Revista Brasileira de Recursos Hídricos
Brazilian Journal of Water Resources
Versão On-line ISSN 2318-0331
RBRH, Porto Alegre, v. 29, e18,
Scientific/Technical Article

https://doi.org/10.1590/2318-0331.292420230131

Investigation of using missing data imputation methodologies effect on the SARIMA model performance: application to average monthly flows

Investigação do efeito do uso de metodologias de imputação de dados faltantes no desempenho do modelo SARIMA: aplicação para vazões médias mensais

Michel Trarbach Bleidorn[1], Isamara Maria Schmidt[1], José Antonio Tosta dos Reis[1], Deysilara Figueira Pani[1], Wanderson de Paula Pinto[2], Carlo Corrêa Solci[1], Antonio Sergio Ferreira Mendonça[1] & Gutemberg Hespanha Brasil[1]

[1] Universidade Federal do Espírito Santo, Vitória, ES, Brasil
[2] Centro Universitário FAVENI, Guarulhos, SP, Brasil

E-mails: michelbleidorn@gmail.com (MTB), isamaraschmidt@gmail.com (IMS), jatreis@gmail.com (JATR), deysi.larapani@gmail.com (DFP), wandersondpp@gmail.com (WPP), csolci13@gmail.com (CCS), anserfm@terra.com.br (ASFM), ghbrasil@terra.com.br (GHB)

Received: November 03, 2023 - Revised: April 12, 2024 - Accepted: April 29, 2024

ABSTRACT

Accuracy in river flows forecasts is crucial for Hydrology, but is challenged by fluviometric data quality. This study investigates the impact of different missing data imputation methods on the Seasonal Autoregressive Integrated Moving Average (SARIMA) model performance. SARIMA $(1,1,1)(0,1,1)_{12}$ was selected using semi-automated criteria, such as lowest AIC, significant parameters (*p-value* < 0.05) and residuals adequacy. This model was then compared with reconstructed series using different imputation methods such as Mean (AM), Median (M), Spline and Stinemann Interpolations, Regional Weighting (RW), Multiple Linear Regression (MLR), Multiple Imputation (MI) and Maximum Likelihood (ML). The data were analyzed considering scenarios of 5, 20 and 40% missing data, following random and block patterns, using data from the Doce River, in Southeast Brazil. Results obtained by the performance indicators and, their respective relative differences, indicated that, univariate (AM and M) and multivariate (PW and RLM) methods limited the model's performance, while univariate Spline and Stine and multivariate IM and ML methods didn't present significant limitations, except Spline for the block pattern. It is concluded that, future predictions accuracy depends, not only on a well-trained and validated model, but also on the appropriate use of missing data imputation methods.

Keywords: Missing data imputation methodologies; Forecast; SARIMA.

RESUMO

A precisão nas previsões de vazão dos rios é crucial para a Hidrologia, mas é desafiada pela qualidade dos dados fluviométricos. Este estudo investiga o impacto de diferentes métodos de imputação de dados faltantes no desempenho do modelo Autoregressivo Integrado de Médias Móveis Sazonal (SARIMA). O modelo SARIMA $(1,1,1)(0,1,1)_{12}$ foi selecionado usando critérios semi-automatizados, como menor AIC, parâmetros significativos (*p-valor* < 0,05) e adequação dos resíduos. Este modelo foi então comparado com séries reconstruídas usando diferentes métodos de imputação, como Média (AM), Mediana (M), Interpolações Spline e Stinemann, Ponderação Regional (RW), Regressão Linear Múltipla (MLR), Imputação Múltipla (MI) e Máxima Verossimilhança (ML). Os dados foram analisados considerando cenários de 5, 20 e 40% de dados faltantes, seguindo padrões aleatórios e de blocos, utilizando dados do Rio Doce, no Sudeste do Brasil. Os resultados obtidos pelos indicadores de desempenho e suas respectivas diferenças relativas, indicaram que, métodos univariados (AM e M) e multivariados (PW e RLM) limitaram o desempenho do modelo, enquanto os métodos univariados Spline e Stine e multivariados IM e ML não apresentaram limitações significativas, exceto Spline para o padrão de blocos. Conclui-se que a precisão das previsões futuras depende, não apenas de um modelo bem treinado e validado, mas também, do uso adequado de métodos de imputação de dados faltantes.

Palavras-chave: Metodologias de imputação de dados faltantes; Previsão; SARIMA.

INTRODUCTION

Although flow forecasting constitutes a relevant focus of interest for Hydrology, its conduct presents substantial challenges, mainly the increase in its quality. Successful forecasts not only make it possible to contribute to important sectors of water planning, such as human supply (Liu et al., 2021), control of floods and droughts (Ahmad et al., 2022), reservoirs operation for water storage and/or hydroelectric power generation, irrigation (Aghelpour et al., 2021) and industry (Schäfer et al., 2016), but they also anticipate water use conflicts and are fundamental to mitigating the impacts of climate change (Musa, 2013).

Among applied forecasting methodologies, physical, conceptual and data-based methods stand out (Apaydin et al., 2021). While the first two require complex variables that are difficult to appropriate, data-based methods offer direct analyzes of the variable of interest, being advantageous in this sense (Khodakhah et al., 2022).

The attributes highlighted by Fu et al. (2019) such as the ability to establish a quantitative interaction between inputs and outputs, fast modeling speed and high prediction accuracy, are essential factors that drive the choice of data-driven methods. For this reason, in recent years, techniques such as artificial intelligence, deep learning, machine learning and time series have been widely adopted in forecasting hydrological variables.

Time series methodology, developed in the 1970s by statisticians Box and Jenkins, which consolidated the class of ARMA (Autoregressive and Moving Average) models, allowed the establishment of several generic models. Seasonal Autoregressive Integrated Moving Average (SARIMA) model stands out in this class, as it considers seasonality, an intrinsic characteristic of flow time series (Abudu et al., 2010; Bayer et al., 2012).

Essentially, time series methodologies efficiency depends on the serial correlation structure. In this way, data quality plays a crucial role in the successful application of these methodologies. In contrast, the incidence of missing data in flow time series imposes limitations on the autocorrelation function, leading to questions about viability and performance of forecast models, generating uncertainty in results and decision making (Giustarini et al., 2016; Dembélé et al., 2019).

The main causes of missing data in flow time series are the hydrometrist absence, failures in the instruments of data collection, loss of annotations, and monitoring termination or interruptions (Gao et al., 2018; Bleidorn et al., 2022). Therefore, there has been a growing interest in recent years in the missing data treatment, including through different imputations approaches (Hamzah et al., 2020). Examples include the works of Tencaliec et al. (2015), that used dynamic regression to impute missing data from the Durance River, France; Dembélé et al. (2019), that applied direct sampling with different hydroclimatic settings to fill gaps in flow data from the Volta River, West Africa; Semiromi & Koch (2019), that used singular spectrum analysis and multichannel to reconstruct groundwater levels in the Ardabil Plain, Iran; Arriagada et al. (2021), that utilized a machine learning algorithm to fill missing data on daily flows in several Chile watersheds; Hamzah et al. (2022), that used equations of multiple chained imputations (IM) to fill missing data of Langat River, Malaysia.

In order to preserve the analysis quality in studies with time series, different authors previously used methodologies to impute missing data. For example, in the study by Pinto et al. (2015), was used the maximum likelihood methodology through the EM (Expectation-Maximization) algorithm to reconstruct the monthly mean flows series from Doce River, Brazil, before adjust the SARIMA $(1,1,1)(1,1,2)_{12}$ model. Bleidorn et al. (2019) utilized the average of the closest neighbors to perform the missing data imputation in the time series from Jucu River, Brazil, and, subsequently, the SARIMA $(1,0,0)(5,1,0)_{12}$ model was fitted. Duarte et al. (2019) imputed missing data of the monthly mean flows series of Manuel Alves da Natividade River, Brazil, using the Kalman filter, and after model adjustment analysis, the SARIMA $(1,0,1)(1,1,4)_{12}$ was the one that presented the best performance. Salame et al. (2019) utilized the IM to fill the missing values of flows and precipitation in the Araguaia Watershed, Brazil and, afterwards, compare the performance of Box and Jenkins approach and artificial neural networks for forecasts of the studied variables. Phan & Nguyen (2020) used linear imputation and/or moving average to fill the missing data of a monitoring station of the Red River, China, with posterior adjustment of ARIMA models. Retike et al. (2022) performed data reconstruction from the Latvian national groundwater level database. After that, the authors adjusted ARMA and ARIMA models to observed and reconstructed data and, evaluated, how the AIC values behaved, since their lowest values indicate the best fit of the models to the data. Using 605 series, the adjusted models evaluation showed, for most series, that the AIC values were significantly improved for the reconstructed series.

Other studies used methodologies to impute missing data and, after that, used the approach of interest in their analyses. For example, Gill et al. (2007) observed that groundwater level predictions using Neural Networks and Support Vector Machines after missing's imputation with the least squares method were close to the observed data performance. Chen et al. (2018) evaluated the influence of imputation missing precipitation data on the performance of hydrological nonpoint pollution (H/NPS) forecasting using the SWAT model in a case study for Daning River watershed, Three Gorges Reservoir Region, China. The authors concluded that the EMB imputation methodology (a combination of the EM and Bootstrap algorithm) was better to the performances of the other two methods analyzed (data augmentation algorithm and meteorological generator).

Although imputation methods application before using interest hydrological models has been recurrently observed, a significant gap in the literature lies in the assessment of their impact on the SARIMA model specific performance. Therefore, this study aims to investigate how different missing data imputation approaches influence both the fit and forecasting ability of the SARIMA time series model. This analysis will be carried out, using average monthly flow data from the Doce river, Brazil, as a case study.

Results obtained in this study have the potential to contribute to improving flow forecasts quality. By offering specific insights into how different imputation techniques affect SARIMA model performance, it will be possible to make more consistent decisions in water resources management. Furthermore, this research will contribute to reducing the uncertainty associated with hydrological forecasts, providing a more solid basis for strategic and operational decision-making.

METHODOLOGY

Study area

Doce River watershed (Figure 1) is located in the Southeast Region of Brazil, situated in the states of Minas Gerais and Espírito Santo, between the parallels 17°45' and 21°15' of south latitude and the medians 39°55' e 43°45' of west longitude. The Doce River has total extension of 853 km and drainage area of 83,465 km^2 (Coelho, 2007), of which 86% belongs to the Minas Gerais State and the remainder (14%) to the Espírito Santo State, being, therefore, an interstate watershed.

The Doce River Watershed population is estimated in 3.5 million inhabitants, distributed in 228 municipalities, being 200 of Minas Gerais and 28 of Espírito Santo. The economic activity in the watershed is very diversified: (i) in farming, stands out traditional crops, coffee culture, sugar cane, beef and dairy cattle breeding, pig farming, among others; (ii) in the agroindustry, it's highlighted the production of sugar and alcohol; (iii) has the largest steelmaking complex of Latin America, to which are associated mining and reforestation companies; (iv) additionally, stand out cellulose and dairy industries, commerce and services related to industrial complexes, as well as electric power generation, with great potential for exploitation (Comitê da Bacia Hidrográfica do Rio Doce, 2022).

Data

The monthly average flow time series data cover the period from 1987 to 2011, totaling 25 years of observations (or 300 months), obtained through the Hydrological Information System (Brasil, 2022), of the National Water Agency and Basic Sanitation (ANA). The Colatina station was used to carry out the imputations, and in cases of multivariate imputation methodologies, support stations were utilized. The support stations were standardized following the same base period (1987 to 2011), and as a prerequisite, the correlation between stations was evaluated, mainly for the central study object station (Colatina). The selected stations are identified in this study as: Fazenda Cachoeira D'Antas (E1), Cachoeira dos Óculos Montante (E2), Belo Oriente (E3), Governados Valadares (E4), Tumiritinga (E5) and Colatina (E6). Figure 1 shows the station spatialization along the watercourse. Information regarding the global positioning characteristics and the stations' drainage area and the drainage area are shown in Table 1.

As a precept for multivariate imputation methods, it is necessary that the data present homogeneity between them. The Pearson correlation coefficient (ϱ) between stations can be seen in Table 2. The lowest correlation value (0.9373) was found between stations E1 and E6, which was expected because they are further away from each other. The high values enable the multivariate imputation methodologies use and that one station can provide reliable information when imputing missing data from another.

Imputation of missing data

Imputation methodologies can be classified in two ways: by the imputation amount and the need to use auxiliary series or not. When it comes to the imputation type, single imputation is that which occurs when missing data are imputed only once). In turn, multiple imputation occurs when missings undergoes numerous imputations and, followed by inference analysis, the appropriate value for imputation is defined. When the imputation method only requires the series of interest in generating information to conduct the imputation, the technique is called univariate. In cases where support series are necessary to carry out the imputation in the series of interest, it is called a multivariate procedure.

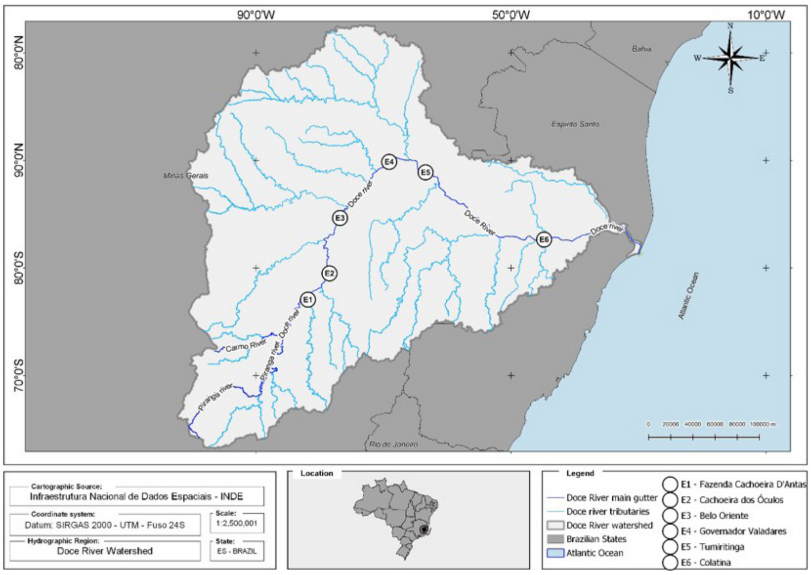

Figure 1. Location of the Doce River Watershed.

Table 1. Stations selected for the study.

Identification	Coordinates (degrees)		Zone	Drainage Area (km^2)
	Lat. (S)	Long. (W)		
E1	19°59’39.84”	42°40’27.84”	23 K	10,100
E2	19°46’36.84”	42°28’35.04”	23 K	15,900
E3	19°19’46.92”	42°22’33.96”	23 K	24,200
E4	18°52’59.16”	41°57’02.88”	24 K	40,500
E5	18°58’15.96”	41°38’30.12”	24 K	55,100
E6	19°32’00.00”	40°37’46.92”	24 K	76,400

Note: Lat.: Latitude; Long.: Longitude.

Table 2. Pearson correlations for monthly mean flow variable data between stations.

	E1	E2	E3	E4	E5	E6
E1	1					
E2	0.9861	1				
E3	0.9792	0.9857	1			
E4	0.9609	0.9688	0.9893	1		
E5	0.9457	0.9535	0.9771	0.9937	1	
E6	0.9373	0.9481	0.9702	0.9838	0.9824	1

The imputation methods utilized in this study were grouped as follows: (i) single univariate imputation, represented by the arithmetic mean (AM), median (M) and interpolations (Spline and Stineman) methodologies; (ii) single multivariate imputation, which refers to Regional Weighting (RW) and Multiple Linear Regression (MLR); and finally, (iii) multiple multivariate imputation, represented by the multiple imputation (IM) and maximum likelihood (ML) methodologies.

Methods of single univariate imputation

Understood as basic approaches for imputing failures, missing data are replaced by mean or median, attributed using the data present (or remaining) in the series of interest. Equations 1 and 2 represent, respectively, AM and M.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$m(x) = \begin{cases} x_{\left(\frac{n+}{}\right)} & \textit{if n is an odd number} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+\right)}}{2} & \textit{if n is an even number} \end{cases} \tag{2}$$

In Equation 1, $\overline{x}$ represents the mean and in Equation 2, $m(x)$ represents the median. In both equations, n is the quantity of data.

Methods of single multivariate imputation

The equations for the MLR and RW methodologies, are represented, respectively, by Equations 3 and 4.

$$X = \beta_0 + \sum_{i=1}^{n} \beta_i Y_i + \varepsilon \tag{3}$$

In Equation 3, X represents the series dependent of the linear equation; β_0 is the linear coefficient vector; β_i represents the angular coefficients; Y_i denotes the independent station; and ε are the residuals of the model.

$$X = \frac{1}{n}\sum_{i=1}^{n} \frac{N_X}{N_i} D_i \tag{4}$$

In Equation 4, N_X and N_i denote, respectively, the monthly mean flow data for the station with missing data to be imputed and the monthly mean flow of order “i” of the neighbor station (m^3.s^{-1}); D_i denotes the observed values of order “i” in the neighbor stations during the month of occurrence in the station with the data to be imputed (m^3.s^{-1}); and n is the number of neighbor stations considered.

Imputation methods by interpolation

In this study two advanced approaches of nonlinear interpolation were applied. The first one, called Spline interpolation, is applied over a Spline function defined in Equation 5.

$$S(x) = \begin{cases} P_0(x), x \in (-\infty, \tau_l); \\ P_j(x), x \in [\tau_j, \tau_{j+1}), j = 1, \ldots r-1; \\ P_0(x), x \in [\tau_r, \infty); \end{cases} \tag{5}$$

where $S:\mathbb{R} \to \mathbb{R}, \{P_0(x), P_2(x), \ldots P_r(x)\}$ is a sequence of cubic polynomials, and $\tau_l < \tau_2 < \ldots < \tau_r$ is a sequence of real numbers called knots of the Spline space.

The second, called Stineman interpolation, utilizes the rational interpolation that is intended to provide better results than Spline interpolation in the case of sudden changes in slope. The equating of Stineman interpolation is described in the following according to the work of Demirhan & Renwick (2018). Consider x_j and y_j to be rectangular coordinates of the j-th point on a curve and let y'_j be the slope of the curve at the j-th point for $j=1,\ldots,n$ and $x_j < x_{j+1}$ for $j=1,\ldots,n-1$. Thereafter, the Stineman interpolation is applied to calculate the interpolated value of y via Algorithm 1.

Algorithm 1:

1. Given x that satisfies the condition $\ddot{u}_j \leq \ \leq \ {}_{j+1}$, calculate the inclination of the line segment that joining two points by $s_j = \frac{(y_{j+1}-y_j)}{(x_{j+1}-x_j)}$.
2. Calculate the ordinate corresponding to x by $y_0 = y_j + s_j(x - x_j)$.
3. Calculate the vertical distance of the point $(x - y_0)$ to a line through $(x_j - y_j)$ with slope y'_j by $\Delta y_j = y_j + y'_j(x - x_j) - y_0$ for j and $j+1$.
4. Calculate the interpolated value by $y = y_0\left[\Delta y_j \Delta y_{j+1}(x - x_{j+1} + x - x_j)\right] / \left[(\Delta y_j - \Delta y_{j+1})(x_{j+1} - x_j)\right] \text{if } \Delta y_j \Delta y_{j+1} < 0$.

To implement Algorithm 1, one needs to know the values of the slopes y'_j. If these are not initially known, it is necessary to apply Algorithm 2 to perform the calculation of theses slopes. For this, let I, J and K be any three consecutive points that satisfy $\left(IJ'\right) > y'_j > \left(JK'\right)$ or $\left(IJ'\right) < y'_j < \left(JK'\right)$, where $(\cdot)$ denotes the slope of the internal segment of the curve.

Algorithm 2:

1. For internal points, the slope is calculated by $y'_j = \frac{(y_j - y_i)\left[(x_k - x_j)^2 + (y_k - y_j)^2\right] + (y_k - y_j)\left[(x_j - x_i)^2 + (y_j - y_i)^2\right]}{(x_j - x_i)\left[(x_k - x_j)^2 + (y_k - y_j)^2\right] + (x_k - x_j)\left[(x_j - x_i)^2 + (y_j - y_i)^2\right]}$.
2. For final points, calculate the slope for the final point m by $y'_m = 2s - y'_j$, where s is the slope of the line segment that joins the points J and the final points.

Method of multiple multivariate imputation

The IM methodology, proposed by Rubin (1987), appeared as a more flexible manner and alternative to the maximum likelihood methods when there exists a large quantity of missing data (Schafer & Graham, 2002). This technique enables the inclusion of uncertainty of imputation in the results, which is the major limitation associated to the single imputation techniques (Nunes et al., 2009). The technique is based on three steps: (i) in the first, are generated m sets of imputed data; (ii) utilizing standardized procedures, m analyses are made in the set of imputed data; and, (iii) the results of the m analyses are combined to obtain the necessary inferences to choose the values considered in the final imputation.

In the first steps, the imputation techniques have to preserve the relation of the missing and present observations and take into account the missing data pattern and the mechanisms of absence. Having realized the m imputations, the step (ii) of the IM can be performed, i.e., the m databases are analyzed by traditional methods of analysis that in this study, is the mean predictive correspondence. As an outcome, it is possible to combine the results and utilize the rules proposed by Rubin (Rubin, 1987; Nunes et al., 2009). In step (iii), from each analysis, the estimates for the interest parameter Q are obtained. Let, $\hat{Q}_j$ for $j=1,2,\ldots,m$, for Q equal to any scalar measure, the combined estimate will be the mean of the individual estimates (Nunes et al., 2009), according to Equation 6.

$$\bar{Q} = \frac{1}{m}{}_{j=1}^{m} \sum_{i=1}^{m} \hat{Q}_j \tag{6}$$

For the combined variance, it is necessary to calculate the variance inside the imputations (Equation 7) and the variance between the imputed databases (Equation 8).

$$\bar{U} = \frac{1}{m}{}_{j=1}^{m} \tag{7}$$

$$B = \frac{1}{m-1}{}_{j=1}^{m} \sum_{i=1}^{m} \left(\hat{Q}_j - \bar{Q}\right)^2 \tag{8}$$

Lastly, the total variance, which is the combined variance, is given by Equation 9.

$$\mathrm{T} = \bar{U} + 1 + \frac{1}{m}B \tag{9}$$

The main idea of the ML methodology is estimating the parameters that would maximize the probability of the distribution of the observed data (with or without the presence of missing values), allowing then, estimate the missing data. The likelihood function is represented as in Equation 10.

$$\mathrm{L}(\Theta \mid \mathrm{Y}) = \Pi \mathrm{f}(Y_i \mid \Theta) \tag{10}$$

where $\mathrm{f}(\mathrm{Y} \mid \Theta)$ is a density function that describes the model responsible for generating the data, Y are the data and Θ is a set of unknown parameters that rules the distribution of Y, from which it is known that ir belongs to Ω_θ. That is, $\mathrm{f}(\mathrm{Y} \mid \Theta)$ is a function of the parameter vector $\Theta \in \Omega_\theta$ given Y, proportional to the density function. Establishing the model and the parameter vector Θ, $\mathrm{f}(\mathrm{Y} \mid \Theta)$ can be used to sample missing values (Allison, 2002).

Due to the impositions associated with the analytical process, numerical methods are useful in the parameter estimation stage. Consolidating itself as an efficient alternative, the EM algorithm is an iterative procedure that consists in repeating two steps: estimation (E) and maximization (M). Consider a dataset with observed and missing data, with density function given by $\mathrm{p}(y^c \mid \Theta)$ whereby, $\mathrm{l}(\Theta, y^c)$ represents the log-likelihood function of the complete and observed data. The algorithm suggests that initially one finds the expected value of the logarithm of the likelihood function (step E) and next its maximum (step M), according to Equation 11:

$$\mathrm{Q}(\Theta \mid \Theta^{(k)}) = \mathrm{E}\left(l^c(\Theta, y^c)\right| y^c, \Theta^{(k)}) \tag{11}$$

In the step M, is aimed to find $\Theta^{(k+1)}$ that maximizes $Q\left(\Theta \mid \Theta^{(k)}\right)$. The process is repeated until convergence is achieved, by means of a stop criterion, such as $\left\|\Theta^{(k+1)}\Theta^{(k)}\right\| < \varepsilon$, when the difference between the estimated values of the parameters in two consecutive iterations is lower than the pre-established.

Mechanisms of missing data

The mechanisms of missing data describe the relations between lost values and the probability of absence, informing the cause of missing in the data. Little & Rubin (2002) define three general theoretical mechanisms extensively utilized in the literature, known as (i) *Missing Completely at Random* (MCAR), (ii) *Missing at Random* (MAR) and (iii) *not Missing at Random* (NMAR). According to Hamzah et al. (2022), in flow data studies, missing data is described as MCAR due to the episode that results in failures not being due to influences from external variables. This mechanism is called ignorable, and there is, therefore, no need for its incorporation in the failure estimation process.

Missing data also can be characterized in patterns (McKnight et al., 2007). The main patterns of missing data are discussed by Hamzah et al. (2022) known as (i) general or random pattern, (ii) unitary non-response pattern and (iii) monotonous pattern. Due to the characteristics of missing data in flow variables being random and, very frequently, with a block pattern, patterns i and ii are of interest in the present study. Once the mechanism and pattern of missing data occurrence were established, artificial failures were simulated following the proportions of 5, 20 and 40% failures. In the block pattern, the following gaps were created:

- 5%: a block of 12 missing observations;
- 20%: four blocks of 12 missing observations in each;
- 40%: eight blocks of 12 missing observations in each.

Considering the perspective of generating consistent results, a total of one thousand (1000) simulation repetitions were replicated for each pattern and percentage of missing data.

SARIMA model

The Seasonal Autoregressive Integrated Moving Average (SARIMA) model is useful because it incorporates the seasonality component in its modeling process, a characteristic present in the Doce River flow regime (Pinto et al., 2015; Bleidorn et al., 2019). Let $Z_t = \{Z_t ; t \in \mathbb{Z}\}$ be a linear process represented by Equation 12:

$$\Phi\left(B^S\right)\phi(B)\nabla^d Z_t = \Theta\left(B^S\right)\theta(B)\varepsilon_t \tag{12}$$

where *s* is called seasonal period of the process and $\varepsilon_t \sim \mathrm{WN}\left(0, \sigma_\varepsilon^2\right)$, in which $\varepsilon_t \sim \mathrm{WN}$ is white noise (WN), defined as a sequence of uncorrelated random variables with zero mean and constant variance over time (Wei, 2006). The operator $\nabla^{\mathbf{d}}$, with $\mathbf{d} = (d, D)$ and d, D non-negative integer numbers that represent, respectively, the number of simple and seasonal differences applied to the process Z_t, defined according to Equation 13:

$$\nabla^{d} = (1-B)^{d}\left(1-B^{S}\right)D. \tag{13}$$

We have that $\Phi\left(z^S\right) = 1 - \Sigma_{i=1}^{P}\Phi_i z^{is}, \phi(z) = 1 - \Sigma_{j=1}^{P}\Phi_j z^j, \Theta\left(z^S\right) = 1 - \Sigma_{k=1}^{Q}\Theta_k z^{ks}$ and $\theta(z) = 1 - \Sigma_{l=1}^{q}\Theta_l z^l$ are polynomials of order $P, p, Q, q \in \mathbb{N}$, respectively, with $z \in \mathbb{C}$, where, $\mathbb{C}$ represents the set of complex numbers, $\mathbb{N}$ denotes the set of natural numbers, and $\{\Phi i\}, \{\phi j\}, \{\Theta k\}$ and $\{\theta l\}$ are sequences of real numbers. The process Z_t with representation given in (12) is denominated seasonal multiplicative ARIMA (SARIMA) of order $(p, d, q) \times (P, D, Q)s$.

Modeling methodology

The SARIMA model is built on the modeling methodology proposed by Box and Jenkins, based on an iterative cycle that contains four steps: (i) identification; (ii) estimation; (iii) residual adequacy; and, (iv) forecasting. In the first step, resources are used to understand the behavior of the series, such as visual analysis of it, of the correlograms (Autocorrelation Function – ACF and Partial Autocorrelation Function – PACF), of the spectral decomposition and the use of tests for detecting trend and seasonality. In this step, adjustment indicators are used, mainly the Akaike Information Criterion (AIC) (Akaike, 1974). In the second step, the parameters of the candidate models are estimated by several approaches, being recurrent the use of the maximum likelihood. In the third step, it is verified that the model is adequate via residual analysis. Having satisfied the above conditions, the model is considered in the steps of adjustment and forecasting.

Semi-automatization of the model choice

Taking into consideration that the model choice is extremely important, the founds in Pinto et al. (2015) and Bleidorn et al. (2019) studies allows to infer the presence of the seasonality component and the characteristic of non-stationarity in the flow series for the same time series studied in this research. Hence, to bypass the non-linear structures and make the series stationary (Box & Jenkins, 1976), it was defined the differencing d and D = 1, and the data transformation via natural logarithm. Pinto et al. (2015) study allows to verified, for the same studied series (E6), a behavior of the correlograms that indicated an autoregressive part (AR) of order 1 and of moving average (MA) of order 2. Given this information, it was possible to define the initial conditions for the semi-automation of model choice.

The semi-automatization was established to remove the subjectivity in the model choice, consequently avoiding the distortion of the effect of using reconstructed databases on modeling and forecasting performance. This formulation consisted of the following steps: (a) all possible combinations resulting from models with a maximum of three parameters in each part of the model (AR and MA of the ordinal a seasonal parts), and the fixed unit differentiation (d and D = 1), considering that this conditioning of at most three parameters of each model component, follows the parsimony principle. The combination of these configurations allowed to generate a total of 254 candidate models. In the next step (b) was to order in increasing magnitude the AIC values. In the third step (c), the analyzed model parameters necessarily should have significance in the level of 99,5% (*p*-valor < 0,05).

Finally, the last step (d), consisted of the analysis of normality and non-autocorrelation. The chosen model was the one that combined the lowest AIC value, significant parameters and normal and non-autocorrelated residuals.

Fitting the model to observed data

Initially, the observed series was used to adjust the SARIMA model, allowing it to be used as a parameter of adjustment analysis to the reconstructed databases. The data were divided into two groups: the first, for adjustment, from January 1987 to December 2011, resulting in 300 months of observations, and the second, from January to December 2012 to evaluate the forecast accuracy one step ahead (12 months). Table 3 presents the AIC and Bayesian (BIC) values (Akaike, 1978) and the normality and non-autocorrelation tests of residuals. Tables 4 and 5 present, respectively, the model's adjustment and prediction performance indicators values.

Performance indicators

The model performance in the adjustment and forecasting stages for the different imputed datasets, were evaluated using the performance indicators Absolute BIAS, Root Mean Squared Error (RMSE), Mean Absolut Percentual Error (MAPE), Nash-Sutcliffe (NSE) (Nash & Sutcliffe, 1970), concordance index (d_2) and Pearson correlation coefficient (ϱ), presented by the Equations 14, 15, 16, 17 and 18, respectively.

The Absolute BIAS quantifies the estimates of underestimation and overestimation with respect to the mean observations.

$$\frac{1}{n}\sum_{i=1}^{N}(x_i - \tilde{x}_i) \tag{14}$$

The RMSE is the quadratic difference between the forecasted or adjusted values and their respective true values. In general, lower values indicate better performance.

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{N}(x_i - \tilde{x}_i)^2} \tag{15}$$

The MAPE is a measure of precision of adjustment and forecasting of a model. Lower values indicate good performance.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(\frac{x_i - \tilde{x}_i}{x_i}\right)\right| \tag{16}$$

ϱ describes the relation between the variables and the closer it is to the extremes (-1 or 1) the stronger the correlation is.

$$\rho \;=\frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N}(y_i - \bar{y})^2\right]}} \tag{17}$$

NSE is used to evaluate the predictive capacity of the hydrological models. The values of NSE vary between -∞ and 1, with values closer to 1 representing good performance.

$$NSE \;=\; 1 - \frac{\sum_{i\ \ 1}^{n}(x_i \quad y_i)^2}{\sum_{i\ \ 1}^{n}(x_i - \overline{x_i})^2} \tag{18}$$

d_2 reflects the concordance between adjustment/forecasting of the model with the observed data, with the desired value of 1 indicating perfect concordance.

$$d_2 = 1 - \left[\frac{\sum_{i=1}^{n}(x_i - \tilde{x}_i)^2}{\sum_{i=1}^{n}\left(\left|x_i - \overline{y_i}\right| + \left|\tilde{x}_i - \overline{y_i}\right|\right)^2}\right] \tag{19}$$

where: n represents the number of observations in the phase of adjustment or forecast; x indicates the observed values; y the adjusted or forecasted values; $\bar{x}$ is the mean of the observations and

Table 3. Statistical tests of normality and correlation of residuals from the model adjusted to observed data.

Model	Adjustment quality		*p*-valor			
	AIC	BIC	S-W	J-B	L-B	B-P
$(1,1,1)(0,1,1)_{12}$	171.4758	186.1137	2.803×10^{-06}	0.0001	0.2880	0.3192

Table 4. Quality-of-fit measurements of selected models.

Quality Measures					
BIAS	RSME	MAPE	NSE	*d*	ϱ
41.9764	21.0488	206.2328	0.6129	0.8599	0.9090

Table 5. Forecast quality measures of selected models.

h	Quality Measures					
	BIAS	RSME	MAPE	NSE	d_2	ϱ
12	-39.4258	167.1951	331.9439	0.4867	0.7673	0.9020

$\bar{y}$ the estimated values in the phase of adjustment or forecasting of the models.

Computational resources

Electronic spreadsheets were used for organizing the data and the software R (R Development Core Team, 2021) was utilized to simulate missing data and its reconstructions, adjustment of the Box and Jenkins methodology using the SARIMA model and the analysis of its performance. The *imputeTS* package was used to carry out the imputations using the Stineman and Spline methodologies and the *mice* and *mtsdi* packages to carry out the imputations using the IM and ML methodologies, respectively.

RESULTS AND DISCUSSION

Descriptive analysis of data

For an initial understanding of the time series under study, some descriptive measures are presented in Table 6. The monthly mean flow was of 805.09 $m^3.s^{-1}$ with standard deviation of 578.48 $m^3.s^{-1}$ and coefficient of variation of 71.85%. The high values of the standard deviation and coefficient of variation indicate that the mean has little representativeness, associated to the seasonal characteristic of the series, as confirmed by the works of Pinto et al. (2015) and Bleidorn et al. (2019) for the same series under study. The series has asymmetry of 1.82 and kurtosis of 3.59, indicating that the distribution is not normal and has heavier tails. The seasonality property of the studied series can be visualized in Figures 2 and 3, with a well-defined interannual variability pattern with periods of greater flows magnitude (November to April) followed by periods of smaller flows magnitude (May to October).

Effect of the imputation methodologies in the quality of adjustment and forecasting

Tables 7 and 8 show the values of the performance indicators of adjustment of the model to the imputed series with random and in block pattern of missing data, respectively. In general, it is possible to verify that, while the percentages of imputed failures increase, the model presents loss of quality, with emphasis on the high values of the indicators BIAS, RMSE, MAPE to lows of the NSE, d_2 and ϱ for the adjustment of the reconstructed series by RW, MLR, M and AM methodologies in both patterns of missing data. For the block pattern, Spline and Stine techniques also limited the model performance. However, it is possible to verify that, even under a critical scenario of failures imputation (40%), in both patterns of missing data, the IM and ML methodologies are efficient to the point that there are no significant changes in the model quality indicators, with low values of BIAS, RMSE, MAPE and high of NSE, d_2 and ϱ.

Table 6. Descriptive measures of Doce River flow.

Descriptive measures	Value
Minimum value ($m^3.s^{-1}$)	194.20
Maximum value ($m^3.s^{-1}$)	3,469.91
Mean ($m^3.s^{-1}$)	805.09
Median ($m^3.s^{-1}$)	582.33
Coefficient of variation (%)	71.85
Standard deviation ($m^3.s^{-1}$)	578.48
Asymmetry	1.82
Kurtosis	3.59

Tables 9 and 10 show the performance indicators in the forecasting step of the model, considering the missing data imputations with random and in block pattern, respectively. Due to the loss of quality in the adjust step, it was expected that the model's performance would be compromised in the forecasting stage for the series reconstructed by RW, MLR, M and AM in both failure patterns and by Stine and Spline methodologies for the pattern in block. Just like in the adjustment step, it is possible to verify low values of the performance indicators BIAS, RMSE, MAPE and high NSE, d_2 and ϱ, for the forecast performance of the model fitted to the series reconstructed by the Stine and Spline methodologies for the random loss pattern and for IM and ML methodologies for both missing data patterns.

Relative difference of the quality of adjustment and forecasting

To facilitate the understanding of the behavior of the quality measures, it is shown in Tables 11 and 12 the relative difference of fit performance measures and in Tables 13 and 14,

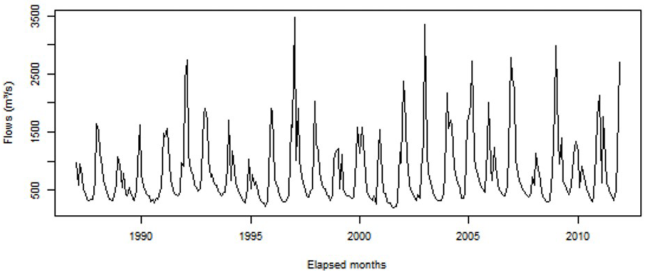

Figure 2. Graphic of the mean flow time series of Doce River.

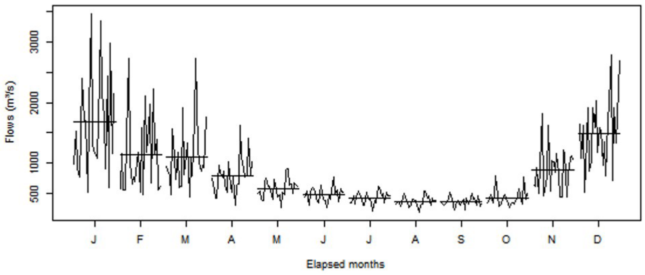

Figure 3. Graphical analysis of seasonality of monthly mean flows time series of Doce River.

Table 7. Quality measures of adjustment for the imputed data with random pattern.

Method	Failures	Quality Measures					
		BIAS	**RMSE**	**MAPE**	**NSE**	d	ϱ
AM	5%	52.7438	21.2943	207.2785	0.6038	0.8485	0.9090
	20%	81.7548	22.7130	219.8330	0.5492	0.7979	0.9092
	40%	111.3557	25.5101	252.2228	0.4314	0.6892	0.9094
M	5%	57.6646	21.3616	207.2529	0.6013	0.8468	0.9092
	20%	101.3630	23.1657	219.6816	0.5311	0.7867	0.9091
	40%	149.0176	26.5386	253.5870	0.3846	0.6588	0.9092
RW	5%	55.2947	21.1764	206.0494	0.6081	0.8559	0.9089
	20%	226.4204	26.0459	262.4324	0.4072	0.7433	0.9062
	40%	552.6584	42.9540	552.6584	-0.6120	0.3112	0.9019
MLR	5%	59.2678	21.2450	206.3229	0.6056	0.8540	0.9087
	20%	236.4774	26.5094	268.8146	0.3860	0.7315	0.9049
	40%	559.7321	43.4211	559.7321	-0.6472	0.2968	0.9004
Spline	5%	40.4013	20.7365	203.1147	0.6243	0.8656	0.9138
	20%	38.9052	19.8956	195.3544	0.6541	0.8779	0.9228
	40%	41.3428	19.4982	189.7887	0.6678	0.8848	0.9285
Stine	5%	42.3592	20.8275	203.6450	0.6210	0.8631	0.9133
	20%	46.1429	20.2043	197.0007	0.6433	0.8701	0.9218
	40%	50.8173	19.8482	193.8725	0.6558	0.8757	0.9262
IM	5%	42.0289	21.0353	206.4785	0.6134	0.8599	0.9088
	20%	42.3063	21.0244	207.3706	0.6138	0.8596	0.9075
	40%	42.5619	21.0404	209.1222	0.6132	0.8587	0.9050
ML	5%	42.0544	21.0425	206.4270	0.6131	0.8598	0.9087
	20%	42.3904	21.0512	207.2645	0.6128	0.8591	0.9073
	40%	42.3017	21.0827	208.9582	0.6116	0.8581	0.9047

Table 8. Quality measures of adjustment for the imputed data with in block pattern.

Method	Failures	Quality Measures					
		BIAS	**RMSE**	**MAPE**	**NSE**	d_2	ϱ
AM	5%	43.4514	21.4090	213.7368	0.5995	0.8522	0.8946
	20%	95.3017	24.8093	235.6318	0.4622	0.7720	0.8686
	40%	70.2809	27.7276	287.0185	0.3282	0.7331	0.7029
M	5%	48.7386	21.5094	213.5885	0.5957	0.8508	0.8974
	20%	113.4883	25.5766	241.4153	0.4284	0.7575	0.8613
	40%	116.3269	28.9237	291.8409	0.2690	0.7169	0.7096
RW	5%	54.4271	21.8005	213.4348	0.5847	0.8501	0.9041
	20%	186.6006	30.1628	303.7369	0.2051	0.7115	0.6380
	40%	324.8904	39.9292	456.2904	-0.3929	0.6162	0.4323
RLM	5%	62.2207	22.4828	223.2875	0.5583	0.8451	0.8820
	20%	166.6538	32.2723	328.2497	0.0900	0.6997	0.5325
	40%	326.9094	40.0867	454.5443	-0.4039	0.6142	0.3920
Spline	5%	47.7851	25.0570	256.5995	0.4514	0.8323	0.8203
	20%	54.9519	26.4562	262.3972	0.3884	0.7882	0.8053
	40%	-210.7843	63.7762	567.0233	-2.5536	0.5045	0.5130
Stine	5%	61.4936	22.5159	220.9544	0.5570	0.8413	0.8783
	20%	85.0125	27.4814	271.9021	0.3401	0.7631	0.7652
	40%	56.0532	29.2919	311.1370	0.2503	0.7337	0.6592
IM	5%	47.2351	21.1062	206.9831	0.6107	0.8579	0.9069
	20%	46.2231	21.4987	210.5306	0.5961	0.8525	0.9052
	40%	34.1526	21.8219	217.7988	0.5839	0.8512	0.8990
ML	5%	45.6403	21.0768	207.1394	0.6118	0.8593	0.9068
	20%	42.8350	21.2958	209.5853	0.6037	0.8575	0.9070
	40%	43.6186	21.8176	214.5231	0.5841	0.8492	0.9008

Table 9. Quality measures of forecasting for imputed data with random pattern.

Method	Failures	Quality Measures					
		BIAS	**RMSE**	**MAPE**	**NSE**	d_2	ϱ
AM	5%	-21.2196	170.1698	331.9334	0.4683	0.7434	0.9020
	20%	51.3489	181.3630	313.7535	0.3961	0.6546	0.9020
	40%	102.4067	196.0498	326.8236	0.2943	0.5259	0.9020
M	5%	-4.4286	171.2643	322.1710	0.4615	0.7359	0.9020
	20%	64.9792	181.2698	306.1544	0.3967	0.6540	0.9020
	40%	149.9682	200.6252	314.7488	0.2610	0.4978	0.9020
RW	5%	-37.9093	167.9537	337.7092	0.4821	0.7635	0.9020
	20%	197.3628	188.5563	311.2692	0.3472	0.6354	0.8951
	40%	569.7914	267.5029	569.7914	-0.3137	0.2562	0.9020
MLR	5%	-21.9195	167.7733	325.2059	0.4832	0.7604	0.9020
	20%	213.1987	191.8399	312.9804	0.3243	0.6166	0.9020
	40%	593.8795	272.0437	593.8795	-0.3587	0.2472	0.9230
Spline	5%	-106.5841	177.2372	386.3699	0.4232	0.7468	0.8811
	20%	-93.3112	173.2579	386.1265	0.4488	0.7494	0.9230
	40%	-79.0259	185.0240	406.6258	0.3715	0.6786	0.9160
Stine	5%	-95.0408	175.4720	374.5865	0.4347	0.7511	0.9020
	20%	-65.6582	172.2744	368.1666	0.4551	0.7440	0.9230
	40%	-34.3310	181.4035	377.5423	0.3958	0.6839	0.9370
IM	5%	-40.9106	167.0590	335.0690	0.4876	0.7677	0.9020
	20%	-70.8730	176.2720	362.1946	0.4295	0.7397	0.8811
	40%	-14.5344	168.3016	323.3994	0.4799	0.7555	0.9020
ML	5%	-47.5680	167.7267	341.4603	0.4835	0.7664	0.9020
	20%	-25.9219	168.3118	329.1422	0.4799	0.7591	0.9020
	40%	-7.3699	168.0394	317.0583	0.4815	0.7555	0.9020

Table 10. Quality measures of forecasting for imputed data with in block pattern.

Method	Failures	Quality Measures					
		BIAS	**RMSE**	**MAPE**	**NSE**	d_2	ϱ
AM	5%	-45.7355	168.5402	344.4893	0.4784	0.7581	0.9020
	20%	127.9092	191.6874	302.3053	0.3254	0.6004	0.8811
	40%	-314.3138	196.4584	563.2575	0.2914	0.7243	0.8671
M	5%	-11.8006	169.8179	323.9684	0.4705	0.7446	0.9020
	20%	177.6978	196.4043	309.4514	0.2918	0.5707	0.8741
	40%	-349.8510	208.5124	605.4242	0.2017	0.6976	0.8251
RW	5%	-27.1742	168.9876	329.6567	0.4757	0.7567	0.9020
	20%	613.2388	277.9278	613.2388	-0.4181	0.2208	0.9090
	40%	-495.4309	273.7219	841.6166	-0.3755	0.0000	0.9020
MLR	5%	-84.9160	168.6021	366.3130	0.4781	0.7715	0.9020
	20%	696.6152	301.1155	696.6152	-0.6646	0.1096	0.8391
	40%	-128.9501	197.0022	481.6821	0.2874	0.6043	0.6293
Spline	5%	-236.5789	183.1726	479.4728	0.3840	0.7659	0.9020
	20%	412.3080	238.9207	424.2963	-0.0479	0.3616	0.8531
	40%	-667.3864	285.5110	859.0869	-0.4965	0.6384	0.7832
Stine	5%	-101.0516	177.7436	387.0325	0.4199	0.7385	0.8811
	20%	533.1416	259.6469	533.1416	-0.2376	0.2940	0.8531
	40%	-271.4511	193.0208	543.8541	0.3160	0.7283	0.8671
IM	5%	-30.6386	166.5110	328.2590	0.4909	0.7654	0.9020
	20%	3.9594	175.3916	322.7458	0.4352	0.7209	0.9020
	40%	-12.8244	181.1161	342.9227	0.3977	0.7054	0.8811
ML	5%	-48.6574	166.2311	338.8334	0.4926	0.7719	0.9020
	20%	23.5632	170.2113	309.1330	0.4681	0.7431	0.9020
	40%	14.5180	177.4495	320.5190	0.4219	0.7133	0.8811

Table 11. Relative difference in the quality of adjustment of the model to the observed and imputed data with random pattern.

Method	Quality Measures						IM	ML
	AM	M	RW	MLR	Spline	Stine		
				5%				
BIAS	25.65	37.37	31.73	41.19	-3.75	0.91	0.13	0.19
RMSE	1.17	1.49	0.61	0.93	-1.48	-1.05	-0.06	-0.03
MAPE	0.51	0.49	-0.09	0.04	-1.51	-1.25	0.12	0.09
NSE	-1.48	-1.89	-0.78	-1.19	1.86	1.32	0.08	0.03
d	-1.33	-1.52	-0.47	-0.69	0.66	0.37	0.00	-0.01
ϱ	0.00	0.02	-0.01	-0.03	0.53	0.47	-0.02	-0.03
				20%				
BIAS	94.76	141.48	439.40	463.36	-7.32	9.93	0.79	0.99
RMSE	7.91	10.06	23.74	25.94	-5.48	-4.01	-0.12	0.01
MAPE	6.59	6.52	27.25	30.35	-5.27	-4.48	0.55	0.50
NSE	-10.39	-13.35	-33.56	-37.02	6.72	4.96	0.15	-0.02
d_2	-7.21	-8.51	-13.56	-14.93	2.09	1.19	-0.03	-0.09
ϱ	0.02	0.01	-0.31	-0.45	1.52	1.41	-0.17	-0.19
				40%				
BIAS	165.28	255.00	1216.59	1233.45	-1.51	21.06	1.39	0.77
RMSE	21.20	26.08	104.07	106.29	-7.37	-5.70	-0.04	0.16
MAPE	22.30	22.96	167.98	171.41	-7.97	-5.99	1.40	1.32
NSE	-29.61	-37.25	-199.85	-205.60	8.96	7.00	0.05	-0.21
d_2	-19.85	-23.39	-63.81	-65.48	2.90	1.84	-0.14	-0.21
ϱ	0.04	0.02	-0.78	-0.95	2.15	1.89	-0.44	-0.47

Table 12. Relative difference in the quality of adjustment of the model to the observed and imputed data with in block pattern.

Method	Quality Measures						IM	ML
	AM	M	RW	MLR	Spline	Stine		
				5%				
BIAS	3.51	16.11	29.66	48.23	13.84	46.50	12.53	8.73
RMSE	1.71	2.19	3.57	6.81	19.04	6.97	0.27	0.13
MAPE	3.64	3.57	3.49	8.27	24.42	7.14	0.36	0.44
NSE	-2.19	-2.81	-4.60	-8.91	-26.35	-9.12	-0.36	-0.18
d_2	-0.90	-1.06	-1.14	-1.72	-3.21	-2.16	-0.23	-0.07
ϱ	-1.58	-1.28	-0.54	-2.97	-9.76	-3.38	-0.23	-0.24
Method	Quality Measures						IM	ML
	AM	M	RW	MLR	Spline	Stine		
				20%				
BIAS	127.04	170.36	344.54	297.02	30.91	102.52	10.12	2.05
RMSE	17.87	21.51	43.30	53.32	25.69	30.56	2.14	1.17
MAPE	14.26	17.06	47.28	59.16	27.23	31.84	2.08	1.63
NSE	-24.59	-30.10	-66.54	-85.32	-36.63	-44.51	-2.74	-1.50
d_2	-10.22	-11.91	-17.26	-18.63	-8.34	-11.26	-0.86	-0.28
ϱ	-4.44	-5.25	-29.81	-41.42	-11.41	-15.82	-0.42	-0.22
				40%				
BIAS	67.43	177.12	673.98	678.79	-602.15	33.54	-18.64	3.91
RMSE	31.73	37.41	89.70	90.45	202.99	39.16	3.67	3.65
MAPE	39.17	41.51	121.25	120.40	174.94	50.87	5.61	4.02
NSE	-46.45	-56.11	-164.11	-165.90	-516.64	-59.16	-4.73	-4.70
d_2	-14.75	-16.63	-28.34	-28.57	-41.33	-14.68	-1.01	-1.24
ϱ	-22.67	-21.94	-52.44	-56.88	-43.56	-27.48	-1.10	-0.90

Table 13. Relative difference in the quality of forecasting of the model to the observed and imputed data with random pattern.

Method	Quality Measures						IM	ML
	AM	**M**	**RW**	**MLR**	**Spline**	**Stine**		
5%								
BIAS	-46.18	-88.77	-3.85	-44.40	170.34	141.06	3.77	20.65
RMSE	1.78	2.43	0.45	0.35	6.01	4.95	-0.08	0.32
MAPE	0.00	-2.94	1.74	-2.03	16.40	12.85	0.94	2.87
NSE	-3.78	-5.18	-0.95	-0.72	-13.05	-10.68	0.18	-0.66
d_2	-3.11	-4.09	-0.50	-0.90	-2.67	-2.11	0.05	-0.12
ϱ	0.00	0.00	0.00	0.00	-2.32	0.00	0.00	0.00
20%								
BIAS	-230.24	-264.81	-600.59	-640.76	136.68	66.54	79.76	-34.25
RMSE	8.47	8.42	12.78	14.74	3.63	3.04	5.43	0.67
MAPE	-5.48	-7.77	-6.23	-5.71	16.32	10.91	9.11	-0.84
NSE	-18.62	-18.49	-28.66	-33.37	-7.79	-6.49	-11.75	-1.40
d_2	-14.69	-14.77	-17.19	-19.64	-2.33	-3.04	-3.60	-1.07
ϱ	0.00	0.00	0.00	-0.77	2.33	2.33	-2.32	0.00
40%								
BIAS	-359.75	-480.38	-1545.22	-1606.32	100.44	-12.92	-63.13	-81.31
RMSE	17.26	19.99	59.99	62.71	10.66	8.50	0.66	0.51
MAPE	-1.54	-5.18	71.65	78.91	22.50	13.74	-2.57	-4.48
NSE	-39.53	-46.37	-164.45	-173.70	-23.67	-18.68	-1.40	-1.07
d_2	-31.46	-35.12	-66.61	-67.78	-11.56	-10.87	-1.54	-1.54
ϱ	0.00	0.00	2.33	0.00	1.55	3.88	0.00	0.00

Table 14. Relative difference in the quality of forecasting of the model to the observed and imputed data with block pattern.

Method	Quality Measures						IM	ML
	AM	**M**	**RW**	**MLR**	**Spline**	**Stine**		
5%								
BIAS	16.00	-70.07	-31.08	115.38	500.06	156.31	-22.29	23.42
RMSE	0.80	1.57	1.07	0.84	9.56	6.31	-0.41	-0.58
MAPE	3.78	-2.40	-0.69	10.35	44.44	16.60	-1.11	2.08
NSE	-1.71	-3.33	-2.26	-1.77	-21.10	-13.73	0.86	1.21
d_2	-1.20	-2.96	-1.38	0.55	-0.18	-3.75	-0.25	0.60
ϱ	0.00	0.00	0.00	0.00	0.00	-2.32	0.00	0.00
20%								
BIAS	-424.43	-550.71	-1655.43	-1866.90	-1145.78	-1452.27	-110.04	-159.77
RMSE	14.65	17.47	66.23	80.10	42.90	55.30	4.90	1.80
MAPE	-8.93	-6.78	84.74	109.86	27.82	60.61	-2.77	-6.87
NSE	-33.14	-40.05	-185.91	-236.55	-109.84	-148.82	-10.58	-3.82
d_2	-21.75	-25.62	-71.22	-85.72	-52.87	-61.68	-6.05	-3.15
ϱ	-2.32	-3.09	0.78	-6.97	-5.42	-5.42	0.00	0.00
40%								
BIAS	697.23	787.37	1156.62	227.07	1592.77	588.51	-67.47	-136.82
RMSE	17.50	24.71	63.71	17.83	70.77	15.45	8.33	6.13
MAPE	69.68	82.39	153.54	45.11	158.80	63.84	3.31	-3.44
NSE	-40.13	-58.56	-177.15	-40.95	-202.01	-35.07	-18.29	-13.31
d_2	-5.60	-9.08	-100.00	-21.24	-16.80	-5.08	-8.07	-7.04
ϱ	-3.87	-8.53	-100.00	-30.23	-13.17	-3.87	-2.32	-2.32

the relative difference for forecasting, considering the performance indicators obtained by the reference model (adjusted to observed data, see Tables 5 and 6).

It is generally inferred that, in the adjustment stage, the quality measures Bias, RMSE and MAPE increase as the proportions of imputed failures were increased. For the model adjusted to the series reconstructed by Spline and Stine methodologies with the random pattern of missing data was observed that the quality of the performance measures increase when the proportions of imputed failures were higher. Considerable quality losses in tuning performance can be attributed to the univariate (AM and M) and multivariate (PW and RLM) single imputation methods. To exemplify, the model adjustment to the series imputed by MLR and RW methodologies showed higher losses of quality, evidenced by the high values of the indicators Bias, RMSE, MAPE and by the decrease in quality of the indicators NSE, d_2 e ϱ.

In the most critical data reconstruction scenario (40%), relative differences in the model adjustment stage reached values of up to 1,233.44%, 106.28%, 171.40%, 205.59%, 65.48% and 0.94% of the respective indicators Bias, RMSE, MAPE, NSE, d_2 and ϱ, for the random pattern of the series reconstructed by the MLR methodology and up to 602.14%, 202.99%, 174.94%, 516.64%, 41.33 and 43.56% for the model adjusted to the series reconstructed by the Spline methodology under the block missing data pattern. Such findings allows to infer that these imputation methodologies do not preserve the series characteristics and compromise the SARIMA model adjustment performance and, consequently, the forecast performance (Table 12).

Studies found that RW, M and AM methodologies tend to underestimate the data variance. By their nature, AM and M are measures of central tendency of the series, being the median preferred when the data get farther of the normal distribution. The authors Ben Aissia et al. (2017), Gao et al. (2018) and Kabir et al. (2020) agree that, despite the simplicity of the methods, reconstruct the loss value using a constant does not reflect the variation that would probably occur if the data would be observed. It is considered important to evaluate the maintenance of the use of RW and MLR methodologies, for limiting the performance of the SARIMA model and, mainly because these are techniques usually applied to reconstruct missing data in hydrological series.

In turn, the reconstructed series by the single univariate (Spline and Stine) and multiple multivariate (IM and ML) imputation methodologies resulted in a good model fitting performance even under critical scenario of losses (40% - exception to Spline applied to the pattern of missing data in block). Small losses of quality of the model performance indicators adjusted to the series reconstructed by the methodologies Stine, IM and ML point up that these approaches preserves the series characteristics (Nunes et al., 2009; Junger & Ponce de Leon, 2015; Bleidorn et al., 2022).

Just as in the adjustment step, the forecasting quality indicators indicate that the series reconstructed by the imputation methodologies AM, M, RW, MLR compromises the model performance, both for the scenario of random missing data and for the block pattern. The model adjusted to the reconstructed series by MLR in the 40% of random failures scenario resulted in relative difference of 1,606.32%, 62.71%, 78.90%, 173.70% e 67.78% in the performance indicators Bias, RMSE, MAPE, NSE and d_2, respectively, and for the reconstructed series by Spline with the block pattern of failures resulted in relative difference values of 1,592.76%, 70.76%, 158.80%, 202.01% e 16.79% for the same indicators.

In cases of series reconstruction using the methodologies Spline (only to random missing data), Stine, IM and ML, as observed in the adjustment phase, minimal quality losses occurred. Even under the most critical data loss scenarios, the model obtained low relative differences, in both missing data patterns, when considering the series reconstructed by the IM and ML methodologies.

Figures 4 and 5 show the graphics of the simulations for the missing data proportion of 40% for the random and in block patterns, respectively. Figures 6 and 7 display a visual comparison of the adjusted model to the observed (black lines) and imputed data (red lines). This comparison shows a good performance of the adjusted model to the reconstructed series by the methodologies Spline, Stine, IM and ML considering the random missing data pattern and Stine, IM and ML for the in-block pattern. The other methodologies, just like suggested the performance indicators, resulted in underestimation of the adjustment to the observed series.

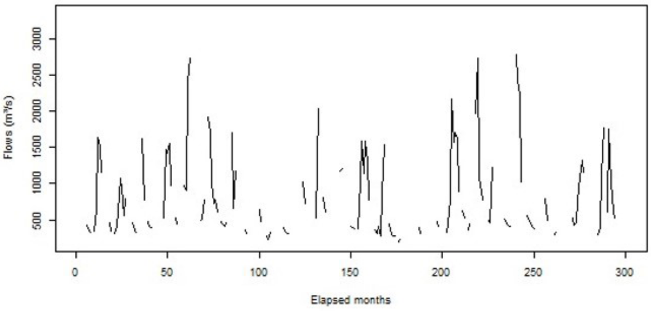

Figure 4. Monthly mean flow time series of station E6, considering 40% of random missing data.

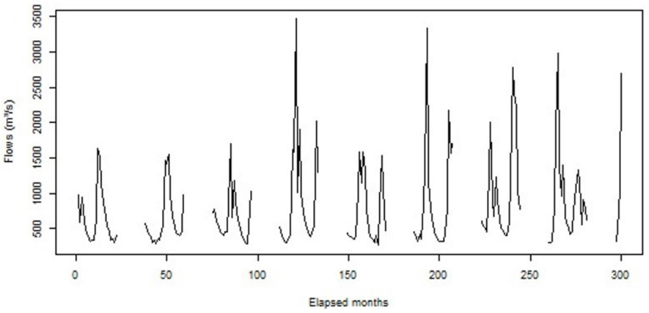

Figure 5. Monthly mean flow time series of station E6, considering 40% of in block missing data.

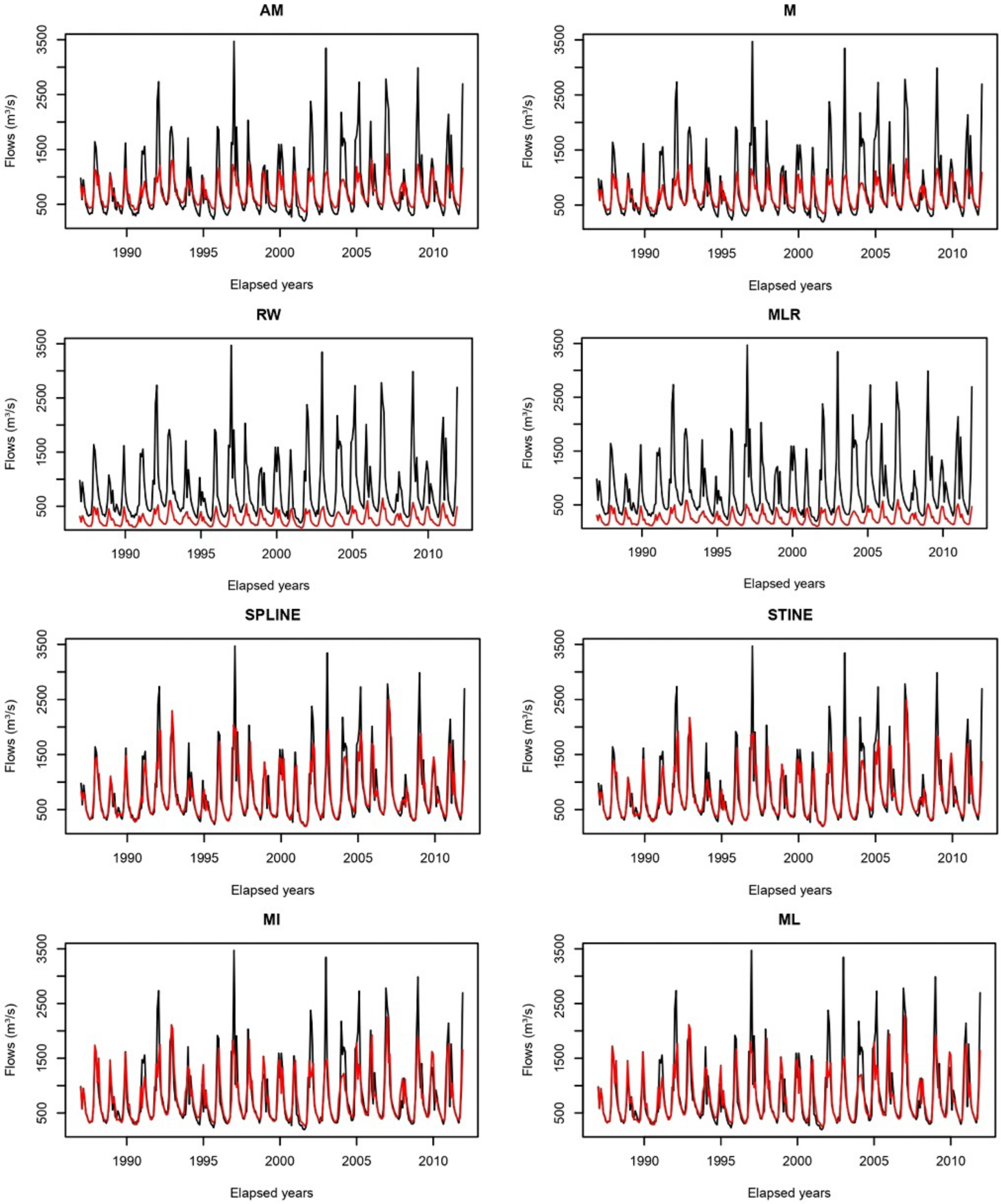

Figure 6. Observed monthly mean flow time series (black line) and model fitted by the imputed series (red line) at station E6, for each imputation methodology, considering 40% of random missing data.

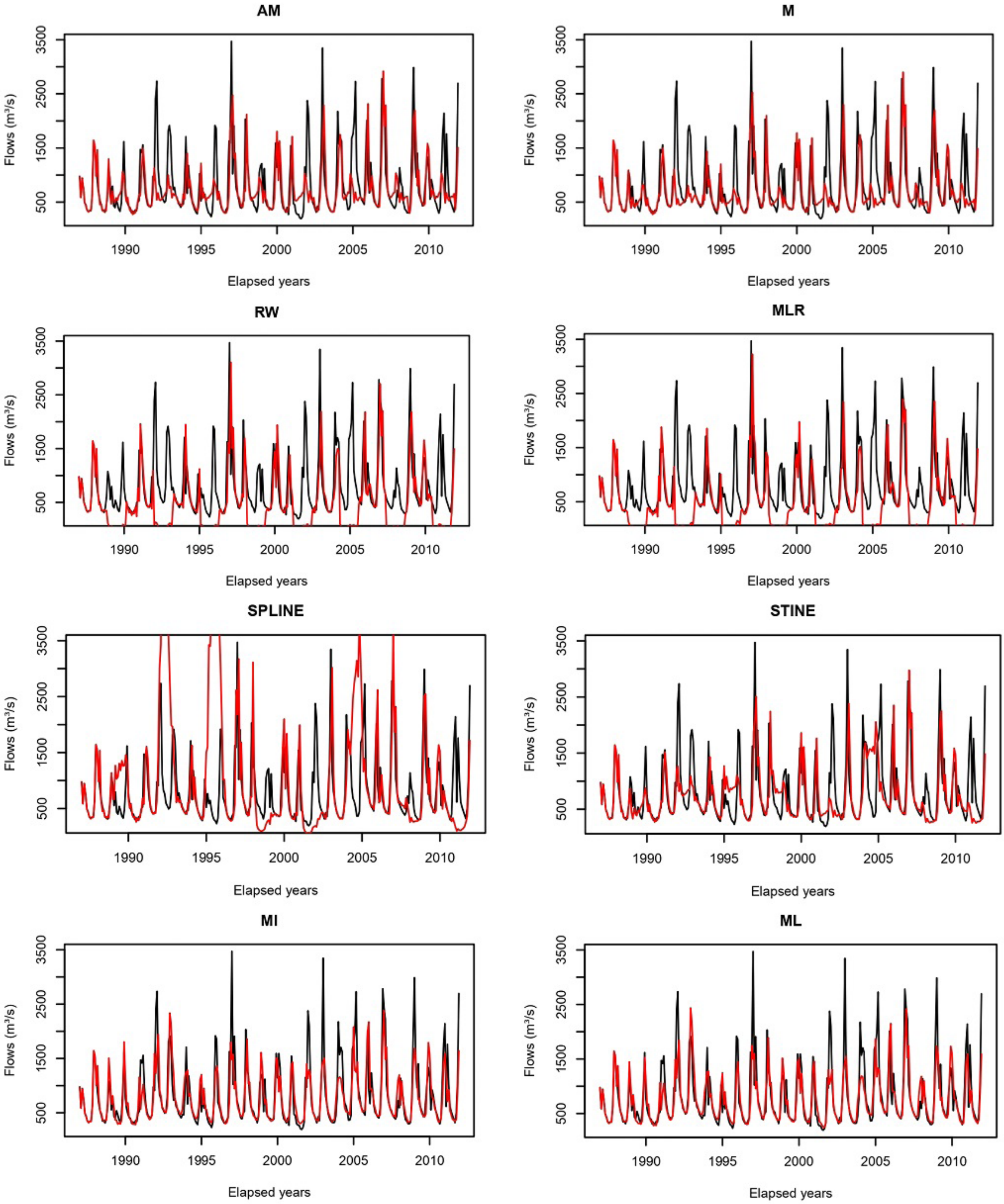

Figure 7. Observed monthly mean flow time series (black line) and model fitted by the imputed series (red line) at station E6, for each imputation methodology, considering 40% of in block missing data.

CONCLUSIONS

This study aimed to evaluate the effect of the use of different missing data imputation methodologies on the fitting and forecasting performance of the SARIMA time series model, considering as a case study the monthly mean flow data of the Doce River in Southeastern Brazil. Different failure proportions simulations under random and block pattern were considered.

A semi-automated approach was considered in the model choice to avoid any subjectivity in its choice. Therefore, the SARIMA $(1,1,1)(0,1,1)_{12}$ model was adjusted to the observed data and served as a quality parameter in the adjustment and forecasting of the series reconstructed by the imputation methodologies. The results indicated that, overall, the model loses quality in adjustment and forecasting as the percentage of imputed missing data increases, especially in the use of the model in the series reconstructed by the mean, regional weighting and multiple linear regression. Therefore, despite being usually used to impute missing data on hydrological variables, it is considered that these methodologies considerably reduce the quality of SARIMA time series model. In contrast, it was verified that the model quality was maintained when applied to the reconstructed series by the Spline (only for the random missing data pattern), Stine, Multiple Imputation and Maximum Likelihood methodologies. These methods, even under extreme conditions of data loss (40%), allowed to preserve the series characteristics.

Whereas data quality is crucial for successful employment of the SARIMA model, the finding results highlight that the quality of flow forecasts can be improved when the missing data processing is carried out using appropriate processing methodologies. The search for increased reliability of the results is useful in engineering projects, risk management, allocation of multiple uses and water between watersheds and in hydroelectric power generation. Now, it is known that the stage of processing missing data with appropriate methodologies must be anticipated before the use of the SARIMA model.

REFERENCES

Abudu, S., Cui, C. L., King, J. P., & Abudukadeer, K. (2010). Comparison of performance of statistical models in forecasting monthly streamflow of Kizil River, China. *Water Science and Engineering*, *3*(3), 269-281. http://doi.org/10.3882/j.issn.1674-2370.2010.03.003.

Aghelpour, P., Bahrami-Pichaghchi, H., & Varshavian, V. (2021). Hydrological drought forecasting using multi-scalar streamflow drought index, stochastic models and machine learning approaches, in northern Iran. *Stochastic Environmental Research and Risk Assessment*, *35*(8), 1615-1635. http://doi.org/10.1007/s00477-020-01949-z.

Ahmad, I., Waseem, M., & Zhang, J. (2022). Developing monthly hydrometeorological timeseries forecasts to reservoir operation in a transboundary river catchment. *Theoretical and Applied Climatology*, *147*(3-4), 1663-1674. http://doi.org/10.1007/s00704-021-03901-9.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. http://doi.org/10.1109/TAC.1974.1100705.

Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. In E. Parzen, K. Tanabe & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 275-280). New York: Springer.

Allison, P. D. (2002). *Quantitative applications in the social sciences: missing data.* Thousand Oaks: SAGE Publications.

Apaydin, H., Sattari, M. T., Falsafian, K., & Prasad, R. (2021). Artificial intelligence modelling integrated with Singular Spectral analysis and Seasonal-Trend decomposition using Loess approaches for streamflow predictions. *Journal of Hydrology*, *600*, 126506. http://doi.org/10.1016/j.jhydrol.2021.126506.

Arriagada, P., Karelovic, B., & Link, O. (2021). Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm. *Journal of Hydrology*, *598*, 126454. http://doi.org/10.1016/j.jhydrol.2021.126454.

Bayer, D. M., Castro, N. M. D. R., & Bayer, F. M. (2012). Modelagem e previsão de vazões médias mensais do rio Potiribu utilizando modelos de séries temporais. *RBRH*, *17*(2), 229-239. http://doi.org/10.21168/rbrh.v17n2.p229-239.

Ben Aissia, M. A. B., Chebana, F., & Ouarda, T. B. (2017). Multivariate missing data in hydrology–Review and applications. *Advances in Water Resources*, *110*, 299-309. http://doi.org/10.1016/j.advwatres.2017.10.002.

Bleidorn, M. T., Pinto, W. P., Braum, E. S., Lima, G. B., & Montebeller, C. A. (2019). Modelagem e previsão de vazões médias mensais do rio Jucu, ES, utilizando o modelo SARIMA. *Irriga*, *24*(2), 320-335. http://doi.org/10.15809/irriga.2019v24n2p320-335.

Bleidorn, M. T., Pinto, W. P., Schmidt, I. M., Mendonça, A. S. F., & Reis, J. A. T. D. (2022). Methodological approaches for imputing missing data into monthly flows series. *Revista Ambiente & Água*, *17*(2), e2795. http://doi.org/10.4136/ambi-agua.2795.

Box, G. E., & Jenkins, G. M. (1976). *Time series analysis: forecasting and control.* Hoboken: John Wiley & Sons.

Brasil. Agência Nacional de Águas – ANA. (2022). *HIDROWEB: sistema de informações hidrológicas.* Brasília. Retrieved in 2023, November 1, from http://hidroweb.ana.gov.br/default.asp

Chen, L., Xu, J., Wang, G., Liu, H., Zhai, L., Li, S., Sun, C., & Shen, Z. (2018). Influence of rainfall data scarcity on non-point source pollution prediction: implications for physically based models. *Journal of Hydrology*, *562*, 1-16. http://doi.org/10.1016/j.jhydrol.2018.04.044.

Coelho, A. L. N. (2007). *Alterações hidrogeomorfológicas no médio-baixo Rio Doce/ES* (Tese de doutorado). Departamento de Geografia, Instituto de Geociências, Universidade Federal Fluminense, Rio de Janeiro. Retrieved in 2022, October 20, from http://www.dominiopublico.gov.br/pesquisa/DetalheObraForm.do?select_action=&co_obra=157909

Comitê da Bacia Hidrográfica do Rio Doce – CBH-Doce. (2022). *A bacia do Rio Doce: caracterização da Bacia.* Governador Valadares. Retrieved in 2023, November 01, from https://www.cbhdoce.org.br/institucional/a-bacia

Dembélé, M., Oriani, F., Tumbulto, J., Mariéthoz, G., & Schaefli, B. (2019). Gap-filling of daily streamflow time series using Direct Sampling in various hydroclimatic settings. *Journal of Hydrology*, *569*, 573-586. http://doi.org/10.1016/j.jhydrol.2018.11.076.

Demirhan, H., & Renwick, Z. (2018). Missing value imputation for short to mid-term horizontal solar irradiance data. *Applied Energy*, *225*, 998-1012. http://doi.org/10.1016/j.apenergy.2018.05.054.

Duarte, V. B. R., Silva, F. D. C. S., Souza, I. V., Silva, M. V. C., Almeida Sousa, H. G., Giongo, M., & Viola, M. R. (2019). Previsão de vazão na bacia hidrográfica do rio Manuel Alves da Natividade utilizando o modelo de séries temporais SARIMA. *Journal of Biotechnology and Biodiversity*, *7*(4), 457-468. http://doi.org/10.20873/jbb.uft.cemaf.v7n4.duarte.

Fu, J., Zhong, P. A., Chen, J., Xu, B., Zhu, F., & Zhang, Y. (2019). Water resources allocation in transboundary river basins based on a game model considering inflow forecasting errors. *Water Resources Management*, *33*(8), 2809-2825. http://doi.org/10.1007/s11269-019-02259-y.

Gao, Y., Merz, C., Lischeid, G., & Schneider, M. (2018). A review on missing hydrological data processing. *Environmental Earth Sciences*, *77*(2), 47. http://doi.org/10.1007/s12665-018-7228-6.

Gill, M. K., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. *Water Resources Research*, *43*(7), 2006WR005298. http://doi.org/10.1029/2006WR005298.

Giustarini, L., Parisot, O., Ghoniem, M., Hostache, R., Trebs, I., & Otjacques, B. (2016). A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. *Environmental Modelling & Software*, *82*, 308-320. http://doi.org/10.1016/j.envsoft.2016.04.013.

Hamzah, F. B., Mohamad Hamzah, F., Mohd Razali, S. F., & El-Shafie, A. (2022). Multiple imputations by chained equations for recovering missing daily streamflow observations: a case study of Langat River basin in Malaysia. *Hydrological Sciences Journal*, *67*(1), 137-149. http://doi.org/10.1080/02626667.2021.2001471.

Hamzah, F. B., Mohd Hamzah, F., Mohd Razali, S. F., Jaafar, O., & Abdul Jamil, N. (2020). Imputation methods for recovering streamflow observation: a methodological review. *Cogent Environmental Science*, *6*(1), 1745133. http://doi.org/10.1080/23311843.2020.1745133.

Junger, W. L., & Ponce de Leon, A. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96-104. http://doi.org/10.1016/j.atmosenv.2014.11.049.

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2020). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, *5*(6), 365-377. http://doi.org/10.1080/23789689.2019.1600960.

Khodakhah, H., Aghelpour, P., & Hamedi, Z. (2022). Comparing linear and non-linear data-driven approaches in monthly river flow prediction, based on the models SARIMA, LSSVM, ANFIS, and GMDH. *Environmental Science and Pollution Research International*, *29*(15), 21935-21954. http://doi.org/10.1007/s11356-021-17443-0.

Little, R. J., & Rubin, D. B. (2002). *Statistical analysis with missing data.* Hoboken: John Wiley & Sons. http://doi.org/10.1002/9781119013563.

Liu, X., Sang, X., Chang, J., & Zheng, Y. (2021). Multi-model coupling water demand prediction optimization method for megacities based on time series decomposition. *Water Resources Management*, *35*(12), 4021-4041. http://doi.org/10.1007/s11269-021-02927-y.

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: a gentle introduction.* New York: Guilford Press.

Musa, J. J. (2013). Stochastic modeling of Shiroro River stream flow process. *American Journal of Engineering Research*, *2*(6), 49-54.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I: a discussion of principles. *Journal of Hydrology*, *10*(3), 282-290. http://doi.org/10.1016/0022-1694(70)90255-6.

Nunes, L. N., Klück, M. M., & Fachel, J. M. G. (2009). Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. *Cadernos de Saúde Pública*, *25*(2), 268-278. http://doi.org/10.1590/S0102-311X2009000200005.

Phan, T.-T.-H., & Nguyen, X. H. (2020). Combining statistical machine learning models with ARIMA for water level forecasting: the case of the Red river. *Advances in Water Resources*, *142*, 103656. http://doi.org/10.1016/j.advwatres.2020.103656.

Pinto, W. P., Lima, G. B., & Zanetti, J. B. (2015). Previsão de regimes de vazões médias mensais do rio Doce, Colatina-Espírito Santo. *Ciência e Natura*, *37*(3), 1-11. http://doi.org/10.5902/2179460X17143.

R Development Core Team. (2021). *R: a language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Retike, I., Bikše, J., Kalvāns, A., Dēliņa, A., Avotniece, Z., Zaadnoordijk, W. J., Jemeljanova, M., Popovs, K., Babre, A., Zelenkevičs, A., & Baikovs, A. (2022). Rescue of groundwater level time series: how to visually identify and treat errors. *Journal of Hydrology*, *605*, 127294. http://doi.org/10.1016/j.jhydrol.2021.127294.

Rubin, D. B. (1987). Procedures with nonignorable nonresponse. In: D. B. Rubin (Ed.), *Multiple imputation for nonresponse in surveys* (pp. 202-240). New York: John Wiley & Sons. http://doi.org/10.1002/9780470316696.ch6.

Salame, C. W., Queiroz, J. C. B., Souza, E. B., Farias, V. J. C., Rocha, E. J. P., & Moura, H. P. (2019). Um estudo comparativo dos modelos Box-Jenkins e Redes Neurais Artificiais na previsão de vazões e precipitações pluviométricas da Bacia Araguaia, Tocantins, Brasil. *Brazilian Journal of Environmental Sciences*, (52), 28-43. http://doi.org/10.5327//Z2176-947820190444.

Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, *7*(2), 147-177. http://doi.org/10.1037/1082-989X.7.2.147.

Schäfer, M. P., Dietrich, O., & Mbilinyi, B. (2016). Streamflow and lake water level changes and their attributed causes in Eastern and Southern Africa: state of the art review. *International Journal of Water Resources Development*, *32*(6), 853-880. http://doi.org/10.1080/07900627.2015.1091289.

Semiromi, M. T., & Koch, M. (2019). Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. *Hydrological Sciences Journal*, *64*(14), 1711-1726. http://doi.org/10.1080/02626667.2019.1669793.

Tencaliec, P., Favre, A. C., Prieur, C., & Mathevet, T. (2015). Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research*, *51*(12), 9447-9463. http://doi.org/10.1002/2015WR017399.

Wei, W. W. S. (2006). Time series analysis. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods in psychology: statistical analysis* (Vol. 2). Oxford: Oxford University Press.

Authors contributions

Michel Trarbach Bleidorn: Conceptualization, data curation, formal analysis, investigation, methodology, software, writing – original draft, review, supervision.

Isamara Maria Schmidt: Conceptualization, data curation, formal analysis, software, writing – review.

José Antonio Tosta dos Reis: Writing – original draft, formal analysis, review, supervision.

Deysilara Figueira Pani: Writing – original draft, Formal analysis, review.

Wanderson de Paula Pinto: Formal analysis, software, review.

Carlo Corrêa Solci: Formal analysis, review.

Antonio Sergio Ferreira Mendonça: Formal analysis, review.

Gutemberg Hespanha Brasil: Formal analysis, review.

Editor-in-Chief: Adilson Pinheiro

Associated Editor: Carlos Henrique Ribeiro Lima