Computational Geosciences (2018) 22:789–814
https://doi.org/10.1007/s10596-018-9726-8

ORIGINAL PAPER

Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data

A. Mostafaie[1] · E. Forootan[2] · A. Safari[1] · M. Schumacher[3]

Received: 22 May 2017 / Accepted: 19 January 2018 / Published online: 20 February 2018

Abstract

Hydrological models are necessary tools for simulating the water cycle and for understanding changes in water resources. To achieve realistic model simulation results, real-world observations are used to determine model parameters within a "calibration" procedure. Optimization techniques are usually applied in the model calibration step, which assures a maximum similarity between model outputs and observations. Practical experiences of hydrological model calibration have shown that single-objective approaches might not be adequate to tune different aspects of model simulations. These limitations can be as a result of (i) using observations that do not sufficiently represent the dynamics of the water cycle, and/or (ii) due to restricted efficiency of the applied calibration techniques. To address (i), we assess how adding daily Total Water Storage (dTWS) changes derived from the Gravity Recovery And Climate Experiment (GRACE) as an extra observations, besides the traditionally used runoff data, improves calibration of a simple 4-parameter conceptual hydrological model (GR4J, in French: modèle du Génie Rural à 4 paramètres Journalier) within the Danube River Basin. As selecting a proper calibration approach (in ii) is a challenging task and might have significant influence on the quality of model simulations, for the first time, four evolutionary optimization techniques, including the Non-dominated Sorting Genetic Algorithm II (NSGA-II), the Multi-objective Particle Swarm Optimization (MPSO), the Pareto Envelope-Based Selection Algorithm II (PESA-II), and the Strength Pareto Evolutionary Algorithm II (SPEA-II) along with the Combined objective function and Genetic Algorithm (CGA) are tested to calibrate the model in (i). A number of quality measures are applied to assess cardinality, accuracy, and diversity of solutions, which include the Number of Pareto Solutions (NPS), Generation Distance (GD), Spacing (SP), and Maximum Spread (MS). Our results indicate that according to MS and SP, NSGA-II performs better than other techniques for calibrating GR4J using GRACE dTWS and in situ runoff data. Considering GD as a measure of efficiency, MPSO is found to be the best technique. CGA is found to be an efficient method, while considering the statistics of the GR4J's 4 calibrated parameters to rank the optimization techniques. The Nash-Sutcliffe model efficiency coefficient is also used to assess the predictive power of the calibrated hydrological models, for which our results indicate satisfactory performance of the assessed calibration experiments.

Keywords Multi-objective calibration · GRACE · GR4J · Conceptual hydrological model · Evolutionary optimization techniques · Daily total water storage

1 Introduction

Hydrological models are important for monitoring, planning, and managing water resources. Their development is

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10596-018-9726-8) contains supplementary material, which is available to authorized users.

✉ E. Forootan
ForootanE@cardiff.ac.uk

[1] School of Surveying and Geospatial Engineering, College of Engineering University of Tehran, Tehran, Iran

[2] School of Earth and Ocean Sciences, Cardiff University, Cardiff, UK

[3] School of Geographical Sciences University of Bristol, Bristol, UK

required to better understand natural processes and assess changes in the water cycle, and their response to climate change and anthropogenic modifications. From a mathematical point of view, hydrological modeling is the process of describing and quantifying a "real-world system" on the basis of forcing and input data, model parameters, and their initial values [31].

Among different hydrological models, rainfall-runoff models, which simply relate water input and output in a basin, have been extensively used for studying water resource management scenarios and prediction purposes specially for the basins without enough data, as well as for investigating the future variations of climate and land use [8, 85]. The conceptual rainfall-runoff models are usually constructed to represent the physical components of a basin. Model parameters are selected in a way that the properties of a basin of interest be represented as realistic as possible. These parameters however cannot be directly measured and should be estimated indirectly within a so-called (parameter) "calibration" procedure [21, 31]. Generally speaking, while calibrating a model, values of its parameters vary within a predefined range to achieve a good agreement between model simulations and real world observations [75].

Automatic optimization techniques are commonly applied nowadays to estimate proper parameter (sets) for hydrological models. For this, objective functions, which are single valued equations, which depend on model parameters and indicate the numerical agreement between model simulations and observed behavior of the basin of interest, are utilized to estimate the "best" values for model parameters.

Practical experiences show that single-objective calibrations are efficient for highlighting a certain property of a system (i.e., a hydrological model in this study, see, e.g., [18, 29]). This however might lead to increasing errors in some other characteristics. In other words, contraction of all differences between model simulations and observations at one variable may cause to cover or underrate the information content of observations and prevent using all available information [82, 83]. In a hydrological context, for example, performing calibration using only runoff observations adjusts model simulation towards a better runoff output, and therefore, this does not guarantee a better simulation of water states, for example, simulated soil moisture or groundwater compartments.

The above limitations can potentially be addressed by applying multi-objective calibration methods, from which classical techniques seek to define a weighted sum of primary single-objective functions, while the weights are defined by users [13]. In order to ensure that all objective functions can be improved without degrading another, recent multi-objective techniques seek to find a representative set of Pareto optimal solutions, and/or quantify the trade-offs in satisfying a number of objectives, and/or finding a single solution that satisfies particular subjective preferences [32, 35, 54, 80]. [87] explained advantages of multi-objective calibrations and showed that these schemes are applicable and ensure desired results in hydrological applications. Since then, this technique has been applied in several hydrological applications, i.e., based on weighting different objective functions, e.g., [26, 49, 51, 53, 65, 79] and based on searching for Pareto sets and population-based search techniques, e.g., [6, 10, 15, 19, 20, 34, 41, 52, 74, 82]. Efstratiadis and Koutsoyiannis [24] reported a comprehensive summary of the studies about multi-objective calibration techniques.

Traditionally, calibration of hydrological models has been implemented using only river runoff observations. For this, only few model parameters are selected to be calibrated since otherwise the well-known equifinality likely happens that can lead to a good fit to observations within the calibration period but the quality of predictions might be low [9, 72]. To mitigate this problem, some studies assessed the application of Total Water Storage (TWS) data from the Gravity Recovery And Climate Experiment (GRACE, [76]) satellite mission, besides in situ runoff data, for calibrating hydrological models. GRACE TWS data represent changes in vertical summation of all surface and sub-surface water storage change and can be used for assessing the evolution of terrestrial water storage changes and climate impact assessment [23, 27, 28]. Therefore, calibrating models against GRACE data will constrain their mass balance, which directly and indirectly improves the simulation of water storage and water fluxes, respectively. (see, e.g., [70–72]). This view will be followed in this study.

Multivariate optimization techniques have been used in previous studies to calibrate hydrological models against GRACE and runoff data. For example, [84] applied the multi-objective calibration framework of ϵ-Non-dominated-Sorting-Genetic-Algorithm-II [14, ϵ-NSGA-II,] to calibrate the WaterGAP Global Hydrology Model [17, WGHM,]. Three large river basins of Amazon, Mississippi, and Congo, where GRACE signals are very strong are considered in their research. In another attempt, [86] applied ϵ-NSGAII to calibrate the Soil and Water Assessment Tool [4, SWAT,] model for basins in the Sub-Saharan Africa. A step-wise calibration method, known as the Differential Evolution Markov Chain Monte Carlo [77], is applied in [58] to calibrate the Hydrological Predictions for the Environment [50, HYPE,] model over the Da River Basin. Finally, [64] applied the Multi-scale Parameter Regionalization (MPR) technique to improve the mesoscale Hydrologic Model ([mHM, [46, 67]) over 83 European basins with a wide range of distinct physiographic and hydrologic regimes.

The GR4J model (in French, modèle du Génie Rural à 4 paramètres au pas de temps Journalier, [62] a simple (4-parameter) daily continues lumped rainfall-runoff model) is selected here to be calibrated over the Danube River Basin (area of about 801,463 km^2). The structure of GR4J is simple, whereas the model requires few necessary input data (precipitation and actual evapotranspiration), and only 4 model parameters need to be set for simulating water storage and water flux. Our motivation to select Danube is due to the fact that its area is considerably large and its hydrological signal is strong enough to be reflected in the GRACE data. In fact, the hydrology of the Danube River Basin is complex, and GR4J cannot simulate it sufficiently. For example, the GR4J version used in this study does not account for snow accumulation and snowmelt, which have a considerable impact in the hydrology of the basin. This selection, however, has been intentional as we want to investigate whether adding GRACE data benefits storage and flux simulation of a hydrological model, even though the model contains obvious limitations in its structure. Once the model is successfully calibrated, it can be used for identifying sources of temporal variability in hydrological patterns of the study region without going through extensive modeling efforts (see examples in [11, 78]).

Most of previous studies apply monthly GRACE data and a priori select an optimization technique to calibrate their hydrological models. Recent studies by, e.g., [60] and [30], however, indicate that GRACE data with higher temporal resolution are beneficial for studying fast-processing weather and hydrological signals. Therefore, in this study, for the first time, GRACE-derived daily Total Water Storage (dTWS) changes, instead of the commonly used monthly or 10-day solutions, and daily in situ runoff observations are used together to calibrate a hydrological model. GRACE dTWS data contain signals related to high frequency mass changes that do not appear in temporally coarser GRACE data, thus, are potentially more appropriate for calibration purposes. We also compare the performance of five different calibration techniques to achieve reliable sets of parameters, from which four of them are widely used as multi-objective evolutionary algorithms for model calibration, including the Non-dominated Sorting Genetic Algorithm II [14, NSGA-II,], the Multi-objective Particle Swarm Optimization [65, MPSO,], the Pareto Envelope-Based Selection Algorithm II [12, PESA-II,], and the Strength Pareto Evolutionary Algorithm II (SPEA-II, [88]). The Combined objective function and Genetic Algorithm (CGA) as in [52] is also implemented to combine two single optimization procedures and calibrate GR4J against GRACE dTWS and runoff data. Selecting a variety of optimization techniques provides an opportunity to assess different aspects of the quality of optimized solutions such as their cardinality, accuracy, and diversity (see the discussions in [66]). Details of the evolutionary optimization techniques and algorithms to implement them are presented in the electronic supporting material (ESM).

In the following, the study region is introduced in Section 2. In Section 3, the GR4J model is introduced, and the relationships between model parameters, input data, and simulated water storage and water fluxes are discussed. In this section, we also introduce the datasets used for forcing and calibration of GR4J. Multi-objective calibration methods and their objective functions are presented in Section 4, while more details can also be found in the ESM. In Section 5, the numerical results of calibration, validation, and comparison of different methods are presented. Finally, the paper is concluded in Section 6.

2 Study region

The Danube River Basin, located in the Central and Eastern Europe, is the biggest river basin of Europe after the Volga River Basin, see Fig. 1. It is shared by 19 countries and is very valuable from environmental, economical, historical, and social prospective. The Danube River originates from the two small rivers of Brigach and Breg in the Black Forest mountains in Germany (i.e., they can be seen in the left side of the basin in Fig. 1). After joining these two rivers, Danube flows along a south-eastern direction with the length of 2780 km, where it is fed by at least 300 tributaries and crosses through Austria, Slovakia, Hungary, Croatia, Serbia, Bulgaria, Moldova, and Ukraine, then on the shores of Romania branches into three main distributaries of Chilia, Sulina, and Saint George within an extensive delta and it discharges to the Black Sea. The mean height of the basin is about 475 m above the sea level and its mean annual precipitation varies between 2300 to 400 mm, respectively in high mountains and in the outlet delta. Mean annual temperature of the whole basin vary from − 6.2 °C (in the high mountains, Sonnblick Observatory in the Alps in Austria) to + 12 °C (lowlands in the middle and lower parts of the basin, see, e.g., [44]). Evaporation is high in the central parts due to the high range of rainfall combined with high temperatures, e.g., 725 and 700 mm/year, respectively, in the Valley Sava, north of Dinarides, and the slopes of the Carpathian Mountains. Rates of evaporation in most of other regions vary between 500 and 650 mm/year, and the lowest rate is reported to be 100 mm/year within the very high mountains of the central Alps. The basin contains glacier-covered mountains thus snow pack accumulation is expected to play a role in generating runoff. This impact, however, has not been accounted for in the GR4J model used here. Using this setup, we will assess whether adding GRACE data in the calibration step can improve simulations

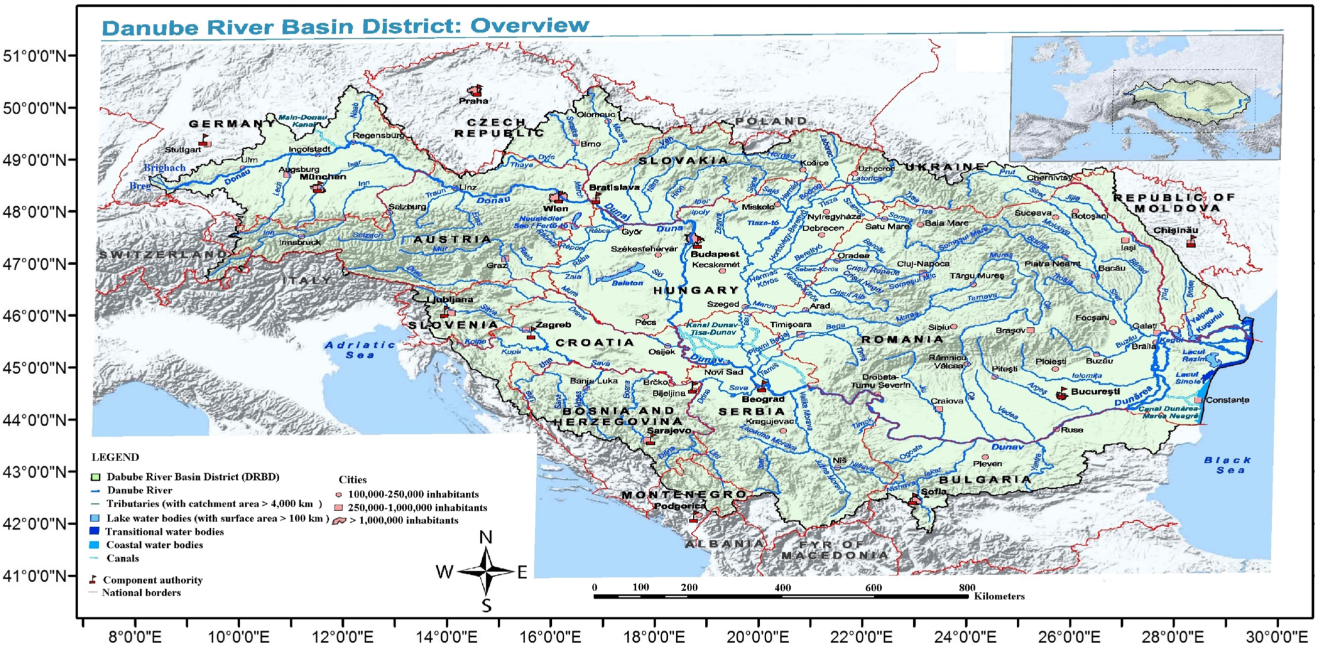


Fig. 1 Map of the Danube River Basin (taken from [25]). This map contains the main information about the Danube river basin, e.g., Danube river path and its main tributaries, topographic, national boundaries, and major cities. The two beginning tributaries of Danube, Brighach and Breg, are depicted in the western part of the basin

of a model with obvious structural limitations. Table 1 summarizes the long-term average of basin water balance components [69]. More environmental details about the region can be found under https://www.icpdr.org/main/.

3 Model and data

3.1 The GR4J model

GR4J is a simple rainfall-runoff model and belongs to the family of hydrological models that focus on the soil moisture compartment [62]. Maximum capacity of production store (X1, mm), groundwater exchange coefficient (X2, mm), maximum capacity of non-linear routing store (X3, mm); and time base of the unit hydrograph (X4, days) are its four parameters ([22]; [61]). The typical input of GR4J is precipitation and evapotranspiration. It can also be calibrated relatively quickly, therefore, various versions of this model have already been successfully used in different regions (e.g., [3, 7, 16, 37, 38, 48, 63]).

The structure of the model is very simple and consists of a soil moisture accounting reservoir, a water exchange function in the production module, two unit hydrographs, and a non-linear routing store in the transfer part of the model. From the 4 parameters, X1 and X2 are related to the water balance and the other parameters are related to the transferring of water. All four parameters accept real numbers, where X1 and X3 are always positive, X4 is greater than 0.5, and X2 accepts zero and also positive and negative numbers [62]. Figure 2 shows the general structure of GR4J and its description can be found in the following.

Given P (an estimation of the catchment precipitation) and E (a mean inter-annual of potential evapotranspiration) as the input for GR4J model, according to the GR4J flowchart (Fig. 2), first the values of net precipitation (Pn) and net evapotranspiration (En) are calculated. Then, the fraction of Pn and En, which goes to the production

Table 1 Long-term average of the Danube River Basin water balance components [69]

Long-term mean values (mm year^{-1})					
Precipitation (P)	Evaporation (E)	Runoff (Q)	Balance Error $d = (P - E - Q)/P$	Runoff Coefficient $a = Q/P$	Discharge $m^3 s^{-1}$
816	547	264	+0.60	0.32	6841

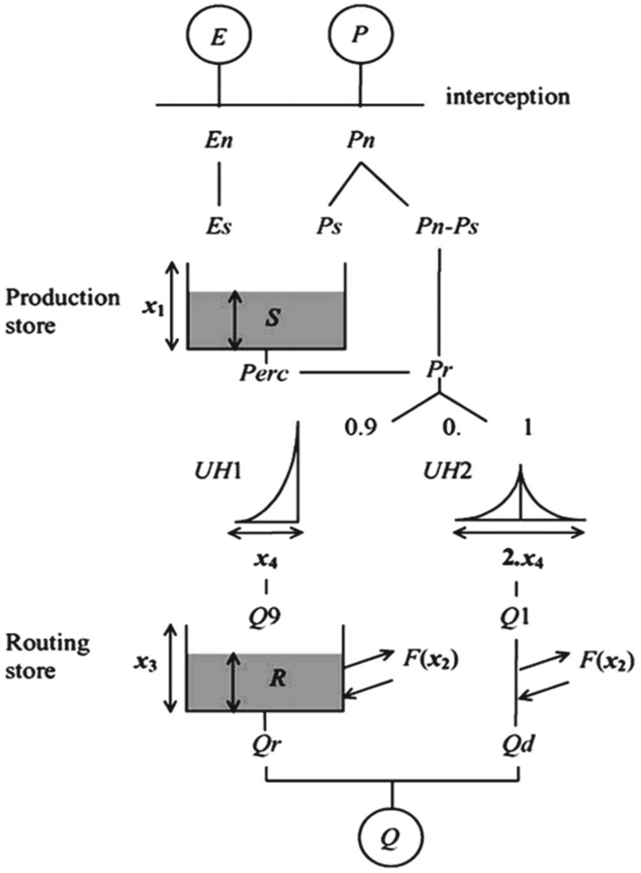


Fig. 2 Digram of the GR4J rainfall-runoff model (taken from [62])

reservoir (respectively Ps and Es), are computed using Eqs. 1 and 2 as

$$Ps = \frac{X1\left(1-(\frac{S}{X1})^2\right).tanh(\frac{Pn}{X1})}{1+\frac{S}{X1}.tanh(\frac{Pn}{X1})}, \tag{1}$$

and

$$Es = \frac{S\left(2-(\frac{S}{X1})^2\right).tanh(\frac{En}{X1})}{1+(1-\frac{S}{X1}).tanh(\frac{En}{X1})}. \tag{2}$$

The production store level is updated through $S = S - Es + Ps$. Then, the amount of percolation ($Perc$) coming from the production store is calculated using Eq. 3, and the production store level is then again updated through $S = S - Perc$. It is predefined that 90% of water quantity that finally reaches to the routing part of the model ($Pr = Perc + (Pn - Ps)$) being routed by the first unit hydrograph (HU1), and the remaining 10% by second unit hydrograph (HU2). These two unite hydrographs depend on the parameter X4. The hydrograms ordinates are calculated from the S curves (the accumulation of the proportion of unit rainfall treated by the hydrogram in function of time), respectively named $SH1$ and $SH2$, which at time t are defined by Eqs. 4 and 5, respectively.

$$Perc = S\left\{1-\left[1+(\frac{4}{9}\frac{S}{X1})^4\right]^{\frac{-1}{4}}\right\}, \tag{3}$$

$$SH1(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ (\frac{t}{X4})^{\frac{5}{2}} & \text{if } 0 < t < X4 \\ 1 & \text{if } t \leq X4 \end{cases}, \tag{4}$$

$$SH2(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ \frac{1}{2}(\frac{t}{X4})^{\frac{5}{2}} & \text{if } 0 < t < X4 \\ 1-\frac{1}{2}(2-\frac{t}{X4})^{\frac{5}{2}} & \text{if } X4 < t < 2X4 \\ 1 & \text{if } t \leq 2X4 \end{cases}, \tag{5}$$

The ordinates of HU1 and HU2 are then obtained from $UH1(j) = SH1(j) - SH1(j-1)$ and $UH2(j) = SH2(j) - SH2(j-1)$, respectively with j being an integer. The output of these unit hydrographs ($Q9$ and $Q1$) is calculated for each time step i, using Eqs. 6 and 7 that are

$$Q9(i) = 0.9\sum_{k=1}^{l} UH1(k).Pr(i-k+1), \tag{6}$$

and

$$Q1(i) = 0.1\sum_{k=1}^{m} UH2(k).Pr(i-k+1), \tag{7}$$

where l= int(X4)+1 and m= int(2.X4)+1, with $int(.)$ representing the integer part. The amount of groundwater loss/gain (F, i.e., groundwater exchange term) is calculated in Eq. 8

$$F = X2(\frac{R}{X3})^{\frac{7}{2}}, \tag{8}$$

which R is the routing store level and F is positive in case of a gain, and negative in case of a loss, or null. The level in the routing store is updated by adding the $Q9$ output of the hydrogram HU1 and F as $R = max(0; R + Q9 + F)$ and then, it empties in an output Qr given in Eq. 9,

$$Qr = R.\left\{1-\left[1+\left(\frac{R}{X3}\right)^4\right]^{\frac{-1}{4}}\right\}. \tag{9}$$

The level in the routing store is updated to $R = R - Qr$ and the output $Q1$ of the hydrogram HU2 goes through the same exchanges to give the flow component Qd=max (0; $Q1 + F$). Finally, the total runoff Q is estimated as

$$Q = Qr + Qd. \tag{10}$$

Selecting a priori values for model parameters is usually done experimentally or according to the literature. Since

GR4J is a conceptual model, the range of its parameters should be modified to enable a realistic simulation of runoff within the study area. Therefore, in this study, to setup the calibration of GR4J over the Danube River Basin, we run numerical experiments, in which the range of model parameters (X1 to X4) is changed and the correlation with existing runoff data is estimated. After several trials, we conclude that the following ranges: 50 mm$< X1 <$4000 mm, -15 mm$< X2 <$15 mm, 20 mm$< X3 < 3500$ mm and 0.5 days$< X4 < 30$ days are suitable for the Danube Basin, i.e. the correlation coefficient between the GR4J's runoff and observations is not negative. Also in this study, the initial values of "production store" and "routing store" are considered as two unknown parameters X5 and X6, respectively, which represents the percentage of the volume of these two stores that are full. Thus, their values can vary between 0 and 100%.

3.2 Calculating Total Water Storage (TWS) changes from GR4J

As mentioned before (in Section 3.1), GR4J contains two storage tanks, two unit hydrographs, as well as percolation and water exchange function for each day of its simulation. GR4J-derived Total Water Storage (TWS) is computed as the amount of water stored in the production store and routing store (respectively S and R) at the end of each day. The other two components are the remaining water in unite hydrographs ($HU1$ and $HU2$) during the simulation process at the end of each day, which are shown here by V and W, respectively. For the study basin, the daily model-based TWS at time t was calculated as:

$$TWS_t = S_t + R_t + V_t + W_t. \tag{11}$$

Basin averaged TWS anomalies ($dTWS$) are computed as a difference between daily TWS (TWS_t) and the temporal mean of TWS ($\overline{TWS}$) during 2003-2010 as:

$$dTWS_t = TWS_t - \overline{TWS}. \tag{12}$$

Figure 3 depicts the simulation calculations using GR4J and indicates how the inputs of this model relate to the simulation results (i.e., runoff and total water storage) regarding the model parameters. It is also shown from which part of the model each component of TWS (named here model states) is taken. This flowchart shows the generation of runoff (Q) and TWS, while indicating the non-linear relationship between the parameters X1 to X4 and the model outputs Q and TWS. Basically by introducing daily precipitation minus evapotranspiration and knowing the current value of production store level (X5 for the beginning of simulation or S from the previous time step) and its maximum level (named X1), it is determined how much water stays in the production store (S) and how much water reaches the routing parts (unit hydrographs shown by HU1 and HU2). The amount of remaining water in the unite hydrographs (HU1 and HU2) during the daily simulation process forms V and W values, which depend on the previous outputs (P_r) and the parameter X4. In the last part, i.e., the non-linear routing store, the amount of water that leaves the basin is calculated, which is related to the X2, X3 and current value of routing store level (X6 for the beginning of simulation or R from the previous time step). The new level of the routing store (R) and its output (Q_r, which is formed in the first unit hydrograph) are calculated and finally the sum of this part and the outgoing water flow of the second unit hydrograph (shown by Q_d) generates the GR4J's final simulated runoff (Q).

3.3 Climate data

To run GR4J, the input data of precipitation and evapotranspiration within the Danube River Basin is extracted from daily 0.5° × 0.5° gridded data of the European Center for Medium range Weather Forecasts (ECMWF). E-OBS daily gridded observations are used for precipitation. Evapotranspiration data are calculated following [36] using temperature observations. The version 8 of these data cover 1950-01-01 to 2012-12-31.

In this study, daily runoff data of Ceatal Izmail station is downloaded from the Global Runoff Data Center (GRDC), Koblenz, Germany, and used for calibration and validation of the GR4J model. The station is located at the outlet of the Danube River Basin with geographical position 45.22 °N and 28.73 °E (http://grdc.bafg.de/). Figure 4 shows the mean precipitation, temperature, and evapotraspiration during 2000–2010. The position of the Ceatal Izmail station and the Danube River is also shown (see also Fig. 1).

3.4 GRACE data

GRACE (Gravity Recovery And Climate Experiment, [76]) satellite mission measures global changes in the Earth's gravity field, which are mainly caused by changes in the terrestrial water storage. GRACE data can be converted to Total Water Storage (TWS) changes, that over continents mainly reflect surface and sub-surface water storage changes. Therefore, GRACE TWS is applied here, beside the in situ river runoff data, to calibrate the GR4J model.

In this study, ITSG-Grace2016 daily GRACE gravity field data, up to degree and order 40 [42], are used to calculate daily TWS changes (dTWS). This data cover 2003 onwards and are updated continuously. Since our forcing data are only available from 2000 to 2010, daily ITSG-Grace2016 data of 2003 to 2010 are converted to basin-averaged dTWS following [47]. Figure 5 shows the time series and its comparison with the observed river runoff

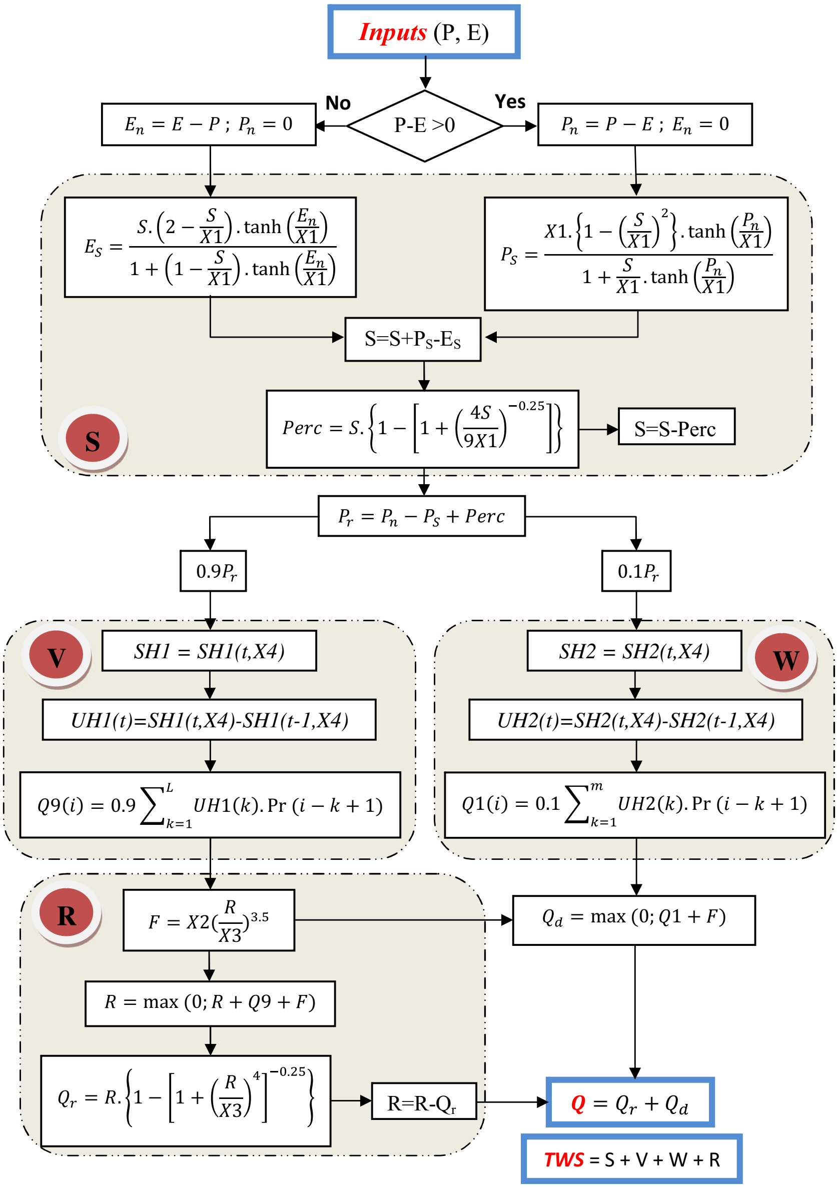

Inputs (P, E)
No
P-E >0
Yes
$E_n = E - P$; $P_n = 0$
$P_n = P - E$; $E_n = 0$
$E_S = \frac{S.\left(2 - \frac{S}{X1}\right).\tanh\left(\frac{E_n}{X1}\right)}{1 + \left(1 - \frac{S}{X1}\right).\tanh\left(\frac{E_n}{X1}\right)}$
$P_S = \frac{X1.\left\{1 - \left(\frac{S}{X1}\right)^2\right\}.\tanh\left(\frac{P_n}{X1}\right)}{1 + \frac{S}{X1}.\tanh\left(\frac{P_n}{X1}\right)}$
S=S+P$_S$-E$_S$
S
$Perc = S.\left\{1 - \left[1 + \left(\frac{4S}{9X1}\right)^4\right]^{-0.25}\right\}$
S=S-Perc
$P_r = P_n - P_S + Perc$
$0.9P_r$
$0.1P_r$
V
$SH1 = SH1(t,X4)$
$UH1(t)=SH1(t,X4)-SH1(t-1,X4)$
$Q9(i) = 0.9 \sum_{k=1}^{L} UH1(k).\Pr(i - k + 1)$
W
$SH2 = SH2(t,X4)$
$UH2(t)=SH2(t,X4)-SH2(t-1,X4)$
$Q1(i) = 0.1 \sum_{k=1}^{m} UH2(k).\Pr(i - k + 1)$
R
$F = X2\left(\frac{R}{X3}\right)^{3.5}$
$Q_d = \max(0; Q1 + F)$
$R = \max(0; R + Q9 + F)$
$Q_r = R.\left\{1 - \left[1 + \left(\frac{R}{X3}\right)^4\right]^{-0.25}\right\}$
R=R-Q$_r$
$Q = Q_r + Q_d$
TWS = S + V + W + R

Fig. 3 Flowchart of the GR4J's computational procedure, which shows how the input observations (precipitation, P, and potential evapotranspiration, E) are related to the outputs of GR4J (i.e., runoff, Q, and Total Water Storage, TWS). After each day, TWS is calculated by summation of existing water on the 4 different parts of GR4J model i.e., $TWS = S + V + W + R$

at the outlet of the Danube River (Ceatal Izmail station). Although the plotted variables have different ranges, a high correlation coefficient of 0.78 is estimated between the time series. This strong correlation of daily river runoff anomalies and TWS anomalies suggests that river water constitutes a large part of the dTWS variation during the peaks of runoff, when other water storages are at (near) capacity and cannot absorb additional inputs of runoff or precipitation [30]. Also from Fig. 5, one can detect a time delay between changes in water storage and river runoff. This is possibly due to the fact GRACE data represents an average storage change in the whole basin but the runoff is measured here at the outlet of the Danube River. Differences in the property of these two measurements and the fact that runoff flows through the basin justifies this delay. We should also mention here that the high correlation between river runoff and dTWS does not harm the calibration of GR4J, since GRACE dTWS data are used to calibrate the sum of water states (S+V+W+R in Fig. 3), and runoff calibrates the remaining water that fluxes out of model (Q in Fig. 3). As a result, the objective functions that are formulated to use these data for model calibration are different from each other as will be shown later.

A summary of the datasets used in this study is reported in Table 2. Since GR4J is a daily model, we try to use daily data in this study. Also the used climate data (forcing data including precipitation and temperature) are gridded datasets with 0.5° × 0.5° spatial resolution, which are averaged over the Danube basin. Precipitation data and calculated evapotraspiration used as input data to run GR4J

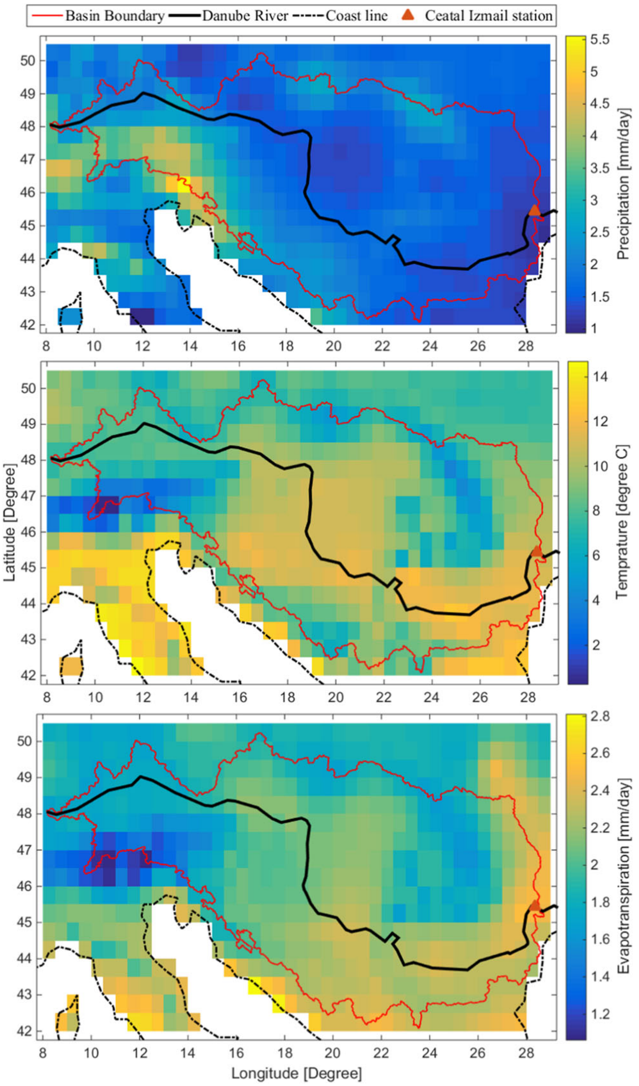


Fig. 4 Average precipitation (mm/day), average temperature (°C) and average evapotranspiration (mm/day) within the Danube River Basin during the study period (2000-2010). The red line defines the basin's boundary, the black line represents the Danube River. The red triangle represents the Ceatal Izmail station position. X and Y axes represent the longitude and latitude in degree (°), respectively

model and runoff together with basin-averaged GRACE dTWS are used for calibrating the model and validating its simulation results.

3.5 Comparing GRACE- and GR4J- Total Water Storage (TWS) changes

Figure 6 shows the dTWS anomalies simulated by GR4J within the Danube River Basin during the study period after calibration of GR4J using only runoff observations (blue) and its difference with GRACE time series of dTWS (that of Fig. 5), which is shown here in red. The results indicate a weak agreement between TWS observations and model simulations (correlation coefficient of 0.52), and their differences reach up to 121.7 mm in high peaks, e.g., in 2006 (see Fig. 6). Therefore, the model must be calibrated before being used for hydrological assessments.

4 Calibration techniques

4.1 Concept of single- and multi-objective calibration

Optimization methods in general can be used in a single-objective or multi-objective mode according to the formulation of problem. The former methods search for the optimum of one specific objective function and are defined as the minimization (or maximization) of a scalar objective function F(θ). Finding the best solution corresponds to the minimum or maximum value of an objective function is the main goal of a single-objective optimization. This type of optimization is a useful tool for exploring the nature of problems, but are not able to provide a set of alternative solutions that trade different objectives against each other. On the contrary, in multi-objective optimization, there is no single optimal solution and the goal is to find a set of solutions, which simultaneously optimize more than one (conflicting) objective functions that measure individual process descriptions. The interaction among different objectives leads to a set of compromised solutions, largely known as the trade-off, non-dominated, non-inferior or Pareto-optimal solutions [68]. According to [53], multi-objective model calibration can be performed on the basis of multi-variable measurements (e.g., groundwater level, river runoff or water content), multi-site measurements (i.e. several measurement sites of the same variable distributed within the catchment), and multi-response modes (i.e. designing objective functions that measure various responses of the hydrological processes such as the general water balance, peak flows, and low flows). A multi-objective calibration problem can be formulated as:

$$\min[\mathrm{F}_1(\theta), \mathrm{F}_2(\theta), ..., \mathrm{F}_m(\theta)], \;\; \theta \in \Theta, \tag{13}$$

where $\mathrm{F}_i(\theta)$, $i = 1, 2, ..., m$ are different objective functions. Equation 13 is a constrained optimization problem since the values of θ are limited to the feasible parameter space Θ and each model parameter has a minimum and maximum. These limits are specified according to physical and mathematical constraints, information about physical characteristics of the system and simulation experiments [45, 52, 53]. Here, θ contains the GR4J's parameters, i.e., X1, X2, X3, X4, X5 and X6, while the parameter space, i.e., Θ, is defined based on the upper and lower bounds of these

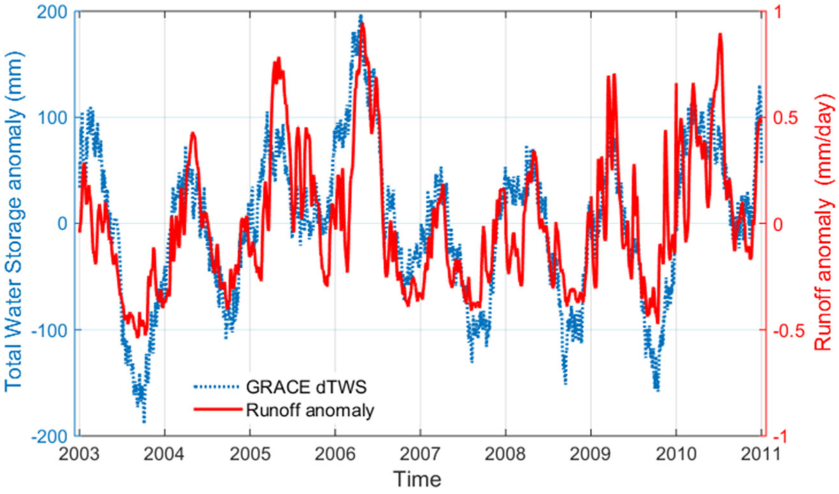


Fig. 5 Daily Total Water Storage changes (dTWS) from GRACE averaged within the Danube River Basin (blue, left y-axis) and runoff anomalies at the Ceatal Izmail station located at the Danube River Basin outlet (red, right y-axis)

parameters defined in Section 3.1. Both single- and multi-objective calibration methods are applied in this study to estimate these parameters, whose corresponding objective functions are introduced in the next section.

4.2 Objective function(s)

In this study, runoff and dTWS changes (anomalies) are applied to calibrate GR4J based on single-objective optimization methods and multi-variable measurements for multi-objective optimization methods, which the Nash-Sutcliffe model efficiency coefficient (NS, [57]) is used to verify the results. NS coefficient, which has extensively been used in hydrological applications, is a normalized measure calculated as:

$$NS = 1 - \frac{(\sum_{i=1}^{n} Y_i^{obs} - Y_i^{sim})^2}{(\sum_{i=1}^{n} Y_i^{obs} - Y^{mean})^2}, \tag{14}$$

where Y_i^{obs} and Y_i^{sim} are observed and simulated values, respectively, and Y^{mean} stands for the temporal mean of n (number of days used for calibration) observations (Y_i^{obs}). In this study, observations (Y_i^{obs}) can be either runoff or GRACE dTWS anomalies and are used separately to calculate two objective functions, i.e., $NS_{\text{dTWS}} = 1 - \frac{(dTWS_i^{\text{GRACE}} - dTWS_i^{\text{simulated}})^2}{(dTWS_i^{\text{GRACE}} - dTWS^{\text{mean GRACE}})^2}$ and $NS_{\text{Q}} = 1 - \frac{(Q_i^{\text{Observed}} - Q_i^{\text{simulated}})^2}{(Q_i^{\text{Observed}} - Q^{\text{mean Observed}})^2}$, where Q stands for runoff. The NS values vary between $-\infty$ and 1.0, where 1.0 is the optimum value and indicates the full compliance between observations and simulations. Values between 0.0 and 1.0 are generally viewed as acceptable levels of performance, whereas negative values indicates that the mean observed value is a better predictor than the simulated value, and generally interpreted as unacceptable performance [56]. Considering (14), it is clear that the optimization using dTWS and runoff likely yields different results. This is because for dTWS one replace the values of Eq. 11 will be used but for runoff we use Eq. 10, which itself depends on model runs as described in Section 3.1. In Eq. 14, the observations $dTWS_i^{\text{simulated}}$ and $Q_i^{\text{simulated}}$ are related to their equivalents simulated by GR4J. In other words, the values of model parameters ($\theta = [X1, X2, X3, X4, X5, X6]^T$) are calculated using the objective functions of $F_1(\theta) = NS_{dTWS}(\theta)$ and $F_2(\theta) = NS_Q(\theta)$.

Table 2 Summary of the datasets used in this study

Product	Type	Spatial Resolution [lat x lon]	Temporal Resolution	Coverage	Data used
GRACE TWS	ITSG-Grace2016	basin averaged	Daily	Global	2003-2010
Climate Data	Forcing (E-OBS)	0.5° x 0.5°	Daily	Europe	2000-2010
	Runoff (GRDC)	In situ Stations	Daily	Cetal-Izmail Station	2000-2010

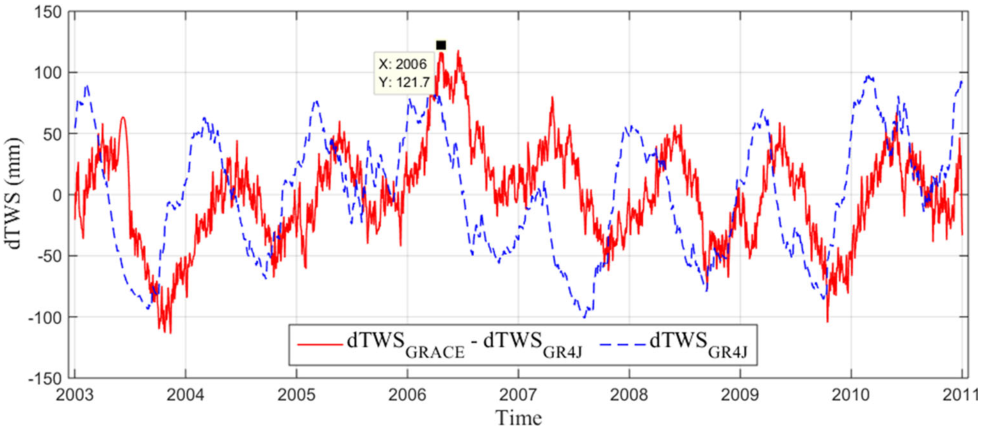


Fig. 6 Daily TWS (dTWS) anomalies simulated by GR4J within the Danube River Basin during January 2003 to December 2010. We used 161.32, 13.43, 914.18 and 12.08 for X1 to X4, respectively, that are estimated by calibrating GR4J against (only) runoff observations. Daily TWS simulated by GR4J is shown in blue. Daily differences of basin averaged GRACE dTWS and those simulated by GR4J are shown in red. A maximum difference of 121.7 (mm) is detected in the 2006 peak. GRACE dTWS time series is shown in Fig. 5

4.3 Evolutionary calibration methods

In this paper, five multi-objective evolutionary algorithms are used, where in Section 4.3.1, we combine two objective functions and in the other four the two objective functions are independently optimized to derive one set of calibrated parameters for the GR4J model (Sections 4.3.2, 4.3.3, 4.3.4, and 4.3.5). Parameter values for implementing these optimization technique are provided in Section 5.

4.3.1 Calibration using a Combined objective function and Genetic Algorithm (CGA)

CGA is a multi-objective technique, which can be considered as an extension of an ordinary genetic algorithm (GA). In other words, using CGA, it is possible to combine a number of single- objective optimization techniques and solve them using the GA process. Here, we construct a combined objective function (COF) using the NS coefficients of runoff and dTWS as

$$COF = \sqrt[2]{(1 - NS_Q)^2 + (1 - NS_{dTWS})^2}. \tag{15}$$

which represent the Euclidean distance to the optimum solution, i.e., $NS_Q = 1$ and $NS_{dTWS} = 1$ with NS being computed from Eq. 14. Equation 15 is optimized by a ordinary GA, whose description can found in the following.

GA starts with an initial population, consisting of individual solutions named chromosomes, which are formed by genes (decision variables). GA repeatedly modifies this population in a way that at each step, random samples are selected from the current population to be parents, and they are used to produce new samples (children) for the next generation. Over successive generations (iterations), populations converge and yield an optimum solution. GA can be applied to solve a variety of optimization problems that are not well suited for standard optimization algorithms, including problems, in which the objective function is discontinuous, non-differentiable, stochastic, or highly nonlinear [2]. Procedures involved in a GA can be summarized in 6 steps that are as follows:

Step (1) Initialization, where the procedure starts by setting the iteration index (t) to 1, and randomly generating N solutions to form the first population ($P1$), and subsequently evaluating the fitness of solutions in $P1$. In the case of hydrological model calibration, a population contains N sets of model parameters. A pre-defined objective function (e.g., NC coefficient in Eq. 14) is used to evaluate each solution.

Step (2) Selection, which allows the fittest individuals to pass their gens to the next generation and improve the general fitness. Two pairs of individuals (parents) with high fitness have higher chance to be selected for reproduction. There are a few different selection methods, but with similar idea, exist, which include the proportional selection, selection based on ranking, tournament selection, and roulette wheel [39].

Step (3) Crossover, which provides the chance to generate new individuals by combining their properties. This is hoped to help achieving an even fitter solution than the one derived from parents.

Step (4) Mutation, which introduces a randomness property into the genes of population and avoids generating identical populations.

Step (5) Fitness assignment, which evaluates and assigns a fitness value to each solution based on evaluating its objective function.

Step (6) If the stopping criterion is satisfied, it terminates the search and returns to the current population, else, sets t=t+1 and goes to Step 2.

4.3.2 Non-dominated Sorting Genetic Algorithm II (NSGA-II)

NSGA-II is a multi-objective evolutionary optimization algorithm, which was introduced by [14]. Elitism is implemented in the NSGA-II's multi-objective search to preserve the best members of generated model parameters. A GA is then applied to generate new sets, and subsequently, the crowding distance (an estimate of the density of solutions surrounding a particular solution), and the crowded comparison operator are respectively applied to estimate the density of solutions in the objective space and guide the selection process towards a uniformly spread Pareto-frontier (see, e.g., [5]). Generally speaking, NSGA-II contains two main features that are (1) the use of a Pareto ranking mechanism to classify solutions and (2) a density estimator known as crowding distance to maintain the diversity among the individual solutions [43]. The 6 steps of GA (in Section 4.3.1) are modified and extended to 8-steps in the NSGA-II algorithm. The procedure is described in Section 1 of the EMS.

4.3.3 Multi-objective Particle Swarm Optimization (MPSO)

Particle Swarm Optimization (PSO) is a search method that yields (a near) optimal solution for an optimization problem [40]. PSO searches the feasible space suing a population of solutions instead of a single point to find the optimal solution. In principle, within PSO, each particle moves around the space based on its position and velocity direction and improves its position using the particles' own experience (cognitive information) and observation of neighbors (social information). Best response are discovered and used to guide the movements in the next iterations. [55] extended the PSO strategy to solve multi-objective problems, thus generating the Multi-objective Particle Swarm Optimization (MPSO) technique. The structure of MPSO implemented in this study follows the one proposed by [1], and the algorithm can be found in Section 2 of the EMS.

4.3.4 Pareto Envelope-Based Selection Algorithm II (PESA-II)

PESA-II is a multi-objective evolutionary optimization algorithm, which uses the mechanism of GA together with a selection based on Pareto envelope and an external archive to store the approximated Pareto solutions. Parents and offspring are selected from the external archive using grids that created based on the geographical distribution of the archive members. Populations of solutions is PESA-II are maintained, whereas the size of internal population is fixed, and that of external population is non-fixed but limited. The task of internal population is to explore new solutions, and to achieve this by the standard evolutionary algorithm processes of reproduction and variation (i.e., crossover and mutation). The purpose of the external population is to store and exploit good solutions, which is done by maintaining a large and diverse set of the non-dominated solutions discovered during the search. Both elitism and diversity preservation mechanisms are considered in PESA-II. In this paper, the implementation steps of PESA-II follow that of [12], which are described in Section 3 of the EMS.

4.3.5 Strength Pareto Evolutionary Algorithm II (SPEA-II)

SPEA-II belongs to the field of evolutionary multiple objective algorithms, which uses an initial and an archive population. At the start, random initial and archive populations with fixed size are generated. The fitness value of each individual in the initial population and archive is calculated per iteration. Next, all non-dominated solutions of initial and external population are copied to the external set of the next iteration (new archive). Using an environmental selection procedure, the size of each archive is set to a predefined limit. Afterwards, mating pool is filled with the solutions resulted from performing binary tournament selection on the new archive set. Finally, crossover and mutation operators are applied to the mating pool and the new initial population is generated. If any of the stopping criteria is satisfied the non-dominated individuals in the new archive forms the Pareto-optimal set. The structure of SPEA-II algorithm following [88] is explained in Section 4 of the EMS.

4.4 Solution selection from Pareto front set

Multi-objective optimization techniques enable us to estimate several sets of non-dominated solutions. To apply this in practice, users need to execute (for example hydrological) models and run optimization techniques with specific goals. Finally, they should be able to select a set of model parameters from the Pareto front set as "calibrated" parameters, which are those with minimum distance (d_R) to the reference solution placed at the top of the optimization ranking as

$$d_R(f, z) = (\sum_{m=1}^{M} |f_m(x) - z_m|)^{1/p}, \tag{16}$$

where z_m is the reference solution of the optimization problem. Since in this study the NS coefficient of runoff and $dTWS$ are used as objective functions and the optimal value of NS is 1, the reference solution corresponds to

$z_1 = 1$ and $z_2 = 1$. In Eq. 16, the Euclidean distance (which implies $p = 2$) is used to estimate distances to the reference solution. Besides f_m are the objective functions and m ($m = 2$) is the number of functions, thus we have: $f_1 = NS_Q$ and $f_2 = NS_{dTWS}$.

4.5 Metrics for comparison of different algorithm results

The quality of results derived from multi-objective optimization, where two or more conflicting objective functions exist, can be assessed by different metrics, which can broadly categorized as accuracy, diversity, and cardinality metrics [59]. In general, an acceptable accuracy can be achieved by implementing a search towards the Pareto-optimal region, while the diversity requires a search along the Pareto-optimal front, and the techniques with larger number of valid results ensure cardinality. Therefore, these three criteria are considered here while discussing (dis)advantages of optimization algorithms applied here. In this study, four common metrics are applied, which include: number of Pareto solutions (NPS), generational distance (GD), spacing (SP), and maximum spread (MS). From these, NPS investigates the cardinality of algorithms, GD represents the accuracy, and the other two metrics indicate the diversity of final solutions of each algorithms.

- Number of Pareto solutions (NPS) indicates how many Pareto optimal solutions are found and is a cardinality metric; thus, a greater NPS number indicates that the algorithm is better suited to achieve this goal.
- Generational distance (GD) is an indicator of mean distance of Pareto solutions from a known Pareto-optimal value as

$$GD = \frac{1}{n_S}(\sum_{i=1}^{n_S} d_i^2)^{1/2}, \tag{17}$$

  where n_S indicates how many sets of non-dominated solutions are found and d_i is the minimum Euclidean distance (measured in objective space) between i-th solution and Pareto-optimal front solution. Thus, better optimization algorithms provide lower GD [81]. Thus, GD has been commonly used in different studies to measure the accuracy of optimization methods [66].
- Spacing (SP), suggested by [73], is a relative distance between consecutive solutions in the obtained non-dominated set. This metric is defined as

$$SP = \sqrt{\frac{1}{n_S}\sum_{i=1}^{n_S}(d_i - \overline{d})^2}, \tag{18}$$

  where d_i and $\overline{d}$ are calculated as

$$d_i = \min_{k \in n_S \wedge k \neq i} \sum_{m=1}^{M}(| f_m^i - f_m^k), \tag{19}$$

  and

$$\overline{d} = \sum_{i=1}^{n_S}(\frac{d_i}{n_S}), \tag{20}$$

  with $\overline{d}$ representing the minimum value of the sum of absolute differences between the i-th solution (f_m^i) and any other solutions (f_m^k) in the non-dominated set. An algorithm with a smaller SP indicates that the solutions are distributed (nearly) uniformly, making it a good measure to evaluate diversity.
- Maximum Spread (MS), is a metric to measure the length of the diagonal of a hyperbox formed by the extreme objective function observed in the non-dominated set. In the case of two objective problems, this metric corresponds to the Euclidean distance between the two extreme solutions (f_m) as

$$MS = \sqrt{\sum_{m=1}^{M}(\max_{i=1:n_S} f_m^i - \min_{i=1:n_S} f_m^i)^2}. \tag{21}$$

Thus, MS measures diversity and range of values covered by the final solutions set of multi-objective optimization algorithms. Larger maximum spread values indicate the better performance of these algorithms.

5 Results

5.1 GR4J parameters and their sensitivity

First, we assess how the variation of GR4J parameters affect its simulation results. Therefore, we change model parameters, and run the model and evaluate the changes in the model outputs and states, i.e., daily runoff, daily TWS anomalies ($dTWS$) and its compartments. As mentioned in Section 3, the initial values of "production store" and "routing store" are also considered as two new unknown parameters X5 and X6, respectively, beside the GR4J's original 4 model parameters X1 to X4 (maximum capacity of production store, groundwater exchange coefficient, maximum capacity of non-linear routing store, and time base of the unit hydrograph). By default, these parameters are initialized by the warm up period of model initialization. Introducing them as new parameters, therefore, allows us to assess their effects (together with other parameters) on model simulations.

Our assessments are implemented by perturbing one parameter and setting the other five to their nominal values. The ensemble of perturbed parameters is created by applying a Monte-Carlo sampling approach as:

$$\alpha_i = \alpha_{nom} + \epsilon_i, \; i = 1:50, \tag{22}$$

where α_{nom} and ϵ_i are respectively the nominal value of each parameter and the generated disturbed value. In Eq. 22, ϵ_i is selected from a normal distribution with the standard deviation d that is calculated as $d = 0.1\ (\min(\alpha_{nom} - \alpha_{min}, \alpha_{max} - \alpha_{nom}))$ with α_{max} and α_{min} being the maximum and minimum of the parameter values, respectively.

Generated ensembles in Eq. 22 are propagated by running the model from January 2003 to December 2007, and the period of 2008–2010 is used for validation. Detailed graphs are presented in the ESM (Figs. 6, 7, 8, 9, 10 and 11) showing the temporal evolution of state simulations due to changes in model parameters. The results are also summarized in Table 3, which indicate that the impact of X1 (maximum capacity of production store) and X3 (maximum capacity of non-linear routing store) on the model derived dTWS changes is bigger than other parameters since they directly change the state variables (compare Figs. 6 and 8 of ESM with the rest). We also show that X1 (Fig. 6 of ESM) affects all components of the dTWS, while X3 (Fig. 8 of ESM) only affects the routing store component. As expected X4 (the time base of the unit hydrograph) has no effect on the production store (S, see Fig. 9 of ESM). Runoff is influenced by all four main parameters (X1 to X4), from which the influence of X4 on the magnitude of runoff is much smaller than the others (compare Fig. 9 of the ESM with the rest). Values in Table 3 also indicate that the new added parameters X5 and X6 affect the simulation of dTWS, but their impacts are less than that of X1 and X3. The additional parameters X5 and X6 are found to be effective only in the first 1–2 years of model simulations (Figs. 10 and 11 of the ESM). Particularly, it can be seen that perturbing X5 changes all model states and it has the biggest impact on the simulated dTWS and S. The greatest influence of X6 is observed on S (and consequently dTWS) and runoff. Nevertheless, one can conclude that the warm up period has an impact on the GR4J's simulation of water storages.

In the following, we calibrate the GR4J model using 6 parameters (i.e., θ is $[X1, X2, X3, X4, X5, X6]^T$ in Eq. 13) separately one time using only runoff data and the other time by applying GRACE dTWS data. Details of parameters selected for running the genetic algorithms are summarized in Table 4. A single-objective calibration is implemented to assess the impact of separate calibration of GR4J against GRACE dTWS and runoff, whose results are shown in Table 5. From numerical values, one can conclude that considering the warm-up period (X5 and X6) might have an impact on the estimation of model parameters but it does not significantly change the NS coefficient using either of runoff or dTWS. It is worth mentioning that we also run the model using 3 years (2000–2003) for warm up,

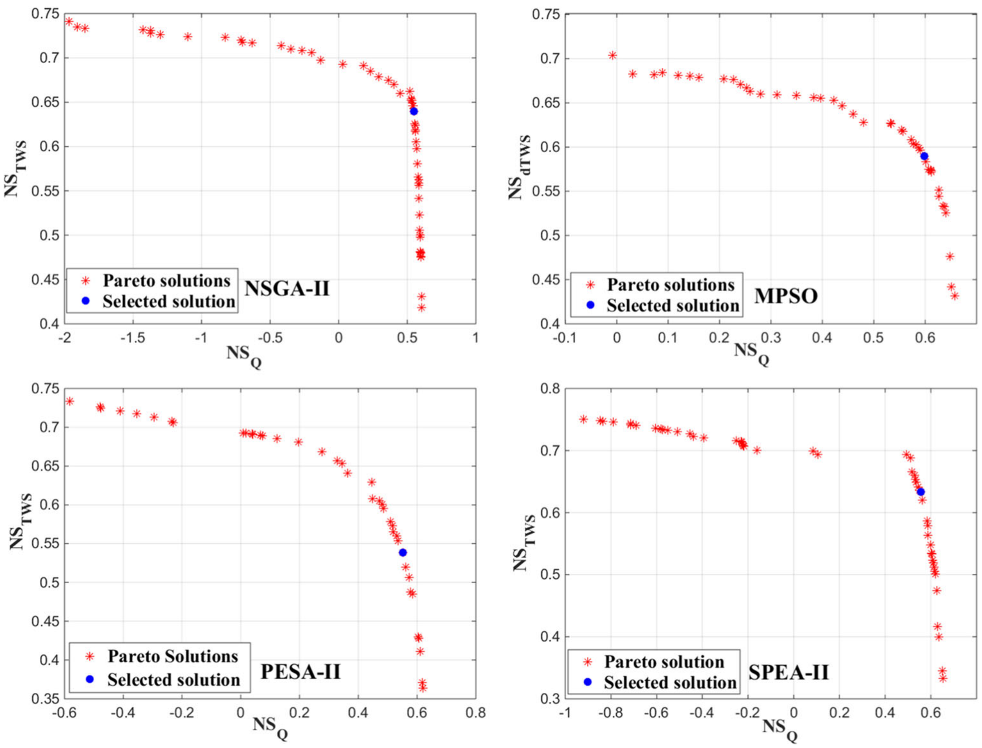
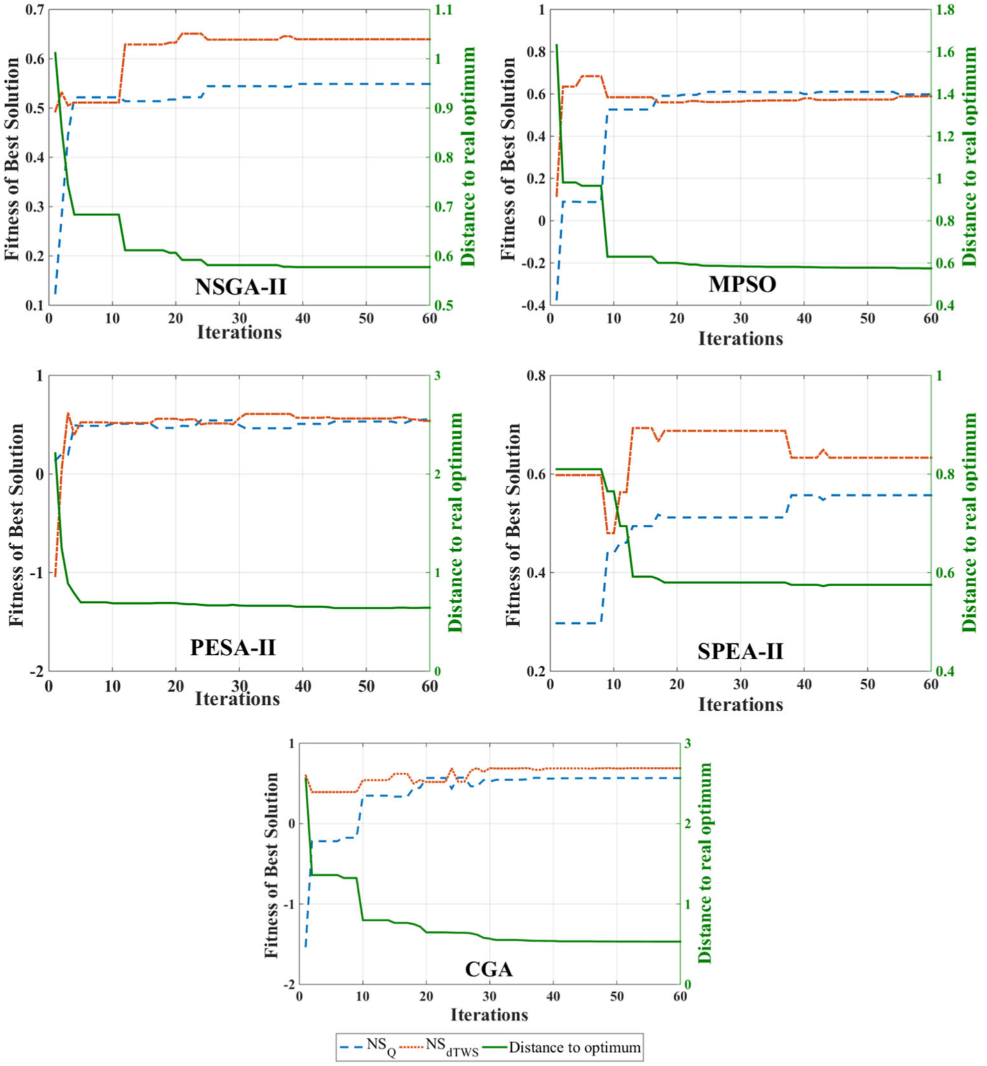


Fig. 7 Last iteration Pareto archive solutions (red strikes) and its selected solution (blue point) of the best run for each algorithm, NSGA-II (top-left), MPSO (top-right), PESA-II (bottom-left), and SPEA-II (bottom-right)

Fig. 8 An overview of variations in the objective function corresponding to the optimum solution (NS from runoff is shown by the blue dashed line and the red dot-line corresponds to that of dTWS), and distance changes of the selected solution objective to the optimal objective through various iterations. Results are orders as NSGA-II (top-left), MPSO (top-right), PESA-II (middle-left), SPEA-II (middle-right), and CGA (bottom)



and calibrate only the 4 original parameters of X1-X4 (i.e., θ is $[X1, X2, X3, X4]^T$ in Eq. 13). Our results indicate that although the estimated parameters during these two process are not the same, the resulted NS coefficients after calibration with runoff and dTWS anomalies are very close, see Table 5. Therefore, we conclude that in application with limited data for warming up the GR4J, the two new parameters (X5 and X6) can be introduced as initial values of storages, and their optimum values can be estimated from the calibration step without harming the accuracy of model simulations.

From the above experiments, we conclude that estimated parameters derived by calibrating against dTWS do not necessarily provide an acceptable runoff simulations and vise versa. For example, we derive a NS coefficient of 0.72 using dTWS data while calibrating 4 parameters (X1-X4). Using these parameters to run GR4J, however, yields runoff simulations with the NS coefficient of -3.16. In another try, we calibrate all 6 parameters against runoff, which yields the NS of 0.61. Running the model using these parameters, however, results in dTWS simulations with the NS of 0.32. These results motivate an application of the multi-objective calibrations methods to estimate model parameters using both dTWS and runoff observations.

5.2 Multi-objective calibration of GR4J and uncertainty of its parameters

In this study, the parameters of algorithms are tuned before the optimization process and the most appropriate parameter values of the multi-objective optimization algorithms for calibrating GR4J are determined by running a series of trial-and-error experiments. All parameters are set to a priori values that are summarized in Table 6. In fact, within different runs, the evolutionary algorithms estimate a variety of solutions and it is very rare that different runs estimate the same solutions due to the essence of evolutionary algorithms. Therefore, to obtain an estimation

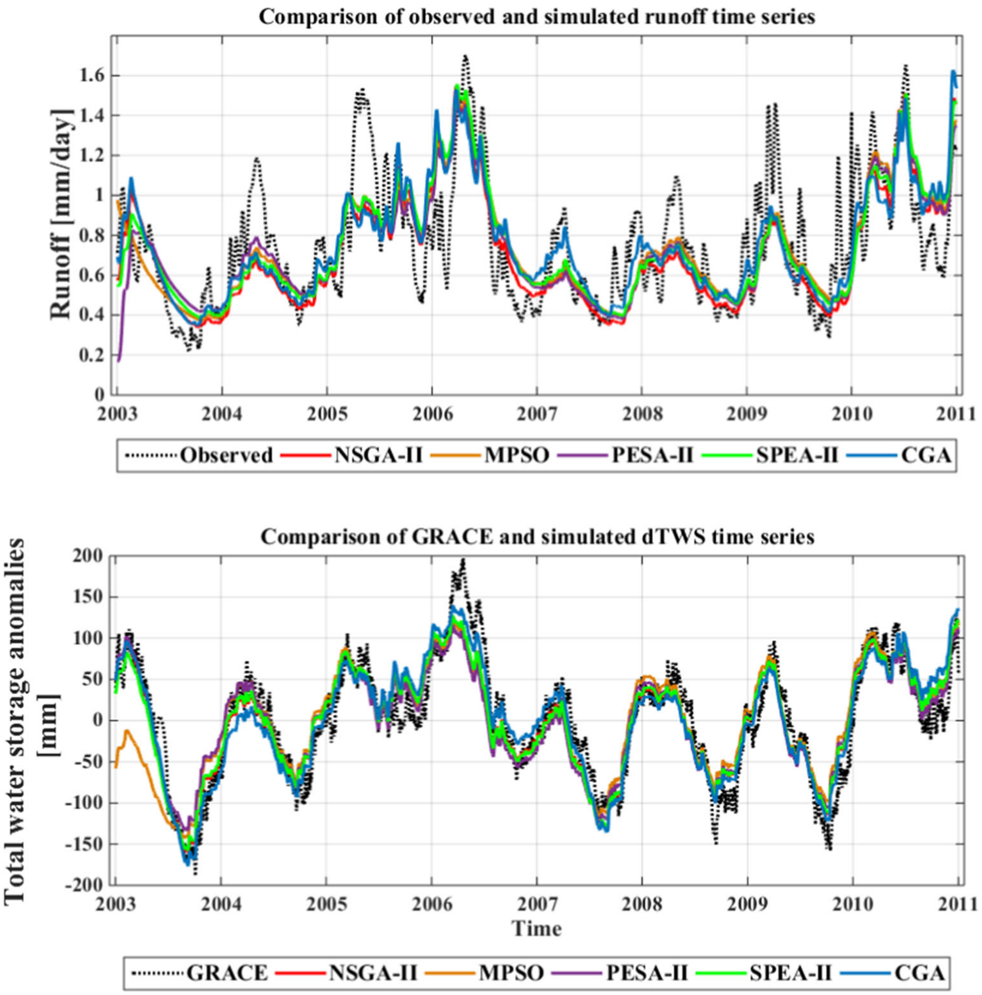


Fig. 9 Top: Simulated runoff derived from GR4J using optimized parameters from different optimization algorithms and their comparisons with in situ runoff (black dotted line). Down: GR4J simulated dTWS anomalies and their comparisons with basin averaged GRACE dTWS (black dotted line). Results correspond to the Danube River Basin during the calibration and validation period (2003-2010)

of parameter uncertainty and to assess the ability of each method to discover the same solutions during various runs, each multi-objective calibration algorithm is run 50 times, which 50 was set based on the trade-off between computational time and simulation accuracy. (details of the optimization setup is summarized in Table 6). [56] showed that model performance can be evaluated as “satisfactory” for a monthly time step if $NS > 0.50$. Since the GR4J simulates daily hydrological outputs, by converting its daily simulation to the monthly time step during some comparison experiments, it is find that NS=0.5 in monthly time step is almost equal to 0.4 in daily one. Therefore, in this study, the model runs with NS more than 0.4 are regarded as satisfactory runs (SR). These runs are then used to calculate uncertainties of the model parameters.

The results during the calibration period 2003–2007 are summarized in Tables 7 and 8. The optimum parameters are then used to run the model during the validation period of 2008–2010. The results indicate that some of the model runs in the validation period are inefficient and their NS coefficient are unacceptable; thus, the statistics are calculated after excluding those runs with $NS < 0.4$. Tables 9 and 10 provide the results of the validation period with σ and $\tilde{\sigma}$ to be standard deviation and normalized standard of the GR4J parameters, respectively.

According to the results in Table 7, one can see that SPEA-II and PESA-II provide the most successful measures for calibrating GR4J. MPSO has the minimum number of satisfactory runs, and SPEA-II result to the maximum number of successful calibration runs, i.e., just 31 runs of the MPSO calibration accepted through the 50 different runs while this number increased to 49 for SPEA-II. For a better comparison of parameter uncertainty, estimated standard deviation for each parameter is divided to its variation range and is called “normal standard deviation” of that parameters (shows by $\tilde{\sigma}$). Estimated normal standard deviation of X4 by all methods has the highest value. In contrast, the smallest standard deviation is found for X3. In summary, Table 7 indicates that the estimated parameter uncertainties of MPSO are greater from those derived from other optimization techniques. Although, the number of

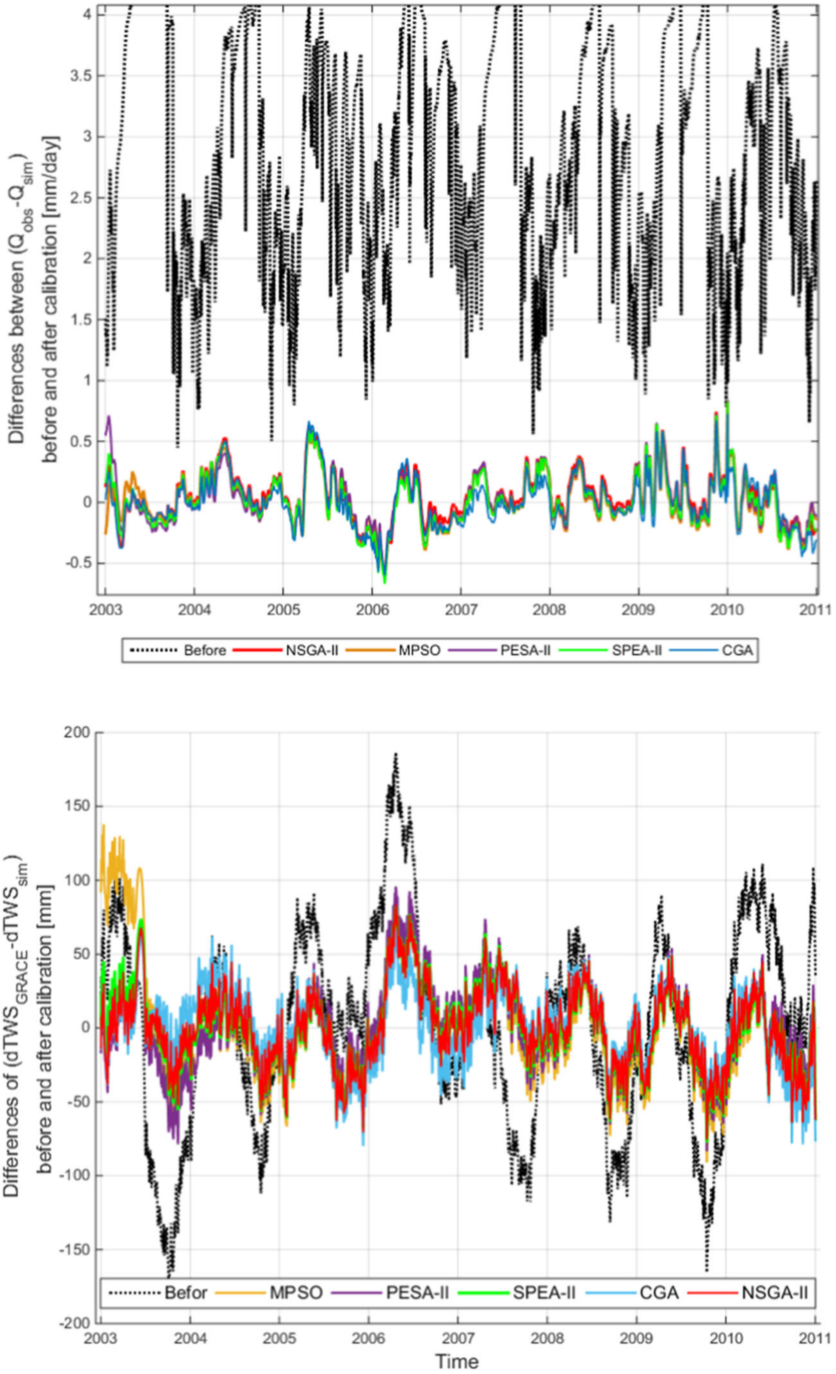


Fig. 10 Top: comparison between GR4J simulated runoff and in situ runoff, before and after calibration by 5 different algorithms. Down: comparison between simulated dTWS anomalies and GRACE derived basin averaged dTWS before and after calibration within the Danube River Basin during calibration and validation periods (2003–2010). The Y axes represent (differences between) runoff and dTWS while the X axes indicate time

successful runs of NSGA-II is found to be less than those of PESA-II and SPEA-II, its other statistics are at the first place with respect to other Pareto based multi-objective algorithms but CGA method depict the better results in all aspects. Table 8 summarizes some statistics that correspond to the values of objective functions (NS_Q and NS_{dTWS}) of final solution by all satisfactory runs. Considering these values, one can conclude that all methods find almost the same NS value for both Q and dTWS and the final solutions are at the same distance to the optimum solution.

Tables 9 and 10 provide similar information as Tables 7 and 8, respectively, but for the validation period (2008-2010). The results indicate that NSGA-II provides the least number of satisfactory runs while PESA-II has the greatest number. Comparing the statistics in Table 10 with Table 8, one can see an increase in the minimum distance of the final solutions to the optimum solutions, which can be expected as the time series of runoff and dTWS are non-stationary and the calibration during a short period might not capture the behavior of their variations adequately enough.

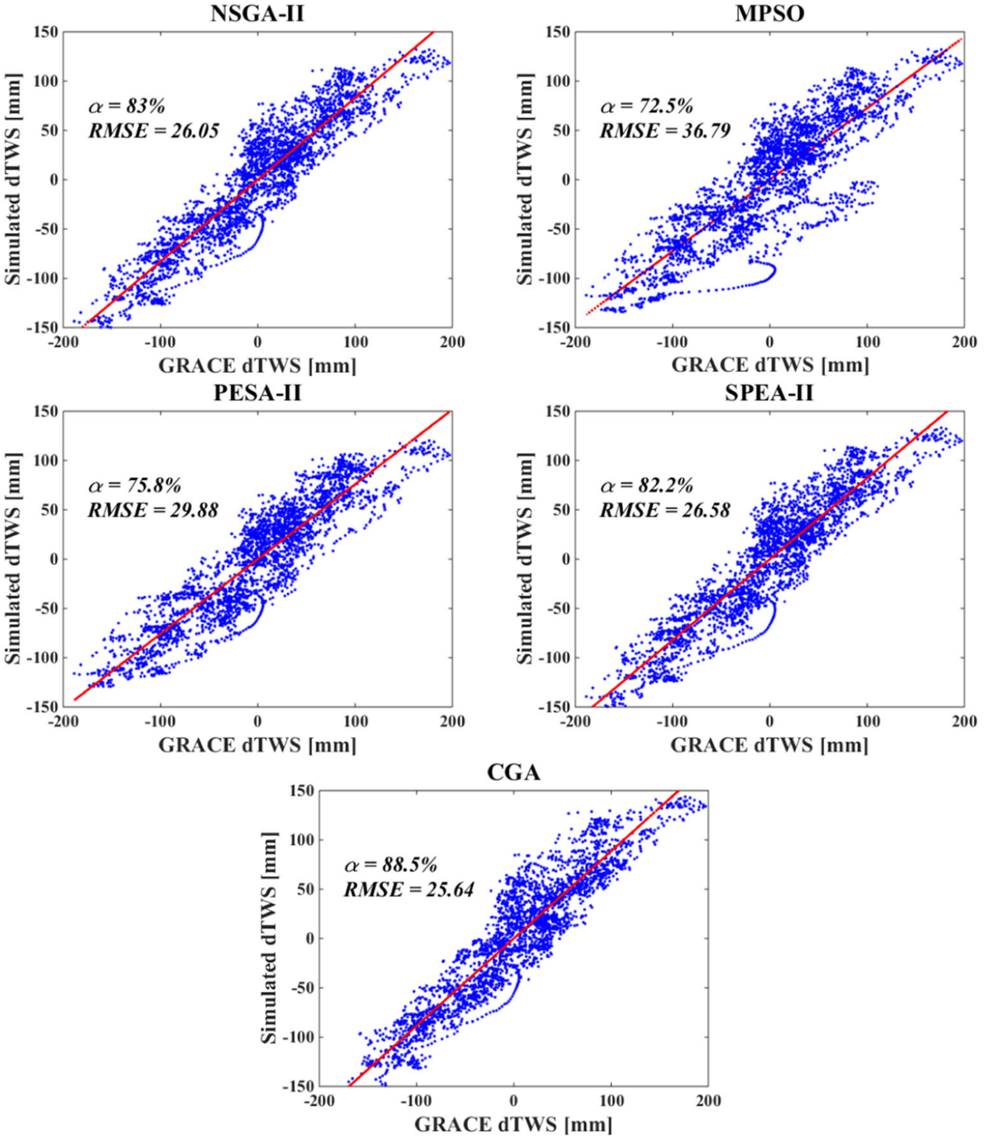


Fig. 11 Bi-plots of simulated dTWS against GRACE derived dTWS using the final calibration values from the five optimization algorithm. Results are ordered as NSGA-II (top-left), MPSO (top-right), PESA-II (middle-left), SPEA-II (middle-right), and CGA (bottom)

Table 3 Minimum and maximum standard deviation of the GR4J's outputs due to propagation of model parameters using an ensemble of size 50

States	Perturbed parameters											
	X1 (mm)		X2 (mm)		X3 (mm)		X4 (days)		X5 (%)		X6 (%)	
	min	max	min	max	min	max	min	max	min	max	min	max
S	3.00	23.83	0	0	0	0	0	0	0	11.92	0	0
R	0	7.45	0.01	1.22	22.29	38.47	0	1.80	0	2.25	0	37.18
V	0	2.40	0	0	0	0	0	2.68	0	1.04	0	0
W	0	0.31	0	0	0	0	0	0.33	0	0.14	0	0
TWS	3.00	22.94	0.01	1.22	22.29	38.47	0	1.58	0	11.97	0	37.18
Q	0	0.16	0.01	0.04	0.01	0.22	0	0.06	0	0.05	0	1.21

At each time, only one parameter is perturbed and other parameters are set to the nominal values. The values of 226.23, 12.77, 587.66, 17.37, 0.32, and 0.34 are considered as nominal values of X1 to X6, respectively

Table 4 Setup of the genetic algorithms used in this study

Population size	50
Generation size	60
Crossover rate	0.7
Selection method	Roulette Wheel
Mutation rate	0.1

5.3 Best model runs and final calibration solutions

Figure 7 displays the objective of Pareto solutions for selected runs derived from the optimization methods and the nearest solution to the reference solution (i.e., [$NS_Q = 1$, $NS_{dTWS} = 1$]) is highlighted in blue. The results indicate a trade-off between NS_Q nad NS_{dTWS} in the objective function space, which means an improvement in one objective requires a degradation in another one. This figures also shows that for all algorithms at the last iteration Pareto archive solutions contain no inferior point, i.e., a point in which improvement can be attained in all the objectives. We find that calibration using NSGA-II results in a much wider trade-off between both objectives. The performance of the Pareto solutions varies between $\sim$0.2 and $\sim$0.6 for runoff and between $\sim$0.42 and $\sim$0.75 for dTWS but the most variations are around 0.5 to 0.6 and around 0.42 to 0.67 respectively for Q and dTWS. Also generally all algorithms depict the same trade-off in dTWS objective function and the main differences correspond to calibrations using runoff data.

Figure 8 shows the behavior of objective functions that correspond to the selected solution (or the best run). The results illustrate the final fit of the model run, i.e., the Euclidean distance to the reference solution, after each iteration. In other words, Fig. 8 illustrates the procedure of the convergence of the four multi-objective algorithms and the CGA used in this study. For all methods after an increasing in the trend of both objective and some fluctuations in the initial iterations, the objective functions are found to oscillate around the average value of two objective functions so that by increasing one the other one decreases. The mean value of the two objective functions is found to be 0.6 except for PESA-II, which suggests the average value of 0.55 and these results accommodate with the final value of objective functions at the last iteration (see Table 11).

In the following, we show the NS coefficient and its evolutions after each iteration correspond to NSGA-II, MPSO, PESA-II and SPEA-II, as well as CGA algorithms. Note that the NS values vary between $-\infty$ and 1 (with 1 being the optimum value), and one might expect an increasing trend in the evolution of NS plots against iteration numbers that shows the convergence of these algorithms to the optimal value. We find that MPSO, PESA-II, and CGA find the optimum solution faster than the other techniques. The distances between selected optimum solutions and the real optimal value are also depicted in Fig. 8, for which we observe a descending trend derived from all methods. The smallest distance is derived for NSGA-II and MPSO.

The values of both objective functions, which represent the fitness of each calibration variable on the calibration and validation periods at the last iteration of best run are summarized in Table 11. Equation 16 is used to select the best run and its final solution from its Pareto archive. [56] showed that model performance can be evaluated as "satisfactory" for a monthly time step if $NS > 0.50$. Calculated performance criterion for all methods is in the range of satisfactory for both calibration and validation periods. The final estimated parameters of each method is also presented in Table 11. It can be seen that

Table 5 Results of calibration (2003-2007) and validation (2008-2010) of the GR4J model using its original 4 parameters with the warm up of 2000–2003, and an independent run with six parameters without warm-up period

Method	Calibration	Estimated parameters						NS	NS of	NS of
	variable	X1 (mm)	X2 (mm)	X3 (mm)	X4 (days)	X5 (%)	X6 (%)	Variable	calibration	Validation
GA with warm up	Runoff	812.33	7.12	206.85	13.86			Runoff	0.56	0.55
								dTWS	0.68	0.37
	dTWS	425.87	-15	1048.95	30			Runoff	-3.16	-4.60
								dTWS	0.72	0.37
GA without warm up	Runoff	294.64	9.67	353.55	11.85	0.46	0.34	Runoff	0.61	0.54
								dTWS	0.32	0.39
	dTWS	456.93	-7.05	2903.83	23.79	0.32	0.23	Runoff	-2.23	-3.79
								dTWS	0.76	0.17

In each case, once runoff and the other time TWS anomalies (dTWS) are separately used for calibration using a single-objective genetic algorithm optimization technique as in Table 4

Table 6 Setup of the multi-objective algorithms

Algorithm	Parameter	value	Algorithm	Parameter	value
NSGA-II	Population size	50	PESA-II	Initial population size	50
	Crossover rate	0.9		Crossover rate	0.8
	Mutation rate	0.1		Mutation rate	0.2
	Selection method	Tournament Selection		External archive size	50
	Number of iterations	60		Number of iterations	60
MPSO	Population size	50	SPEA-II	Initial population size	50
	Cognitive constant	1		Crossover rate	0.7
	Social constant	2		Mutation rate	0.3
	External archive size	50		External archive size	50
	Number of iterations	60		Number of iterations	60

Table 7 Statistics derived from the 50 different runs of the multi-objective calibration algorithms

	NSGA-II	MPSO	PESA-II	SPEA-II	CGA
SR	33	31	47	49	49
$\sigma_{X1}(mm)/\tilde{\sigma}_{X1}$	367.568 / 0.09	352.980 / 0.09	278.257 / 0.07	312.257 / 0.08	300.10 / 0.07
$\sigma_{X2}(mm)/\tilde{\sigma}_{X2}$	2.283 / 0.08	3.620 / 0.12	2.823 / 0.09	2.933 / 0.10	2.12 / 0.07
$\sigma_{X3}(mm)/\tilde{\sigma}_{X3}$	116.479 / 0.03	216.306 / 0.06	6.359 / 0.05	3.881 / 0.05	2.21 / 0.03
$\sigma_{X4}(day)/\tilde{\sigma}_{X4}$	3.359 / 0.11	10.636 / 0.36	6.359 / 0.21	3.881 / 0.13	2.21 / 0.07
$\sigma_{X5}(\%)/\tilde{\sigma}_{X5}$	3.7 / 0.04	32.3 / 0.32	14.5 / 0.15	6.6 / 0.07	3.3 / 0.03
$\sigma_{X6}(\%)/\tilde{\sigma}_{X6}$	4.3 / 0.04	12.3 / 0.12	4.5 / 0.05	3.8 / 0.04	3.0 / 0.03

The runs with $NS < 0.4$ are excluded from our uncertainty estimation. The calibrated values of model parameters using each algorithm are reported in Table 11

Table 8 Calculated minimum, maximum and mean of NS_Q, NS_{dTWS} and distance to optimum NS using satisfactory runs (SR) of calibration period

Variable	Statistic	NSGA-II	MPSO	PESA-II	SPEA-II	CGA
NS_Q	Min	0.40	0.42	0.44	0.46	0.49
	Mean	0.51	0.53	0.53	0.54	0.56
	Max	0.57	0.63	0.59	0.59	0.60
NS_{dTWS}	Min	0.45	0.46	0.53	0.60	0.58
	Mean	0.66	0.59	0.63	0.65	0.64
	Max	0.70	0.71	0.70	0.70	0.71
d	Max	0.74	0.75	0.67	0.62	0.60
	Mean	0.60	0.63	0.60	0.58	0.57
	Min	0.53	0.55	0.55	0.54	0.53

Table 9 Statistics of the satisfactory runs during the validation period

	NSGA-II	MPSO	PESA-II	SPEA-II	CGA
SR	13	22	40	36	34
$\sigma_{X1}(mm)/\tilde{\sigma}_{X1}$	108.93 / 0.03	65.872 / 0.02	91.608 / 0.02	85. 60 / 0.02	105.91 / 0.03
$\sigma_{X2}(mm)/\tilde{\sigma}_{X2}$	1.47 / 0.05	1.965 / 0.07	1. 78 / 0.06	1.69 / 0.06	1.2 / 0.04
$\sigma_{X3}(mm)/\tilde{\sigma}_{X3}$	100.99 / 0.03	132.756 / 0.04	127.049 / 0.04	105.57 / 0.03	103.00 / 0.03
$\sigma_{X4}(day)/\tilde{\sigma}_{X4}$	2.98 / 0.10	10.630 / 0.36	6.101 / 0.21	3.73/ 0.13	2.21 / 0.07
$\sigma_{X5}(\%)/\tilde{\sigma}_{X5}$	3.0 / 0.03	38.4 / 0.38	15.6 / 0.16	7.0 / 0.07	6.0 / 0.06
$\sigma_{X6}(\%)/\tilde{\sigma}_{X6}$	3.0 / 0.03	11.2 / 0.11	3.5 / 0.03	2.0 / 0.02	2.0 / 0.02

The runs with $NS < 0.4$ are excluded from the calculation of uncertainty values

Table 10 Calculated minimum, maximum and mean of NS_Q, NS_{dTWS} and distance to optimum NS using satisfactory runs (SR) of validation period

Variable	Statistic	NSGA-II	MPSO	PESA-II	SPEA-II	CGA
NS_Q	Min	0.45	0.43	0.41	0.41	0.46
	Mean	0.51	0.50	0.52	0.50	0.52
	Max	0.55	0.56	0.56	0.54	0.55
NS_{dTWS}	Min	0.41	0.44	0.43	0.41	0.41
	Mean	0.49	0.53	0.52	0.52	0.53
	Max	0.54	0.56	0.56	0.55	0.55
d	Max	0.78	0.74	0.77	0.76	0.76
	Mean	0.71	0.69	0.70	0.70	0.66
	Min	0.66	0.64	0.65	0.65	0.64

different parameter sets are estimated after applying various optimization methods, which indicate that this hydrological model calibration problem has no unique solution.

In the following, we assess the four multi-objective techniques NSGA-II, MPSO, PESA-II and SPEA-II, as well as CGA to calibrate GR4J model, and the results of calibration are compared with the observed GRACE TWS anomalies and the in situ runoff data (see Fig. 9). The results indicate that the simulations fairly well catch the peaks of both dTWS and runoff time series, for example, the differences (observation minus simulation) in high peaks of dTWS are reduced in all methods, i.e., the value of 121.7 mm in 2006 reduced to 83.48, 88.33, 95.9, 82.71, and 66.84 after calibration the model using NSGA-II, MPSO, PESA-II, SPEA-II and CGA, respectively. In Table 11, the numerical statistics are presented, which indicate that the five optimization techniques are only slightly different.

Differences between GRACE dTWS and in situ runoff and their corresponding simulated values before and after calibration are depicted in Fig. 10. The runoff results (Fig. 10 (top)) indicate that the magnitude of differences is reduced by 25% (75% improvement). For dTWS, the improvement reaches to 50%, see Fig. 10 (bottom). We still find seasonality in the residual time series, which indicates that calibration only is not able to account for basins complex hydrological processes. Nevertheless, a comparison of Figs. 9 and 10 indicates that calibration considerably mitigates errors in amplitude and timing of model simulations.

In order to illustrate the detailed differences between model simulations and observations, in Fig. 11 we show the bi-plots of simulated GR4J dTWS against GRACE dTWS and in Fig. 12 simulated runoff values against in situ observations. These figures justify which optimization technique results in less errors. The results indicate that applying CGA and MPSO respectively provides the best and worst dTWS results with the overall fitness, i.e., the normalized root mean square error, of 88.5 and 72.5% and the RMSE values of 25.64 and 36.79 (mm/day), see Fig. 11. Calibration results for the runoff simulations from the five assessed optimization techniques are found to be very similar, i.e., the overall fitness of 61.4% (PESA-II) to 66.6% (NSGA-II), and the RMSEs are found to be ∼ 0.2 mm/day see Fig. 12. Generally, the match of runoff simulations with observations are found to be less than those of dTWS.

Finally, as an example, the GR4J's runs after implementing the NSGA-II optimization are shown in Fig. 13. In each

Table 11 Final NS coefficient during the calibration (2003–2007) and validation (2008–2010) periods

Method	$NS_{Calibration}$		$NS_{Validation}$		Estimated Parameters					
	Q	dTWS	Q	dTWS	X1(mm)	X2(mm)	X3(mm)	X4(days)	X5(%)	X6(%)
NSGA-II	0.55	0.64	0.52	0.51	548.72	8.49	299.81	12.90	49	35
MPSO	0.60	0.59	0.54	0.53	312.90	14.95	690.80	17.51	17	33
PESA-II	0.55	0.53	0.53	0.52	317.00	14.29	669.66	16.67	77	22
SPEA-II	0.56	0.63	0.51	0.53	481.44	11.29	423.30	16.40	49	32
CGA	0.57	0.69	0.55	0.55	950.67	6.18	169.55	13.94	44	40

Results correspond to the five methods assessed here

Fig. 12 Bi-plots of simulated runoff against observed runoff using the final calibration values from each optimization algorithm. Results are orders as NSGA-II (top-left), MPSO (top-right), PESA-II (middle-left), SPEA-II (middle-right), and CGA (bottom)

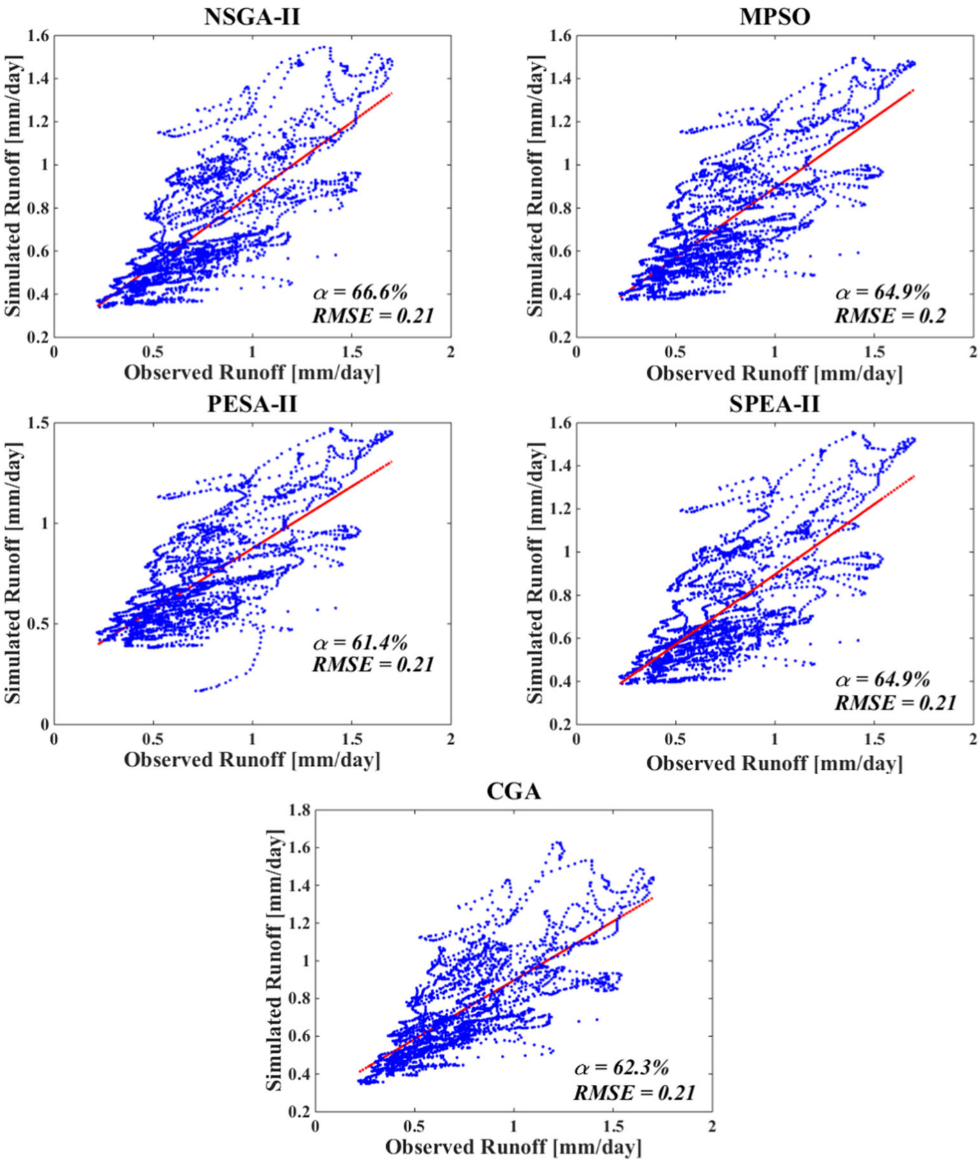


Fig. 13 Modeled and observed runoff (top), modeled and GRACE dTWS (bottom). In both plots, model parameters are estimated using the NSGA-II method (blue: observations, black: best simulation run, green: all accepted runs)

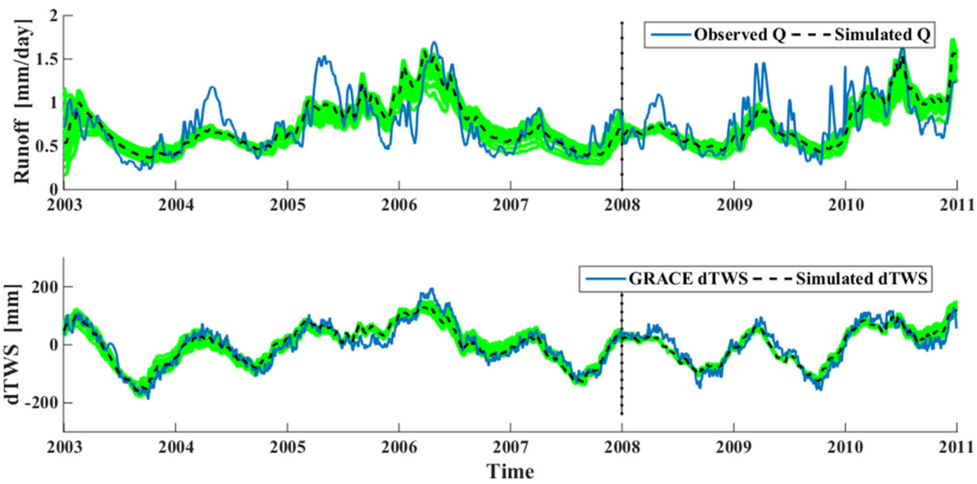


Table 12 Calculated metrics for different multi-objective calibration methods

Method	NPS	GD	SP	MS
NSGA-II	50	0.1731	1.7E-16	2.5946
MPSO	50	0.1155	7.1E-16	2.0399
PESA-II	50	0.1262	3.8E-16	0.9311
SPEA-II	50	0.1608	5.2E-16	0.8503

plot, we show those solutions that correspond to the NS of greater than 0.4 for both dTWS and Q to estimate the uncertainties. The best simulated run is plotted in red lines and the blue lines correspond to the observations. From the results, one can conclude that the runoff simulations exhibit bigger fluctuations (compared to the observations) than those of dTWS, indicating the fact that the GR4J's runoff simulations are more sensitive to the model parameters. However, the implemented calibration techniques show satisfactory skills to simulate both dTWS and runoff values with acceptable accuracy.

5.4 Performance metrics and comparisons

In the following, statistics derived from implementing each introduced optimization method is discussed with the summary presented in Table 12. The number of discovered non-dominated solutions of all methods is found to be 50, which indicate all algorithms provide similar cardinality. Considering the values of generational distances (GDs), we find that MPSO provides the lowest distance between the Pareto-optimal solutions and estimated Pareto solutions and that of NSGA-II yields the biggest. Therefore, the MPSO's solutions are found to be the nearest ones to the reference solutions and provide the best accuracy. NSGA-II is found to provide better statistics, while considering SP (spacing metric) and MS (maximum spread), which indicate although this algorithm provides slightly worse accuracy, compared to the other techniques, it results in better diversity. The values in Tables 7 and 9 are used to rank the quality of the optimization techniques that are shown in Table 13, from which the CGA, NSGA-II, PESA-II, SPEA-II, and MPSO methods can be ranked from the first to the fifth place. Comparing the ranks in the calibration and validation steps, the first conclusion of this table is that the best estimated solutions during calibration may not necessarily provide the best results in the validation period. This means that it is impossible to find a unique parameter set that gives the best simulation and evaluation results, but instead several parameter sets give similarly good model results when evaluated by the observations in terms of different performance criteria [33, also see the discussions in].

As it often happens in calibration experiences, the performance of all optimization methods is worsen in the validation period. for example, according to Tables 7 and 9, percentage of the satisfactory runs in the calibration are found to be 66, 62, 94, 98, and 98%, which respectively decreases to 39, 71, 85, 73, and 69%. Results are reported for the NSGA-II, MPSO, PESA-II, SPEA-II, and CGA methods, respectively. The results of all optimizations techniques are however statistically significant and acceptable.

6 Summary and conclusions

Hydrological model calibration is an essential computational process to ensure the consistency of model simulations with real world observations. However, limited ability of observations in representing the complexity of the water cycle and also restricted ability of calibration (optimization) techniques introduce limitations in constructing models with reliable simulation/forecast skills. To address these problems, we assess the application of daily Total Water Storage (dTWS) changes derived from the Gravity Recovery And

Table 13 Ranking of methods based on the standard deviation of main 4 parameters of GR4J and satisfied runs of model

Parameter	Period	NSGA-II	MPSO	PESA-II	SPEA-II	CGA
Satisfactory runs	Calibration	3	4	2	1	1
	Validation	5	4	1	2	3
X1	Calibration	5	4	1	3	2
	Validation	5	1	3	2	4
X2	Calibration	2	5	3	4	1
	Validation	2	5	4	3	1
X3	Calibration	2	5	4	3	1
	Validation	1	5	4	3	2
X4	Calibration	2	5	4	3	1
	Validation	2	5	4	3	1

Climate Experiment (GRACE) and daily in situ runoff data to calibrate the 4-parameter hydrological model GR4J. We first show that using a single-objective calibration method that only considers either dTWS or runoff data for calibrating GR4J does not satisfy an accurate simulation of the other variable. Table 5 represents the results of calibration of GR4J using a simple single-objective genetic algorithm. Therefore, five multi-objective techniques are applied to calibrate GR4J against both GRACE and runoff data.

Four evolutionary optimization techniques, including NSGA-II, MPSO, PESA-II, and SPEA-II, and the Combined objective function and Genetic Algorithm (CGA) are tested for calibration. The results indicate that all of the assessed optimization techniques provide satisfactory performance in both simulation (2003–2007) and validation (2008–2010) periods. We, however, use a number of quality metrics that are discussed in the previous sections (see the numerical results, e.g., in Table 12) to rank the four evolutionary optimization techniques that create Pareto-frontier. In summary, according to the diversity based metrics (i.e., maximum spread, MS, and spacing, SP), NSGA-II method selected as the best method. MPSO is ranked first according to the accuracy metric (i.e., generational distance, GD). Finally, the performance of all algorithms is found the same, while considering the cardinality measure (i.e., number of Pareto solutions, NPS).

In Table 13, the rankings of the five optimization algorithms are summarized that correspond to the estimation of the main 4 parameters of GR4J, which are summarized in Tables 7 and 9. In other words, here we consider the statistics (e.g., satisfactory runs and standard deviations) that are estimated for the GR4J's 4 parameters to rank the optimization techniques. The results indicate that CGA generally performs the best and MPSO is placed in the last rank for calibrating GR4J using GRACE dTWS and in situ runoff data.

Among the calibrated parameters, we observe that the standard deviation of X1 and X3 is the highest in both calibration and validation periods, which likely indicate the sensitivity of the model simulations to these parameters, see Tables 7 and 9. We also add two new parameters that account for the initial states, correspond to the production store and the routing store. These two parameters are calibrated to avoid the model warm-up period. Our results indicate that both parameters can be successfully calibrated and their effect on the model simulations can only be detected during the first two years of the model runs. According to Table 11, as expected, no unique set of optimum values is found for the GR4J's parameters indicating that there is no unique solution for this optimization problem.

After calibration, the GR4J model satisfactory describes the mean runoff behavior. But runoff extremes and low flow periods were not properly estimated. This is likely due to the limitation of the current implementation of GR4J in representing hydro-meteorological processes like snow accumulation and snow-melt. Therefore, we conclude that although adding GRACE dTWS can improve model simulations, to achieve an efficient simulation of runoff and to a less extent water storage, the structure of GR4J model must be improved. In the light of these results, we do not recommend this version of GR4J to study flood inundation or extremes in the Danube River Basin.

As discussed in the method section, the uncertainty analysis of this paper only shows a minimum level of uncertainty associated with the model structure. This uncertainty is shown in terms of parameters ranges of the Pareto front solutions, and/or a band of model simulations. Therefore, it is desirable to adopt more robust methodologies to account for different sources of uncertainty such as input data, parameters, and model structure. A possible solution is to extend the experiments by involving error probabilities in the calibration procedure.

Acknowledgments The authors are grateful to Professor Clint N. Dawson (Editor in Chief) and an anonymous reviewer for providing valuable comments, which we used to improve the quality of this manuscript. We also would like to thank Dr Laurent Longuevergne (Université de Rennes1) for helping us to setup the GR4J model over the Danube River Basin. We are grateful to the GRACE data products and river discharge data that are freely available from their providers.

References

1. Abido, M.A.: Multiobjective particle swarm optimization for environmental/economic dispatch problem. Electr. Power Syst. Res. **79**(7), 1105–1113 (2009). https://doi.org/10.1016/j.epsr.2009.02.005
2. Akwir, N.A., Chedjou, J.C., Kyamakya, K.: Neural-Network-Based Calibration of Macroscopic Traffic Flow Models. Recent Advances in Nonlinear Dynamics and Synchronization (Pp 151-173). Springer, Cham (2018). https://doi.org/10.1007/978-3-319-58996-1
3. Andréassian, V., Parent, E., Michel, C.: Using a parsimonious rainfall-runoff model to detect non-stationarities in the hydrological behavior of watersheds. J. Hydrol. **279**(1-4), 458–463 (2003)
4. Arnold, J.G., Fohrer, N.: SWAT2000: current capabilities and research opportunities in applied watershed modeling. Hydrol. Process. **19**(3), 563–572 (2005). https://doi.org/10.1002/hyp.5611
5. BASU: Dynamic economic emission dispatch using nondominated sorting genetic algorithm-II (2008). https://doi.org/10.1016/j.ijepes.2007.06.009

6. Bekele, E.G., Nicklow, J.W.: Multi-objective automatic calibration of SWAT using NSGA-II. J. Hydrol. **341**(3), 165–176 (2007). https://doi.org/10.1016/j.jhydrol.2007.05.014
7. Bennett, J.C., Robertson, D.E., Ward, P.G., Hapuarachchi, H.P., Wang, Q.J.: Calibrating hourly rainfall-runoff models with daily forcings for streamflow forecasting applications in meso-scale catchments. Environ. Model Softw. **76**, 20–36 (2016). https://doi.org/10.1016/j.envsoft.2015.11.006
8. Beven, K.J.: Rainfall-runoff modelling: the primer. Wiley, New York (2011). https://doi.org/10.1002/9781119951001
9. Beven, K., Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrol. **249**(1), 11–29 (2001). https://doi.org/10.1016/S0022-1694(01)00421-8
10. Boyle, D.P., Gupta, H.V., Sorooshian, S.: Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods. Water Resour. Res. **36**(12), 3663–3674 (2000). https://doi.org/10.1029/2000WR900207
11. Broderick, C., Matthews, T., Wilby, R.L., Bastola, S., Murphy, C.: Transferability of hydrological models and ensemble averaging methods between contrasting climatic periods. Water Resour. Res. **52**(10), 8343–8373 (2016). https://doi.org/10.1002/2016WR018850
12. Corne, D.W., Jerram, N.R., Knowles, J.D., Oates, M.J.: PESA-II: Region-based selection in evolutionary multiobjective optimization. In: Proceedings of the 3rd Annual Conference on Genetic and Evolutionary Computation, pp. 283–290. Morgan Kaufmann Publishers Inc. (2001)
13. Deb, K.: Multi-Objective Optimization using Evolutionary Algorithms, vol. 16. Wiley, New York (2001)
14. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.A.M.T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. **6**(2), 182–197 (2002). https://doi.org/10.1109/4235.996017
15. Deckers, D.L., Booij, M.J., Rientjes, T.H., Krol, M.S.: Catchment variability and parameter estimation in multi-objective regionalisation of a rainfall-runoff model. Water Resour. Manag. **24**(14), 3961–3985 (2010). https://doi.org/10.1007/s11269-010-9642-8
16. Demirel, M.C., Booij, M., Hoekstra, A.: The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. Hydrology and Earth System Sciences (2015). https://doi.org/10.5194/hess-19-275-2015
17. Döll, P., Kaspar, F., Lehner, B.: A global hydrological model for deriving water availability indicators: model tuning and validation. J. Hydrol. **270**(1-2), 105–134 (2003). https://doi.org/10.1016/S0022-1694(02)00283-4
18. Duan, Q.: Global optimization for watershed model calibration. Calibration of watershed models, 89-104, pp. 89–104. American Geophysical Union, Washington DC (2003). https://doi.org/10.1029/WS006
19. Dumedah, G., Berg, A.A., Wineberg, M., Collier, R.: Selecting model parameter sets from a trade-off surface generated from the non-dominated sorting genetic algorithm-II. Water Resour. Manag. **24**(15), 4469–4489 (2010). https://doi.org/10.1007/s11269-010-9668-y
20. Dumedah, G.: Formulation of the evolutionary-based data assimilation, and its implementation in hydrological forecasting. Water Resour. Manag. **26**(13), 3853–3870 (2012). https://doi.org/10.1007/s11269-012-0107-0
21. Eckhardt, K., Arnold, J.G.: Automatic calibration of a distributed catchment model. J. Hydrol. **251**(1), 103–109 (2001). https://doi.org/10.1016/S0022-1694(01)00429-2
22. Edijatno, D.E., Oliveria Nascimento, N.I.L.O., Yang, X., Makhlouf, Z., Michel, C.: GR3J: a daily watershed model with three free parameters. Hydrol. Sci. J. **44**(2), 263–277 (1999). https://doi.org/10.1080/02626669909492221
23. Eicker, A., Forootan, E., Springer, A., Longuevergne, L., Kusche, J.: Does GRACE see the terrestrial water cycle ‘intensifying’? J. Geophys. Res.-Atmos. **121**, 733–745 (2016). https://doi.org/10.1002/2015JD023808
24. Efstratiadis, A., Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a review. Hydrol. Sci. J–J Des Sciences Hydrologiques **55**(1), 58–78 (2010). https://doi.org/10.1080/02626660903526292
25. Food and Agriculture Organization of the United Nations: FAO. FAO Water Report 34 (2009)
26. Foglia, L., Hill, M.C., Mehl, S.W., Burlando, P.: Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. Water Resour. Res. **45**, 6 (2009). https://doi.org/10.1029/2008WR007255
27. Forootan, E., Rietbroek, R., Kusche, J., Sharifi, M.A., Awange, J., Schmidt, M., Omondi, P., Famiglietti, J.: Separation of large scale water storage patterns over Iran using GRACE, altimetry and hydrological data. Remote Sens. Environ. **140**, 580–595 (2014). https://doi.org/10.1016/j.rse.2013.09.025
28. Forootan, E., Safari, A., Mostafaie, A., Schumacher, M., Delavar, M., Awange, J.L.: Large-scale total water storage and water flux changes over the arid and semiarid parts of the Middle East from GRACE and reanalysis products. Surv. Geophys. **38**(3), 591–615 (2017). https://doi.org/10.1007/s10712-016-9403-1
29. Gan, T.Y., Biftu, G.F.: Automatic calibration of conceptual rainfall-runoff models: optimization algorithms, catchment conditions, and model structure. Water Resour. Res. **32**(12), 3513–3524 (1996). https://doi.org/10.1029/95WR02195
30. Gouweleeuw, B.T., Kvas, A., Grüber, C., Gain, A.K., Mayer-Gürr, T., Flechtner, F., Güntner, A.: Daily GRACE gravity field solutions track major flood events in the Ganges-Brahmaputra Delta, Hydrol. Earth Syst. Sci. Discuss., in review (2017). https://doi.org/10.5194/hess-2016-653
31. Gupta, H.V., Sorooshian, S., Yapo, P.O.: Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour. Res. **34**(4), 751–763 (1998). https://doi.org/10.1029/97WR03495
32. Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F.: Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. **377**(1), 80–91 (2009). https://doi.org/10.1016/j.jhydrol.2009.08.003
33. Güntner, A.: Improvement of global hydrological models using GRACE data. Surv. Geophys. **29**(4-5), 375–397 (2008). https://doi.org/10.1007/s10712-008-9038-y
34. Guo, J., Zhou, J., Zou, Q., Liu, Y., Song, L.: A novel multi-objective shuffled complex differential evolution algorithm with application to hydrological model parameter optimization. Water Resour. Manag. **27**(8), 2923–2946 (2013). https://doi.org/10.1007/s11269-013-0324-1
35. Hall, J.W., Tarantola, S., Bates, P.D., Horritt, M.S.: Distributed sensitivity analysis of flood inundation model calibration. J. Hydraul. Eng. **131**(2), 117–126 (2005). https://doi.org/10.1061/(ASCE)0733-9429(2005)131:2(117)
36. Hargreaves, G.H., Samani, Z.A.: Estimating potential evapotranspiration. J. Irrig. Drain Engr., ASCE **108**(IR3), 223–230 (1982)
37. Harlan, D., Wangsadipura, M., Munajat, C.M.: Rainfall-Runoff Modeling of citarum hulu river basin by using GR4j. In: Proceedings of the World Congress on Engineering 2010, pp. 1607–1611 (2010)
38. Hublart, P., Ruelland, D., Atauri, I.G.D.C., Ibacache, A.: Reliability of a conceptual hydrological model in a semi-arid Andean catchment facing water-use changes. Proc. IAHS **371**, 203–209 (2015). https://doi.org/10.5194/piahs-371-203-2015

39. Jebari, K., Madiafi, M.: Selection methods for genetic algorithms. Int. J. Emerg. Sci. **3**(4), 333–344 (2013)
40. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: IEEE International Conference on Neural Networks, IV, Piscataway, pp. 1942–1948 (1995)
41. Khu, S.T., Savic, D., Liu, Y.: Evolutionary-based multiobjective meta-model approach for rainfall-runoff model calibration. Geophys. Res. Abstr. **7**, 09858 (2005)
42. Klinger, B., Mayer-Gürr, T.: The role of accelerometer data calibration within GRACE gravity field recovery: Results from ITSG-Grace2016. Advances in Space Research (2016). https://doi.org/10.1016/j.asr.2016.08.007
43. Konak, A., Coit, D.W., Smith, A.E.: Multi-objective optimization using genetic algorithms: a tutorial. Reliab. Eng. Syst. Saf. **91**(9), 992–1007 (2006). https://doi.org/10.1016/j.ress.2005.11.018
44. Kovács, P.: Characterization of the runoff regime and its stability in the Danube catchment. In: Hydrological Processes of the Danube River Basin: Perspectives from the Danubian Countries, edited by Brilly, Mitja, pp. 143–173. Springer, Netherlands (2010). isbn=978-90-481-3423-6, https://doi.org/10.1007/978-90-481-3423-6_5
45. Kuczera, G.: Efficient subspace probabilistic parameter optimization for catchment models. Water Resour. Res. **33**(1), 177–185 (1997). https://doi.org/10.1029/96WR02671
46. Kumar, R., Samaniego, L., Attinger, S.: Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. Water Resour. Res. 49 (2013). https://doi.org/10.1029/2012WR012195
47. Kusche, J., Eicker, A., Forootan, E.: Analysis tools for GRACE and related data sets, theoretical basis. The International Geoscience Programme (IGCP). IGCP 565: Supporting water resource management with improved Earth observations, www.igcp565.org/workshops/Johannesburg_2011/kusche_LectureNotes_analysistools.pdf (2011)
48. Le Lay, M., Galle, S., Saulnier, G.M., Braud, I.: Exploring the relationship between hydroclimatic stationarity and rainfall-runoff model parameter stability: a case study in West Africa. Water Resour. Res. **43** (7), (2007). https://doi.org/10.1029/2006WR005257
49. Li, X., Weller, D.E., Jordan, T.E.: Watershed model calibration using multi-objective optimization and multi-site averaging. J. Hydrol. **380**(3), 277–288 (2010). https://doi.org/10.1016/j.jhydrol.2009.11.003
50. Lindström, G., Pers, C., Rosberg, J., Strömqvist, J., Arheimer, B.: Development and testing of the HYPE (Hydrological Predictions for the Environment) water quality model for different spatial scales. Hydrol. Res. **41**(3-4), 295–319 (2010). https://doi.org/10.2166/nh.2010.007
51. Lu, D., Ye, M., Meyer, P.D., Curtis, G.P., Shi, X., Niu, X.F., Yabusaki, S.B.: Effects of error covariance structure on estimation of model averaging weights and predictive performance. Water Resour. Res. **49**(9), 6029–6047 (2013). https://doi.org/10.1002/wrcr.20441
52. Madsen, H.: Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. J. Hydrol. **235**(3), 276–288 (2000). https://doi.org/10.1016/S0022-1694(00)00279-1
53. Madsen, H.: Parameter estimation in distributed hydrological catchment modeling using automatic calibration with multiple objectives. Advan. Water Res. **26**(2), 205–216 (2003). https://doi.org/10.1016/S0309-1708(02)00092-1
54. Matott, L.S., Babendreier, J.E., Purucker, S.T.: Evaluating uncertainty in integrated environmental models: a review of concepts and tools. Water Resour. Res. 45 6 (2009). https://doi.org/10.1029/2008WR007301
55. Moore, J., Chapman, R.: Application of particle swarm to multiobjective optimization. Department of Computer Science and Software Engineering, Auburn University (1999)
56. Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L.: Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE **50**(3), 885–900 (2007). https://doi.org/10.13031/2013.23153
57. Nash, J.E., Sutcliffe, J.V.: River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. **10**(3), 282–290 (1970). https://doi.org/10.1016/0022-1694(70)90255-6
58. Ning, S., Ishidaira, H., Wang, J.: Calibrating a hydrological model by step-wise method using GRACE TWS and discharge data. J. Jpn. Soc. Civ. Eng., Ser. B1 (Hydraul. Eng.) **71**(4), 85–90 (2015)
59. Okabe, T., Jin, Y., Sendhoff, B.: A critical survey of performance indices for multi-objective optimisation Evolutionary Computation. In: The 2003 Congress on CEC'03, vol. 2, pp. 878–885. IEEE (2003). https://doi.org/10.1109/CEC.2003.1299759
60. Omondi, P., et al.: Changes in temperature and precipitation extremes over the Greater Horn of Africa region from 1961 to 2010. Int. J. Climatol. **34**(4), 1262–1277 (2014). https://doi.org/10.1002/joc.3763
61. Perrin, C.: Towards an Improvement of a Lumped Rainfall-Runoff Model through a Comparative Approach (Doctoral Dissertation, Ph D Thesis, Université Joseph Fourier, Grenoble) (2000)
62. Perrin, C., Michel, C., Andréassian, V.: Improvement of a parsimonious model for streamflow simulation. J. Hydrol. **279**(1), 275–289 (2003). https://doi.org/10.1016/S0022-1694(03)00225-7
63. Perrin, C., Andréassian, V., Rojas Serna, C., Mathevet, T., Le Moine, N.: Discrete parameterization of hydrological models: evaluating the use of parameter sets libraries over 900 catchments. Water Resour. Res. **44** (8), (2008). https://doi.org/10.1029/2007WR006579
64. Rakovec, O., Kumar, R., Attinger, S., Samaniego, L.: Improving the realism of hydrologic model functioning through multivariate parameter estimation. Water Resour. Res. **52**, 7779–7792 (2016). https://doi.org/10.1002/2016WR019430
65. Reddy, M.J., Nagesh Kumar, D.: Multi-objective particle swarm optimization for generating optimal trade-offs in reservoir operation. Hydrol. Process. **21**(21), 2897–2909 (2007). https://doi.org/10.1002/hyp.6507
66. Riquelme, N., Von Lücken, C., Baran, B.: Performance metrics in multi-objective optimization. In: Computing Conference (CLEI), 2015 Latin American, pp. 1–11. IEEE (2015). https://doi.org/10.1109/CLEI.2015.7360024
67. Samaniego, L., Kumar, R., Attinger, S.: Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. Water Resour. Res. **46**, W05523 (2010). https://doi.org/10.1029/2008WR007327
68. Savic, D.: Single-objective vs. multiobjective optimisation for integrated decision support (2002)
69. Schiller, H., Miklós, D., Sass, J.: The Danube River and its basin physical characteristics, water regime and water balance. In: Hydrological Processes of the Danube River Basin, pp. 25–77. Springer, Netherlands (2010). https://doi.org/10.1007/978-90-481-3423-6_2
70. Schumacher, M., Eicker, A., Kusche, J., Müller Schmied, H., Döll, P.: Covariance analysis and sensitivity studies for GRACE assimilation into WGHM. In: Rizos, C., Willis, P. (eds.) IAG 150 Years. International Association of Geodesy Symposia, vol. 143, pp. 241–247. Springer, Cham (2015). https://doi.org/10.1007/1345_2015_119
71. Schumacher, M., Kusche, J., Döll, P.: A systematic impact assessment of GRACE error correlation on data assimilation in hydrological models. J. Geod. **90**, 537 (2016). https://doi.org/10.1007/s00190-016-0892-y

72. Schumacher, M., Forootan, E., van Dijk, A.I.J.M., Müller Schmied, H., Crosbie, R.S., Kusche, J., Döll, P.: Improving drought simulations within the Murray-Darling Basin by combined calibration/assimilation of GRACE data into the waterGAP global hydrology model. Remote Sens. Environ. **204**, 212–228 (2018). https://doi.org/10.1016/j.rse.2017.10.029
73. Scott, J.R.: Fault Tolerant Design Using Single and Multi-Criteria Genetic Algorithms Master's Thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology (1995)
74. Shafii, M., Smedt, F.D.: Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. Hydrol. Earth Syst. Sci. **13**(11), 2137–2149 (2009). https://doi.org/10.5194/hess-13-2137-2009
75. Sorooshian, S., Gupta, V.K. In: Singh, V.P. (ed.): Model Calibration, Chapter 2 in Computer Models of Watershed Hydrology, pp. 23–68. Water Resources Publications Highlands Ranch, Littleton (1995)
76. Tapley, B.D., Bettadpur, S., Watkins, M., Reigber, C.: The gravity recovery and climate experiment: mission overview and early results. Geophys. Res. Lett. **31**, L09607 (2004). https://doi.org/10.1029/2004GL019920
77. Ter Braak, C.F.T.: A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. Stat. Comput. **16**, 239–249 (2006). https://doi.org/10.1007/s11222-006-8769-1
78. Taye, M.T., Willems, P.: Identifying sources of temporal variability in hydrological extremes of the upper Blue Nile basin. J. Hydrol. **499**, 61–70 (2013). https://doi.org/10.1016/j.jhydrol.2013.06.053
79. Tiedeman, C.R., Green, C.T.: Effect of correlated observation error on parameters, predictions, and uncertainty. Water Resour. Res. **49**(10), 6339–6355 (2013). https://doi.org/10.1002/wrcr.20499
80. Van Werkhoven, K., Wagener, T., Reed, P., Tang, Y.: Sensitivity-guided reduction of parametric dimensionality for multi-objective calibration of watershed models. Adv. Water Resour. **32**(8), 1154–1169 (2009). https://doi.org/10.1016/j.advwatres.2009.03.002
81. Veldhuizen, D.A.V., Lamont, G.B.: Multiobjective Evolutionary Algorithm Research: a History and Analysis. Dep of Electrical and Computer Engineering, Air Force Institute of Technology, Tech. Rep. (1998)
82. Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S.: Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resour. Res. 39 8 (2003). https://doi.org/10.1029/2002WR001746
83. Wagener, T.: Evaluation of catchment models. Hydrol. Process. **17**(16), 3375–3378 (2003). https://doi.org/10.1002/hyp.5158
84. Werth, S., Güntner, A., Petrovic, S., Schmidt, R.: Integration of GRACE mass variations into a global hydrological model. Earth Planet. Sci. Lett. **277**(1), 166–173 (2009). https://doi.org/10.1016/j.epsl.2008.10.021
85. Williams, J.R. In: Singh, V.P. (ed.): The EPIC Model. Chapter 25 in Computer Models of Watershed Hydrology, pp. 909–1000. Water Resources Publications Highlands Ranch, Littleton (1995)
86. Xie, H., Longuevergne, L., Ringler, C., Scanlon, B.R.: Calibration and evaluation of a semi-distributed watershed model of Sub-Saharan Africa using GRACE data. Hydrol. Earth Syst. Sci. **16**(9), 3083–3099 (2012). https://doi.org/10.5194/hess-16-3083-2012
87. Yapo, P.O., Gupta, H.V., Sorooshian, S.: Multi-objective global optimization for hydrologic models. J. Hydrol. **204**(1), 83–97 (1998). https://doi.org/10.1016/S0022-1694(97)00107-8
88. Zitzler, E., Laumanns, M., Thiele, L.: SPEA 2: Improving the Strength Pareto Evolutionary Algorithm. Technical Report 103 Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH) Zurich (2001)