Earth Science Informatics (2025) 18:382
https://doi.org/10.1007/s12145-025-01885-y

RESEARCH

Convergence and final performances of optimization algorithms for rainfall- runoff model calibration based on the number of function calls

Zeynep Beril Ersoy[1,2] · Okan Fistikoglu[1] · Umut Okkan[2] · Berkay Derin[2]

Received: 7 November 2024 / Accepted: 13 April 2025 / Published online: 22 April 2025

Abstract
This study investigates the final performance and convergence behavior of 14 optimization algorithms, three of which are hybrids that combine derivative-based local search algorithms with several metaheuristics, for calibrating seven conceptual rainfall-runoff models (CRRMs) over two watersheds in Turkey. The study employs three objective functions: Nash–Sutcliffe Efficiency (NS), log-transformed NS (LNS), and Kling-Gupta Efficiency. The TOPSIS multi-criteria decision-making tool was used to rank the algorithms based on their performance across various levels of number of objective function calls (NOFCs). The study findings highlight the differential responses of optimization algorithms to NOFC variations and offer valuable insights into selecting suitable algorithms for CRRM calibration. Accordingly, as the NOFCs increased from 2500 to 10,000, differential evolution (DE) variants demonstrated remarkable adaptability, emerging as the top performers. In contrast, conventional metaheuristics struggled with improvements despite the increase in function calls, primarily due to premature convergence issues. Moreover, particle swarm optimization (PSO) variants endowed with mutation or derivative schemes performed well at lower NOFCs, but as more extensive exploration became necessary, they showed diminishing returns. The study underscores the superiority of DE variants, which include more complex mutation schemes or are derivative-based, in long-run scenarios where both computational efficiency and calibration accuracy are important.

Keywords Conceptual rainfall-runoff models · Model calibration · Hybrid algorithms · The number of objective function calls · TOPSIS

Introduction

Describing and modeling the relationship between rainfall and runoff at the watershed scale is challenging due to several meteorological, land cover, and soil-related parameters that vary spatially and temporally. It is widely acknowledged that models cannot be overly comprehensive in practice and must offer a general approximation of various physical processes in a basin, and within this context, conceptual rainfall-runoff models are often preferred (Deng and Wang 2021). Even though conceptual rainfall-runoff models (CRRMs) are sensitive to hydro-meteorological variability and temporal resolution, it cannot be claimed that they demonstrate worse runoff simulation performance than sophisticated distributed models (Vansteenkiste et al. 2014). Nevertheless, CRRMs, albeit simpler and requiring less data than physical models, still need to be calibrated for precise runoff simulation (Goswami and O'Connor 2007).

With the developments in informatics, the number of papers examining and comparing optimization algorithms for rainfall-runoff model calibration has begun to increase (Piotrowski et al. 2017; Qin et al. 2018; Okkan and Kirdemir 2020; Napiorkowski et al. 2023). Direct search methods such as simplex and pattern search algorithms have been some of the initial strategies evaluated in the automatic calibration of CRRMs (e.g., Hendrickson et al. 1988; Gan and Biftu 1996). In addition, Newton-type derivative-based algorithms have been investigated during the calibrating stages (e.g., Gupta and Sorooshian 1985), but some lack of robustness of such algorithms

✉ Zeynep Beril Ersoy
zeynepberil.ersoy@balikesir.edu.tr

[1] Department of Civil Engineering, Hydraulic Division, Dokuz Eylul University, İzmir, Turkey

[2] Department of Civil Engineering, Hydraulic Division, Balikesir University, Balikesir, Turkey

has been noted, and the main reason for this seems to be poor conditioning of the response surface of the parameter space of the models (Hendrickson et al. 1988). Thereafter, understanding the non-convex structure of the calibration problem within CRRMs, the sensitivities of both derivative and direct search algorithms to initial parameter values, and the predominance of their tendency to be trapped in local minima have led hydrological model users to prefer global optimization algorithms (GOAs) based on swarm intelligence or stochastic evolutionary algorithms. Studies on various genetic algorithm (GA) schemes and shuffled complex evolution (SCE) algorithms in the calibration of CRRMs are obviously prevalent (e.g., Duan et al. 1994; Gan and Biftu 1996; Franchini et al. 1998; Cooper et al. 2007; Wu et al. 2012). Moreover, research papers examining particle swarm optimization (PSO), differential evolution (DE), covariance matrix adaptation evolution strategy (CMAES), harmony search (HS), cuckoo search (CS), invasive weed colonization (IWC), artificial bee colony (ABC), dynamically dimensioned search (DDS), and more recent metaheuristics in calibrating various CRRMs have also gained attention (e.g., Goswami and O'Connor 2007; Arsenault et al. 2014; Huang et al. 2014; Piotrowski et al. 2017, 2019; Okkan and Kirdemir 2020; Napiorkowski et al. 2023).

In rainfall-runoff modelling, present perceptions mostly favor GOAs due to their overall robustness. The ability of GOAs to capture the optimal solution within multiple runs stands out compared to that of derivative algorithms (e.g., Okkan and Kirdemir 2020). Despite their clear benefits reported, GOAs may display late convergence when calibrating CRRMs. Yet most of them can be considered equivalent in terms of final performances, making it difficult to identify which one is superior (Piotrowski et al. 2017). Besides, these algorithms can be computationally costly depending on the number of objective function calls (hereinafter referred to as NOFCs) assigned and are vulnerable to parameter dimensionality as well (Qin et al. 2018). Therefore, some hydrologists have made some modifications to derivative algorithms, motivated by the need for less computational intensity and faster convergence. For example, Qin et al. (2018) offered a robust algorithm by incorporating several heuristics into the conventional Gauss–Newton algorithm to enhance its effectiveness in exploring optimal solutions and performing well with challenging objective functions. Additionally, Okkan and Kirdemir (2020) embedded the Levenberg–Marquardt (LM) algorithm, which requires first-order partial derivatives of the residual errors, into standard PSO in a nested manner, dramatically improving the robustness characteristics of each algorithm used in this hybridization. All these approaches aspire to attain a comparable level of robustness as familiar GOAs while providing a notably reduced computational cost (i.e., NOFCs), thus resulting in enhanced overall efficacy.

As can be understood from CRRM calibration experiences, the ranking of not only GOAs but also derivative ones hybridized with various metaheuristics depends largely on NOFCs and on convergence rates as well. To the best of our knowledge, the issue of figuring out the appropriate NOFCs has been rarely addressed in CRRM studies. Arsenault et al. (2014) studied the average number of function calls required for a given algorithm to reach roughly the best value of the objective function throughout the entire search. Their work revealed that near-optimal results were typically achieved within the first half of the search process, with around half of cases reaching this level within the 5000 function calls. Similarly, Lespinas et al. (2018) found that in the calibration of the ecohydrological model VELMA through the DDS algorithm, the objective function improved rapidly up to 300 function calls, but after this point it enhanced slowly until 10,000 function calls. In another recent study in which Piotrowski et al. (2019) calibrated two CRRMs with three modern metaheuristics, when using more than 10,000 function calls, merely a minor enhancement in model performances has been detected, regardless of study area or evaluated GOAs.

Motivated by all these findings, we made it our mission to select suitable algorithms that converge quickly during calibration of different CRRMs and deliver robust final performances without incurring excessive NOFCs. Despite the increasing use of automatic optimization algorithms in the calibration phase of various hydrological models, the convergence and ultimate performance of the algorithms have not been thoroughly inspected in relation to the NOFCs. The scant number of papers have concentrated on the efficacy of individual algorithms rather than undertaking a systematic comparison of both hybrid and traditional ones under varying NOFC conditions. This study addresses this gap by conducting a systematic evaluation of both the convergence behavior and final performance of hybrid metaheuristic approaches that are developed with derivative schemes in comparison to familiar global optimization algorithms over multiple lumped hydrological models. Moreover, another distinctive aspect of this study is that it has employed a straightforward framework based on Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) in order to rank the algorithms used in model calibration according to different metrics.

In this context, independent experiments were performed on various CRRMs through given calibration algorithms running under different NOFCs, and the algorithms were subjected to a multiple criteria decision-making process. More specifically, the objectives of this study are to.

- benchmark the upgraded versions of the derivative LM algorithm that utilize several metaheuristics against well-known GOAs such as GA, SCE, two PSO variants, five DE variants, ABC, and gray wolf optimizer (GWO) over seven CRRMs, three objective functions, and two watersheds by setting NOFCs to 500, 1000, 2500, 5000, and 10,000,
- make empirical inferences about what the appropriate number of function calls would be from traces of optimization trajectories collected and,
- generate the decision matrix through the final performance statistics (30-run mean values) and convergence rates pertaining to each algorithm run for different objective functions and to rank the algorithms within themselves via the TOPSIS.

This study presents a novel investigation into the comparative convergence behavior and ultimate performance of hybrid metaheuristic approaches, enhanced with derivative schemes, against conventional metaheuristics under varying NOFC conditions, applied across multiple lumped models. A distinguishing aspect of the work is the integration of the TOPSIS method to systematically rank the calibration algorithms, thereby offering a robust framework for evaluating algorithmic suitability.

The rest of the paper is structured as follows: Sect. 2 first describes the study region and data. Afterward, it delivers key information about the selected CRRMs, the objective functions, the calibration algorithms utilized, the metrics for testing their robustness, and the TOPSIS algorithm implemented. Section 3 reports the results of the case study and covers the discussions regarding the findings. The final section draws the main conclusions.

Material and methods

Study area and data

Two watersheds, namely, Tahtali and Beydag, which are located in the Kucuk Menderes River Basin (KMRB) in western Turkey, are considered in this study (Fig. 1). Major agricultural operations have recently been made for KMRB, leading to higher water demand despite limited water availability. Therefore, the region has become one of the hotspot study sites in Turkey where drought monitoring studies and irrigation water management models are performed by various researchers (e.g., Pusatli et al. 2009; Eris et al. 2020). In this regard, efforts to establish hydrological models for available watersheds in the KMRB and calibrate them optimally might provide a basis for state-owned organizations in the region.

The flow gauging stations representing Tahtali and Beydag watersheds are station D06A007 with a drainage area

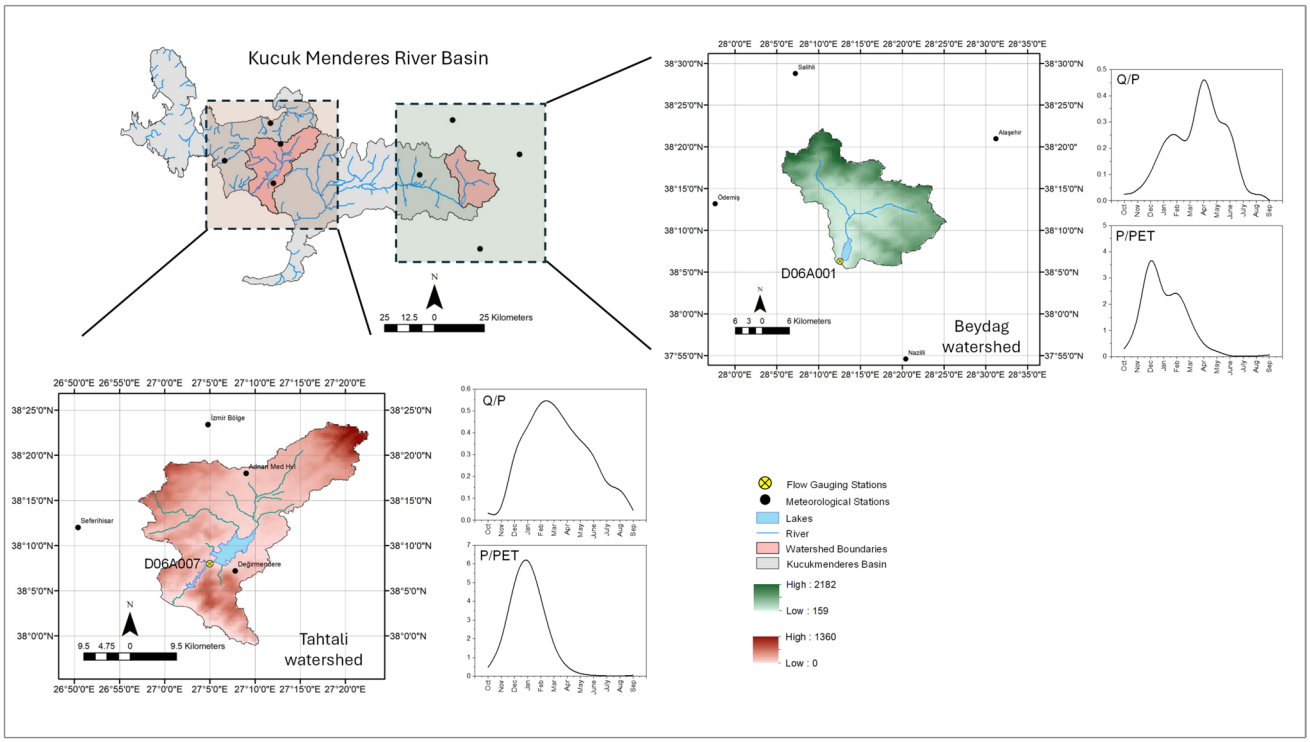

Fig. 1 The two study sites in the Kucuk Menderes River Basin, Turkey, and the location of the stations used

of 513 km^2, and station D06A001 with a drainage area of 445 km^2, respectively. After the dams were built on the relevant tributaries, these flow gauging stations were no longer operated. Yet, natural monthly streamflow data regarding D06A007 and D06A001 stations during the 1970–1988 and 1987–1999 water year periods could be obtained from the General Directorate of State Hydraulic Works of Turkey. In the calibration of the CRRMs for Tahtali watershed, the first 10 years of 19-year natural streamflow data of station D06A007 were evaluated. For the Beydag watershed, the first 7 years of the 13-year streamflow data compiled were used in calibration. Thiessen-weighted precipitation (P) and areal mean temperature (T_{mean}) data for the mentioned periods were also compiled at monthly time scale to prepare hydrological model inputs. When calculating potential evapotranspiration (PET) values, which serve as the secondary input to the CRRMs, one may utilize empirical equations like Penman–Monteith, which necessitate comparatively more data. However, employing empirical equations based on mean temperature also instills adequate confidence in the calibration of lumped models. Although PET estimates derived from an imperfect equation may differ substantially from those derived from the reference formula Penman–Monteith, these biases can be mitigated by adjusting the conceptual parameters to the inputs during the calibration process of the hydrological model (e.g., Oudin et al. 2005; Seiller and Anctil 2016; Okkan et al. 2024). Therefore, PET inputs were estimated utilizing the temperature-based Kharrufa equation, which was locally calibrated, as implemented by Xu and Singh (2001). Furthermore, according to streamflow monitoring periods, the mean annual temperature and PET regime for both watersheds are around 16.0 °C and 1450 mm, respectively. The mean annual precipitation in the Tahtali watershed is 825 mm, which indicates that the region is predominantly characterized by a dry sub-humid climate. The Beydag watershed received nearly 485 mm of annual mean precipitation during the streamflow observation period, which implies that the study site is in a semi-arid climatic zone. For the Tahtali watershed, one-third of the total precipitation constituted total runoff, of which two-thirds were observed in the winter season. Nevertheless, the Beydag watershed experienced its maximal surface runoff coefficient in April, while the overall runoff accounted for a mere 20% of the precipitation, as can be deduced from Fig. 1.

CRRMs

In the study, seven lumped CRRMs with different runoff partitioning structures that have been successfully exerted under both stationary and changing climate conditions were selected (Fig. 2). They are the Témez model (Pérez-Sánchez et al. 2022), the dynamic water balance model (Dynwbm) (Zhang et al. 2008), the abcde model obtained by adding a parameter into the abcd model (Okkan and Kirdemir 2020), the Gr2 m model (Mouelhi et al. 2006), the Australian water balance model (Boughton 2004), a re-adapted version of the Guo model (Pérez-Sánchez et al. 2019), and the Thornthwaite water balance model (Twbm), in which groundwater storage is incorporated (Elçi et al. 2010). These models, which necessitate P and PET inputs, depict the catchment hydrology through a set of parametric functions that stand for conceptual soil moisture, groundwater, and routing storages. Parameter names and their value ranges used in the calibration of the CRRMs are provided as supplementary material (see Table S1). All models were applied on a monthly time scale so that they could also produce input for the operating simulation of the two existing dam reservoirs. The equations employed to simulate runoff components and other intermediate variables for these models are given in Figs. S1–S7. The variables represented by their acronyms in these figures are also explicated in Table S2.

Objective functions

The Nash–Sutcliffe efficiency (NS) (Nash and Sutcliffe 1970) and the Kling-Gupta efficiency (KGE) (Gupta et al. 2009), which emphasize the general concordance between simulated runoff values and observations, are widely employed model assessment criteria. It has been pointed out that the NS and KGE put more weight on high flows, while the logarithmic version of Nash–Sutcliffe efficiency (hereinafter referred to as LNS) prioritizes low flows (Pushpalatha et al. 2012; Deng and Wang 2021). Thus, this study employed NS, LNS, and KGE metrics to assess how optimization algorithms responded to different objective functions. The values of three metrics range from -∞ to 1 (perfect fit), and the related calculation formulas are summarized in Table 1. Since the optimization algorithms in this study were set up to be adaptable to minimization problems, the following objective functions were evaluated while calibrating CRRMs referred to in the previous subsection:

$$f_1 = 1 - NS \tag{1}$$

$$f_2 = 1 - LNS \tag{2}$$

$$f_3 = 1 - KGE \tag{3}$$

Although the theoretical minimum value of each objective function is 0, the true minimum value is unknown and contingent upon the model and the data used.

Overview of optimization algorithms used

In the study, real-coded GA, SCE, two PSO variants, five DE variants consisting of different mutation schemes, ABC, and

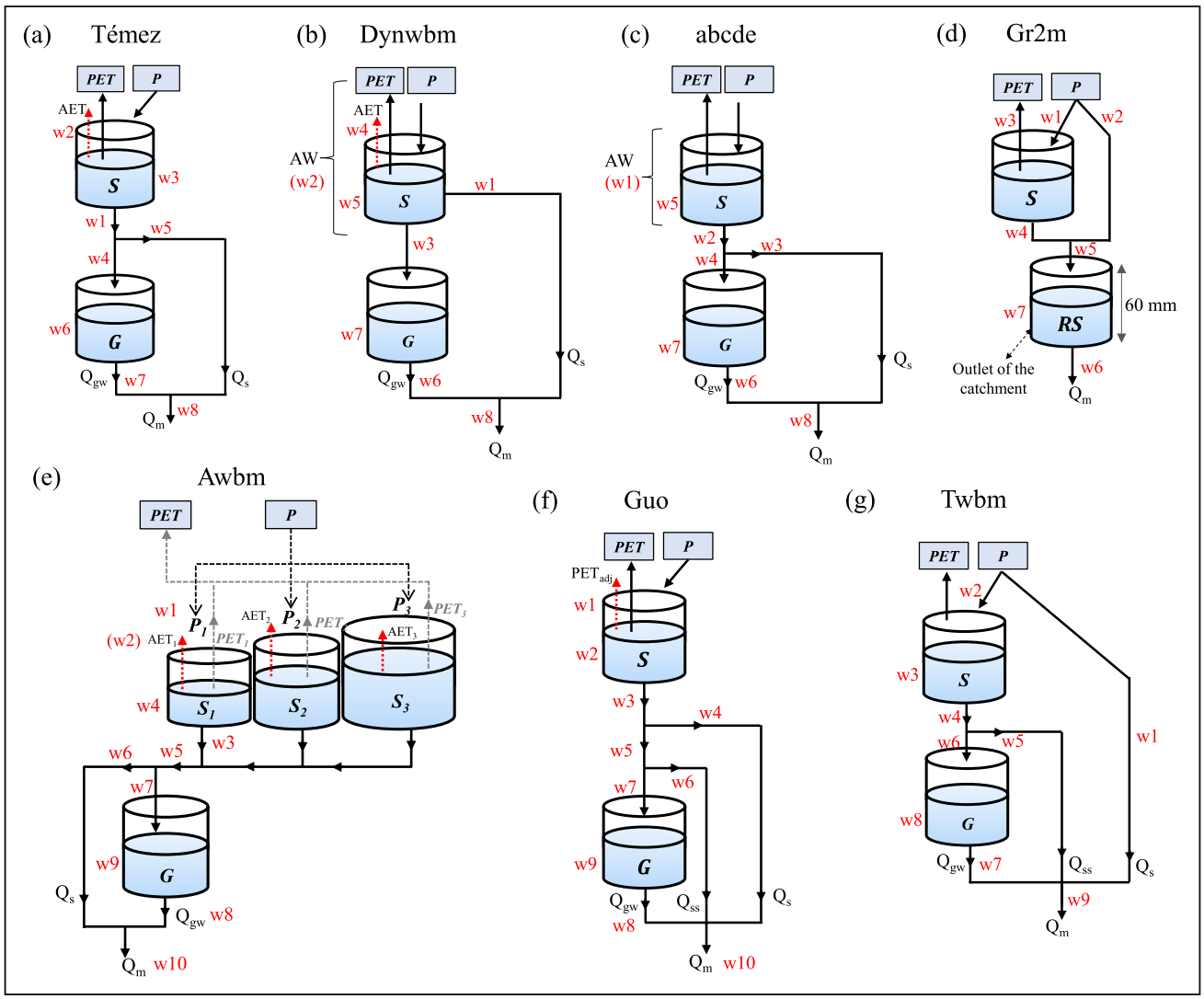

Fig. 2 Conceptual representations of seven CRRMs in different runoff partitioning structures. Details of the models and the meanings of the symbols associated with this figure are provided as *Supplementary Material*

Table 1 Model evaluation criteria together with allowed ranges and perfect match values

Criteria formula	Allowable range	Perfect match
$NS = 1 - \frac{\sum_{i=1}^{n}(Q_{o,i}-Q_{m,i})^2}{\sum_{i=1}^{n}(Q_{o,i}-\overline{Q_o})^2}$	(-∞ to 1)	1
$LNS = 1 - \frac{\sum_{i=1}^{n}(\log(Q_{o,i})-\log(Q_{m,i}))^2}{\sum_{i=1}^{n}(\log(Q_{o,i})-\overline{\log(Q_o)})^2}$	(-∞ to 1)	1
$KGE = 1 - \sqrt{(R-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$	(-∞ to 1)	1

where $Q_{o,i}$ and $Q_{m,i}$ are the observed and simulated runoff at the i^{th} month, respectively; n is the number of data points for the calibration or validation period; R is the correlation coefficient between the observed and simulated runoff values; $\overline{Q_o}$ is the mean values of the runoff observations; α is the ratio of standard deviations of the simulated and observed monthly runoff; β is the ratio of mean values of the simulated and observed monthly runoff

GWO were employed. These were chosen due to their adaptability to the calibration of CRRMs and their reputation in the field of soft computing. As one might expect, there would be sensitive parameters that govern all optimization algorithms used. We either conducted several experiments in tuning the hyperparameters or fixed them in runs by referring to past experiences or recommendations (Table S3). Accordingly, the control parameters assigned for the GA, PSO (chaotic random inertia weighted version), and ABC algorithms are identical to those in the study performed by Okkan and Kirdemir (2020). For the elitist-mutated variant of PSO (referred to as PSOm), the procedure proposed by Kang and Zhang (2016) was followed. The performance of SCE is contingent upon the proper setting of only a few control parameters, and the default values recommended by Duan et al. (1994) were the preferable ones in this study.

The GWO algorithm, which is one of the novel metaheuristics that has not yet been preferred in CRRM calibration, was applied as suggested by Mirjalili et al. (2014). Moreover, five DE variants with different mutation schemes were adopted (Table S4). In the original DE algorithm (DE1), in any gth generation, three chromosomes with different sequence numbers are randomly selected from the population ($i = 1, 2,\ldots,N_{POP}$), except for chromosome i, and the mutant vector is obtained with a scaling factor SF. The DE2 variant launches the mutation process by combining the optimal solution x_{best} with two random chromosomes. The other variant DE3 is an adapted version of the standard one with five random chromosomes and two scaling factors (i.e., SF_1 and SF_2). Similarly to DE2, variant DE4 implements a local search strategy that revolves around x_{best}, but it contains two scaling factors. The DE5 variant concentrates on the local search around the ith chromosome, considering both the two random chromosomes in the population pool and the finest chromosome. In the DE variants, all operators except mutation are implemented in the same manner as in standard DE1, and the recommendation by Leon and Xiong (2014) was followed in the selection of their control parameters.

All of the algorithms above mentioned undertake global optimization, yet they incur computational expense by conducting a stochastic search. On the other hand, deterministic ones like derivative-based algorithms converge fast but may become trapped in local minima. Thus, new frameworks have been also adopted that blend the advantages of stochastic and derivative algorithms while circumventing their shortcomings (Noel 2012; Okkan and Kirdemir 2020). As per the same intention, the Levenberg–Marquardt (LM) algorithm, which operates first-order partial derivatives of the residual errors, was incorporated into PSO, DE, and GWO in a nested form for this study. The proposed hybrid algorithms follow a sequential nested structure, where the metaheuristic component facilitates global exploration, while the LM allows for refining the best solution and iteratively updates the competent solution within the population. This combination aims to reduce the number of function calls required for convergence compared to that of a standalone metaheuristic while also soothing the risk of getting stuck in local minima that derivative-based methods can encounter, preventing premature stagnation. While Fan et al. (2004) employed similar methodology that comprised the coupling of PSO with the derivative-free Nelder–Mead method to analyze certain multimodal test functions, the hybridization procedure followed in the present study is schematized as a pseudo-code in Fig. S8.

Consequently, a total of 14 optimization algorithms, three of which are hybrid types (i.e., HPSO, HDE, and HGWO), were employed to conduct calibrations of seven CRRMs. As each optimization algorithm used is population-based, the initial stage entails the establishment of a population of candidate solutions. Each parameter within these solutions is randomly assigned a value within its feasible range (Table S1). While calibrating each CRRM, the algorithms were run 30 times with those starting populations each time, and the population size N_{POP} was set to 50 for each run. The calibration experiments were repeated with a variety of NOFCs, including 500, 1000, 2500, 5000, and 10,000, to derive interpretations of the appropriate number of function calls from multiple optimization attempts. It should be noted that these values are set depending upon the iterative internal process of the algorithms and whether they are subjected to LM-based hybridization.

Performance rating of calibration algorithms

The study first obtained 30-run mean values of performance metrics (i.e., F_m) for different NOFC sets, as was carried out by Piotrowski et al. (2019). Additionally, it is essential to measure the rate at which algorithms converge to the optimal value during iterations. According to He and Lin (2016), the convergence rate of an algorithm for t generations is

$$C(r,t) = 1 - \left(\left| \frac{F_{opt} - F(1)}{F_{opt} - F(0)} \right| \times \cdots \times \left| \frac{F_{opt} - F(t)}{F_{opt} - F(t-1)} \right| \right)^{1/t} \equiv 1 - \left(\left| \frac{F_{opt} - F(t)}{F_{opt} - F(0)} \right| \right)^{1/t} \quad (4)$$

where $F(t)$ is the objective function value reached when the t^{th} iteration is terminated. F_{opt} is the most suitable among the optimal results achieved by the algorithms.

The above expression serves to determine the geometric convergence rate of any algorithm during given iterations for the r^{th} run. Given that initializing the algorithms with random parameter solutions would produce distinct $F(0)$ sets, the 30-run mean of convergence rates (C_m) was regarded as descriptive statistics. To summarize, the final performance measure F_m along with C_m could be obtained under a certain NOFC condition, while calibrating a conceptual model using any algorithm.

The utilization of a multi-criteria decision-making procedure was found necessary to analyze $14 \times 3 \times 2 = 84$ performance metric values compiled for any CRRM and NOFC variation. At this juncture, given the decision-making matrix (D), it is feasible to rank the calibration algorithms by means of TOPSIS method.

$$D = \begin{vmatrix} F_m(1,f_1) & F_m(1,f_2) & F_m(1,f_3) & C_m(1,f_1) & C_m(1,f_2) & C_m(1,f_3) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ F_m(j,f_1) & F_m(j,f_2) & F_m(j,f_3) & C_m(j,f_1) & C_m(j,f_2) & C_m(j,f_3) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ F_m(k,f_1) & F_m(k,f_2) & F_m(k,f_3) & C_m(k,f_1) & C_m(k,f_2) & C_m(k,f_3) \end{vmatrix} \quad (5)$$

where the 30-run mean value of the pertinent objective function derived by calibrating any conceptual model with the j^{th} algorithm that considers the minimization of f_1 is denoted as $F_m(j,f_1)$. For the other elements of matrix D, the rows of which are $k = 14$, the same notation is applicable.

While it is considered an effective instrument for ranking hydrological models and global climate models (e.g., Raju and Kumar 2015), TOPSIS, the details of which are given above, was adapted for this study in order to grade optimization algorithms in the calibration of each CRRM.

The weighted normalized decision matrix is obtained through multiplication of the $z_{j,i}$ matrix, which is generated by normalizing each column in the D, with weights w_i (Eq. 6).

$$v_{j,i} = w_i z_{j,i}, i = 1, 2, \ldots, 6 \quad (6)$$

From that matrix, the ideal (V_i^+) and non-ideal (V_i^-) values for each criterion are determined. Accordingly, in the first three columns recapitulating the F_m criteria, V^+ and V^- are the minimum and maximum values of the relevant columns, respectively. Since the last three columns display the mean convergence rates, their maximums are the most ideal ones for the relevant criterion. Afterwards, the Euclidean distances to the ideal and non-ideal points are calculated for each candidate algorithm as follows:

$$d_j^+ = \sqrt{\sum_{i=1}^{6} \left(v_{j,i} - V_i^+\right)^2} \quad (7)$$

$$d_j^- = \sqrt{\sum_{i=1}^{6} \left(v_{j,i} - V_i^-\right)^2} \quad (8)$$

Lastly, the relative closeness to ideal values can be determined (Eq. 9), which allows for the evaluation of the overall performance of the j^{th} optimization algorithm under any chosen NOFC while calibrating a CRRM.

$$C_j = \frac{d_j^-}{d_j^+ + d_j^-}, 0 < C_j < 1, j = 1, 2, ..., 14 \quad (9)$$

The C_j values were sorted in descending order, resulting in the rank of the highest value being 1. While applying TOPSIS, the weighting scenarios in Table 2 were evaluated.

Results and discussion

Our study examines a total of 2940 variants: 7 conceptual models ×3 objective functions ×14 calibration algorithms ×5 maximum number of function calls ×2 watersheds. All algorithm experiments were carried out in the MATLAB environment to minimize the objective functions. Considering the comprehensive chain of combinations, we initially focused on inspecting the final and convergence performances of the optimization algorithms in relation to the NOFCs (Sect. 3.1–3.6). Subsequently, we assessed the CRRMs from a hydrological standpoint (Sect. 3.7) and addressed the limitations of the study (Sect. 3.8).

To what extent do final performances vary depending on NOFCs?

Prior research has indicated that there is no apparent relationship between the performance of the conceptual models on validation data and the NOFCs (Piotrowski et al. 2019).

Table 2 The criteria weights evaluated while applying TOPSIS methodology

weighting scenarios	F_m			C_m		
	f_1	f_2	f_3	f_1	f_2	f_3
S1	**1.0**					
S2		**1.0**				
S3			**1.0**			
S4				**1.0**		
S5					**1.0**	
S6						**1.0**
S7	**1/3**	**1/3**	**1/3**			
S8				**1/3**	**1/3**	**1/3**
S9	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**

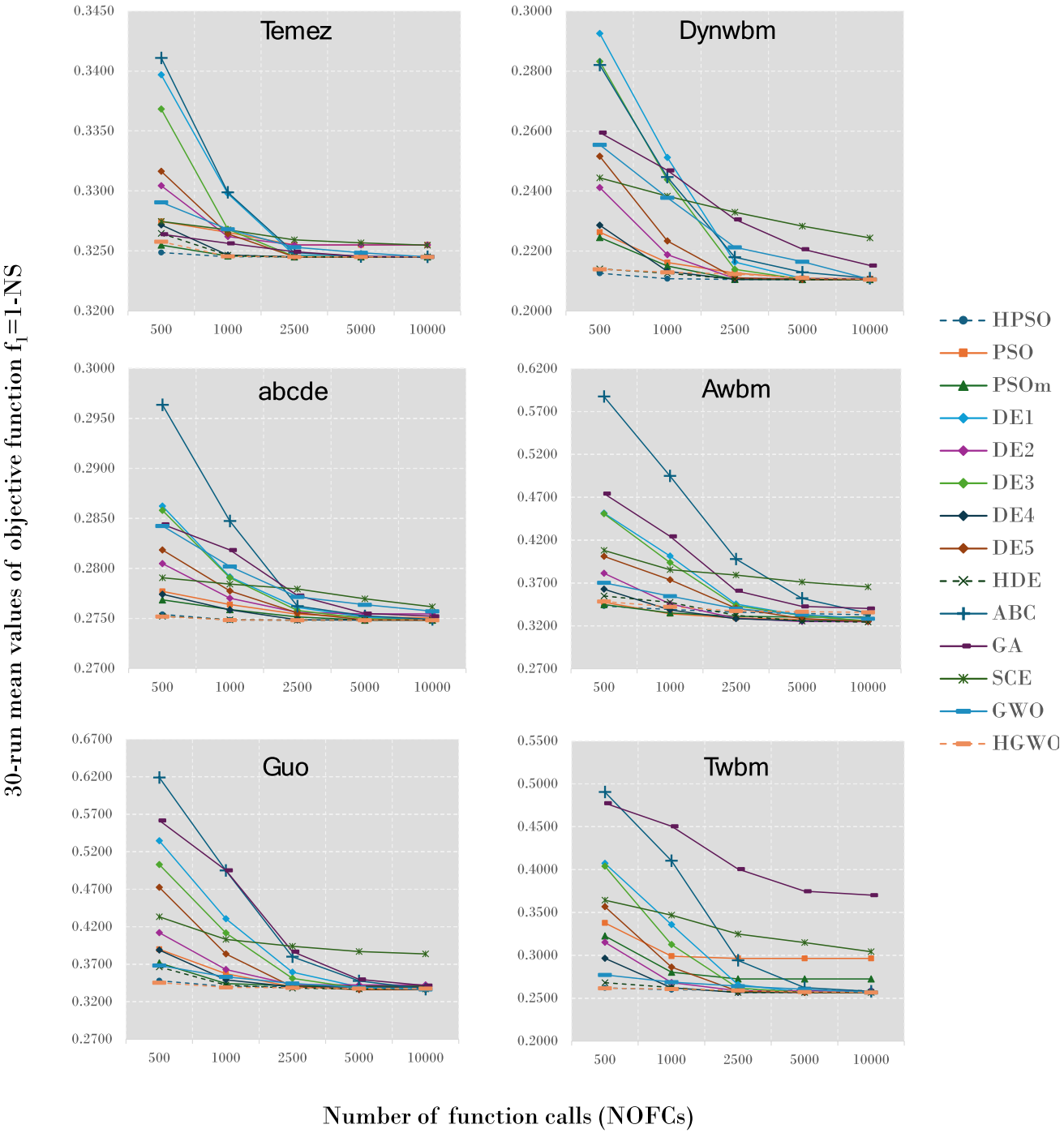

Fig. 3. 30-run mean values of the objective function f_1 over different conceptual models adapted for the Beydag watershed when the maximum number of function calls were set to 500, 1000, 2500, 5000, and 10,000 during calibration. It was plotted using the values in Table S4, except for those of Gr2m

Thus, only results that included metrics from the calibration period data were taken into consideration when evaluating the calibration skills of the 14 optimization algorithms. In this section, the final performances of each calibration algorithm, which were operated under distinct NOFC sets, were separately provided for three objective functions. Yet, the fact that we obtained results from a multitude of variants led us to include some of them in the Supplementary material. Tables S5–S10 contain a summary of the results, featuring the mean values of the objective functions (i.e., F_m statistics) for all variants that were analyzed, calculated across 30 independent runs.

To facilitate the visualization of the 30-run mean values of the objective function f_1 in relation to varying values of the NOFCs, a representative graph was provided (Fig. 3). It is immediately apparent from Fig. 3 and supplementary tables that the variability between algorithms in terms of F_m decreases as NOFC increases, which is in line with the findings of Piotrowski et al. (2017). Undoubtedly, these results may not always reflect the same circumstance, as revealed by Arsenault et al. (2014), since the choice of algorithms will influence how the calibration turns out. Our findings also demonstrate that the variability in question is contingent upon CRRM. For instance, the optimal result was easily achieved by any algorithm for two-parameter Gr2m, even using 500 function calls. Hence, the results of this model are not included in Fig. 3. As can be seen from the same figure, in the calibration of models such as abcde, and Temez, most of the algorithms have tended to give alike final performances as the NOFCs increased. However, the algorithms' responses did not appear to be identical, even when they operated with 10,000 function calls for the rest of the CRRMs. Given that calibration algorithms may need to be executed with greater NOFCs when obtaining the optimal objective function value for any given CRRM, it would be revealing to determine the appropriate number of function calls from these calculated F_m statistics. In this regard, it was observed from Tables S5–S10 that, as of the

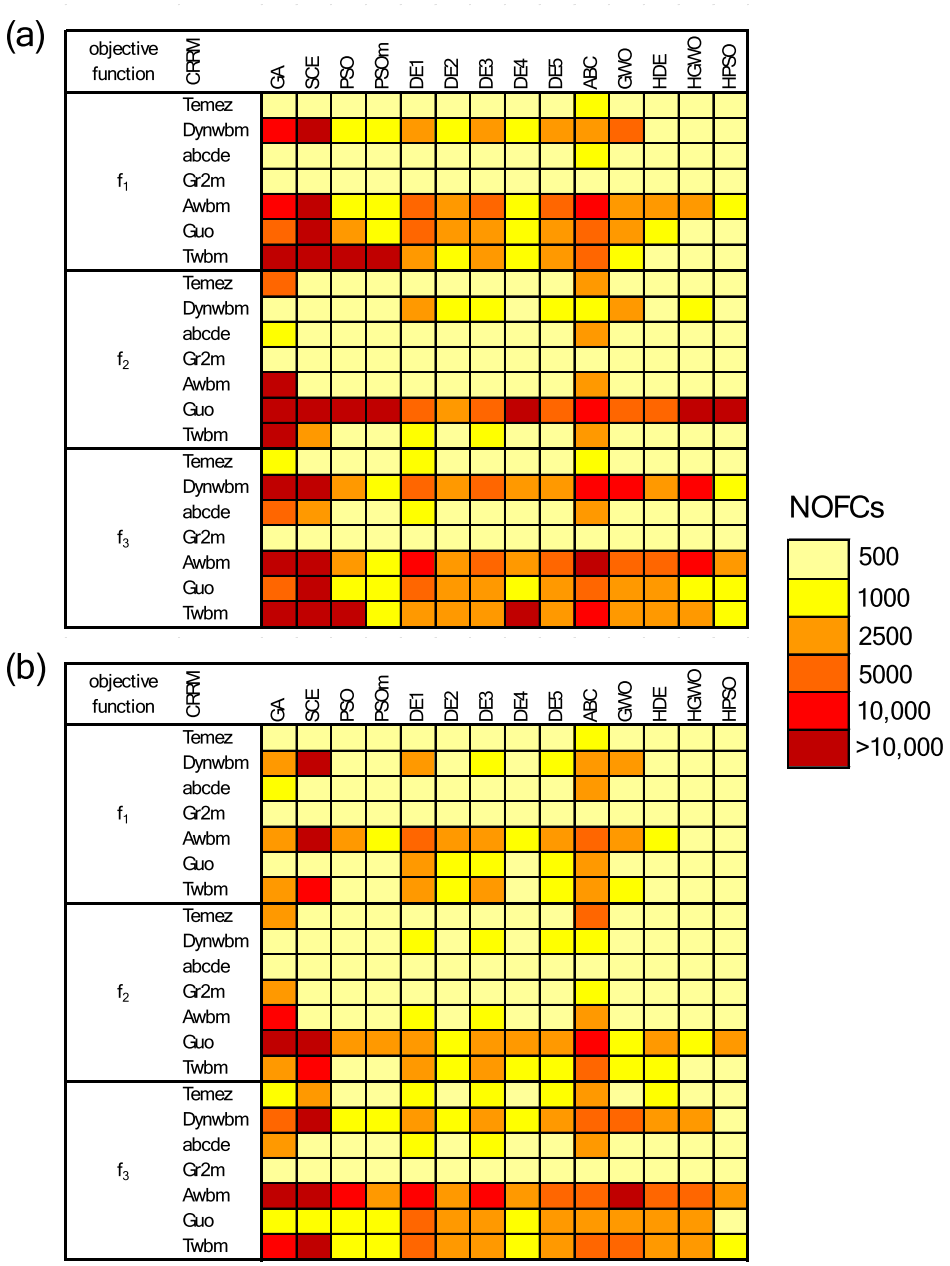

Fig. 4 The number of function calls deemed sufficient for each calibration algorithm over different cases (a: Beydag watershed experiments, b: Tahtali watershed experiments)

NOFCs in the rows corresponding to the bold values, the related optimization algorithm has produced values that are close to the optimal objective function value captured for the conceptual model. An allowable tolerance level for deviation from the optimal objective function value was set at 5.0% during the process of deciding the adequate number of function calls. In Fig. 4, the NOFCs that are considered sufficient for each calibration algorithm in various cases are denoted. Referring to Fig. 4, the following observations can be drawn:

- In models such as Gr2m, Temez, and abcde which were calibrated based on NS maximization (i.e., f_1 minimization), all algorithms aside from ABC and GA produced results that were mostly identical to the optimal value with 500 to 1000 function calls. This may be attributable to the fact that these models have fewer parameters or have marginal interactions between parameters in comparison to other CRRMs utilized.
- Since the rainfall-runoff relationships for the Tahtali watershed were better captured, the optimization algorithms faced fewer challenges in calibrating the models. As a result, in many cases tested over the Tahtali watershed, they tended to converge to the optimum with slightly fewer NOFCs compared to those of the Beydag watershed.
- What is particularly obvious in Fig. 4 is that hybrid algorithms (i.e., HDE, HGWO, and HPSO) have yielded satisfactory F_m statistics versus their basic counterparts for even minimum of NOFCs, given the median values of function calls between models. While Piotrowski et al. (2019) and Arsenault et al. (2014) reported that using 3000–5000 function calls could be sufficient for calibration data in most cases, the hybrid algorithms secured a high level of confidence in model calibration, even when running with one-fifth of those NOFCs, making us query whether longer calibration efforts might be a waste of time. This inference is, of course, unique to NS maximization and may not be applicable to other objective functions as a rule.
- In models including Dynwbm, Awbm, Guo, and Twbm, the statistics of reaching the optimal solutions did not seem satisfactory even if NOFC is set to 10,000 in at least one or the GA, and SCE algorithms. This was even more evident in the Beydag watershed experiments, and especially in the Guo model, which adopted LNS maximization (i.e. f_2 minimization) and the Twbm, which prioritized high flow simulation through NS maximization.
- Most of the algorithms being operated with LNS maximization have captured the optimal solution with 500–1000 function calls across the models. But the fact that monthly hydrological models are generally weak in low flow simulations may pose additional challenges to the algorithms, as emphasized by Deng and Wang (2021). In this regard, the limited capability of some models (i.e., Guo) to simulate low flows may have pushed the algorithms to employ more function calls during calibrations.
- It is apparent that while maximizing the KGE that combines the three components, the inter-model variability of the NOFCs required for the algorithms increased, and their generalization became more elusive as well. That is, balancing three aspects during calibration can be more intricate than concentrating solely on the variance, as is the case with NS. Nonetheless, it looks promising that HPSO, one of the hybrids, and DE4, and PSO variants, among the conventional GOAs, are capable of producing robust results with reasonable numbers of function calls.

As for summarizing this subsection, it is obvious that algorithms need to be executed with different NOFCs to achieve stable final performances. Besides, the results get influenced by the study area (i.e., in situ observations), objective function, and conceptual model selection, which all exacerbate uncertainty on algorithm settings. Indeed, it is noteworthy that the HPSO and some basic GOAs (e.g., PSOm, and DE4) succeeded in achieving optimal results with less calibration effort in most cases, with a few exceptions. Although all these give a preliminary inference in general, the weighting in TOPSIS would serve as a bridge for the integration of criteria with varying performances, thus undertaking grading for all optimization algorithms used.

Understanding the optimization algorithms' final performance via TOPSIS

The weighting scenarios to be used when evaluating the decision matrix are specified in Table 2. If grading is solely based upon the final performance (i.e., F_m), irrespective of the rate of convergence, it would be feasible to interpret the rankings using the S1, S2, and S3 scenarios, which assess each objective function separately, as well as the S7 scenario, which assigns equal importance to all three objective functions. The implementation of TOPSIS yielded C_j values that were employed to rank the calibration algorithms within themselves for each CRRM. To prevent the ranking from being significantly influenced by minor differences in those values, each value was rounded to one decimal place prior to being arranged in descending order. In this study, the highest rank (i.e., 1) was assigned to the largest C_j value, and in the event of tied values, the highest rank among the tied positions was uniformly allocated to each of the identical values, as illustrated in Fig. S9. This figure exemplifies the ranking of the optimization algorithms that calibrate CRRMs with NS maximization in terms of the F_m statistics. It is also clear from Fig. S9 that hybrid-type algorithms have given high

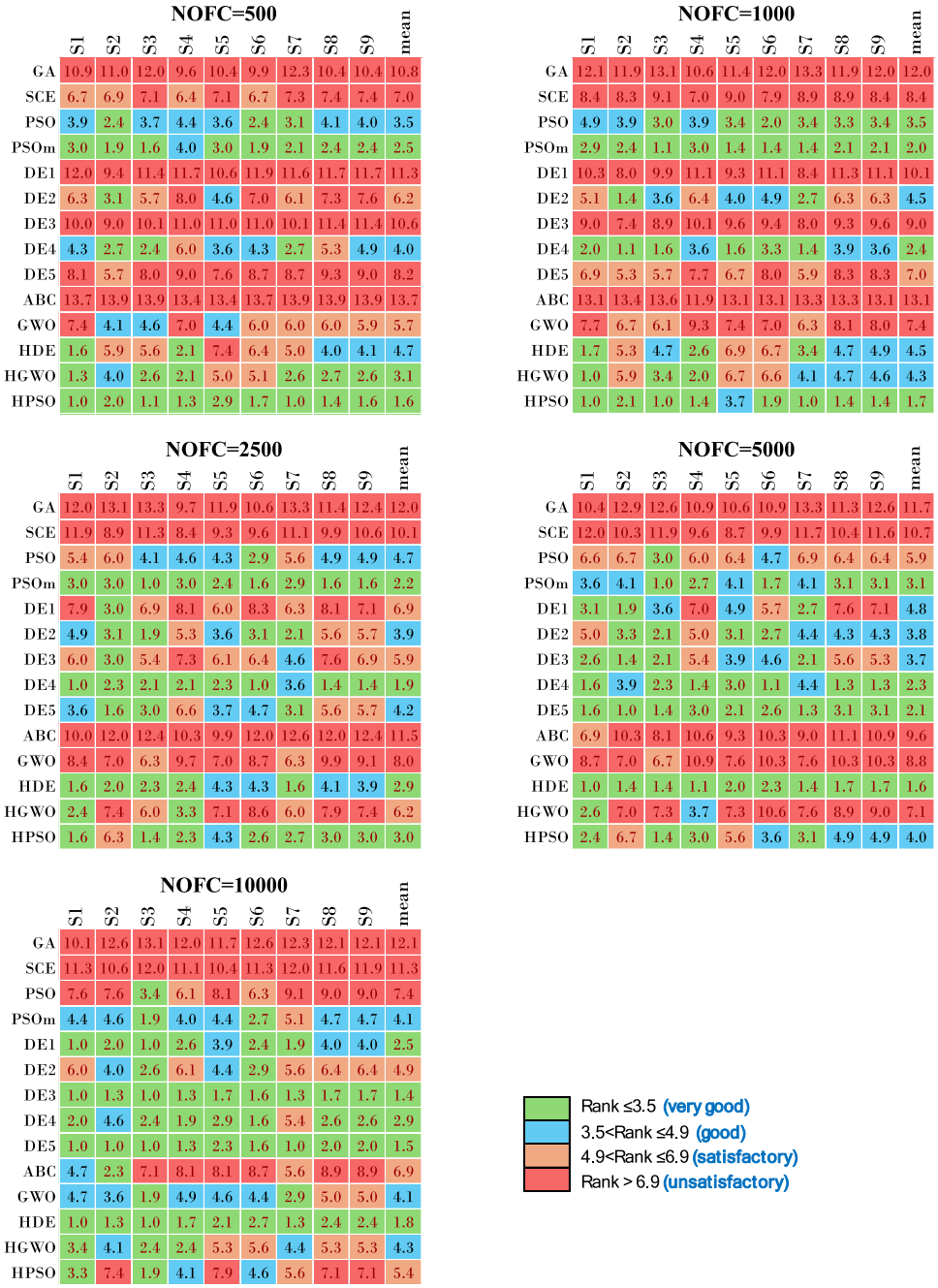

Fig. 5 Ratings of calibration algorithms operated with varying numbers of function calls over the Beydag watershed for multiple TOPSIS scenarios. The last column of each diagram contains the overall mean of the ranks derived from all scenarios

ranks, and the inter-model variation of these rankings is comparatively marginal. Yet, it has also been found that the rank values pertaining to the algorithms may differ based on either the NOFC level or the TOPSIS scenario used. Therefore, in our study, where many different variations emerged, we focused on overall mean values of the ranks assigned by the algorithm to various CRRMs in order to forge a unified perspective on TOPSIS scenarios that were conducted at increasing NOFC levels. Accordingly, the overall ratings of calibration algorithms that were applied at varying numbers of function calls over the Beydag watershed for nine TOPSIS scenarios are provided in Fig. 5. Besides, Fig. S10 is as in Fig. 5, but it pertains to the Tahtali watershed.

To facilitate comprehension of the final performance of optimization algorithms and to enhance the readability of Fig. 5 and Fig. S10, boxplots were prepared using the average rank values regarding the scenarios S1, S2, S3, and S7 (see Fig. 6 and Fig. S11). One of the notable observations from these figures is that HPSO, PSOm, and DE4 are much more prominent throughout these four scenarios, particularly when evaluating 500 and 1000 function calls. This is also in accordance with the previous inferences that were

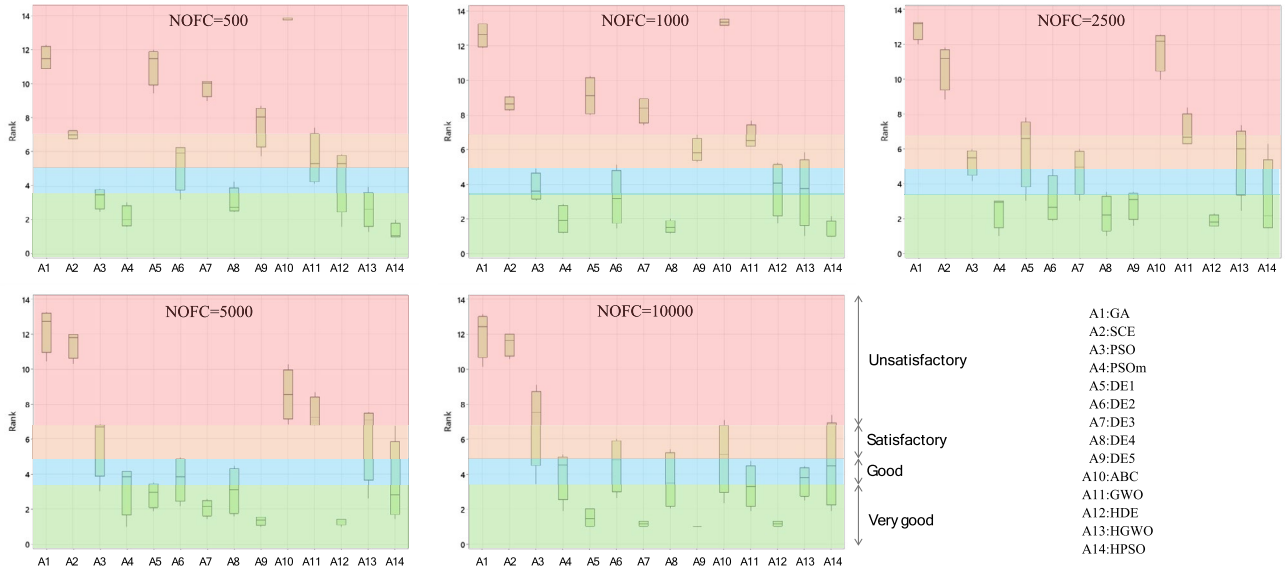

Fig. 6 Boxplot comprising the ranks obtained from the S1, S2, S3, and S7 scenarios for the Beydag watershed. That is, each box diagram was created using four rank values

made for Fig. 4. It has been emphasized by Piotrowski et al. (2019) that the inclusion of excessive function calls does not always enhance the calibration performance and may even slightly deteriorate the performance regarding the validation data, leading to a form of overfitting. Accordingly, these algorithms, which can be executed with a minimal number of function calls, may allow for modeling reliability during both calibration and validation periods.

Another key finding is that as the number of function calls increases from 2500 to 10,000, the DE variants, especially DE3, DE5, and hybrid HDE, shift to the *very good* category and even emerge as the top performers in all four scenarios. Conversely, GA and SCE did not seem to benefit as much from an increase in function calls. In fact, it is reported that the adaptability of these well-known types is typically restricted by the premature convergence issue rather than by the choice of NOFC (Pandey et al. 2014). Moreover, it is promising that the ranking of ABC tends to improve slightly as NOFC increases, and perhaps it might compete with others as the calibration time gets longer. In summary, there might be some potential reasons why certain algorithms leverage NOFC effects differently, resulting in distinct performance profiles:

i. While certain algorithms need to be subjected to extensive parameter tuning to perform well when running with a limited number of function calls, as the number of function calls goes up, these algorithms may have more opportunity to find better solutions regardless of their fine-tuning needs.
ii. Some algorithms, such as HPSO and PSOm, which are equipped with mutation or derivative schemes focusing on rapidly converging to the local optima, can outperform others under a small number of function calls. Yet, their adaptation abilities may deteriorate somewhat when they are forced to do much more exploration, which means visiting utterly new locations within the search space.

However, it should be noted that the above detection varies depending on which algorithms the modifications are made to. Given that DE algorithms have been previously demonstrated as a more viable option for hydrological model calibration compared to PSO variants (Napiorkowski et al. 2023), it might not be unexpected that some DE variants, such as HDE and DE5, dynamically adapt to their control parameters due to increments in NOFCs and thus excel in longer calibrations.

Convergence speed analyses

It has been reported in the literature that many optimization algorithms can reach solutions of comparable quality, which raises the question of whether they are likewise quick. Nonetheless, we implemented innovative hybrid strategies and versatile mutation schemes with rapid convergence features and sought to figure out whether their convergence speed exceeded that of well-known GOAs. While doing this, we again performed TOPSIS under

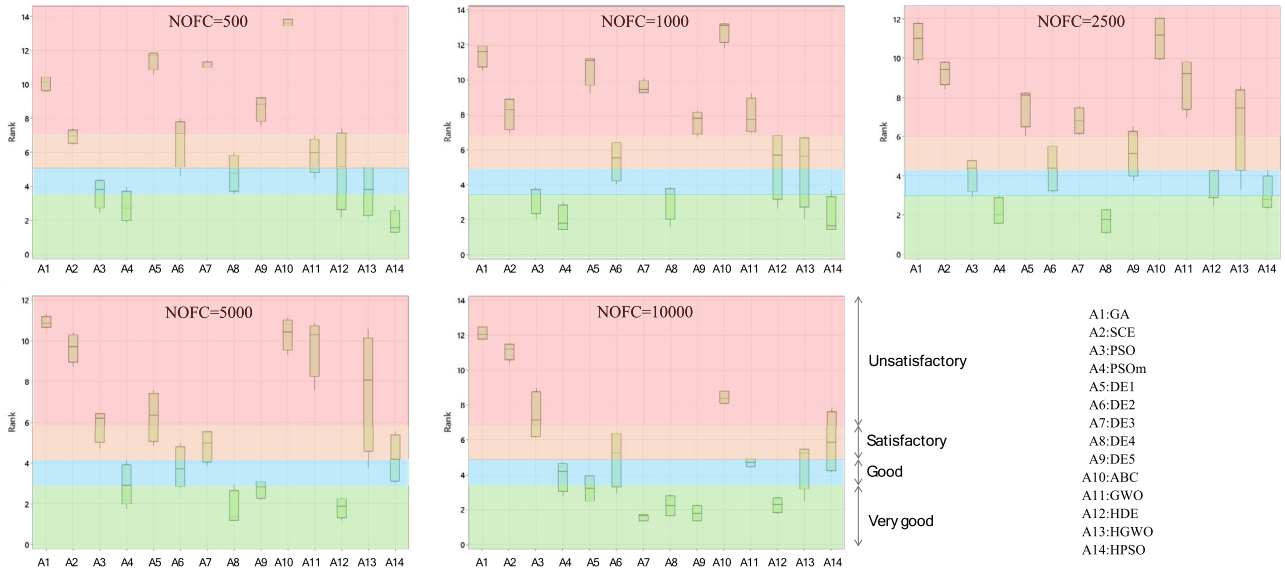

Fig. 7 Boxplot depicting the rank values of the S4, S5, S6, and S8 scenarios, which are explanatory for grading the algorithms' convergence speeds over the Beydag watershed

several conditions. Accordingly, scenarios S4, S5, and S6 ranked the optimization algorithms for each objective function based on their convergence speed, while scenario S8 repeated similar operations, giving equal weights to metrics regarding the three objective functions. Due to the nature of different optimization algorithms, their convergence speed rankings could be influenced in several ways as the number of NOFCs increases, as evidenced by Fig. 5 and Fig. S10. Besides, boxplots were once again prepared to more easily interpret the algorithm ranking regarding scenarios S4, S5, S6, and S8 versus assigned NOFCs (see Fig. 7 and Fig. S12). The subsequent remarks can be made considering all of these figures:

- The increase in convergence speed with more function calls has become more prominent in DE variants. Especially when the number of function calls is set to 2500 or more, the DE4, which adopts a different mutation strategy, has gained the capability to search for neighboring solutions more densely around promising solutions, making it stand out in terms of the exploitation mechanism. Furthermore, DE3, DE4, DE5, and HDE exhibited no stagnation at 10,000 function calls, emerging as the most convergent for nearly all four scenarios.
- Even though both HPSO and HDE used derivative-free strategies for global search and then switched to the same LM process for local refinement, they displayed varying convergence rates. This is due to the fact that they need different NOFCs to balance the trade-offs between the two intertwined stages of hybridization. Especially, HPSO made rapid progress toward the optimal solutions for 500–1000 function calls; however, it exhibited diminishing returns in terms of convergence speed as the number of NOFCs shifted up. Contrary to this, HDE required much more function calls to efficiently exploit the gradient information on its own. These have shown that to influentially transition between derivative-free and derivative methods within the hybrid structure and, hence, overcome issues that slow down convergency, it might be necessary to employ a variable number of function calls, ranging from 500 to 10,000.
- Another significant observation is that PSOm maintained its ranking in the very good category for convergence up to 2500 function calls. With increasing function calls, PSOm may have stagnated around the local optima and encountered challenges in exploring new regions of the search space. However, it was able to reach the convergence provided by DE4, using fewer function calls, and was found to be as fast as HPSO and even more adaptable to more intricate objective functions such as KGE.

Ranking optimization algorithms by considering both criteria

Thus far, it has been revealed that the scenario selection notably influences one or more of the hybrid algorithms or DE variants. In this subsection, TOPSIS facilitated the algorithm ranking process by considering both the final performance and the convergence speed simultaneously. TOPSIS operated with scenario S9 allowed for the incorporation of equal weighting for each criterion, enabling us to avoid the probable biases that could arise from the use of any singular

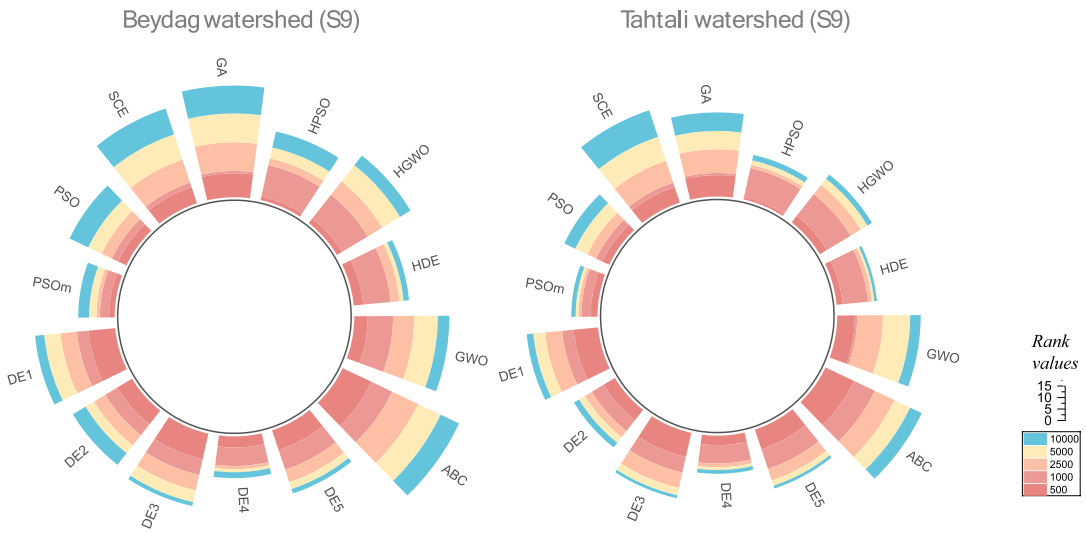

Fig. 8 Performance ratings of each algorithm at various NOFC levels based on all possible criteria. As the size of the slices in any bar expands, the algorithm's success rate diminishes

scenario. Figure 8 provided a more nuanced understanding of the performance of each optimization algorithm in comparison to others by utilizing the S9 scenario rank values from Fig. 5 and Fig. S10. Accordingly, GA, SCE, GWO, and ABC were the worst performers for nearly all NOFCs and presumably displayed a weak capacity for navigating the solution space, which, in turn, impeded their general efficacy.

Additionally, DE4, DE5, and HDE demonstrated an increasing trend of performance enhancement up to 5000 function calls, after which they kept their gradings in the *very good* category. In contrast, HPSO outperformed these DE variants during the initial phases (i.e., 500 and 1000 function calls), making it a superior option for situations where optimal calibration time would be a consideration. But it was unable to maintain this advantage in longer runs in which more NOFCs were chosen. On the other hand, PSOm appeared to be more balanced at function calls ranging from 500 to 5000, whereas HPSO, HDE, and certain DE variants offered superior ratings in more limited NOFC ranges. All these findings are somewhat consistent with those that were previously made in Sects. 3.2 and 3.3. Essentially, the fact that optimization algorithms, such as HPSO, HDE, DE4, DE5, and PSOm, are graded as *very good* at certain NOFC levels in terms of both final and convergence performances is due to their well-balanced behaviors. Cuevas et al. (2014) highlighted that achieving a balance between visiting new points, i.e., exploration, and refining previously visited locations is one of the expected features of optimization algorithms. Therefore, the aforementioned methodologies may be considered viable in this context. Nevertheless, the model user's choice should be guided by the specific demands of the calibration task. To summarize, it is possible to assert that both PSOm and HPSO would be rather ideal if the CRRMs are parsimonious, as in this study. In the event that the hydrological model calibration is influenced by problem dimensionality and non-convexity, HDE, DE4, and DE5 might be more suitable, provided that there is sufficient computational capacity and time.

Overall hypothesis testing for NOFCs

Given that the 14 optimization algorithms evaluated in this study are all stochastic based, it is usual that the solutions pertaining to the hydrological model calibrations may vary between NOFC conditions and even with respect to each other. As seen in the study, any optimization algorithm run under a certain NOFC setting could give 30 independent results. Having stored these for all variations, non-parametric Mann–Whitney U (MW) test was applied for pairwise comparison at 5% significance level, and it was questioned whether the performances of the algorithms under different objective functions came from the same distribution. Accordingly, the frequencies of the cases of being from the same distribution (H_0 hypothesis) are shown in Fig. 9.

As can be clearly seen from Fig. 9, while the variability between the algorithms'outputs is more pronounced at low NOFC levels (i.e. 1000), the ultimate performance similarity between the algorithms presents a more consistent pattern when NOFC is set to 10,000. On the other hand, when watershed-based comparison is made, the readily establishment of rainfall-runoff relationships over Tahtali watershed makes the differences between the algorithms less marked. In addition, GA and SCE algorithms could not show similarity with the other algorithms even at high NOFCs, and this is line with those mentioned in the previous sections. However,

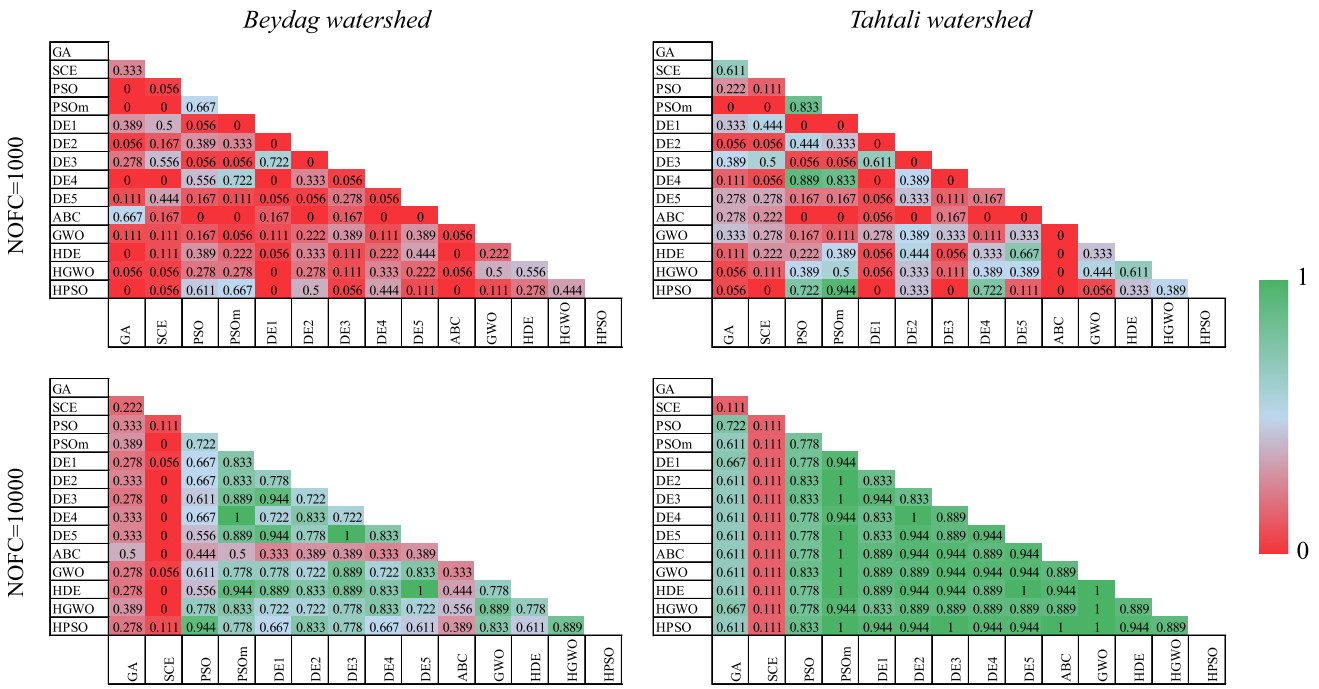

Fig. 9 Pairwise similarity matrix of optimization algorithms based on MW results over two NOFCs. Green shades indicate high similarity, while red shades indicate significant differences

it should be noted that MW can examine the distribution identity of the ultimate performances of the algorithms. Considering that there are cases where the convergence rate may weaken at high NOFCs, it is usual that the algorithm performance rankings attributed in the previous sections do not completely overlap with the MW results.

Final performance stability via bootstrap confidence interval analysis

To further assess the reliability and consistency of the final performance of the calibration algorithms, bootstrap confidence intervals (CIs) were derived on the means of the normalized indices under two sample NOFC conditions (1000 and 10,000). The 95% confidence intervals obtained using 10,000 bootstrap samples could give a preliminary idea about the stability and uncertainty of the calibration algorithms at two different NOFC levels. The results for the Beydag watershed are presented in Fig. 10, while those for the Tahtali case are shown in Fig. S13.

According to the findings, especially in the Beydag watershed, setting NOFC to 1000 gives relatively wide CIs as expected, which re-emphasizes the uncertain performance of some algorithms at different objective functions. A similar trend is observed for the Tahtali case, yet the findings seem to be slightly more stable than those in Beydag. As NOFC increases up to 10,000, the CIs tend to narrow and indicate that higher NOFCs often contribute to more stable solutions. However, while increasing NOFC levels increased stability, they did not completely eliminate the variability in the final performance of the algorithms.

Performance comparison of CRRMs

As each algorithm has the potential to produce optimal parameter sets in at least one of the available runs, it would be beneficial to conduct a more thorough examination of the results from a hydrological standpoint. Therefore, this section investigated the trade-offs between LNS and NS and inquired as to which conceptual models generated more robust simulations. Investigating these trade-offs helps to identify the CRRMs that not only capture the general behavior of the watershed system but also ensure low–high flow accuracy, ultimately leading to more hydrologically sound decisions for the study regions.

The extent to which a calibration based on LNS maximization affects the metrics obtained with NS-based calibration is questioned through the r^2 criterion proposed by Nash and Sutcliffe (1970). This criterion refers to the proportion of the initial variance unaccounted for by the model calibrated with NS maximization, and the extent to which this is then attempted to be accounted for by the model calibrated with LNS maximization can be examined (Eqs. 10 and 11).

$$r^2(NS) = 100 \times (NS_{f2} - NS_{f1})/(1 - NS_{f1}) \quad (10)$$

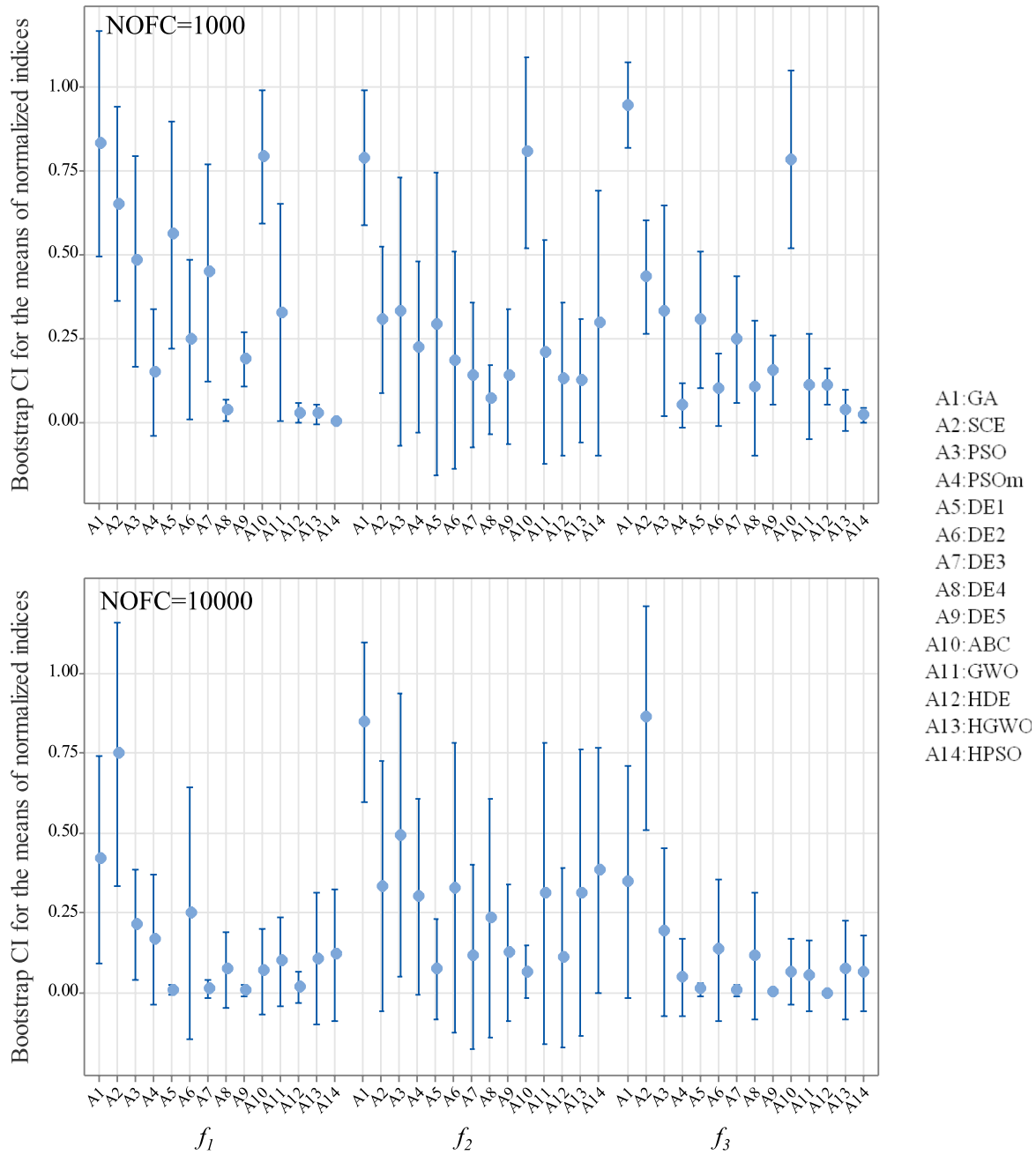

Fig. 10 Bootstrap confidence intervals (CIs) for the means of normalized indices in Beydag watershed under two NOFCs

$$r^2(LNS) = 100 \times (LNS_{f2} - LNS_{f1})/(1 - LNS_{f1}) \quad (11)$$

Besides, we interpreted the NS and LNS ratings of the CRRMs according to Moriasi et al. (2007): <0.50 (Unsatisfactory), 0.501–0.65 (Satisfactory), 0.651–0.75 (Good), 0.751–1.00 (Very Good). Table 3 summarizes general performance ratings provided by the models for the calibration and validation periods when NS or LNS maximization are chosen as the objective function. Accordingly, most of the models performed in the Beydag watershed are assessed as *good* as for their NS performances during the calibration period when the objective function is set to f_1. Although there are several CRRMs that can be rated as *very good* in respect to the same criterion during the validation period for this study region, it is important to note that the Dynwbm based on the Budyko framework yielded *very good* simulations during both periods. The fact that all models calibrated in the Tahtali watershed with the same objective function have produced runoff simulations that fell into the *very good* category in all periods indicates that the rainfall-runoff

Table 3 Comparative performance metrics of models calibrated using NS and LNS maximization, highlighting calibration and validation ratings in (a) Beydag watershed and (b) Tahtali watershed

(a)

f	*Criteria*	abcde	Awbm	Dynwbm	Gr2m	Guo	Temez	Twbm
f_1	NS_{cal}	0.725	0.678	0.790	0.654	0.664	0.676	0.743
	NS_{val}	0.823	0.673	0.793	0.804	0.731	0.743	0.722
f_2	NS_{cal}	0.472	0.459	0.727	0.286	0.211	0.180	0.255
	NS_{val}	0.436	0.551	0.707	0.532	0.065	0.480	0.470

f	*Criteria*	abcde	Awbm	Dynwbm	Gr2m	Guo	Temez	Twbm
f_1	LNS_{cal}	0.462	0.268	0.587	0.193	0.240	0.227	0.333
	LNS_{val}	0.493	0.315	0.591	0.354	0.337	0.255	0.265
f_2	LNS_{cal}	0.688	0.560	0.692	0.605	0.644	0.486	0.614
	LNS_{val}	0.743	0.646	0.770	0.727	0.663	0.527	0.635

(b)

f	*Criteria*	abcde	Awbm	Dynwbm	Gr2m	Guo	Temez	Twbm
f_1	NS_{cal}	0.848	0.857	0.812	0.851	0.844	0.808	0.845
	NS_{val}	0.924	0.882	0.934	0.946	0.862	0.895	0.861
f_2	NS_{cal}	0.577	0.786	0.640	0.616	0.388	0.629	0.813
	NS_{val}	0.636	0.902	0.856	0.754	0.376	0.883	0.839

f	*Criteria*	abcde	Awbm	Dynwbm	Gr2m	Guo	Temez	Twbm
f_1	LNS_{cal}	0.801	0.496	-0.623	-0.011	-0.003	0.650	0.864
	LNS_{val}	0.858	0.880	-1.405	0.891	0.570	0.656	0.912
f_2	LNS_{cal}	0.828	0.668	0.827	0.658	0.739	0.826	0.885
	LNS_{val}	0.877	0.912	0.825	0.701	0.899	0.910	0.918

Very Good Good Satisfactory Unsatisfactory

relationship can be better established for this watershed. Several factors can be responsible for the disparities in the performance of CRRMs over two neighboring watersheds, even if they are in proximity. As noted by Xu and Vandewiele (1994), increasing the calibration data length from five years to ten years in monthly rainfall-runoff models can markedly enhance the models' capacity to generate stable simulations. They also highlighted that extending the calibration sample length from 10 years to 15 or 20 years did not yield a notable enhancement. The relatively short length of the natural streamflow observations for the Beydag watershed where the calibration process carried out with only 7 years of data may have resulted in slightly poorer simulation performance for this region.

Table 3 clearly demonstrates that most of the models tend to exhibit unsatisfactory LNS performances for the Beydag watershed, provided that they are calibrated so as to maximize NS. This is because this objective function focuses on high flows that contribute more to the overall variance. For the Beydag watershed, the Dynwbm demonstrated a distinct superiority over the other models regarding NS statistics and produced LNS values that can be considered *satisfactory*,

with a value of nearly 0.6. Besides, in the Tahtali watershed, the parameter estimates of the abcde and Twbm models calibrated through NS maximization were found to precisely simulate low flow conditions during both the calibration and validation periods. Olsen et al. (2013) stated that modifications to model structures may be more essential than the selection of objective functions for enhancing low flow simulations. The incorporation of an additional parameter into the original groundwater storage functions within the Dynwbm, abcde, and Twbm models appears to indirectly align with the inference posited by Olsen et al. (2013).

While the objective function was f_2, the CRRMs were oriented towards minimizing errors in low flow simulations at the expense of potentially failing to represent high flow events, and this choice resulted in an apparent deterioration in the NS values (Table S11). This is attributed to the tendency of peaks to flatten out through the logarithmic transformation of streamflow data (Krause et al. 2005). Nevertheless, the degree to which LNS maximization influences NS statistics is not always the case, and the trade-offs may be pretty minor for certain models. For example, the NS values yielded by the Dynwbm calibrated with LNS maximization in the Beydag watershed during the calibration and validation periods are 0.727 and 0.707, respectively, which are deemed to be of *good* quality. As for the Tahtali watershed, the calibration exercises conducted in a manner that prioritizes low flows have resulted in the models being categorized as either *good* or *very good* in terms of LNS performances, which is a more plausible outcome than that of the Beydag watershed.

Given that the reservoir operation processes in the two watersheds where monthly rainfall-runoff relationships are modeled with various lumped approaches are conducted independently of the provision of ecological flows during dry periods, it is anticipated that the inflows simulated with only the conceptual parameters providing NS maximization will constitute sufficient input to the operational studies of the existing reservoirs. Particularly, even when f_1 is chosen as the sole objective function, the Dynwbm for the Beydag watershed and the abcde and Twbm for the Tahtali watershed are the models that provide the NS-LNS balance and can allow for a robust representation of the overall hydrological regime.

Study limitations

Although this study covers a broad scope, it has certain limitations. One key limitation is the fixed NOFCs used in the experiments. While multiple NOFC levels (i.e., 500, 1000, 2500, 5000 and 10,000) were tested, the study neglects to investigate adaptive NOFC strategies. Some auto-tuning strategies embedded into evolutionary algorithms using fuzzy logic control have adopted dynamic strategies to balance exploration and exploitation phases, reducing unnecessary function calls by adjusting control parameters based on the convergence behavior (Lin and Gen 2009). Those adaptive strategies can improve the calibration efficiency of CRRMs by avoiding premature convergence and assuring comprehensive exploration of the solution space. By restricting NOFC to predefined values, this study may not fully leverage the potential of adaptive strategies, where fewer or more function calls might be required depending on model complexity. Furthermore, while this research focuses on conceptual rainfall-runoff models, future studies should investigate the effectiveness of such adaptive strategies in calibrating more physically based, parameter-intensive models, where optimizing the number of function evaluations could be even more critical due to their more complex and demanding calibration requirements.

Conclusions

This study explored the calibration of seven CRRMs using 14 optimization algorithms across two watersheds in Turkey, benchmarking both hybrid and standard metaheuristic algorithms in terms of final performance and convergence speed. Unlike previous studies that focus solely on individual algorithm efficiency, this research systematically benchmarks hybrid metaheuristic algorithms against conventional optimization methods in the calibration of seven CRRMs. This allows for a comprehensive comparative assessment across different algorithmic families. Prior studies often overlook the relationship between NOFC and calibration performance. This study uniquely explores how different NOFC levels (500–10,000) influence convergence behavior and final performance, providing valuable insights into the computational efficiency of calibration processes. The results align with, and in some cases diverge from, previous research, providing novel insights into model calibration practices. The study yielded the subsequent empirical findings:

1. Consistent with Qin et al. (2018) and Okkan and Kirdemir (2020), hybrid algorithms incorporating derivative-based strategies achieve robust solutions with significantly fewer function calls compared to conventional metaheuristics. This demonstrates the computational advantages of hybridization in reducing calibration effort while maintaining high accuracy.
2. As the number of function calls (NOFCs) increased, DE variants proved highly adaptable, outperforming other algorithms. In contrast, traditional metaheuristics faced challenges due to premature convergence. While some PSO variants performed well with fewer function calls, their relative effectiveness declined as more extensive exploration was required. Overall, DE variants, particu-

larly those with advanced mutation schemes or derivative-based approaches, excelled in balancing efficiency and accuracy in long-run scenarios.

3. This study is one of the first applications of the TOPSIS decision-making framework to rank optimization algorithms for hydrological model calibration. This multi-criteria evaluation approach enables a balanced assessment of both final performance and computational efficiency, which has not been systematically implemented in previous hydrological optimization studies.
4. Achieving a balance between high and low flow simulations was also integral to the success of model calibration. In the present study, the Dynwbm model effectively balanced both NS and LNS metrics within the Beydag watershed, while the abcde and Twbm models exhibited a similar equilibrium in the Tahtali watershed, reflecting robust performance under varying flow conditions. Olsen et al. (2013) emphasized that structural modifications to models are often more influential in enhancing low flow simulations than the mere selection of alternative objective functions. By integrating additional parameters related to groundwater storage in these models, we indirectly responded to the recommendation of Olsen et al. (2013), facilitating these models'capacity to establish a comprehensive balance between NS and LNS metrics.

In summary, this study focuses on investigating the impact of NOFCs on convergence and final performance trade-offs, which has been rarely addressed in previous hydrological model calibration studies. Questioning how different optimization algorithms respond to changing NOFC conditions on hydrological models run with various objective functions and the TOPSIS strategy adopted to rank these algorithms according to the multiple scenarios combining different metrics will provide meaningful insights to hydrological model users. Nevertheless, future research should explore the application of adaptive NOFC strategies, particularly in more complex hydrological models, despite studies confirming that lumped and distributed models can lead to similar accuracy (e.g., Vansteenkiste et al. 2014).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12145-025-01885-y.

Author Contribution Z.B.E and U.O. wrote and revised the main manuscript text. O.F. and U.O. supervised the study. Z.B.E., U.O. and B.D. prepared the MATLAB codes, all figures, and tables. U.O. provided funding. All authors reviewed the manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). This study was funded by the Scientific and Technological Research Council of Turkey under Grant No.122Y083.

Data Availability No datasets were generated or analysed during the current study.

Declarations

Competing Interests The authors declare no competing interests.

References

Arsenault R, Poulin A, Côté P, Brissette F (2014) Comparison of stochastic optimization algorithms in hydrological model calibration. J Hydrol Eng 19(7):1374–1384. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000938

Boughton W (2004) The Australian water balance model. Environ Model Softw 19(10):943–956. https://doi.org/10.1016/j.envsoft.2003.10.007

Cooper VA, Nguyen VTV, Nicell JA (2007) Calibration of conceptual rainfall–runoff models using global optimisation methods with hydrologic process-based parameter constraints. J Hydrol 334(3–4):455–466. https://doi.org/10.1016/j.jhydrol.2006.10.036

Cuevas E, Echavarría A, Ramírez-Ortegón MA (2014) An optimization algorithm inspired by the States of Matter that improves the balance between exploration and exploitation. Appl Intell 40:256–272. https://doi.org/10.1007/s10489-013-0458-0

Deng C, Wang W (2021) A two-stage partitioning monthly model and assessment of its performance on runoff modeling. J Hydrol 592:125829. https://doi.org/10.1016/j.jhydrol.2020.125829

Duan Q, Sorooshian S, Gupta VK (1994) Optimal use of the SCE-UA global optimization method for calibrating watershed models. J Hydrol 158(3–4):265–284. https://doi.org/10.1016/0022-1694(94)90057-4

Elçi A, Karadaş D, Fıstıkoğlu O (2010) The combined use of MODFLOW and precipitation-runoff modeling to simulate groundwater flow in a diffuse-pollution prone watershed. Water Sci Technol 62(1):180–188. https://doi.org/10.2166/wst.2010.215

Eris E, Cavus Y, Aksoy H, Burgan HI, Aksu H, Boyacioglu H (2020) Spatiotemporal analysis of meteorological drought over Kucuk Menderes River Basin in the Aegean Region of Turkey. Theoret Appl Climatol 142(3):1515–1530. https://doi.org/10.1007/s00704-020-03384-0

Fan SKS, Liang YC, Zahara E (2004) Hybrid simplex search and particle swarm optimization for the global optimization of multimodal functions. Eng Optim 36(4):401–418. https://doi.org/10.1080/03052150410001685 21

Franchini M, Galeati G, Berra S (1998) Global optimization techniques for the calibration of conceptual rainfall-runoff models. Hydrol Sci J 43(3):443–458. https://doi.org/10.1080/02626669809492137

Gan TY, Biftu GF (1996) Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions,

and model structure. Water Resour Res 32(12):3513–3524. https://doi.org/10.1029/95WR02195
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. https://doi.org/10.1623/hysj.52.3.432
Gupta VK, Sorooshian S (1985) The automatic calibration of conceptual catchment models using derivative-based optimization algorithms. Water Resour Res 21(4):473–485. https://doi.org/10.1029/WR021i004p00473
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
He J, Lin G (2016) Average convergence rate of evolutionary algorithms. IEEE Trans Evol Comput 20(2):316–321. https://doi.org/10.1109/TEVC.2015.2444793
Hendrickson JD, Sorooshian S, Brazil LE (1988) Comparison of Newton-type and direct search algorithms for calibration of conceptual rainfall-runoff models. Water Resour Res 24(5):691–700. https://doi.org/10.1029/WR024i005p00691
Huang X, Liao W, Lei X, Jia Y, Wang Y, Wang X, Jiang Y, Wang H (2014) Parameter optimization of distributed hydrological model with a modified dynamically dimensioned search algorithm. Environ Model Softw 52:98–110. https://doi.org/10.1016/j.envsoft.2013.09.028
Kang L, Zhang S (2016) Application of the elitist-mutated PSO and an improved GSA to estimate parameters of linear and nonlinear Muskingum flood routing models. PLoS ONE 11(1):e0147338. https://doi.org/10.1371/journal.pone.0147338
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
Leon M, Xiong N (2014) Investigation of mutation strategies in differential evolution for solving global optimization problems. In Artificial Intelligence and Soft Computing: 13th International Conference, ICAISC 2014, Zakopane, Poland, June 1–5, 2014, Proceedings, Part I 13 (pp. 372–383). Springer International Publishing. https://doi.org/10.1007/978-3-319-07173-2_32
Lespinas F, Dastoor A, Fortin V (2018) Performance of the dynamically dimensioned search algorithm: influence of parameter initialization strategy when calibrating a physically based hydrological model. Hydrol Res 49(4):971–988. https://doi.org/10.2166/nh.2017.139
Lin L, Gen M (2009) Auto-tuning strategy for evolutionary algorithms: balancing between exploration and exploitation. Soft Comput 13:157–168. https://doi.org/10.1007/s00500-008-0303-2
Mirjalili S, Mirjalili SM, Lewis A (2014) Grey wolf optimizer. Adv Eng Softw 69:46–61. https://doi.org/10.1016/j.advengsoft.2013.12.007
Moriasi DN, Arnold JG, Van Liew MW, Binger RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50(3):885–900. https://doi.org/10.13031/2013.23153
Mouelhi S, Michel C, Perrin C, Andréassian V (2006) Stepwise development of a two-parameter monthly water balance model. J Hydrol 318(1–4):200–214. https://doi.org/10.1016/j.jhydrol.2005.06.014
Napiorkowski JJ, Piotrowski AP, Karamuz E, Senbeta TB (2023) Calibration of conceptual rainfall-runoff models by selected differential evolution and particle swarm optimization variants. Acta Geophys 71(5):2325–2338. https://doi.org/10.1007/s11600-022-00988-0
Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—A discussion of principles. J Hydrol 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6
Noel MM (2012) A new gradient based particle swarm optimization algorithm for accurate computation of global minimum. Appl Soft Comput 12(1):353–359. https://doi.org/10.1016/j.asoc.2011.08.037
Okkan U, Kirdemir U (2020) Towards a hybrid algorithm for the robust calibration of rainfall–runoff models. J Hydroinf 22(4):876–899. https://doi.org/10.2166/hydro.2020.016
Okkan U, Fistikoglu O, Ersoy ZB, Noori AT (2024) Analyzing the uncertainty of potential evapotranspiration models in drought projections derived for a semi-arid watershed. Theoret Appl Climatol 155:2329–2346. https://doi.org/10.1007/s00704-023-04817-2
Olsen M, Troldborg L, Henriksen HJ, Conallin J, Refsgaard JC, Boegh E (2013) Evaluation of a typical hydrological model in relation to environmental flows. J Hydrol 507:52–62. https://doi.org/10.1016/j.jhydrol.2013.10.022
Oudin L, Hervieu F, Michel C, Perrin C, Andréassian V, Anctil F, Loumagne C (2005) Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. J Hydrol 303(1–4):290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026
Pandey HM, Chaudhary A, Mehrotra D (2014) A comparative review of approaches to prevent premature convergence in GA. Appl Soft Comput 24:1047–1077. https://doi.org/10.1016/j.asoc.2014.08.025
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(10):2872. https://doi.org/10.3390/su11102872
Pérez-Sánchez J, Senent-Aparicio J, Jimeno-Sáez P (2022) The application of spreadsheets for teaching hydrological modeling and climate change impacts on streamflow. Comput Appl Eng Educ 30(5):1510–1525. https://doi.org/10.1002/cae.22541
Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ, Osuch M, Kundzewicz ZW (2017) Are modern metaheuristics successful in calibrating simple conceptual rainfall–runoff models? Hydrol Sci J 62(4):606–625. https://doi.org/10.1080/02626667.2016.1234712
Piotrowski AP, Napiorkowski JJ, Osuch M (2019) Relationship between calibration time and final performance of conceptual rainfall-runoff models. Water Resour Manage 33:19–37. https://doi.org/10.1007/s11269-018-2085-3
Pusatli OT, Camur MZ, Yazicigil H (2009) Susceptibility indexing method for irrigation water management planning: applications to K. Menderes river basin Turkey. J Environ Manag 90(1):341–347. https://doi.org/10.1016/j.jenvman.2007.10.002
Pushpalatha R, Perrin C, Moine NL, Andréassian V (2012) A review of efficiency criteria suitable for evaluating low-flow simulations. J Hydrol 420–421:171–182. https://doi.org/10.1016/j.jhydrol.2011.11.055
Qin Y, Kavetski D, Kuczera G (2018) A robust Gauss-Newton algorithm for the optimization of hydrological models: From standard Gauss-Newton to robust Gauss-Newton. Water Resour Res 54(11):9655–9683. https://doi.org/10.1029/2017WR022488
Raju KS, Kumar DN (2015) Ranking general circulation models for India using TOPSIS. J Water Climate Change 6(2):288–299. https://doi.org/10.2166/wcc.2014.074
Seiller G, Anctil F (2016) How do potential evapotranspiration formulas influence hydrological projections? Hydrol Sci J 61(12):2249–2266. https://doi.org/10.1080/02626667.2015.1100302
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. https://doi.org/10.1016/j.jhydrol.2014.01.050
Wu SJ, Lien HC, Chang CH (2012) Calibration of a conceptual rainfall–runoff model using a genetic algorithm integrated with runoff estimation sensitivity to parameters. J Hydroinf 14(2):497–511. https://doi.org/10.2166/hydro.2011.010
Xu CY, Singh VP (2001) Evaluation and generalization of temperature-based methods for calculating evaporation. Hydrol Process 15(2):305–319. https://doi.org/10.1002/hyp.119

Xu CY, Vandewiele GL (1994) Sensitivity of monthly rainfall-runoff models to input errors and data length. Hydrol Sci J 39(2):157–176. https://doi.org/10.1080/02626669409492731

Zhang L, Potter N, Hickel K, Zhang Y, Shao Q (2008) Water balance modeling over variable time scales based on the Budyko framework - Model development and testing. J Hydrol 360(1–4):117–131. https://doi.org/10.1016/j.jhydrol.2008.07.021