Methodology for Estimating Streamflow by Water Balance and Rating Curve Methods Based on Logistic Regression

Tarcila Neves Generoso1 · Demetrius David da Silva1 · Ricardo Santos Silva Amorim1 · Lineu Neiva Rodrigues1,2 · Erli Pinto dos Santos1

Received: 23 March 2022 / Accepted: 7 July 2022 / Published online: 1 August 2022
© The Author(s), under exclusive licence to Springer Nature B.V. 2022

Abstract
Both water balance (WB) and rating curve (RC) are methods for estimating streamflow. The first is mostly used to estimate reservoir outflows, while the second is usually adopted in hydrometeorological network streamflow gauges. While WB uses hourly collected data, the RC estimates streamflow using current water level and extrapolation techniques. The objective of this study was to analyze variations in the reservoir’s hourly outflow at Queimado Hydroelectric Power Plant (HPP Queimado) and to propose a method to evaluate whether the estimate of the daily outflows, obtained by the WB method, is similar to the flow values obtained at a conventional station. The logistic regression (LR) model was used because it is a method that adopts binary, categorically dependent variables to identify the event of interest. The results showed that the values of streamflow, obtained from an average of two daily readings, were a good representation of the flows in the region. The LR was able to identify atypical data, especially in the rainy season. This means that data consistency analysis can be faster and safer, when adequately employed and considering the proposed conditions, contributing to both management policies and the management of water resources.

Keywords Extrapolation · Modeling · Outflow · Streamflow gauge station · Water balance method

1 Introduction

According to Wannasin et al. (2021), hydroelectric plants can have several constructive purposes, in addition to focusing on energy generation. The authors clarified that water accumulation reservoirs can mitigate floods and droughts through flow regulation, by reducing flood peaks and increasing minimum flows. In Brazil, Joint Resolution ANEEL/ANA n° 03 (August 10, 2010) determined that, in situations where regulating reservoirs are present at hydroelectric plants, hydrometric stations should monitor rainfall, water level and flow and they should be automated and telemeasured, recording data at a maximum interval of one hour.

* Tarcila Neves Generoso
tarcila_neves@yahoo.com.br

Extended author information available on the last page of the article
The amount of water that enters and leaves a reservoir over a given period of time, is normally calculated by the water balance (WB) method. For each plant in the system, the volume stored at the end of a time interval must equal the sum of its initial volume and the incremental inflow volume (Westin et al. 2021). To determine the outflows by this method, it is necessary to obtain two variables: the turbined flow (removed from the reservoir by the penstock that feeds the water to the turbines for energy generation) and the spilled flow (which is released by the reservoir spillways and does not pass through the turbines (Molina 2016)).

The requirement to collect data at hourly intervals in hydroelectric plants allows 24 readings to be taken, distributed throughout the day, which increases the accuracy of average daily estimates because it minimizes the errors caused by flow variability and helps us understand the dynamic demand for electricity in the region. Thus, the average daily flow readings become more representative, in relation to the flow variation regime.

On the other hand, in areas where there are no hydroelectric power plants, the most used method for estimating flows in watercourses is the rating curve (RC), which is based on the measurement of water height, or water level, in the river. The construction of the RC or calibration curve depends on the determination of different readings of the water level and the corresponding measurements of streamflow; the latter is dependent on the average flow velocity and on the wet area considered in cross section. This area, however, may change over time, as the characteristics of the riverbed change over the years, which means that the RC may lose its representativeness, requiring frequent field flow measurement campaigns (Manfreda 2018). The adjustment of a reliable key curve is sometimes difficult to establish due to the associated uncertainties (Ibeje 2018).

The relation between water level and streamflow, measured at points over time to obtain the RC, can be expressed by means of an equation, which allows continuous water level readings to be converted into streamflow series. At stations belonging to the Brazilian National Hydrometeorological Network (RHN) of the Brazilian National Agency for Water and Basic Sanitation (ANA), water level readings are taken twice daily (7:00 AM and 5:00 PM – Brasília local time: GMT-3) and the average value of the two readings is taken to be representative for the purposes of average daily flow estimation.

In situations with maximum flows, however, it is common to apply methodologies for extrapolating the RC. Usually, there are no flow measurements associated with the most critical periods of flooding which, due to risks that field technicians may be exposed to.

Therefore, it is important to evaluate the estimated flows by considering the comparison between the WB and RC methods, since both are recognized and applied worldwide. This study assumed that the WB method would have fewer limitations than the RC method, mainly due to the nature of the extrapolations necessary to obtain the maximum flows in the RC method. As explained by Peña-Arancibia et al. (2015), there is, especially in the highest flow values, for several stations, the high potential uncertainty caused by the extrapolation of the curves.

In view of the above, this study aimed to: i) analyze the behavior of hourly outflow from the reservoir of the Queimado Hydroelectric Power Plant (Queimado HPP); and ii) propose a methodology, based on logistic regression (LR), to evaluate the daily outflow from the Queimado HPP, estimated by the WB method and compared to the daily flows of the conventional streamflow gauge station (estimated by the RC method) located downstream of the reservoir.
2 Material and Methods

2.1 Study Area

This study was carried out using information from the reservoir of the Queimado Hydroelectric Power Plant (Queimado HPP), located in the Preto River Basin (10,325 km²), which is within the São Francisco River Basin, localized in Brazil. Queimado HPP (Fig. 1) is owned by a consortium of two energy companies: the Companhia Energética de Minas Gerais (CEMIG) and the Companhia Energética de Brasília. The Plant started operating in 2004, has a contribution area of approximately 3,657 km², and its reservoir has a useful volume of 389.46 hm³.

Immediately downstream of the reservoir there is a conventional ANA streamflow gauge station, ‘Fazenda Limeira (4260000)’, with a drainage area equivalent to 3,890 km²; labeled Q1 in Fig. 1. The distance of the Q1 station to the section where the reservoir outflow occurs, Qd, is approximately 10 km and there is no significant contribution from tributaries or withdrawal of water in this stretch.

The region has two well-defined seasons, with the rainy season starting in November and ending in April and the dry season, with low incidence of rainfall, between May and October.

2.2 Hydrological Data

Data on the hourly outflows from the Queimado HPP were obtained from CEMIG so that, for each day, there are 24 hourly readings, starting at 01:00 AM and ending at 00:00 of the subsequent day, from 2004 to 2019. Due to inconsistencies found in the series of hourly readings, an adjustment was made to the series so that the time series, for each day, would be properly aligned.

Fig. 1 Location of the study area, with emphasis on the outflow measurement section (Qd) of the Queimado reservoir and the Fazenda Limeira station (42460000), represented by Q1
flows in the period from 2004 to 2005, probably justified by the instabilities resulting from the beginning of operations at the reservoir in 2004, it was decided to exclude these data and only work with the period from 2006 to 2019.

Within the working period, days with at least one gap during the day were also excluded, since the objective was to precisely evaluate the variability of flow throughout the day.

The historical data and the RC of the Q1 streamflow gauge station, were obtained from the Hidroweb platform (https://www.snihr.gov.br/hidroweb/serieshistoricas), from the RHN. However, this station only showed coherent data until 2014, so the period from 2006 to 2014 was established for comparative analysis with the outflow data from the reservoir (Qd).

In order to support the analyses between the sections of interest, rainfall data from the Limeira Farm rain gauge station (01647008), located near the section of the Q1 streamflow gauge station, were also obtained from the Hidroweb platform for the period 2006 to 2014.

2.3 Analysis of the Hourly Behavior of Outflows

Due to the significant variability of the outflow from the reservoir (Qd) over a day, the flow behavior over a 24 h period was evaluated using the average monthly value of each of the times considered, for the period 2006 to 2019, with the aim of identifying the times when the highest amplitudes of water turbining (due to electricity production) are obtained. Possible differences between the mean of the 24 h total and the measurements at the times of 7:00 AM and 5:00 PM (used in the RHN) were also evaluated, to check if there are considerable differences in the behavior of the flows throughout the day.

2.4 Comparison of Methods for Flow Estimation

For this analysis, a criterion for data selection was applied, considering the identification and removal of incompatible values between the outflows (Qd) and those of the Q1 station. Thus, data that showed a Q1/Qd ratio ≤ 0.90 were removed, since the basic premise assumed that stations located downstream tend to have daily flow values higher than those of upstream sections, which was reinforced in the present study as no significant water abstractions were identified between the Queimado HPP and the Q1 station. Therefore, 10% was the limit of permissiveness incorporated, because of the level of accuracy of the equipment and possible unidentified captures along the segment of the watercourse between the two sections considered. In addition, data whose Q1/Qd ratios behaved as outliers in the monthly boxplot analysis were also eliminated. The percentage of data eliminated after considering the above-mentioned criteria was about 16%.

Next, the outflow data (Qd) and the data from the conventional ANA streamflow gauge station (Q1) (in m³·s⁻¹) were converted to specific streamflow (in L·s⁻¹·km⁻²) (qd and q1, respectively). This enabled a comparison of the values obtained with the different flow estimation methods for the different drainage areas.

In order to evaluate the methodologies more uniformly, the same time interval was considered when obtaining the data, both in the outflow and in the conventional station of ANA. Therefore, for this analysis, the specific daily streamflows were taken into account, obtained from the average of the outflow values observed at 7:00 AM and 5:00 PM (qd7-17), as well as the daily specific streamflows of the downstream fluviometric station (q1), also obtained by readings at 7:00 AM and 5:00 PM.
Initially, an analysis of the distribution of the outflow and streamflow data from the conventional streamflow gauge station was performed via a quantile-quantile plot (qq-plot). This tool was explained by Rodu and Kafadar (2021) and it compares two distributions by plotting the quantiles of one distribution against those of the other, and comparing them with a reference distribution, so that the presence of deviations from the straight line indicate differences in the shapes of the two distributions.

Next, the original flow series (qd\textsubscript{7-17} and q1) were tested to check whether or not they could be classified as being statistically similar. This was achieved by applying the F test, in the modified form of Graybill (1976). The Graybill F test is recommended when the objective is to know whether the estimates obtained by an equation (Y\textsubscript{j}) are statistically equal to the observed values (Y\textsubscript{1}) or whether the estimates of a given variable, obtained by two methods of evaluation (Y\textsubscript{1} = Y\textsubscript{j}), are statistically equal. In this case, the series of outflow data, qd\textsubscript{7-17}, was considered to be the standard method, with the flows estimated at the ANA station (q1) as an alternative estimation method. The Graybill F test was applied by testing the Hypothesis H\textsubscript{0}: \beta\textsubscript{0} = 0 and \beta\textsubscript{1} = 1. In the present study, a 1% significance level (0.01) was considered. This test was applied using features of the forestmangr package (Braga et al. 2021).

After verifying the statistical differences between the series, they were subjected to a new F test, now considering only the observations in which the modulus of the difference between q1 and qd\textsubscript{7-17} (absolute difference = |q1 - qd\textsubscript{1-17}|) was lower than or equal to the standard deviation of q1, in order to prove whether or not they would become statistically similar.

This analysis was important, so that the application of LR models could be continued. LR is a widely used method for modeling dichotomous results (Strzelecka et al. 2020) and was already being used in hydrological studies, as can be seen in the study by Fathi et al. (2021). The model allows the examination of the influence of many independent variables on a dependent variable. The dependent variable takes only two values: ‘1’ means that an event of interest has occurred (qd\textsubscript{7-17} and q1 showed discrepant) and its absence means that this variable assumes a value of ‘0’ (qd\textsubscript{7-17} and q1 showed ‘similar’ behavior).

The first stage dealt with the imposition of a decision limit, whose function was to allow the algorithm to be able to separate the two classes. Next, the parameters of LR were estimated and the model was fitted by selecting the independent variables, using two different forms of sampling: random and stratified.

The value of the decision limit required for the fit of the LR model was initially based on the value of the standard deviation of q1, previously used for filtering the flow series when the F test was applied.

The LR parameters were estimated using the maximum likelihood function which, according to King and Zeng (2001), is known to suffer from some bias for small samples. When the number of cases of interest is less than 5% of the total samples, the model can frame them as a rare phenomenon, so it is necessary to apply specific corrections to deal with the situation, thus reducing the bias (King and Zeng 2001). For this reason, it was decided to apply the maximum penalized likelihood method proposed by Firth (1993), whose objective was to produce fine estimates for the parameters of the model, introducing a small bias in the score function.

When conducting the analyses, it was verified that the decision limit initially adopted, so that the regression model could be fitted, was insufficient for the maximum likelihood function (which estimates the parameters of the models) to be able to converge. As a result, a new criterion was adopted to increase the number of cases related to discrepant flows and then two thirds (2/3) of the standard deviation of q1 were considered.
Thus, there was a small increase in the discrepant flows (from 19 to 46) which, although apparently not very significant, was enough for the model to be able to converge.

Subsequently, the technical requirements for the LR modeling were verified and the number of variables and observations to be tested were determined.

The independent variables tested, highlighting those effectively used were: Daily average specific outflow from the Queimado HPP reservoir (2 readings – 7:00 AM and 5:00 PM) – $qd_{7,17}$; Difference between $q_1$ and $qd_{7,17}$ – $Dif$ and Julian day of the year on which the observation was recorded – $d$. The independent variables discarded in the modeling after preprocessing were: Daily average specific streamflow of station 42460000 (2 readings) – $q_1$; Daily average specific outflow from the Queimado HPP reservoir (24 readings) – $qd$ and Julian month of the year in which the observation was recorded – $m$.

As LR is very sensitive to multicollinearity in the independent variables, those covariates that presented Pearson’s correlation coefficient greater than or equal to 0.95 were discarded, using tools from the caret package (Kuhn 2020).

In the machine learning procedures, the data were divided into: training, where the hyperparameters of the penalized LR model are fitted, and testing, where the best fit model is used to classify a set of data not known to it, so as to contain every month of the year (this means that there was no risk of fitting a biased regression model).

Therefore, two strategies for the division of data were adopted. The first performed stratified sampling, with the aim of containing the same distribution of the data, with flow values between the lowest and the highest magnitudes being recorded in the series, in the training and test sets. The second strategy performed random sampling, with 70% of the data used for training and 30% used for testing; this division was also adopted by Gharib and Davies (2021). In both approaches, the dataset was divided using features of the rsample package (Silge et al. 2021).

The metrics used to evaluate the models were accuracy (which represents the proportion of cases that were correctly predicted) and the Kappa coefficient (which represents the consistency or agreement of results when the measurement or examination is repeated under equal conditions (Fraga-Maia and Santana 2005). According to Kafy et al. (2021), a Kappa coefficient greater than 0.61 can be considered a good fit. Also, the importance of each of the variables in the models was evaluated using features of the vip package: vip was used to analyze the performance of the models (Accuracy and Kappa) when the variables were omitted through permutation.

Subsequently, the best fit model for LR was adopted and used to classify the flows from a data set that had never been worked by the model (test). Thus, it was possible to produce the contingency matrix, where the layout of a table is constructed to visualize the performance of the algorithm, allowing the flow values whose classification is assigned with the value 0 (‘similar’) to be filtered.

After classification in the test set, the Graybill F test was applied again, to check whether the observations classified by the model with a value of ‘0’in the flow series, could be considered as being statistically similar.

The entire process of analysis, visualization and modeling was carried out using the R programming language (Team 2021) and RStudio software. Modeling, training, validation and testing of the models were performed using features of the caret (Kuhn 2020) and glmnet (Friedman et al. 2010) packages, which have a Penalized Logistic Regression implementation.
Finally, for the data whose classification was assigned with the value 1 (‘attention’), analyses were carried out in association with the RC and the rainfall data obtained in the rain gauge station (01647008), located near the site where the q1 streamflow gauge station is positioned.

3 Results and Discussion

3.1 Analysis of the Hourly Behavior of the Outflow from the Queimado HPP Reservoir

The results obtained from the monthly hourly averages of the outflow from the Queimado HPP reservoir are presented in Fig. 2. These historical series account for the rainy seasons (November to April) and for the dry seasons (May to October) for the period 2006 to 2019.

It can be seen in Fig. 2a that, in general, the outflows show no sudden hourly variations within each month of the rainy period (November to April), which means that the production of electricity remains virtually constant throughout the day, in each of the months in this period. However, a slight increase in the average magnitude of the flows can be seen in some months (November, February, March and April), in the afternoon and night periods rather than the morning, which is still not very expressive (2.5 to 5.0%) considering the magnitude of the regularized flows downstream (approximately 40 m$^3$.s$^{-1}$).

Figure 2b shows the variation of the hourly flows downstream of the Queimado HPP reservoir in the dry season (May to October) and it is possible to note relatively different behavior compared to that of the flow in the rainy season. The analysis shows lowest flow values predominantly occur in the morning period but, despite the variations identified in the dry season, the largest mean difference between the highest and lowest hourly flows in each month does not exceed 4 m$^3$.s$^{-1}$.

However, a comparative evaluation of the behavior of the mean outflow over 24 h and the outflow values calculated from the average of 7:00 AM and 5:00 PM (currently performed by ANA), showed that there are almost no differences in the values from the two time intervals. Therefore, it can be said that the results show good representativeness regarding the behavior of the flows when using the average of the values obtained at 7:00 AM and 5:00 PM.

Fig. 2 Variation of the average hourly outflow from the Queimado HPP reservoir, observed throughout the day, considering the rainy season (a) and dry season (b) of the Preto River Basin
Although the results indicate that there will probably not be any significant problems, in relation to the flow estimates in ANA stations located downstream of the Queimado HPP reservoir due to the observation times (7:00 AM and 5:00 PM), it is important to evaluate whether differences can occur according to the methodology adopted, to quantify the flows in the RHN.

3.2 Comparison of Methods for Flow Estimation

Figure 3 shows the normality graph obtained by the qq-plot, where it was possible to analyze the flow data estimated by the WB method and those of the RC, indicating that there is dispersion between the values, especially when they exceed about 30 L.s⁻¹.km⁻².

When applying the F test in the data series, it was found that the original series of qd7-17 and q1, for the analysis period, were characterized as statistically different. As a result of this, the F test was applied again, considering only the observations in which the modulus of the difference between q1 and qd7-17 was lower than or equal to the standard deviation of q1. It was confirmed that they became statistically similar (Fig. 4).

The number of observations was 2766 in Fig. 4a and 2747 in Fig. 4b. Thus, 19 observations were identified as discrepant between the series, causing changes in the degree of agreement between them, since in (a) the series are statistically different and in (b) they became statistically equal.

After the verifications presented, the fitted models of LR via stratified sampling and random sampling achieved good results for both the training and the testing steps, as shown in Table 1.

Table 1 shows that the Accuracy values of the model, considering the two samplings, reached percentages close to 100%, which is an excellent indication. Regarding Kappa values, there were reductions in the stratified sampling, from the training to the testing phase.

![Figure 3](image)

**Fig. 3** Graph of the normal distribution of the observed values for qd7-17 and for q1, for the period 2006 to 2014
The sharp reduction in this type of sampling indicates the presence of overfitting and bias. On the other hand, in the models using random sampling, the Kappa index increased, indicating the generalization capacity of the model in this approach. In addition, the Kappa

![Scatter plots between qd7-17 and q1 (x, y) resulting from the Graybill F test, where: a represents the statistically different original series; b represents the series statistically equal (*β_0 = 0, **β_1 = 1) after applying the filtering the samples by the standard deviation, and graphs of statistical equality between the methods of the RC and WB from the identification and removal of observations classified as 'attention', considering: stratified sampling (c) and random sampling (d)](image)

Table 1  Statistical indices obtained of the LR model using stratified sampling and random sampling

| Steps     | Sampling | Statistical Indices |
|-----------|----------|---------------------|
|           | Accuracy | Kappa               |
| Training  | Stratified | 0.994  | 0.817 |
|           | Random    | 0.995  | 0.789 |
| Testing   | Stratified | 0.992  | 0.621 |
|           | Random    | 0.996  | 0.901 |
values found for the two models are greater than 0.61 and, therefore, classified as good (Kafy et al. 2021).

In an attempt to check whether the applied modeling would be able to identify the observations classified as ‘attention’ and ensure statistical equality between the specific streamflows qd_{7-17} and q_1, the Graybill F test was re-applied to the data classified as ‘0’ in the LR and it was possible to verify the effectiveness of the methodology using the two samplings, as shown in Fig. 4.

It can be also observed that the coefficients of determination were equal to 0.96 and 0.95 in the stratified sampling and random sampling, respectively, indicating the existence of a high correlation between the flow values obtained by the two methodologies.

In the training phase, the testing to evaluate the importance of each of the variables used in the LR model (for both stratified and random sampling) verified that the variable related to the day of the year (d) was the one that had the least importance. It stands out, however, which does not mean that its presence is not able to improve the performance of the model.

In the training phase, the testing to evaluate the importance of each of the variables used in the LR model (for both stratified and random sampling) verified that the variable related to the day of the year (d) was the one that had the least importance, however, which does not mean that its presence is not able to improve the performance of the model.

When evaluating only the data which were considered as ‘attention’ by the LR, 14 cases were quantified for the test set when random sampling was considered and 5 cases were quantified when stratified sampling was used.

Figure 5 clearly shows the cases found by the LR model with a classification ‘1’ (i.e., ‘attention’) so that most of the incompatibilities found for the flow values occurred predominantly in the rainy season, i.e., between November and April.

The highest values of qd_{7-17} and q_1 found for the period 2006 to 2014 were estimated at 48.2 and 49.9 L.s^{-1}.km^{-2} (176.2 and 194.2 m^3.s^{-1}, respectively). Therefore, it should be noted that, when LR was performed using random sampling, the model identified situations of ‘attention in q_1 for low flows, i.e., incompatibility between the methodologies of the RC and the WB, even in situations where the magnitude of the values is considered low (as observed in Fig. 3), less than 30 L.s^{-1}.km^{-2} and close to the average of 11.7 L.s^{-1}.km^{-2}. Thus, these values are most likely positioned below the upper range of extrapolation of the RC.
This does not mean that there was an error in the modeling, since there may be other reasons why the differences between the estimates may have occurred, since the variable of greatest importance for the model was the difference between \( q_{d7-17} \) and \( q_1 \). The model draws attention to flows that do not show similar behavior to each other and it is up to the analyst to verify the reason for the occurrence of such a situation; it may or may not be incorrectly estimated data.

The modeling performed using random sampling was more sensitive to the existing differences, from a pre-established limit between the estimates in the two analyzed segments, than when stratified sampling was used. For the latter, the ‘attention’ data were only found for specific streamflows above 30 \( \text{L.s}^{-1}.\text{km}^{-2} \). Thus, the classification capacity and behavior of a model depends significantly on the sampling set of data selected for its training.

Figure 5 also presents the analysis of rainfall behavior in the rainfall season close to the Q1 station and, as can be seen, the discrepancies found between the flows \( q_{d7-17} \) and \( q_1 \) are not associated with the occurrence of rainfall. In most of the cases identified, there was no incidence of rain in the region that could influence these changes.

However, in the modeling using random sampling (Fig. 5b) for two of the 14 days classified with incompatible flow values, there may have been some influence of rainfall due to the magnitude of the values recorded (days 02/19/2007 and 04/07/2013), but it is still not possible to affirm with certainty, since rainfall values could not be obtained in the drainage area upstream of the site where the \( q_{d7-17} \) data were collected, so that a more appropriate comparison could be made.

Finally, when analyzing the flow data effectively measured in the field, for the calibration of the RC in the section where the conventional streamflow gauge station of ANA is located during the period analyzed in this study (Fig. 6), it was found that the measured flow values did not exceed 90 \( \text{m}^3.\text{s}^{-1} \), i.e., approximately 24 \( \text{L.s}^{-1}.\text{km}^{-2} \). This frames the values above 30 \( \text{L.s}^{-1}.\text{km}^{-2} \) within an extrapolation zone and, therefore, with higher probabilities of estimation errors.

![Fig. 6 Measurements of water level and streamflow performed in the cross-section of the Q1 station to construct the valid RCs for the periods 2003 to 2007, 2007 to 2011 and 2011 to 2014](image)
In view of the analyses performed, it was found that the methodology that uses the RC method had, for most of the flow data considered, estimates very similar to the flow values estimated by the WB method, which reinforces the effectiveness of the method to obtain this flow measurement.

Nevertheless, it is important to mention the need for preliminary analysis of the data, seeking to remove gross inconsistencies that could greatly influence the performance of the results.

Although estimates of discrepant flows between the RC and WB methods were found in a few cases, these were proved by the statistical differences between the series and, therefore, could be corrected.

Therefore, the methodology proposed in the present study showed the potential for identifying data that may have inconsistencies related to problems resulting from both the extrapolation of the RC, assumed to be predominant in this study, and by any other factors. Other sources of errors may be associated with downstream boundary condition location and effects, changes in channel geometry, roughness variations, backwater effects and the proximity of tributaries (Quintero et al. 2021).

The great differential of the proposed methodology is to make the process of consistency of flow estimates by the RC method less laborious and more reliable, provided that it is feasible to compare the data to those from a standard method, which has less variability of hydraulic parameters, such as WB. Thus, the process of correcting inconsistent or incoherent hydrological information may be performed in a less laborious and subjective manner.

4 Conclusions

The estimated streamflow data at ANA conventional stations, based on the average between two daily readings taken at 7:00 AM and 5:00 PM, show good representation of the behavior of the streamflow in the Preto River Basin.

The daily flow values estimated by the WB method (reservoir flow) and by the rating curve method (flow at the ANA flow measuring station located downstream) present statistically similar estimates for the most part. The biggest differences found in the flows by the WB and classification curve methods occurred mainly for the maximum flows, due to the use of indirect methods for extrapolation of the classification curve.

The LR method was able to identify discrepant streamflow measurements, by considering the comparison between the WB and RC methods, so that it can be used as a methodology for consistent flow data for situations in which there is the possibility of comparing the data with those obtained by standard methods.

The study contributes to making consistent analysis less laborious and safer.

Acknowledgements The authors would like to thank the Department of Agricultural Engineering (DEA) and as the Universidade Federal de Viçosa, for supporting this research.

Author Contributions Tarcila Neves Generoso, Demetrius David da Silva, Ricardo Santos Silva Amorim, Lineu Neiva Rodrigues, Erli Pinto dos Santos participated in the creation, design, material preparation, data collection and analysis.
Funding The authors thank the Coordination of Higher Education Personnel Improvement (CAPES) – Finance Code: 001 and the National Council for Scientific and Technological Development (CNPq) for the scholarship grants.

Availability of Data and Materials Agência Nacional de Águas (ANA).

Declarations

Ethical Approval Not applicable.

Consent to Participate All authors have consented to participate in the publication process.

Consent to Publish All authors agreed to publish this manuscript.

Competing Interests The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Demetrius David da Silva reports that financial support was provided by Coordination of Higher Education Personnel Improvement (CAPES) and by the National Council for Scientific and Technological Development (CNPq).

References

Braga SR, de Oliveira MLR, Gorgens EB (2021) forestmangr: Forest mensuration and management. R package version 0.9.4. https://cran.r-project.org/package=forestmangr

Fathi MM, Awadallah AG, Awadallah NA (2021) Estimation of regional sub-daily rainfall ratios using SKATER algorithm and logistic regression. Water Resour Manag 35:555–571. https://doi.org/10.1007/s11269-020-02730-1

Firth D (1993) Bias reduction of maximum likelihood estimates. Biometrika 80(1):87–38. https://doi.org/10.1093/biomet/80.1.27

Fraga-Maia H, Santana VS (2005) Concordância de informações de adolescentes e suas mães em inquérito de saúde. Rev Saúde Púb 39(3):430–437. https://doi.org/10.1590/S0034-89102005000300014

Friedman JH, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22. https://doi.org/10.18637/JSS.V033.I01

Gharib A, Davies EGR (2021) A workflow to address pitfalls and challenges in applying machine learning models to hydrology. Adv Water Resour. https://doi.org/10.1016/j.adwres.2021.113359

Graybill FA (1976) Theory and application of the linear model. Duxbury Press, Belmont, p 704p

Ibeje AO (2018) Use of lower-order and higher-order streams in modelling the rating curve. Water Resour Manag 33:819–830. https://doi.org/10.1007/s11269-018-2145-8

Kafy A-A, Faisal A-A, Shuvo RM, Naim MNH, Sikdar MS, Chowdhury RR, Islam MA, Sarker MHS, Khan MHH, Kona MA (2021) Remote sensing approach to simulate the land use/land cover and seasonal land surface temperature change using machine learning algorithms in a fastest-growing megacity of Bangladesh. Remote Sens Appl Soc Environ 21. https://doi.org/10.1016/j.rsase.2020.100463

King G, Zeng L (2001) Logistic regression in rare events data. Polit Anal 9:137–163. https://doi.org/10.1093/oxfordjournals.pan.a004868

Kuhn M (2020) Caret: classification and regression training (6.0–86). https://cran.r-project.org/package=caret

Manfreda S (2018) On the derivation of flow rating curves in data-scarce environments. J Hydrol 562:151–154. https://doi.org/10.1016/j.jhydrol.2018.04.058

Molina XZB de la C (2016) Avaliação da acurácia do modelo de simulação a usinas individualizadas baseadas no pós-operativo. Dissertação. Universidade Estadual de Campinas, Campinas/SP, 215p

Peña-Arancibia JL, Zhang Y, Pagendam DE, Viney NR, Lerat J, Van Dijk AIJM, Vaze J, Frost AJ (2015) Streamflow rating uncertainty: characterization and impacts on model calibration and performance. Environ Model Softw 63:32–44. https://doi.org/10.1016/j.envsoft.2014.09.011

Quintero F, Rojas M, Muste M, Krajewski WF, Perez G, Johnson S, Anderson A, Hunemuller T, Cappuccio B, Zogg J (2021) Development of synthetic rating curves: Case study in Iowa. J Hydrol Eng 26(1). https://doi.org/10.1061/(ASCE)HE.1943-5584.0002022

Rodu J, Kafadar K (2021) The q-q Boxplot. J Comput Graph Stat 31(1):26–39. https://doi.org/10.1080/10618600.2021.1938586
Silge J, Chow F, Kuhn M, Wickham H (2021) rsample: General Resampling Infrastructure. R package version 0.1.0. https://cran.r-project.org/package=rsample
Strzelecka A, Kurdyś-Kujawska A, Zawadzka D (2020) Application of logistic regression models to assess household financial decisions regarding debt. Procedia Comput Sci 176:3418–3427. https://doi.org/10.1016/j.procs.2020.09.055
Team RC (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Wannasin C, Brauner CC, Uijlenhoet R, Van Verseveld WJ, Weerts AH (2021) Daily flow simulation in Thailand Part II: Unraveling effects of reservoir operation. J Hydrol Reg Stud. https://doi.org/10.1016/j.ejrh.2021.100792
Westin LGF, Conceição LR, Bortoni EC, Marcato ALM, Ribeiro CBDM, Honório LDM (2021) Evaluating the impact of streamflow rating curve precision on firm energy of hydropower plants. Water 13(8). https://doi.org/10.3390/w13081016

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

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Tarcila Neves Generoso1 · Demetrius David da Silva1 · Ricardo Santos Silva Amorim1 · Lineu Neiva Rodrigues1,2 · Erli Pinto dos Santos1

Demetrius David da Silva
demetrius@ufv.br
Ricardo Santos Silva Amorim
rsamorim@ufv.br
Lineu Neiva Rodrigues
lineu.rodrigues@embrapa.br
Erli Pinto dos Santos
erli.santos@ufv.br

1 Department of Agricultural Engineering, Universidade Federal de Viçosa, University Campus, Peter Henry Rolfs Avenue, Viçosa, Minas Gerais 36570-900, Brazil
2 Brazilian Agricultural Research Corporation (EMPRAPA Cerrados), Planaltina, DF 020 Km18, Brazil