Methodology for estimating streamflow by water balance and rating curve methods based on logistic regression

Tarcila N. Neves Generoso (✉ tarcila_neves@yahoo.com.br)
Universidade Federal de Vícosa

Demetrius David da Silva
Universidade Federal de Vícosa

Ricardo Santos Silva Amorim
Universidade Federal de Vícosa

Lineu Neiva Rodrigues
Brazilian Agricultural Research Corporation: Empresa Brasileira de Pesquisa Agropecuária

Erli Pinto dos Santos
Universidade Federal de Vícosa

Research Article

Keywords: Extrapolation, modeling, outflow, streamflow gauge station

Posted Date: April 19th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1482060/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

The water balance method is used, in Brazil, to estimate the outflow from the regularization reservoirs, based on hourly data. The rating curve method, normally adopted in conventional streamflow gauge stations of the hydrometeorological network national, the values are estimated from the average of two daily readings of the water level elevation and requires the use of extrapolation techniques to the estimation of values outside the validation limits. The objective of study was to analyze the variation of the reservoir’s hourly outflows Queimado Hydroelectric Power Plant (HPP Queimado) and to propose a method to evaluate whether the estimate of the daily outflows, obtained by the water balance method, are similar to the flow values obtained at a streamflow gauge station of the ANA network. The logistic regression model was used because it is a method that adopts binary categorical dependent variables to identify the event of interest. The results showed that the average values of streamflow obtained between the observations made from the average of 2 or 24 readings daily were close, so that the estimated data for two daily readings show good representation regarding the variability of the flows in the region. The water balance and rating curve methods showed similar statistical performance for most of the streamflow data considered. The logistic regression was able to identify atypical data in the series of analyzed flows, especially in the rainy season, which are mainly associated with the methods for extrapolation of the rating curve.

1. Introduction

Hydroelectric power plants may have several construction purposes because, according to Wannasin et al. (2021), in addition to the focus on power generation, the water accumulation reservoirs of the plants can mitigate flood and drought events by regularizing the flow, with lower flood peaks and higher minimum flows.

Cheng et al. (2019) comment that flow control in hydroelectric power plants is extremely important for several reasons, including the evaluation of performance regarding the operation, monitoring of equipment, guarantee of the conditions necessary for the preservation of a healthy aquatic environment downstream of the dam, and also to meet specific water release parameters to which they are subjected and which may vary seasonally and/or due to specific conditions (floods or droughts).

It is essential that the monitoring of operation in these structures be done appropriately, at regular intervals, to understand the behavior and, consequently, to support decision-making. In Brazil, the ANEEL/ANA Joint Resolution No. 03 of August 10, 2010, determines that, in situations where there is the presence of hydroelectric power plants with regularization reservoirs, hydrometric stations with rainfall, water level and streamflow monitoring must be automated and telemetered, so that the data collected by these stations can be recorded at a maximum interval of one hour.

In these cases, the amount of water that enters and leaves a reservoir, in a certain period of time, is calculated by the water balance method, and for the determination of outflows it is necessary to obtain
two variables: the turbined flow, removed from the reservoir by the penstock that conducts water to the turbines for the generation of energy, and the spilled flow, released by the reservoir spillways and which does not pass through the turbines and, therefore, does not generate energy (Molina, 2016).

The requirement for data collection at hourly intervals in hydroelectric power plants allows 24 readings distributed throughout the day to be performed, which increases the accuracy of daily average estimates, since it minimizes errors caused by variability of flow and also helps to understand the dynamics of the demands for electricity in the region. Thus, it is possible to evaluate the times of higher and lower water turbining, which contributes for the average daily flow readings to be more representative regarding the flow variation regime.

On the other hand, in segments where there is no presence of hydroelectric power plants, the most used method for estimating flows in watercourses is the rating curve, which is based on the measurement of data of water height or water level of the river. The construction of the rating curve or calibration curve depends on the determination of different readings of the water level, through a liminometric ruler, and the corresponding measurements of streamflow in loco, normally made with the use of a hydraulic windlass, for the purpose of determining the average velocity and the wet area survey of the cross section. This area, however, may change over time, as the characteristics of the riverbed change over the years, which means that the rating curve may lose its representativeness, hence requiring frequent field flow measurement campaigns (Manfreda, 2018).

The relation between water level and streamflow, measured at points over time to obtain the rating curve, 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 National Hydrometeorological Network (Rede Hidrometeorológica Nacional - RHN), under the responsibility of the National Agency for Water and Basic Sanitation (Agência Nacional de Águas e Saneamento Básico - ANA), the reading of water level is done daily at two times, 7 h and 17 h, and the average value of the two readings is used as representative for the purpose of average daily flow estimation.

In situations with maximum flows, however, it is common to apply methodologies for extrapolating the rating curve, since usually, due to operational problems, there are no flow measurements associated with the most critical periods of floods, which according to Grison and Kobiyama (2009) is justified by the danger for the hydrometry team to perform measurements in these situations where the flow rate is high.

Therefore, it is important to evaluate the estimated flows considering the comparison between the water balance and rating curve methods, since both are recognized and applied worldwide. Thus, the study assumed that the water balance method would have fewer limitations than the rating curve method, mainly due to the nature of extrapolations that are necessary to obtain the maximum flows in the rating curve method. As explained by Peña-Arancibia et al. (2015), there is heteroscedasticity (different variances) between the measured flow and the values estimated by the rating curves, especially in the highest flow values, for several stations, highlighting the high uncertainty potentially 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, to evaluate the daily outflow from the Queimado HPP, estimated by the water balance method, compared to the daily flows of the conventional streamflow gauge station, estimated by the rating curve method, located downstream of the said reservoir.

2. Material And Methods

2.1. Study area

The study was carried out using information related to the reservoir of the Queimado Hydroelectric Power Plant (Queimado HPP), located in the Preto River Basin (10,325 km²), sub-basin of the Paracatu River, located within the São Francisco River Basin. The Preto River Basin covers part of the Federal District (13%) and the states of Goiás (22%) and Minas Gerais (65%).

The Queimado HPP (Fig. 1), owned by the consortium between the Minas Gerais Energy Company (Companhia Energética de Minas Gerais - CEMIG) and the Brasília Energy Company (Companhia Energética de Brasília - CEB), is located in Palmital de Minas, a district belonging to the municipality of Cabeceira Grande, Minas Gerais. The Plant started operating in 2004, has a contribution area of approximately 3,657 km², its reservoir has a useful volume of 389.46 hm³, and its operation must respect the multiple uses.

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

The region has two well-defined seasons, with rainy season starting in November and ending in April, and 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 1:00 h and ending at 24:00 h, from 2004 to 2019. Due to the inconsistencies found in the series of hourly flows for 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 work with the period from 2006 to 2019.

Within the working period, days with gaps in at least one time of the day were also excluded, since the absence of one of these data could compromise the analyses, especially those whose objective was precisely to evaluate the variability of flows throughout the day.
The historical data and the rating curve of the Q1 streamflow gauge station, located immediately downstream of the Queimado HPP reservoir, were obtained from the Hidroweb platform (https://www.snirh.gov.br/hidroweb/serieshistoricas), which includes data from the RHN, under the responsibility of ANA. However, this station only showed coherent data until 2014, so the period from 2006 to 2014 was established for comparative analysis with the data of outflow from the reservoir (Qd).

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

### 2.3. Analysis of the hourly behavior of outflows

Due to the variability of the outflow from the reservoir (Qd) over the days is very large, the flow behavior over the 24-hour period was evaluated using the average monthly value of each of the times considered, for the period from 2006 to 2019, seeking to identify the times when the highest amplitudes of water turbining due to electricity production are obtained. Possible differences between the means of the 24 hours and the time of 7 h and 17 h (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 Q1/Qd ratio ≤ 0.90 were removed, since the basic premise assumed was that stations located downstream tend to have daily flow values higher than those of upstream sections, which is 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 according to the level of accuracy of the equipment and possible unidentified captures along the segment of watercourse between the two sections considered. In addition, data whose Q1/Qd ratios behaved as outliers by the monthly boxplot analysis were also eliminated. The percentage of data eliminated after considering the above-mentioned criteria was about 16%, so that 2766 data were used out of a total of 3287 available data.

Next, the data of outflow (Qd) and from the conventional streamflow gauge station of ANA (Q1), in m$^3$.s$^{-1}$, were converted to specific streamflow, in L.s$^{-1}$.km$^{-2}$ (qd and q1, respectively), in order to enable the comparison of the values obtained with the different flow estimation methods for the different drainage areas, namely: rating curve (adopted by ANA to estimate Q1) and water balance (adopted by CEMIG to estimate Qd).

In order to evaluate the methodologies more uniformly, the same time interval was considered to obtain the data, both in the outflow and in the conventional station of ANA. Thus, for this analysis, the daily specific streamflows were taken into account, obtained from the average of the values observed of outflow at 7 h and 17 h (qd$_{7-17}$), as well as the daily specific streamflows of the downstream fluvimetric station (q1), also obtained by readings at 7 h and 17 h.
Initially, an analysis of the distribution of the data of outflow and the streamflow from the conventional streamflow gauge station was performed, via quantile-quantile plot (qq-plot). This tool was proposed by Wilk and Gnanadesikan (1968) and, as explained by Nobre et al. (2017), is a plot used to observe whether two sets of data belong to the same probability distribution, in which the points are formed by the sample quantiles, so that if the points align in the result, forming a line, the two samples can be considered the same.

Next, the original flow series (qd\textsubscript{7−17} and q1) were tested to check whether or not they could be classified as 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 (\(Y_j\)) are statistically equal to the observed values (\(Y_1\)) or whether the estimates of a given variable, obtained by two methods of evaluation (\(Y_1 = Y_j\)), are statistically equal, which justifies their application in the present study.

If the estimates of \(Y_1\) and \(Y_j\) are exactly the same after fitting a linear regression, \(Y_j = \beta_0 + \beta_1 Y_1 + \epsilon\), the coefficients of the equation would be simultaneously \(\beta_0 = 0\) and \(\beta_1 = 1\), defining a line with 45°, passing through the origin.

Thus, the series of outflow data, obtained from the water balance method (qd\textsubscript{7−17}), was considered here as the standard method, with the flows estimated at the ANA station (q1), by the rating curve method, as an alternative estimation method. The Graybill F test was applied, as shown in Eq. 1, testing the Hypothesis \(H_0: \beta_0 = 0\) and \(\beta_1 = 1\).

\[
F(\text{H}_0) = \frac{(\beta - \theta) \cdot (Y_1 Y_1) (\beta - \theta)}{2 \text{RMS}}
\]

Equation 1

where:

\[
\beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}
\]

Eq. 2

\[
\theta = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

Equation 3
\[
(Y_1 Y_1) = \begin{bmatrix}
 n & \sum Y_1 \\
\sum Y_1 & \sum Y_1^2
\end{bmatrix}
\]

Equation 4

RMS is the residual mean square of the regression \(Y_j = \beta_0 + \beta_1 Y_1\). When \(F(H_0) \geq F_\alpha(2, n-2)\), the hypothesis was rejected. When \(F(H_0) < F_\alpha(2, n-2)\), the hypothesis was not rejected, assuming the similarity between the two methods evaluated, that is, \(Y_1 = Y_j\) at a significance level \(\alpha\). 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 again subjected to a new F test, now considering only the observations in which the modulus of the difference between \(q_1\) and \(qd_{7-17}\) (\(absolutedifference = |q_1 - qd_{7-17}|\)) was lower than or equal to the standard deviation of \(q_1\), in order to prove whether or not they would become statistically similar. This analysis was important so that the application of logistic regression models could be continued. According to Fernandes et al. (2020), the logistic regression model adopts binary categorical dependent variables to identify the event of interest, being a methodology that allows estimating the probability associated with its occurrence. The same authors explain that an event of interest can be classified as “1” and its absence as “0”, so that the logistic regression can be interpreted as a particular case of generalized linear models in which the dependent variable is dichotomous. In the present study, estimates that assumed value 1 were framed in a condition in which \(qd_{7-17}\) and \(q_1\) showed discrepant behavior, and those that assumed value 0 were considered as data that showed “similar” behavior.

To implement the logistic regression model, it is necessary to adopt some preliminary criteria that enable its application. Therefore, 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 logistic regression 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 logistic regression model was initially based on the value of the standard deviation of \(q_1\), previously used for filtering the flow series when the F test was applied.

The logistic regression parameters are estimated using the maximum likelihood function, which according to King and Zeng (2001) is known to suffer from some bias for small samples. The degree of bias is strongly dependent on the number of cases less frequent in both categories. 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; Fernandes et al. 2020). For this reason, it was decided to apply the maximum penalized
likelihood method proposed by Firth (1993), whose objective is 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 logistic regression modeling were verified, where the number of variables and observations to be tested were determined. Since logistic regression is very sensitive to multicollinearity in independent variables, those covariates that showed Pearson’s correlation coefficient greater than or equal to 0.95 were eliminated using 

caret

package tools (Kuhn, 2020).

Table 1 presents the number of independent variables tested, highlighting those effectively used and those discarded in the modeling after preprocessing.

<table>
<thead>
<tr>
<th>Independent Variables</th>
<th>Representation</th>
<th>Situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily average specific streamflow obtained from the ANA streamflow gauge station 42460000, considering two readings.</td>
<td>q1</td>
<td>Discarded</td>
</tr>
<tr>
<td>Daily average specific outflow from the Queimado HPP reservoir obtained considering 24 readings.</td>
<td>qd</td>
<td>Discarded</td>
</tr>
<tr>
<td>Daily average specific outflow from the Queimado HPP reservoir obtained considering 2 readings (7 and 17 h)</td>
<td>qd&lt;sub&gt;7,17&lt;/sub&gt;</td>
<td>Selected</td>
</tr>
<tr>
<td>Julian day of the year on which the observation was recorded</td>
<td>d</td>
<td>Selected</td>
</tr>
<tr>
<td>Julian month of the year in which the observation was recorded</td>
<td>m</td>
<td>Discarded</td>
</tr>
<tr>
<td>Difference between q1 and qd&lt;sub&gt;7,17&lt;/sub&gt;</td>
<td>Dif</td>
<td>Selected</td>
</tr>
</tbody>
</table>

In order to ensure that the selection of the sampling data for training the models was made with greater representativeness, considering especially the rainiest periods of the year, when the extrapolations of the rating curve probably occur, the data sets were divided in the machine learning procedures for the strategies of: training, where the hyperparameters of the penalized logistic regression model are fitted; and test, where the best fitted model is used to classify a set of data not known to it, so as to contain every month of the year so that there was no risk of fitting a biased regression model.

Thus, two strategies of data division were adopted: one performing stratified sampling, aiming to contain the same distribution of the data, with flow values between the lowest and the highest magnitudes
recorded in the series, in the training and test sets; and the second strategy performing random sampling, both with 70% of the amount of data for training and 30% for test, a division also adopted by Gharib and Davis (2021). In both approaches, the dataset was divided using features of the \textit{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 a measure that expresses the consistency or agreement of results when the measurement or examination is repeated under equal conditions (Fraga-Maia and Santana, 2005).

The Kappa coefficient evaluates the degree of agreement between the classification and the actual value of the same sample and its value can range from −1 to 1. According to Al Kafy et al. (2021), a Kappa coefficient greater than 0.61 can be considered a good fit.

The importance of each of the variables used in the modeling was also evaluated by the vip function, of the \textit{vip} package. In this case, the function analyzed the performance of the model (Accuracy and Kappa), when the variables were, through permutation, omitted.

Subsequently, the best fit model for logistic regression 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 “0”, present in the flow series, could be considered as statistically similar.

The entire process of analysis, visualization and modeling was carried out with features of the R programming language (R Core Team, 2021), using RStudio software. Modeling, training, validation and testing of the models were performed using features of the packages \textit{caret} (Kuhn, 2020) and \textit{glmnet} (Friedman, Hastie and Tibshirani, 2010), which has 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 rating curve and the rainfall data obtained in the rain gauge station (01647008), located near the site where the q1 streamflow gauge station is positioned, aiming to improve the understanding about these behaviors.

\section{3. Results And Discussion}

\subsection*{3.1. Analysis of the hourly behavior of the outflow from the Queimado HPP reservoir}
The results obtained from the monthly hourly averages for the outflow from the Queimado HPP reservoir, for the rainy season (November to April) of the historical series relative to the period from 2006 to 2019, are presented in Fig. 2.

It is possible to observe in Fig. 2 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 can be noticed in the average magnitude of the flows in some months (November, February, March and April), in the afternoon and night periods compared to the morning, but still with average differences on the order of 1 to at most 2 m$^3$.s$^{-1}$, which is not very expressive (2.5 to 5%) considering the magnitude of the regularized flows downstream (approximately 40 m$^3$.s$^{-1}$). Figure 3 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 a relatively different behavior compared to that of the flow in the rainy season (Fig. 2).

A more carefully analysis of the behavior of hourly flows in the dry period (Fig. 3), shows slight variations between the different times of the day, as in the months of May (higher values between 9 and 24 hours), June, July, August, September and October (higher values from 19 hours).

The lowest flow values occur almost predominantly 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 the value of 4 m$^3$.s$^{-1}$. Although the greatest differences in hourly flow values were found between 7 and 19 hours, approximately, a comparative evaluation of the behavior of the mean outflow over 24 hours and the outflow values calculated from the average of 7 and 17 hours, as is currently done by ANA, showed that there are almost no differences in the values for the two time intervals (mean difference of 1.1 m$^3$.s$^{-1}$), so 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 and 17 hours.

As an example, Fig. 4 proves that, except for some isolated points, there are no considerable differences in the behavior of the outflow from the reservoir when considering the averages of 24 hours of reading or the average of readings taken at 7 and 17 hours, precisely due to the insignificant change in the behavior of the flows throughout the day in the Queimado HPP.

The results indicated that probably there will be no significant problems in relation to the flow estimates in ANA stations located downstream of the Queimado HPP reservoir due to observation times (7 and 17 hours). However, it is important to evaluate whether differences can occur according to the methodology adopted to quantify the flows in the RHN, since the method used to estimate the flows (rating curve) is different from that adopted (water balance) to obtain the outflow from the water accumulation reservoirs.

### 3.2. Comparison of methods for flow estimation

Figure 5 shows the normality graph obtained by the qq-plot, where it was possible to analyze the flow data estimated by the water balance method and those of the rating curve, indicating that there is
dispersion between the values, especially when they exceed about 30 \( \text{L.s}^{-1}.\text{km}^{-2} \).

When applying the F test in the data series, it was found that the original series of \( q_{d7-17} \) and \( q_1 \), for the analysis period considered, 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 \( q_1 \) and \( q_{d7-17} \) was lower than or equal to the standard deviation of \( q_1 \), it was possible to confirm that they became statistically similar, as shown in Fig. 6.

The number of observations is 2766 in Fig. 6a and 2747 in Fig. 6b. This means that 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 logistic regression via stratified sampling and random sampling achieved good results for both the training and the test steps, as shown in Table 2.

<table>
<thead>
<tr>
<th>Steps</th>
<th>Sampling</th>
<th>Statistical Indices</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Accuracy</td>
<td>Kappa</td>
</tr>
<tr>
<td>Training</td>
<td>Stratified</td>
<td>0.994</td>
<td>0.817</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.995</td>
<td>0.789</td>
</tr>
<tr>
<td>Test</td>
<td>Stratified</td>
<td>0.992</td>
<td>0.621</td>
</tr>
<tr>
<td></td>
<td>Random</td>
<td>0.996</td>
<td>0.901</td>
</tr>
</tbody>
</table>

Table 2 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 from the training to the test phase in the stratified sampling. The sharp reduction in this type of sampling indicates the presence of overfitting and bias. On the other hand, in the modeling using random sampling, the Kappa index increased, indicating the generalization capacity of the model in this approach. In addition, the Kappa values found for the two models are greater than 0.61, being classified as good (Al 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 \( q_{d7-17} \) and \( q_1 \), the Graybill F test was applied again to the data classified as “0” in the logistic regression, and it was possible to verify effectiveness of the methodology using the two samplings, as shown in Fig. 7.

In addition to the statistical equality verified for the selected series, it can be 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, through the test to evaluate the importance of each of the variables used in the logistic regression model, for both stratified and random sampling, it was verified that the variable related to the day of the year (d) was the one that had the least importance, so that its removal would cause the lowest loss of performance in the modeling compared to the other variables tested. It stands out, however, which does not mean that its presence is not able to improve the performance of the model.

The variable related to the difference (Dif) between q1 and q_{d,7-17} was the most important for the modeling and the differences that caused greater influence on the classification decision process of the model (codes 0 or 1) occurred for flow values associated with the greatest magnitudes, that is, in the rainy season.

When evaluating only data that were considered as “attention” by the logistic regression, 14 cases were quantified for the test set when random sampling was considered and 5 cases were quantified when stratified sampling was used.

Figure 8 clearly shows the cases found by the logistic regression model with classification “1”, that is, “attention”, so that most of the incompatibilities found for the flow values occurred predominantly in the rainy season, that is, between November and April.

The highest values of q_{d,7-17} and q1 found for the period from 2006 to 2014 were estimated at 48.2 and 49.9 L.s^{-1}.km^{-2}, respectively, representing flows corresponding to 176.2 and 194.2 m^{3}.s^{-1}. Therefore, it is possible to note that, when logistic regression was performed using random sampling, the model identified situations of “attention” in q1 for low flows, that is, incompatibility between the methodologies of the rating curve and the water balance even in situations where the magnitude of the values is considered low, that is, as observed in Fig. 5, 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 rating curve.

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 (e.g. observation errors or alteration in the flow section), since the variable of greatest importance for the model was the difference between q_{d,7-17} and q1. 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 situation, and it may or may not be an 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 data of “attention” were found only for specific streamflows above 30 L.s^{-1}.
Thus, the classification capacity and behavior of a model depends significantly on the sampling set of data selected for its training.

Figure 8 also presents the analysis of the behavior of rainfall in the rain gauge station near the Q1 station, in order to check whether the discrepancies found between the flows \( q_{d7-17} \) and \( q_1 \) could or not be associated with the occurrence of rains.

As can be seen in Fig. 8, in general, rainfall cannot be considered as the main factor related to the discrepancy between the daily flow values between the segments, since in most of the cases identified there was no incidence of rainfall in the region that could influence these changes.

However, in the modeling using random sampling (Fig. 8b), in 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 rating curve in the section where the conventional streamflow gauge station of ANA is located during the period analyzed in this study (Fig. 9), it was found that the measured flow values did not exceed 90 m\(^3\).s\(^{-1}\), that is, something around 24 L.s\(^{-1}\).km\(^{-2}\), framing the values above 30 L.s\(^{-1}\).km\(^{-2}\) within an extrapolation zone and therefore with higher probabilities of estimation errors.

In view of the analyses performed, it was found that the methodology that uses the rating curve method had, for most of the flow data considered, estimates very similar to the flow values estimated by the water balance 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.

Moreover, although in few cases, estimates of discrepant flows between the rating curve and water balance methods were found, proven by the statistical differences between the series and, therefore, they can be corrected (Fig. 8).

Thus, the methodology proposed by the present study showed potential for identifying data that may have inconsistencies related to problems resulting from both the extrapolation of the rating curve, assumed to be predominant in this study, and by any other factors that should be carefully checked.

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

4. Conclusions

The variation of the hourly outflows from the Queimado HPP reservoir has low daily variability throughout the months of the year, and the greatest amplitude of the data occurs between the morning and night periods in the dry season.

The estimated streamflow data at ANA conventional stations, based on the average between two daily readings taken at 7 and 17h, show good representation of the behavior of the streamflow in the Preto river basin.

The daily flow values estimated by the water balance method (outflow from the reservoir) and by the rating curve method (flow in the streamflow gauge station of ANA, located downstream) have statistically similar estimates for the minimum and average flows, pointing to a good compatibility between the methods.

The greatest differences found in the estimation of flows by the water balance and the rating curve methods occurred mostly in the rainy season, when maximum flows occur. The differences found in this study were associated with the application of methods for extrapolation of the rating curve, since for its calibration, no measured flow values greater than 24 L.s$^{-1}$.km$^{-2}$ were verified in the watercourse section.

The logistic regression method was able to identify atypical data in the historical series of analyzed flows, considering the comparison between the water balance and rating curve methods, so it can be used as a methodology of consistency of flow data for situations in which there is the possibility of comparing the data with those obtained by a standard method.

Declarations

Acknowledgements The authors thank the Department of Agricultural Engineering (DEA) and the Center of Reference in Water Resources (CRRH), all of the Universidade Federal de Viçosa for supporting the researchers.

Funding The authors thank the Coordination of Higher Education Personnel Improvement (CAPES), award number: 001, and National Council for Scientific and Technological Development (CNPq) for the scholarship granted.

Authors Contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Tarcila Neves Generoso, Demetrius David da Silva, Ricardo Santos Silva Amorim, Lineu Neiva Rodrigues and Erli Pinto dos Santos. The first draft of the
manuscript was written by Tarcila Neves Generoso and all authors commented on previous versions of the manuscript.

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

**Ethical Approval** Authors agreed to the ethical approval needed to publish this manuscript

**Consent to Participate** Authors have consent to participate in the publication process.

**Consent to Publish** 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 financial support was provided by Coordination of Higher Education Personnel Improvement (CAPES) and by National Council for Scientific and Technological Development (CNPq).

**References**


Figures

Figure 1

Location of the study area, with emphasis on the outflow measurement section (Qd) of the Queimado reservoir and the Fazenda Limeira fluviometric station (42460000), represented by Q1
**Figure 2**

Variation of the average hourly outflow from the Queimado HPP reservoir, observed throughout the day, considering the rainy season (November to April) of the Preto River Basin.

**Figure 3**

Behavior of the average outflow from the Queimado HPP reservoir, observed for the 24 hours of the day, considering the dry season (May to October) from the Preto River basin.

**Figure 4**

Comparison of the behavior of the outflow from the Queimado HPP reservoir when considering the daily estimates from the averages of 24 readings ($Q_d$) and obtained based on the average of two readings at 7 and 17 hours ($Q_{d7-17}$) for the years 2006 (a) and 2011 (b).

**Figure 5**

Graph of normality distribution of the observed values of specific outflows at 7 and 17 hours ($q_{d7-17}$) and of the specific daily streamflows of the downstream fluvimetric station ($q_1$), also obtained by readings.
at 7 and 17 hours, for the period from 2006 to 2014

Figure 6

Scatter plots between $q_{d7-17}$ and $q_1$ ($x, y$) resulting from the Graybill F test at 1% significance level, where: (a) represents the statistically different original series and (b) represents the series after applying the criterion of filtering the samples by the standard deviation, characterized as statistically equal ($^{*}\beta_0 = 0$, $^{**}\beta_1 = 1$)

Figure 7

Graphs of statistical equality between the methods of the rating curve (used in obtaining $q_1$) and water balance (used in obtaining $q_{d7-17}$) from the identification and removal of observations classified as
“attention”, at 1% significance level, considering the stratified sampling (a) and random sampling (b) as well as the evaluation of rainfall data in a rain gauge station located near Q1.

**Figure 8**

Observations classified by the logistic regression model as “attention”, considering the stratified sampling (a) and random sampling (b) as well as the evaluation of rainfall data in a rain gauge station located near Q1.

**Figure 9**

Measurements of water level and streamflow performed in the cross-section of the ANA Q1 streamflow gauge station to construct the valid rating curves for the periods from 2003 to 2007, 2007 to 2011 and 2011 to 2014.