https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions
Open Access

Randomized Block Nonparametric Temporal Disaggregation of Hydrological Variables

RB-NPD (version1.0) - model development

Taesam Lee[1] and Taha B. M. J. Ouarda[2]

Key words: Disaggregation, Nonparametric Model, Flood, Great Lakes, Random Block

[1]Department of Civil Engineering, Gyeongsang National University, Jinju, South Korea

[2]Canada Research Chair in Statistical Hydro-Climatology, National Institute for Scientific Research, INRS-ETE, 490 de la Couronne, Quebec (QC), Canada G1K 9A9

Corresponding Author: Taesam Lee,
501 Jinju-daero, Jinju, Gyeongsangnam-do, 660-701, South Korea,
E-mail: tae3lee@gnu.ac.kr, Tel: +82 55 772 1797

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Abstract

Stochastically simulated data have been employed for hydrological variables in critical water-related risk management. The simulated data can be utilized to assess the existing flood protection structure and future mitigation frameworks. Disaggregation of the simulated annual data to a lower time scale is often required since water resource management and flood mitigation plans should be done in a fine scale such as a monthly or quarter-monthly. In the current study, the randomized random block length was proposed for the nonparametric disaggregation model since one of the major weakness points for the nonparametric disaggregation model is repetition of similar patterns in the disaggregated data. Furthermore, long-term dependence structure was also mainly focused to preserve since consistent high-flow results devastating damages to inundated area. The proposed model was compared with the existing parametric and nonparametric disaggregation models. The annual net basin supplies (NBS) of the Lake Champlain–Richelieu River (LCRR) Basin was employed to test the performance of the proposed model by reproducing the critical statistics of the 2011 flood in the LCRR Basin. The 2011 flood occurred and was sustained for a few months. The results show that the existing parametric and nonparametric models have limitations and shortcoming and do not provide sufficient temporal dependence. In contrast, the proposed random block-based nonparametric disaggregation (RB-NPD) model with further model enhancement by the genetic algorithm mixture illustrates that the proposed RB-NPD model can be a comparable alternative and that its enhancement is suitable for disaggregating the annual NBS data for the LCRR Basin.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

1. Introduction

Disaggregation models for hydrological variables have been developed in a number of studies to downscale simulated data of a coarse time scale to a fine time scale since water management should be performed in a fine time scale such as monthly or quarter-monthly. Valencia and Schaake (1973) proposed a parametric disaggregation model. The major shortcoming of the model by Valencia and Schaake (1973) is that there is no consideration for the previous year. Mejia and Rousselle (1976) improved the model by including the additional term for the last month of the previous year. However, the models require a significant number of parameters. To avoid parsimoniousness, some models have been proposed in a number of studies (Stedinger and Vogel, 1984; Lane and Frevert, 1990; Santos and Salas, 1992) (Santos and Salas, 1992). Furthermore, Koutsoyiannis and Manetas (1996) proposed the accurate adjusting procedure (AAP), which integrates a model for the higher scale (e.g., yearly) and a model for the lower scale (e.g., monthly) by matching the generated sequences at each time scale.

Alternative nonparametric disaggregation methods have been proposed in a number of studies (Srikanthan and Mcmahon, 1982; Porter and Pink, 1991; Tarboton et al., 1998; Prairie et al., 2007; Lee et al., 2010; Lee and Jeong, 2014; Lee and Park, 2017). Prairie et al. (2007) employed the K-nearest neighbor resampling technique, and Lee et al. (2010) improved the model by including the genetic algorithm (GA) mixture.

In disaggregation models of hydrologic variables, higher time scale data (here, annual) are disaggregated into lower time scales (here, monthly) according to the relationship between the annual and monthly data. It is relatively easy to preserve the inner-annual relationship (i.e., month-to-month). However, the interannual relationship of the disaggregated monthly data cannot be

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

easily captured since a disaggregation procedure is commonly performed with the current annual data value without the condition of the previous and future relationships. This might result in a discontinuity between the current monthly data and the previous (also following) year for the disaggregated monthly data. It is, in general, not very problematic for disaggregated data since the interannual relationship contains data at a higher time scale (i.e., annual). However, the interannual relationship is critical for the reproduction of a certain event, such as the 2011 flood in the Lake Champlain–Richelieu River (LCRR) Basin, which lasted approximately 3 months. An extreme event that consistently occurs over a long time cannot be generated unless the interannual relationship is appropriately considered.

The U.S. and Canadian governments launched an initiative to identify how flood forecasting, preparedness and mitigation can be improved in the LCRR Basin. Synthetic net basin supply (NBS) series of the LCRR Basin are crucial to evaluate the adequacy of flood risk mitigation measures and management strategies under a number of potential hydrological scenarios that might occur in the future. Furthermore, the regulations and management plans for the LCRR Basin have been performed on a monthly or quarter-monthly scale. Therefore, an appropriate temporal disaggregation model to provide more specific information for the basin should be applied, if any, or developed to meet the specific statistical characteristics of the 2011 flood in the LCRR Basin.

In the current study, comparable existing parametric and nonparametric models were tested to disaggregate the annual NBS data for the LCRR Basin. The performance of the existing models was carefully examined. Furthermore, a novel approach based on nonparametric techniques was proposed to improve the performance of the existing models, especially for the reproduction of the critical statistics related to the 2011 flood event in the LCRR Basin. Specifically, efforts were

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions

made to find and devise a disaggregation model that appropriately captures the interannual relationship in disaggregated data.

The present report is organized as follows. A mathematical description of the employed models is described in Section 2. The model development is described in Section 3. The data description and application methodology are presented in Section 4. The results of the compared models and the proposed model are shown in Section 5, followed by a summary and conclusion in Section 6.

2. Mathematical Background

2.1. Parametric Disaggregation

Valencia and Schaake (1973) introduced the basic temporal disaggregation model for annual flows to seasonal flows, followed by its extended version by Mejia and Rousselle (1976), defined as:

$$\mathbf{Y}_t = \mathbf{A}X_t + \mathbf{B}\boldsymbol{\varepsilon}_t \quad (1)$$

where X_t is the annual time series at year *t*, $\mathbf{Y}_t$ is the seasonal data for year *t*, τ =1..., n_m as $\mathbf{Y}_t = \left[Y_{t,1}, Y_{t,2}, \ldots, Y_{t,\tau}, \ldots, Y_{t,n_m}\right]^T$, and n_m is the number of seasons. **A** and **B** are $n_m \times 1$ and $n_m \times n_m$ parameter matrices, respectively. $\boldsymbol{\varepsilon}_t$ is the $n_m \times 1$ column noise vector uncorrelated with each element distributed as a standard normal. Mejia and Rousselle (1976) included an additional term to preserve the lag-1 correlation between the current year and the past year as:

$$\mathbf{Y}_t = \mathbf{A}X_t + \mathbf{B}\boldsymbol{\varepsilon}_t + \mathbf{C}\mathbf{Y}_{t,n_m} \quad (2)$$

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

where **C** is the $n_m \times 1$ parameter vector. These disaggregation models suffer from the parsimonious problem since the models require too many parameters, sometimes more than observations.

To avoid this drawback, Lane and Frevert (1990) proposed the condensed version of the parametric disaggregation model on a one-season-at-a-time basis as:

$$Y_{t,\tau} = \mathrm{A}_\tau X_t + \mathrm{B}_\tau \varepsilon_{t,\tau} + \mathrm{C}_\tau \mathrm{Y}_{t,\tau-1} \quad (3)$$

where, A_τ, B_τ, and C_τ are parameters at each season τ. These parameters can be estimated by applying the covariance matrices as:

$$\widehat{\mathrm{A}}_\tau = [\mathbf{S}_{\mathrm{YX}}(\tau,\tau) - \mathbf{S}_{\mathrm{YY}}(\tau,\tau-1)\mathbf{S}_{\mathrm{YY}}^{-1}(\tau-1,\tau-1)\mathbf{S}_{\mathrm{YX}}(\tau-1,\tau)] \cdot [\mathbf{S}_{\mathrm{XX}}(\tau,\tau) - \mathbf{S}_{\mathrm{XY}}(\tau,\tau-1)\mathbf{S}_{\mathrm{YY}}^{-1}(\tau-1,\tau-1)\mathbf{S}_{\mathrm{YX}}(\tau-1,\tau)]^{-1} \quad (4)$$

$$\widehat{\mathrm{C}}_\tau = [\mathbf{S}_{\mathrm{YY}}(\tau,\tau-1) - \widehat{\mathrm{A}}_\tau \mathbf{S}_{\mathrm{XY}}(\tau,\tau-1)]\mathbf{S}_{\mathrm{YY}}^{-1}(\tau-1,\tau-1) \quad (5)$$

$$\widehat{\mathrm{B}}_\tau \widehat{\mathrm{B}}_\tau^T = \mathbf{S}_{\mathrm{YY}}(\tau,\tau) - \widehat{\mathrm{A}}_\tau \mathbf{S}_{\mathrm{XY}}(\tau,\tau) - \widehat{\mathrm{C}}_\tau \mathbf{S}_{\mathrm{YY}}^{-1}(\tau-1,\tau) \quad (6)$$

where, $\mathbf{S}_{\mathrm{YX}}(a,b)$ indicates a covariance between $Y_{t,a}$ and $X_{t,b}$. Note that $\widehat{\mathrm{B}}_\tau$ is estimated from the $\widehat{\mathrm{B}}_\tau \widehat{\mathrm{B}}_\tau^T$ with either using eigenvalue and eigenvectors or estimating a lower triangular form matrix (Bras and Rodriguesz-Iturbe, 1994). To meet the additive condition, adjustment must be made as:

$$Y_{t,\tau}^* = Y_{t,\tau} \times X_t / \sum_{\tau=1}^{n_m} Y_{t,\tau} \quad (7)$$

2.2. Nonparametric Disaggregation (NPD)

Lee et al. (2010) proposed a nonparametric disaggregation model based on k-nearest neighbor resampling and a genetic algorithm for streamflow applications and further developed it for daily

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

precipitation (Lee and Jeong, 2014; Lee and Park, 2017). The procedure is briefly described as follows:

Let the annual, x_t, and seasonal observations $\mathbf{y}_t = [y_{t,1}, \ldots, y_{t,n_m}]$ and t=1...,N, where N is the record length. In addition, X_t is the target annual variable. The objective is to disaggregate the annual time series X_t to the seasonal time series $\mathbf{Y}_t = [Y_{t,1}, \ldots, Y_{t,n_m}]$.

Assumed that the number of nearest neighbors, k, is already known, the temporal disaggregation procedure is as follows:

(1) The distances between the target annual value X_t and the observed annual variable are estimated as:

$$D_i = \begin{bmatrix} X_t - x_i \\ Y_{t-1,n_m} - y_{i-1,n_m} \end{bmatrix}^{\mathrm{T}} \Xi \begin{bmatrix} X_t - x_i \\ Y_{t-1,n_m} - y_{i-1,n_m} \end{bmatrix} \qquad i = 2, \ldots, N \qquad (8)$$

where the distances are measured for i=2,…, N, and Ξ is the variance–covariance matrix of $[x_i, y_{i-1,n_m}]$. Here, the target annual value is considered. In addition, the simulated seasonal value of the last month of the previous year is also taken into account to preserve the dependence of the previous set, as in Lane's model in Eq.(3).

(2) The estimated distances from Step (1) are arranged in ascending order, the first k distances (i.e., the smallest k values) are selected, and the time indices of the smallest k distances are reserved.

(3) One of the stored k time indices is randomly chosen with the weighting probability given by:

$$w_m = \frac{1/m}{\sum_{j=1}^{J} 1/j}, \qquad m = 1, \ldots, k \qquad (9)$$

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions
Open Access
EGU

(4) The seasonal values of the selected time index (denoted as p) are assigned from Step (3) as $\mathbf{y}_p = [y_{p,1}, \dots, y_{p,n_m}]$.

(5) The following steps are executed for GA mixing:

(5-1) Reproduction: One additional time index is selected using Steps (1) through (4) and this index is denoted as p^*. The corresponding seasonal values are obtained, $\mathbf{y}_{p*} = [y_{p*,1}, \dots, y_{p*,n_m}]$. The subsequent two GA operators use the two selected vectors, $\mathbf{y}_p$ and $\mathbf{y}_{p*}$.

(5-2) Crossover: Each element $y_{p,\tau}$ is replaced with $y_{p^*,\tau}$ at the crossover probability P_c, as:

$$Y_{t,\tau} = \begin{cases} y_{p^*,\tau} & \text{if } r < P_c \\ y_{p,\tau} & \text{otherwise} \end{cases} \quad (10)$$

where r is a uniform random number between 0 and 1.

(5-3) Mutation: Each element (i.e., each season, τ=1,…, n_m) is replaced with the one chosen from all observations of this season with the mutation probability P_m, i.e.,

$$Y_{t,\tau} = \begin{cases} y_{a,\tau} & \text{if } r < P_m \\ y_{p,\tau} & \text{otherwise} \end{cases} \quad (11)$$

where $y_{a,\tau}$ is selected from $[y_{1,\tau}, \dots y_{N,\tau}]$ with equal probability for i=1,…,N.

(6) The GA mixed values are adjusted as follows to preserve the additive condition as in Eq. (7).

(7) Steps (1)-(5) are repeated until the target data are generated.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

The characteristics of the mutation probability P_m and the crossover probability P_c were studied well by Lee et al. (2010) and Lee (2008). In the current study, two probabilities were used as tuning parameters to manipulate preservation of the historical statistics for the generated monthly time series. The selection of the number of nearest neighbors (k) has been studied (Lall and Sharma, 1996; Lee and Ouarda, 2011). The most common and simplest selection method was applied in the current study by setting $k = \sqrt{N}$. This heuristic approach has commonly been employed in simulation studies with KNNR (Lall and Sharma, 1996; Lee and Ouarda, 2011; Lee et al., 2017; Lee and Ouarda, 2019).

2.3. Normal Copula Standardization

Seasonal variables, especially in hydroclimatological fields, are commonly skewed. A number of transformation methods have been attempted, such as box-cox, log, and gamma transformation. Among others, copula normal standardization can be a good alternative due to its simplicity and preservation of the marginal statistics as follows:

$$Z_t = F_{\Phi}^{-1}[F_Y(Y;\boldsymbol{\theta})] \tag{12}$$

where F_Y is the selected distribution for the Y variable, such as gamma, and F_{Φ}^{-1} is the inverse standard normal distribution. With this normalization, the result variable (F_Z) definitely has a standard normal distribution. For a marginal distribution, gamma was chosen since this distribution has been commonly used in hydroclimatological variables and fitted well to positively nonnegative skewed variables. Applications have been made in the statistical downscaling in climate change studies (Lee and Singh, 2018). Its back-transformation can be performed by:

$$Y_t = F_Y^{-1}[F_{\Phi}(Z;\boldsymbol{\theta})] \tag{13}$$

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

3. Model Development

3.1. Random Block-based Nonparametric Disaggregation (RB-NPD)

To avoid discontinuation between the current year and following year, the random block length was used instead of the fixed length (i.e., l=12). The proposed method is similar to the original NPD model. However, both the current and following years for annual data must be considered in selecting the candidate. It can be described as follows: (1) to generate the block length (l) from a discrete distribution; (2) to estimate the distances between the observed and generated data, such as the current year and following year values of the observed and target annual data; and (3) to mix the selected block and one additional block. A detailed description is provided in Figure 1:

i. A block length, L_B, is generated randomly from a discrete distribution (e.g., geometric or Poisson) for the length of the following seasonal values that follow $Y_{t,\tau}$. Among other distributions, a Poisson distribution is used because the distribution shape is close to a gaussian distribution centered on the mean. More information on the selection of this discrete distribution in block bootstrapping can be found in previous studies (Lee and Ouarda, 2012; Lee and Ouarda, 2019). The Poisson distribution with its parameter (α) is:

$$L_B \sim \frac{e^{-\alpha}\alpha^{l-1}}{(\alpha-1)!} \qquad l = 1,2,\dots. \tag{14}$$

Note that the parameter (α) is the mean of L_B. The parameter for this Poisson distribution (τ) was set as the number of seasons used (e.g., α =12 for monthly) so that the average block length was the same as the number of seasons. In Figure 1, l=10 is generated.

ii. Distances are estimated to collect close observations to the current status with KNNR as follows:

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

$$D_i = \begin{bmatrix} X_t - x_i \\ Y_{t,\tau-1} - y_{i,\tau-1} \end{bmatrix}^{\mathrm{T}} \Xi_\tau^{-1} \begin{bmatrix} X_t - x_i \\ Y_{t,\tau-1} - y_{i,\tau-1} \end{bmatrix} \qquad i = 1, \ldots, N \tag{15}$$

$$D_i = \begin{bmatrix} X_t - x_i \\ X_{t+1} - x_{i+1} \\ Y_{t,\tau-1} - y_{i,\tau-1} \end{bmatrix}^{\mathrm{T}} \Xi_\tau'^{-1} \begin{bmatrix} X_t - x_i \\ X_{t+1} - x_{i+1} \\ Y_{t,\tau-1} - y_{i,\tau-1} \end{bmatrix} \qquad i = 2, \ldots, N-1 \tag{16}$$

Eq. (15) or (16) should be used according to whether the generated block length overrides the next year. For example, in Figure 1, the block length l=10 overrides the following year since the starting season for the current simulation is τ=11, and Eq. (16) must be employed. Ξ_τ and Ξ'_τ are the variance–covariance matrices for the considered elements as follows:

$$\Xi_\tau = \begin{bmatrix} \mathrm{var}(X_t) & \mathrm{cov}(X_t, Y_{t,\tau-1}) \\ \mathrm{cov}(X_t, Y_{t,\tau-1}) & \mathrm{var}(Y_{t,\tau-1}) \end{bmatrix} \tag{17}$$

$$\Xi_\tau = \begin{bmatrix} \mathrm{var}(X_t) & \mathrm{cov}(X_t, X_{t+1}) & \mathrm{cov}(X_t, Y_{t,\tau-1}) \\ \mathrm{cov}(X_t, X_{t+1}) & \mathrm{var}(X_{t+1}) & \mathrm{cov}(X_{t+1}, Y_{t,\tau-1}) \\ \mathrm{cov}(X_t, Y_{t,\tau-1}) & \mathrm{cov}(X_{t+1}, Y_{t,\tau-1}) & \mathrm{var}(Y_{t,\tau-1}) \end{bmatrix} \tag{18}$$

The first element (j=1) in both equations and the last element in Eq. (16) are omitted because the data are not available.

iii. From the k numbers of the smallest distances among j = 2, …, N (or N-1), one of the time indices is chosen with the probability in Eq. (9). Assume that the selected points are p, and its following sequence is obtained. For example, l=10 in Figure 1 and the following sequence is selected as the generated sequence: $[Y_{t-1,11}, Y_{t-1,12}, Y_{t,1} \ldots, Y_{t,8}] = [y_{p-1,11}, y_{p-1,12}, y_{p,1} \ldots, y_{p,8}]$.

iv. For the GA crossover, one more sequence is chosen with the steps above (ii)-(iii) and the additional time index is assumed as p^* in Figure 1 as $[y_{p^*-1,11}, y_{p^*-1,12}, y_{p^*,1} \ldots, y_{p^*,8}]$. Each

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

element of the first chosen sequence is replaced with the crossover probability as in Eq. (10). In Figure 1, the elements [$y_{p\text{-}1,12}$, $y_{p,5}$, $y_{p\text{-}1,7}$] are replaced with [$y_{p^*\text{-}1,12}$, $y_{p^*,5}$, $y_{p^*\text{-}1,7}$]

v. For the GA mutation, each element is substituted into the gamma random number with the probability P_m (i.e., $r<P_m$, where r is a uniform random number between 0 and 1), i.e.,

$$Y_{t,\tau}^{new} \sim Gamma\ (\alpha_\tau, \beta_\tau) \tag{19}$$

where α_τ and β_τ are parameters that can be estimated by fitting the gamma distribution to the seasonal data, i.e., $y_{t,\tau}$ and t=1...,N. In Figure 1, $Y_{t,2}$ is simulated from Eq. (19).

vi. Simulated seasonal data at each year are adjusted to meet the additive condition as in Eq.(7).

vii. Steps (i)-(vi) are repeated until the required data are disaggregated.

3.2. Model Enhancement

Further consideration was tested to preserve the interconnection in the mutated values from Eq.(19) by replacing the value with the following condition:

$$Z_{t,\tau} = \begin{cases} Z_{t,\tau}^{new} & \text{if } Z_{t,\tau-1}Z_{t,\tau}^{new} + Z_{t,\tau}^{new}Z_{t,\tau+1} > Z_{t,\tau-1}Z_{t,\tau} + Z_{t,\tau}Z_{t,\tau+1} \\ Z_{t,\tau} & \text{otherwise} \end{cases} \tag{20}$$

where $Z = \Phi^{-1}[F_Y(Y); \alpha_\tau, \beta_\tau]$. Note that the condition of $Z_{t,\tau-1}Z_{t,\tau}^{new} + Z_{t,\tau}^{new}Z_{t,\tau+1} > Z_{t,\tau-1}Z_{t,\tau} + Z_{t,\tau}Z_{t,\tau+1}$ indicates that the newly proposed value ($Z_{t,\tau}^{new}$) has a higher correlation than the original value ($Z_{t,\tau}$) since Z is the standard normal variable and its multiplication indicates the correlation. This modification is named after 'the enhanced correlation algorithm in crossover' and denoted as 'ECAco'.

Furthermore, we tested an additional algorithm by simulating the mutation value as

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions

$$Z_{t,\tau}^{new} = a_1 Z_{t,\tau-1} + a_2 \tilde{Z}_t + \varepsilon_t \quad (21)$$

$$Y_{t,\tau}^{new} = F_Y^{-1}[\Phi(Z_{t,\tau}^{new}); \alpha_\tau, \beta_\tau] \quad (22)$$

where $Z_{t,\tau-1}$ is the standard normal variable transformed from $Y_{t,\tau-1}$ with the gamma distribution and $\tilde{Z}_t$ is the standard normal variable of X_t. This enhancement of the parametric simulation algorithm in mutation is denoted as 'PSAm'.

4. Data Description and Application Methodology

4.1. Data Description

The annual and monthly data of the net basin supply (NBS) series for the Lake Champlain–Richelieu River system (LCRR) were applied in the current study. The water supplies to a lake or a river are referred to as NBS, and they are estimated with both component-based and residual-based methods (Croley and Lee, 1993). The component-based NBS series is used due to its accuracy and popularity in the literature (Fagherazzi et al., 2011; Ouarda and Charron, 2019). The LCRR Basin has an area of 23,900 km^{2,} with approximately 84% of the basin in northeastern New York and northwestern Vermont in the US and 16% in Quebec in Canada, as shown in Figure 2. In the spring of 2011 in the LCRR Basin, the worst flooding ever recorded in the past 100 years occurred, which damaged homes, businesses, and farms. Several annual stochastic series for the NBS were simulated in different studies, and further disaggregation to monthly data is required since the regulations and water management were performed on a monthly or quarter-monthly time scale.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

4.2. Application Methodology

A number of disaggregation models were considered in the current study, including the parametric VS and MR models and the NPD model (the original nonparametric disaggregation model from Lee et al. (2010)). Furthermore, the proposed RB-NPD model was fully tested with/without the enhanced algorithms in the GA, ECAco and PSAm. In application, the RB-NPD model implies the model without applying the enhanced algorithms and the RB-NPD model with ECAco (called the selective model) is only the enhancement applied in the crossover while the RB-NPD model with ECAco+ PSAm (called hybrid model) presents both the enhancements applied to the crossover and mutation process. Note that the parametric VS and MR models were applied to the copula transformed data with Eq. (12) and the disaggregated data for these models were back-transformed with Eq. (13). Additionally, the Lane model was also tested. However, its performance was not satisfactory and was much worse than those of the VS and MR models. Therefore, its results were not included in this manuscript.

To check the performance of the disaggregation models considered in the current study, the observed annual NBS data were disaggregated, and 200 series were produced. Note that the proposed disaggregation model is based on the simulation technique, and an infinite number of series can be produced from the simulation-based disaggregation model. The key statistics of the disaggregated data were estimated and presented by boxplots following the comparison with the observed data. In a boxplot, the boxes represent the interquartile range (IQR), and the whiskers extend up to 1.5 IQR. The horizontal line inside the box shows the median of the data. Data beyond the whiskers (1.5 IQR) are indicated by a plus sign (+).

Furthermore, the interesting feature of the disaggregated NBS is the reproduction of the observed high values, especially in a consistent manner. In other words, the flood in the LCRR

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Basin in 2011 continued for a few months. It is important to generate these continuous high values for a few months in the disaggregated data. Therefore, the key statistics of the accumulated data up to six months were also tested. Note that the n-month accumulation was performed by averaging the monthly data of the previous months. For example, 3 months accumulation for Month 2 (i.e., τ=2) is the average value of the $Y_{t-1,12}$, $Y_{t,1}$, and $Y_{t,2}$.

5. Results

5.1. Parametric Disaggregation

5.1.1. Valencia–Schaake (VS) model

In Figure 3, the observed annual NBS data (top panel) and the observed and disaggregated monthly NBS data are presented (bottom panel), indicating that the disaggregated data reproduce the variability of the observed monthly data with a higher maximum than the observed data. Figure 4 presents the basic statistics of the disaggregated monthly data with boxplots and the statistics of the data observed by the dotted line with cross markers. The figure illustrates that the mean and standard deviation are reproduced as well as the extrema (i.e., maximum and minimum), while significant underestimation is found in the skewness.

This underestimation of the skewness results stems from the fact that the gamma marginal distribution employed with the copula transformation was not good enough to reproduce these statistics. Three gamma distributions including the location parameter were also tested to reproduce this statistic. However, the location parameter induces another problem: no smaller values than the location parameter were simulated if the location parameter is greater than zero, and negative values were simulated when it is smaller than zero. Other distributions also have

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

similar problems. Furthermore, the original transformation method, such as box-cox, log, and power transformation, produces a larger problem of generating exceptionally large values and negative values after back-transformation (Jeong and Lee, 2015). Additionally, the extreme statistics are underestimated in several months, as shown in the right middle panel in Figure 4. This might be induced from the underestimation of the skewness by the normal copula transformation with the gamma marginal distribution. High skewness cannot be always reproduced in a gamma distribution, and it affects the magnitude and frequency of extreme events in the disaggregated data.

The marginal cumulative distribution function and probability distribution function for the disaggregated data with the VS model and observed data are presented in Figure 5 and Figure 6, respectively. As shown in Figure 5, high disaggregated values are lower than those observed with the same CDF in most months (i.e., the blue thick solid line is located on the left side of the dotted red line with a cross marker, especially in months 1, 2, 6, 8, and 9). The skewness as well as the maximum of these months are highly underestimated, as shown in the left and right middle panels in Figure 4.

The PDFs in Figure 6 show that the disaggregated marginal distribution does not match the observed one. The observed PDF is more positively skewed than the median of the 200 disaggregated series. This indicates that the disaggregated data have lower skewness than the observed data. Furthermore, the observed PDF presents a thicker tail than the disaggregated median, implying that the disaggregated maximum should be lower than the observed maximum. As shown in Figure 4, the disaggregated data from the VS model often underestimate the maximum of the observed data.

Another critical point of the performance of the VS model is in the lag-1 autocorrelation (ACF1). The ACF1 of the first month is not preserved as in Figure 4. The ACF1 of the first month indicates the relationship between the last month of the previous year and the first month of the current year (i.e., $\mathrm{corr}(Y_{t-1,n_m}, Y_{t,1})$). This is because there is no term considering the previous year in the VS model in Eq. (1). This is one of the major reasons why the MR mode was devised, as in Eq. (2). Therefore, the extremes of the accumulated data are expected to be underestimated.

As shown in Figure 7, the maximum values of the accumulated monthly data up to six months disaggregated with the VS model are underestimated in a number of months. For example, the maximum values of Months 5 and 6 are generally underestimated in most accumulated datasets, as are Months 1 and 2. The reproduction of this statistic for these months is important since the current study started from the 2011 flood that occurred on April 13 and lasted 67 days until June 19. The major cause of this underestimation might be the discontinuity between the previous year and the marginal gamma distribution.

5.1.2. Mejia–Rousselle (MR) model

To avoid the discontinuity of the VS model with the previous year, the MR model includes an additional term by taking the last month of the previous year into account, as in Eq. (2). Its performance improvement can be observed in the lag-1 correlation of the basic statistics in the right bottom panel in Figure 8. The first month of ACF1 was improved. Even if it was just one simple improvement, it implies that the disaggregated data are connected to the previous year. The same behavior as the VS model can be observed for the other statistics since the same normal copula transformation with the gamma marginal distribution was applied to this MR model. In Figure 9, the maximum of the accumulated data presents little improvement compared to that of

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

the VS model shown in Figure 7. The underestimation for Months 5 and 6 still often occurred in the MR model.

5.2. Nonparametric Disaggregation

5.2.1. Original NPD Model

The original NPD model by Lee et al. (2010) was tested. For the NPD model, the results with P_c=0.1 and P_m=0.01 for the GA mixture in Eqs. (10) and (11) were presented. Rather high values of these probabilities produced more diverse scenarios. The results of the key statistics are shown in Figure 10. All statistics were reproduced well from the NPD model except the slight underestimation of ACF1, which was not significant. The slight underestimation of ACF1 was induced from the GA mixture. Lowering the probabilities could lead to less diverse disaggregated scenarios since these probabilities control the magnitude of mixing and mutating the scenarios from the observed sequences. Combined with the additive adjustment, totally new values and patterns could be produced from the NPD model.

The maximum of the accumulated data presents better performance than the parametric model, as shown in Figure 11. The statistics for Month 5 improved in this model, while those for Month 6 were still underestimated. Additionally, the other months, such as Months 2 and 1, also improved. This improvement might be induced from the marginal distribution. As shown in Figure 12 and Figure 13 for the CDF and PDF, respectively, the marginal distribution of the disaggregated monthly data reproduced the observed marginal distribution well, especially comparable to the results of the parametric model in Figure 5 and Figure 6. Since the NPD model does not use parameters, especially for a marginal distribution, the characteristics of the observed marginal

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

distribution were preserved well. This feature remains in the other NPD model (i.e., RB-NPD) as well.

There were some tangible improvements in the original NPD model compared to the parametric model. However, the NPD model always disaggregated the annual data with a 12-month length basis. The applied combinations (called blocks) for a year were always from one of the observed monthly combinations for a year. In other words, one of the 12 month observed combinations must always be taken for the disaggregation (as $y_{t,1}, y_{t,2}, \dots, y_{t,12}$). Even if the GA mixture is applied to overcome this feature, it still suffers from this limitation. This might result in a weak lagged correlation, especially in the early few months, such as Months 1 and 2. Figure 14 presents the lagged correlation starting at each month, i.e., lag-5 correlation at Month 2 is $\mathrm{corr}(Y_{t-1,n_m-4}, Y_{t,2})$. The lagged correlations at Months 1-4 were underestimated in a number of lags, while those of the other months were preserved well.

5.2.2. *Proposed RB-NPD Model*

To avoid the fixed length and the effects on the lagged correlation, the random block-based NPD model (i.e., RB-NPD) was devised in the current study. Instead of the fixed 12-month block by Lee et al. (2010), the block length was randomly selected from a Poisson distribution as in Eq.(14). The parameter α=12 was used to make the average of the block length the same as the original NPD model. This random block allows the change point of the block to be different from the first month of a year in the original NPD model. By changing the block length into a random block, the distance and its related covariance matrix must be changed at each month. Furthermore, the adjustment to meet the additive condition must be made whenever the random block length overrides the following year.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

The key statistics of the disaggregated data with the RB-NPD model are presented in Figure 15. Note that relatively high probabilities of crossover and mutation for the GA mixture were employed as P_m=0.1 and P_c=0.3 to produce diverse disaggregated scenarios. Further testing was performed to check whether the employed probabilities were feasible with root mean square error (RMSE), especially the maximum for accumulated data. One of the major objectives was to produce extreme events such as the 2011 flood that had consistently high NBS values for a long period.

As shown in Figure 16, increasing the crossover probability (Pc) did not lower the RMSE. Since a high Pc can be beneficial for producing diverse scenarios, P_c =0.3 can be acceptable. In addition, a lower P_m value might be better for all accumulated data except Acc-2 shown in the panel in Figure 16(b). With P_c =0.3, P_m =0.1 is the best choice for Acc-1 (see the blue dotted line with circle in Figure 16 (a)). Additionally, the choice of P_m =0.1 is an acceptable choice, showing the second lowest RMSE for all accumulated data with P_c =0.3. Note that P_m =0.01 can be a good alternative, but this might lower the role of the ECAco and PSAm algorithms in the model enhancement. Therefore, P_m=0.1 and P_c=0.3 were applied to the following results. Diverse values of P_m and P_c were also tested in the simulation, and no significantly better results were found.

All observed statistics were preserved well except ACF1. The underestimation of ACF1 was induced by the high probabilities of mutation and crossover. The enhancement was made by choosing high correlation in the crossover as in Eq. (20), denoted as the enhanced correlation algorithm in the crossover (ECAco, see the model development section). Note that $Z_{t,\tau-1}Z_{t,\tau} + Z_{t,\tau}Z_{t,\tau+1}$ indicates the correlation since for variables a and b, corr(a,b)=cov(a,b)/std(a)std(b) and std(a) and std(b) are one and mean(a) and mean(b) are zero for standard normal variables (i.e.,

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

$\mathrm{corr}(a, b) = E[a \cdot b]$). Further enhancement was made in the mutation by replacing the monthly sequence with the sequence generated by the parametric simulation algorithm, called PSAm.

The basic statistics of the enhanced RB-NPD model with ECAco and PSAm are presented in Figure 17. The figure shows that all statistics were reproduced well, including ACF1. Even slight overestimation of ACF1 is shown. The maximum of the accumulated data in Figure 18 was better preserved in this model compared to the original NPD model in Figure 11, especially Months 5 and 6, which were the most critical months when floods occurred in the LCRR Basin. Furthermore, the lagged correlation at each month in Figure 19 was better preserved than that of the NPD model in Figure 14. In particular, the lagged correlations of Months 1, 2, and 3 improved in the enhanced RB-NPD model.

Further model testing was performed with wavelet analysis (Foufoula-Georgiou and Kumar, 1994; Grinsted et al., 2004) for the RB-NPD models (1) without enhancement (Basic), with ECAco only (Selective), and with ECAco and PSAm (Hybrid). The magnitude-squared coherences (C^2) between the observed data and the disaggregated data from the RB-NPD models are presented in Figure 20. At lower frequencies, strong coherences can be observed, and this is rational since the aggregated data (annual NBS) of the disaggregated data (monthly NBS) are the same as those observed from the additive condition. Figure 21 presents the magnitude-squared coherence for the selected frequencies with high magnitudes. This illustrates that the selective and hybrid models show higher coherence than the basic model and shows that the selective and hybrid algorithms mimic the spectral frequency behavior of the observed data. The results indicate that the proposed algorithm can be a reasonable alternative to disaggregate the annual NBS data to monthly or quarter-monthly scale data.

6. Summary and Conclusions

Based on the 2011 flood in the LCRR Basin, the assessment of the existing flood protection structure and future mitigation frameworks requires simulation scenarios. Furthermore, the scenarios must be on a monthly or quarter-monthly scale. Here, the disaggregation model development was made in the current study focusing on preserving the consistent extreme event of the 2011 flood that was sustained for more than three months.

The existing parametric models, VS and MR, and the nonparametric model of the NPD were tested. The VS and MR models employed the normal copula transformation with the gamma marginal distribution instead of the traditional log, box-cox, or power transformation to avoid producing exceptionally large values and negative values. The results were reasonable, but the skewness and maximum statistics were underestimated. Furthermore, the maximum of the accumulated data was not reproduced well, especially for Months 5 and 6, which were critically related to flood events in the LCRR Basin. In contrast, the NPD model reproduced all basic statistics, including skewness and maximum statistics. However, the lagged correlation at each month was underestimated, especially in the first few months (i.e., Months τ=1, 2, and 3) due to the fixed number of blocks, 12 months, in the disaggregation procedure. Additionally, the maximum of the accumulated data was not appropriately reproduced, especially in Month 6.

To overcome the shortages induced by the fixed number of blocks, the random block was suggested in the NPD model as the RB-NPD model. The proposed RB-NPD model varies the starting month of the block, while the starting month of the original NPD model for the block is always Month 1 (τ=1). Further model enhancement was made in the GA mixture to improve the cross-correlation by adding the ECAco and PSAm algorithms. The enhanced RB-NPD model

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

reproduced the lagged correlation as well as the maximum of the accumulated data, especially in Months 5 and 6, which are critical for the purpose of the current disaggregation model. Furthermore, the disaggregated data with the enhanced RB-NPD model present better preservation of the lagged correlation at each month than the NPD model, as well as spectral coherence.

Therefore, the results indicate that the proposed RB-NPD model can be a comparable alternative and that its enhancement is suitable for disaggregating the annual NBS data for the LCRR Basin.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

Abbreviations

Abbreviations

LCRR	Lake Champlain–Richelieu River
NBS	Net Basin Supply
VS	Valencia and Schaake
MR	Mejia and Rousselle
NPD	Nonparametric Disaggregation
GA	Genetic Algorithm
KNNR	K-Nearest Neighbor Resampling
CDF	Cumulative Distribution Function
PDF	Probability Density Function
RB-NPD	Random Block-based Nonparametric Disaggregation
PSAm	Parametric Simulation Algorithm in mutation
ECAco	Enhanced Correlation Algorithm in crossover

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions

7. Code and data availability

The model code and example data are available at Mendeley Data in <https://data.mendeley.com/datasets/jrrwbc4cx6/1>. The net basin supply (NBS) observed data are available from the International Joint Commission (IJC, https://www.ijc.org/en) upon request

Competing interests

The author declares that they have no conflict of interest.

Author Contribution

L.T. performed writing and collection data as well as programming while O.T. carried out the research plan and supervising.

Acknowledgement

The first author acknowledges that this research was partially supported by a grant (2022-MOIS63-001) of Cooperative Research Method and Safety Management Technology in National Disaster funded by Ministry of Interior and Safety(MOIS, Korea).

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

References

Bras, R.L., Rodriguesz-Iturbe, I., 1994. Random Functions and Hydrology. Dover Books on Advanced Mathematics. Dover, New York, 559 pp.

Croley, T.E., II, Lee, D.H., 1993. EVALUATION OF GREAT LAKES NET BASIN SUPPLY FORECASTS. JAWRA Journal of the American Water Resources Association, 29(2): 267-282. DOI:10.1111/j.1752-1688.1993.tb03207.x

Fagherazzi, L., Salas, J.D., Sveinsson, O., 2011. Stochastic Modeling and Simulation of the Great Lakes System, International Upper Great Lakes Quebec.

Foufoula-Georgiou, E., Kumar, P. (Eds.), 1994. Wavelets in Geophysics. Wavelet analysis and its applications. Academic Press, San Diego, CA, 373 pp.

Grinsted, A., Moore, J.C., Jevrejeva, S., 2004. Application of the cross wavelet transform and wavelet coherence to geophysical times series. Nonlinear Processes in Geophysics, 11(5-6): 561-566.

Jeong, C., Lee, T., 2015. Copula-based modeling and stochastic simulation of seasonal intermittent streamflows for arid regions. Journal of Hydro-Environment Research, 9(4): 604-613. DOI:10.1016/j.jher.2014.06.001

Koutsoyiannis, D., Manetas, A., 1996. Simple disaggregation by accurate adjusting procedures. Water Resources Research, 32(7): 2105-2117.

Lall, U., Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resources Research, 32(3): 679-693.

Lane, W.L., Frevert, D.K., 1990. Applied Stochastic Techniques, Personal Computer, Version 5.2, User's Manual, U.S. Bureau of Reclamation, Denver, Colorado.

Lee, T., Jeong, C., 2014. Nonparametric statistical temporal downscaling of daily precipitation to hourly precipitation and implications for climate change scenarios. Journal of Hydrology, 510: 182-196. DOI:10.1016/j.jhydrol.2013.12.027

Lee, T., Ouarda, T.B.M.J., 2011. Identification of model order and number of neighbors for k-nearest neighbor resampling. Journal of Hydrology, 404(3-4): 136-145.

Lee, T., Ouarda, T.B.M.J., 2012. Stochastic simulation of nonstationary oscillation hydroclimatic processes using empirical mode decomposition. Water Resources Research, 48(2). DOI:doi:10.1029/2011WR010660

Lee, T., Ouarda, T.B.M.J., 2019. Multivariate Nonstationary Oscillation Simulation of Climate Indices With Empirical Mode Decomposition. Water Resources Research, 55(6): 5033-5052. DOI:10.1029/2018WR023892

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Lee, T., Ouarda, T.B.M.J., Yoon, S., 2017. KNN-based local linear regression for the analysis and simulation of low flow extremes under climatic influence. Climate Dynamics, 49(9-10): 3493-3511. DOI:10.1007/s00382-017-3525-0

Lee, T., Park, T., 2017. Nonparametric temporal downscaling with event-based population generating algorithm for RCM daily precipitation to hourly: Model development and performance evaluation. Journal of Hydrology, 547: 498-516. DOI:10.1016/j.jhydrol.2017.01.049

Lee, T., Salas, J.D., Prairie, J., 2010. An enhanced nonparametric streamflow disaggregation model with genetic algorithm. Water Resources Research, 46(8). DOI:10.1029/2009WR007761

Lee, T., Singh, V.P., 2018. Statistical Downscaling for Hydrological and Environmental Applications, 1. CRC press, Boca Raton, FL, 181 pp.

Lee, T.S., 2008. Stochastic simulation of hydrologic data based on nonparametric approaches, Ph. D. Dissertation, Colorado State University, Fort Collins, CO., USA, 346 pp.

Mejia, J.M., Rousselle, J., 1976. Disaggregation Models in Hydrology Revisited. Water Resources Research, 12(2): 185-186.

Ouarda, T.B.M.J., Charron, C., 2019. Frequency analysis of Richelieu River flood flows, Lake Champlain flood level and NBS to the Richelieu River basin, Final project report, INRS-ETE, Quebec, Canada.

Porter, J.W., Pink, B.J., 1991. A method of synthetic fragments for disaggregation in stochastic data generation, Int. Hydrol. and Water Resour. Symp. The Institution of Engineers, Australia, Canberra, pp. 187-191.

Prairie, J., Rajagopalan, B., Lall, U., Fulp, T., 2007. A stochastic nonparametric technique for space-time disaggregation of streamflows. Water Resources Research, 43(3): W03432. DOI:W03432

10.1029/2005wr004721

Santos, E.G., Salas, J.D., 1992. Stepwise Disaggregation Scheme for Synthetic Hydrology. Journal of Hydraulic Engineering-Asce, 118(5): 765-784.

Srikanthan, R., Mcmahon, T.A., 1982. Simulation of Annual and Monthly Rainfalls - a Preliminary-Study at 5 Australian Stations. Journal of Applied Meteorology, 21(10): 1472-1479.

Stedinger, J.R., Vogel, R.M., 1984. Disaggregation Procedures for Generating Serially Correlated Flow Vectors. Water Resources Research, 20(1): 47-56.

Tarboton, D.G., Sharma, A., Lall, U., 1998. Disaggregation procedures for stochastic hydrology based on nonparametric density estimation. Water Resources Research, 34(1): 107-119.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

Valencia, D., Schaake, J.C., 1973. Disaggregation Processes in Stochastic Hydrology. Water Resources Research, 9(3): 580-585.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Figures

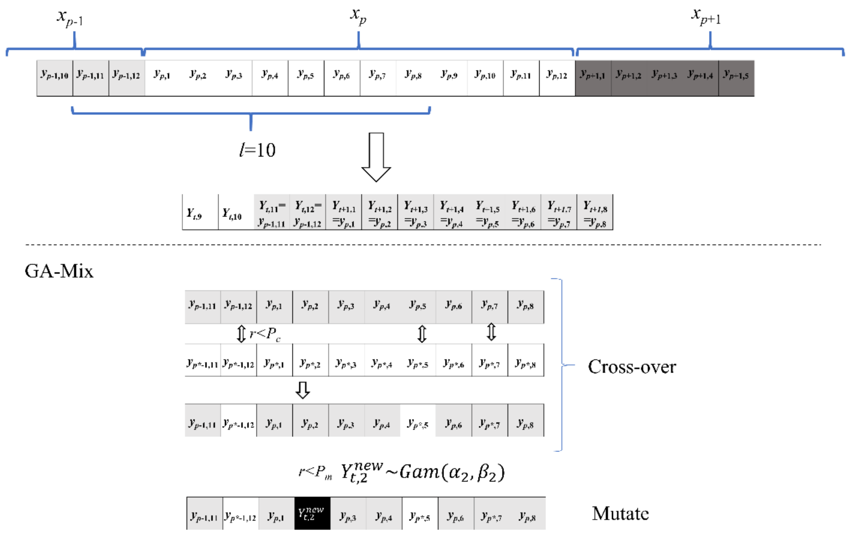
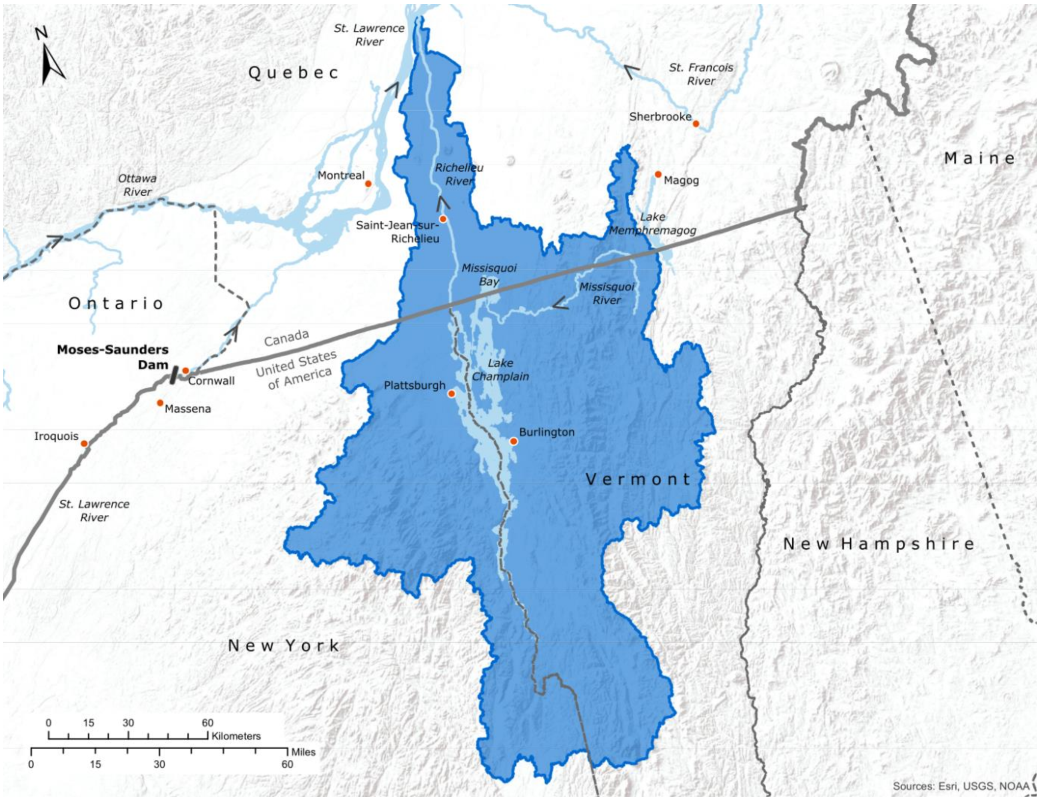


Figure 1. Diagram of the proposed random block-based nonparametric disaggregation (RB-NPD) from the current study.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

573



574 Figure 2. Map of the LCRR Basin. Note that dark blue represents the whole area of the LCRR
575 Basin, light blue inside dark blue represents Lake Champlain, and the northward line from Lake
576 Champlain represents the Richelieu River. Note that the map was provided by the International
577 Joint Commission.

578

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions
Open Access

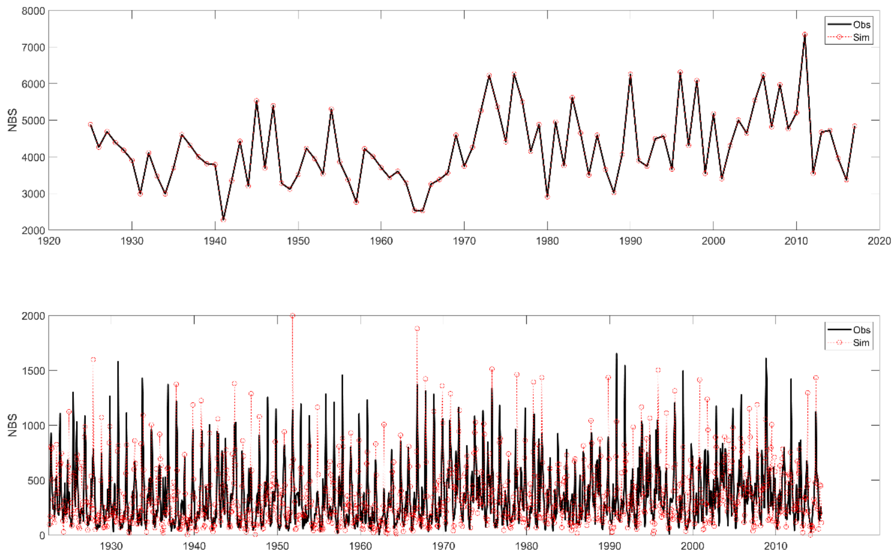


Figure 3. Time series of the annual (top panel) and monthly (bottom panel) data for the observed (thick solid line) and disaggregated (dotted line with circle) simulation data with the VS model.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions
Open Access

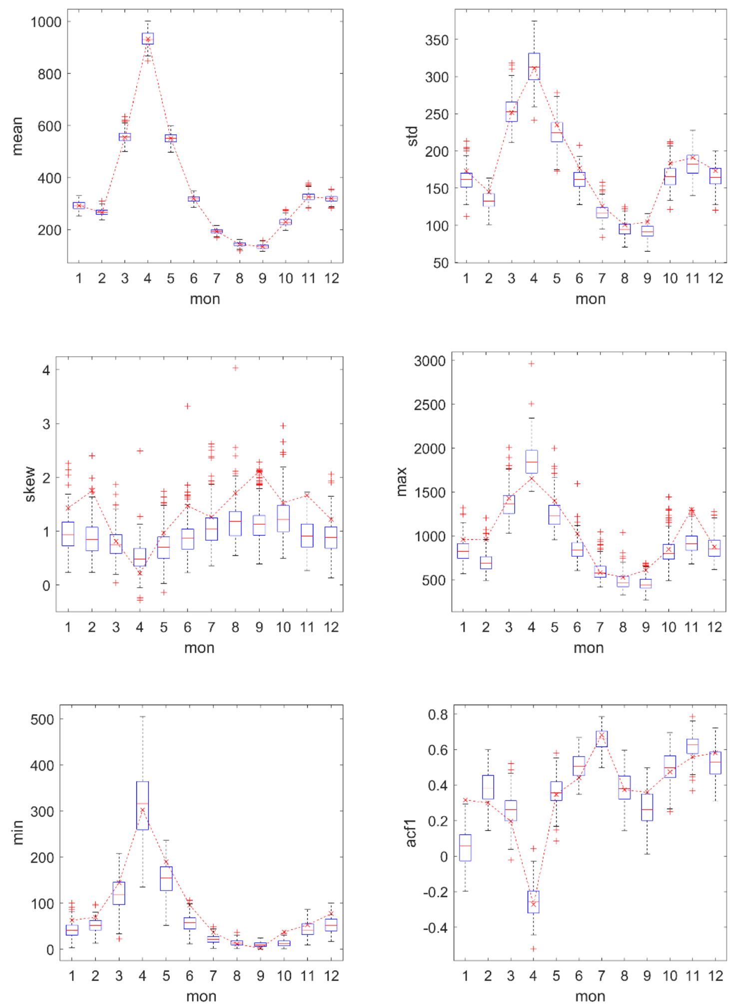
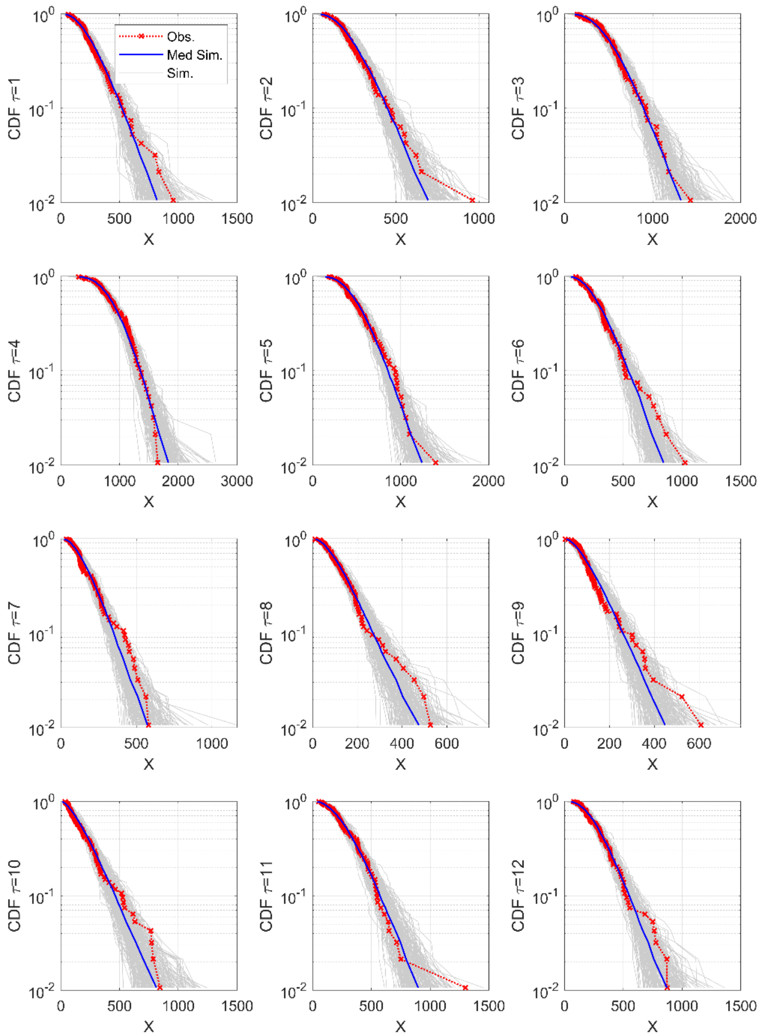


Figure 4. Boxplots of the basic statistics of the disaggregated monthly data from the annual data with the VS model for the NBS of the LCRR basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.). The boxes represent the interquartile range (IQR), and the whiskers extend to 1.5 IQR. The horizontal lines inside the boxes depict the data median. Data beyond the whiskers (1.5 IQR) are shown by a plus sign (+).

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions



589

590 Figure 5. Cumulative distribution function (CDF) of the disaggregated data in each month with
591 the VS model from the annual NBS of the LCRR Basin. Note that the observed data are
592 represented with the dotted line and cross marker (.x.); (2) the 200 disaggregated simulation
593 series are shown with thin gray lines, while their median is represented with the thin blue line;
594 and (3) τ indicates the month from 1 to 12.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

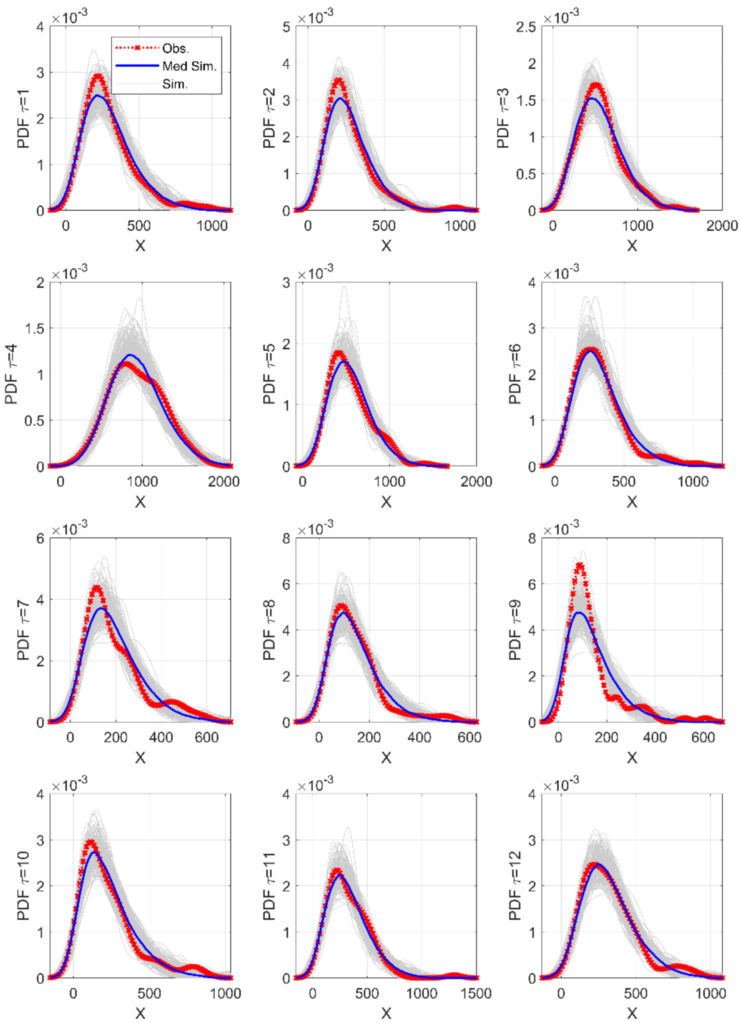
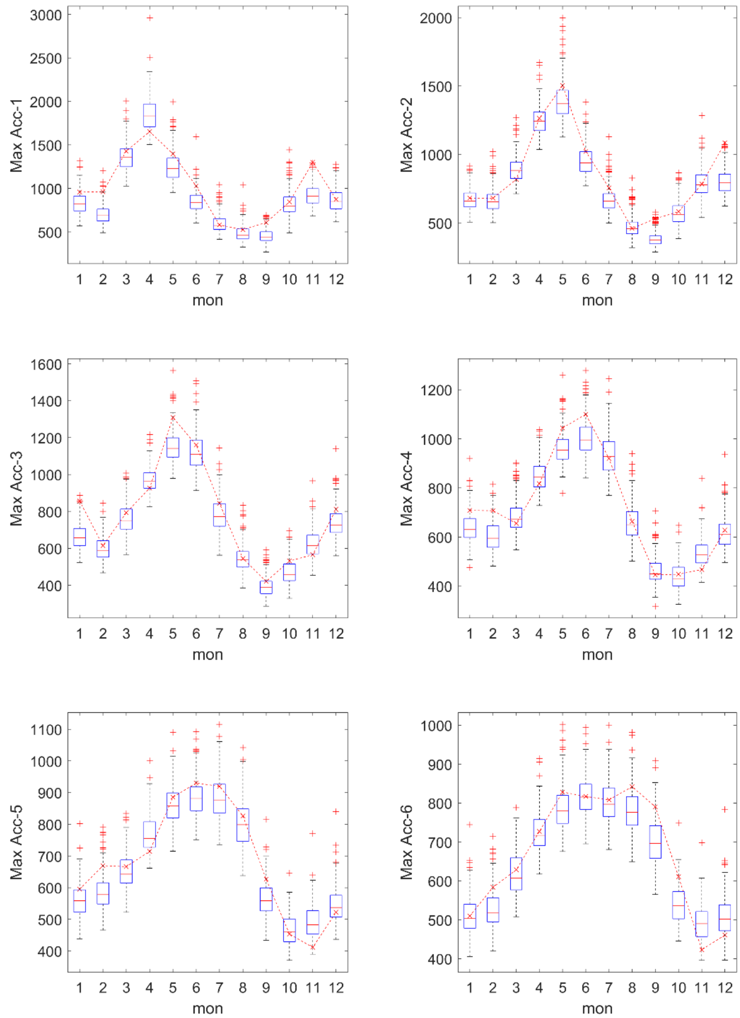


Figure 6. Probability density function (PDF) of the disaggregated data at each month with the VS model with P_m=0.01 and P_c=0.1 from the annual NBS of the LCRR Basin. Note that the observed data are represented with the dotted line and cross marker (.x.); and (2) the 200 disaggregated simulation series are shown with thin gray lines, while their median is represented with the thick blue line.



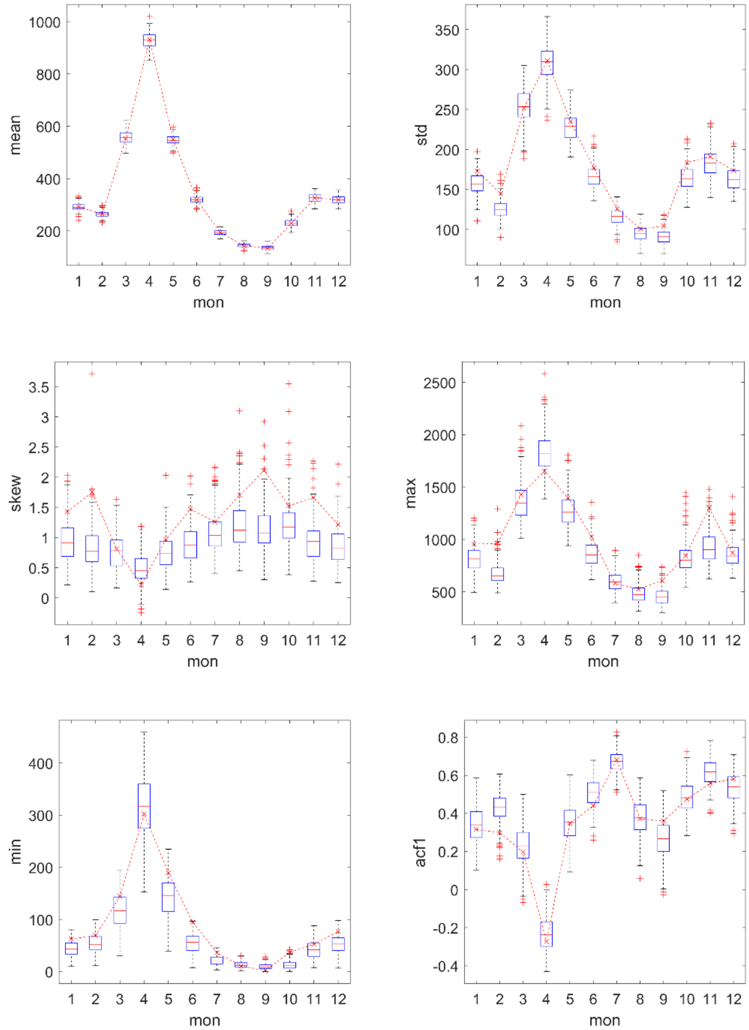
601

602 Figure 7. Boxplots of the maximum of the accumulated data for 1-6 months at each month of the
603 disaggregated data with the VS model from the annual to the monthly NBS of the LCRR Basin.
604 Note that the statistics of the observed data are also represented with the dotted line and cross
605 marker (.x.); (2) the accumulation was performed for the previous months. For example, the acc-
606 4 data at Month 6 are obtained by summing the monthly data of 6, 5, 4, and 3 months.

607
608

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions
Open Access



609
610 Figure 8. Same as Figure 4 but for downscaled data from the MR model.
611
612

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions
Open Access

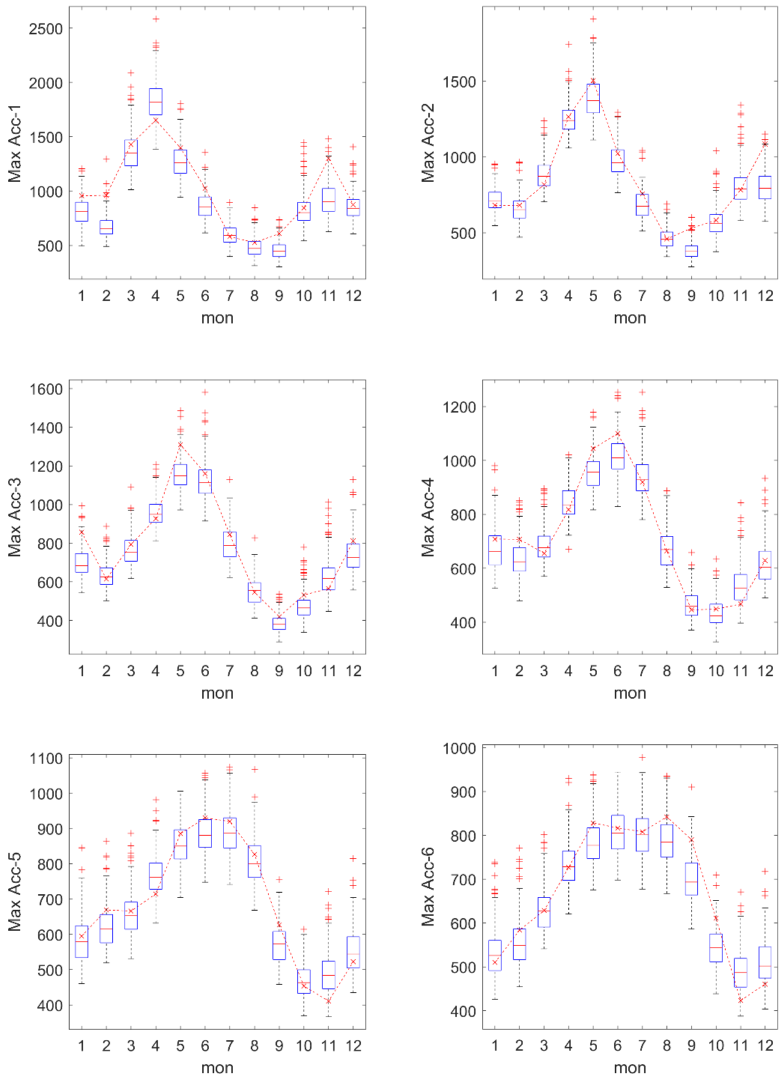


Figure 9. Same as Figure 7 but for downscaled data from the MR model.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions
Open Access

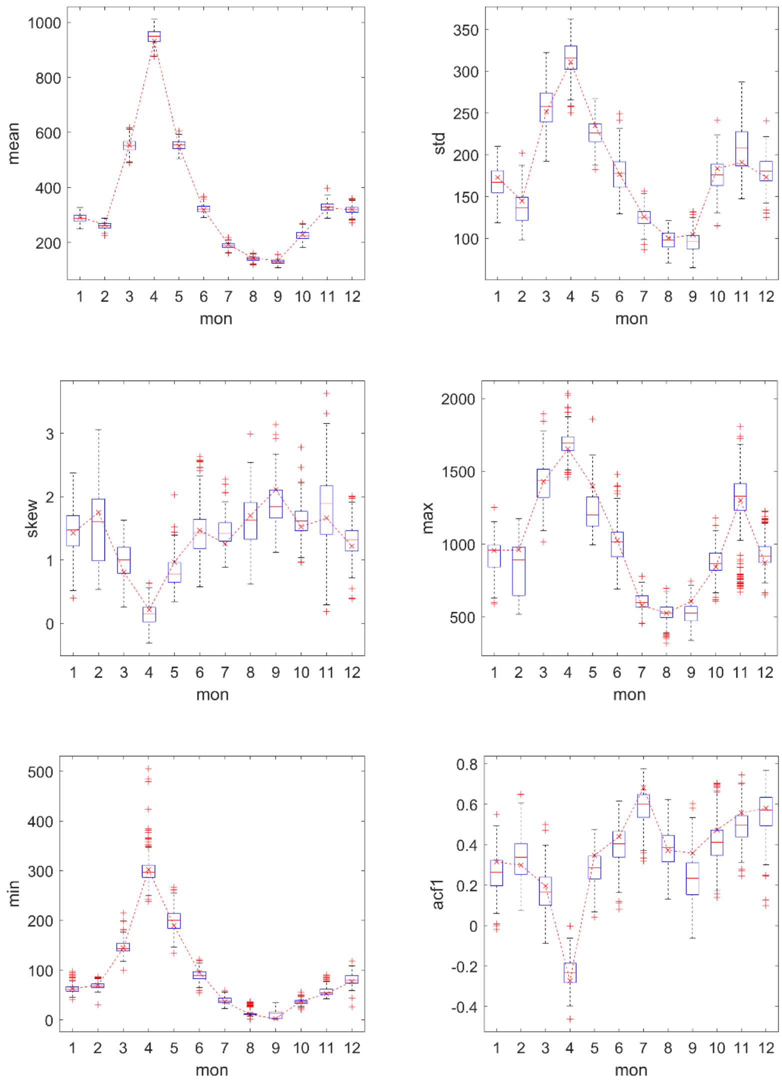


Figure 10. Boxplots of the basic statistics of the disaggregated monthly data from the annual data with the NPD model with P_c=0.1 and P_m=0.01 for the NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.).

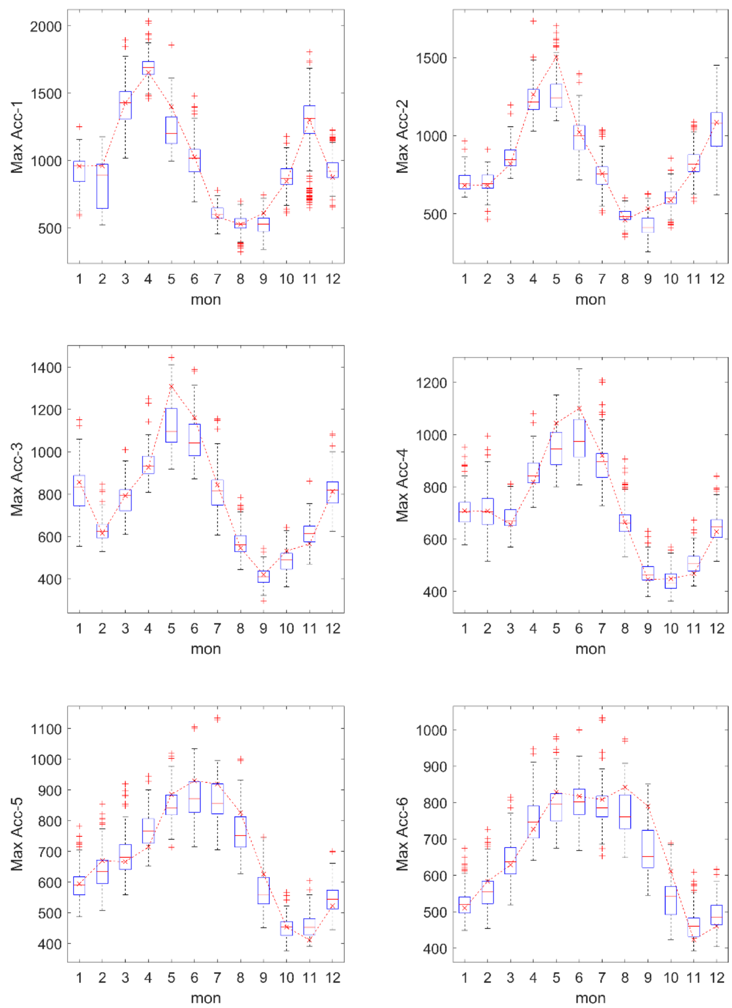


Figure 11. Boxplots of the maximum of the accumulated data for 1-6 months at each month of the disaggregated data by the NPD model with P_c=0.1 and P_m=0.01 from the annual to the monthly NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.); and (2) the accumulation was made for the previous months. For example, the acc-4 data at Month 6 are obtained by summing the monthly data of 6, 5, 4, and 3 months.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

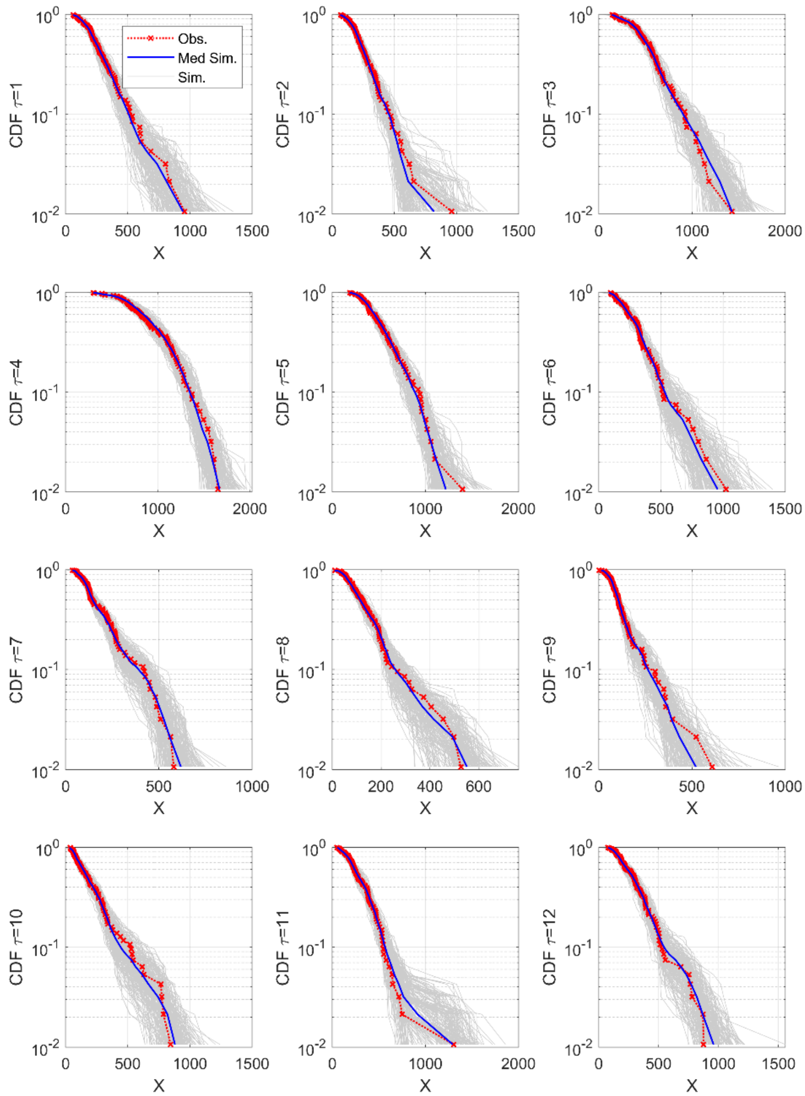


Figure 12. Cumulative distribution function (CDF) of the disaggregated data at each month with the NPD model with P_m=0.01 and P_c=0.1 from the annual NBS of the LCRR Basin. Note that the observed data are represented with the dotted line and cross marker (.x.); and (2) the 200 disaggregated simulation series are shown with the thin gray lines, while their median is represented with the thick blue line.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

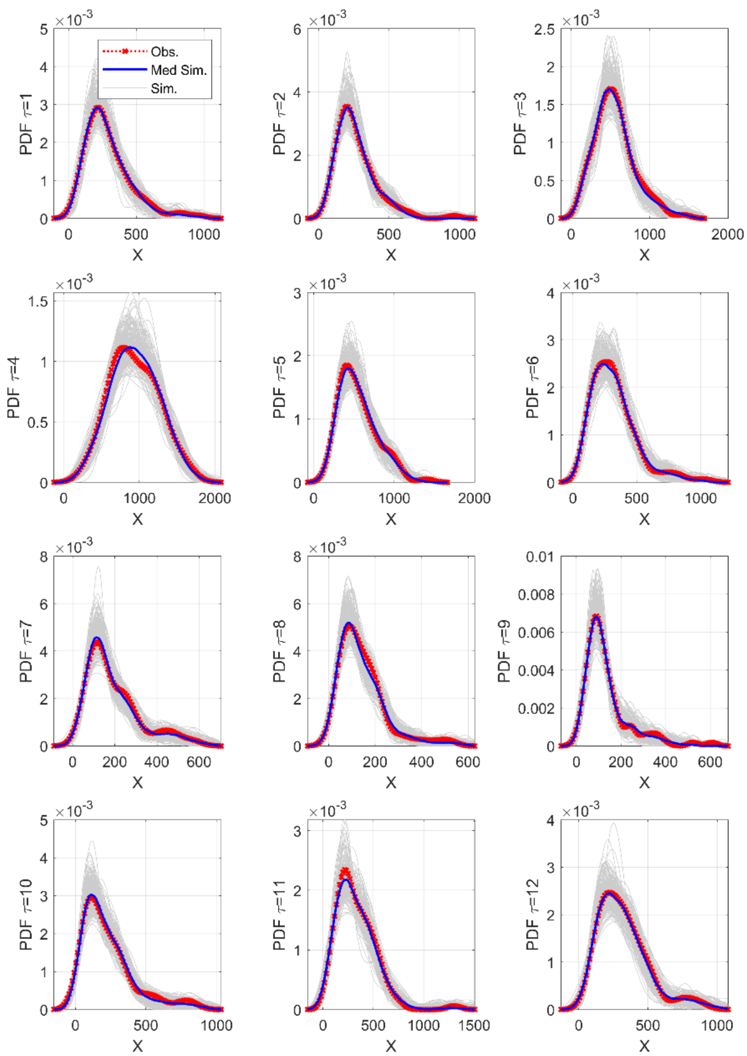


Figure 13. Probability density function (PDF) of the disaggregated data at each month with the NPD model with P_m=0.01 and P_c=0.1 from the annual NBS of the LCRR Basin. Note that the observed data are represented with the dotted line and cross marker (.x.); and (2) the 200 disaggregated simulation series are shown with the thin gray lines, while their median is represented with the thick blue line.

Geoscientific Model Development
Discussions
Open Access

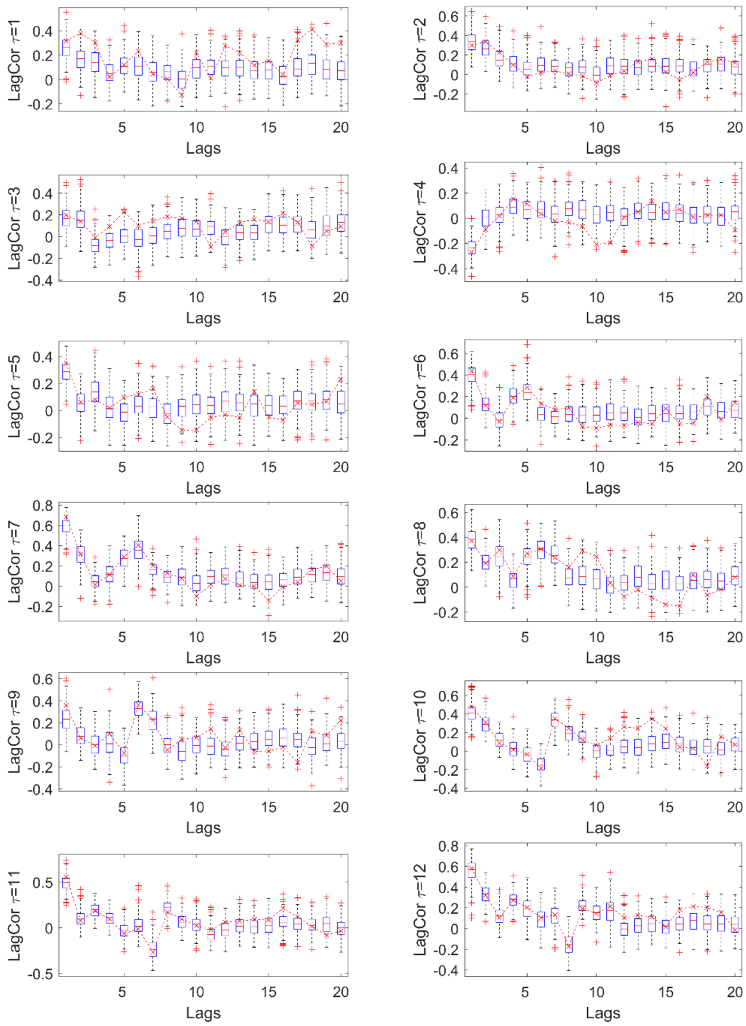


Figure 14. Boxplots of the lagged correlation for the disaggregated monthly data with the NPD model with P_m=0.01 and P_c=0.1 from the annual NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.); and (2) the lagged correlation was estimated at each month. For example, the lag-2 correlation at τ=1 was estimated with the Month-1 data of the current year and the Month-11 data of the previous year.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

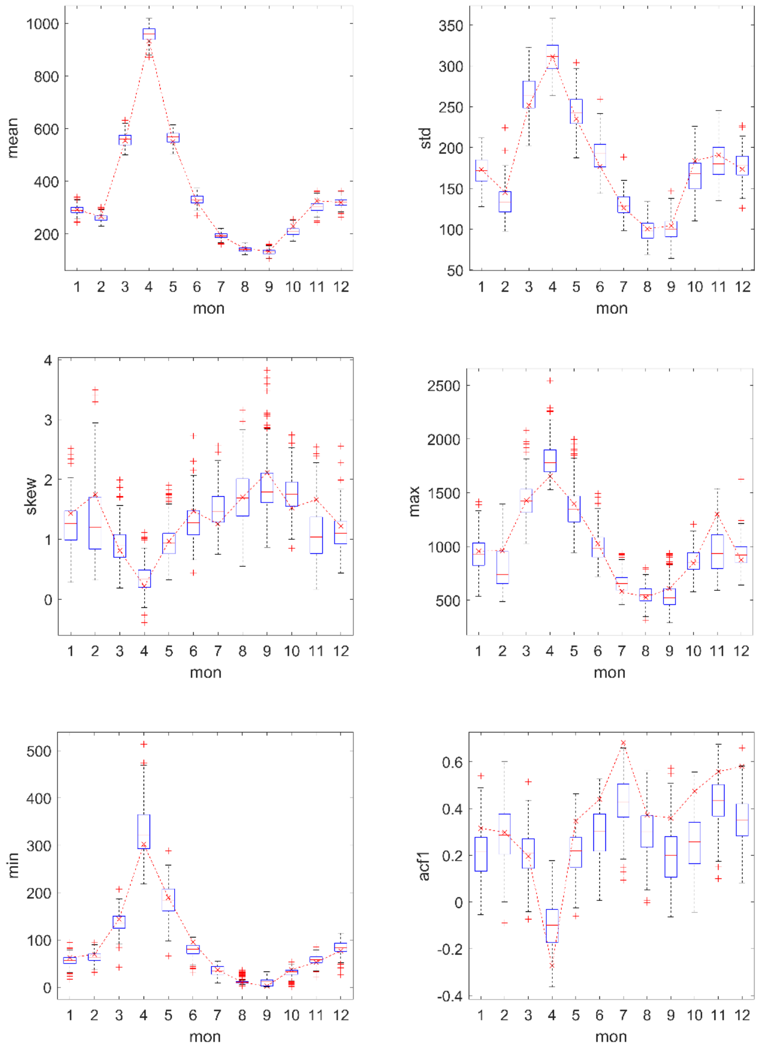


Figure 15. Boxplots of the basic statistics of the disaggregated monthly data from the annual data with the RB-NPD model with P_c=0.3 and P_m=0.1 for the NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.).

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

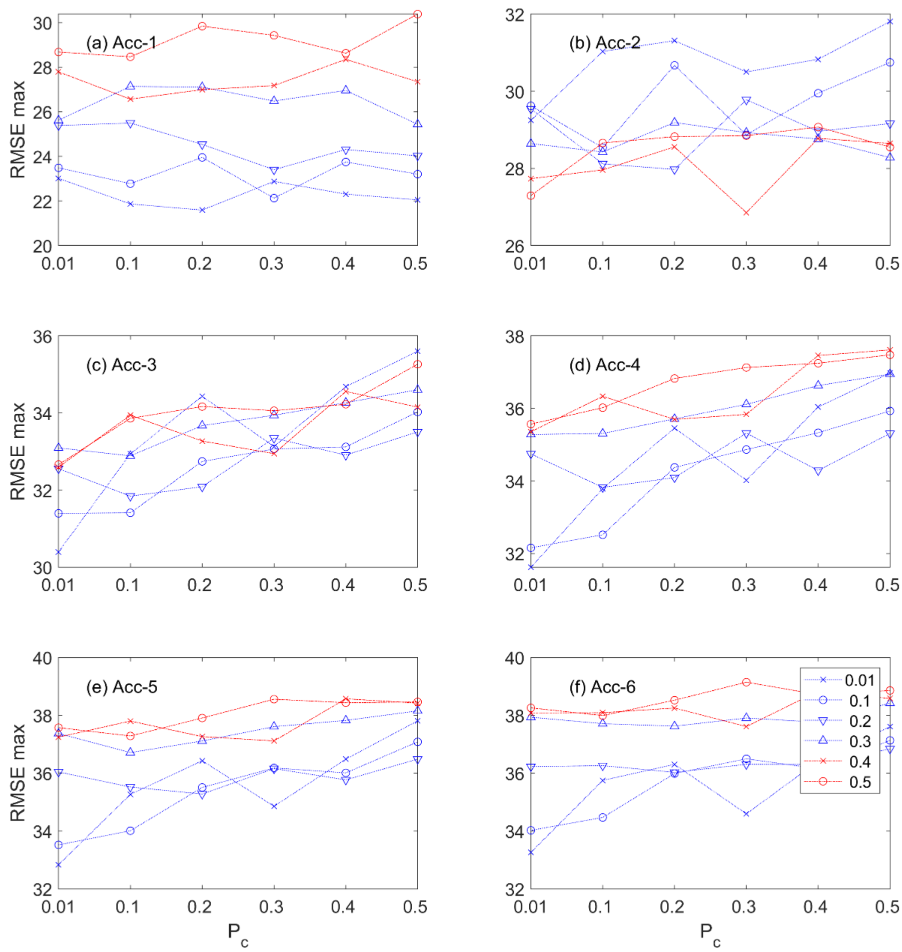


Figure 16. Root mean square error between the observed maximum values for the average of the accumulated monthly NBS and the simulated maximum values with different crossover and mutation probabilities of 0.01, 0.1, 0.2, 0.3, 0.4, and 0.5.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions

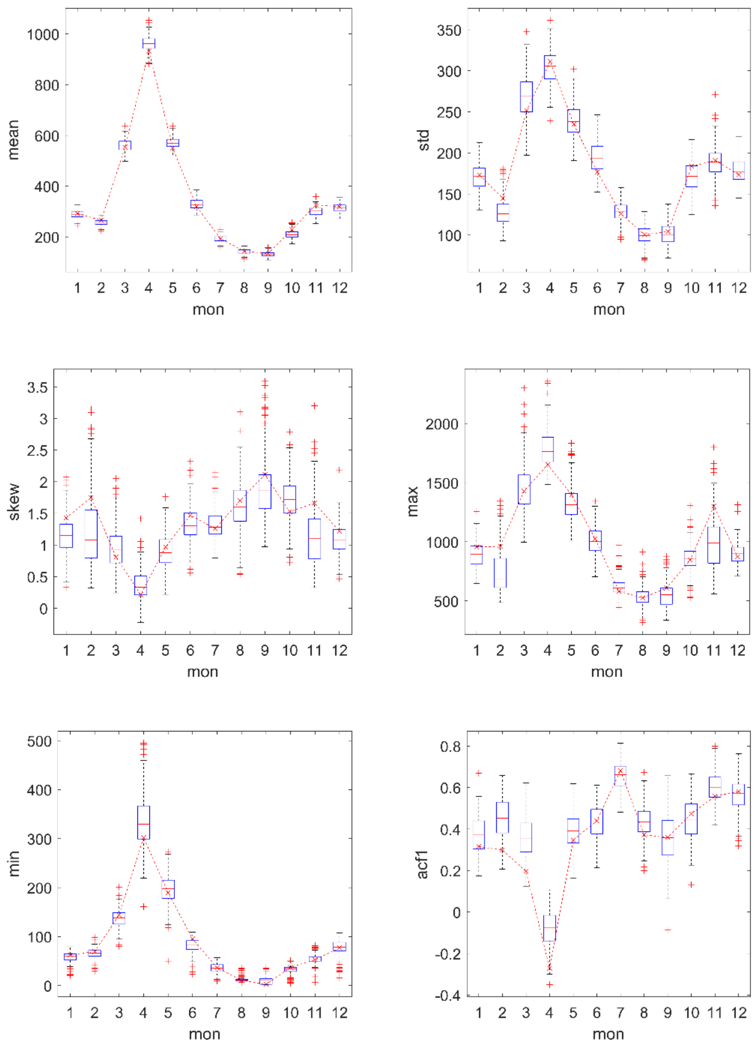


Figure 17. Boxplots of the basic statistics of the disaggregated monthly data from the annual data for the RB-NPD model with ECAco+PSAm (hybrid model) as well as P_c=0.3 and P_m=0.1 for the NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.).

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

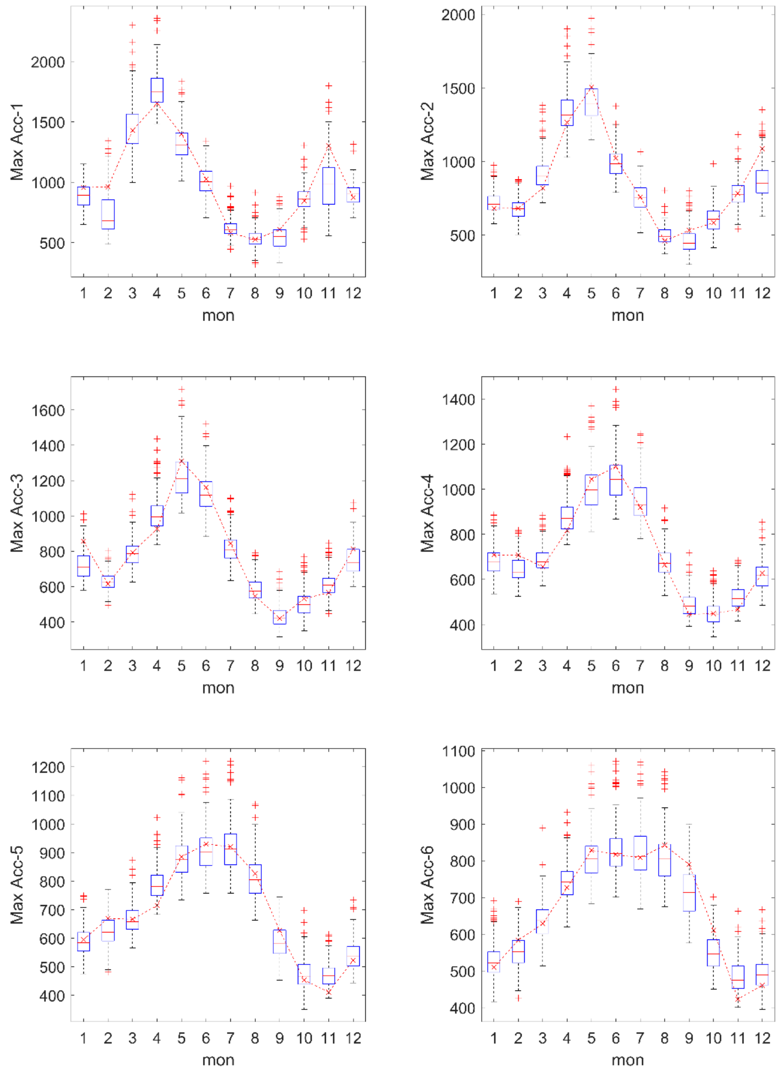


Figure 18. Boxplots of the maximum of the accumulated data for 1-6 months at each month of the disaggregated data from the RB-NPD model with the ECAco+PSAm as well as P_c=0.3 and P_m=0.1 from the annual to the monthly NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.); and (2) the accumulation was made for the previous months. For example, the acc-4 data at Month 6 are obtained by summing the monthly data of 6, 5, 4, and 3 months.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

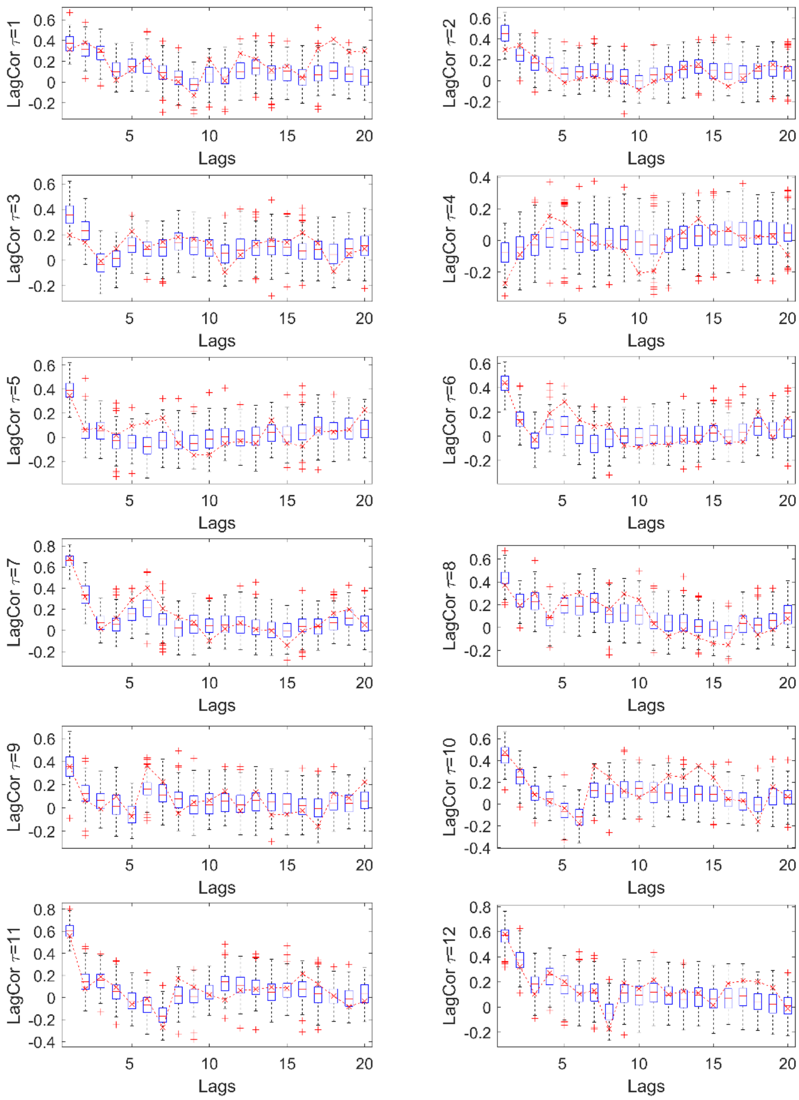


Figure 19. Boxplots of the lagged correlation for the disaggregated monthly data for the RB-NPD model with the ECAco+PSAm as well as P_c=0.3 and P_m=0.1 from the annual NBS of the LCRR Basin. Note that the statistics of the observed data are also represented with the dotted line and cross marker (.x.); and (2) the lagged correlation was estimated at each month. For example, the lag-2 correlation at τ=1 was estimated with the Month-1 data of the current year and the Month-11 data of the previous year.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development
Discussions

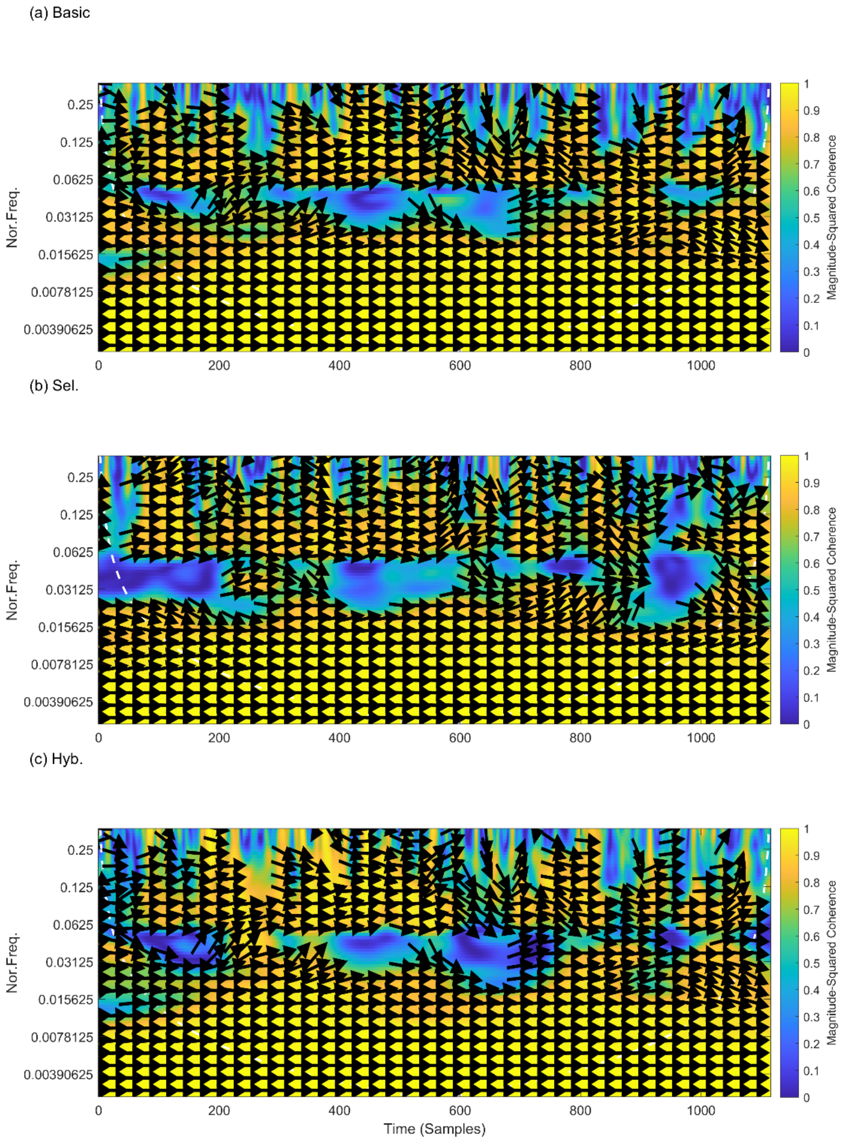


Figure 20. Magnitude-squared coherence (C^2) of all frequencies between the observed monthly NBS and the example of the disaggregated data from the RB models of the (a) basic, (b) selective, and (c) hybrid algorithms. Note that (1) $\mathrm{C}^2(f) = |S_{xy}(f)|/S_{xx}(f) \cdot S_{yy}(f)$, where $S_{ab}(f)$ is the cross power spectrum of two signals, *a* and *b*, at frequency *f*; (2) very strong coherence can be seen in lower normalized frequencies; (3) lower frequency indicates long-term variability; and (4) this is sound since the disaggregated data have the same annual values from the additive condition.

https://doi.org/10.5194/gmd-2022-274
Preprint. Discussion started: 26 January 2023

Geoscientific Model Development Discussions
Open Access

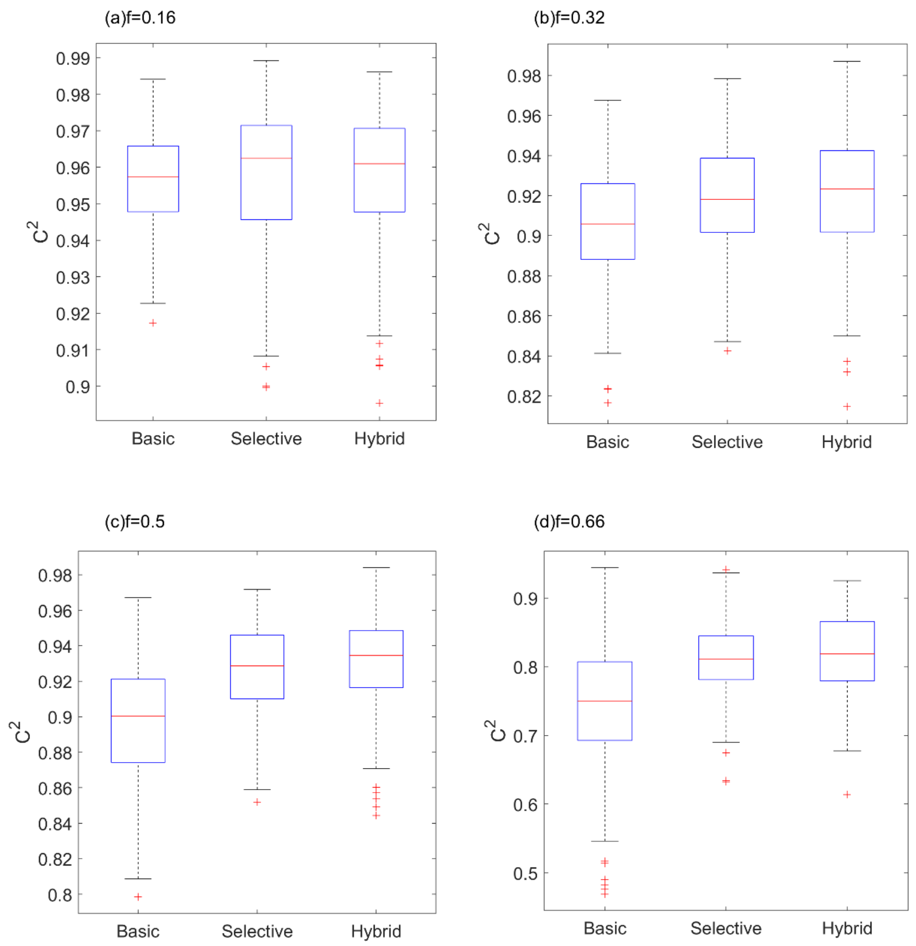


Figure 21. Magnitude-squared coherence (C^2) of selected high frequencies (f=0.16, 0.32, 0.5, and 0.66) between the observed monthly NBS and the disaggregated from the RB models of the basic, selective, and hybrid models. Note that high coherence indicates that the model mimics the spectral frequencies of the observed data well.