Water Resources Management (2019) 33:4509–4524
https://doi.org/10.1007/s11269-019-02368-8

Joint Optimization of Conceptual Rainfall-Runoff Model Parameters and Weights Attributed to Meteorological Stations

Adam P. Piotrowski[1] · Marzena Osuch[1] · Jarosław J. Napiorkowski[1]

Received: 14 May 2019 / Accepted: 26 September 2019 /
Published online: 6 November 2019

Abstract
Conceptual lumped rainfall-runoff models are frequently used for various environmental problems. To put them into practice, both the model calibration method and data series of the area-averaged precipitation and air temperature are needed. In the case when data from more than one measurement station are available, first the catchment-averaged meteorological data series are usually obtained by some method, and then they are used for calibration of a lumped rainfall-runoff model. However, various optimization methods could easily be applied to simultaneously calibrate both the aggregation weights attributed to various meteorological stations to obtain a lumped meteorological data series and the rainfall-runoff model parameters. This increases the problem dimensionality but allows the optimization procedure to choose the data that are most important for the rainfall-runoff process in a particular catchment, without a priori assumptions. We test the idea using two conceptual models, HBV and GR4J, and three mutually different, relatively recently proposed Evolutionary Computation and Swarm Intelligence optimization algorithms, that are applied to three catchments located in Poland and northwestern USA. We consider two cases: with and without the model error correction applied to the rainfall-runoff models. It is shown that for the calibration period, joint optimization of the weights used to aggregate the meteorological data and the parameters of the rainfall-runoff model improves the results. However, the results for the validation period are inconclusive and depend on the model, error correction, optimization algorithm, and catchment.

Keywords Lumped conceptual rainfall-runoff modelling · Mean aerial precipitation · HBV · GR4J · Thiessen polygons · Model calibration

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11269-019-02368-8) contains supplementary material, which is available to authorized users.

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

Extended author information available on the last page of the article

1 Introduction

Relationships between the meteorological and hydrological data are frequently captured by conceptual rainfall-runoff models. Many conceptual rainfall-runoff models are lumped, which means that they use only a single precipitation and air temperature data series to represent the whole catchment. Despite vulnerability to temporal resolution (Jie et al. 2018), lumped conceptual hydrological models may perform just as well as complex models (Bhadra et al. 2010; Lobligeois et al. 2014). However, if more than a single measurement station is available within, or close to the catchment of interest, the data series must somehow be averaged prior to application of the lumped model. This may be done by various kinds of methods (Singh and Chowdhury 1986; Vaze et al. 2011), among which the Thiessen Polygons (Thiessen and Alter 1911) are probably the most popular and widely used. The final version of the rainfall series may have an impact not only on the performance of the model, but also on proper representation of the catchment's critical rainfall (Yuan et al., 2019), and hence flood management procedures.

Even though parameters of conceptual rainfall-runoff models generally have some physical representation, to be applied for a specific catchment they almost always require calibration on measured data. Often some automatic procedures are used for this task (Goswami and O'Connor 2007; Vrugt et al. 2013; Arsenault et al. 2014; Piotrowski et al. 2017a; Peng et al. 2018). The performance of rainfall-runoff models depends to some extent on the calibration procedures used. When the proper global search heuristic optimization procedure is chosen, one may use it not only for calibration of model parameters, but also for choosing and managing the measurement data. Some similar ideas have already been applied to rainfall-runoff modelling. For example, Anctil et al. (2006) showed that the binary Genetic Algorithm may help in choosing which rain gauges should be used for catchment-averaging of rainfall data when artificial neural networks are applied for rainfall-runoff modelling. Girons Lopez and Seibert (2016) used a version of the Genetic Algorithm with a steepest gradient method to calibrate the lumped HBV model (Bergström 1976; Lindström et al. 1997) for nine streamflow gauge locations in the chosen catchment. Arsenault et al. (2015) showed that a global-search optimization algorithm may be used to calibrate weights for multimodel averaging for rainfall-runoff purposes. Osuch (2015) applied the SCEM-UA algorithm (Vrugt et al. 2003) for joint calibration of the HBV parameters and the weights attributed to four rainfall stations located within the catchment of interest.

Unfortunately, the approaches presented in the mentioned studies were generally applied ad-hoc, without in-depth comparison with alternative methods. In the present paper we perform detailed tests to verify whether calibration algorithms may efficiently be used to optimize together the parameters of the lumped conceptual rainfall-runoff model and the weights attributed to different meteorological stations that are used to aggregate meteorological data for the catchments of interest. It is very important for practitioners to learn whether the burden of additional computations – coupled calibration of model parameters and weights attributed to different stations – is indeed worth the effort.

In this paper we consider two versions of the joint calibration method. They differ by whether weights attributed to meteorological stations have to sum up to 1. If not, the calibration procedure is able to make corrections to the amount of precipitation measured, but may also over- or under-estimate the air temperature, and hence evaporation. The results are compared with those obtained when the classical Thiessen Polygons (Thiessen and Alter 1911) and simple averaging methods (Maidment 1993) are used to obtain a catchment-wide lumped precipitation and air temperature data series, prior to calibration of the rainfall-runoff models.

The impact of joint calibration of the model parameters and weights attributed to meteorological stations may depend on other methodological settings, or catchment-specific conditions. Hence, we focus our attention on catchments located in temperate climatic zones, while examining various methodological issues. More specifically, two conceptual lumped rainfall-runoff models, HBV (Bergström 1976; Lindström et al. 1997) and GR4J (Perrin et al. 2003), and three optimization algorithms (MDE_pBX (Islam et al. 2012), SPS-L-SHADE-EIG (Guo et al. 2015) and GLPSO (Gong et al. 2016)), are used for modelling at three catchments of comparable size (the Biala Tarnowska and the Suprasl catchment from Poland and the Cedar catchment from Washington state, USA), and results are discussed separately when the error correction procedure (Refsgaard 1997) is or is not used. To find out which methodological factors used in the study (model, calibration algorithm, error correction procedure, or method of attributing weights to meteorological stations) are of major importance for modelling performance, we apply ANOVA (von Storch and Zwiers 2001) with the post-hoc Tukey test (Tukey 1949).

2 Materials and Methods

Classically, lumped rainfall-runoff models require a catchment-aggregated time series of precipitation and air temperature. In this paper we aim to verify the possibility of joint calibration of rainfall-runoff model parameters and weights attributed to each meteorological station in order to obtain an area-averaged precipitation and air temperature time series.

2.1 Meteorological Data Averaging

In this study four methods, applied independently to obtain a lumped meteorological time series, are tested. They are denoted as w1 to w 4:

- w1: simple averaging of measurements observed at each considered (see Suppl. Table 1 and Suppl. Figs. 1–3, available as Supplementary Material in on-line version of the paper) meteorological station (Maidment 1993);
- w2: the Thiessen Polygons (Thiessen and Alter 1911), probably the most widely used method (Groisman and Legates 1994) that often gives results of similar quality to the isohyetal method and other classical approaches (Shaw and Lynn 1972; Singh and Chowdhury 1986);
- w3: an optimization algorithm simultaneously calibrates weights attributed to each meteorological station with parameters of the rainfall-runoff models. Weights are bound between 0 and 1, and the sum of weights must add up to 1 (technically, temporary weights are set by an algorithm to values within [0,1], then all weights are summed and the values of the temporary weights are divided by this sum).
- w4: is similar to w3, weights are also bound within [0,1], but do not have to add up to 1, hence the optimization algorithm may “correct” the amount of precipitation (and air temperatures) for the whole catchment. This may be especially important for precipitation during the cold season, as snowfall is under-catched by most classical measurement

techniques (Goodison et al. 1998). However, summer rainfall undercatch may also play some role (Taskinen and Soderholm 2016).

As in versions w3 and w4 the weights for data aggregation are calibrated together with rainfall-runoff model parameters, and the efficiency of the approach will depend on the model, the number of its parameters, and the calibration algorithm used. Hence, we perform tests with different models and calibration procedures. Apart from the Suprasl catchment, the same weights are used to aggregate both precipitation and the air temperature data series. In the case of the Suprasl catchment, air temperature measurements were available only from Bialystok, other stations solely provided precipitation data.

2.2 Rainfall Runoff Models

Two popular lumped conceptual rainfall-runoff models are used in this research, HBV and GR4J. We use the same versions of HBV and GR4J with the snow module that have been used and discussed in Piotrowski et al. (2017a). The HBV version used has 13 parameters, the GR4J – seven (three of the snow module and four of the basic GR4J). Model parameters and their boundary values used during calibration are defined in Suppl. Table 2. The total number of parameters to be found for each model by the calibration procedure (problem dimensionality) is given in Suppl. Table 3. Note that problem dimensionality depends on the model, catchment and the version of meteorological data averaging.

The rainfall-runoff models are run on a daily basis, the river runoff at time $t+1$, noted y_{t+1}, is simulated based on catchment-averaged precipitation, air temperature and potential evapotranspiration (*Ep*) data from the previous day *t*. Potential evapotranspiration is computed according to the classical Hamon (1961, 1963) method.

2.2.1 Error Correction Procedure

In this study we consider both 1) the results obtained directly from HBV or GR4J models (we call them “raw” further in the paper), and 2) results updated by means of a linear error correction procedure (Refsgaard 1997; Madsen et al. 2000). We wish to find out whether the impact of the meteorological data aggregation method on modeling performance is affected by error correction, which is frequently used in practical applications. Let us denote the output from the model (HBV or GR4J) for *t*-th day by ${y^m}_t$. We may use it directly as a simulated value of runoff (${y^s}_{t+1} = {y^m}_{t+1}$, the “raw” case), or apply an error correction procedure, in which the past outputs from the HBV or GR4J are added as exogenous inputs to the linear regression error model:

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

where $\varepsilon^s_t = y_t - y^m_t$ is the prediction error. Hence, in the second case the final simulated runoff is calculated as

$$y^s_{t+1} = y^m_{t+1} + \varepsilon^s_{t+1} \tag{2}$$

where ${y^s}_{t+1}$ is the simulated value of runoff for *t*-th day, considered to be a final modelling output. Following Madsen et al. (2000) and (Piotrowski et al. 2017a), δ is set to 3 days, hence the model from eq. (1) has the following form

$$\varepsilon_{t+1}^{s} = a_0 + a_1\varepsilon_t^s + a_2\varepsilon_{t-1}^s + a_3\varepsilon_{t-2}^s + a_4 y_{t+1}^m + a_5 y_t^m + a_6 y_{t-1}^m \tag{3}$$

with linear parameters a_0-a_6, that are fitted to the calibration data by means of the least squares approach.

2.3 Calibration Methods

The performance of conceptual rainfall-runoff models calibrated by different optimization methods may be uneven (Goswami and O'Connor 2007; Arsenault et al. 2014; Piotrowski et al. 2017a). Hence, three different optimization algorithms are applied separately in this study to each model on every catchment, namely MDE_pBX, SPS-L-SHADE-EIG and GLPSO.

Control parameters of each specific algorithm, including population size, were kept the same as suggested in the source papers. In the case of SPS-L-SHADE-EIG, the variant and control parameter settings considered as default in Guo et al., (2015) are used. Problem dimensionalities for each case considered in this paper are given in Suppl. Table 3. The initial population was generated randomly from uniform distribution within the bounds set individually for each parameter of every model (HBV or GR4J, see Suppl. Table 2). As the results may depend on the maximum number of function calls allowed during calibration (Piotrowski et al. 2017b), we set this value to 30,000 for all considered cases. To obtain statistics that would allow proper comparison of the methodologies used, calibration of each model by every algorithm on each catchment considered in this study was repeated 30 times (hence 30 runs of each calibration algorithm was made for every considered case).

2.4 Comparison Criteria

As objective function the mean square error (*MSE*) is used in this study

$$MSE(\mathbf{p},\mathbf{w}) = \frac{1}{N}\sum_{t=1}^{N}\left(y_t - y_t^s(\mathbf{p},\mathbf{w})\right)^2 \tag{4}$$

where $y_t^s(\mathbf{p},\mathbf{w})$ is a simulated runoff for time t, $y_t$is a measured value of runoff, and N is the number of daily data in a particular data set (calibration or validation) for a particular catchment, $\mathbf{p}$ is a vector of model parameters, and $\mathbf{w}$ is a vector of weights used to obtain the aggregated meteorological data series. Hence, minimization problems are solved by optimization algorithms. In addition, we compute the values of Nash-Sutcliffe (NSC) coefficient (Nash and Sutcliffe, 1970)

$$NSC(\mathbf{p},\mathbf{w}) = 1 - \frac{\sum_{t=1}^{N}\left(y_t - y_t^s(\mathbf{p},\mathbf{w})\right)^2}{\sum_{t=1}^{N}\left(y_t - \overline{y}\right)^2} \tag{5}$$

where $\overline{y_t}$ is a mean of N measured runoff values. Contrary to *MSE*, the higher value of *NSC*, the better performance of the model, and 1 means perfect fit. To compare the results the full ranges of *MSE* values obtained in 30 runs are discussed, and the information on the mean performance and respective standard deviations is given.

2.5 N-Way Analysis of Variance

To find which factors (model, calibration algorithm, error correction procedure or the method to attribute weights to meteorological stations) are most important for the results of rainfall-runoff modeling, the 4-way analysis of variance (ANOVA, von Storch and Zwiers 2001) has been used with the post-hoc Tukey test (Tukey 1949) in a way similar to what was presented in Osuch et al. (2017). The spread of biases induced by each of the four factors is defined as

$$y_{abcd} = \mu + M_a + C_b + E_c + A_d + \varepsilon_{abcd} \tag{6}$$

where y_{abcd} is the mean bias of the modeled runoff for the a-th model (HBV or GR4J), b-th is the calibration algorithm (GLPSO, MDE_pBX or SPS-L-SHADE-EIG), c-th is the variant of error correction (with or without the error correction procedure) and d-th is the method to attribute weights to meteorological stations (w1, w2, w3, w4). In eq. (6) μ is the overall mean, M_a, C_b, E_c, A_d are the principal contributions to the overall variance from the model, calibration method, error correction procedure and method to attribute weights to meteorological stations, respectively, and ε represents an unexplained variance. Interactions between factors were not considered. The analyses were done by the Type III sums of squares ANOVA, separately for mean MSE values from the calibration and validation data sets.

To test the significance of difference in the mean response for the four groups considered, the post-hoc Tukey test was used for the results obtained from the ANOVA. The results from Tukey's test are shown as a graph of population marginal mean estimates (Searle et al. 1980) and the comparison intervals are represented by a continuous line extending from a particular symbol. According to the test, the means of two groups are significantly different (at $\alpha = 0.05$) if their intervals do not overlap.

2.6 Study Area and Data Sets

In this paper we use data collected from three catchments: the Suprasl River (Poland), the Biala Tarnowska River (Poland) and the Cedar River (Washington State, USA). In Suppl. Figs. 1-3 the location of the catchments and meteorological stations are depicted. The Suprasl River catchment is lowland, the other two are mountainous. All catchments are of comparable size, a few hundred square kilometres, and are located in temperate climatic conditions, between 47°N and 54°N. In winter and spring time, the melting of snow is an important factor for runoff generation at each catchment. Other important information on each catchment and the data set used are given in Suppl. Table 1.

The data available from each catchment are divided chronologically into non-overlapping calibration and validation sets (see column 3 in Suppl. Table 1). During model optimization the calibration set, composed of roughly 70% of available data, is used. The validation set is considered as independent data which are not used during calibration at all. The role of the validation set is to verify the quality of the calibrated models. The first year (365 days) of the calibration data is considered as a run-up period and is not used to compute the value of the objective function. In this study the validation set is always composed of the data collected later than data used for the calibration (a chronological approach is used).

The hydro-climatic conditions observed in each catchment are shown, separately for the calibration and validation data periods, in Suppl. Fig. 4. Mean, minimum and maximum values of the mean monthly air temperatures, the mean monthly flow, and the sum of the monthly precipitation are given. The illustrated meteorological conditions represent the catchment-averaged values, assuming a simple average from all available meteorological stations considered for a particular catchment. We may see that for each catchment there are no major differences in hydrological and climatic conditions between the calibration and validation sets, apart from the lack of extremely cold periods during January and February in the Suprasl and Biala Tarnowska catchments in the validation data set (such cold periods were observed in the calibration set).

For the Cedar and the Biala Tarnowska catchments, precipitation and air temperature data from three and five meteorological stations are available, respectively. For the Suprasl catchment, precipitation data come from five stations, but air temperature is measured only at a single site (Bialystok). Hence, the catchment-averaged precipitation is computed for each catchment based on all relevant stations, but air temperature is averaged (using the same weights as precipitation) over the catchment only in the case of the Biala Tarnowska and Cedar catchments. Some considered meteorological stations are located close to, but outside of the proper catchment; their importance for the rainfall-runoff modelling will be assessed by appointing weights using specific procedures considered in this study.

3 Results

The 30-run mean and standard deviation values of *MSE* and *NSC* for each tested variant and catchment are given in Supplementary Tables 4–9. However, we discuss results by referring to the figures, which aim at presenting the most important findings more clearly. First, we discuss the differences of weights attributed to meteorological stations obtained from various methods. Then, we briefly mention the main differences in lumped meteorological data created with such weights. Finally, the performance of models using each considered approach is debated.

3.1 Impact of Calibration on Weights Attributed to Meteorological Stations

For each catchment, the variability of weights over 30 runs of the optimization procedure attributed to each station by the four aggregation methods (w1-w4) is given separately for every model and calibration algorithm by box-plots in Fig. 1. We suggest Fig. 1, and each of the sub-figures, should be read by rows. In each row the spread of weights attributed by methods w1-w4 to a particular meteorological station (s1-s5) is given (separately for each calibration algorithm). Weights are always limited within [0,1] an interval. However, weights attributed to each station by methods w1, w2, and w3 do sum up to 1; in the case of w4 they do not (see section 2.1 for clarification). Definitions of meteorological stations are given in the caption to Fig. 1. For each configuration (catchment, model, calibration algorithm, method of meteorological data averaging) the calibration was repeated 30 times. However, for the first two methods of meteorological data averaging (w1-w2) the weights were deterministically attributed to each meteorological station, irrespective of the calibration algorithm (hence we have just one set of weights per catchment for version w1, and the other one for version w2). In contrast, when weights attributed to meteorological stations are calibrated by means of optimization algorithms, the results may differ in each run, hence we obtain a distribution of weights attributed to each station in 30 runs of the algorithm.

When weights are set by the simple average method (w1), the weight attributed to each station equals 0.2 (or 0.33 in the case of the Cedar catchment), due to the number of

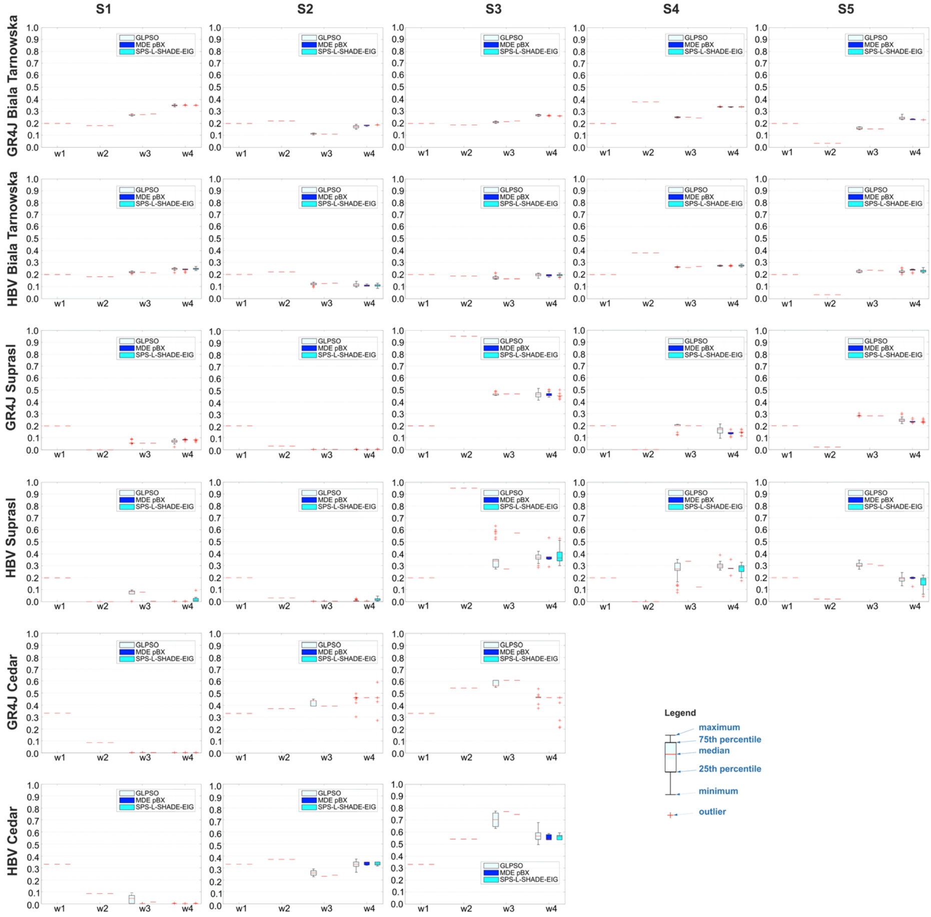

Fig. 1 Box plots showing the variability of weights attributed by each method (w1-w4) to every meteorological station (see section 3.1 for the explanations). In the case of the Biala Tarnowska and Suprasl catchments there are five meteorological stations (represented by columns s1-s5); in the case of the Cedar River catchment there are three meteorological stations (represented by columns s1-s3). Results are shown separately for the HBV and GR4J models. The order of meteorological stations is the same as the order in Table 1: for Biala Tarnowska: s1 – Biecz; s2 – Krynica; s3 – Nowy Sacz; s4 – Tarnow; s5 – Wysowa; for Suprasl: s1 – Bialystok; s2 – Zablodow; s3 – Grodek; s4 – Suprasl; s5 – Szudzialowo; for Cedar River: s1 – Seattle-Tacoma Int's Airp.; s2 – Cedar Lake; s3 – Landsburg

meteorological stations taken into account. When Thiessen Polygons are used, weights are the same as given in Suppl. Table 1 (and do not differ for the HBV and GR4J models). For the w3 and w4 methods, 30 values are obtained for each weight; in these cases weights may also differ quite noticeably depending on whether the HBV or GR4J model is used. For example compare weights attributed to station 3 of the Suprasl catchment by versions w3 and w4 when HBV or GR4J are used. The differences come out because the weights are calibrated together with model parameters, hence they depend on how well the model parameters are fitted.

Generally, from Fig. 1 we see that when weights are calibrated by an algorithm together with model parameters, they frequently differ from weights obtained by Thiessen Polygons. For mountainous catchments (Cedar and Biala Tarnowska) the differences are modest, often within a ± 0.1 margin; moreover, weights obtained in particular runs are similar. A few exceptions are observed for stations s2 and s3 of the Cedar catchment, especially when the w4 method is used. In the case of the flat Suprasl catchment much more diverse weights are attributed to stations s3-s5. They depend on which model and calibration algorithm is applied, but also may highly differ for each run, especially when the HBV model is used. This may be due to many local optima in the search space, or may indicate a poor convergence of the algorithm. Physically, it is probably the effect of flat orography, which does not make any area of the catchment more prone to higher rainfall. Hence, in a lowland catchment weights become fitted to a limited number of high rainfall data that may be distributed randomly across the catchment.

If we sum up the weights attributed by method w4 (where weights do not have to sum up to 1) to all stations in each run, we obtain: in the case of the Biala Tarnowska catchment values between 0.99–1.06 for the HBV model, but much larger, 1.35–1.38 for GR4J; in the case of the Cedar catchment often the values are within 0.86–0.96 for both models (but in a very few cases the weights attributed to GR4J summed up to just 0.64); in case of the Suprasl catchment the values were between 0.81–0.89 for HBV and 0.88–0.96 for GR4J. Hence, in the majority of cases the calibrated weights suggest that precipitation that contributes to runoff may be lower than the measured one in the case of the Suprasl and the Cedar catchments; on the contrary, precipitation seems to be underestimated in the Biala Tarnowska catchment. We are puzzled by the large differences obtained for the Biala Tarnowska catchment when the HBV or GR4J models are used.

3.2 Impact of Calibration on the Lumped Meteorological Time Series

Because weights attributed to meteorological stations by calibration algorithms in each run do differ, and are often different from those obtained by Thiessen Polygons, some differences in the lumped meteorological data series created with those weights are expected. To illustrate this, in Suppl. Figs. 5–6 we show an example of a lumped air temperature series generated by each spatial aggregation method (versions w1-w4) for 100 day periods from a validation data set. Suppl. Figs. 5–6 show examples from the Biala Tarnowska and the Cedar River catchment. For the Suprasl catchment we have just one station with air temperature measurements and other stations provide a precipitation time series only. We have selected the periods that are representative for each catchment; showing longer data on a single figure makes the differences barely recognizable. Note that when weights were calibrated by optimization algorithms (w3-w4), we show only the highest (max in Suppl. Figs. 5–6) and the lowest (min in Suppl. Figs. 5–6) lumped air temperature for a particular day, among those proposed by all three algorithms for both models (hence we show the highest and the lowest air temperatures for a particular day from 180 generated series, which come up from 3 algorithms, 2 models, and 30 runs). Accordingly, in Suppl. Figs. 7–9, we illustrate examples of a lumped precipitation time series generated by each method for a 23 day long period with frequent precipitation (the period is shorter to make differences in precipitation recognizable).

From Suppl. Fig. 5 we see that for the Biala Tarnowska catchment the lumped air temperatures aggregated by weights obtained by a simple average (w1), Thiessen Polygons (w2), or optimization algorithms (w3) are similar, with the exception of the variant when the weights do not have to sum up to 1 (w4). Lumped air temperatures aggregated with weights proposed by variant w4 lead to diverse air temperatures, which may be more extreme (hot in summer, cold in winter) than air temperatures aggregated by other methods. This is because the calibrated weights, if not forced to sum up to 1, frequently sum up to values over 1.3 when the GR4J model is used on this catchment.

In the case of the Cedar River catchment (Suppl. Fig. 6), the results obtained by version w4 are the most different to the others. However, the lumped air temperatures aggregated with weights that sum up to 1 that were obtained by different methods (w1-w3) also differ for this catchment, sometimes even by a few degrees C. Hence, the impact of the method which attributes weights to meteorological stations on lumped air temperatures seems larger for the Cedar catchment than for the Biala Tarnowska catchment.

In the case of lumped precipitation (Suppl. Figs. 7–9) the differences are similar to those observed for lumped air temperatures. For the Suprasl and the Cedar catchments, the lumped precipitation series are more diverse than in the case of the Biala Tarnowska catchment. The main differences are observed when weights generated by the optimization algorithm do not have to sum up to 1. However, the differences between the linear (w1) and Thiessen Polygons (w2) methods are also large, especially for the Suprasl and the Cedar River catchment, which is due to the specific locations chosen.

The differences between the lumped precipitation time series created with weights obtained from different methods are higher than the differences in lumped air temperatures. This is an effect of highly uneven precipitation across the catchments. The lumped meteorological series that are most different to the others are obtained when calibration algorithms are used to calibrate weights attributed to each meteorological station, and the weights do not have to sum up to 1 (w4). However, the precipitation time series obtained by weights that do sum up to 1 but were achieved from various methods may also differ noticeably; the highest precipitation generated may even be over 100% higher than the lowest for a rainy day, see the example from the Cedar catchment in Suppl. Fig. 8.

3.3 Impact of Calibration of Weights Attributed to Meteorological Stations on Modeling Performance

It is important to verify whether the impact of the chosen method of attributing weights to meteorological stations may be considered significant. To address that issue, an N-way ANOVA with four factors (model, calibration algorithm, error correction procedure and method to attribute weights to meteorological stations) is applied for each catchment, separately for calibration and validation data, as discussed in section 2.6. The ANOVA test is based on mean *MSE* values from 30 runs performed for each considered case. Table 1 shows the contributions to the total variance by each among four factors, and associated *p* values (the higher the p value, the less significant the factor). For very low p values we set them to 0, as they indicate significance of particular factor anyway. Figure 2 shows the results of the post-hoc Tukey (1949) test for mean *MSE* separately for the calibration and validation data. Note that the mean *MSE* of biases of particular groups of a specific factor are said to be significantly different (at $\alpha = 0.05$) if their intervals do not overlap. From Table 1 and Fig. 2 we may find that the calibration algorithm turned out to be unimportant in all catchments, considering both calibration and validation sets. This may indicate that all three calibration methods were chosen properly, or that some other methodological factors are so important, that they "hide" the impact of the optimization method. The most important factor is whether the error correction procedure is used or not; according to Table 1, its relative contribution to the total variance is over 66%, in some cases even 95%. However, according to Tukey's test (Fig. 2) all factors apart from the calibration algorithm are significant, and the way the weights attributed

Table 1 Relative contributions (in %) of four factors (the error correction procedure, model, calibration algorithm and method of attributing weights to meteorological stations) to the total variance according to the 4-way ANOVA test for mean MSE, performed separately for calibration and validation data from each catchment

catchment	data set	error correction procedures		models		optimization algorithms		weights attribution methods		unexplained variance (ε)
		relative contribution (%)	p value	relative contribution (%)	p value	relative contribution (%)	p value	relative contribution (%)	p value	relative contribution (%)
Biala Tarnowska	calibration	90.08	0.00	0.91	0.00	0.01	0.84	7.62	0.00	1.38
	validation	85.77	0.00	5.57	0.00	0.12	0.25	6.84	0.00	1.71
Cedar	calibration	94.80	0.00	1.78	0.00	0.00	0.98	0.46	0.12	2.95
	validation	99.39	0.00	0.03	0.07	0.00	0.88	0.25	0.00	0.33
Suprasl	calibration	97.82	0.00	0.76	0.00	0.01	0.90	0.32	0.01	1.09
	validation	68.49	0.00	1.07	0.10	0.00	0.99	15.48	0.00	14.96

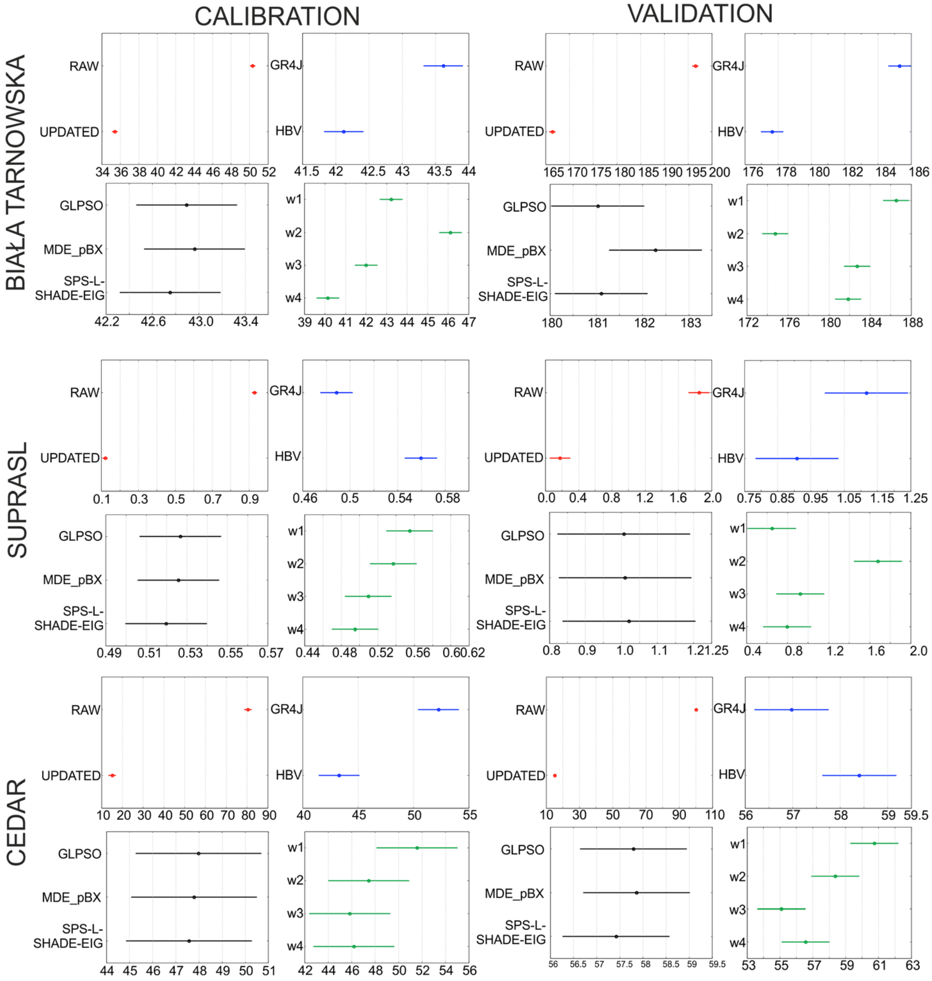

Fig. 2 Variability of mean MSE values over 30 runs of the optimization procedure for Biala Tarnowska (top two sub-plots), Suprasl (middle two sub-plots) and Cedar (bottom two sub-plots) separately for the calibration (left sub-plots) and validation (right sub-plots) sets, as a function of the model (top right panel of each sub-plot), bias correction procedure (top left panel of each sub-plot), calibration algorithm (bottom left panel of each sub-plot) and method of attribution of weights to meteorological stations (bottom right panel of each sub-plot) based on 4-way ANOVA and Tukey's honestly significant difference criterion

to each meteorological station are set seems to be the second most important. For the validation data set the difference between attributing the weights to meteorological stations by means of Thiessen Polygons or by calibrating them together with model parameters is always significant when weights have to sum up to 1 (intervals of w2 and w3 do not overlap). This is an important finding, even if not always confirmed by the results from Tukey's test for the calibration data sets.

The variability of *MSE* for raw data (without an error correction procedure) obtained for each catchment with the use of every model, the calibration algorithm and the method of attributing weights to meteorological stations, are shown by box-plots in Fig. 3 for the calibration and validation data sets. Similar results after the error correction procedure are

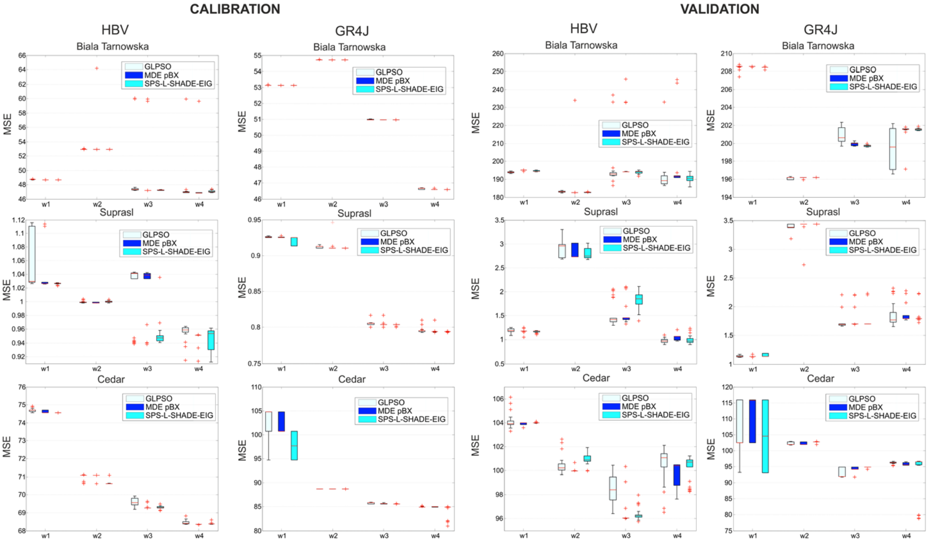

Fig. 3 Box-plots showing variability of mean square errors (MSE) obtained in 30 runs for raw (without error correction) calibration (left half) and validation (right half) data for each catchment. Results obtained by each calibration algorithm using every method to attribute weights to meteorological stations are given separately

given in Fig. 4. The mean and standard deviation values of *MSE* and NSC for each considered variant are also given in Suppl. Tables 4–9.

From Fig. 3 we see that for the calibration data it is always better to optimize weights attributed to meteorological stations together with model parameters than setting them by Thiessen Polygons or a simple average when the GR4J model is used. In the case of HBV the

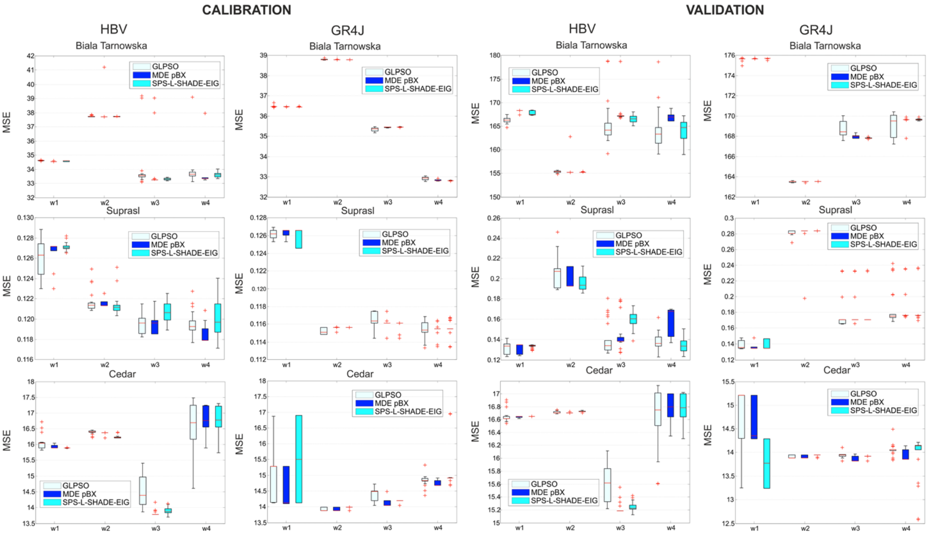

Fig. 4 Box-plots showing variability of mean square errors (MSE) obtained in 30 runs for calibration (left half) and validation (right half) data for each catchment when an error correction procedure is applied. Results obtained by each calibration algorithm using every method to attribute weights to meteorological stations are given separately

results depend on the calibration algorithm, but when SPS-L-SHADE-EIG is used, the joint calibration of parameters and weights attributed to meteorological stations is almost always (one exception among 30 runs for the Suprasl catchment when the algorithm stuck in a poorer local minimum) the best choice. Moreover, results are almost always better when the calibrated weights do not have to sum up to 1 (*MSE* for w4 is lower than for w3).

This positive outcome for the calibration data is, however, not necessarily confirmed on the validation data sets (see Fig. 3). For validation data, calibrating the weights attributed to meteorological stations together with model parameters is the best choice for the Cedar catchment irrespective of the model used, especially when weights have to sum up to 1. When weights do not have to sum up to 1, joint calibration of weights and HBV parameters leads to the best results for the Suprasl catchment, but when GR4J is used, the simple mean leads to better results. For the Biala Tarnowska catchment, the Thiessen Polygons method is the best choice. This finding for the Biala Tarnowska is interesting, as it shows that the large differences in sums of weights observed for version w4 (when the weights were calibrated together with the HBV or GR4J models and did not have to sum up to 1) discussed in section 3.1 are of minor practical importance (as Thiessen Polygons leads to the best results for this catchment anyway) and may be due to some kind of overfitting in the case of the GR4J model (for which weights summed up to a value much higher than 1 when w4 was used). Hence, finding the best approach of attributing the weights to meteorological stations based on the validation data is difficult; results depend on the catchment and, to lesser degree, the rainfall-runoff model used. However, apart from the Biala Tarnowska catchment, the joint calibration of model parameters and weights attributed to meteorological stations may be recommended.

When an error correction procedure is used to improve the results, the outcome becomes completely blurred. The joint calibration of weights attributed to meteorological stations and model parameters is the best choice according to the validation data when an error correction procedure is used (Fig. 4) only when HBV is applied on the Cedar catchment (assuming that weights have to sum up to 1). When GR4J is used for this catchment the results are inconclusive, and significantly depend on the calibration algorithm. For other catchments either the simple average (on the Suprasl catchment), or Thiessen Polygons (on the Biala Tarnowska catchment) leads to better results.

4 Conclusions

In the present study we have carefully researched whether the joint calibration of lumped rainfall-runoff model parameters and the weights attributed to various meteorological stations is efficient. The final outcome is, unfortunately, not as clear as readers would hope for. Joint optimization is advised for the calibration data, highlighting the potential usefulness of optimization algorithms. The effectiveness of the method is also partly confirmed on validation data, as long as an error correction procedure is not applied. However, when one uses an error correction procedure, there is no justification for calibrating the model parameters and the weights attributed to meteorological stations jointly.

There are no clear links between the catchment and the performance of the discussed methodology. As the method performs well for calibration data, but not necessarily for validation data, its usefulness depends on both data quality and weather patterns that dominate major runoff events during different periods, not the simple characteristics of specific catchments.

Acknowledgements The research published in this paper has been financed from the National Science Centre, Poland, grant number 2016/21/B/ST10/02516 (2017-2020) and supported within statutory activities No 3841/E-41/S/2019 of the Ministry of Science and Higher Education of Poland.

Compliance with Ethical Standards

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

References

Anctil F, Lauzon N, Andréassian V, Oudin L, Perrin C (2006) Improvement of rainfall-runoff forecasts through mean areal rainfall optimization. J Hydrol 328:717–725

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

Arsenault R, Gatien P, Renaurd B, Brissette F, Martel JL (2015) A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J Hydrol 529:754–767

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

Bhadra A, Bandyopadhyay A, Singh R, Raghuwanshi NS (2010) Rainfall-runoff modeling: comparison of two approaches with different data requirements. Water Resour Manag 24:37–62

Girons Lopez M, Seibert J (2016) Influence of hydro-meteorological data spatial aggregation on streamflow modeling. J Hydrol 541:1212–1220

Goodison BE, Louie PYT, Yang D (1998) WMO solid precipitation measurement intercomparison: final report. Instruments and observing methods report 67, WMO/TD-no. 872. World Meteorological Organization, Geneva, Switzerland

Gong YJ, Li JJ, Zhou Y, Li Y, Chung HSH, Shi YH, Zhang J (2016) Genetic learning particle swarm optimization. IEEE Trans Cybernet 46(10):2277–2290

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

Groisman PY, Legates DR (1994) The accuracy of United States precipitation data. Bull Am Meteorol Soc 75(2): 215–227

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

Hamon WR (1961) Estimation potential evapotranspiration. J Hydraul Div Proc ASCE 87(HY3):107–120

Hamon WR (1963) Computation of direct runoff amounts from storm rainfall. Int Assoc Sci Hydrol 63:52–62

Islam SM, Das S, Ghosh S, Roy S, Suganthan PN (2012) An adaptive differential evolution algorithm with novel mutation and crossover strategies for global numerical optimization. IEEE Trans Syst Man Cybernet, Part B –Cybernet 42(2):482–500

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

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

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

Madsen H, Butte MB, Khu ST, Liong SY (2000) Data assimilation in rainfall–runoff forecasting. In: fourth international conference on hydroinformatics, 23–27 July 2000, Cedar Rapids, IA., Iowa City, IA: University of Iowa, College of Engineering, USA

Maidment DR (ed) (1993) Handbook of hydrology. McGraw-Hill, New York, USA

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

Osuch M (2015). Sensitivity and uncertainty analysis of precipitation-runoff models for the middle Vistula Basin, Chapter in GeoPlanet: Earth and Planetary Sciences, 61–81, https://doi.org/10.1007/978-3-319-18854-6_5

Osuch M, Lawrence D, Meresa HK, Napiorkowski JJ, Romanowicz RJ (2017) Projected changes in flood indices in selected catchments in Poland in the 21st century. Stoch Env Res Risk A 31(9):2435–2457

Peng T, Zhou J, Zhang C, Sun N (2018) Modelling and combined application of orthogonal chaotic NSGA-II and improved TOPSIS to optimize a conceptual hydrological model. Water Resour Manag 32(11):3781–3799

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

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

Piotrowski AP, Napiorkowski MJ, Napiorkowski JJ, Rowiński PM (2017b) Swarm intelligence and evolutionary algorithms: performance versus speed. Inf Sci 384:34–85

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

Searle SR, Speed FM, Milliken GA (1980) Population marginal means in the linear model: an alternative to least squares means. Am Stat 34(4):216–221

Shaw EM, Lynn PP (1972) Areal rainfall evaluation using two surface fitting techniques. Hydrol Sci Bull 17(4): 419–433

Singh VP, Chowdhury PK (1986) Comparing some methods of estimating mean aerial rainfall. J Am Water Resour Assoc 22(2):275–282

Taskinen A, Soderholm K (2016) Operational correction of daily precipitation measurements in Finland. Boreal Environ Res 21:1–24

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

Tukey J (1949) Comparing individual means in the analysis of variance. Biometrics 5(2):99–114

Vaze J, Post DA, Chiew FHS, Perraud JM, Teng J, Viney NR (2011) Conceptual rainfall–runoff model performance with different spatial rainfall inputs. J Hydrometeorol 12:1100–1112

von Storch H, Zwiers F (2001) Statistical analysis in climate research. Cambridge University Press, Cambridge

Vrugt JA, Gupta HV, Bouten W, Sorooshian S (2003) A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour Res 39(8):1201. https://doi.org/10.1029/2002WR001642

Vrugt JA, ter Braak CJF, Diks CGH, Schoups G (2013) Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: theory, concepts and applications. Adv Water Resour 51:457–478

Yuan WL, Liu MQ, Wan F (2019) Calculation of critical rainfall for small-watershed flash floods based on the HEC-HMS hydrological model. Water Resour Manag 33(9):2555–2575

Affiliations

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

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