HYDROLOGICAL SCIENCES JOURNAL
2018, VOL. 63, NO. 12, 1745–1758
https://doi.org/10.1080/02626667.2018.1541090

Kernel distributed residual function in a revised multiple order autoregressive model and its applications in hydrology

Nesa Ilich[a], Amr Gharib[b] and Evan G. R. Davies[b]

[a]Optimal Solutions Ltd., Calgary, Alberta, Canada; [b]Department of Civil and Environmental Engineering, University of Alberta, Edmonton, Alberta, Canada

ABSTRACT
This paper describes a new approach to fill missing data in hydrologic series. Based on a multiple-order autoregressive model, our algorithm represents the random term with an empirical distribution function that includes different parameters for the low, medium and high ranges of the modelled hydrologic variable. The algorithm involves a corrective mechanism that preserves the original statistical distribution of the series that are filled, while also eliminating the possibility of obtaining negative values for low flows. The algorithm requires multiple correlated hydrologic time series with sufficient data to permit accurate calculation of their statistical properties. It ensures that both the original statistical dependence among the data series and the statistical distribution functions will be preserved after the missing data had been filled. The model has been tested using 15 streamflow series in the Upper Bow River watershed in Alberta, Canada.

ARTICLE HISTORY
Received 22 November 2017
Accepted 13 September 2018

EDITOR
A. Castellarin

ASSOCIATE EDITOR
A. Langousis

KEYWORDS
regression; autoregressive models; residual function; missing data; hydrologic time series

1 Introduction

Gaps in historic streamflow data series pose a significant problem for hydrologists and water resources practitioners. Long and continuous data sets allow more reliable estimates of the expected range of water availability during various critical periods, and they form important input to studies that require continuous data points for multiple locations in a river basin within a common time period. While desirable, the availability of such data is rare, since hydrometric stations in a river basin have commenced or terminated operations at different times, and data monitoring has historically been interrupted due to budgetary constraints or equipment malfunctions, resulting in random occurrence of missing data records at various stations.

Estimating missing data in time series is not only of interest to hydrologists, but also to all professionals who need lengthy continuous observed time series as input to their analyses. This need has prompted many researchers to look for ways to fill the missing data. Moffat *et al.* (2007) compared 15 different missing-data algorithms to estimate missing values of net ecosystem CO_2 exchange (NEE) in eddy-covariance time series and evaluated their performance for different artificial gap scenarios based on a set of 10 benchmark datasets from six forested sites in Europe. They did not find an obvious single winner for all benchmarks, but their work did evaluate both classical as well as new and emerging techniques. Specifically, the classical non-linear regression techniques (NLRs), look-up table (LUT), marginal distribution sampling (MDS), and semi-parametric model (SPM) methods generally performed well, while the artificial neural network based techniques (ANNs) were found to be only slightly better than the other techniques on some benchmarks. The simple interpolation technique of mean diurnal variation (MDV) also produced acceptable results, while some new and sophisticated techniques, including the dual unscented Kalman filter (UKF), the multiple imputation method (MIM), the terrestrial biosphere model (BETHY), one of the ANNs and one of the NLRs, tended to develop biased estimates. Tardivo and Berti (2013) developed criteria for selecting the best predictor stations for in-filling missing daily streamflow data using a dense network of nearby hydrologic stations. In addition to station selection, much of their work was devoted to algorithm selection, but did not provide any firm preference among them. While Elshorbagy *et al.* (2002) had some success in experimenting with the use of chaos theory, earlier approaches that apply classical statistical models primarily based on correlation, regression and moving

CONTACT Nesa Ilich nilich@optimal-solutions-ltd.com

average using various enhancement techniques continue to be used (Gyau-Boakye and Schultz 1994, Simonovic 1995), and have seen continuous improvement, as attested by recent work of Gottschalk *et al.* (2015) and Tencaliec *et al.* (2015).

The ideas in this paper rely on recent developments in statistics that have produced several successful algorithms that preserve statistical correlation between randomly generated variables and their original statistical distributions. Iman and Conover (1982) introduced the first of these algorithms. They were subsequently used in hydrology and improved in various ways (Ilich 2009), and several recent publications have discussed their use to enhance development of stochastic hydrologic time series (Ilich 2013). Although earlier work by Simonovic (1995) provides a foundation for establishing goals and evaluation criteria for such applications – such as the preservation of both the original statistical distribution function, as well as the statistical dependence with the adjacent hydrologic time series – the literature lacks publications that apply such methods to in-filling of missing data, where the final outcome is a mix of existing data and modelled results that were used for in-filling.

This paper begins by explaining the concept of developing correlated random variables with arbitrary statistical distributions, which is extended to cover the mix of existing data (that remain unchanged) and missing data that are filled. An application of the algorithm is demonstrated for the Bow River Basin system of Alberta, Canada, which contains 15 locations with time series of naturalized flows on the main stem and its tributaries, some of which have missing records for multiple years. The first case study validates the algorithm, which was applied to periods where historic data were intentionally removed, such that the simulated in-filled series could be compared with the actual historic data. In the second case study, an extended version of the algorithm in-fills missing data for 14 stations in reference to only one principal downstream station, where data at this station were obtained through disaggregation of annual flow estimates based on tree ring data. Here, annual flow estimates for the 1111–1911 period were obtained from a regressive relationship of the historic flows and tree rings data from multiple trees developed for the 1912–2013 period. These estimates were completed as part of a separate study by Sauchyn and Ilich (2017), so their development is not described in detail in this paper. However, important conceptual ideas related to the disaggregation of annual flows are provided here for completeness and clarity. This approach helped complete over 904 years of weekly flows at 15 locations that preserve all the relevant weekly flow statistics of the last 85 years of the historic record. The City of Calgary has used these weekly series as input for its modelling of drought management operational scenarios.

2 Conceptual basis of the proposed algorithm

The proposed algorithm requires that, for each time step with missing data for a particular series, at least one other hydrometric station with available data can be used as a point of reference. If all hydrometric stations within the available subset lack data for certain time periods, then either an alternative technique (usually a combination of autoregressive and moving average approaches) must be applied first to in-fill the missing data on one of the stations, or a new station must be included in the subset. This new station may be more remote, but must have a reasonable statistical correlation that can be used to in-fill the missing data.

In general terms, multiple order auto regression model has the following formulation:

$$y_{i,t} = \sum_{t=1}^{m} \sum_{j=1}^{n} b_{j,t} x_{j,t} + c_0 + \sigma_i \qquad (1)$$

This formulation predicts the dependent variable $y_{i,t}$ so as to combine the spatial cross-correlation between correlated data at various hydrometric stations (or locations *j*) with one or more time lags *t*. The above model is subdivided into three terms on the right hand side of the equation: (a) a linear form of independent variables $x_{j,t}$ and their coefficients $b_{j,t}$, presented as the sum product; (b) a regression constant term c_o; and (c) random term σ_i which is normally distributed with the mean of zero and standard deviation equal to the standard error of regression, i.e. $\sigma_i = N[0, \varepsilon_i]$. Each part will be discussed in more detail below.

The sum-product term contains the influence of spatial and temporal statistical dependence. The model evaluates these dependences on an equal footing, based exclusively on the value of their correlation coefficients. For data in-filling algorithms of average weekly flows, it is typically sufficient for practical purposes to focus on high cross-correlation in spatial terms. The most significant temporal dependence found in autocorrelation is usually met by association with the independent variables (stations that have complete data). However, this may not always be the case, and the model should be able to also replicate autocorrelation functions. The number of independent stations *n* in model construction is a matter of judgment. It depends on the number of available stations and their correlation coefficients with the particular station with missing data, and can be set by applying principal component analyses to the correlation matrix either by setting a low-correlation threshold below which the correlation is no longer hydrologically meaningful, or by limiting the size of the pool of independent stations (parameter n in Equation (1)), and placing in it the stations with the highest correlations to the station with missing data. The above model can be constructed for each series with missing data by first calculating the correlation matrix and then using the correlation coefficients to calculate the regression coefficients b_j with respect to the selected station with missing data (Ilich 2009). There are several important limitations of this model defined by Equation (1) that should be addressed and corrected so as to ensure its productive use in hydrology.

The regression constant c_o may or may not have a physical meaning. For large river basins with year-round flows, this constant should represent the absolute minimum flow; however, if used in this manner, it is usually highly inaccurate. For example, in many cases a positive value of c_o produces an artificially high minimum flow, especially on smaller streams where actual flows may reach very small or zero values. In contrast, a negative regression-constant value may cause the model to produce negative flows for periods with extremely low independent flows associated with independent regression variables. Neither the artificial minimum flow that can never be violated, nor the negative river flows calculated due to a negative regression constant are a desirable outcome of standard regression models that have been used in hydrology. In the past, such outcomes were usually corrected manually by the practitioners.

Finally, the random term σ_i is a product of the standard error of the estimate and a standardized normally-distributed random variable with a mean of zero and a standard deviation of ε_i. Frequent use of this term within a linear model typically causes the final model results also to be normally distributed, despite the fact that the distribution of hydrologic series is asymmetric. Recognizing this disadvantage, some researchers have therefore applied alternative approaches. For example, Efstratiadis *et al.* (2014) used a three-parameter gamma distribution instead of normal distribution, which is a better representation of the asymmetric distribution of residuals. The approach in this work is similar, but uses instead an empirical kernel distribution (Parzen 1962) that has been further refined into different forms as a function of the argument, as explained in the section below entitled “Empirical distribution of the residual function”.

In general, standard linear regression does not guarantee that simulated data share the same statistical properties as the original flow series. In contrast, one of the constraints of the proposed algorithm is to preserve the basic statistics of the original series at a given site, along with the spatial and temporal statistical dependence with the other associated data series. The algorithm for generating correlated random variables is outlined next, so that the case studies presented in this paper are easier to understand.

The first step in the algorithm identifies the data series to be used as a starting reference for all other series. In many instances, the most downstream hydrometric station has the most complete records, and is also correlated to all upstream stations, so the choice is obvious. However, where this is not the case, a dominant starting series may be determined as: (a) the series with the highest correlations to all other series; or (b) the series with the most complete record. These two criteria can be lumped into the following statistic:

$$R_i = w\frac{1}{n}\sum_{j=1}^{n} \mathrm{\rho}_{i,j} + (1-w)\mathrm{f_i} \qquad (2)$$

where $0 \leq w \leq 1$, $0 \leq w \leq 1$.

The statistic R_i refers to station i and it represents the weighted sum of the average of all correlations between station i and the other stations j, $\rho_{\mathrm{i,j}}$, and the fraction of time, $\rho_{\mathrm{i,j}}$, that data are missing for station i. The weight factor w is set arbitrarily by the user to give greater importance either to the length of the missing record or to the strength of the statistical dependence. In most cases, a value of 0.5 for w is reasonable. This statistic can be used to provide a relative ranking of all available data series, and helps to determine the order in which the missing data stations will be filled. Once the missing values for the first series in this list are generated, they are used as an independent station to fill the next series in the list, and the same holds for all other subsequent series – in other words, once a series is completed, it is added to the pool of independent variables from which the next series in the list is filled until all remaining series are completed. In general, the algorithm is data-driven, with only one user-defined parameter (a correlation threshold below which statistical dependence between stations i and j is deemed to be insignificant), and it proceeds through the following steps:

(1) Read and conduct initial processing of input data (a time series of available data, in which missing values can be designated as −999.0) and apply a user-supplied correlation threshold.

Next, calculate correlation matrices based on existing data and estimate the kernel density functions of the residuals based on all available data.

(2) Determine the data-filling sequence for all stations.

(3) Fill the missing data for the first station in the list.

(4) Check that the results of Step 3 comply with the historic statistical distribution and correlations to the previously-generated stations. If necessary, conduct iterative fine-tuning.

(5) Return to Step 3 to process the next station in the list, until all stations in the list have been completed.

3 Empirical distribution of the residual function

Figure 1 shows the residuals from an experimental simulation, which uses two typical correlated hydrologic series based on a standard regressive cross-correlative relationship between two stations in a medium-sized catchment. These series are modelled as:

$$y_t = b_t x_t + c_0 + \sigma_i \tag{3}$$

The empirical distribution in Fig. 1 is based on observed data. It can be closely represented by a kernel empirical distribution function. Originally defined by Parzen (1962), kernel density estimators are data driven non-parametric functions that define a probability density function. They eliminate the need to fit parameters of one of the known mathematical functions that are typically used in statistics to represent probability density functions. Mathematically, they take the form:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{4}$$

where $f(x)$ is the probability density function of random variable x, n is the sample size, and h is the bandwidth that would correspond to the size of the bin typically used when constructing probability density functions based on empirical data. Various functional forms have been proposed for K, and typically the Gaussian form is one of the most popular, based on the general assumption that the data points x within the $+/-\ h$ band from the x_i point are locally distributed normally, although the entire distribution of $f(x)$ may be asymmetrical. Theoretically, $h \to 0$ when $n \to \infty$, and the principal issue with kernel distribution is to determine the size of bandwidth h. Many empirical data driven formulas for h have been proposed, and Scott (1979) proposed an algorithm to define h based on minimizing the integrated mean square error of the estimated histogram and the actual data sample. The use of the kernel density estimator has been widespread in hydrology in the last two decades.

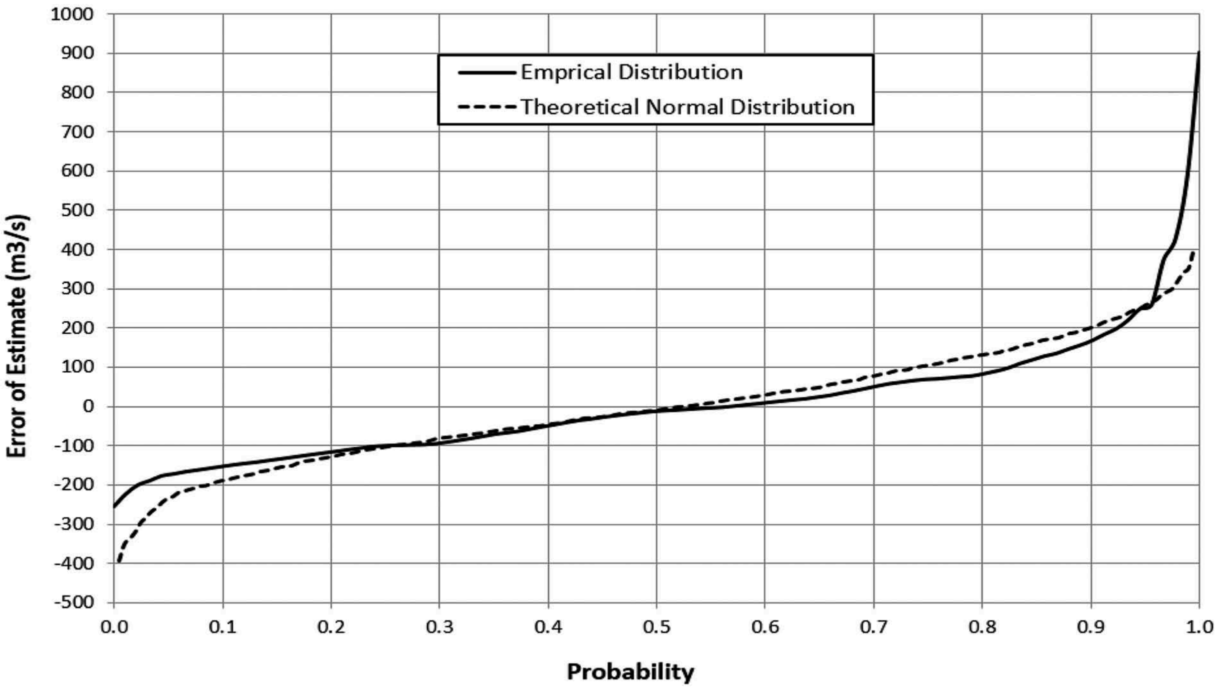

Figure 1. Empirical and theoretical normal distribution of regression residuals.

As stated by Sharma *et al.* (1997), empirical cumulative distribution plots can be approximated with a localized weighted moving-average of the available sorted data. In general, empirical distribution functions can be viewed as a close approximation of the observed data. However, a regression model shown in Equation (3) tends to skew the original distribution of the dependent variable y_t since the random term in regression is normally distributed. The graphs in Fig. 1 illustrate typical differences between the observed data and the data obtained using a standard regression with normally distributed residuals defined by Equation (3). This difference is visible in the lower 20 percentiles and in the upper 40 percentiles (probabilities higher than 0.6 as shown in Fig. 1) of the theoretical normal distribution curve, which demonstrates that normally distributed residuals negatively affect the correspondence of the distribution function of the simulated series to the empirical distribution. To preserve the target distribution function, one can develop a model that replaces random residuals based on normal distribution $f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$ with those from an empirical distribution function, *F*, such that the series is now calculated as:

$$y_t = b_t x_t + c_0 + F \qquad (5)$$

Due to its irregular shape and the inclusion of both positive and negative values, it is appropriate to model the function of residuals *F* using the kernel distribution as its empirical distribution. The residuals represent the differences in probabilities between the two curves depicted in Fig. 1. A kernel-type distribution is the preferred choice for the functional form of *F*, because of the unpredictability of its shape and the presence of negative values, and because it is guaranteed to fit the historic data in the probability range for which the data are available. Further, the use of this distribution typically eliminates the need to run the goodness-of-fit tests that are required for fitting theoretical statistical distributions.

When analysing the empirical residual functions for streamflow data, it is important to note that the empirical distribution of residuals depends on the data range. For the same station *i*, residuals for high flows would have a higher standard deviation and different distribution than for low flows. Rather than trying to develop a functional form that depicts the change in the range of confidence limits, in the first approximation it is possible to partition the residual error function into several distinct functions, each related to the target values of the dependent variable. We tested several subdivisions and found that division of the dependent variable into three segments, i.e. the lower 33 percentile, median (33–67 percentile), and the upper sub-set (67–100 percentile), provides excellent results. The corresponding residual cumulative distributions are shown in Fig. 2.

Figure 2 is based on a data sample obtained from two hydrometric stations on the Bow River near Calgary. This partitioning of the above regression model produces three equivalent models, based on

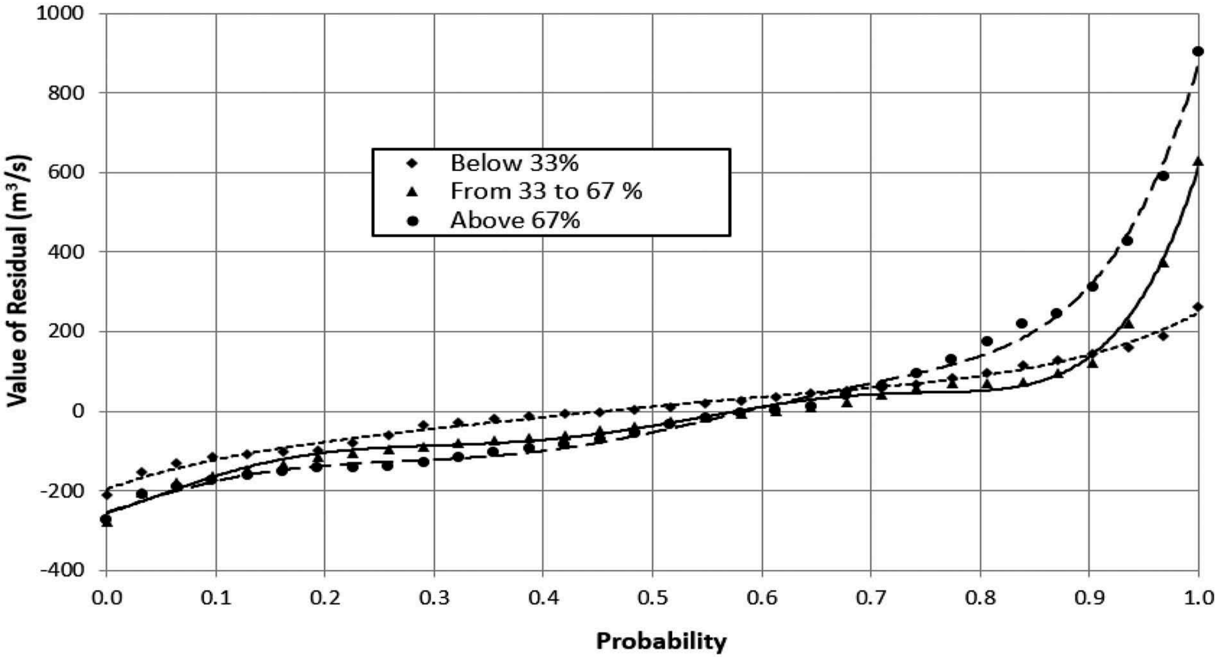

Figure 2. Breakdown of the residual function into three equivalent functions.

the expected range (represented as the probability, $P(x_t)$) of the predictor:

$$y_t = b_t x_t + c_0 + \varepsilon_i F_1, \text{where } P(x_t) < \frac{1}{3} \quad (6)$$

$$y_t = b_t x_t + c_0 + \varepsilon_i F_2, \text{where } \frac{1}{3} \leq P(x_t) \leq \frac{2}{3} \quad (7)$$

$$y_t = b_t x_t + c_0 + \varepsilon_i F_3, \text{where } P(x_t) > \frac{2}{3} \quad (8)$$

The benefits of model partitioning are as follows:

a. For the lower 33 percentile values of the predictor, the expected random variation (function F_1) is reduced compared to the higher ranges of the predictor. Function F_1 appears to be almost symmetrical, with a shape that closely resembles a normal distribution.

b. For median values of the predictor, the expected range of random variation is asymmetrical, with high positive values in the top 10 percentiles. These high values are more than double the range of the bottom 10 percentiles, which shows the gradual change of the random term as a function of predicted value y_t.

c. For the top 33 percentile values of the predictor, the residual shows a much higher likelihood of significantly overestimating (rather than underestimating) the regression target, and has a more pronounced asymmetry.

Note that the partitioning regression model is not limited to producing three sub-models; however, a set of experimental models divided into four and five similar partitions did not meaningfully improve our results. Sub-division of the predictor also need not be limited to partitions of equal size, and partition size is not critical to the functionality of the proposed algorithm. Even without subdivision of the regression model, the algorithm would function, but would require more iterative fine-tuning steps, as explained below. At this point, the results showed it was sufficient to divide the predictor variable into three sub-sets that correspond to low, moderate and high flows.

Figure 3 compares statistical distributions of the historic natural flow data points represented as dots with a standard regression model with a normally-distributed residual. The solid line in Fig. 3 was created by a model defined in Equation (3) where independent variable x_t represents flow at an upstream hydrometric station in reference to the downstream station y_t. It demonstrates the changes to the distribution function of historic natural flows imposed by the standard regression model with respect to the target defined by the historic data.

Figure 4 demonstrates that the change of the random term from Equation (3) to the terms in Equations (5), (6) and (7) provides a regression fit that is acceptable and that does not require any additional fine-tuning in most (over 90%) of the numerical experiments in this study. The generated frequency distribution functions of modeled series (solid lines in Figs. 3 and 4) are produced from samples of 1000 generated variables, using the Weibull plotting position formula.

Note that, although the refined model (Fig. 4) produces results that are much closer to the desired distribution function, it may still generate imperfect extreme high or low flow values. For example, the model could generate negative low-flow values for streams with small positive values in the historic record, and negative residuals could have an absolute value higher than the

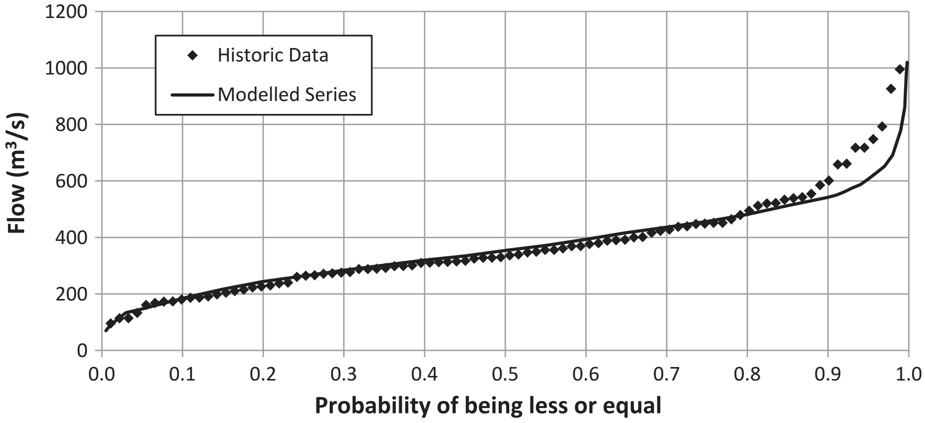

Figure 3. Standard regression term: historic and generated cumulative frequency distribution.

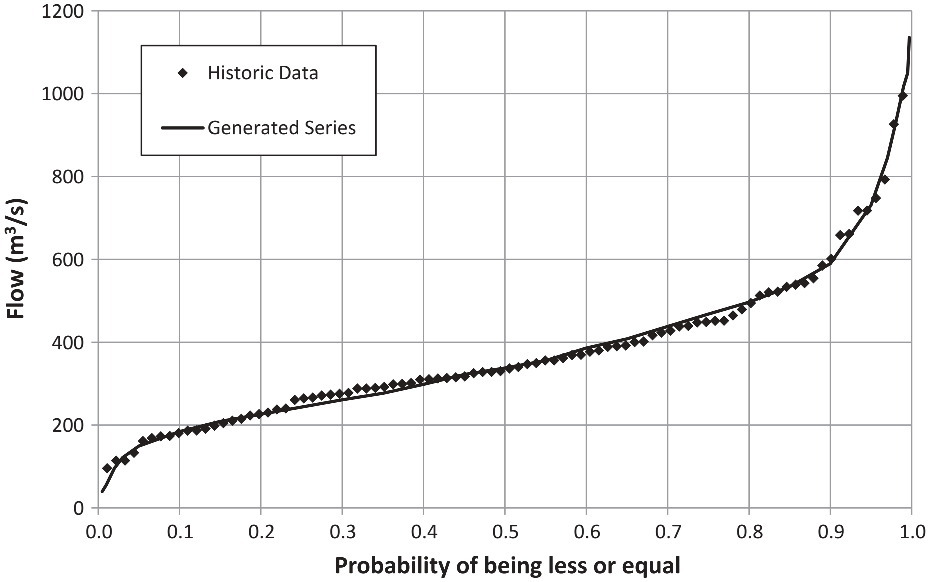

Figure 4. Regression model with a modified residual function: historic and generated cumulative frequency distribution.

predicted low-flow targets. Therefore, function F_1 has a smaller expected value than F_2 and F_3 for negative residuals, which minimizes the likelihood of high absolute-value residuals, but it does not eliminate it. To address this, the algorithm proceeds to fine-tuning steps (typically only one such step is sufficient) which was borrowed from earlier work (Ilich 2009), and which is briefly explained in the following section.

4 Preserving historical statistical distribution

Following the initial application of the regression model with partitioned residuals to generate the missing data series, the next step is to establish whether the dependent variable y_t generated by the proposed model follows the desired statistical distribution. The first indication that a correction is necessary is the presence of negative values within the y_t series. These negative values could have been avoided by using the logarithmic transformation of the original data and developing the regression model (4) using the transformed data. However, this approach does not guarantee that the reverse anti-log transformation required at the end of the process would preserve the original historic distribution of each data series with missing data.

Fitting historic data with statistical distribution of the complete series after the gaps were filled may include common statistical tests, such as the sum of least squares of the differences between the cumulative statistical distribution for the historic data and the completed series after data in-filling. If this difference is sufficiently small, missing data for the current variable y_t have been filled successfully and the algorithm proceeds to the next variable. Otherwise, an adjustment phase begins that may require one or more iterative steps. This adjustment applies an established algorithm, described in detail by Ilich (2009), that generates correlated random variables while retaining the desired marginal distribution – this is its first application to fill missing data. The algorithm is based on the theorem that two arrays have the highest possible correlation when their individual members are sorted in the same sequence, as originally postulated and proved by Whitt (1976). That approach was the basis of the first successful algorithm for generating correlated random variables with arbitrary statistical distributions (Iman and Conover 1982). Steps of the adjustment procedure are briefly outlined here; for a detailed description, see Ilich (2009). In general, this adjustment involves random generation of the entire target series such that it conforms with the desired statistical distribution and omits negative values. Further, some of its elements are used to remove biases associated with standard regression models mentioned earlier in the paper (i.e. undesirable positive threshold or occasional negative values). The algorithm consists of the following steps:

a. Fit the model coefficients described in Equations (1) and (4) with historic data to determine:
   - regression coefficients $b_{j,t}$ and c_j. in Equation (1). These are determined by first developing a correlation matrix calculated on the basis of the available historical data at all stations. The correlation matrix is then used as input for calculation of the regression matrix coefficients in Equation (1).

- Construct h_t in Equation (4) based on the available historic data, and determine the statistical distribution of variable y_t for all stations by taking into account the available historic data.
- Evaluate residual functions F_1, F_2 and F_3 for all stations that are identified as dependent (the ones that require in-filling of missing data).

b. Generate a complete series that fills the missing data for the first variable y_t by using the regression model as per Equations (6), (7) and (8). Check if the difference α between the regression coefficient r^2 of the generated series y_t and the target regression r_t^2 from the original historic series is within a prescribed tolerance limit (e.g. $\alpha = | r_t^2 - r^2 |/r_t^2 \le 0.05$). If so, and if there are not negative values generated within the y_t series, move to the next series with missing data, or exit if there are no more series to process.

c. If $\alpha > 0.05$ or if there were negative values generated in series y_t in Step 2, use the following correctional procedure: generate a desired number of flow data realizations of the random variable y_t', which follows the kernel based distribution function constructed using the available historic flow data and the extreme value distribution that fills the tail ends. The approach proposed by Moon *et al.* (1993) is used to combine the kernel based distribution with the theoretical extreme value distributions on each tail end, which are filled using extreme value type I and III distribution for high and low probability ends of the statistical distribution function. This step will ensure there are no negative values in the generated series and that the substitute series has a desirable statistical distribution. This step is executed only once, other steps will use different permutations of random variable y_t' until a satisfactory solution is found.

d. Determine the rank of the original data points in both series y_t and y_t' and replace the previously-estimated missing y_t data points with the y_t' data points that correspond to the same rank in both series, where rank refers to the ordered number in their sorted sequence. This will produce an update y_t^u to the missing values of y_t, where possible negative values are replaced by the small non-negative values obtained in Step 1 that fit the desired distribution function, without any material change to the correlation obtained in Step 1 based on the theorem postulated by Whitt (1976).

e. Check if the difference α between the regression coefficient r^2 of the updated series y_t^u and the historic series h_t is within a prescribed tolerance limit ($\alpha \le 0.05$). If so, the y_t^u solution is accepted and the algorithm moves to Step 1 for the next variable. If not, the random term F in the general model is modified such that the target correlations are increased for new estimates y_t'' by replacing the modified random term F with $(1-\alpha)F$, which would reduce the regression error term, and consequently increase the correlation coefficient:

$$y_t'' = \sum_{j=1}^{n} \mathrm{b}_{\mathrm{j,t}}\mathrm{x}_{\mathrm{j,t}} + \mathrm{c}_0 + (1-\alpha)F \qquad (9)$$

a. Replace the values of y_t' with a new y_t'' and repeat steps 4–6 until the convergence criterion $\alpha \le 0.05$ is satisfied. This will ensure compliance with both the target correlation structure, as well as with the desired statistical distribution. Normally, convergence is achieved in a single iteration.

The above algorithm has already been tested on a variety of statistical distributions (Ilich 2009), although its earlier applications in previous studies utilized standard normal distribution σ_i, of the residual function, which often required several iterations to achieve convergence. The use of data-driven residual functions F_1, F_2 and F_3 significantly simplifies this process and typically generates a close match to the desired statistical distribution fit after a single iteration.

5 Description of the test procedure and its application to the Bow River Basin, Alberta, Canada

A map of the Upper Bow River Basin with the relevant hydrometric stations is shown in Fig. 5. Most Bow River Basin runoff originates in the upstream portion of the basin, and the portion downstream from the confluence with the Highwood River is non-contributing. Therefore, only one station from the downstream section is included in this study; that station, at Crowfoot Creek (not shown in Fig. 5), represents only about 0.6% of the total annual Bow River Basin flow. Further, most flow series used as input in this study are naturalized flow records, where the effect of regulation was removed in previous studies. Flows for the Bow River at Banff are a sum of natural flows at

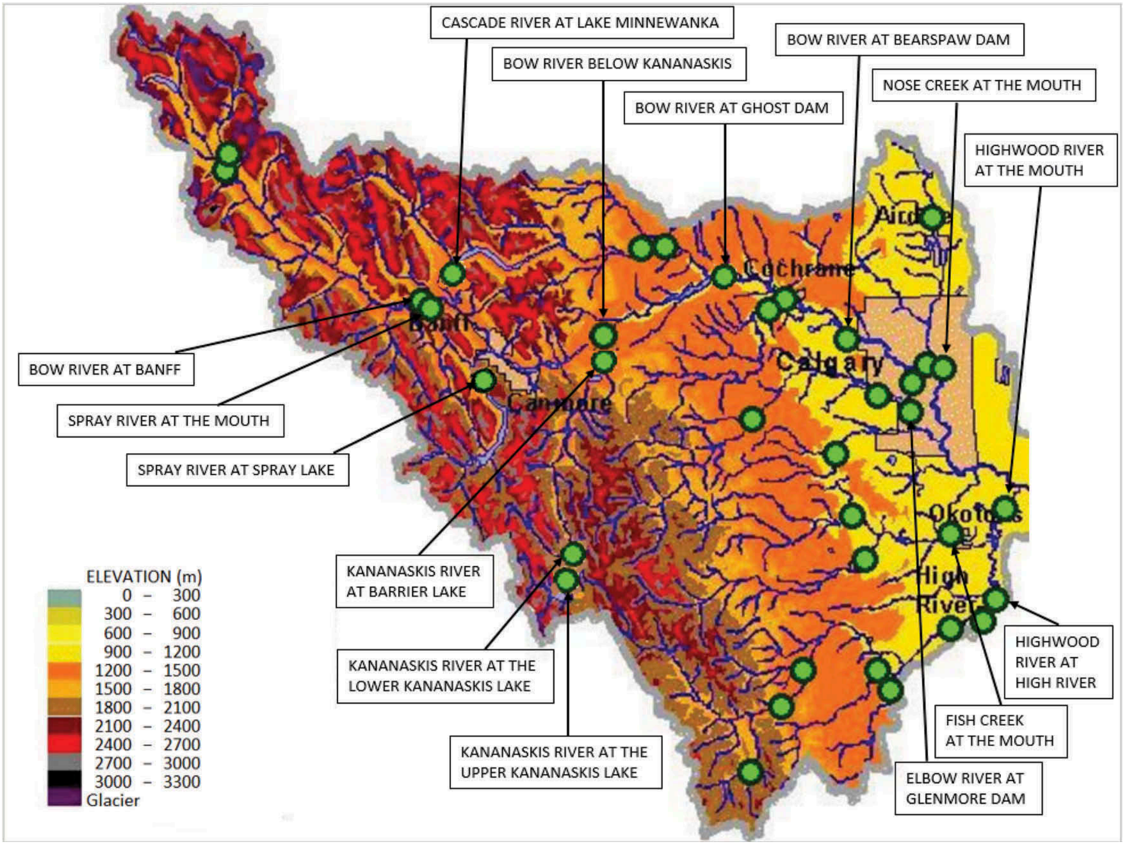
CASCADE RIVER AT LAKE MINNEWANKA
BOW RIVER BELOW KANANASKIS
BOW RIVER AT BEARSPAW DAM
NOSE CREEK AT THE MOUTH
BOW RIVER AT GHOST DAM
HIGHWOOD RIVER AT THE MOUTH
BOW RIVER AT BANFF
SPRAY RIVER AT THE MOUTH
SPRAY RIVER AT SPRAY LAKE
KANANASKIS RIVER AT BARRIER LAKE
KANANASKIS RIVER AT THE LOWER KANANASKIS LAKE
KANANASKIS RIVER AT THE UPPER KANANASKIS LAKE
HIGHWOOD RIVER AT HIGH RIVER
FISH CREEK AT THE MOUTH
ELBOW RIVER AT GLENMORE DAM
Aird
Cochrane
Banff
Calgary
Canmore
Okotoks
High River
ELEVATION (m)
0 – 300
300 – 600
600 – 900
900 – 1200
1200 – 1500
1500 – 1800
1800 – 2100
2100 – 2400
2400 – 2700
2700 – 3000
3000 – 3300
Glacier

Figure 5. Upper Bow River Basin with locations of selected hydrometric stations.

the Spray River tributary and the Bow River at Banff above the confluence with the Spray River.

The record length for the selected hydrometric stations varies both in terms of the starting date and the frequency and duration of missing data. Table 1 summarizes the data availability and shows that more than half of the stations have complete data between 1930 and 2014, in part because of earlier efforts by Alberta Environment and Parks (AEP, a Provincial Government Agency in charge of water resources management) to naturalize historic flow series, where sporadic missing data for some stations were filled by routing flows from upstream stations and adding estimates of local runoff. The main gaps are associated with different starting dates of hydrometric-station operation.

The accuracy of the data in-filling process can be ascertained by removing selected years from the historic series, generating data estimates using other stations where data are available, and then comparing the estimates with the actual historic record. Such comparisons should yield particularly close results if other nearby stations are highly correlated to the station where missing data were artificially created for test purposes. To this end, a cross-correlation matrix for data at all hydrometric stations is given in Table 2, where the station numbers are linked to the names in Table 1 in the same order of appearance. Since the approach is based on the statistical dependence of the selected station on data from other stations in the same basin, the quality of the results would be expected to be higher where a strong correlation exists between the generated data and the reference stations used for data in-filling. For example, correlation between the Bow

Table 1. Data availability in the Upper Bow River Basin.

Station no.	Station	Years with missing data in 1930–2014 period
1	Bow River at Bearspaw Reservoir	
2	Bow River at Ghost Dam	
3	Bow River below Kananaskis confluence	
4	Kananaskis River at Barrier Lake	
5	Bow River below Spray River confluence	
6	Lake Minnewanka Inflow	
7	Spray Lake Inflow	1931, 1932, 1939–1975
8	Upper Kananaskis Inflow	1930–1974
9	Lower Kananaskis Inflow	1930–1950
10	Fish Creek at the mouth*	1951–1955
11	Elbow River at Glenmore Dam	
12	Highwood River at High River*	
13	Highwood River at the mouth*	1930–1974
14	Crowfoot Creek at the mouth*	1930–1950
15	Nose Creek at the mouth*	

*Denotes tributaries without flow regulation

Table 2. Cross-correlation coefficients of historic data at all hydrometric stations.

	Stn 1	Stn 2	Stn 3	Stn 4	Stn 5	Stn 6	Stn 7	Stn 8	Stn 9	Stn 10	Stn 11	Stn 12	Stn 13	Stn 14	Stn 15
Stn 1	1.000	0.996	0.988	0.893	0.968	0.952	0.953	0.910	0.456	0.795	0.766	0.685	0.858	0.175	0.142
Stn 2		1.000	0.995	0.890	0.978	0.951	0.963	0.914	0.419	0.771	0.750	0.668	0.864	0.166	0.131
Stn 3			1.000	0.886	0.985	0.943	0.969	0.911	0.378	0.740	0.732	0.657	0.871	0.158	0.119
Stn 4				1.000	0.843	0.833	0.936	0.923	0.464	0.755	0.750	0.646	0.889	0.123	0.110
Stn 5					1.000	0.925	0.979	0.898	0.310	0.697	0.693	0.628	0.855	0.155	0.106
Stn 6						1.000	0.928	0.846	0.441	0.818	0.796	0.701	0.823	0.156	0.138
Stn 7							1.000	0.878	0.430	0.697	0.725	0.692	0.855	0.157	0.177
Stn 8								1.000	0.387	0.680	0.649	0.623	0.871	0.163	0.171
Stn 9									1.000	0.738	0.617	0.636	0.390	0.201	0.272
Stn 10										1.000	0.900	0.757	0.668	0.210	0.198
Stn 11											1.000	0.797	0.663	0.177	0.151
Stn 12												1.000	0.644	0.168	0.163
Stn 13													1.000	0.117	0.159
Stn 14														1.000	0.247
Stn 15															1.000

River at Bearspaw Dam in Calgary (Station 1) and at the Ghost Dam (Station 2) some 35 km upstream is very high, at 0.996, which results in good agreement between the historic and predicted (simulated) data when Station 1 data are used to in-fill missing values in the Station 2 record. Figure 6 compares the Bow River historic flows for the 2010–2012 period, with the flows generated under the assumption the 2010–2012 data were missing. It shows that the proposed model can predict missing flow data with reasonable accuracy for instances where data are available at adjacent stations that are highly correlated with the target station. A similar close agreement occurs when multiple stations are used to fill the missing data. Stations numbered 1–9 are located within the City of Calgary or upstream of it, while the remaining stations (10–14) are located downstream of the City of Calgary. Stations 14 and 15 have very small flows compared to other stations which exist only during the open flow season. Although only the positive flows were used to calculate correlations, these two stations have low correlations with other stations.

Further, in some instances of high correlation between the selected station and several other stations in the basin, historic data for certain years occasionally reveal odd patterns, which implies either a systematic error in the measurements or, more likely, uncertainties in the calculations for naturalized flows. An example of such a pattern is visible at the Spray River at Spray Lake (Station 7 in Table 2), where the historic series included flow naturalization based on historic lake levels and inflow data. In this case, uncertainties may be related to the lake level measurements or the absence of accurate precipitation data for this lake (the nearest meteorologic station is some 40 km away). Also, a large lake volume in combination with relatively small inflows implies that even small errors in recorded water levels can have significant effects on the accuracy of the calculated natural flows. The same test which assumed missing data for several selected years and then compared the results of predicted values to the actual historic record revealed excellent agreement in most years, but a probable anomaly in the recession limb of the 2011 hydrograph, as depicted in Fig. 7.

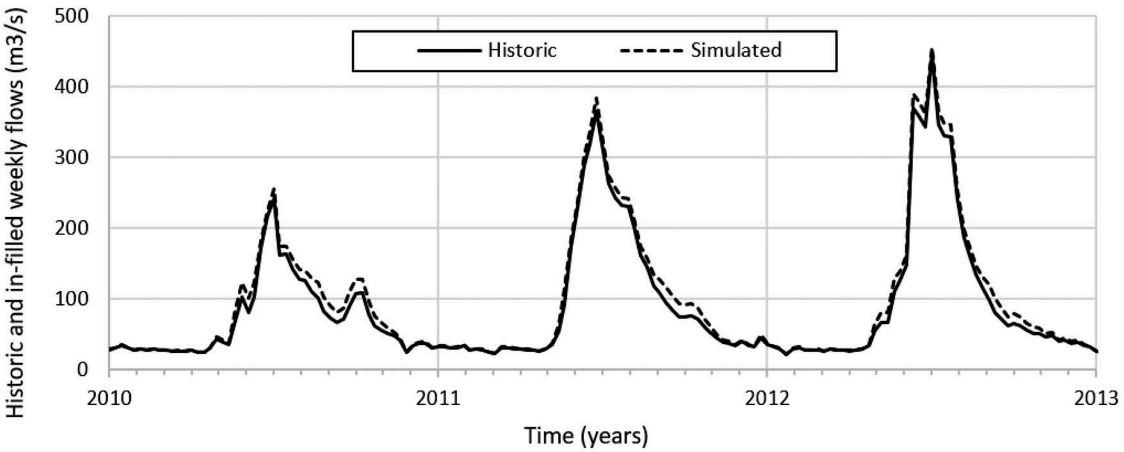

Figure 6. Bow River at Ghost Lake: historic vs simulated flows.

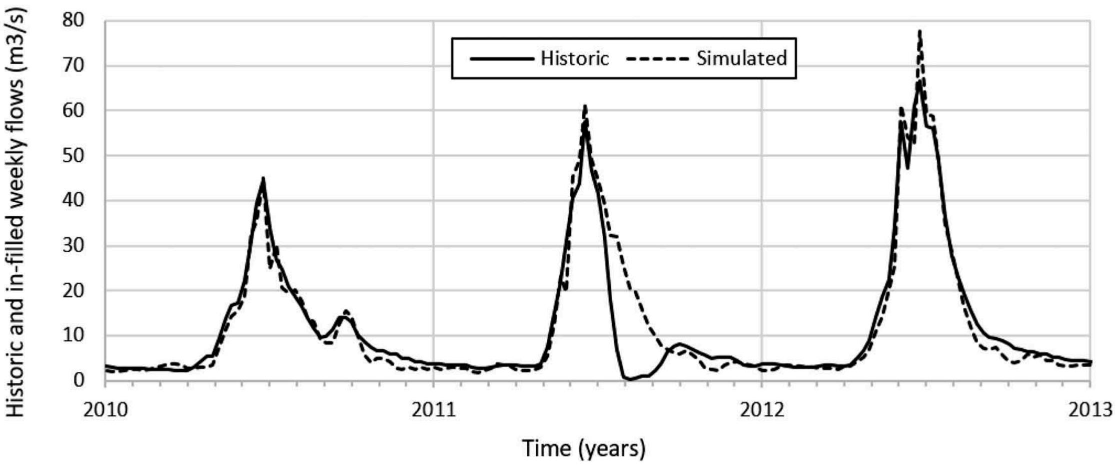

Figure 7. Spray River at Spray Lake: historic vs simulated flows.

Here, the historic natural flows in August of 2011 are close to zero in spite of all adjacent stations showing significantly higher flows. Typical flows in this period for other years are over 10 m^3/s. This result indicates a possible need to re-examine the lake balance calculations used to reconstruct the natural flows for 2011.

6 Extension to data generated with tree-ring proxy series

The proposed algorithm achieves the stated goals of (1) producing complete continuous series for each station with the same statistical distribution as the original series, while also (2) preserving other statistical properties such as cross-correlations and autocorrelations. This section illustrates its extension to the results of Sauchyn *et al.* (2014), and Sauchyn and Ilich (2017). Sauchyn *et al.* (2014) used tree-ring data as a proxy to develop annual flow estimates at several locations in Alberta, including the City of Calgary. Figure 8 shows the fit between historic and simulated annual flows for the City of Calgary obtained from a regression model that used statistical dependence between tree-ring data and annual flows over the past 100 years.

This statistical relationship used the tree-ring data to reconstruct annual flows back to the year 1111. However, because annual flows at a single site are of limited use, it was necessary to first break down this annual series into weekly values using a decomposition algorithm developed for that purpose (Sauchyn and Ilich 2017), and then to use the resulting weekly 1111–2014 flow series at the City of Calgary as a basis for in-filling missing data for the remaining 14 upstream locations, including both the main stem and the tributaries. This paper presents only the general conceptual approach used to decompose the annual flows using the tree-ring proxy.

The principal aim of the flow data extension is to develop a statistical relationship between the tree-ring thickness data and the annual flow volumes for the historic years when both are available, as shown in

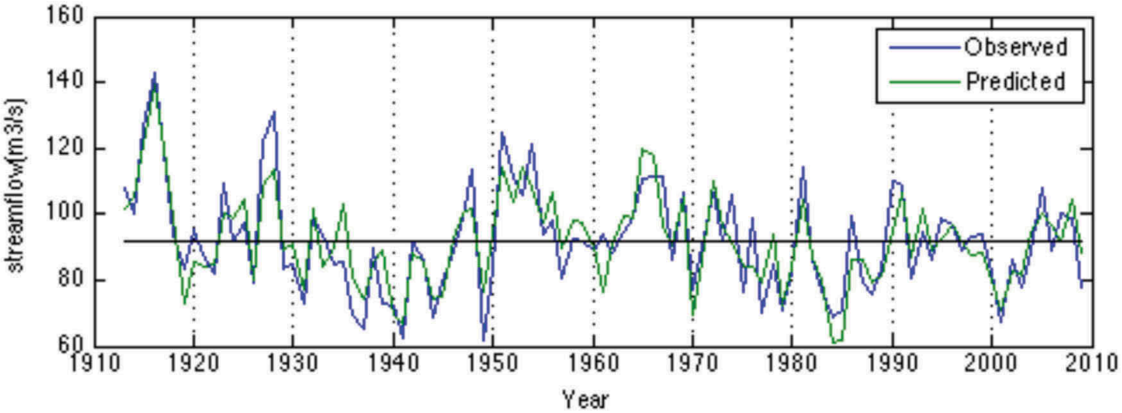

Figure 8. Annual flows at the Bow River at Calgary (observed and simulated using tree rings). Reprinted with permission from Sauchyn *et al.* (2014).

Fig. 8. This relationship can then be used to back-cast the sequence of annual flows based on a given tree-ring series, which may exceed 1000 years for some trees. Such annual series are of value because they contain information on the duration of multi-year droughts and wet periods that are typically more severe than those in a much-shorter historic record. Of particular interest are the duration and severity of dry year sequences that may be longer than those in the historic record.

As an example, after the extended annual flow estimates were generated from the tree-ring data, they were decomposed into weekly flows through a relatively simple boot-strapping method. Sauchyn and Ilich (2017) provide further information on the procedure, which is briefly summarized here:

(1) Generate a large number (typically 10,000 years) of hypothetical years of weekly flow series such that the weekly flows are mutually correlated from week 1 to week 52 and follow the same weekly flow distribution function. This step uses an approach described in previous publications (Ilich 2009, 2013).

(2) Select and re-arrange a subset of 904 years of data from the pool of all 10 000 generated years that satifies the following two criteria simultaneously:
   - The sum of all weekly flows are close to the annual flow volume predicted from the tree-ring data;
   - The correlations between the transitional weeks – those between the previous and current years – are similar to the historic correlations for the same transitional weeks (e.g. weeks 48–52 of year *i* – 1 should be correlated to weeks 1–4 of year *i* with similar correlation coefficients as in the historic series).

A number of verification statistics to verify success of the above procedure are presented in the referenced publication (Sauchyn and Ilich 2017). An example of the approach here uses estimates of weekly flows generated for the Bow River at Calgary from 1111 to 2014 as a basis to in-fill missing weekly flows for the other 14 stations to the same year, thus completing a weekly flow series for 15 stations over the 904-year period. Such long-term data should form a valuable input to various river basin planning studies, particularly because they provide more variability than the standard historic data series. The statistical results of this exercise are summarized in the following tables, which compare both cross-correlations and autocorrelations between the historic and the entire generated series. Close proximity of the cross-correlation coefficients between the historic (Table 2) and simulated (Table 3) series is demonstrated by comparing the entries in these two tables. This comparison indicates that the in-filled flow series have preserved the statistical correlation structure between the stations, which is one of the key goals of the algorithm. Other relevant statistics, such as the comparison of means and standard deviations, are shown in Table 4, while Tables 5 and 6 compare the autocorrelation of historic and simulated series for up to 10 weeks at all 15 stations. For instances where close to 100 years of historic data have reasonably close mean values to the means of the simulated annual flows based on the tree-ring data, a close correspondence is expected between the weekly statistics of the historic and in-filled series.

7 Conclusions and recommendations

This paper proposed a relatively simple and effective algorithm for the in-filling of missing hydrologic data, based on its statistical dependence with other available hydrologic series. Although essentially based on multiple regression, the algorithm removes typical shortcomings

Table 3. Cross-correlation coefficients of simulated data (1111–2014) at all hydrometric stations.

Sim.	Stn 1	Stn 2	Stn 3	Stn 4	Stn 5	Stn 6	Stn 7	Stn 8	Stn 9	Stn 10	Stn 11	Stn 12	Stn 13	Stn 14	Stn 15
Stn 1	1.000	0.995	0.992	0.953	0.983	0.975	0.936	0.925	0.495	0.806	0.821	0.776	0.871	0.196	0.121
Stn 2		1.000	0.997	0.950	0.988	0.974	0.939	0.927	0.479	0.795	0.803	0.752	0.870	0.192	0.118
Stn 3			1.000	0.944	0.993	0.967	0.941	0.923	0.451	0.772	0.790	0.740	0.872	0.185	0.115
Stn 4				1.000	0.923	0.939	0.896	0.938	0.509	0.814	0.821	0.775	0.858	0.185	0.113
Stn 5					1.000	0.954	0.937	0.906	0.431	0.749	0.767	0.717	0.860	0.183	0.113
Stn 6						1.000	0.922	0.897	0.508	0.840	0.842	0.796	0.848	0.185	0.115
Stn 7							1.000	0.906	0.397	0.728	0.758	0.721	0.865	0.173	0.112
Stn 8								1.000	0.538	0.755	0.767	0.745	0.881	0.198	0.126
Stn 9									1.000	0.709	0.619	0.607	0.493	0.232	0.140
Stn 10										1.000	0.904	0.864	0.705	0.234	0.133
Stn 11											1.000	0.915	0.729	0.209	0.121
Stn 12												1.000	0.707	0.212	0.118
Stn 13													1.000	0.169	0.113
Stn 14														1.000	0.510
Stn 15															1.000

Table 4. Comparison of relevant statistics of the historic and simulated series.

	Stn 1	Stn 2	Stn 3	Stn 4	Stn 5	Stn 6	Stn 7	Stn 8	Stn 9	Stn 10	Stn 11	Stn 12	Stn 13	Stn 14	Stn 15
							Historic data (1930–2014)								
Min	12.19	8.32	4.81	0.62	7.87	0.00	0.20	0.13	0.00	1.02	0.31	0.00	0.00	0.00	0.00
Avg	86.75	82.85	74.19	14.91	53.05	8.31	11.67	7.63	0.85	9.34	12.38	20.37	3.66	1.00	0.53
Max	1058.8	902.03	775.49	220.88	424.69	130.04	121.62	69.97	65.24	249.80	258.16	629.58	36.20	58.43	27.28
StDev	80.12	76.67	72.53	12.78	56.17	9.65	13.43	7.33	2.19	10.87	19.79	34.80	3.57	2.63	1.34
							Simulated data (1111–1929)								
Min	4.28	6.76	3.07	0.48	4.81	0.00	0.20	0.08	0.00	1.21	0.31	0.00	0.00	0.00	0.00
Avg	87.14	82.70	74.29	14.86	53.20	8.31	12.33	7.60	1.02	10.07	14.15	24.02	3.79	1.00	0.68
Max	1073.47	1023.88	925.40	231.98	557.62	164.42	152.31	88.61	80.44	327.12	293.70	803.64	48.39	74.65	35.68
StDev	86.00	77.82	74.11	12.92	57.44	9.83	13.85	7.54	2.46	11.09	19.83	35.03	3.58	2.60	1.44

Table 5. Historic autocorrelations for all stations (1930–2014).

Lag	Stn 1	Stn 2	Stn 3	Stn 4	Stn 5	Stn 6	Stn 7	Stn 8	Stn 9	Stn 10	Stn 11	Stn 12	Stn 13	Stn 14	Stn 15
1	0.90	0.91	0.91	0.86	0.92	0.86	0.84	0.90	0.44	0.75	0.86	0.81	0.81	0.44	0.81
2	0.81	0.81	0.82	0.75	0.83	0.74	0.78	0.82	0.36	0.63	0.71	0.67	0.69	0.17	0.61
3	0.70	0.71	0.71	0.64	0.73	0.64	0.72	0.74	0.28	0.53	0.59	0.55	0.58	0.12	0.49
4	0.60	0.61	0.61	0.53	0.63	0.53	0.65	0.65	0.21	0.42	0.46	0.43	0.48	0.06	0.38
5	0.49	0.50	0.49	0.41	0.52	0.43	0.58	0.55	0.14	0.33	0.34	0.32	0.38	0.03	0.31
6	0.39	0.39	0.39	0.31	0.41	0.32	0.52	0.46	0.13	0.26	0.24	0.24	0.28	0.03	0.26
7	0.29	0.29	0.28	0.21	0.31	0.23	0.45	0.35	0.11	0.20	0.15	0.16	0.18	0.04	0.22
8	0.20	0.20	0.19	0.13	0.21	0.16	0.40	0.26	0.10	0.15	0.08	0.10	0.09	0.03	0.20
9	0.12	0.12	0.11	0.06	0.13	0.09	0.35	0.18	0.06	0.10	0.02	0.05	0.02	0.02	0.18
10	0.05	0.05	0.04	0.00	0.05	0.03	0.31	0.11	0.03	0.06	–0.02	0.01	–0.04	0.02	0.16

Table 6. Autocorrelations of simulated series (1111–2014) for all stations.

Lag	Stn 1	Stn 2	Stn 3	Stn 4	Stn 5	Stn 6	Stn 7	Stn 8	Stn 9	Stn 10	Stn 11	Stn 12	Stn 13	Stn 14	Stn 15
1	0.90	0.90	0.90	0.82	0.90	0.85	0.87	0.81	0.56	0.73	0.76	0.70	0.85	0.52	0.75
2	0.80	0.81	0.80	0.71	0.81	0.74	0.76	0.73	0.27	0.55	0.57	0.51	0.70	0.15	0.47
3	0.70	0.71	0.71	0.61	0.72	0.63	0.66	0.64	0.18	0.42	0.44	0.38	0.58	0.07	0.29
4	0.59	0.61	0.60	0.51	0.61	0.53	0.56	0.54	0.14	0.34	0.34	0.29	0.49	0.06	0.18
5	0.49	0.50	0.50	0.41	0.51	0.43	0.46	0.45	0.10	0.26	0.25	0.21	0.39	0.04	0.12
6	0.39	0.40	0.40	0.32	0.41	0.33	0.36	0.35	0.06	0.19	0.18	0.15	0.30	0.03	0.08
7	0.29	0.31	0.30	0.24	0.31	0.24	0.27	0.27	0.04	0.14	0.12	0.10	0.22	0.03	0.06
8	0.21	0.22	0.22	0.17	0.23	0.17	0.19	0.19	0.02	0.09	0.08	0.06	0.15	0.04	0.05
9	0.14	0.15	0.14	0.11	0.15	0.10	0.11	0.12	0.01	0.06	0.04	0.03	0.09	0.03	0.04
10	0.07	0.07	0.07	0.05	0.07	0.04	0.05	0.06	0.00	0.03	0.01	0.01	0.03	0.02	0.03

of regression techniques associated with fixed thresholds and occasional negative values. Most importantly, the algorithm preserves the statistical distribution of the predicted series, as well as all other relevant statistical dependence among related hydrologic time series, as demonstrated by the numerical examples presented in this paper.

The proposed approach can be both robust and effective. It can also be combined with other models, such as one that generates weekly flow estimates that comply with annual tree-ring signals, so as to give lengthy hypothetical flow series that preserve statistical distributions of historic flows and their correlation structure. For the Bow River Basin, 904 years of weekly streamflows were in-filled at 14 stations based on a single station whose data were generated using tree-ring data and the available historic records.

The main shortcoming of the algorithm is that it cannot handle time periods that lack data points for all stations within a study region. However, this problem can be rectified through input data preparation that uses other existing methods to complete data for at least one dominant station that is highly correlated to other stations.

Acknowledgments

The authors would like to thank Alberta Innovates, the City of Calgary and the Bow River, Eastern, Raymond, St. Mary River, Taber and Western Irrigation Districts in Southern Alberta for their support.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Funding for this reaserch has been provided by Alberta Innovates - Technology Futures, the City of Calgary, several irrigation districts in Alberta named in the acknowledgments list, and Optimal Solutions Ltd. of Calgary.

ORCID

Evan G. R. Davies http://orcid.org/0000-0003-0536-333X

References

Efstratiadis, A., *et al.*, 2014. A multivariate stochastic model for the generation of synthetic time series at multiple time scales reproducing long-term persistence. *Environmental Modelling & Software*, 62, 139–152. doi:10.1016/j.envsoft.2014.08.017

Elshorbagy, A., Simonovic, S.P., and Panu, U.S., 2002. Estimation of missing streamflow data using principles of chaos theory. *Journal of Hydrology*, 255 (2002), 123–133. doi:10.1016/S0022-1694(01)00513-3

Gottschalk, L., *et al.*, 2015. Interpolation of monthly runoff along rivers applying empirical orthogonal functions: application to the Upper Magdalena River Colombia. *Journal of Hydrology*, 528 (2015), 177–191. doi:10.1016/j.jhydrol.2015.06.029

Gyau-Boakye, P. and Schultz, G.A., 1994. Filling gaps in runoff time series in West Africa. *Hydrological Sciences Journal*, 39 (6), 621–636. doi:10.1080/02626669409492784

Ilich, N., 2009. A matching algorithm for generation of correlated random variables with arbitrary distribution functions. *European Journal of Operational Research*, 192 (2), 468–478. doi:10.1016/j.ejor.2007.09.024

Ilich, N., 2013. An effective three-step algorithm for multi-site generation of weekly stochastic hydrologic time series. *Hydrological Sciences Journal*, 59 (1–2), 2014. doi:10.1080/02626667.2013.822643

Iman, R. and Conover, W., 1982. A distribution free approach to inducing a rank correlation among input variables. *Communications in Statistics – Simulation and Computation*, 11 (3), 311–334. doi:10.1080/03610918208812265

Moffat, A., *et al.*, 2007. Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes, Agric. *Forest Meteorology*. doi:10.1016/j.agrformet.2007.08.011

Moon, Y.U., Lall, U., and Bosworth, K., 1993. A comparison of tail probability estimators for flood frequency analyses. *Journal of Hydrology*, 151, 343–363. doi:10.1016/0022-1694(93)90242-2

Parzen, E., 1962. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33 (3), 1065–1076.

Sauchyn, D. and Ilich, N. 2017. Nine hundred years of weekly streamflows: stochastic downscaling of ensemble tree-ring reconstructions. *Water Resources Research*. doi:10.1002/2017WR021585, Wiley on-line library.

Sauchyn, D., *et al.*, 2014. Dendrohydrology in Canada's western interior and applications to water resource management. *Journal of Hydrology*. doi:10.1016/j.jhydrol.2014.11.049

Scott, D., 1979. Op optimal and data-based histograms. *Biometirka*, 66 (3), 605–615. doi:10.1093/biomet/66.3.605

Sharma, A., Tarboton, D.G., and Lall, U., 1997. Streamflow Simulation: a nonparametric approach. *Water Resources Research*, 33 (2), 291–308. doi:10.1029/96WR02839

Simonovic, S.P., 1995. Synthesizing missing streamflow records on several Manitoba streams using multiple non-linear standardized correlation analysis. *Hydrological Sciences Journal*, 40 (2), 183–203. doi:10.1080/02626669509491403

Tardivo, G. and Berti, A., 2013. The selection of predictors in a regression-based method for gap filling in daily temperature datasets. *International Journal of Climatology*, 34, 1311–1317. doi:10.1002/joc.3766

Tencaliec, P., *et al.*, 2015. Reconstruction of missing daily streamflow data using dynamic regression models. *Water Resources Research*, 51, 9447–9463. doi:10.1002/2015WR017399

Whitt, W., 1976. Bivariate distributions with given marginals. *The Annals of Statistics*, 4, 1280–1289. doi:10.1214/aos/1176343660