Theoretical and Applied Climatology
https://doi.org/10.1007/s00704-023-04459-4

RESEARCH

Estimating river discharge from rainfall satellite data through simple statistical models

Paula Birocchi[1] · Danilo Augusto Silva[1] · Dalton Kei Sasaki[1] · Marcelo Dottori[1]

Received: 8 January 2023 / Accepted: 10 April 2023

Abstract

Quantitative knowledge of river discharge measurements is essential for understanding coastal and estuarine dynamics and salinity variations. However, direct measurements are scarce for a large portion of rivers in Brazil. In this study, five simple models based on remote sensing and local rainfall datasets (MERGE) for the Ribeira de Iguape catchment are used to estimate the Valo Grande Channel (VGC) discharge on the southeastern coast of Brazil. These models use linear, quadratic, exponential, and two different multiple linear regression methods. The predicted VGC discharge time series resulting from each model is compared with the estimated time series based on in situ data from the Water and Electric Energy Department (DAEE in Portuguese). The estimated time series presented reasonable results, with skills varying from 0.84 to 0.92 and Nash–Sutcliffe efficiency (NSE) indices varying from 0.54 to 0.75, with the highest values corresponding to the multiple linear regression models. This methodology allowed us to reproduce longer time series at a daily frequency, as well as monthly averages between 2000 and 2020.

1 Introduction

The lack of measured river runoff data is evident in Brazil, and this deficiency imposes considerable limitations on the reliability of prediction systems, especially when studying estuarine regions. This kind of information is essential for constructing and calibrating numerical models for oceanic predictions (Marta-Almeida et al. 2021). As an example, an important effort was made by Carvalho et al. (2018) to construct a monthly climatology dataset of river discharges along the Brazilian continental shelf. These authors advised that future studies should focus on individual shelf regions, since near-real-time runoff data are still very rare and, when implemented, do not provide information for long periods of time.

The water volume received by a watershed and, therefore, the river runoff depends on the climatic conditions; soil characteristics; vegetation coverage; human, agricultural, and industrial activities; evapotranspiration in the capture region; and the interactions between these factors and other factors (Coleman and Wright 1971). When rainfall water does not infiltrate into the soil and runs across the land surface, it results in runoff through streams, rivers, lakes, or reservoirs (Perlman 2016). Surface runoff is a leading process in the hydrological cycle that connects precipitation to surface reservoirs (Sitterson et al. 2018). River flows balance the hydrological cycle by returning excess rainfall to the oceans and regulating how much freshwater flows through catchments (Sitterson et al. 2018). Runoff data are relevant for monitoring water resources and solving water quality and quantity problems such as flood forecasting and ecological and biological relationships within watersheds (Kokkonen et al. 2001). River discharge is also the main controller of contaminant dispersion and transport due to excessive nutrients and pesticides from agricultural lands being washed through catchments during rainy periods (Sitterson et al. 2018). The acquisition of runoff data can help water resource managers account for pollution in water resources (Sitterson et al. 2018).

In Brazil, river runoff is usually determined once a day at monitoring stations with linimetric measurements (relative height from sea level), which are converted to discharge amounts using a calibration curve. In rivers that flow into the sea, this monitoring work is carried out upstream of the headwaters of the estuary, where the movement is unidirectional and tides do not interfere with river flow. Other direct measurement methods can also be employed for this

✉ Paula Birocchi
paula.birocchi@usp.br

[1] Instituto Oceanográfico, Universidade de São Paulo, São Paulo, Brazil

Published online: 02 May 2023

purpose, such as the use of current data, the discharge of known effluents in a partial area of the water basin, or the use of semiempirical equations to determine discharge. All these methods have been thoroughly explained in past literature, for instance, in Miranda et al. (2012).

However, due to the lack of discharge data, modeling serves as an important method for predicting river discharge from rainfall data. Accurately representing rainfall both spatially and temporally is important for rainfall-runoff modeling, as rainfall is commonly one of the main model inputs (Faurès et al. 1995). In terms of their spatial structures, catchment processes in rainfall-runoff models can be divided into lumped, semidistributed, and fully distributed processes (Sitterson et al. 2018). Spatial variability is not considered in the outputs of lumped models, while semidistributed models reflect some spatial variability. Fully distributed models process spatial variability and generate runoff for each grid cell (Sitterson et al. 2018). In terms of the model structure, rainfall-runoff models may be classified into physically based, conceptual, and empirical models (Devia et al. 2015; Sitterson et al. 2018). Physically based models include explicit physical mechanisms involved in the hydrological cycle but are limited by the meteorological input data, high computational costs, and calibration challenges (Wood et al. 2011). These models are also limited by the need for a large number of parameters and are site-specific (Sitterson et al. 2018). Conceptual models are based on simplified equations that represent the water storage in a given catchment and do not consider spatial variability within the catchment (Sitterson et al. 2018). Empirical models usually consider the nonlinear relationships between inputs and outputs within a black-box concept. The best application of these models is in catchments with a lack of data, with runoff being the only required output. Such models can present highly accurate predictions with rapid run times (Sitterson et al. 2018). Some examples of these models include the regression equations and machine learning models involved in artificial and deep neural networks (Devia et al. 2015; Sitterson et al. 2018), and Artificial Neural Networks (ANNs), Fuzzy Logic, and Genetic Algorithm (GA) (Dwarakish and Ganasri, 2015). Empirical relationships were also used to estimate river discharge from satellite data in the Yangtze River (China) (Sichangi et al. 2018), and from precipitation and temperature data in the Colorado River (USA) (Vano and Lettenmaier 2014).

As worldwide examples, we can cite some rainfall-runoff models, such as the conceptual and semidistributed TOPMODEL (Topography Based Hydrological Model) (Devia et al. 2015) and HBV (Hydrologiska Byrans Vattenavdelning) (Bergstrom 1976; Devia et al. 2015) models and the complex physically based SWAT (Soil and Water Assessment Tool) and MIKE SHE (Systeme Hydrologique European) models (Devia et al. 2015). The MGB-IPH rainfall-runoff model (Collischonn 2007) has been used for large-scale basins in South America (Allasia et al. 2006). When comparing the model results with the data collected at riverine gauging stations in the Taquari-Antas basin, Collischonn et al. (2007) found Nash–Sutcliffe efficiency (NSE) coefficient values varying from 0.40 to 0.84. The FFBP rainfall-runoff empirical model was applied in the Liebien River (Taiwan) presenting a R^2 score value of 0.97 (Chen et al. 2013). Najafi and Moradkhani (2016) used model combinations with empirical relationships to forecast the discharge of four rivers and obtained NSE coefficient values varying from 0.50 to 0.96. A comparison between three empirical models was made by Belvederesi et al. (2020), including a regression model, for which the authors obtained NSE values from 0.65 to 0.78. Three empirical models were also applied by Sahoo et al. (2019) to forecast low flows in three stations in the Mahanadi river basin (India). The authors obtained NSE coefficient values from 0.56 to 0.97, depending on the region and model. The NSE coefficient measures the efficiency E proposed by Nash and Sutcliffe (1970) as one minus the sum of the absolute squared difference between the modeled and observed values normalized by the variance of the observed values during the investigated period. Model simulations can be judged as satisfactory if the NSE value is higher than 0.50 (Moriasi et al., 2007).

In Brazil, the Storm Water Management Model (SWMM) (Rossman et al. 2010) has been applied for the drainage area of Belo Horizonte, Minas Gerais, producing an average Nash–Sutcliffe coefficient of 0.72 (Rosa et al. 2020). The performance of the SWMM was good, but its application required not only rainfall data but also the topographic data and land use map of the region and was thus a much more complex model than the empirical model that uses only rainfall data as the input. Another example is the probability-distributed model (PDM, Moore 2007), which has been applied to 5 basins in the southeastern region of Brazil. The input data required for this model are rainfall and potential evapotranspiration, and the model uses a specific and complex formulation.

The majority of the models exemplified above use additional physical parameters beyond rainfall as inputs and are therefore considered complex, as they take into account the hydrological processes of the studied catchment. Here, we propose the application of regression equations in our models, using rainfall as the single input with the objective of obtaining runoff. We apply these regression models to an important estuarine region with scarce freshwater discharge measurements along the Brazilian coast: the Cananéia-Iguape estuarine lagoon system (CIELS, hereafter).

The CIELS spreads for more than 70 km along the coast of the state of São Paulo, hosts relatively large populations of cetaceans (Santos and Rosso 2007; Geise et al. 1999; Filla et al. 2012) and fishes (Barcellini et al. 2013, Curcho et al. 2009), and is a reproduction area for several other organisms (Barioto et al. 2017; Stanski et al. 2018; Bochini et al. 2019; Galvão et al. 2000). However, considering the ecological importance of this system, the freshwater runoff of the Valo Grande Channel (VGC), the main freshwater source for the CIELS, has been poorly measured over time. As an example of runoff measurements obtained in the CIELS, Bérgamo (2000) presented monthly discharge estimates using the river input data from the drainage basins of the Ribeira de Iguape River and next to Cananéia. The author concluded that the seasonal freshwater variations in the systems of both drainage basins followed the seasonal fluctuations in rainfall in the region. Additionally, in 1955 and 1965, the currents in the VGC and the hourly instantaneous discharge in a complete tidal cycle (12h25) were measured (GEOBRÁS 1966). Estimates of the average daily discharge of the VGC region in a 12-year period (1954 to 1965) were also produced based on discharge data collected from the Três Barras station along Ribeira de Iguape River, located upstream of the VGC (GEOBRÁS 1966).

In the present study, rainfall data are applied as the unique fully distributed inputs from the Ribeira de Iguape watershed in empirical and simplified statistical models to estimate the lumped time-series discharge of the VGC. A particular advantage of these statistical models is that limited hydrological data (except rainfall and runoff data) are demanded without considering the other physical variables representing the hydrological process of the studied catchment, such as topography, land use maps, or evaporation.

The scientific question addressed herein is as follows: what is the performance of a simple statistical model based on rainfall in predicting the discharge of a water basin? We hypothesize that simple statistical models that use the rainfall data from MERGE, which combines satellite-derived and local precipitation data, are capable of generating daily discharge time series with a good accuracy. Such detailed runoff outputs would improve the quality of CIELS numerical simulations aiming to reproduce its physical characteristics, such as currents and salinity, and would have application potential for other estuarine systems, as well as supporting other research subjects with this essential information. The model results can also support decision-making in the areas of water resource planning and management. In addition, these results can assist urban planners and managers in undertaking the necessary measures to address extreme high-flow predictions.

In the next section, we describe the main physical characteristics of the CIELS, followed by a description of the methodology in Section 2. The results and discussion are explained in Sections 3 and 4, respectively, and we present the conclusion in Section 5.

1.1 Study area

The Ribeira de Iguape watershed, the main contributor to the VGC, occupies the southeastern portion of São Paulo state and the eastern portion of Paraná state between the latitudes of 23°50′ and 25°30′ S and longitudes of 46°50′ and 50° 00′ W. Currently, the Ribeira de Iguape basin is an unhindered catchment with no constructed dams (CBH-RB 2008; IBAMA 2016). The catchment covers a total area of 24.980 km^2 inside São Paulo and Paraná states (DAEE 1998). In the Iguape region, part of the discharge of the Ribeira de Iguape River is diverted to Mar Pequeno through the VGC, which is an artificial connection constructed between 1828 and 1830 (Moraes 1997). The channel originally presented a width of approximately 5 m, but erosion over time has enlarged it to a 250-m width.

The southern coast of the São Paulo climate is defined as tropical and humid with rainfall related to the seasons, with wet summers and dry winters (Ma et al. 2011). This region is influenced by the South Atlantic Convergence Zone (ZCAS), which is a semipermanent feature characterized by a NW–SE-oriented band of condensation and nebulosity (Satyamurti et al. 1998). This feature is responsible for most of the precipitation in South America during summertime (Ma et al. 2011) and can occur in the Ribeira de Iguape watershed region. The passage of cold front systems also influences the precipitation regime in this area, mostly occurring during wintertime (Stech and Lorenzetti 1992). Frontal systems associated with precipitation lead to the absence of a true dry season in the region, generating relatively well-distributed precipitation throughout the year, with the highest and lowest rainfall measured in summer and winter, respectively (GEOBRÁS 1966). The seasonal rainfall variability in southeastern Brazil can cause negative environmental, social, and economic impacts due to anomalous precipitation and water availability (Zhang et al. 2018). During 2014 and 2015, this region suffered one of the most severe droughts since 1960, leading to insufficient hydroelectric power generation throughout the entire country and the depletion of water reservoirs in the metropolitan region of São Paulo (Zhang et al. 2018).

In the Iguape region, relative humidity presents values greater than 70% throughout the year, with an annual mean rainfall of 1555 mm (GEOBRÁS 1966). In the Cananéia-Iguape estuary, the annual average precipitation is 2200 mm, with a maximum monthly average precipitation of 266.9 mm occurring between January and March (Oliveira et al. 2009) and a minimum monthly average precipitation of 95.3 mm occurring in July and August (Oliveira et al. 2009).

The Ribeira de Iguape River is the main tributary discharging freshwater into the VGC (Afonso 2006; Marta-Almeida et al. 2021). The freshwater discharge of this river presented daily mean values between 84 and 1.601 m^3 s^{-1} from 1954 to 1965 (GEOBRÁS 1966). The monthly mean discharge values in the VGC range from 240 m^3 s^{-1} in August to 550 m^3 s^{-1} in February (Ambrosio 2016).

2 Methods

2.1 A daily precipitation dataset: MERGE

To evaluate the VGC discharge, we used the MERGE precipitation dataset (Rozante et al. 2010). This dataset is produced and distributed in Brazil by the Center for Weather Forecasting and Climate Studies (CPTEC) at the National Institute for Space Research (INPE). MERGE applies a Barnes objective analysis (Barnes 1973), combining data from meteorological stations distributed over the Brazilian territory and satellite precipitation data (Rozante et al. 2010). The daily rainfall data from MERGE used in this study have a spatial resolution of 0.1° in the Ribeira do Iguape watershed (Fig. 2) and span the period from June 2000 to December 2020. The watershed comprises 224 MERGE grid points (Fig. 2).

2.2 Valo Grande Channel discharge estimate

The Três Barras mouth (Fig. 1) is the location where the Ribeira de Iguape River bifurcates in the VGC and continues to flow through the other branch. The VGC total discharge results from approximately 75% of the total contribution of the Três Barras mouth (GEOBRÁS 1966), corresponding to approximately 95.6% of the main drainage represented by the discharge of three main affluents, namely, the Jacupiranga, Ribeira do Iguape, and Pariquera-Açu rivers (Fig. 1).

To estimate the VGC discharge and to train and evaluate the statistical models (Section 2.3), we applied a series of approximations to represent the best possible estimates without direct measurements. Data from the Water and Electric Energy Department (DAEE in Portuguese) database collected from 2011 to 2019 were used, representing the most recent available measurements of the Jacupiranga (Q_J) and Ribeira do Iguape (Q_R) fluviometric stations (Fig. 1) discharges. Gaps in the time series ranged from 2 to 5 days and were filled using linear interpolation. It is important to mention that Q_R was measured in the city of Registro (Fig. 1), which is located approximately 70 km from the connection to the Três Barras mouth, resulting in a lag of approximately 36 h between the measurements taken at Registro and at this location (Pisetta 2006; Pisetta 2010). As the discharge data at the fluviometric station are daily means, we assumed that the Ribeira de Iguape River discharge values (Q_R) corresponded to the measurements obtained on the previous day at the Jacupiranga River (Q_J) to calculate the Três Barras discharge (Q_{TB}). This same method was applied by Pisetta (2006, 2010), thus introducing a 24-h lag. As an example, for the calculation of Q_{TB} on 14 April, we used discharge values from 13 April for Q_R and from 14 April for Q_J. Finally, we estimated the Pariquera-Açú River discharge (Q_{PA}) by adding 10% of the Jacupiranga River discharge, which, on average, corresponded with the measurements well (GEOBRÁS 1966).

Q_{TB} was computed using the following equation, which includes all three discharge measurements:

$$Q_{TB} = Q_R + 1.1\ Q_J$$

Notably, Q_{PA} is embedded in the Q_J coefficient.

Since the VGC discharge is approximately 75% of Q_{TB}, we estimated the Valo Grande artificial channel discharge (Q_{VGC}) as follows:

$$Q_{VGC} = 0.75\ Q_{TB}$$

2.3 Models

Five different models were implemented: linear regression, quadratic regression, exponential regression, and two distinct multiple linear regression (MLR) models. These models were developed using the Python scikit-learn library (Pedregosa et al. 2011) by applying the ordinary least squares (OLS) linear regression method.

We randomly trained the models using 67% of the rainfall data from MERGE and the VGC discharge data as our training subsets. The other portion (33%) was used as the test subset to validate these models. During the training stage, we shuffled the dataset input to each model 2000 times and selected the moment where the root mean squared error (RMSE) was minimized. In these cases, the shuffling ranged from 0 to 2000, and the minimum RMSE was achieved at 1506 shuffling times for the linear, quadratic, and exponential models. This method was utilized to obtain a set containing 67% of the total time series that presented the minimum error possible. This is inspired in standard procedure of more advanced machine learning models, where the training algorithm adjusts a model coefficient by iterating repeatedly the training dataset through the model. The improvement comes from reducing the loss values over the iterations, due to improved coefficient values. Since we cannot readjust a multiple linear regression by retraining the algorithm, we instead shuffled the training dataset in a reproducible way in order to get a training dataset that provides a model with good performance in both training and test data. As both evaluations had similar skill, we can affirm the model produces good results without overfitting.

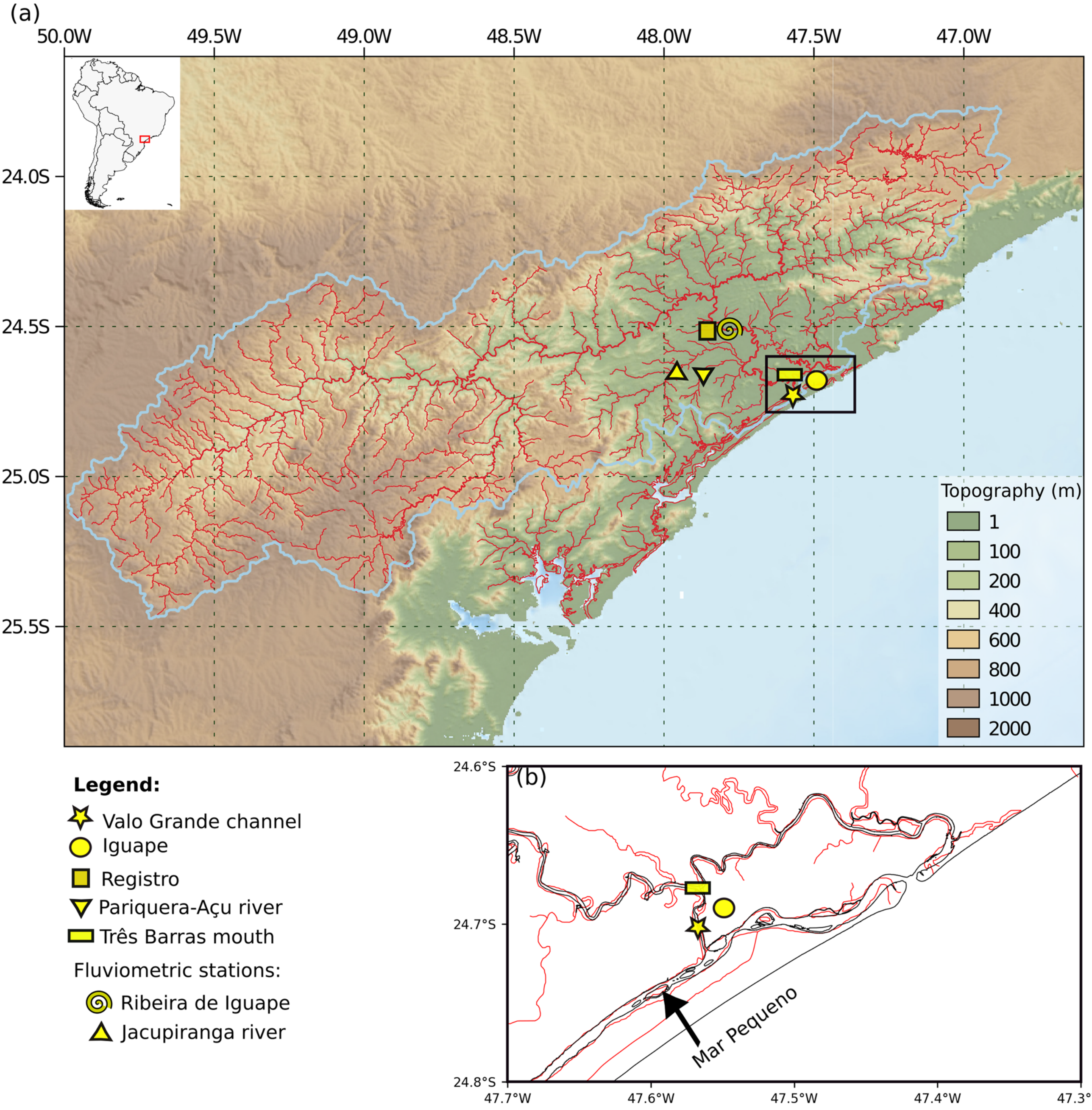

Fig. 1 (**a**) Topography of the study area, including the Ribeira de Iguape watershed boundary (blue contour). The thin red lines represent the rivers in the region. (**b**) Magnified view of the Valo Grande Channel region. The coastline is presented with a black line

After splitting the data into training and validation portions, the data were linearized using the corresponding methods in the quadratic and exponential models. For the linearization of these models, we fitted the 67% training portion of the VGC discharge data into quadratic and exponential functions using the root-mean square and log, respectively. This process generated transformed VGC discharge values as outputs, which were then used for the linear regression.

The RMSE, skill (Willmott 1981), and coefficient of determination (CD) (or R^2 score) were computed for the models. We used the test subset and observed VGC discharge values to calculate these parameters, and from now on, we will call this process the test subset validation step. The CD is the percentage of variation that is described by the linear regression line. This term represents the proportion of variation in the dependent variable (discharge) that

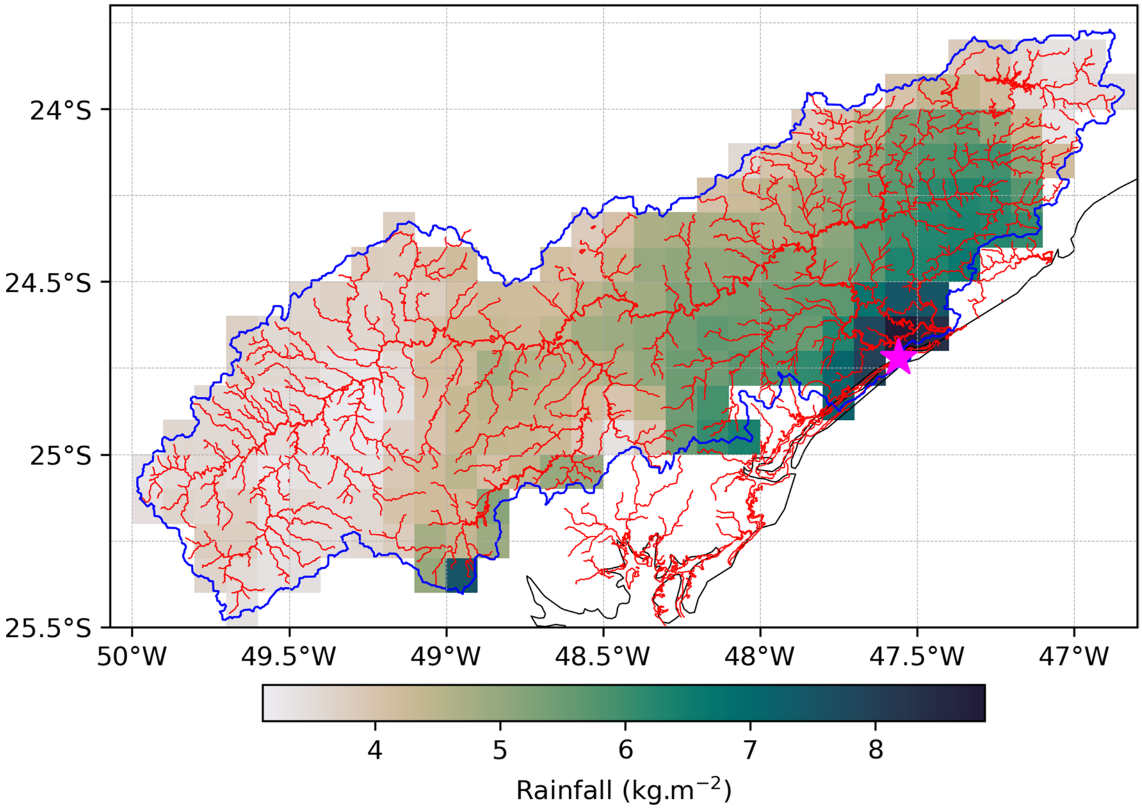

Fig. 2 Mean rainfall data available from each MERGE cell, totalizing 224 grid points, from June 2000 to December 2020 for the Ribeira do Iguape watershed. The blue contour line indicates the watershed boundaries and the magenta star indicates the VGC location

is predictable from the independent variables (rainfall data points from MERGE) and is defined as follows:

$$\mathrm{CD} = 1 - \left(\mathrm{SE}_{\mathrm{line}}/\mathrm{SE}_{\mathrm{d}}\right)$$

where $\mathrm{SE}_{\mathrm{line}}$ is the total squared error between the data points and the line fitted by the linear regression model, representing the distance between the data and regression line, and SE_{d} is the total variation in the predicted discharge. The CD value varies from 0 to 1, where 1 indicates a perfect fit of the model with the data.

For the linear, quadratic, and exponential regression models, the VGC discharge was estimated as a function of the average rainfall (R) over the area and can be expressed as follows:

$$Q_{\mathrm{VGC}} \ = a + b * R \ \text{(linear)}$$

$$Q_{\mathrm{VGC}} = a * e^{b} * \mathrm{R} \ \text{(exponential)}$$

$$Q_{\mathrm{VGC}} = a + b * (R)2 \ \text{(quadratic)}$$

where a is an offset or a constant linear coefficient and b is the linear, quadratic, or exponential coefficient, depending on the model, determined by the least square fitting method mentioned above.

The rainfall (R) values used in the linear, quadratic, and exponential models were computed as follows:

$$R = \left(\sum\nolimits_{n=1}^{224} R_n \mathrm{FT}\right)/n$$

where n is the number of grid points (224) from MERGE and $R_{\ n}^{\ \mathrm{FT}}$ is the accumulated rainfall over a specific and fixed period of time on each grid point. This time was chosen based on the maximum correlation found between accumulated precipitation and runoff, as shown in Section 3.1.

The other two models, which are classified as MLR models, predict the value of one dependent variable based on two or more independent variables (Shoaib et al. 2018). In these models, we used the 224 grid points of accumulated rainfall from the MERGE dataset and treated them as independent variables. They cover the entire Ribeira de Iguape watershed area (Fig. 2). Each pixel (or grid point) is associated with a time series of rain, which was considered as a variable in the multiple linear regression model. Spatial grid points with missing values were removed. We considered two MLR models for comparison. The first one considered the accumulated rainfall in a fixed period of time for every grid point (R^{FT}) according to the highest correlation found between the spatial-mean MERGE rainfall value and the VGC discharge. In this case, the equation used to predict the VGC discharge is given as follows:

$$Q_{\mathrm{VGC}} \ = a + \sum_{n=1}^{224} \left(b_n * R^{\mathrm{FT}}\ n\right)$$

Note that, unlike the linear model, which has a single linear coefficient for the whole domain, in this model, each grid point has a specific linear coefficient (b_n). However, the period of time over which the accumulated rainfall was computed was the same for every grid point. This specific time was the same as that applied for the first three models, as presented in Section 3.1.

The second MLR model was very similar to the first but also considered a different period of time to compute the accumulated rainfall at each grid point, also based on the highest correlation obtained between the rainfall and discharge at each grid point. In this case, the equation used to estimate the VGC discharge is given as follows:

$$Q_{\mathrm{VGC}} \ = a + \sum_{n=1}^{224} \ \left(b_n * R^{VT}\ n\right)$$

where R^{VT} (to avoid confusion with R^{FT}) is the accumulated rainfall at each grid point computed in different time periods. These specific time periods were chosen according to the maximum correlation between the accumulated rainfall and runoff and are presented in Fig. 6.

The minimum RMSEs were achieved at shuffling times equal to 736 and 1513 for the MLR models using R^{FT} and R^{VT}, respectively. For the multiple regression models, we also applied the statsmodel library (Seabold et al. 2010) to obtain the spatial distributions of the correlation, p-value, and standard error associated with the OLS method. The standard error represents the average distance that the observed values fall from the regression line. The p-value is a measurement of how likely a coefficient is to be calculated through our model by chance (McAleer 2020). For example, a p-value of 0.378 indicates that there is a 37.8% chance that the independent variable (rainfall) has no effect on the dependent variable (VGC discharge), and our results are produced by chance.

To quantitatively compare the MERGE rainfall data with the DAEE discharge data and generate the model analysis results, we transformed the MERGE data from their original unit, kg m^{-2} day^{-1}, to m^{3} s^{-1} (the same unit used for the discharge data). We converted the unit of rainfall MERGE data considering the area of the MERGE grid cells and a complete day (24 h, which was converted to seconds), so we multiplied it by the grid cell area and divided it by the water density, equal to 1000 kg m^{-3}.

We obtained the predicted VGC discharge time series from each of the models. Then, we compared these predictions with the time series estimated based on data collected from DAEE using the method proposed by GEOBRÁS (1966) and described in Section 2.2. This process, defined here as the time series validation process, compares the modeled and observed VGC discharge time series by calculating the skill score, RMSE, CD, Pearson's correlation coefficient, and Nash–Sutcliffe efficiency (NSE) index (Nash and Sutcliffe 1970).

In general, the MERGE rainfall data is organized in a set of grid cells covering the Ribeira de Iguape watershed which drains to the Valo Grande channel and then into the South Atlantic Ocean. The methodology assigns each grid point as an independent variable in the multiple regression model, with the discharge estimated at the mouth of the VGC from combined estimates of discharge at three rivers located inland—the Jacupiranga, the Ribeira de Iquape, and the Pariquera-Acu rivers. The regression coefficients developed for each grid point assign the fraction of discharge at the VGC associated with that grid cell, thus providing a model of the VGC discharge, the main tributary of the studied watershed.

This study demonstrates a method that uses distributed rainfall data to predict the Valo Grande channel discharge, which has limited river gauge information. Considering the same period of MERGE data, there were two considerable gaps in Q_{R} available data, the first from 2000 to 2003 and the second from 2006 to 2010, showing that missing data is recurrent for river discharge in the region. Consequently, our Q_{VGC} estimate started in 2011. The main goal of our analysis was to provide a complete time series of river discharge estimation for the Valo Grande channel covering possible gaps and scarcity of river discharge data. In order to accomplish this goal, we used a free and time continuous (from 2000) available rainfall data product (MERGE).

3 Results

3.1 Accumulated rainfall (R)

The accumulated rainfall (R) is defined as the sum of precipitation over a given period and a given area. R^7 is presented here as the accumulated rainfall over 7 days, as this term presented the maximum correlation with the discharge data in the correlation analysis, as explained above. Henceforth, R^7 corresponds to the accumulated rainfall fixed on time (R^{FT}). Figure 3 shows the time series of discharge and R^7.

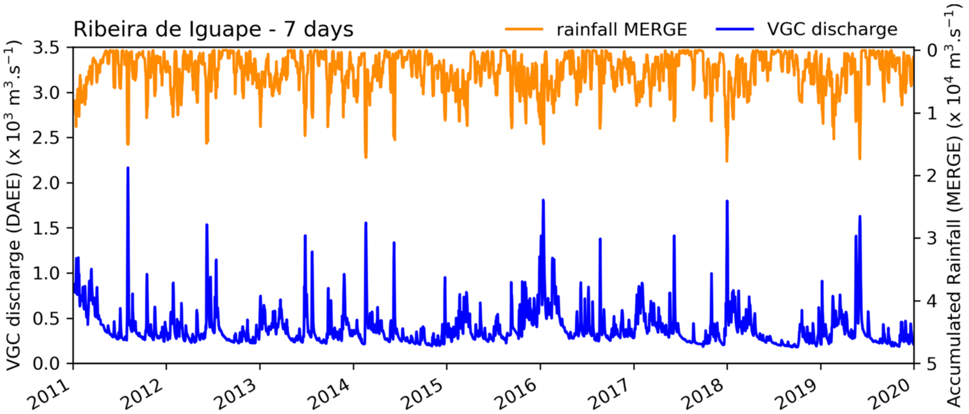

Fig. 3 R^7 (in orange) and discharge series for Valo Grande (in blue) estimated by the GEOBRÁS (1966) method. Note that the precipitation time series is vertically inverted

This quantity presents a correlation of 0.735 and p-value < 0.01 with the discharge data estimated for the Valo Grande Channel using the method described in Section 2.2 (Fig. S1, available in the Electronic Supplementary Material).

The river discharge peaks coincided with the rainfall peaks, indicating the influence of rainfall in modulating discharge (Fig. 3). For instance, intense rainfall occurred at the beginning of 2011, reaching rates over 10,000 m^3 s^{-1} before decreasing progressively until June. This same pattern was observed in the VGC discharge, with a peak flux of almost 1200 m^3 s^{-1} at the beginning of 2011 and a decreasing flux until the end of June. Interestingly, all rainfall peaks approximately coincided with the river discharge peaks, showing the coherence between the two datasets.

3.2 Regression models

In this section, we present the model results. We focus on the MLR models, as they were considerably more reliable than the linear, quadratic, and exponential models. Details about the statistical differences among these models are presented in Section 3.3, and the scientific reasons and explanation of these differences are presented in Section 4. The detailed results of the linear, quadratic, and exponential models are available in the Electronic Supplementary Material (Figs. S2 and S3).

3.2.1 Multiple linear regression models

We present two different MLR models. As we found the highest correlation between discharge and accumulated rainfall in 7 days (R^7), in the first multiple regression model, we used MERGE R^7. In the second one, we used the accumulated rainfall varying in time (R^{VT}), depending on the highest correlation for each grid cell of MERGE.

The multiple regression model using R^7 resulted in minimum, maximum, and mean standard errors of 0.003, 0.064, and 0.009, respectively. In 99% of the MERGE grid cells, the standard errors were lower than 0.03, except in the three grid cells in the southern region of the catchment (Fig. S4-a). In general, the errors associated with the regression results were low in almost the entire domain. The p-values were lower than 0.05 for 28% of the grid points (63 grid points of 224 in total) (Fig. S4 -b).

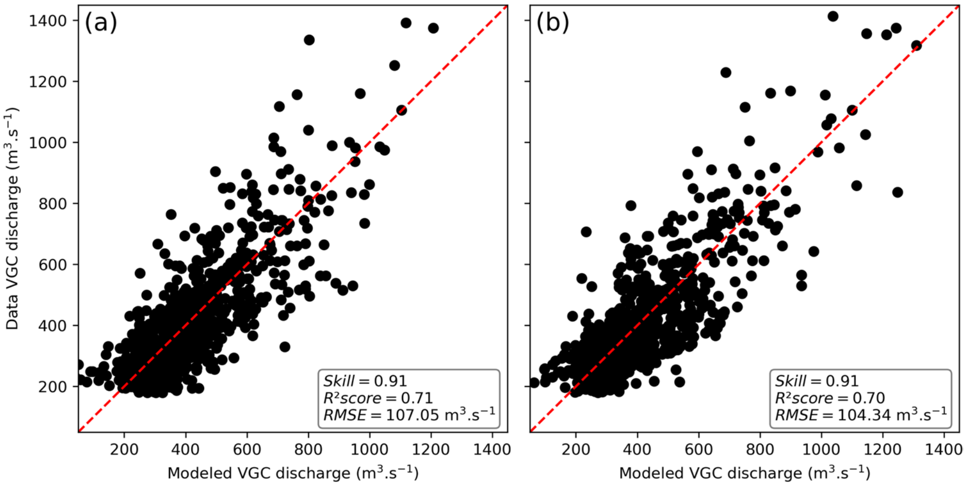

Fig. 4 VGC discharge results of the multiple regression models using R^7 (**a**) and R^{VT} (**b**) for the test subset validation. The red lines show the perfect fit between the model results and data

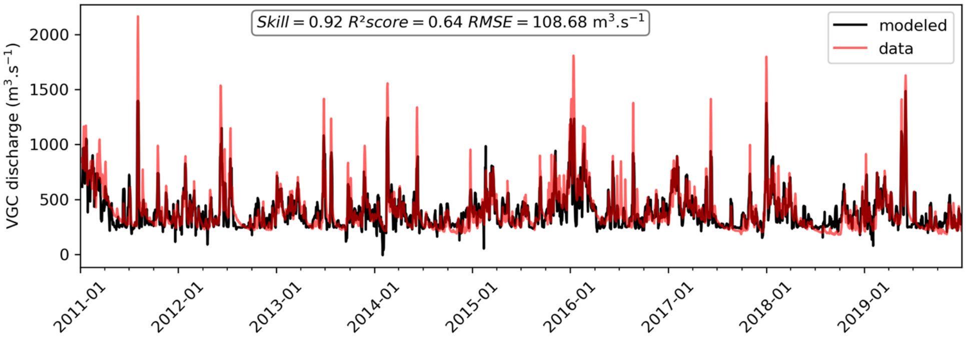

Fig. 5 VGC discharge time series produced by the MLR model using R^7 (in black) and estimated by the method proposed by GEOBRÁS (1966) (in red)

The black circles in Fig. 4 represent each value included in the test subset validation of the multiple regression models using R^7 (Fig. 4a) and R^{VT} (Fig. 4b). The multiple regression model using R^7 presented a skill of 0.91, R² score equal to 0.71, and RMSE of 107.05 m^3 s^{-1} (Fig. 4a). The multiple regression model with R^{VT} presented a skill of 0.91, CD of 0.70, and RMSE of 104.34 m^3 s^{-1} (Fig. 4b). The validation of the models showed that both R^7 and R^{VT} were able to reproduce the discharge data considerably well for both low and high values (Fig. 4a and b, respectively).

The time-series validation of the multiple regression model using R^7 showed that the model represented the discharge reasonably well (Fig. 5), presenting a skill of 0.92, CD of 0.64, and RMSE of 108.68 m^3 s^{-1}. The modeled time series reflected the main discharge patterns, including the peaks observed in January and June, as well as the lowest values present in the observations. Notably, the seasonal patterns were also present in the reconstructed time series (Fig. 5).

The best model results were produced by the R^{VT}-considering model, using the highest correlation coefficient between each rainfall grid cell of MERGE and the discharge time series. The maximum correlation coefficients (Fig. 6a) were found between 6 and 9 days of accumulated rainfall (called R^6 and R^9, respectively) (Fig. 6b).

The average standard error of R^{VT} was 0.009. For all MERGE grid cells, the standard error was lower than 0.03, with the exception of two grid cells in the southern region of the watershed, showing that the errors associated with the regression were low (Fig. S5-a). The *p*-values were lower than 0.05 for 33% of the grid points, or 73 grid points (Fig. S5-b).

For the time-series validation, we found a skill of 0.92, CD of 0.67, and RMSE of 106.35 m^3 s^{-1} (Fig. 7). The modeled time series reflected both the main patterns of and seasonal variability in the data (Fig. 7), similar to the previously tested multiple regression model (Fig. 5).

The results obtained for both multiple regression models were similar. The modeled VGC discharge was better represented in the multiple regression estimates (Fig. 4) than in the linear, quadratic, and exponential simulations (Figs. S2 and S3). All models underestimated discharge values higher than 1000 m^3 s^{-1}, with the best results provided by the MLR models.

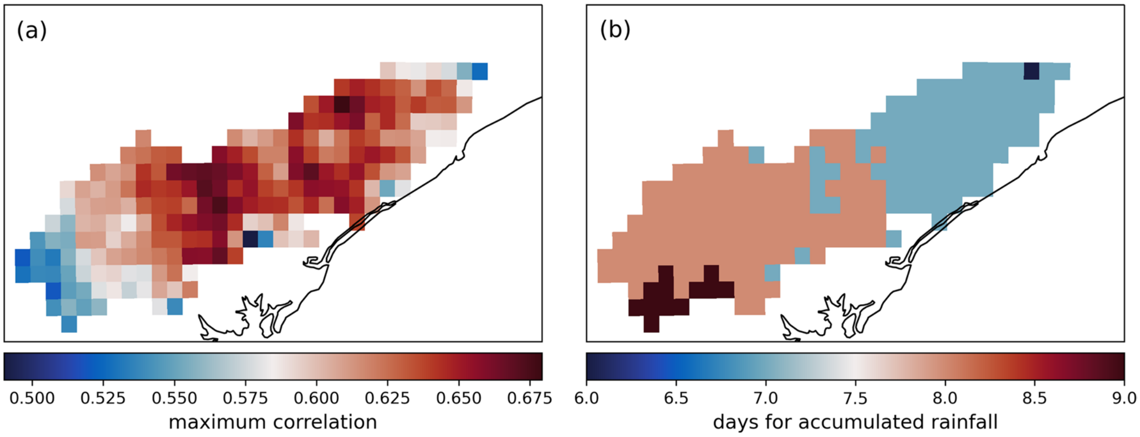

Fig. 6 (**a**) Maximum correlation coefficient at each grid cell using *R* on *X* days, according to panel (b). (**b**) Number of days of accumulated rainfall used to detect the highest correlation coefficients in panel (a)

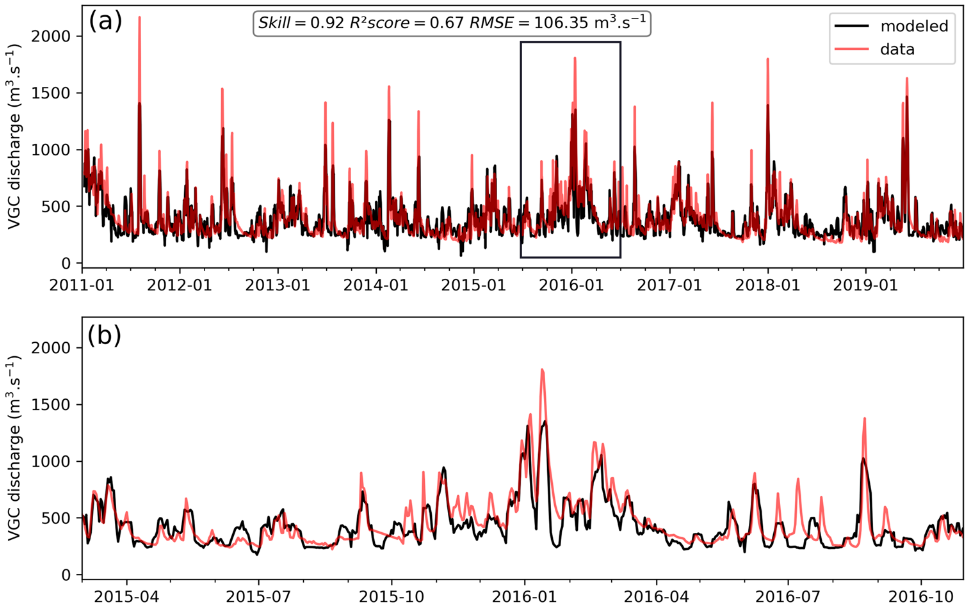

Fig. 7 (**a**) VGC discharge time series provided by the multiple linear regression model using accumulated rainfall R^{VT} (in black) and estimated by the method proposed by GEOBRÁS (1966) (in red). (**b**) Magnified view for the period indicated by the black box in (a). The multiple regression model using R^{VT} presented a skill of 0.92, CD of 0.67, and RMSE of 106.35 m^3 s^{-1} and therefore satisfactorily represented the data

3.3 Model comparison

In this section, we evaluate all model results using the test subset and time-series validations. Table 1 shows the skill, coefficient of determination (CD), and RMSE values derived in the test subset validation. Table 2 shows the results of the time-series validation for all models, including the skill, coefficient of determination, RMSE, Pearson's correlation coefficient, and the Nash–Sutcliffe efficiency (NSE) values.

In the test subset validation, the skill values were above 0.8 in all cases, with the highest values corresponding to 0.91 for the MLR models (Table 1). The CDs of the linear, quadratic, and exponential models showed values of approximately 0.5, and the RMSE values were between approximately 122 m^3 s^{-1} and 125 m^3 s^{-1} (Table 1). For the test subset validation of the MLR models, the CD values were equal to 0.71 and 0.70, and the RMSE values were equal to 107.05 m^3 s^{-1} and 104.34 m^3 s^{-1} for the R^7 and R^{VT} models, respectively (Table 1).

From the time-series validation results, we found a skill of approximately 0.83, CDs between 0.54 and 0.57, RMSEs from 136 to 142 m^3 s^{-1}, and Pearson's correlation coefficients between 0.73 and 0.77 for the linear, quadratic, and exponential models (Table 2). For both MLR models, we found a skill of 0.92 for the time-series validation (Table 2). In this validation, the CD values were equal to 0.64 and 0.67, the RMSE values were 108.68 m^3 s^{-1} and 106.35 m^3 s^{-1}, and Pearson's correlation coefficients were 0.85 and 0.86 for the MLR models using R^7 and R^{VT}, respectively (Table 2). Pearson's correlation coefficients of the 5 models presented *p*-values lower than 0.01.

Table 1 Skill, coefficient of determination (CD), and RMSE (m^3 s^{-1}) values derived in the test subset validations of the linear, quadratic, and exponential models

Model	Skill	CD	RMSE (m^3 s^{-1})
Linear	0.83	0.51	125.26
Exponential	0.81	0.52	124.26
Quadratic	0.82	0.53	122.66
Multiple linear regression (R^7)	0.91	0.71	107.05
Multiple linear regression (R^{VT})	0.91	0.70	104.34

The NSE index presented an improvement of up to 28% for the multiple regression models when compared to the linear, quadratic, and exponential models. The highest NSE index was equal to 0.75, which was found for the multiple regression model using R^{VT} (Table 2). The multiple regression model using R^7 presented an NSE index of 0.74, while the linear, quadratic, and exponential models presented NSE indices equal to 0.54, 0.56, and 0.57, respectively (Table 2).

Figure 8 clearly shows that all of the regression models underestimate high flows, and either underestimate or overestimate the extreme low flows. However, statistical analysis in this study still shows that these models present very good estimates, even for extreme events with differences, in these cases, that do not exceed 22%. Comparing the QQplot derived for each model, the data quantiles were better reproduced by the MLR models considering R^7 and R^{VT} (Fig. 8d and e, respectively) than by the linear, quadratic, and exponential models (Fig. 8a, b, and c, respectively). The linear, quadratic, and exponential models were able to effectively represent VGC discharge values between 250 and 600 m^3 s^{-1}, with the model results closely coinciding with the perfect-fit lines (in red) in the QQ plots (Fig. 8a, b, and c). Increased river

Table 2 Results of the time-series validation. The skill score, coefficient of determination (CD), RMSE (m^3 s^{-1}), and Pearson's correlation coefficient values between each model result and the discharge time series estimated for the Valo Grande using the method proposed by GEOBRÁS (1966) from 2011 to 2019 are shown

Model	Skill	CD	RMSE (m^3 s^{-1})	Pearson's correlation and *p*-value	Nash–Sutcliffe efficiency (NSE) index
Linear	0.84	0.54	141.61	0.73 – <0.01	0.54
Exponential	0.84	0.57	136.38	0.77 – <0.01	0.57
Quadratic	0.83	0.56	137.63	0.76 – <0.01	0.56
Multiple linear regression (R^7)	0.92	0.64	108.68	0.85 – <0.01	0.74
Multiple linear regression (R^{VT})	0.92	0.67	106.35	0.86 – <0.01	0.75

discharge (>600 m^3 s^{-1}) is associated with a degradation of the results of these models, as the models tended to underestimate peak values, presenting RMSEs varying from 136 to 141 m^3 s^{-1} at discharge values above 600 m^3 s^{-1}. This limitation was more evident in the comparison between the modeled and observed VGC discharge time series (Figs. S3) but was reduced in the multiple regression models (Fig. 8d and e), which exhibited better agreement between the model results and data during high-flow events above 600 m^3 s^{-1}, with RMSE values of 106.30 m^3 s^{-1} and 103.70 m^3 s^{-1} for the MLR models considering R^7 and R^{VT}, respectively. These results indicate improvements between 22 and 27% for the MLR models compared to the previous ones. The multiple regression models fit the data well at discharge values between 240 and 900 m^3 s^{-1} (Fig. 8d and e). In these cases, values lower than 240 m^3 s^{-1} and above 900 m^3 s^{-1} were overestimated and underestimated, respectively, and generally increased with the discharge amount. Nevertheless, these underestimations detected in the multiple regression models were lower than those in the linear, quadratic, and exponential models, showing the clear improvement of the multiple regression models compared to the other cases. A specific period of the VGC discharge time series containing high and low peaks predicted by the models (Fig. S8-a and S8-b, respectively) indicated that the multiple regression model using R^{VT} better reproduced the discharge data than the other models. Both MLR models presented satisfactory results and could thus be used to predict good VGC discharge estimates.

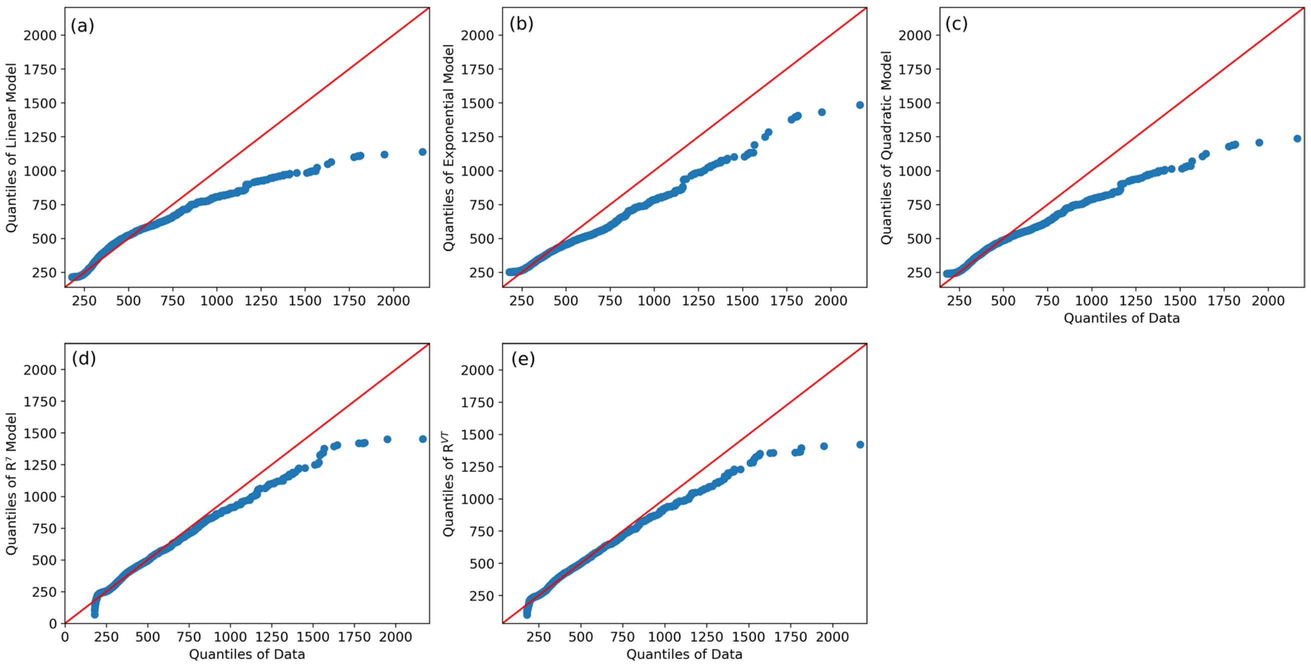

Fig. 8 QQ plots for all investigated models, including (**a**) the linear model, (**b**) the exponential model, (**c**) the quadratic model, (**d**) the MLR model considering R^7, and (**e**) the MLR model considering R^{VT}

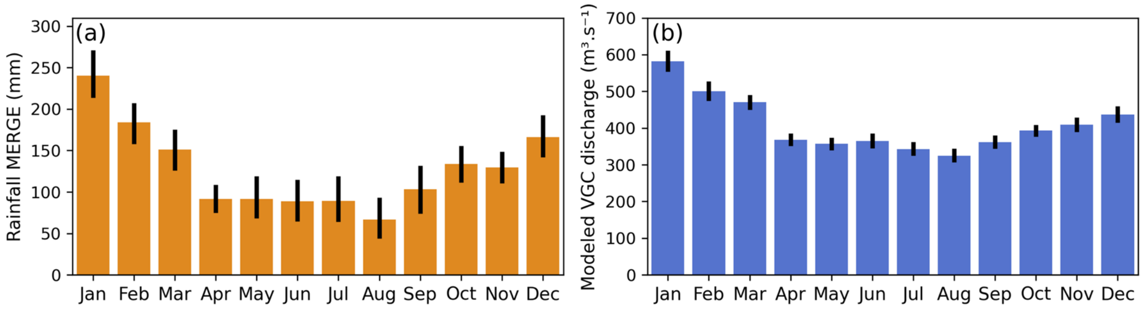

Fig. 9 (**a**) Monthly mean rainfall (orange bars) and 95% confidence interval (black vertical lines) for the Ribeira do Iguape watershed using the MERGE data recorded from June 2000 to December 2020. (**b**) Monthly mean VGC discharge (blue bars) and 95% confidence interval (black vertical lines) from June 2000 to December 2020. These results were generated from the multiple regression model considering R^{VT}

3.4 Reconstruction of complete time series and seasonal comparisons (2000–2020)

Based on the results of the previous sections, the MLR model considering R^{VT} presented the most accurate predictions. Thus, we applied this model to predict a complete time series of VGC discharges (Fig. S6). Since the MERGE data start in June 2000, the time series spans from this month until December 2020 (Fig. S6).

The seasonal precipitation patterns in the Ribeira de Iguape watershed and this reconstructed VGC discharge series were evaluated by the monthly averages considering a 95% confidence level (Fig. 9). The monthly mean rainfall from the MERGE dataset presented the highest values from October to March (Fig. 9a), varying from 130 mm (in November), with a confidence interval from 111 to 149 mm, to 241 mm (in January), with a confidence interval from approximately 74 to 133 mm. The highest mean rainfall and discharge values occurred in January (241 mm and 582 m^3 s^{-1}, respectively), and the lowest values occurred in August (66 mm and 325 m^3 s^{-1}, respectively), considering the period from 2000 to 2020 (Fig. 9). From April to July, the rainfall and VGC discharge values presented relatively small variations, oscillating between 88 and 93 mm, and between 342 and 368 m^3 s^{-1}, respectively (Fig. 9). These values began increasing in September, with monthly mean values of 100 mm and 380 m^3 s^{-1}, within confidence intervals of approximately 74 to 133 mm, and 370 to 390 m^3 s^{-1}, for rainfall and VGC discharge, respectively, achieving their maximum peaks in January.

Basically, the seasonal variability present in rainfall modulated its variability in the VGC discharge results. Both the rainfall and VGC discharge series presented seasonal variabilities, with the highest values found in summer (December, January, and February) and in March, with values greater than 150 mm and 470 m^3 s^{-1}, respectively. The lowest values were found in winter (June, July, and August), at 89 mm and 365 m^3 s^{-1}, with the lowest values found in August, at 66 mm and 325 m^3 s^{-1} for rainfall and discharge, respectively (Figs. 3 and 9).

The highest variabilities were observed in September, January, and July, with confidence intervals ranging from 74 to 133 mm, from 211 to 270 mm, and from 61 to 117 mm, respectively (Fig. 9a). For the monthly mean VGC discharge values (Fig. 9b), the confidence intervals ranged from 111 to 149 m^3 s^{-1} in October and from 560 to 603 m^3 s^{-1} in January. From March to November, the monthly mean VGC discharge values presented lower variabilities than those in December, January, and February, showing confidence interval range values in October from 383 and 405 m^3 s^{-1} and in June from 350 to 380 m^3 s^{-1}. During December, January, and February, the confidence intervals ranged from 420 to

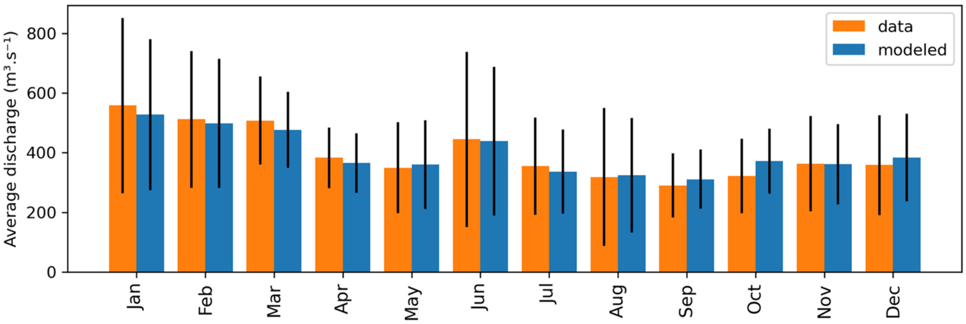

Fig. 10 Monthly mean VGC discharge and standard deviation (black vertical lines) values of the VGC discharge series estimated using the GEOBRÁS (1966) method (orange bars) and with the multiple regression model considering R^{VT} (blue bars) from 2011 to 2019

453 m^3 s^{-1}, 560 to 603 m^3 s^{-1}, and from 481 to 520 m^3 s^{-1}, respectively (Fig. 9b).

The climatology of the VGC discharge was computed using the daily discharge estimates obtained with the multiple regression model considering R^{VT}. These results were compared to the monthly mean VGC discharge data obtained from the daily discharge data available from 2011 and 2019 (Fig. 10). For this period, when comparing the monthly mean VGC discharges between the multiple regression model considering R^{VT} and the data, we found a skill of 0.96, RMSE of 51.61 m^3 s^{-1}, and Pearson's correlation coefficient of 0.95 (with a *p*-value smaller than 0.01). Therefore, the VGC discharge monthly means were well correlated with the monthly means obtained with the multiple regression model considering R^{VT} (Fig. 10), allowing us to extend the climatological period to the period in which MERGE data are available. The highest monthly mean values were found during January in both data and modelled series, at 570 m^3 s^{-1} and 530 m^3 s^{-1}, respectively (Fig. 10). The lowest values were detected during September and reached 280 m^3 s^{-1} and 300 m^3 s^{-1} in the data and model series, respectively (Fig. 10). The best fit between the data and modelled monthly means occurred in November, followed by in June and August (Fig. 10). From January to April, and also in July, the model predictions showed smaller values than the data, while during May, August, September, October, and December, the model overestimated the discharge values compared to the VGC data (Fig. 10). The standard deviation values were similar between the data and multiple regression model considering R^{VT} (Fig. 10). The highest standard deviations were found in January (from approximately 250 to 850 m^3 s^{-1}) and June (from approximately 180 to 750 m^3 s^{-1}), followed by in February and August, which presented values between 260 and 720 m^3 s^{-1} and between 90 and 550 m^3 s^{-1}, respectively (Fig. 10).

High anomalous values were detected from 2011 to 2019 in June (Fig. 10). Curiously, during this period, June presented higher monthly mean values (approximately 450 m^3 s^{-1}) than May and July (approximately 340 m^3 s^{-1} and 320 m^3 s^{-1}, respectively) (Fig. 10). In addition, this pattern contrasted the observed monthly mean June discharge from 2000 to 2010 and in 2020 (Fig. 11). This was the result of the high daily variability (Fig. S7) and the presence of discharge peaks in the VGC discharge during June in 6 of the 9 years from 2011 to 2019 (Fig. S7). These anomalous discharge values in June were caused by the high precipitation in the same period, which followed a consistent pattern starting in 2012 and followed by 2013, 2014, 2016, 2017, and 2019 (Fig. S7). These June anomalies were also found by Marta-Almeida et al. (2021) but were not investigated further in this work, as the topic was out of the scope of this study.

In addition, we presented the monthly mean values from 2000 to 2020 in the predicted time series using daily estimates from the multiple regression model considering R^{VT} and from 2011 to 2019 compared to the observed VGC discharges from the DAEE (Fig. 11). The VGC discharge values estimated from the DAEE data from 2011 to 2019 (orange bars in Fig. 11l to t) were effectively represented by the multiple regression model considering R^{VT} (blue bars in Fig. 11l to t). Twelve of 20 years presented the highest monthly mean VGC-modeled discharge values in January, varying from approximately 340 to 850 m^3 s^{-1} (Fig. 11). The highest monthly VGC discharge values occurred during summer (considering January, February, and March), detected in 2001, 2006, 2007, 2011, 2016, 2017, and 2018 (Fig. 11b, g, h, l, q, r, and s, respectively). In December, high VGC discharge values were also noticed in 2001, 2007, 2008, 2010, 2015, and 2020, varying from 400 to 850 m^3 s^{-1} (Fig. 11b, h, i, k, p, and u, respectively). The lowest values were detected during winter (June, July, and August) only in 2002, 2006, and 2010 (Fig. 11c, g, and k, respectively).

High standard deviation values varying from approximately 300 to 700 m^3 s^{-1} were detected in the climatology for June, July, and August (winter) (Fig. 10). These variabilities were detected due to highly anomalous VGC discharge values mainly in June. This was evident in 2012, 2013, 2014, 2016, 2017, and 2019 (Fig. 11m, n, o, q, r, and t, respectively); in some cases, discharge values similar or even superior to the summer-month values were achieved. Such a finding clearly occurred in June of 2012 (Fig. 11m) and 2019 (Fig. 11t), when the VGC discharge values reached 750 m^3 s^{-1} and 610 m^3 s^{-1}, respectively. In June, we also detected high standard deviations with average values for this month, at 295 m^3 s^{-1} and 190 m^3 s^{-1} for the model results and data, respectively. Anomalous VGC discharge values were also detected in August 2011, reaching 700 m^3 s^{-1} and 600 m^3 s^{-1} for the data and model results, respectively (Fig. 11l). The highest standard deviations were found in December, January, and February, with average values reaching 297 m^3 s^{-1} and 274 m^3 s^{-1} (in January) for the model results and data (Fig. 11), respectively.

4 Discussion

4.1 Accumulated rainfall

The good coherence values between R^7 and the VGC discharge peaks (Fig. 3) and the maximum correlation found between these terms (Fig. S1) allowed us to use R^7 as the input for the linear, quadratic, exponential, and multiple linear regression considering R^{FT} models. We also applied R as a spatially and temporally varying field (R^{VT}) in one of the multiple regression models. The idea of using accumulated rainfall as a single input was to produce simple and reliable statistical models for predicting discharge. Since

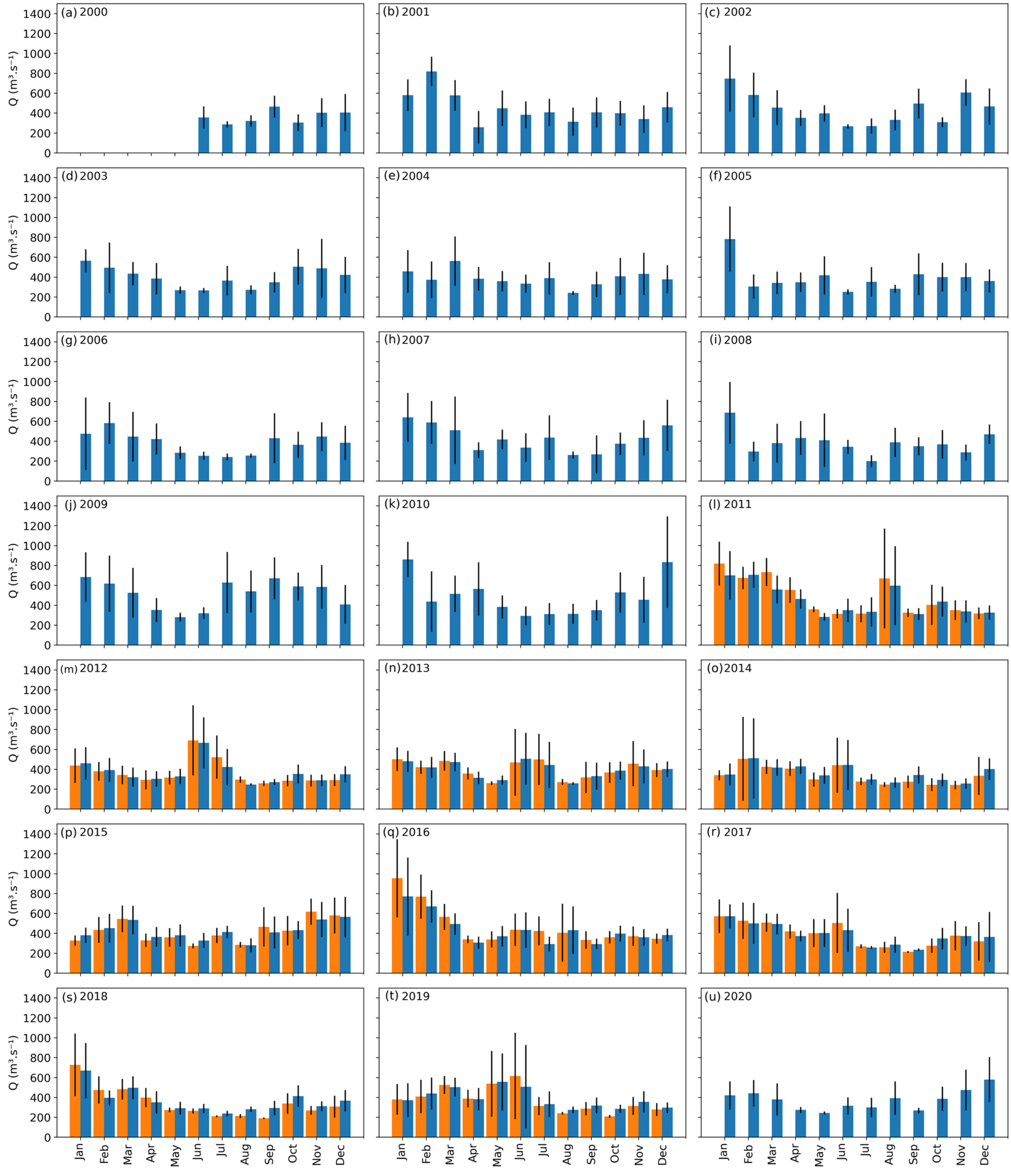

Fig. 11 Modeled monthly mean VGC discharge (in blue) series derived from MERGE rainfall data recorded from 2000 to 2020 and the (in orange) monthly mean VGC discharge series estimated with DAEE data from 2011 to 2019, separated by year and month, including the standard deviation for each month (black vertical lines)

accumulated precipitation was the only input of these models, these data were essential for this study and controlled all the results. Ribeira de Iguape is a catchment with no dams (CBH-RB 2008; IBAMA 2016), in which the rainfall directly influences the unhindered river discharge. This rainfall/river relationship with no human intervention makes our methodology feasible. A decrease in the freshwater contribution from the Jaguaribe River was detected due to the intensification of dam construction along this catchment (ANA 2008; Dias et al. 2013). The performance and outputs of the analyzed models are discussed in the next paragraphs.

A similar pattern of coincident peaks at subseasonal scales was also detected between the rainfall and runoff data in the HBV rainfall-runoff model (Bergström 1976) at the Leaf River catchment in MS, USA, with high rainfall events being associated with peaks in the runoff time series (Abebe et al. 2010). In this case, seasonal fluctuations in both rainfall and runoff time series were detected similarly to our results. Although these series exhibited comparable seasonal variations, it is important to mention that because they are different river basins located in opposite hemispheres, the periods of highest (and lowest) values are expected to differ and, therefore, exhibit distinct peak occurrences. For the Leaf River catchment, the highest values were observed between April and June. The performance and outputs of the analyzed models are discussed in the next sections.

Melesse et al. (2003) adopted distributed rainfall data in an empirical rainfall-runoff model, but considered only a sparse network of rain gauge stations. The authors stated that accurate rainfall data for the catchment is very critical to predict the stream flow and they recommend the use of spatially distributed rainfall data. Rainfall spatial distribution using four gauge stations was obtained by Kim et al. (2003) who applied a storm runoff model. There was a limitation of using only four gauge stations and other rainfall detection methods, such as radar techniques, are suggested to better represent spatial variations for the whole catchment (Kim et al. 2003). Spatially distributed rainfall together with other parameters (roughness and land cover) was also used by Melesse and Graham (2004) to predict the storm runoff. However, the authors considered a fixed period (5 days) to estimate the accumulated rainfall in each grid cell, differing from our methodology in which we calculated the accumulated rainfall with the number of days varying in the domain. Moreover, Melesse and Graham (2004) owned a temporal sparse data, and not a complete rainfall time series as we did. Other studies applied empirical rainfall-runoff models using the rainfall with time lags (Sarkar and Kumar 2012, Lohani et al. 2014, Ahani et al. 2018, Dariane and Azimi 2018). An empirical model was developed by Lohani et al. (2014) with a 16-h lag rainfall data implemented as the input. Sarkar and Kumar (2012) adopted the rainfall with lags varying from 0 to 22 h as input in an ANN empirical model to simulate the runoff in the catchment of Ajay river (India). Moreover, Sarkar and Kumar (2012) and Ahani et al. (2018) used a single station with available rainfall data, and Lohani et al. (2014) obtained the spatially averaged rainfall, not considering the data distribution and variability in the catchment area, as we are considering in our multiple linear regression models.

4.2 Regression models

Peak values (both high and low) are better represented in the MLR models considering R^7 and R^{VT} than in the linear, quadratic, and exponential models and are especially well-represented by the R^{VT}-considering model. In general, the modeled VGC discharge presented RMSE values between 15 and 17% lower (Fig. 4) than those output by the previous models (Fig. S2); however, discharge values higher than 1000 $\mathrm{m}^3\ \mathrm{s}^{-1}$ continued to be underestimated (Fig. 4). A possible explanation is that MERGE underestimates the rainfall peaks when compared with the TRMM satellite data (Rozante et al., 2010). In other words, more realistic data input to our regression model is likely to enhance the results. Such under representations limit the regression models for flood studies and very low flow conditions. This underestimation was also detected in a rainfall-runoff empirical FFBP model applied in the Linbien River (Taiwan) (Chen et al. 2013). These findings have also been reported for more sophisticated models, such as ANNs models, including feed forward back propagation (FFBP), radial basis function-based (RBF), and generalized regression neural networks (GRNN) models, and in an MLR model (Cigizoglu and Alp 2004). Nevertheless, these models were capable of reproducing the runoff observations with reasonable MSE values varying from 40 to 105 $\mathrm{m}^6\ \mathrm{s}^{-2}$ (Cigizoglu and Alp 2004). The FFBP, RBF, and MLR models generated similar MSE values, showing that MLR models can produce results as accurate as ANN model results (Cigizoglu and Alp 2004). The underestimation of high-flow occurrences by the models developed here (Fig. 8) was also detected by Kratzert et al. (2018) for both RNN and LSTM models, although the LSTM model outperformed the RNN model. Nevertheless, this LSTM model was capable of reproducing seasonal patterns and fluctuations, similar to our R^7 and R^{VT} regression models. It is important to mention that Kratzert et al. (2018) used three different datasets that consisted of 7 different meteorological variables (the day length, precipitation, shortwave downward radiation, maximum and minimum temperature, snow-water equivalent, and humidity) as inputs to the LSTM model, while our results used only rainfall as an input and obtained similar NSE coefficients for the R^7 and R^{VT} regression model results. This evidence shows that we have implemented much simpler models, with a unique input, with the same robustness in predicting the runoff of a catchment as the LSTM model.

Comparing the *p*-values and the coefficient values obtained in the multiple linear regression models (Fig. S4 and Fig. S5), we observed that the grid points with coefficient magnitude higher than 0.025 were significant (with *p*-value < 0.05). For grid points where the coefficient magnitude was lower than 0.01, the *p*-value was higher than 0.05, showing that these grid points were not significant for the multiple linear regression model results. This explains the high *p*-values obtained in the domain (Fig. S4 and Fig. S5). Excluding these no significant grid points (with *p*-value > 0.05) from the model results, generated a very similar time series for the VGC discharge, with RMSE differences of 0.05% in relation to our results, showing no significant differences. It can be explained by the fact that the errors probably are compensating each other over the domain. The *p*-values were satisfactory, as the models were capable of reasonably reproducing the complete time series for the VGC discharge, the main information that we were interested in.

Both MLR models presented here reproduced the seasonal fluctuations present in the runoff data (Figs. 5 and 7). The highest runoff peaks were detected during the summer months and in June during some years. Both low- and high-flow events were well-represented by these models (Figs. 5 and 7), especially the model considering R^{VT} (Fig. 7b). In the R^{VT} model, the field of the maximum correlation coefficient was not equally distributed throughout the catchment (Fig. 6) because precipitation is not homogeneous across this domain (Fig. 2). The maximum correlation was associated with R^8 and R^9 in the southern portion, while in the northern part, the maximum correlation was associated with R^6 and R^7. These spatial patterns show that the water associated with precipitation in the southern region takes a longer time (~8–9 days) to reach the VGC than the water derived from rain in the northern area (~6–7 days) (Fig. 6). These spatial differences are associated with the Ribeira de Iguape watershed topography (Fig. 1). The southern region presents higher topography, with elevation values ranging from 800 to 1000 m and a maximum elevation of approximately 2000 m. On the other hand, the northern region presents lower topography, with flatter terrain ranging in elevation from 1 to 100 m (Fig. 1). A catchment with abrupt topography usually presents more permeable bedrock substrates and usually has a larger water storage capacity than a watershed with gentler topography (Sayama et al. 2011). Therefore, an abrupt-topography catchment can sustain lower streamflow responsiveness to precipitation (e.g., Sayama et al. 2011; Wang et al. 2018; Jiang et al. 2012; Muñoz-Villers and McDonnell 2013). Our results are consistent with the findings of previous studies, showing a lower streamflow responsiveness to precipitation in the southern region (steeper topography) of the watershed compared to the northern region (flatter topography) (Fig. 1).

Among the analyzed models, the temporally varying multiple regression model (R^{VT}) was the most reliable in predicting the discharge time series in the Ribeira de Iguape watershed (Table 2), outperforming the other models. This was a consequence of employing all available spatial information of MERGE over the watershed when computing the multiple regression models, in contrast with the linear, exponential, and quadratic regression models, which used only the time series of average spatial precipitation. The skill values of the multiple regression model results were approximately 10% higher than those of the other models. The CDs of the multiple regression models also increased by at least 18% relative to the other models. The RMSE was reduced by at least 30 m^3 s^{-1} and Pearson's correlation coefficient increased by 16% when applying the multiple regression models. Both multiple regression models presented the highest skill among the models (0.92). However, the R^{VT} regression model results were marginally better, presenting a CD of 0.67, RMSE of 106.35 m^3 s^{-1}, and Pearson correlation of 0.86 (*p*-value<0.01) compared to the R^7 regression model results, with a CD equal to 0.64, RMSE of 108.68 m^3 s^{-1}, and Pearson correlation of 0.85 (*p*-value <0.01). Higher CD and lower RMSE values, varying from 0.76 to 0.83 and from 0.5 to 2.5 m^3 s^{-1}, respectively, were estimated for the Jhelum catchment in India (Dar 2017). These results could be associated with the area of this catchment (~8600 km²), which is almost three times smaller than that of the Ribeira de Iguape catchment.

NSE coefficients above 0.74 were found for the MLR models, which increased their performance compared to the other models and showed the robustness of the R^7 and R^{VT} regression models (Table 2). Comparable NSE coefficients were found using the SWMM model at the Belo Horizonte catchment (NSE equal to 0.72, Rosa et al. 2020) and in the Taquari-Antas basin at Rio Grande do Sul, Brazil (NSEs varying from 0.40 to 0.84, Collischonn et al. 2007). Similar NSE coefficients with mean and median values of 0.68 and 0.72, respectively, were also obtained for 241 catchments from the CAMELS dataset using the long short-term memory (LSTM) network, a special type of recurrent neural network (Experiment 3 in Kratzert et al. 2018).

The Ribeira de Iguape catchment is the main source (60%) of freshwater on the southeastern coast of Brazil (Afonso 2006; Marta-Almeida et al. 2021) and therefore strongly influences the salinity in the South Brazil Bight due to the freshwater flowing through the VGC. A discharge time series was produced for the Valo Grande Channel for the period from 2000 to 2020 resulting from the most reliable model (Fig. S6). In the time-series results, two moments presented negative discharges, which is physically impossible.

This is a limitation of the model and should thus be considered depending on the model application (Fig. S6). Nevertheless, negative flow estimations were also detected in other rainfall-runoff models, including RBF neural networks and MLR models (Cigizoglu and Alp 2004). With the predicted time series, we qualitatively provided and validated the monthly mean discharge (Fig. 10) and presented the monthly climatology obtained for the VGC discharge (Fig. 11).

The resulting RMSE values found for our MLR models were at least 50% lower than the RMSE values of a similar rainfall-runoff model applied for the Madhya Pradesh watershed in India (Patel et al. 2016). Our MLR models presented similar NSE and correlation coefficient values compared to the most reliable LSTM model developed by Xiang et al. (2020). Our results were also consistent with those of other rainfall-runoff models. As a comparison, the semidistributed HBV-96 version conceptual model (Lindström et al. 1997) presented $R^2{}_{\mathrm{V}}$ (Lindström 1997) values ranging from 0.6 to 0.8. In addition, our NSE values derived for the MLR models were higher than the NSE values found in 604 basins in the USA using a coupled Snow-17 snow model and the Sacramento Soil Moisture Accounting Model (Newman et al. 2015). For the calibration period, 90% (604) of 671 basins showed NSE values greater than 0.55, and in 34% (225) of the basins, the model provided NSEs higher than 0.8 (Newman et al. 2015).

The monthly mean rainfall information from the MERGE dataset presented the highest values from October to March (Fig. 9a), which is consistent with the period described by Carvalho et al. (2018). The maximum river discharge during summer for the Ribeira de Iguape can be explained by the rainfall distribution in this region (Carvalho et al. 2018). The VGC runoff follows this same pattern (Fig. 9b). Generally, from September to March, the river discharges between Florianópolis and Bertioga, including in Ribeira do Iguape, Paranaguá, Itapocú, and São João, are approximately twice as large as those in the drier months around July (Marta-Almeida et al. 2021).

The seasonal variability is evident in the modeled VGC discharge (Fig. 11), with intense runoff events occurring in the summer, mainly in January and December (Fig. 11), with high standard deviation values varying from 300 to 800 $\mathrm{m}^3\ \mathrm{s}^{-1}$ in the climatology of December, January, and February (Fig. 10). Discharge peaks were also observed during winter (mainly in June, with some in August), with high standard deviation values varying from 300 to 700 $\mathrm{m}^3\ \mathrm{s}^{-1}$ in the climatology for June, July, and August (Fig. 10). These intense discharge events were associated with high rainfall events. In general, the highest discharge peaks were observed on 4 August 2011, at 2164 $\mathrm{m}^3\ \mathrm{s}^{-1}$, followed by 13 January 2016, at 1806 $\mathrm{m}^3\ \mathrm{s}^{-1}$, and 1 January 2018, at approximately 1797 $\mathrm{m}^3\ \mathrm{s}^{-1}$ (Fig. 11). Marta-Almeida et al. (2021) found that the highest peaks occurred in February, June, and December in 2014 but did not observe highest peaks in these months in other years (except for a low peak in July 2016). We observed peaks in February and June of 2014 but not in December of 2014 (Fig. 11). The reconstructed discharge time series and the monthly climatology presented here can be used in oceanic models to represent the discharge from the Ribeira de Iguape watershed system in estuarine and coastal hydrodynamic forecast and hindcast simulations. These applications should consider that, in general, the modeled discharge peaks present some underrepresentation of reality, which might limit some studies related to extreme events.

5 Concluding remarks

In this work, simple models based on MERGE rainfall data were shown to be capable of providing information about discharge in the Ribeira de Iguape watershed. We evaluated 5 different models that use rainfall data from MERGE to provide discharge estimations for the VGC. We applied linear regression, quadratic, and exponential models and two different MLR models. In the first multiple regression model, R was estimated for 7 days (R^7), and in the second model, the number of days used to compute R varied spatially (R^{VT}); in addition, we calculated the standard errors and p-values for each grid cell from the MERGE precipitation data. The VGC discharge time series predicted by each model was compared with estimated discharges from data collected from the DAEE. Among the applied models, the multiple regression with R^{VT} was the most accurate (RMSE of 103.70 $\mathrm{m}^3\ \mathrm{s}^{-1}$) and had the highest coefficient of determination (CD of 0.75), skill (0.92), and NSE (0.75) values. We attribute this better representation of the VGC discharge to this type of statistical model considering individual grid points from the MERGE rainfall data instead of using spatially averaged rainfall, as the other models did. Moreover, the multiple regression model that considered R^{VT} was more reliable than the regression that considered a constant period of accumulation (R^7). This finding is the result of the first model also considering varying days of accumulation based on the highest correlation value between the R and VGC discharge.

Discharge measurements are important for constructing and calibrating numerical models for oceanic predictions, leading to improved salinity fields and ocean dynamic simulations in continental shelf and estuarine regions. The VGC discharge time series and the monthly climatology results presented here can be included in operational oceanic models at small and mesoscales in the region and can be used to calibrate numerical simulations. In addition, it is fundamental to estimate the dispersion of materials and pollutants through rivers and channels and to evaluate their impacts. The results presented here can also be used to investigate the hydrodynamics in the CIELS and how pollutants and other substances are transported in this system.

The methodology presented here can be replicated in other watersheds around the world if sufficiently long training datasets are available. The models require as inputs precipitation maps constructed over time for a given watershed. This information is used to set the MLR models. MLR can be considered a powerful statistical tool to generate highly simple and reliable rainfall-runoff models. Such an MLR model can provide similar results compared to ANN models, such as FFBPNN and GRNN models (Turhan 2021). MLR models have also presented similar results as ANN models when applied together with wavelet transformation (Partal 2017).

The models applied in this study presented some limitations, particularly regarding the detection of high and extremely low flows, which relates to the input data limitations itself (MERGE). However, the underestimation of peaks detected in our models is also present in more sophisticated models, as ANN models (Cigizoglu and Alp 2004, Chen et al. 2013). Although we detected grid cells with no significance in the multiple linear regression models, their presence did not impact our final results. The main advantage of the applied methods was to provide a reasonable estimate of river discharge for a period when there were no measurements. Some other advantages were providing a spatial distribution (map) for the variables in the model and also a continuous time series of the discharge. As a future work, regression models for each of the discharge gauging locations (Jacupiranga, Ribeira do Iguape, and Pariquera-Açu) can be developed for comparison and validation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00704-023-04459-4.

Acknowledgements The authors thank the CPTEC/INPE for providing the MERGE dataset, downloaded through the FTP server at http://ftp.cptec.inpe.br/modelos/tempo/MERGE/GPM/DAILY. We also thank the CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico) for providing the PhD scholarship for Paula Birocchi during the development of this study and Ligia Dias Araujo for helping to create Fig. 1.

Code availability The Python codes used here can be requested via email to paula.birocchi@usp.br.

Author contribution All authors contributed to the study's conception, design, and investigation. The methodology was chosen and the formal analysis was made by Paula Birocchi, Danilo Augusto Silva, and Dalton Kei Sasaki. The software usage and validation were made by Paula Birocchi. The visual elements of the paper were made by Paula Birocchi and Dalton Kei Sasaki. The first draft of the manuscript was written by Paula Birocchi and all authors commented on previous versions of the manuscript. All authors contributed to writing, reviewing, and editing the manuscript. All authors read and approved the final manuscript.

Funding Author Paula Birocchi has received a PhD scholarship from CNPQ (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

Data availability The MERGE datasets used and analyzed during the current study are available at http://ftp.cptec.inpe.br/modelos/tempo/MERGE/GPM/DAILY. For more information about the data results for the implemented models, please contact Paula Birocchi.

Declarations

Ethics approval Not applicable

Consent to participate The authors Paula Birocchi, Danilo Augusto Silva, Dalton Kei Sasaki, and Marcelo Dottori consent to participate in this article.

Consent for publication Not applicable

Competing interests The authors declare no competing interests.

References

Abebe NA, Ogden FL, Pradhan NR (2010) Sensitivity and uncertainty analysis of the conceptual HBV rainfall–runoff model: implications for parameter estimation. J Hydrol 389(3-4):301–310. https://doi.org/10.1016/j.jhydrol.2010.06.007

Afonso CM (2006) A paisagem da Baixada Santista: urbanizacao, transformacão e conservacão. Editora Universidade de São Paulo: FAPESP, São Paulo, p 310

Ahani A, Shourian M, Rad PR (2018) Performance assessment of the linear, nonlinear and nonparametric data driven models in river flow forecasting. Water Resour Manage 32(2):383–399. https://doi.org/10.1007/s11269-017-1792-5

Allasia DG, Silva B, Collischonn W, Tucci CEM (2006) Large basin simulation experience in South America. In: Sivapalan M, Wagener T, Uhlenbrook S, Zehe E, Lakshmi V, Liang X, Tachikawa Y, Kumar P (eds) Proc. Brazil Symp., April 2005Prediction in Ungauged Basins: Promises and Progress. IAHS Publ. 303. IAHS Press, Wallingford, UK, pp 360–370

Ambrosio BG (2016) Dinâmica da desembocadura lagunar de Cananéia, litoral sul do estado de São Paulo (Master dissertation). Universidade de São Paulo

ANA (2008) National Water Agency. Historic Outflows From Jaguaribe River. www.hidroweb.ana.gov.br (retrieved 20.10.2008).

Barcellini VC, Motta FS, Martins AM, Moro PS (2013) Recreational anglers and fishing guides from an estuarine protected area in southeastern Brazil: socioeconomic characteristics and views on fisheries management. Ocean Coast Manag 76:23–29. https://doi.org/10.1016/j.ocecoaman.2013.02.012

Barioto JG, Stanski G, Grabowski RC, Costa RC, Castilho AL (2017) Ecological distribution of Penaeus schmitti (Dendrobranchiata: Penaeidae) juveniles and adults on the southern coast of São Paulo state Brazil. Marine Biology Research 13(6):693–703. https://doi.org/10.1080/17451000.2017.1287923

Barnes SL (1973) Mesoscale objective analysis using weighted time-series observations. In: NOAA Tech. Memo. ERL NSSL-62. National Severe Storms Laboratory, Norman, OK, p 60

Belvederesi C, Dominic JA, Hassan QK, Gupta A, Achari G (2020) Short-term river flow forecasting framework and its application in cold climatic regions. Water 12(11):3049. https://doi.org/10.3390/w12113049

Bérgamo AL (2000) Características da hidrografia, circulação e transporte de sal: Barra de Cananéia, Sul do mar de Cananéia e Baía do Trapandé (Doctoral dissertation. Universidade de São Paulo

Bergström S (1976) Development and application of a conceptual runoff model for Scandinavian Catchments. Report RHO 7, Swedish Meteorological and Hydrological Institute, Norrkoping, Sweden, 134

Bochini GL, Stanski G, Castilho AL, da Costa RC (2019) The crustacean bycatch of seabob shrimp Xiphopenaeus kroyeri (Heller, 1862) fisheries in the Cananéia region, southern coast of São Paulo, Brazil. Reg Stud Mar Sci 31:100799. https://doi.org/10.1016/j.rsma.2019.100799

Carvalho OJ, Aguiar W, Cirano M, Genz F, Amorim FND (2018) A climatology of the annual cycle of river discharges into the Brazilian continental shelves: from seasonal to interannual variability. Environ Earth Sci 77:192. https://doi.org/10.1007/s12665-018-7349-y

CBH-RB (Comitê da Bacia Hidrográfica do Ribeira de Iguape e Litoral Sul) (2008) Plano Diretor de Recursos Hídricos da Unidade de Gerenciamento N° 11: Bacia Hidrográfica do Ribeira de Iguape e Litoral Sul. Fundespa e Fundo Estadual de Recursos Hidrícos https://sigrh.sp.gov.br/public/uploads/documents/7082/plano_bacia_ugrhi-11_2008-2011.pdf. Accesse 1 Oct 2022

Chen SM, Wang YM, Tsou I (2013) Using artificial neural network approach for modelling rainfall–runoff due to typhoon. J Earth Syst Sci 122:399–405. https://doi.org/10.1007/s12040-013-0289-8

Cigizoglu HK, Alp M (2004) Rainfall-runoff modelling using three neural network methods. In: Rutkowski L, Siekmann JH, Tadeusiewicz R, Zadeh LA (eds) Artificial Intelligence and Soft Computing - ICAISC 2004. ICAISC 2004, Lecture Notes in Computer Science(), vol 3070. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-24844-6_20

Coleman JM, Wright LD (1971) Analysis of major rivers and their deltas: procedures and rationale, with two examples. Louisiana State University Press, Baton Rouge, p 125

Collischonn W, Allasia D, Da Silva BC, Tucci CE (2007) The MGB-IPH model for large-scale rainfall—runoff modelling. Hydrol Sci J 52(5):878–895. https://doi.org/10.1623/hysj.52.5.878

Curcho MRSM, Farias LA, Baggio SR, Fonseca BC, Nascimento SMD, Bortoli MC, Braga ES, Fávaro DIT (2009) Mercury and methylmercury content, fatty acids profile, and proximate composition of consumed fish in Cananéia, São Paulo, Brazil. Rev Inst Adolfo Lutz 68(3):442–450

DAEE (Departamento Estadual de Águas e Energia Elétrica) (1998) Bacia Hidrográfica do Ribeira do Iguape - Plano de Ação Para o Controle das Inundações e Diretrizes Para o Desenvolvimento do Vale, p 68

Dariane AB, Azimi S (2018) Streamflow forecasting by combining neural networks and fuzzy models using advanced methods of input variable selection. J Hydroinformatics 20(2):520–532. https://doi.org/10.2166/hydro.2017.076

Dar LA (2017) Rainfall-runoff modeling using multiple linear regression technique. International Journal for Research in Applied Sciences, Engineering and Technology, 5(7), 214–218.

Devia GK, Ganasri BP, Dwarakish GS (2015) A review on hydrological models. Aquatic Procedia 4:1001–1007. https://doi.org/10.1016/j.aqpro.2015.02.126

Dias FDS, Castro BM, Lacerda LDD (2013) Continental shelf water masses off the Jaguaribe River (4S), northeastern Brazil. Cont Shelf Res 66:123–135. https://doi.org/10.1016/j.csr.2013.06.005

Dwarakish GS, Ganasri BP (2015) Impact of land use change on hydrological systems: a review of current modeling approaches. Cogent Geoscience 1(1):1115691. https://doi.org/10.1080/23312041.2015.1115691

Faurès JM, Goodrich DC, Woolhiser DA, Sorooshian S (1995) Impact of small-scale spatial rainfall variability on runoff modeling. J Hydrol 173(1-4):309–326. https://doi.org/10.1016/0022-1694(95)02704-S

Filla GDF, Oliveira CIBD, Gonçalves JM, Monteiro-Filho ELDA (2012) The economic evaluation of estuarine dolphin (Sotalia guianensis) watching tourism in the Cananéia region, south-eastern Brazil. Int J Green Econ 6(1):95–116

Galvão MSN, Pereira OM, Machado IC, Heriques MB (2000) Reproductive characters of the oyster Crassostrea brasiliana from mangroves of Cananéia estuary, São Paulo, Brazil. Bol Inst Pesca 26:147–162 ISSN: 0046-9939

Geise L, Gomes N, Cerqueira R (1999) Behaviour, habitat use and population size of Sotalia fluviatilis (Gervais, 1853) (Cetacea, Delphinidae) in the Cananéia estuary region, São Paulo, Brazil. Rev Bras Biol 59(2):183–194. https://doi.org/10.1590/S0034-71081999000200002

GEOBRÁS (GEOBRÁS S/A - Engenharia e Fundações) (1966) Complexo VaIo Grande- Mar Pequeno - Rio Ribeira de Iguape. Relatório para o serviço do Vale do Ribeira. São Paulo, DAEE. 2 v

IBAMA (Instituto Brasileiro do Meio Ambiente e dos Recursos Naturais Renováveis) (2016) Indeferimento do pedido de licença prévia para a UHE Tijuco Alto. Ministério do Meio Ambiente https://site-antigo.socioambiental.org/sites/blog.socioambiental.org/files/nsa/arquivos/document.pdf. Accessed 15 Sept 2022

Jiang R, Woli KP, Kuramochi K, Hayakawa A, Shimizu M, Hatano R (2012) Coupled control of land use and topography on nitrate-nitrogen dynamics in three adjacent watersheds. Catena 97:1–11. https://doi.org/10.1016/j.catena.2012.04.015

Kim SJ, Kwon HJ, Jung IK, Park GA (2003) A comparative study on grid-based storm runoff prediction using Thiessen and spatially distributed rainfall. Paddy Water Environ 1:149–155. https://doi.org/10.1007/s10333-003-0023-2

Kokkonen T, Koivusalo H, Karvonen T (2001) A semi-distributed approach to rainfall-runoff modelling—a case study in a snow affected catchment. Environ Model Software 16(5):481–493. https://doi.org/10.1016/S1364-8152(01)00028-7

Kratzert F, Klotz D, Brenner C, Schulz K, Herrnegger M (2018) Rainfall–runoff modelling using long short-term memory (LSTM) networks. Hydrol Earth Syst Sci 22(11):6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Lindström G (1997) A simple automatic calibration routine for the HBV model. Nordic Hydrology 28:153–168. https://doi.org/10.2166/nh.1997.0009

Lindström G, Gardelin M, Johansson B, Persson M, Bergström S (1997) Development and test of the distributed HBV-96 hydrological model. J Hydrol 201:272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

Lohani AK, Goel NK, Bhatia KKS (2014) Improving real time flood forecasting using fuzzy inference system. J Hydrol 509:25–41. https://doi.org/10.1016/j.jhydrol.2013.11.021

Ma HY, Mechoso CR, Xue Y, Xiao H, Wu CM, Li JL, De Sales F (2011) Impact of land surface processes on the South American warm season climate. Climate Dynam 37(1):187–203. https://doi.org/10.1007/s00382-010-0813-3

McAleer T (2020) Interpreting linear regression through statsmodels .summary(). Available at: https://medium.com/swlh/interpreting-linear-regression-through-statsmodels-summary-4796d359035a. Accessed 10 Oct 2021

Miranda LB, Castro BM, Kjerfve B (2012) Princípios de Oceanografia Física de Estuários, second edn. Universidade de São Paulo (USP), São Paulo, SP, Brazil

Marta-Almeida M, Dalbosco A, Franco D, Ruiz-Villarreal M (2021) Dynamics of river plumes in the South Brazilian Bight and South Brazil. Ocean Dynamics 71(1):59–80. https://doi.org/10.1007/s10236-020-01397-x

Melesse AM, Graham WD, Jordan JD (2003) Spatially distributed watershed mapping and modeling: GIS-based storm runoff response and hydrograph analysis: Part 2. J Spat Hydrol 3(2):1–28

Melesse AM, Graham WD (2004) Storm runoff prediction based on a spatially distributed travel time method utilizing remote sensing and GIS 1. Am J Water Resour 40(4):863–879. https://doi.org/10.1111/j.1752-1688.2004.tb01051.x

Moraes RP (1997) Transporte de chumbo e metais associados no Rio Ribeira de Iguape, São Paulo, Brasil. Masters Dissertation. Instituto de Geociências, Universidade Estadual de Campinas, Campinas, p 94

Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50(3):885–900. https://doi.org/10.13031/2013.23153

Moore RJ (2007) The PDM rainfall–runoff model. Hydrol Earth Syst Sci 11(1):483–499. https://doi.org/10.5194/hess-11-483-2007

Muñoz-Villers LE, McDonnell JJ (2013) Land use change effects on runoff generation in a humid tropical montane cloud forest region. Hydrol Earth Syst Sci 17:3543–3560. https://doi.org/10.5194/hess-17-3543-2013

Najafi MR, Moradkhani H (2016) Ensemble combination of seasonal streamflow forecasts. J Hydrol Eng 21(1):04015043. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250

Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—a discussion of principles. J Hydrol 10(3):282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Newman AJ, Clark MP, Sampson K, Wood A, Hay LE, Bock A et al (2015) Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol Earth Syst Sci 19(1):209–223. https://doi.org/10.5194/hess-19-209-2015

Oliveira J, Braga ES, Jesus SC, Abrahao FF, Santos GF, Chiozzini V (2009) Assessment of natural radium isotopes and sup (222) Rn in water samples from Cananeia-Iguape estuarine complex, Sao Paulo. In: International nuclear atlantic conference; meeting on nuclear applications, 9th; meeting on reactor physics and thermal hydraulics, 16th; meeting on nuclear industry, 1st, September 27 - October 2, 2009, Rio de Janeiro, RJ. Proceedings... Sao Paulo: ABEN, 2009, 2009. Available at: http://repositorio.ipen.br/handle/123456789/12324. Accessed in: 11-05-2022.

Patel S, Hardaha MK, Seetpal MK, Madankar KK (2016) Multiple linear regression model for stream flow estimation of Wainganga River. Am J Sci Eng 2(1):1–5. https://doi.org/10.11648/j.ajwse.20160201.11

Partal T (2017) Wavelet regression and wavelet neural network models for forecasting monthly streamflow. J Water Clim Chang 8(1):48–61. https://doi.org/10.2166/wcc.2016.091

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: Machine learning in Python. J Mach Learn Res 12:2825–2830

Perlman H (2016) The water cycle- USGS Water Science School. https://water.usgs.gov/edu/watercycle.html. Accessed 13 Jul 2022

Pisetta M (2006) Transporte de sedimentos por suspensão no sistema estuarino-lagunar de Cananéia-lguape (SP). Master in Chemical and Geological Oceanography - Instituto Oceanográfico, Universidade de São Paulo (IO–USP), São Paulo, p 179 (Master dissertation

Pisetta M (2010) Análise do processo de distribuição do material particulado emsuspensão e metais associados no sistema Cananéia-Iguape (SP). Instituto Oceanográfico, Universidade de São Paulo, p 175 PhD Thesis

Rosa DWB, Nascimento NO, Moura PM, Macedo GD (2020) Assessment of the hydrological response of an urban watershed to rainfall-runoff events in different land use scenarios – Belo Horizonte, MG, Brazil. Water Sci Technol 81(4):679–693. https://doi.org/10.2166/wst.2020.148

Rossman LA (2010) Storm water management model user's manual, version 5.0. National Risk Management Research Laboratory, Office of Research and Development, US Environmental Protection Agency, Cincinnati, p 276

Rozante JR, Moreira DS, Gonçalves LGG, Vila DA (2010) Combining TRMM and surface observations of precipitation: technique and validation over South America. Weather and Forecasting 25:885–894. https://doi.org/10.1175/2010WAF2222325.1

Sahoo BB, Jha R, Singh A, Kumar D (2019) Application of support vector regression for modeling low flow time series. KSCE J Civ Eng 23:923–934. https://doi.org/10.1007/s12205-018-0128-1

Sarkar A, Kumar R (2012) Artificial neural networks for event based rainfall-runoff modeling. J Water Resource Prot 4(10):891. https://doi.org/10.4236/jwarp.2012.410105

Santos MDO, Rosso S (2007) Ecological aspects of marine tucuxi dolphins (Sotalia guianensis) based on group size and composition in the Cananéia estuary, southeastern Brazil. Lat Am J Aquat Ma:71–82. https://doi.org/10.5597/lajam00110

Satyamurti P, Nobre C, Dias PLS (1998) South America. In: Karoly DJ, Vicent DJ (eds) Meteorology of the Southern Hemisphere. American Meteorological Society, Boston, pp 119–139

Sayama T, McDonnell JJ, Dhakal A, Sullivan K (2011) How much water can a watershed store? Hydrol Process 25:3899–3908. https://doi.org/10.1002/hyp.8288

Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference 57(61):10–25080

Sichangi AW, Wang L, Hu Z (2018) Estimation of river discharge solely from remote-sensing derived data: an initial study over the Yangtze river. Remote Sens (Basel) 10(9):1385. https://doi.org/10.3390/rs10091385

Sitterson J, Knightes C, Parmar R, Wolfe K, Avant B, Muche M (2018) An overview of rainfall-runoff model types. 9th International Congress on Environmental Modelling and Software. https://scholarsarchive.byu.edu/iemssconference. Accessed 22 Aug 2022

Shoaib M, Shamseldin AY, Khan S, Khan MM, Khan ZM, Sultan T, Melville BW (2018) A comparative study of various hybrid wavelet feedforward neural network models for runoff forecasting. Water resources management 32(1):83–103. https://doi.org/10.1007/s11269-017-1796-1

Stanski G, Carvalho MM, Garcia JR, Goncalves GR, Costa RC, Castilho AL (2018) Geographical variation and local environment effects in the reproductive output and fecundity of the shrimp Exhippolysmata oplophoroides (Decapoda: Caridea) in southeastern Brazil. Invertebrate Reproduction & Development 62(2):119–124. https://doi.org/10.1080/07924259.2018.1448305

Stech JL, Lorenzzetti JA (1992) The response of the South Brazil Bight to the passage of wintertime cold fronts. J Geophys Res Oceans 97(C6):9507–9520. https://doi.org/10.1029/92JC00486

Turhan E (2021) A comparative evaluation of the use of artificial neural networks for modeling the rainfall-runoff relationship in water resources Management. Journal of Ecological Engineering 22(5). https://doi.org/10.12911/22998993/135775

Vano JA, Lettenmaier DP (2014) A sensitivity-based approach to evaluating future changes in Colorado River discharge. Clim Change 122:621–634. https://doi.org/10.1007/s10584-013-1023-x

Wang C, Shang S, Jia D, Han Y, Sauvage S, Sánchez-Pérez JM, Kuramochi K, Hatano R (2018) Integrated effects of land use and topography on streamflow response to precipitation in an agriculture-forest dominated northern watershed. Water 10(5):633. https://doi.org/10.3390/w10050633

Willmott CJ (1981) On the validation of models. Physical geography 2(2):184–194. https://doi.org/10.1080/02723646.1981.10642213

Wood EF, Roundy JK, Troy TJ, van Beek LPH, Bierkens MFP, Blyth E, de Roo A, Döll P, Ek M, Famiglietti J, Gochis D, van de Giesen N, Houser P, Jaffé PR, Kollet S, Lehner B, Lettenmaier DP, Peters-Lidard C, Sivapalan M et al (2011) Hyperresolution global land surface modeling: meeting a grand challenge for monitoring

Earth's terrestrial water. Water Resour Res 47:W05301. https://doi.org/10.1029/2010WR010090

Xiang Z, Yan J, Demir I (2020) A rainfall-runoff model with LSTM-based sequence-to-sequence learning. Water Resour Res 56(1). https://doi.org/10.1029/2019WR025326

Zhang R, Cuartas LA, Carvalho LVC, Leal KRD, Mendiondo EM, Abe N, Birkinshaw S, Mohor SG, Seluchi ME, Nobre CA (2018) Season-based rainfall–runoff modelling using the probability-distributed model (PDM) for large basins in southeastern Brazil. Hydrol Process 32(14):2217–2230. https://doi.org/10.1002/hyp.13154

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