Water Resources Management (2019) 33:19–37
https://doi.org/10.1007/s11269-018-2085-3

Relationship Between Calibration Time and Final Performance of Conceptual Rainfall-Runoff Models

Adam P. Piotrowski[1] • Jaroslaw J. Napiorkowski[1] • Marzena Osuch[1]

Received: 29 March 2018 / Accepted: 13 August 2018 /
Published online: 22 August 2018

Abstract

Various methods are used in the literature for calibration of conceptual rainfall-runoff models. However, very rarely the question on the relation between the number of model runs (or function calls) and the quality of solutions found is asked. In this study two lumped conceptual rainfall-runoff models (HBV and GR4J with added snow module) are calibrated for five catchments, located in temperate climate zones of USA and Poland, by means of three modern variants of Evolutionary Computation and Swarm Intelligence optimization algorithms with four different maximum numbers of function calls set to 1000, 3000, 10,000 and 30,000. At the calibration stage, when more than 10,000 function calls is used, only marginal improvement in model performance has been found, irrespective of the catchment or calibration algorithm. For validation data, the relation between the number of function calls and model performance is even weaker, in some cases the longer calibration, the poorer modelling performance. It is also shown that the opinion on the model performance based on different popular hydrological criteria, like the Nash-Sutcliffe coefficient or the Persistence Index, may be misleading. This is because very similar, largely positive values of Nash-Sutcliffe coefficient obtained on different catchments may be accompanied by contradictory values of the Persistence Index.

Keywords Rainfall-runoff modelling · HBV · GR4J · Model calibration · Differential Evolution · Particle swarm optimization

1 Introduction

Although relations between elements of the hydrological cycle are well understood, their quantitative description is still a challenge. Many hydrologists are interested in processes that describe a relation between precipitation and runoff at the catchment scale. However, depending

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11269-018-2085-3) contains supplementary material, which is available to authorized users.

✉ Adam P. Piotrowski
adampp@igf.edu.pl

[1] Institute of Geophysics, Polish Academy of Sciences, Ks. Janusza 64, 01-452 Warsaw, Poland

on many meteorological, land cover and soil-related factors that are variable at small spatial and temporal scales, the quantitative description of the processes that relates precipitation to runoff at the catchment is difficult. It is frequently assumed that models which could be used in practice cannot be too detailed (Bergström 1991) and must roughly approximate many physical processes that occur in the catchment-scale environment – such models are frequently called "conceptual" rainfall-runoff models. If the models simplify all processes at the catchment scale, they are called "lumped". It turns out that for rainfall-runoff modelling lumped conceptual models are considered to perform not worse than more complex distributed models (Vansteenkiste et al. 2014; Lobligeois et al. 2014), even though they are vulnerable to temporal resolution (Jie et al. 2018) and variability of hydro-meteorological conditions (Poncelet et al. 2017). Although conceptual models are much simpler and require much less information than physical ones, they still have some theoretical background, contrary to purely data based-models like artificial neural networks, stochastic transfer functions or nearest neighbourhood search approaches (Pechlivanidis et al. 2011). However, irrespective of which model is used, it needs to be calibrated. Even if model parameters have a physical representation, they almost never can be precisely measured (Beven 2012). The problem of rainfall-runoff model calibration, considering various sources of uncertainty or not, has been discussed in various reviews (Beven 2012; Refsgaard 1997). In this study, attention is focused on single-objective automatic Swarm Intelligence or Evolutionary Algorithms that aim at finding the best set of parameters for a given model and catchment and become very popular in water-related studies (Tayfur 2017). Even though it is well known that due to the concept of equifinality (Beven 2012) the solution found would strongly depend on the model representation, objective function or data, and would not have "general" meaning (Vrugt et al. 2008; Merz et al. 2011; Osuch et al. 2015), such approach to model calibration remains widespread, as discussed below.

During the last decade a number of papers aimed at comparison among optimization algorithms applied to rainfall-runoff model calibration have been published (Goswami and O'Connor 2007; Arsenault et al. 2014, Piotrowski et al. 2017a). Such studies generally showed that many algorithms perform similarly well and no best method may be determined. However, the performance of optimization algorithm depends largely on the number of allowed function calls (or model runs). The number of function calls is a measure of method's computational time independent of the code, computer or language used. As shown in two recent methodological papers (Posik et al. 2012; Piotrowski et al. 2017b) the ranking of optimization algorithms highly depends on the number of available function calls. This is inevitably linked with the problem of computational efficiency. Unfortunately, in hydrological modelling the problem of proper setting the number of function calls is rarely discussed. In many papers, the number of model runs used by calibration procedure is even not clearly stated. If the number of function calls is given in the paper, values used may vary severely. For example, a million function calls were used to calibrate air-to-stream temperature models with four to eight parameters by means of Particle Swarm Optimization (PSO) in Toffolon and Piccolroaz (2015), what seems a large exaggeration. As many as five million function calls are used in Bi et al. (2016) for water distribution system optimization problems, however, this time such large number could be justified by the fact that even 1000-dimensional problems were tackled. On the other hand, Wang et al. (2010) set the number of function calls to only 100 when using a simple Genetic Algorithm and Shuffled Complex Evolution (SCE-UA) method for calibration of grid-based distributed rainfall–runoff model with four parameters. Such examples show how large differences may be spotted in the literature, often without any justification.

In some hydrological studies, the relation between the performance of the conceptual rainfall-runoff model and the number of function calls is considered, often referring to the graphically-illustrated convergence speed (Tolson and Shoemaker 2007; Arsenault et al. 2014; Piotrowski et al. 2017a). In such figures the relation between the quality measure of the best solution found so far is plotted against the already performed number of function calls. However, from such illustrations published in various studies aiming at rainfall-runoff modeling different conclusions could be drawn. From Arsenault et al. (2014) one may learn that some algorithms initially converge quicker than the others, but such methods often perform worse than the best ones at the end of the search. A similar result was obtained by Jeon et al. (2014), but for another goal, namely calibration of long-term hydrologic impact assessment model. On the contrary, although in Tolson and Shoemaker (2007) and Piotrowski et al. (2017a) also differences in convergence speed between various algorithms applied for calibration of conceptual rainfall-runoff models are noted, finally vast majority of methods reach almost equal performance. This may suggest that most "good" methods may lead to the solutions of equal quality if the number of function calls is large enough, what again puts attention to both the concepts of equifinality (Beven 2012) and algorithm efficiency.

In Arsenault et al. (2014) study one may also find another approach to this topic. It shows after how many function calls the particular algorithm on average reaches the 95% of the best value of the Nash-Sutcliffe coefficient (Nash and Sutcliffe 1970) found during the whole search (e.g. 25,000 function calls). It turns out that such close-to-the-best results are often found during the first half of the search, and in about 50% of cases – during the first 5000 function calls. Hence, is the longer search just a waste of time?

Studies discussed above addressed the problem of how fast the particular method converges when the predefined number of allowed function calls is preset and anyway used. Would such results be similar to those obtained when different maximum numbers of function calls are assumed? For example, when setting the maximum number of function calls to 20,000, Jeon et al. (2014) found that the variant of Genetic Algorithm they apply perform better than SCE-UA method (Duan et al. 1992) until 5150 function evaluations is used, but SCE-UA is better afterwards. Does it means that when one sets the maximum number of function calls to 5000, Genetic Algorithm would still perform better? Here two points have to be addressed.

First, best solution found so far by the algorithm cannot deteriorate during a run. However, if a number of runs is performed with some number of function calls, and then the same number of runs is repeated independently with a larger number of function calls, during repeated runs better solution may be found sooner, or only worse solutions may be found. Hence, if someone performs 100 runs for, say, 10,000 function calls, never the average performance after 5000 calls would be better than those after 10,000 calls (and almost always solutions found after 10,000 calls *will* be better). However, if one performs 100 runs with 5000 function calls each, and then independently performs 100 runs with 10,000 function calls each, it *may* happen that the average performance found when 5000 function calls were allowed turns out *better* than in the case when the number of function calls was set to 10,000. This may even be not rare for problems on which algorithms are easily trapped in a local optimum, hence additional function calls are simply wasted.

Second, if the optimization algorithm is developed in a "smart" way, its behaviour at a particular stage of the run would differ, depending on the remaining computational budget. Such algorithms search differently when the number of allowed function calls is low then when it is high. For example, in variants of PSO with inertia weight (Shi and Eberhart 1998), the inertia weight is decreasing linearly with time such that the lower the maximum number of function calls is, the quicker the decrease of inertia weight. In Successful-history based Differential Evolution with linear population size reduction (L-SHADE, Tanabe and Fukunaga

2014), a winner of IEEE Conference on Evolutionary Computation (IEEE CEC) in 2014, population size is linearly reduced with the number of function calls from 18·D (where D is the problem dimensionality) at the beginning to just four individuals at the end of the search. Also, Dynamically Dimensioned Search (Tolson and Shoemaker 2007), developed for hydrological applications, has been designed in order to quicken the search when the computational budget is low and performs more explorations when it is high. It results in different behavior, and hence relatively small differences in final results achieved, when the preset number of function calls varied between 200 and 2000 (see Tolson and Shoemaker 2007).

Hence, to verify how many function calls are needed to solve efficiently the problem of conceptual rainfall-runoff calibration when advanced Evolutionary Algorithms are used, instead of referring to plots showing convergence speed, *independent* tests must be performed with different numbers of function calls. Moreover, it needs to be verified whether any relation found between the number of function calls used and the model performance holds not only for the data used for calibration but also for validation ones. This is the main goal of the present study, in which two lumped conceptual rainfall-runoff models, HBV (Bergström 1976; Lindström et al. 1997) and GR4J (Perrin et al. 2003) with snow module (Piotrowski et al. 2017a), are calibrated by means of three optimization algorithms with four different maximum numbers of function calls, set to 1000, 3000, 10,000 and 30,000 (each about three times larger than the previous one). The aim of this paper is to find a relationship between the modeling performance and the number of function calls when the advanced optimization procedures are used.

In addition to the main goal, differences between the perception of the same results depending on the hydrological criterion used are emphasized. Among a few criteria applied in the paper, we put attention to two specific ones, namely Persistence Index (PI) (Kitanidis and Bras 1980) and Nash-Sutcliff coefficient (NSC) (Nash and Sutcliffe 1970). According to both PI and NSC criteria, the best fit is obtained when NSC or $PI = 1$. However, each criterion measures something else. Using the model is a better choice than assuming that the forecasted runoff will be the same as the recent observation in the case of PI, or the same as long-term mean in the case of NSC, as long as the value of the criterion is above 0. As both criteria are frequently used, this paper researches how much they may differ in practice.

2 Models, Study Sites, Data and Methods

In this study two lumped conceptual rainfall-runoff models, HBV and GR4J, are used. For the description of both models, see main source papers: Bergström (1976) and Lindström et al. (1997) in case of the HBV and Perrin et al. (2003) in case of the GR4J. A specific version of the HBV (with 13 parameters) and the GR4J with snow module (with seven parameters, four from the basic GR4J and three from snow module) that are used in this study are discussed in Piotrowski et al. (2017a). Both models have found numerous practical applications (Lindström et al. 1997; Beven 2012; Tian et al. 2013).

Both models are applied at the daily time scale. The river runoff (y_{t+1}) at time $t+1$ is simulated based on the precipitation (R_t), air temperature (T_t) and evapotranspiration (E_t) (related to air temperature according to Hamon's (1961) method) data from previous day t. Hydrological data are always collected from the single gauge station in this study, but, depending on the catchment considered, the meteorological data may come from a single meteorological station located within the catchment, or a few different stations located within, or close to the catchment. Time series from each catchment are divided into the calibration and the validation sets. Calibration set, composed of roughly 70% of available data, is used during model optimization; the validation set is used only to verify the quality of calibrated models on the independent data. However, the first year (365 days) of the calibration data is considered as a warm-up period and is not used to compute the value of the objective function.

In this study, runoff values simulated by HBV or GR4J models are either considered directly (such results are called "raw" further in the paper), or after applying error correction procedure (Refsgaard 1997; Madsen and Skotner 2005; Liu et al. 2016). In the second case, after termination of the calibration procedure, the raw results from the HBV and the GR4J models are updated by means of linear regression with exogenous inputs; in this version, the past forecasts from the raw HBV or the raw GR4J predictions are added as exogenous inputs to the linear regression error model:

$$\varepsilon^s_{t+1} = f\left(\varepsilon^s_t, \varepsilon^s_{t-1}, \dots, \varepsilon^s_{t-\delta+1}, y^m_{t+1}, y^m_t, \dots, y^m_{t-\delta}\right) \tag{1}$$

where m (model) denotes the HBV or the GR4J and $\varepsilon^s_t = y_t - y^m_t$ is the prediction error. The final simulated runoff is calculated as $y^s_{t+1} = y^m_{t+1} + \varepsilon^s_{t+1}$. Here δ is set to 3 days for both the HBV and the GR4J.

Five catchments are considered in this study: the lowland Suprasl (Poland) and the Irondequoit Creek (New York state, USA), the hilly Fanno Creek (Oregon, USA), and the mountainous Biala Tarnowska (Poland) and the Cedar River (Washington state, USA). Their basic characteristics and descriptions are given in Table 1 (data sources for USA catchments are: US Geological Survey (USGS) and National Centers for Environmental Information, National Oceanic and Atmospheric Administration (NOAA); for Polish catchments: Institute of Meteorology and Water Management (IMGW)). Table 1 also includes detailed, site-specific information on data sets used, gaps in data, and splitting observation series into the calibration and validation sets.

In case of the Biala Tarnowska and the Cedar River catchments, the lumped precipitation and air temperature data series are obtained by Thiessen Polygons (Thiessen and Alter 1911) method from measurements collected at three and five stations, respectively (see Table 1). In case of the Suprasl River, the air temperature data come from Bialystok station only, but lumped precipitation data set is based on measurements from five meteorological stations. For the Fanno Creek and the Irondequoit Creek, all meteorological data come from a single meteorological station.

Using just one calibration method could bias the results. Hence, calibration of each model for every catchment and with each considered number of function calls is performed by three optimization algorithms: Modified Differential Evolution with p-best crossover (MDE_pBX, Islam et al. 2012), Successful parents selecting L-SHADE with eigenvector-based crossover (SPS-L-SHADE-EIG, Guo et al. 2015) and Genetic Learning Particle Swarm Optimization (GLPSO, Gong et al. 2016). Control parameters of all three optimization algorithms are set as suggested in the source papers (in the case of SPS-L-SHADE-EIG, the so called "default" variant of control parameter settings has been used, see Guo et al. 2015). Population size has been set to 100 for MDE_pBX (Islam et al. 2012; Piotrowski 2017), 50 for GLPSO (Gong et al. 2016) and follows linear reduction scheme from 19D (where D is a problem dimensionality; $D = 13$ for the HBV and 7 for the GR4J) to 4 in SPS-L-SHADE-EIG (Guo et al. 2015). Note that in SPS-L-SHADE-EIG algorithm the values of population size are related to the computational budget, hence the method *will* behave differently when large and small numbers of maximum function calls are preset.

To verify the impact of the number of allowed function calls, all calibration experiments are repeated with four different values of function calls, set to 1000, 3000, 10,000 and 30,000. The

Table 1 Brief description of five considered catchments

Catchment name	Gauge name	Calibration/validation data periods	Location of the gauge station	Meteorological stations/percentage of catchment coverage according to Thiessen Polygons method	Catchment size (km^2)	Elevation (lowest – highest point) (m a.s.l.)	Description of catchments, climatic conditions and data sets
Biala Tarnowska	Koszyce Wielkie	01.01.1971–31.12.1999 / 01.01.2000–30.10.2010	Poland 49°58′48″N 20°56′39″E	Biecz/18% Krynica/22% Nowy Sacz/19% Tarnow/38% Wysowa/3% (percentages refer to both precipitation and air temperature)	957	208–997 (mountainous)	Southern and central parts are mountainous, northern part relatively flat. Wet summers with daily mean air temperatures ≈ 19 °C at the gauge; snowy winters with daily mean air temperatures ≈ −2 °C at the gauge; colder in the mountainous south. Catchment covered mostly by mountains, forests, pastures and agriculture. Snow on the ground mainly between December and March. There are no gaps in the data.
Suprasl	Zaluki (only upper part of large catchment is considered)	06.01.1977–31.12.1992 / 01.01.1993–31.10.2000	Poland 53°09′29″N 23°32′42″E	Bialystok/0% Zabludow/3% Grodek/95% Suprasl/0% Szudzialowo/2% (percentages refer to precipitation only, air temperature available only from Bialystok)	345	≈ 150 (flat lowland)	Mostly dry summers with occasional rainfalls, daily mean air temperatures ≈ 17 °C; snowy winters with daily mean air temperatures ≈ −3 °C; uniform climatic conditions in the whole catchment. Covered by forests, pastures and agriculture; snow on the ground mainly between November and March. There are no gaps in data sets.
Cedar River	Renton, WA (USGS 12119000)	01.12.2002–30.04.2009 / 01.07.2009–31.01.2017	Washington state, USA	Seattle–Tacoma International Airport WA	477	17–1780 (mountainous)	Catchment covered mostly by mountains and forests in the east, agriculture and residential areas in the west. Wet summers with daily

Table 1 (continued)

Catchment name	Gauge name	Calibration/validation data periods	Location of the gauge station	Meteorological stations/percentage of catchment coverage according to Thiessen Polygons method	Catchment size (km^2)	Elevation (lowest – highest point) (m a.s.l.)	Description of catchments, climatic conditions and data sets
			47°29′12″N 122°11′43″ W	(USW00024233) /9% Cedar Lake WA (USC00451233) /37% Landsburg WA (USC00454486) /54% (percentages refer to both precipitation and air temperature)			mean temperatures ≈ 19 °C at the gauge; mild winters with daily mean air temperatures ≈ 5 °C at the gauge; much colder in the eastern, mountainous part. Snow on the ground at higher elevations mainly from November to April, but rare in the lower part of the catchment. The catchment is partly regulated by Chester Moor Lake. Many data are lacking for May and June 2009, hence the gap between calibration and validation period. However, the model was computing runoff also for that period using approximated data, to avoid another run-off period. Excluding May and June 2009, there are no gaps in runoff data. However, there are gaps in three meteorological stations; together over 60 precipitation values and almost 120 air temperature values are missing. The longest gap in precipitation data is 9 days long for Landsburgh, 2 days long for Cedar Lake and 1 day long for Seattle-Tacoma Int'l Airport. In

Table 1 (continued)

Catchment name	Gauge name	Calibration/validation data periods	Location of the gauge station	Meteorological stations/percentage of catchment coverage according to Thiessen Polygons method	Catchment size (km^2)	Elevation (lowest – highest point) (m a.s.l.)	Description of catchments, climatic conditions and data sets
							air temperature data there are no gaps for Seattle-Tacoma Int'l Airport, the longest gap for Cedar Lake is 5 days long and for Landsburg there is a large, 60-days long gap. For particular station gaps are filled, according to data availability by one of the following methods: 1. using linear approximation from two other stations, 2. by linear approximation among the closest previous and following measurements at the same station, 3. (only in case of precipitation data) by setting precipitation to 0.
Fanno Creek	Durham, OR (USGS 14206950)	07.09.2002–31.12.2011 / 01.01.2012–03.03.2017	Oregon state, USA 45°24′13”N 122°45′13”W	Portland International Airport (USW00024229) (the only meteorological station used for this catchment)	82	35–320 (hilly)	Wet summers with daily mean temperatures ≈ 21 °C; mild winters with daily mean air temperatures ≈ 5 °C. Climatic conditions are relatively uniform in the whole catchment. Catchment covered mostly by residential areas, both in its hilly and flat parts. Snow on the ground is very rare. There are no gaps in the data.
Irondequoit Creek		01.10.1995–31.12.2010 /	New York state, USA	Rochester International	368	≈ 155 (flat lowland)	Precipitation uniform through the year, daily mean air temperatures

Table 1 (continued)

Catchment name	Gauge name	Calibration/validation data periods	Location of the gauge station	Meteorological stations/percentage of catchment coverage according to Thiessen Polygons method	Catchment size (km^2)	Elevation (lowest – highest point) (m a.s.l.)	Description of catchments, climatic conditions and data sets
	Penfield, NY (USGS 0423205010)	01.01.2011–31.01.2017	43°08′42”N 77°30′43”W	Airport (USW00014768) (the only meteorological station used for this catchment)			≈ 22 °C; snowy winters with daily mean air temperatures ≈ −4 °C; uniform climatic conditions in the whole catchment. Catchment covered mostly by residential areas and agriculture. Snow on the ground mainly between December and March. Apart from air temperature measurement at a single day (which is filled by setting mean value from previous and following observation) there are no gaps in the data.

idea is that in various tests the numbers of function calls differ roughly by a factor of three. Tests are independent, what is especially important for SPS-L-SHADE-EIG that scale population size with the remaining number of function calls.

In this study, all tests are repeated 30 times, each time with different, randomly generated initial populations. This gives a sample of 30 solutions for each considered variant ("variant" means, for example, calibration of the raw HBV model by means of GLPSO for the Cedar River with the maximum number of function calls set to 30,000). The total number of variants considered in this paper is 240: 2 models × 2 error correction procedure (used or not) × 3 optimization algorithms × 4 maximum numbers of function calls × 5 catchments.

Three criteria are considered in this study:

- *MSE*: mean square error (being the objective function used for calibration):

$$MSE = \frac{1}{N}\sum_{t=1}^{N}\left(y_t - y_t^s\right)^2 \tag{2}$$

where y_t^s is a simulated runoff for time t, y_t is a measured value of runoff, and N is the number of daily data in the particular data set (calibration or validation) for particular catchment;

- *NSC*: the Nash-Sutcliffe coefficient (Nash and Sutcliffe 1970):

$$NSC = 1 - \frac{\sum_{t=1}^{N}\left(y_t - y_t^s\right)^2}{\sum_{t=1}^{N}\left(y_t - \overline{y_t}\right)^2} \tag{3}$$

where $\overline{y_t}$ is a mean of N measured runoff values;

- Persistence Index coefficient (Kitanidis and Bras 1980):

$$PI = 1 - \frac{\sum_{t=L+1}^{N}\left(y_t - y_t^s\right)^2}{\sum_{t=L+1}^{N}\left(y_t - y_{t-L}\right)^2} \tag{4}$$

where L is a lead time, set to 1 in this study.

To compare the results the 30-runs mean values of every criterion for each variant are computed, together with associated standard deviations. The *MSE* is used for calibration as the objective function, *NCS* and *PI* are computed after each model is calibrated.

3 Results

Detailed results, including 30-run mean and standard deviation values of *MSE*, *PI* and *NSC* criteria for all tested variants, are given in Supplementary Tables 1–10 that are available online as Supplementary Material. To facilitate reading, the relation between the number of function calls and the model performance is presented in Figs. 1, 2, 3, 4 and 5, where values of *MSE*, *PI*,

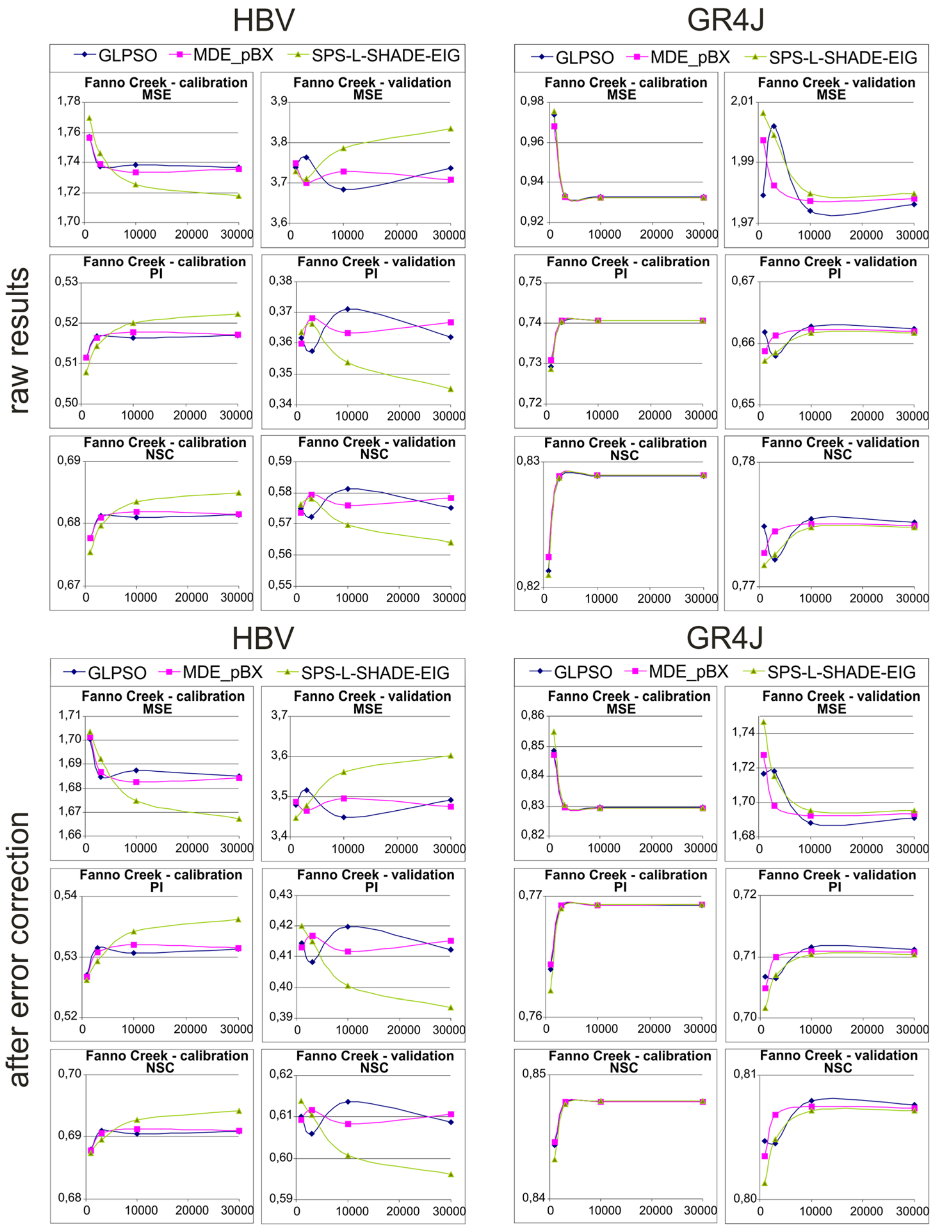

Fig. 1 Relation between the performance of HBV (left two columns) and GR4J (right two columns) models (vertical axis) without (upper three rows) and with (lower three rows) error correction procedure for Fanno Creek, Oregon, USA, and the maximum number of function calls used during calibration (horizontal axis). MSE (mean square error), PI (persistence index) and NSC (Nash-Sutcliffe coefficient) obtained when the maximum numbers of function calls were set to 1000, 3000, 10.000 and 30.000 (experiments were performed independently). Each experiment was repeated 30 times; 30-runs average values are shown

and *NSC* obtained for both the calibration and the validation data are shown. Due to space limitation, the results obtained with the use of error correction procedure are illustrated

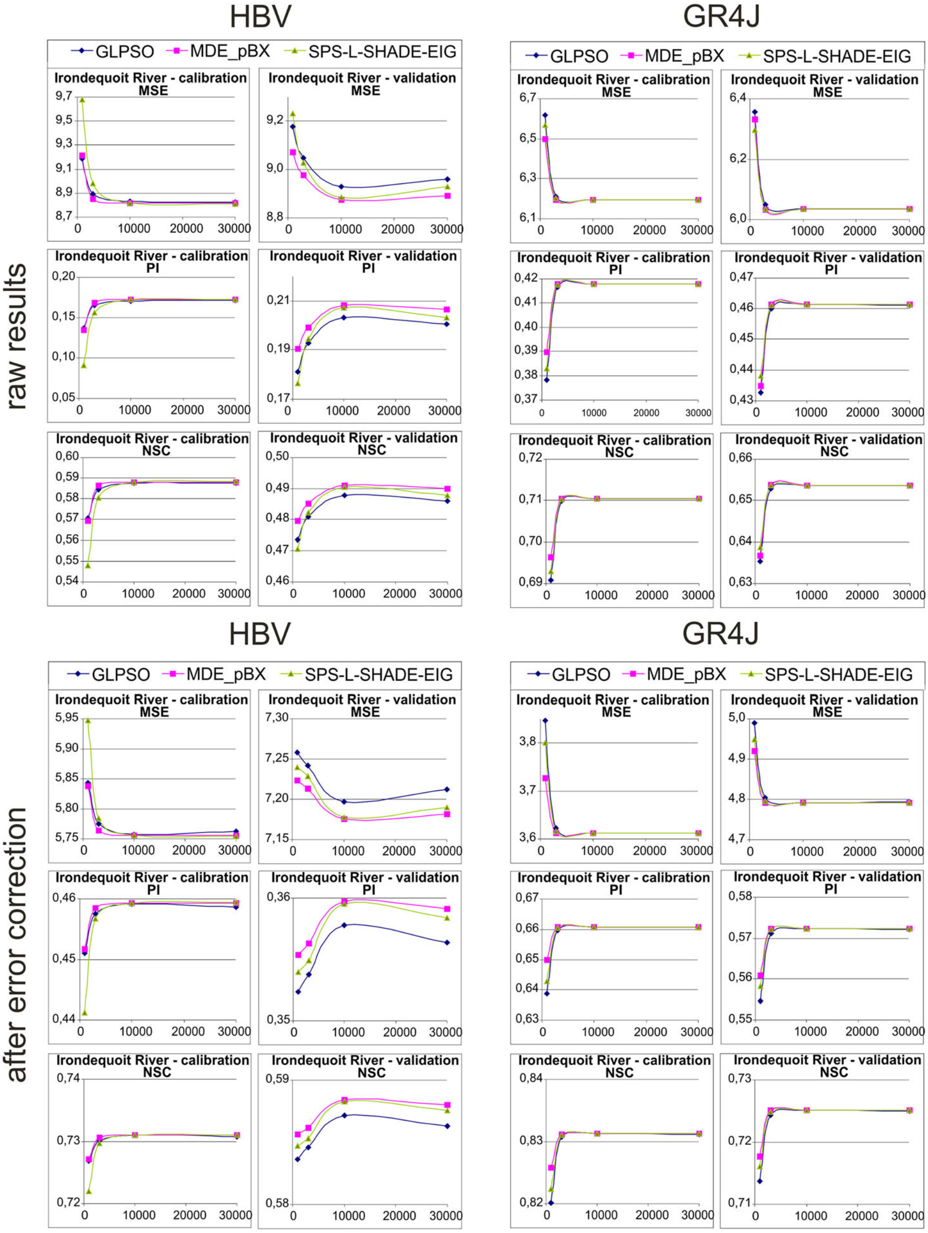

Fig. 2 Relation between the performance of HBV (left two columns) and GR4J (right two columns) models (vertical axis) without (upper three rows) and with (lower three rows) error correction procedure for Irondequoit River, New York, USA, and the maximum number of function calls used during calibration (horizontal axis). MSE (mean square error), PI (persistence index) and NSC (Nash-Sutcliffe coefficient) obtained when the maximum numbers of function calls were set to 1000, 3000, 10.000 and 30.000 (experiments were performed independently). Each experiment was repeated 30 times; 30-runs average values are shown

graphically only for two selected catchments (Fanno Creek, Fig. 1, and Irondequoit Creek, Fig. 2); results for all catchments are given in Supplementary Tables 1–10.

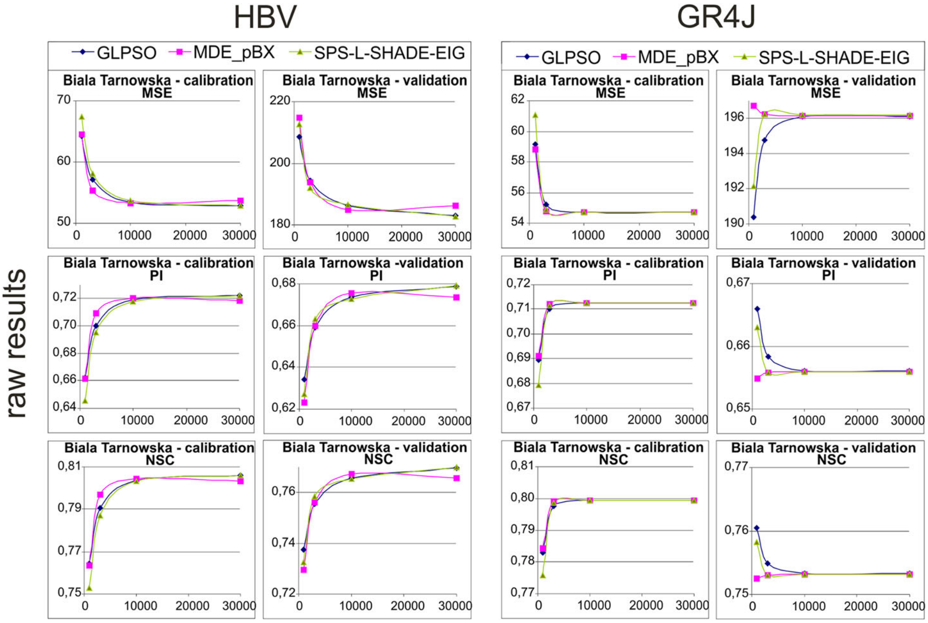

Fig. 3 Relation between the performance of raw HBV (left two columns) and GR4J (right two columns) models (vertical axis) and the maximum number of function calls used during calibration (horizontal axis). Results for the Biala Tarnowska catchment obtained when the maximum numbers of function calls were set to 1000, 3000, 10.000 and 30.000 (experiments were performed independently). Each experiment was repeated 30 times; 30-runs average values are shown

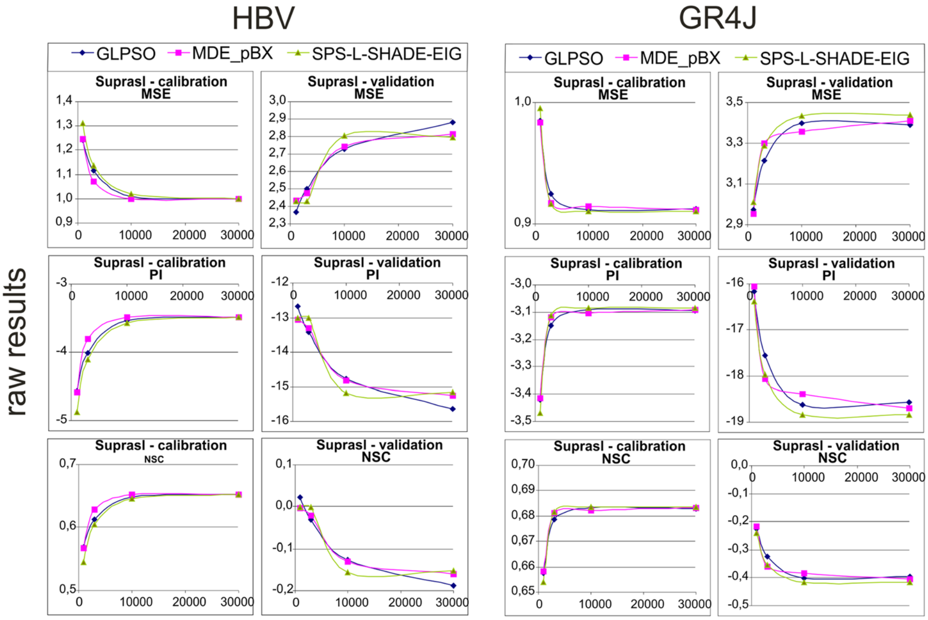

Fig. 4 Relation between the performance of raw HBV (left two columns) and GR4J (right two columns) models (vertical axis) and the maximum number of function calls used during calibration (horizontal axis). Results for the Suprasl catchment obtained when the maximum numbers of function calls were set to 1000, 3000, 10.000 and 30.000 (experiments were performed independently). Each experiment was repeated 30 times; 30-runs average values are shown

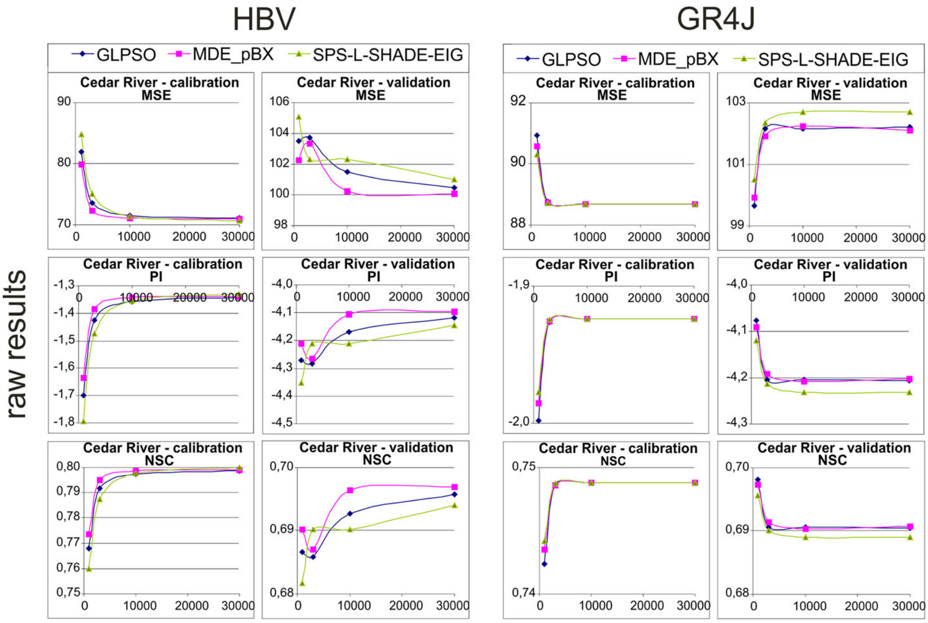

Fig. 5 Relation between the performance of raw HBV (left two columns) and GR4J (right two columns) models (vertical axis) and the maximum number of function calls used during calibration (horizontal axis). Results for the Cedar River catchment obtained when the maximum numbers of function calls were set to 1000, 3000, 10.000 and 30.000 (experiments were performed independently). Each experiment was repeated 30 times; 30-runs average values are shown

3.1 Number of Function Calls Versus Model Performance

Anyone who expects that the longer conceptual model is calibrated, the better results are obtained may be disappointed by our results. Based on performed experiments, this is not true for the independent data, even if often holds for the calibration ones.

In the case of the Fanno Creek (OR, USA) results much differ for the calibration and the validation data sets (see Fig. 1 and Suppl. Tables 1–2). At calibration stage, the performance of the HBV and the GR4J models, both raw and with error correction procedure, is better when 10,000 function calls were allowed than when their values were limited to 1000 or 3000. These results rather do not depend on the calibration algorithm. However, using 30,000 function calls do not improve results for the calibration data when GLPSO or MDE_pBX calibration methods are used. SPS-L-SHADE-EIG is able to take advantage of additional time, but the improvement is observed only for the raw HBV or the raw GR4J models.

The picture gets worse during model validation (see Fig. 1 and Suppl. Tables 1–2). Depending on the model, calibration algorithm, and whether the error correction procedure is applied or not, any number of function calls may turn out the best choice for validation data (even just 1000, as in the case of the HBV with error correction procedure and SPS-L-SHADE-EIG used as calibration method, see the second column from the left in lower part of Fig. 1).

The possible explanation is that there are large differences between hydro-meteorological conditions during calibration and validation periods. Hence, too long calibration of rainfall-runoff models by means of a good method leads to a kind of overfitting – a term well known from artificial neural networks (Geman et al. 1992), where it means fitting the models that are general approximators (Hornik et al. 1989) not only to the signal, but also to a noise present in a data sample. Although conceptual rainfall-runoff models are not general approximators, if calibration takes long enough, their parameters may be fitted to the noise achieving good performance for calibration, but much poorer for validation data.

Results obtained for the Irondequoit River (NY, USA) are much simpler (see Fig. 2 and Suppl. Tables 3–4). For both calibration and validation data 10,000 function calls are the best choice, irrespective of the model, calibration algorithm or error correction procedure. Using 30,000 function calls do not improve results for the calibration data, and slightly deteriorates performance on validation data, what may again point out at some form of overfitting.

In the case of the Biala Tarnowska River catchment (Fig. 3, Suppl. Tables 5–6) impact of the maximum number of function calls on the performance depends on the model. In case of the HBV model the results show a relatively simple pattern: when SPS-L-SHADE-EIG or GLPSO are used for calibration and error correction procedure is not applied, setting the number of function calls to 30,000 is the best choice, both for calibration and validation data. The differences in *MSE*, *PI* or *NSC* when the number of function calls is set to 10,000 or 30,000 are marginal. If error correction procedure is used, or if MDE_pBX algorithm is applied, 10,000 function calls may perform better. In case of the GR4J model, results are similar in most cases, but with notable exceptions. As seen in the last column of Fig. 3, when GLPSO or SPS-L-SHADE-EIG algorithms are used and error correction procedures are not applied, setting the maximum number of function calls to only 1000 may be the best choice for validation data. Longer calibration may lead to a quick decrease in model performance on the validation set. If error correction procedure is applied, 1000 function calls may still be the best choice for validation data, but only when GLPSO is used for calibration of the GR4J model (see Suppl. Table 6). Hence, results obtained for the Biala Tarnowska River confirm non-intuitive findings discussed earlier for the Fanno Creek that in some circumstances the shorter calibration time is, the better performance on independent data may be achieved.

In the case of the Suprasl River (see Fig. 4 and Suppl. Tables 7–8) the longer optimization, the better results for calibration period, but worse for validation one. This means that for this flat catchment the parameters calibrated on one period may hardly be representative for the other one.

In case of the Cedar River (WA, USA) results are inconclusive, especially for the validation data (Fig. 5, Suppl. Tables 9–10). Any number of function calls between 1000 and 30,000 may lead to the best results for the validation set, depending on the methodological variant considered. For the calibration data, there is only a marginal difference between the final performance of models calibrated for 10,000 and 30,000 function calls, and not much worse results are obtained when the number of function calls equals 3000.

Overall, according to both criteria which are comparable among catchments (*NSC* and *PI*), the worst results for validation data are obtained for flat, snow-fed Suprasl catchment; the best results are obtained for mountainous and relatively climatically homogenous Biala Tarnowska catchment.

The above discussion may lead to three relatively general conclusions:

1. Although in this study five daily time series (14–39 years long) are used, the modelling performances for calibration and validation data are often highly uneven. The differences in *MSE*, *NSC* and *PI* values for calibration and validation data are especially high for the Fanno Creek (OR, USA), the Biala Tarnowska and the Suprasl Rivers (Poland), to a smaller degree in case of the Cedar River (WA, USA).

2. For the vast majority of tested variants, the more function calls are available, the better results are obtained for the calibration data. As experiments with different numbers of function calls are independent, obtaining worse results after longer calibration for the calibration data set is possible, and indeed occasionally observed. However, almost always the differences between results obtained in experiments with the maximum numbers of function calls set to 10,000 and 30,000 are marginal. In many cases, 3000 function calls may be sufficient for calibration data, but this is not a rule.
3. Even though for calibration data better results may be obtained when more function calls are used, this does not result in better performance for validation data. Depending on the catchment and various details of modelling methodology used, the best results for validation data may be obtained when the number of function calls is set to any considered value (1000, 3000, 10,000 or 30,000). No general relation has been found between the performance of the rainfall-runoff model on validation data and the number of function calls. This is a very unfortunate conclusion that may, however, at least partly suggest why in the literature some hydrological models are being calibrated by hundreds, others by millions of function calls. This conclusion is different from much more optimistic one found for various tests in Evolutionary Computation-related literature (Posik et al. 2012; Piotrowski et al. 2017b), as here the performance of the specific model on validation data that come from environmental measurements is considered, what was not the case in that two papers.

3.2 Performance measures: Persistence Index versus Nash-Sutcliffe Coefficient

In hydrological literature frequently various criteria are used to evaluate the performance of a rainfall-runoff model. However, two popular ones, *NSC* or *PI*, are rarely used together. Both *PI* and *NSC* are maximization criteria, with the optimum equal to 1. Both allow negative values. However, in *PI* negative value means that it is a better choice to use current flow as the future one than using the model, in *NSC* negative value means that it is better to set long-term averaged flow value than using the model (Schaefli and Gupta 2007). As both criteria leave the reader with an impression that the positive values indicate the usefulness of the model, contrary to the negative, and that the values close to 1 indicate that the model is successful, it may be interesting to compare them side by side.

As may be seen from Figs. 1, 2, 3, 4 and 5 and Suppl. Tables 1–10, *NSC* is almost always much higher than *PI*. It is easier to win the comparison against an average flow than against the most recent observation. However, in some cases, the differences may be substantial and mislead the readers. For example, consider situations in Fig. 5. For both the HBV and the GR4J models applied to the Cedar River without error correction procedure *NSC* values are highly positive (0.68–0.80), but *PI* values are highly negative, showing that both models have no predictive skills.

As seen in Fig. 3, the values of *NSC* similar to those found for the Cedar River are obtained for the Biala Tarnowska River (0.73–0.81). However, for this catchment *PI* values are not only nonnegative, but also well above 0.6. Hence, the information from both *PI* and *NSC* criteria leads to fully contradictory impression for the Biala Tarnowska River and the Cedar River: *NSC* suggests equally good performance on both catchments, according to *PI* models perform well on Biala Tarnowska catchment, but very poorly on Cedar River catchment. This may be the effect of climatic non-homogeneities within the Cedar River catchment, which is divided into two much different parts: very mild, flat, lowland west which generally lacks snowy and frosty conditions in winter, and mountainous east, with frequent snow and frost.

The above discussion should be seen as a warning to not overestimate the importance of *NSC* or *PI* values. Especially *NSC* seems to be a doubtful criterion, as it frequently suggests that the model with little predictive skills leads to respectful results, what has been observed also in McCuen et al. (2006), Jain and Sudheer (2008) and Lin et al. (2017).

4 Conclusions

This paper aims at studying the impact of the assumed number of function calls to be used during calibration of the lumped conceptual rainfall-runoff model on the final performance. Tests with different numbers of function calls (1000, 3000, 10,000 and 30,000) are performed independently, hence longer calibration does not necessarily imply better results. Two models are tested (HBV and GR4J), each applied with or without error correction procedure (Madsen and Skotner 2005), and with three calibration procedures (GLPSO, MDE_pBX and SPS-L-SHADE-EIG) at five catchments (the mountainous Biala Tarnowska, Poland and Cedar River, WA, USA; the hilly Fanno Creek, OR, USA; the lowland Irondequoit Creek, NY, USA and Suprasl, Poland) located in temperate climatic conditions. Research is based on 14–39 years long daily data that are divided into calibration and validation parts.

For various catchments, substantial differences in modelling performances for the calibration and the validation data may be observed, what is in agreement with Merz et al. (2011) and Osuch et al. (2015). Such results have large implication for the number of function calls that should efficiently be used during conceptual rainfall-runoff model calibration. For the calibration data set, often the more function calls are used, the better results are obtained. However, differences between results obtained when 10,000 and 30,000 function calls are used are often meaningless. For validation data, this is often not the case. Depending on the catchment and methodological details, the best results for validation data may be obtained with any considered number of function calls, from 1000 to 30,000. Sometimes the longer optimization is, the better results are obtained for the calibration set, but the poorer for the validation one.

Such lack of consistent relation between the rainfall-runoff model performance and the length of model calibration may clarify the total mess in the hydrological literature, where some models are calibrated for hundreds, other for millions of function calls. As no rules may be observed, each time practitioners must look for the best setting for their specific application. Hence, when calibrating conceptual rainfall-runoff models for practical applications, tests with at least two different numbers of function calls, one very small and one moderate (e.g. 1000 and 10,000) are advised, to see if the choice affect the results; in each case calibration should be performed at least a few times to verify consistency of the results.

In the future, it is recommended to verify how the choice of the stopping conditions of Markov Chain Monte Carlo (MCMC) methods (see Vrugt et al. 2008) affects the distributions of found solutions for conceptual rainfall-runoff models.

The intuitive opinion on the model performance is frequently based on the hydrological criteria like the Nash-Sutcliffe coefficient (*NSC*) or the Persistence Index (*PI*). Although such indexes measure something else, both are maximized, both have the maximum in 1 and both suggest that model is “useful” if its value is positive (higher than 0). In this paper, it is found that the opinion on model performance based on the two criteria may be misleading, as similar largely positive values of *NSC* (≈0.7) observed on two different catchments may be accompanied by very contradictory *PI* values (≈0.6 at one catchment and highly negative at the other).

Acknowledgments The research published in this paper has been financed from the National Science Centre, Poland, grant number 2016/21/B/ST10/02516 (2017-2020).

Compliance with Ethical Standards

Conflict of Interest Authors declare that they have no conflict of interest.

References

Arsenault R, Poulin A, Côte P, Brissette F (2014) Comparison of stochastic optimization algorithms in hydrological model calibration. J Hydrol Eng 19(7):1374–1384

Bergström S (1976) Development and application of a conceptual runoff model for Scandinavian catchments. Norrköping: Svergies Meteorologiska och Hydrologiska Institut, SMHI Report RHO 7:134

Bergström S (1991) Principles and confidence in hydrological modeling. Nord Hydrol 22:123–136

Beven K (2012) Rainfall-runoff modeling. The Primer. Wiley-Blackwell, UK, 2nd Eds. 472p

Bi WW, Maier HR, Dandy GC (2016) Impact of starting position and searching mechanism on evolutionary algorithm convergence rate. J Water Resour Plan Manag 142(9):04016026. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000655

Duan QY, Sorooshian S, Gupta VK (1992) Effective and efficient global optimization for conceptual rainfall–runoff models. Water Resour Res 28(4):1015–1031

Geman G, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. Neural Comput 4: 1–58

Gong YJ, Li JJ, Zhou Y, Li Y, Chung HSH, Shi YH, Zhang J (2016) Genetic Learning Particle Swarm Optimization. IEEE Transactions on Cybernetics 46(10):2277–2290

Goswami M, O'Connor KM (2007) Comparative assessment of six automatic optimization techniques for calibration of a conceptual rainfall–runoff model. Hydrol Sci J 52(3):432–449

Guo SM, Tsai JSH, Yang CC, Hsu PH (2015) A self-optimization approach for L-SHADE incorporated with eigenvector-based crossover and successful-parent-selecting framework on CEC 2015 benchmark set. In: Proc. IEEE Congress on Evolutionary Computation, Sendai, Japan, pp. 1003–1010

Hamon WR (1961) Estimation potential evapotranspiration. Journal of the Hydraulics Division, Proceedings of the ASCE 87(HY3):107–120

Hornik K, Stinchcombe M, White H (1989) Multilayer feed forward networks are universal approximators. Neural Netw 2:359–366

Islam SM, Das S, Ghosh S, Roy S, Suganthan PN (2012) An adaptive Differential Evolution algorithm with novel mutation and crossover strategies for global numerical optimization. IEEE Transactions on Systems, Man, and Cybernetics, Part B–Cybernetics 42(2):482–500

Jain SK, Sudheer KP (2008) Fitting of hydrologic models: A close look at the Nash-Sutcliffe Index. J Hydrol Eng 13(10):981–986

Jeon JH, Park CG, Engel BA (2014) Comparison of performance between Genetic Algorithm and SCE-UA for calibration of SCS-CN surface runoff simulation. Water 6:3433–3456

Jie MX, Chen H, Xu CY, Zeng Q, Chen J, Kim JS, Guo SL, Guo FQ (2018) Transferability of conceptual hydrological models across temporal resolutions: approach and application. Water Resour Manag 32:1367–1381

Kitanidis PK, Bras RL (1980) Real-time forecasting with a conceptual hydrologic model. 2: application and results. Water Resour Res 16(6):1034–1044

Lin F, Chen XW, Yao HX (2017) Evaluating the use of Nash-Sutcliffe efficiency coefficient in goodness-of-fit measures for daily runoff simulation with SWAT. J Hydrol Eng 22(11):05017023. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001580

Lindström G, Johansson B, Persson M, Gardelin M, Bergström S (1997) Development and test of the distributed HBV-96 hydrological model. J Hydrol 201:272–288

Liu ZJ, Guo SL, Zhang HG, Liu D, Yang G (2016) Comparative study of three updating procedures for real-time flood forecasting. Water Resour Manag 30:2111–2126

Lobligeois F, Andreassian V, Perrin C, Tabary P, Loumagne C (2014) When does higher spatial resolution rainfall information improve streamflow simulation? An evaluation using 3620 flood events. Hydrol Earth Syst Sci 18:575–594

Madsen H, Skotner C (2005) Adaptive state updating in real-time river flow forecasting – a combined filtering and error forecasting procedure. J Hydrol 308:302–312

McCuen RH, Knight Z, Cutter AG (2006) Evaluation of the Nash-Sutcliffe efficiency index. J Hydrol Eng 11(6): 597–602

Merz R, Parajka J, Blöschl G (2011) Time stability of catchment model parameters: Implications for climate impact analyses. Water Resour Res 47:W02531. https://doi.org/10.1029/2010WR009505

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I — A discussion of principles. J Hydrol 10(3):282–290

Osuch M, Romanowicz RJ, Booij MJ (2015) The influence of parametric uncertainty on the relationships between HBV model parameters and climatic characteristics. Hydrol Sci J 60(7–8):1299–1316

Pechlivanidis IG, Jackson BM, McIntyre NR, Wheater HS (2011) Catchment scale hydrological modeling: A review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications. Global NEST Journal 13(3):193–214

Perrin C, Michel C, Andreassian V (2003) Improvement of a parsimonious model for streamflow simulation. J Hydrol 279:275–289

Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ, Osuch M, Kundzewicz ZW (2017a) Are modern metaheuristics successful in calibrating simple conceptual rainfall–runoff models? Hydrol Sci J 62(4): 606–625

Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ, Rowinski PM (2017b) Swarm Intelligence and Evolutionary Algorithms: Performance versus speed. Inf Sci 384:34–85

Piotrowski AP (2017) Review of Differential Evolution population size. Swarm and Evolutionary Computation 32:1–24

Poncelet C, Merz R, Merz B, Parajka J, Oudin L, Andreassian V, Perrin C (2017) Process-based interpretation of conceptual hydrological model performance using a multinational catchment set. Water Resour Res 53: 7247–7268

Posik P, Huyer W, Pal L (2012) A comparison of global search algorithms for continuous black box optimization. Evol Comput 20(4):509–541

Refsgaard JC (1997) Validation and intercomparison of different updating procedures for real-time forecasting. Nord Hydrol 28:65–84

Schaefli B, Gupta HV (2007) Do Nash values have value? Hydrol Process 21:2075–2080

Shi Y, Eberhart RC (1998) A modified particle swarm optimizer. In: Proceeding in IEEE Congress on Evolutionary Computation (CEC), pp. 69–73

Tanabe R, Fukunaga A (2014) Improving the search performance of SHADE using linear population size reduction. In: 2014 Proceedings of IEEE Congress on Evolutionary Computation, pp. 1658–1665

Tayfur G (2017) Modern optimization methods in water resources planning, engineering and management. Water Resour Manag 31:3205–3233

Thiessen AH, Alter JC (1911) Precipitation averages for large areas. Mon Weather Rev 39:1082–1084

Tian Y, Xu YP, Zhang XJ (2013) Assessment of climate change impacts on river high flows through comparative use of GR4J, HBV and Xinanjiang models. Water Resour Manag 27:2871–2888

Toffolon M, Piccolroaz S (2015) A hybrid model for river water temperature as a function of air temperature and discharge. Environ Res Lett 10:114011. https://doi.org/10.1088/1748-9326/10/11/114011

Tolson BA, Shoemaker CA (2007) Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. Water Resour Res 43:W01413. https://doi.org/10.1029/2005WR004723

Vansteenkiste T, Tavakoli M, van Steenbergen N, de Smedt F, Batelaan O, Pereira F, Willems P (2014) Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. J Hydrol 511:335–349

Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA (2008) Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour Res 44:W00B09. https://doi.org/10.1029/2007WR006720

Wang YC, Yu PS, Yang TC (2010) Comparison of genetic algorithms and shuffled complex evolution approach for calibrating distributed rainfall–runoff model. Hydrol Process 24:1015–1026