Research Article

Performance of rainfall–runoff models in reproducing hydrological extremes: a case of the River Malaba sub-catchment

Ambrose Mubialiwo[1,2] · Adane Abebe[1,3] · Charles Onyutha[2]

Received: 7 November 2020 / Accepted: 17 March 2021 / Published online: 31 March 2021

Abstract

Daily River Malaba flows recorded from 1999 to 2016 were modelled using seven lumped conceptual rainfall–runoff models including AWBM, SACRAMENTO, TANK, IHACRES, SIMHYD, SMAR and HMSV. Optimal parameters of each model were obtained using an automatic calibration strategy. Mismatches between observed and modelled flows were assessed using a total of nine "goodness-of-fit" metrics. Capacity of the models to reproduce historical hydrological extremes was assessed through comparison of amplitude–duration–frequency (ADF) relationships or curves constructed based on observed and modelled flow quantiles. Generally, most of the hydrological models performed better for high than low flows. ADF curves of both high and low flows for various return periods from 5 to 100 years were well reproduced by AWBM, SAC, TANK and HMSV. ADF curves for high and low flows were poorly reproduced by SIMHYD and SMAR, respectively. Overall, AWBM performed slightly better than other models if both high and low flows are to be considered simultaneously. The deviations of these models were larger for high than low return periods. It was found that the choice of a "goodness-of-fit" metric affects how model performance can be judged. Results from this study also show that when focusing on hydrological extremes, uncertainty due to the choice of a particular model should be taken into consideration. Insights from this study provide relevant information for planning of risk-based water resources applications.

Keywords Rainfall–runoff model · "Goodness-of-fit" measures · River Malaba sub-catchment · Uganda · Hydrological extremes · High flows · Low flows

1 Introduction

Heavy rainfall events have increasingly been experienced in various regions, while other areas of the world receive light or no rains as reported by the Intergovernmental Panel on Climate Change IPCC [1]. This results in floods, landslide, drought, hurricanes, hence, causing distressing damages and losses to public life and property. The study area of the present work (River Malaba sub-catchment) is not exceptional to these extreme weather events. River Malaba sub-catchment in Eastern Uganda has fertile soils which support agriculture and livestock grazing. In addition, River Malaba provides for other economic activities including fishing and fish farming. The River Malaba sub-catchment within the Mpologoma catchment has been affected by rainfall-induced landslides in the highland districts of Manafwa and Bududa. Within the same area, annually, disastrous floods have been experienced in the low-lying districts of Butaleja and Manafwa [2, 3]. These

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42452-021-04514-7.

✉ Ambrose Mubialiwo, mambroze@gmail.com | [1]Africa Center of Excellence for Water Management, Addis Ababa University, P.O. Box 1176, Addis Ababa, Ethiopia. [2]Department of Civil and Environmental Engineering, Kyambogo University, P.O. Box 1, Kyambogo, Kampala, Uganda. [3]Department of Water Resources and Irrigation Engineering, Arba Minch University, PO Box 21, Arba Minch, Ethiopia.

disasters could be linked to impacts of human factors on the sub-catchment hydrology. Besides, the study area hydrometeorology could be influenced by climate variability. Consequently, the possibility of Uganda "Vision 2040" targets has been compromised as noted by the Ministry of Water and Environment, MWE [3, 4]. Some of the dramatic events include (1) the floods of December 2019 with at least 4 death and over 2,000 people displaced [5]; (2) the October 2018 severe floods and landslides in Bududa, displacing 858 people, 51 death and a total of 12,000 people affected [6]; (3) the severe landslides of March 2010 killed over 400 people, displacing 5,000 people in Bududa district [7], and over 33,000 households affected in Butaleja [8]. These events tend to be punctuated by long dry spells, for instance, after the floods and landslide in 2014, there was a long dry spell [9]. In May 2012, flooding resulted in bursting of River Malaba banks affecting over 200 families in Malaba town council and Osukuru sub-county, Tororo district. These events destroyed acres of food crops and resulted in at least 4 death [10]. Furthermore, the report by Reliefweb [11] categorised the Elgon sub-region (where the study area is located) as one of the most affected areas by the 2007 floods that resulted in destruction of several infrastructure, particularly roads, bridges and buildings, killed human beings, and wrecked crops. This event affected almost the entire Uganda.

To facilitate predictive planning and operation of risk-based water resources management within the study area, where rain-fed agriculture and livestock grazing are major economic activities, there is need to perform hydrodynamic modelling. Hydrodynamic modelling of weather events such as floods, necessitates understanding the hydrological processes with focus on the extremes. Hydrological modelling can be performed using either lumped conceptual, semi-distributed or distributed models. Distributed models consider the spatial distribution of rainfall, evapotranspiration and watershed characteristics at a resolution normally selected by the modeller to reflect the spatio-temporal variability of runoff. Some of the distributed (physically based) hydrological models include the Gridded Surface Subsurface Hydrologic Analysis (GSSHA) [12], Systeme Hydrologique Europeen, "SHE" [13], European hydrological system model (MIKE-SHE) [14], modular modelling system (MMS) [15]. Some models are not physically based but rather semi-distributed, e.g. Soil and Water Assessment Tool (SWAT) [16] which is operated on hydrological response unit (HRU) and necessitates parameter calibration. Whereas the physically based (distributed) models have better computational capacity and are robust, with well-implemented numerical methods, their application particularly, in rainfall–runoff simulations is still inadequate. Huge amount of data inputs may undesirably result in a more complex model which may lead to high prediction uncertainty especially if a model has large number of parameters [17].

On the other hand, lumped conceptual models are based on average spatial characteristics of the system, whose basis is to simulate flow at the outlet of the catchment [18, 19]. Examples of lumped conceptual models include the Australian water balance model (AWBM) [20], Sacramento (SAC) [21], TANK [22], Identification of Unit Hydrographs and Component Flows from Rainfall, Evaporation and Stream-flow data (IHACRES) [23, 24], SIMHYD [25], soil moisture accounting and routing (SMAR) [26], hydrological model focusing on sub-flows' variation (HMSV) [27], Pitman model [28], hydrological engineering centre-hydrological modelling system (HEC-HMS) model [29], hydrologiska byråns vattenbalansavdelning (HBV) model [30]. The high capacity to simulate runoff with easy to use methods, and the minimum data requirement, has made the lumped conceptual models prevalent to distributed models for rainfall–runoff modelling. Lumped conceptual models have widely been applied [31]. Rainfall–runoff modelling can be done based on event-based or continuous approach. Recently, there is a transition from event-based continuous hydrological modelling. For instance, Grimaldi et al. [32] applied the Continuous Simulation Model for Small Ungauged Basins (COSMO4SUB) to ungauged catchment and the results were comparable with the event-based approach. Similarly, several studies have reported superior performance of Artificial Neural Network (ANN) models [31, 33, 34]

Within Uganda (where the study area is located), several studies [35–37] have applied lumped conceptual hydrological models. However, it was noticeable that most studies applied a single hydrological model except for Onyutha et al. [36] that compared the performance of seven lumped conceptual models in the simulation of daily River Kafu flows. However, this study was conducted in western Uganda and not in the eastern part of Uganda (where the study area is located). Furthermore, most of these studies applied either one or very few (maximum of 3) "goodness-of-fit" measures. The selection and application of a particular model and "goodness-of-fit" measure from the several can result in a huge bias while concluding on the worthiness of the obtained model results [38]. This could be attributed to the varying structures and parameters amongst different model [39]. In addition, a particular "goodness-of-fit" measure may not provide information on some analyses components such as model residuals, making them inadequate in assessing model performance [38, 40].

Reliable hydrological modelling results are vital for decision makers to avoid profligate expenditures resulting from wrongly informed predictive planning. Prior to conducting this research article, studies conducted to evaluate several lumped conceptual models' performance based on multiple "goodness-of-fit" statistics to simulate hydrological extremes in River Malaba sub-catchment were lacking. Therefore, to address the above research gap, this study evaluated the performance of seven rainfall–runoff models in simulating hydrological extremes of River Malaba sub-catchment while assessing model performance using nine "goodness-of-fit" metrics.

2 Study area, data and models

2.1 The study area

The River Malaba sub-catchment (Fig. 1) has a drainage area of approximately 3500 km^2. The sub-catchment is part of the Mpologoma catchment under the Kyoga Water Management Zone (KWMZ), stretching between latitudes 0° 19′ N and 1° 07′ N and longitudes 33° 37′ E and 34° 37′ E. The sub-catchment drainage area is transboundary, shared between Uganda (about 69% or 2395 km^2) and Kenya (around 31% or 1100 km^2). River Malaba sub-catchment is comprised of River Malaba, formed by two tributaries of Lwakhakha and Makalisi which are later joined by the Lumbaka/Kibimba tributary.

The basin has a flow outlet at Budumba with latitude = 0° 49′ N and longitude = 33° 47′ N. The River Malaba sub-catchment extends from Mount Elgon at about 4320 m above sea level traversing through the districts of Bududa, Manafwa, Tororo, Butaleja, and discharging into River Mpologoma at the shores of Lake Kyoga at about 1000 m above sea level. The core land use in the sub-catchment is rain-fed subsistence agriculture employing approximately 85% of the population. Practically, the region beyond the Mount Elgon forested area consists of agricultural and grassland, fallow land and isolated woodlots. Land use changes in the sub-catchment ecosystem have adversely changed the river Malaba hydrological flow regimes [41]. The major soils types in the sub-catchment are Petric Plinthosols and Gleysols. The other types include Lixic ferralsols, Acric ferrasls and Nitisols [42, 43]. The rainfall over the study area occurs in two seasons with the first and more intense from March to May (MAM), while the second and highly variable occurs between October and December (OND). The basin receives an average annual rainfall of about 1375 mm, though the districts of Bududa and Manafwa receive slightly higher rains (on average 1800 mm per annum). The basin climate is partly affected by the existence of large water bodies (such as Lake Victoria and Lake Kyoga) and the mountain Elgon slope breezes that tend to affect the afternoon convection [44]. It is clear that the altitude is higher in the highlands of Mount Elgon. The sub-catchment wettest and driest periods occur from March to April and June to October, respectively, with an average temperature range of 15.8 to 30.6 °C [45].

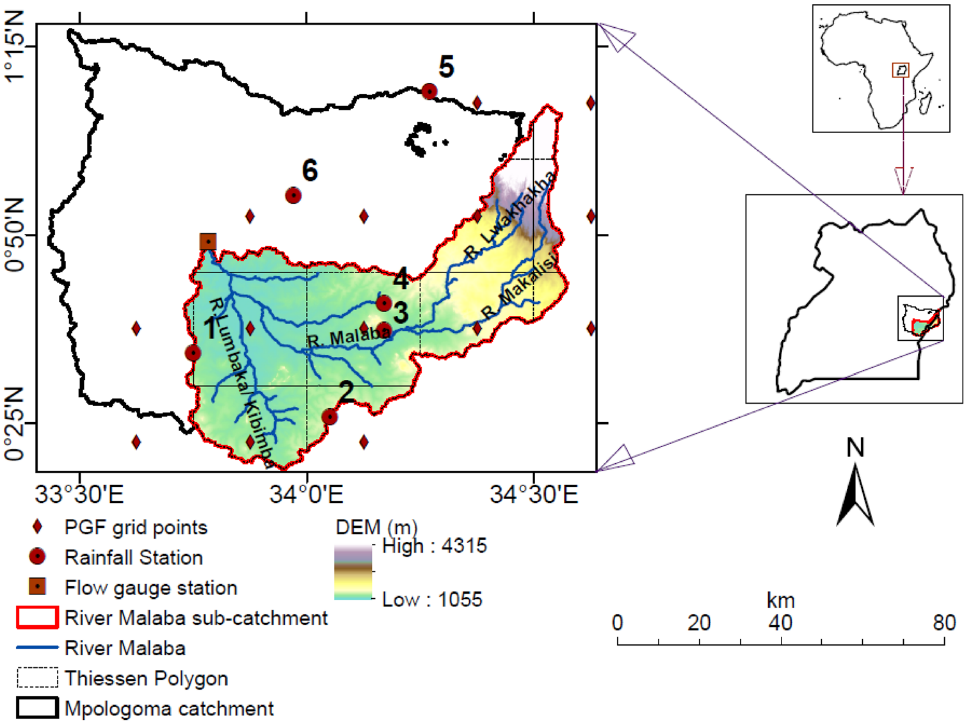


Fig. 1 Location of River Malaba sub-catchment. The background map is the Digital Elevation Model (DEM) acquired via http://srtm.csi.cgiar.org/ (accessed: 08 February, 2021)

2.2 Data

2.2.1 Rainfall and potential evapotranspiration data

As noted in the previous study by Mubialiwo et al. [46], the study area has poor distribution of meteorological stations, with the existing ones having few data available in the recent years in addition to uncertain and questionable quality [47, 48]. Therefore, the daily precipitation, minimum (T_{min}) and maximum (T_{max}) temperature data in a gridded (0.25° × 0.25°) form of the Princeton Global Forcing (PGF) [49] were obtained from http://hydrology.princeton.edu/data/pgf/v3/0.25deg/ (accessed: 07 February 2021). The PGF data was previously applied by Zhang et al. [50] to evaluate performance of four models in streamflow simulation. Like any other reanalysis datasets, PGF has a few shortcomings including bias and random errors which are attributed to numerous factors such as sampling rate, inadequate spreading of sensors and uncertainties in the rainfall retrieval algorithms [51–53]. Furthermore, the study by Mubialiwo et al. [46] revealed that the PGF data overestimated and/or underestimated the oscillation highs and lows from observed rainfall over the Mpologoma catchment (where the study area is located). This evidenced existence of bias in the PGF data. Therefore, prior to the use of PGF data, it was thought vital to perform bias correction [53], using observed rainfall data from the six stations in and around the study area (Table 1). For each station, the annual rainfall statistical metrics (coefficient of variation (C_{v}), skewness (C_{s}), actual excess kurtosis (C_{k}) and annual mean rainfall (AMR)) were determined. From Table 1, the values of C_{v} varied from 0.12 to 0.28, which represent a modest variability on a year to year basis. Under ideal situation, the values of C_{s} and C_{k} are expected to be equal to zero for a normal distribution. However, from Table 1, data at the 6 stations are, on average, somewhat positively skewed ($Cs = 0.54$) and platykurtic ($C_{\text{k}} = -0.04$). The highest values of C_{k} and C_{s} were obtained at Sukulu VTRO, while Butaleja prison exhibited the smallest values. The AMR varied between 1020.79 and 1640.66 mm. The highest AMR was observed at Dabani catholic while Butaleja prison exhibited the lowest value AMR value.

With the exception of station 4 (Tororo), the remaining stations have data ending in the 1980s. This is attributed to the non-functionality of weather stations subsequent to the civil war that started in 1981, resulting in the breakdown of many measuring stations across Uganda according to the Japan International Cooperation Agency [54].

The missing values in the observed rainfall data were in-filled using the Inverse Distance Weighted (IDW) interpolation [55], a technique previously applied by Mubialiwo et al. [46]. The bias correction was done using the simple multiplicative bias correction method [56]. This method was adopted because of the poor distribution of the rain gauge networks within the study area. The study by Tian et al. [57] compared the two common error models (additive and multiplicative) and recommended the use of multiplicative bias correlation method for bias removal. The monthly bias correction factor B_{cf} was used to adjust the daily PGF precipitation data. The B_{cf} was obtained as follows:

$$B_{\text{cf}} = \frac{P_{\text{obs},i}}{P_{\text{PGF},i}} \tag{1}$$

where $P_{\text{obs},i}$ and $P_{\text{PGF},i}$ are station-based and PGF-based rainfall data at monthly time scale, respectively.

One bias factor was calculated for each month in a year and applied to daily data. The study by Saber and Yilmaz [58] applied a similar approach with monthly bias factors used to correct hourly data.

The daily bias corrected PGF rainfall was computed as follows:

$$P_{\text{PGF\_after}(x,y),T_i} = B_{\text{cf}} \times P_{\text{PGF\_before}(x,y),T_i} \tag{2}$$

where $P_{\text{PGF\_before}(x,y),T_i}$ and $P_{\text{PGF\_after}(x,y),T_i}$ are the PGF products for day T_i at grid (x, y) before and after bias removal, respectively.

Monthly bias factors were computed considering the station(s) that is(are) located in a particular grid cell or the closest station(s). It is noticeable from Table 1 that only the

Table 1 Rainfall station, their coordinates, data record period and statistical metrics

S/N	Station name	Coordinate		Data record		Statistical metric			
		Lat. [°]	Long. [°]	From	To	C_{k} [-]	C_{s} [-]	C_{v} [-]	AMR [mm]
1	Bugiri	0.57	33.75	1948	1975	0.2	0.63	0.26	1389.21
2	Dabani Catholic	0.43	34.05	1948	1983	−0.52	0.66	0.17	1640.66
3	Sukulu VTRO	0.62	34.17	1963	1987	0.29	0.86	0.19	1543.84
4	Tororo	0.68	34.17	1981	2016	0.07	0.34	0.12	1506.39
5	Bugusege Coffee	1.15	34.27	1948	1982	-0.21	0.73	0.17	1533.22
6	Butaleja Prison	0.92	33.97	1948	1982	-0.07	0.04	0.28	1020.79

Lat latitude, *Long* longitude

Tororo station had data of corresponding period to the PGF-based rainfall.

The computed bias factor at Tororo station did not exhibit noticeable variability for the period before and after 1983. Therefore, since all stations were from the same region, they were presumed that they exhibit minimal variation in the bias factors. Consequently, the available periods of station-based data (Table 1) were used to compute monthly bias correction factors, and applied to daily PGF data from 1999 to 2016. This approach was previously used by Piani et al. [59] on statistical bias correction for daily precipitation in regional climate models over Europe. In the study [59], bias correction factors were calculated using data from 1961 to 1970 and applied to data of a different period (from 1991 to 2000).

In the current study, the catchment-wide rainfall averages were obtained using the Thiessen polygon [60] (Fig. 1) constructed from the 14 grid points. It is noticeable that 5 grid points fall inside the study area while an additional nine are situated outside but very close (on average less than 10 km from the sub-catchment boundary). All the 14 grids were used in obtaining the average rainfall over the sub-catchment. The sub-catchment daily rainfall time series are shown in Fig. 2a.

The sub-catchment *PET* (mm/day) (Fig. 2b) was computed using the Hargreaves formula [61, 62] (Eq. 3). The method requires mainly minimum and maximum temperatures as the key inputs. The Hargreaves method, was recently applied by Mubialiwo et al. [46] in Mpologoma catchment and Onyutha et al. [63] in KWMZ, where the study area is located.

$$\mathrm{PET} = 0.0023\left(T_{\mathrm{mean}} + 17.8\right)\left(T_{\max} - T_{\min}\right)^{0.5}\mathrm{Ra} \tag{3}$$

where R_a measures the incoming extra-terrestrial radiation (in W/m^2), estimated based on the location's latitude and the calendar day of the year, T_{mean} in °C is the mean temperature.

2.2.2 Observed flow data

Average daily River Malaba flow series measured at Budumba gauge station (with station ID 82,217, latitude = 0º49′N and longitude = 33° 47′ E) (Fig. 1) from 1999 to 2016 was obtained from the Uganda Ministry of Water and Environment (MWE). The data was checked for quality assurance using visual inspection and statistical methods to ensure only satisfactory and quality data is used in the research. Only eighteen (18) years of recent flow data (from 1999 to 2016) for River Malaba were used (Fig. 2c). Their selection was linked to the anticipated studies of flood analysis in the study area (requiring recent information) that will be based on output from this study. Nevertheless, the eighteen years of data were considered very sufficient because longer calibration data series do not certainly yield better model performance [64]. The study by Li et al. [64] revealed that only eight years of data are adequate to get a stable approximation of model performance. Since rainfall–runoff modelling required the same period of rainfall, PET and flow data, the period from 1999 to 2016 is considered here. This is because PGF-based data currently ends in 2016.

Similar to the observed rainfall data, statistical metrics in the PGF-based rainfall, PET and observed flow were computed as shown in Table 2. Rainfall, PET and observed flow were negatively skewed (negative C_s values) and platykurtic (negative C_k values) compared with the normal distribution, for which it is anticipated that $C_s = 0$ and $C_k = 0$. However, their C_v values indicated that there is slight variation on a year to year basis. The average annual rainfall

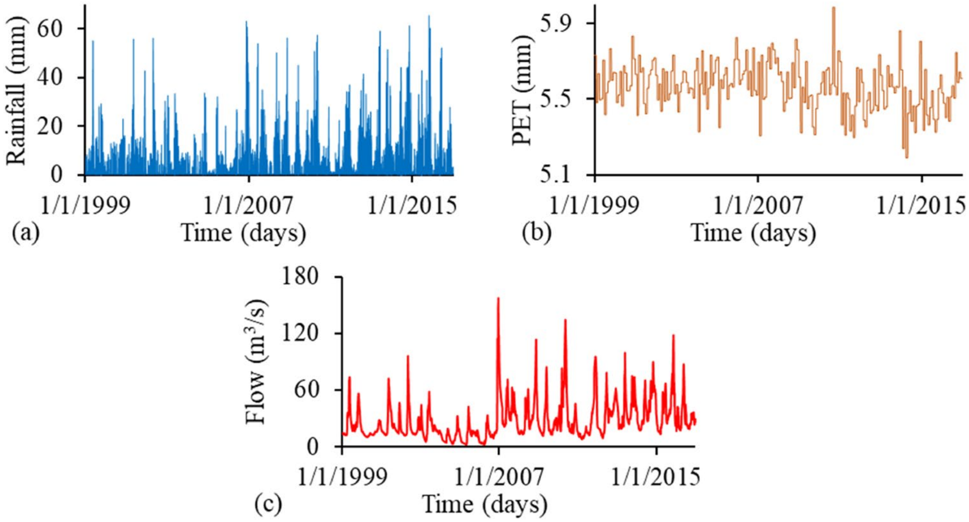


Fig. 2 Time series of **a** daily rainfall, **b** daily PET and **c** daily observed flow used in rainfall–runoff modelling

Table 2 Statistical metrics in PGF rainfall, PET and observed flow

Data	Statistical metric			
	C_k [–]	C_s [–]	C_v [–]	AMV
PGF rainfall	−0.7	−0.42	0.29	1125.25 mm
PET	−0.9	−0.65	0.01	5.57 mm
Observed flow	−1.04	−0.22	0.34	10,436.15 m^3s^{-1}

AMV annual mean value

was 1125.25 mm with mean annual PET of 5.57 mm and annual mean flow of 10,436.15 m^3/s.

3 Methodology

3.1 *Rainfall*-runoff modelling

This study used seven lumped conceptual rainfall–runoff models. Of the seven models, the six are internationally well-known including the Australian Water Balance Model AWBM [20], SACRAMENTO [21], TANK [22], Identification of Unit Hydrographs and Component Flows from Rainfall, Evaporation and Stream (IHACRES) flow data [23, 24], SIMHYD [25], and Soil Moisture Accounting and Routing SMAR [26]. These six models were obtained from the "eWater Toolkit" of the Cooperative Research Centre for Catchment Hydrology in Australia via http://www.toolkit.net.au/ (accessed: 08 February 2021). The seventh model or Hydrological Model focusing on Sub-flows Variation HMSV [27] was accessed freely via https://sites.google.com/site/conyutha/tools-to-download (accessed: 10 February 2021). These models were selected because they (1) are freely available online and (2) were found to be robust for rainfall–runoff modelling under various climatic conditions as demonstrated in several recent studies [27, 36, 65–69].

Here is a brief mention of each model's parameters. Detailed description of the models and their parameters including sensitivity analyses are included in the Supplementary Material (Sub-Sect. M1.1–M1.7, Table M1 and Figs. M9–M11). AWBM has 8 parameters, but three of them (Baseflow index, baseflow recession constant, surface flow recession constant) are considered the major ones. SAC model has 17 parameters, with three designed for direct runoff simulation, other 3 for water capacity in upper zone, 2 for percolation into lower zones, while the remaining 9 are designed for water capacity in the lower zone. TANK model has 18 parameters categorised under 7 classes. The model has 3 major parameters (i.e. water levels in the tanks, height of outlets at tanks and runoff coefficients. IHACRES model has 11 parameters. SIMHYD model has 9 parameters, with 4 sensitive ones (i.e. infiltration coefficient and shape, interflow coefficient, and base flow coefficient). SMAR model has five water balance parameters and 4 routing parameters. HMSV has total of 10 parameters. Four parameters are for baseflow, 2 for interflow and 4 for overland flow simulation. A combination of the 10 parameters is used to calibrate the full model to simulate the total runoff. Higher values of recession constants imply delayed contributions of different components to the total runoff. Higher values of initial soil moisture storage indicate faster contribution of runoff from the catchment.

Modelling was done using meteorological data (rainfall and potential evapotranspiration) as described in Sect. 2.2.1. Prior to inputting the data into the models, it was converted into formats required by each of the seven rainfall–runoff models. The Initial model parameters for each model as provided in Podger [70] (for AWBM, TANK, SAC, SIMHYD and SMAR), Croke et al. [23] (for IHACRES), and Onyutha [27] (for HMSV) were set and the model ran to generate outputs. Sensitivity analysis on the model parameters was done prior to calibration. This was done to establish the contribution of a particular parameter variation to model output in order to identify which parameters have a great or less impact on the model response. This study adopted the use of local sensitivity analysis (LSA) method because it is simple, fast and can yield results similar to the global sensitivity analysis (GSA) [71]. The LSA method focuses of the impact on model output caused by a single parameter while other parameters are fixed. The corresponding Nash–Sutcliffe efficiency (NSE) [72] values were obtained and NSE curves plotted for each parameter.

The model parameters were calibrated using the observed river flow data (sub-Sect. 2.2.2) from 01/01/1999 to 31/12/2009, until there was a reasonable match between the simulated and observed flow. Model calibration can be based on manual or automatic strategy. With manual calibration, there is a trial and error adjustment of parameters based on the modellers' visual inspection of simulated and observed values, making it is very difficult to yield hydrologically sound results, and the process is tedious, time-consuming, especially for models with many parameters [73]. When comparing models, manual calibration also yields subjective results. Thus, in this study, calibration was done using the automatic calibration strategy with model parameters automatically adjusted following systematic search algorithms based on the set objective function. In the rainfall–runoff models' frameworks, Nash–Sutcliffe efficiency (NSE) [72] was used as the optimisation objective function, while the models' performance was further assessed based on other eight "goodness-of-fit "metrics as described shortly next. Five models (AWBM, SACRAMENTO, TANK, SIMHYD and SMAR) were calibrated based on shuffled complex algorithm (SCE) [74]. The calibration of IHACRES followed the approach described by Croke et al. [23]. Similarly, the HMSV was automatically calibrated based on the Generalised Likelihood Uncertainty Estimation (GLUE) developed by Beven and Binley [75]. HMSV framework allows a "step-wise" calibration strategy, first by calibrating parameters based on each of the sub-flows (baseflow, interflow and overland flow). The full model run is then performed combining parameters from the sub-flow models. This approach allows modelling of quick flows focusing on high flows while baseflow considers variation of the low flows. The models were validated using the observed river flow data for a record outside the calibration period (from 01/01/2010 to 31/12/2016). During validation, the same parameters and corresponding values used for calibration were maintained.

The models performance were evaluated based on nine widely used statistical indicators, including**,** Nash–Sutcliffe efficiency (NSE) [72]**,** model average bias (MAB) (%), Ratio of Root Mean squared error to the maximum event (RRM), relative efficiency (R_e) [76], index of agreement (I_a) [77], coefficient of determination (R) [78], mean absolute error (MAE), coefficient of model accuracy (CMA) [38] and Kling–Gupta efficiency (KGE) [79] as shown in Eqs. (4) to (12). For all the seven model, NSE was used as the optimisation objective function for calibration. Consider Q_{obs}, Q_{sim}, $\overline{Q}_{\text{obs}}$, and $\overline{Q}_{\text{sim}}$ as the observed, modelled, mean of observed and mean modelled flows, respectively. Furthermore, take $Q_{\max}$, σ_{obs}, σ_{sim} and n, to denote the maximum observed flows, standard deviation in observed, standard deviation in modelled flows and sample size, respectively. Lastly, consider r as the rank-based Spearman correlation coefficient between Q_{obs} and Q_{sim}. The various "goodness-of-fit" metrics were computed using:

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n} \left(Q_{\text{sim},i} - Q_{\text{obs},i}\right)^2}{\sum_{i=1}^{n} \left(Q_{\text{obs},i} - \overline{Q}_{\text{obs}}\right)^2} \tag{4}$$

$$\text{MAB} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Q_{\text{sim},i} - Q_{\text{obs},i}}{Q_{\text{obs},i}} \times 100 \right) \tag{5}$$

$$\text{RRM} = \frac{1}{Q_{\max}} \left(\frac{1}{n} \sum_{i=1}^{n} \left(Q_{\text{sim},i} - Q_{\text{obs},i}\right)^2 \right)^{0.5} \tag{6}$$

$$R_{\text{e}} = 1 - \frac{\sum_{i=1}^{n} \left|Q_{\text{sim},i} - Q_{\text{obs},i}\right|^{f_p}}{\sum_{i=1}^{n} \left|Q_{\text{obs},i} - \overline{Q}_{\text{obs}}\right|^{f_p}} \tag{7}$$

$$I_{\text{a}} = 1 - \frac{\sum_{i=1}^{n} \left|Q_{\text{obs},i} - Q_{\text{sim},i}\right|^{f_p}}{\sum_{i=1}^{n} \left(\left|Q_{\text{sim},i} - \overline{Q}_{\text{obs}}\right| + \left|Q_{\text{obs},i} - \overline{Q}_{\text{obs}}\right| \right)^{f_p}} \tag{8}$$

$$R = \frac{\sum_{i=1}^{n} \left(Q_{\text{obs},i} - \overline{Q}_{\text{obs}}\right)\left(Q_{\text{sim},i} - \overline{Q}_{\text{sim}}\right)}{\sqrt{\sum_{i=1}^{n} \left(Q_{\text{obs},i} - \overline{Q}_{\text{obs}}\right)^2} \times \sqrt{\sum_{i=1}^{n} \left(Q_{\text{sim},i} - \overline{Q}_{\text{sim}}\right)^2}} \tag{9}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} \left(Q_{\text{obs},i} - Q_{\text{sim},i}\right)}{n} \tag{10}$$

$$\text{CMA} = |r| \times \frac{\min\left(\overline{Q}_{\text{sim}}^2, \overline{Q}_{\text{obs}}^2\right)}{\max\left(\overline{Q}_{\text{sim}}^2, \overline{Q}_{\text{obs}}^2\right)} \times \frac{\sum_{i=1}^{n} \min\left(\left(Q_{\text{obs},i} - 3\overline{Q}_{\text{obs}}\right)^2, \left(Q_{\text{sim},i} - 3\overline{Q}_{\text{obs}}\right)^2 \right)}{\sum_{i=1}^{n} \max\left(\left(Q_{\text{obs},i} - 3\overline{Q}_{\text{obs}}\right)^2, \left(Q_{\text{sim},i} - 3\overline{Q}_{\text{obs}}\right)^2 \right)} \tag{11}$$

$$\text{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\text{sim}}}{\sigma_{\text{obs}}} - 1 \right)^2 + \left(\frac{\overline{Q}_{\text{sim}}}{\overline{Q}_{\text{obs}}} - 1 \right)^2} \tag{12}$$

where f_p is the power factor.

NSE demonstrates how fit the simulation mimics the observation, and it varies between –∞ and 1.0, with the value of 1.0 denoting a perfect match. A value greater than 0.5 for NSE is considered acceptable [80]. MAB and MAE denote the bias and mean error magnitude between modelled and observed values. The value of MAB and MAE equal to 0 signifies an unbiased model [81]. RRM was considered in this study instead of the root mean squared error because it is unitless and has a small value, making it suitable to compare with MAB. The values of R_e and I_a range between 0 and 1. A value of 1 designates perfect match between simulated and observed, while 0 implies total divergence. I_a is known for its insensitivity to systematic model overestimation and underestimations [77]. In this study, the values of $f_p = 1$ were used to balance between high flows and low flows [76, 81]. However, the value of f_p can be greater than 1 if the model evaluation is focused on high flows. R quantifies how much the observed dispersion is explained by the simulation. R ranges between 0 and 1.

A value of 1 means perfect prediction of the modelled dispersion to the observed, and 0 indicates no correlation. This indicator has a weakness, as it only measures the dispersion, hence it is not recommended to be used solely [76]. The values of CMA range from 0 to 1 with CMA equal to 1 signifying a perfect model, while CMA equal to 0 indicates that the model simulations do not match the observations [38]. Similar to NSE, a value of KGE equal to one denotes a perfect agreement between observations and simulation [82].

3.1.1 Comparison of the model performance based on the "nine goodness-of-fit" measures

For the three periods of calibration (01/01/1999 to 31/12/2009), validation (01/01/2010 to 31/12/2016), and entire data period (01/01/1999 to 31/12/2016), models were ranked from 1 to 7 denoting the best and worst performance, respectively. Ranking was initially done based on individual "goodness-of-fit" statistic. For instance, a model with the highest (lowest) KGE was allocated rank 1 (7), indicating best (worst) model. This was done for all the nine metrics. The summation of ranks for the three periods (calibration, validation and entire period) were obtained for the nine metrics (NSE, MAB, MAE, RRM, $R_{e,}$ $I_{a,}$ R, CMA, and KGE). The model with the smallest (largest) sum of ranks for each individual metric was considered the best (worst).

3.2 Amplitude–duration frequency analyses

To facilitate planning, design and operation of various water management projects against weather events (such as floods, drought) and/or to reduce human and economic losses, there is need to understand the hydrological extremes and their frequency at selected temporal scales [83, 84]. This can be through extreme value analysis on the hydrological time series for different aggregation levels, which generates Amplitude–Duration–Frequency (ADF) relationships. ADF for discharge and rainfall are called Flow–Duration Frequency (FDF) and Intensity–Duration Frequency (IDF), respectively. In this study, only the FDF are developed for both observed and modelled flow for each of the seven hydrological models. Aggregation simply converts fine resolution data into coarser time units, e.g. from daily to monthly, which hydrologically implies representing a delayed response of a watershed [85]. Here, aggregated hydrological time series were obtained by use of *n*-day moving window. Aggregation levels are selected with the consideration of appropriate water resources management aspects such as floods and drought in agriculture and domestic water supply. The considered aggregation levels ranged from 1 day to 3 months (1, 3, 5, 7, 10, 30, 60, 90 days) for high flows and 1 day to 1 year (1, 10, 30, 90, 150, 180, 240, 365 days) for low flows. The adopted aggregation levels were applied and/or recommended in previous studies [81, 86, 87].

While analyses of high flows were based on aggregation of original flow series (X), for low flows, the original flows were transformed as $(1/X)$. The transformed flow series now follows exponential or Generalised Pareto Distribution (GDP) [88] as an alternative of the Weibull or Fréchet distribution. Besides, the transformation makes it possible to perform extreme value analyses on both high and low flows following the same approach. Taking x_{t}, β and λ as the threshold, scale and shape parameters, respectively, and considering $1/X$ as H, calibration of the exponential extreme value distribution, above the defined threshold h_{t}, follows the following expressions as clearly described by Onyutha [85].

$$G(h) = P\left\{H \le h \middle| H \ge h_t\right\} = 1 - \exp\left\{-\frac{\left(h - h_t\right)}{\beta}\right\}. \tag{13}$$

Assuming $x_{\mathrm{t}} = 1/h_{\mathrm{t}}$, Eq. (13) becomes:

$$\begin{aligned} G(h) &= P\left\{X \le x \middle| X \ge x_{\mathrm{t}}\right\} = P\left\{H \le h \middle| H \ge h_{\mathrm{t}}\right\} \\ G(h) &= 1 - P\left\{H \le h \middle| H \ge h_{\mathrm{t}}\right\} \end{aligned} \tag{14}$$

$$\begin{aligned} G(h) &= \exp\left(-\frac{h - h_{\mathrm{t}}}{\beta}\right) = \exp\left(-\frac{x_{\mathrm{t}} - x}{x_{\mathrm{t}} x \beta}\right) \\ G(h) &= P\left\{X \le x \middle| X \le x_{\mathrm{t}}\right\} = \exp\left(-\frac{x_{\mathrm{t}} - x}{x_{\mathrm{t}} x \beta}\right). \end{aligned} \tag{15}$$

Considering the values lower than x_{t}, Eq. (15) matches the Fréchet distribution $G(x) = \exp\left(-x^{-\kappa}/\beta\right)$ where $\kappa = 1$.

The subsequent step after aggregation of the time series was the extraction of independent hydrological extreme events. This was done for each aggregation level, for both high and low flows. Two main approaches exist for extracting extreme hydrological events and these include the Peak-Over-Threshold (POT) and the Annual Maxima Method (AMM) techniques [89, 90]. While the AMM is simple and generates extreme events that have high independence, the number of events can be few particularly for the short data record length as they generate only one event per year. On the other hand, the POT technique generates a satisfactory number of extreme events above the set threshold [91]. Nevertheless, closely successive flood peaks could actually be a one flood, because the damage results from the highest and the related peaks may only have indirect contribution effects [90]. In this study, both the AMM and POT approaches were used. The AMM-generated independent events were used to compare the model performance in reproducing annual maxima and minima flows, while the POT-generated independent events were used for frequency analysis of hydrological extremes (in construction of FDF curves).

Several tools for extracting POT values exist including the WETSPRO: Water Engineering Time Series Processing tool [92], and Frequency Analyses considering Non-Stationarity (FAN-Stat) [93]. The study by Mubialiwo et al. [46] revealed the presence of trends and shifts in the rainfall and potential evapotranspiration over the study area. Since the same meteorological datasets (rainfall and potential evapotranspiration) were used in this study as inputs in the hydrological models, the generated flows were presumed to exhibit non-stationarity. Therefore, the FAN-Stat tool that considers non-stationarity was adopted to obtain the independent POT value. The FAN-Stat tool was downloaded freely online via https://sites.google.com/site/conyutha/tools-to-download (accessed: 10 February 2021). Using the extracted POT events, the extreme value distribution (EVD) was fitted to the independent extreme high and low flow events. According to Segers [94], in extreme value theory, a conditional probability distribution of independent extreme events follows a Generalised Pareto Distribution (GPD), if only values above an appropriately high threshold x_t are used such that:

$$G(x) = 1 - \exp\left\{ -\frac{(x - x_t)}{\beta} \right\} \text{ for } \lambda = 0 \tag{16}$$

and

$$G(x) = 1 - \left\{ 1 + \lambda \frac{(x - x_t)}{\beta} \right\}^{-1/\lambda} \text{ for } \lambda \neq 0 \tag{17}$$

for values of $\lambda = 0$, the generated shape of the distribution tail is "normal" but when $\lambda > 0$ or $\lambda < 0$, the tail is heavy or light, respectively. For normal tail ($\lambda = 0$), the GPD matches the exponential distribution.

Following the above concept, the distribution tail analysis in River Malaba sub-catchment is done for the high and low extreme river flow events. The weighted linear regression method recently applied by Baig et al. [95] was used to determine the EVD parameters. To compute the EVD parameters in Eq. (16), exponential quantile plot with $-\ln[1 - G(x)]$ in the abscissa and x in the ordinate is adopted. The distribution assumes a straight line with a slope equal to β that can be computed using Eq. (18) by implementing the weighting factors suggested in Hill [96].

$$\beta_\mathrm{t} = \frac{1}{t - 1}\left[\sum_{i=1}^{t-1} (x_\mathrm{t}) \right] - x_\mathrm{t} \tag{18}$$

where t denotes the number of POT events above the selected threshold x_t.

On the other hand, for parameters of a GPD as in Eq. (17), the plot with $-\ln[1 - G(x)]$ in the abscissa and x in the ordinate is used. The GPD appears as a line and the slope approximated to λ.The slope value of β in the GPD expression (Eq. 17) can be computed using Eq. (19), and the shape parameter λ can be estimated by the least square weighted linear regression assuming the weights suggested in Hill [96] as shown in Eq. (20).

$$\beta_\mathrm{t} = \lambda_\mathrm{t} \times x_\mathrm{t} \tag{19}$$

$$\lambda_t = \frac{1}{t - 1}\left[\sum_{i=1}^{t-1} \ln(x_i) \right] - \ln(x_t). \tag{20}$$

The finest value of x_t is selected from the exponential Q–Q plot at the point with minimum mean squared error (MSE). The MSE of the respective weighted linear regressions in EVD and GPD can be obtained from Eqs. (21, 22), respectively.

$$\mathrm{MSE}_\mathrm{t} = \frac{1}{t - 1}\left[\sum_{i=1}^{t-1} \left(-\ln\left(\frac{i}{t}\right) \right)^{-1} . \left(x_i - x_\mathrm{t} - \beta_\mathrm{t} \ln\left(\frac{t}{i}\right) \right)^2 \right] \tag{21}$$

$$\mathrm{MSE}_t = \frac{1}{t - 1}\left[\sum_{i=1}^{t-1} \left(-\ln\left(\frac{i}{t}\right) \right)^{-1} . \left(\ln\left(\frac{x_i}{x_t}\right) - \lambda_t \ln\left(\frac{t}{i}\right) \right)^2 \right]. \tag{22}$$

After calibration of the distribution and establishing the parameters, flow quantiles were computed. By taking n to be the data record length in years (in this case 18 years) and r as the rank of the generated POT extreme events (with 1 allocated to the highest POT values), the theoretical return period T based on the calibrated distribution was computed (Eq. 23). Similarly, the empirical return period (Eq. 24) was determined.

$$T(\text{years}) = \left(\frac{n}{t}\right)\frac{1}{1 - G(X_T)} = \frac{n}{t \times \exp\left(-\frac{(x - x_\mathrm{t})}{\beta}\right)} \tag{23}$$

$$T(\text{years}) = \left(\frac{n}{r}\right) \tag{24}$$

where X_T is the flow value corresponding to T obtained from Eq. (24). t was previously defined as the rank of the threshold values. $[1 - G(x)]$ is the fitted EVD.

To carry out an extrapolation of the flow quantiles, Eqs. (25) and (26) can be used for exponential distribution and generalise Pareto Distribution, respectively.

$$X_\mathrm{T} = x_\mathrm{t} + \beta\{\ln(T)\} - \ln\left(\frac{n}{t}\right) \text{ for } \lambda = 0 \tag{25}$$

$$X_T = \exp\left(\ln\left(x_t\right) + \lambda\left\{\ln(T) - \ln\left(\frac{n}{t}\right)\right\}\right) \text{ for } \lambda \neq 0. \quad (26)$$

The flow quantiles were extrapolated to a particular return period (e.g. 25, 50, 100 years) for various purposes of water resources engineering applications such as determining the return period of an historical flow event. It is worth noting that in this study, extrapolation of quantiles is based on data of 18 years (1999 to 2016). However, the study by Schulz K, Bernhardt M [97] discovered that extrapolations of quantiles for higher return periods (say 100 years) based on series of short-term can bring about large uncertainties. To minimise the possible uncertainties resulting from a small sample size, extrapolation of quantiles should not be for a return period greater than three times the record length (in years) of the data being used for analysis [87]. Using the calibrated distribution parameters, the FDF relationships comprising of the accumulated values of flow for all the aggregation levels at three return periods (25, 50 and 100 years) were developed for both high and low flows. Lastly, the average model biases in replicating the observed high and low-flow quantiles at different aggregation levels were computed based on Eq. (5).

4 Results and discussion

4.1 Model performance for the calibration, validation and entire data period

Figure 3 shows the graphical comparison of observed and simulated river flow time series from the seven rainfall–runoff models. The set of parameters used to attain results in Fig. 3 are included in Supplementary Material M2 (see Table M1). It is noticeable from Table M1 that optimised parameters were all with in the allowable ranges for each model. There is noticeable underestimation of the peaks especially by TANK, IHACRES, SMAR and HMSV (Fig. 3c–d, f–g). There was an underestimation of low flows by IHACRES, SIMHYD and SMAR (Fig. 3d–f).

It is worth noting that calibration of a model based on total catchment water balance allocates more value to the high flows than low flows [81]. Nonetheless, generally all the models well reproduced the pattern in the observed flow.

Figure 4 shows ranking of the model performance based on NSE (Fig. 4a) and KGE (Fig. 4b) "goodness-of-fit" statistics. Additional information of model ranking (based on the remaining seven metrics (MAB, RRM, Re, Ia, R, MAE, CMA)) can be obtained in Supplementary Material M2 Fig. M8a–g. Generally, AWBM had the smallest rank (1) hence, exhibiting the best performance (Fig. 4a,b and Supplementary Material Fig. M8a–f)). However, when based on CMA (Supplementary Material M2 Fig. M8g, AWBM ranked third after SAC and Tank. In several past studies [36, 81], AWBM exhibited superlative performances. On the other hand, SMAR exhibited unsatisfactory performance with NSE of 0.46, 0.28 and 0.44 for calibration, validation and entire period, respectively (Fig. 4 and Table 3). All models had NSE values above 0.50 (except SMAR).

The nine "goodness-of-fit" measures used for the calibration, validation and entire periods of the seven models are shown in Table 3. All models (except SMAR) exhibited laudable performance in simulating flows in the catchment. The values of NSE varied between 0.46 to 0.83, 0.28 to 0.81 and 0.44 to 0.84 for calibration, validation and entire period, respectively. The best performance was obtained with the entire period, followed by the calibration period. Statistically, the best model performance would be shown by MAB of 0%. However, in this study, the MAB values varied amongst the models with some exhibiting negative, while others showed positive. This discrepancy in the MAB value amongst the models could be attributed to the varying model structures used to convert rainfall into runoff. The values of RRM ranged between 0.048–0.085 and 0.069–0.133, for calibration and validation, respectively. Of the seven models, AWBM had the smallest value, while SMAR exhibited the largest values of RRM (Table 3). Based on R_e and I_a, the best performance of a model would be indicated by a value of 1. Nevertheless, highest R_e and I_a values of 0.595 and 0.798, respectively, were obtained by the AWBM. The lowest R_e and I_a values of 0.242 and 0.635, respectively, were realised by the SMAR. This further attests the unsatisfactory performance of SMAR. The values of R ranged between 0.712–0.917 and 0.552–0.901, respectively, for calibration and validation periods. This indicates that for all the models, the modelled dispersion to the observed was generally well predicted. The MAE values did not vary much amongst the models, except for SIMHYD and SMAR that slightly varied from other models. The values were within the ranges of 5.0–9.6 m^3s^{-1} and 6.6–14.2 m^3s^{-1}, respectively, for the calibration and validation periods.

By using the CMA, the best model performance would be shown by the value of 1. In this study, SAC had the highest value (0.791) followed by HMSV (0.725) for the calibration period (Table 3). Considering the validation period, SAC still had the highest value (0.807) followed by TANK (0.745). For both calibration and validation period, SMAR had the smallest values of 0.466 and 0.293, respectively. Considering the KGE metric, AWBM performed best followed by SAC and TANK while SMAR performed last for all the periods. Generally, it can be concluded that the SMAR model performed worst, while AWBM performed

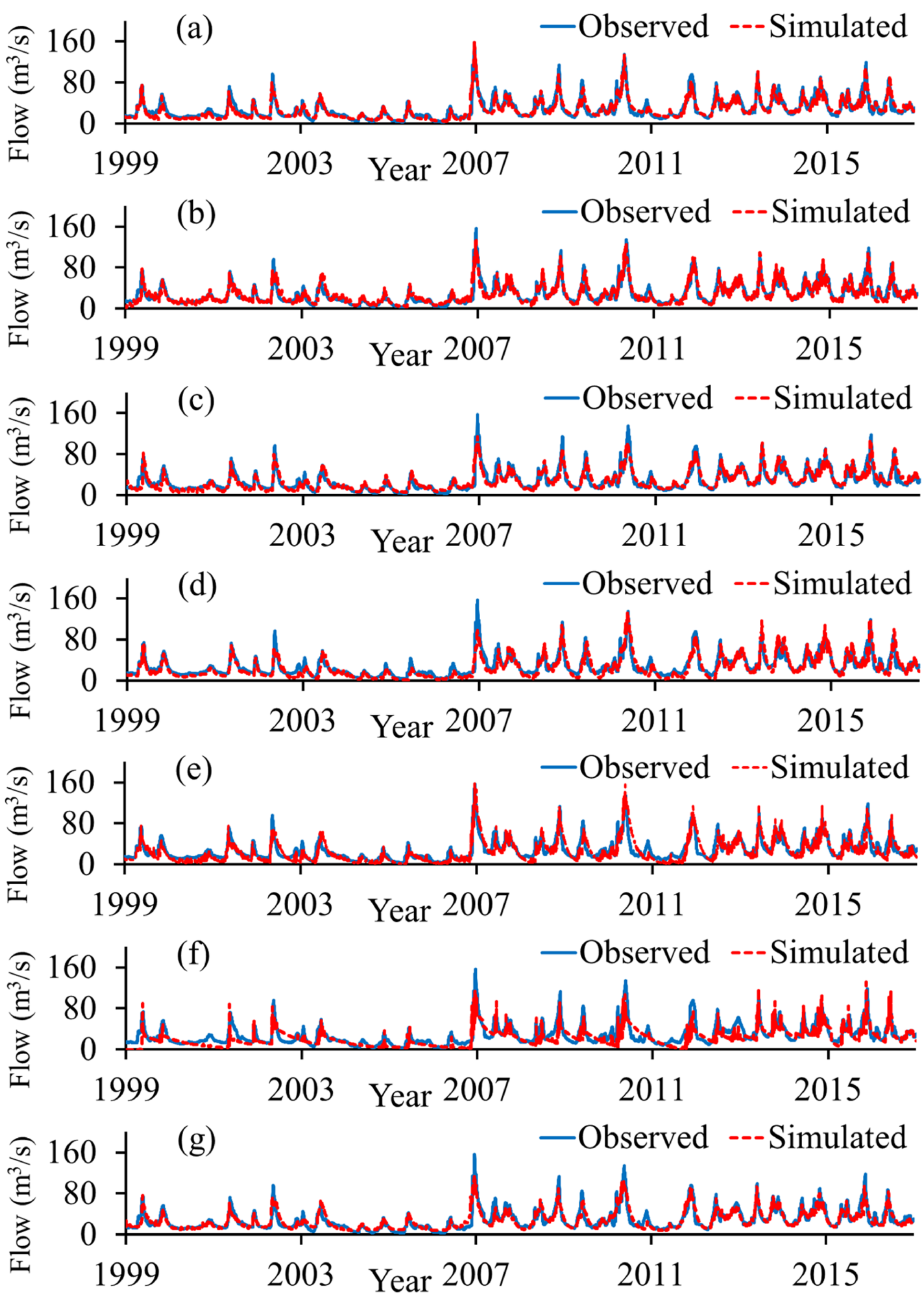


Fig. 3 Observed discharge and simulated flows using **a** AWBM, **b** SAC, **c** TANK, **d** IHACRES, **e** SIMHYD, **f** SMAR, **g** HMSV

finest based on the nine considered statistical indicators. However, in some studies, e.g. Onyutha et al. [36], SMAR generally performed better than SAC and TANK models. This could be attributed to the varying spatial resolution of catchments such as the size, climate, geology, landscape and hydrological data [70, 81]. In some cases, the lumped conceptual models have been found to perform better than physically based models. For instance in the study by Jaiswal et al. [98], for calibration, AWBM and TANK yield NSE values of 0.76 and 0.84, respectively, while SWAT had NSE of 0.75. However, in other areas, some rainfall–runoff models can exhibit unsatisfactory performances. For instance, the study by Pérez-Sánchez et al. [68], that compared six hydrological balance models (Témez, ABCD, GR2M, AWBM, GUO-5p and Thornthwaite-Mather) in several basins within Spain, AWBM model did not perform well.

Figure 5 shows the performance of models based on compiled values of ranking from all the nine "goodness-of-fit" measures constrained to the catchment total water balance. This followed a procedure described in Sect. 3.1.1. Production of Fig. 5 considered the

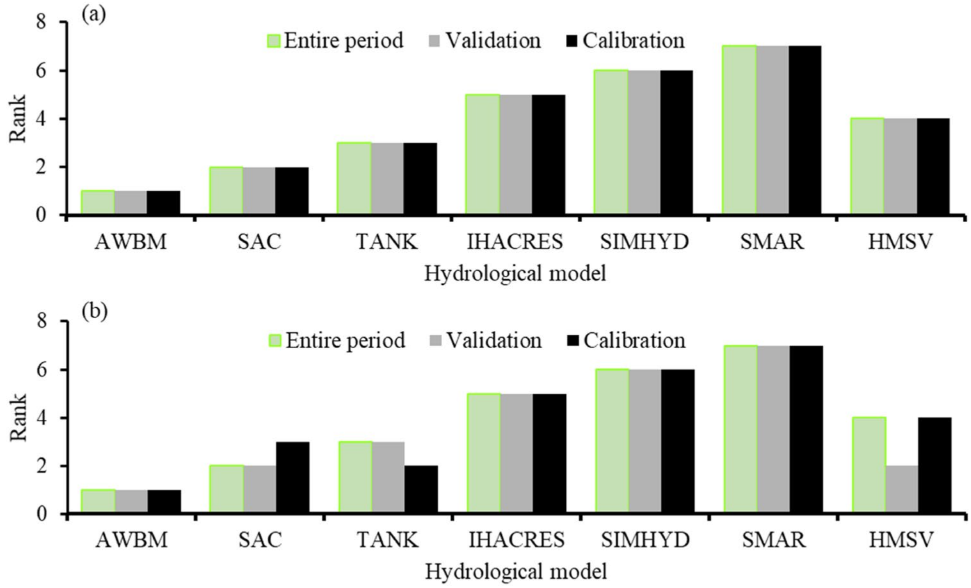


Fig. 4 Model performance in terms on ranking based on **a** Nash–Sutcliffe Efficiency (NSE), **b** Kling–Gupta efficiency (KGE). Ranks 1 and 7 denote best and worst performing model, respectively

performance of models during calibration, validation and entire period (Table 3). It is noticeable from Fig. 5 that AWBM performed best followed by SAC, while SMAR performance last. It should be noted that performance of these models can vary based on the catchment of application. For instance, while TANK was amongst the best 3 models in this study, the study by Onyutha et al. [36], that compared the performance of several lumped conceptual models in the simulation of daily River Kafu flows, TANK did not perform better. In the same study [36], SMAR generally performed well.

4.2 Comparison of model performance in simulating annual maxima and annual minima flow events

Figures 6 and 7, respectively, show the performance of models to simulate annual maxima and minima flow events in each year. For a superlative model, all the observed and simulated scatter points would be plotted along the 45° line.

It is observed that all the hydrological models simulated high flows (Fig. 6), better than low flows (Fig. 7). AWBM model performed better in simulating annual maxima flow events (Fig. 6a), followed by SACRAMENTO (Fig. 6b), while IHACRES performed last (Fig. 6d). The annual maxima observed flow events beyond 100 ms^{-1} were slighted underestimated by HMSV and TANK models (Fig. 6c,g). It is evident that the annual minima flow events in each year were largely overestimated by AWBM, SMAR and HMSV (Fig. 7a,f and g). However, SAC, IHACRES, SIMHYD models underestimated the low flow events (Fig. 7b,d,e). It is only TANK model exhibited a somewhat realistic balanced over- and underestimation with scatter points falling on both sides of the 45° line (Fig. 7c). Generally, HMSV model exhibited the best performance (Fig. 7g), while the SMAR performed last (Fig. 7f) in simulating annual minima.

The performance of hydrological models based on compiled values of ranking from all the nine "goodness-of-fit" measures considering the annual maxima and minima flows in each year is shown in Figs. 8 and 9. Like in Fig. 5, Figs. 8 and 9 are obtained based on the procedure explained in Sect. 3.1.1. Similar to the observations in Table 3 and Figs. 3, 4 and 5, AWBM model performed better than other models in simulating the annual maxima series, except when based on the MAB statistical indicator (Fig. 8). Similarly, SAC model performed second best. While Fig. 5 shows that SIMHYD and SMAR performed last, these two models exhibited commendable performance in simulating annual maxima flows. Instead, IHACRES, TANK and HMSV portrayed the worst performances. By considering the annual minima series, largely HMSV had the best performance, followed by AWBM and TANK (Fig. 9). Similar to the observations in Fig. 5, SMAR, SIMHYD and IHACRES performed last in simulating annual minima flow events (Fig. 9).

4.3 Amplitude–duration frequency analyses

In this study, the extreme value distribution tail analysis at all aggregation levels showed a normal tail for the exponential quantile–quantile (Q–Q) plot for both high and low flow events. For instance, results of the calibrated exponential distribution obtained at 1-day aggregation level for high and low flows are shown in Figs. 10 and 11, respectively. In the same figures, empirical and extrapolated

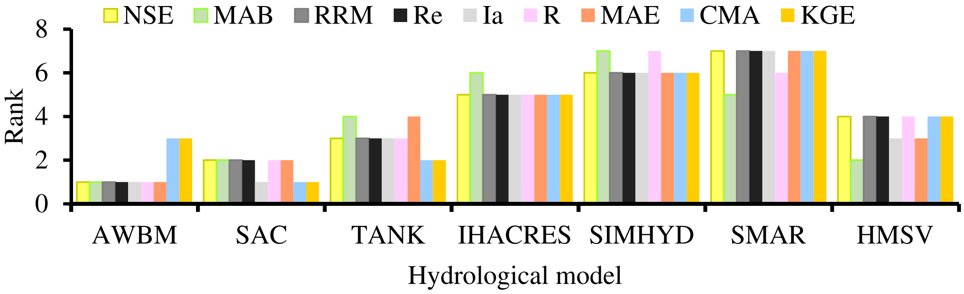
Table 3 Statistical performance evaluation of the models

S/N	Model	Calibration	Validation	Entire period
Nash–Sutcliffe efficiency (NSE)				
1	AWBM	0.828	0.808	0.837
2	Sacramento (SAC)	0.808	0.795	0.822
3	TANK	0.799	0.791	0.816
4	IHACRES	0.741	0.700	0.749
5	SIMHYD	0.630	0.502	0.612
6	SMAR	0.458	0.281	0.436
7	HMSV	0.778	0.787	0.804
Model average bias (MAB, %)				
1	AWBM	−3.121	4.910	0.186
2	Sacramento (SAC)	11.478	-4.935	4.721
3	TANK	3.555	9.625	6.053
4	IHACRES	−23.452	-6.020	-16.274
5	SIMHYD	−24.060	-10.563	-18.502
6	SMAR	−2.433	25.300	8.986
7	HMSV	12.010	-3.065	5.802
Ratio of root mean squared error to maximum event (RRM)				
1	AWBM	0.048	0.069	0.053
2	Sacramento (SAC)	0.051	0.071	0.055
3	TANK	0.052	0.072	0.056
4	IHACRES	0.059	0.086	0.065
5	SIMHYD	0.070	0.111	0.081
6	SMAR	0.085	0.133	0.098
7	HMSV	0.055	0.072	0.058
Relative efficiency (Re)				
1	AWBM	0.595	0.594	0.627
2	Sacramento (SAC)	0.569	0.580	0.608
3	TANK	0.547	0.567	0.692
4	IHACRES	0.477	0.460	0.510
5	SIMHYD	0.338	0.310	0.377
6	SMAR	0.242	0.130	0.250
7	HMSV	0.547	0.570	0.590
Index of agreement (I_d)				
1	AWBM	0.798	0.782	0.809
2	Sacramento (SAC)	0.783	0.795	0.803
3	TANK	0.774	0.774	0.794
4	IHACRES	0.758	0.756	0.776
5	SIMHYD	0.706	0.679	0.714
6	SMAR	0.635	0.479	0.615
7	HMSV	0.765	0.783	0.791
Coefficient of determination (R)				
1	AWBM	0.917	0.901	0.918
2	Sacramento (SAC)	0.901	0.902	0.909
3	TANK	0.896	0.892	0.903
4	IHACRES	0.895	0.881	0.895
5	SIMHYD	0.856	0.797	0.843
6	SMAR	0.712	0.552	0.685
7	HMSV	0.882	0.893	0.898
Mean absolute error (MAE) (m^3s^{-1})				
1	AWBM	5.079	6.609	5.709
2	Sacramento (SAC)	5.411	6.842	6.000
3	TANK	5.688	7.040	6.245
4	IHACRES	6.571	8.764	7.474
5	SIMHYD	8.316	11.299	9.544
6	SMAR	9.510	14.160	11.425
7	HMSV	5.689	6.962	6.213
Coefficient of model accuracy (CMA)				
1	AWBM	0.707	0.730	0.761
2	Sacramento (SAC)	0.791	0.807	0.839
3	TANK	0.705	0.745	0.834
4	IHACRES	0.560	0.726	0.688
5	SIMHYD	0.519	0.548	0.573
6	SMAR	0.466	0.293	0.554
7	HMSV	0.725	0.719	0.737
Kling–Gupta efficiency (KGE)				
1	AWBM	0.888	0.890	0.900
2	Sacramento (SAC)	0.854	0.843	0.883
3	TANK	0.873	0.826	0.879
4	IHACRES	0.779	0.804	0.837
5	SIMHYD	0.756	0.745	0.772
6	SMAR	0.669	0.459	0.653
7	HMSV	0.821	0.847	0.841

quantiles are shown. In all cases, linearity behaviour in the quantiles was realised towards the tail of the distributions. While the exponential plot was adopted to obtain the linear behaviour of the quantiles, log-transformation was applied to the return period *T* on the *x*-axis. Towards the tail of the distribution, the mismatch between the empirical and modelled quantiles gest systematically larger than those for small return periods. This tends to result from the influence of flooding on flow measurement stemming from the bias in rating curve extrapolation or the difference between the river discharge and the catchment rainfall–runoff discharge [87]. Censoring out outliers prevents underestimation of flow which could result from assuming a light instead of normal tail. Assuming a light tail may result in underestimation of design quantiles for sizing water infrastructures such as bridges. In the Supplementary Material M3 Figs. M12 to M17, exponential Q–Q plots obtained at different aggregation levels for both high (5, 30 and 90 days) and low flows (30, 180 and 365 days) are provided.

Considering the high flows (Fig. 10), the frequency curve for SIMHYD and SMAR showed substantial deviations from the observed curve for the estimated flows especially at higher return periods, while AWBM, SAC, TANK, IHACRES and HMSV frequency curves did not display substantial deviations. From Fig. 10, the return period

Fig. 5 Assessment of model performance based on compiled values of ranking from all the nine "goodness-of-fit" considering the overall water balance. The best performing model is ranked 1 while the worst takes rank 7



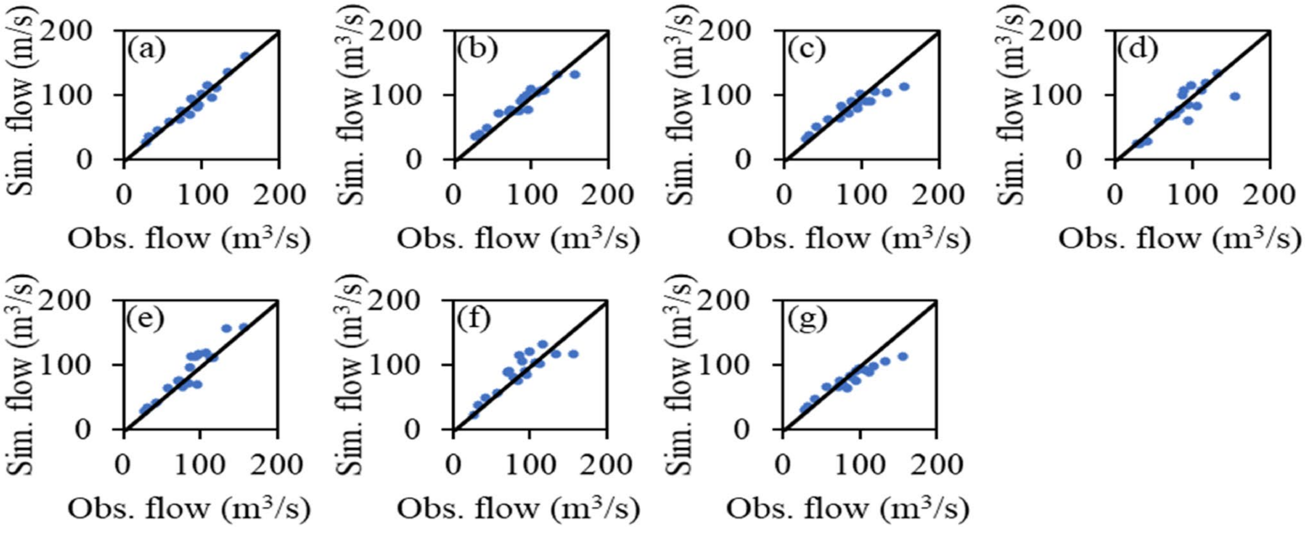


Fig. 6 Model performance assessment based on comparison of observed and simulated annual maxima flows in each year for **a** AWBM, **b** SAC, **c** TANK, **d** IHACRES, **e** SIMHYD, **f** SMAR, **g** HMSV, Sim. stands for simulated, while Obs. means observed

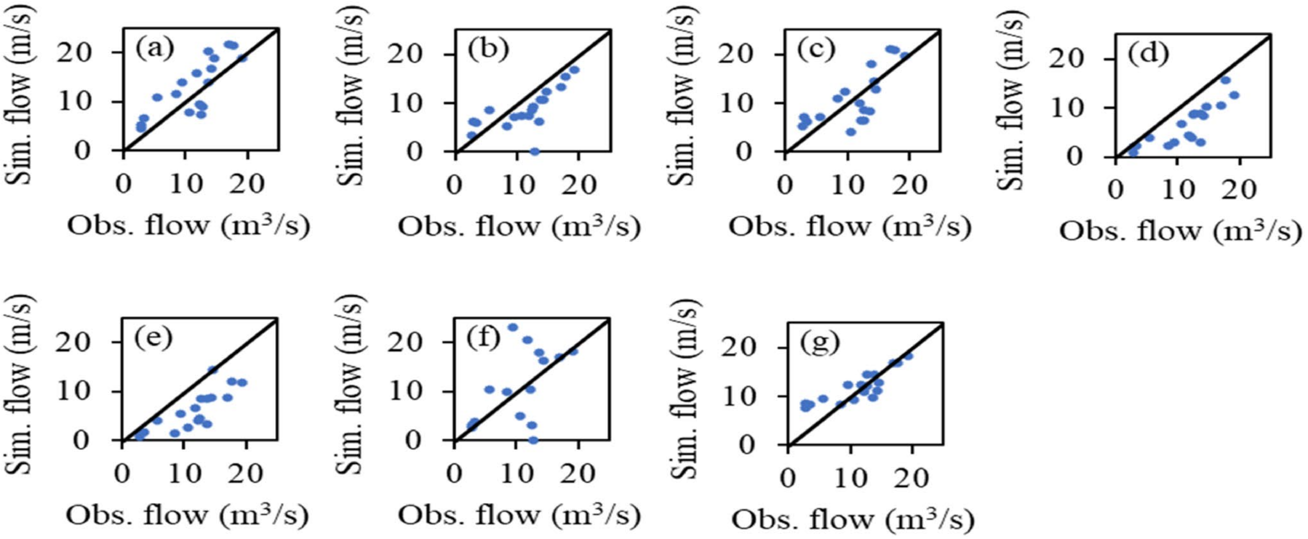


Fig. 7 Model performance assessment based on comparison of observed and simulated annual minima flows in each year for **a** AWBM, **b** SAC, **c** TANK, **d** IHACRES, **e** SIMHYD, **f** SMAR, **g** HMSV; Sim. stands for simulated, while Obs. means observed

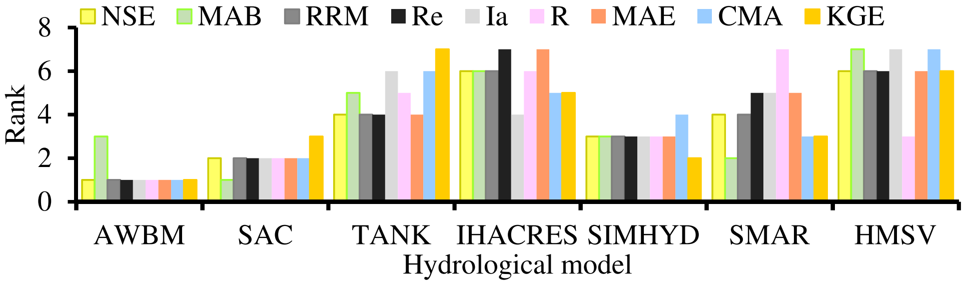


Fig. 8 Assessment of model performance based on compiled values of ranking from all the nine "goodness-of-fit" considering the annual maxima flows. The best performing model is ranked 1 while the worst takes rank 7

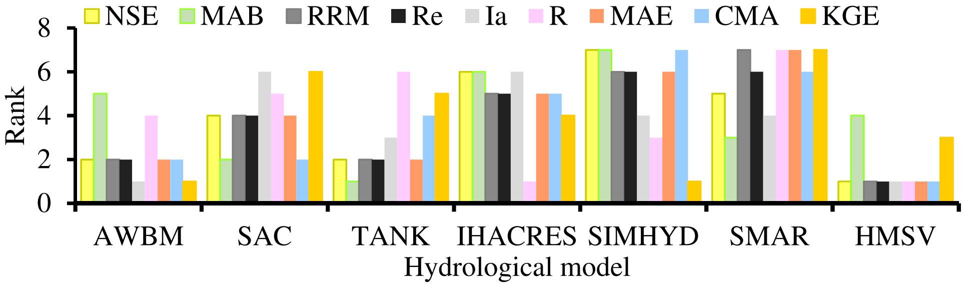


Fig. 9 Assessment of model performance based on compiled values of ranking from all the nine "goodness-of-fit" considering the annual minima flows. The best performing model is ranked 1 while the worst takes rank 7

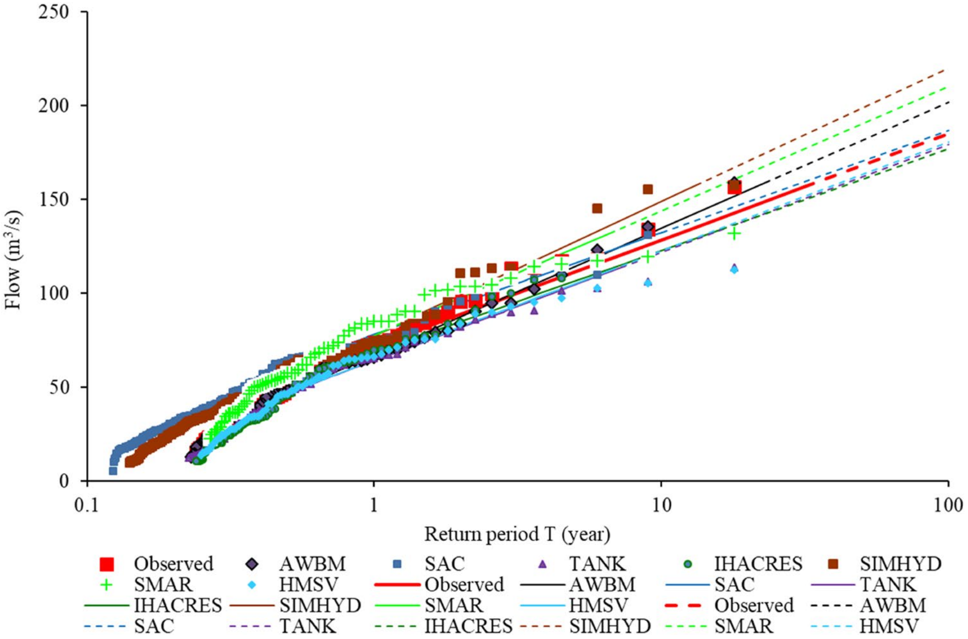


Fig. 10 Comparison of exponential quantile–quantile plot for the high flow POT events considering 1-day aggregation level. The markers represent the empirical, solid lines signify theoretical (calibrated distribution), while dashed lines represent extrapolated quantiles

for an average 1-day flood of 150 m^3s^{-1} is 24 years based on the observed flow curve, while the same event has a return period of 17 years for an evaluation based on AWBM model. However, when considering the HMSV model, the same event will have a return period of 30 years, while for the SIMHYD model, the return period is 10 years.

It is noticeable that for low flows (Fig. 11), the actual quantiles were attained after back transformation of the (1/X). In view of the low flows (Fig. 11), SMAR, SIMHYD and IHACRES flow curves exhibited large deviations from the observed quantile even at return periods less than 1 year. The AWBM, SAC, TANK and HMSV flow curves have close agreement with the observed quantiles throughout all the return periods (despite slight deviations on the extrapolated quantiles). The extrapolated quantiles for high flows can be relevant for vigilant flood analysis which can guide the planning and designing of risk-based water engineering structures such as bridges and slipways. Similarly, the extrapolated quantiles for low flows can be relevant for cautious drought analysis which is key in water resources management aspects

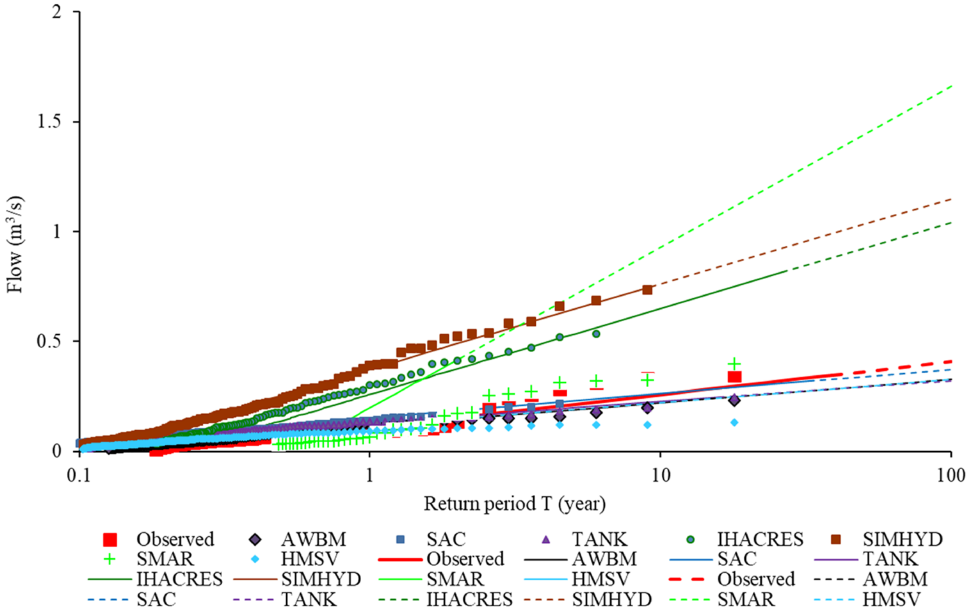


Fig. 11 Comparison of exponential quantile–quantile plot for the low flow POT events considering 1-day aggregation level. The markers represent the empirical, solid lines signify theoretical (calibrated distribution), while dashed lines represent extrapolated quantiles

such as determining irrigation water for agriculture and domestic water supply.

The ADF relationships considering all the aggregation levels are shown in Figs. 12 and 13 for high and low flows, respectively. These were generated from the quantiles projected based on the EVD. For high flows, it is shown that SMAR (Fig. 12f) overestimated the high-flow quantiles for the aggregation level of one day. However, for aggregation levels higher than 3 days, SMAR underestimated the high-flow quantiles, except at 60 and 90 days. For all aggregation levels, SIMHYD (Fig. 12e) overestimated the high-flow quantiles particularly for the return periods of 25 and 100 years. For all aggregation levels, considering the AWBM, SAC, TANK, HMSV (Fig. 12a–c, g), the observed and simulated high-flow quantiles were generally in close agreement. The underestimation of high-flow quantiles at higher aggregation levels by SMAR could be attributed to the model inadequacy in capturing higher quantiles in the extreme value distribution tail. Similarly, the overestimation or underestimation of flow quantiles at one-day aggregation level could be linked to the high noise in flow time series at low aggregation levels, resulting in uncertainties in calibration of the extreme value distribution.

Considering the low flows, IHACRES, SIMHYD and SMAR (Fig. 13d–f) underestimated the low-flow quantiles. However, for AWBM, SAC, TANK, HMSV (Fig. 13a–c, d), the quantiles from the observed and simulated low flows were comparable, except at aggregation level of one day with noticeable overestimation.

The model biases in reproducing high and low quantiles at different aggregation levels are shown in Tables 4 and 5, reactively. From Table 4, SIMHYD exhibited positive biases at all aggregation level indicating an over estimation of high-flow quantiles. Except for the one-day aggregation level, SMAR showed an underestimation of high-flow quantiles. For low flows (Table 5), IHACRES, SIMHYD and SMAR showed large negative biases at all aggregation levels signifying an underestimation of low-flow quantiles. HMSV showed positive biases throughout (except for 150 days aggregation level), and on average it exhibited the smallest biases compared to other models.

4.4 Explanation of the differences in performance of the various models

As earlier noted, all the seven rainfall–runoff models (AWBM, SAC, TANK, IHACRES, SIMHYD, SMAR and HMSV) simulated the high flows better than the low flows. The study Staudinger et al. [99], that assessed the impact of model structure on low flow simulation, established that most rainfall–runoff models poorly reproduce low flows because their structures are largely designed to mimic high flows. Li et al. [100] also stressed that the structure of a model may influence its performance. However, in this study, the variability in model performance could not evidently be attributed to the differences in model structures. This is because various models performed either better under the consideration of total water balance and/or

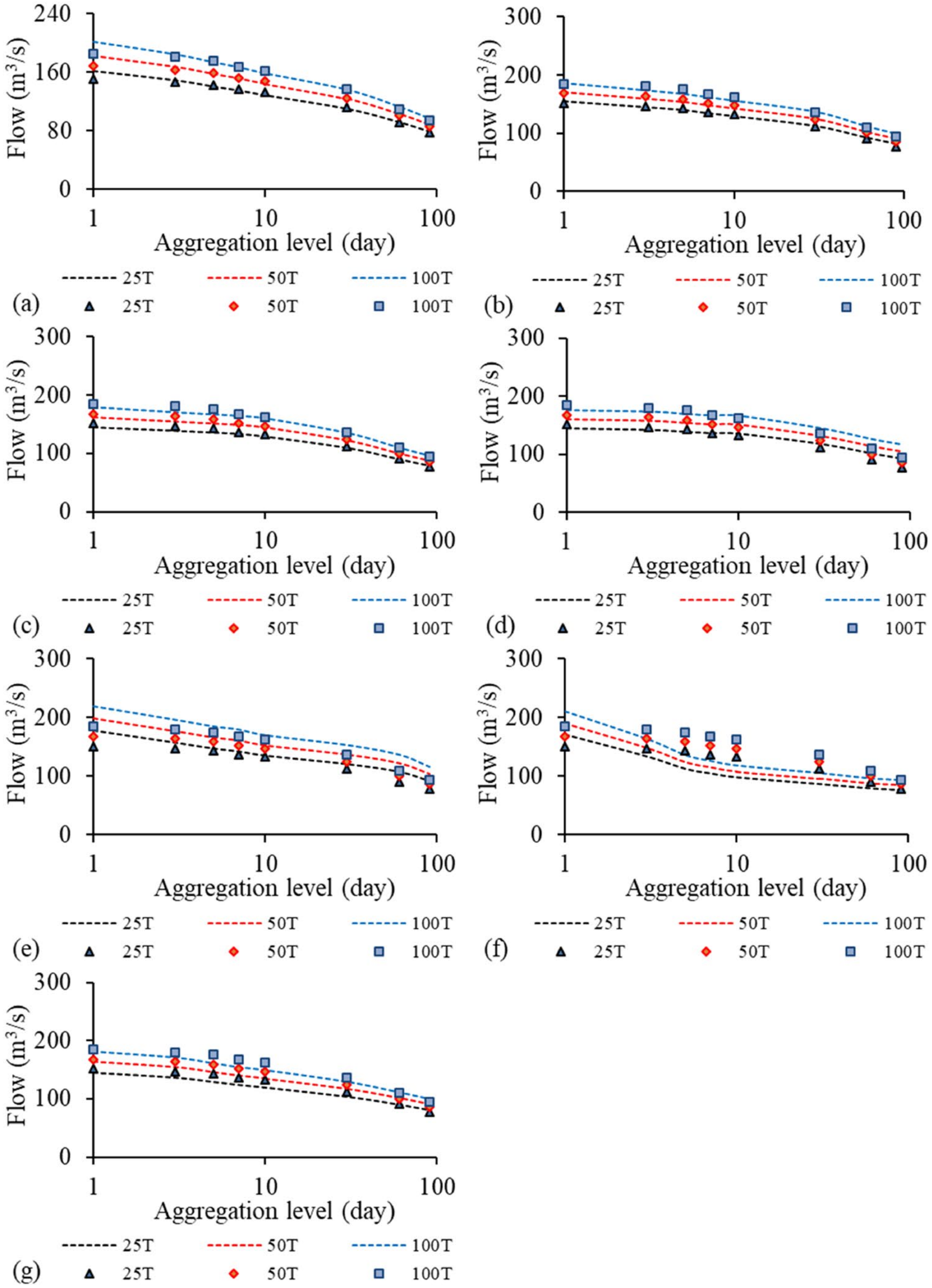


Fig. 12 The FDF plots for high extreme observed and simulated flows obtained at different aggregation levels (1–90 days) for **a** AWBM, **b** SAC, **c** TANK, **d** IHACRES, **e** SIMHYD, **f** SMAR, **g** HMSV obtained at 25-, 50- and 100-year return periods. In the legend, dashed lines denote simulated flow, while the markers signify observed flow

extreme hydrological conditions. Generally, four models (AWBM, SAC, TANK and HMSV) performed well regarding simulating flows in the catchment. Considering the total water balance, AWBM had the highest NSE values of 0.828 and 0.808 for calibration and validation periods, respectively. In addition, AWBM had the lowest MAB value of 0.186% for the entire period (1999–2016). Equally, AWBM performed better than other models in simulating the annual maxima flows. However, the model came second after HMSV in simulating the annual minima flows.

Model performance in simulating flows could as well be associated with model parameters [101]. For instance, the influence of low flows in most squared residual-based objective functions is low. Models' parameters exhibited varying sensitivity. In supplementary material M2 Figs. M9–M11, a sample of parameter sensitivity analyses for AWBM, SACRAMENTO, SMAR can be found. AWBM was highly sensitive to higher values (close to 1.0) of base flow recession constant and also relatively sensitive to surface flow recession constant, storage capacity of third store and fraction of catchment area for the

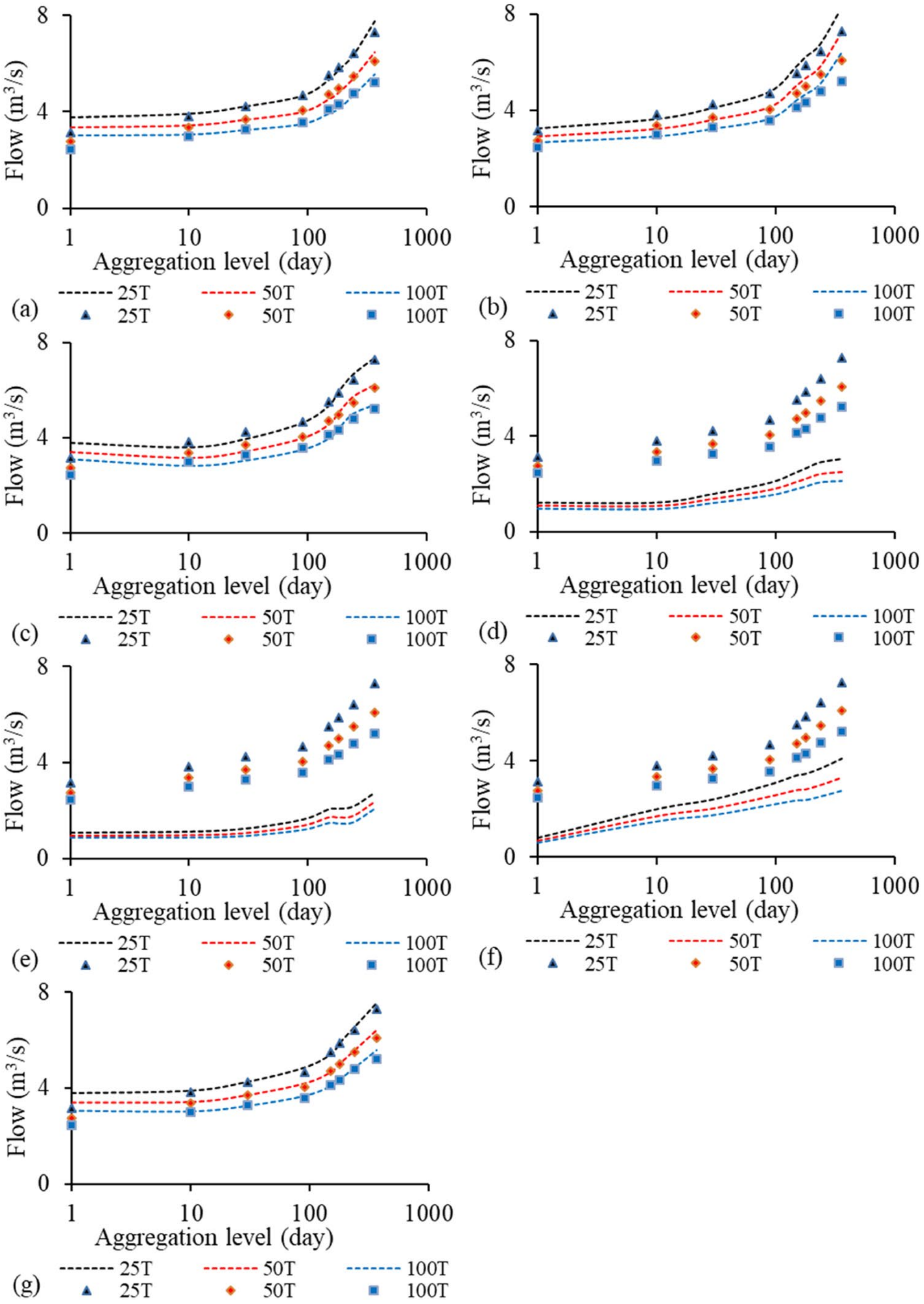


Fig. 13 The FDF plots for low extreme observed and simulated flows obtained at different aggregation levels (1–365 days) for **a** AWBM, **b** SAC, **c** TANK, **d** IHACRES, **e** SIMHYD, **f** SMAR, **g** HMSV obtained at 25-, 50- and 100-year return periods. In the legend, dashed lines denote simulated flow, while the markers signify observed flow

Table 4 Model biases (%) in simulating high-flow quantiles

Model	Aggregation level [day]							
	1	3	5	7	10	30	60	90
AWBM	7.99	1.96	−1.26	−0.52	−2.44	−0.83	2.26	2.68
SAC	1.37	− 2.52	− 3.36	− 2.35	− 3.14	0.37	2.04	5.39
TANK	− 3.79	− 5.47	− 5.02	− 2.02	− 2.03	− 1.62	− 0.38	2.10
IHACRES	− 4.50	− 3.29	− 3.01	0.09	2.75	6.06	13.58	21.72
SIMHYD	18.11	8.00	4.25	5.83	3.41	9.79	21.26	20.35
SMAR	13.11	− 9.27	− 21.93	− 23.96	− 26.58	− 22.74	− 12.58	− 1.31
HMSV	− 2.98	− 5.96	− 8.80	− 8.23	− 8.72	− 6.24	0.41	5.23

Table 5 Model biases (%) in simulating low-flow quantiles

Model	Aggregation level [day]							
	1	10	30	90	150	180	240	365
AWBM	22.35	3.21	−0.06	−0.69	−3.76	−3.45	−1.64	6.42
SAC	6.38	−3.26	−2.06	2.32	6.13	6.96	5.95	18.33
TANK	26.67	−4.84	−6.73	−1.63	−3.68	0.98	4.67	2.76
IHACRES	−60.73	−67.99	−62.82	−56.64	−55.70	−55.43	−55.63	−58.62
SIMHYD	−64.89	−70.35	−70.65	−65.77	−63.29	−65.35	−67.36	−61.28
SMAR	−74.72	−48.89	−44.85	−37.26	−40.54	−42.99	−44.93	−45.38
HMSV	22.70	1.65	0.00	3.30	−1.45	0.80	1.90	5.20

first store. However, fraction of catchment area for the second store and storage capacity of second store were almost insensitive parameters with minimal variation in the NSE values (supplementary material Fig. M9). The baseflow index, baseflow and surface runoff recession constant parameters in AWBM could justify the laudable performance of the model in reproducing both high and low flows. Previous studies [36, 81, 102] revealed good performance of AWBM. However, in some cases, the AWBM may not yield better results. For instance, in the study by Pérez-Sánchez et al. [68], that compared six hydrological models including the AWBM, revealed poor performance of the AWBM. Conclusively, it can be stated that the AWBM structure is tailored towards better capturing high flows than low flows.

Similarly, the SACRAMENTO model with a set of seventeen parameters had the second-best performance in simulating flows during calibration and validation and mimicking of the annual maxima flows. Some of the SACRAMENTO parameters exhibited high sensitivity only at small values (e.g. additional fraction of pervious area, exponential percolation rate) while other had high sensitivity at higher values (e.g. fraction of base flow which is groundwater flow, lower zone free water primary maximum). Other parameters were almost insensitive (e.g. upper zone Free water maximum, fraction of water unavailable for transpiration) (supplementary material Fig. M10).

The performance of SIMHYD model with nine parameters was not adequately better for both low and high flow simulation. The study by Li et al. [100], that compared three hydrological models (HBV, SIMHYD and XAJ), revealed lower efficiency of the SIMHYD model. While SMAR model generally had well-identified parameters, three parameters (unit hydrograph linear routing, unit hydrograph linear routing component, infiltration rate) exhibited low sensitivity (supplementary material Fig. M11). This might have resulted in the lower performance of the model. In some studies (e.g. Bashar [103]), SMAR model yielded sufficiently better results. The performance of a model may not be associated with the number of parameters it has [104]. This is because, SACRAMENTO model that could be regarded as being over-parameterised with 17 parameters, performed better than other models with relatively fewer parameters, e.g. SIMHYD (9 parameters), IHACRES (11 parameters), SMAR (9 parameters) and HMSV (10 parameters). Besides, AWBM model with the smallest number of parameters (eight), performed best overall. Although the calibration of HMSV was based on the "step-wise" (subflow variation) strategy (calibration baseflow, interflow and overland flows separately), the model did not perform well in simulating annual maxima, despite it performing best in simulating annual minima flows. The commendable performance of HMSV model in simulating low flows could be attributed to the base flow parameter and baseflow recession constant. The credible ranking of HMSV in reproducing annual maxima flow could also be linked to the overland and interflow parameters in the model.

Interestingly, the performance of a model may as well be attributed to the selected "goodness-of-fit" measures [38, 81]. For instance, it is noticeable that for annual minima flows, HMSV generally performed better than TANK model. However, when based on the MAB "goodness-of-fit" metric, TANK model displayed better performance than HMSV. Similarly, for annual maxima flows, AWBM performed far better than other models. However, when based on the MAB, SAC and SMAR models performed better than AWBM. Besides, under the consideration of total water balance, using the CMA and KGE statistical indicators, SAC and TANK model performed better than the AWBM, despite AWBM displaying overall superlative performance. Hence, to strongly conclude on the efficacy of a particular rainfall–runoff model, it is vital to assess its performance using various "goodness-of-fit" statistics as implemented in this study.

5 Conclusion

Previous studies conducted to evaluate several conceptual rainfall–runoff models (AWBM, SACRAMENTO, TANK, IHACRES, SIMHYD, SMAR and HMSV) performances based on multiple "goodness-of-fit" metrics in simulating hydrological extremes in River Malaba sub-catchment are lacking. This study analysed seven lumped conceptual rainfall–runoff models based on nine "goodness-of-fit" measures in simulating hydrological extreme of River Malaba sub-catchment. Because of the poor distribution of meteorological stations, with the existing ones having few data available in the recent years, in addition to the uncertain and questionable quality, this study was based on PGF meteorological dataset recorded from 1999 to 2016. Bias correction in the PGF data was performed using the simple multiplicative bias correction method. Daily River Malaba observed flow time series at Budumba station were used in model calibration and validation. Considering the total water balance, largely the models (except SMAR) well reproduced the observed flow pattern. All the models (except AWBM, TANK and HMSV), largely underestimated the base flow. Generally, AWBM performed better than other models with NSE of 0.83 for calibration, while SMAR performed last with NSE of 0.46.

The seven rainfall–runoff models simulated the high flows better than the low flows. This could be attributed to the structure of most rainfall–runoff models tailored at reproducing high flows better than the low flows. By designing hydrological model structures flexible in capturing both high and flow flows could avert the situation. The AWBM model performed better than other models in simulating annual maxima flow events. The annual minima flow events in each year were largely overestimated by AWBM, SMAR and HMSV, but underestimated by the SAC, IHACRES, SIMHYD models. Largely HMSV ranked number one, followed by the AWBM and TANK in simulating annual minima flow events. SMAR, SIMHYD and IHACRES did not exhibit satisfactory performance in simulating annual minima flows.

Flow–Duration Frequency (FDF) analyses yielded a normal tail of the exponential Q–Q plot at the selected aggregation levels for both low and high flows. The frequency curves for SIMHYD and SMAR showed large deviations from the observed curve especially at higher return periods, while AWBM, SAC, TANK, IHACRES and HMSV frequency curves did not display substantial deviations. For the same flood event, different models yielded varying return periods. For instance, the return period for an average 1-day flow of 150 m^3s^{-1} is 24 years based on the observed flow curve, while the same event has a return period of 17 years for an evaluation based on AWBM model. However, using the HMSV model, the same event will have a return period of 30 years, while for the SIMHYD model, the return period is 10 years. The discrepancy in mimicking the observed frequency curves by different models demonstrates how implicit determination of the return period for a particular flow event can vary depending on the considered hydrological model. For the low flows, SMAR, SIMHYD and IHACRES flow curves exhibited large deviations from the observed curve even at return periods less than 1 year. The AWBM, SAC, TANK and HMSV flow curves had close agreement with the observed curves throughout all the return periods.

The simulated high-flow quantiles from AWBM, SAC, TANK, HMSV and the observed were generally in close agreement. However, SMAR and SIMHYD underestimated and overestimated high-flow quantiles, respectively. IHACRES, SIMHYD and SMAR underestimated the low-flow quantiles. With AWBM, SAC, TANK, HMSV, the quantiles from the observed and simulated low flows were comparable. All models exhibited varying biases (positive and/or negative) in reproducing the extreme flow quantiles. The differences in biases designate the influence of model selection for hydrological extreme analysis. The biases could be attributed to errors in model inputs (in this case, rainfall and PET) or in the observed flows used for model calibration and validation. Errors in observed flow could be due to flooding influence or due to wrong extrapolation made by the rating curve [105]. Flooding influence results in down bending of the rating curve to lower water levels hence leading to underestimation for higher flows. Besides, the errors in estimating the EVD parameters especially the slope parameter β might have contributed to the biases. However, in spite of accurate estimation of parameters, an extrapolation outside the calibration and validation periods of a distribution can be very erroneous.

By simultaneously considering the overall assessment of high and low flows, AWBM slightly performed better than other models in simulating hydrological extreme in River Malaba sub-catchment, while SMAR ranked last. A few limitations to this study and recommendations for future research studies are worth stating. The study adopted the computation of bias correction factors for PGF data without distinguishing between seasons but rather monthly values were used. It is likely that omitting seasonal difference in the analysis might yield biased results due to blind temporal relationship. To even out the possible uncertainties in the computed bias factors, it is recommended that future research studies consider seasons and temporal changes. Besides, the uncertainties in the use of uncalibrated empirical Hargreaves method for PET estimation should be quantified in future research. Future research studies may also evaluate the hydrological models' performance without bias correction of PGF data. The study by Mubialiwo et al. [46] revealed the presence of trends and shifts in the rainfall and potential evapotranspiration over the study area. Since the same meteorological datasets (rainfall and potential evapotranspiration) were used in this study as the inputs in the hydrological models, the generated flow is presumed to contain trends and sub-trends, which could indicate possible surge of flooding events and/or severity of hydrological droughts. However, this was not analysed in this study and is recommended for investigation in the future research studies. Furthermore, the impact of human factors (e.g. land use) and climate variability on water resources in the River Malaba sub-catchment should be investigated by future research.

Even with the above-mentioned limitations (which may be addressed by the suggested future research studies), this study provides relevant information for planning of risk-based water resources applications. The flow data from the best hydrological model (AWBM) can be used for many applications in water resources management. For instance, the FDF relationships can be used to construct rainfall–runoff design hydrographs as inputs in the hydrodynamic flood model to simulate discharge at different locations along the river. The simulated discharge can then be applied in the design of flood protection systems and calculation of flood maps for different return periods in the study area. Besides, FDF relationships can be adopted to estimate the return periods of historical flow events. Worth noting is that for the same catchment, even models of the same family (e.g. the "e-Water Toolkit" RRLs) can produce varying results and their performance can differ based on the selected "goodness-of-fit" measures. Therefore, it is necessary to compare results from various models prior to selection of a particular model to support decision regarding water resources management application, e.g. simulating flows that can support floods or drought analysis.

Acknowledgments The research was financed through the Ph.D. scholarship awarded to the first author by the Africa Center of Excellence for Water Management, Addis Ababa University, Ethiopia (Grant No. ACEWM/GSR/4334/11). The authors acknowledge the use of Princeton Global Forcings (PGF) gridded data [91]. The observed rainfall used for bias correction in the PGF dataset was obtained from the Uganda Ministry of Water and Environment. The six hydrological models (AWBM, SACRAMENTO, TANK, IHACRES, SIMHYD and SMAR) were obtained from the "eWater Toolkit" of the Cooperative Research Centre for Catchment Hydrology in Australia via the link http://www.toolkit.net.au/ (accessed: 08 February 2021). The seventh model (HMSV) and the FAN-Stat tool were accessed freely via https://sites.google.com/site/conyutha/tools-to-download (accessed: 10 February 2021). We greatly appreciate the Department of Civil and Environmental Engineering, Kyambogo University, Uganda for offering office space and computer laboratory services.

Data availability Data used in this study can be accessed on request from the corresponding author.

Declarations

Conflicts of interest The authors declare no competing interests concerning the publication of this work.

References

1. IPCC (2018) Global warming of 15 °C, an IPCC special report on the impacts of global warming of 15 °C above pre-industrial levels and related global greenhouse gas emission pathways. In: The context of strengthening the global response to the threat of climate change. Intergovernmental Panel on Climate Change, Switzerland
2. Mayega RW, Tumuhamye N, Atuyambe L, Okello D, Bua G, Ssentongo J, Bazeyo W (2015) Qualitative assessment of resilience to the effects of climate variability in the three communities in Uganda. RAN secretariat and east african resilience innovation lab (EA RILab), Kampala, Uganda
3. Ministry of Water and Environment (2018) Mpologoma catchment management plan. Ministry of Water and Environment, Kampala, Uganda
4. Ministry of Water and Environment (2015) Water and environment sector performance report 2015. Ministry of Water and Environment, Kampala, Uganda
5. Floodlist (2020) Uganda–deadly floods and landslides in eastern region (updated). http://floodlist.com/africa/uganda-floods-bududa-sironko-december-2019. Accessed 10 Jan 2020
6. ACAPS (2020) Uganda: flooding and landslides in Bududa. https://www.acaps.org/sites/acaps/files/products/files/20181018_acaps_start_briefing_note_uganda_flooding_and_landslides_in_bududa.pdf. Accessed 15 Oct 2020
7. Atuyambe LM, Ediau M, Orach CG, Musenero M, Bazeyo W (2011) Land slide disaster in eastern Uganda: rapid assessment of water, sanitation and hygiene situation in Bulucheke camp, Bududa district. Environ Health 10:1–22. https://doi.org/10.1186/1476-069X-10-38
8. OCHA Uganda (2020) Eastern uganda landslides and floods situation report#3. https://reliefweb.int/sites/reliefweb.int/files/resources/81D1CDAB49713514C12576EA003098A5-Full_Report.pdf. Accessed 3 Oct 2020
9. Markandya A, Dale N, Garcia J, Langoya C, Monkhouse C (2015) Economic assessment of the impacts of climate change in uganda: arabica coffee production in the Mount Elgon region (Bududa District). In: Ministry of water and environment. https://cdkn.org/wp-content/uploads/2015/11/Uganda_CC-Economics_Mount-Elgon_case-study.pdf. Accessed 13 Oct 2020
10. Reliefweb (2012) Floods ravage gardens in Tororo. In: OCHA. https://reliefweb.int/report/uganda/floods-ravage-gardens-tororo
11. Reliefweb (2007) Uganda: floods OCHA situation report no. 7. In: OCHA. https://reliefweb.int/report/uganda/uganda-floods-ocha-situation-report-no-7
12. Downer CW, Ogden FL, Martin WD, Harmon RS (2002) Theory, development, and applicability of the surface water hydrologic

model CASC2D. Hydrol Process 16:255–275. https://doi.org/10.1002/hyp.338

13. Abbott MB, Bathurst JC, Cunge JA, O'connell P, Rasmussen J (1986) An introduction to the European hydrological system—Systeme Hydrologique Europeen, "SHE", 1: History and philosophy of a physically-based, distributed modelling system. J Hydrol 87:45–59
14. Refsgaard J, Storm B (1995) MIKE SHE. In: Singh VP (ed) Computer models of watershed hydrology. Water Resources Publications, Highlands Ranch, Colorado, USA, pp 809–846
15. Leavesley GH, Restrepo PJ, Markstrom SL, Dixon M, Stannard LG (1996) The modular modeling system (MMS): User's manual. Geological survey (U.S.) Open-File Report 96–151, Denver, Colorado
16. Arnold J, Moriasi D, Gassman P, Abbaspour K, White M, Srinivasan R, Santhi C, Harmel R, Van Griensven A, Van Liew M, Kannan N, Jha M (2012) SWAT: model use, calibration, and validation. Trans ASABE 55:1491–1508
17. Pande S, Arkesteijn L, Savenije HHG, Bastidas LA (2014) Hydrological model parameter dimensionality is a weak measure of prediction uncertainty. Hydrol Earth Syst Sci Discuss 11:2555–2582. https://doi.org/10.5194/hessd-11-2555-2014
18. Asadi A (2013) The comparison of lumped and distributed models for estimating flood hydrograph (study area : Kabkian basin). J Electron Commun Eng Res 1:7–13
19. Tassew BG, Belete MA, Miegel K (2019) Application of HEC-HMS model for flow simulation in the Lake Tana Basin: the case of gilgel abay catchment, upper Blue Nile Basin, Ethiopia. Hydrology 6:1–17. https://doi.org/10.3390/hydrology6010021
20. Boughton W (2004) The Australian water balance model. Environ Model Softw 19:943–956. https://doi.org/10.1016/j.envsoft.2003.10.007
21. Burnash R (1995) The NWS river forecast system-catchment modeling. In: Singh V (ed) Computer models of watershed hydrology. Water Resources Publication, Colorado, pp 311–366
22. Sugawara M (1995) Tank model. In: Singh VP (ed) Computer models of watershed hydrology. Water Resources Publications, Littleton, CO, USA, pp 165–214
23. Croke B, Andrew F, Spate J, Cuddy S (2005) IHACRES user guide. Technical report 2005/19. Second edition. iCAM, School of resources, environment and society, The Australian National University, Canberra, Australia
24. Jakeman AJ, Littlewood IG, Whitehead PG (1990) Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments. J Hydrol 117:275–300. https://doi.org/10.1016/0022-1694(90)90097-H
25. Porter JW, McMahon TA (1971) A model for the simulation of Streamflow data from climatic records. J Hydrol 13:297–324
26. O'Connell PE, Nash JE, Farrell JP (1970) River flow forecasting through conceptual models part II—the Brosna catchment at Ferbane. J Hydrol 10:317–329. https://doi.org/10.1016/0022-1694(70)90221-0
27. Onyutha C (2019) Hydrological model supported by a step-wise calibration against sub-flows and validation of extreme flow events. Water 11(244):1–23. https://doi.org/10.3390/w11020244
28. Pitman WV (1973) A mathematical model for generating monthly river flows from meteorological data in South Africa. University of the Witwatersrand, Hydrological Research Unit
29. Ford D, Pingel N, DeVries JJ (2008) Hydrologic modeling system-HEC-HMS-applications guide. US Army Corps of Engineers Hydrologic Engineering Center, Davis, CA, USA
30. Bergström S (1992) The HBV model-its structure and applications. SMHI, Norrköping Sweden
31. Adnan RM, Liang Z, Heddam S, Kermani MZ, Kisi O, Li B (2020) Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs. J Hydrol. https://doi.org/10.1016/j.jhydrol.2019.124371
32. Grimaldi S, Nardi F, Piscopia R, Petroselli A, Apollonio C (2020) Continuous hydrologic modelling for design simulation in small and ungauged Basins: a step forward and some tests for its practical use. J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.125664
33. Aghelpour P, Varshavian V (2020) Evaluation of stochastic and artificial intelligence models in modeling and predicting of river daily flow time series. Stoch Environ Res Risk Assess 34:33–50. https://doi.org/10.1007/s00477-019-01761-4
34. Ebtehaj I, Zeynoddin M, Bonakdari H (2020) Discussion of " Comparative assessment of time series and artificial intelligence models to estimate monthly streamflow: a local and external data analysis approach" by Saeid Mehdizadeh, Farshad Fathian, Mir Jafar Sadegh Safari and Jan F. Adamowski J Hydrol. https://doi.org/10.1016/j.jhydrol.2020.124614
35. Ednah MC (2018) Sustainable water use in lake Edward-George Basin: a case study of river mubuku-sebwe sub-catchments. MSc. Thesis, Department of Environmental Management, Makerere University
36. Onyutha C, Amollo CJ, Nyende J, Nakagiri A (2021) Suitability of averaged outputs from multiple rainfall-runoff models for hydrological extremes: a case of River Kafu catchment in East Africa. Int J Energy Water Res 5:43–56. https://doi.org/10.1007/s42108-020-00075-4
37. Onyutha C, Willems P (2018) Investigation of flow-rainfall co-variation for catchments selected based on the two main sources of River Nile. Stoch Environ Res Risk Assess 32:623–641. https://doi.org/10.1007/s00477-017-1397-9
38. Onyutha C (2020) From R-squared to coefficient of model accuracy for assessing "goodness-of-fits". Geosci Model Develop Discuss. https://doi.org/10.5194/gmd-2020-51
39. Tegegne G, Park DK, Kim Y (2017) Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. J Hydrol Reg Stud 14:49–66. https://doi.org/10.1016/j.ejrh.2017.10.002
40. Zhou Y, Guo S, Chang F (2019) Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts. J Hydrol 570:343–355. https://doi.org/10.1016/j.jhydrol.2018.12.040
41. Jiang B, Bamutaze Y, Pilesjö P (2014) Climate change and land degradation in Africa: a case study in the Mount Elgon region, Uganda. Geo-spatial Inf Sci 17:39–53. https://doi.org/10.1080/10095020.2014.889271
42. Barasa B, Kakembo V, Mwololo Waema T, Laban M (2017) Effects of heterogeneous land use/cover types on river channel morphology in the Solo River catchment, Eastern Uganda. Geocarto Int 32:155–166. https://doi.org/10.1080/10106049.2015.1132480
43. Kitutu MG, Muwanga A, Poesen J, Deckers JA (2009) Influence of soil properties on landslide occurrences in Bududa district, Eastern Uganda. Afr J Agric Res 4:611–620
44. Camberlin P (2009) Nile Basin climates. In: Dumont HJ (ed) The nile: Origin, environments, limnology and human use. Springer, Berlin, Germany, pp 307–333
45. Barasa B, Kakembo V, Mugagga F, Egeru A (2013) Comparison of extreme weather events and streamflow from drought indices and a hydrological model in River Malaba, Eastern Uganda. Int J Environ Stud 70:940–951. https://doi.org/10.1080/00207233.2013.862463
46. Mubialiwo A, Onyutha C, Abebe A (2020) Historical rainfall and evapotranspiration changes over mpologoma catchment

in uganda. Adv Meteorol 2020:1–19. https://doi.org/10.1155/2020/8870935

47. Van GA, Ndomba P, Yalew S, Kilonzo F (2012) Critical review of SWAT applications in the upper Nile Basin countries. Hydrol Earth Syst Sci 16:3371–3381. https://doi.org/10.5194/hess-16-3371-2012
48. Onyutha C, Willems P (2017) Space-time variability of extreme rainfall in the river Nile basin. Int J Climatol 37(14):4915–4924. https://doi.org/10.1002/joc.5132
49. Sheffield J, Goteti G, Wood EF (2006) Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modeling. J Clim 19:3088–3111. https://doi.org/10.1175/JCLI3790.1
50. Zhang Y, Zheng H, Chiew FHS, Peña-Arancibia J, Zhou X (2016) Evaluating regional and global hydrological models against streamflow and evapotranspiration measurements. Am Meteorol Soc 17(3):995–1010. https://doi.org/10.1175/JHM-D-15-0107.1
51. Ehret U, Zehe E, Wulfmeyer V, Warrach-Sagi K, Liebert J (2012) HESS opinions "should we apply bias correction to global and regional climate model data?" Hydrol Earth Syst Sci 16:3391–3404. https://doi.org/10.5194/hess-16-3391-2012
52. Nair S, Srinivasan G, Nemani R (2009) Evaluation of multi-satellite TRMM derived rainfall estimates over a western state of India. J Meteorol Soc Jpn 87:927–939. https://doi.org/10.2151/jmsj.87.927
53. Sharifi E, Saghafian B, Steinacker R (2018) Bias correction of satellite precipitation products based on concept of copula. Geophys Res Abstr 20
54. JICA (2011) The development study on water resources development and management for Lake Kyoga Basin in the Republic of Uganda. Kampala
55. Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 23rd national conference. Harvard College-Cambridge, Massachusetts, pp 517–524
56. Nielsen JP (1998) Multiplicative bias correction in kernel hazard estimation. Scand J Stat 25:541–553
57. Tian Y, Huffman GJ, Adler RF, Tang L, Sapiano M, Maggioni V, Wu H (2013) Modeling errors in daily precipitation measurements: additive or multiplicative? Geophys Res Lett 40:2060–2065. https://doi.org/10.1002/grl.50320
58. Saber M, Yilmaz KK (2018) Evaluation and bias correction of satellite-based rainfall estimates for modelling flash floods over the mediterranean region: application to Karpuz River Basin, Turkey. Water 10:1–24. https://doi.org/10.3390/w10050657
59. Piani C, Haerter JO, Coppola E (2010) Statistical bias correction for daily precipitation in regional climate models over Europe. Theor Appl Climatol 99:187–192. https://doi.org/10.1007/s00704-009-0134-9
60. Thiessen AH (1911) Precipitation averages for large areas. Mon Weather Rev 39:1082–1084
61. Hargreaves GH, Samani ZA (1982) Estimation of potential evapotranspiration. J Irrigs Drain Div 108:225–230
62. Hargreaves GH, Samani ZA (1985) Reference crop evapotranspiration from temperature. Trans Am Soc Agric Eng 1:96–99
63. Onyutha C, Acayo G, Nyende J (2020) Analyses of precipitation and evapotranspiration changes across the Lake Kyoga Basin in East Africa. Water 12:1–23. https://doi.org/10.3390/w12041134
64. Li C, Wang H, Liu J, Yan D, Yu F, Zhang L (2010) Effect of calibration data series length on performance and optimal parameters of hydrological model. Water Sci Eng 3:378–393. https://doi.org/10.3882/j.issn.1674-2370.2010.04.002
65. Birhanu D, Kim H, Jang C, Park S (2018) Does the complexity of evapotranspiration and hydrological models enhance robustness ? Sustainability 10:1–34. https://doi.org/10.3390/su10082837
66. Chelangat C, Abebe A (2021) Reservoir operation for optimal water use of Kabalega reservoir in Uganda. Int J Energy Water Resour. https://doi.org/10.1007/s42108-020-00109-x
67. Goodarzi MS, Amiri BJ, Azarneyvand H, Khazaee M, Mahdianzadeh N (2020) Assessing the performance of a hydrological tank model at various spatial scales. J Water Manag Model 29:1–8. https://doi.org/10.14796/JWMM.C472
68. Pérez-Sánchez J, Senent-aparicio J, Segura-Méndez F, Pulido-velazquez D, Srinivasan R (2019) Evaluating hydrological models for deriving water resources in Peninsular Spain. Sustainability 11(2872):1–36. https://doi.org/10.3390/su11102872
69. Tiwari H, Balvanshi A (2020) Hydrological modelling of Bina River Basin in Madhya Pradesh, India. Glob J Eng Sci 6:1–7. https://doi.org/10.33552/GJES.2020.06.000639
70. Podger G (2004) User guide-rainfall runoff library, catchment modelling toolkit. CRC for Catchment Hydrology, Australia
71. Link KG, Stobb MT, Di PJ, Neeves KB, Fogelson AL, Sindi SS, Leiderman K (2018) A local and global sensitivity analysis of a mathematical model of coagulation and platelet deposition under flow. PLoS ONE 13:1–38. https://doi.org/10.1371/journal.pone.0200917
72. Nash J, Sutcliffe JV (1970) River flow forecasting through conceptual models: part I—a discussion of principles. Hydrology 10:282–290
73. Amir MSII, Khan MMK, Rasul MG, Sharma RH, Akram F (2013) Automatic multi-objective calibration of a rainfall runoff model for the Fitzroy Basin, Queensland, Australia. Int J Environ Sci Dev 4:313–315
74. Duan QY, Gupta VK, Sorooshian S (1993) Shuffled complex evolution approach for effective and efficient global minimization. J Optim Theory Appl 76:501–521
75. Beven KJ, Binley AM (1992) The future role of distributed models: Model calibration and predictive uncertainty. Hydrol Process 6:279–298
76. Krause P, Boyle DP, Bäse F (2005) Comparison of different efficiency criteria for hydrological model assessment. Adv Geosci 5:89–97. https://doi.org/10.5194/adgeo-5-89-2005
77. Willmott CJ (1981) On the validation of models. Phys Geogr 2:184–194. https://doi.org/10.1080/02723646.1981.10642213
78. Bartlett RF (1993) Linear modelling of pearson's product moment correlation coefficient: an application of Fisher's Z transformation. J R Stat Soc 42:45–53
79. Gupta HV, Kling H, Yilmaz KK, Martinez GF (2009) Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J Hydrol 377:80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003
80. Santhi C, Arnold JG, Williams JR, Dugas WA, Srinivasan R, Hauck LM (2001) Validation of the SWAT model on a large river basin with point and nonpoint sources. J Am Water Resour Assoc 37:1169–1188. https://doi.org/10.1111/j.1752-1688.2001.tb03630.x
81. Onyutha C (2016) Influence of hydrological model selection on simulation of moderate and extreme flow events : a case study of the Blue Nile Basin. Adv Meteorol 2016:1–28. https://doi.org/10.1155/2016/7148326
82. Knoben WJM, Freer JE, Woods RA (2019) Technical note: inherent benchmark or not? comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol Earth Syst Sci 23:4323–4331. https://doi.org/10.5194/hess-23-4323-2019
83. Devkota S, Shakya MN, Sudmeier-rieux K, Jaboyedoff M, Van WCJ, Mcadoo BG, Adhikari A (2018) Development of monsoonal rainfall empirical model for data-scarce situations: the case of the central-Western Hills (Panchase region) of Nepal. Hydrology 5:1–27. https://doi.org/10.3390/hydrology5020027

84. Nhat LM, Tachikawa Y, Takara K (2006) Establishment of Intensity-duration-frequency curves for precipitation in the Monsoon Area of Vietnam. Ann Disaster Prev Res Inst Kyoto Univ 49:93–103
85. Onyutha C (2012) Statistical modelling of FDC and return periods to characterise QDF and design of threshold of hydrological extremes. J Urban Environ Eng 6:132–148. https://doi.org/10.4090/juee.2012.v6n2.132148
86. Insitute of Hydrology (1980) Low flow studies report1. Insitute of Hydrology, Wallingford, United Kingdom
87. Onyutha C, Willems P (2013) Uncertainties in flow-duration-frequency relationships of high and low flow extremes in Lake Victoria Basin. Water 5:1561–1579. https://doi.org/10.3390/w5041561
88. Pickands J (1975) Statistical inference using extreme order statistics. Ann Stat 3:119–131
89. Lang M, Ouarda TBMJ, Bobée B (1999) Towards operational guidelines for over-threshold modeling. J Hydrol 225:103–117. https://doi.org/10.1016/S0022-1694(99)00167-5
90. Langbein W (1949) Annual floods and the partial-duration flood series. Trans Am Geophys Union 30:879–881
91. Far SS, Wahab AKA (2016) Evaluation of peaks-over-threshold method. Ocean Sci Discuss. https://doi.org/10.5194/os-2016-47
92. Willems P (2009) Environmental modelling & software a time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. Environ Model Softw 24:311–321. https://doi.org/10.1016/j.envsoft.2008.09.005
93. Onyutha C (2017) On rigorous drought assessment using daily time scale: Non-stationary frequency analyses, revisited concepts, and a new method to yield non-parametric indices. Hydrology 4(48):1–43. https://doi.org/10.3390/hydrology4040048
94. Segers J (2005) Generalized Pickands estimators for the extreme value index. J Stat Plan Inference 128:381–396. https://doi.org/10.1016/j.jspi.2003.11.004
95. Baig MRI, Ahmad IA, Shahfahad TM, Rahman A (2020) Annals of GIS Analysis of shoreline changes in Vishakhapatnam coastal tract of Andhra Pradesh, India: an application of digital shoreline analysis system (DSAS). Ann GIS 26:361–376. https://doi.org/10.1080/19475683.2020.1815839
96. Hill BM (1975) Institute of mathematical statistics is collaborating with JSTOR to digitize, preserve, and extend access to the annals of statistics. ® www.jstor.org. Ann Stat 3:1163–1174
97. Schulz K, Bernhardt M (2016) The end of trend estimation for extreme floods under climate change? Hydrol Process 30:1804–1808. https://doi.org/10.1002/hyp.10816
98. Jaiswal RK, Ali S, Bharti B (2020) Comparative evaluation of conceptual and physical rainfall–runoff models. Appl Water Sci 10:1–14. https://doi.org/10.1007/s13201-019-1122-6
99. Staudinger M, Stahl K, Seibert J, Clark MP, Tallaksen LM (2011) Comparison of hydrological model structures based on recession and low flow simulations. Hydrol Earth Syst Sci 15:3447–3459. https://doi.org/10.5194/hess-15-3447-2011
100. Li H, Beldring S, Xu C (2015) Stability of model performance and parameter values on two catchments facing changes in climatic conditions. Hydrol Sci J 60:1317–1330. https://doi.org/10.1080/02626667.2014.978333
101. Zhang G, Xie T, Zhang L, Hua X, Liu F (2017) Application of multi-step parameter estimation method based on optimization algorithm in sacramento model. Waste Manag 9:1–21. https://doi.org/10.3390/w9070495
102. Yu B, Zhu Z (2015) A comparative assessment of AWBM and SimHyd for forested watersheds. Hydrol Sci J 60(7–8):1200–1212. https://doi.org/10.1080/02626667.2014.961924
103. Bashar KE (2012) Comparative performance of soil moisture accounting approach in continuous hydrologic simulation of the Blue Nile. Nile Basin Water Sci Eng J 5:1–10
104. Jakeman AJ, Hornberger GM (1993) How much complexity is warranted in a rainfall-runoff model? Water Resour Res 29:2637–2649
105. Horner I, Renard B, Le CJ, Branger F, McMillan HK, Pierrefeu G (2018) Impact of stage measurement errors on streamflow uncertainty. Water Resour Res 54:1952–1976. https://doi.org/10.1002/2017WR022039

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.