Synthetic simulation of spatially-correlated streamflows: Weighted-modified Fractional Gaussian Noise

Cristián Chadwick[1], Frederic Babonneau[2,3], Tito Homem-de-Mello[4], Agustín Letelier[1]

[1]Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, Diagonal Las Torres 2640, Peñalolén, Santiago, Chile

[2]Kedge Business School, 680 Cr de la Libération, 33405 Talence, France

[3]ORDECSYS, 4 Place de l'Etrier, CH-1224 Chêne-Bougeries, Suisse

[4]School of Business,Universidad Adolfo Ibáñez, Diagonal Las Torres 2640, Peñalolén, Santiago, Chile

Key Points:

- We propose a Weighted modified Fractional Gaussian Noise (WmFGN) model that addresses both temporal and spatial correlations simultaneously.
- The method searches for an optimal convex combination of the spatial and temporal correlation matrices according on the user's priority.
- Our results on a Chilean basin demonstrate that WmFGN represents a significant improvement over existing methods in preserving correlations.

Corresponding author: Cristián Chadwick, `cristian.chadwick@uai.cl`

Abstract

Stochastic methods have been typically used for the design and operations of hydraulic infrastructure. They allow decision makers to evaluate existing or new infrastructure under different possible scenarios, giving them the flexibility and tools needed in decision making. In this paper, we present a novel stochastic streamflow simulation approach able to replicate both temporal and spatial dependencies from the original data in a multi-site basin context. The proposed model is a multi-site extension of the modified Fractional Gaussian Noise (mFGN) model which is well-known to be efficient to maintain periodic correlation for several time lags, but presents shortcomings in preserving the spatial correlation. Our method, called Weighted-mFGN (WmFGN), incorporates spatial dependency into streamflows simulated with mFGN by relying on the Cholesky decomposition of the spatial correlation matrix of the historical streamflow records. As the order in which the decomposition steps are performed (temporal then spatial, or vice-versa) affects the performance in terms of preserving the temporal and spatial correlation, our method searches for an optimal convex combination of the resulting correlation matrices. The result is a Pareto-curve that indicates the optimal weights of the convex combination depending on the importance given by the user to spatial and temporal correlations. The model is applied to Bio-bio River basin (Chile), where the results show that the WmFGN maintains the qualities of the single-site mFGN, while significantly improving spatial correlation.

1 Introduction

Stochastic methods have been typically used to improve and evaluate the design and operation of existing or new hydraulic infrastructures, e.g., the evaluation of reservoir performance using stochastic streamflows by Hashimoto et al. (1982). Stochastic streamflow generation allows the evaluation of infrastructure, under different scenarios, of usage, public policies, operation, and even under climate change conditions (Kirsch et al., 2013). Due to the new challenges that water resources are facing, such as climate change, as well as changes in public policies in the changing world, robust stochastic methods able to simulate synthetic streamflows consistent with historical records, capable of incorporating possible changes are required.

Multiple synthetic streamflow generation models have been developed in the literature, to answer both to scientific and decision maker needs of scenarios evaluation. These stochastic generation models work at a single or multi-site scale to replicate the statistical behaviour of streamflows. The advantage of multi-site models is that they can evaluate scenarios over an entire basin at the same time. There have been several discussions in literature to determine the most complete stochastic multi-site streamflow model, without getting to consensus (Srinivas & Srinivasan, 2005). To the best of our knowledge, methods always fail for simulating muti-site streamflows on at least one dimension, e.g., temporal correlation, capture of seasonality, spatial correlation, or long-run dependencies. In this paper, we present a novel stochastic streamflow simulation approach able to replicate both temporal and spatial dependencies from the original data in a multi-site basin context.

Initial studies in stochastic hydrology were based on booststrap techniques (Efron & Tibshirani, 1994), generating time-series from the random sampling with replacement of historical records that lost any autocorrelation specific to the original series. This strategy was followed by several variants such as the method of moving blocks Bootstrap (Vogel & Shallcross, 1996; Srinivas & Srinivasan, 2005) and nearest neighbor Bootstrap (Lall & Sharma, 1996). The former method only partially corrects the autocorrelation issues, and the latter depends on the availability of historical data, which is a drawback if one wants to simulate stochastic change conditions as projected in (IPCC, 2021). In parallel to Bootstrap methods, the family of autoregressive (AR) models arose as a first order Markovian model (Thomas Harold, 1962). These methods later evolved with multiple related works (Matalas, 1967; Moreau & Pyatt, 1970; Jettmar & Young, 1975; Young & Jettmar, 1976) to formalize the p-order AR (AR(p)) models (Box et al., 2015), and the autoregressive moving average model (ARMA). Autorregressive models adequately incorporate autocorrelation in the time-series, but they assume that the autocorrelation is constant in time. This is an important limitation for the simulation of shorter time step streamflows (e.g. less than one year), as autocorrelation does change over the year, due to seasonality. In view of the above, periodic autoregressive models (PAR(p)) have been proposed (Pagano, 1978; Parzen & Pagano, 1979; Salas et al., 1982), which are AR(p) models using sets of autocorrelations specific to each time period (e.g., weekly or monthly). However, even when the PAR(p) manages to circumvent the AR(p) models autocorrelation problem, doubts arise as to how long the period should be (e.g. monthly or seasonally), or which parameter estimation methodology should be used (Noakes et al., 1985).

More recently, Copula-based autorregresive models have been proposed for multi-site runoff synthetic generation (Chen et al., 2015; Lee & Salas, 2011; Hao & Singh, 2013; de Almeida Pereira & Veiga, 2019; Pereira et al., 2017; Reddy & Ganguli, 2012). A major strength of the Copula-based models is their flexibility given that they adjust the copulas to historical input data by using marginal distribution functions. These functions allow to simulate streamflows with scarce available information, showing great sensitivity in the identification of nonlinear dependencies in the sampling, maintaining the structural benefits and limitations of the PAR(p) or ARMA models. A monthly copulas model has been proposed in Xu et al. (2022) for flow forecasting that is highly capable of predicting future short and medium-term flows in non-stationary contexts. Note that flow forecasting is used for decision making, but some strategic decisions require long-term simulations, which are not addressed by flow forecasting methods.

Attempts to integrate both temporal and spatial correlations for synthetic runoff generation have been proposed with trivariate copulas by Chen et al. (2015), which simulates first a single streamflow (Lee & Salas, 2011), and then adds the multi-site correlation. Trivariate copulas are able to preserve cross-correlation between different tributaries at lag 0, and consistently replicate historical characteristics of the different sites such as mean, variance and autocorrelations in lags 1 and 2 (with larger but acceptable differences in the latter). However, similarly to Hao and Singh (2013), the marginal properties of the copula cannot be directly estimated from data. They must be numerically approximated, which is a drawback in the use of the models as stated in (de Almeida Pereira & Veiga, 2019). Another application was developed by Pereira et al. (2017) through a two-stage model in which simulations for different sites (39 hydropower plants) are generated independently with a PAR(p) model. The spatial correlations are then incorporated in a second stage by means of vine-copulas as proposed in (Erhardt et al., 2015). In (de Almeida Pereira & Veiga, 2019), the authors developed a multi-site flow simulator based on copula autoregressive (COPAR) model previously used in economics (Brechmann & Czado, 2015). The COPAR model has a periodic component and directly solve the temporal and spatial relationships of the different tributaries with a multi-dimensional copula. As in Chen et al. (2015), it ensures spatial correlations in the simulations close to the historical ones up to lag 2, and mean autocorrelations consistent with the historical ones up to lag 5 (except for some months). The Copulas and other autoregressive based models have the drawback of not being sensitive enough to replicate high historical temporal correlation (Kirsch et al., 2013).

Other methodologies used in hydrology that deal with the simulation of temporal and spatial features simultaneously are introduced by Tsoukalas et al. (2018a, 2018b). These authors design a new family of Nataf-based models which is an extension of Nataf's joint distribution models (Nataf, 1962) initially implemented to generate random vectors with arbitrary distributions in independent series but with cross-correlation. This process starts with the generation of random data from Gaussian copulas to then transform the marginal distribution with the inverse cumulative distribution function. The SMARTA (Symmetric Moving Average (neaRly) To Anything) model (Tsoukalas et al., 2018b) expands the capabilities of a Symmetric Moving Average (SMA), from just Gaussian distribution to almost any distribution. The SMA models are able to replicate short-run and long-run time dependencies in univariate as well as multivariate context, but are unable to incorporate cyclostationary correlation structures (i.e., seasonality or periodicity in temporal correlation). A model which has several of the qualities of SMARTA and is able to capture the cyclostationary correlation structure is SPARTA (Stochastic Periodic AutoRegressive To Anything) (Tsoukalas et al., 2018a). SPARTA just as SMARTA uses a Nataf-based model, but it starts with a PAR(p) model, instead of a SMA one. These gives SPARTA the capability of simulating cyclostationary correlation, but it also loses the capability of the SMARTA of simulating long-run time dependencies.

The long-run dependencies (LRD) are known as Hurst phenomenon (Koutsoyiannis, 2002), which is measured with the Hurst coefficient index (H). The higher the magnitude of the index, the higher the prevalence of significant autocorrelation at very high lags (e.g. 100 lags). A statistical model capable to capture and replicate the Hurst phenomenon is the Fractional Gaussian Noise (FGN) method (Mandelbrot & Van Ness, 1968; Mandelbrot & Wallis, 1968, 1969). The FGN was originally proposed as a mathematical approach to emulate long-range dependencies seen in normally distributed data, which had immediate implications in hydrology. Unfortunately it was concluded that FGN fails in simulations longer than 100 time periods (McLeod & Hipel, 1978). Although the FGN can properly simulate the frequency of extreme events (Mandelbrot & Wallis, 1968), the previously mentioned drawback was a dead end, until Kirsch et al. (2013) proposed the modified Fractional Gaussian Noise (mFGN), and solved the period barrier that hampered the use of FGN. The mFGN is able to generate univariate time series of infinite length, while replicating the cyclostationary correlation of it. With this model, Kirsch et al. (2013) demonstrated that one can use mFGN to simulate several years of streamflow preserving its correlation structure. The streamflow generation with mFGN is performed by transforming the Gaussian generated data with a Log-Normal distribution. The method allows for additive changes in the mean and standard deviation of the time series, thereby making it a powerful and useful tool for climate variability and change studies. The mFGN approach has the advantage over autorregresive models, because it captures high levels of autocorrelation, cyclostationary correlation, as well as the Hurst phenomenon (Kirsch et al., 2013).

The mFGN proposed by Kirsch et al. (2013) has a good performance in replicating a single-site streamflow, but, to the best of our knowledge, there is only one study that tries to extend mFGN to a multi-site context, without a successful result (Herman et al., 2016). Herman et al. (2016) increase the likelihood of drought events by increasing the weight of low streamflows in the distribution successfully. Nonetheless, they try to extend the mFGN into a multi-site method by using a bootstrap resampling technique of historical data, which is able to preserve historical temporal correlations, but it presents some difficulties in preserving spatial correlation. Although there are autorregresive multi-site models, in a single-site streamflow generation, the mFGN has shown to outperform autorregresive models such as AR(p) or PAR(p) (Kirsch et al., 2013), hence the extension of mFGN into a multi-site method would allow preserving its benefits in multi-site streamflow generation.

Our main objective is to build a novel stochastic streamflow generator, which we call *Weighted-modified Fractional Gaussian Noise* (WmFGN), which is able to replicate historical time (i.e. short-run and long-run dependencies, as well as cyclostationary correlation) and space dependencies from the original data. The WmFGN is an extension of the mFGN into a multi-site method. WmFGN relies on the Cholesky decomposition of the spatial correlation matrix of the historical streamflow records, which is then used to add spatial correlation to streamflow time series simulated with mFGN. As the order in which the decomposition steps are performed (i.e., temporal then spatial, or vice-versa) affects the final result, our method searches for an optimal convex combination of the resulting matrices. The result is a Pareto-curve that indicates the optimal weights of the convex combination depending on the relative importance of spatial and temporal correlations given by the hydrological modeler. This framework represents an expansion of the mFGN to the multi-streamflow case, which is useful for long term energy planning input, climate change assessment, water utility management, and other already proven applications in which synthetic streamflow time series are required.

The paper is structured as follows. In Section 2 we present an in-depth explanation of the proposed framework and its origins, moving on in Section 3 to a case study in the Chilean Bio-bio river basin where the WmFGN is applied. The results are discussed in Section 4, and Section 5 presents concluding remarks about the capabilities of the proposed model.

2 Methodology

In this section we describe the *Weighted-mFGN* methodology we propose. Before that, we recall the FGN and mFGN methodologies proposed in the literature, upon which we build our approach. In what follows we shall assume that the monthly streamflow follows a log-normal distribution, which is a common assumption in the literature as streamflows do indeed tend to follow such a distribution in practice.

2.1 Fractional Gaussian Noise

We start by describing the Fractional Gaussian Noise (FGN) method (Mandelbrot & Van Ness, 1968; Mandelbrot & Wallis, 1968, 1969). Consider a matrix $\widehat{\mathbf{Y}}$ which is populated with N years of historic inflow data in such a way that the hydrological years are set as rows, and each month is a column (it is implied that months are treated as independent processes $\widehat{Y}_i^j = [\widehat{Y}_1^j, ..., \widehat{Y}_N^j]$, where the superscript $j \in= 1, \ldots, J$ stands for the jth month, and the subscript i indexes the years in the data):

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{\mathbf{Y}}_{1,1} & \widehat{\mathbf{Y}}_{1,2} & \widehat{\mathbf{Y}}_{1,3} & \cdots & \widehat{\mathbf{Y}}_{1,J} \\ \widehat{\mathbf{Y}}_{2,1} & \widehat{\mathbf{Y}}_{2,2} & \widehat{\mathbf{Y}}_{2,3} & \cdots & \widehat{\mathbf{Y}}_{2,J} \\ \widehat{\mathbf{Y}}_{3,1} & \widehat{\mathbf{Y}}_{3,2} & \widehat{\mathbf{Y}}_{3,3} & \cdots & \widehat{\mathbf{Y}}_{3,J} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \widehat{\mathbf{Y}}_{N,1} & \widehat{\mathbf{Y}}_{N,2} & \widehat{\mathbf{Y}}_{N,3} & \cdots & \widehat{\mathbf{Y}}_{N,J} \end{bmatrix} \tag{1}$$

As a first step, the matrix $\widehat{\mathbf{Y}}$ in (1) is transformed to resemble a Normal distribution in each of its months, and then it is standardized. More specifically, let $\widetilde{\mathbf{Y}}$ be defined such that each element in $\widetilde{\mathbf{Y}}$ is the natural logarithm of $\widehat{\mathbf{Y}}$. The means and variances corresponding to each month column in $\widetilde{\mathbf{Y}}$ (denoted by $\tilde{\mu}_j$ and $\tilde{\sigma}_j^2$, respectively) are also calculated, consolidating a new whitened seasonality matrix $\mathbf{Y}$ defined as follows:

$$\mathbf{Y}_{i,j} = \frac{\widetilde{\mathbf{Y}}_{i,j} - \tilde{\mu}_j}{\tilde{\sigma}_j}, \qquad i = 1, \ldots, N, \; j = 1, \ldots, J. \tag{2}$$

The next step is to generate a new matrix $\mathbf{X}$ of size $N_s \times J$ with independent random samples from a Normal(0,1) distribution, where N_s is the number of years to simulate. This matrix $\mathbf{X}$ is called the uncorrelated synthetic inflow matrix. To introduce the time dependencies of the original streamflow time-series, the matrix $\Sigma := \mathrm{Corr}(\mathbf{Y})$ is computed, which is the square and symmetric correlation matrix of the original rearranged time-series containing the pairwise correlation coefficients between all the months. The Cholesky decomposition of Σ is then computed as:

$$\Sigma = Q^T Q. \tag{3}$$

The Cholesky decomposition in (3) is the key step of the FGN because with the resulting upper triangular matrix Q the uncorrelated synthetic inflow matrix can be adjusted to capture the historic monthly temporal correlations, i.e. one computes

$$\mathbf{Z} := \mathbf{X}Q. \tag{4}$$

The output matrix $\mathbf{Z}$ is of size $N_s \times J$. Note that $\mathrm{Corr}(\mathbf{Z}) \approx \mathrm{Corr}(\mathbf{Y})$ as desired, thereby preserving the temporal correlation between months of each year, but not the correlations across years. Finally, $\mathbf{Z}$ is transformed back into the original space of streamflows by computing

$$\overline{\mathbf{Z}}_{i,j} := \tilde{\mu}_j + \mathbf{Z}_{i,j}\,\tilde{\sigma}_j \tag{5}$$

$$\widehat{\mathbf{Z}}_{i,j} := \exp(\overline{\mathbf{Z}}_{i,j}). \tag{6}$$

2.2 Modified Fractional Gaussian Noise (mFGN)

The FGN approach described above provides a clean and simple of way to incorporate temporal correlations. One deficiency of the method, however, is that only considers the correlations between months *within the same year*. To overcome that issue, (Kirsch et al., 2013) propose a modification to the method that overlaps 6-month periods. More specifically, let $\mathbf{Y}$ be the matrix constructed in (1). Then, a new matrix $\mathbf{Y}'$

is built (see Figure 1a) so that the row corresponding to the ith year in $\mathbf{Y}'$ contains the last six months of year i plus the first six months of year $i+1$ in $\mathbf{Y}$ (note that $\mathbf{Y}'$ is one row shorter than $\mathbf{Y}$). That is, $\mathbf{Y}'$ can be constructed by applying a linear operator $\mathcal{F}$ to $\mathbf{Y}$ as follows. Let

$$\mathbf{T} \ := \begin{bmatrix} 0_{6\times 6} & I_{6\times 6} \\ I_{6\times 6} & 0_{6\times 6} \end{bmatrix} \tag{7}$$

and define the swapped data matrix $\mathbf{S} := \mathbf{Y}\,\mathbf{T}$. Define $\mathbf{S_1}$ and $\mathbf{S_2}$ as the left and right halves of $\mathbf{S}$, i.e.,

$$\mathbf{S} \ = \ [\mathbf{S_1} \,|\, \mathbf{S_2}]\,.$$

Now define the $N-1 \times N$ matrices

$$\mathbf{I_1} := \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & 1 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{I_2} := \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

Then we have that

$$\mathbf{Y}' \ = \ \mathcal{F}(\mathbf{Y}) \ := \ [\mathbf{I_1}\,\mathbf{S_1} \,|\, \mathbf{I_2}\,\mathbf{S_2}]\,.$$

Let Q' be the matrix corresponding to the Cholesky decomposition of Corr($\mathbf{Y}'$).

Now, consider as before a matrix $\mathbf{X}$ of size $N_s \times J$ with independent random samples from a Normal(0,1) distribution, where N_s is one year more than the ones to be simulated, and the matrix Q corresponding to the Cholesky decomposition of Corr($\mathbf{Y}$). Then, a new matrix $\mathbf{X}'$ of size $N_s - 1 \times J$ is constructed by applying the linear operator $\mathcal{F}$ defined above to $\mathbf{X}$, i.e., $\mathbf{X}' := \mathcal{F}(\mathbf{X})$, and one computes

$$\mathbf{Z}^1 := \mathbf{X}Q, \quad \mathbf{Z}^2 := \mathbf{X}'Q'. \tag{8}$$

The final matrix $\mathbf{Z}$ of simulated values is then built by using the right-most columns of $\mathbf{Z}^1$ and $\mathbf{Z}^2$, as indicated in Figure 1b.

2.3 Weighted modified Fractional Gaussian Noise (WmFGN)

2.3.1 Spatial correlation integration into temporally correlated data

The mFGN method described above yields excellent results in the sense that the corresponding simulated series preserve the temporal correlation of the data. The method, however, falls short of representing *spatial* correlations adequately. To circumvent this

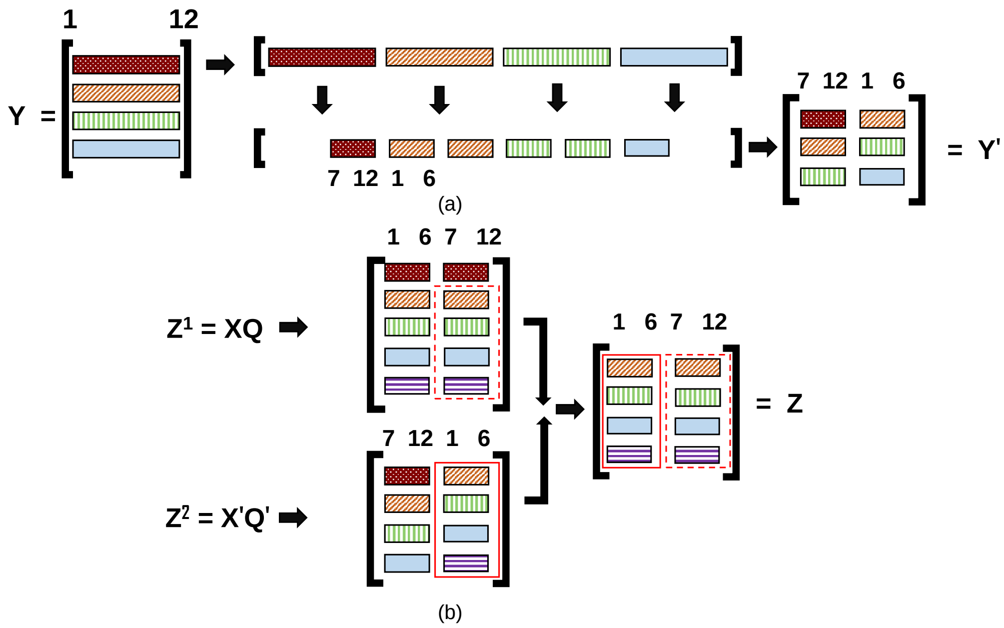


Figure 1. (a) Example process of how to retrieve **Y'** out of **Y** matrix (equivalent to the obtention of **X'** out of **X**. (b) Demonstration of how to build **Z**. Monthly scaled version of the mFGN process developed in Kirsch et al. (2013)

limitation, Kirsch et al. (2013) propose a modification that uses the same random "seed" when simulating correlated basins. The spatial approach proposed by Kirsch et al. (2013) consists in applying the mFGN as described in Section 2.2, but with a slight modification when building **X**. Instead of using random Normal(0,1) numbers to fill **X**, one *bootstraps* values from the historical data **Y**. That is, for each month/year one wants to simulate, a year is selected randomly from the historical data and the corresponding month of that year is used for **X**. The spatial correlation is then imposed by making sure the historical bootstrapped seed corresponds to the same month and year in each site.

To illustrate the idea, suppose we want to generate simulated values for the months of January, February and March in the year 2025 at two nearby sites, and that **Y** contains the monthly records of both sites from 1981 to 2020. Then, a random selection of years for **X** might choose 1992, 2005, and 1987 for January, February and March, respectively, and the corresponding **Y** values of those months/years are used for *both* sites. That is, by denoting by $\mathbf{X}^k_{i,j}$ the simulated streamflow for month j of year i at site k (and similarly for the historical data $\mathbf{Y}$) we have:

$$\begin{aligned} \mathbf{X}^1_{2025,1} &:= \mathbf{Y}^1_{1992,1}, & \mathbf{X}^2_{2025,1} &:= \mathbf{Y}^2_{1992,1} \\ \mathbf{X}^1_{2025,2} &:= \mathbf{Y}^1_{2005,2}, & \mathbf{X}^2_{2025,2} &:= \mathbf{Y}^2_{2005,2} \\ \mathbf{X}^1_{2025,3} &:= \mathbf{Y}^1_{1987,3}, & \mathbf{X}^2_{2025,3} &:= \mathbf{Y}^2_{1987,3} \end{aligned}$$

Note that each site will have its own set of simulated values $\mathbf{X}$, but the values will be correlated because the historical data is spatially correlated. Nevertheless, the mFGN distorts the spatial correlation as reported by Herman et al. (2016). Because of the limitations of mFGN in preserving spatial correlation, we propose an alternative method, as we describe next.

To introduce spatial correlation to the independently simulated streamflows of each site, we shall consider a three-dimensional version of the normalized historical inflow data matrix $\mathbf{Y}$ defined in (1)-(2) so that the third dimension corresponds to each site (see Figure 2). Denote the new structure as $\mathcal{Y}$, which has dimension $N \times J \times K$, where K is the total number of sites and denote by $\mathbf{Y}^k$ the normalized historical inflow data matrix for site k. We then have that $\mathcal{Y} = [\mathcal{Y}_{ijk}]$, where

$$\mathcal{Y}_{ijk} \;:=\; \mathbf{Y}^k_{i,j},\; i = 1, \ldots, N,\; j = 1, \ldots, J,\; k = 1, \ldots, K. \tag{9}$$

The main idea of our procedure is described as follows. First, we create matrices $\mathbf{U}^1, \ldots, \mathbf{U}^J$, each of dimension $N \times K$, such that each $\mathbf{U}^j$, $j = 1, \ldots, J$, is a slice of $\mathcal{Y}$ *in the dimension of time*, i.e., $\mathbf{U}^j = [\mathbf{U}^j_{ik}]$, where

$$\mathbf{U}^j_{ik} \;:=\; \mathcal{Y}_{i,j,k},\; i = 1, \ldots, N,\; k = 1, \ldots, K. \tag{10}$$

As a second step, we calculate the spatial correlation matrix $\mathrm{Corr}(\mathbf{U}^j)$ and its upper triangular Cholesky decomposition matrix R^j (of dimension $K \times K$), i.e.,

$$(R^j)^T (R^j) \;=\; \mathrm{Corr}(\mathbf{U}^j). \tag{11}$$

Next, we construct a three-dimensional matrix $\mathcal{Z}$ similarly to $\mathcal{Y}$, but using the matrices $\mathbf{Z}^k$ of *simulated* data constructed in Section 2.2 for each site k instead of the normalized data matrices $\mathbf{Y}^k$. As before, we define matrices $\mathbf{V}^1, \ldots, \mathbf{V}^J$, each of dimension $N_s \times K$, such that each $\mathbf{V}^j$, $j = 1, \ldots, J$, is a slice of $\mathcal{Z}$ in the dimension of time, i.e..

$$\mathbf{V}^j_{ik} \;:=\; \mathcal{Z}_{i,j,k},\; i = 1, \ldots, N_s,\; k = 1, \ldots, K. \tag{12}$$

The key step of our procedure is the calculation of the matrices

$$\mathbf{W}^j \;:=\; \mathbf{V}^j R^j,\; j = 1, \ldots, J. \tag{13}$$

Such a step incorporates the spatial correlation into the simulated data for each month. Finally, we construct a three-dimensional matrix $\mathcal{W}$ as

$$\mathcal{W}_{ijk} \;:=\; \mathbf{W}^j_{i,k},\; i = 1, \ldots, N_s,\; j = 1, \ldots, J,\; k = 1, \ldots, K. \tag{14}$$

The matrix $\mathcal{W}$ now contains our simulated data for all sites and all months, which takes into account both temporal and spatial correlations, *in that order*. We shall call this procedure *mFGNS*, which is illustrated in Figure 2.

2.3.2 Reverting the order: temporal correlation integration into spatially correlated data

The mFGNS procedure proposed in Section 2.3.1 makes clear that spatial correlation is incorporated into the simulated data *after* accounting for temporal correlation. One could, however, invert the order in which we apply the correlations. That is, starting with the full normalized data matrix $\mathcal{Y}$ constructed in (9), we can first construct matrices $\mathbf{U}^1, \ldots, \mathbf{U}^J$ as in (10) and their respective Cholesky decomposition matrices R^j as in (11). The next step is to generate a new matrix $\widetilde{\mathbf{X}}$ of size $N_s \times K$ with independent random samples from a Normal(0,1) distribution, where N_s is the one year more than the number of years to simulate. Now, by using the Cholesky decomposition matrix R^j, the uncorrelated synthetic matrix $\widetilde{\mathbf{X}}$ can be adjusted to capture the spatial correlation for each month j, i.e. one computes

$$\widetilde{\mathbf{V}}^j \;:=\; \widetilde{\mathbf{X}} R^j. \tag{15}$$

Note that $\widetilde{\mathbf{V}}^j$ (which has dimension $N_s \times K$) contains the spatially correlated simulated data for month j. We then construct the three-dimensional matrix $\widetilde{\mathcal{V}}$ as

$$\widetilde{\mathcal{V}}_{ijk} \;:=\; \widetilde{\mathbf{V}}^j_{i,k},\; i = 1, \ldots, N_s,\; j = 1, \ldots, J,\; k = 1, \ldots, K, \tag{16}$$

and define $\widetilde{\mathbf{Z}}^1, \ldots, \widetilde{\mathbf{Z}}^K$, each of dimension $N_s \times J$ as slices of $\widetilde{\mathcal{V}}$ *in the dimension of space*, i.e., for each $k = 1, \ldots, K$ we have

$$\widetilde{\mathbf{Z}}^k_{ij} \;:=\; \widetilde{\mathcal{V}}_{i,j,k},\; i = 1, \ldots, N_s,\; j = 1, \ldots, J. \tag{17}$$

Then, for each $k = 1, \ldots, K$, we apply the mFGN procedure of Section 2.2 with $\widetilde{\mathbf{Z}}^k$ in place of $\mathbf{X}$, thereby yielding a matrix $\widetilde{\mathbf{W}}^k$ which incorporates temporal correlation into

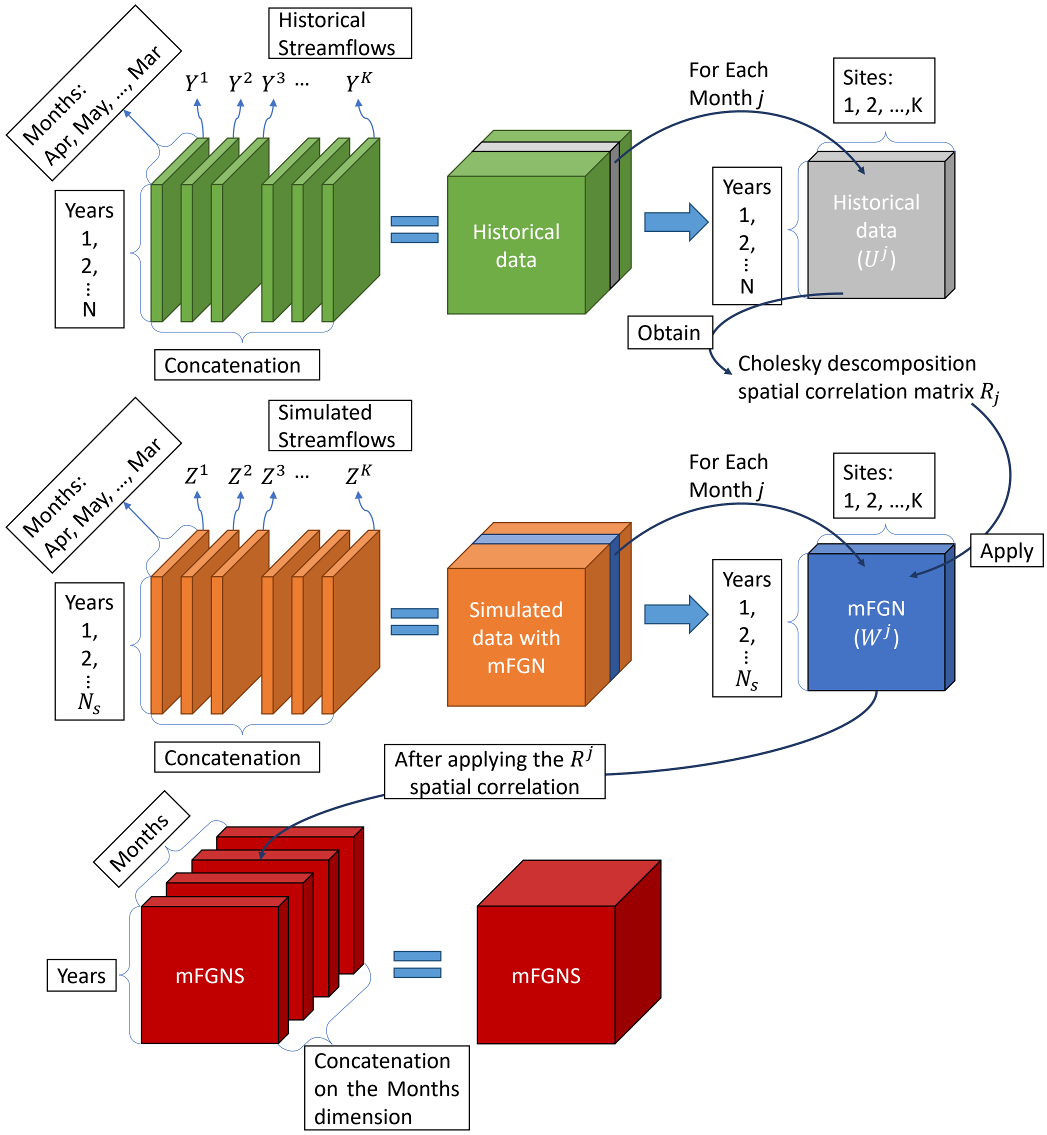


Figure 2. Schematics of mFGNS

the (spatially correlated) simulated data for site k. Finally, we construct a three-dimensional matrix $\widetilde{\mathcal{W}}$ as

$$\widetilde{\mathcal{W}}_{ijk} := \widetilde{\mathbf{W}}^k_{i,j}, \; i = 1, \ldots, N_s, \; j = 1, \ldots, J, \; k = 1, \ldots, K. \tag{18}$$

228 The matrix $\widetilde{\mathcal{W}}$ now contains our simulated data for all sites and all months, which takes
229 into account both spatial and temporal correlations, *in that order*. We shall call this pro-
230 cedure *SmFGN*, which is illustrated in Figure 3.

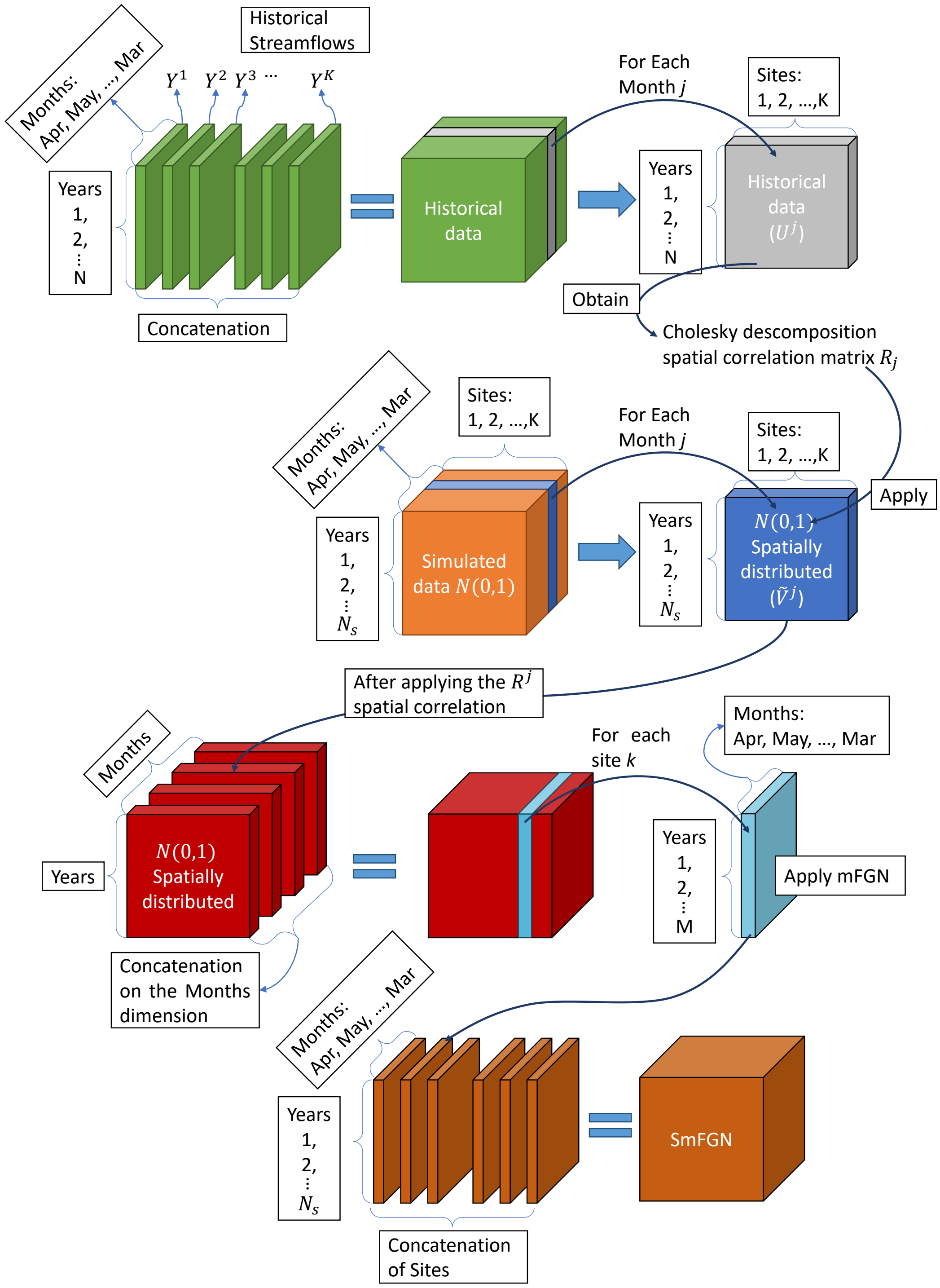


Figure 3. Schematics of SmFGN

2.3.3 Combining the mFGNS and SmFGN approaches

As discussed earlier, the procedures mFGNS and SmFGN described in the previous sections both aim at the same goal, which is to incorporate spatial correlation into the mFGN approach. The two procedures, however, lead to different simulated results, as the order in which the spatial and temporal correlations are considered does indeed matter. As we shall see in Section 4, the dimension that is considered first (spatial or temporal) is worse represented in the simulated data than the dimension that comes second.

It is natural then to consider a *weighted average* of the simulated data generated by the two procedures, a procedure we shall call *Weighted mFGN* (WmFGN for short). Note that for the WmFGN procedure to work, the random numbers $\mathbf{X}$ used in the mFGNS and SmFGN procedures must be the same. More specifically, we consider the three-dimensional matrices $\mathcal{W}$ and $\widetilde{\mathcal{W}}$ defined respectively in (14) and (18), and define, for $\alpha \in [0,1]$,

$$\widehat{\mathcal{W}}(\alpha) \ := (1-\alpha)\mathcal{W} + \alpha\widetilde{\mathcal{W}}. \tag{19}$$

Our goal is to find the value of α such that the spatial and temporal correlations induced by $\widehat{\mathcal{W}}(\alpha)$ are closest to the corresponding correlations of the historical data. With that in mind, we define the following error metrics:

$$\Delta^s_{\text{avg}}(\alpha) := \text{mean spatial error} \ = \ \frac{1}{J}\sum_{j=1}^{J} \frac{1}{K^2}\sum_{\ell,k=1}^{K} \left|\text{Corr}(\mathbf{Y}^j)_{\ell k} - \text{Corr}(\widehat{\mathbf{W}}^j(\alpha))_{\ell k}\right|, \tag{20}$$

$$\Delta^t_{\text{avg}}(\alpha) := \text{mean temporal error} \ = \ \frac{1}{K}\sum_{k=1}^{K} \frac{1}{J^2}\sum_{j,\ell=1}^{J} \left|\text{Corr}(\mathbf{Y}^k)_{j\ell} - \text{Corr}(\widehat{\mathbf{W}}^k(\alpha))_{j\ell}\right|, \tag{21}$$

$$\Delta^s_{\max}(\alpha) := \text{max. spatial error} \ = \ \max_{j=1,\ldots,J} \ \max_{\ell,k=1,\ldots,K} \left|\text{Corr}(\mathbf{Y}^j)_{\ell k} - \text{Corr}(\widehat{\mathbf{W}}^j(\alpha))_{\ell k}\right|, \tag{22}$$

$$\Delta^t_{\max}(\alpha) := \text{max. temporal error} \ \max_{k=1,\ldots,K} \ \max_{j,\ell=1,\ldots,J} \left|\text{Corr}(\mathbf{Y}^k)_{j\ell} - \text{Corr}(\widehat{\mathbf{W}}^k(\alpha))_{j\ell}\right|. \tag{23}$$

In the above equations, $\mathbf{Y}^j$ denotes a slice of $\mathcal{Y}$ across the month j, $\mathbf{Y}^k$ denotes a slice of $\mathcal{Y}$ across the site k, and similarly for $\widehat{\mathbf{W}}^j(\alpha)$ and $\widehat{\mathbf{W}}^k(\alpha)$.

The metrics defined in (20)-(23) measure the correlation error in four different ways—spatial or temporal error, mean or maximum error. Suppose the decision maker is interested in minimizing both the spatial and temporal errors, but one of the dimensions is more important than the other. Such preference can be represented by a (user-defined) parameter $\lambda \in [0,1]$ such that the temporal error has weight λ whereas the spatial error has weight $1-\lambda$. We can then define two optimization problems to find the optimal α:

$$\text{Objective 1:} \quad \min_{\alpha\in[0,1]} \lambda\Delta^t_{\text{avg}}(\alpha) + (1-\lambda)\Delta^s_{\text{avg}}(\alpha) \tag{24}$$

$$\text{Objective 2:} \quad \min_{\alpha\in[0,1]} \lambda\Delta^t_{\text{max}}(\alpha) + (1-\lambda)\Delta^s_{\text{max}}(\alpha). \tag{25}$$

Note that, in either case, the optimal α^* is a function of the user-defined parameter λ. Once α^* is found, by using (19) the simulated data is then defined as $\widehat{\mathcal{W}}(\alpha^*)$.

Remark: The metrics defined in (20)-(23) can be interpreted in terms of vector norms on matrices (see, e.g., Horn and Johnson (2012)). To see that, given a square matrix $A_{M\times M}$, for $p \geq 1$ define the ℓ_p-*norm*

$$\|A\|_p \;:=\; \left(\sum_{i=1}^{M}\sum_{j=1}^{M}|A_{ij}|^p\right)^{1/p}.$$

As customary, the above definition can be extended to $p=\infty$ as follows:

$$\|A\|_\infty \;:=\; \max_{i=1,\ldots,M}\max_{j=1,\ldots,M}|A_{ij}|.$$

Define now the following error metrics:

$$\Delta^s_p(\alpha):=\text{spatial error} \;=\; \left\|\left[v_j \,:\, v_j = \left\|\text{Corr}(\mathbf{Y}^j) - \text{Corr}(\widehat{\mathbf{W}}^j(\alpha))\right\|_p\right]\right\|_p, \tag{26}$$

$$\Delta^t_p(\alpha):=\text{temporal error} \;=\; \left\|\left[u_k \,:\, u_k = \left\|\text{Corr}(\mathbf{Y}^k) - \text{Corr}(\widehat{\mathbf{W}}^k(\alpha))\right\|_p\right]\right\|_p, \tag{27}$$

In the above equations, $\mathbf{Y}^j$ denotes a slice of $\mathcal{Y}$ across the month j, $\mathbf{Y}^k$ denotes a slice of $\mathcal{Y}$ across the site k, and similarly for $\widehat{\mathbf{W}}^j(\alpha)$ and $\widehat{\mathbf{W}}^k(\alpha)$. We see that (26)-(27) coincides with (22)-(23) when $p=\infty$. Moreover, when $p=1$, (20) is equivalent to (26) divided by JK^2, whereas (21) is equivalent to (27) divided by KJ^2. We use (20)-(23) as they are more intuitive to formulate, but the interpretation as vector norms on matrices opens the possibility to measure the error with different values of p—for instance, $p=2$ which corresponds to the well-known Frobenius norm.

In the next sections we present a case study to illustrate the application of the WmFGN procedure described above, and compare the results with those obtained by using the approach of (Kirsch et al., 2013).

3 Case study

To test the *WmFGN* methodology, we simulate streamflows from the Bio-bio river basin depicted in Figure 4. The Bio-bio river basin, located in Southern Chile, presents an average annual precipitation of 1,330 mm, which leads to mean daily discharges of 960 m^3/s (Grantham et al., 2013). The basin has a significant urban area, which includes the city of Concepción, a large percentage of forest plantations (∼20% of the land cover), and has been historically important for the country due to its hydroelectric production (Grantham et al., 2013). The Bio-bio river basin represents almost 40% of the hydroelectric potential of Chile, a country that is historically known for the importance of its hydroelectric sector (CNE, 2023). In addition, the Biobio region (i.e., formed by the Biobio basin, small coastal basins near Biobio and also used to include the northern Ñuble region, see Figure 4) supplies about 10% of Chile's urban drinking water consumption (Molinos-Senante & Donoso, 2021).

Given the importance of hydroelectricity for Chile and the Bio-bio river basin, we decided to perform our numerical analysis on the streamflows of that basin used by the National Electric Coordinator (NEC) (CEN, 2021). The data includes weekly streamflows (i.e., considering four weeks per month) between the hydrological years 1960/61 and 2018/19, note that hydrological years start in April in this region, of the rivers of interest for the NEC (e.g. inflows of hydro-power plants). After filtering the NEC streamflow database by location, the weekly flows were aggregated into monthly time series. Then, the flows were filtered to identify those for which most of their months (i.e., at least 9 out of 12) had a log-normal distribution. Finally, nine sites remained for the Bio-bio river basin, which are those presented in Figure 4. Statistics for these rivers are presented in Table 1, which include location, annual streamflow mean and standard deviation.

The seasonal variation of the monthly mean and standard deviation of the streamflows, for the period 1960/61-2018/19, are presented in Figures 5 for three representative rivers (Abanico, El Toro and Ralco). Seasonal variations for the remaining six rivers are given in the Supplementary material (Figures S1 and S2). As can be seen in these figures, most locations, regardless of the streamflows magnitudes, present a double peak in the Winter months (Jun-July) and in Spring (October to December). The first one is related to a pluvial peak, given that most of the precipitation falls during Winter, while the second peak is related to snow-melt. Hence, the Bio-bio river basins has a mixed nivo-pluvial flow.

Some recent challenges of the Bio-bio river basin have been related to both floods and droughts. Hence, developing proper hydrological modeling is of importance for the basin. The Bio-bio river basin suffered a 100-year flood during Winter of 2006 (Gironás et al., 2021). On the other hand Bio-bio is undergoing a “megadrought”, which corresponds to an uninterrupted event of below-average precipitation years since 2010 (Boisier et al., 2016). The megadrought is a phenomenon that has affected other basins as well (Barría et al., 2021), having an impact for more than a decade over Central-Southern Chile (Garreaud et al., 2020, 2021). Also, climate change projections over Central-Southern Chile indicate that precipitations should decrease in the future (Chadwick et al., 2018; Araya-Osses et al., 2020; Chadwick et al., 2023).

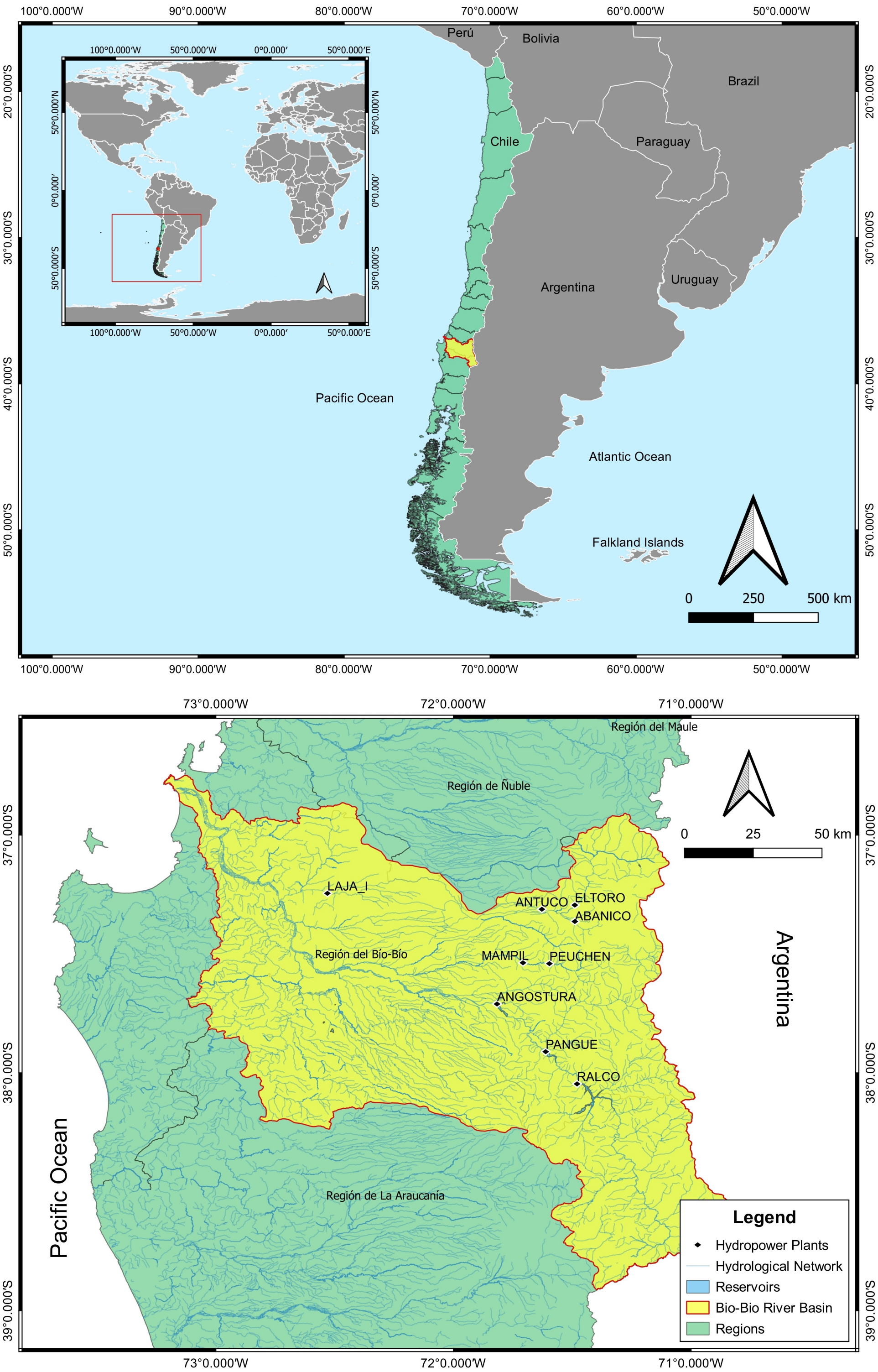


Figure 4. Bio-bio river basin overview.

Table 1. Locations of the streamflows with their annual mean, and standard deviation

Site Number	River	Lat.	Lon.	Mean (m^3/s)	Std. Dev. (m^3/s)
1	Laja I	-37.24	-72.53	15.09	6.96
2	Angostura	-37.71	-71.81	131.49	39.46
3	Antuco	-37.31	-71.63	49.05	15.07
4	Abanico	-37.36	-71.50	4.46	1.46
5	El Toro	-37.29	-71.50	61.67	16.61
6	Ralco	-38.04	-71.48	249.45	68.92
7	Pangue	-37.91	-71.61	28.17	11.49
8	Mampil	-37.53	-71.70	21.75	5.31
9	Peuchen	-37.54	-71.59	35.50	9.17

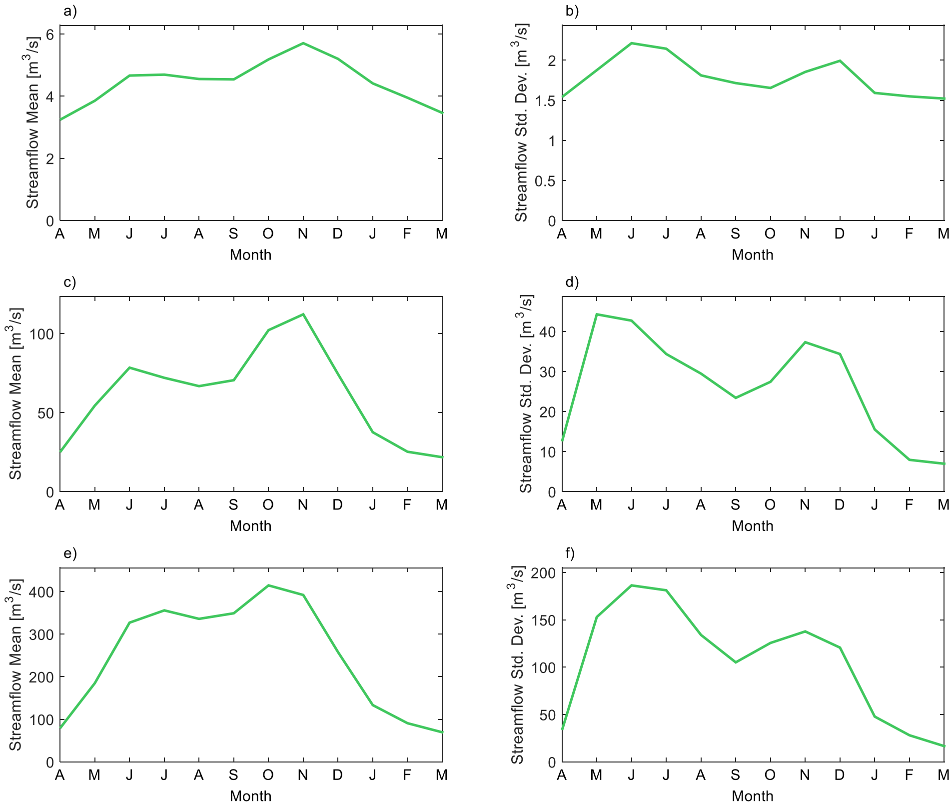


Figure 5. Monthly mean (a, c, and e) and standard deviation (b, d, and f) of the streamflows of Abanico (a, and b), El Toro (c, and d), and Ralco (e, and f) rivers.

4 Numerical results

4.1 Optimal parameters from WmFGN

The performance of the proposed WmFGN is measured in term of its capability of preserving the original temporal and spatial correlations in the observed data. The measurements of correlation errors of the simulated data are computed in Eqs. (20) to (23) for different values of α and plotted in Figures 6a and 6b. The performance is also compared against that of the mFGN, which is not dependent on α and presents difficulties with replicating the spatial correlation of the observed data. The temporal and spatial correlation errors for the mean and maximum error metrics for mFGN are computed using similar expressions as in (20)-(23), but with the matrix $\widehat{\mathcal{W}}(\alpha)$ replaced with the matrix corresponding to the mFGN method with spatial correlation added via re-sampling, as discussed in Section 2.3.1. We shall denote the resulting correlation errors for mFGN by δ^s_{avg}, δ^t_{avg}, $\delta^s_{\max}$ and $\delta^t_{\max}$, using a notation analogous to that in (20)-(23).

Figures 6c and 6d depict the Pareto frontier of spatial and temporal correlation errors. As expected, there is a trade-off between obtaining good performance in the spatial correlation and good performance in the temporal correlation, both in terms of the mean error (Figures 6a and 6c) and the maximum error (Figures 6b and 6d).

For this specific problem, we see in Figures 6b and 6d that the WmFGN, with a value of α around 0.5, shows considerable reduction of the maximum error in the spatial correlation compared to mFGN (which coincides with the spatial error of WmFGN with $\alpha = 1$), with almost no increase in the temporal correlation error. On the other hand, we observe in Figures 6a and 6c that there is a range of values of α (between around 0.75 and 0.98) for which both spatial and temporal mean errors for WmFGN are smaller than the mFGN errors, that is, for that metric WmGFN is superior to mFGN in both spatial and temporal dimensions.

Although the results are specific for the Bio-bio basin, they represent a clear illustration of how the WmFGN presents an improvement over mFGN in terms of preserving both spatial and temporal correlations, in addition to allowing for more flexibility.

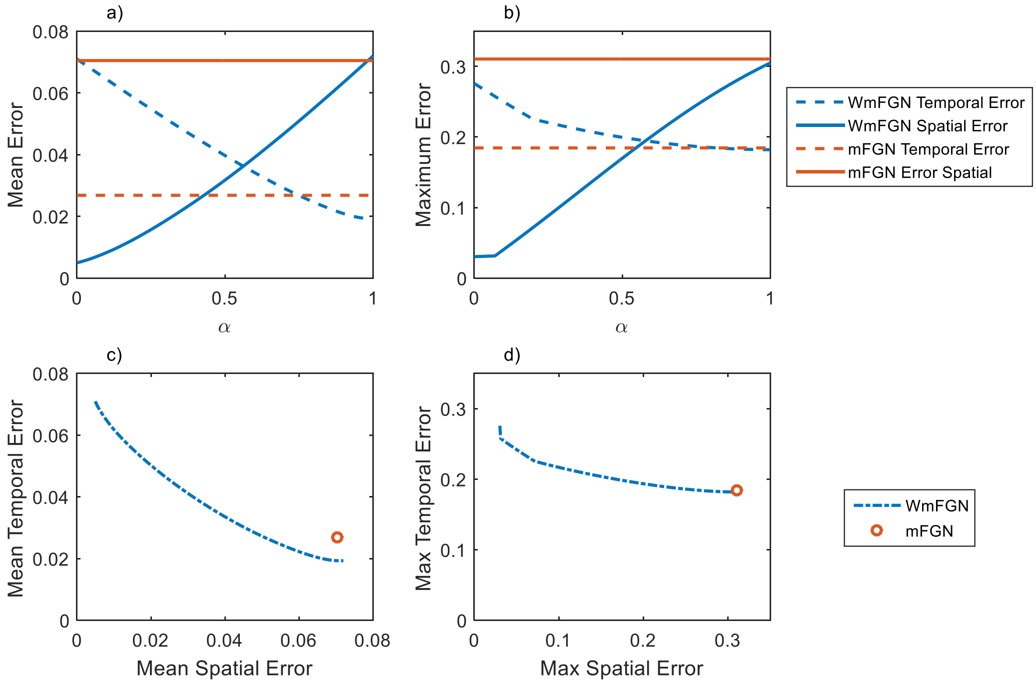


Figure 6. Mean (a) and maximum (b) spatial and temporal correlation errors as a function of α; (c) and (d) depict spatial vs. temporal error for the mean and maximum error metrics, respectively.

When finding the optimal weight α balancing mFGNS and SmFGN in (19) that minimizes the weighted sum of spatial and temporal mean (resp. maximum) correlation errors with model (24) (resp. (25)) under different user-defined parameters λ, we obtain a curve as displayed in Figure 7a (resp. Figure 7b). As discussed earlier, the value of λ allows the user of WmFGN to prioritize between the spatial (i.e., λ=0) or temporal (i.e., λ=1) correlation. Not surprisingly, the optimal value of α coincides with λ at the extreme cases——after all, if the user is only concerned with spatial correlation (i.e., chooses $\lambda = 0$) then the best combination between mFGNS and SmFGN is really just using mFGNS which gives the highest priority to spatial correlation, and that corresponds to taking $\alpha = 0$ in (19). An analogous argument holds for the case where temporal correlation is preferred.

The optimal values of the objective functions 1 (Eq. (24)) and 2 (Eq. (25)) are presented in Figures 7c and 7d, respectively, for different values of the user-defined parameter λ. For comparison, we also compute the values of Objectives 1 and 2 for the mFGN.

This amounts to calculating

$$\text{Objective 1:} \quad \lambda\delta^t_{\text{avg}} + (1-\lambda)\delta^s_{\text{avg}} \tag{28}$$

$$\text{Objective 2:} \quad \lambda\delta^t_{\text{max}} + (1-\lambda)\delta^s_{\text{max}}, \tag{29}$$

where δ^t_{avg}, δ^s_{avg}, δ^t_{max} and δ^s_{max} are the correlation errors for mFGN, as defined earlier. Note that for objective function 1, WmFGN always yields lower values than mFGN (Figure 7c), whereas for objective function 2, WmFGN yields lower values than mFGN for most values of λ, except when λ is close to 1 in which case both WmFGN and mFGN coincide (Figure 7d). The advantage of using WmFGN over mFGN increases as the user gives higher importance of spatial correlation over temporal one, which is visually represented as the increasing gap between the objective functions in Figures 7c and 7d, as λ approaches zero.

Deciding on an appropriate value for λ will depend on the user's priority for correctly simulating spatial or temporal correlation, which will eventually define the associated value for α. Nevertheless, if the user has similar priorities for both correlations, and wants to decide which λ to use, an option could be simply using λ=0.5. Interestingly, as seen in Figure 7a, such a value corresponds to taking $\alpha = 0.33$, that is, giving twice the weight to mFGNS relatively to SmFGN. Another option could be equating the temporal and spatial errors, which for the mean error criterion yields a value of α of 0.562 (Figure 6a), whereas for the maximum error criterion it yields a value of $\alpha = 0.578$ (Figure 6b). These values of α correspond to taking $\lambda = 0.574$ and $\lambda = 0.842$ for the mean and maximum error criteria, respectively (Figure 7a and 7b). Note also that these values of λ are the maximizers of the WmFGN objective functions in Figures 7c and 7d, and represent a change in the concavity of the α-curves (Figures 7a and 7b).

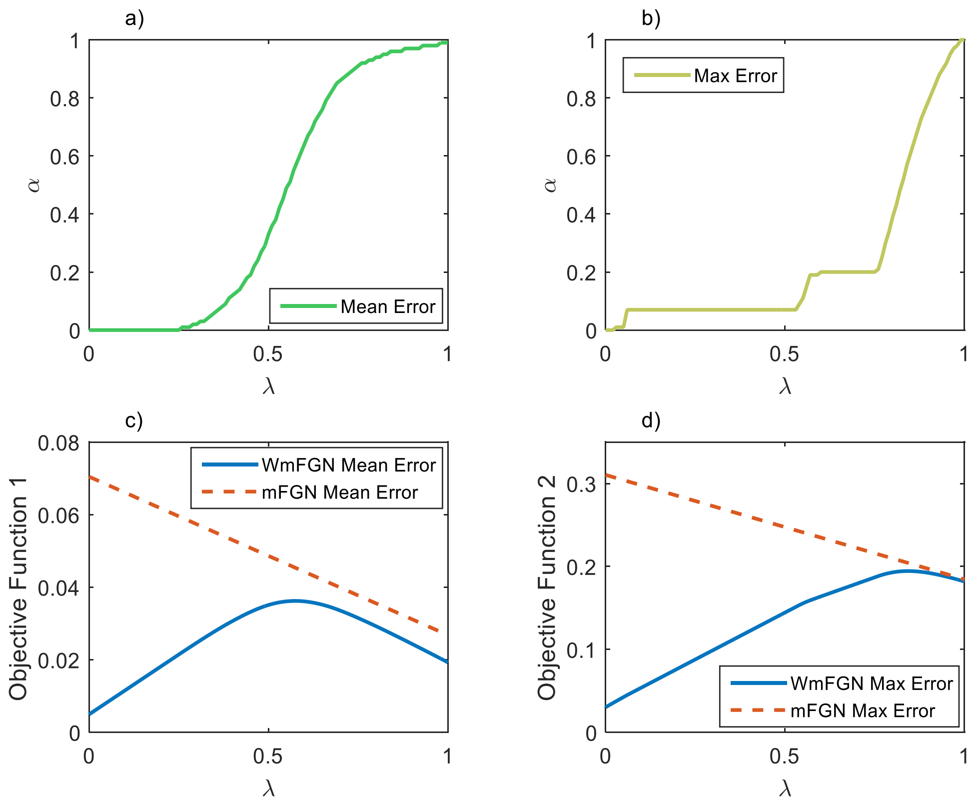


Figure 7. Optimal values of α as a function of λ for the mean error (a) and maximum error (b) criteria; (c) and (d) depict the optimal objective function values in (24) and (25), respectively, as a function of λ.

4.2 Illustration of the behaviour of the correlations

One advantage of the WmFGN approach, compared to other methods proposed in the literature (including mFGN) is that it tailors the procedure according to the importance of temporal vs spatial correlation specified by the user, which in this case is accomplished by means of the parameter λ in (24) and (25). Figures 8d-8h and 9d-9h illustrate that flexibility, displaying the correlations calculated from the simulated data generated by WmFGN using the mean error metric (Objective 1). The figures depict the temporal correlation of among months for a representative river (Ralco), and the spatial correlation among locations for a representative month (November), respectively, for different values of the parameter λ. In addition to the actual correlations, Figures 8i-8m and 9i-9m show the correlation errors with respect to observed data. For the sake of comparison, Figures 8a-8c and 9a-9c show the correlations of observed data, the correlations calculated from data simulated for mFGN, and the associated correlation errors.

The figures demonstrate that WmFGN accomplishes what it proposes to do. For values of $\lambda \geq 0.75$ (priority to temporal correlation), we see in Figure 8 that the temporal correlation errors are indeed small. These errors increase as λ decreases. In Figure 9 we see the opposite effect—the spatial errors are small for $\lambda \leq 0.25$ (priority to spatial correlation), and increase as λ increases.

The figures also corroborate the previous conclusions about the benefit of the WmFGN approach over mFGN for preserving both spatial and temporal correlations. The mFGN procedure—which by construction prioritizes temporal correlation— presents very low temporal correlation errors as shown in Figure 8c, at the expense of high spatial correlation errors (see Figure 9c). This is in line with the results of previous studies (Herman et al., 2016). However, a comparison between the correlation errors for mFGN and for WmFGN with $\lambda = 1$ in Figures 8m and 9m (which is the comparable case where full priority is given to temporal correlation) shows that the temporal correlation errors for WmFGN are in fact smaller than those for mFGN, and the spatial correlation errors are similar. Moreover, by introducing flexibility via the λ parameter, the WmFGN approach allows the user to "sacrifice" some of the precision in the temporal correlation in order to increase the precision in the spatial correlation—a flexibility that is not present in mFGN.

The above discussion is based on the results corresponding to Objective 1 (mean error criterion). Similar conclusions can be obtained by examining Figures 10 and 11, which display the results corresponding to Objective 2 (maximum error criterion). Note also that the results discussed above for the chosen representative river and month apply similarly to the other rivers and months considered in this paper as shown in Figures S3 to S40 in the Supplementary Material.

Although the choice of error metric to be used (mean or maximum error, corresponding to Objective 1 and 2, respectively) is problem-specific and depends on the priorities of the user, there are some general recommendations. For example, when comparing the WmFGN temporal correlation errors in Figures 8i-8m and 10i-10m, we see that the latter are more sensitive to changes in the user-defined parameter λ—indeed, the errors are similar up to $\lambda = 0.75$, and then they change considerably for $\lambda = 1$. The correlation errors in Figure 8 change more smoothly and hence are not so sensitive to small changes in λ. Such behavior is also illustrated in Figures 7a and 7b, where we see a smoother curve for the case of the mean error metric. These properties, if desired, would favor the use of the mean error metric over the maximum one.

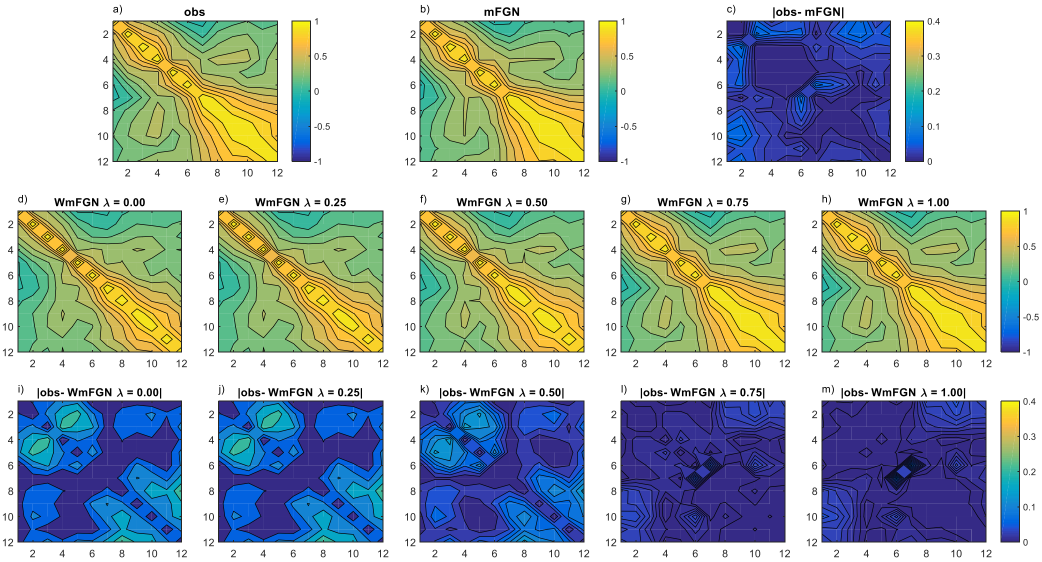


Figure 8. Pairwise temporal correlations of the 12 months of the year (i.e., the first month of the hydrological year is April), for the Ralco river, for different time series: a) observed, b) mFGN, c) absolute difference between observed and mFGN, d-h) WmFGN with different λ values, subjected to objective function 1, and i-m) absolute difference between the observed and WmFGN with different λ values.

5 Conclusions

Hydrology has used for several years the synthetic simulation of hydroclimatic variables in different problems. Several reasons make it attractive to extrapolate historical records, or to have the capability of analyzing the behaviour of infrastructure under conditions different from the historical ones. When evaluating the design of new water infrastructure such as reservoirs or water facilities, stochastic methods have been used. These tools have also shown to be useful in the evaluation of the operation of current infrastructure. In addition, due to challenges as climate variability and change, it does not suffice to evaluate new and existing infrastructure under historical conditions. For this reason, the synthetic simulation of streamflows that are not only consistent with historic

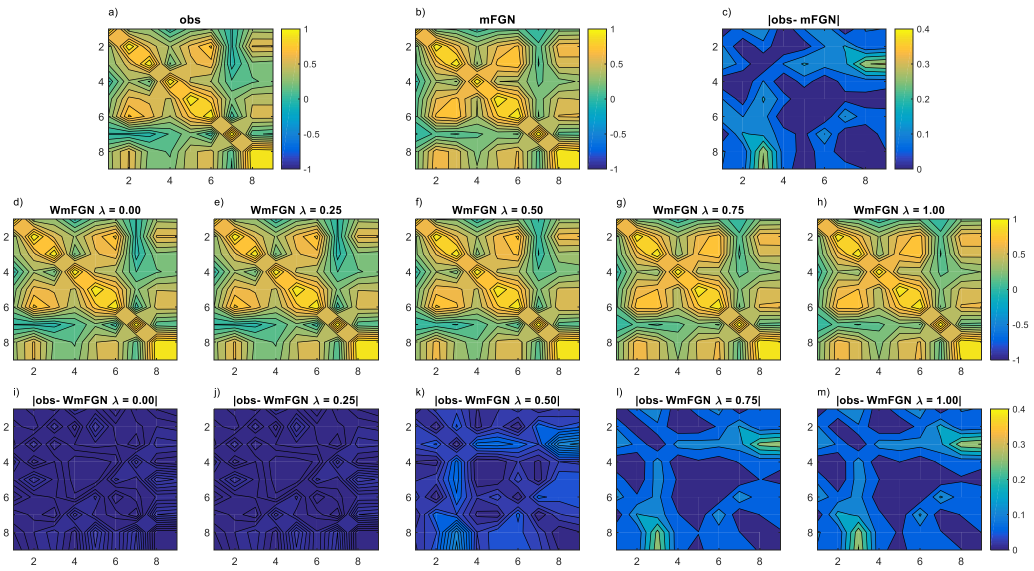


Figure 9. Pairwise spatial correlations of the nine river sites (i.e., the sites use the numbering from Table 1), for the month of November, for different time series: a) observed, b) mFGN, c) absolute difference between observed and mFGN, d-h) WmFGN with different λ values, subjected to objective function 1, and i-m) absolute difference between the observed and WmFGN with different λ values.

statistical properties, but also adjustable to future conditions is necessary. The stochastic models of the family of Fractional Gaussian Noise (FGN) have great potential for this.

The FGN approaches have shown to be capable of capturing long term memory in time series. Unfortunately, the original FGN procedure is not able to simulate infinite time series; that changed when the Modified FGN (mFGN) method was developed. The mFGN procedure is capable of simulating infinite time series that recreate the seasonal or periodic correlation structure, overcoming the major limitation of FGN. Also, mFGN has shown to replicate the temporal correlations of the data. However, mFGN is not well suited to represent the spatial correlation structure required to simulate several streamflows at the same time.

In this paper we have proposed a new method, called Weighted mFGN (WmFGN), that addresses both temporal and spatial correlations *simultaneously*. Our numerical experiments for a basin in Chile demonstrate that the WmFGN procedure represents a significant improvement in preserving the spatial correlation, when compared against mFGN. Moreover, since there is a trade-off in terms of representing the spatial and temporal cor-

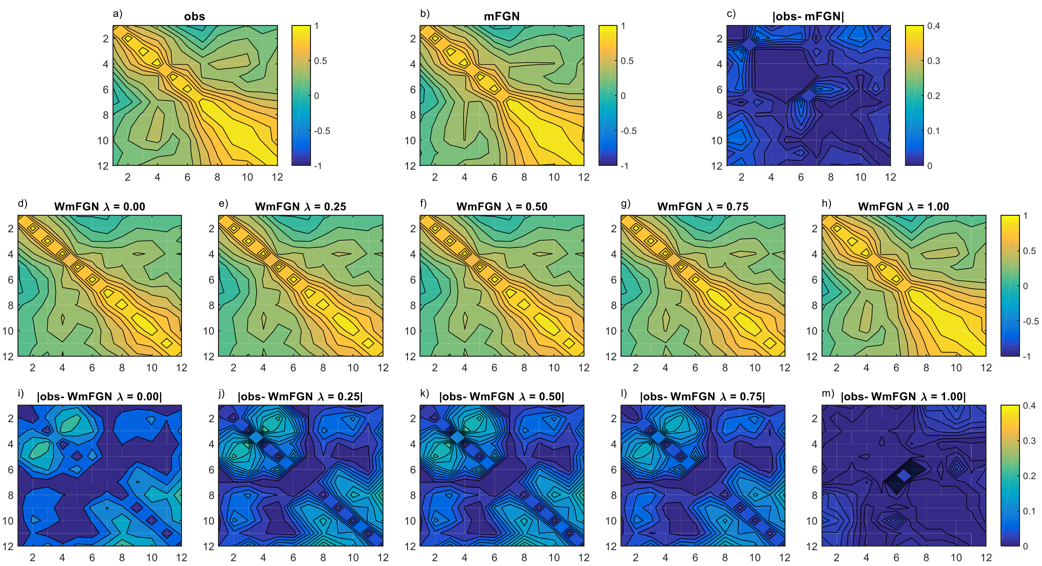


Figure 10. Pairwise temporal correlations of the twelve months of the year (i.e., the first month of the hydrological year is April), for the Ralco river, for different time series: a) observed, b) mFGN, c) absolute difference between observed and mFGN, d-h) WmFGN with different λ values, subjected to objective function 2, and i-m) absolute difference between the observed and WmFGN with different λ values.

relations, the method allows the user to specify the importance of one type of correlation over the other, and tailors the method for that choice by optimizing over some internal parameters. To the best of our knowledge, no other method in the literature addresses this trade-off in a systematic way.

Regardless of the trade-off, in our experiments the WmFGN procedure outperforms mFGN, even when temporal correlation is prioritized. Moreover, the higher the priority of the spatial correlation specified by the user, the higher the benefit of using WmFGN over mFGN.

As discussed earlier, the proposed approach requires the user to specify the importance of temporal correlation over the spatial one, by means of a parameter λ such that the weight of temporal correlation is λ whereas the weight of spatial correlation is $1-\lambda$. In the absence of a preference, the user can give equal weights to both correlations (i.e., choose $\lambda = 0.5$). Note however that such a choice does not imply that the errors in both correlations (with regards to observed data) are the same; thus, another possi-

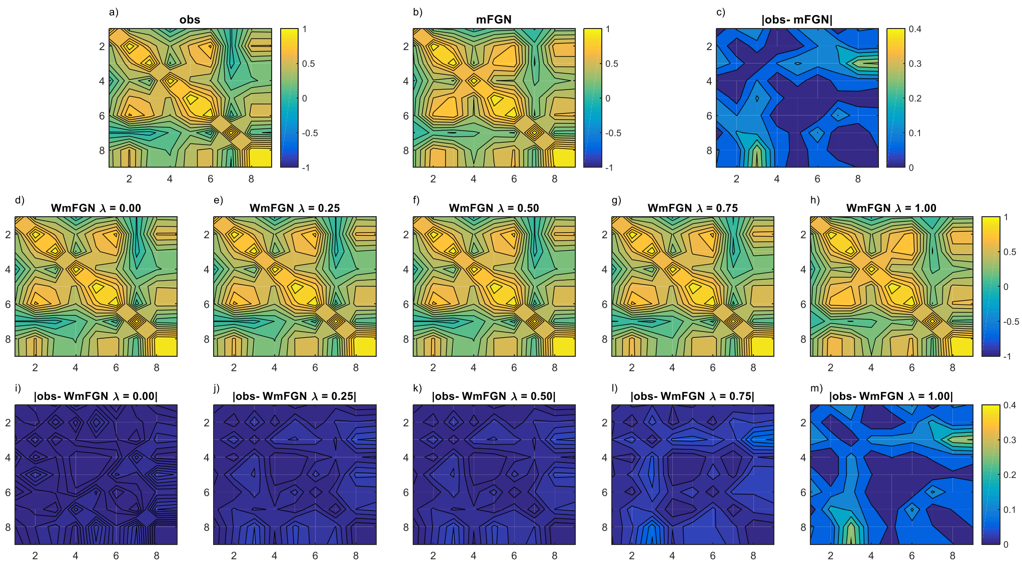


Figure 11. Pairwise spatial correlations of the nine river sites (i.e., the sites use the numbering from Table 1), for the month of November, for different time series: a) observed, b) mFGN, c) absolute difference between observed and mFGN, d-h) WmFGN with different λ values, subjected to objective function 2, and i-m) absolute difference between the observed and WmFGN with different λ values.

ble choice for the user is to impose that the errors in both correlations be equal, and let the method compute the corresponding value of λ automatically.

Finally, it is important to remember that our conclusions about the performance of WmFGN are based on the numerical experiments we have conducted. Future studies should further test the WmFGN in other basins, and also with different climates. Moreover, different error metrics can be tested; for instance, the remark in Section 2.3 suggests that other vector norms on matrices (or, more generally, other matrix norms) can be used.

Acknowledgments

This research was funded by grant ANILLO ACT 192094. We also acknowledge grants FONDECYT de Iniciación 11220952. We thank the National Electric Coordinator for (NEC, https://www.coordinador.cl/reportes-y-estadisticas/#Estadisticas) for the availability of the data.

References

Araya-Osses, D., Casanueva, A., Román-Figueroa, C., Uribe, J., & Paneque, M. (2020). Climate change projections of temperature and precipitation in chile based on statistical downscaling. *Climate Dynamics*, *54*, 4309-4330. doi: https://doi.org/10.1007/s00382-020-05231-4

Barría, P., Chadwick, C., Ocampo-Melgar, A., Galleguillos, M., Garreaud, R., Díaz-Vasconcellos, R., ... Poblete-Caballero, D. (2021). Water management or megadrought: what caused the chilean aculeo lake drying? *Regional Environmental Change*, *21*(19). doi: https://doi.org/10.1007/s10113-021-01750-w

Boisier, J. P., Rondanelli, R., Garreaud, R. D., & Muñoz, F. (2016). Anthropogenic and natural contributions to the southeast pacific precipitation decline and recent megadrought in central chile. *Geophysical Research Letters*, *43*(1), 413–421. doi: https://doi.org/10.1002/2015GL067265

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control.* John Wiley & Sons.

Brechmann, E. C., & Czado, C. (2015). Copar—multivariate time series modeling using the copula autoregressive model. *Applied Stochastic Models in Business and Industry*, *31*(4), 495–514. doi: https://doi.org/10.1002/asmb.2043

CEN. (2021). *Reportes, estadísticas y plataformas de uso frecuente, developed by: Coordinador eléctrico nacional.* Retrieved 15.11.2021, from `\url{https://www.coordinador.cl/reportes-y-estadisticas/#Estadisticas}`

Chadwick, C., Gironás, J., Gonzalez-Leiva, F., & Aedo, S. (2023). Bias-adjustment to preserve the changes in variability: The unbiased mapping of gcm changes to local stations. *Hydrological Science Journal*, Accepted. doi: https://doi.org/10.1080/02626667.2023.2201450

Chadwick, C., Gironás, J., Vicuña, S., Meza, F., & McPhee, J. (2018). Using a statistical preanalysis approach as an ensemble technique for the unbiased mapping of gcm changes to local stations. *Journal of Hydrometeorology*, *19*(9), 1447-1465. doi: https://doi.org/10.1175/JHM-D-17-0198.1

Chen, L., Singh, V. P., Guo, S., Zhou, J., & Zhang, J. (2015). Copula-based method for multisite monthly and daily streamflow simulation. *Journal of Hydrology*, *528*, 369–384. doi: https://doi.org/10.1016/j.jhydrol.2015.05.018

CNE. (2023). *Reporte capacidad instalada generación, marzo 2023, developed by:*

Comisión nacional de energía. Retrieved 15.04.2023, from `\url{https://www.cne.cl/normativas/electrica/consulta-publica/electricidad/}`

de Almeida Pereira, G. A., & Veiga, Á. (2019). Periodic copula autoregressive model designed to multivariate streamflow time series modelling. *Water Resources Management*, *33*(10), 3417–3431. doi: https://doi.org/10.1007/s11269-019-02308-6

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap.* CRC press. doi: https://doi.org/10.1201/9780429246593

Erhardt, T. M., Czado, C., & Schepsmeier, U. (2015). R-vine models for spatial time series with an application to daily mean temperature. *Biometrics*, *71*(2), 323–332. doi: https://doi.org/10.1111/biom.12279

Garreaud, R., Alvarez-Garreton, C., Brichivich, J., Boisier, J., Christie, D., Galleguillos, M., ... Zambrano-Bigiarini, M. (2021). The 2010-2015 megadrought in central chile: impacts on regional hydroclimate and vegetation. *Hydrological and Earth System Sciences*, *21*, 6307-6327. doi: https://doi.org/10.5194/hess-21-6307-2017

Garreaud, R., Boisier, J., Rondanelli, R., Montecinos, A., Sepúlveda, H. H., & Veloso-Aguila, D. (2020). The central chile mega drought (2010–2018): a climate dynamics perspective. *International Journal of Climatology*, *40*(1), 421–439. doi: https://doi.org/10.1002/joc.6219

Gironás, J., Bunster, T., Chadwick, C., & Fernández, B. (2021). *Floods. in: Fernández, b., gironás, j. (eds) water resources of chile. world water resources, vol 8.* Springer.

Grantham, T., Figueroa, R., & Prat, N. (2013). Water management in mediterranean river basins: a comparison of management frameworks, physical impacts, and ecological responses. *Hydrobiologia*, *719*, 451-482. doi: https://doi.org/10.1007/s10750-012-1289-4

Hao, Z., & Singh, V. P. (2013). Modeling multisite streamflow dependence with maximum entropy copula. *Water Resources Research*, *49*(10), 7139–7143. doi: https://doi.org/10.1002/wrcr.20523

Hashimoto, T., Stedinger, J., & Loucks, D. (1982). Reliability, resiliency, and vulnerability criteria for water resource system performance evaluation. *Water resources research*, *18*(1), 14-20. doi: https://doi.org/10.1029/

WR018i001p00014

Herman, J. D., Zeff, H. B., Lamontagne, J. R., Reed, P. M., & Characklis, G. W. (2016). Synthetic drought scenario generation to support bottom-up water supply vulnerability assessments. *Journal of Water Resources Planning and Management*, *142*(11), 04016050. doi: https://doi.org/10.1061/(ASCE)WR.1943-5452.0000701

Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis, 2nd ed.* Cambridge University Press.

IPCC. (2021). *Summary for Policymakers. In: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [MassonDelmotte, V., P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou (eds.)]* (Tech. Rep.). Cambridge University Press. In Press.

Jettmar, R., & Young, G. (1975). Hydrologic estimation and economic regret. *Water Resources Research*, *11*(5), 648–656. doi: https://doi.org/10.1029/WR011i005p00648

Kirsch, B. R., Characklis, G. W., & Zeff, H. B. (2013). Evaluating the impact of alternative hydro-climate scenarios on transfer agreements: Practical improvement for generating synthetic streamflows. *Journal of Water Resources Planning and Management*, *139*(4), 396–406. doi: https://doi.org/10.1061/(ASCE)WR.1943-5452.0000287

Koutsoyiannis, D. (2002). The hurst phenomenon and fractional gaussian noise made easy. *Hydrological Sciences Journal*, *47*(4), 573–595. doi: https://doi.org/10.1080/02626660209492961

Lall, U., & Sharma, A. (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water resources research*, *32*(3), 679–693. doi: https://doi.org/10.1029/95WR02966

Lee, T., & Salas, J. D. (2011). Copula-based stochastic simulation of hydrological data applied to nile river flows. *Hydrology Research*, *42*(4), 318–330. doi: https://doi.org/10.2166/nh.2011.085

Mandelbrot, B. B., & Van Ness, J. W. (1968). Fractional brownian motions, fractional noises and applications. *SIAM review*, *10*(4), 422–437.

Mandelbrot, B. B., & Wallis, J. R. (1968). Noah, joseph, and operational hydrology. *Water resources research*, *4*(5), 909–918.

Mandelbrot, B. B., & Wallis, J. R. (1969). Computer experiments with fractional gaussian noises: Part 1, averages and variances. *Water resources research*, *5*(1), 228–241.

Matalas, N. C. (1967). Mathematical assessment of synthetic hydrology. *Water Resources Research*, *3*(4), 937–945. doi: https://doi.org/10.1029/WR003i004p00937

McLeod, A. I., & Hipel, K. W. (1978). Preservation of the rescaled adjusted range: 1. a reassessment of the hurst phenomenon. *Water Resources Research*, *14*(3), 491–508.

Molinos-Senante, M., & Donoso, G. (2021). *Domestic uses of water. in: Fernández, b., gironás, j. (eds) water resources of chile. world water resources, vol 8.* Cambridge University Press.

Moreau, D. H., & Pyatt, E. E. (1970). Weekly and monthly flows in synthetic hydrology. *Water Resources Research*, *6*(1), 53–61. doi: https://doi.org/10.1029/WR006i001p00053

Nataf, A. (1962). Determination des distribution don t les marges sont donnees. *Comptes Rendus de l Academie des Sciences*, *225*, 42–43.

Noakes, D. J., McLeod, A. I., & Hipel, K. W. (1985). Forecasting monthly riverflow time series. *International Journal of Forecasting*, *1*(2), 179–190. doi: https://doi.org/10.1016/0169-2070(85)90022-6

Pagano, M. (1978). On periodic and multiple autoregressions. *The Annals of Statistics*, *6*(6), 1310 – 1317. doi: https://doi.org/10.1214/aos/1176344376

Parzen, E., & Pagano, M. (1979). An approach to modeling seasonally stationary time series. *Journal of Econometrics*, *9*(1-2), 137–153. doi: https://doi.org/10.1016/0304-4076(79)90100-3

Pereira, G. A., Veiga, Á., Erhardt, T., & Czado, C. (2017). A periodic spatial vine copula model for multi-site streamflow simulation. *Electric Power Systems Research*, *152*, 9–17. doi: https://doi.org/10.1016/j.epsr.2017.06.017

Reddy, M. J., & Ganguli, P. (2012). Bivariate flood frequency analysis of upper godavari river flows using archimedean copulas. *Water Resources Management*, *26*(14), 3995–4018. doi: https://doi.org/10.1007/s11269-012-0124-z

Salas, J. D., Boes, D. C., & Smith, R. A. (1982). Estimation of arma models with seasonal parameters. *Water Resources Research*, *18*(4), 1006–1010. doi: https://doi.org/10.1029/WR018i004p01006

Srinivas, V., & Srinivasan, K. (2005). Hybrid moving block bootstrap for stochastic simulation of multi-site multi-season streamflows. *Journal of Hydrology*, *302*(1-4), 307–330. doi: https://doi.org/10.1016/j.jhydrol.2004.07.011

Thomas Harold, A. (1962). Mathematical synthesis of streamflow sequences for the analysis of river basin by simulation. *Design of water resources-systems*, 459–493. doi: https://doi.org/10.4159/harvard.9780674421042.c15

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. doi: https://doi-org.ezproxy.cul.columbia.edu/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. doi: https://doi-org.ezproxy.cul.columbia.edu/10.1029/2017WR022462

Vogel, R. M., & Shallcross, A. L. (1996). The moving blocks bootstrap versus parametric time series models. *Water resources research*, *32*(6), 1875–1882. doi: https://doi.org/10.1029/96WR00928

Xu, P., Wang, D., Wang, Y., & Singh, V. P. (2022). A stepwise and dynamic c-vine copula–based approach for nonstationary monthly streamflow forecasts. *Journal of Hydrologic Engineering*, *27*(1), 04021043. doi: :https://doi.org/10.1061/(ASCE)HE.1943-5584.0002145

Young, G., & Jettmar, R. (1976). Modeling monthly hydrologic persistence. *Water Resources Research*, *12*(5), 829–835. doi: https://doi.org/10.1029/WR012i005p00829

Figure 1.

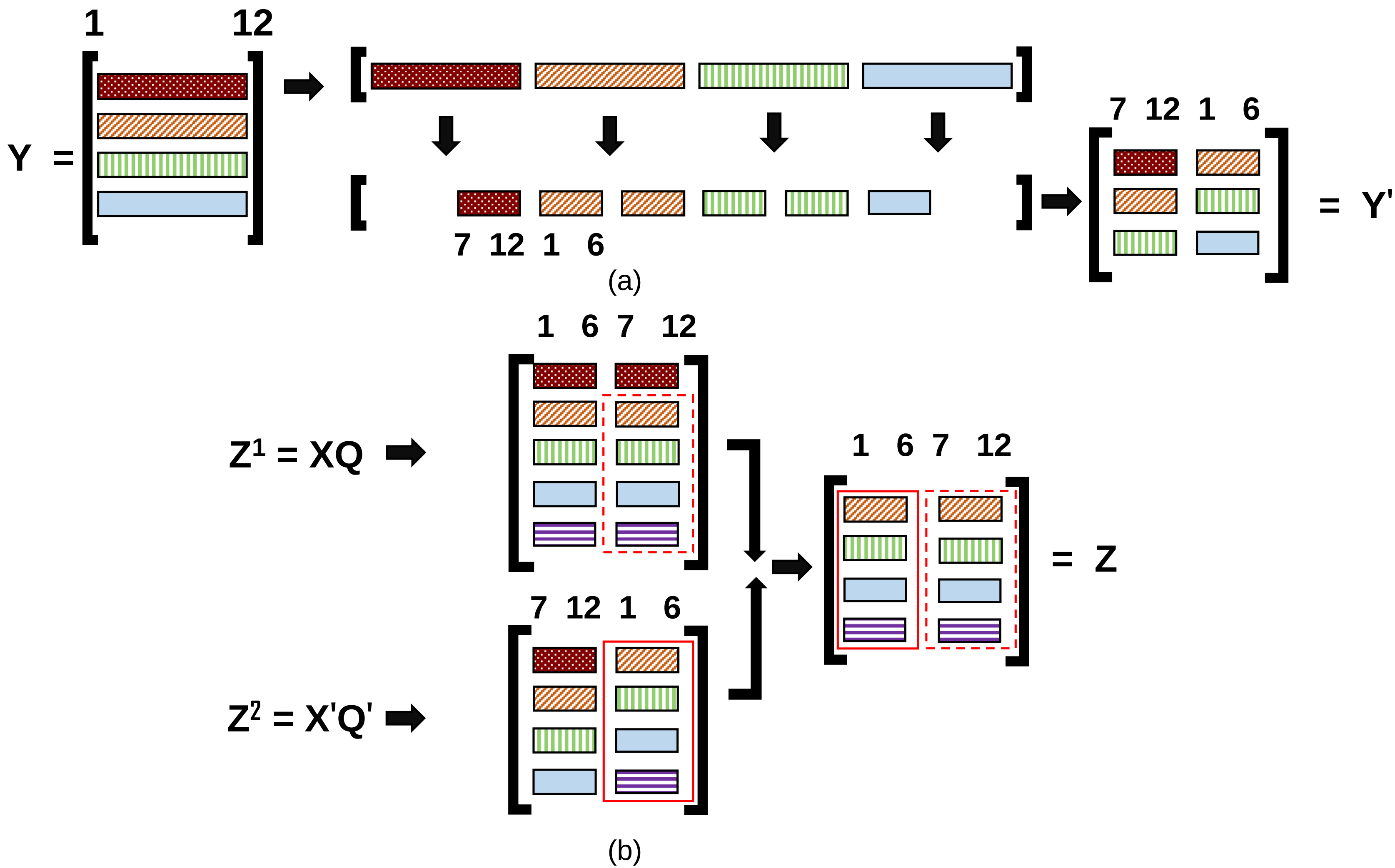

1
12
Y =
7 12 1 6
= Y'
(a)
1 6 7 12
Z^1 = XQ
7 12 1 6
Z^2 = X'Q'
= Z
(b)

Figure 2.

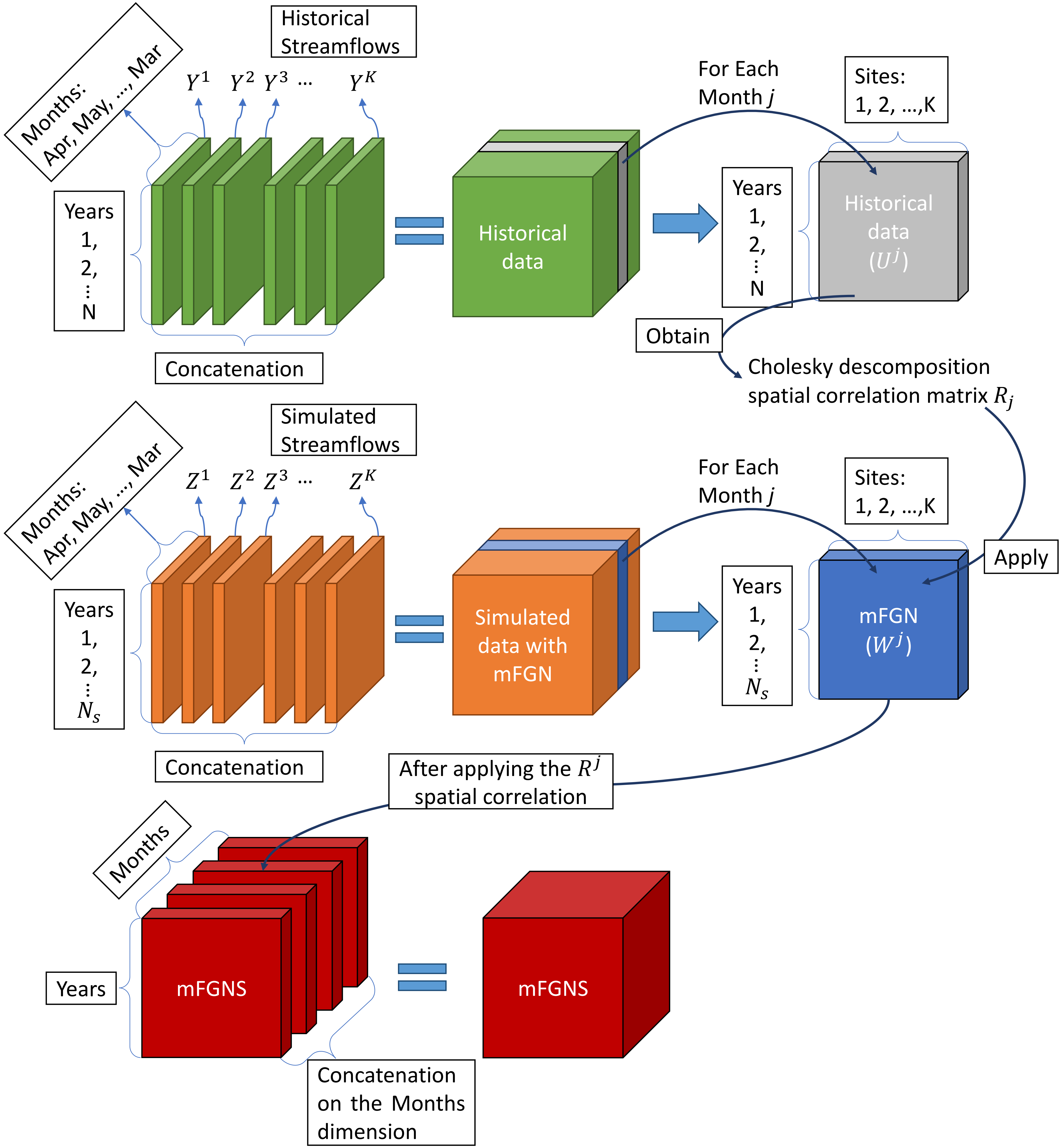

Months: Apr, May, ..., Mar
Historical Streamflows
Y^1 Y^2 Y^3 ... Y^K
Years 1, 2, ⋮ N
Concatenation
Historical data
For Each Month j
Sites: 1, 2, ...,K
Years 1, 2, ⋮ N
Historical data (U^j)
Obtain
Cholesky descomposition spatial correlation matrix R_j
Months: Apr, May, ..., Mar
Simulated Streamflows
Z^1 Z^2 Z^3 ... Z^K
Years 1, 2, ⋮ N_s
Concatenation
Simulated data with mFGN
For Each Month j
Sites: 1, 2, ...,K
Years 1, 2, ⋮ N_s
mFGN (W^j)
Apply
After applying the R^j spatial correlation
Months
Years
mFGNS
mFGNS
Concatenation on the Months dimension

Figure 3.

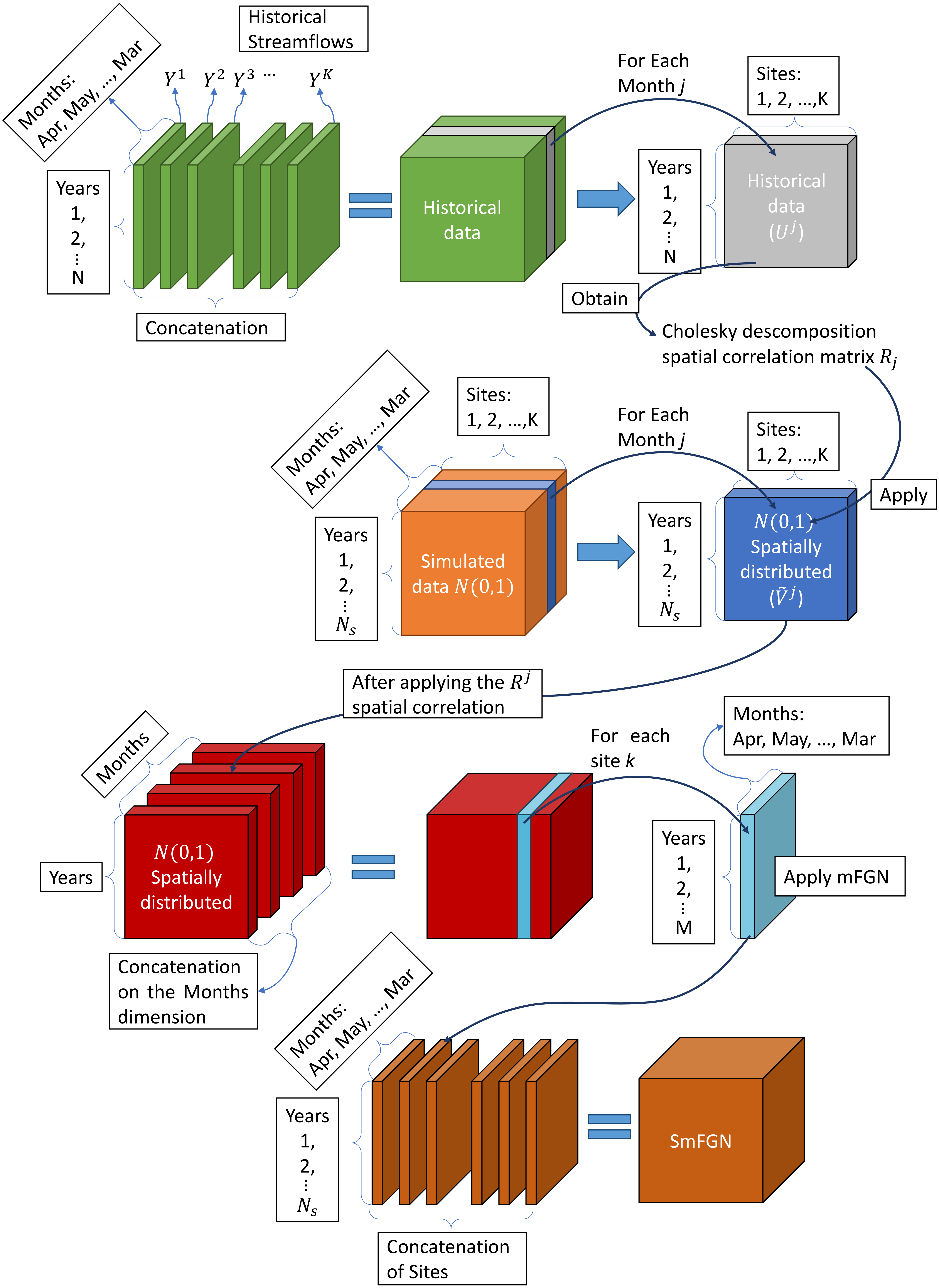
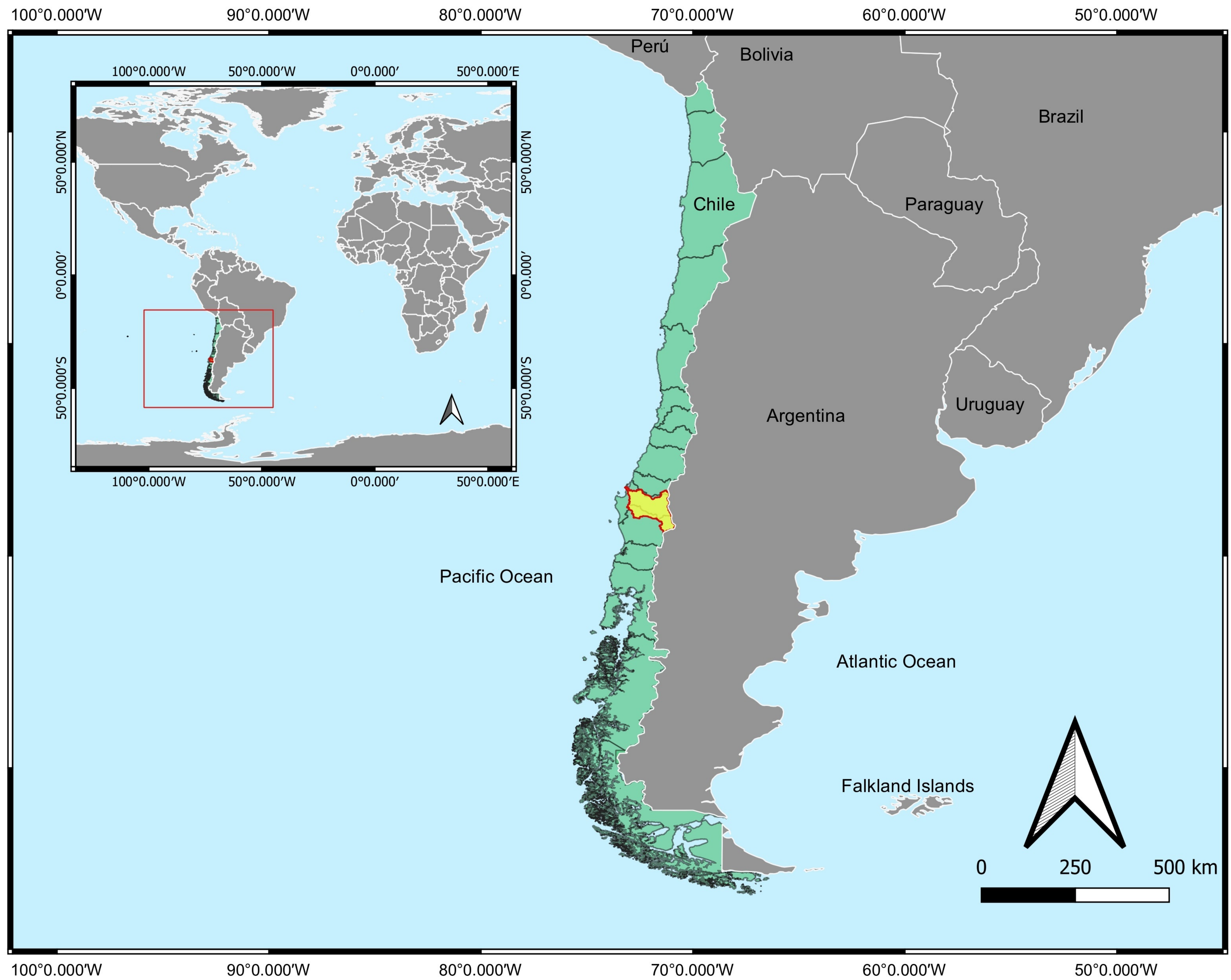

Months: Apr, May, ..., Mar
Historical Streamflows
Y^1 Y^2 Y^3 ... Y^K
Years 1, 2, ⋮ N
Concatenation
Historical data
For Each Month j
Sites: 1, 2, ...,K
Years 1, 2, ⋮ N
Historical data (U^j)
Obtain
Cholesky descomposition spatial correlation matrix R_j
Months: Apr, May, ..., Mar
Sites: 1, 2, ...,K
Years 1, 2, ⋮ N_s
Simulated data $N(0,1)$
For Each Month j
Sites: 1, 2, ...,K
Years 1, 2, ⋮ N_s
$N(0,1)$ Spatially distributed ($\tilde{V}^j$)
Apply
After applying the R^j spatial correlation
Months
Years
$N(0,1)$ Spatially distributed
Concatenation on the Months dimension
For each site k
Months: Apr, May, ..., Mar
Years 1, 2, ⋮ M
Apply mFGN
Months: Apr, May, ..., Mar
Years 1, 2, ⋮ N_s
SmFGN
Concatenation of Sites

Figure 4.


100°0.000′W
90°0.000′W
80°0.000′W
70°0.000′W
60°0.000′W
50°0.000′W
20°0.000′S
30°0.000′S
40°0.000′S
50°0.000′S
Perú
Bolivia
Brazil
Chile
Paraguay
Argentina
Uruguay
Pacific Ocean
Atlantic Ocean
Falkland Islands
0
250
500 km
100°0.000′W
50°0.000′W
0°0.000′
50°0.000′E
50°0.000′N
0°0.000′
50°0.000′S

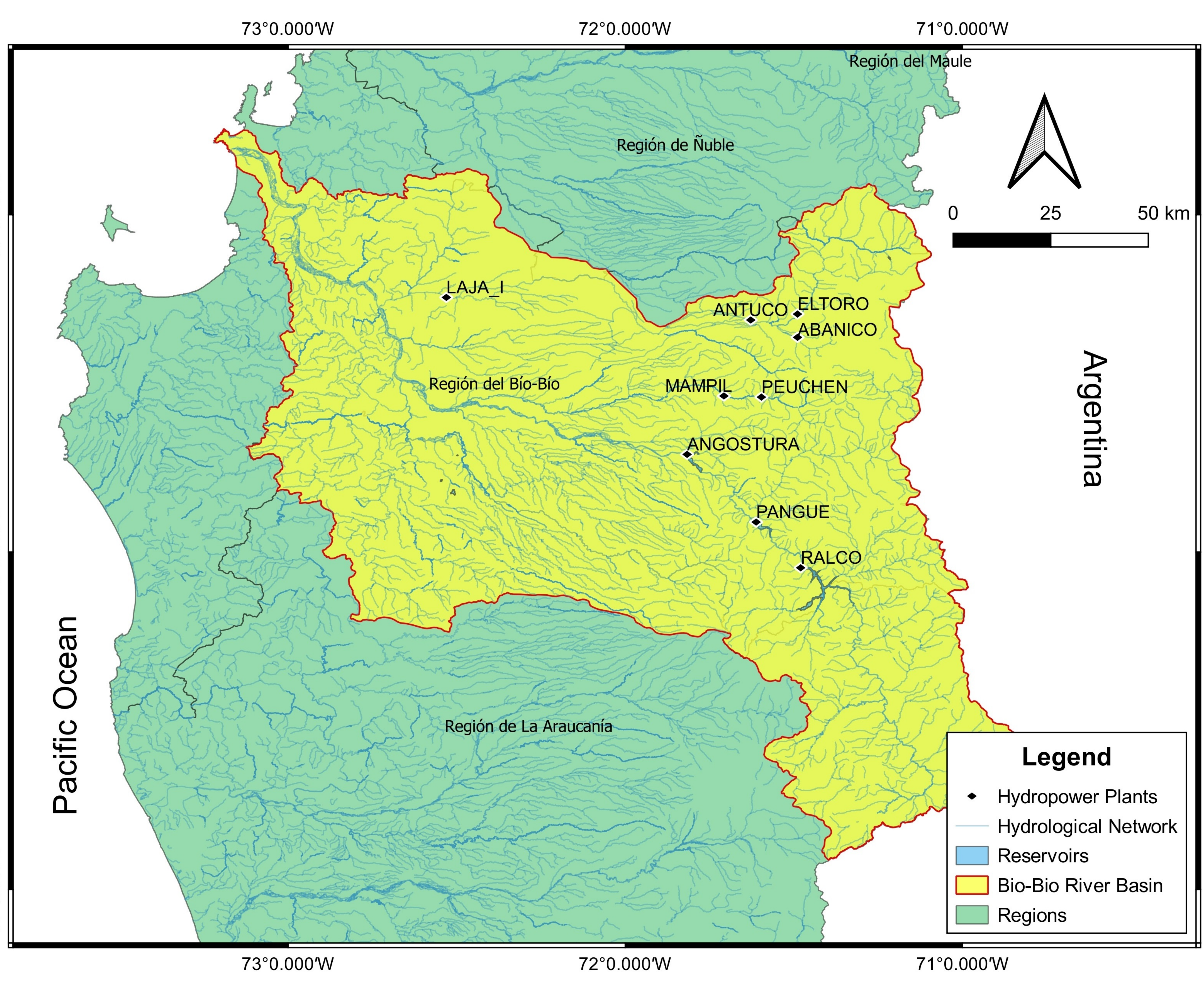

73°0.000′W
72°0.000′W
71°0.000′W
37°0.000′S
38°0.000′S
39°0.000′S
Región del Maule
Región de Ñuble
0
25
50 km
LAJA_I
ANTUCO
ELTORO
ABANICO
Región del Bío-Bío
MAMPIL
PEUCHEN
ANGOSTURA
PANGUE
RALCO
Argentina
Pacific Ocean
Región de La Araucanía
Legend
Hydropower Plants
Hydrological Network
Reservoirs
Bio-Bio River Basin
Regions

Figure 5.

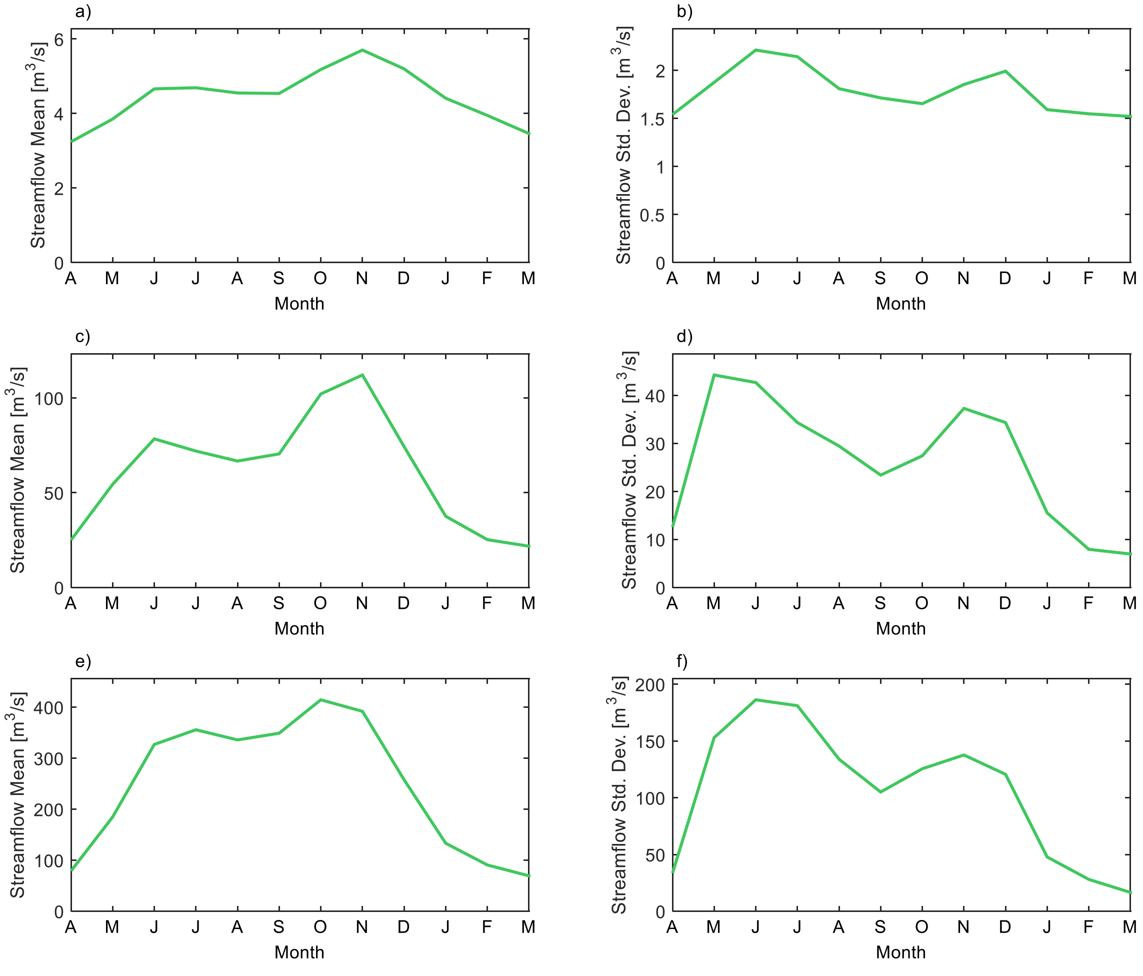
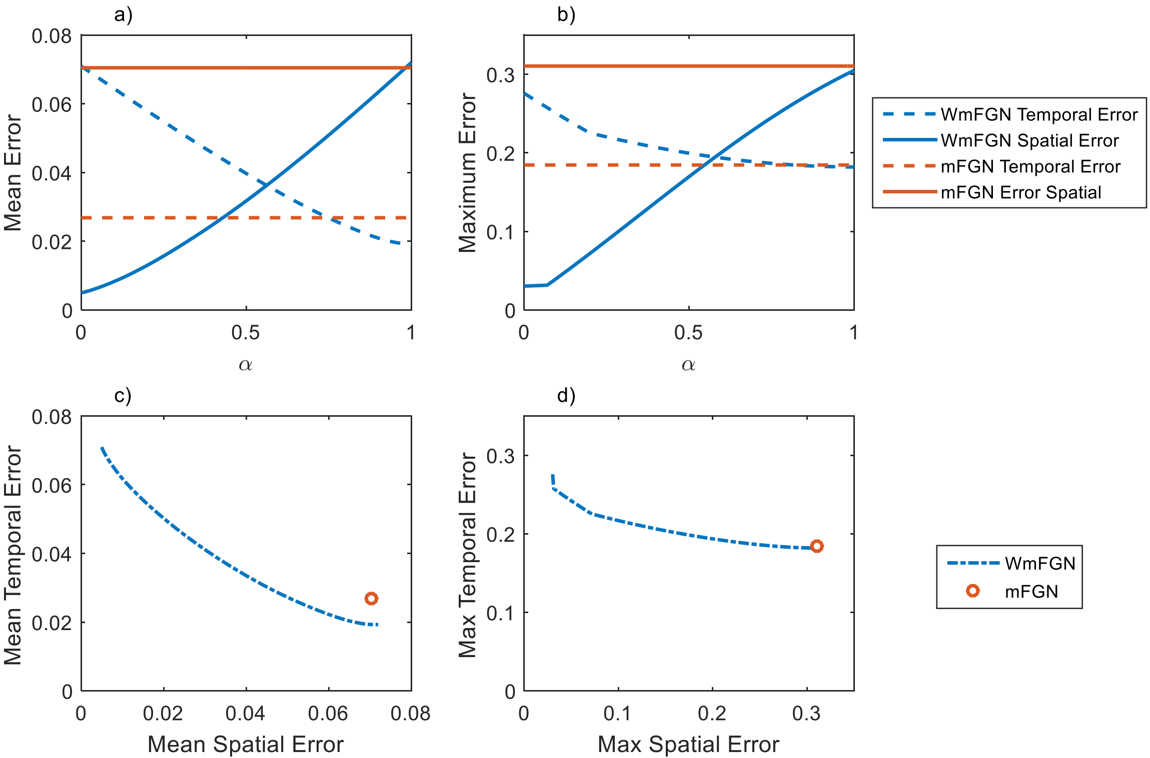
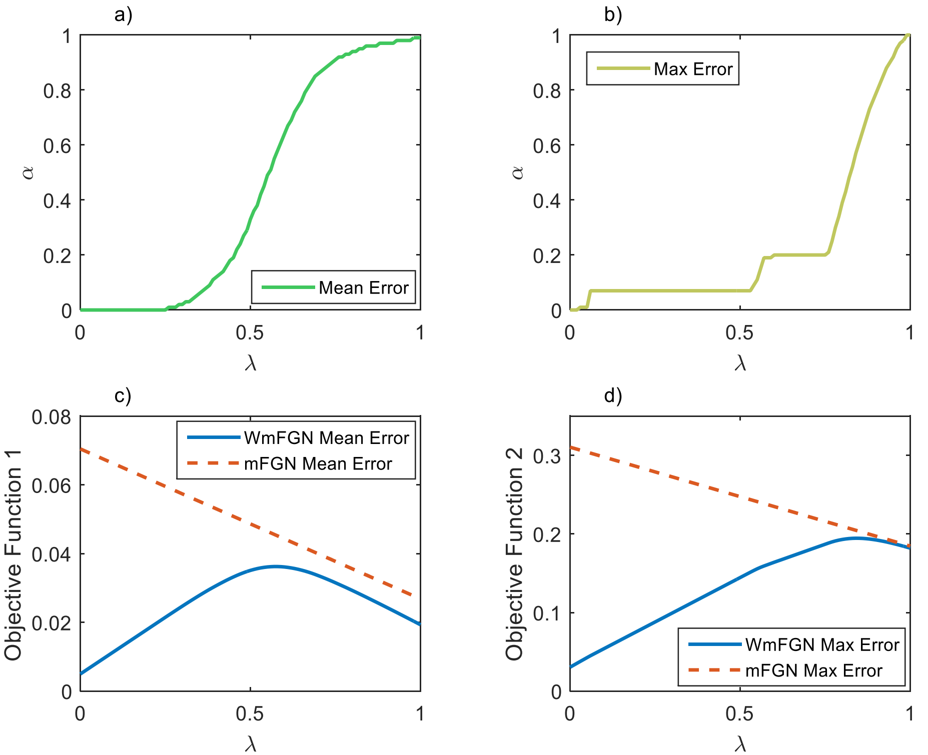

a)
b)
c)
d)
e)
f)
Streamflow Mean [m^3/s]
Streamflow Std. Dev. [m^3/s]
Month
A M J J A S O N D J F M

Figure 6.



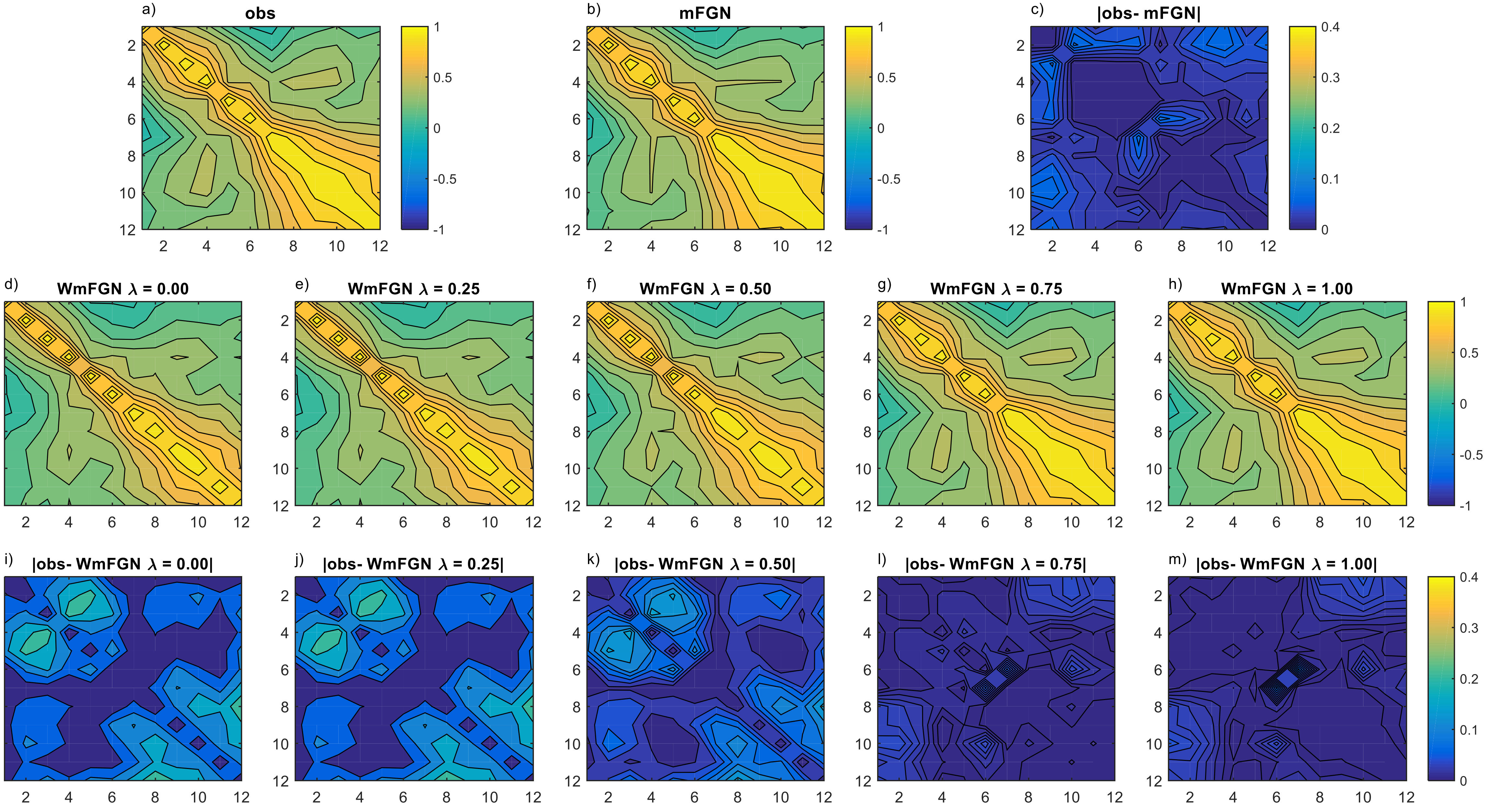
a)
b)
c)
d)
Mean Error
Maximum Error
Mean Temporal Error
Max Temporal Error
Mean Spatial Error
Max Spatial Error
α
WmFGN Temporal Error
WmFGN Spatial Error
mFGN Temporal Error
mFGN Error Spatial
WmFGN
mFGN

Figure 7.



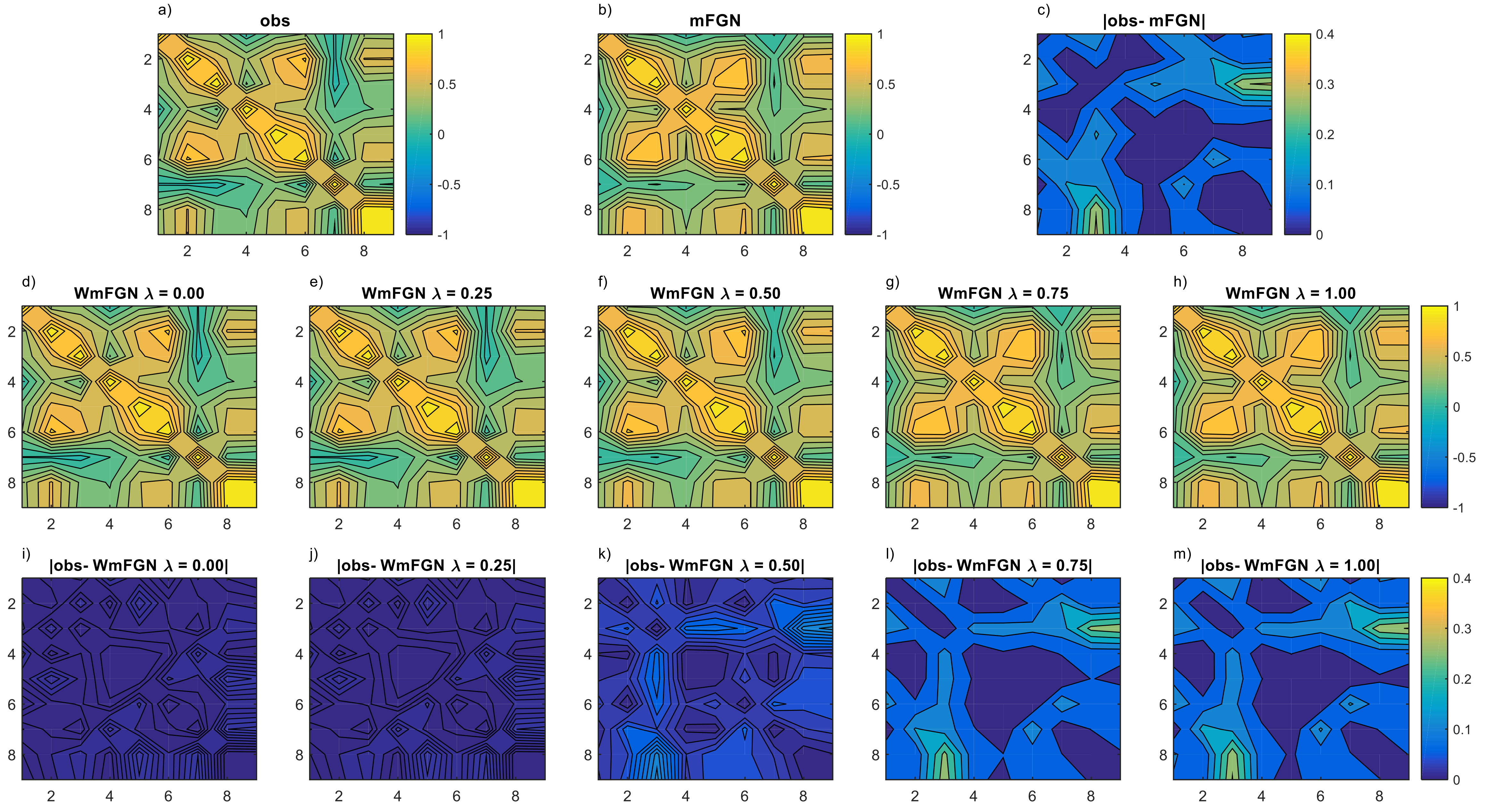
a)
b)
c)
d)
Mean Error
Max Error
WmFGN Mean Error
mFGN Mean Error
WmFGN Max Error
mFGN Max Error
α
λ
Objective Function 1
Objective Function 2
0
0.2
0.4
0.6
0.8
1
0.5
0.02
0.04
0.06
0.08
0.1
0.3

Figure 8.



a) obs
b) mFGN
c) |obs- mFGN|
d) WmFGN λ = 0.00
e) WmFGN λ = 0.25
f) WmFGN λ = 0.50
g) WmFGN λ = 0.75
h) WmFGN λ = 1.00
i) |obs- WmFGN λ = 0.00|
j) |obs- WmFGN λ = 0.25|
k) |obs- WmFGN λ = 0.50|
l) |obs- WmFGN λ = 0.75|
m) |obs- WmFGN λ = 1.00|

Figure 9.



a)
obs
b)
mFGN
c)
|obs- mFGN|
d)
WmFGN λ = 0.00
e)
WmFGN λ = 0.25
f)
WmFGN λ = 0.50
g)
WmFGN λ = 0.75
h)
WmFGN λ = 1.00
i)
|obs- WmFGN λ = 0.00|
j)
|obs- WmFGN λ = 0.25|
k)
|obs- WmFGN λ = 0.50|
l)
|obs- WmFGN λ = 0.75|
m)
|obs- WmFGN λ = 1.00|

Figure 10.

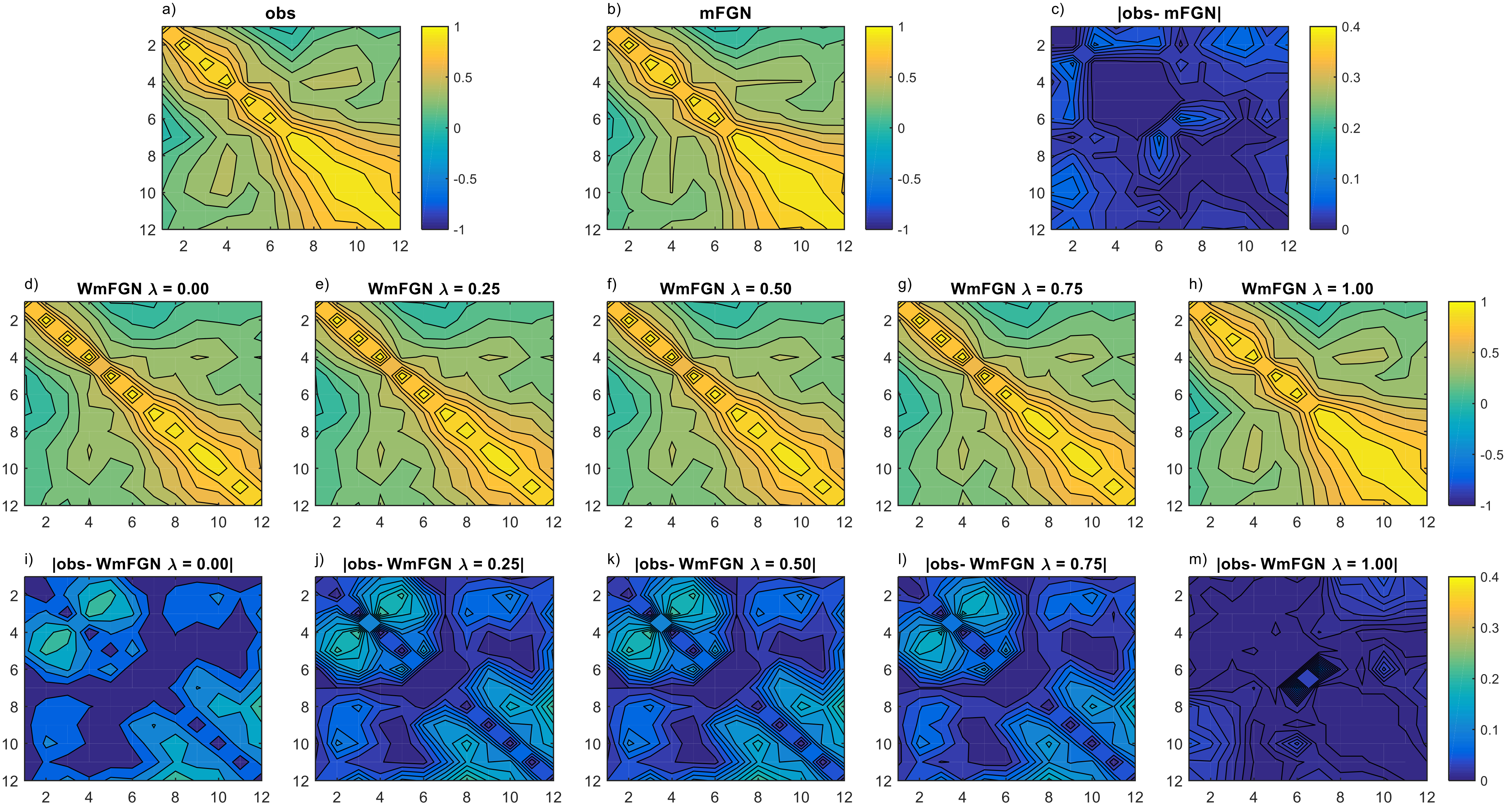
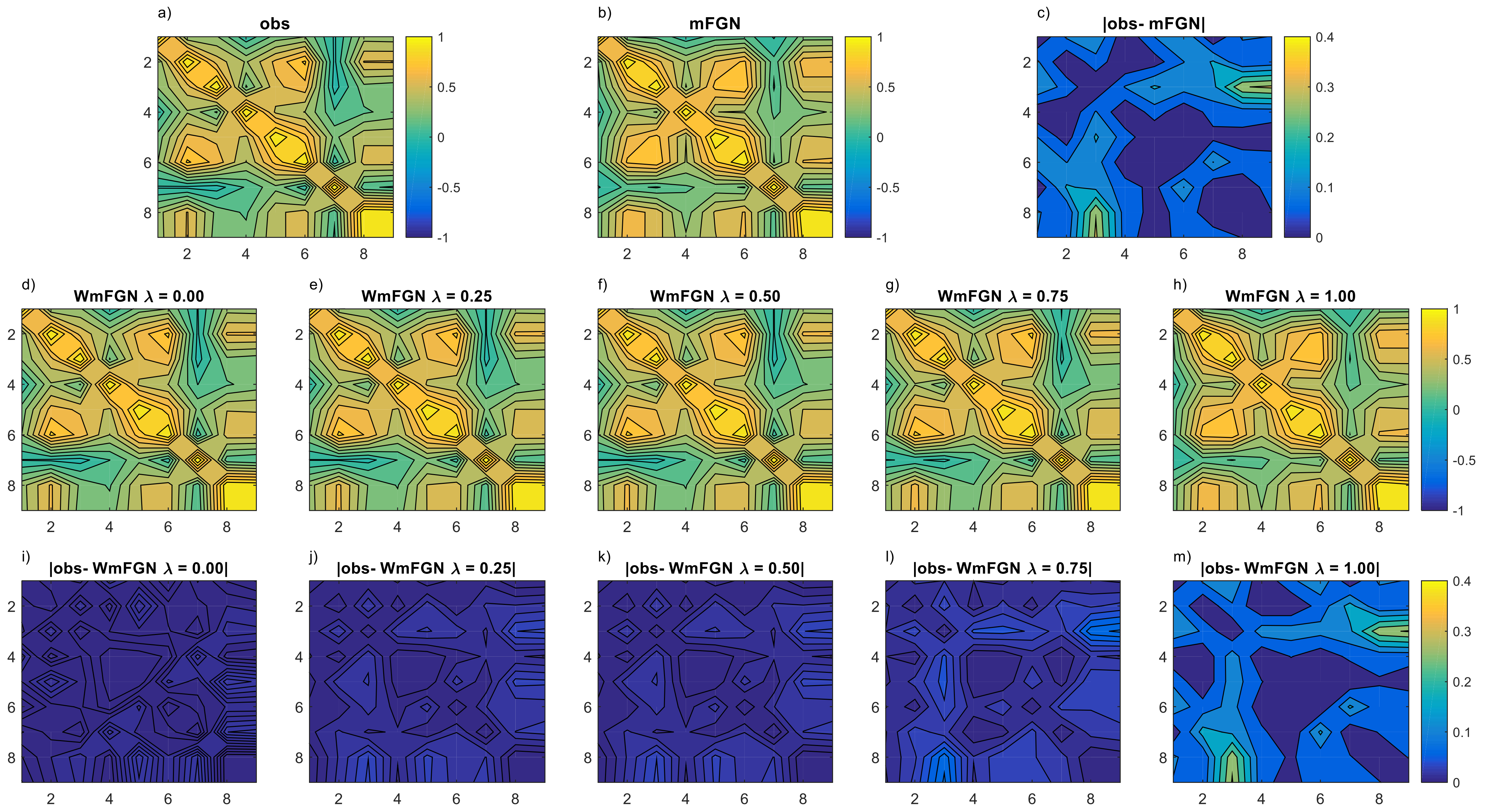

a) obs
b) mFGN
c) |obs- mFGN|
d) WmFGN λ = 0.00
e) WmFGN λ = 0.25
f) WmFGN λ = 0.50
g) WmFGN λ = 0.75
h) WmFGN λ = 1.00
i) |obs- WmFGN λ = 0.00|
j) |obs- WmFGN λ = 0.25|
k) |obs- WmFGN λ = 0.50|
l) |obs- WmFGN λ = 0.75|
m) |obs- WmFGN λ = 1.00|

Figure 11.


a)
obs
b)
mFGN
c)
|obs- mFGN|
d)
WmFGN λ = 0.00
e)
WmFGN λ = 0.25
f)
WmFGN λ = 0.50
g)
WmFGN λ = 0.75
h)
WmFGN λ = 1.00
i)
|obs- WmFGN λ = 0.00|
j)
|obs- WmFGN λ = 0.25|
k)
|obs- WmFGN λ = 0.50|
l)
|obs- WmFGN λ = 0.75|
m)
|obs- WmFGN λ = 1.00|