



## Gap-filling of daily precipitation and streamflow time series: a method comparison at random and sequential gaps

Letícia Lopes Martins, Wander Araújo Martins, Iam Caio de Abreu Rodrigues, Ana Carolina Freitas Xavier, Jener Fernando Leite de Moraes & Gabriel Constantino Blain

**To cite this article:** Letícia Lopes Martins, Wander Araújo Martins, Iam Caio de Abreu Rodrigues, Ana Carolina Freitas Xavier, Jener Fernando Leite de Moraes & Gabriel Constantino Blain (2023) Gap-filling of daily precipitation and streamflow time series: a method comparison at random and sequential gaps, *Hydrological Sciences Journal*, 68:1, 148-160, DOI: [10.1080/02626667.2022.2145200](https://doi.org/10.1080/02626667.2022.2145200)

**To link to this article:** <https://doi.org/10.1080/02626667.2022.2145200>



[View supplementary material](#)



Published online: 15 Dec 2022.



[Submit your article to this journal](#)



Article views: 183



[View related articles](#)



[View Crossmark data](#)



## Gap-filling of daily precipitation and streamflow time series: a method comparison at random and sequential gaps

Letícia Lopes Martins <sup>a</sup>, Wander Araújo Martins <sup>a</sup>, Iam Caio de Abreu Rodrigues <sup>a</sup>, Ana Carolina Freitas Xavier <sup>b</sup>, Jener Fernando Leite de Moraes <sup>c</sup> and Gabriel Constantino Blain <sup>d</sup>

<sup>a</sup>Post-graduation Program in Tropical and Subtropical Agriculture, Agronomic Institute, Campinas, Brazil; <sup>b</sup>Water Resources, Vale Institute of Technology, Belém, Brazil; <sup>c</sup>Soil Research Department, Agronomic Institute, Campinas, Brazil; <sup>d</sup>Ecophysiology and Biophysics Department, Agronomic Institute, Campinas, Brazil

### ABSTRACT

This study compared the performance of seven gap-filling methods in daily streamflow and precipitation data and assessed the maximum number of gap days on which the methods perform well. Random (occurring randomly in the data series) and sequential (consecutive days with missing record) gaps were considered. Results show that the type of gaps affects the performance of the methods for gap-filling streamflow and precipitation data. Concerning random gaps, the best methods for streamflow were autoregressive integrate moving average and spline interpolation. For precipitation, the best methods were inverse distance weighting and linear regression (LR). Regarding sequential gaps, LR and multiple regression perform well in gap-filling up to 60 consecutive days in streamflow series. The other methods perform well up to 15 consecutive gap days. For precipitation series, the methods performed well up to seven consecutive missing days. For longer gaps (15, 30, 45 and 60 days), the methods performed poorly.

### ARTICLE HISTORY

Received 5 April 2022  
Accepted 7 October 2022

### EDITOR

A. Fiori

### ASSOCIATE EDITOR

A. Gires

### KEYWORDS

hydrological data gaps;  
missing values; climate data  
gaps; data filling methods

## 1 Introduction

Long-running streamflow and precipitation series are essential for several hydrological and climatological assessments including the study of extreme events (Fontolan *et al.* 2019, Tirupathi *et al.* 2019, Xavier *et al.* 2019, Tabari 2020), hydrological and climatic modelling (Strehmel *et al.* 2016, Vigiak *et al.* 2016, Veldkamp *et al.* 2018, Tabari 2020), and management of water resources (Wang *et al.* 2015). These historical series are also essential for quantifying water surplus and shortage in a region because availability and quality of water may affect agricultural planning and other public policies of water management. However, missing data remain a major limitation for several hydro-meteorological studies (Oriani *et al.* 2020). Although this limitation occurs in many areas of the globe, it is particularly critical in developing countries (Wang *et al.* 2015, Dembélé *et al.* 2019, Peleg *et al.* 2020) where resources are often limited for data collection and equipment maintenance and replacement (Dembélé *et al.* 2019).

Missing data are typically caused by equipment failures, data transmission issues, data transcription issues (in manual stations), extreme weather events and other causes (Pappas *et al.* 2014, Dembélé *et al.* 2019). The key issue is that many hydrological data analysis methods have little or no tolerance for missing values (Pappas *et al.* 2014, Wang *et al.* 2015, Zhang and Post 2018). Thus, the presence of such gaps in streamflow and precipitation series may prevent a reliable characterization of the hydrological regime in a region (Zhang and Post 2018). The presence of missing values in a data series can be classified into two types: random and sequential (or consecutive). Random gaps occur randomly in the data series at different

percentages. Sequential or consecutive gaps (e.g. consecutive days with no data record) can last for days, weeks, months, or years, and often occur due to poor or lacking equipment maintenance or equipment failures (Kanda *et al.* 2018, Zhang and Post 2018). This feature further emphasizes the need to select the best available gap-filling method for each variable and for the type of missing data.

Several methods have been proposed for gap-filling in streamflow and precipitation data (Pappas *et al.* 2014, Wang *et al.* 2015, Zhang and Post 2018, Beguería *et al.* 2019, Dembélé *et al.* 2019). They range from simple methods, such as average and simple linear regression, to more complex methods, such as artificial neural networks (Lepot *et al.* 2017, Coutinho *et al.* 2018, Luna *et al.* 2020). The most frequently used methods for gap-filling in streamflow and precipitation time series are simple, multiple and logistic regression (Lepot *et al.* 2017, Mfwango *et al.* 2018), polynomial or spline interpolation (Shope and Maharjan 2015, Lepot *et al.* 2017, Brubacher *et al.* 2020), simple nearest neighbour (Oriani *et al.* 2020), interpolation techniques (Chen and Liu 2012, Ismail and Ibrahim 2017, Oriani *et al.* 2020), ordinary kriging (Lepot *et al.* 2017, Oriani *et al.* 2020), hydrological modelling (Zhang and Post 2018), autoregressive models (Lepot *et al.* 2017, Ren *et al.* 2019), and satellite data applications (Ekeu-wei *et al.* 2018). More complex techniques such as non-linear mathematical programming, dynamic state-space models, artificial neural networks (Khalil *et al.* 2001, Kim and Pachepsky 2010, Coutinho *et al.* 2018, Ren *et al.* 2019, Vega-Garcia *et al.* 2019) and hybrid methods (process-based and statistics tools) are also used.

In addition to these methods, there is the adaptation and creation of new methods as proposed by Pappas *et al.* (2014), Zhang and Post (2018), Dembélé *et al.* (2019), Vega-Garcia *et al.* (2019) and others. Dembélé *et al.* (2019) developed a method of direct sampling for filling streamflow gaps; the study was carried out in different climatic zones and hydrological regimes and for different upstream-downstream relations among the gauging stations used for gap-filling. Pappas *et al.* (2014) proposed a quick method for gap-filling in hydro-meteorological data. This innovative method is used in the absence of neighbouring stations; however, it is applied for a small number of missing data. Zhang and Post (2018) used two hydrological models ("Génie Rural à 4 paramètres Journalier" (GR4J) and the lumped conceptual daily rainfall-runoff model (SIMHYD)) for gap-filling in streamflow data, considering different percentages (5, 10 and 20%) of missing data. Vega-Garcia *et al.* (2019) applied cascade-correlation neural networks for gap-filling in daily streamflow data at different types of gaps.

All these gap-filling studies added important advances to the hydro-meteorological literature; however, there is still a lack of studies comparing the performance of these methods for filling distinct types of gaps and defining the maximum percentage of missing data, or maximum number of sequential gap days, for which these methods perform well. Kanda *et al.* (2018) is one of the few examples of a study that assessed the performance of gap-filling methods for distinct types of gaps. Although this latter study provides evidence supporting the hypothesis that the performance of such methods is affected by the type of the gaps (sequential or random), it addressed only meteorological variables (precipitation and air temperature). To the authors' best knowledge, there no study evaluating this hypothesis for hydrological variables, such as streamflow.

In addition to the points already mentioned, it is known that the performance of gap-filling methods is also related to the temporal (autocorrelation or serial correlation) and/or spatial (cross-correlation) dependence of hydro-meteorological variables. Precipitation series, for example, exhibit high temporal and spatial variability (Kanda *et al.* 2018, Oriani *et al.* 2020). On the other hand, streamflow series tend to present more homogeneous characteristics in space and time, and a good performance may be achieved using methods with temporal and spatial dependence. Thus, the analysis of the spatial and temporal dependence of the series is a key point for choosing the most appropriate method for gap-filling. In this sense, we hypothesize that the performance of gap-filling methods in precipitation and streamflow series is influenced by the presence of temporal or spatial correlation in the series, as well as by the type (random or sequential) and percentage of gaps.

In this context, the objective of this study was to (i) compare the performance of seven simple and widely used gap-filling methods (linear and multiple regression, spline interpolation, autoregressive integrate moving average (ARIMA), inverse distance weighting, k-nearest neighbour and artificial neural network combined with ARIMA) when applied to precipitation and streamflow daily series with distinct types of gaps (random or sequential); (ii) define the maximum number of days of sequential gaps in which it is possible to use these methods; and (iii) recommend the most appropriate methods for each gap type and variable (streamflow and precipitation).

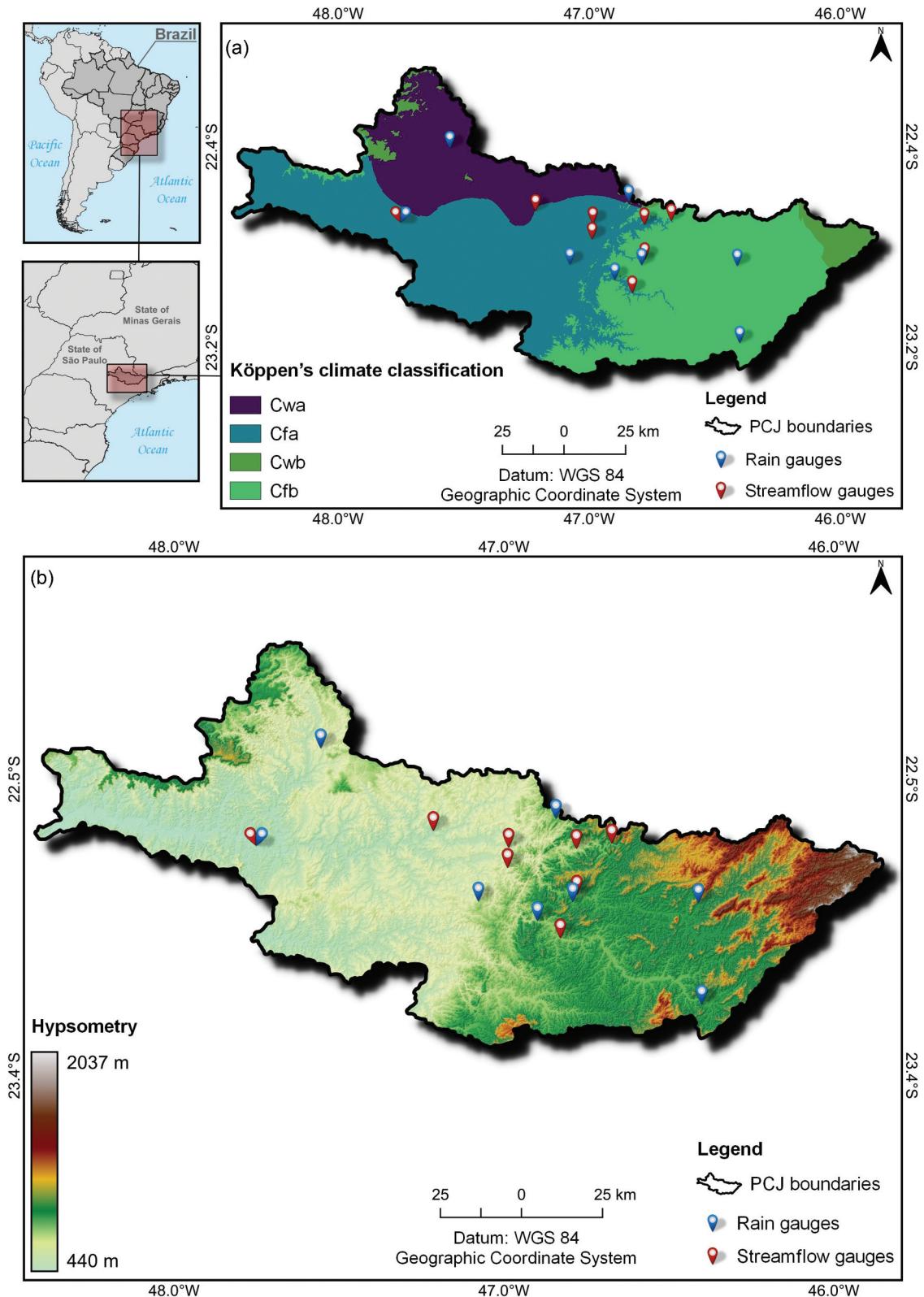
## 2 Study area and data

### 2.1 Study area

The testing region was the watershed of the Piracicaba, Capivari, and Jundiaí (PCJ) rivers situated in the states of São Paulo and Minas Gerais, Brazil (Fig. 1(a)). This region has intense economic, agricultural, and industrial activities (CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí 2018). The PCJ is one of the most important basins in the state of São Paulo, covering 15 320 km<sup>2</sup> and containing 76 cities and approximately 5.8 million habitants (CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí 2020). The Cantareira System, which is one of the biggest water supply systems in Brazil, is situated in the PCJ area. This system is of great importance as it is responsible for supplying water to the highly populated metropolitan region of São Paulo city (~8.8 million people; CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí 2020). Since 1961, this water reservoir has experienced an increase in the severity and frequency of droughts (Nobre *et al.* 2016).

Due to its great importance in providing water to one of the most important regions of Brazil, the PCJ has been the object of studies related to hydrological modelling (Lopes *et al.* 2020) and extreme weather (Taffarello *et al.* 2016, Cruz *et al.* 2017, Lemos *et al.* 2020). The recurrence of water crises in the state of São Paulo since 2013 has increased the need for such investigations. This scenario has led to the region's heightened vulnerability as temperatures have risen and precipitation has decreased (Nobre *et al.* 2016). In this context, studies addressing water management, droughts, and climate extremes are critical in the PCJ and will have an impact on Brazil more generally, given the country's reliance on this sector.

The PCJ presents anisotropic characteristics, mainly in topography, climate, and soil classes (CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí 2018). Anisotropy refers to the different patterns of spatial variability of a variable. In other words, it is related to different patterns of spatial dependence (Clark and Harper 2000). According to the Köppen-Geiger classification (see Fig. 1(a); Alvares *et al.* 2013), the PCJ watershed has four types of climates: Cfa (humid sub-tropical oceanic without dry season and with hot summer); Cfb (humid sub-tropical oceanic without dry season and with temperate summer); Cwa (humid sub-tropical with dry winter and hot summer) and Cwb (humid sub-tropical with dry winter and temperate summer). Annual precipitation ranges from 1195 to 1605 mm, with the highest precipitation rates occurring in the headwater areas (CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí 2018). The altitude ranges from 440 to 2037 m, and the stations used in this study and the relief in which they are situated are depicted in Fig. 1(b). According to CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí (2018), the predominant soils are: Acrisols (50.5%), Ferrasols (35.6%), Leptsols (4.9%), Gleysols (1.4%), Cambisols (0.7%), and Nitisols (0.4%). As also pointed out by CBH-PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí (2018), 22.5% of the area is occupied by native vegetation and forest remnants, 24% is pasture and 22% is devoted to agricultural activities.



**Figure 1.** Study area including rain and streamflow gauges: (a) localization by Köppen's climate classification (Cfa (humid sub-tropical oceanic without dry season and with hot summer); Cfb (humid subtropical oceanic without dry season and with temperate summer); Cwa (humid subtropical with dry winter and hot summer) and Cwb (humid sub-tropical with dry winter and temperate summer)) (Alvares et al. 2013); (b) location of the streamflow and precipitation gauges in relation to the topography of the study area.

## 2.2 Data

Daily precipitation and streamflow data from 1949 to 1979 were selected. We chose this time period because it contains

the greatest number of precipitation and fluviometric stations (eight precipitation and fluviometric stations) with no missing data (Fig. 1(b)). Precipitation data were collected from the

**Table 1.** Streamflow and precipitation database used in this study.

Streamflow stations	Long-term mean ( $m^3 \cdot s^{-1}$ ) <sup>1</sup>	Drainage area ( $km^2$ )	Rain station	Annual accumulated precipitation (mm) <sup>1</sup>
3d02	6.46	387.00	D3-002	1226
3d03	32.00	2490	D3-018	1396
3d06	27.10	1920	D3-031	1255
3d09	32.70	1950	D3-046	1300
4d01	54.00	3.394	D4-012	1152
4d07	145.00	10 918	D4-044	1217
62 615 000	32.60	2180	D4-061	1230
62 625 000	14.50	663.00	E3-099	1156

<sup>1</sup>For the period from 1949 to 1979.

Department of Water and Electricity (DAEE, available at <http://www.hidrologia.daee.sp.gov.br/>), while streamflow data were obtained from the National Water Agency of Brazil (ANA, available at <http://www.snh.gov.br/hidroweb/>) and DAEE. Table 1 presents further information on the streamflow and precipitation series, respectively. In the Supplementary material, the statistical properties of both precipitation and streamflow series are shown (see Supplementary material, Figs S1–S2 and Tables S1–S2).

### 3 Methods

In this study seven methods for gap-filling were considered: linear regression (LR), multiple regression (MR), spline interpolation (or cubic spline regression), autoregressive integrated moving average (ARIMA), inverse distance weighting (IDW), k-nearest neighbour (k-NN) and artificial neural network (recurrent neural networks (RNN) combined with ARIMA). These methods were chosen because of their widespread use (Pappas *et al.* 2014, Vega-Garcia *et al.* 2019, Oriani *et al.* 2020) and because previous studies indicated that they perform well for both streamflow and precipitation data. Streamflow and precipitation data with random and sequential gaps were used to test all methods.

LR, MR, k-NN, spline and IDW are methods related to the spatial dependence (cross-correlation) of the time series. The ARIMA and RNN methods depend on autocorrelation. Thus, Pearson's correlation (Wilks 2011) and Moran's index (Equation 1; Moran 1950) were used to detect cross-correlation and to establish the donor stations. The autocorrelation function (Wilks 2011) was used to estimate the temporal dependence in the time series. Both autocorrelation and cross-correlation analyses were performed at the 5% significance level.

$$I = \frac{n \sum_{i=1}^n \sum_{j=1}^n w_{ij} z_i z_j}{S_0 \sum_{i=1}^n z_i^2} \quad (1)$$

where  $w_{ij}$  = the spatial weight between observations  $i$  and  $j$ ;  $z_i$  = the deviation of an attribute for observation  $i$  from its mean ( $x_i - \bar{X}$ );  $n$  = the number of observations; and  $S_0$  = the sum of all the spatial weights (Equation 1.1):

$$S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij} \quad (1.1)$$

#### 3.1 Linear regression

LR is a method frequently used for gap-filling in hydrometeorological series (Alley and Burns 1983, Beyad and Maeder 2013,

Mfwango *et al.* 2018). It is based on the linear relationship between two stations – a station with a gap and a station without a gap – where the latter is called the donor station (Alley and Burns 1983). This method was based on the following steps: (i) perform an LR between each donor station and the station with gaps (Equation 2); (ii) choose as the donor station the one with the highest Pearson's coefficient.

$$y = \beta_0 + \beta x \quad (2)$$

where  $y$  = station with the gap (target station),  $x$  = donor station,  $\beta_0$  = intercept with the y-axis and  $\beta$  = slope (regression coefficient).

#### 3.2 Multiple regression

Multiple regressions are also widely used for gap-filling in hydro-meteorological series. MR (Equation 3) was applied considering that the streamflow and precipitation values of the target station were dependent on the two closest donor stations (Beyad and Maeder 2013, Lepot *et al.* 2017, Mfwango *et al.* 2018).

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \quad (3)$$

where  $y_i$  = target station;  $x_1$  = donor station 1;  $x_2$  = donor station 2; and  $i$  is the time step.

#### 3.3 ARIMA models

This method uses autocorrelation (temporal persistence) to predict missing data (Ren *et al.* 2019, Dorich *et al.* 2020). It is commonly used for time series prediction (Alsharif *et al.* 2019, Kassem *et al.* 2020), but has also been used for gap-filling in time series (Granado *et al.* 2015, Lepot *et al.* 2017, Ponkina *et al.* 2021). For time series with weak autocorrelation, such as precipitation series, this method is expected to perform poorly (Dorich *et al.* 2020). ARIMA is defined by Equation (4). Lepot *et al.* (2017) recommended the following steps to apply the ARIMA method: (i) apply the Kalman filter; (ii) compute the missing values; (iii) estimate the p and q parameters; (iv) apply a smoothing algorithm for the missing data. In this study, the “na\_Kalman” function and the “auto.arima” method from the “ImputeTS” (Moritz *et al.* 2017) R package was employed for gap-filling using autoregressive models.

$$\theta_p(B)(1 - B)^d x_t = \Phi_q(B)w_t \quad (4)$$

where  $\theta_p$  and  $\Phi_q$  = polynomials of orders  $p$  and  $q$ , respectively;  $x_t$  = a time series;  $B$  = the backward shift operator (lag

operator; Equation 4.1);  $w_t$  = a white noise series; and  $d$  = the difference order.

$$B_{x_t} = x_t - 1 \quad (4.1)$$

By successively applying B, it follows that (Equation 4.2):

$$B_{x_t}^n = x_t - n \quad (4.2)$$

### 3.4 Inverse distance weighting

IDW is one of the most used interpolation methods for gap-filling (Chen and Liu 2012, Kanda et al. 2018, Oriani et al. 2020, Chinasho et al. 2021). This method assumes that data coming from sites that are close to each other are more similar than those coming from larger distances apart. This method weights the observed data according to distance so that the greatest weights are assigned to the locations closest to the missing value. The missing value is calculated using Equation (5) (Cressman 1959, Shepard 1968).

$$Y_0 = \frac{\sum_{i=1}^n Y_i / d_i}{\sum_{i=1}^n \left(\frac{1}{d_i}\right)} \quad (5)$$

where  $Y_0$  = the estimated missing value;  $Y_i$  = the observed value at the  $i^{\text{th}}$  station;  $d_i$  = the Euclidean distance between target station and the  $i^{\text{th}}$  donor station; and  $n$  = the power exponent.

### 3.5 Spline interpolation

Spline interpolation uses a cubic polynomial function that splits the dataset into several *knots* and fits each knot to a separate model (Shope and Maharjan 2015). A spline function of  $n$  degree is a type of piecewise polynomial, so each  $n$  degree is applied to successive intervals of the points (North and Livingstone 2013, Shope and Maharjan 2015). For this reason, the values obtained using this method are usually influenced by the values measured at neighbouring stations. This method uses a polynomial function between each pair of tabulated points with non-locally determined coefficients; this feature ensures global smoothness in the interpolated function up to a certain order of derivatives (North and Livingstone 2013, Shope and Maharjan 2015, Brubacher et al. 2020). The difference between linear interpolation and spline interpolation is that the first interpolation is performed considering two measured values for gap-filling, while spline interpolation incorporates values from several measurements for gap-filling. Since the spline function interpolates all data points, it is given by Equation (6).

$$S(x_j) = z(x_j) \quad (6)$$

where  $z(x_j)$  are the values of the variables at points  $x_j$ .

In spline interpolation the essential idea is to find a curve that connects data points, by adjusting the piecewise function of the form (Equation 6.1):

$$S(x) = \begin{cases} s_1(x) & \text{if } x_1 \leq x < x_2 \\ s_2(x) & \text{if } x_2 \leq x < x_3 \\ \vdots \\ s_{n-1}(x) & \text{if } x_{n-1} \leq x < x_n \end{cases} \quad (6.1)$$

where  $s_i$  is a third-degree polynomial defined by Equation 6.2:

$$s_i(x) = \alpha_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i \quad (6.2)$$

for  $i = 1, 2, \dots, n-1$

where  $\alpha_i, b_i, c_i$  and  $d_i$  are the function coefficients. For  $n$  points, there are  $n-1$  cubic functions to find, with  $\alpha_i, b_i, c_i$  and  $d_i$  coefficients. So, the first and second derivatives of these  $n-1$  equations are essential for this calculation (Equations 6.3 and 6.4):

$$s'_i(x) = 3\alpha_i(x - x_i)^2 + 2b_i(x - x_i) + c_i \quad (6.3)$$

$$s''_i(x) = 6\alpha_i(x - x_i) + 2b_i \quad (6.4)$$

for  $i = 1, 2, \dots, n-1$

Thus, a spline function  $S(x)$  should meet the condition that  $S(x)$  for the observed points must be equal to  $z(x)$  and the smoothing  $I(S)$  (Equation 6.5) should be as small as possible.

$$I(S) = \min \quad (6.5)$$

### 3.6 K-nearest neighbour

k-NN is one of the most commonly used methods for gap-filling in hydro-meteorological data series (Lepot et al. 2017, Beguería et al. 2019). This method relies on the proximity of neighbouring stations for gap-filling in the station with missing data, and it imputes missing values using the weighted average of the neighbouring stations. In this study, the distance between nearby stations was based on Gower distance (Gower 1971). The distance between stations is defined by Equation (7). The k-NN method creates a reference series for each location of interest. The reference series is formed by the weighted average of the values observed at neighbouring stations (Beguería et al. 2019). In this study,  $k = 5$  (number of neighbouring stations) was used, because it led to the best performance of this method.

$$d_{i,j} = \frac{\sum_{k=1}^p w_k \delta_{i,j,k}}{\sum_{k=1}^p w_k} \quad (7)$$

where  $w_k$  = weight;  $\delta_{i,j,k}$  = contribution of the  $k^{\text{th}}$  variable; and  $p$  = the number of neighbouring stations. In this study, the absolute distance was divided by the total range (Equation 7.1):

$$\delta_{i,j,k} = |x_{i,k} - x_{j,k}| / r_k \quad (7.1)$$

where  $x_{i,k}$  = value of the  $k^{\text{th}}$  variable of the  $i^{\text{th}}$  observation; and  $r_k$  = is the range of the  $k^{\text{th}}$  variable.

### 3.7 Artificial neural network

Artificial neural networks have been studied since the 1990s as a promising method for filling in missing data and predicting future data (Khalil et al. 2001, Ren et al. 2019). In this study, we

used RNNs, which are a type of neural network used in sequential data. RNNs are designed to recognize patterns in a data series, so projections made by an RNN consider past observations. For this reason, RNNs are known as memory neural networks (Giles *et al.* 1994, Ren *et al.* 2019). In this study, we used the temporal series already filled by the ARIMA method, thus this method is based on the combination of ARIMA and RNN multilayer perceptron. This type of neural network architecture is more effective and parsimonious compared to networks with a single layer (Haykin 2001). The objective of this combination was to verify whether there would be improvements in the prediction of missing values with the combination of two methods that consider memory in temporal series. We followed these steps for the prediction of missing values using the RNNs: (i) specification of the autocorrelation lag (memory of the series); (ii) normalization of the data for neural network training; (iii) splitting the dataset into training and testing; (iv) neural network training and definition of the number of epochs; (v) neural network prediction; (vi) denormalization of the data. The following R packages were used to perform this method: RNN (Quast and Fichou 2016) and NNFOR (Kourentzes 2019).

### 3.8 Generation of missing data

The performance of the gap-filling algorithms was evaluated using artificially generated random and sequential gaps in the original series. The random gaps were set to generate series with 5, 10, 15 and 20% of missing data in respect to the total length of the series. These percentages were defined based on the studies of Kanda *et al.* (2018), Ismail and Ibrahim (2017) and Zhang and Post (2018). The gaps were generated using the R package “missMethods” (Rockel 2020), which employs an approach developed by Santos *et al.* (2019) for creating random gaps. The missing completely at random (MCAR) mechanism was used for such a purpose. MCAR generates series based on a predefined percentage of missing records. Therefore, the maximum probability of missing values is defined by this percentage. Missing values are randomly defined using the Bernoulli distribution.

The sequential gaps were generated considering different numbers of sequential days with no record. Sequential gaps (2, 3, 4, 5, 6, 7, 15, 30, 45 and 60 consecutive days with no data record) are usually observed in precipitation and streamflow series (Kanda *et al.* 2018, Dembélé *et al.* 2019). These sequential gaps correspond to 6, 8, 11, 13, 15, 18, 32, 50, 59 and 68% of the total sample size of the series. We assumed that series with more than 70% gaps are usually removed from most of the studies (Pappas *et al.* 2014, Wang *et al.* 2015, Dembélé *et al.* 2019). In addition, as demonstrated in section 4, the performance of the models decreases as the percentage (or number of days with no data record) increases. Each sequential gap was randomly generated, i.e. there was no temporal pattern of these gaps in the series. The performance of the methods was evaluated using the statistical methods presented in section 3.9.

### 3.9 Assessing the performance of each gap-filling method

Considering accuracy as the average correspondence between individual predictions and the observed data (Wilks 2011), widely used metrics were applied to assess the performance

of the gap-filling methods. The absolute mean error (AME), mean squared error (MSE) and root-mean-square-error (RMSE) are widely used scalar measures of accuracy for continuous data (Vlček and Huth 2009, Wilks 2011, Wilcke *et al.* 2013, Pierce *et al.* 2015, Pereira *et al.* 2018, Xavier *et al.* 2019), and they can be calculated according to Equations (8–10):

$$AME = \frac{\sum_{i=1}^n |V_{obs_i} - V_{est_i}|}{n} \quad (8)$$

$$MSE = \frac{\sum_{i=1}^n |V_{obs_i} - V_{est_i}|^2}{n} \quad (9)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n |V_{obs_i} - V_{est_i}|^2}{n}} \quad (10)$$

where  $V_{est_i}$  = estimated values;  $V_{obs_i}$  = observed data; and  $n$  = number of missing values estimated.

The Pearson determination coefficient ( $R^2$ ) and Willmott's index of agreement were also used to assess the performance of the gap-filling models. Willmott's index was used in its modified version ( $D_{mod}$ ; Willmott *et al.* 1985) and was calculated according to Equation (11):

$$D_{mod} = 1 - \frac{\sum_{i=1}^n |P_i - O_i|}{\sum_{i=1}^n (|P_i - \bar{O}| + |O_i - \bar{O}|)} \quad (11)$$

where  $P$  = the predicted values and  $O$  = the observed values.

An advantage of  $D_{mod}$  over its original version ( $D$ ) is that it approaches its maximum value ( $D_{mod}= 1$ ), which indicates a complete agreement between estimates and observations, more slowly as  $P$  approaches  $O$  (Equation 11). Therefore, when compared to  $D$ , the  $D_{mod}$  version provides greater separation when comparing distinct models. The  $D_{mod}$  is also less sensitive to outliers (Willmott *et al.* 2012). As recommended by Willmott *et al.* (1985) and Pereira *et al.* (2018), 95% confidence intervals – specified through bootstrap estimates – were estimated for all accuracy/precision measures to enhance their reliability.

## 4 Results and discussion

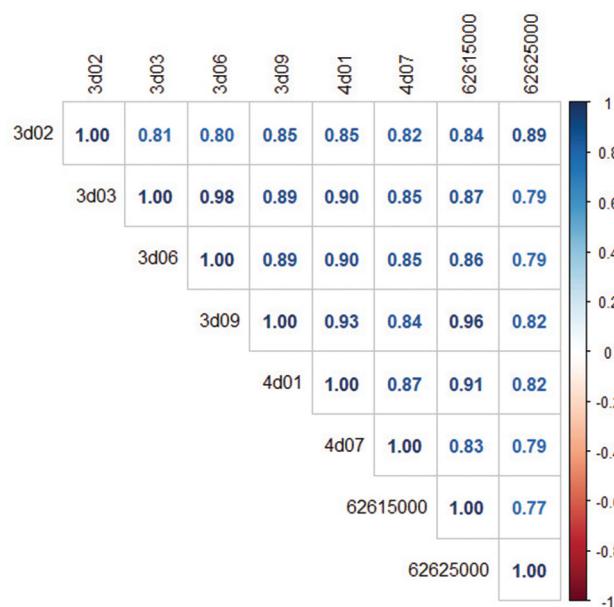
### 4.1 Cross-correlation and autocorrelation in streamflow and precipitation series

The Moran's I method indicated the presence of significant positive cross-correlation among the precipitation series and among the streamflow series. Considering the value of Moran's I, the streamflow series ( $I = 0.48$ ) presented a higher level of spatial dependence than the precipitation series ( $I = 0.23$ ). This latter statement is also in line with the values of the Pearson's coefficient observed for the precipitation series (Fig. 2), which varied from 0.34 to 0.74 (117.6% variation) and those observed for the streamflow series (Fig. 3), which varied from 0.77 to 0.98 (27.3% variation). The Moran's I statistics and the Pearson coefficients also suggest that the effect of the anisotropic features of the PCJ region are more notable in the precipitation series than in the streamflow series.

With regards to the autocorrelation function, the coefficients depicted in Fig. 4 show that the daily streamflow data



**Figure 2.** Pearson's correlation coefficient of daily precipitation for all weather stations.



**Figure 3.** Pearson's correlation coefficient of daily streamflow for all fluvimetric stations.

present a high level of temporal persistence. The autocorrelation coefficients for all station presented values higher than 0.9 for lag = 1, which gradually decreased as the lag increased. This feature indicates that the daily streamflow data of the PCJ's rivers are highly dependent on the data observed on the previous day, as also observed by Tigabu *et al.* (2020) and Duvert *et al.* (2015). As pointed out by Tigabu *et al.* (2020), the strong autocorrelation in streamflow series is directly related to the high predictability of these series. On the other hand, the pattern of the correlogram of the precipitation series (Fig. 5) shows low levels of temporal persistence. At lag = 1, most of the series presented autocorrelation coefficients lower than 0.30 (Fig. 5). The results depicted in Figs 4 and 5 suggest that gap-filling methods based on spatial dependence or/and

temporal persistence are expected to perform well when applied to daily streamflow. For gap-filling in precipitation series, the methods based on temporal persistence of the data are expected to perform poorly.

#### 4.2 Streamflow series

The results found in this study indicated that the type of gap (random or sequential) affects the performance of the gap-filling methods for streamflow. Furthermore, while the percentage of random gaps did not affect the performance of the methods, the number of days with sequential gaps did affect their performance.

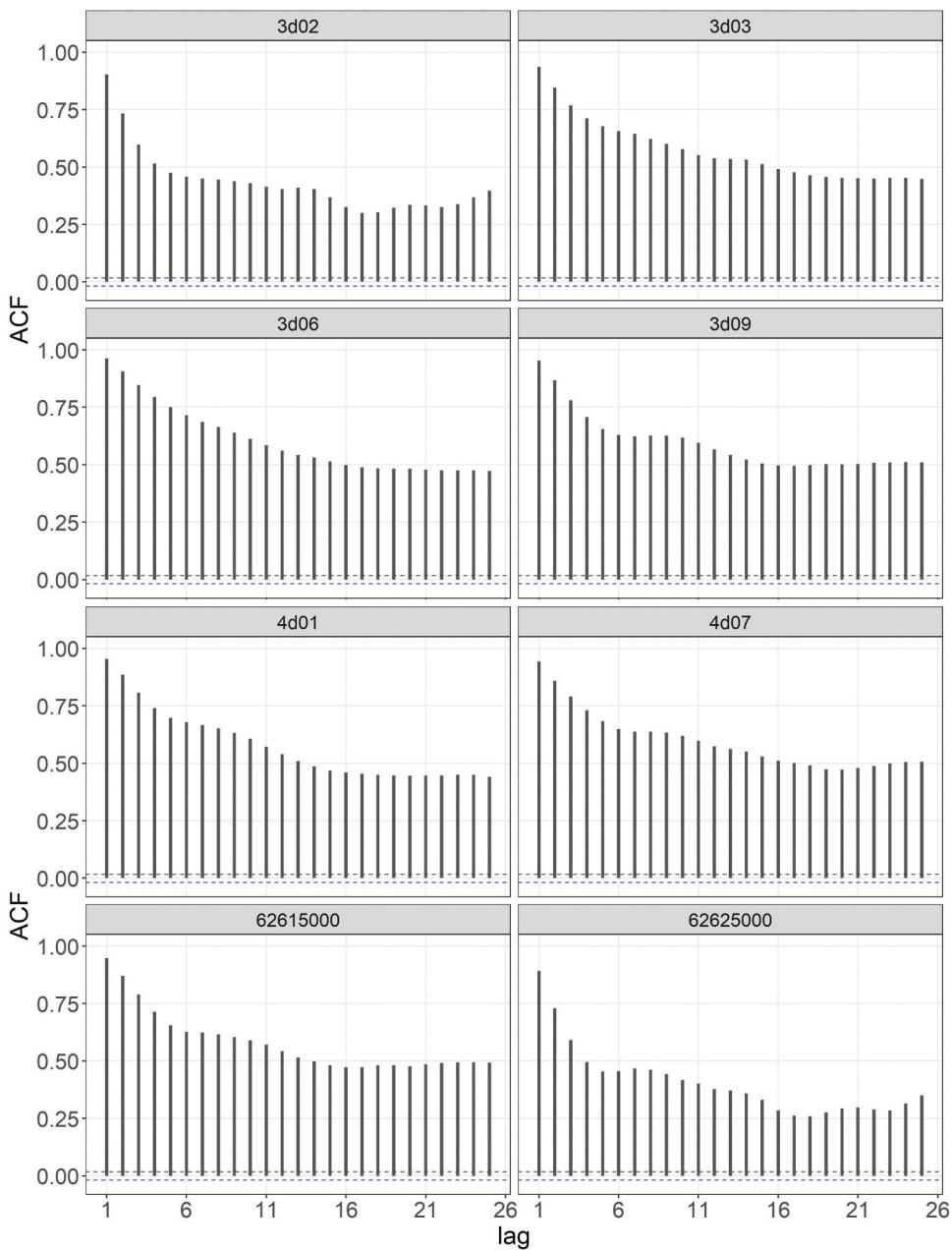
With regards to random gaps and considering the 95% confidence interval, the spline and ARIMA methods presented equivalent performance at all gap percentages. The performance of the methods at different percentages of random gaps (5, 10, 15 and 20%) is presented in Fig. 6 (see also Supplementary material, Tables S3 and S7). The analyses based on  $D_{mod}$  indicated that all methods performed well for gap-filling streamflow data (except IDW, with  $D_{mod}$  values below 0.7), since the values for  $D_{mod}$  remained above 0.7 (Fig. 6).

The results depicted in Fig. 6 indicate that the percentage of random gaps considered in this study (5, 10, 15 and 20%) did not significantly affect the performance of the gap-filling methods. These results are in line with those of Kanda *et al.* (2018) and Ismail and Ibrahim (2017) and allow us to identify the spline and ARIMA methods as the best for filling random gaps. The LR, MR and k-NN methods also performed well in filling random gaps, with  $D_{mod}$  values above 0.80 at all percentages of missing data.

The results for sequential gaps are depicted in Fig. 7 (see also Supplementary material, Tables S4, S8 and S9), and they do not permit the identification of a single best method for gap-filling distinct sequential gaps. The methods showed similar performances for gap-filling from two to seven gap days (Fig. 7(a)).  $D_{mod}$  values were higher than 0.7 from two to seven gap days, except for IDW which showed values around 0.6. The LR and MR methods performed well for all percentages of missing records, including the largest gaps (Fig. 7(b)). However, the other methods showed poor performance after 15 days of gaps. Spline and k-NN performed well in filling short gaps (2, 3, 4, 5, 6, and 7 days). However, they performed poorly in filling longer gaps, with values as high as 0.29.

With regards to sequential gaps, the number of missing data affected the performance of the gap-filling methods (Fig. 7 (a) and (b)). As the number of consecutive days with gaps increases, the performance of the autocorrelation-dependent method (ARIMA) decreases, so the cross-correlation-dependent methods perform better in this situation. For 30, 45 and 60 gap days, all methods (apart from LR and MR) showed poor performance.

As observed for random gaps, the IDW presented the worst performance among all methods. Although this method relies on the presence of spatial dependence among the series, it also performed poorly for gap-filling streamflow data, which presented significant levels of cross-correlation. Considering the studies of Tung (1983) and Kanda *et al.* (2018), this feature

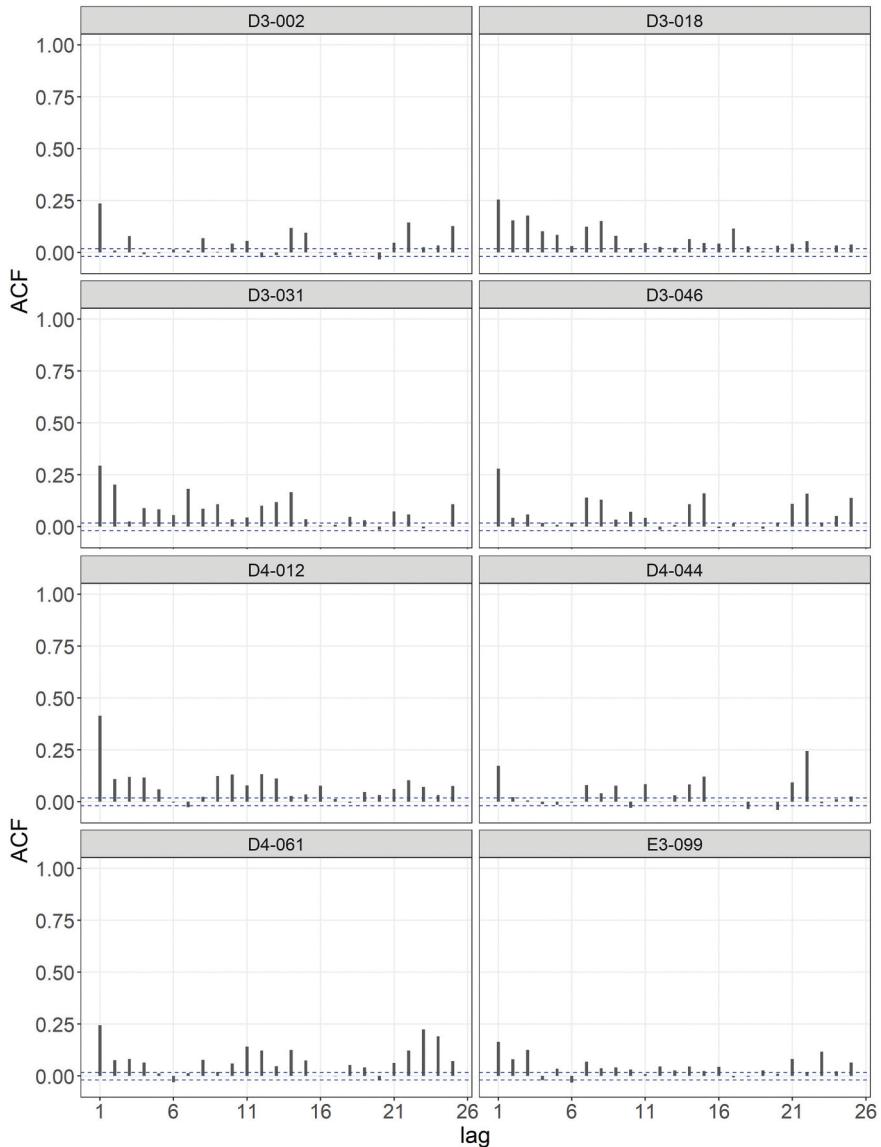


**Figure 4.** Autocorrelation function (ACF) of daily streamflow for all fluvimetric stations. The dotted line represents the white noise limits.

suggests that the IDW was not able to capture the anisotropy features of the PCJ watershed described in section 2.1. We emphasize that although the IDW gave the worst performance in comparison to the other methods, its  $D_{mod}$  values were higher than 0.75 in most stations. The lowest values were observed at stations 3d02, 4d07 and 62 625 000 (see Supplementary material, Table S7) because these stations have long-term mean streamflows and drainage area different from the others, which decreased the performance of this method, since it considers the similarity between donor and recipient stations.

Streamflow series present high cross-correlation and autocorrelation (as described in section 4.1), which agrees with the performance of the best methods proposed in this study. Based

on the results presented here, one can infer that in the presence of random gaps there is no difference in the performance of methods that consider autocorrelation and cross-correlation, since the best methods were ARIMA (temporal dependence), spline, LR and MR (spatial dependence). However, with sequential gaps, it is expected that the method with the highest autocorrelation would perform worse. Nevertheless, due to the high autocorrelation existing in the streamflow data (as described in section 4.1), the ARIMA method (highly dependent on autocorrelation) performed well in filling sequential gaps up to 15 gap days. These results can also be regarded as evidence favouring the use of ARIMA and spline methods for random gaps and LR and MR for filling sequential gaps in streamflow series with cross-correlation and autocorrelation.



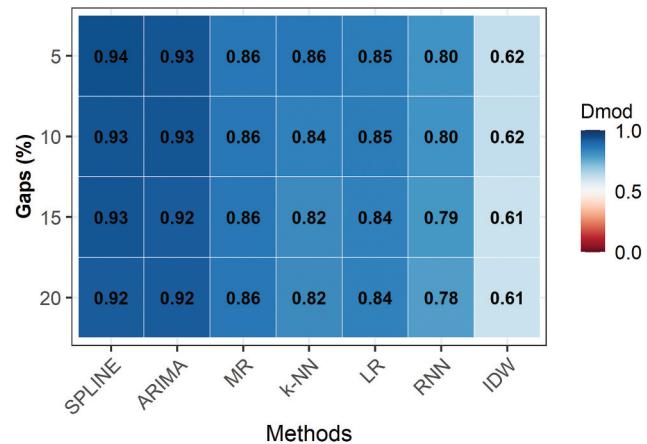
**Figure 5.** Autocorrelation function (ACF) of daily precipitation for all weather stations. The dotted line represents the white noise limits.

#### 4.3 Precipitation series

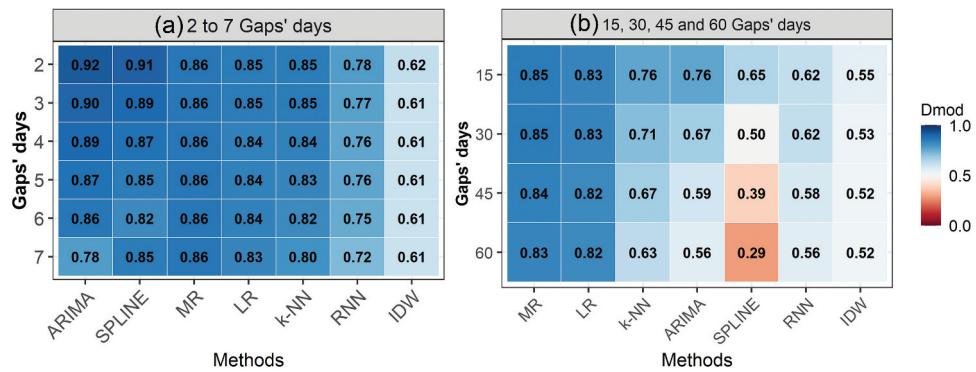
The results found in this study showed that the type of gap affects the performance of the gap-filling methods for precipitation data series. In addition, the percentage of random gaps and the number of sequential gap days also influenced the performance of the methods evaluated in this study.

With random gaps the best methods were IDW and LR for 5, 10, 15 and 20% of gaps and k-NN for 5, 10 and 15% of gaps, considering a 95% confidence interval (Fig. 8; see Supplementary material, Tables S5 and S10). The other methods showed poor performance for filling random precipitation gaps ( $D_{mod}$  below 0.64). As can be seen in Fig. 8, the  $D_{mod}$  values for the RNN method were less than 0.2 at all stations and at all percentages of gaps, hence indicating its poor performance as a gap-filling method.

Although differences were observed in the performance of the methods with respect to the percentage of gaps, the methods that presented the best performance were those using



**Figure 6.** Willmott's index modified ( $D_{mod}$ ) values of different methods when estimating missing random gaps in streamflow data. Tested methods: SPLINE, spline interpolation; ARIMA, autoregressive integrated moving average; MR, multiple regression; k-NN, k-nearest neighbour; LR, linear regression; RNN, recurrent neural networks combined with ARIMA and IDW, inverse distance weighting.



**Figure 7.** Willmott's index modified ( $D_{mod}$ ) values of different methods when estimating missing sequential gaps in streamflow data: (a) 2–7 gap days; (b) 15, 30, 45 and 60 gap days. Tested methods: SPLINE, spline interpolation; ARIMA, autoregressive integrated moving average; MR, multiple regression; k-NN, k-nearest neighbour; LR, linear regression; RNN, recurrent neural networks combined with ARIMA and IDW, inverse distance weighting.

interpolation techniques (IDW for 5, 10, 15 and 20% gaps and k-NN for 5, 10 and 15% gaps). These results suggest that methods that use interpolation techniques for gap-filling in random gaps for precipitation series are better than other techniques. Our results agree with previous studies such as Ismail and Ibrahim (2017), Kanda *et al.* (2018) and Oriani *et al.* (2020), which verified that interpolation methods for gap-filling perform well in hydro-meteorological series.

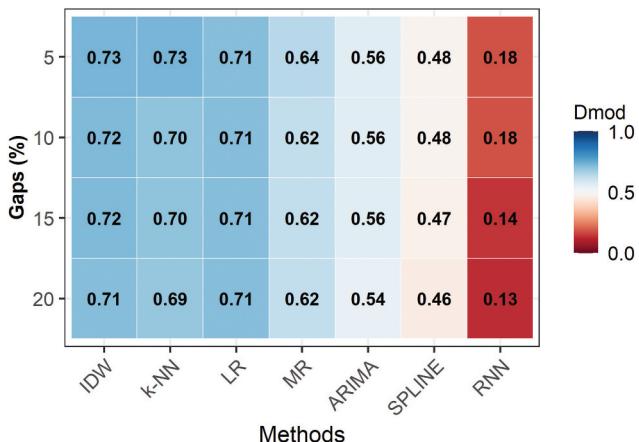
We point out that although the performance of the IDW method was contradictory in the gap-filling methods in streamflow (worst performance) and precipitation series (best performance), the value was similar for both variables (lower than 0.62 in streamflow series and higher than 0.71 in precipitation series). This fact supports the hypothesis raised in section 4.1, that anisotropy exerts a greater influence on precipitation series than on streamflow in the PCJ.

Gap-filling methods performed poorly for sequential gaps, with  $D_{mod}$  values ranging from 0.04 to 0.72 (Fig. 9 (a) and (b)). Overall, LR performed well up to seven sequential gap days. k-NN's performance was satisfactory up to six consecutive gap days. The other methods showed poor

performance for all gaps, making it impossible to use them for gap-filling in precipitation series (see Supplementary material, Tables S6, S11 and S12).

The results shown in this section favour the use of LR and k-NN methods for sequential gap-filling in precipitation series. As pointed out by Oriani *et al.* (2020), when there are a low number of stations available, simpler methods for gap-filling tend to outperform complex methods. LR was the least complex method tested in this study; however, it was the one that showed the best results for gap-filling in precipitation series. Moreover, in areas that present homogeneity in the precipitation pattern, k-NN is recommended as a method for gap-filling, since this method relies on nearby stations with similar characteristics to identify similar patterns and make predictions (Wu 2009, Oriani *et al.* 2020).

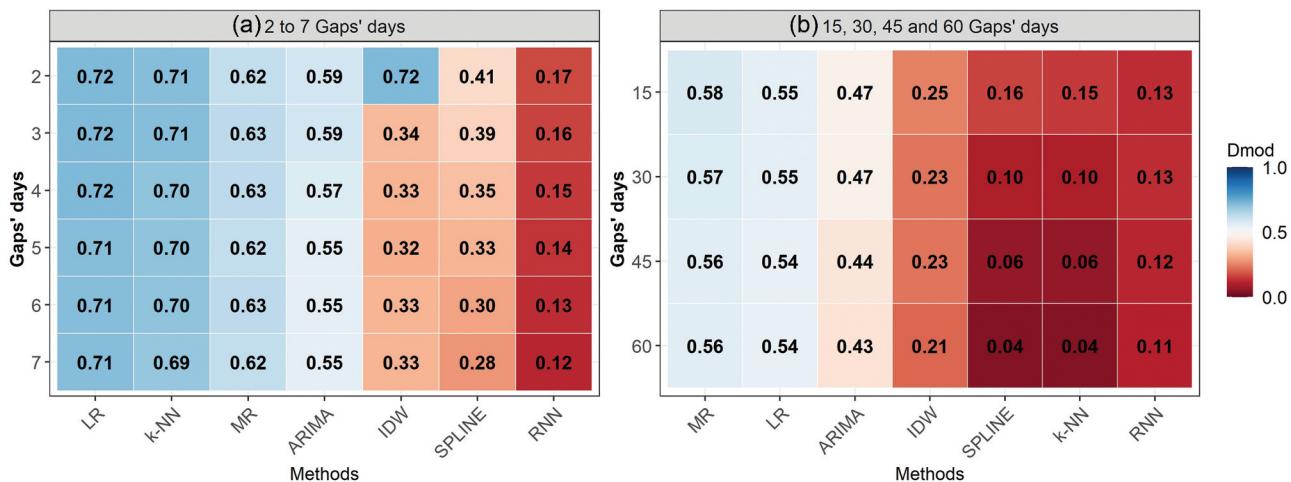
The method with the worst performance on random and sequential gaps was RNN, which may be related to the fact that it is based on temporal dependence (Giles *et al.* 1994). As described in section 4.1, the precipitation time series used in this study presented weak autocorrelation. Thus, the use of RNNs for gap-filling in precipitation data in PCJ is not recommended.



**Figure 8.** Willmott's index modified ( $D_{mod}$ ) values of different methods when estimating random gaps in precipitation data. Tested methods: SPLINE, spline interpolation; ARIMA, autoregressive integrated moving average; MR, multiple regression; k-NN, k-nearest neighbour; LR, linear regression; RNN, recurrent neural networks combined with ARIMA and IDW, inverse distance weighting.

#### 4.4 Recommendations for the use of gap-filling methods in future studies

Although this study was carried out in only one test region, the results obtained are relevant for other regions of the globe. This statement is based on the fact that we evaluated the performance of gap-filling methods that have been widely used in hydrological and meteorological studies throughout the world. At this point, it is worth mentioning that although the performance of any gap-filling method may vary depending on the region in which it is applied, the procedures proposed in this study for generating missing data (random and sequential) and the statistical procedures used to assess the performance of the gap-filling methods may be regarded as a reference for future studies. Furthermore, the relationship between methods' performances and cross-correlation/autocorrelation may also be used in these future studies to select candidate methods or to explain their performance.



**Figure 9.** Willmott's index modified ( $D_{mod}$ ) values of different methods when estimating sequential gaps in precipitation data: (a) 2–7 gap days; (b) 15, 30, 45 and 60 gap days. Tested methods: SPLINE, spline interpolation; ARIMA, autoregressive integrated moving average; MR, multiple regression; k-NN, k-nearest neighbour; LR, linear regression; RNN, recurrent neural networks combined with ARIMA and IDW, inverse distance weighting.

The results found in this study are in line with the hypothesis that the performance of gap-filling methods in precipitation and streamflow series is affected by the presence of temporal or spatial correlation and by the type (random or sequential) and percentage of gaps. The best results were obtained when filling random gaps. This statement holds for both precipitation and streamflow series, since the methods performed well at all gap lengths considered in this study. On the other hand, filling sequential gaps is a challenge. In this case, the performance of all methods decreased as the number of missing records increased. Moreover, the maximum number of consecutive days with gaps in which the methods perform well was different for precipitation and streamflow series. For precipitation series, gap-filling can be carried out for gaps of up to seven consecutive days using the LR method. Due to high cross-correlation and autocorrelation in streamflow series, gap-filling can be carried out for gaps of up to 15 consecutive days using the ARIMA and k-NN methods, and for gaps of up to 60 consecutive days using the LR and MR methods.

## 5 Conclusions

This study assessed the performance of distinct gap-filling methods in precipitation and streamflow daily data, considering different types of gaps (random and sequential). The results indicated that the percentage of random gaps does not affect the performance of the methods for gap-filling in streamflow, but it does influence the performance of the methods in precipitation series. However, in sequential gaps, the number of missing days affects the performance of the methods in both streamflow and precipitation series.

The results found in this study support the hypothesis that the presence of temporal or spatial correlation of the series, as well the type of gaps (random or sequential) affects the performance of the gap-filling methods. The best methods for random gap-filling in precipitation and streamflow data were spline and ARIMA for streamflow and IDW (for 5, 10, 15 and

20% gaps) and k-NN (for 5, 10 and 15% gaps) in precipitation series. Regarding sequential gaps in streamflow, the best methods were spline, ARIMA, LR, MR, k-NN and RNN for up to 15 days of gaps and MR and LR up to 60 consecutive gap days. For sequential gaps in precipitation data, the best methods were LR up to seven gap days and k-NN up to six gap days. In general, the methods dependent on cross-correlation and autocorrelation showed good performance for gap-filling in streamflow series. For precipitation series the best methods are those that consider cross-correlation.

## Acknowledgements

We thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior PhD Scholarship for Letícia Lopes Martins and Wander Araújo Martins; and CNPq for fellowship for Gabriel Constantino Blain (Process 307616/2019-3).

## Disclosure Statement

No potential conflict of interest was reported by the authors.

## Funding

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and National Council for Scientific and Technological Development (CNPq) Process [307616/2019-3].

## ORCID

Letícia Lopes Martins <http://orcid.org/0000-0002-0299-3005>  
 Wander Araújo Martins <http://orcid.org/0000-0002-4608-9464>  
 Iam Caio de Abreu Rodrigues <http://orcid.org/0000-0003-1000-1122>  
 Ana Carolina Freitas Xavier <http://orcid.org/0000-0002-9732-8779>  
 Jener Fernando Leite de Moraes <http://orcid.org/0000-0003-1112-4204>  
 Gabriel Constantino Blain <http://orcid.org/0000-0001-8832-7734>

## References

- Alley, W.M. and Burns, A.W., 1983. Mixed-station extension of monthly streamflow records. *Journal of Hydraulic Engineering*, 109 (10), 1272–1284. doi:10.1061/(ASCE)0733-9429(1983)109:10(1272).
- Alsharif, M.H., Mohammad, K.Y., and Jeong, K., 2019. Time series ARIMA model for prediction of daily and monthly average global solar radiation: the case study of Seoul, South Korea. *Symmetry*, 11 (2), 240. doi:10.3390/sym11020240.
- Alvares, C.A., et al., 2013. Köppen's climate classification map for Brazil. *Meteorologische Zeitschrift*, 22 (6), 711–728. doi:10.1127/0941-2948/2013/0507.
- Beguería, S., et al., 2019. Gap filling of monthly temperature data and its effect on climatic variability and trends. *Journal of Climate*, 32 (22), 7797–7821. doi:10.1175/JCLI-D-19-0244.1.
- Beyad, Y. and Maeder, M., 2013. Multivariate linear regression with missing values. *Anal Chim Acta*, 796 (796), 38–41. doi:10.1016/j.aca.2013.08.027.
- Brubacher, J.P., Oliveira, G.G., and Guasselli, L.A., 2020. Preenchimento de Falhas e Espacialização de Dados Pluviométricos: desafios e Perspectivas. *Revista Brasileira de Meteorologia*, 35 (4), 615–629. doi:10.1590/0102-77863540067.
- CBH- PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí, 2018. *Primeira Revisão do Plano das Bacias Hidrográficas dos rios Piracicaba, Capivari e Jundiaí 2010-2020* (Relatório final – revisão 5 – Tomo I). Piracicaba: Profill e Rhamá.
- CBH- PCJ Comitê das bacias hidrográficas dos rios Piracicaba, Capivari e Jundiaí, 2020. *Relatório da Situação dos Recursos Hídricos 2020* (ano base 2019) (versão simplificada). Piracicaba: Agência das Bacias PCJ.
- Chen, F.-W. and Liu, C.-W., 2012. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment*, 10 (3), 209–222. doi:10.1007/s10333-012-0319-1.
- Chinasho, A., et al., 2021. Evaluation of seven gap-filling techniques for daily station-based rainfall datasets in South Ethiopia. *Advances in Meteorology*, 22. doi:10.1155/2021/9657460.
- Clark, I. and Harper, W.V., 2000. *Practical geostatistics*. Scotland: Geostokos (Escosse) Limited.
- Coutinho, E.R., et al., 2018. Application of Artificial Neural Networks (ANNs) in the gap filling of meteorological time series. *Revista Brasileira de Meteorologia*, 33 (2), 317–328. doi:10.1590/0102-7786332013.
- Cressman, G.P., 1959. An operational objective analysis system. *Monthly Weather Review*, 87 (10), 367–374. doi:10.1175/1520-0493(1959)087<0367:AOOAS>2.0.CO;2.
- Cruz, P.P.N., et al., 2017. Hydrological modeling of the Ribeirão das Posses – an assessment based on the Agricultural Ecosystem Services (AgES) watershed model. *Revista Ambiente & Água*, 12, 3. doi:10.4136/ambi-agua.2073.
- Dembélé, M., et al., 2019. Gap-filling of daily streamflow time series using direct sampling in various hydroclimatic settings. *Journal of Hydrology*, 569, 573–586. doi:10.1016/j.jhydrol.2018.11.076.
- Dorich, C.D., et al., 2020. Global research alliance N<sub>2</sub>O chamber methodology guidelines: guidelines for gap-filling missing measurements. *Journal of Environmental Quality*, 49 (5), 1186–1202. doi:10.1002/jeq2.20138.
- Duvert, C., et al., 2015. Correlation and spectral analyses to assess the response of a shallow aquifer to low and high frequency rainfall fluctuations. *Journal of Hydrology*, 527, 894–907. doi:10.1016/j.jhydrol.2015.05.054.
- Ekeu-wei, I.T., Blackburn, G.A., and Pedruco, P., 2018. Infilling missing data in hydrology: solutions using satellite radar altimetry and multiple imputation for data-sparse regions. *Water*, 10 (10), 1483. doi:10.3390/w10101483.
- Fontolan, M., et al., 2019. Using climate change models to assess the probability of weather extremes events: a local scale study based on the generalized extreme value distribution. *Bragantia*, 78 (1), 146–157. doi:10.1590/1678-4499.2018144.
- Giles, C.L., Kuhn, G.M., and Williams, R.J., 1994. Dynamic recurrent neural networks: theory and applications. *IEE Transactions on Neural Networks*, 5 (2), 153–156. doi:10.1109/TNN.1994.8753425.
- Gower, J.C., 1971. A general coefficient of similarity and some of its properties. *Biometrics*, 27 (4), 857–871. doi:10.2307/2528823.
- Granado, J.P., Garrido, R., and Suárez, C., 2015. MIARMA: a minimal-loss information method for filling gaps in time series. *Astronomy & Astrophysics*, 57. doi:10.1051/0004-6361/201425056.
- Haykin, S., 2001. *Redes Neurais: princípios e prática*. 2nd ed. Porto Alegre: Bookman.
- Ismail, W.N.W. and Ibrahim, W.Z.W.Z., 2017. Estimation of rainfall and streamflow missing data for Terengganu, Malaysia by using interpolation technique methods. *Malaysian Journal of Fundamental and Applied Sciences*, 13 (3), 213–217. doi:10.11113/mjfas.v13n3.578.
- Kanda, N., et al., 2018. Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of Karakoram Himalaya. *Meteorological Applications*, 25 (3), 337–349. doi:10.1002/met.1699.
- Kassem, A.A., Raheem, A.M., and Khidir, K.M., 2020. Daily streamflow prediction for Khazir River Basin using ARIMA and ANN models. *Zanco: Journal of Pure and Applied Sciences*, 32 (3), 30–39. doi:10.21271/ZJPAS.32.3.4.
- Khalil, M., Panu, U.S., and Lennox, W.C., 2001. Groups and neural networks based streamflow data infilling procedures. *Journal of Hydrology*, 241 (3–4), 153–176. doi:10.1016/S0022-1694(00)00332-2.
- Kim, J.-W. and Pachepsky, Y., 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. *Journal of Hydrology*, 394 (3–4), 305–314. doi:10.1016/j.jhydrol.2010.09.005.
- Kourentzes, N., 2019. *nnfor: Time Series Forecasting with Neural Networks*. R package version 0.9.6. Available from: <https://CRAN.R-project.org/package=nnfor> [Accessed 20 June 2021].
- Lemos, M.C., et al., 2020. Building on adaptive capacity extreme events in Brazil: water reform, participation, and climate information across four river basins. *Regional Environmental Change*, 20 (2), 53. doi:10.1007/s10113-020-01636-3.
- Lepot, M., Aubin, J.B., and Clemens, F.H.L.R., 2017. Interpolation in time series: an introductory overview of existing methods. *Their Performance Criteria and Uncertainty Assessment*. *Water*, 9, 796. doi:10.3390/w9100796.
- Lopes, T.R., et al., 2020. Hydrological modeling for the Piracicaba River basin to support water management and ecosystem services. *Journal of South American Earth Sciences*, 103, 102752. doi:10.1016/j.jsames.2020.102752.
- Luna, A.M., et al., 2020. Assessing the best gap-filling technique for river stage data suitable for low capacity processors and real-time application using IoT. *Sensors*, 20 (21), 6354. doi:10.3390/s20216354.
- Mfwango, L.H., Salim, C.J., and Kazumba, S., 2018. Estimation of missing river flow data for hydrologic analysis: the case of Great Ruaha river catchment. *Hydrology: Current Research*, 9, 299. doi:10.4172/2157-7587.1000299.
- Moran, P., 1950. A test for the serial independence of residuals. *Biometrika*, 37 (1–2), 178–181. doi:10.1093/biomet/37.1-2.178.
- Moritz, S. and Bartz-Beielstein, T., 2017. imputeTS: time series missing value imputation in R. *The R Journal*, 9 (1), 207–218. doi:10.32614/RJ-2017-009.
- Nobre, C.A., et al., 2016. Some characteristics and impacts of the drought and water crisis in Southeastern Brazil during 2014 and 2015. *Journal of Water Resource and Protection*, 8 (2), 2. doi:10.4236/jwarp.2016.82022.
- North, R.P. and Livingstone, D.M., 2013. Comparison of linear and cubic spline methods of interpolating lake water column profiles. *Limnology and Oceanography: Methods*, 11 (4), 213–224. doi:10.4319/lom.2013.11.213.
- Oriani, F., et al., 2020. Missing data imputation for multisite rainfall networks: a comparison between geostatistical interpolation and pattern-based estimation on different terrain types. *Journal of Hydrometeorology*, 21, 10. doi:10.1175/JHM-D-19-0220.1.
- Pappas, C., Papalexiou, S.M., and Koutsoyiannis, D., 2014. A quick gap filling of missing hydrometeorological data. *Journal of Geophysical Research: Atmospheres*, 119 (15), 9290–9930. doi:10.1002/2014JD021633.

- Peleg, N., et al., 2020. Downscaling climate projections over large and data sparse regions: methodological application in the Zambezi River Basin. *International Journal of Climatology*, 40 (15), 6242–6264. doi:[10.1002/joc.6578](https://doi.org/10.1002/joc.6578).
- Pereira, H.R., et al., 2018. On the performance of three indices of agreement: an easy-to-use r-code for calculating the Willmott indices. *Bragantia*, 77 (2), 394–403. doi:[10.1590/1678-4499.2017054](https://doi.org/10.1590/1678-4499.2017054).
- Pierce, D.W., et al., 2015. Improved bias correction techniques for hydrological simulations of climate change. *Journal of Hydrometeorology*, 16 (6), 2421–2442. doi:[10.1175/jhm-d-14-0236.1](https://doi.org/10.1175/jhm-d-14-0236.1).
- Ponkina, E., et al., 2021. Do ARMA models provide better gap filling in time series of soil temperature and soil moisture? The case of arable land in the Kulunda Steppe, Russia. *Land*, 10 (6), 579. doi:[10.3390/land10060579](https://doi.org/10.3390/land10060579).
- Quast, B.A. and Fichou, D., 2016. *rnn: a recurrent neural network in R*. v.0.81 Available from: <http://qua.st/rnn> [Accessed 10 June 2021].
- Ren, H., et al., 2019. Using deep learning to fill spatio-temporal data gaps in hydrological monitoring networks. *Hydrology and Earth System Sciences*. [preprint]. doi:[10.5194/hess-2019-196](https://doi.org/10.5194/hess-2019-196).
- Rockel, T., 2020. *missMethods: methods for missing data*. R package version 0.2.0. Available from: <https://CRAN.R-project.org/package=missMethods>. [Accessed 20 January 2021].
- Santos, M.S., et al., 2019. Generating synthetic missing data: a review by missing mechanism. *IEEE Access*, 7, 11651–11667. doi:[10.1109/ACCESS.2019.2891360](https://doi.org/10.1109/ACCESS.2019.2891360).
- Shepard, D., 1968. A two-dimensional interpolation function for irregularly-spaced data. *23rd ACM National Conference*. Princeton, NJ: Brandon Systems Press, 517–524.
- Shope, C. and Maharjan, G.R., 2015. Modeling spatiotemporal precipitation: effects of density, interpolation, and land use distribution. *Advances in Meteorology*. 25. doi:[10.1155/2015/174196](https://doi.org/10.1155/2015/174196).
- Strehmel, A., et al., 2016. Field data-based implementation of land management and terraces on the catchment scale for an eco-hydrological modelling approach in the Three Gorges Region, China. *Agricultural Water Management*, 175, 43–60. doi:[10.1016/j.agwat.2015.10.007](https://doi.org/10.1016/j.agwat.2015.10.007).
- Tabari, H., 2020. Climate change impact on flood and extreme precipitation increases with water availability. *Scientific Reports*, 10 (1), 13768. doi:[10.1038/s41598-020-70816-2](https://doi.org/10.1038/s41598-020-70816-2).
- Taffarello, D., et al., 2016. Field investigations of the 2013–14 drought through quasi-quantitative freshwater monitoring at the headwaters of the Cantareira System, Brazil. *Water International*, 41 (5), 5. doi:[10.1080/02508060.2016.1188352](https://doi.org/10.1080/02508060.2016.1188352).
- Tigabu, T.B., et al., 2020. Statistical analysis of rainfall and streamflow time series in the Lake Tana Basin, Ethiopia. *Journal of Water and Climate Change*, 11 (1), 258–273. doi:[10.2166/wcc.2018.008](https://doi.org/10.2166/wcc.2018.008).
- Tirupathi, C., Shashidhar, T., and Srinivasan, R., 2019. Analysis of rainfall extremes and water yield of Krishna River basin in under future climate scenarios. *Journal of Hydrology: Regional Studies*. 19. doi:[10.1016/j.ejrh.2018.10.004](https://doi.org/10.1016/j.ejrh.2018.10.004).
- Tung, Y.K., 1983. Point rainfall estimation for a mountainous region. *Journal of Hydraulic Engineering*, 109 (10), 1386–1393. doi:[10.1061/\(ASCE\)0733-9429\(1983\)109:10\(1386\)](https://doi.org/10.1061/(ASCE)0733-9429(1983)109:10(1386).
- Vega-Garcia, C., Decuyper, M., and Alcázar, J., 2019. Applying cascade-correlation neural networks to in-fill gaps in Mediterranean daily flow data series. *Water*, 11 (8), 1691. doi:[10.3390/w11081691](https://doi.org/10.3390/w11081691).
- Veldkamp, T.I., et al., 2018. Human impact parameterizations in global hydrological models improve estimates of monthly discharges and hydrological extremes: a multi-model validation study. *Environmental Research Letters*, 13 (5), 055008. doi:[10.1088/1748-9326/aab96f](https://doi.org/10.1088/1748-9326/aab96f).
- Vigiak, O., et al., 2016. Impact of current riparian land on sediment retention in the Danube River Basin. *Sustainability of Water Quality and Ecology*, 8, 30–49. doi:[10.1016/j.swaqe.2016.08.001](https://doi.org/10.1016/j.swaqe.2016.08.001).
- Vlček, O. and Huth, R., 2009. Is daily precipitation Gamma-distributed? Adverse effects of an incorrect use of the Kolmogorov-Smirnov test. *Atmospheric Research*, 93 (4), 759–766. doi:[10.1016/j.atmosres.2009.03.005](https://doi.org/10.1016/j.atmosres.2009.03.005).
- Wang, D., et al., 2015. A hybrid wavelet analysis-cloud model data-extending approach for meteorologic and hydrologic time series. *Journal of Geophysical Research: Atmospheres*, 120 (9), 4057–4071. doi:[10.1002/2015JD023192](https://doi.org/10.1002/2015JD023192).
- Wilcke, R.A.I., Mendlik, T., and Gobiet, A., 2013. Multi-variable error correction of regional climate models. *Climatic Change*, 120 (4), 871–887. doi:[10.1007/s10584-013-0845-x](https://doi.org/10.1007/s10584-013-0845-x).
- Wilks, D.S., 2011. *Statistical methods in the atmospheric sciences*. 3rd ed. Oxford, UK: Academic Press.
- Willmott, C.J., et al., 1985. Statistics for the evaluation of model performance. *Journal of Geophysical Research*, 90 (C5), 8995–9005. doi:[10.1029/JC090iC05p08995](https://doi.org/10.1029/JC090iC05p08995).
- Willmott, C.J., Robeson, S.M., and Matsuura, K.A., 2012. A refined index of model performance. *International Journal of Climatology*, 32 (13), 2088–2094. doi:[10.1002/joc.2419](https://doi.org/10.1002/joc.2419).
- Wu, J., 2009. A novel artificial neural network ensemble model based on k-nearest neighbor nonparametric estimation of regression function and its application for rainfall forecasting. *Int. Joint Conf. on Computational Sciences and Optimization*. Sanya, China, 41–53. doi:[10.1109/CSO.2009.307](https://doi.org/10.1109/CSO.2009.307).
- Xavier, A.C.F., et al., 2019. Stationary and nonstationary detection of extreme precipitation events and trends of average precipitation from 1980 to 2010 in the Paraná River basin, Brazil. *International Journal of Climatology*, 40, 1–16. doi:[10.1002/joc.6265](https://doi.org/10.1002/joc.6265).
- Zhang, Y. and Post, D., 2018. How good are hydrological models for gap-filling streamflow data?. *Hydrology and Earth System Sciences*, 22 (8), 4593–4604. doi:[10.5194/hess-22-4593-2018](https://doi.org/10.5194/hess-22-4593-2018).