Banco de México

Documentos de Investigación

Banco de México

Working Papers

N° 2013-09

Semiparametric Copula-Based Stochastic Weather Generator

Miriam Juárez-Torres
Banco de México

James W. Richardson
Texas A&M University

Dmitry Vedenov
Texas *A*&M University

July 2013

Documento de Investigación
2013-09

Working Paper
2013-09

Semiparametric Copula-Based Stochastic Weather Generator*

Miriam Juárez-Torres†
Banco de México

James W. Richardson‡
Texas A&M University

Dmitry Vedenov§
Texas A&M University

Abstract: Stochastic Weather Generators (SWGs) try to replicate the stochastic patterns of climatological variables characterized by high dimensionality, non-normal probability density functions and non-linear dependence relationships. However, conventional SWGs usually typify weather variables with not always justified probability distributions assuming linear dependence between variables. This research proposes an alternative SWG that introduces the advantages of the copula modeling into the replication of stochastic weather patterns. The semiparametric copula-based SWG introduces more flexibility allowing researcher to model non-linear dependence structures independently of the marginals involved. Also, it can better model tail dependence, which would result in a more accurate reproduction of extreme weather events.
Keywords: Weather generator, Archimedian copulas, Brownian Bridge, Extreme weather events replication.
JEL Classification: Y4.

Resumen: Los Generadores Estocásticos de Clima (GSC) tratan de replicar los patrones estocásticos de las variables climatológicas, las cuales se caracterizan por alta dimensionalidad, funciones de densidad de probabilidad no normales y patrones de dependencia no lineales. Sin embargo, usualmente los GECs representan a las variables climatológicas con distribuciones de probabilidad difícilmente justificables y asumen dependencia lineal entre las variables. Esta investigación propone un GEC alternativo que introduce las ventajas de las copulas en la reproducción de los patrones estocásticos de clima. El GEC semiparamétrico basado en copulas introduce mayor flexibilidad, lo cual permite modelar las estructuras de dependencia no lineal independientemente de las distribuciones marginales involucradas. Asimismo, el GEC propuesto captura la dependencia en las colas de la distribución, lo cual resulta en una reproducción más exacta de los eventos climáticos extremos.
Palabras Clave: Generador de clima, Copulas Arquimedianas, Puente browniano, Reproducción de eventos climáticos extremos.

*The authors want to thank to Michael Longnecker for his comments. Miriam Juarez-Torres acknowledges the support of CONACYT and COMEXUS Becas Fulbright-Garcia Robles for funding her doctoral studies.
† Dirección General de Investigación Económica. Email: mjuarez@banxico.org.mx.
‡ Department of Agricultural Economics, Texas A&M University. Email: jwrichardson@tamu.edu.
§ Department of Agricultural Economics, Texas A&M University. Email: vedenov@tamu.edu.

1. Introduction

Climatological variables are complex systems, characterized by high dimensionality, non-normal probability density functions and non-linear dependence relationships. Temperature belongs to bounded and skewed distributions, and precipitation is ruled by nonlinear and highly variable, in space and time, physical processes. The high zero frequency in precipitation probability distribution explains its discontinuity between zero and nonzero observations and its mixed discrete and continuous temporal correlations.

SWGs are numerical models that try to reproduce the statistical properties from the observed historical climate series: maximum temperature, minimum temperature and precipitation. SWG parameters comprise a concise summary of climate behavior and use Monte Carlo methods as a random number generator for simulation. In theory, these models can generate long synthetic weather series that preserve the statistical properties of the original data observed in a broad variety of climates and regions.

SWGs have numerous applications; they have been widely used as input in crop simulation models because of their ability to generate missing data and to produce long series to allow good estimates of the probability of extreme events that affect crop yield.[1] Also, their parameters can be interpolated to generate synthetic daily data for unobserved locations and they are frequently used in climate changes studies for impact evaluation.

In the last decade, the desire for the more accurate replication of stochastic patterns has led to the application of copula methods in the modeling of natural hazards.

The main objective of this research is to propose an alternative SWG that introduces the advantages of the copula modeling into the replication of stochastic weather patterns. This SWG approach models, in cross-section, the dependence structure between the probability distributions of the climatological variables, while the dynamics is reproduced by a Brownian bridge stochastic process that emulates the daily time stochastic behavior of the weather variables. Thus, the joint distribution of weather

[1] In terms of their use in crop growth models, the SWGs are used in the yield simulation for the ratemaking process of new crop insurance schemes when no historical yield series are available.

variables incorporates their non-linear dependence structure and more accurately reproduces the extreme weather patterns.

Furthermore, another important objective is to carry out a comprehensive evaluation of the proposed SWG in terms of its replication power of weather patterns through the climate simulation of three weather stations with highly differentiated climatic patterns across the United States and, in comparison with Richardsons SWG, to show the capabilities and the limitations of this approach.

2. Background and Related Work

SWGs are stochastic numerical models that reproduce the observed climate series by preserving their statistical properties. SWGs are not forecasting algorithms, neither are they deterministic weather models that numerically integrate partial differential equations. Traditional modeling of climate variables, on SWG, relies on a multivariate distribution, which is usually characterized jointly under the same parametric family and their pattern of dependence is assumed to be linear. Usually this multivariate approach is limited by its dependence structure (it only considers linear dependence), with a high number of parameters and dismisses additional information regarding their individual behavior (Schölzel and Friederichs 2008).

One of the most widely used SWG with this approach is the Richardson's SWG (Richardson, 1981). This SWG considers for each variable a dependence structure (serial correlation) characterized by a first order linear autoregressive model. The model configuration awards a primary role to precipitation, preserving the dependence on time, the correlation between variables, and the seasonal characteristics in actual weather data for the location. Precipitation is characterized in two steps. The first stage characterizes the precipitation occurrence process using a Markov chain exponential model with the two states: wet and dry. The probability of rain is conditioned on the wet or the dry

status of the previous day, which exhibits persistence or positive serial autocorrelation. [2] So, wet and dry runs tend to clump together in volume more strongly than could be expected by chance. Second, daily precipitation amount, given a wet day, is supposed to be independently determined by an exponential distribution (Richardson 1981). The inputs for the model are monthly means and coefficients of variation for each variable.

According to Wilks and Wilby (1999), although Richardson´s model operates on a daily time step, their simulations do not show longer-term variations. Their random realizations show a lower monthly mean temperature and lower monthly accumulated precipitation than the observed weather data. The implicit rigidity in the dependence pattern prevents the model from capturing the variability and the long-term changes in the climatological process.

3. The Semiparametric Copula-Based Stochastic Weather Generator

In recent years, the simulation of multivariate data using the copula approach, the extreme value analysis and the modeling of more complex dependence structures have increased in climatological phenomena analysis. This study proposes an alternative SWG, based on copula methodology, to simulate the weather variables: precipitation, maximum temperature and minimum temperature. The main objective of this research is to apply the copula conditional mixture approach to capture more accurately their nonlinear dependence structure and the occurrence of extreme events.

The copula approach has several advantages. In particular, its flexibility allows researchers to model dependence structures between random variables independently of the marginal involved and the different treatments on dependence structures for extreme events, common in weather variables. Multivariate probability distribution resulting from the copula approach might capture more accurately long-term changes in the

[2] The behavior of the Markov chain is ruled by the transition probabilities that specify the conditional probabilities for the system to be in each of its possible states during the next time period. In a first order Markov chain, the transition probabilities controlling for the next stage of the system depend only on the current state of the system (Wilks, 2011).

hydrologic cycle and weather patterns of specific regions because it can model different patterns of dependency and joint extreme events. The extreme dependence captured by extreme value copulas allows calculating the return period of a given event and reproducing more accurately the extreme weather events occurrence such as droughts, hail, heavy rain, cold spells and heat waves.

Basically the idea of modeling climatic variables using copula methods relies on the dynamics of these variables. Every year weather observations follow a well-known cycle: high temperature realizations in summers and low temperature realizations during winters. Although weather realizations are assumed stochastic, their differences between one day and the next one are not far. For example, usually the temperature on June 1st is at most two or three degrees different from June 2nd, or even on June 5th. Furthermore, the climate dependence patterns might not be adequately modeled assuming linear dependence. Figure 1 shows an association between maximum temperature realizations and minimum temperature realizations as well as a connection between lower realizations in minimum temperature and rainfall occurrence.

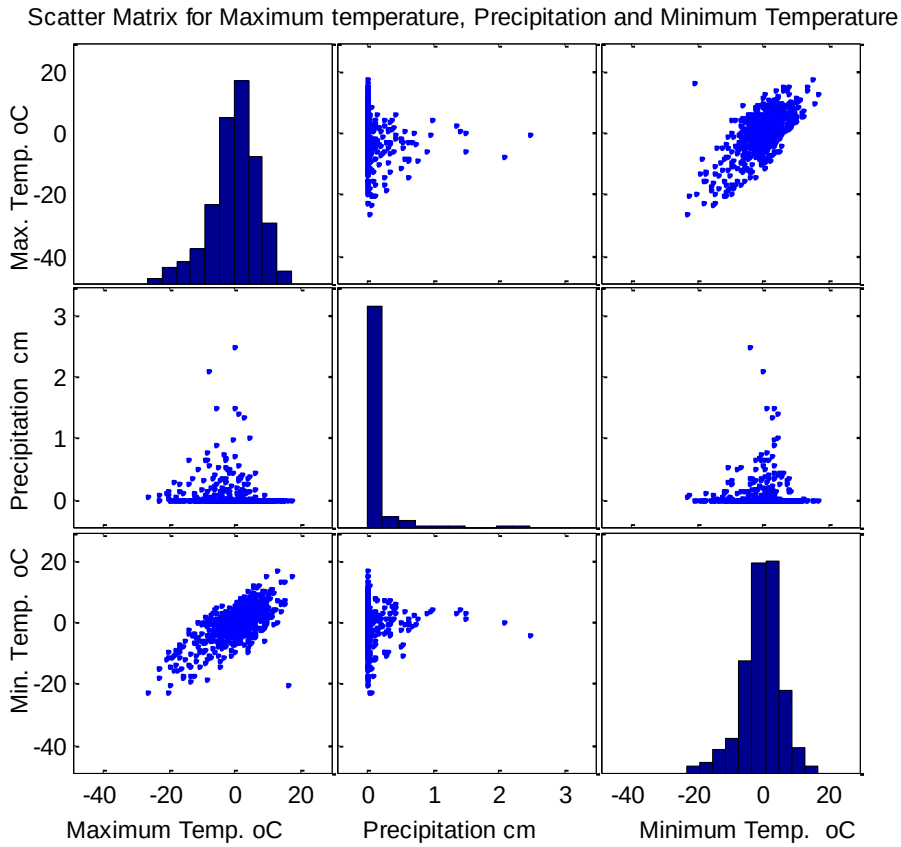


Figure 1. Bidimensional scatter matrix for weather variables from the weather station located in Conrad, Pondera County, Montana.

Although copula approach could capture more accurately the dependence patterns, its estimation on daily basis leads to a dimensionality problem. In this context, dimensionality problem refers to the copula modeling of all dates throughout a year for three variables (maximum temperature, minimum temperature and precipitation), which implies the simultaneous modeling of hundreds of variables. So, the volume of the space increases so fast that the available data become sparse, diminishing the reliability of the estimation.

In the most general case, each daily realization of a given weather variable would be treated as a separate random variable, so that modeling $\boldsymbol{k}$ random variables over a

period of $\boldsymbol{m}$ days would require the construction of a $\boldsymbol{kxm}$ dimensional joint distribution. The treatment proposed by this research for the dimensionality problem is the selection of the observations with the highest average anomalies per month to perform the copula estimation, thus reducing the number of random variables to 12 dates per year, or one per month.[3]

For the selected dates, the parameters of the probability distribution for the marginals are estimated individually, and the parameters of the trivariate copula are estimated jointly. Thus, the copula simulated weather variables describe their joint probability distributions at the selected time which are the bordering conditions of the climate stochastic dynamic simulation. The daily dynamic of weather series is emulated by a random walk described by a reversible geometric Brownian Motion that reflects the intertemporal dynamic of weather variables evolving on a path forward through time. Basically the semiparametric copula-based SWG structure would impose only a joint dependence structure and the Brownian motion process between the simulated structures to emulate their daily time stochastic dynamics.[4]

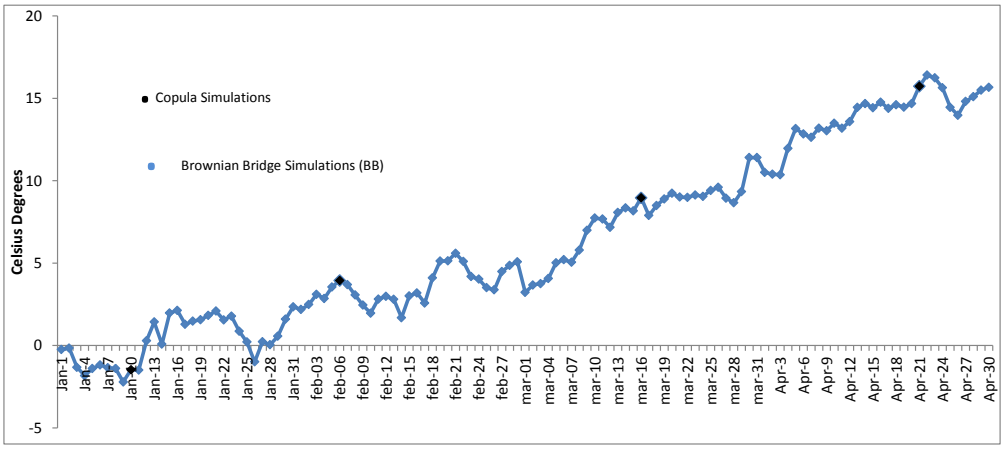


Figure 2. The semiparametric copula-based SWG incorporates the copula simulations as the bordering conditions of the climate stochastic dynamic simulation and the Brownian Bridge reproduces the daily dynamics of weather variables.

[3] Anomalies measure deviations over a certain period of time (month, season or year) from the long-term climate statistics, relating to their calendar period.

[4] The Brownian motion has been adopted as the probabilistic model for numerous natural phenomena such as weather and hydrology because it describes the random movement of particles in multidimensional space and its simplicity.

The strong connection between random weather variables and their daily sequence validate the assumption of the Brownian Bridge stochastic process interpolating the Copula realizations (see Figure 2). The Brownian Bridge results from the conditioning of the Brownian motion on its endpoints and its potential advantage relies on its use with variance reduction techniques and low discrepancy methods (Glasserman 2010). Turvey (2005) applied a similar idea for daily pricing of weather insurance for ice wine in the Niagara Peninsula of Southern Ontario.

Although there are numerous applications of copula modeling to hydrology and climate, to our knowledge this is the first application of copula methods for a multivariate SWG.

4. Methods and General Theory about Copulas

Copulas are joint cumulative distribution functions that describe dependencies among distributions independent of their univariate marginals which are $U(0,1)$ (Joe 1997). In terms of an m-dimensional distribution F with marginal cumulative distribution functions $(F_1, \dots, F_m)$, and a j^{th} univariate margin F_j, the copula associated with F is a distribution function $C: [0,1]^m \rightarrow [0,1]$ that satisfies

$$F(x) = C\ (F_1(X_1), \dots, F_m(X_m)), \qquad x \in R^m \tag{1}$$

Besides, if F is a continuous *m*-variate distribution function with univariate margins $(F_1, \dots, F_m)$, and quantile functions $(F_1^{-1}, \dots, F_m^{-1})$, then

$$C(u) = F(F_1^{-1}(u_1), \dots, F\ _m^{-1}(u_m)) \tag{2}$$

In terms of multivariate weather data simulation, copula representation is more than convenient because of their probabilistic interpretation. The Sklar theorem states that if all $(F_1, \dots, F_m)$ are continuous, then copula C^m is uniquely determined on the range of $(F_1, \dots, F_m)$. As a consequence, the joint probability density of multivariate

distributions can be presented as the product of the marginal probability densities and the copula density, which is the canonical representation (Cherubini et al. 2004).

$$f_x(\boldsymbol{x}) = \quad c_x\big(F_1(x_1), F_2(x_2), \dots, F_m(x_m)\big) \; . \; \prod_{j=1}^{m} f_j\big(x_j\big) \tag{3}$$

where

$$c_x\big(F_1(x_1), F_2(x_2), \dots, F_m(x_m)\big) = \frac{\delta(C(F_1(x_1), F_2(x_2), \dots, F_m(x_m)))}{\delta F_1(x_1), \delta F_2(x_2), \dots, \delta F_m(x_m)} \tag{4}$$

Two important implications are derived from Sklar theorem. First, the independent representation from marginals of the copula defines the dependence structure in the multivariate structure (Nelsen 2006). This separation between marginal distributions and dependence creates the flexibility to use marginals from different types of distributions that describe better the multivariate phenomena. The second implication is the possibility of simulating random variables with the same probability distributions as original data and preserving the dependence structure through the copula.

Copula Families

One direct application of the copula methods is the simulation of dependent variables, and diverse forms of dependence can be modeled using different copula families specifications. The copula families characterize dependence functional forms related to properties that include reflection symmetry, extreme value copula, multivariate extendibility, as well as dependence properties (Joe 1997). Each copula family or class is represented by its density and conditional distribution function and the parameter or a vector of parameters.

In particular, the Archimedean family is widely used in the modeling of climate and hydrological phenomena because of their simplicity, their broad variety in dependence structures and the modeling of multivariate extremes. For example, Frank copula can model both negative and positive association, and Gumbel copula can only

describe positive forms of association. Also, Frank copula does not capture tail dependence, as opposed to Gumbel copula. For the SWG, Gumbel copula can model the positive association and the concentration of observations in the extreme upper right corner of the Figure 1 that captures the distribution of the maximum temperature and minimum temperature observations.

In general terms, Archimedean copulas can be constructed by a generator function $\varphi: I \rightarrow R^m$. The most frequent source of generators for m-dimensional Archimedean copulas are the Laplace inverse transformations for distribution functions. This generator function (φ) must be strictly monotone, continuous, decreasing, convex with $\varphi(1) = 0$ to guarantee its existence and to allow the multivariate extension of the copula, see details in Cherubini et al. (2004). Given a generator φ and its pseudo-inverse,[5] the function $C: [0,1]^n \rightarrow [0,1]$ expresses an Archimedean copula C^A.

$$C^A(u_1, \dots, u_m) = \varphi^{-1}(\varphi(u_1) + \cdots + \varphi(u_m)) \tag{5}$$

The three Archimedian copulas used in this research are Gumbel, Clayton and Frank. There are important reasons for their use. Gumbel Clayton and Frank copula classes have been widely used in hydrology and climate modeling. These copula specifications are easily constructed and allow modeling different characteristics of the climate processes, such as tail dependence, that accounts for extreme dependence.

The extreme value Gumbel copula belongs to the Gumbel-Hougaard family. Gumbel copula is completely monotone, has upper tail dependence, extreme value copula and partial multivariate extension.[6]

$$Generator \quad \varphi(u) = (-\ln(u))^{\alpha} \tag{6}$$

$$\varphi^{-1}(t) = \exp\left(-t^{\frac{1}{\alpha}}\right) \tag{7}$$

[5] The pseudo-inverse function φ^{-1}, in composition with the generator, gives the identity, see Cherubini et al. (2004) for a detailed explanation.

[6] Extreme value distributions and their three types (Gumbel, Frétchet and the Weibull) provide the only non-degenerated limit laws for adequate transformed maxima of identical and independently distributed random variables. For a detailed reference in this issue, consult Embrechts et al. (2001).

$$C(u_1, u_2, \dots, u_m) = \exp\left\{-[\sum_{i=1}^{m}(-\ln u_i)^{\alpha}]^{\frac{1}{\alpha}}\right\} \quad for \;\; \alpha > 1 \tag{8}$$

Clayton *m*-copula, belongs to Clayton family, is completely monotone and owns lower and upper tail dependence.

$$Generator \quad \varphi(u) = \mathrm{u}^{-\alpha} - 1 \tag{9}$$

$$\varphi^{-1}(t) = (t+1)^{-\frac{1}{\alpha}} \tag{10}$$

$$C(u_1, u_2, \dots, u_m) = [\sum_{i=1}^{m}(\mathrm{u_i}^{-\alpha} - n + 1)]^{-\frac{1}{\alpha}} \quad for \;\; \alpha > 0 \tag{11}$$

Frank *m*-copula, belongs to Frank family, is completely monotone, has reflection symmetry, partial multivariate extension and extension to negative dependence, see Cherubini et al. (2004) for more details.

$$Generator \quad \varphi(u) = ln\left(\frac{\exp(-\alpha u)-1}{\exp(-\alpha)-1}\right) \tag{12}$$

$$\varphi^{-1}(t) = -\frac{1}{\alpha}\ln\left(1 + e^{t}(e^{-\alpha} - 1)\right) \tag{13}$$

$$C(u_1, u_2, \dots, u_m) = -\frac{1}{\alpha}\ln\left\{1 + \frac{\prod_{i=1}^{m}(e^{-\alpha u_i}-1)}{(e^{-\alpha}-1)^{m-1}}\right\} \quad for \;\; 0 \le \alpha \le \infty \; when \; n \ge 3 \tag{14}$$

Mixtures of Conditional Distributions

The conditional mixture method introduces additional flexibility in the model extending bivariate copulas to an arbitrary dimension, while it models different dependence patterns in the multivariate distribution. This technique allows additional flexibility because it can interpolate from perfect conditional negative dependence to perfect conditional positive dependence, with conditional independence in between (Salvadori et al. 2007).

For the SWG, the a priori application of one copula family in the modeling of the three weather variables could reduce the accuracy of the copula representation because the dependence specification for the temperature variables could not fit well the dependence structure between temperatures and precipitation. In this sense, the conditional mixture allows one to model the dependence pattern by pairs of variables

capturing the best dependence structure in each pair of variables using the conditional sampling method.

M-variate distributions can be constructed based on $m-1$ dimensional margins, which must have $m-2$ variables in common. If one is given, $(F_{12}, F_{23}, \dots, F_{m-1m})$, $m \geq 3$, it is possible to build an m-variate distribution starting with the trivariate distribution $F_{i,i+1,i+2} \in F(F_{i,i+1}, F_{i+1,i+2})$, next the four-variate distributions from $F_i, \dots, F_{i+3} \in F(F_{i,i+1,i+2}, F_{i+1,i+2,i+3})$ and so on. There is a bivariate copula C_{ij} associated with the (i, j) bivariate margin of the m-variate distribution. For (i, j) with $|j-i| > 1$, C_{ij} measures the amount of conditional dependence in the i^{th} and j^{th}variables, given those variables with indices in between (Joe 1997). Following Joe (1997), the next equation shows the construction of a trivariate copula family.

$$F_{123}(\boldsymbol{x}) = \int_{-\infty}^{x_2} C_{13}\left(F_{1|2}(x_1 \mid x_2) F_{3|2}(x_3 \mid x_2)\right) F_2(dx_2), \tag{15}$$

The conditional probability distribution functions are regular. The arguments of the integrand are conditional cumulative distribution functions ($F_{1|2}$ and $F_{2|3}$) obtained from F_{12} and F_{23}. They can be written in terms of copulas because by construction, equation (15) is a trivariate distribution with univariate margins F_1, F_2, F_3 and bivariate margins F_{12} and F_{23}. C_{13} can be interpreted as a copula representing the amount of conditional dependence between the first and third univariate margins given the behavior of the second (Joe 1997). This method can be extended recursively to an *m*-dimensional copula.

The trivariate copula can be derived directly by using the integral representation in equation (15) and Sklar's theorem. The application of the conditional mixture method for the estimation of a trivariate copula can include diverse marginal distributions and different copula family specifications.

$$C_{123}(u_1, u_2, u_3) = \int_0^{u_2} C_{13}\left(\frac{\partial C_{12}(u_1, x)}{\partial u_2}, \frac{\partial C_{23}(x, u_3)}{\partial u_2}\right) \mathrm{dx} \tag{16}$$

The estimation and simulation of copulas is possible by the calculation of partial derivatives, as the following equations show.

$$\begin{aligned} c_{123}(u_1, u_2, u_3) &= \frac{\partial^3 C(u_1, u_2, u_3)}{\partial u_1 \partial u_2 \partial u_3} \\ &= \frac{\partial^2 C_{13}\left(\frac{\partial C_{12}(u_1, x)}{\partial u_2}, \frac{\partial C_{23}(x, u_3)}{\partial u_2}\right)}{\partial u_1 \partial u_3} \times \frac{\partial^2 C_{12}(u_1, u_2)}{\partial u_1 \partial u_2} \\ &\times \frac{\partial^2 C_{23}(u_2, u_3)}{\partial u_2 \partial u_3} \\ &= c_{13}\left(\frac{\partial C_{12}(u_1, u_2)}{\partial u_2}, \frac{\partial C_{23}(u_2, u_3)}{\partial u_2}\right) \times c_{12}(u_1, u_2) \times c_{23}(u_2, u_3) \end{aligned} \tag{17}$$

with

$$\frac{\partial C_{ij}\left(u_i, u_j, \theta_{ij}\right)}{\partial u_i} \tag{18}$$

$$c_{ij}\left(u_i, u_j\right) = \frac{\partial^2 C_{ij}\left(u_i, u_j, \theta_{ij}\right)}{\partial u_i \partial u_j} \tag{19}$$

Thus, different specification families can be used to give more flexibility to the specification. Three different parameters substituting equation (18), and (19) into (17) result in a three-variable-three parameter copula density $c_{123}\left(u_1, u_2, u_3;\ \theta_1, \theta_{2,}\theta_{3,}\right)$. This expression can be used to estimate the parameter values by Maximum Likelihood or they can be rewritten in terms of their Laplace transformation representations for the Archimedian Copulas. Equations (20) and (21) detail the partial derivatives of the Laplace representation that simplifies the parameter estimation.

$$C_k\left(u_k \middle| u_1, \dots, u_{k-1}\right) = \frac{\varphi^{-1(k-1)}[\varphi(u_1) + \varphi(u_2) + \dots + \varphi(u_k)]}{\varphi^{-1(k-1)}[\varphi(u_1) + \varphi(u_2) + \dots + \varphi(u_{k-1})]} \quad (20)$$

$$\frac{\partial^{k-1} C_{k-1}(u_1, \dots u_k)}{\partial u_1, \dots, \partial u_{k-1}} = \varphi^{-1(k-1)}[\varphi(u_1) + \varphi(u_2) + \dots + \varphi(u_{k-1})] . \prod_{j=1}^{k-1} \varphi^{(1)}(u_j) \quad (21)$$

The canonical representation for the multivariate density function, in equation (3), allows decomposing the statistical modeling of copulas in two steps: first the identification and modeling of the marginal distributions; and second, the estimation of the suitable copula function. This procedure can be generalized to mainly three methods: the Exact Maximum Likelihood (EML) method, the inference for the marginal (IFM) method and the canonical maximum likelihood (CML) method.

5. Brownian Bridge

The Brownian Bridge reproduces the daily dynamics of the weather variables through the generation of high quality sequences that outline the paths of the Brownian motion process. The simulated realizations of the copula will act as milestones (or borderline conditions), while the Brownian bridge will sample the sequence of weather variables between such observations using Monte Carlo methods. Brownian Bridge has its origin in Brownian motion stochastic process.

Brownian motion is a continuous-time stochastic process $\{W(t),\ 0 \leq t \leq T\}$ that, in general terms, describes the random movement of particles in multidimensional space. Brownian motion is a continuous function on $[0, T]$ such that maps $t \rightarrow W(t)$ with probability 1; centered $W(0) = 0$; and it has independent increments normally distributed, $[W(t) - W(s)] \sim \mathrm{N}(0, \mathrm{t} - \mathrm{s})$ for any $0 \leq \mathrm{s} < \mathrm{t} \leq \mathrm{T}$.[7] A Brownian bridge constructed from a Brownian motion with drift μ, is the same as the

[7] Brownian motion is an exact method because the joint distribution of the simulated values $[W(t_1),\dots,W(t_n)]$ is the same for the joint distribution of the corresponding Brownian motion at $[t_1,\dots,t_n]$, see Glasserman (2010) for more details.

one constructed from a standard Brownian motion, only the first step (sampling the rightmost point) would change. Instead of sampling $W(t_n)$ from $N(0, t_n)$, it would be sampled from $N(\mu t_n, t_n)$.

The conditional distribution of $W(t_1), \dots, W(t_{n-1})$ given $W(t_n)$ is the same for all values of μ (Glasserman 2010). Let $Z_1, \dots, Z_n$ be independent standard normal random variables, so a Brownian motion on $[0, T]$ with $t_0 = 0$ and $W(0) = 0$ can be generated with the subsequent values as

$$W(t_{i+1}) = W(t_i) + \sqrt{t_{i+1} - t_i}\, Z_{i+1}, \quad i = 0, \dots, n-1 \tag{22}$$

However, this research carries out an implementation of this process in a discrete version. The Brownian bridge $B(t)$ is a continuous-time stochastic process that describes the conditional probability distribution of a short memory Brownian motion model $W(t)$, with the condition $B(1) = 0$.[8] Still, Brownian Bridge has stationary but non-independent increments which are the result of conditioning the final value to be canceled in the considered interval. Given the Brownian motion $\{W(t),\ 0 \le t \le T\}$ then the Brownian Bridge independent of $W(T)$ is expressed as

$$B(t) = W(t) - \frac{t}{T} W(T), \qquad t \in [0, T] \tag{23}$$

When the Brownian bridge realizations satisfy an initial point $B(t) = x$ and a final point $B(T) = y$, the Brownian Bridge can be expressed as

$$B_{0,x}^{T,y}(t) = x + W(t) - \frac{t}{T}(W(T) - y + x) \tag{24}$$

See Appendix A for a detailed explanation of Brownian Bridge treatment and construction.[9]

[8] The Martingale property is:

$$E[W(t_n)|W(t_{n-1}), W(t_{n-2}), \dots, W(t_1)] = W(t_{n-1}), \quad t_1 < t_2 < \cdots < t_n$$

[9] The Brownian Bridge Matlab program generates the underlying Brownian motion process by successive increments.

Simulation Methods for the SWG

Monte Carlo method is a fundamental component of the semiparametric copula-based SWG that generates independent sequences under the distributional assumptions defined and provides variance reduction.

In the copula simulation, Salvadori et al. (2007) and Cherubini et al. (2004) provide a straightforward method Monte Carlo based on Sklar's Theorem with a nested structure for the simulation of multivariate copula vectors by calculating partial derivatives of the copulas. Eventually only composite functions of partial derivatives for bivariate copulas are evaluated. Assume that $\boldsymbol{F}$ is a multivariate distribution with continuous marginals $(F_1, F_2, \dots, F_m)$ that can be represented by an m-copula, C^m. Then, the generation of a vector $(X_1, X_2, \dots, X_m) \sim \mathrm{F}$, can be done by simulating a vector $(U_1, U_2, \dots, U_m) \sim \mathrm{C}$, where the random variables U_i's are Uniform [0,1]. Because copulas are invariant to monotonic transformations, the simulated random vector $\boldsymbol{X}$ has the same dependence structure as vector $\boldsymbol{U}$.

Initially with the bivariate copula with known parameters, the idea is to generate pairs (u_1, u_2) of [0,1] uniformly distributed random variable samples U_1 and U_2whose joint distribution is C using the conditional distribution for the random variable U_2 at a given u_1 of U_1:

$$c_1(u_2) = \Pr(U_2 \leq u_2 \mid \mathrm{U}_1 = \mathrm{u}_1)$$

Where the conditional functions are the partial derivatives of the copula $C_1(u_2)$.

$$c_1(u_2) = \Pr(U_2 \leq u_2 \mid U_1 = u_1) = \frac{\delta C}{\delta\, u_1} = C_{u1}(u_2)^{10} \tag{25}$$

Then, two independent uniform random variable samples $(u_1, u_2) \in [0,1]$ are generated, where u_1 is the first draw. Next, the quasi-inverse function of $c_{u1}(u_2)$ that

[10] $c_1(u_2)$ is a non-decreasing function and exist for all $u_2 \in [0,1]$

depends on the copula´s parameter and u_1 is calculated. So, the second draw (u_2) is generated from

$$u_2 = c_1^{-1}(u_2)$$

Successively to simulate u_3, sampled from U_3 and consistent with the joint distribution function C and the previously sampled u_1, u_2

$$c_3\left(u_3 \mid u_1, u_2\right) = Pr\{U_3 \leq u_3 \mid U_1 = u_1, U_2 = u_2\} \tag{26}$$

$$C_3\left(u_3 \mid u_1, u_2\right) = \frac{\delta_{u_1,u_2} C(u_1, u_2, u_3)}{\delta_{u_1,u_2} C(u_1, u_2)} \tag{27}$$

Thus, to simulate u_3 from $C_3\left(u_3 \mid u_1, u_2\right)$, one must draw u_3 from $U(0,1)$ from which $u_3 = C_3^{-1}\left(u_3 \mid u_1, u_2\right)$ can be obtained through the equation $u_3 = C_3\left(u_3 \mid u_1, u_2\right)$ by numerical rootfinding.

For the simulation of the Brownian Bridge that emulates the daily dynamics of the weather series, Monte Carlo or quasi-Monte Carlo methods generate random sequences to outline the paths of the Brownian motion process, by sampling points acting as the milestones (Brandimonte 2006). The property of stationary independent increments of the Brownian Bridge makes the simulation process equivalent to the random variable generation from a specific infinitely divisible distribution (Glasserman 2010). Because Brownian bridge relies on Brownian motion, it exhibits centered Gaussian distribution, and Martingale and Markovian properties that aggregate more persistence in the simulated weather series.

6. Generation Methodology

The methodology to simulate weather variables using the semiparametric copula-based SWG comprises eight stages. The first step includes the date selection of the observations with the highest average monthly anomaly. The second step consists in detrending the original daily weather observations. The third step relies on the selection

of the probability distributions for the marginals. The forth step involves the estimation of parameters and the determination of the best specification for copula mixture. The fifth step is the estimation of trivariate normal parameters for the Brownian bridge. The sixth step consists of the copula simulation for the selected dates. The seventh step is the Brownian bridge generation to emulate daily dynamics of the weather variables. Finally, the eighth step comprises the incorporation of the trend into the daily simulated weather variables.

The semiparametric copula-based SWG is applied to simulate weather for three locations with highly differentiated weather patterns across the United States: Conrad, Montana; Spokane, Washington; and Temple, Texas. The climatological information was obtained from the National Oceanic and Atmospheric Administration (NOAA) website.[11] The available data for Conrad and Spokane weather stations was 50 years (1960-2010) of daily observations. For Temple weather station, 42 years of daily data was available. The selection of dates was carried out by choosing 12 observations per year (one per month) according to the highest absolute deviations from historical monthly means.

Variable Detrending

The modeling of the weather variables requires the decomposition of the series when some sequential or cyclical patterns are observed. The standard methodology consists of decomposing the series in long-term trend, seasonal behavior and white noise. Harmonic analysis is useful to extract the fluctuations and variations in the series, using sin and cosine functions.[12] Application of harmonic series requires three adjustments (Wilks 2011). First, the fundamental frequency term $w_1 = 2\Pi/n$ rescales proportionally time to angular measure, i.e. specifies the fraction of the full cycle over the whole data series (given n, the length of the data is considered as a full cycle of 360^o or 2π radians in angular measure). Second, the amplitude (C_1) is the determination of

[11] http://gis.ncdc.noaa.gov/map/cdo/?thm=themeDaily

[12] These periodic functions have repetitive patterns every 2 π radians or 360^o and they oscillate around their average value of zero and attain maximum values of +1 and minimum of -1. The cosine function is maximized at 0^o, 360^o and so on, the sine function is maximized at 90^o, 450^o and so on.

the stretching or compressing of the cosine or sine into the range of the data. Third, the phase angle or phase shift (ϕ) that makes the lateral adjustment of the harmonic function.

$$Y_t = \bar{y} + C_1 \cos\left(\frac{2\Pi t}{n} - \phi\right) + C_1 \sin\left(\frac{2\Pi t}{n} - \phi\right) \tag{28}$$

$$C_1 \cos\left(\frac{2\Pi t}{n} - \phi\right) = A_1 \cos\left(\frac{2\Pi t}{n}\right) + B_1 \sin\left(\frac{2\Pi t}{n}\right) \tag{29}$$

Where $A_1 = C_1 \cos(\phi)$ and $B_1 = C_1 \sin(\phi)$ are the amplitudes of an upshifted cosine and sine waves. The parameters A_1 and B_1 are calculated by using standard regression methods.

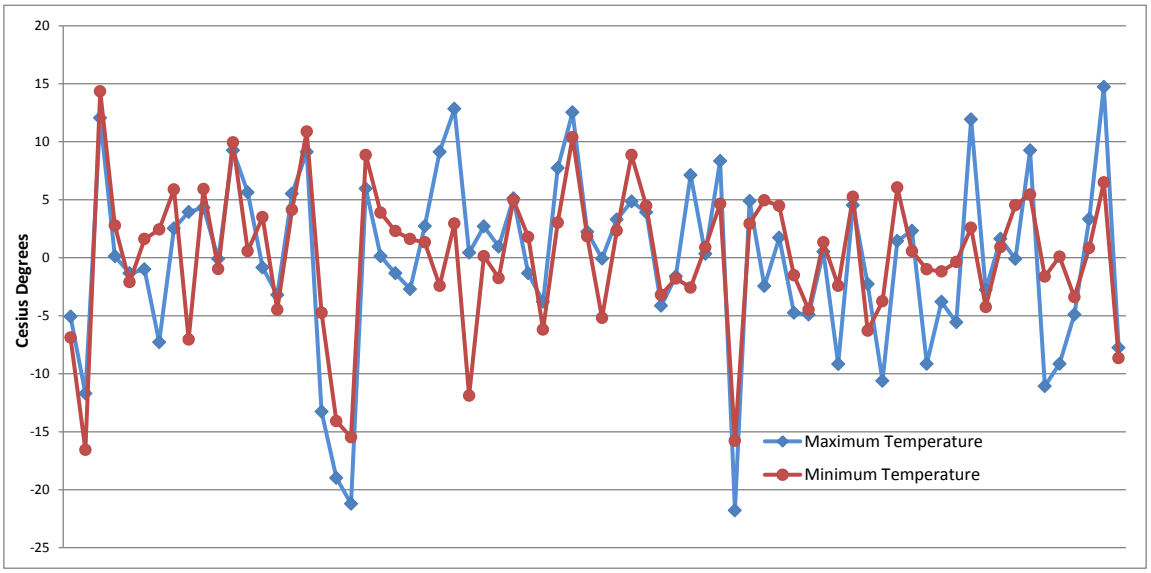


Figure 3. Detrended maximum temperature and minimum temperature anomalies, Conrad, Pondera County, Montana

This detrending technique was applied to daily weather observations, considering a cycle of 365 days, for maximum temperature and minimum temperature for the three

weather stations.[13] Once the trend was removed from these data series, specific dates were selected for the copula parameter estimation. Figure 3 shows the application of this method for the maximum temperature monthly anomalies with data from Conrad, Montana weather station.

Selection Process for Marginal Distributions

Parametric distributions have been widely used to model climate variables. In parametric density fitting, the criterion of selection for the best fit distribution rely on Maximum Likelihood as a competitive indicator of goodness of fit, especially if the parametric densities have the same number of parameters. However, when the number of parameters differs, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) can derive a conclusion about fitting the distributions.

However, the pitfall of the parametric approach is the a priori assumption of the parametric functional form of the variable to be estimated. Misspecification often occurs because restrictive assumptions can result in a misrepresentative characterization of the true density, thus producing erroneous estimates that lead to unsound inference. Frequently Gaussian distributions are used for modeling temperature with the Box-Cox transformation. Gamma distribution is suitable to model precipitation, but estimation is complex because parameters do not exactly correspond to the moments of the distribution (Wilks 2011).

Nonparametric characterization of the marginal distribution is a potential option because of its flexibility. Nonparametric representation requires more data to achieve the same grade of precision as a parametric model and some regularity conditions such as smoothness and differentiability (Wand and Jones 1995).

[13] The coefficients of the regressions for maximum and minimum temperature were significant explaining 66% of the behavior of both variables in Conrad, 80% and 70% respectively in Spokane and 69% and 75% respectively in Temple. In precipitation, any trend specification was significant for all the stations.

$$\hat{f}(x;h) = \frac{1}{nh}\sum_{i=1}^{n} K\left\{\frac{(x - x_i)}{h}\right\} \tag{30}$$

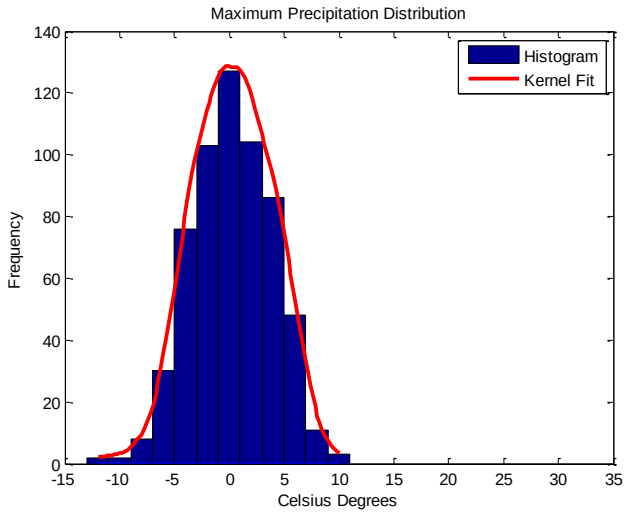
Where K is a function that satisfies$\int K(x)dx = 1$, which is the kernel and h is bandwidth or window width and is a positive number. K is chosen to be a unimodal probability density function that is symmetric about zero ensuring that $\hat{f}(x;h)$ is a density (Wand and Jones 1995). For a given sample size n, if h is small, the resulting estimator will have a small bias but a large variance. Conversely, if h is large, the resulting estimator will have a small variance but large bias. Minimization of the Mean Square Error (MSE) – which is the error measure of the estimation of the density at a single point of the density kernel function – is a consequence of the bandwidth optimal selection, which requires the balance of the bias squared and the variance terms. Although, there is a broad variety of kernel functions (uniform, triangular, biweight, triweight, Epanechnikov, Gaussian), by simplicity this research will focus on Gaussian kernel with emphasis on the choice of the bandwidth.

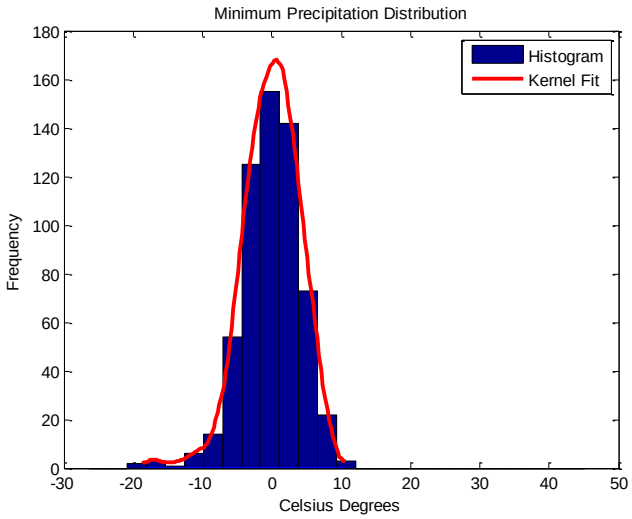
Table 1. Parametric distributions fit for weather variables, three weather stations

Distribution	α	β	- Σ log L	AIC	BIC
	Conrad MT1974, Pondera County, Montana				
			Normal		
Maximum Temperature	0.40	5.16	1,835.79	-3,667.58	-3,658.79
Precipitation	0.06	0.22	-64.48	132.95	141.75
Minimum Temperature	-0.57	5.90	1,916.06	-3,828.12	-3,819.33
			Extremevalue (Gumbel)		
Maximum Temperature	3.05	7.64	2,028.13	99,999.00	99,999.00
Precipitation	0.21	0.53	387.32	-770.64	-761.84
Minimum Temperature	2.22	5.38	1,920.69	3,986.14	3,994.93
			Exponential		
Precipitation	0.06		-1,071.89	2,145.79	2,150.18
	Spokane WA, Spokane County, Washington				
			Normal		
Maximum Temperature	0.29	3.62	1,622.86	-3,241.72	-3,232.93
Precipitation	0.12	0.30	128.04	-252.08	-243.29
Minimum Temperature	-0.09	4.16	1,705.92	-3,407.85	-3,399.05
			Extremevalue (Gumbel)		
Maximum Temperature	2.08	3.42	1,651.93	3,467.29	3,476.08
Precipitation	0.30	0.52	421.92	-839.85	-831.05
Minimum Temperature	1.88	3.68	1,703.00	3,551.31	3.56E+03
			Exponential		
Precipitation	0.12		-685.35	1,372.70	1,377.10
	Temple TX, Bell County, Texas				
			Normal		
Maximum Temperature	0.36	3.90	1,398.35	-2,792.71	-2,784.27
Precipitation	0.24	0.89	656.70	-2,928.01	-1,300.95
Minimum Temperature	-0.40	4.47	1,466.00	-2,928.01	-2,919.56
			Extremevalue (Gumbel)		
Maximum Temperature	2.30	4.20	1,456.83	3,102.95	3,111.39
Precipitation	0.86	2.08	1,019.89	-2,035.78	-2,027.34
Minimum Temperature	1.84	4.66	1,519.48	3,213.06	3,221.50
			Exponential		
Precipitation	0.24		-205.38	412.76	416.98

Table 1 shows the results of the parametric estimation for the marginals. Although the AIC and the BIC show the normal distribution as the best parametric specification for maximum temperature and minimum temperature, and the extreme value distribution for the precipitation, these distributions hardly provide a good

description of the data.[14] The kernel distribution attains the best fit for maximum and minimum temperature, while the large volume of weather data provides reliability on non-parametric estimations that usually captures more accurately the probability in the tails of the distribution.





[14] Parameters for other parametric distributions were fitted, but the results were even poorer than those in Table 1.

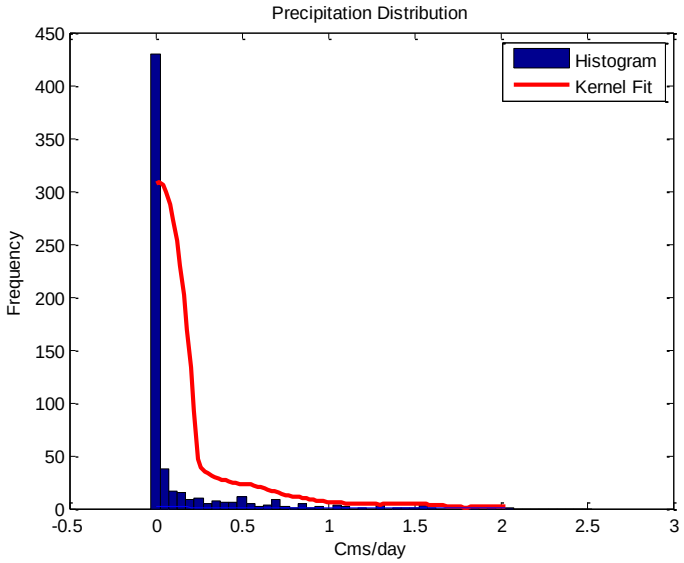


Figure 4. Probability kernel distribution fit: Conrad, Pondera County, Montana

Copula Family Selection

Usually the goodness of fit test (GOF) helps to determine if the observed data are well-modeled by the specified dependence structure of the multivariate distribution for the specific family of parametric copulas. However, the development of a GOF test for the mixture copula exceeds the primary objective of this research. This research uses one reasonable criterion for the selection of the appropriate copula family based on a ranking that measures the likeness of that sample coming from a given distribution. Although maximum likelihood cannot be the criterion because parametric distributions are unknown, it is possible to use maximum likelihood as ranking base measure for all component mixtures that denote improvements in flexibility.

Although important advances have been attained in GOF test, the formal methodology is just recently emerging and the developments for the multivariate case are inconclusive (Genest et al. 2009). Most of the progress has been done for a one-dimensional test, for the multivariate case the value of the statistics depends on the order

of conditioning of the variables. So, different conditioning decisions could lead to different results (Genest et al. 2009).

Table 2. Copula mixture estimation for climatic variables by weather stations

Copula Mixture	Param. 1	Param. 2	Param. 3	- Σ log L	AIC	BIC
	Conrad MT1974, Pondera County, Montana					
Clayton	0.001000			0.041399	1.917202	6.314132
Frank	0.001000			0.025487	1.949025	6.345955
Gumbel	**1.100000**			**5.732520**	**-5.244449**	**-9.465039**
Clayton, Clayton,Clayton	0.001000	0.077741	-0.048485 *	-1.036122	8.072244	21.263033
Clayton, Clayton,Gumbel	1.022122	-0.164136	0.846158 *	-19.985389	45.970779	59.161568
Frank, Frank, Frank	0.001000	-1.099259	-1.418806	-13.851869	33.703739	46.894528
Frank, Frank, Clayton	0.001000	0.862911	-0.044400 *	-1.105269	8.210538	21.401327
Frank, Frank, Gumbel	1.000204	-1.297100	0.832434	-20.887645	47.775289	60.966078
Gumbel,Gumbel, Gumbel	1.513458	0.952485 *	0.751978 *	-25.867954	57.735909	70.926698
Gumbel, Frank, Clayton	0.813972 *	0.925946	-0.043871	-4.338217	14.676433	27.867222
Gumbel, Frank, Gumbel	0.995191	1.000036	0.842727 *	-25.061201	56.122402	69.313191
	Temple TX, Bell County, Texas					
Clayton	0.00100			0.05886	6.10288	1.88229
Frank	0.00100			0.02635	6.16790	1.94731
Gumbel	**1.10000**			**-41.69724**	**-29.63757**	**-25.41698**
Clayton, Clayton,Clayton	0.00100	0.14990	-0.07667 *	-3.51428	25.69033	13.02856
Clayton, Clayton,Gumbel	0.00100	0.04713	0.77590 *	-32.56102	83.78380	71.12203
Frank, Frank, Frank	0.00100	1.21032	-2.14424 *	-29.12381	76.90940	64.24763
Frank, Frank, Clayton	0.00100	1.73605	-0.07553 *	-5.27117	29.20411	16.54234
Frank, Frank, Gumbel	0.00100	1.61078	0.79008 *	-34.17259	87.00695	74.34518
Gumbel,Gumbel, Gumbel	0.90662 *	1.11905	0.77290 *	-41.69724	102.05625	89.39448
Gumbel, Frank, Clayton	0.87410	1.75801	-0.07423 *	-7.30074	33.26324	20.60147
Gumbel, Frank, Gumbel	0.90662 *	1.11897	0.77243 *	-40.54667	99.75512	87.09335
	Spokane WA, Spokane County, Washington					
Clayton	0.00100			0.11772	1.76457	6.16150
Frank	0.00100			0.02776	1.94448	6.34141
Gumbel	**1.10000**			**-32.98232**	**-37.37925**	**-32.98232**
Clayton, Clayton,Clayton	0.00100	0.21398	-0.12842	-11.13180	28.26361	41.45439
Clayton, Clayton,Gumbel	0.00100	1.13314	0.89037 *	-12.96967	31.93935	45.13014
Frank, Frank, Frank	0.00100	0.19995	0.87737	-16.54238	39.08476	52.27555
Frank, Frank, Clayton	0.00100	1.77027	-0.10693 *	-13.77605	33.55209	46.74288
Frank, Frank, Gumbel	0.00100	1.72906	-1.00565 *	-16.30425	38.60849	51.79928
Gumbel,Gumbel, Gumbel	0.00100 *	1.85597	0.89791 *	-34.14157	74.28315	87.47394
Gumbel, Frank, Clayton	0.73401 *	1.83158	-0.13861	-34.72642	75.45285	88.64364
Gumbel, Frank, Gumbel	0.81940 *	2.74101	0.86815 *	-33.25106	72.50213	85.69292

Note: * Parameters that do not satisfy monotonicity conditions.

It is impossible to evaluate all copula mixtures combinations. However, in this context the maximum likelihood provides some discernment about the applicability of a

particular distribution to every sample. Table 2 shows the AIC and BIC statistics for the considered weather stations. The best specifications are attained by the one-parameter-Gumbel copula family.

7. Statistical validation for the semiparametric copula-based SWG

A comparative evaluation of the semiparametric copula-based SWG performance versus the Richardson's SWG performance is carried out to learn about their strengths and limitations.

Two-Sample Kolmogorov-Smirnov Test

Because any parametric functional form is used, a non-parametric two sample Kolmogorov-Smirnov test can be used to evaluate the performance of the Copula method to replicate the distribution for the weather series. The two-sample Kolmogorov-Smirnov test (KS) is applied to Copula simulations to compare the $c.d.f.$ of the generated weather series vs. the $c.d.f.$ of original observed weather data at each one of the three locations. In this context, this non-parametric test compares two unknown $c.d.f.$s: F for the observed data and G for the simulation, quantifying the distance between the empirical distribution functions of the two samples through the test statistic in the following expression

$$D_n = \sup |F_{n1} - G_{n2}| \tag{31}$$

Where $F_{\_n1}$ is the empirical $c.d.f.$from a sample of n_1 data values (observed weather data) and $G_{\_n2}$ is the empirical cdf from a sample of n_2 data values (simulated data).

The null hypothesis is H_0: $F_{n1} = G_{n2}$. The fit is measured by the statistic $D_n$with its asymptotic distribution and the limiting distribution $\sqrt{n}D_n$ is distribution free, in consequence, the reasonable criterion is to reject H_0 if D_n is large (Mood et al. 1974). The critical values were obtained based on a 120-year sample of draws for every one of the three locations using both SWGs and the two-sample Kolmogorov-Smirnov test. In

large samples, critical values for the statistic D_n are determined by simulation.

Table 3. Two-sample Kolmogorov-Smirnov test for selected dates

(P-Values)

	Copula Based SWG			Richardson's SWG		
Selected Dates	**Conrad, Montana**			**Conrad, Montana**		
	Max Temp	**Precipitation**	**Min Temp**	**Max Temp**	**Precipitation**	**Min Temp**
January, 9th	0.028 *	0.267	0.002 **	0.226	0.560	0.022 *
February, 1st	0.001 **	0.998	0.000 **	0.167	0.989	0.062
March, 3rd	0.453	0.999	0.525	0.433	0.610	0.085
April, 3rd	0.737	0.995	0.242	0.448	0.662	0.794
May, 1st	0.372	0.999	0.095	0.028 *	0.081	0.073
June, 1st	0.000 **	0.072	0.019 *	0.019 *	0.053	0.267
July, 1st	0.001 **	0.071	0.046 *	0.628	0.404	0.794
August, 7th	0.011 *	0.009 **	0.112	0.190	0.404	0.069
September, 5th	0.571	0.886	0.225	0.062	0.696	0.139
October, 1st	0.744	0.995	0.306	0.017 *	0.880	0.145
November, 4th	0.819	0.969	0.063	0.104	0.960	0.056
December, 1st	0.763	0.669	0.489	0.867	0.960	0.464
	Copula Based SWG			**Richardson's SWG**		
	Spokane, Washington			**Spokane, Washington**		
January, 9th	0.069	0.001 **	0.107	0.074	0.008 **	0.000 **
February, 1st	0.159	0.050 *	0.001 **	0.005 **	0.008 **	0.000 **
March, 3rd	0.113	0.024 *	0.304	0.001 **	0.326	0.000 **
April, 3rd	0.012 *	0.804	0.017 *	0.055	0.100	0.000 **
May, 1st	0.077	0.362	0.058	0.673	0.010 **	0.000 **
June, 1st	0.089	0.765	0.360	0.718	0.000	0.056
July, 1st	0.082	0.999	0.011 *	0.673	0.050	0.098
August, 7th	0.007 **	0.530	0.000 **	0.000 **	0.999	0.000 **
September, 5th	0.691	0.701	0.730	0.039	0.368	0.000 **
October, 1st	0.022 *	0.739	0.172	0.026 *	0.308	0.000 **
November, 4th	0.023 *	0.004 **	0.123	0.001 **	0.001	0.000 **
December, 1st	0.000 **	0.000 **	0.371	0.006 **	0.000	0.000 **
	Copula Based SWG			**Richardson's SWG**		
	Temple, Texas			**Temple, Texas**		
January, 9th	0.035 *	0.393	0.015 *	0.082	0.166	0.860
February, 1st	0.013 *	0.049 *	0.040 *	0.407	0.211	0.945
March, 3rd	0.057	0.232	0.074	0.377	0.673	0.339
April, 3rd	0.007 **	0.959	0.530	0.709	0.860	0.860
May, 1st	0.113	0.001 **	0.203	0.356	0.231	0.186
June, 1st	0.000 **	0.114	0.013 *	0.252	0.403	0.087
July, 1st	0.000 **	0.989	0.000 **	0.015 *	0.915	0.108
August, 7th	0.013 *	0.980	0.003 **	0.200	0.999	0.067
September, 5th	0.002 **	0.985	0.000 **	0.356	0.915	0.209
October, 1st	0.111	0.965	0.343	0.915	0.761	0.938
November, 4th	0.046 *	0.866	0.199	0.274	0.403	0.615
December, 1st	0.967	0.994	0.142	0.399	0.575	0.549

Note: * Reject H_0 at 5% significance level

** Reject H_0 at 1% significance level

Table 3 shows the p-values for the selected dates generated. Simulated dates that reject the H_0 are marked with asterisks, in these cases the probability distribution of the simulated weather data does not correspond to the probability distribution of the observed data. In the case of the semiparametric copula-based SWG simulations the H_0 is rejected in 37% of the cases, while in the case of Richardson SWG is 27%.

This rate of rejection in the case of the semiparametric copula-based SWG can be attributed to the fact that the KS test tends to be more sensitive near the center (median values) of the distribution than at the tails. The cases of rejection for the Richardson's SWG are concentrated in simulated data of the Spokane weather station, whose distributions show more rounded peaks than in Temple, Texas weather station. In contrast, the cases of rejection for the semiparametric copula-based SWG are concentrated in Temple, Texas weather station, whose distribution has a more acute peak around the mean.

Quantile Analysis

Quantiles of the distributions are calculated to analyze in detail the differences in the distributions for the simulated weather series versus the observed weather series. A 120-year simulation was performed to carry out the quantile analysis. The quantiles of a distribution are points taken at regular intervals c.d.f. function that provides nonparametric estimators of their population counterparts based on a set of independent observations $\{X_1, X_2, X_3\}$ from the distribution F. Quantile of the distribution F is defined by the following expression:

$$Q(p) = F^{-1}(p) = \inf\{x: F(x) \geq p\}, \qquad 0 < p < 1 \tag{32}$$

Let $\{X_{(1)}, X_{(2)}, \dots, X_{(n)}\}$ denote the order statistics of $\{X_1, X_2, X_3\}$ and let $\widehat{Q_\iota}(p)$ denote the i^{th} sample quantile.

Table 4. Comparative quantile analysis for three locations[15]

	Precipitation Amount, cm			Maximum Temperature, °C			Minimum Temperature, °C		
Quantile	**Observed**	**Copula**	**Richardson**	**Observed**	**Copula**	**Richardson**	**Observed**	**Copula**	**Richardson**
				Conrad, Montana					
0.025	0.000	0.000	0.000	-13.300	-8.430	-10.509	-25.000	-19.554	-23.416
0.05	0.000	0.000	0.000	-7.800	-5.078	-6.691	-21.100	-16.894	-19.987
0.1	0.000	0.000	0.000	-1.100	-1.570	-2.270	-15.600	-13.919	-15.950
0.2	0.000	0.000	0.000	4.400	2.882	3.499	-9.400	-10.824	-10.628
0.3	0.000	0.000	0.000	7.800	6.621	7.845	-5.600	-8.173	-6.548
0.4	0.000	0.000	0.000	11.100	10.101	11.656	-2.800	-5.286	-3.130
0.5	0.000	0.001	0.000	14.400	13.668	15.179	-0.600	-2.231	-0.109
0.6	0.000	0.002	0.000	18.300	17.021	18.574	2.200	0.850	2.653
0.7	0.000	0.004	0.000	21.700	20.708	21.926	5.000	4.014	5.170
0.8	0.000	0.012	0.000	25.000	24.521	25.360	7.800	7.213	7.532
0.9	0.203	0.227	0.189	28.900	29.184	29.218	10.000	10.269	10.171
0.975	0.864	0.700	0.862	32.800	35.305	34.067	13.300	14.168	13.460
				Spokane, Washington					
0.025	0.000	0.000	0.000	-5.000	-4.310	-9.620	-13.300	-9.946	-21.603
0.05	0.000	0.000	0.000	-2.200	-2.193	-6.219	-10.000	-8.247	-18.306
0.1	0.000	0.000	0.000	0.600	0.513	-2.214	-6.100	-6.355	-14.364
0.2	0.000	0.000	0.000	3.900	4.117	3.000	-2.800	-3.920	-9.242
0.3	0.000	0.000	0.000	7.200	7.296	6.939	-1.100	-1.713	-5.409
0.4	0.000	0.000	0.000	10.000	11.101	10.356	0.600	0.707	-2.121
0.5	0.000	0.002	0.000	13.300	14.906	13.668	2.800	3.323	0.757
0.6	0.000	0.004	0.000	17.200	18.782	16.967	5.000	5.676	3.400
0.7	0.025	0.013	0.000	21.100	22.232	20.440	7.800	7.952	5.899
0.8	0.127	0.142	0.057	25.000	25.227	24.114	10.000	10.098	8.255
0.9	0.406	0.385	0.310	29.400	28.778	28.416	12.800	12.460	10.881
0.975	1.036	0.847	1.051	33.900	33.279	33.821	16.700	15.418	14.170
				Temple, Texas					
0.025	0.000	0.000	0.000	5.600	8.628	5.834	-3.900	-1.949	-3.887
0.05	0.000	0.000	0.000	8.300	10.808	8.879	-1.700	-0.089	-1.718
0.1	0.000	0.000	0.000	12.800	13.258	12.514	1.100	1.579	0.916
0.2	0.000	0.000	0.000	17.200	16.658	17.173	4.400	4.265	4.598
0.3	0.000	0.000	0.000	21.100	19.400	20.764	7.700	6.622	7.802
0.4	0.000	0.000	0.000	23.900	22.198	23.771	10.600	9.397	10.868
0.5	0.000	0.000	0.000	26.700	25.397	26.549	13.900	12.355	13.943
0.6	0.000	0.000	0.000	28.900	28.454	29.058	17.200	15.529	16.841
0.7	0.000	0.000	0.000	31.700	30.986	31.440	19.900	18.297	19.272
0.8	0.025	0.025	0.000	33.300	33.538	33.684	21.700	20.588	21.130
0.9	0.533	0.802	0.618	35.600	36.810	36.122	22.800	23.421	22.878
0.975	2.769	2.901	2.767	37.800	41.595	39.152	23.900	26.973	25.023

Table 4 shows values of the weather variables for different quantiles of the distribution. The semiparametric copula-based SWG generates weather series

[15] In weather stations the minimum reported amount of precipitation is 0.0254 cm (0.01 inches).

significantly closer to the original observed data. Although the reproduction of the weather patterns is consistent, the replication of the climate is comparatively better for the station of Spokane, Washington and Temple, Texas than for Conrad, Montana. The values of the lower percentiles are more accurate in the case of the simulations generated by the semiparametric copula-based SWG. This result could be attributed to the property of Gumbel´s copula to capture upper tail dependence of the distribution.[16]

Statistical Analysis of the Simulated Weather Series

The validation of a weather generator based only on the analysis of their moments distribution (mean, standard deviation, skewness and kurtosis) is insufficient. A more accurate description of the occurrence of precipitation by season provides key information to evaluate the performance of the semiparametric copula-based SWG. For such purpose 28-day period indicators were calculated for both, the generated and the observed weather data series. Next, mean values of accumulated precipitation amounts (cm), mean number of rainy days, mean minimum temperature and mean maximum temperature per period were calculated.

[16] Several weather series were generated using different copula families and, in some cases, the results were substantially different in terms of the weather patterns reproduction from observed data.

Table 5. Average rainfall amount and average number of rainy days by 28-day period

	Precipitation Amount in Cms.			Number of Rainy Days		
Period	**Observed**	**Copula**	**Richardson**	**Observed**	**Copula**	**Richardson**
			Conrad, Montana			
1	0.976	1.590	0.858	5.020	4.625	3.067
2	0.798	1.502	0.801	4.120	4.708	2.720
3	1.193	1.573	1.148	4.580	5.125	3.367
4	2.111	2.002	1.853	5.040	5.525	3.780
5	3.335	1.903	3.722	6.620	5.767	5.187
6	6.672	1.272	5.227	9.800	4.542	6.787
7	4.086	2.323	3.960	7.420	5.958	5.293
8	2.483	2.535	2.681	5.440	6.542	4.507
9	2.925	1.950	2.565	5.900	5.058	4.573
10	2.156	1.357	1.897	4.960	4.367	3.513
11	1.037	1.699	1.137	3.660	4.558	2.533
12	1.127	1.703	1.108	4.420	4.242	3.007
13	1.036	2.210	0.901	4.700	5.925	2.947
			Spokane, Washington			
1	4.390	3.053	1.233	12.314	8.958	4.253
2	3.807	3.196	1.149	10.824	9.575	4.020
3	3.477	2.946	1.555	10.275	6.233	5.093
4	3.035	2.978	2.760	9.118	4.967	6.267
5	3.123	3.106	5.468	8.627	5.408	8.187
6	3.872	2.295	6.699	8.765	4.042	9.080
7	1.805	2.287	5.888	5.118	3.533	8.300
8	0.977	1.531	4.010	3.216	3.350	6.153
9	1.710	1.594	3.651	4.647	2.625	6.080
10	1.659	3.588	2.715	5.314	5.258	5.433
11	2.839	4.311	1.596	7.353	6.642	3.860
12	5.558	3.675	1.159	12.667	6.683	3.573
13	5.443	3.111	1.223	12.824	8.717	4.060
			Temple, Texas			
1	4.999	8.728	5.098	6.561	6.608	5.380
2	6.792	7.112	5.753	6.951	5.208	5.540
3	5.552	5.798	5.480	6.512	4.408	5.340
4	5.297	4.364	5.947	5.927	4.383	5.030
5	10.555	6.271	9.093	6.976	5.125	5.450
6	9.318	10.320	9.153	6.171	6.017	5.000
7	5.472	8.680	6.428	4.634	5.442	3.840
8	4.265	6.061	4.283	3.220	5.700	2.950
9	6.122	6.591	5.720	4.439	5.833	3.890
10	8.556	4.532	9.753	5.902	5.250	5.110
11	9.223	5.943	7.331	6.024	5.117	4.360
12	6.554	9.549	6.826	5.634	6.283	4.820
13	6.209	10.246	7.352	5.976	7.333	5.240

Table 5 shows that the simulated mean precipitation amounts do not differ significantly from the values obtained from the observed data. However, the replication of the volume of rainfall is more accurate for locations with higher amounts of water such as Temple, Texas than in locations with low levels of rainfall during the year. The average number of days per period generated by the semiparametric copula-based SWG was in general terms close to the observed data. However, the semiparametric copula-based SWG shows certain inflexibility in replicating the amounts of water and the recurrence of rain periods in highly variable precipitation patterns.

Table 6. Average maximum temperature and average minimum temperature by 28-day period

	Maximum Temperature °C			Minimum Temperature °C		
Period	Observed	Copula	EPIC	Observed	Copula	EPIC
			Conrad, Montana			
1	0.03	0.48	1.16	-13.47	-13.06	-13.04
2	3.38	1.23	3.26	-10.59	-12.20	-10.89
3	6.73	5.35	6.99	-7.85	-9.15	-7.70
4	12.62	11.15	12.61	-2.92	-4.32	-3.15
5	17.37	17.12	17.79	1.60	1.28	1.69
6	21.56	22.44	21.86	6.19	5.72	5.88
7	25.90	26.45	25.88	9.07	8.41	8.87
8	28.48	27.05	27.57	10.24	8.66	9.45
9	25.86	24.62	25.50	7.92	7.04	7.42
10	20.04	19.89	20.19	2.78	2.77	2.80
11	14.10	13.20	14.09	-1.93	-2.85	-2.20
12	5.74	6.81	6.86	-8.12	-7.93	-7.59
13	1.03	2.74	1.83	-11.91	-11.39	-12.05
			Spokane, Washington			
1	0.45	0.66	1.88	-5.60	-6.12	-11.35
2	3.77	1.43	2.89	-3.65	-5.74	-9.63
3	8.16	5.37	5.69	-1.42	-3.23	-6.87
4	12.64	11.23	10.35	1.06	0.47	-2.57
5	17.31	17.91	15.24	4.38	4.85	1.95
6	21.51	23.70	19.89	8.24	8.91	6.02
7	26.21	27.31	24.46	11.32	11.25	9.30
8	29.90	27.81	26.95	13.77	11.45	10.31
9	26.38	25.51	24.73	11.14	10.25	8.55
10	20.96	21.55	19.00	6.61	7.95	3.99
11	12.91	15.47	13.21	1.44	4.04	-0.88
12	4.71	9.58	7.06	-1.99	-0.05	-5.88
13	0.57	4.30	2.69	-5.44	-3.69	-10.35
			Temple, Texas			
1	13.83	15.45	14.57	1.85	2.44	2.29
2	16.31	15.23	16.14	3.70	2.55	3.66
3	19.99	17.15	19.84	7.13	4.66	7.16
4	24.13	21.25	24.24	11.60	8.96	11.54
5	27.75	26.57	27.88	15.90	14.02	15.93
6	31.05	31.54	31.02	19.51	19.01	19.33
7	33.80	34.69	33.57	21.83	22.08	21.53
8	35.40	35.65	35.03	22.54	22.88	22.34
9	34.81	34.03	34.37	21.98	21.42	21.48
10	30.44	30.75	30.51	17.75	17.73	17.87
11	25.29	26.36	25.71	12.59	13.41	12.82
12	19.70	21.48	20.74	7.35	8.72	7.98
13	15.26	17.54	15.95	3.26	4.58	3.76

The same analysis is applied for the daily simulated temperatures. Table 6 shows the mean maximum temperature and the mean minimum temperature for 120 years of generated series and for the observed weather series. The means for the maximum and minimum temperature in the three weather stations are close to the observed data. The differences in averages can be mainly attributed to the detrending technique by harmonic analysis. Both SWGs reproduce significantly close weather patterns in the three weather stations. However, there is no conclusive evidence about how to rank the accurateness of these models.

Table 7. Annual average temperature and number of days of extreme events by, weather station

Weather Variable	Observed	Copula	Richardson
Conrad			
Maximum Temperature, ^{o}C	40.60	47.50	49.99
Minimum Temperature, ^{o}C	-27.20	-28.53	-45.58
Days ≥ 35 ^{o}C	1.10	5.08	6.38
Days ≤ 0^{o}C	183.84	205.13	183.93
Spokane			
Maximum Temperature, ^{o}C	42.20	42.02	51.42
Minimum Temperature, ^{o}C	-24.40	-24.56	-43.37
Days ≥ 35 ^{o}C	6.41	3.27	6.17
Days ≤ 0^{o}C	138.27	113.95	172.73
Temple			
Maximum Temperature, ^{o}C	43.30	55.30	48.32
Minimum Temperature, ^{o}C	-14.40	-9.68	-16.23
Days ≥ 35 ^{o}C	54.61	55.43	52.11
Days ≤ 0^{o}C	31.49	18.92	29.14

Table 7 summarizes the ability of the semiparametric copula-based SWG to reproduce the distribution of annual extreme temperatures in minimum temperature and maximum temperature series. The comparative analysis of the generated and the observed data in Table 7 confirms that semiparametric copula-based SWG reproduces much closer the patterns of extreme events in weather series. Both, semiparametric copula-based SWG and Richardson's SWG, show about the same number of days with extreme temperatures; however, the semiparametric copula-based SWG shows a better

replication in magnitude of the temperature extreme events of the observed data for the three weather stations.

8. Summary and Conclusions

Stochastic Weather Generators are an essential tool for the generation of weather series. However, conventional SWGs characterize climate variables with parametric probability distributions and linear dependence between, when in fact they are complex and their probability density function is close to non-normal distributions. This structure leads to the simulation of inaccurate climatic variables and to the sub replication of extreme weather events from observed data.

The proposed semiparametric copula-based SWG more accurately replicate some aspects of the observed weather patterns, such as the nonlinear dependence structure and the occurrence of extreme events between precipitation, maximum temperature, and minimum temperature. The idea of modeling climatic variables using copula methods relies on the behavior and structure of these variables. Copula methods provide the flexibility to model nonlinear dependence structures between random variables independent of the marginal distributions involved. The marginal distributions were modeled by the non-parametric kernel smoothing specification. The large volume of data provides reliability on non-parametric estimations and captures more accurately the probability in the tails of the distribution.

Instead of estimating parameters for 365 dates, one per day throughout the year, the alternative to treat the dimensionality problem in copula estimation was the selection of 12 dates, those with the highest historical monthly average anomaly. Thus, the weather series simulated by copula methods are the bordering conditions of the weather stochastic simulator, while the Brownian Bridge uses Monte Carlo methods to replicate the daily dynamic of weather variables evolving on a path forward through time. Although the numerous specifications were tested, the final specification was the one-parameter Gumbel family.

Finally the statistical tests performed on simulated weather, showed that semiparametric copula-based SWG can perform an acceptable replication of the observed weather patterns and an accurate reproduction of the extreme weather event patterns. Although in general there is no conclusive evidence about the superiority of the semiparametric copula-based SWG, one of its remarkable characteristics is the accurate representations on magnitudes of extreme weather events in temperatures.

References

Brandimonte, P. 2006. *Numerical Methods in Finance and Economics. A MATLAB-Based Introduction*. 2nd ed. New York: John Wiley and Sons.

Cherubini, U., E. Luciano and W. Vecchiato. 2004. *Copula Methods in Finance*. West Sussex: John Wiley & Sons.

Cohen, A. C. 1991. *Truncated and Censored Samples: Theory and Applications*. Statistics, Textbooks and Monographs. New York: John Wiley and Sons.

Embrechts, P., F. Lindskog, A. McNeil. 2001. *Modeling Dependence with Copulas and Applications to Risk Management. Handbook of Heavy Tailed Distributions in Finance*, ed. S. Rachev, San Diego: Elsevier.

Favre, A.-C., S. El Adlouni, L. Perreault, N. Thi´emonge and B. Bobee. 2004. "Multivariate Hydrological Frequency Analysis Using Copulas." *Water Resources Research* (40): W01101.

Genest, C. and A.-C Favre. 2007. "Everything you Always Wanted to Know About Copula Modeling but Were Afraid to Ask." *Journal of Hydrologic Engineering 1*2 (4): 347–368.

Genest, C., B. R´emillardb and D. Beaudoinc. 2009. "Goodness-of-Fit Tests for Copulas: A Review and a Power Study". *Insurance: Mathematics and Economics* (44): 199–213

Glasserman, P. 2010. *Monte Carlo Methods in Financial Engineering.* Applications of Mathematics. Stochastic Modeling and Applied Probability 53. New York: Springer.

Huynh, H. T., V. S. Lai and I. Soumaré. 2008. *Stochastic Simulation and Applications in Finance with MATLAB® Programs*. West Sussex: The Wiley Finance Series.

Joe, H. 1997. *Multivariate Models and Dependence Concepts*. London: Chapman &Hall.

L'ecuyer, P. 2004. *Random Number Generation.* Working paper, D´epartement d'Informatique et de Recherche Op´erationnelle, University of Montreal. http://www.iro.umontreal.ca/~lecuyer (last accessed: 5/11/2011)

Mood, A., F. A. Graybill and D.C. Boes. 1974. *Introduction to the Theory of Statistics.* 3rd ed. Singapore: McGraw Hill International Editions.

Nelsen, R. 2006. *An Introduction to Copulas*. Series in Statistics. New York: Springer Verlag.

Richardson, C. W. 1981. "Stochastic Simulation of Daily Precipitation, Temperature and Solar Radiation." *Water Resources Research* (17): 182-190.

Salvadori, G., C. Michele, N. T. Kottegoda and R. Rosso. 2007. *Extremes in Nature*. *An Approach Using Copulas*. New York: Springer.

Schölzel, C., and P. Friederichs. 2008. "Multivariate Non-normally Distributed Random Variables in Climate Research Introduction to the Copula Approach." *Nonlinear Processes in Geophysics* 15 (5): 761–772.

Turvey, C.G. 2005. "The Pricing of Degree-Day Weather Options." *Agricultural Finance Review* (65): 59-86.

Wand, M. C. and M. P. Jones. 1995. *Kernel Smoothing*. New York: Chapman & Hall.

Wilks, D. 1990. "Maximum Likelihood Estimation for the Gamma Distribution Using Data Containing Zeros." *Journal of Climate* (3): 1495-1501.

Wilks, D. 2011. *Statistical Methods in the Atmospherical Sciences*. International Geophysic Series; v. 100. 3rd ed. San Diego: Elsevier.

Wilks, D. and R. Wilby. 1999. "The Weather Generation Game: A Review of Stochastic Weather Models." *Progress in Physical Geography* 23(3): 329-357.

Appendix A

Brownian Bridge Treatment and Construction

The Brownian Bridge construction involves a process that begins with the generation of the final value $W(t_n)$ then filling in the intermediate values' amounts by simulating a Brownian Bridge from $0 = W(0)$ to $W(t_n)$. Next, $W(t_{\lceil \frac{n}{2} \rceil})$ is sampled, and values between times $t_{\lceil \frac{n}{2} \rceil}$ and t_n are filled in to simulate the Brownian Bridge from $W(t_{\lceil \frac{n}{2} \rceil})$ to $W(t_n)$ and so on.

In particular, the Brownian Bridge includes the generation of a tridimensional Brownian motion process $\boldsymbol{X}$ with one of the variates truncated to emulate the precipitation behavior, where the drift $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ must reflect such circumstances, $\boldsymbol{X} \sim \mathrm{BM}(\mu, \Sigma)$. Thus, the Brownian motion process $\boldsymbol{X}$ must be added to $W_i(t_n)$ at the first step of the independent one-dimensional construction to each one of the coordinates (Glasserman 2010).

$$X(t_{i+1}) = X(t_i) + \mu(t_{i+1} - t_i) + \sqrt{t_{i+1} - t_i}\, B\, Z_i, \quad i = 0, \dots, n-1 \tag{33}$$

Where $BB^T = \Sigma$ and Z is an independent $N(0,1)$. Thus, the drift $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are estimated by fitting the historical weather variables (maximum temperature, minimum temperature and precipitation) using the maximum likelihood estimation method. The parameters estimation is from a population with single truncated sample, normal p.d.f. and the truncation point at zero. Cohen (1991) shows the analytical solutions for $\bar{x}$ and σ, derived using maximum likelihood estimation. When restriction occurs only in one of the variates of the multivariate distribution, such as in the case of precipitation; say $X = (x_1, x_2, x_3)$ is the trivariate distribution with the following p.d.f equation.

$$f(\boldsymbol{X}) = 2\pi^{-3/2} \left| \Sigma^{ij} \right|^{-1/2} \exp^{(-1/2)(x-\mu)' \Sigma^{-1} (x-\mu)} \tag{34}$$

For left truncated samples, the analytical solutions for the estimates μ_1 and $\sigma_1$have closed form solutions. As Cohen (1991) shows solutions for x_1 (truncated variate) can be calculated only from marginal data of x_1, without considering any other variate.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} \tag{35}$$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n} \tag{36}$$

$$\theta(\xi) = \frac{Q(\xi)}{Q(\xi) - \xi'} \tag{37}$$

$$Q(\xi) = \frac{\phi(\xi)}{1 - \Phi(\xi)} \tag{38}$$

$$\mu = \bar{x} - \theta(\xi)(\bar{x} - T) \tag{39}$$

$$\sigma^2 = s^2 + \theta(\xi)(\bar{x} - T) \tag{40}$$

Where n is the number of truncated rain-rate samples, $\theta(\xi)$ is the auxiliary estimation function, and $\phi(\xi)$ and $\Phi(\xi)$ are probability distribution function and cumulative distribution function of the standard normal distribution, respectively. The parameters estimation for the remaining two variates and their correlation coefficients is the following. Consult Cohen (1991) for more details.

$$\hat{\mu}_j = \overline{x_j} - \overline{r_j}\,\frac{s_j}{s_1}\,(\overline{x_1} - \hat{\mu}_1), \tag{41}$$

$$\hat{\sigma}_j = \overline{s_j}\sqrt{\frac{1-\hat{\lambda}\,\left(1-\overline{r_{ij}^2}\right)}{1-\hat{\lambda}}} \tag{42}$$

$$\hat{\rho}_{ij} = \frac{\overline{r_{ij}} - \hat{\lambda}\left(\overline{r_{ij}} - \overline{r_{1i}}\,\overline{r_{1j}}\right)}{\sqrt{\left[1-\hat{\lambda}\left(1-\overline{r_{1i}^2}\right)\right]\left[1-\hat{\lambda}\,\left(1-\overline{r_{1j}^2}\right)\right]}} \tag{43}$$

For $i = 1,2,\dots,p-1,\ \ j = 1,2,\dots,p, i < j.$

$$\hat{\lambda} = 1 - \frac{\overline{s_1^2}}{\overline{\sigma_1^2}} \tag{44}$$

Since by definition $r_{ij} = 1$, the last equation for $i = 1$ becomes

$$\hat{\rho}_{1j} = \frac{\overline{r_{ij}}}{\sqrt{\left[1-\hat{\lambda}\left(1-\overline{r_{1j}^2}\right)\right]}} \tag{45}$$