion of principles. *Journal of Hydrology* 10 (3). [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6).
- Nkiaka, E., Taylor, A., Dougill, A. J., Antwi-Agyei, P., Fournier, N., Bosire, E. N., Konte, O., Lawal, K. A., Mutai, B., Mwangi, E., Tichehurst, H., Toure, A., & Warnaaars, T. (2019). Identifying user needs for weather and climate services to enhance resilience to climate shocks in sub-Saharan Africa. In *Environmental Research Letters* (Vol. 14, Issue 12). Institute of Physics Publishing. Doi: 10.1088/1748-9326/ab4dfe.
- Norbiato, D., Borga, M., Dinala, R., 2009. Flash flood warning in ungauged basins by use of the flash flood guidance and model-based runoff thresholds. *Meteorological Applications* 16 (1). <https://doi.org/10.1002/met.126>.
- Peirce, C.S., 1884. The numerical measure of the success of predictions. In *Science* (vol. 4, Issue 93) <https://doi.org/10.1126/science.ns-4.93.453-a>.
- Pfannerstill, M., Guse, B., Fohrer, N., 2014. Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *Journal of Hydrology* 510. <https://doi.org/10.1016/j.jhydrol.2013.12.044>.
- Raulino, J.B.S., Silveira, C.S., Lima Neto, I.E., 2021. Assessment of climate change impacts on hydrology and water quality of large semi-arid reservoirs in Brazil. *Hydrological Sciences Journal* 66 (8). <https://doi.org/10.1080/02626667.2021.1933491>.
- Reed, S., Schaake, J., Zhang, Z., 2007. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrology* 337 (3–4). <https://doi.org/10.1016/j.jhydrol.2007.02.015>.
- 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., Christensen, J.H., 2014. A framework for testing the ability of models to project climate change and its impacts. *Climatic Change* 122 (1–2). <https://doi.org/10.1007/s10584-013-0990-2>.
- Shafii, M., Tolson, B.A., 2015. Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resources Research* 51 (5). <https://doi.org/10.1002/2014WR016520>.
- SMHI (2023). "Information about Warnings." SMHIs Vädervarningar, 26 Oct. 2021, www.smhi.se/kunskapsbanken/meteorologi/varningar-och-meddelanden. Accessed 16 May 2023.
- Strömqvist, J., Arheimer, B., Dahné, J., Donnelly, C., Lindström, G., 2012. Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale. *Hydrological Sciences Journal* 57 (2). <https://doi.org/10.1080/02626667.2011.637497>.
- Teegavarapu, R.S.V., Sharma, P.J., Lal Patel, P., 2022. Frequency-based performance measure for hydrologic model evaluation. *Journal of Hydrology* 608. <https://doi.org/10.1016/j.jhydrol.2022.127583>.
- Thielen, J., Bartholmes, J., Ramos, M.H., De Roo, A., 2009. The European flood alert system - part 1: Concept and development. *Hydrology and Earth System Sciences* 13 (2). <https://doi.org/10.5194/hess-13-125-2009>.
- 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., Vaze, J., 2015. Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments. *Hydrol. Sci. J.* 60, 1184–1199. <https://doi.org/10.1080/02626667.2014.967248>.
- Törnqvist, L., Vartia, P., Vartia, Y.O., 1985. How should relative changes be measured? *American Statistician* 39 (1). <https://doi.org/10.1080/00031305.1985.10479385>.
- Vaughan, C., Dessai, S., 2014. Climate services for society: Origins, institutional arrangements, and design elements for an evaluation framework. *Wiley Interdisciplinary Reviews: Climate Change* 5 (5). <https://doi.org/10.1002/wcc.290>.
- Vincent, K., Archer, E., Henriksson, R., Pardoe, J., Mittal, N., 2020. Reflections on a key component of co-producing climate services: Defining climate metrics from user needs. *Climate Services* 20. <https://doi.org/10.1016/j.cliser.2020.100204>.
- Westerberg, I.K., Guerrero, J.L., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S., Freer, J.E., Xu, C.Y., 2011. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences* 15 (7). <https://doi.org/10.5194/hess-15-2205-2011>.
- Westerberg, I.K., Wagener, T., Coxon, G., McMillan, H.K., Castellarin, A., Montanari, A., Freer, J., 2016. Uncertainty in hydrological signatures for gauged and ungauged catchments. *Water Resources Research* 52 (3). <https://doi.org/10.1002/2015WR017635>.
- Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* 30 (1). <https://doi.org/10.3354/cr030079>.
- World Meteorological Organization (WMO). (2021). Guidelines on Seasonal Hydrological Prediction.
- Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research* 44 (9). <https://doi.org/10.1029/2007WR006716>.
- Younis, J., Anquetin, S., Thielen, J., 2008. The benefit of high-resolution operational weather forecasts for flash flood warning. *Hydrology and Earth System Sciences* 12 (4). <https://doi.org/10.5194/hess-12-1039-2008>.