Copula autoregressive methodology for multi-lag,multi-site simulation of rainfall

Andres Felipe Ramirez[1] and Carlos Felipe Valencia[1]

[1]University of los Andes

November 26, 2022

Abstract

This work presents a methodology for the synthetic generation of rainfall time series based on the copula autoregressive methodology with multiple lags and for multiple sites. In this model, the multivariate time series is decomposed using pairwise copula functions to represent the whole cross-dependence, spatial and temporal structure of the data. We explore the advantages of using this nonlinear method over more traditional approaches that as an intermediate step transform the data to a normal distribution or usually omit the zero mass characteristics of the data. The use of copulas gives flexibility to represent the serial variability of the observed data on the simulation and allows for more control of the desired properties. We use discrete zero mass density distributions to assess the nature of rainfall, alongside a vector generalized linear model for the evaluation of time series distributions and their time dependence in multiple locations. We found that the copula autoregressive methodology models in a satisfactory manner the characteristics of the data, including its zero mass characteristics. These results will help to better understand the fluctuating nature of rainfall and also help to understand the underlying stochastic process.

Copula autoregressive methodology for multi-lag, multi-site simulation of rainfall

Andrés Felipe Ramírez[1], Carlos Felipe Valencia[1]

[1]University of los Andes, Cra 1 E 19A-40, Bogota, Colombia. Center for optimization and applied probability (COPA)

Key Points:

- We applied a multi-site and multi-lag methodology for simulation of rainfall time series based on copula functions.
- We use bivariate Copula functions to model the discrete-continuous distribution, allowing to model the occurrence and amount of rainfall.
- Marginal distributions of rainfall are estimated using a VGLM considering discrete-continuous distributions including temporal covariables.

Corresponding author: Andrés Felipe Ramírez, `af.ramirez12@uniandes.edu.co`

Corresponding author: Carlos Felipe Valencia, `cf.valencia@uniandes.edu.co`

Abstract

This work presents a methodology for the synthetic generation of rainfall time series based on the copula autoregressive methodology with multiple lags and for multiple sites. In this model, the multivariate time series is decomposed using pairwise copula functions to represent the whole cross-dependence, spatial and temporal structure of the data. We explore the advantages of using this nonlinear method over more traditional approaches that as an intermediate step transform the data to a normal distribution or usually omit the zero mass characteristics of the data. The use of copulas gives flexibility to represent the serial variability of the observed data on the simulation and allows for more control of the desired properties. We use discrete zero mass density distributions to assess the nature of rainfall, alongside a vector generalized linear model for the evaluation of time series distributions and their time dependence in multiple locations. We found that the copula autoregressive methodology models in a satisfactory manner the characteristics of the data, including its zero mass characteristics. These results will help to better understand the fluctuating nature of rainfall and also help to understand the underlying stochastic process.

Keywords— Rainfall simulation, Copula autoregressive model, Multi-site rainfall simulation, multi-lag rainfall simulation, Vector Generalized Linear Model

1 Introduction

Stochastic generation of synthetic rainfall time series is a fundamental tool in areas such as hydrology, agriculture, engineering, climate and energy, where it can be used for the analysis of complex systems using numerical implementations (D. S. Wilks & Wilby (1999); Srikanthan & McMahon (2001)). A correct representation of the stochastic nature of rainfall includes its spatial and temporal properties. These characteristics allow for the study of several phenomena without the necessity of real intervention; moreover, they support the optimal design of larger systems that otherwise would require difficult physical or mathematical models. Furthermore, synthetic generation in multiple locations is applied more widely in applications than the single-location counterpart despite the requirement for the representation of the spatial dependence between the locations. For example, in energy production and the integration of renewable sources, the correct understanding and modeling of the stochastic process for rainfall is crucial for the evaluation of complementary viability and reliability in power systems (Tinaikar (2013)).

The problem of synthetically generating daily rainfall time series has been approached with different methodologies during the last 40 years. A detailed description and classification of the models used can be consulted in D. S. Wilks & Wilby (1999), Srikanthan & McMahon (2001), Mehrotra et al. (2006) or Vaittinada Ayar et al. (2020). One taxonomy is to separate the models between the single-site models that are concerned with the correct specification of the statistical properties of observed time series to reproduce them with simulated data and the models that extend these properties to the multi-site problem on which the spatial relations have to be incorporated.

The difficulty of the rainfall stochastic process is that it represents two underlying events, the occurrence and the amount. When seeing rainfall as a single process, the marginal distribution at each time should have a mass on zero, that is, the probability of the amount being equal to zero is greater than 0. Some methodologies approach this problem using a two-step mechanism that first simulates the occurrence (for example using a first-order discrete Markov Chain) and conditionally on this they simulate the amount of rainfall (e.g. Richardson (1981), D. Wilks (1998), Mhanna & Bauwens (2012),). As a result of this method, the random variables that model the amount of rainfall are conditionally independent from the ones that model the occurrence. For example, if there is rain on two consecutive days, the amount of rain on the second day is independent from the amount on the first day. This could undermine the representation of some statistical properties such as the autocorrelation function when looking at the series conditionally on the occurrence. One alternative is to discretize the stochastic process and then use a first or second order Markov Chain (e.g. Haan et al. (1976)); however, the main repercussion is the less accurate representation of the amount of rainfall in the marginal densities and the addition in computational effort needed when using a large number of states. Another branch of models uses resampling techniques such as bootstrapping in order to approximate the distribution of observed series (e.g.Buishand & Brandsma (2001)). Other strategies are known to assume that the non-zero part of the dataset might have normal marginal densities (Zhou et al. (2020), Ahn (2020), Ayar et al. (2020)). Once the new data is normal, the derived methodologies take advantage of the well known linear ARIMA (autoregressive integrated moving average) stochastic process to construct theoretically robust statistical models. With normal time series, it is possible to incorporate temporal and spatial dependence in the first two moments of the distributions; however, this strategy depends on the marginal densities and the threshold used for censored values and positive rainfall measurements. This approach might poorly estimate several characteristics of the actual process (Zhou et al. (2019)). Furthermore, rainfall data is unlikely to correlate symmetrically for low and high values (Bárdossy et al. (2017)) and the required transformation from the marginals to the normal distribution (gaussianization) conditions the behavior of the series (Sarmiento et al. (2018)).

The use of copula functions to model the time or spatial dependence of rainfall time series has been explored more during the last years, where several applications in hydrology use copula functions to assess the complexity of the data (e.g. Kao & Govindaraju (2008), Bárdossy & Pegram (2009), Zhang & Singh (2019), Serinaldi (2009a), Serinaldi (2009b)). The benefits of using the copula function is that they can estimate the underlying stochastic process and structure between several rainfall time series independently of their marginal distribution functions (Vandenberghe et al. (2010), Balistrocchi & Bacchi (2011)). This provides enough flexibility on the relations among variables to overcome some limitations of linear Gaussian processes that rely on specific transformations (Bárdossy et al. (2017), Sarmiento et al. (2018)). Several studies, such as the one conducted by Zhang & Singh (2007) suggest that the copula based distribution function fits the dependence structure of observed rainfall characteristics data series better than the multivariate normal probability distribution.

Copula functions are naturally defined for continuous variables to accommodate the estimation for zero inflated distributions. Authors like Serinaldi (2009a), Serinaldi (2009b) and Li et al. (2013) use a partition in quadrants of the bivariate uniform distributions to generate conditional distributions. These partitions are adaptation of the ideas of Shimizu (1993) and Herr & Krzysztofowicz (2005) applied to the copula model for time series. Although these copula models are a great tool to model the series dependence structure up to lag 1 in an autoregressive process, the spatial dependence of multi-site time series is not usually modeled directly from the between series and conditional series dependence. These dependence structures in daily rainfall time series make it difficult to model over the range of observations using only linear correlation coefficients, such as Pearson correlation. Instead, dependence methodologies such as Kendall's τ are often used to model the correlation between locations (i.e Serinaldi (2009a), Serinaldi (2009b)).

In this paper, we propose a fully copula based methodology to model time series of rainfall within an autoregressive process with multiple lags and with spatial dependence based also on copula functions. That is, the whole stochastic system can be expressed as a multivariate model decomposed in pairwise copula functions using the COPAR methodology (Brechmann et al. (2015)), which bases its calculations on generating all the pair copula construction needed for the between series and conditional series dependence among the different time series for each of the sites following an R-vine structure. In addition, we employ a vector generalized linear model (VGLM) to fit a zero inflated continuous distribution in terms of temporal variables such as the month of the year. With this method, it is possible to estimate the marginal distributions in one model avoiding the possible over-fitting that can be created when partitioning the sample and estimating the marginal distributions for each month.

The structure of this paper goes as follows. The methodology used as well as the contributions made to the state of the art are described in Section 2. A case study is presented in Section 3 along with the evaluation of the results. Finally, section 4 presents a discussion about the methodology used and the future work, as well as the conclusions.

2 Methodology

The rainfall synthetic data generator is based on a model that has two main components: (i) marginal distribution estimation through a VGLM for mixtures of densities with discrete mass distributions (Section 2.1), and (ii) a multivariate copula algorithm for modeling the temporal and spatial dependence in one statistical model using decomposition in bivariate uniform densities (Section 2.2).

The rainfall time series is not a stationary process given that, according to a specific location, there are seasons and inter-annual cycles that have to be considered. These seasonal components of the series directly affect the distribution of the marginal random variable. Contrary to Gaussian time series, on which the seasonal effect is modeled on the mean, the copula model allows for a more flexible temporal effect in some (or all) the parameters of the distributions. The reason for this is that the temporal and spatial dependencies are modeled through the uniform variables resulting from using the inverse cumulative probability function transformation. Using a VGLM to estimate all rainfall marginal densities permits the use of all available data to include the temporal effects as independent variables. Furthermore, the occurrence process (rain or no-rain) is embedded in the marginal distribution by using a mixture model that combines the positive part with the probability mass on zero. Once the multivariate rainfall time series is standardized to be uniform $[0, 1]$ for all values, it is modeled by the copula functions to express the whole joint probability distribution. This distribution has many components (dimensions) that would make it difficult to estimate without simplification rules to avoid the curse of dimensionality. We use a decomposition in a R-vine structure and limit the numbers of lags that are significant, as done in a linear autoregressive (AR) model. This approach extends the previous models in Serinaldi (2009a) and Li et al. (2013) that only permit one lag. The same principle can be used to model the multivariate distribution applied to the spatial dependence. The copula autoregressive methodology models the entire dependence of the time series differing from more traditional tools for modeling time series such as autoregressive processes (AR) that transform any distribution to normal.

2.1 Rainfall marginal distribution

The marginal distributions of ground measured rainfall is modeled using a vector generalized linear model (VGLMs). Unlike the classical generalized linear models (GLMs), there is not a restriction on the number of parameters in the distributions, and they are purposely general to allow greater utility (Yee, 2015). Consider the data set expressed as (x_i, y_i) for $i = 1, \ldots, n$, where x_i is a vector of p explanatory variables and y_i is the response for the observation i. The objective is to fit a generalized regression model involving the parameter

ν_j, where $j = 1, \ldots, J$. The VGLM model, estimates each one of the parameters in the generalized regression as the function of a linear combination of the explanatory variables:

$$g_j(\nu_j) = \beta_j^T x = \beta_{(j)1} X_1 + \beta_{(j)2} X_2 + \cdots + \beta_{(j)p} X_p \quad , \tag{1}$$

where g_j is a parameter link function. In this study, we estimated the parameters of the VGLMs with the Interactively Re-weighted Least Squares (IRLS) algorithm. For a detailed explanation of the estimation of the VGLM, the reader is referred to Yee (2015)

For modeling the marginal probability distribution of rainfall, we selected the zero adjusted gamma distribution, due to the zero mass characteristics presented in the data. This is a special case of zero adjusted distributions on zero and the positive real line. The zero adjusted gamma distribution is a combination of a discrete value 0 with probability ν and a gamma GA(μ,σ) distribution with probability (1-ν), where μ is the mean and σ is the square root of the dispersion parameter according to the exponential family factorization. The probability function of the zero adjusted gamma distribution denoted by ZAGA($\mu(t)$,$\sigma(t)$,$\nu(t)$) is given by:

$$f_y(y \mid \mu(t), \sigma(t), \nu(t)) = \begin{cases} \nu(t) & \text{y} = 0 \\ (1-\nu(t)) f_w(y \mid \mu(t), \sigma(t)) & \text{y} > 0 \end{cases} \tag{2}$$

For $0 \leq y(t) < \infty$, where $\mu(t) > 0$, $\sigma(t) > 0$, $0 < \nu(t) < 1$ and W $\sim$ GA(μ(t),σ(t)) has a gamma distribution with density:

$$f_w(y \mid \mu(t), \sigma(t)) = \frac{y^{\frac{1}{\sigma^2}-1} \exp\left(\frac{-y}{\sigma^2 \mu}\right)}{(\sigma^2 \mu) \Gamma(\frac{1}{\sigma^2})} \quad \text{y} > 0 \quad , \tag{3}$$

where the dependence of the parameters on t is left implicit for simplicity. The mean of the ZAGA distribution is $(1-\nu(t))\mu$ and the variance is $(1-\nu(t))\mu^2(\nu + \sigma^2)$. The independent variable t represents the yearly seasons that change the marginal distributions. In our case, we model t as the month of the year using a factor with values (1-12). The default link functions that relate the parameter (μ,σ,ν) depending on the time are:

$$log(\mu) = \beta_\mu^T t \tag{4}$$

$$log(\sigma) = \beta_\sigma^T t \tag{5}$$

$$log(\nu) = \beta_\nu^T t \tag{6}$$

2.2 The copula autoregressive model

The COPAR model proposed by Brechmann et al. (2015) exploits the flexibility of vine copulas for non-linear and asymmetric modeling of serial and between-series dependence. The fundamental pieces to build the autoregressive model are the bivariate copulas, which are distributions on the unit square $[0,1]^2$ such that both marginals are uniform U(0,1). The theorem by Sklar (1959) explains that for any given variables X and Y with joint distribution $F_{X,Y}(x,y)$ and marginals cumulative distribution functions (CDF) $F_X(x)$ and $F_Y(y)$ respectively, there exists a unique copula function $C_{XY}(\cdot,\cdot)$ that connects $F_{X,Y}(\cdot,\cdot)$ to $F_X(\cdot)$ and $F_Y(\cdot)$ via $F_{X,Y}(x,y) = C_{XY}(F_X(x), F_Y(y))$. Therefore, the information in the joint distribution is decomposed into that of the marginal distributions and the one of the copula function, where the latter captures the entire dependence structure between X and Y (Chen & Fan, 2006). For a detailed explanation on copula theory the reader is referred to Nelsen (1999).

2.2.1 Pair-copula decomposition for univariate time series

Let $\{X_t\}_{t=1,\cdots,T}$ be a univariate stationary time series of continuously distributed data, with marginal distribution function F_{X_t} and density function f_{X_t}. The joint distribution of $\{X_t\}$ can be decomposed by selecting models for the conditionals as:

$$f(x_1,\cdots,x_T) = f(x_1)\prod_{\mathrm{t}=2}^{T} f(x_t \mid x_{t-1},\cdots,x_1). \tag{7}$$

Smith et al. (2010) outline that this expression can be used to obtain a general decomposition in terms of bivariate copulas, as for every $s<t$ exists a copula density $c_{t,s}$ such that:

$$\begin{aligned} f(x_t, x_s \mid x_{t-1},\cdots,x_{s+1}) = c_{t,s}(F(x_t \mid x_{t-1},\cdots,x_{s+1}), F(x_s \mid x_{t-1},\cdots,x_{s+1}))\cdot \\ f(x_t \mid x_{t-1},\cdots,x_{s+1}), f(x_s \mid x_{t-1},\cdots,x_{s+1})\ , \end{aligned} \tag{8}$$

where $F(X_t \mid X_{t-1},\cdots,X_{s+1})$ and $F(X_s \mid X_{t-1},\cdots,X_{s+1})$ are the conditional distributions functions of X_t and X_s respectively. This expression is the density of the Sklar theorem, conditional upon $\{X_{t-1},\cdots,X_{s+1}\}$. Rearranging terms in Equation (8) gives:

$$\begin{aligned} f(x_t \mid x_{t-1},\cdots,x_s) = c_{t,s}(F(x_t \mid x_{t-1},\cdots,x_{s+1}), F(x_s \mid x_{t-1},\cdots,x_{s+1}))\cdot \\ f(x_t \mid x_{t-1},\cdots,x_{s+1}). \end{aligned} \tag{9}$$

By recursive conditioning on $s = 1; 2; \cdots, t-1$ we obtain:

$$f(x_t \mid x_{t-1}, \cdots, x_1) = \prod_{j=1}^{t-2} \{c_{t,j}(F(x_t \mid x_{t-1}, \cdots, x_{j+1}), F(x_j \mid x_{t-1}, \cdots, x_{j+1}))\} \cdot$$
$$c_{t,t-1}(F(x_t), F(x_{t-1}))f(x_t). \tag{10}$$

For notation simplicity, denote $u_{t|j} = F(X_t \mid X_{t-1}, \cdots, X_j)$ and $u_{j|t} = F(X_j \mid X_t, \cdots, X_{j+1})$ as the projections backwards and forwards $t-j$ steps, respectively. Also, by denoting $u_{t|t} = F(X_t)$, the joint density function in equation (7) can be written as:

$$f(x_1, \cdots, x_T) = \prod_{t=2}^{T} \left[\prod_{j=1}^{t-1} \{c_{t,j}(u_{t|j+1}, u_{j|t-1})\} f(X_t) \right] f(X_1). \tag{11}$$

This decomposition of the joint density function of the time series allows the methodology to achieve very flexible models, as no restrictions are necessary in the selection of the copula families. Nonetheless, copulas corresponding to the same time lag must be identical to assure that the simulated time series is stationary. This particular pair-copula decomposition is known as D-vine and takes part in a more general class of decomposition known as R-vines, which is a graphic theoretic model to establish which pair-copulas are included in the decomposition of a time series. An R-vine on d variables is a sequence of $d-1$ linked trees (connected acyclic graphs) that satisfy several conditions. R-vine theory is described in detail by Bedford (2001) and H Joe (2011).

2.2.2 The COPAR model for multivariate time series

Although the model works in general for any number of dimensions, we explain the copula autoregressive model for two univariate time series $\{X_t\}_{t=1,\cdots,T}$ and $\{Y_t\}_{t=1,\cdots,T}$ jointly distributed at time point $t = 1, \cdots, T$ for brevity in the exposition. A flexible multivariate distribution of $\{X_t\}$ and $\{Y_t\}$ is presented, which allows the nonlinear dependence-serial as well as between-series dependence structure to be constructed. The model is based on a particular R-vine structure that has the following components (Brechmann et al., 2015).

1. Marginal distributions F_X and F_Y of $\{X_t\}$ and $\{Y_t\}$
2. An R-vine for the serial and between-series dependence of $\{X_t\}$ and $\{Y_t\}$, with the selection of the following pair-copulas:

(a) Serial dependence of $\{X_t\}$:

$$X_s, X_t \mid X_{s+1}, \cdots, X_{t-1} \quad \forall \quad 1 \le s \le t \le T \tag{12}$$

(b) Between series dependence:

$$X_s, Y_t \mid X_{s+1}, \cdots, X_t \quad \forall \quad 1 \leq s \leq t \leq T \tag{13}$$

and

$$Y_s, X_t \mid X_1, \cdots, X_{t-1}; Y_{s+1}, \cdots, Y_{t-1} \quad \forall \quad 1 \leq s \leq t \leq T \tag{14}$$

(c) Conditional serial dependence of $\{Y_t\}$:

$$Y_s, Y_t \mid X_1, \cdots, X_t; Y_{s+1}, \cdots, Y_{t-1} \quad \forall \quad 1 \leq s \leq t \leq T \tag{15}$$

As mentioned before, pair-copulas with equal lag must be identical. It is important to notice that $\{X_t\}$ has a pivotal role in this modeling approach because the serial dependence of $\{X_t\}$ is modeled unconditionally, meanwhile that of $\{Y_t\}$ is specified conditionally on $\{X_t\}$. The selection of the copulas is performed sequentially since, for example, $c_{X_t X_{t-2}|X_{t-1}}$ depends on the copula $c_{X_t X_{t-1}}$. Finally, the order of the model (k) is selected if all pair-copulas corresponding to a lag length greater than k are independence copulas.

2.2.3 Pair-copula estimation of variables with zero mass

One of the contributions of this paper is the inclusion of variables with zero mass in the copula autoregressive modeling for time series shown in section 2.2.2. In order to model these particular distributions, we selected a zero adjusted gamma distribution for the daily rainfall data, which is a mixture distribution that combines a Bernoulli and a Gamma probability distribution, dependent on the parameter $\nu(t)$ which models the probability of zero or non-zero values. This distribution is explained in depth in section 2.1. The zero mass characteristics of the data makes the density probability function $f_Y(y)$ to have a jump discontinuity on zero, (Figure 1). For modeling the copula joint distribution, however, it is necessary to evaluate this function and calculate the respective distribution function $F_Y(y)$, since the results of this calculation should be a continuous uniform variable. Moreover, to generate synthetic data, the inverse function F_Y^{-1} has to be applied on random numbers. To address this issue, we adopt the approach described in Faugeras (2012), from which the bivariate copula estimation of the autoregressive model explained in section 2.2 can be assessed with some fundamental considerations. In particular, for the points that have a mass value at $\nu(t)$, a uniform variable is generated when evaluating F_Y to fill the gaps created by the jumps. In this case, it is clear that $F_Y(0) = \nu$, and therefore, for the purpose of evaluation and simulation, it is replaced by a uniform random variable between 0 and

ν. Figure 2 presents the evaluation of $F_Y(y)$ for all the points in Figure 1, where $\nu = 0.6$. Also, Figure 2 presents the respective histogram. Figure 3 show the counterparts after using the randomization procedure. The proposed strategy to use copula functions with mixed random variables with probability mass in zero is explained in algorithm 1.

for $t \leftarrow 1$ **to** T **do**

$$u_{t|t} = F_y = \begin{cases} \nu(t) & \text{y} = 0 \\ \nu(t) + (1-\nu(t))\frac{\gamma(\sigma(t)^{-2}, y\mu(t)^{-1}\sigma(t)^{-2})}{\Gamma(\sigma(t)^{-2})} & \text{y} > 0 \end{cases} \tag{16}$$

end

for $t \leftarrow 1$ **to** T **do**

  if $u_{t|t} = \nu(t)$ **then**

    $u_{t|t} = runif(0, \nu(t))$

  end

end

Algorithm 1: Synthetic non zero uniform data generation

Where $F_y = (y \mid \mu(t), \sigma(t), \nu(t))$ is the distribution function of the zero adjusted gamma distribution for $0 \leq y(t) < \infty$, with $\mu(t) > 0$, $\sigma(t) > 0$, $0 < \nu(t) < 1$. Also, $runif$ is a random uniform variable between 0 and $\nu(t)$.

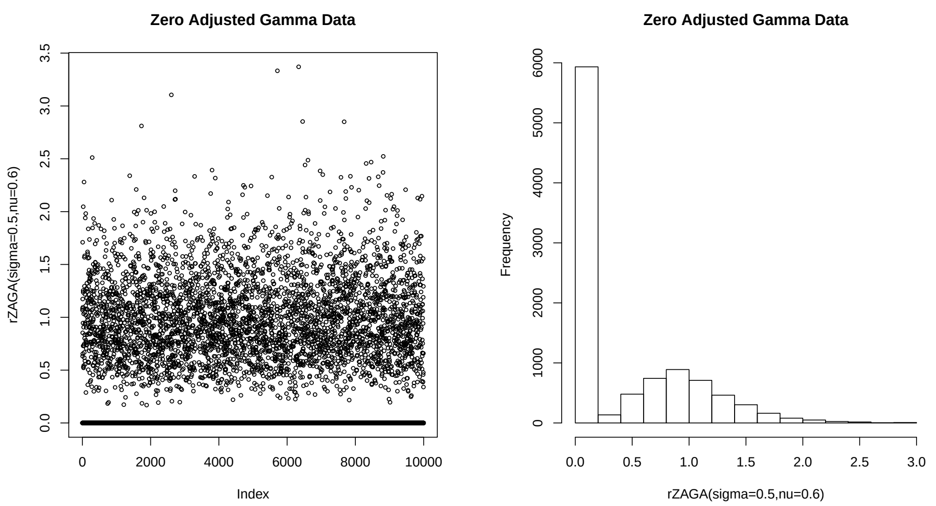

Figure 1. Scatter plot (LEFT) and histogram (RIGHT) of density function for distribution ZAGA(sigma=0.5, ν=0.6)

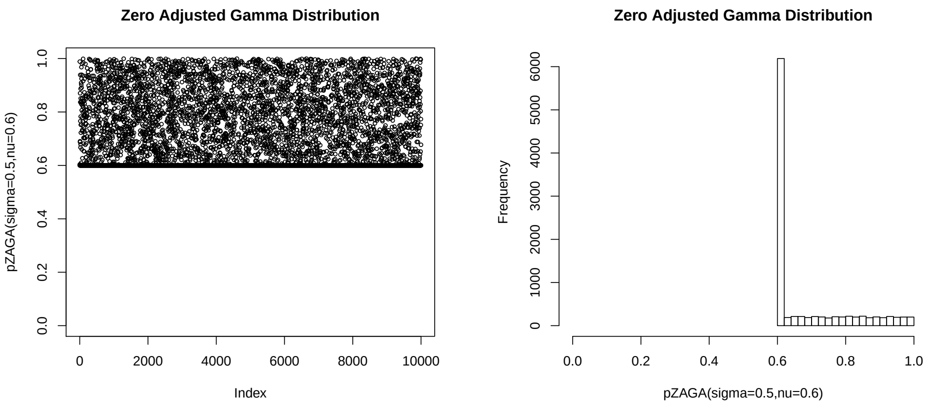

Figure 2. Scatter plot (LEFT) and histogram (RIGHT) of CDF function w/o Algorithm 1 for distribution ZAGA(sigma=0.5, ν=0.6)

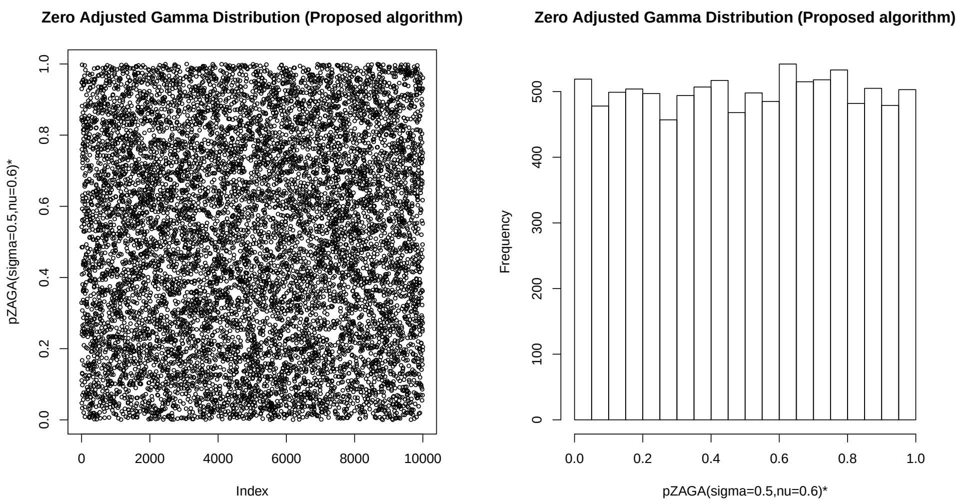

Figure 3. Scatter plot (LEFT) and histogram (RIGHT) of CDF function w/ Algorithm 1 for distribution ZAGA(sigma=0.5, ν=0.6)

If X and Y are two jointly distributed variables with probability mass on zero, the probability distribution function (CDF) $F_{XY}(x, y)$ may be defined conditionally on four cases with respective probabilities: $P_{00} = P(X = 0, Y = 0)$, $P_{10} = P(X > 0, Y = 0)$, $P_{01} = P(X = 0, Y > 0)$ and $P_1 = P(X > 0, Y > 0)$. Therefore, it is possible to write it as:

$$F_{XY}(x, y) = P_{00} + P_{10}H_X(x) + P_{01}H_Y(y) + P_{11}H_{XY}(x, y) \ , \tag{17}$$

where $H_{XY} = C^{11}_{XY}(F^1_X(x), F^1_Y(y))$ can be modeled as a bivariate continuous copula function (conditional on both variables being positive), with $F^1_X(x) = P(X \leq x|X > 0, Y > 0)$ and $F^1_Y(y) = P(Y \leq y|X > 0, Y > 0)$. According to this, the resulting bivariate copula function to represent $F_{XY}(x, y)$ can be also decomposed in quadrants. That is, if X and Y are both marginally distributed $ZAGA$ with $p_x = \nu(X) = P(X = 0)$, and $p_y = P(Y = 0)$, then $F_{X,Y} = C_{XY}(u, v)$, where $u = F_X(x)$, and $v = F_X(y)$. Let $c_{XY}(\cdot, \cdot)$ be the correspondent bivariate density of $C_{XY}(\cdot, \cdot)$, then it would look as Figure 4. With this function, it is possible to define the conditional cumulative distributions of Y given X, and their respective inverse, using the splitting mechanism described in Herr & Krzysztofowicz (2005) and Serinaldi (2009a).

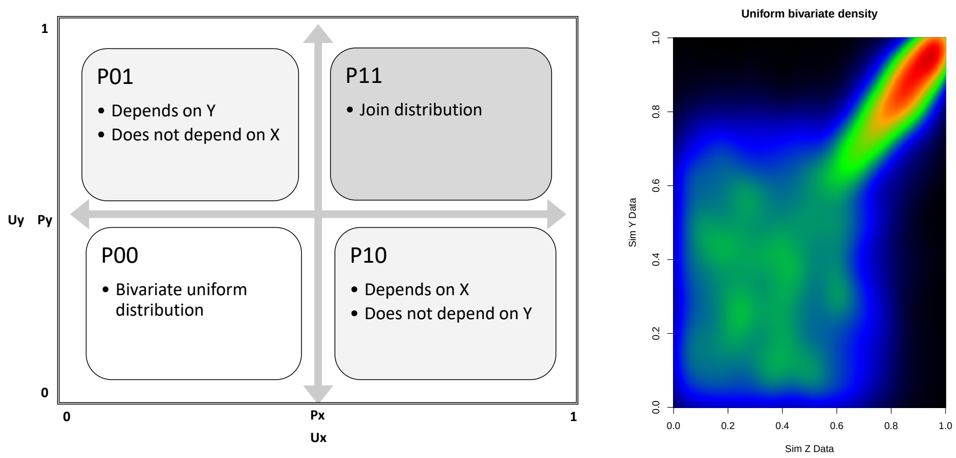

Figure 4. Bivariate copula density for two variables with ZAGA distributions. In the left, the quadrants are represented. In the right, the empirical bivariate copula density for real rainfall data

Note that the same partition can be applied directly to the density $c_{XY}(u, v)$, where:

$$c_{XY}(u, v) = P_{00} + P_{10}c_u\left(\frac{u - p_x}{1 - p_x}\right) + P_{01}c_v\left(\frac{v - p_y}{1 - p_y}\right) + \quad (18)$$

$$P_{11}c^{11}_{XY}\left(\frac{1 - p_x}{P_{11}}\left[\left(\frac{u - p_x}{1 - p_x}\right) - C_u(u)\frac{P_{10}}{1 - p_x}\right], \frac{1 - p_y}{P_{11}}\left[\left(\frac{v - p_y}{1 - p_y}\right) - C_v(v)\frac{P_{01}}{1 - p_y}\right]\right), \quad (19)$$

where $c^{11}_{XY}(u, v)$ is the corresponding density function of $C^{11}_{XY}(u, v)$, that is defined on the condition that both variables X and Y are greater than zero. C_u and C_v are the cumulative distribution functions of c_u and c_v respectively, where

$$C_u(u) = P\left(F_X(X) \le \frac{u-p_x}{1-p_x}\Big| X>0, Y=0\right) = P\left(p_x Z + (1-p_x)F_X^1\left[H_X^{-1}(Z)\right] \le \frac{u-p_x}{1-p_x}\right), \tag{20}$$

where $Z \sim Uniform(0,1)$. Analogously,

$$C_v(v) = P\left(F_Y(y) \le \frac{v-p_y}{1-p_y}\Big| X=0, Y>0\right) = P\left(p_y Z + (1-p_y)F_Y^1\left[H_Y^{-1}(Z)\right] \le \frac{v-p_y}{1-p_y}\right) . \tag{21}$$

2.3 Discrete copula CDF and inverse CDF

The at-site rainfall daily generator is based on the simulation from mixed discrete-continuous conditional distribution, which is deduced from a copula-based discrete-continuous conditional bivariate distribution. Moreover, this study is an extension of the research done by Serinaldi (2009b), Serinaldi (2008), Serinaldi (2009a), Herr & Krzysztofowicz (2005). The main contribution to these methodologies is that when we model the temporal and spatial (multi-site) conditional series dependencies, we model the whole multivariate time-series with copula relations (COPAR methodology). The resulting methodology models the dependence structure directly from the conditional dependence generated in the construction of the R-vine pair-copula structure following the specified structures in Brechmann et al. (2015). For an in depth description and mathematical background in the construction of bivariate copula-based discrete-continuous conditional distribution and the copula autoregressive methodology, the reader is referred to the previous authors mentioned.

The main building block for simulating time series based on a copula based autoregressive model is the conditional bivariate copula. For example, for a stationary univariate time series, let $C_{t,t-k}(v|v_k)$ be the bivariate copula of the variable v with its respective lag k, for instance, when k takes the value of 1, we define the copula of $v = F_X(X_t)$ with $v_1 = F_X(X_{t-1})$. To generate the value v conditioned on v_1, it is necessary to define the function:

$$h_{t,t-1} = \frac{\delta C_{t,t-1}(v, v_1)}{\delta v_1} , \tag{22}$$

and its respective inverse $h^{-1}_{t,t-1}$. Note that according to the time series decomposition presented in equation 11, when $k > 1$, $h_{t,t-k}$ correspond to the derivative of the conditional bivariate copula $C_{t,t-k}(v, v_k|v_{k+1}, \ldots, v_1)$. When the original variables have mass on zero, these derivatives have to be adapted in order to account for the partition in quadrants (Figure 4). This section illustrates the estimation of the conditional copula CDF ($h_{t,j}$) and its inverse ($h_{t,j}^{-1}$) for the mixed discrete-continuous copula, and then how to use this function in order to build the multivariate time series synthetic generation algorithm. The construction of the conditional CDF is presented in Algorithm 2, and conditional inverse CDF is explained in Algorithm 3, for $v = F_x(x_t|x_{j+1},\ldots,x_{t-1})$ and $v_j = F_x(x_j|x_{j+1},\ldots,x_{t-1})$.

if $v_j \leq p_{v_j}$ **then**
    if $v \leq p_v$ **then**
        $h_{t,j} = \frac{P_{00}}{P_{00}+P_{01}}\left(\frac{v}{p_v}\right)$
    else
        $h_{t,j} = \frac{P_{00}}{P_{00}+P_{01}} + \frac{P_{01}}{P_{00}+P_{01}}C_v\left(v\right)$
    end
else
    $\psi = \frac{P_{10}}{P_{10}+P_{11}}c_{v_j}\left(v_j\right)$
    if $v \leq p_v$ **then**
        $h_{t,j} = \left(\frac{v_j}{p_{v_j}}\right)\psi$
    else
        $\tilde{v} = \frac{1-p_v}{P_{11}}\left[\left(\frac{v-p_v}{1-p_v}\right) - C_v\left(v\right)\frac{P_{10}}{1-p_v}\right]$
        $\tilde{v}_j = \frac{1-p_{v_j}}{P_{11}}\left[\left(\frac{v_j-p_{v_j}}{1-p_{v_j}}\right) - C_{v_j}\left(v_j\right)\frac{P_{10}}{1-p_{v_j}}\right]$
        $h_{t,j} = \psi + (1-\psi)\frac{\delta C^{11}_{X_t,X_j}(\tilde{v},\tilde{v}_j)}{\delta \tilde{v}_j}$
    end
end

Algorithm 2: Continuous-Discrete bivariate copula-based CDF.

if $v_j \leq p_{v_j}$ **then**
  if $v \leq P_{00}/(P_{00}+P_{01})$ **then**
    $h_{t,j}^{-1} = \frac{P_{00}+P_{01}}{P_{00}} p_v v$
  else
    $h_{t,j}^{-1} = C_v^{-1}\left(v - \frac{\frac{P_{00}}{P_{00}+P_{01}}}{1-\frac{P_{00}}{P_{00}+P_{01}}}\right)$
  end
else
  $\xi = \frac{P_{10}}{P_{10}+P_{11}} c_{v_j}(v_j)$
  if $v \leq \xi$ **then**
    $h_{t,j}^{-1} = p_v v/\xi$
  else
    $\tilde{v} = G_v^{-1}(v)$
    $\tilde{v_j} = \frac{1-p_{v_j}}{P_{11}}\left[\left(\frac{v_j - p_{v_j}}{1-p_{v_j}}\right) - C_{v_j}(v_j)\frac{P_{10}}{1-p_{v_j}}\right]$
    $h_{t,j}^{-1} = \xi + (1-\xi)\frac{\delta C_{X_t,X_j}^{11}(\tilde{v},\tilde{v_j})}{\delta \tilde{v_j}}^{-1}$
  end
end

Algorithm 3: Continuous-Discrete bivariate copula-based inverse CDF.

Where $p_v = P_{00} + P_{01}$ and $p_{v_j} = P_{00} + P_{10}$, and:

$$G_v(a) = \frac{1-p_v}{P_{11}}\left[\left(\frac{a-p_v}{1-p_v}\right) - C_v(a)\frac{P_{10}}{1-p_v}\right]. \tag{23}$$

The discrete copula distribution function and the inverse function are conditionally estimated from the probabilities of each of the quadrants and the respective empirical distributions. That is, C_u and C_v can be estimated directly from the uniform variables. For the top right quadrant (where both variables are continuous), the estimation is done with parametric copulas using the correspondent conditional values. Thus, $\frac{\delta C^{11}(u,v)}{\delta u}$ is the copula function where both marginals take non-zero values. We limit the parametric families used for the P_{11} quadrant to be: 0 = independence copula, 1 = Gaussian copula, 2 = Student t copula (t-copula), 3 = Clayton copula, 4 = Gumbel copula, 5 = Frank copula and 6 = Joe copula. However, the analysis could be performed with any family, in particular copula families with long tails such as the Plackett copula.

2.3.1 Simulation algorithm

This section presents the simulation algorithm based on the copula autoregressive model. First, we introduce the algorithm for univariate time series from Smith et al. (2010), after that, we explain the adjustment that was assessed for two jointly observed time series. This procedure can be extended to any number of variables and eany number of lags.

Going back to the D-vine decomposition shown in section 2.2.1, the critical aspect is the evaluation of $u_{t|j+1}$ and $u_{j|t-1}$ from Equation (11). In this regard, it is worth mentioning the following property H Joe (2011): let $u_1 = F(X_1)$ and $u_2 = F(X_2)$ be conditional distribution functions, and $F(X_1, X_2) = C(u_1, u_2; \theta)$ where C is a bivariate copula function with parameters θ, then $F(X_1 \mid X_2) = h(u_1 \mid u_2; \theta)$ is defined as in equation 22. For $j \leq t$, application of this property gives the following recursive relationships:

$$u_{t|j} = F(x_t \mid x_{t-1}, \cdots, x_j) = h_{t,j}(U_{t|j+1} \mid U_{j|t-1}; \theta_{t,j}) \tag{24}$$

$$u_{j|t} = F(x_j \mid x_t, \cdots, x_{j+1}) = h_{t,j}(U_{j|t-1} \mid U_{t|j+1}; \theta_{t,j}) \ . \tag{25}$$

From Equations (24) and (25), all values of $u_{j|t}$ and $u_{t|j}$ can be obtained, since they correspond to a forward and backward recursion respectively. A more detailed explanation of this process is presented in algorithm 4.

for $t \leftarrow 1$ **to** T **do**
  Set $u_{t|t} = F(x_t)$
end
for $k \leftarrow 1$ **to** $T-1$ **do**
  for $i \leftarrow k+1$ **to** T **do**

$$\text{Backward recursion} : u_{i|i-k} = h_{i,i-k}(u_{i|i-k+1} \mid u_{i-k|i-1}; \theta_{i,i-k})$$

$$\text{Forward recursion} : u_{i-k|i} = h_{i,i-k}(u_{i-k|i+1} \mid u_{i|i-k+1}; \theta_{i,i-k})$$

  end
end

Algorithm 4: Recursion Algorithm for determining $u_{j|t}$ and $u_{t|j}$

The conditional distribution function of x_t given the previous values, can be obtained from Equation (24) as $F(x_t \mid x_{t-1}, \cdots, x_1) = u_{t|1} = h_{t,1}(u_{t|2} \mid u_{1|t-1}; \theta_{t,1})$ where $u_{t|2} = F(x_t \mid x_{t-1}, \cdots, x_2)$. Recursively, the conditional distribution function of X_t given the previous variables on the series can be expressed as:

$$F(x_t \mid x_{t-a}, \cdots, x_1) = h_{t,1} \ \ o \ \ h_{t,2} \ \ o \ \ \cdots \ \ h_{t,t-1} \ \ o \ \ F(x_t)\ . \tag{26}$$

For evaluating all $h_{t,j}$ functions, the values of $u_{1|t-1}, \cdots, u_{t-1|t-1}$ must be computed, hence the importance of Algorithm 4. Assuming that the series is Markovian of order p, the distribution function can be simplified as $F(x_t \mid x_{t-1}, \cdots, x_1) = F(x_t \mid x_{t-1}, \cdots, x_{t-q})$ where $q = \min(p, t-1)$. Therefore, uniformly distributed random numbers between 0 and 1 (w_t) are generated and a realization of the time series is computed as $x_t = F^{-1}(w_t \mid x_{t-1}, \cdots, x_{t-q})$. This methodology for univariate time series simulation is presented in algorithm 5.

for $t \leftarrow 1$ **to** T **do**
  Generate $\mathrm{w}_t \sim$ Uniform(0,1)
  if $t = 1$ **then**
    Set $\mathrm{x}_1 = F^{-1}(w_1)$
  else
    Set $\mathrm{x}_t = F^{-1} \ o \ h^{-1}_{t,t-1} \ o \ \cdots \ o \ h^{-1}_{t,t-q}(w_t \mid u_{i-q|i-1})$
  end
end

Algorithm 5: Simulation Algorithm for univariate time series

As the objective of this study is to simulate multi-site rainfall time series using the estimation of the COPAR model, algorithm 5 must be modified for simulating a multivariate time series. Although we present the methodology for a bivariate case, this procedure can be extended to any number of variables as stated before. First consider $\{X_t\}$ as the time series of a particular rainfall station and $F_X(\cdot)$ as its marginal distribution function. Also consider Y_t as the series of a different rainfall station for $t = 1, \cdots, T$ and $F_V(\cdot)$ as its marginal distribution function. We re-order the elements of both time series into a univariate series $W = (W_1, \cdots, W_N)$ with $N = 2T$ in which we interpolate the values of both series i.e. $W = (X_1, Y_1, \cdots, X_T, Y_T)$. Recalling the model from section 2.2.2, we fit a COPAR (k) model to the data in which the serial dependence of $\{X_t\}$ is modeled unconditionally and that of Y_t is modeled conditionally on $\{X_t\}$. Then, we store the copulas in two categories. The first one ($C^{(1)}_{t,j}$) corresponds to the copulas from Equations (12) and (14), i.e. the ones for modeling $\{X_t\}$ given previous values of $\{X_t\}$ and $\{Y_t\}$, whereas the second category ($C^{(2)}_{t,j}$) corresponds to the copulas from Equations (13) and (15), i.e. the ones for modeling $\{Y_t\}$ given previous values of $\{X_t\}$ and $\{Y_t\}$. The simulation algorithm is modified in the following manner:

for $t \leftarrow 1$ **to** $2T$ **do**

$$\text{Generate } \ \omega_t \sim \text{Uniform}(0,1)$$

$$\text{Set } \ \text{q} = \min(((k+1)*2)-1, t-1)$$

$$\text{Set } \ \text{m} = \text{t} \mod 2$$

  if $t = 1$ **then**

    Set $\text{x}_1 = F_V^{-1}(\omega_1)$

  else

    if $m = 1$ **then**

      Set $\text{x}_t = F_V^{-1} \ o \ h_{t,t-1}^{-1(1)} \ o \ \cdots \ o \ h_{t,t-q}^{-1(1)}(\omega_t \mid u_{i-q|i-1})$

    else

      Set $\text{x}_t = F_\theta^{-1} \ o \ h_{t,t-1}^{-1(2)} \ o \ \cdots \ o \ h_{t,t-q}^{-1(2)}(\omega_t \mid u_{i-q|i-1})$

    end

  end

end

Algorithm 6: Simulation Algorithm for bivariate time series

3 Implementation and Analysis of Results

In this section, we present the case study as well as the results and the comparisons between the observed and simulated rainfall time series. We analyze rainfall time series from ground measured data in three stations in the province of Trentino, Italy, with locations shown in 5. The first site is located at Pergine Valsugana (Convento) **(Variable Z)** in Latitude (46.01°N) and Longitude (11.23°E) with an elevation of 475m. The second site is located at Cles (Convento) **(Variable Y)** in Latitude (46.35°N) and Longitude (11.02°E) with an elevation of 665m. Finally, the last site is located at Trento (Laste) **(Variable X)** in Latitude (46.07°N) and Longitude (11.01°E) with an elevation of 312m. The province of Trentino is a region with multiple variable meteorological characteristics. These ground stations measured data that consist of daily time series from 1961 to 1990. The data set trentino is available in the R software package RMAWGEN (Cordano & Eccel, 2017). These stations were chosen in particular because of the heterogeneity of the statistics of the records and also due to their high spatial correlation. The resolution of the data is 0.1mm, so that any observation less than this threshold is taken to be a dry day. The data was preprocessed and cleaned for outliers and missing observations. The methodologies or algorithms proposed in this paper do not depend on the selected data; this data was selected with the purpose of showcasing the characteristics of the observed data compared to the simulated data.

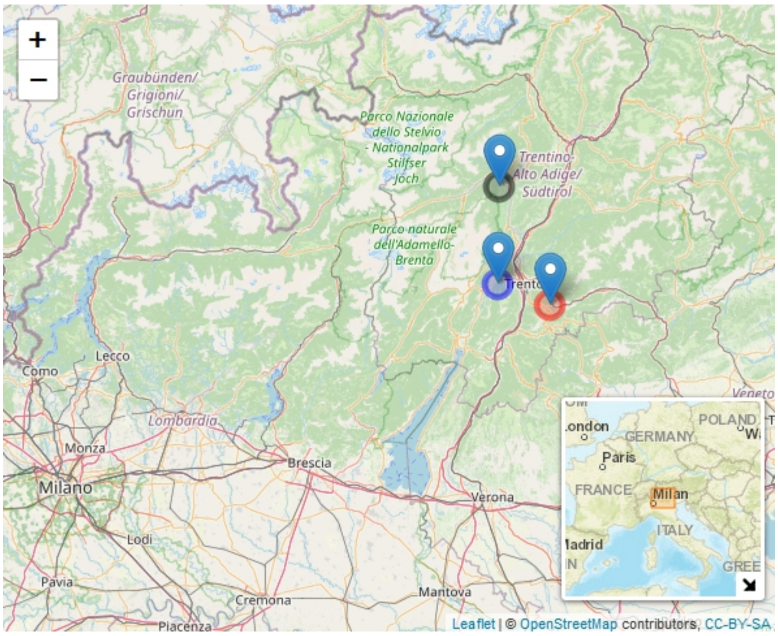

Figure 5. Case study stations in Trento, Italy

The first part of the results consists of the analysis of the marginal distribution estimations. Figure 6 shows the results for the parameter estimation for each of the parameters needed in the ZAGA Distribution using the VGLM methodology. Parameters μ and σ are related to the gamma distribution (shape and scale respectively), and parameter ν is related to the probability of the Bernoulli distribution that models the occurrence portion of the process. These results show the seasonality of the process throughout the months of the year; moreover, we can see that this site in particular presents a lower probability of rainfall at midyear, and a higher amount of rainfall in the rest of the months of the year. This phenomena is also present in the observed and simulated data. Figure 7 present the density plots for the non-zero values of the rainfall data for the observed and simulated time series. We can see that the simulation data correctly models the density function of the monthly non-zero rainfall time series. Although the max values per month are sometimes underestimated due to the restrictions imposed by the gamma distribution used to model the non-zero characteristics, we can see that the methodology correctly models the reduction of rainfall amount at midyear and the increase of rainfall amount for the other months. Further improvement in this matter can be assessed and a further discussion can be found in Section 4.

A further analysis can be conducted when analyzing the mean, standard deviation and density of the data alongside its characteristics. Table 2 presents the sample mean and standard deviation of the rainfall time series for the Pergine Valsugana site (variable Z) for both observed (Obs) and simulated (Sim) values discriminated for every month for all the years of the study. We also calculated the column $m(mean)$ and $m(sd)$ which are the relative error of the simulation compared to the observed values for the mean and standard deviation on each of the months (Sarmiento et al., 2018). The discrepancies of the mean and standard deviation (SD) could be caused by the difference of the maximum values modeled by the zero adjusted gamma distribution as explained before; however, for most of the months the mean and standard deviation of rainfall is captured adequately. The respective results for the Cles and Trento sites (Variables Y and X) can be found in Section 5. An in depth analysis of the monthly max values for the observed and simulated data is presented in Figure 8. This figure shows the empirical distributions of the monthly maximum values for the observed and simulated rainfall for the variable Z. When looking at the monthly maximum values, the simulation method has a tendency to produce some sparse values. Although more research needs to be done on this issue, we hypothesize that this is an effect of the type of continuous gamma distribution used to model the non-zero values of the distribution, since the actual process of the data has longer tails. The respective results for the Cles and Trento sites (Variables Y and X) can be found in Section 5; moreover, as stated before, an in depth discussion of this matter is conducted in Section 4. Figure 9 include a time-series plot, a box-plot and an histogram of the rainfall for the daily, monthly and yearly observed data as well as the simulated data for the Pergine Valsugana site using the methodology proposed in this paper. These results show that the distribution of the variables could be reproduced using this methodology and also the methodology could model the non-stationarity characteristics of the series. The respective results of Variables Y and X for this analysis can be found in Section 5.

Following these results, another key factor for comparing the simulated and observed data is observing the autocorrelation and partial autocorrelation functions as well as the correlation between sites. Figure 10 show the autocorrelation and partial autocorrelation function for the sites of the study. This figure shows that the two lags modeled are simulated correctly; moreover, the correlation between sites is presented in Table 1. Finally, Figure 11 show the 2D density plots of the bivariate distribution for variables Z and Y. The correspondent 2D density plots for the other bivariate distributions (Z - X and Y - X) can be found in Section 5.

	Pergine Valsugana **(Variable Z)**	Cles **(Variable Y)**	Trento **(Variable X)**
(Variable Z)	1.0000	0.8270	0.7605
(Variable Y)	0.8270	1.0000	0.8535
(Variable X)	0.7605	0.8535	1.0000
(Variable Z)	1.0000	0.7732	0.7143
(Variable Y)	0.7732	1.0000	0.8522
(Variable X)	0.7143	0.8522	1.0000

Table 1. Observed (TOP) and simulated (BOTTOM) data linear correlation for the 3 sites

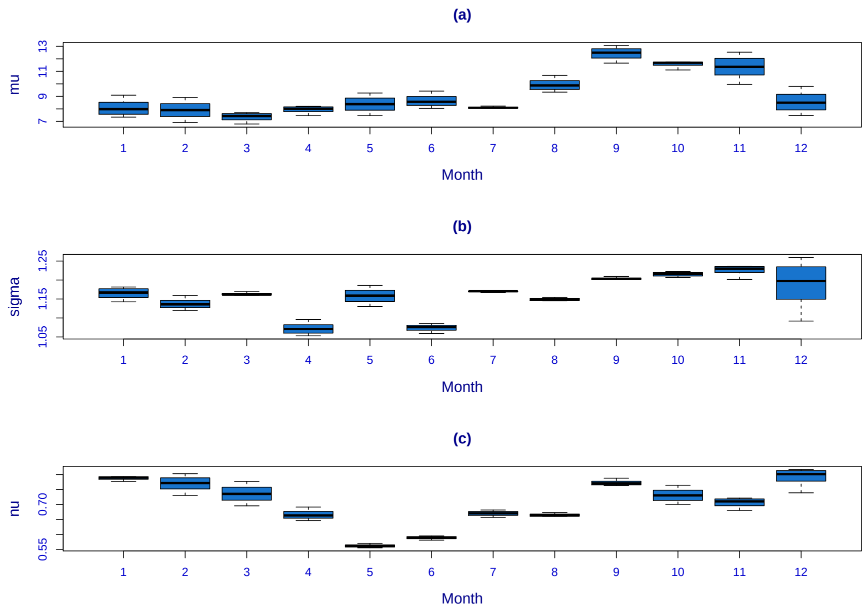

Figure 6. VGAM parameter estimation of ZAGA distribution for Pergine Valsugana

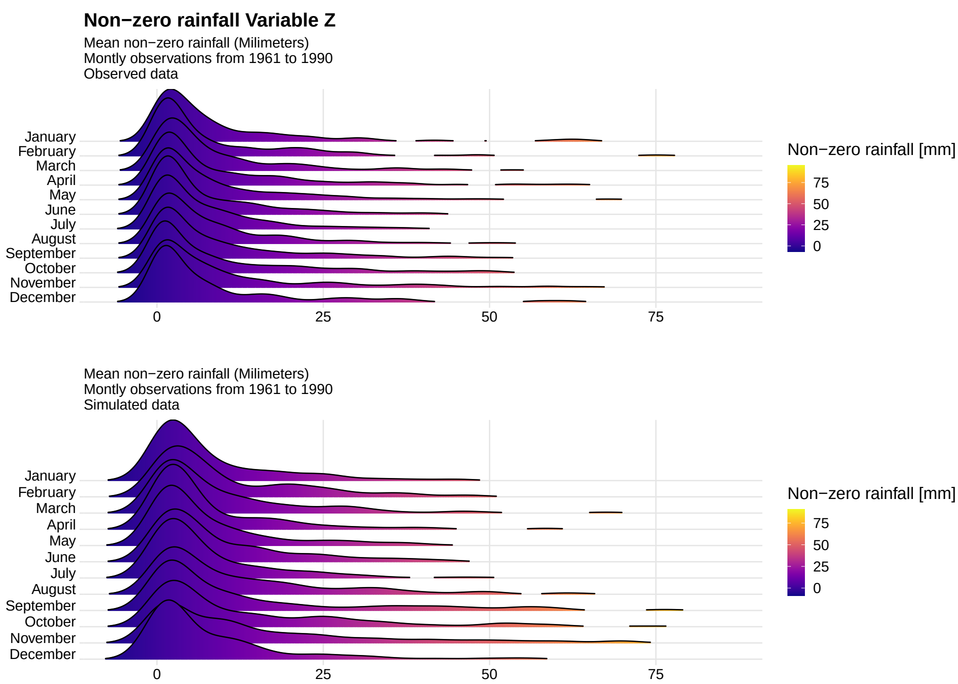

Figure 7. Observed (TOP) and Simulated (BOTTOM) non-zero rainfall mean density plot Pergine Valsugana

	Month	Obs mean	Obs sd	Sim Mean	Sim sd	m(mean)	m(sd)
1	Jan	1.82	6.68	1.46	4.55	-0.20	-0.32
2	Feb	2.09	8.70	1.99	7.43	-0.05	-0.15
3	Mar	2.40	7.60	2.43	6.92	0.01	-0.09
4	Apr	3.21	8.95	2.05	5.62	-0.36	-0.37
5	May	3.77	9.16	2.95	7.73	-0.22	-0.16
6	Jun	3.23	6.99	2.53	6.68	-0.22	-0.04
7	Jul	2.90	7.11	2.66	7.37	-0.08	0.04
8	Aug	3.16	7.97	3.67	8.74	0.16	0.10
9	Sep	2.79	8.96	2.42	7.90	-0.13	-0.12
10	Oct	3.03	9.85	2.94	10.71	-0.03	0.09
11	Nov	3.36	10.03	3.18	9.60	-0.05	-0.04
12	Dec	1.75	6.75	0.98	5.02	-0.44	-0.26

Table 2. Montly mean and sd comparisson Pergine Valsugana (Convento)

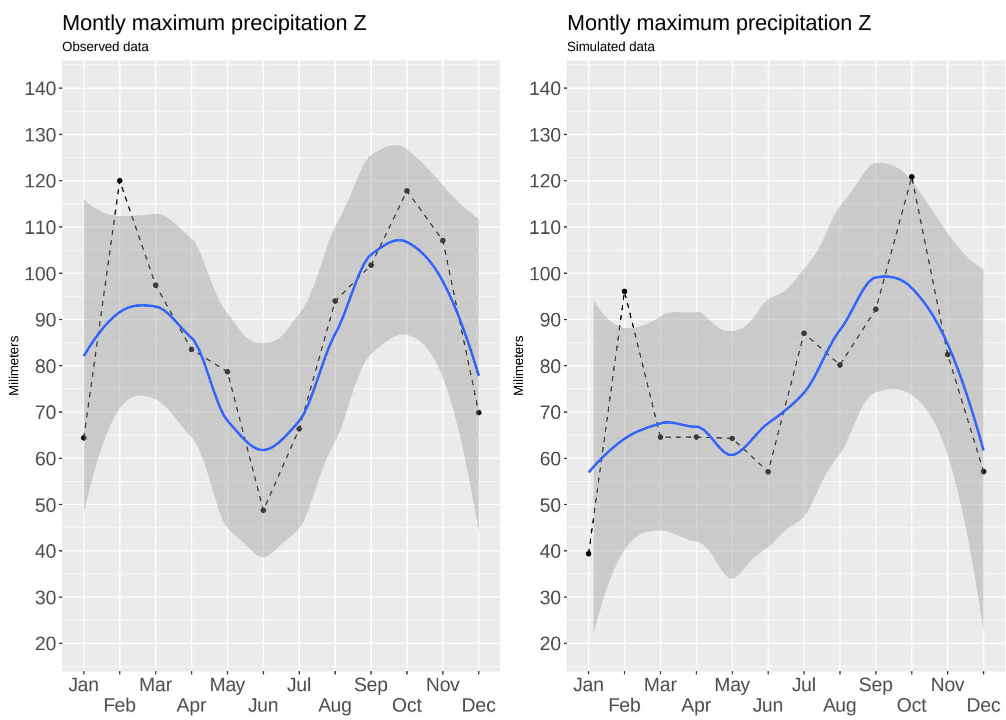

Figure 8. Monthly maximum rain fall plot Pergine Valsugana

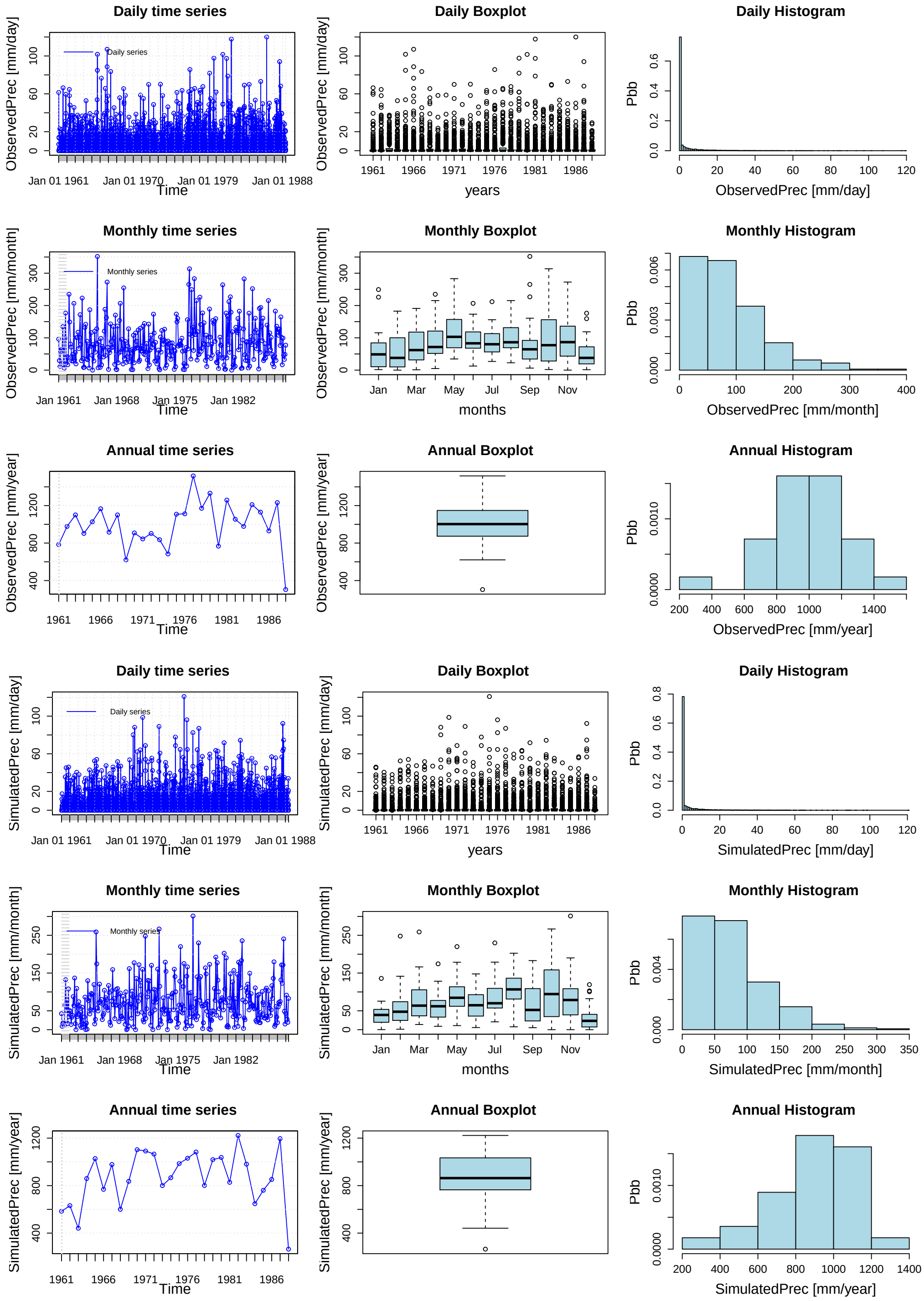

Figure 9. Observed (TOP) and simulated (BOTTOM) monthly plots Pergine Valsugana (Convento)

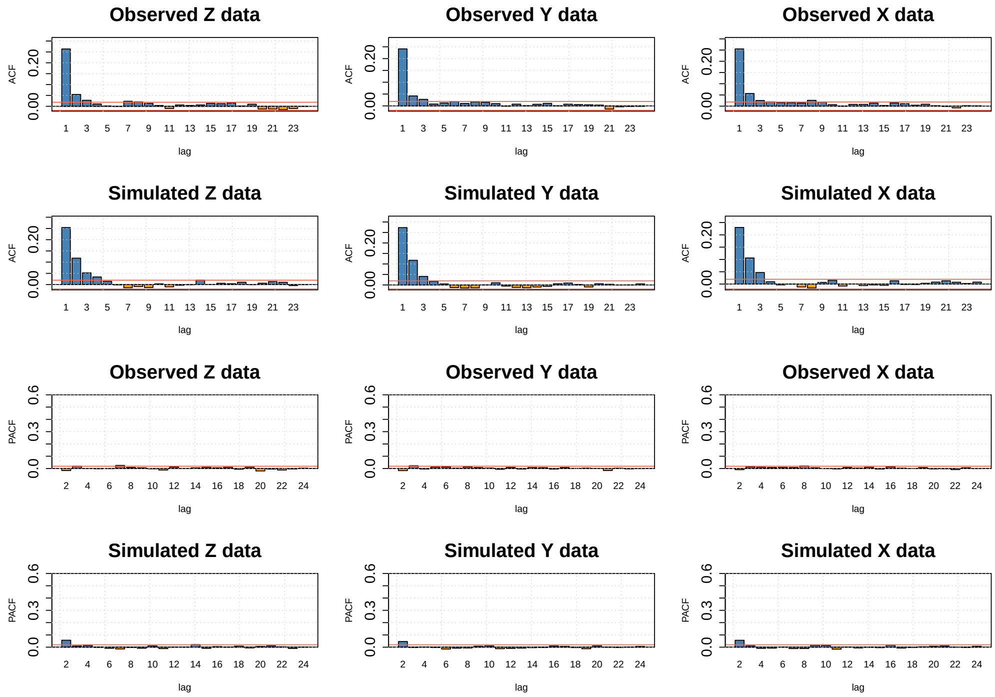

Figure 10. Simple (ACF) and Partial autocorrelation function (PACF) plots for observed and simulated data.

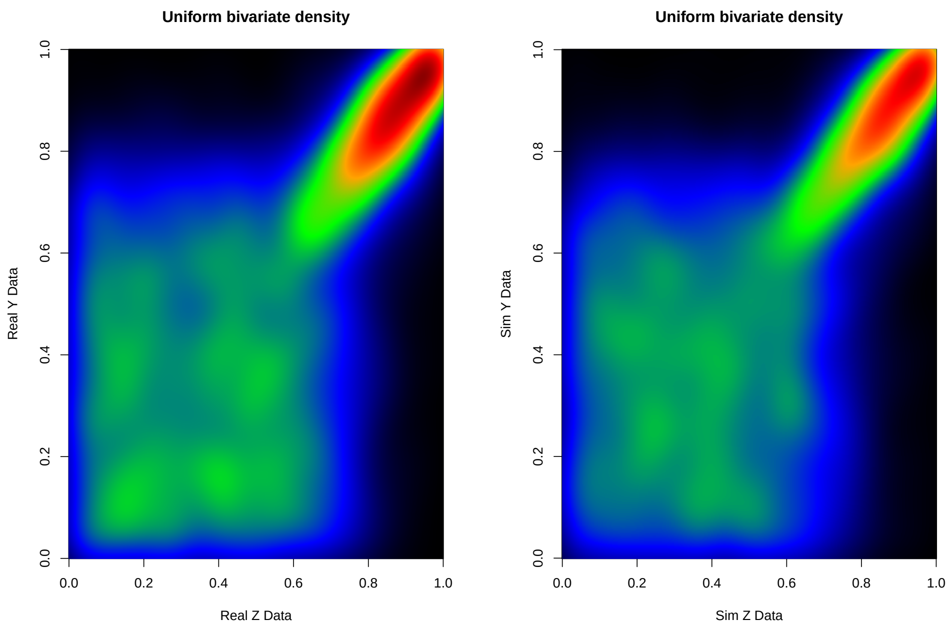

Figure 11. Uniform bivariate density observed (LEFT) and simulated (RIGHT) data variables Z - Y

4 Discussion and Conclusions

We proposed a rainfall time series simulation strategy using the copula autoregressive methodology (COPAR) proposed by Brechmann et al. (2015) implemented into the discrete-continuous conditional bivariate estimation problem. The results presented in this paper are promising and might help to better understand the rainfall simulation phenomena. Overall, the stochastic properties of the series are replicated in terms of the spatial and temporal dependence, as well as the marginal distributions. Using copula functions to model the pairwise dependence of all variables provides a flexible framework to generate more realistic series. This is true in particular because: (i) the model can be estimated in one stage, without separating the occurrence and the amount processes; (ii) copula functions may exploit non-linear relations that are flexible and solve one of the fundamental problems when using transformations of linear (gaussian) time series; and (iii) the whole multivariate time series can be integrally modeled using the same copula based model, no matter the number of dimensions (locations) and the number of lags in the autoregressive process, meaning that no post-estimation introduction of spatial dependency is necessary.

It is important to notice that in our research, we found that the gamma distribution underestimates the maximum values of the simulated rainfall series. This could be improved by using a different zero adjusted distribution that better fits the non-zero long tailed characteristics of the data. Future research in this matter would be to explore the properties of the proposed model with hybrid exponential and generalized Pareto distributions, as proposed in Li et al. (2013). In addition, the exploration of copula functions could be more flexible to account for spatial and temporal tail dependence, such as meta-elliptical copulas (Genest et al., 2007) or v-transformed copulas (Bárdossy & Pegram, 2009).

A main difference from previous approaches used to model rainfall time series is that the methodology we propose allows us to estimate the inner and cross dependencies between the different sites directly from the conditional dependencies of the rainfall observations. This estimation is assessed by one model that accounts for all the dependence structures without the need of partitioning the estimation procedure into several models. Although this is an advantage in terms of the reduction of overfitting, the estimation of the pair copula construction can be computationally expensive.

The construction of all the pair copula functions needed for this study as well as the simulation procedure had a total computational time of around 30 minutes using a 2.2 Ghz

Intel core i7-4702MQ 7 Gen processor with 16Gb of RAM. An automated algorithm for generating the complete pair-copula R-vine structure for each month and the simulation structure required was also generated. When modeling the discrete-continuous conditional bivariate distribution, we use the empirical estimation of the distribution function for the joint marginal P_{01} and P_{10} distributions. A further improvement in the method could be gained by upgrading this approach, for instance, by using right skewed distributions with longer tails. However, it is important to state that the results found using the empirical approach were satisfactory.

Finally, we performed different experiments using non parametric Bernstein copula functions, but the computational time was severely compromised; moreover, the grid used for the discretization of the unitary square may have resulted in some sparse large values that could imply over estimations. Nevertheless, after using parametric copula functions or combinations of these, we mitigated the issue and improved the computational time significantly. For future work, this methodology could be applied to a random fields model where rainfall series could be interpolated in a certain space correctly maintaining the stochastic properties and nature of the data.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or non-for-profit sectors. However, the authors express their gratitude to the colleagues in the Industrial Engineering Department at University of los Andes for their valuable comments. The dataset used to validated the methodology proposed in the paper is available in the R software package RMAWGEN available in the CRAN repository (Cordano & Eccel, 2017). The results provided in this study can be replicated with this data as explained in the text.

Declaration of interests: None.

References

Ahn, K.-H. (2020). Coupled annual and daily multivariate and multisite stochastic weather generator to preserve low- and high-frequency variability to assess climate vulnerability. *Journal of Hydrology*, *581*, 124443. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022169419311783` doi: https://doi.org/10.1016/j.jhydrol.2019.124443

Ayar, P. V., Blanchet, J., Paquet, E., & Penot, D. (2020). Space-time simulation of precipitation based on weather pattern sub-sampling and meta-gaussian model. *Journal of Hydrology*, *581*, 124451. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022169419311862` doi: https://doi.org/10.1016/j.jhydrol.2019.124451

Balistrocchi, M., & Bacchi, B. (2011). Modelling the statistical dependence of rainfall event variables through copula functions. *Hydrology & Earth System Sciences*, *15*(6).

Bárdossy, A., et al. (2017). *Asymmetric dependence based spatial copula models: empirical investigations and consequences on precipitation fields.* Stuttgart: Eigenverlag des Instituts für Wasser-und Umweltsystemmodellierung . . . .

Bárdossy, A., & Pegram, G. (2009). Copula based multisite model for daily precipitation simulation. *Hydrology & Earth System Sciences*, *13*(12).

Bedford, C. R. M., Tim. (2001). Probability density decomposition for conditionally dependent random variables modeled by vines. *Journal of Econometrics*, *32*(1), 245–268. Retrieved from `http://www.rogermcooke.net/rogermcooke_files/aivines.pdf`

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

Buishand, T. A., & Brandsma, T. (2001). Multisite simulation of daily precipitation and temperature in the rhine basin by nearest-neighbor resampling. *Water Resources Research*, *37*(11), 2761–2776.

Chen, X., & Fan, Y. (2006). Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, *130*(2), 307-335. Retrieved from `https://doi.org/10.1016/j.jeconom.2005.03.004`

Cordano, E., & Eccel, E. (2017). Rmawgen: Multi-site auto-regressive weather generator [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=RMAWGEN` (R package version 1.3.7)

Faugeras, O. P. (2012, November). *Probabilistic constructions of discrete copulas.* Retrieved from https://hal.archives-ouvertes.fr/hal-00751393 (working paper or preprint)

Genest, C., Favre, A.-C., Béliveau, J., & Jacques, C. (2007). Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resources Research*, *43*(9).

Haan, C., Allen, D., & Street, J. (1976). A markov chain model of daily rainfall. *Water Resources Research*, *12*(3), 443–449.

Herr, H. D., & Krzysztofowicz, R. (2005). Generic probability distribution of rainfall in space: The bivariate model. *Journal of Hydrology*, *306*(1-4), 234–263.

H Joe, D. K. (2011). *Dependence modeling: Vine copula handbook.* World Scientific.

Kao, S.-C., & Govindaraju, R. S. (2008). Trivariate statistical analysis of extreme rainfall events via the plackett family of copulas. *Water Resources Research*, *44*(2).

Li, C., Singh, V. P., & Mishra, A. K. (2013). A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation. *Water Resources Research*, *49*(2), 767–789.

Mehrotra, R., Srikanthan, R., & Sharma, A. (2006). A comparison of three stochastic multi-site precipitation occurrence generators. *Journal of Hydrology*, *331*(1-2), 280–292.

Mhanna, M., & Bauwens, W. (2012). Stochastic single-site generation of daily and monthly rainfall in the middle east. *Meteorological Applications*, *19*(1), 111–117.

Nelsen, R. (1999). *Introduction. in: An introduction to copulas. lecture notes in statistics.* New York, NY: Springer. Retrieved from https://doi.org/10.1007/978-1-4757-3076-0_1

Richardson, C. W. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water resources research*, *17*(1), 182–190.

Sarmiento, C., Valencia, C., & Akhavan-Tabatabaei, R. (2018). Copula autoregressive methodology for the simulation of wind speed and direction time series. *Journal of Wind Engineering and Industrial Aerodynamics*, *174*, 188–199.

Serinaldi, F. (2008). Analysis of inter-gauge dependence by kendall's τ k, upper tail dependence coefficient, and 2-copulas with application to rainfall fields. *Stochastic Environmental Research and Risk Assessment*, *22*(6), 671–688.

Serinaldi, F. (2009a). Copula-based mixed models for bivariate rainfall data: an empirical study in regression perspective. *Stochastic environmental research and risk assessment*, *23*(5), 677–693.

Serinaldi, F. (2009b). A multisite daily rainfall generator driven by bivariate copula-based mixed distributions. *Journal of Geophysical Research: Atmospheres*, *114* (D10).

Shimizu, K. (1993). A bivariate mixed lognormal distribution with an analysis of rainfall data. *Journal of Applied Meteorology*, *32* (2), 161–171.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut Statistique de l'Université de Paris*, *8*.

Smith, M., Min, A., Almeida, C., & Czado, C. (2010). Modeling longitudinal data using a pair-copula decomposition of serial dependence. *Journal of the American Statistical Association*, 1467-1479. Retrieved from `https://doi.org/10.1198/jasa.2010.tm09572`

Srikanthan, R., & McMahon, T. (2001). Stochastic generation of annual, monthly and daily climate data: A review. *Hydrol. Earth Syst. Sci*, *5*, 653–670.

Tinaikar, A. (2013, 01). Harvesting energy from rainfall. *International Journal of Renewable and Sustainable Energy*, *2*, 130. doi: 10.11648/j.ijrse.20130203.18

Vaittinada Ayar, P., Blanchet, J., Paquet, E., & Penot, D. (2020). Space-time simulation of precipitation based on weather pattern sub-sampling and meta-gaussian model. *JHyd*, *581*, 124451.

Vandenberghe, S., Verhoest, N., & De Baets, B. (2010). Fitting bivariate copulas to the dependence structure between storm characteristics: A detailed analysis based on 105 year 10 min rainfall. *Water resources research*, *46* (1).

Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *journal of Hydrology*, *210* (1-4), 178–191.

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

Yee, T. W. (2015). *Vector generalized linear and additive models*. Springer. Retrieved from `https://doi.org/10.1007/978-1-4939-2818-7`

Zhang, L., & Singh, V. P. (2007). Bivariate rainfall frequency distributions using archimedean copulas. *Journal of Hydrology*, *332* (1), 93 - 109. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022169406003386` doi: https://doi.org/10.1016/j.jhydrol.2006.06.033

Zhang, L., & Singh, V. P. (2019). Rainfall frequency analysis. In *Copulas and their applications in water resources engineering* (p. 367–395). Cambridge University Press. doi: 10.1017/9781108565103.011

Zhou, L., Meng, Y., & Abbaspour, K. C. (2019). A new framework for multi-site stochastic rainfall generator based on empirical orthogonal function analysis and hilbert-huang transform. *Journal of Hydrology*, *575*, 730–742.

Zhou, L., Meng, Y., Lu, C., Yin, S., & Ren, D. (2020). A frequency-domain nonstationary multi-site rainfall generator for use in hydrological impact assessment. *Journal of Hydrology*, *585*, 124770. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022169420302304` doi: https://doi.org/10.1016/j.jhydrol.2020.124770

5 Supporting information

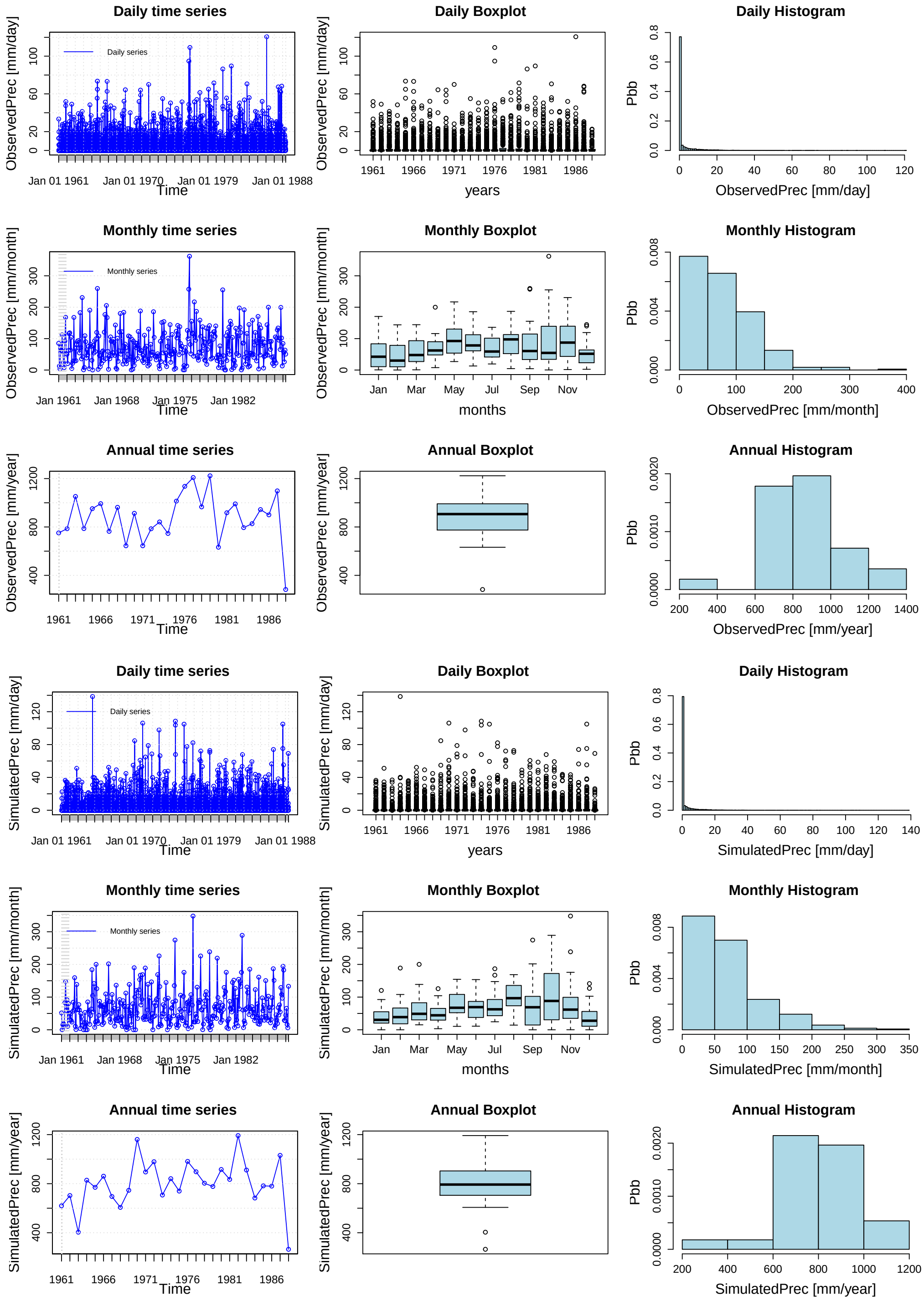

Figure 12. Observed and Simulated monthly plots Y variable

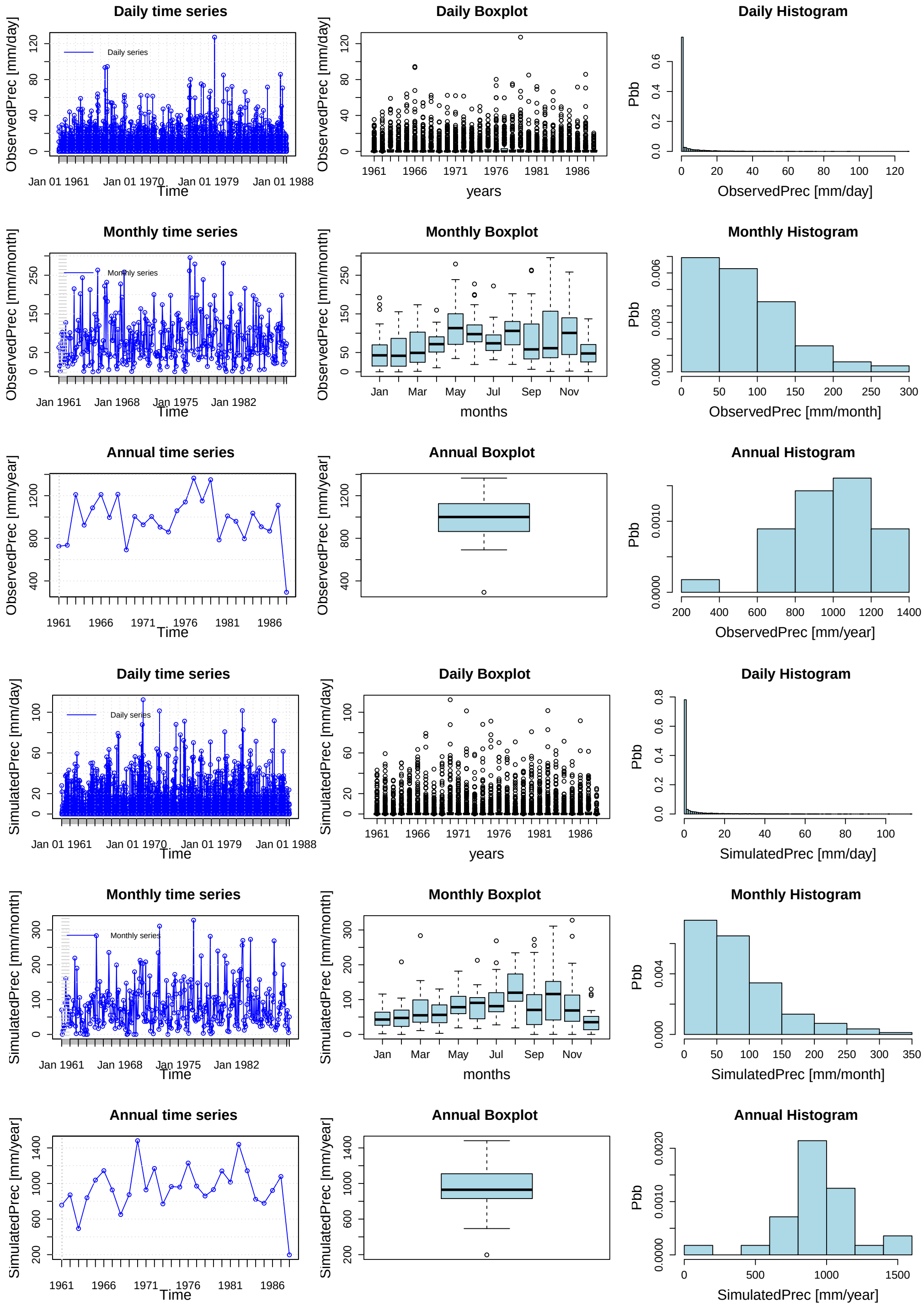

Figure 13. Observed and Simulated monthly plots X variable

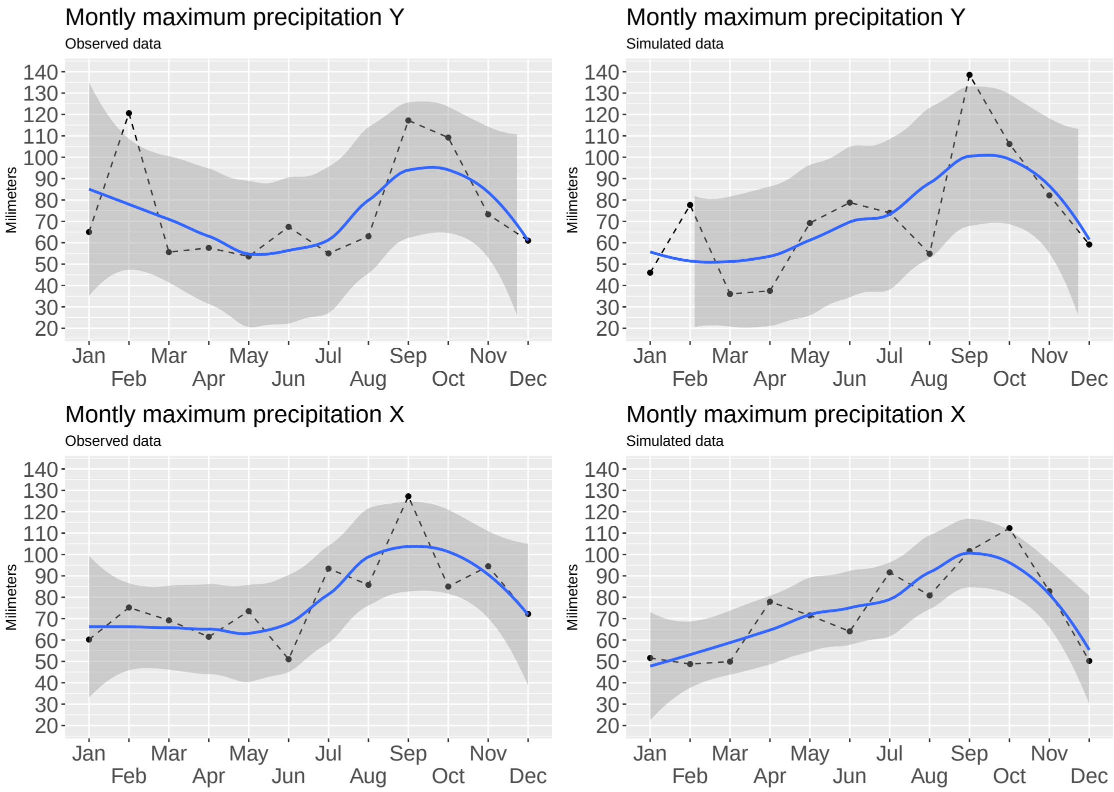

Figure 14. Monthly maximum rain fall plot Cles and Trento sites

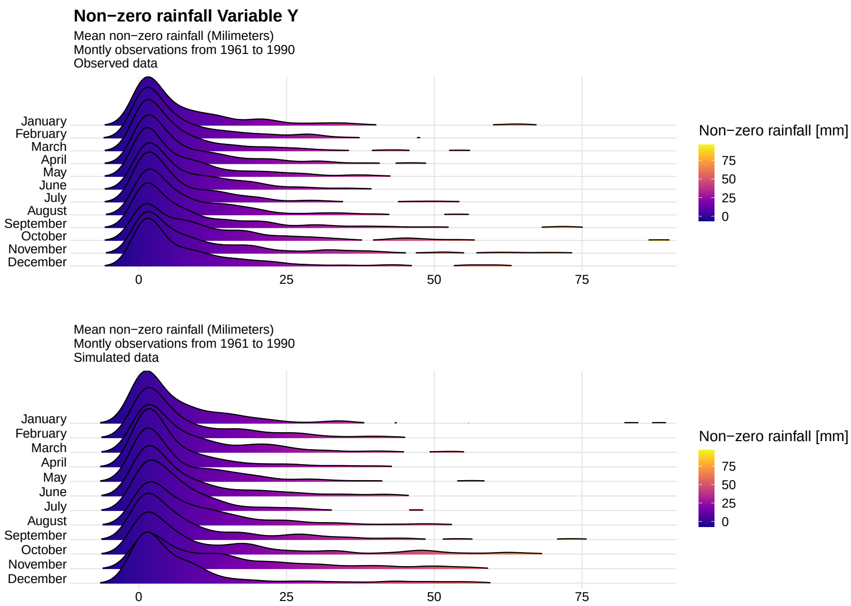

Figure 15. Observed and simulated non-zero rainfall mean density plot Cles

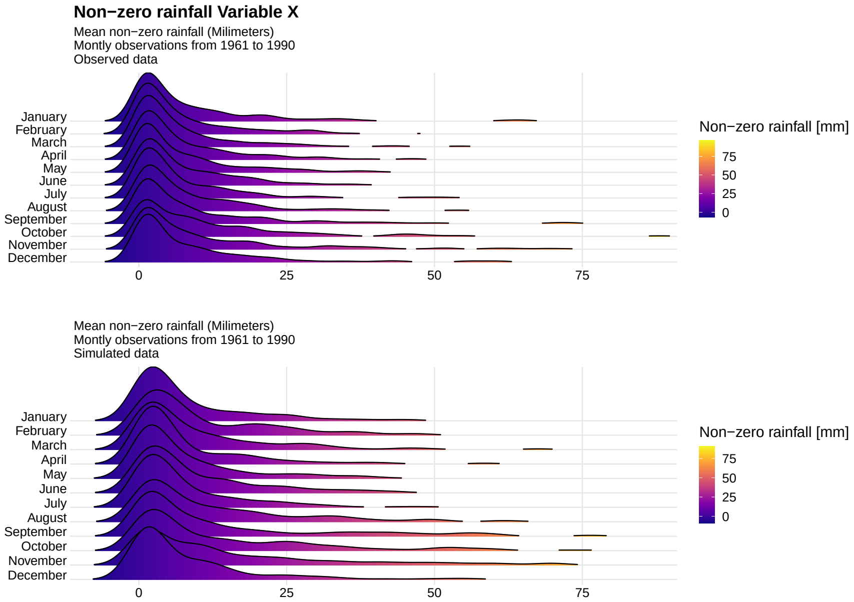

Figure 16. Observed and simulated non-zero rainfall mean density plot Trento

	Month	Obs mean	Obs sd	Sim Mean	Sim sd	m(mean)	m(sd)
1	Jan	1.66	5.97	1.42	4.72	-0.15	-0.21
2	Feb	1.62	6.35	1.66	6.38	0.03	0.01
3	Mar	1.85	5.65	1.79	4.98	-0.03	-0.12
4	Apr	2.54	6.50	1.69	4.69	-0.34	-0.28
5	May	2.97	6.91	2.42	6.81	-0.19	-0.01
6	Jun	3.04	6.91	2.33	6.92	-0.23	0.00
7	Jul	2.43	6.31	2.37	6.17	-0.02	-0.02
8	Aug	2.80	6.95	3.14	7.09	0.12	0.02
9	Sep	2.55	9.10	2.67	10.24	0.05	0.13
10	Oct	3.02	9.33	3.16	11.46	0.05	0.23
11	Nov	3.14	9.14	3.22	9.92	0.02	0.09
12	Dec	1.70	6.09	1.09	5.07	-0.36	-0.17

Table 3. Montly mean and sd comparisson Y variable

	Month	Obs mean	Obs sd	Sim Mean	Sim sd	m(mean)	m(sd)
1	Jan	1.71	5.61	1.68	4.97	-0.02	-0.11
2	Feb	1.81	5.94	1.75	5.93	-0.03	-0.00
3	Mar	1.96	5.99	2.24	6.47	0.14	0.08
4	Apr	2.66	6.53	2.05	5.95	-0.23	-0.09
5	May	3.65	8.03	2.76	7.23	-0.24	-0.10
6	Jun	3.55	7.02	2.85	7.58	-0.20	0.08
7	Jul	2.67	7.03	3.15	8.48	0.18	0.21
8	Aug	3.35	8.47	4.38	10.03	0.31	0.18
9	Sep	2.82	9.39	2.90	10.10	0.03	0.08
10	Oct	3.12	9.09	3.35	11.33	0.07	0.25
11	Nov	3.35	9.48	3.53	10.42	0.05	0.10
12	Dec	1.74	5.72	1.07	4.45	-0.39	-0.22

Table 4. Montly mean and sd comparisson X variable

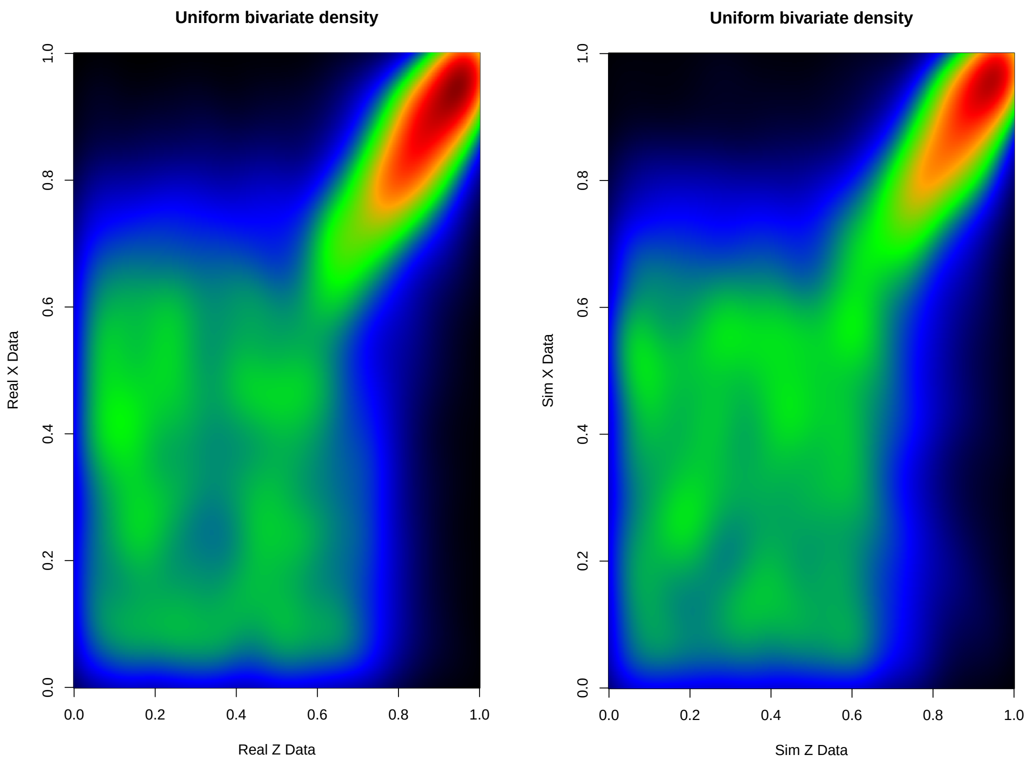

Figure 17. Uniform bivariate density observed (LEFT) simulated (RIGHT) data variables Y - X

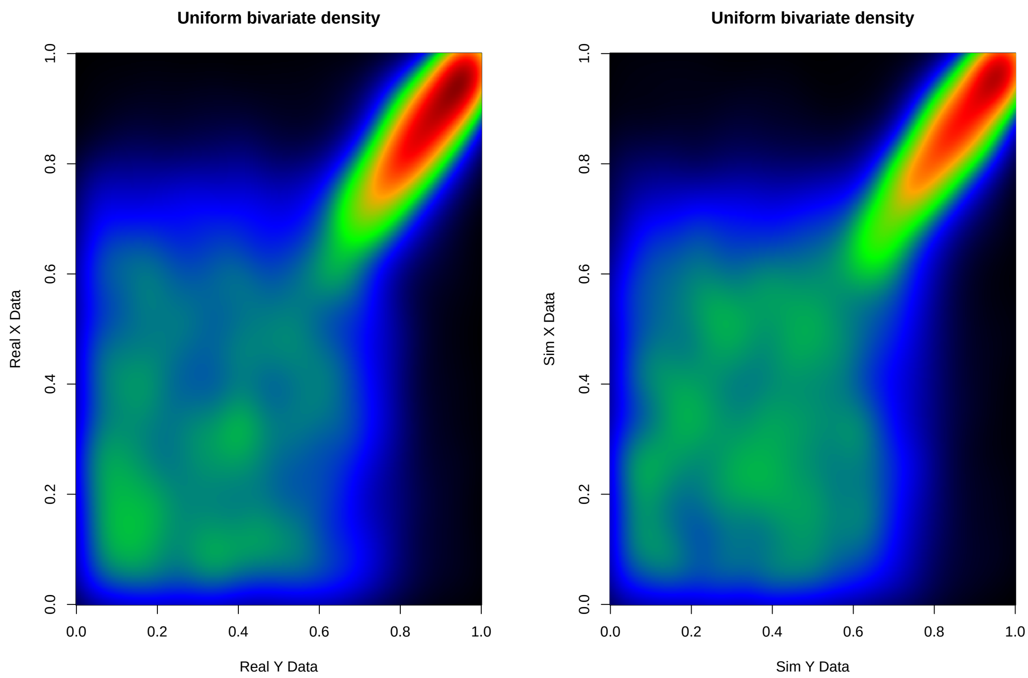

Figure 18. Uniform bivariate density observed (LEFT) and simulated (RIGHT) data variables Y - X