Multivariate and multi-scale generator based on non-parametric stochastic algorithms

Đurica Marković, Siniša Ilić, Dragutin Pavlović, Jasna Plavšić and Nesa Ilich

ABSTRACT

A method for generating combined multivariate time series at multiple locations and at different time scales is presented. The procedure is based on three steps: first, the Monte Carlo method generation of data with statistical properties as close as possible to the observed series; second, the rearrangement of the order of simulated data in the series to achieve target correlations; and third, the permutation of series for correlation adjustment between consecutive years. The method is non-parametric and retains, to a satisfactory degree, the properties of the observed time series at the selected simulation time scale and at coarser time scales. The new approach is tested on two case studies, where it is applied to the log-transformed streamflow and precipitation at weekly and monthly time scales. Special attention is given to the extrapolation of non-parametric cumulative frequency distributions in their tail zones. The results show a good agreement of stochastic properties between the simulated and observed data. For example, for one of the case studies, the average relative errors of the observed and simulated weekly precipitation and streamflow statistics (up to skewness coefficient) are in the range of 0.1–9.2% and 0–5.4%, respectively.

Key words | cross-correlation, hydrologic time series, non-parametric methods, serial correlation, stochastic data generation

Đurica Marković (corresponding author)
Siniša Ilić
Faculty of Technical Sciences,
University of Priština at Kosovska Mitrovica,
Kneza Miloša 7, 38220 Kosovska Mitrovica,
Serbia
E-mail: *djurica.markovic@pr.ac.rs*

Dragutin Pavlović
Jasna Plavšić
Faculty of Civil Engineering,
University of Belgrade,
P.O. Box 42, 11120 Belgrade,
Serbia

Nesa Ilich
Optimal Solutions Ltd,
7128-5 Street NW, Calgary, AB T2 K 1C8,
Canada

INTRODUCTION

Long hydrologic time series are required for effective water resources system planning, design, and operation. However, those are often too short, unreliable, or non-existent. In these situations, various methods can be used for generating synthetic time series of sufficient length with richer regimes (e.g., containing more extreme values compared to those found in short observed series), while keeping the existing statistical properties of the original series intact. The majority of these methods have been used to generate a single type of time series, e.g., streamflow, precipitation, or temperature. More recently, there are examples of novel stochastic simulation methods capable of dealing with multivariate stationary or cyclo-stationary processes of any time scale with any marginal distribution and correlation structure. Some of these methods are based on a non-parametric approach (Ilich 2014; Srivastav & Simonovic 2015), and some are based on a parametric approach (Tsoukalas *et al.* 2018a, 2018b; Kossieris *et al.* 2019).

Hazen (1914) was probably the first to use the notion of synthetic time series in hydrology. He generated a 300-year long synthetic hydrologic series combining data from 14 watercourses. Since then, many other approaches emerged for a hydrologic series generation.

Generating a time-dependent hydrologic series is much more complex than generating independent series since its goal is to preserve not only the statistical distribution function of the original sample but also its autocorrelation function for all significant lags. The well-known model by

doi: 10.2166/hydro.2019.071

Thomas & Fiering (1962) with serially correlated flows was the first model of this kind used for monthly flow generation at a single site. This type of model reproduces the essential statistical characteristics of the series but may lead to unrealistic dependence patterns when combined with non-Gaussian white noise (Tsoukalas *et al.* 2018c). The problem becomes more difficult for the multivariate and/or multisite generation (e.g., streamflows at multiple gauging stations or streamflows and precipitation) where the interstation dependence (i.e., cross-correlation) has also to be preserved in the generated series. The first multisite stochastic flow generation model was developed by Fiering (1964).

A number of models for stochastic hydrological time series generation are based on a stochastic processes approach, such as the autoregressive moving average (ARMA) models (Box & Jenkins 1970). Despite the advantages of the autoregressive and the moving average group of models (including ARMA and autoregressive integrated moving average (ARIMA)), they suffer from the 'short memory' problems, meaning that the serial correlation function quickly diminishes with the time lag (Koutsoyiannis 2000). This approach involves the simultaneous fitting of a large number of parameters related to the joint marginal probability distribution functions in order to comply with the spatial and temporal covariance structure of the shorter historic time series. A detailed review of the application of the Box–Jenkins approach in hydrology is presented by Salas *et al.* (1980). However, recent papers (Tsoukalas *et al.* 2018a, 2018b) introduce methods that can be used for the simulation of non-Gaussian univariate and multivariate stationary processes capable of preserving any correlation structure and marginal distribution at any scale. The method was also applied for simulating a non-physical process (water demand) at fine scales from 1 h up to 1 min (Kossieris *et al.* 2019).

In the last 20 years, many authors have developed the non-parametric methods for simulating hydrological processes. This became possible by the emergence of new mathematical procedures and methods, and the advances in computational power and software tools. The methods mostly used are the moving block bootstrap (Srinivas & Srinivasan 2005), K-nearest neighbour (K-NN; Sharif & Burn 2006, 2007), or kernel-based methods (Sharma *et al.* 1997). The main advantage of these methods is that they do not rely on the parameter estimates, while they suffer from the inability to extrapolate the probability distribution beyond the observed data.

Multisite streamflow series generation requires a stochastic model capable of reproducing the relevant statistical characteristics of the observed data series. Ideally, the model should be capable of working with selected time discretization (e.g., day, week, or month) and also preserve the key statistical characteristics at coarser time scales (e.g., annual). Furthermore, it should be able to extrapolate sensibly the distribution tails for a particular time discretization. Finally, the model also needs to preserve the serial and cross-correlation structure for each time scale, as well as the intra-annual cycle. All these requirements were discussed in detail by Moran (1970), Salas *et al.* (1980), Koutsoyiannis (2005), and Srinivas & Srinivasan (2005).

Stochastic methods are also used for generating precipitation time series. As precipitation is generally modelled as an intermittent stochastic process, the models need to simulate both precipitation occurrence and intensities/depths in time. Compared to streamflow generation methods, they have to reproduce additional observed data characteristics, such as precipitation occurrence, duration, or the distribution of consecutive wet and dry days. Modelling intensities/depths in stochastic precipitation models is identical to modelling streamflow distributions. For the occurrence of dry and wet spells, two types of models are commonly used: Markov chain or renewal process based (Wilks & Wilby 1999). Those based on the Markov chains are often used to specify the state of each spell as wet or dry. These models have been applied to data from various climatic regions and series lengths; however, the structure of the model has to be adjusted to the local conditions for each case study. In addition to the above-mentioned two-part models, resampling models, transition probability matrix models, and modifications of ARMA-type models (e.g., using normalization transformations or non-Gaussian white noise) are also used for generating rainfall (Srikanthan & McMahon 2001). A good review of the topic is given by Srikanthan & McMahon (2001), Haberlandt *et al.* (2011), and Serinaldi & Kilsby (2014).

Harrold *et al.* (2003a, 2003b) used the non-parametric approach for modelling single-site daily rainfall occurrences and rainfall amounts for a 140-year long rainfall record at

Sydney, Australia. Their rainfall simulation model is based on the K-NN resampling method, where the Markov model was used to generate sequences of dry and wet states (Harrold *et al.* 2003a). The model preserves short- and long-term time series characteristics, i.e., seasonal, annual, and multi-annual properties of observed data series. Mehrotra *et al.* (2006) also applied a multisite K-NN model for precipitation generation at 30 stations in Australia, along with other two parametric generators, while Basinger *et al.* (2010) used a non-parametric procedure based on bootstrapped Markov chains for precipitation occurrence and resampling from observed data for precipitation amounts.

In addition to the methods for generating single-variate hydrometeorological series, there is a need to develop approaches for generating multivariate series. Such a stochastic model is developed by Srivastav & Simonovic (2014, 2015); this model uses the maximum entropy principle and the bootstrap method to generate multiple variables at multiple sites. It reproduces data statistics, keeping the spatial and temporal structure of data interdependence. The bootstrap method is implemented through the K-NN approach for data generation. The model is tested on daily data (precipitation, maximum, and minimum air temperature) from 22 gauging stations in the Thames River catchment (Ontario, Canada). However, the method does not preserve the serial correlation between two consecutive years.

Unlike in some water resources areas related to modelling, such as river hydraulics, where there are universally accepted modelling tools, such as HEC-RAS, there is no similar tool in stochastic hydrology, i.e., there is no universally accepted multivariate time series generation model for simultaneous modelling of flows and precipitation that is widely used by hydrologists around the world. In the works by Ilich & Despotovic (2008), Ilich (2014), and Marković *et al.* (2015), a different approach to the generation of stochastic streamflow series is developed that presents an essential departure from the previously established methods. The proposed method consists of three steps: (1) independent data sets for the given time step are generated using the Monte Carlo method, in which the statistical distribution functions of the observed series are fully maintained, (2) data from the individual data sets are then rearranged to induce serial and cross-correlation coefficients of the observed series, and (3) annual streamflows are rearranged to adjust their serial correlation for time intervals that cross-connect two consecutive years. Such an approach has not been proposed by other studies. Moreover, to our best knowledge, other approaches do not deal explicitly with correlation between data in the transition from one year to another, which is, in our methodology, done in Step 3 by re-ordering whole years in the generated weekly/monthly series. Ilich & Despotovic (2008) have applied this methodology to weekly streamflows. Ilich (2014) has introduced the intermittent precipitation series along with the continuous weekly streamflow series in the simulation procedure. Marković *et al.* (2015) made further improvements in order to enhance the method's performance by employing the logarithmic transformation to data in order to reduce skewness coefficient and the effect of outliers and by including additional control to simulate the persistence of extremely low summer and autumn flows in dry years.

This paper builds on the previous work of Ilich & Despotovic (2008), Ilich (2014), and Marković *et al.* (2015) by expanding the methodology for combined generation of streamflow and precipitation time series. The improvement of the methodology lies in introducing a new method for the extrapolation of distribution tails, which is different from the use of parametric distributions in Ilich & Despotovic (2008) and Ilich (2014). The main advantages of the proposed methodology are: (1) starting from the shortest time step considered, the methodology ensures that statistics are preserved for all larger steps, (2) the method preserves the serial correlation between two consecutive years, (3) both continuous and intermittent time series can be generated, and (4) the procedure is completely automated with a set of default agreement criteria. The application of the methodology in this paper includes streamflow and precipitation data in Canada and Serbia, but the method can be used for any combination of hydrologic and/or weather variables. While Marković *et al.* (2015) generated streamflow data for both Canada and Serbia, in this paper streamflow and precipitation data are jointly generated. By comparing these two sets of results, the efficiency of the generating algorithm is evaluated in terms of multivariate applications.

The next section gives an overview of the proposed methodology. It is followed by its application to two data sets of weekly flows, one from Serbia (three hydrologic

stations and one meteorological station) and one from Canada (seven hydrologic stations and four meteorological stations). The last section provides discussion and conclusions with recommendations for further improvements.

METHODOLOGY

Hydrological time series represent continuous natural processes and are defined in practical applications in a discrete form of average flows or total precipitation for a selected time scale, such as day, week, or month. They are modelled as stochastic processes characterized by probability distributions and low-order summary statistics (i.e., mean, variance, and skewness coefficient), and correlation structures.

The non-parametric stochastic generation method used in this study is formulated so as to respect the principle that the generated synthetic series should have distribution functions and a correlation structure very similar to those of the observed series. In order to achieve this, statistics such as the mean, standard deviation, and skew at each time step should be preserved in the generated series, and the serial and cross-correlations should match the observed for any significant lag. Annual statistics of the simulated series, such as the annual mean, standard deviation, and serial and cross-correlations, should also match the annual statistics of the observed series.

The procedure of stochastic streamflow generation relies on the assumptions that observed data represent the natural hydrologic regime. This means that it should be free from any effects of regulation, such as an upstream reservoir operation or diversion structures, and that the observed process at each time step has a unique statistical distribution that should be matched in the simulated series. This distribution function can be represented either by a theoretical parametric distribution that fits the data well or by using an empirical distribution, such as the non-parametric kernel-based distributions. The reason for using the non-parametric probability distributions is to avoid specifying any particular parametric distribution in the data generation process. A possible probability distribution model can be based on combining the non-parametric approach within the range of the observed data with a parametric distribution at tails, with smoothed inter-range transitions (Ilich 2014).

The observed data that represent the input for the generation procedure are organized in a matrix $\mathbf{X}$, as shown in Figure SM1 in the Supplementary Material. The number of rows in the matrix is equal to the number of years n in the record. This matrix consists of K blocks of columns for each of the K stations considered. If, for example, weekly data are considered, each column in a block contains streamflow series for 1 week. For K stations, the total number of variables, i.e., columns in matrix $\mathbf{X}$, is $M = 52\,K$ for weekly data or $M = 12\,K$ for monthly data. Thus, the matrix $\mathbf{X}$ is given with:

$$\mathbf{X} = [x_{ij}], \qquad i = 1, 2, \ldots, n, \qquad j = 1, 2, \ldots, M \tag{1}$$

The columns X_j $(j = 1, 2, \ldots, M)$ of the matrix $\mathbf{X}$ represent the series for each selected time step:

$$X_j = [x_{ij}], \qquad i = 1, 2, \ldots, n \tag{2}$$

For the given input matrix $\mathbf{X}$, the correlation matrix $\mathbf{C}$ of size $M \times M$ contains correlation coefficients ρ_{ij} between two columns X_i and X_j (see Figure SM2 in the Supplementary Material):

$$\rho_{ij} = \mathrm{Corr}(X_i, \quad X_j), \qquad i, \quad j = 1, 2, \ldots, \quad M \tag{3}$$

Diagonal elements ρ_{ij}, when $i = j$, are equal to 1. Non-diagonal elements of matrix $\mathbf{C}$ represent either serial correlation coefficients (for a single station) or cross-correlation coefficients (interstation dependence). For example, for weekly data, $\rho_{1,20}$ is the serial correlation between flows in 1st and 20th week at station 1, while $\rho_{2,72}$ is the cross-correlation between the 2nd week flow at station 1 and the 20th week flow at station 2.

The three steps of the proposed procedure for data generation are described in the sequel providing a basic theoretical background for each step (the full details are presented in Marković et al. (2015)) and using the pseudo-codes to clarify the method. The procedure is described for weekly flows, but it is equally valid for other temporal discretizations.

Step 1 – generation of independent data sets

The first step is to generate N years of random weekly data having the statistical distributions for each time step as close as possible to the target statistics of the historical series

represented in the matrix $\mathbf{X}$ of size $n \times M$, where n is the number of years of the observed data and M is the total number of columns (i.e., $M = 52\,K$ for K stations). This step includes compiling the observed series and their log-transformation (to mitigate the skewness intrinsic in the data), defining the target statistics (observed mean value, standard deviation, and coefficient of skewness) in the log-space for each week at every station, and then running the Monte Carlo procedure for generating data from the observed distributions. In order to avoid the logarithm of zero, log-transformation of zero precipitation is increased by a constant of 1 mm. For basins exhibiting zero flows, the same would be applied. Generated data from this step are stored in the resulting matrix $\mathbf{G}$ of the generated independent data sets, which has M columns and N rows (in our study, $N = 1{,}000$), but in general N can be as large as necessary.

The probability distributions of the observed data for each week are defined using the non-parametric kernel approach combined with an extrapolation algorithm for the distribution tails. The advantage of the non-parametric approach is that it lends itself to a completely automatic procedure, which is a desirable feature. However, the non-parametric kernel distributions perform poorly outside the range of the observed data. The idea for distribution function extrapolation in the tail sections in this article originates from the work of Scholz (1995). This extrapolation method linearizes distribution tails by utilizing linear dependence of a variate (e.g., streamflow) on the standard variate of a theoretical distribution when plotted on a probability paper. However, depending on the sample data and the existence of outliers, extrapolating the lower tail could produce negative values, while extrapolating the upper tail could yield generated values much greater than the maximum observed value. Both cases are undesirable.

To overcome the drawbacks of Scholz's approach, a different heuristic algorithm is applied here for extrapolating the distribution tails. The linear extrapolation is applied to the log-transformed variate $Y = \ln X$ plotted against the standard normal variate z (Figure 1). The developed algorithm assumes that the upper and lower tail extrapolating lines must lie within the confidence interval of the observed distribution. However, the confidence interval of the non-parametric distribution function

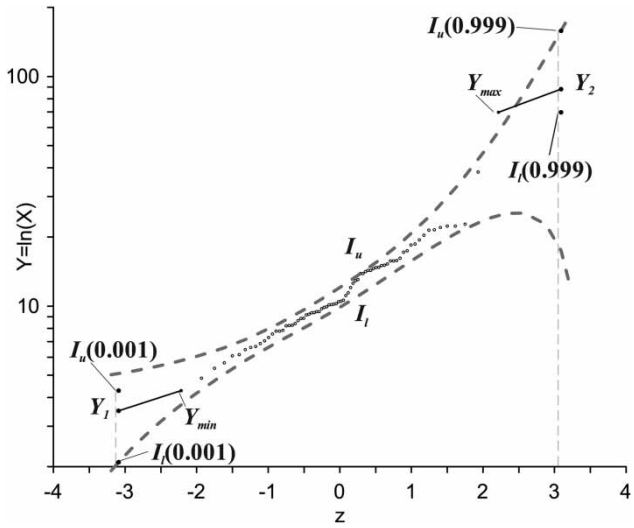

Figure 1 | Extrapolation of distribution tails applied in the model; the observed $Y_{\min}$ and $Y_{\max}$ are connected to randomly selected points (crosses) from the 90% confidence interval of 0.1% and 99.9% GEV quantiles.

cannot be constructed outside of the observed data range. For this reason, the confidence interval limits outside of the observed range are estimated by assuming the general extreme value (GEV) distribution. The GEV parameters are estimated by the method of L-moments for each observed weekly series according to the formulae given by Rao & Hamed (2000). Each extrapolation line is determined by two points (Figure 1). At the lower tail, the first point is defined by the log-transformed minimum observed value $Y_{\min}$, and the second point is a randomly selected value Y_1 from the 90% confidence interval of the 0.1% GEV quantile (i.e., for the cumulative distribution function (CDF) or CDF value of 0.1% or standard normal variate $z = -3.09$). Similarly, at the upper tail, the first point is the log-transformed maximum observed value $Y_{\max}$, and the second point is a randomly selected value Y_2 from the 90% confidence interval of the 99.9% GEV quantile (with $z = 3.09$). The 'randomness' of the choice of the points Y_1 and Y_2 from the 90% confidence interval (I_l, I_u) is restricted by the logical conditions: Y_1 cannot be greater than $Y_{\min}$ and Y_2 cannot be smaller than $Y_{\max}$. These constraints can be formalized as:

$$I_l(0.001) < Y_1 < \min\{Y_{\min}, \quad I_u(0.001)\} \tag{4}$$

$$\max\{Y_{\max}, \quad I_l(0.999)\} < Y_2 < I_u(0.999) \tag{5}$$

The random values Y_1 and Y_2 are obtained from the restricted ranges given in Equations (4) and (5) by multiplying the range span by the uniformly distributed random number from the [0,1] interval and by adding the product to the lower range limit $I_l(0.999)$ at the upper tail, or subtracting it from the upper limit $I_u(0.001)$ at the lower tail. The outermost points (i.e., selected GEV quantiles) are linearly connected to the log-transformed minimum and maximum observed values $Y_{\min}$ and $Y_{\max}$. These linear dependencies on the log-normal probability plot are used for random sampling outside the observed data range (as shown in Figure 1).

The algorithm for Step 1 is presented by the pseudo-code for Step 1, which generates M data vectors by random sampling from the non-parametric distributions of the observed vectors using the pre-set criteria for agreement of the observed and simulated data statistics (mean, variance, and skewness). The generation process ends when the generated statistics are close enough to the observed ones, as defined by specified criteria for each statistic. We have chosen to restrain the error in mean logarithmic flows to 0.001 (corresponding to an error of 0.1% in the original data space) for generating the first 10,000 data. If a desired mean value is not obtained from the first 10,000 data, the tolerance limit is relaxed to 0.003. Similarly, the tolerance limit in the skew of the logarithmic flows is 0.03 or 0.05 after 10,000 simulations.

Algorithm 1: Pseudo-code for Step 1 – generating random data vectors

1: Load n years of historical data in matrix **X** (size $n \times M$)
2: Pre-process **X** (perform logarithmic transformation and sort each column independently) and store in matrix **XLS**
3: Initialize output matrix **G** (size $N \times M$) for N years of generated uncorrelated data ($g_{ij} = 0$)
4: **for** $j = 1$ to M // for each column vector j in **XLS**
5: calculate the observed ratio of zero values p_0 in column j
6: calculate target statistics for non-zero values
7: define target CDF by calculating observed non-parametric CDF and extrapolating the tails from data in column j
8: **for** $i = 1$ to N // for each year i to be generated
9: generate random number u from the range (0, 1)
10: **if** ($u < p_0$) **then** $g_{ij} = 0$ **else** $g_{ij} =$ inverse of target CDF for u
11: end **for** (next i) // generated data vector G_j created
12: calculate statistics for non-zero values in the generated vector G_j
13: **if** statistics for G_j match target statistics **then** continue to line 23 (next j);
14: **else**
15: set new count $k = 1$; set maximum number of iterations
16: **while** statistics do not match target statistics **or** maximum number of iterations is reached
17: generate new data value gg (like in lines 9–11)
18: create trial data vector G_j by replacing g_{kj} by gg
19: calculate new statistics for non-zero values in trial data vector G_j
20: **if** new statistics are better than old statistics **then** approve a replacement on k^{th} position and set the new statistics is target statistics
21: **else** continue generation process with $k = k + 1$ (go to line 16)
22: end **while**
23: end **for** (next j)
24: post-process data in **G** from log-transformed to original data space

If the generated series does not fulfil the specified criteria after the first N simulations, the algorithm would continue to generate the $(N+1)$st data value and to evaluate statistics of the series in the range $[2, N+1]$ by comparing it to those of the series in the range $[1, N]$. The process of generating one additional data value and sequential comparison of updated generated statistics with the observed ones is continued for each data vector until the specified criterion is met. At the end of the process, N years of log-transformed data are generated for each data vector, having the marginal distribution that corresponds to that of the observed vector. The generated series are then transformed back from the log space to the original data space and stored in matrix **G**.

Generating precipitation data takes into account that precipitation is an intermittent process and that the CDF $F(x)$ of a precipitation vector consists of two parts: probability p_0 of zero precipitation in one time interval (i.e., dry interval occurrence) and the conditional CDF of precipitation depth during the wet interval $F_1(x)$ weighted by the

wet interval probability ($1–p_0$):

$$F(x) = p_0 + (1 - \ p_0) \cdot F_1(x) \quad (6)$$

Generating precipitation data, therefore, has two stages: (1) assessing the dry interval probability p_0 and the distribution of precipitation depths in wet intervals $F_1(x)$ from the observed data and (2) random sampling of precipitation depths by sampling a random number u from the uniform [0,1] distribution, evaluating F_1 that satisfies Equation (6) for $F(x) = u$, and finally estimating the corresponding precipitation depth quantile as $x_u = F_1^{-1}[(u - \ p_0)/(1 - \ p_0)]$. The remaining procedure is identical to generating streamflow data.

Step 2 – adjusting the correlation structure of the generated series

The data vectors generated in Step 1 for each week or month represent uncorrelated streamflow or precipitation series, but they should also have the appropriate correlation structure of the observed series in order to describe realistically the natural hydrologic or precipitation regime at given locations. The correlation structure includes the serial correlation between weekly and monthly data at each site and cross-correlation between the sites. In the case of streamflows, it is also important that the persistence of low flows within an extremely dry year is maintained in the generated time series, leading to the occurrence of extremely low annual flow.

The algorithm for Step 2 is divided into two parts. The first part deals with data rearrangement to match the correlation of the observed weekly data (Algorithm 2.1), while the second part serves two purposes: it improves the fit between the distributions of the observed and generated annual minima and allows user to control the fraction of extremely dry years in the generated data set (Algorithm 2.2).

Algorithm 2.1: Pseudo-code for the first part of Step 2 – adjusting the correlation structure

```
load matrices X and G from Step 1
for M column vectors in matrix X, calculate correlation matrix C (of size M × M) end for
if C is not a positive definite matrix, then calculate the closest positive definite matrix and store it in C
apply the Iman–Conover method to rearrange elements in G with a correlation matrix closest to C
```

Algorithm 2.2: Pseudo-code for the second part of Step 2 – adjusting extremely dry years

```
set the number n_d of extremely dry years in generated data
for each station
  create vector AO of annual sums of observed weekly data in X for n years
  create vector AG of annual sums of generated weekly data in G for N years
  find the smallest value AO_min in AO
  find the smallest n_d values in AG and their positions IAG
    for i = 1 to n_d // perform a loop with respect to indices IAG in G
    while AG(IAG_i) > AO_min // while generated annual sum in row IAG_i is greater than the minimum observed annual sum
      find column j with the maximum value in row IAG_i of G and store data cell position pos1
      find row k with the minimum value in column j of G and store data cell position pos2
      swap the values between positions pos1 and pos2
      recalculate AG(IAG_i)
    end while
  end for (next i)
end for (next station)
```

In the first part of Step 2, the algorithm of Iman and Conover (ICA) (Iman & Conover 1982) is used for data permutations within the generated vectors to achieve target correlation structure. The matrix **G** resulting from Step 1 is the input for the algorithm, and its columns are the series to be rearranged. The observed data matrix **X** is used here to calculate the observed correlation matrix **C**, which is set as a target correlation matrix for ICA. The ICA application was presented in detail in Marković *et al.* (2015).

Considering that the purpose of the proposed stochastic method is to provide input for the optimal design of reservoir storage and/or optimal reservoir operation, it is important that the generated series covers a wide range of

input data and includes events that could be critical for reservoir operation, such as long droughts. These events from the lower or the upper tail of flow distributions are not present in the observed series but are expected to emerge within N years, which is usually much greater than the number of years with observations. The critical events are very wet or dry years. The dry years with the total annual runoff below the observed minimum are more critical for water allocation. Although the methodology generally yields the minimum generated streamflow lower than the minimum weekly observed ones, the previously described rearrangement for achieving the target correlation structure may not produce a series with extremely dry year(s) in which low flows persist over longer durations.

For this reason, the algorithm of Ilich (2014) is upgraded for additional rearrangement of the simulated data set so that it contains a number of extremely dry years. This is achieved by additional swaps of the smallest weekly flows, while keeping previously achieved correlation structure, as explained by Marković *et al.* (2015) and shown in Algorithm 2.2 (code lines 35–42). One additional rearrangement yields one extremely dry year, but the procedure can be repeated for an arbitrary number n_d of extremely dry years with n_d smallest annual flows. The same procedure of additional rearrangement can be applied for the extreme wet years if they are of interest for the reservoir operation management.

Step 3 – adjusting the correlation of weekly flows from one year to another and of mean annual flows

Serial correlation of weekly or monthly hydrologic time series for different lags should not only be preserved within one year but also from one year to another. For example, flows in weeks 1, 2, etc. in a year are dependent on flows in weeks 50, 51, and 52 from the previous year. Such correlations in the observed data should, therefore, be reflected in the generated data. Also, annual streamflows also exhibit correlations that should be maintained in the generated series. These two requirements can be achieved by rearranging complete years (i.e., rows in matrix **G**) with already arranged weekly streamflows (Ilich & Despotovic 2008). By doing so in Step 3 of the methodology, the generated random variates are effectively converted into time series with the required correlation structure.

Algorithm 3 shows the pseudo-code for rearranging generated data to adjust the serial correlation of weekly data in the transition from one year to another and to adjust the serial correlation of the aggregated annual data. If s represents the index of the last time interval in a year ($s = 52$ for weekly data), then for any station from the given data set $\rho_{s,1}$ is the observed serial correlation coefficient between the 52nd week of the current year and the 1st week of the next year. Similarly, $\rho_{s-1,1}$ describes the correlation between week 51 in the current year and week 1 in the next year, etc. Performing an additional rearrangement to adjust serial correlation over the time index range [s – 1, 2] accounts for two time lags. The rearrangement criterion for station k is to minimize the statistic D_k representing the sum of squared differences between observed and simulated transitional correlations up to lag 2 (Ilich & Despotovic 2008):

$$D_k = (\rho^G_{s-1,1} - \rho_{s-1,1})^2 + (\rho^G_{s,1} - \rho_{s,1})^2 + (\rho^G_{s,2} - \rho_{s,2})^2 \quad (7)$$

where G denotes correlation coefficients in the generated data. The above statistic can be expanded to include correlations for any number L of weeks at the end and the beginning of the year.

The correlation structure of the observed annual flows or precipitation also has to be preserved in the simulated series. Similarly, if RAO_l and RAG_l denote annual serial correlation coefficients for lag l for the observed and generated data sets, respectively, the criteria D_k can be expanded by the term, which measures the goodness of fit of the annual serial correlations up to lag m:

$$D_k = \sum_{p=1}^{L} \sum_{q=s-L+p}^{s} (\rho^{kG}_{q,p} - \rho^{k}_{q,p})^2 + \sum_{l=1}^{m} (RAO^k_l - RAG^k_l)^2 \quad (8)$$

where q and p are indices of weeks in the transition from one year to another, and s is the number of weeks in a year. The serial correlation of weekly data can generally be adjusted up to an arbitrary lag L, while the annual serial correlation is adjusted up to the lag $m = N/4$, where N is the number of data years in the observed series, as recommended by Box & Jenkins (1970). For all

gauging stations, composite criteria statistic can be introduced as the sum of all D_k values, where K is the number of stations:

$$D = \sum_{k=1}^{K} D_k \tag{9}$$

The rearrangement of rows in matrix **G** is performed until D is sufficiently small, i.e., smaller than a pre-set value D_0. To find an appropriate order of years (i.e., rows in matrix **G**) that satisfies the transitional weekly and annual correlations, the algorithm in this step combines forward and backward searches for substitute rows, starting from the first and the last rows of **G** simultaneously. The algorithm stops at the first encounter of satisfied criteria for statistic D.

Algorithm 3: Pseudo-code for Step 3 – adjusting transitional weekly correlation and annual correlation

```
load matrices X and G from Step 2
set the value for the number L of ending/starting weeks in a year to be included in the adjustment
set the value for the number m of lags in annual serial correlation function to be included in the adjustment
set the value for the tolerance limit D0 for the criteria statistic D
for each station k of K
  find transitional weekly correlations:
    from X extract L last columns with rows from 1 to n – 1 and L first columns with rows from 2 to n
    from G extract L last columns with rows from 1 to N – 1 and L last columns with rows from 2 to N
    calculate L(L + 1)/2 correlation coefficients rho_qp between the extracted columns in each: X and G
  calculate D1(k) as the sum of differences between observed and generated rho_qp for all lags
  find annual correlations:
    create vector AO of annual sums of observed weekly data in X for n years
    create vector AG of annual sums of generated weekly data in G for N years
    calculate autocorrelation functions RAO and RAG of observed/generated annual data up to lag m
  calculate D2(k) as the square sum of differences between RAO and RAG for all lags
end for (next station)
calculate statistic D = D1 + D2
start rearrangement algorithm on matrix G: set initial best statistic DB = D
for i_asc = first year to last year with increment +1
  for i_desc = last year to first year with increment −1
    if i_asc <> i_desc
      trial swap of data values between rows i_asc and i_desc
      recalculate correlation coefficients and statistic D
      if D < DB then accept trial swap and set DB = D
      if DB < D0 then break
    end if // i_asc <> i_desc
  end for (next i_desc)
end for (next i_asc)
```

APPLICATION

Models and data sets

The presented method for multivariate, multisite, and multi-temporal stochastic hydrologic generation is applied to two data sets, one from Serbia and one from Canada, consisting of streamflow and precipitation data series from a different number of stations. For both data sets, two models are applied: (1) model for the generation of streamflow series, denoted here as MG-Q, and (2) model for the generation of streamflow and precipitation series, denoted as MG-QP. Both models are applied for two time scales: weekly and monthly (symbolized by letters w and m, respectively; e.g., MG-Q(w) is the model for generating streamflows on a weekly scale).

The Serbian data set comprises daily data from three hydrologic stations (Devići, Mlanča, and Ušće) on the Studenica River and meteorological station Kraljevo with the precipitation data (upper part of Figure 2). The streamflow data represent natural flows because there are no water control facilities on the Studenica River. Before the application, data were subjected to quality control procedures. Minor gaps were filled using the regression analysis with other stations. The record is 49 years long from 1964 to 2012.

The Canadian study region is the Oldman River in Southern Alberta with two of its tributaries, Waterton River and St Mary River (lower part of Figure 2).

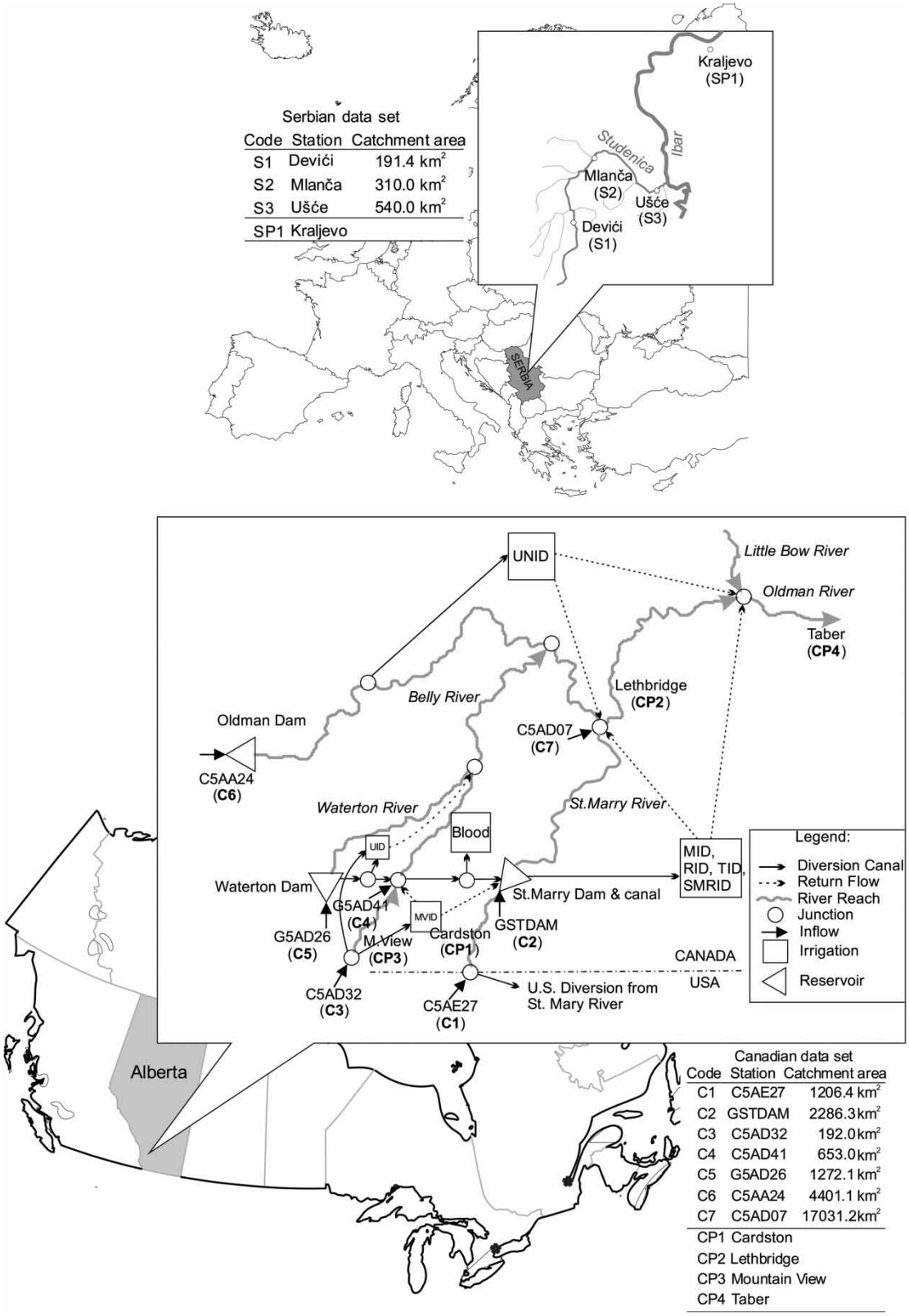

Figure 2 | The map of the study area in Serbia (top) and Canada (bottom) – short codes and full names for the stations used in the model application.

Naturalized weekly flows were obtained from Alberta Environment's natural flow database, with an available record from 1912 to 2001 and for precipitation from 1928 to 2001. Ilich (2014) used this data set as an example for his original procedure. Table SM1 in the Supplementary Material summarizes the information for all stations.

The presented method for multivariate stochastic generation is coded in the MATLAB environment according to Algorithms 1, 2, and 3 and executed on various computing machines from laptop to desktop PCs. Our experience is that the execution is substantially dependant on the number of variables that are of interest (streamflows, precipitation, temperatures, etc.), the number of gauging stations, and the length of the simulation time step. In the case of application of the MG-QP model to the Canadian data set (seven streamflow and four precipitation stations) at weekly time scale, the computational time is as follows: 1 h for Step 1, 3 min for Step 2, and 20 h for Step 3. Faster execution would be possible if the code were implemented in computer languages that can be compiled.

The paper of Marković *et al.* (2015) presented the results of simulations involving only the MG-Q model (streamflow data generation only) for Canadian and Serbian data sets. This paper presents the results for the MG-QP model that includes both streamflow and precipitation data from Canada and Serbia. These two sets of results enable comparing the efficiency of the algorithm in generating streamflows by taking into account either streamflow dependence structure only or streamflow–precipitation-dependence structure.

Results

Results for Step 1 – generation of random series

The distributions of the generated weekly vectors obtained by the MG-QP model are almost identical to the observed ones. Figure 3(a) shows the empirical distributions of the observed and simulated 10th-week precipitation for the Serbian precipitation station SP1. For comparison reasons, some of the most commonly used parametric distribution functions (Gumbel, Pearson 3, log-Pearson 3, and two-parameter gamma) are also applied to the data in Figure 3(a). It can be seen that the employed parametric distributions in this example do not have sufficient flexibility to describe the data at distribution tails, while the non-parametric distributions provide that the generated data have almost the same empirical distribution as the observed data. Also, the non-parametric distributions are more appropriate at the lower tail, where some parametric distributions would yield negative values. The same results for stations CP1 and S2 are given in Figures SM3 and SM4, leading to the same conclusions.

The good fit of the distributions of the observed and generated vectors also leads to a good fit in the vectors' statistics. The means, standard deviations, and skew coefficients of weekly precipitation are almost identical for the observed and simulated series, as shown in plots (b), (c), and (d) of Figure 3. For example, the relative errors in mean weekly flows/precipitation data are in the range of 0.2–6.4% for station S1 (mean 2.1%), 0.1–6.2% for station S2 (mean 2.2%), 0–5.9% for station S3 (mean 2.3%), and 0.1–7.9% for station SP1 (mean 2.7%). Complete results on errors in means are given in Table SM2, showing that the errors for the shorter Serbian data set are comparable with those for the longer Canadian data set.

The generated data sets have greater maxima than the observed ones, as expected in the longer series (Figure SM6). Similarly, simulated minimum flows are smaller than the observed, as shown by Marković *et al.* (2015). With zero being the most frequent minimum value in the observed precipitation series, the same is the case in the simulated series. Also, the percentages of zero values in the observed and the generated precipitation series are very similar (panel (c) in Figure SM6).

The same conclusions can be made about good reproduction of the distributions of observed monthly vectors. Figure 4 compares these distributions using the box plots. The errors in mean monthly data are in the range of 0.1–4.8% for Serbian stations and 0.0–4.4% for Canadian stations (Table SM3).

Results for Step 2 – serial and cross-correlation

The data rearrangement resulting from the application of the ICA results in a good fit between the observed and generated correlation structure. Lag 1 and lag 2 serial correlations for stations SP1 and S2 are compared in Figure SM7. Equally good results are obtained for higher lags and all stations. It

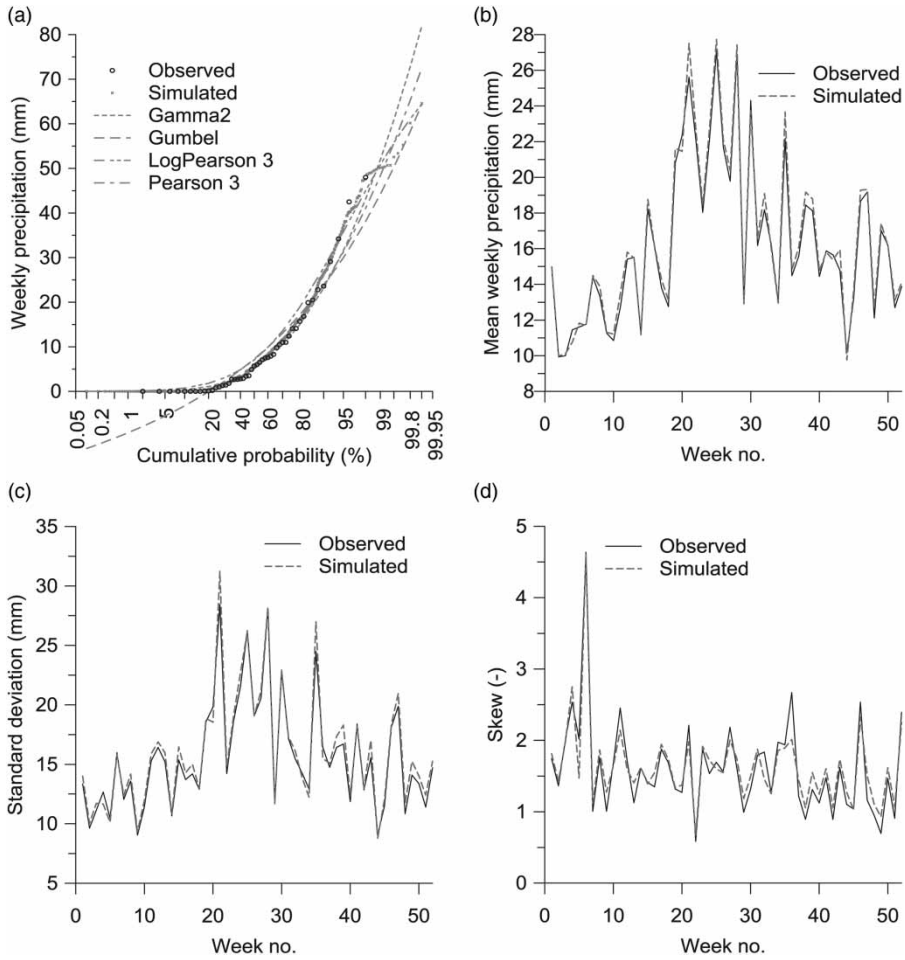

Figure 3 | Model MG-QP(*w*), precipitation station SP1: (a) empirical distributions of the observed and simulated precipitation for week 10 compared to four commonly used distributions fitted to the observed data; (b–d) observed and simulated means, standard deviations, and skew coefficients of data vectors for each week.

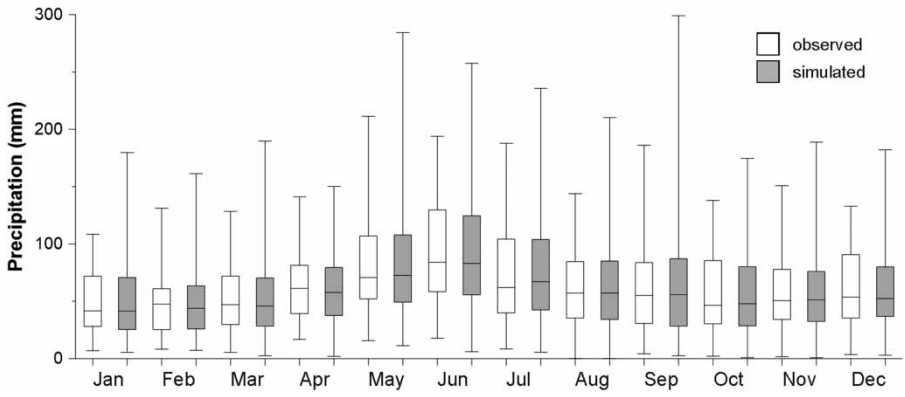

Figure 4 | Model MG-QP(*m*), box-and-whiskers plot of the observed (white) and simulated (grey) monthly precipitation at meteorological station SP1.

is important to notice that the algorithm reproduces not only high correlations but also the small ones, which are below the significance level. The average and maximum differences of the observed and simulated correlation coefficients for weekly data (derived from the correlation matrices for corresponding data) are 0.035 and 0.273 for Serbian

stations, respectively, and 0.033 and 0.308 for Canadian stations, respectively.

For the monthly data, the reproduction of autocorrelation is also good (Figures SM8). The average and maximum differences of the observed and simulated correlation coefficients for monthly data are 0.021 and 0.118 for Serbian data, and 0.022 and 0.169 for Canadian data.

Results for Step 3 – transitional weekly correlation and annual correlation

The data rearrangements in Step 3 lead to an adjustment of the correlation coefficients in the year-to-year transition and therefore at the end of this step, the generated data represent the time series with the completely reproduced autocorrelation function (ACF) of the observed time series. The simulation results show that the transitional year-to-year correlations for weekly data are well simulated (Table SM4). The differences between the observed and generated transitional correlations are generally very small (in average 0.036), but the greatest differences (up to 0.383) are attributed to Serbian hydrologic stations. As a result, the ACFs of the observed and generated data are in good agreement. The examples of the ACFs for weekly precipitation are given in Figure 5, showing that the correlation structure is preserved even for small correlations close to zero. Similarly, a comparison of the cross-correlation functions for weekly data at selected stations (Figures SM9 and SM10) also shows good agreement.

When aggregated on a coarser temporal scale, the generated data are comparable to the aggregated observed data in terms of the annual statistics, distributions, and correlation structure. This is shown by aggregating generated weekly data to a 4-week scale and an annual scale. An example of the observed and simulated annual precipitation distribution functions is shown in Figure SM11. This figure also illustrates the effect of additional treatment at the end of Step 2 over the years with low annual precipitation, which results in a better agreement of the lower distribution tail.

The main statistics for the weekly streamflow data aggregated to the 4-week scale for one station are presented in Figure SM12, also showing good agreement. The comparison of the statistics of the annual streamflows and precipitation

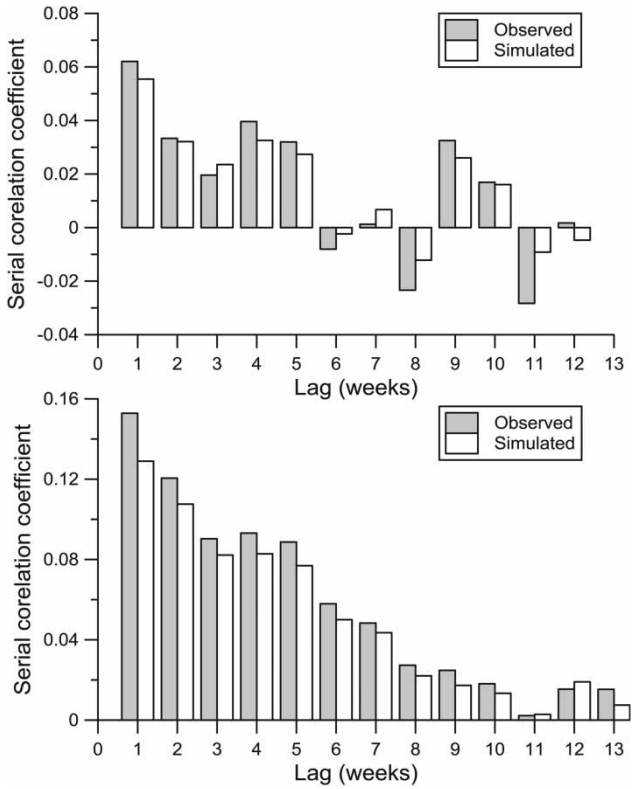

Figure 5 | Model MG-QP(*w*), comparison of serial correlation functions of the observed and simulated weekly precipitation at station SP1 (top) and CP1 (bottom).

aggregated from weekly data is given in Table SM5, showing remarkable agreement. Differences in the means do not exceed 2.3% and 2.9% for Serbian and Canadian stations, respectively, while the differences in standard deviations are almost negligible for flows and somewhat greater for precipitation due to its more random nature.

Serial correlation is also preserved in the aggregated series. Annual ACFs of weekly precipitation aggregated to annual scale for two stations are shown in Figure 6. Statistically insignificant correlations are well reproduced in the simulated series for up to lag 12. The cross-correlation of the annually aggregated weekly data is also preserved (Table SM6). The average and maximum deviations of the observed and simulated cross-correlations are 0.008 and 0.048 for Serbian data, respectively, while the corresponding values for the Canadian data are 0.016 and 0.08 (these are slightly greater because more precipitation stations were included in imposing the correlation structure). Figure SM13 presents ACFs of weekly data aggregated to the 4-week scale.

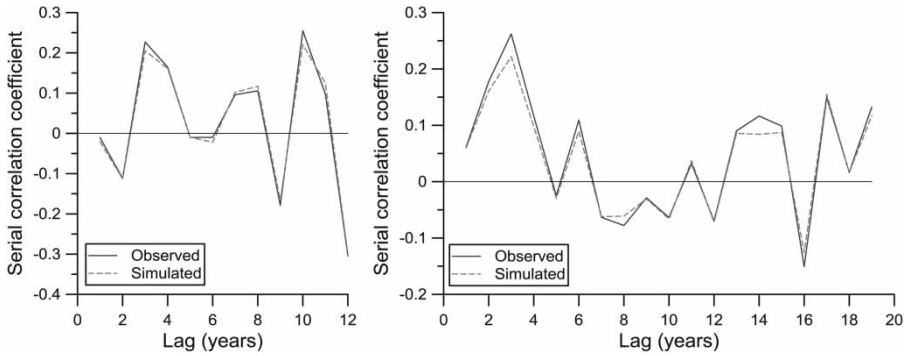

Figure 6 | Model MG-QP(*w*), ACFs of annually aggregated weekly data for stations SP1 (left) and CP1 (right).

In the 1,000-year long generated series, the annual extreme values should exceed those found in the observed series. The generated and the observed annual minima or maxima should generally be evaluated in terms of their distributions. To avoid deciding on the goodness of fit of the theoretical distributions to the observed and generated data, we compare the ranges of theoretical quantiles obtained by fitting some of the commonly used theoretical distributions to both observed and simulated annual data (we used log-normal, Gumbel, Pearson 3, log-Pearson 3, and two-parameter gamma distributions). The ranges of theoretical quantiles of the minimum and maximum annual weekly streamflows for station C1 are compared in Figure SM14. The ranges of theoretical quantiles of generated maxima mostly overlap with those for the observed maxima, although are somewhat wider. The ranges of theoretical quantiles of generated minima also mostly overlap with those for the observed minima and can be lower than their observed counterparts for greater probabilities. This indicates the direction for future improvement of the model.

The effects of the rearrangement algorithms in Steps 2 and 3 can also be seen through marginal improvements in achieving the dependence structure of the generated data after Step 1, Step 2, and Step 3. This is illustrated for weekly ACFs, transitional weekly correlations, and serial correlation of weekly streamflows aggregated to annual scale in Figures SM15, SM16, and SM17.

The results for monthly data show equally good agreement of transitional year-to-year correlations (Table SM7) and complete ACFs (Figure SM18). Comparison of the statistics of the annually aggregated monthly data is shown in Tables SM8 and SM9.

Comparison of MG-QP and MG-Q models

By comparing the results for the streamflows simulated by the MG-QP model presented in this paper with the results of the simulations with the MG-Q model presented in Marković *et al.* (2015), no significant differences in the model performance can be seen. For example, empirical distributions of observed and simulated series, observed and simulated weekly flow means, standard deviations, and skew coefficients are almost the same for both models (Figures SM4 and SM5). Also, the relative errors in the means of weekly streamflows by the MG-Q model range from 0.0% to 3.65%, which is virtually the same as with the MG-QP model. Additional comparisons of the results of two models are given in the Supplementary Material (Figures SM19, SM20, and SM21), showing that the model performance is not deteriorated with the introduction of a greater number of variables and more complicated dependence structure of the multivariate setup.

CONCLUSIONS

This paper presents the development and application of the stochastic model for generating simultaneous multivariate hydrological time series for a weekly or monthly temporal scale. The following are the main characteristics of the proposed methodology:

- It uses non-parametric distributions coupled with the extrapolation algorithm for data generation and non-parametric rearrangement algorithms to achieve the target correlation structure.

- The heuristic extrapolation algorithm provides a robust solution for extrapolating tails and allows fully automated execution of the algorithm.
- The methodology ensures that the empirical statistic properties of the processes are preserved to a satisfactory degree at the simulation time scale as well as at coarser time scales (e.g., by aggregating from weekly to monthly or annual scale).
- The method preserves the serial correlation on the transition from one year to another.
- Both continuous and intermittent hydrological time series can be generated.
- The generation process is based on the log-transformed data in order to reduce the effect of outliers and avoid negative generated values.
- The procedure is completely automated with a set of default agreement criteria.

The results derived from the two independent data sets (from Serbia and Canada) show that the model can satisfactorily reproduce the probability distributions of multivariate observed series. This is evident from the good match between the main statistics (mean, variance, and skewness coefficient) of the generated and the observed data series. For example, the average relative errors of the observed and simulated weekly precipitation and streamflow series are in the range of 0.1–9.2% and 0–5.4%, respectively (Table SM2), for the Canadian case study. The agreement is achieved by a careful application of non-parametric probability distributions on log-transformed observed data and by using the developed algorithm for the extrapolation of the non-parametric probability distribution.

The logarithmic transformation of the observed data mitigates the influence of outliers and/or skew in data on the resulting long synthetic data series. The algorithm for the extrapolation of the non-parametric probability distribution uses the linear extrapolation of the CDFs using the log-normal probability plot. The extrapolation is performed in the range of the 90% confidence interval of the GEV probability distribution for the 1,000-year quantiles. This algorithm enables equally successful simultaneous generation of long streamflow and precipitation series in a hydrologically homogeneous region.

Two model setups that are considered, one based solely on streamflow data (presented in Marković *et al.* (2015)) and another based on streamflow and precipitation data (presented in this paper), generate a series of almost identical stochastic and marginal characteristics to those observed.

Further research should go in the direction of algorithm refinement regarding computational efficiency for a large number of gauging sites with long records and short time steps (e.g., daily time step). Another improvement can be found in the development of a more efficient method for the optimization algorithm in Step 3.

ACKNOWLEDGEMENT

The authors would like to thank Republic Hydrometeorological Service of Serbia for providing data on streamflows and precipitation for this study and to three anonymous reviewers who helped to improve the manuscript significantly. Code availability: The authors are open to considering source code sharing request subject to internal procedures among the authors. Data availability: Data are available at request from the first author (djurica.markovic@pr.ac.rs) under condition that they are not distributed to the third parties.

SUPPLEMENTARY DATA

The Supplementary Data for this paper is available online at http://dx.doi.org/10.2166/hydro.2019.071.

REFERENCES

Basinger, M., Montalto, F. & Lall, U. 2010 A rainwater harvesting system reliability model based on nonparametric stochastic rainfall generator. *Journal of Hydrology* **392**, 105–118.

Box, G. & Jenkins, G. 1970 *Time Series Analysis Forecasting and Control*. Holden-Day, Oakland, CA.

Fiering, M. B. 1964 Multivariate technique for synthetic hydrology. *Journal of the Hydraulics Division* **90** (5), 43–60.

Haberlandt, U., Hundecha, Y., Pahlow, M. & Schumann, A. H. 2011 Rainfall generators for application in flood studies. In: *Flood Risk Assessment and Management*. Springer, Dordrecht, The Netherlands, pp. 117–147.

Harrold, T., Sharma, A. & Sheather, S. 2003a A nonparametric model for stochastic generation of daily rainfall occurrence. *Water Resources Research* **39** (10), 1300.

Harrold, T., Sharma, A. & Sheather, S. 2003b A nonparametric model for stochastic generation of daily rainfall amounts. *Water Resources Research* **39** (12), 1343.

Hazen, A. 1914 Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Association of Civil Engineers* **77**, 1539–1669.

Ilich, N. 2014 An effective three-step algorithm for multi-site generation of stochastic weekly hydrological time series. *Hydrologic Sciences Journal* **59** (1), 1–14.

Ilich, N. & Despotovic, J. 2008 A simple method for effective multi-site generation of stochastic hydrologic time series. *Stochastic Environmental Research and Risk Assessment* **22** (2), 265–279.

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

Kossieris, P., Tsoukalas, I., Makropoulos, C. & Savic, D. 2019 Simulating marginal and dependence behaviour of water demand processes at any fine time scale. *Water* **11** (5), 885. https://doi.org/10.3390/w11050885.

Koutsoyiannis, D. 2000 A generalized mathematical framework for stochastic simulation and forecast of hydrologic time series. *Water Resources Research* **36** (6), 1519–1533.

Koutsoyiannis, D. 2005 Stochastic simulation of hydrosystems. In: *Water Encyclopedia*. John Wiley & Sons, Inc. https://doi.org/10.1002/047147844X.sw913.

Marković, Đ., Plavšić, J., Ilich, N. & Ilic, S. 2015 Non-parametric stochastic generation of streamflow series at multiple locations. *Water Resources Management* **29** (13), 4787–4801.

Mehrotra, R., Srikanthan, R. & Sharma, A. 2006 A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology* **331**, 280–292.

Moran, P. 1970 Simulation and evaluation of complex water systems operations. *Water Resources Research* **6** (6), 1737–1742. https://doi.org/10.1029/WR006i006p01737.

Rao, R. & Hamed, K. 2000 *Flood Frequency Analysis*. CRC Press, Boca Raton.

Salas, J. D., Delleur, J. W., Yevjevich, V. & Lane, W. 1980 *Applied Modeling of Hydrologic Time Series*. Water Resources Publications, Littleton, CO, USA.

Scholz, F. 1995 *Nonparametric Tail Extrapolation*. Boeing Information & Support Services. Available from: http://faculty.washington.edu/fscholz/Reports/ISSTECH-95-014.pdf (accessed 20 February 2019).

Serinaldi, F. & Kilsby, C. G. 2014 Simulating daily rainfall fields over large areas for collective risk estimation. *Journal of Hydrology* **512**, 285–302. https://doi.org/10.1016/j.jhydrol.2014.02.043.

Sharif, M. & Burn, D. 2006 Simulating climate change scenarios using an improved K-nearest neighbor model. *Journal of Hydrology* **325**, 179–196.

Sharif, M. & Burn, D. 2007 Improved K-nearest neighbor weather generating model. *Journal of Hydrologic Engineering* **12** (1), 42–51.

Sharma, A., Tarboton, D. & Lall, U. 1997 Streamflow simulation: a nonparametric approach. *Water Resources Research* **33** (2), 291–308.

Srikanthan, R. & McMahon, T. 2001 Stochastic generation of annual, monthly and daily climate data: a review. *Hydrology and Earth System Sciences* **5** (4), 653–670.

Srinivas, V. & Srinivasan, K. 2005 Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows. *Journal of Hydrology* **302**, 307–330.

Srivastav, R. & Simonovic, S. 2014 An analytical procedure for multi-site, multi-season streamflow generation using maximum entropy bootstrapping. *Environmental Modelling & Software* **59**, 59–75.

Srivastav, R. & Simonovic, S. 2015 Multi-site, multivariate weather generator using maximum entropy bootstrap. *Climate Dynamics* **44** (11–12), 3431–3448.

Thomas Jr., H. A. & Fiering, M. B. 1962 Mathematical synthesis of streamflow sequences for the analyses of river basins by simulation. In: *The Design of Water Resources Systems*. Harvard University Press, Cambridge, MA, pp. 459–493.

Tsoukalas, I., Efstratiadis, A. & Makropoulos, C. 2018a Stochastic periodic autoregressive to anything (SPARTA): modeling and simulation of cyclostationary processes with arbitrary marginal distributions. *Water Resources Research* **54** (1), 161–185. https://doi.org/10.1002/2017WR021394.

Tsoukalas, I., Makropoulos, C. & Koutsoyiannis, D. 2018b Simulation of stochastic processes exhibiting any-range dependence and arbitrary marginal distributions. *Water Resources Research* **54** (11), 9484–9513.

Tsoukalas, I., Papalexiou, S., Efstratiadis, A. & Makropoulos, C. 2018c A cautionary note on the reproduction of dependencies through linear stochastic models with non-Gaussian white noise. *Water* **10** (6), 771. https://doi.org/10.3390/w10060771.

Wilks, D. & Wilby, R. 1999 The weather generation game: a review of stochastic weather models. *Progress in Physical Geography* **23** (3), 329–357.

First received 26 March 2019; accepted in revised form 5 August 2019. Available online 10 September 2019