Journal of Hydrology 598 (2021) 126454

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine learning algorithm

Pedro Arriagada [a,*], Bruno Karelovic [b], Oscar Link [c]

[a] Department of Environmental Engineering, Faculty of Environmental Sciences, Universidad de Concepción, Chile
[b] Department of Computer Science, Faculty of Engineering, Universidad de Concepción, Chile
[c] Department of Civil Engineering, Faculty of Engineering, Universidad de Concepción, Chile

ARTICLE INFO

This manuscript was handled by Andras Bardossy, Editor-in-Chief,

Keywords:
MissForest
Gaps filling
Daily flow
Discharge records

ABSTRACT

Complete hydrological time series are crucial for water and energy resources management and modelling in a changing climate. The reliability of the non-parametric stochastic machine learning algorithm MissForest was assessed for gap-filling of daily streamflow time series in a data-scarce region with strong climatic variability. A total of 1,586 reconstructions of streamflows for 1970–2016 were analyzed. Overall, MissForest performed satisfactorily to well, allowing a precise and reliable simulation of the missing data quickly and automatically. MissForest performance increased with the number of predictor records and record length, achieving satisfactory results with 20 or more records having 15 or more years in length. Reconstructed daily streamflow time series of rivers with natural flow regimes were simulated with good performance, which slightly decreased for discharge magnitude alterations by runoff inputs from urbanized areas and water diversion for irrigation. In cases of severe alterations of the flow regime, such as by hydropeaking, MissForest failed at filling daily streamflow series gaps. Reconstructed hydrographs allow analysis of streamflow change and variability and their interactions with key climatic variables.

1. Introduction

Complete hydrological time series are crucial for management and modeling of water, energy and other natural resources in a changing climate (Arriagada et al., 2019; Tencaliec et al., 2015). Data gaps cause difficulties in data interpretation, ineffective model calibration, unreliable timing of peak flows and biased statistics (Dembélé et al., 2019; Starrett et al., 2010), but are inherent to daily streamflow series for a number of reasons related to limited economic resources and political conflicts, such as sporadic operation of gauge stations, blackouts of the measuring devices, effects of extreme weather events, limited access to download data from loggers located in remote areas, scarcity of observers and human errors (Dembélé et al., 2019; Elshorbagy et al., 2000; Harvey et al., 2012). Incomplete streamflow series are more frequent and gaps are generally longer in developing countries (Amisigo and van de Giesen, 2005; Dembélé et al., 2019; Gyau-Boakye and Schultz, 1994; Sidibe et al., 2018), where so-called data-scarce regions, i.e., areas with a gauge density below World Meteorological Organization standards (WMO, 2008), also occur. At the same time, these regions are under the greatest pressure to develop water use infrastructure (Vörösmarty et al., 2010).

Daily streamflow time series of rivers and streams with natural flow regimes (Poff et al., 1997) are more suitable for successful application of gap-filling methods (e.g., Dembélé et al., 2019; Sidibe et al., 2018; Vega-Garcia et al., 2019), as the flow regime is characterized by the magnitude, frequency, duration, timing and rate of change, which respond to global and regional climate drivers such as modes of climate variability, jet streams, storm tracks and atmospheric rivers, as well as to river basin characteristics such as land uses, geology, vegetation and topography (McGregor, 2019). By contrast, in regulated rivers, alteration of one or more attributes of the flow regime due to human activities such as energy production, flood protection, irrigation, industrial and recreational activities and urbanization can introduce heavy artificial influences (e.g., Mackay et al., 2014) and significantly complicate automatic computation of the missing streamflow values from neighboring gauge stations (Harvey et al., 2010, 2012). In analysis of streamflow gauge series on a large spatial scale, both classes of streamflow data are sometimes mixed, coming from rivers with natural and regulated flow regimes, challenging gap-filling methods (Dembélé et al., 2019; Sidibe et al., 2018). Especially, when the region lacks basic information on

* Corresponding author.
E-mail address: parriagada@udec.cl (P. Arriagada).

https://doi.org/10.1016/j.jhydrol.2021.126454
Received 27 March 2020; Received in revised form 7 May 2021; Accepted 11 May 2021
Available online 14 May 2021

Table 1
Location, geomorphological and climate data for each basin in the study area.

Basin	Latitude(° ‘)	Longitude(° ‘)	Area (km^2)	Maximum altitude(m)	Predominant Climate according to Köppen Classification	Flow regime*	Annual precipitation(mm) **	Annual mean river flow (m^3/s)
Maipo	32°55′-34°18′ S	69°48′-71°38′ W	15,273	6,546	Csa-Csb	Snowmelt	650	134
Rapel	33°54′-35°00′ S	70°01′-71°51′ W	13,766	5,138	Csa-Csb	Snowmelt-rain	882	169
Mataquito	34°48′-35°38′ S	70°24′-72°11′ W	6,332	4,058	Csb	Snowmelt-rain	1373	113
Maule	35°06′-36°35′ S	70°21′-72°27′ W	21,052	3,931	Csb	Snowmelt-rain	1400	495
Itata	36°12′-37°20′ S	71°02′-72°52′ W	11,326	3,178	Csb	Snowmelt-rain	1764	331
Biobío	36°52′-38°54′ S	70°50′-73°12′ W	24,369	3,487	Csb	Rain	1873	971
Imperial	37°49′-38°58′ S	71°27′-73°30′ W	12,668	3,066	Csb-Cfb	Rain	2056	264
Toltén	38°36′-39°38′ S	71°24′-73°14′ W	8,448	3,710	Cfb	Rain	2062	540
Valdivia	39°18′-40°12′ S	71°36′-73°24′ W	10,244	2,824	Cfb	Rain	2592	546
Bueno	39°54′-41°17′ S	71°40′-73°43′ W	15,366	2,410	Cfb	Rain	2861	394

* fromValdés-Pineda et al. (2014)
** estimated from DGA (2016)

water uses and extractions to perform a proper flow classification for clustering or filtering out reference stations, a non-parametric gap filling method is desired.

Techniques for infilling missing flow data vary from simple interpolation to models and complex statistical analysis (Gyau-Boakye and Schultz, 1994). A classification of existing methods for infilling gaps in streamflow time series according to their mathematical complexity was provided by Harvey et al. (2012), who distinguished six classes of methods, namely manual inference, serial interpolation techniques, scaling factors, equipercentile techniques, linear regression and

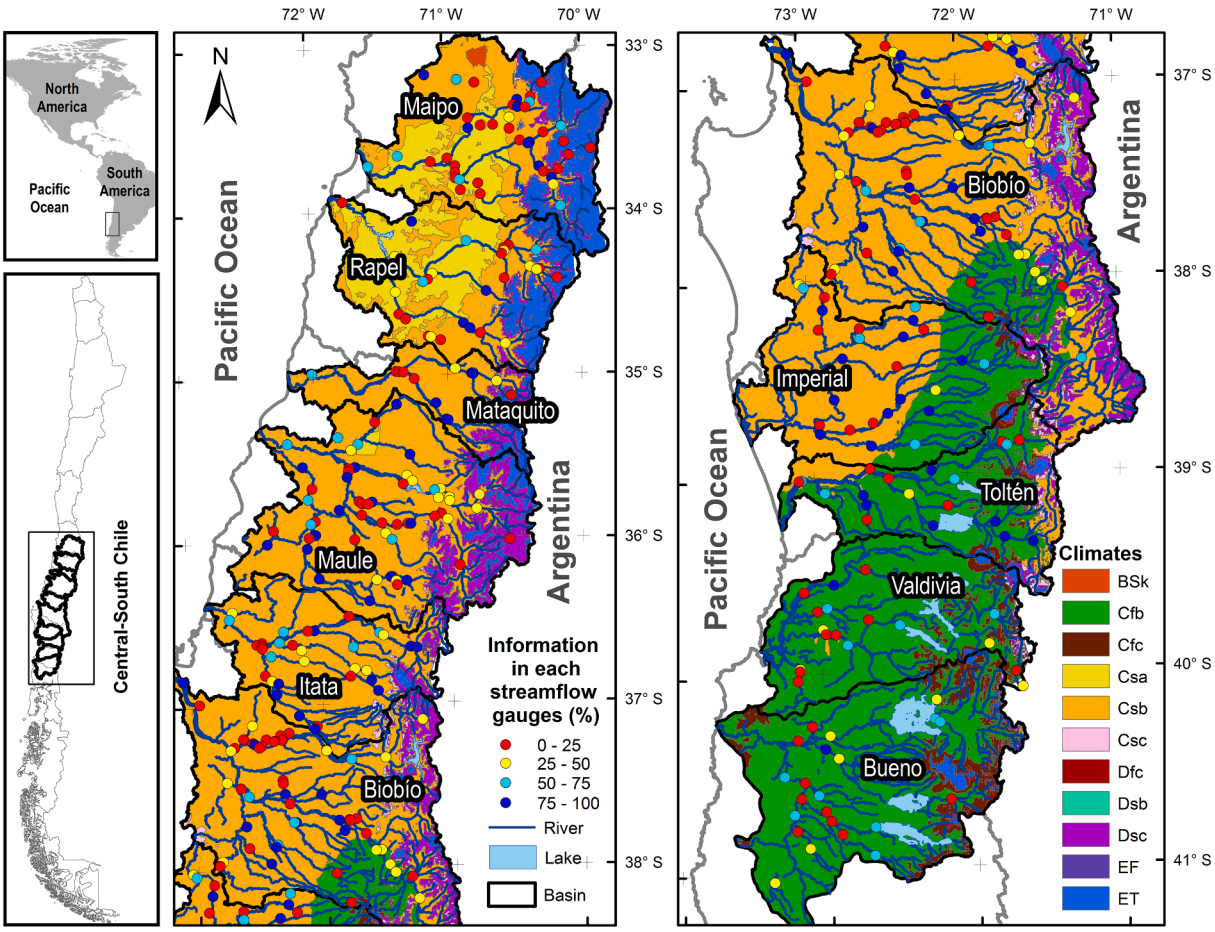

Fig. 1. Location of the study area, climates, and locations of streamflow gauges, including the data availability for the 1970–2016 period.

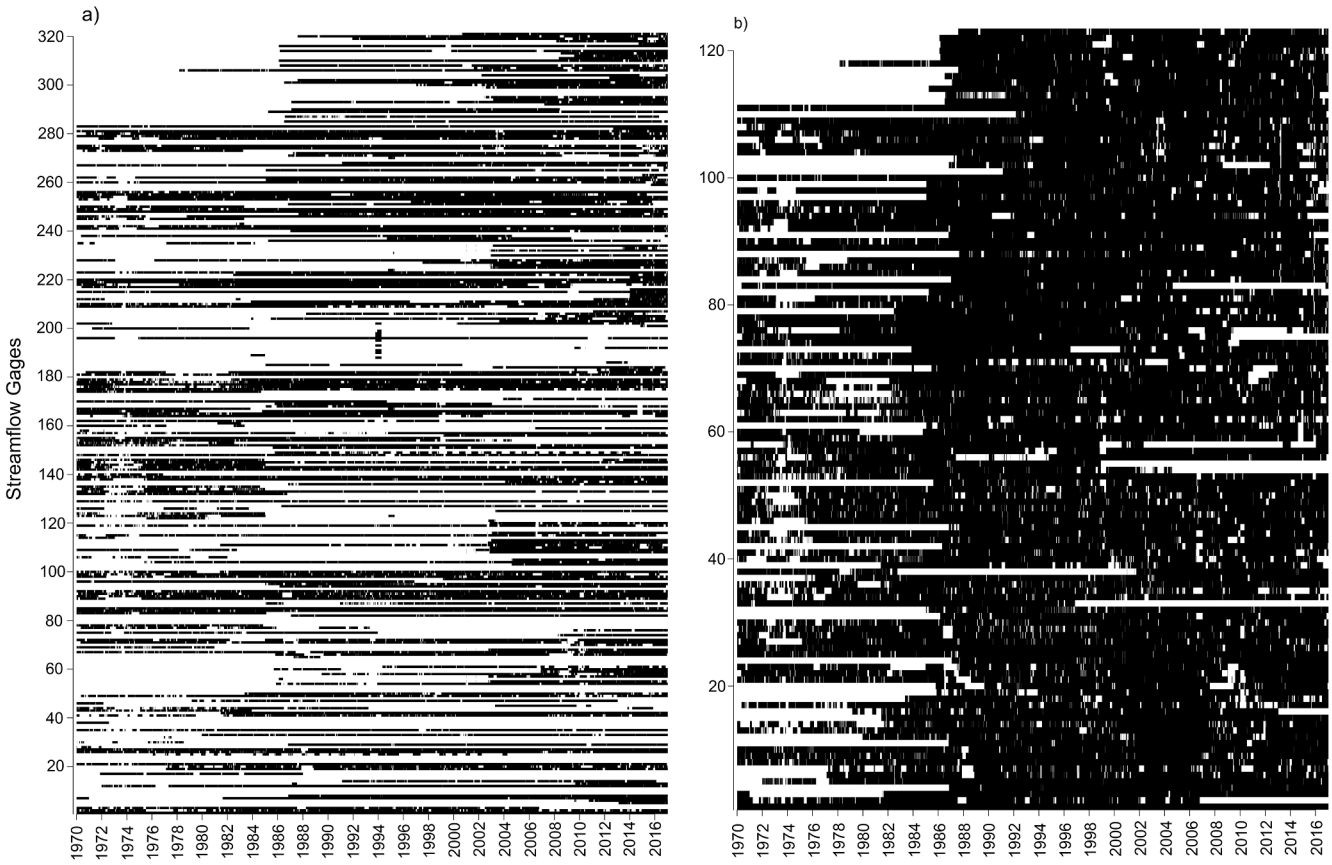

Fig. 2. Available and missing data: a) all streamflow gauges in study area, b) 122 streamflow gauges data records > 50% complete in the study period (1970–2016).

hydrological modelling. Further, a number of machine learning methods have been applied to infill missing flow data, including artificial neuronal networks (e.g., Ben Aissia et al., 2017; Kim et al., 2015; Mwale et al., 2012; Vega-Garcia et al., 2019), random forest models (Petty and Dhingra, 2018; Sidibe et al., 2018) and stochastic non-parametric methods such as direct sampling (Dembélé et al., 2019).

In particular, random forest by Breiman (2001) is a non-parametric machine learning algorithm for data simulation based on a combination of tree predictions. It was extended by Stekhoven and Bühlmann (2012) to the MissForest algorithm for missing value imputation in mixed-type data series. Missforest takes a different approach than random forest by recasting the missing data problem as a prediction problem. Data are imputed by regressing each variable in turn against all other variables and then predicting missing data for the dependent variable using the fitted forest (Tang and Ishwaran, 2017). Potential advantages of MissForest models over other alternatives for infilling daily streamflow data in large regions are: (1) they can quickly handle large amounts of data and the missing data imputation is unsupervised and automatic, avoiding the determination of predictor stations (Sidibe et al., 2018); (2) they can handle multiple data gaps in the series (Tang and Ishwaran, 2017); (3) they are easy to implement in computational languages such as R, as they don't require initial setting and calibration of parameters (Muñoz et al., 2018); and (4), they achieve competitive predictive performance and are computationally efficient, making them suitable for real-world prediction tasks (Sidibe et al., 2018).

Random forest has been applied in different scientific contexts such as medicine (Deshmukh et al., 2019; Stekhoven and Bühlmann, 2012; Waljee et al., 2013), sensitive information protection (Marino et al., 2019) and food chemistry (Tao et al., 2019). In water resources (see Tyralis et al., 2019), random forests have recently been tested for reconstruction of monthly streamflows in regions with different climates (Sidibe et al., 2018) and for flash-flood forecasting (Muñoz et al., 2018). Modern methods, such as MissForest, have superior performances than older traditional methods as has been shown by Waljee et al. (2013). Particularly, MissForest outperforms the well-known k-nearest neighbors (KNN) method (Tang and Ishwaran, 2017; Troyanskaya et al., 2001), and parametric MICE (Tang and Ishwaran, 2017; Van Buuren, 2007). In hydrology, MissForest have a performance similar to modern methods such as Multivariate Imputation by Chained Equations (MICE; Van Buuren and Oudshoorn, 1999) as have been shown for infilling monthly streamflow data by Sidibe et al. (2018). The infilling of gaps in daily streamflow series in regions with different climates is more challenging than previous applications because streamflow temporal and spatial variability is higher. Clearly, the gap-filled data are more accurate when examining trends at the annual scale, followed by the monthly scale and, finally, the daily scale, at which the results are the least satisfactory (Zhang and Post, 2018).

The aim of the present work is to assess the reliability of the machine learning algorithm MissForest for automatic gap-filling of daily streamflow time series in a large, data-scarce region with a variety of different climates, using all available gauges at both, regulated and unregulated rivers, and streams.

2. Materials and methods

2.1. Study area

The study area of 152,351 km^2 includes 10 watersheds located in central Chile, between latitudes 32°55′ and 41°17′ S, and longitudes 69°48′ and 73°43′ W. The population of about 12,316,144 inhabitants (71% of the country; INE, 2018) is concentrated in a few big cities located in the Central Valley or coastal plain. Table 1 shows the important properties of the watersheds and Fig. 1 shows a map of the study area.

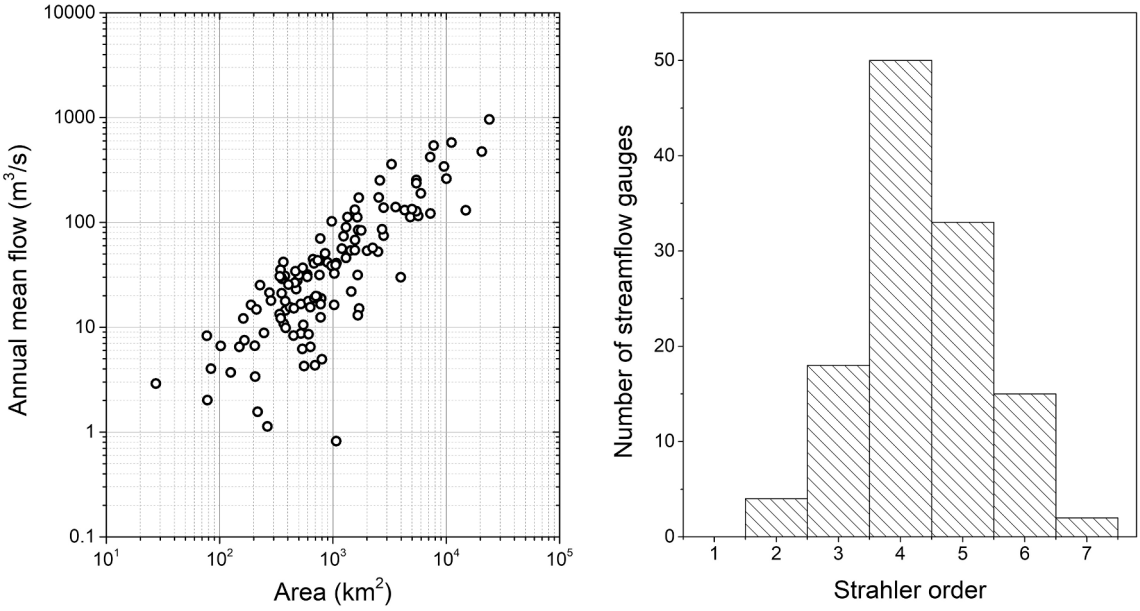

Fig. 3. Scatterplot of gauged catchment area versus mean flow with a dot for each gauge (left), and Strahler order of the reach at each gauge versus the number of gauges (right).

In the northern part of the study area, from Maipo to Biobío, the climate is dominated by the Pacific anticyclone (Garreaud et al., 2009; Valdés-Pineda et al., 2014). According to the Köppen classification (Beck et al., 2018), the climates present in this area are: Arid steppe cold (BSk); Temperate dry and hot summer (Csa); Temperate dry and warm summer (Csb); Temperate dry and cold summer (Csc); Temperate no dry season and warm summer (Cfb); Cold dry and warm summer (Dsb); Cold dry and cold summer (Dsc); Polar frost (EF) and Polar tundra (ET). The predominant climate is Csb (see Fig. 1). In the southern part of the study area, from Imperial to Bueno, the climate is dominated by the southern westerlies (Garreaud et al., 2009; Valdés-Pineda et al., 2018). According to the Köppen classification (Beck et al., 2018), the climates present in this area are: Temperate dry and warm summer (Csb); Temperate dry and cold summer (Csc); Temperate no dry season and warm summer (Cfb); Temperate no dry and cold summer (Cfc); Cold no dry season and cold summer (Dfc); Cold dry and cold summer (Dsc) and Polar tundra (ET). The predominant climate is Cfb (see Fig. 1). The Biobío and Imperial watersheds are located in a climatic transition area with mixed influence of the southeast Pacific anticyclone and the westerlies (Falvey and Garreaud, 2009; Valdés-Pineda et al., 2018). The climate variability is influenced by oscillations, with different periods such as the El Niño-Southern Oscillation (Escobar and Aceituno, 1998), the Pacific Decadal Oscillation (Montecinos and Aceituno, 2003), and the Antarctic Oscillation (Urrutia et al., 2011; Valdés-Pineda et al., 2018). About 30% of the winter storms are warm winter rainstorms caused by atmospheric rivers (Garreaud, 2013).

According to the Soil Taxonomy classification system, most of the study area is covered by six orders, namely Alfisols, Entisols, Inceptisols, Mollisols, Ultisols, and Vertisols, as well as Andisol and Histosol series (Bonilla and Johnson, 2012). Land surface slope values range from close to zero in the Central Valley – a geological depression with an approximately 70-km-wide plain formed between the Andes and the coastal range, extending south from Valparaíso (north of Maipo basin) for about 1000 km to the Araucanía Region – to 0.65 (m/m) in the Andes (Carretier et al., 2018). The study area includes most of the cultivated and productive land in the country, with the majority of farms (72%) and national forest area (54%) (Bonilla and Johnson, 2012). The rainfall regime, soil properties, high slopes and land uses make the study area particularly vulnerable to erosion processes (Bonilla and Johnson, 2012; Ellies, 2000). At the same time, the study area includes 91 of the 148 existing hydropower plants in Chile, with a total power of 5.05 GW, i.e., 76% of the national installed hydropower; in addition to these existing projects, 30 new hydropower plants with a total of 0.65 GW are under environmental evaluation or construction. Moreover, the exploitable hydropower of the study region has been estimated at about 12 GW, spread among 1200 sites, most of which are located in the Andes or the piedmont region (Arriagada et al., 2019). There is clearly a conflict between conservation of freshwater environments and both hydropower and irrigation (Habit et al., 2019; Laborde et al., 2016).

Table 2
Streamflow gauges in the study area.

Basin	Number ofSFG	Number ofdecommissioned SFG	Number ofSFG having over 50% of data in the study period	SFG density (km^2 per gauge) *
Maipo	46	17	16	332 (955)
Rapel	30	16	7	459 (1,967)
Mataquito	14	7	6	452 (1,055)
Maule	57	30	22	369 (957)
Itata	33	17	16	343 (708)
Biobío	55	30	15	443 (1,625)
Imperial	26	13	16	487 (792)
Toltén	19	13	11	445 (768)
Valdivia	18	9	5	569 (2,049)
Bueno	22	19	8	698 (1,921)

* Numbers in parentheses shows the SFG density based only on gauges with information that is at least 50% complete.

2.2. Streamflow data

In the study area streamflow is monitored at 320 gauges administrated by the National Water Agency (Dirección General de Aguas, DGA). Fig. 1 shows the location of the study area, climates, and locations of streamflow gauges (SFG), including the data availability for the 1970–2016 period.

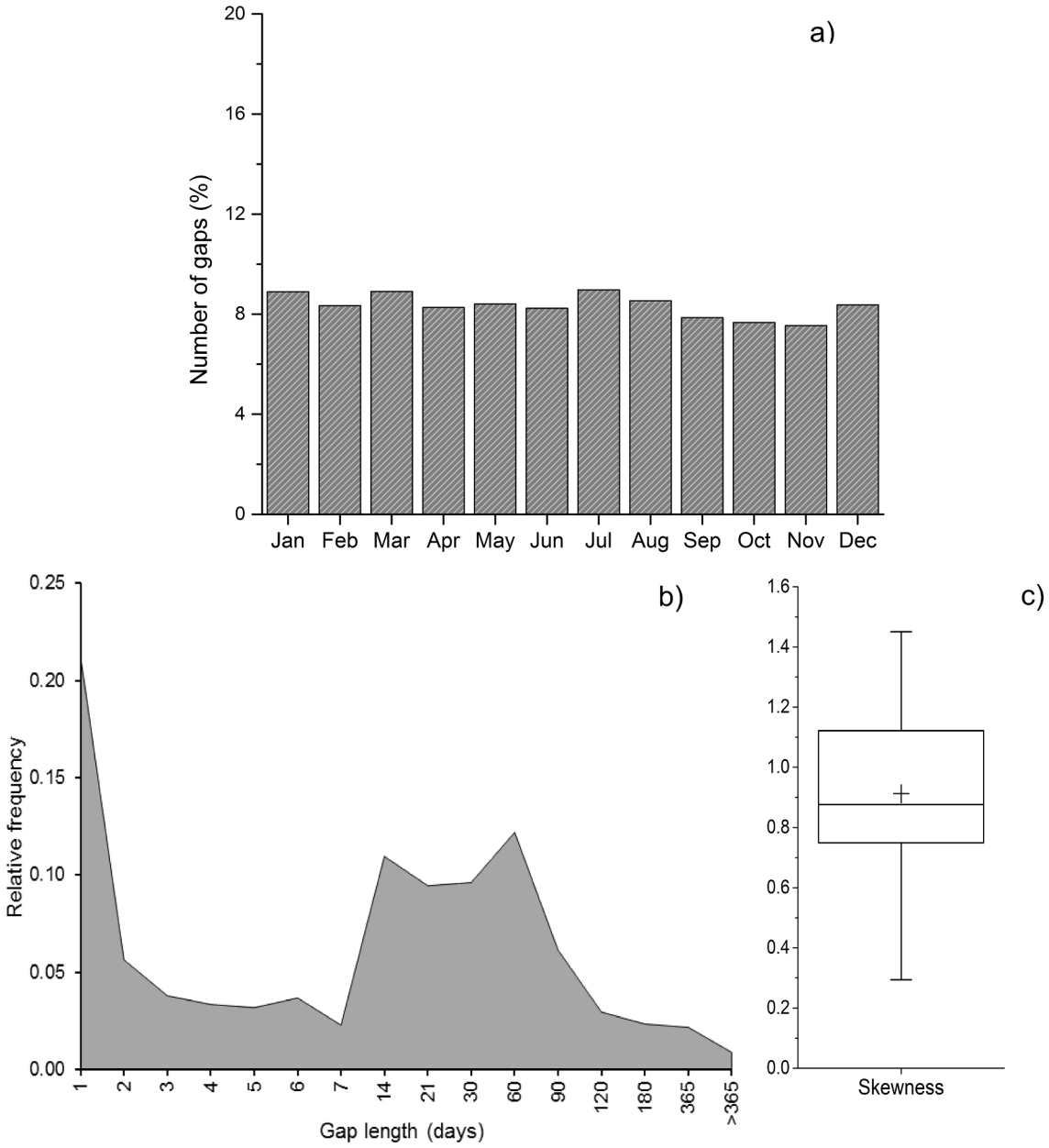

Fig. 4. Distribution of missed data along the months of the year (a) relative frequency of gap lengths (b), and skewness of daily streamflow series (c).

Fig. 2 shows the available and missing data at the 122 streamflow gauges with data records that are at least 50% complete in the study period.

Even though the overall streamflow gauge density in the study area, 456 km^2 per gauge station, is in compliance with WMO standards (WMO, 2008), there are 171 decommissioned gauges and only 122 gauges present data over 50% of the time between 1970 and 2016. The length of data required would depend on the intended purpose. Herein, we set the threshold for the acceptable percentage of missing data to consider a gauge station usable at 50%, which is quite challenging for any gap filling method. Consequently, considering only the 122 usable gauges, the gauge density is 1,238 km^2 per gauge, which is below WMO standards, and thus the study region is data scarce. Fig. 3 shows the scatterplot of gauged catchment area versus mean flow with a dot for each gauge, and the Strahler order of the reach at each gauge versus the number of gauges to inform about hydrological variability between gauges and sizes of gauged sub-catchments.

Table 2 shows the number of streamflow gauges in the study area, distinguishing between decommissioned gauges and those having records over 50% complete during the study period, and the gauge density.

Fig. 4 shows the distribution of missed data along the months of the year (a), relative frequency of gap lengths in the records (b), and skewness of daily streamflow series (c).

Missing data in the usable 122 gauges are uniformly distributed. Gap length was typically 1 day, or between 7 and 60 days. Skewness of daily streamflows at the 122 gauges used in our study presented a maximum skewness of 1.45, a mean of 0.91, a median of 0.88, and a minimum of 0.29. Positive skewness values indicate that the gauged records are all right skewed.

2.3. *The MissForest algorithm*

Random forests (RF; Breiman, 2001) grow many decision trees and average their results. Each node within each decision tree chooses a random subset of the variables and applies the bootstrap aggregation

technique, i.e., given the training set a random sample is selected m times with replacement of the training set and a decision tree to these samples is fitted. Thus, the correlation between the trees is reduced and better results are achieved. Stekhoven and Bühlmann (2012) extended the RF to the MissForest algorithm (MF) for missing value imputation in mixed-type data. MF consists of training a random forest iteratively on observed variables for prediction of the missing values. Defining $X = (X_{1,\ldots,}X_p)$ as a data set of $n \times p$ dimensions, corresponding to p streamflow gauges with n recorded daily streamflows. For a given gauge X_s, let $i_{miss}(s)$ be the set of days where station s presents missing values. Then, the dataset is separated into four parts:

$Y_{obs}^{(s)}$: The observed streamflow values at gauge X_s

$Y_{mis}^{(s)}$: The missing values at gauge X_s

$X_{obs}^{(s)}$: The observed streamflow at another gauge in days $\{1, \ldots, n\} \setminus i_{miss}(s)$

$X_{mis}^{(s)}$: The missing streamflow at another gauge in days $i_{miss}(s)$

Note that $X_{obs}^{(s)}$ can have missing values and $X_{mis}^{(s)}$ can contain observed streamflows.

Our goal is to fill the missing values $Y_{mis}^{(s)}$. To do so, the main idea is to train a random forest to predict $Y_{obs}^{(s)}$ from $X_{obs}^{(s)}$ and then to use this trained random forest to predict our missing values at gauge X_s ($Y_{mis}^{(s)}$) from $X_{mis}^{(s)}$. Nevertheless, there could be some missing values in $X_{mis}^{(s)}$ and $X_{obs}^{(s)}$, in which case we should fill these values as a first step as follows: the average recorded daily streamflows at each gauge X_t during the study period are imputed to each missing value of gauge t.

Now, gauges are sorted by first identifying those with less missing data. For each value X_s, the missing values are imputed by fitting a random forest with input $X_{obs}^{(s)}$ and output $X = (X_{1,\ldots,}X_p)$. Next, missing values $Y_{mis}^{(s)}$ are predicted by the trained random forest with input $X_{mis}^{(s)}$. The imputation procedure is repeated until the difference between the newly imputed data and the previous one increases for the first time. More precisely, let X_k^{imp} be the previously imputed data in the k-th iteration and X_{k+1}^{imp} be the updated imput in the (k + 1)-th iteration. The difference (Δ) is calculated as follows:

$$\Delta_k = \frac{\sum i \in X\left(X_{k+1}^{imp} - X_k^{imp}\right)^2}{\sum i \in X\left(X_{k+1}^{imp}\right)} \tag{3}$$

The stop criterion is met as soon as Δ_{k+1} is larger than Δ_k. One thousand regression trees were used in all computations based on previous experiences by Arriagada et al. (2019), and the maximum number of iterations was set to hundred, i.e., a sufficiently large number to ensure fulfilment of the convergence criterion in Eq. (3). The algorithms were implemented using R 3.6.1 (R Core team, 2019), the hyfo (v1.4.0; XU, 2018), hydroGOF (v0.3–10; Zambrano, 2017), lubridate (v1.7.4; Grolemund and Wickham, 2011) and MissForest (v1.4; Stekhoven, 2013) packages.

2.4. Synthetic missing data scenarios and method performance

Missing daily streamflows in the study area distributed uniform over the year (see Fig. 4a). Two types of artificial gaps were generated, namely a) Removed contiguous segments: at each gauge only a segment (having lengths of 7, 14, 21, 30, 60, 180, and 365 days) was randomly removed from the entire record (1970–2016); b) Removed single data points: observed values (30, 60, 90, 120, 180 and 365 days) were randomly removed from the entire record (1970–2016) at each of the gauges. MissForest was applied to infill the gaps contained in the records together with the artificial gaps. Our analysis includes 13 reconstructions of the 1970–2016 period at each of the 122 streamflow gauges (which included gauges at rivers and streams with natural and altered flows), i.e., 1.586 simulations. The reconstruction of a given

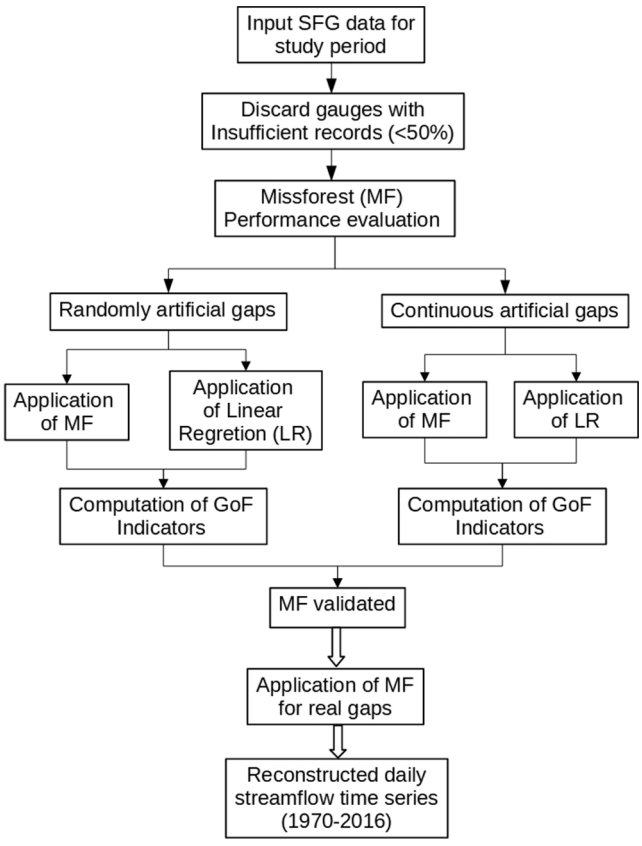

Fig. 5. Flowchart illustrating the workflow step by step.

scenario (e.g., removed continuous segment having 14 days in length) was repeated 122 times, which are considered replicates of the experiments.

The performance of MissForest at infilling daily streamflow data was tested by comparing the filled series with the observed data using goodness-of-fit indicators (GoF): coefficient of determination (R^2), the percent bias (*PBIAS*), and the Kling-Gupta efficiency (*KGE*) (Kling et al., 2012):

$$R^2 = \left[\frac{\sum_{i=1}^{n}(O_i - \mu_o)(S_i - \mu_s)}{\sqrt{\sum_{i=1}^{n}(O_i - \mu_o)^2}\sqrt{\sum_{i=1}^{n}(O_s - \mu_s)^2}}\right] \tag{4}$$

$$PBIAS = \left[\frac{\sum_{i=1}^{n} S_i - O_i}{\sum_{i=1}^{n} O_i}\right] x100 \tag{5}$$

$$KGE = 1 - \sqrt{(r-1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{6}$$

$$\beta = \frac{\mu_s}{\mu_o} \tag{7}$$

$$\gamma = \frac{\sigma_s/\mu_s}{\sigma_o/\mu_o} \tag{8}$$

where O and S are the observed and simulated data, μ is the mean, σ is the standard deviation, r is the correlation coefficient between simulated and observed data, β is the bias ratio and, finally, γ is the variability ratio. The optimal value of R^2 and *KGE* is one, while the optimal value of PBIAS is zero. The threshold values for satisfactory, good and very good performance are: $0.60 < R^2 \leq 0.75$, $0.75 < R^2 \leq 0.85$, $R^2 > 0.85$ and $\pm 15 >$ PBIAS $\geq \pm 10$, $\pm 10 >$ PBIAS $\geq \pm 5$, and PBIAS $< \pm 5$ (Moriasi et al., 2015). Knoben et al. (2019) distinguished two classes of performance according to KGE, namely good for KGE > -0.41 and bad for KGE <

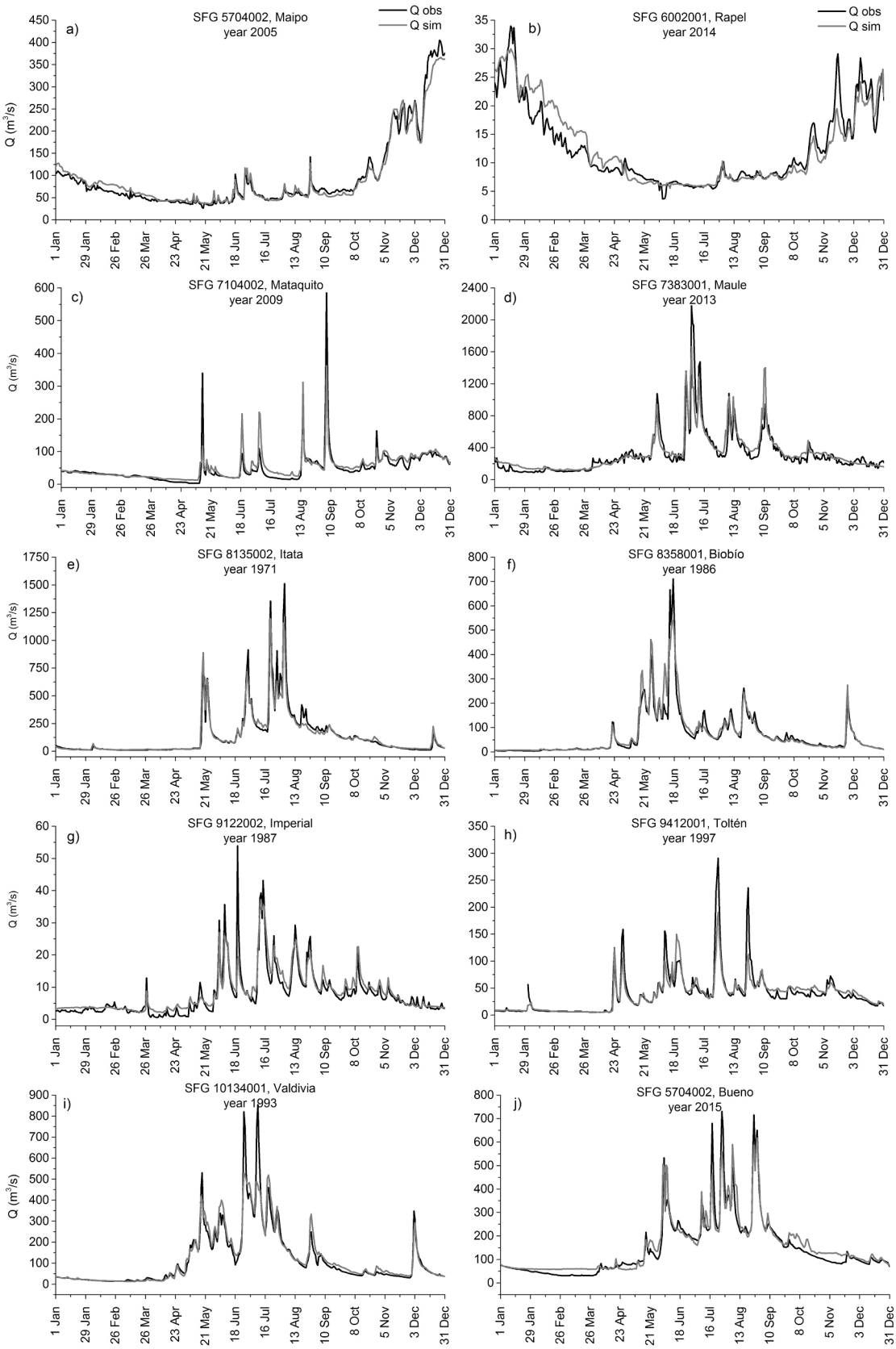

Fig. 6. Observed and simulated hydrographs in streamflow gauges located in the a) Maipo River in the Andes, with a Csb climate; b) Rapel River in the Andes, with a Csb climate; c) Mataquito River in the Andes, with a Csa climate; d) Maule River in the coastal plain, with a Csb climate; e) Itata River in the Central Valley, with a Csb climate; f) Biobío River in the coastal plain, with a Csb climate; g) Imperial River in the Andes with a Cfb climate, h) Toltén River in the Andes, with a Cfb climate; i) Valdivia River in the Central Valley, with a Cfb climate; and j) Bueno River in the Andes, with a Cfb climate.

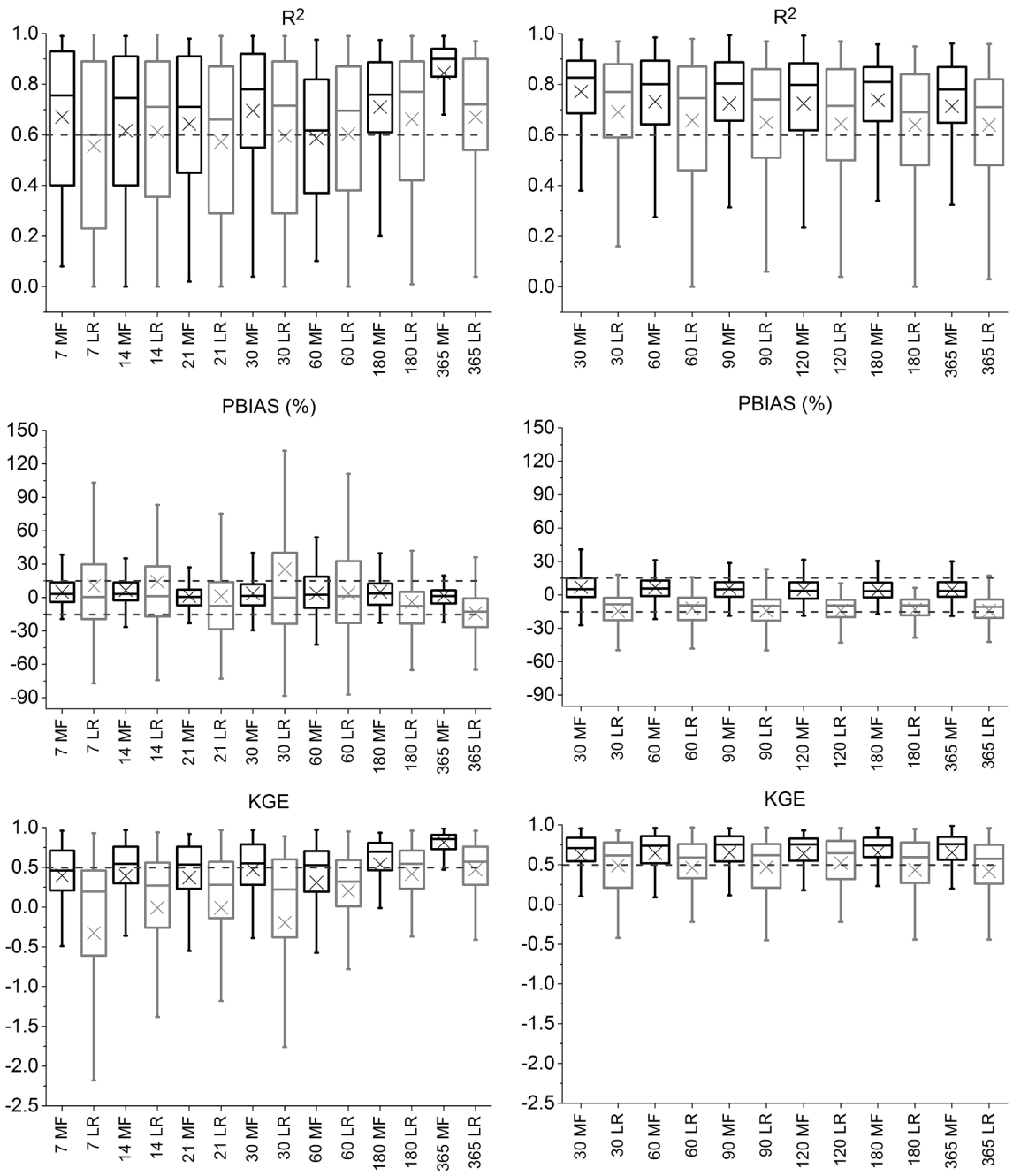

Fig. 7. Performance of MissForest (MF, black boxplots) and Linear Regression in log–log space (LR, gray boxplots) for removed contiguous segments (left) and removed single data points (right). The crosses indicate the mean value, and the whiskers show the 5th and 95th percentiles. Dashed lines correspond to the limit for a satisfactory performance.

-0.41.

The method was benchmarked through a comparison of the obtained results with those computed applying the linear regression (LR) in log–log space method. The linear regression was computed using all coincident non-missing days at the gauge of interest and the nearest gauging station. The nearest gauging station was that with the shortest linear distance to the gauge of interest. Fig. 5 shows a flow diagram for automatic gap-filling of daily streamflow data using MissForest.

To investigate if MissForest is sensitive to the number of non-missing days within the record which is being filled, we determined how long does a record have to be (how many non-missing days) before MissForest can be used to infill all remaining missing data, by keeping (for training) different numbers of non-missing days (2, 3, 4, 5, 10, 15, 20, 30 and 47 years) whilst predicting on different numbers of simulated missing days (30,180 and 365 days). We also evaluated the accuracy of predicting missing flows related to the number of records included in MissForest, incrementally increasing the number of gauges used as predictors (6, 12, 18, 24, 37, 49 and 61 gauges which correspond to 5, 10, 15, 20, 30, 40 and 50% of the available gauges).

2.5. *Effects of altered flows on MissForest performance*

A generalized analysis of gauges at rivers and streams with altered versus unaltered flows was not feasible. The study region lacks sufficient basic information for a classification representing the degree and type of flow alteration. However, the effects of altered flows on MissForest

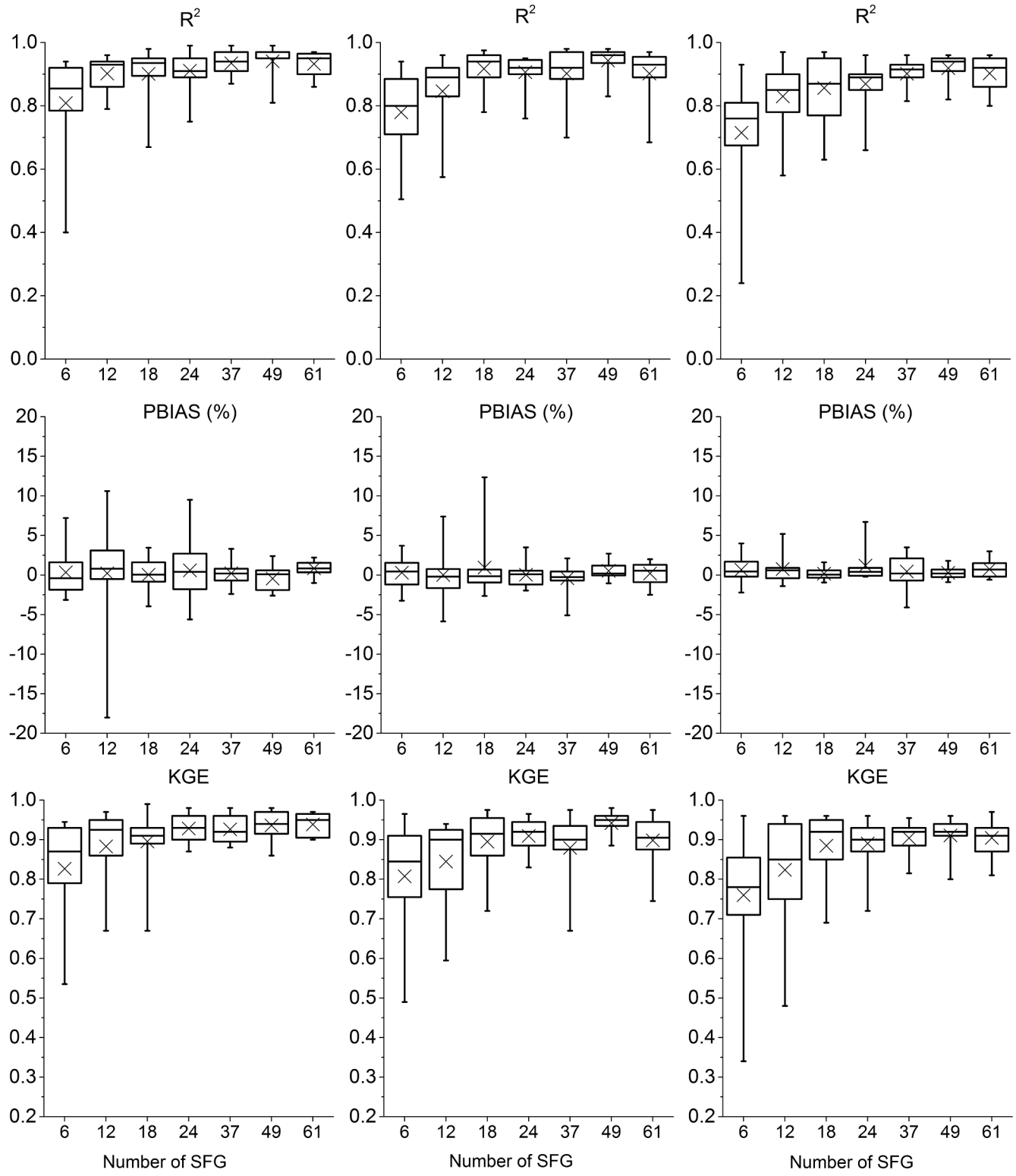

Fig. 8. MissForest performance for different numbers of non-missing days (2, 3, 4, 5, and 47 years) whilst predicting on different numbers of simulated missing days 30 (Left), 180 (center) and 365 (right).

performance were investigated by creating a 365-day-long gap, in a year missing less than 20% of data at selected gauges where human alterations, such as water diversion to an irrigation channel, surface runoff inputs from urbanized areas, and upstream hydropower operation occurred. Again, MissForest performance from the set of 122 gauges (that contained both altered and unaltered gauges) was tested by comparing filled and observed series using goodness-of-fit indicators R^2, *PBIAS*, and *KGE*.

2.6. Reconstruction of streamflow records

As an application case, MissForest was used for the reconstruction of streamflow records at the 122 gauges located in the study area for the 1970–2016 period.

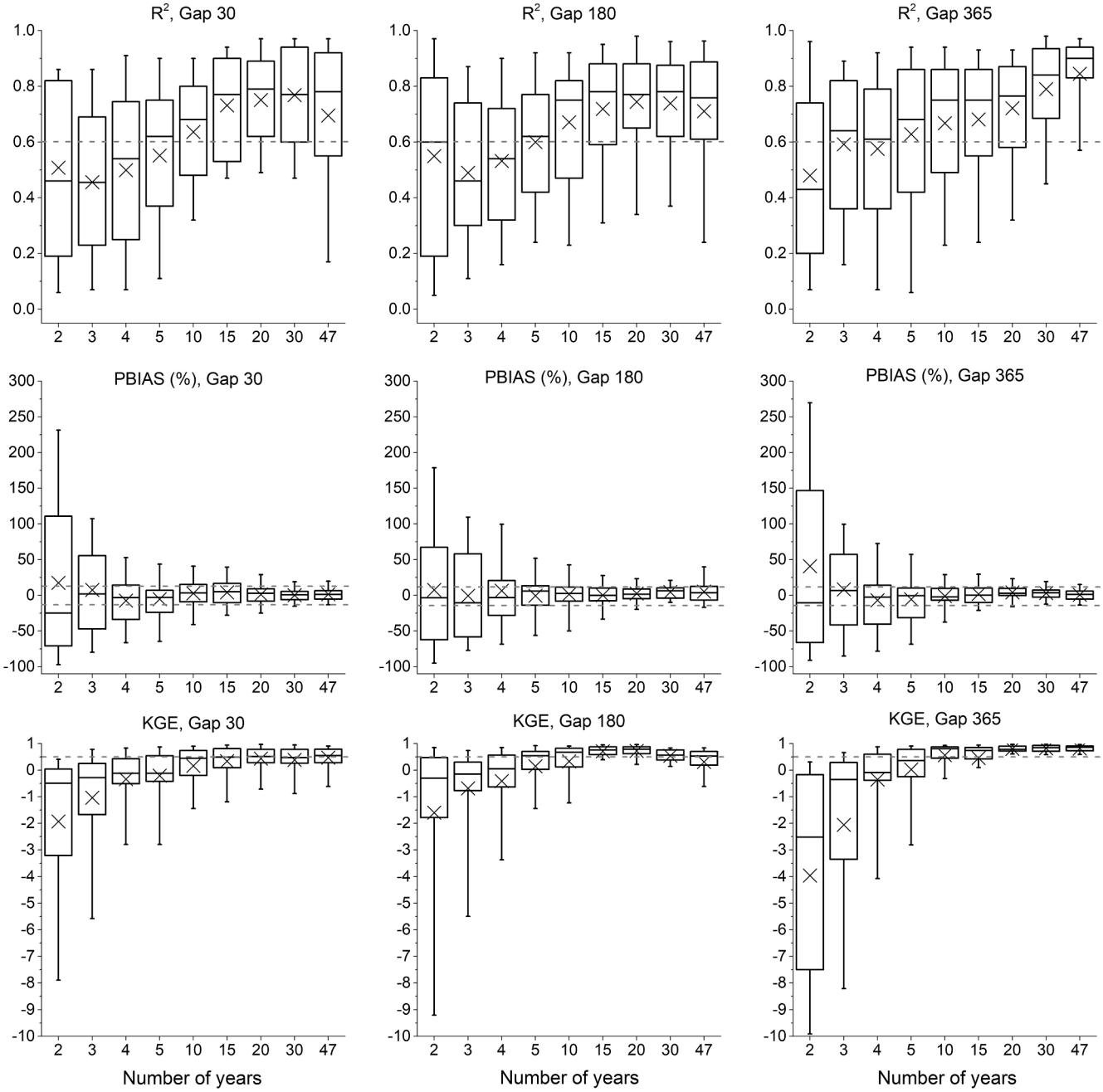

Fig. 9. R^2, *PBIAS* and *KGE* of MissForest using different number of records as predictors for filling gaps of 2% (left), 5% (middle), and 50% (right) of data randomly taken out at a given gauge.

3. Results

MissForest was applied to the subset of streamflow gauges with less than 50% missing data, i.e., 122 streamflow gauges here. In all presented computations ten or fewer iterations were needed to satisfy the convergence criterion in Eq. (3).

3.1. MissForest performance at gap-filing of daily streamflow time series

Fig. 6 shows observed and simulated hydrographs in streamflow gauges located in different climates and geographic units.

Overall, the shape of the observed hydrograph is well reproduced, with good timing and representation of the annual seasonality in all cases, suggesting that climate diversity and the geographic units in which the different gauges are located are not important controls for MissForest performance. The simulated hydrographs match observed low flows in rainfall-dominated flow regimes (dry season corresponds to summer from December 21 to March 21) as well as high flows in watersheds with rain, snow and/or glacial melting.

Fig. 7 shows the performance of MissForest and Linear Regression in log–log space for infilling removed contiguous segments and single data points.

MissForest performed well at infilling single data points and contiguous segments (average values of $R^2 > 0.6$, *PBIAS* $< \pm 15\%$ and *KGE* > 0.5 in both cases). Mean values of R^2, *PBIAS,* and *KGE* were different in both cases (Mann-Whitney U Test, $U = 6.0$, $p = 0.038$), and slightly higher for the infilling of single data points, illustrating that for MissForest temporal correlations are important. Mean values of R^2, *PBIAS,* and *KGE* did not change significantly with the number of removed single data points (Mann-Whitney U Test by R^2, $U > 7456$, $p >$

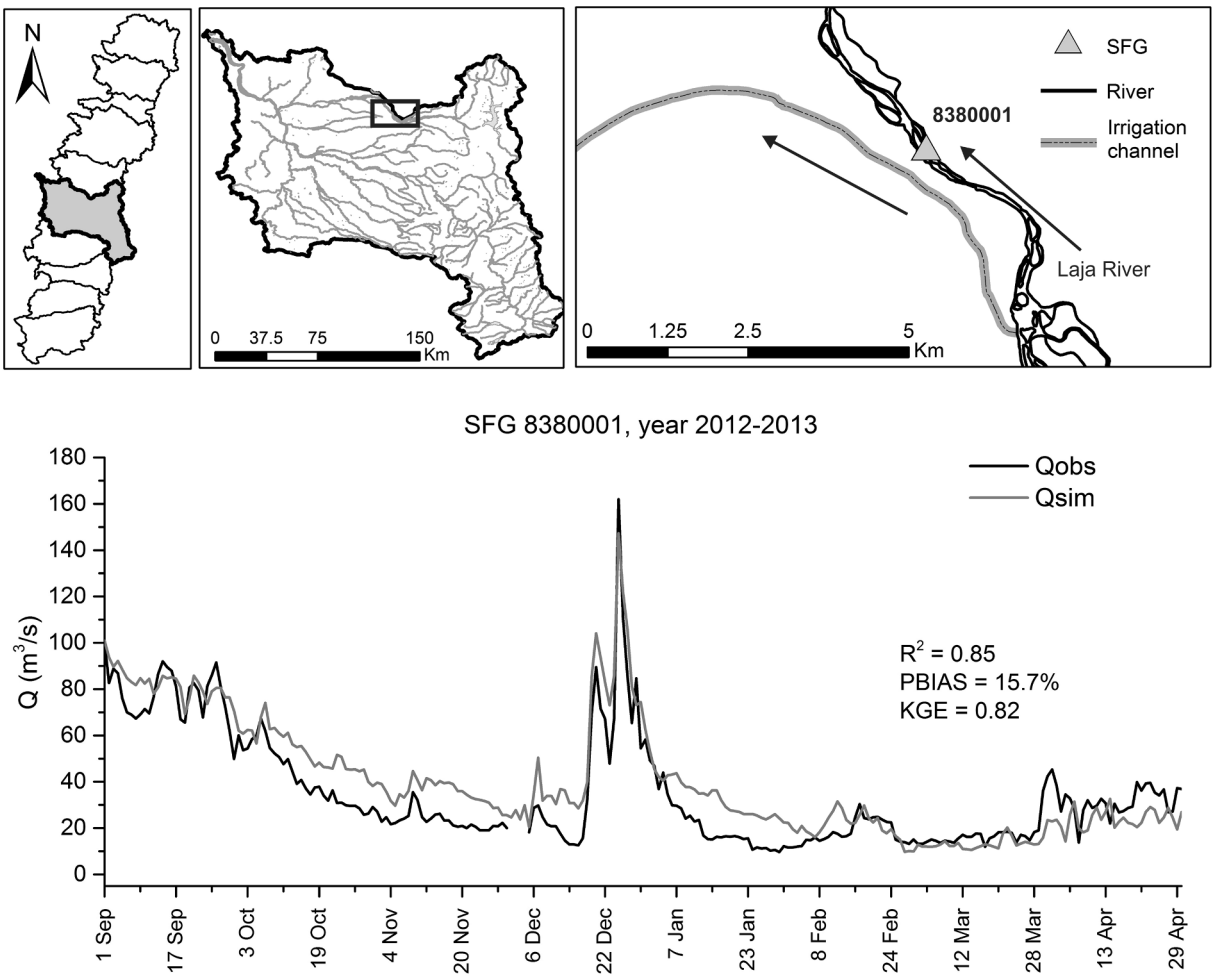

Fig. 10. Observed and simulated hydrographs at a gauge located downstream of a water diversion for irrigation.

0.224; *PBIAS*, U > 7239, p > 0.539; and *KGE* U > 7068, p > 0.497), nor with the length of removed continuous segments for lengths up to 60 days (Mann-Whitney U Test by R^2, U > 6774, p > 0.162; *PBIAS*, U > 7366, p > 0.386; and *KGE* U > 6729, p > 0.196). However, R^2 and *KGE* presented a significant increase for removed continuous segments longer than 60 days (Mann-Whitney U Test by R^2, U > 4276, p less than 0.0007; and *KGE* U > 3480, p < 0.0008). These results suggest that MissForest performance is not highly sensitive to the amount of missing data. Dispersion of R^2, *PBIAS,* and *KGE* was important in all cases and higher when infilling contiguous segments. These high dispersion values represent important differences in the quality of reconstructed hydrographs at the different gauges, suggesting that external factors such as altered flow regimes play an important role in MissForest performance for infilling daily streamflow time series; thus, such cases are analyzed in further detail in section 3.2.

MissForest outperformed the Linear Regression in log–log space. On average, all performance indices were better, and dispersion was smaller. Compared to the Linear Regression in log–log space, MissForest produced higher minimum and maximum performances. However, our analysis also shows that Linear Regression produces reasonably good results for infilling gaps of 1-day duration (see Fig. 7 right column). This means that the data from one nearby gauge station can be used to fill gaps in this case. In addition, MissForest tend to overestimate the infilling gaps of 1-day duration (see PBIAS in the right column of Fig. 7), in line with results by Janitza & Hornung (2018) who showed that this is a typical error in regressions trees when multiple predictors for infilling a single point area used.

Fig. 8 shows the performance of MissForest for different numbers of non-missing days.

In case of short records, i.e., 2 to 5 years long, MissForest performance improved considerably with the number of non-missing days for missing data imputation. When records get longer than 15 years, however, PBIAS and KGE tend to converge to a constant value in the range of satisfactory or better. KGE results indicate that often>15 years-worth of non-missing days are needed to produce adequate performance

Fig. 9 shows the R^2, *PBIAS* and *KGE* of MissForest using different number of records as predictors for filling gaps of 2% (365), 5% (858), and 50% (8584) of data randomly taken out at a given gauge.

MissForest performance increased with the number of predictor gauges, from a minimum of 6 gauges and up to 24 gauges. After that, i.e., when>25 gauges are used as predictors, the MissForest performance indices tend to converge to a constant value.

3.2. *Effects of altered flows on MissForest performance*

Our aim is to evaluate MissForest for gap filling at large, data scarce regions. Such regions lack basic information about water uses and extractions and illegal water diversions can even worsen the problem. Thus, our goal is to provide a method to predict missing flow values regardless of whether they are from altered or unaltered gauges. Even though, the presence of gauges with altered flows is a plausible reason for a reduced performance of MissForest at filling data gaps there, a generalized analysis of altered versus unaltered flows is not feasible, as the study region lacks sufficient basic information for a proper flow classification. Alternatively, we present three illustrative cases to exemplify how altered flows can affect the performance of the method.

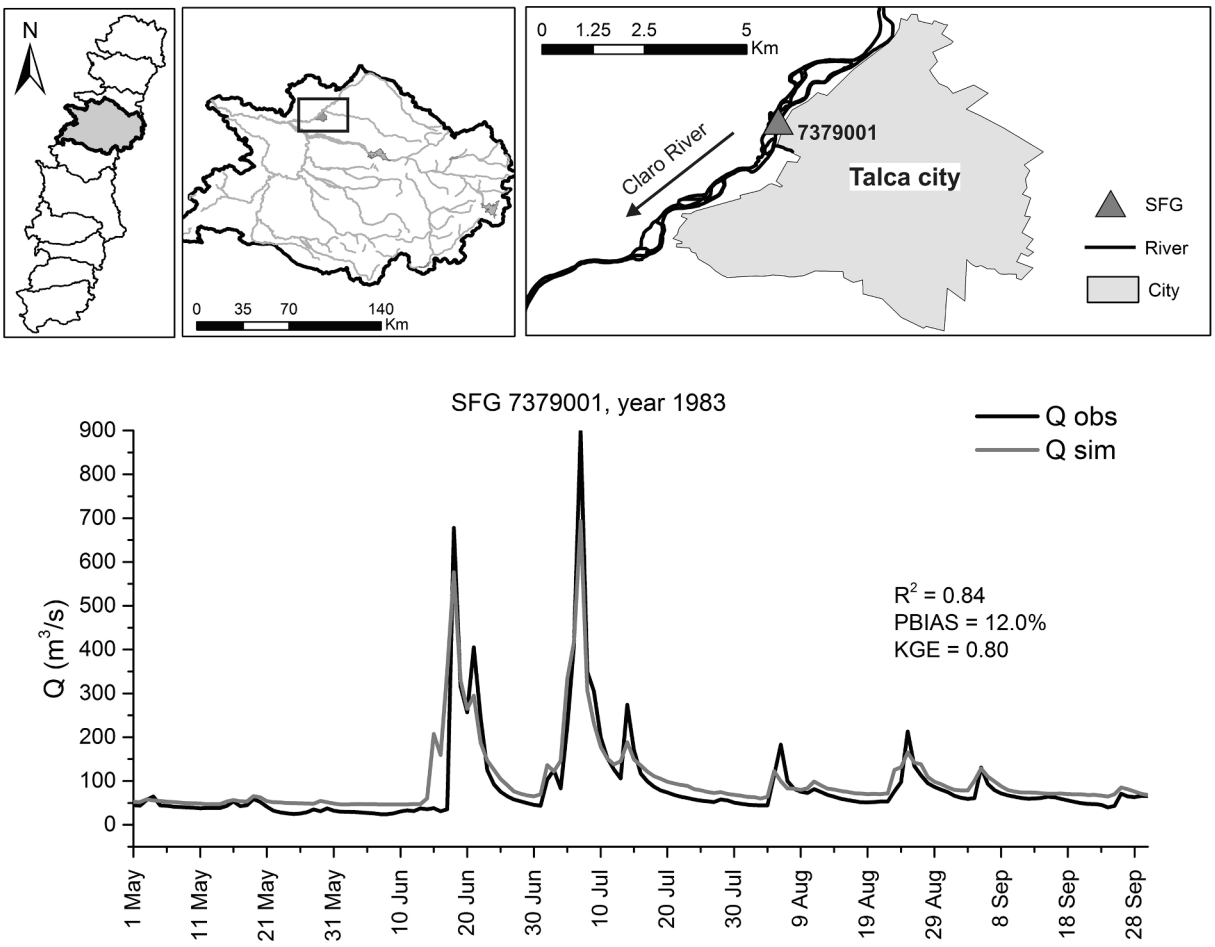

Fig. 11. Observed and simulated hydrographs at a gauge downstream of urbanized areas.

3.2.1. Water diversion for irrigation

The study area includes most of the cultivated and productive land in the country, including most of its farms, and thus it presents several water intakes placed in rivers and streams to divert water for irrigation. The Laja River (Biobío basin) presents significant water diversions for agriculture (Mardones and Vargas, 2005). The irrigation period is usually from October 15 to March 31, altering the natural flow regime downstream of the water intakes. Fig. 10 shows measured and simulated hydrographs at a gauge downstream of a water intake that diverts water for irrigation during the irrigation period (October-March).

Clearly, the water diversion for irrigation altered the natural flow regime, diminishing the discharges during the irrigation period. In this case, the simulated flows are higher than the observed flows. However, MissForest performance is satisfactory to good (R^2 = 0.85, PBIAS = 15.70%, KGE = 0.82), showing that systematic alterations of discharge can still be followed by MissForest, at least in terms of trends. Remarkably, a small flood that occurred around Christmas forced the water intakes to close, and the streamflow imputation was correct in this period.

3.2.2. Runoff inputs from urbanized areas

Areas that are impervious due to urbanization increase surface runoff and thus water contribution from these areas increases peak discharges of rivers during storms. Fig. 11 shows measured and simulated hydrographs at a gauge located downstream of rainwater inputs form the city of Talca, near the Claro River (Maule basin).

Simulated peak discharges are clearly lower than observed discharges during the storms, and consequently during the recession limb of floods the simulated discharges are higher than the observed discharges, evidencing the effects of urbanization on river discharges. In such cases, MissForest performance decreased to satisfactory (R^2 = 0.84, PBIAS = 12.0%, KGE = 0.80).

3.2.3. Discharge regulation for hydropower

Hydropower dams can alter their discharge several times a day to meet peak electricity demand, resulting in the alteration of downstream flow, including changes in magnitude, duration, timing, rate of change (upramping and downramping rate) and frequency. Fig. 12 shows measured and simulated hydrographs at two gauges in the Maule River, located up- and downstream of the Colbún dam (400 MW), i.e., with natural and altered flow regimes, respectively, in 2013.

Upstream of the dam, the flow regime is natural, and the performance of MissForest was clearly good (R^2 = 0.8, PBIAS = 11.8%, KGE = 0.84). However, downstream of the dam, streamflows change according to energy production to satisfy variable demand with a high stochastic component, and MissForest performance declined (R^2 = 0.32, PBIAS = 33%, KGE = 0.42), evidencing difficulties in the imputation of values in these cases.

3.3. Reconstruction of streamflow records

Complete daily streamflow time series are crucial for water, energy and natural resources management. As overall MissForest presented satisfactory to good performance at gap-filling of daily streamflow time series, existing records at the 122 gauges located in the study area were reconstructed. Fig. 13 shows the observed and reconstructed hydrographs over the 1970–2016 study period at the nearest gauges to the mouth of the ten studied watersheds: a) Maipo, b) Rapel, c) Mataquito,

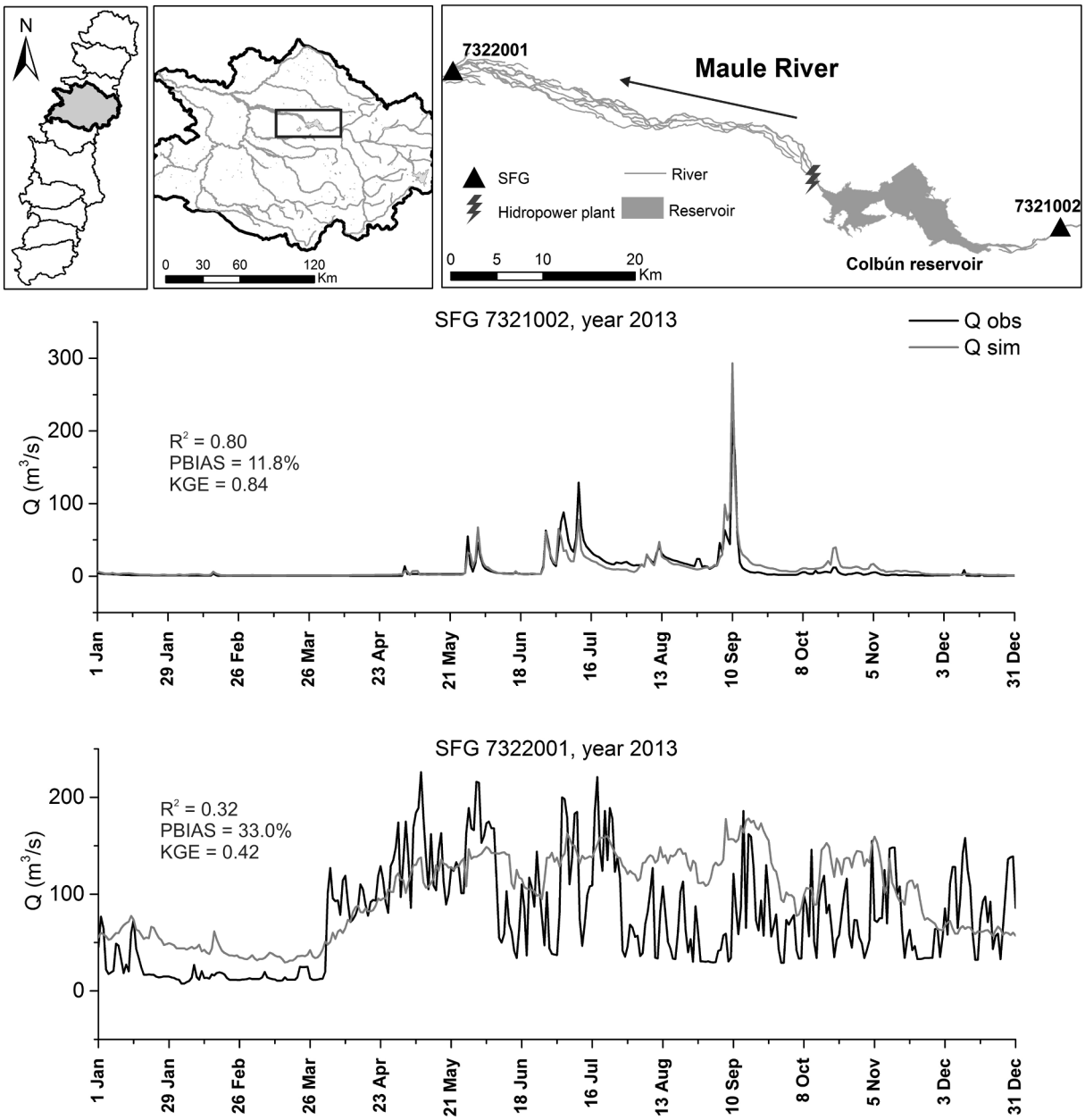

Fig. 12. Observed and simulated hydrographs at two gauges located up- and downstream of the Colbún dam (400 MW), respectively.

d) Maule, e) Itata, f) Biobío, g) Imperial, h) Toltén, i) Valdivia and j) Bueno.

The predicted hydrographs allow the analysis of streamflow change and variability and their interactions with key climatic variables such as precipitation, temperature and potential evapotranspiration in central Chile between 1970 and 2016.

4. Discussion

Researchers often set a threshold for the acceptable percentage of missing data to consider a gauge station usable. For instance, thresholds of 1% (Petrone et al., 2010), 5% (Ukkola et al., 2016), 10% (Déry et al., 2009), 15% (Liu and Zhang, 2017), and 20% (Lopes et al., 2016) have been adopted in previous studies. Further, in the presented study, we adopted a threshold of 50%, which allowed us to work with 122 out of 324 (38%) existing gauge stations. Under such conditions, we showed that the study area is a data-scarce region with poor availability of daily streamflow records, well below the desired standards recommended by the World Meteorological Organization (WMO, 2008). The very low gauge density and data availability presented a challenging scenario for gap-filling methods. As the study area lacks basic information for a proper classification of streamflows into natural or regulated, the aim of this study was to blindly apply MissForest to fill all the gaps using all available gauges with sufficient information, i.e., percentage of missing data less than 50%.

Our missing data randomly occurred in the time-series, and thus they were assumed to be uniformly distributed. However, it could happen that under certain conditions, (e.g., during and after floods because of physical disruption to equipment), missed data present a distribution different to uniform. In such cases, it is expected that predictions of real missing days are on average higher than real non-missing days. We followed the proposed method by Stekhoven & Bülmann (2012); consequently, we didn't apply any transformation to the data sets. However, MissForest can produce severely biased regression estimates and downward biased confidence interval coverages for highly skewed variables in nonlinear models (Hong and Lynn, 2020). Thus, the method

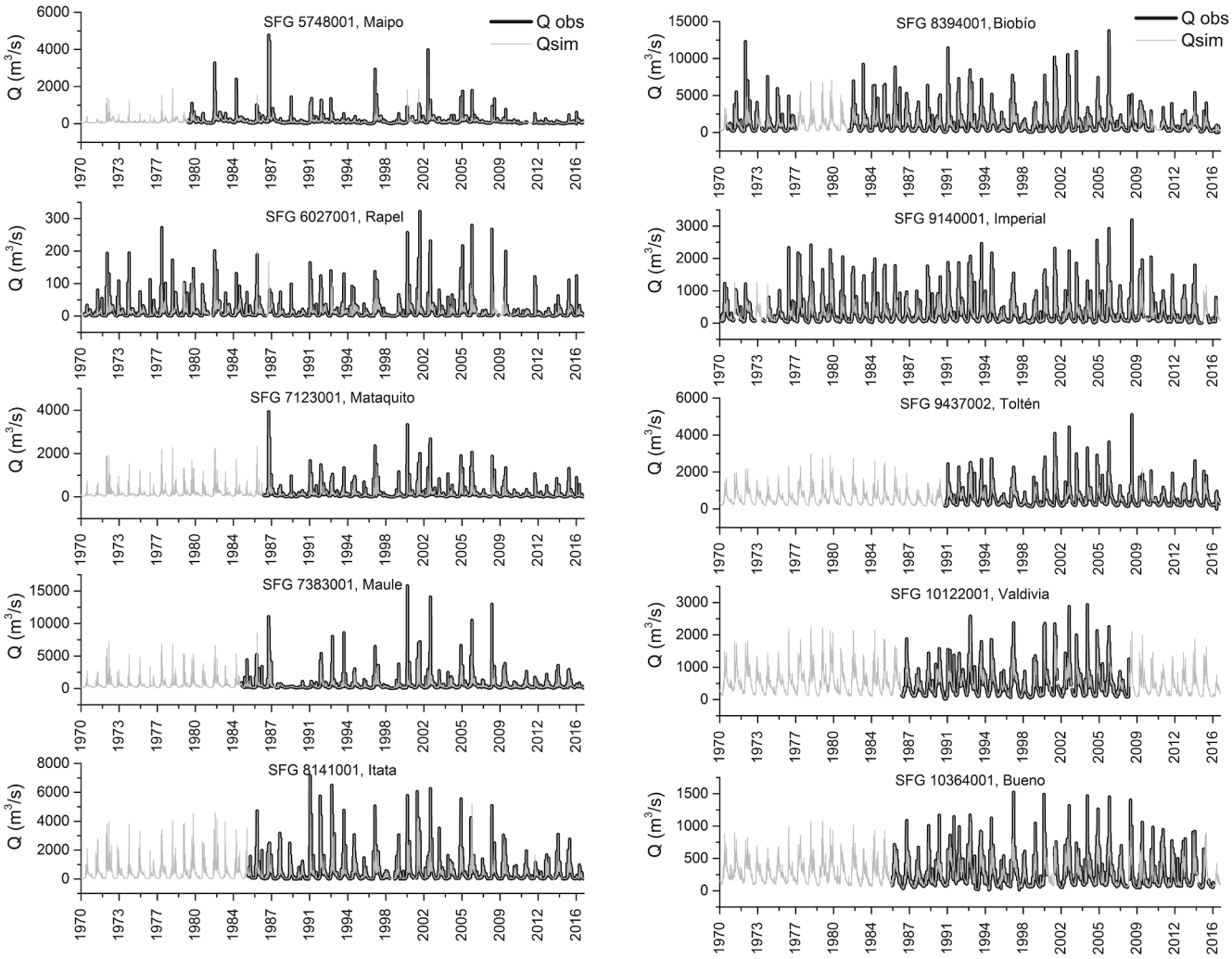

Fig. 13. Observed and reconstructed hydrographs over the 1970–2016 study period at the mouths of the ten studied watersheds: a) Maipo, b) Rapel, c) Mataquito, d) Maule, e) Itata, f) Biobío, g) Imperial, h) Toltén, i) Valdivia and j) Bueno.

has to be applied with some caution. Specifically, data should not be highly skewed, and data gaps should not distribute with bias. In this sense, variation of daily streamflow is not expected to produce highly skewed distributions (see e.g., Blum et al., 2017).

The problem of gaps in data series may be solved theoretically by completing daily flow records from existing data at nearby gauging stations, either upstream or downstream of the same watercourse, although in most existing methods the choice of predictor station may be a critical factor affecting the results (Harvey et al., 2010). Data-driven models such as MissForest are purely empirical and do not consider the complex physical laws in the real world, but as they depend only on the information content in the hydrological data, they are usually easier to develop (Vega-Garcia et al., 2019) and integrate into hydrological information systems, which, combined with a suite of numerical models – physical, statistical or socio-economic – comprise a decision support system for water, energy and natural resources management such as SHEM (Petty and Dhingra, 2018).

According to the goodness-of-fit indicators coefficient of determination (R^2), percent bias (PBIAS) and the Kling-Gupta efficiency (KGE), MissForest achieved satisfactory to good performance, similar to the results of Sidibe et al. (2018) for monthly streamflows in West and Central Africa and Dembélé et al. (2019) for daily streamflows in West Africa using the Direct Sampling method. Remarkably, the performance achieved by MissForest at gap-filling daily streamflows in a data scarce region – central Chile in this case – was comparable to that achieved with alternative methods in data-rich regions such as Mediterranean Europe. For instance, Vega-Garcia et al. (2019) selected 5 out of 240 gauges in the Ebro watershed that presented unimpaired, natural flow regimes with a reliable data range of 30 years of daily weather and flow records and no more than three gaps, achieving R = 0.7–0.8 with an advanced ANN model.

In our results, the presence of altered flows provides a plausible explanation for those gauges where the MissForest method performed worst. However, a generalized analysis of altered versus unaltered flows was not feasible, as the study region lacks sufficient basic information for a proper flow classification. Future work will further analyze whether excluding gauges at streams with altered flow regimes increases predictive performance for gauges at streams with natural flow regime, and whether excluding the "unaltered" gauges increases predictive performance for "altered" gauges. As an indicative example, we selected three illustrative cases to show how altered flows can affect the performance of the method: MissForest performance declined for altered flow regimes such as reduced streamflows due to water diversion for irrigation during the dry season and increased streamflows due to surface runoff inputs from urban areas. In such cases, i.e., when the natural flow regime is changed mostly in terms of magnitude, but maintains other properties like frequency, timing and rate of change, MissForest performance is still satisfactory. Severe alterations to the flow regime such as hydropeaking impeded acceptable performance of MissForest for missing-value imputation. The alteration of the natural flow regimes of

the studied rivers and streams partly explains the high dispersion observed in MissForest performance. In a heavily modified environment, the hydrological effects of human activity can exceed those caused by climate variability (Somorowska and Łaszewski, 2019); consequently, our future work will concentrate on the automatic reconstruction of altered daily streamflow series.

5. Conclusion

MissForest, a non-parametric stochastic machine learning algorithm, was applied to infill gaps in daily streamflow time series and its performance was assessed. A total of 1,586 reconstructions of streamflows for the 1970–2016 period were developed using data records from 122 gauge stations located in different regulated and unregulated rivers and streams in 11 climatic regions throughout central Chile.

Reconstructed daily streamflow time series of rivers with natural flow regimes were simulated with good performance, with similar quality to that attained in reconstruction of monthly streamflow time series or by applying alternative methods in data-rich regions. Reconstruction of altered flows was more challenging for gap-filling methods. In these cases, MissForest performance slightly decreased for discharge magnitude alterations such as those caused by runoff inputs from urbanized areas and water diversion at water intakes for irrigation. In cases of severe flow regime alterations such as hydropeaking, MissForest failed at filling the gaps in daily streamflow series.

Overall, MissForest presented satisfactory to good performance ($R^2 > 0.6$, $PBIAS \pm 15\%$, $KGE > 0.5$), allowing a precise and reliable simulation of the missing data, quickly and automatically, making it suitable for applications in large, data-scarce regions with different climates. MissForest performance increased with the number of predictor records and record length, achieving satisfactory results with 20 or more records having 15 or more years in length.

The reconstructed hydrographs for 1970–2016 allow the analysis of streamflow change and variability and their interactions with key climatic variables such as precipitation, temperature and potential evapotranspiration in central Chile.

CRediT authorship contribution statement

Pedro Arriagada: Conceptualization, Methodology, Software, Investigation, Writing - review & editing. **Bruno Karelovic:** Methodology, Software, Validation. **Oscar Link:** Conceptualization, Validation, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the Universidad de Concepción for providing the institutional support to conduct this research and the financial support of the PREGA research program through project N° 4503152513.

References

Amisigo, B.A., van de Giesen, N.C., 2005. Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series, with examples from the Volta Basin, West Africa. Hydrol. Earth Syst. Sci. Discuss. 2, 449–481. https://doi.org/10.5194/hessd-2-449-2005.

Arriagada, P., Dieppois, B., Sidibe, M., Link, O., 2019. Impacts of Climate Change and Climate Variability on Hydropower Potential in Data-Scarce Regions Subjected to Multi-Decadal Variability. Energies 12, 2747. https://doi.org/10.3390/en12142747.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future köppen-geiger climate classification maps at 1-km resolution. Sci. Data 5, 1–12. https://doi.org/10.1038/sdata.2018.214.

Ben Aissia, M.A., Chebana, F., Ouarda, T.B.M.J., 2017. Multivariate missing data in hydrology – Review and applications. Adv. Water Resour. 110, 299–309. https://doi.org/10.1016/j.advwatres.2017.10.002.

Bonilla, C., Johnson, O., 2012. Soil erodibility mapping and its correlation with soil properties in Central Chile. Geoderma 189–190, 116–123. https://doi.org/10.1016/j.geoderma.2012.05.005.

Blum, A., Archfield, S., Vogel, R., 2017. On the probability of daily streamflow in the United States. Hydrol. Earth Syst. Sci. 21, 3093–3103. https://doi.org/10.5194/hess-21-3093-2017.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Carretier, S., Tolorza, V., Regard, V., Aguilar, G., Bermúdez, M.A., Martinod, J., Guyot, J. L., Hérail, G., Riquelme, R., 2018. Review of erosion dynamics along the major N-S climatic gradient in Chile and perspectives. Geomorphology 300, 45–68. https://doi.org/10.1016/j.geomorph.2017.10.016.

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. J. Hydrol. 569, 573–586. https://doi.org/10.1016/j.jhydrol.2018.11.076.

Déry, S.J., Stahl, K., Moore, R.D., Whitfield, P.H., Menounos, B., Burford, J.E., 2009. Detection of runoff timing changes in pluvial, nival, and glacial rivers of western Canada. Water Resour. Res. 45, 1–11. https://doi.org/10.1029/2008WR006975.

Deshmukh, H., Papageorgiou, M., Kilpatrick, E.S., Atkin, S.L., Sathyapalan, T., 2019. Development of a novel risk prediction and risk stratification score for polycystic ovary syndrome. Clin. Endocrinol. (Oxf) 90, 162–169. https://doi.org/10.1111/cen.13879.

DGA (Direccion General de Aguas in spanish), Water atlas of Chile 2016 Atlas del Agua de Chile Available in http://bibliotecadigital.ciren.cl/handle/123456789/26705 (accessed March 2020).

Ellies, A., 2000. Soil erosion and its control in Chile - An overview. Acta Geol. Hisp. 35, 279–284.

Elshorbagy, A.A., Panu, U.S., Simonovic, S.P., 2000. Group-based estimation of missing hydrological data: I. Approach and general methodology. Hydrol. Sci. J. 45, 849–866. https://doi.org/10.1080/02626660009492388.

Escobar, F., Aceituno, P., 1998. Influencia del fenómeno ENSO sobre la precipitación nival en el sector andino de Chile Central, durante el invierno austral. Bull. Inst. Fr. Etudes Andin. 27, 753–759.

Falvey, M., Garreaud, R.D., 2009. Regional cooling in a warming world: Recent temperature trends in the southeast Pacific and along the west coast of subtropical South America (1979–2006). J. Geophys. Res. Atmos. 114, 1–16. https://doi.org/10.1029/2008JD010519.

Janitza, S., Hornung, R., 2018. On the overestimation of random forest's out-of-bag error. PLoS ONE. 13 (8), e0201904 https://doi.org/10.1371/journal.pone.0201904.

Garreaud, R., 2013. Warm winter storms in central chile. J. Hydrometeorol. 14, 1515–1534. https://doi.org/10.1175/JHM-D-12-0135.1.

Garreaud, R.D., Vuille, M., Compagnucci, R., Marengo, J., 2009. Present-day South American climate. Palaeogeogr. Palaeoclimatol. Palaeoecol. 281, 180–195. https://doi.org/10.1016/j.palaeo.2007.10.032.

Grolemund, G., Wickham, H., 2011. lubridate: Make dealing with dates a little easier. R package version 1 (7), 4. https://CRAN.R-project.org/package=lubridate.

Gyau-Boakye, P., Schultz, G.A., 1994. Filling gaps in runoff time series in west africa. Hydrol. Sci. J. 39, 621–636. https://doi.org/10.1080/02626669409492784.

Habit, E., García, A., Díaz, G., Arriagada, P., Link, O., Parra, O., Thoms, M., 2019. River science and management issues in Chile: Hydropower development and native fish communities. River Res. Appl. 35, 489–499. https://doi.org/10.1002/rra.3374.

Harvey, C., Dixon, H., Hannaford, J., 2010. Developing best practice for infilling daily river flow data, in: BHS Third International Symposium, Managing Consequences of a Changing Global Environment. pp. 1–8. https://doi.org/10.7558/bhs.2010.ic119.

Harvey, C.L., Dixon, H., Hannaford, J., 2012. An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK. Hydrol. Res. 43, 618–636. https://doi.org/10.2166/nh.2012.110.

Hong, S., Lynn, H., 2020. Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction. BMC Medical Research Methodology. 20, 199. https://doi.org/10.1186/s12874-020-01080-1.

INE (Instituto nacional de estadística in spanish), 2018. Synthesis of results of the 2017 census. Síntesis de resultados del censo 2017. Available in https://www.censo2017.cl/descargas/home/sintesis-de-resultados-censo2017.pdf (accessed March 2020).

Kim, M., Baek, S., Ligaray, M., Pyo, J., Park, M., Cho, K.H., 2015. Comparative studies of different imputation methods for recovering streamflow observation. Water (Switzerland) 7, 6847–6860. https://doi.org/10.3390/w7126663.

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424–425, 264–277. https://doi.org/10.1016/j.jhydrol.2012.01.011.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol. Earth Syst. Sci. Discuss. 1–7 https://doi.org/10.5194/hess-2019-327.

Laborde, A., González, A., Sanhueza, C., Arriagada, P., Wilkes, M., Habit, E., Link, O., 2016. Hydropower Development, Riverine Connectivity, and Non-sport Fish Species: criteria for Hydraulic Design of Fishways. River Res. Appl. 32, 1949–1957. https://doi.org/10.1002/rra.3040.

Liu, J., Zhang, Y., 2017. Multi-temporal clustering of continental floods and associated atmospheric circulations. J. Hydrol. 555, 744–759. https://doi.org/10.1016/j.jhydrol.2017.10.072.

Lopes, A., Chiang, J., Thompson, S., Dracup, J., 2016. Trend and uncertainty in spatial-temporal patterns of hydrological droughts in the Amazon basin. Geophys. Res. Lett. 43, 1–8. https://doi.org/10.1002/2016GL067738.

Mackay, S.J., Arthington, A.H., James, C.S., 2014. Classification and comparison of natural and altered flow regimes to support an Australian trial of the Ecological Limits of Hydrologic Alteration framework. Ecohydrology 7, 1485–1507. https://doi.org/10.1002/eco.1473.

Mardones, M., Vargas, J., 2005. Efectos hidrológicos de los usos eléctrico y agrícola en la cuenca del río Laja (Chile centro-sur). Rev. Geogr. Norte Gd. 33, 89–102.

Marino, S., Zhou, N., Zhao, Y., Wang, L., Wu, Q., Dinov, I.D., 2019. HDDA: DataSifter: statistical obfuscation of electronic health records and other sensitive datasets. J. Stat. Comput. Simul. 89, 249–271. https://doi.org/10.1080/00949655.2018.1545228.

McGregor, G.R., 2019. Climate and rivers. River Res. Appl. 1–22 https://doi.org/10.1002/rra.3508.

Montecinos, A., Aceituno, P., 2003. Seasonality of the ENSO-related rainfall variability in central Chile and associated circulation anomalies. J. Clim. 16, 281–296. https://doi.org/10.1175/1520-0442(2003)016<0281:SOTERR>2.0.CO;2.

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and Water Quality Models: Performance Measures and Evaluation Criteria. Trans. ASABE 58, 1763–1785. https://doi.org/10.13031/trans.58.10715.

Muñoz, P., Orellana-Alvear, J., Willems, P., Célleri, R., 2018. Flash-flood forecasting in an andean mountain catchment-development of a step-wise methodology based on the random forest algorithm. Water (Switzerland) 10. https://doi.org/10.3390/w10111519.

Mwale, F.D., Adeloye, A.J., Rustum, R., 2012. Infilling of missing rainfall and streamflow data in the Shire River basin, Malawi - A self organizing map approach. Phys. Chem. Earth 50–52, 34–43. https://doi.org/10.1016/j.pce.2012.09.006.

Petrone, K.C., Hughes, J.D., Van Niel, T.G., Silberstein, R.P., 2010. Streamflow decline in southwestern Australia, 1950–2008. Geophys. Res. Lett. 37, 1–7. https://doi.org/10.1029/2010GL043102.

Petty, T.R., Dhingra, P., 2018. Streamflow Hydrology Estimate Using Machine Learning (SHEM). J. Am. Water Resour. Assoc. 54, 55–68. https://doi.org/10.1111/1752-1688.12555.

Poff, N.L., Allan, J.D., Bain, M.B., Karr, J.R., Prestegaard, K.L., Richter, B.D., Sparks, R.E., Stromberg, J.C., 1997. The natural flow regime. Bioscience 47, 769–784. https://doi.org/10.2307/1313099.

R Core Team, 2019. R: A Language and Environment for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

Sidibe, M., Dieppois, B., Mahé, G., Paturel, J.E., Amoussou, E., Anifowose, B., Lawler, D., 2018. Trend and variability in a new, reconstructed streamflow dataset for West and Central Africa, and climatic interactions, 1950–2005. J. Hydrol. 561, 478–493. https://doi.org/10.1016/j.jhydrol.2018.04.024.

Somorowska, U., Łaszewski, M., 2019. Quantifying streamflow response to climate variability, wastewater inflow, and sprawling urbanization in a heavily modified river basin. Sci. Total Environ. 656, 458–467. https://doi.org/10.1016/j.scitotenv.2018.11.331.

Starrett, S.K., Heier, T., Su, Y., Bandurraga, M., Tuan, D., Starrett, S., 2010. An example of the impact that filled-in peakflow data can have on flood frequency analysis, in: World Environmental and Water Resources Congress 2010: Challenges of Change - Proceedings of the World Environmental and Water Resources Congress 2010. pp. 2451–2455. https://doi.org/10.1061/41114(371)252.

Stekhoven, D.J., Bühlmann, P., 2012. Missforest-Non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118. https://doi.org/10.1093/bioinformatics/btr597.

Stekhoven, D., 2013. missForest: Nonparametric missing value imputation using random forest. R package version 1, 4. https://CRAN.R-project.org/package=missForest.

Tang, F., Ishwaran, H., 2017. Random forest missing data algorithms. Stat. Anal. Data Min. 10, 363–377. https://doi.org/10.1002/sam.11348.

Tao, N., Chen, Y., Wu, Y., Wang, X., Li, L., Zhu, A., 2019. The terpene limonene induced the green mold of citrus fruit through regulation of reactive oxygen species (ROS) homeostasis in Penicillium digitatum spores. Food Chem. 277, 414–422. https://doi.org/10.1016/j.foodchem.2018.10.142.

Tencaliec, P., Favre, A., Prieur, C., Tencaliec, P., Favre, A., Prieur, C., Mathevet, T., 2015. Reconstruction of missing daily streamflow data using dynamic regression models. Water Resour. Res. 51, 9447–9463. https://doi.org/10.1002/2015WR017399.

Troyanskaya, O., Cantor, M., Sherlock, G., et al., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics. 17 (6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.

Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history inwater resources. Water (Switzerland) 11. https://doi.org/10.3390/w11050910.

Ukkola, A.M., Keenan, T.F., Kelley, D.I., Prentice, I.C., 2016. Vegetation plays an important role in mediating future water resources. Environ. Res. Lett. 11 https://doi.org/10.1088/1748-9326/11/9/094022.

Urrutia, R.B., Lara, A., Villalba, R., Christie, D.A., Le Quesne, C., Cuq, A., 2011. Multicentury tree ring reconstruction of annual streamflow for the Maule River watershed in south central Chile. Water Resour. Res. 47, 1–15. https://doi.org/10.1029/2010WR009562.

Valdés-Pineda, R., Cañón, J., Valdés, J.B., 2018. Multi-decadal 40- to 60-year cycles of precipitation variability in Chile (South America) and their relationship to the AMO and PDO signals. J. Hydrol. 556, 1153–1170. https://doi.org/10.1016/j.jhydrol.2017.01.031.

Valdés-Pineda, R., Pizarro, R., García-Chevesich, P., Valdés, J.B., Olivares, C., Vera, M., Balocchi, F., Pérez, F., Vallejos, C., Fuentes, R., Abarza, A., Helwig, B., 2014. Water governance in Chile: Availability, management and climate change. J. Hydrol. 519, 2538–2567. https://doi.org/10.1016/j.jhydrol.2014.04.016.

Van Buuren, S., Oudshoorn, K., 1999. Flexible Multivariate Imputation by MICE. TNO Prevention Center, Leiden, The Netherlands.

Van Buuren, S., 2007. Multiple imputation of discrete and continuous data by fully conditional specification. Stat. Methods Med. Res. 16, 219–242. https://doi.org/10.1177/0962280206074463.

Vega-Garcia, C., Decuyper, M., Alcázar, J., 2019. Applying Cascade-Correlation Neural Networks to In-Fill Gaps in Mediterranean Daily Flow Data Series. Water 11, 1691. https://doi.org/10.3390/w11081691.

Vörösmarty, C.J., McIntyre, P.B., Gessner, M.O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S.E., Sullivan, C.A., Liermann, C.R., Davies, P.M., 2010. Global threats to human water security and river biodiversity. Nature 467, 555–561. https://doi.org/10.1038/nature09440.

Waljee, A.K., Mukherjee, A., Singal, A.G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., Higgins, P.D.R., 2013. Comparison of imputation methods for missing laboratory data in medicine. BMJ Open 3, 1–8. https://doi.org/10.1136/bmjopen-2013-002847.

WMO, 2008. Guide to Hydrological Practices. Volume I: Hydrology-From Measurement to Hydrological Information. Hydrological Sciences Journal 56 (1), 196–197. https://doi.org/10.1080/02626667.2011.546602.

XU, Y. hyfo: Hydrology and Climate Forecasting R package version 1 4 2018 https://CRAN.R-project.org/package=hyfo.

Zambrano, M., 2017. hydroGOF: Goodness-of-Fit Functions fo comparison of simulated and observed hydrological time series. R package version 0.3-10. https://CRAN.R-project.org/package=hydroGOF.

Zhang, Y., Post, D., 2018. How good are hydrological models for gap-filling streamflow data? Hydrol. Earth Syst. Sci. 22, 4593–4604. https://doi.org/10.5194/hess-22-4593-2018.