Reconstruction of missing daily streamflow data using dynamic regression models

Patricia Tencaliec, Anne-Catherine Favre, Clémentine Prieur, Thibault Mathevet

AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2015WR017399

Reconstruction of missing daily streamflow data using dynamic regression models

Patricia Tencaliec[1,2,3], Anne-Catherine Favre[4,5,6], Clémentine Prieur[1,2,3], and Thibault Mathevet[7]

[1]Grenoble Alpes University, LJK, Grenoble, France, [2]CNRS, LJK, Grenoble, France, [3]Inria Project/Team AIRSEA, France, [4]Grenoble Alpes University, LTHE, Grenoble, France, [5]CNRS, LTHE, Grenoble, France, [6]IRD, LTHE, Grenoble, France, [7]Electricité de France, Division Technique Générale, Grenoble, France

Key Points:
- Dynamic regression models are used to reconstruct missing streamflow data
- Dynamic regression models solve the problem of correlated residuals
- Daily discharge data are reconstructed using only streamflow information

Correspondence to:
P. Tencaliec,
patricia.tencaliec@imag.fr

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

Received 17 APR 2015
Accepted 9 NOV 2015
Accepted article online 12 NOV 2015
Published online 10 DEC 2015

Abstract River discharge is one of the most important quantities in hydrology. It provides fundamental records for water resources management and climate change monitoring. Even very short data-gaps in this information can cause extremely different analysis outputs. Therefore, reconstructing missing data of incomplete data sets is an important step regarding the performance of the environmental models, engineering, and research applications, thus it presents a great challenge. The objective of this paper is to introduce an effective technique for reconstructing missing daily discharge data when one has access to only daily streamflow data. The proposed procedure uses a combination of regression and autoregressive integrated moving average models (ARIMA) called dynamic regression model. This model uses the linear relationship between neighbor and correlated stations and then adjusts the residual term by fitting an ARIMA structure. Application of the model to eight daily streamflow data for the Durance river watershed showed that the model yields reliable estimates for the missing data in the time series. Simulation studies were also conducted to evaluate the performance of the procedure.

1. Introduction

In hydrology, the main variables that describe the hydrological functioning of watersheds are air temperature, precipitation, soil moisture, and streamflow. Numerous research and operational applications, such as water resources management, extreme flood or drought predetermination, streamflow forecast, and climate variability analysis, require reliable time series. Since extreme events are seldom by definition, long and continuous time series are necessary, allowing a more accurate analysis of watershed operation.

Due to technical or maintenance issues, long hydrometric data production and management are a difficult task and, eventually, gaps in the data set arise, e.g., measurement stations can be damaged during flood events. These missing intervals in the time series represent a loss of information and can cause erroneous summary data interpretation or unreliable scientific analysis.

Consequently, in order to obtain reliable and accurate information from the data, these gaps must be filled. The estimation of missing intervals, also known in the literature as imputation [*Schneider*, 2001], infilling [*Harvey et al*., 2012], or reconstruction [*Kim and Pachepsky*, 2010], represents a great challenge in hydrology and geosciences in general.

The reconstruction of missing streamflow data is a problem studied from decades ago and, even nowadays, it continues to be a challenge. There are several methods reported in the literature. Among these, we remind the works of *Hirsch* [1979] and *Wallis et al*. [1991] that discuss infilling approaches for daily data using data from the nearby station(s), i.e., the missing values of a target station are replaced with the weighted values from a neighbor station. The weights are then computed as the ratio between the drainage area of the target and neighbor station, or the ratio between the monthly mean flow data of the two stations. Other references, like the works of *Raman et al*. [1995] and *Woodhouse et al*. [2006], recommend the use of regression analysis for reconstructing the missing data. More recent studies present procedures for filling missing hydrological data by using state-space models with Estimation-Maximization (EM) algorithm as in *Amisigo and van de Giesen* [2005] or approaches that involve artificial neural networks as presented in *Khalil et al*. [2001]; *Elshorbagy et al*. [2002]; or *Coulibaly and Baldwin* [2005].

Reviews studies by *Gyau-Boakye and Schultz* [1994] and *Harvey et al*. [2012] summarize and compare several methods used for infilling flow data. *Gyau-Boakye and Schultz* [1994] compare 10 widely known techniques including interpolation, recursive models, autoregressive models, regression, and nonlinear models. Their results show that the model choice is influenced by the length of the estimation period or by the season, but on average, interpolation and multiple regression models yield good results. In *Harvey et al*. [2012], there is an extended description of approaches used in hydrology for missing data imputation or prediction, along with an applied comparison of simple and multiple regression models. The authors proved that one can have a better accuracy if multiple input variables are included.

In this study, the dynamic regression models (DRMs) [see *Pankratz*, 1991 or *Box and Jenkins*, 1976, for details] were applied to estimate the missing flow data. The DRM estimates an output variable based on one or multiple input variables and also adjusts the correlation from the remainder part (residuals) by fitting an autoregressive integrated moving average (ARIMA) structure.

The DRMs have been used before by *Tsay* [1984] to model the monthly highway traffic volume in Taiwan, by *Greenhouse et al*. [1987] to fit biological rhythm data, by *Miaou* [1990] to estimate the water demand in some states from USA, or, more recently, by *Bercu and Proïa* [2013] to forecast energy consumption.

We have seen earlier that previous works addressed the infilling of flow data by using the multiple linear regression [*Gyau-Boakye and Schultz*, 1994; *Woodhouse et al*., 2006; *Harvey et al*., 2012] or even the simple linear regression with residual modeling [*Raman et al*., 1995], but none approached the problem as a multiple linear regression with residual modeling. While the models found in the literature address only one aspect of the prior problem formulation, i.e., either the inclusion of multiple inputs or the modeling of the residuals from a regression with one input, the use of DRM solves both points. Therefore, the novelty of the present study is that it allows to handle multiple inputs, and additionally, to model the residuals with an ARIMA process for streamflow data infilling. Also, the output from one period can be associated with inputs not only from the same period, but also from a past time. Therefore, with DRM we can use lagged input variables.

The main objective of this research is to reconstruct streamflow data containing missing intervals of various lengths using the DRM approach. We want to address a particular case when one has available for the analysis only streamflow data. This consideration is founded as, frequently, we do not have access to long-historical data (i.e., first decades of the 20th century or earlier) for other variables, like i.e., precipitation. Therefore, we want to present an approach that takes as input variables just the streamflow data from several correlated hydrometric stations.

The paper is organized as follows. Section 2 provides a short presentation of the data and an exploratory analysis to better understand the watershed behavior. In the third section, we address the theoretical background of the technique, the methodology, and the approach used for validating the model. Section 4 is devoted to the case study on the Durance watershed, with a discussion on the performance of the estimated models. The accuracy of the models is also measured on simulations. We end by giving the conclusions of our study in section 5.

2. Data Presentation and Exploratory Analysis

2.1. Data Presentation

The data used in this study come from the Durance watershed. Situated in the South-East region of France, the Durance river is the second largest tributary of Rhone, after Saône, with a length of more than 300 km and a catchment area of more than 14,000 km^2. It is defined by its many uses and the natural basin, becoming one of the most important rivers in Southern France. The entire watershed offers many purposes like hydropower generation, irrigations, water supply for cities like Marseille and Aix-en-Provence or tourism near the lakes. Furthermore, due to its mixed climatological environment (from a nival regime in the North-East area to a mediterranean-pluvial in the South area), along with the geographical and functional complexity, the analysis of the Durance river is challenging.

The Durance watershed is divided into three geographical areas: upper, middle, and lower basin. The upper Durance is characterized by a mountainous area with abrupt valleys, the middle part has a lower altitude and the valleys are wider, while the lower Durance is the smallest and composed mainly of dry lowland, but

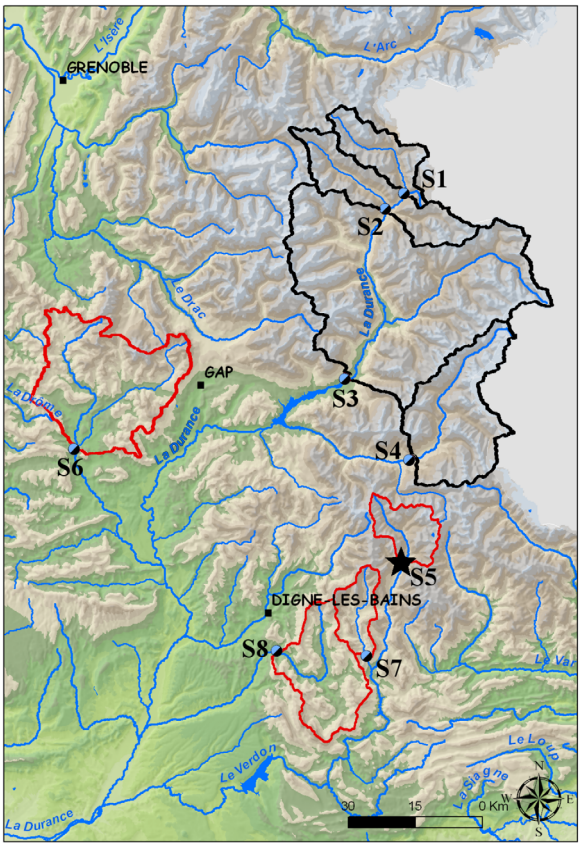

Figure 1. Location and drainage area of the selected stations from the Durance watershed (black and red contours suggest the clusters of PAM classification from section 2.2).

it still remains in a hilly area. There are more than 50 hydrometric stations within the watershed, but for this study we selected eight stations situated in the upper and middle regions of the Durance that have a data sample longer than 100 years, which is seldom in hydrology. The location of the stations and their main characteristics are presented in Figure 1 and Table 1.

The observations for the flow data are provided by Electricité de France (EDF) or the HYDRO database system (http://www.hydro.eaufrance.fr/). We used in this study the daily flow measurements starting from 1904 until 2010, thus 107 years.

The measurement installations are situated on the rivers and most of them provide natural flow data. These stations were installed at the beginning of the 20th century in order to help the French administration issue flood alerts [*Imbeaux*, 1892] and improve the understanding of the hydroelectric potential of the Durance watershed. An extensive part of these streamflow time series (i.e., the early decades) had to be restored from different archives through a documentary research, see *Kuentz* [2013]; *Kuentz et al.* [2013, 2014] for details. These studies provide an extended characterization of the hydrometeorological variability of the Durance watershed during the last century, and also give an historical review about the measurement procedures at each station. In the early period (1904–1909), the river stage measurements were made by daily human observation, then from 1910 to 1950 by using a limnograph (device for automatically recording the water level) and last, since 1980, by using an electronic data logger.

This difference in measurements can create homogeneity issues and, thus, have an impact on the analysis of the streamflow data. To address this aspect, we followed the two-step approach introduced by *Wijngaard et al.* [2003]. Same workflow for detecting inhomogeneity was applied later by *Kang and Yusof* [2012] for a hydrometeorological data set with missing values. The approach consists of (1) applying four homogeneity tests: standard normal homogeneity test, Buishand range test, Pettitt test, and von Neumann ratio test to evaluate the series and, then, (2) classifying these tests results into three classes: useful (homogeneous data), doubtful, and suspect (inhomogeneous data). The details of each test and the steps of the approach

Table 1. Main Characteristics of the Selected Stations

Code	Station Name	In Service from	Location	Altitude (m)	Area (km^2)	# Missing Data	% Missing Data
S1	Durance (Val-des-Près)	1917	Upper	1360	203	9217	24%
S2	Durance (Briançon)	1905	Upper	1187	548	4900	13%
S3	Durance (La Clapière)	1903	Upper	787	2170	5903	15%
S4	Ubaye (Barcelonnette)	1903	Upper	1132	549	1207	3%
S5	Verdon (Colmars)	1903	Middle	1230	158	3340	9%
S6	Buech (Chambons)	1905	Middle	662	723	5473	14%
S7	Issole (Saint-André-les-Alpes)	1904	Middle	931	137	9711	25%
S8	Asse (Clue de Chabrières)	1906	Middle	605	375	7067	18%

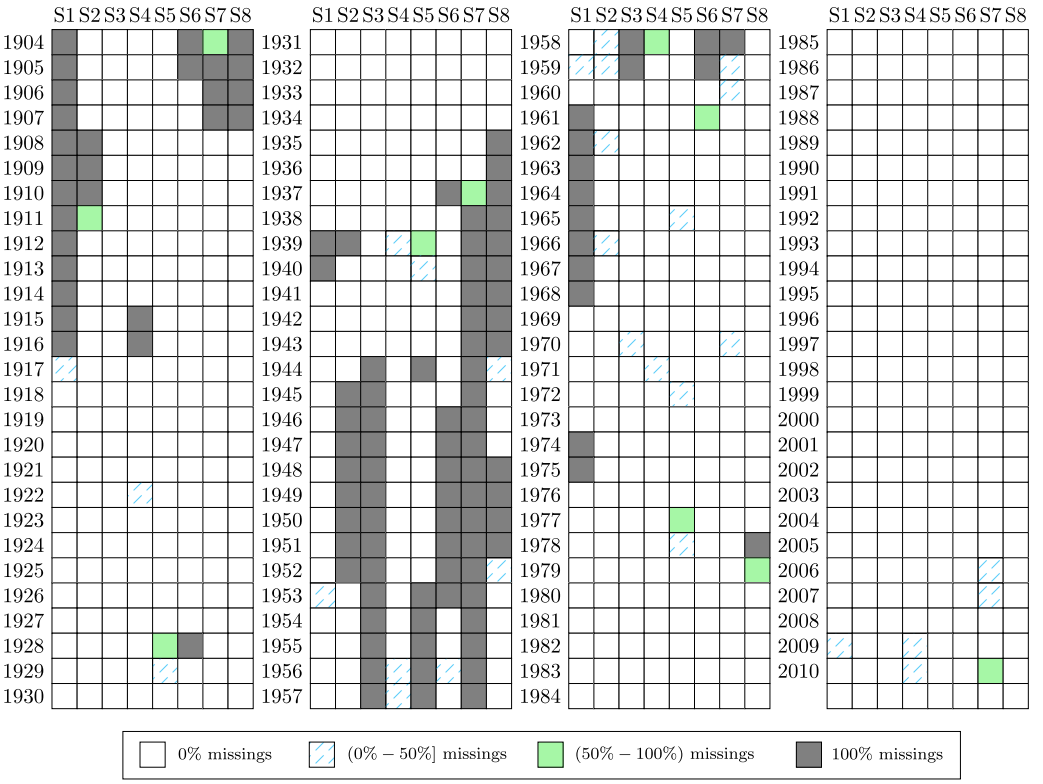

Figure 2. Missing data pattern from 1904 to 2010 for the Durance watershed.

can be found in the two references mentioned above. The testing variable used is the annual maxima, as suggested by *Kang and Yusof* [2012] in their study of daily rainfall data.

The results show that all eight stations from Durance watershed are classified as "useful," at a 5% significance level, thus we can say that the data are homogeneous.

The missing data for the Durance watershed are mainly due to absence of human reading (early period) and technical/maintenance issues, or disturbances during the 2nd World War [*Kuentz et al.*, 2014]. Consequently, these data contain a large number of missing points, especially at the beginning of the period and around 1940–1960. The percentage of missing data for the eight stations ranges from 3% to 25%. In Figure 2, one can find the pattern of the missing data for each station for the entire period 1904–2010.

2.2. Exploratory Data Analysis

To determine the spatial and temporal relationships and correlations between the eight stations, an exploratory data analysis was used to determine possible similarities among variables (stations) and, eventually, to group them based on their properties. This part is important as it offers an initial selection for the input variables in the regression models.

First, we look at the monthly mean flow (hydrological regime) and distinguish the behavior of the station. An illustration of the hydrological regimes for the eight stations is shown in Figure 3. It can be seen that each station has two periods of high flow. For the stations from upper Durance (S1–S4), there is a peak on the first half of autumn (due to strong rainfall events) and another one, much higher, at the beginning of the summer (due to snowmelt). Meanwhile, the stations from middle Durance (S5–S8) have one peak at the end of autumn (due to rainfall) and the second one around middle/end spring (due to an early snowmelt). The results are consistent with the climate of that area and the elevation ranges of those watersheds. The stations from upper Durance are located in a rocky mountain area, where, besides the rainfalls in autumn, most of the precipitations fall as snow. For this reason, we will have a glacial-snow regime characterized by very high flow at the beginning of the summer due to snow and glaciers melt and dry winter (low flow). As

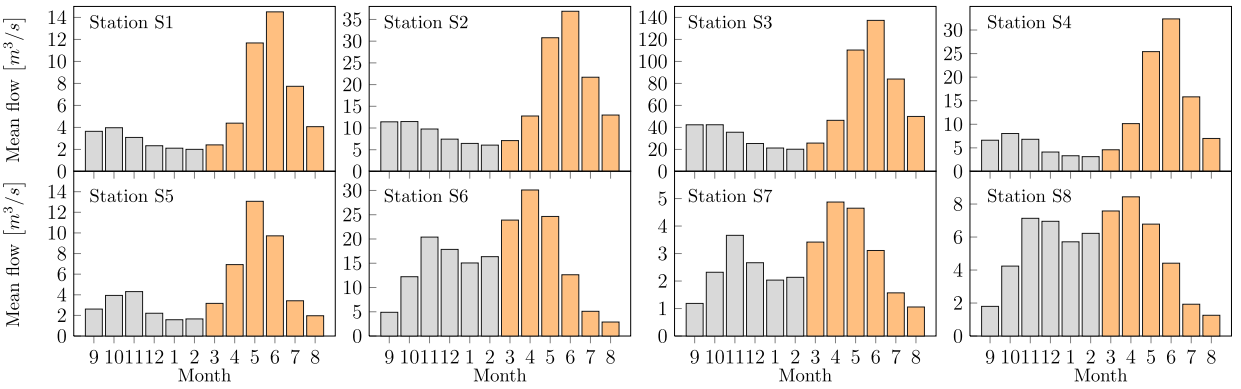

Figure 3. Hydrological regimes (monthly mean flow) for the eight stations of the Durance watershed (cold season in gray, warm season in orange). Note: the months are ordered from September to August for a clearer illustration of the two seasons.

we descend in altitude (middle Durance), we will observe that the autumn rain is increasing and lasting until the first part of the winter; then, the snowmelt process is starting earlier, like in May for S5 and in April for S6, S7, S8. This specific behavior is called rain-snow regime. Therefore, each regime displays two seasons: autumn-winter (cold season), defined by rain (less in upper Durance, more in middle Durance), and spring-summer (warm season), defined by snowmelt (earlier or later depending on the altitude).

The above statements were also validated by statistical analysis of correlation and cluster analysis. The correlation matrix of the daily flow data is computed using only the complete cases of the data set (only observations that have information for all the stations, i.e., 49.7% of the data). The chosen criterion is Spearman's rank correlation coefficient. It is a nonparametric rank statistic, similar to Pearson's correlation coefficient, and it measures the dependence between two variables as a monotonic function; for more details, the reader is referred to *Lehmann and D'Abrera* [2006].

The results, described in Table 2, show that all the coefficients are positive with strong correlation ($>$0.8) between the group of stations S1–S4 and the group S6–S8. Station S5 is a particular case; it has a higher value in relation with S4 and S7, but all its values are very close to each other. Assessment of the correlation for each of the two seasons (cold, warm) defined above, show that for the cold season there is a decrease in dependence for both upper and middle Durance and S5 tends to go more with the middle Durance stations, while for the warm season the groups upper and middle Durance are better split, but station S5 still remains an "in-between" station.

Next, we consider a clustering technique, called partitioning around medoids (PAM), to classify the stations based on their spatial/temporal characteristics. The idea of this approach is to divide the data set into groups so that the distance between them is minimized. It is very similar to the well-known *k*-means technique, but, in the case of PAM, each center (called medoid) is the point itself, so a member of the group, not a mean value like in the *k*-means case. The detailed procedure of the technique can be found in

Table 2. Daily Flow Correlation Matrix for All Data (on the Left Table), Cold Season (Upper Right Table, in Bold), and Warm Season (Lower Right Table, in Italic)

	S1	S2	S3	S4	S5	S6	S7	S8
S1	1.00	0.92	0.93	0.87	0.66	0.08	0.28	0.05
	S2	1.00	0.93	0.87	0.66	0.08	0.29	0.04
		S3	1.00	0.90	0.68	0.09	0.31	0.04
			S4	1.00	0.76	0.22	0.46	0.19
				S5	1.00	0.57	0.73	0.53
					S6	1.00	0.82	0.85
						S7	1.00	0.85
							S8	1.00

	S1	S2	S3	S4	S5	S6	S7	S8
		0.79	**0.83**	**0.72**	**0.51**	**0.11**	**0.21**	**0.10**
S1			**0.83**	**0.74**	**0.56**	**0.15**	**0.28**	**0.13**
S2	*0.96*			**0.82**	**0.61**	**0.15**	**0.30**	**0.13**
S3	*0.96*	*0.96*			**0.67**	**0.26**	**0.44**	**0.27**
S4	*0.91*	*0.91*	*0.92*			**0.57**	**0.70**	**0.57**
S5	*0.63*	*0.63*	*0.63*	*0.74*			**0.80**	**0.84**
S6	*0.04*	*0.04*	*0.03*	*0.18*	*0.62*			**0.86**
S7	*0.21*	*0.21*	*0.20*	*0.36*	*0.72*	*0.86*		
S8	*0.02*	*0.02*	*0.00*	*0.15*	*0.56*	*0.85*	*0.87*	

Kaufman and Rousseeuw [1990]. To choose the relevant number of clusters and to determine if a station is well classified, we will use the silhouette coefficient, introduced by *Rousseeuw* [1987] and defined below:

$$s_i = 1 - \frac{d_{i,c(i)}}{\delta_{i,-c(i)}} \tag{1}$$

where $d_{i,c(i)}$ represents the intracluster distance between medoid $c(i)$ and station i, and $\delta_{i,-c(i)}$ corresponds to the smallest distance between station i and all the other medoids except $c(i)$. To compute PAM performance for the k clusters, the average silhouette index is used: $s(k) = \frac{\sum_{i=1}^{n} s_i}{n}$, where n is the number of stations.

We applied PAM classification on our daily flow data (S1–S8) by using two and three clusters. When using two clusters the data are classified as Group 1 = {S1, S2, S3, S4} and Group 2 = {S5, S6, S7, S8}, having the medoids S3 and S7. An illustration can be seen in Figure 1 (Group 1 with black contour and Group 2 with red). This division is exactly the geographical split upper-middle Durance. When looking at the silhouette coefficients for each station, we notice that S5 has a negative, but close to zero value, meaning that it may be not well classified in Group 2. In the case with three clusters, the stations are classified as follows: Group 1 = {S1, S2, S3, S4}, Group 2 = {S5}, and Group 3 = {S6, S7, S8}, with the medoids set {S3, S5,S8}. These results are supported also by the hydrological regimes and the correlation matrix, S5 being an "in-between" station with a special behavior. The application of PAM on the cold and warm season subset yields the same output, but with higher or lower value for the silhouette coefficient.

In conclusion, when trying to classify the eight stations, it is clear that the hazy behavior of station S5 makes the grouping a little bit uncertain, while the remaining stations preserve the geographical division of upper and middle Durance.

These relationships will be used later in the choice of explanatory variables in our regression models (see section 4.1).

3. Statistical Modeling: Theory and Methodology

In this section, we describe the models proposed for infilling flow data. Considering their simplicity, the suggested models make a good balance between the quality of the estimates and the model's complexity.

We start by briefly introducing the theory behind the dynamic regression model and its particular cases, and then we address the methodology used for the estimation and validation of the model. For a more detailed outline of the theoretical background, the readers are referred to *Box and Jenkins* [1976], *Pankratz* [1991], or *Makridakis et al.* [1998].

3.1. Theory: Dynamic Regression Models

Dynamic regression models [*Pankratz*, 1991], also called transfer function models by *Box and Jenkins* [1976], are a class of statistical models that describe the relationship between a response variable and one or more explanatory variables using a dynamic form. In time series analysis, the decision and final impact of a change in a variable may have some delays, so it is important to examine these relationships not only at the present time, but also at some other lags.

In order to avoid any confusion for the remaining part of this paper, we will use the formulation "residual term/residuals" for the difference between observations and the estimates of the regression part of the model, and "error term/errors" for the white noise process in the (S)ARIMA model.

A dynamic regression model states how a response variable (Y_t) is related to present and past values of one or more explanatory variables ($X_{t,1}, \ldots, X_{t,l}$). Besides this, it allows for the residual term of the regression to be modeled with a seasonal autoregressive integrated moving average (SARIMA) model.

A SARIMA model is an extension of the well-known ARIMA model [see *Box and Jenkins*, 1976, for details] that addresses seasonality. Therefore, apart from the relationships between observations of successive periods, SARIMA incorporates the relationships between observations at certain period distance, for example a week, a quarter, etc. (seasonal part). A short notation for this model is SARIMA$(p,d,q)(P,D,Q)_s$, where p = nonseasonal autoregressive (AR) order, d = nonseasonal differencing, q = nonseasonal moving average

(MA) order, P = seasonal autoregressive (SAR) order, D = seasonal differencing, Q = seasonal moving average (SMA) order, and s = number of time units per season.

The general dynamic regression model formulation, in terms of the backshift operator B (defined as $B^i Y_t = Y_{t-i}$), with I explanatory variables and a SARIMA$(p,d,q)(P,D,Q)_s$ model for the residuals, is

$$Y_t = \beta_0 + \alpha_1(B)X_{t,1} + \ldots + \alpha_I(B)X_{t,I} + Z_t \tag{2}$$

$$\phi(B)\phi_s(B^s)\nabla^d\nabla_s^D Z_t = \theta(B)\theta_s(B^s)e_t \tag{3}$$

In the formulation (2) of the regression terms, we have β_0 as a constant and the polynomials $\alpha_i(B) = \frac{\omega_i(B)}{\delta_i(B)}B^{b_i}$, with $\omega_i(B) = \omega_{i,0} - \sum_{j=1}^{m_i}\omega_{i,j}B^j$ and $\delta_i(B) = 1 - \sum_{j=1}^{r_i}\delta_{i,j}B^j$. The group (m_i, r_i) represents the orders of the two polynomials $\omega_i(B)$ and $\delta_i(B)$, and b_i is a delay factor. The polynomials $\alpha_i(B)$ (called so far *transfer functions*) represent how Y_t reacts over a time period to a change in $X_{t,i}$. For more details, refer to section 3.2.

In the formulation (3) of the residual term Z_t, e_t represents the tth observation of a white noise process (error term). The operators ∇^d and ∇_s^D are used in case of nonstationary series, and they represent the differencing of order d for the nonseasonal part, respectively, the differencing of order D for the seasonal part with s time units per season; they actually stand for $\nabla^d = (1-B)^d$ and $\nabla_s^D = (1-B^s)^D$, respectively. The new resulted time series is called "integrated" series.

Furthermore, on the equation of the residual term in (3), we have the polynomials of the SARIMA model for the nonseasonal part ($\phi(B)$, $\theta(B)$) and seasonal part ($\phi_s(B^s)$, $\theta_s(B^s)$), as follows:

$\phi(B) = 1 - \sum_{i=1}^{p}\phi_i B^i$ (linear combination of the past p values of the residual term Z_t), where the parameters ϕ_i are called the AR terms (usually referred as AR(p));

$\theta(B) = 1 - \sum_{i=1}^{q}\theta_i B^i$ (linear combination of the past q values of the errors e_t), where the parameters θ_i are called the MA terms (usually referred as MA(q)).

Similarly, we can define the seasonal components of the SARIMA model, namely SAR(P) and SMA(Q), by using the polynomials $\phi_s(B^s)$ and $\theta_s(B^s)$ this time with a s time units lag.

3.2. Methodology: Model Estimation

We will use for computing the model parameters of daily flow data the methodology proposed by *Pankratz* [1991], because it is suited for multiple explanatory variables, compared to the one proposed by *Box and Jenkins* [1976] that is not. The entire procedure is illustrated schematically in Figure 4. The basic idea behind DRM is that it involves a two-part setup: the regression and the residuals modeled with (S)ARIMA. The first step (initialization) is to choose a proxy model for both parts. As reported in *Pankratz* [1991], it is recommended to start with a large enough number of lags for each explanatory variable in the regression and, additionally, to consider a low-order model for the residuals (AR(1)/AR(2)). The estimates of the parameters and the errors of the initial-proxy model are then analyzed and, if necessary, a new model is identified and estimated again. At the end of the procedure, the errors of the selected model must be a white noise process.

The procedure for identifying the new model (step D in Figure 4) requires to find first the order of both the linear transfer functions and (S)ARIMA. For the (S)ARIMA models, the order identification is done by analyzing the sample autocorrelation and partial autocorrelation coefficients, a well-known approach of *Box and Jenkins* [1976], and it will not be theoretically detailed here (see case-study results with details in section 4.1-ARIMA models).

The transfer function order identification ((b_i, m_i, r_i) of the α_i polynomials in (2)) is done by examining the pattern of the coefficients for each explanatory variable. There are some identification rules, with reference to the theoretical functions, reported in *Pankratz* [1991], as follows:

1. b_i (*dead time*, time elapsed until the explanatory variable affects the response variable) represents the number of lags that are zero on the first position(s).
2. The denominator factor $\delta_i(B)$ represents the *decay pattern* and the order r_i of this polynomial is given by:
   2.1. $r_i = 0$—no decay in the pattern of the coefficients
   2.2. $r_i = 1$—exponential decay pattern of the coefficients
   2.3. $r_i = 2$—complex decay pattern of the coefficients ($r_i > 2$ is very rare)

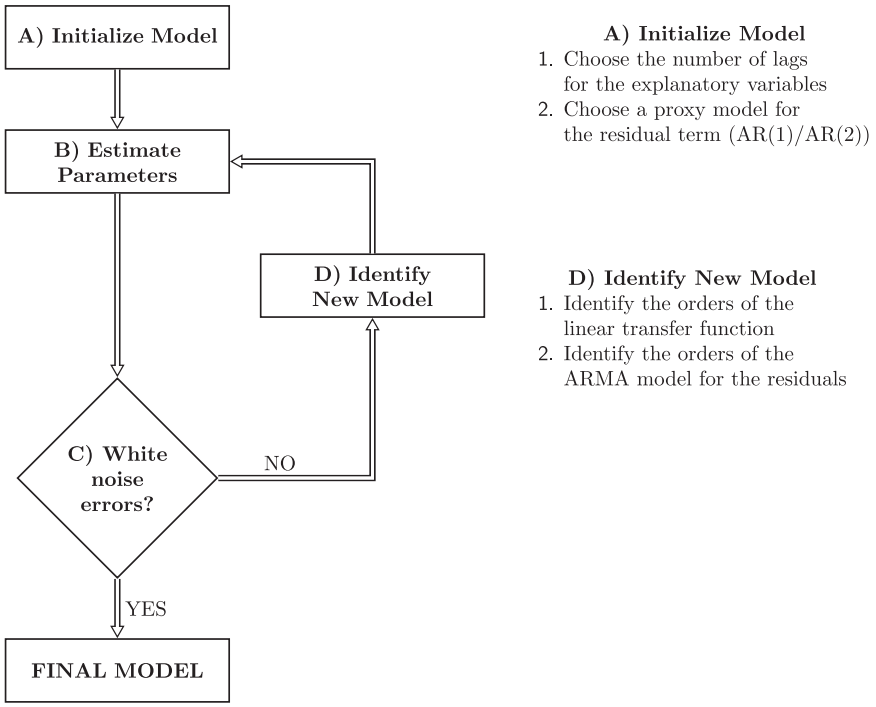

Figure 4. Schematic representation of the estimation methodology.

3. The numerator factor $\omega_i(B)$ captures the *unpatterned spikes* (not part of the decay pattern) in the coefficients' representation and the *decay start-up* value(s). The order of this polynomial is $m_i = u_i + r_i - 1$, where u_i represents the unpatterned coefficients.

   3.1. if $r_i > 0$, then u_i is the number of nonzero parameters before the decay starts
   3.2. if $r_i = 0$, then all the nonzero parameters are considered unpatterned.

The estimation of the parameters could be done by using the ordinary least squares technique, if the moving average part of the ARIMA model is not introduced. In case the MA component is required, the problem becomes impossible to solve as the values for the past errors are unobservable. Consequently, maximum likelihood estimation (MLE) could be used in this case for the parameter estimation.

The stationarity of the variables was checked using two procedures, the Augmented Dickey-Fuller (ADF) unit root test introduced by *Said and Dickey* [1984] and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) stationarity test proposed by *Kwiatkowski et al.* [1992], both of them being tested at a lag order $p = int\left[12\left(\frac{n}{100}\right)^{1/4}\right]$, as suggested by *Schwert* [1989], where n is the sample size. Simple methods exist for transforming a nonstationary series into a stationary one. For instance, series not constant in mean can be differentiated. On the other hand, log-transformation can be used for series not constant in variance. However, other more complex nonstationarity scenarios can be encountered where these approaches are not suitable.

3.3. Methodology: Model Validation

Once the model is defined using the procedure above, one should test its performance and validate it by using a test data set, different from the one used in the estimation. The performance of the model is computed by comparing the data from the test set with the values estimated by the model. Then, the model is validated by comparing it with several other models (i.e., simpler models, other category models, benchmark models, etc.).

In order to measure the efficiency, we use the Kling-Gupta Efficiency (KGE). This criterion was introduced by *Gupta et al.* [2009] and it represents a decomposition of the Nash-Sutcliffe Efficiency (NSE), introduced by *Nash and Sutcliffe* [1970], in terms of three components: correlation, bias, and variability. The general formulation for the KGE is

Table 3. Explanatory Variables Included in the Modeling: Before (Full Model) and After (Reduced Model) Removing Colinearity, for All, Cold, and Warm Season Data

		S1	S2	S3	S4	S5	S6	S7	S8
All data and Warm season	**Full model**	S2,S3,S4	S1,S3,S4	S1,S2,S4	S1,S2,S3,S5	S4,S7	S7,S8	S5,S6,S8	S6,S7
	Reduced model	S3	S1,S4	S1,S4	S1,S5	S4,S7	S7,S8	S5,S6,S8	S6,S7
Cold season	**Full model**	S2,S3,S4	S1,S3,S4	S1,S2,S4	S1,S2,S3	S7	S7,S8	S5,S6,S8	S6,S7
	Reduced model	S3	S1,S4	S1,S4	S3	S7	S7,S8	S5,S6,S8	S6,S7

$$KGE = 1 - \sqrt{(\rho-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{4}$$

where $\rho = \frac{Cov(Y_t^{obs}, Y_t^{est})}{\sigma_{obs}\sigma_{est}}$, $\alpha = \frac{\sigma_{est}}{\sigma_{obs}}$, $\beta = \frac{\mu_{est}}{\mu_{obs}}$ (μ and σ represent the mean and standard deviation of a series and *est* and *obs* stand for estimation and observation set). KGE ranges from $-\infty$ to 1, the closer to 1 the more accurate the model is.

4. Application on the Durance Watershed

The method discussed in section 3.1 is now applied to the eight stations of the Durance watershed for the daily flow measurements of 107 years. There are two main parts in this section. First, we present the results for the model identification and parameters estimation, following with the model validation in the second part.

In order to estimate the parameters of the models, we used the longest part of the data set that has no missing values, namely the last years. Therefore, we will use a sequence of 22 years (1980–2001).

Before modeling, in order to reduce the effect of the outliers (extreme events that create tails in the distribution of the data) and to build a near normal distribution of the observations, we transformed the data, so instead of modeling the raw time series, we used the log-transformed one.

4.1. Model Identification and Parameters Estimation

We have seen in section 2 that there is a very strongly correlated group of stations in the upper Durance (S1-S4) and one in the middle Durance with stations S6–S8. Also, we have concluded that this relationship might change when different subsets were analyzed (entire data, cold season data, warm season data).

As a result, each station was considered to have as explanatory variables in the regression at least all the other stations from the same groups (groups defined in section 2.2). Particular attention is given to station S5 that has an unclear status. In this case, we look at the correlation (see Table 2) and selected as explanatory variables only the stations that have a coefficient greater than 0.7. Consequently, S5 will be used as an explanatory variable for the stations S4 and S7 in the all year period and warm season, while for the cold season, S5 will explain only station S7 and, in return, S5 was modeled only by these stations, i.e., S4 and S7 for all data and warm season, and S7 for cold season. The models deduced from the above consideration are called *full models*. An outline of the full models for each station and type of data set/seasons (entire data/no season split, cold season, warm season) is given in Table 3 (rows 1 and 3). The models may thus differ according to the type of season, so we considered two approaches for the estimation: one that uses a single-model (no season split denoted by M.NS), and the other one that uses a double-model (two-season split denoted by M.2S).

4.1.1. Multicolinearity

Results in Table 3 show that the full model for each station has multiple explanatory variables. Due to the fact that in our exploratory analysis from section 2.2, we have encountered very high correlated stations, we now want to examine if we are in the case of multicolinearity (almost perfect linear relationship among explanatory variables) and thus of faulty parameters' estimation, among others. The reader is referred to Gujarati and Porter [2008] for detailed definitions and consequences of multicolinearity in regression. We measured the strength of the multicolinearity by computing the Variance Inflation Factor (VIF). This index measures how much the variance of estimated regression coefficients is increased when compared to having uncorrelated variables; see *Kutner et al*. [2004] and Gujarati and Porter [2008] for more details. When multicolinearity is found, viz., the computed VIFs are greater than 5 as suggested in *Eng et al*. [2005];

Montgomery et al. [2012], we drop the variable with the highest VIF (among the one with $VIF > 5$) and then reiterate the process until all remaining variables have $VIF \leq 5$.

In our case-study, the results show that there is evidence of colinearity between the group of stations {S1,S2,S3,S4} from the upper Durance. This brings us a *reduced-form model*, as presented in Table 3 (rows 2 and 4). This reduced-form model is estimated and validated later in the study.

4.1.2. Stationarity

After applying the ADF and KPSS tests for the data from 1980 to 2001 (model estimation data set) and looking at their resulted p value, it seemed clear that all the stations are stationary as all the p values for the KPSS test (with the null hypothesis $\mathcal{H}_0$:stationarity) are greater than 0.05 and for the ADF test ($\mathcal{H}_0$:not stationary) less than 0.05.

4.1.3. Model Initialization

We considered, for each station, six lags $(t, t-1, \ldots, t-5)$ for each explanatory variable included in the regression according to the reduced-form model and an AR(1) for the residuals. We conducted the modeling for both models (no-season split denoted by M.NS and two-season split denoted by M.2S).

4.1.4. Parameter Estimation

The estimation of the models is performed by means of the maximum likelihood estimation (MLE) approach introduced by *Gardner et al*. [1980], specifically the MLE via Kalman filter technique. Readers can find a good discussion about this approach in *Ripley* [2002].

The errors of the proxy model were checked and, as they were not a white noise process, a new model was needed.

4.1.5. Linear Transfer Function

We analyzed the patterns of the estimated regression parameters (remember that we considered six lags at each input variable) for each explanatory variables. Considering the rules presented in section 3.2, the general conclusion is that we have no dead time for none of the input variables, so $b_i = 0$ for all stations, and we have no decay pattern as well, so $r_i = 0$, for $i = \{1, \ldots, l\}$. Given that $r_i = 0$ (no pattern), it means that all the parameters are unpatterned, so $m_i = u_i + r_i - 1$, where u_i = nonzero unpatterned coefficients. In our case-study, the behavior of the unpatterned coefficients was discovered to be as follows: the first coefficient (lag-0) highly significant, the second one (lag-1) close to zero but still significant, while the remaining four (lag-2,3,4,5) nonsignificantly different from zero. Therefore, we considered worth modeling both options: $u_i = 1/m_i = 0$ and $u_i = 2/m_i = 1$ (0- and 1-lag for each explanatory variable) and decide about the best one in the validation section.

4.1.6. ARIMA Model for the Residuals

Analyzing the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots (available upon request), it was found that stations S2 and S3 have a weak weekly seasonality. Analyzing in more detail these time series, it was discovered that the periodicity starts in 1966. We found that in December 1965 a dam was installed upstream of station S2, called Pont-Baldy, so that the water is retained and released every week.

Evaluating the ACF/PACF plots and considering the significant spikes, we proposed for station S2 and S3 a $SARIMA(p, d, q)(P, D, Q)_7$, where $d = D = 0$ (due to stationary data), $p, q \leq 5$ and $P, Q \leq 1$, while for the remaining stations we chose an ARIMA(p,d,q), where $d = 0$, $p \leq 5$ and $q \leq 5$.

Therefore, we have tried several models for the residuals and refitted the DRM accordingly. The selection of the (S)ARIMA model was made by looking at the Akaike Information Criterion (AIC) [*Akaike*, 1974] and the Bayesian Information Criterion (BIC) [*Schwarz*, 1978]. As the procedure is straightforward and due to space concerns, we did not show the results obtained for each model, so we report only the resulted best models:

1. S1: ARIMA$(1, 0, 4)$, invariant of the subset used (all data, cold, or warm season)
2. S2: SARIMA$(1, 0, 2)(1, 0, 1)_7$ for all data and cold season and SARIMA$(2, 0, 2)(1, 0, 1)_7$ for warm season data
3. S3: SARIMA$(3, 0, 1)(1, 0, 1)_7$ for all data and SARIMA$(2, 0, 2)(1, 0, 1)_7$ for cold and warm season data
4. S4-S8: ARIMA$(2, 0, 2)$, invariant of the subset used (all data, cold, or warm season).

One last point we focused on in the SARIMA model identification was whether the models for S2 and S3 are multiplicative or additive, considering they also have a seasonal part. For illustration, we take the SARIMA$(1, 0, 2)(1, 0, 1)_7$ model of the cold season from S2. This model has the following mathematical formulation for the residuals Z_t:

$$Z_t = \phi_1 Z_{t-1} + \phi_{s,1} Z_{t-7} - \phi_1 \phi_{s,1} Z_{t-8} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_{s,1} e_{t-7} + \theta_1 \theta_{s,1} e_{t-8} + \theta_2 \theta_{s,1} e_{t-9} \quad (5)$$

which is equivalent with an ARIMA$(8,0,9)$ where the AR lags 2,3,4,5,6 $(\phi'_2, \ldots, \phi'_6)$ and MA lags 3,4,5,6 $(\theta'_3, \ldots, \theta'_6)$ are set to zero, i.e.:

$$Z_t = \phi'_1 Z_{t-1} + \phi'_7 Z_{t-7} + \phi'_8 Z_{t-8} + e_t - \theta'_1 e_{t-1} - \theta_2 e_{t-2} - \theta'_7 e_{t-7} - \theta'_8 e_{t-8} - \theta'_9 e_{t-9} \quad (6)$$

As reported in *Suhartono* [2011], a multiplicative SARIMA assumes that the parameters related to the nonseasonal and seasonal combination (i.e., parameters $\phi'_8, \theta'_8, \theta'_9$ from (6)) are significant and that they are equal with the multiplication between the parameters of the nonseasonal and seasonal components (i.e., $\phi_1 \cdot \phi_{s,1}$, $\theta_1 \cdot \theta_{s,1}$, $\theta_2 \cdot \theta_{s,1}$ from (5)). The results show that all the estimated multiplicative parameters (i.e., $\hat{\phi}'_8, \hat{\theta}'_8, \hat{\theta}'_9$) are significant and, moreover, the multiplication of the nonseasonal and seasonal estimated coefficients ($\hat{\phi}_1 \cdot \hat{\phi}_{s,1}$, $\hat{\theta}_1 \cdot \hat{\theta}_{s,1}$, $\hat{\theta}_2 \cdot \hat{\theta}_{s,1}$) prove to be each time inside the 95% confidence interval for the estimated $\phi'_8, \theta'_8, \theta'_9$ (e.g., $\hat{\phi}'_8 \pm 1.96 \cdot \sigma$). We concluded, thus, that a multiplicative SARIMA was suited for both S2 and S3.

4.1.7. Model Checking

We analyzed the errors of the fitted models and checked if they are a white noise process (zero mean, finite variance, and independence). The Ljung-Box test [*Box and Pierce*, 1970; *Box and Jenkins*, 1976], with the null hypothesis of independence, was used to test the serial correlation. The results show that all the models present independent errors (i.e., all *p* values of the test are near 0.9), mean close to zero and finite variance.

4.2. Model Validation and Performance Evaluation

In order to measure the performance, but also the stability of the estimations, we take into consideration two situations:

1. The data for the explanatory variables in the regression are all present (complete-covariates model).
2. The data for the explanatory variables in the regression are partially or totally missing (missing-covariates model).

For each situation, we used three different test sets each containing four years of daily flow data, that is: 1918–1921, 1931–1934 and 2002–2005 (parameters estimation was performed on the period 1980-2001).

The performance of the models was then compared with a simpler, but common method of reconstructing missing meteorological data [*Hirsch*, 1979; *Wallis et al*., 1991; *Bárdossy and Pegram*, 2014], the nearest-neighbors technique (NN). This method allows the infilling of missing data for a station by taking information from neighbor stations (transferred directly or weighted). For this study, we used as neighbors the input variables initially selected for the regression part of our full model with all data (first row in Table 3). The missing values of the target station were obtained by weighting the neighbor station(s) with the ratio of daily mean flow of the target station over the daily mean flow of each neighbor.

Beside the NN, we used also for comparison continuous streamflow time series over the 1904–2010 period, obtained from meteorological data reconstruction (ANATEM method) [see *Kuentz et al*., 2015, for details] and rainfall-runoff (RR) modeling [*Kuentz*, 2013; *Kuentz et al*., 2013]. This approach includes a complex reconstruction of streamflow data taking into account meteorological input variables, along with a snow-accumulation and melt-process modeling.

To ensure the reliability of our estimations, we end this section with the performance results of the estimated models on simulated data.

4.2.1. Validation When the Data for the Explanatory Variables are all Present (Complete-Covariates Model)

The accuracy of the models was investigated through the KGE criterion described in section 3.3. The results are illustrated in Table 4.

One important aspect that must be emphasized with respect to the KGE results, is that they are rather persistent over the three test sets, meaning that, although in the parameters estimation step we used only the last period of the data, the models behave the same at the beginning or middle part of the 107 years timespan. We have one exception at station S2, that performs better for period 2002–2005 under model M.2S.0lag, but nevertheless, the KGE index is close to model M.2S.1lag that was selected in the other two test periods.

Table 4. KGE Results for the Validation of the Test-Period 1918–1921, 1931–1934, and 2002–2005 for (a) the Complete-Covariates Models and the Two Alternative Methods of Infilling, NN, and ANATEM-RR, and (b) the Missing-Covariates Models (Results Shown Only for the Best-Case Complete-Covariates Models)

		S1	S2	S3	S4	S5	S6	S7	S8
Period 1918–1921[a]									
Complete-covariates model	M.NS.0lag	0.778	0.763	0.879	0.627	**0.857**[a]	0.685	0.845	0.787
	M.2S.0lag	0.863	0.873	0.935	0.764	0.835	0.713	0.844	**0.794**[a]
	M.NS.1lag	0.827	0.795	0.886	0.712	0.812	0.744	**0.902**[a]	0.751
	M.2S.1lag	**0.866**[a]	**0.890**[a]	**0.941**[a]	**0.811**[a]	0.786	**0.776**[a]	0.893	0.784
	NN	***0.926***[b]	***0.904***[b]	*0.656*	*0.652*	*0.702*	*0.724*	*0.815*	*0.522*
	ANATEM-RR	*0.751*	*0.627*	*0.871*	*0.593*	*0.770*	*0.730*	*0.561*	*0.220*
Missing-covariates model	Scenario 1	0.866	0.890	0.941	0.811	0.857	0.679	0.753	0.674
	NAs[c]	*1461*	*0*	*0*	*0*	*0*	*731*	*1311*	*1461*
	Scenario 2	0.866	0.890	0.941	0.811	0.726	0.702	0.799	0.688
	NAs	*1*	*731*	*1461*	*0*	*730*	*1096*	*1461*	*517*
Period 1931–1934[a]									
Complete-covariates model	M.NS.0lag	0.839	0.826	0.761	0.664	**0.722**[a]	0.651	0.749	0.722
	M.2S.0lag	0.910	0.914	0.823	0.833	0.635	0.671	0.740	**0.741**[a]
	M.NS.1lag	0.888	0.856	0.767	0.733	0.662	0.705	**0.796**[a]	0.691
	M.2S.1lag	**0.912**[a]	**0.923**[a]	**0.838**[a]	**0.871**[a]	0.591	**0.729**[a]	0.777	0.734
	NN	*0.897*	*0.769*	***0.907***[b]	*0.743*	*0.543*	*0.519*	*0.793*	*0.410*
	ANATEM-RR	*0.861*	*0.831*	*0.707*	*0.731*	*0.516*	***0.773***[b]	*0.495*	*0.261*
Missing-covariates model	Scenario 1	0.912	0.923	0.838	0.871	0.722	0.654	0.553	0.542
	NAs	*1461*	*0*	*0*	*0*	*0*	*731*	*1311*	*1461*
	Scenario 2	0.912	0.923	0.838	0.871	0.688	0.595	0.650	0.570
	NAs	*1*	*731*	*1461*	*0*	*730*	*1096*	*1461*	*517*
Period 2002–2005[a]									
Complete-covariates model	M.NS.0lag	0.718	0.898	0.853	0.769	**0.780**[a]	0.636	0.809	0.850
	M.2S.0lag	0.804	**0.946**[a]	0.919	0.905	0.724	0.673	0.810	**0.890**[a]
	M.NS.1lag	0.769	0.930	0.858	0.859	0.709	0.694	**0.870**[a]	0.809
	M.2S.1lag	**0.812**[a]	0.933	**0.931**[a]	**0.936**[a]	0.672	**0.743**[a]	0.867	0.876
	NN	*0.774*	*0.694*	*0.885*	*0.839*	*0.587*	*0.591*	*0.769*	*0.805*
	ANATEM-RR	***0.823***[b]	*0.873*	*0.880*	*0.889*	*0.755*	***0.829***[b]	*0.804*	*0.706*
Missing-covariates model	Scenario 1	0.812	0.933	0.931	0.936	0.780	0.670	0.715	0.673
	NAs	*1461*	*0*	*0*	*0*	*0*	*731*	*1311*	*1461*
	Scenario 2	0.811	0.933	0.930	0.936	0.780	0.696	0.717	0.788
	NAs	*1*	*731*	*1461*	*0*	*730*	*1096*	*1461*	*517*

[a]Best model out of the four estimated models.
[b]Cases when NN or ANATEM-RR performs better compared to our best-case model for each station.
[c]*NAs* = number of missing values.

First, the results show that six stations, i.e., all stations from upper Durance (S1–S4) and two stations from middle Durance (S6, S7), have a better fit with model that includes 1-lag for the explanatory variables, while only two stations from middle Durance (S5, S8) work better with model that has 0-lag. The results are coherent when looking at the hydrological regimes and the characteristics of the stations. These stations (S1–S4, S6, S7) are situated at a high altitude and have mainly a regime influenced by snowmelt and low temperature, so it is probable that some delay may appear for the flow. Considering station S5 and S8, the absence of lags is due to the fact that the watershed has a small drainage area (S5) or is characterized by a lowland basin (S8).

Second, for some stations (S1–S4, S6, S8), the models with two-season split are selected, while for the others (S5, S7) the ones with no-season split. There is no clear hydrological explanation for this behavior. It is to be noted that we look only at some characteristics, like hydrological regimes, watershed surface, and altitude; we can have other influential factors that may drive these two stations.

In Table 5, one can find a summary of the selected models for each station, along with their estimated parameters.

The superiority of our approach is emphasized when comparing the KGE results of our models with the ones from the two approaches mentioned earlier, NN and ANATEM-RR. The results reveal that, except some isolated cases (three for NN and three for ANATEM-RR), our approach performs better in each case. An important aspect that must be highlighted is that with DRM the efficiency of the models (KGE) is never lower than 0.72, while NN and ANATEM-RR, due to lack of robustness, reduce up to a level of 0.41 and 0.22, respectively, as seen in Table 4.

Table 5. Summary of the Selected Models for Each Station

	Selected **Models**[a]		Model Parameters										
S1	M.2S.1lag	**Cold**	β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	ϕ_1	θ_1	θ_2	θ_3	θ_4			
			−1.37	0.69	0.05	0.99	0.02	−0.10	−0.06	−0.03			
		Warm	β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	ϕ_1	θ_1	θ_2	θ_3	θ_4			
			−1.32	0.78	−0.03	0.97	0.15	−0.06	−0.07	−0.04			
S2	M.2S.1lag	**Cold**	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$	
			1.29	0.47	−0.03	0.16	0.06	0.89	−0.25	−0.16	0.94	−0.48	
		Warm	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	ϕ_2	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$
			1.08	0.68	0.06	0.12	0.02	0.10	0.62	0.59	−0.17	0.83	−0.36
S3	M.2S.1lag	**Cold**	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	ϕ_2	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$
			2.18	0.51	−0.03	0.35	0.04	1.45	−0.46	−0.56	−0.11	0.87	−0.81
		Warm	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S4}$	$\omega_{2,1}^{S4}$	ϕ_1	ϕ_2	θ_1	θ_2	$\phi_{s,1}$	$\theta_{s,1}$
			2.16	0.54	0.07	0.32	0.07	0.47	−0.49	−0.51	−0.10	0.90	−0.87
S4	M.2S.1lag	**Cold**	β_0	$\omega_{1,0}^{S3}$	$\omega_{1,1}^{S3}$	ϕ_1	ϕ_2	θ_1	θ_2				
			−1.61	0.98	−0.03	1.50	−0.52	−0.58	−0.14				
		Warm	β_0	$\omega_{1,0}^{S1}$	$\omega_{1,1}^{S1}$	$\omega_{2,0}^{S5}$	$\omega_{2,1}^{S5}$	ϕ_1	ϕ_2	θ_1	θ_2		
			0.58	0.56	0.14	0.27	0.10	1.65	−0.66	−0.61	−0.16		
S5	M.NS.0lag		β_0	$\omega_{1,0}^{S4}$	$\omega_{2,0}^{S7}$	ϕ_1	ϕ_2	θ_1	θ_2				
			−0.55	0.62	0.56	1.53	−0.54	−0.73	−0.04				
S6	M.2S.1lag	**Cold**	β_0	$\omega_{1,0}^{S7}$	$\omega_{1,1}^{S7}$	$\omega_{2,0}^{S8}$	$\omega_{2,1}^{S8}$	ϕ_1	ϕ_2	θ_1	θ_2		
			1.04	0.51	0.03	0.51	0.12	1.37	−0.40	−0.45	−0.15		
		Warm	β_0	$\omega_{1,0}^{S7}$	$\omega_{1,1}^{S7}$	$\omega_{2,0}^{S8}$	$\omega_{2,1}^{S8}$	ϕ_1	ϕ_2	θ_1	θ_2		
			1.07	0.58	0.07	0.39	0.07	1.62	−0.63	−0.59	−0.14		
S7	M.NS.1lag		β_0	$\omega_{1,0}^{S5}$	$\omega_{1,1}^{S5}$	$\omega_{2,0}^{S6}$	$\omega_{2,1}^{S6}$	$\omega_{3,0}^{S8}$	$\omega_{3,1}^{S8}$	ϕ_1	ϕ_2	θ_1	θ_2
			−0.17	0.29	0.03	0.10	−0.01	0.31	0.07	1.37	−0.37	−0.53	−0.11
S8	M.2S.0lag	**Cold**	β_0	$\omega_{1,0}^{S6}$	$\omega_{2,0}^{S7}$	ϕ_1	ϕ_2	θ_1	θ_2				
			−0.20	0.37	0.74	1.49	−0.51	−0.73	−0.01				
		Warm	β_0	$\omega_{1,0}^{S6}$	$\omega_{2,0}^{S7}$	ϕ_1	ϕ_2	θ_1	θ_2				
			−0.29	0.31	0.80	1.55	−0.57	−0.66	−0.10				

[a]Corresponding mathematical formulations (illustration of S1 and S2-cold, similar for the other stations):
S1:

$$Y_t^{S1} = \beta_0 + \omega_{1,0}^{S3} X_t^{S3} + \omega_{1,1}^{S3} X_{t-1}^{S3} + Z_t$$
$$Z_t = \phi_1 Z_{t-1} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_3 e_{t-3} - \theta_4 e_{t-4}$$

S2-cold:

$$Y_t^{S2} = \beta_0 + \omega_{1,0}^{S1} X_t^{S1} + \omega_{1,1}^{S1} X_{t-1}^{S1} + \omega_{2,0}^{S4} X_t^{S4} + \omega_{2,1}^{S4} X_{t-1}^{S4} + Z_t$$
$$Z_t = \phi_1 Z_{t-1} + \phi_{s,1} Z_{t-7} - \phi_1 \phi_{s,1} Z_{t-8} + e_t - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \theta_{s,1} e_{t-7} + \theta_1 \theta_{s,1} e_{t-8} + \theta_2 \theta_{s,1} e_{t-9}$$

An illustration of the reconstructed series (considering the best DRM for each station) versus the observed one for the period 2002–2005 can be seen in Figure 5. One can notice that the reconstructions of stations S1, S4, and S5 do not reproduce completely the peak flows, but the recessions are good. Stations S2 and S3 catch very well the peaks, but the weekly fluctuations (stronger at S2, see zoomed areas (a) and (c) in Figure 5) decrease in estimation performance for the long-term reconstructions (see zoomed sectors (a) and (b)). Regarding the other stations, S6 and S7 have mainly well-modeled reconstructions, while station S8 has some overestimated peaks. These aspects should be further studied and addressed in a future research.

The other two periods have similar graphs.

4.2.2. Validation When the Data for the Explanatory Variables are Partially or Totally Missing (Missing-Covariates Model)

There are cases when the complete-covariates model from the previous section cannot be applied as the data for the explanatory variables are missing. The purpose of this section is to test how the proposed models behave in this case. Therefore, in order to be able to apply the estimated (complete-covariates) models, we use the weighted values from the correlated-neighbor stations (i.e., same procedure as in the case of NN estimation, presented at the beginning of section 4.2). When all the covariates are missing, we use the daily mean (mean of the nonmissing values for a certain day for that stations)

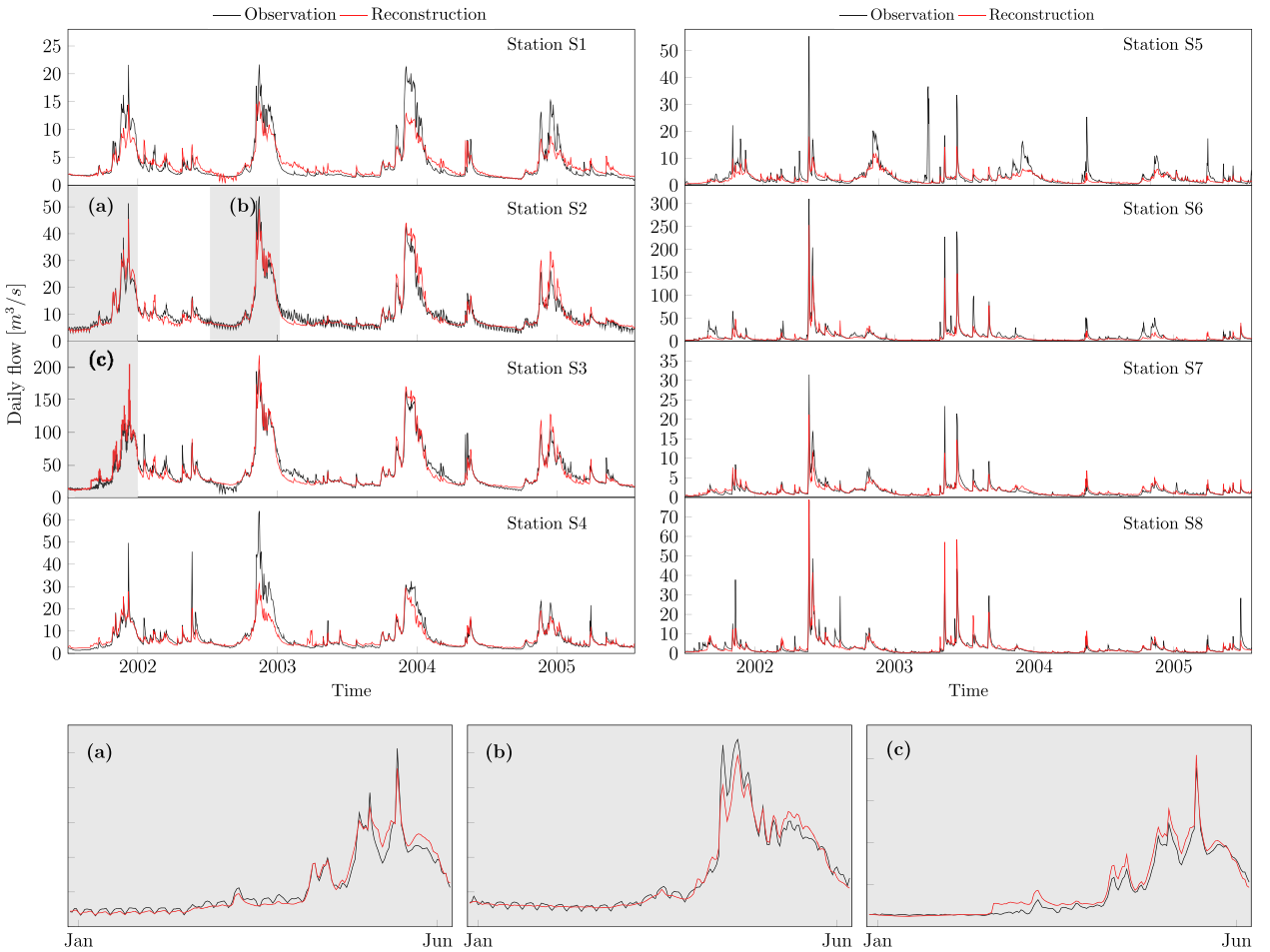

Figure 5. Daily flow estimations versus observations for period 2002–2005 for the Durance watershed, along with three zoomed areas (a), (b), and (c).

In order to validate this procedure, we use only the best models selected for the complete-covariates model study (see Table 5). Then, for each station at a time, we overlay on each test set (1918–1921, 1931–1934, 2002–2005) the pattern of missing values from two periods, 1904–1907 (denoted Scenario 1) and 1951–1954 (denoted Scenario 2). The advantage of this procedure is that we created two scenarios with missing input variables, but we have also the observations in order to test the accuracy. In order to have the best possible output, we proceed first with the stations with the fewest missing values, finishing the reconstruction with the station with the largest number of missing data.

As it was seen from the previous section, the proposed technique of reconstructing streamflow data yields very good results when all input variables are available (complete-covariates model), surpassing the performance of more complex models like ANATEM-RR. However, the KGE results for the missing-covariates models in Table 4 show that by replacing the missing values in the input variables with the weighted values from the correlated neighbors, we slightly decrease in performance, but, overall, the KGE is still above 0.5.

4.2.3. Validation on Simulated Data

In the previous section, we validated our DRMs using a deterministic procedure, thus providing a unique KGE value for each model and station. However, as the used infilling models are stochastic, a single run of the model might not provide enough information about the KGE. Therefore, it is recommended to run the model several times and treat the KGE as a random variable.

In this case, we simulated daily streamflow data for the eight stations for the period 2002–2005. For each station, we started by randomly generating *nsim* (*nsim* = 50) different white noise sequences for the error terms (e_t in (3)) and used them, along with the already estimated (S)ARIMA parameters (see section 4.1), to

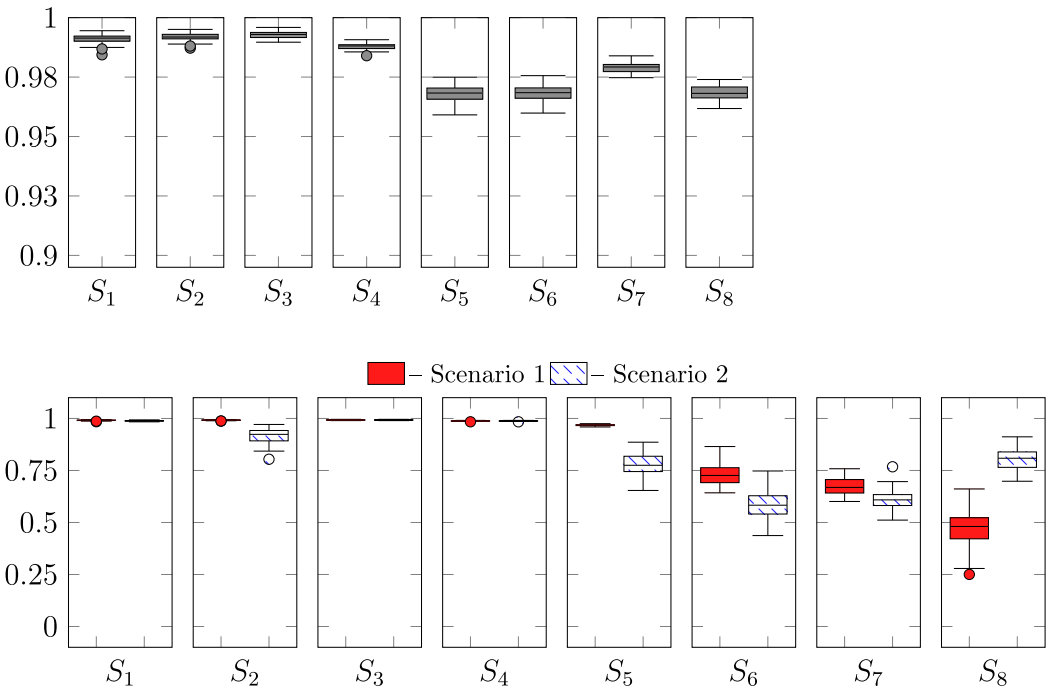

Figure 6. Box-plots of KGE for the *nsim* (*nsim* = 50) simulations. Note: (top plot) Results for the complete-covariates model; (bottom plot) the results for the missing-covariates model for both scenarios (Scenario 1 = overlay periods 1904–1907 and Scenario 2 = overlay period 1951–1954).

create *nsim* residuals series (Z_t in (3)). Then, using the input variables (from the observed daily streamflow series, period 2002–2005) and the previously estimated regression parameters, we performed *nsim* daily streamflow simulations (denoted sim_i, $i=1,\ldots,nsim$).

Afterwards, considering sim_i, we followed the same modeling procedure as in the case of observed streamflow: estimate the parameters and find the fitted series (denoted fit_i, $i=1,\ldots,nsim$). The performance of the estimations was computed with KGE, between fit_i and sim_i data.

Due to space reasons and the similarities between outputs, we discuss and illustrate just the results for model M.NS.1lag. The validation methodology for the simulated data is similar to the one used in sections 4.2.1 and 4.2.2 when we worked with observed daily streamflow data.

Therefore, Figure 6 (top), the validation of the complete-covariates model on simulated data shows that we have a very good performance in the upper Durance (average KGE above 0.99 and variability smaller than 0.002) and a slightly smaller one in middle Durance (average KGE above 0.96 and variability smaller than 0.003), behavior that, in fact, reinforces the statements from the validation on observed data. In Figure 6 (bottom), the validation for missing-covariates model is illustrated for the two scenarios mentioned in the previous section. Attention must be paid during analyzing these plots because the two scales are very different. Once again, it is shown that we decrease in performance when we replace the missing input variables with the weighted values of the correlated neighbors, but the average KGE remains, mainly, above 0.5.

Finally, the reconstructed series for the eight stations can be seen in Figure 7. The reconstructions show once more that in case of an infilling using the complete-covariates model (all covariates are present) the estimations are extremely good (see stations S1,S2,S3, where the percentage of missing data estimated using the complete-covariates model out of the total missing points is 99.69%, 59.94%, and 94.93%, respectively). They slightly decrease in performance when we deal with missing explanatory variables in the model, see the case of the stations from the middle Durance (S5,S6,S7,S8, where the above mentioned percentage decreases to 27.07%, 17.78%, 7.81%, and 19.03%, respectively).

The computations were performed with the R Software, using the packages: *stats* (general-main computations), *iki.dataclim* (homogeneity tests), *cluster* (PAM exploratory analysis), *tseries* (stationarity analysis), and *forecast* (DRM fit and prediction).

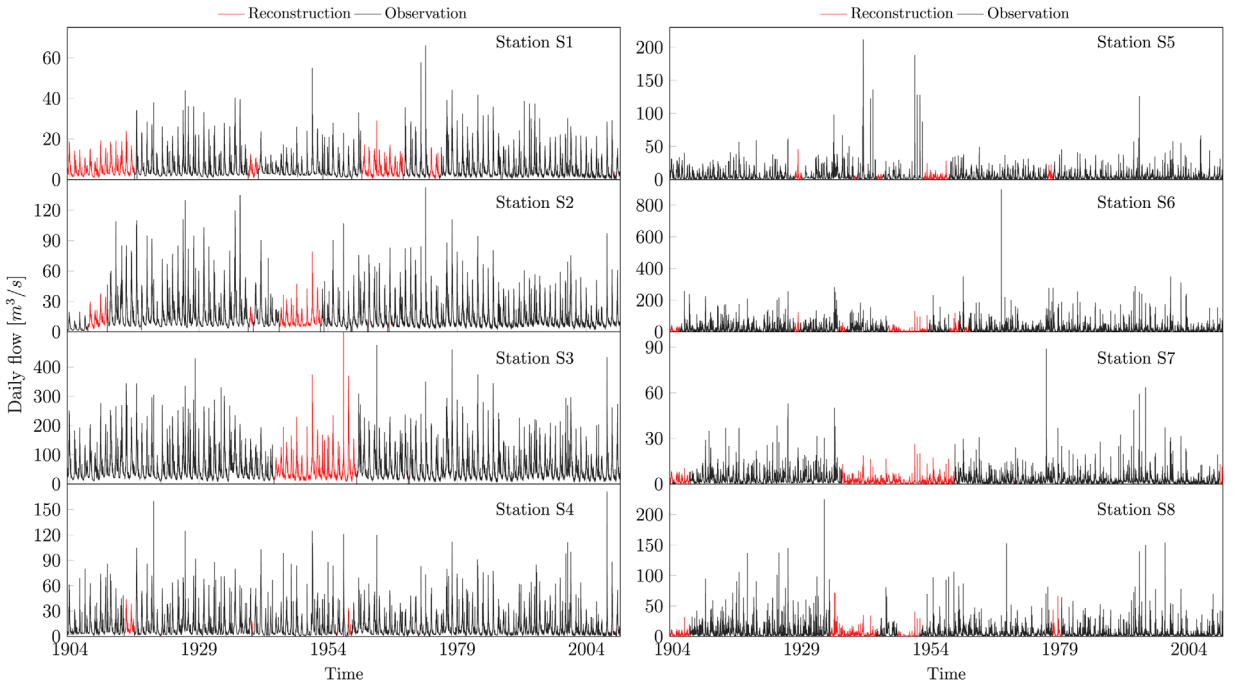

Figure 7. Daily flow reconstructed series of the eight stations of the Durance watershed.

5. Conclusions

Complete records of flow data are very important and critical to a sustainable management of water resources. During the past decades, researchers have developed techniques to reconstruct these series using a variety of methods such as linear and nonlinear models, parametric and nonparametric approaches, etc.

In this study, we present a way of reconstructing daily streamflow data by using dynamic regression. The method uses the linear relationship between the correlated stations at different lags and it adjusts the residuals by fitting a (S)ARIMA model. The proposed reconstruction technique addresses the case when one has access only to daily streamflow time series data and has not available other measurements, i.e., precipitation. It is an accessible approach and it can handle even large amount of observations in a short run-time period. Apart from this, our study was performed on a large watershed characterized by several hydrological regimes and various data quality issues, so it brings a solid and complex analysis.

The results of the application on the eight stations of the Durance river show that dynamic regression models outperform two other modeling approaches, nearest neighbor technique and a more complex meteorological model (ANATEM-RR). When measuring the accuracy of the estimates, it was proven that the choice of the model is highly dependent on the station's characteristic and hydrological regimes and no generalization can be made for all stations. In other words, we have seen that for all the stations from upper Durance we have chosen the dynamic regression model with 1-lag explanatory variables and two-season model, but for the middle Durance, as the stations are more mixed, we have models with or without past lags included and with or without seasonal models. We have also showed that even if we are in the case of missing covariates in the regression, the models can perform well by replacing the missing covariate value with the weighted values of the correlated neighbors or the daily mean when all covariates are missing. However, this action will produce less variability in those parts of the time series and one can have less accurate outputs for the extreme values. More robust methods should be used for a better accuracy in this case.

In conclusion, we introduced in this study a method for reconstructing hydrological data that is very general, flexible, and requires only streamflow data. Based on the results obtained for the Durance watershed, if the model is estimated meticulously, good results can be obtained for any hydrological regime or station type.

Acknowledgments
The authors are grateful to the Editor, Associate Editor, and reviewers for their constructive comments and suggestions which led to significant improvements in the paper. This work was supported by the French national program LEFE/INSU. We thank EDF and Anna Kuentz for providing the data set used in the application part. Because of confidentiality issues, the entire data set cannot be released, but a partial one can be obtained from the public HYDRO database (www.hydro.eaufrance.fr).

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Autom. Control*, *19*(6), 716–723.
Amisigo, B. A., and N. C. van de Giesen (2005), Using a spatiotemporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series, *Hydrol. Earth Syst. Sci.*, *9*(3), 209–224.
Bárdossy, A., and G. Pegram (2014), Infilling missing precipitation records: A comparison of a new copula-based method with other techniques, *J. Hydrol.*, *519*, 1162–1170.
Bercu, S., and F. Proïa (2013), A SARIMAX coupled modelling applied to individual load curves intraday forecasting, *J. Appl. Stat.*, *40*(6), 1333–1348.
Box, G. E. P., and G. M. Jenkins (1976), *Time Series Analysis: Forecasting and Control*, 575 pp., Holden-Day, San Francisco.
Box, G. E. P., and D. A. Pierce (1970), Distribution of residual autocorrelations in autoregressive-integrated moving average time series models, *J. Am. Stat. Assoc.*, *65*(332), 1509–1526.
Coulibaly, P., and C. K. Baldwin (2005), Nonstationary hydrological time series forecasting using nonlinear dynamic methods, *J. Hydrol.*, *307*(1-4), 164–174.
Elshorbagy, A., S. Simonovic, and U. Panu (2002), Estimation of missing streamflow data using principles of chaos theory, *J. Hydrol.*, *255*, 123–133.
Eng, K., G. D. Tasker, and P. Milly (2005), An analysis of region-of-influence methods for flood regionalization in the Gulf-Atlantic rolling plains, *J. Am. Water Resour. Assoc.*, *41*(1), 135–143.
Gardner, G., A. Harvey, and G. Phillips (1980), An algorithm for exact maximum likelihood estimation of average models by means of Kalman filtering, *J. R. Stat. Soc. Ser. C*, *29*(3), 311–322.
Greenhouse, J. B., R. E. Kass, and R. S. Tsay (1987), Fitting nonlinear models with ARMA errors to biological rhythm data, *Stat. Med.*, *6*(2), 167–183.
Gujarati, D., and D. Porter (2008), *Basic Econometrics*, 5th ed., 922 pp., McGraw-Hill, N. Y.
Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*(1–2), 80–91.
Gyau-Boakye, P., and G. Schultz (1994), Filling gaps in runoff time series in West Africa, *Hydrol. Sci. J.*, *39*(6), 621–636.
Harvey, C., H. Dixon, and J. Hannaford (2012), An appraisal of the performance of data-infilling methods for application to daily mean river flow records in the UK, *Hydrol. Res.*, *43*(5), 618–636.
Hirsch, R. (1979), An evaluation of some record reconstruction techniques, *Water Resour. Res.*, *15*(6), 1781–1790.
Imbeaux, E. (1892), *La Durance: Régime, Crues et Inondations*, 200 pp., Vve Ch. Dunod, Paris.
Kang, H. M., and F. Yusof (2012), Homogeneity tests on daily rainfall series, *Int. J. Contemp. Math. Sci.*, *7*(1), 9–22.
Kaufman, L., and P. J. Rousseeuw (1990), *Finding Groups in Data: An Introduction to Cluster Analysis, Wiley Ser. Probab. Stat.*, 368 pp., John Wiley, USA.
Khalil, M., U. Panu, and W. Lennox (2001), Groups and neural networks based streamflow data infilling procedures, *J. Hydrol.*, *241*, 153–176.
Kim, J.-W., and Y. A. Pachepsky (2010), Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation, *J. Hydrol.*, *394*(3–4), 305–314.
Kuentz, A. (2013), Un siècle de variabilité hydro-climatique sur le bassin de la Durance, PhD thesis, AgroParisTech., Paris.
Kuentz, A., T. Mathevet, J. Gailhard, C. Perret, and V. Andréassian (2013), Over 100 years of climatic and hydrologic variability of a Mediterranean and mountainous watershed: The Durance River, in *Cold and Mountain Region Hydrological Systems Under Climate Change: Towards Improved Projections Proceedings*, pp. 19–25, Int. Assoc. of Hydrol. Sci., Gothenburg, Sweden.
Kuentz, A., T. Mathevet, D. Coeur, C. Perret, J. Gailhard, L. Guérin, Y. Gash, and V. Andréassian (2014), Historical hydrometry and hydrology of the Durance river watershed, *La Houille Blanche*, *4*, 57–63.
Kuentz, A., T. Mathevet, J. Gailhard, and B. Hingray (2015), Building long-term and high spatiotemporal resolution precipitation and air temperature reanalyses by mixing local observations and global atmospheric reanalyses: The ANATEM method, *Hydrol. Earth Syst. Sci. Discuss.*, *12*(1), 311–361.
Kutner, M. H., C. Nachtsheim, and J. Neter (2004), *Applied Linear Regression Models*, 4th ed., 701 pp., McGraw-Hill, Boston.
Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin (1992), Testing the null hypothesis of stationarity against the alternative of a unit root, *J. Economet.*, *54*(1-3), 159–178.
Lehmann, E. L., and H. J. M. D'Abrera (2006), *Nonparametrics: Statistical Methods Based on Ranks*, 463 pp., Springer.
Makridakis, S. G., S. C. Wheelwright, and R. J. Hyndman (1998), *Forecasting: Methods and Applications*, 656 pp., John Wiley.
Miaou, S.-P. (1990), A stepwise time series regression procedure for water demand model identification, *Water Resour. Res.*, *26*(9), 1887–1897.
Montgomery, D. C., E. A. Peck, and G. G. Vining (2012), *Introduction to Linear Regression Analysis*, 5th ed., 672 pp., John Wiley, Hoboken, N. J.
Nash, J., and J. Sutcliffe (1970), River flow forecasting through conceptual models part I - A discussion of principles, *J. Hydrol.*, *10*(3), 282–290.
Pankratz, A. (1991), *Forecasting with Dynamic Regression Models*, 400 pp., Wiley-Interscience, USA.
Raman, H., S. Mohan, and P. Padalinathan (1995), Models for extending streamflow data: A case study, *Hydrol. Sci. J.*, *40*(3), 381–393.
Ripley, B. D. (2002), Time series in R 1.5.0, *R J.*, *2*(2), 2–7.
Rousseeuw, P. (1987), Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.*, *20*, 53–65.
Said, S. E., and D. A. Dickey (1984), Testing for unit roots in autoregressive-moving average models of unknown order, *Biometrika*, *71*(3), 599–607.
Schneider, T. (2001), Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values, *J. Clim.*, *14*(5), 853–871.
Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.
Schwert, G. W. (1989), Tests for unit roots: A Monte Carlo investigation, *J. Bus. Econ. Stat.*, *7*(2), 147–159.
Suhartono (2011), Time series forecasting by using seasonal autoregressive integrated moving average: Subset, multiplicative or additive model, *J. Math. Stat.*, *7*(1), 20–27. [Available at http://thescipub.com/html/10.3844/jmssp.2011.20.27.]
Tsay, R. (1984), Regression models with time series errors, *J. Am. Stat. Assoc.*, *79*(385), 118–124.
Wallis, J. R., D. P. Lettenmaier, and E. F. Wood (1991), A daily hydroclimatological data set for the continental United States, *Water Resour. Res.*, *27*(7), 1657–1663.
Wijngaard, J. B., A. M. G. Klein Tank, and G. P. Konnen (2003), Homogeneity of 20th century European daily temperature and precipitation series, *Int. J. Climatol.*, *23*(6), 679–692.
Woodhouse, C. A., S. T. Gray, and D. M. Meko (2006), Updated streamflow reconstructions for the Upper Colorado River Basin, *Water Resour. Res.*, *42*, W05415, doi:10.1029/2005WR004455.